HYBRID DENOISING-SCREENING FOR HIGH-DIMENSIONAL CONTAMINATED DATA

Liming Wang[1], Peng Lai[1], Chen Xu[2,3] and Xingxiang Li[3]

[1] *Nanjing University of Information Science & Technology,*

[2] *Peng Cheng Laboratory and* [3] *Xi'an Jiaotong University*

Supplementary Material

This document contains supplementary material to the main text of the article. Section S1 discusses the comparison of HDS with least trimmed squares (LTS) and mean-shifted-model-based outlier detection (MSMOD). Section S2 provides a detailed algorithm for the practical implementation of HDS. In Section S3, an additional simulation study is presented, including Example 2 and 3. Specifically, Example 2 conducts a detailed sensitivity analysis and Example 3 further evaluates the effectiveness and robustness of HDS on some challenging examples. Lastly, Section S4 presents the proofs of all theoretical results in the main text.

S1 The comparison of HDS with LTS and MSMOD

In this section, we compare HDS with LTS and MSMOD. Some notations are first introduced for ease of expression. Given a series $a_1, a_2, \ldots, a_n$, let $R(a_i) = \sum_{k=1}^{n} I(a_k \leq a_i) - \sum_{k \neq i}^{n} I(a_k = a_i, i < k)$ be the rank of a_i. Define the residual $r_i(\boldsymbol{\beta}) = y_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}$. For a fixed value of $\boldsymbol{\beta}$, let $r_{(i)}(\boldsymbol{\beta}) = r_k(\boldsymbol{\beta})$ satisfying $R(|r_k(\boldsymbol{\beta})|) = i$. As stated in the following Proposition 1, the proposed HDS in (2.3) is equivalent to the restricted versions of LTS and MSMOD.

Proposition 1. *For a fixed L, any optimal $\boldsymbol{\beta}$ solution of problem (2.3) is the optimal solution of the following LTS and MSMOD problems.*

$$\min_{\boldsymbol{\beta}} \sum_{i=1}^{L} [r_{(i)}(\boldsymbol{\beta})]^2 \text{ subject to } \|\boldsymbol{\beta}\|_0 \leq K, \tag{S1.1}$$

$$\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} (y_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta} - \xi_i)^2 \text{ subject to } \|\boldsymbol{\xi}\|_0 \leq n - L \text{ and } \|\boldsymbol{\beta}\|_0 \leq K, \tag{S1.2}$$

where $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_n)^{\mathrm{T}}$. Conversely, any optimal solutions of problem (S1.1) or (S1.2) is the optimal $\boldsymbol{\beta}$ solution of (2.3).

In (S1.1) and (S1.2), the L observations with the smallest residuals are selected for parameter estimation; this echoes the motivation of HDS. However, different expressions of the objective function may lead to diverse algorithms and statistical inferences. For example, to solve LTS in (S1.1), commonly-used methods include PROGRESS Algorithm by Rousseeuw and Leroy (1987) and Feasible Set Algorithm by Hawkins (1994). In these methods, many candidate subsets of $\{1, \ldots, n\}$ are evaluated, and the candidate subset with the smallest objective function value provides an approximate LTS estimate. In general, many candidate subsets are accounted for to obtain a promising solution; this leads the estimation to be very time-consuming, especially when n or p is large. Regarding L_0-based MSMOD in (S1.2), it can be optimized by iterative shrinkage thresholding algorithm (similar to equation (7.1) in She and Owen (2011)) or AOS strategy like HDS. In this paper, we choose (2.3) as our objective because we can design an efficient algorithm and theoretical inferences with the help of auxiliary weight parameters $\{v_i\}_{i=1}^n$.

To our knowledge, (S1.1) or (S1.2) has not been proposed for robust regression or outlier detection for high-dimensional regression, not to mention the joint feature screening. In fact, the penalized LTS and MSMOD are more frequently considered to deal with high-dimensional data containing potential outliers in recent years. To be specific,

$$\text{penalized LTS:} \qquad \min_{\boldsymbol{\beta}} \sum_{i=1}^{L} [r_{(i)}(\boldsymbol{\beta})]^2 + P_\lambda(\boldsymbol{\beta}), \tag{S1.3}$$

$$\text{penalized MSMOD:} \qquad \min_{\boldsymbol{\beta}} \sum_{i=1}^{n} (y_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta} - \xi_i)^2 + P_\lambda(\boldsymbol{\beta}) + P_\lambda(\boldsymbol{\xi}), \tag{S1.4}$$

where $P_\lambda(\cdot)$ is a specific penalty function. The commonly-used penalty functions for penalized LTS (S1.3) include l_1 penalty proposed by Alfons et al. (2013) and elastic-net penalty proposed by Kurnaz et al. (2018). For the penalized MSMOD in (S1.4), She and Owen (2011) suggests using nonconvex penalty functions, such as Hard penalty and SCAD penalty. In theory, (S1.3) or (S1.4) also can be considered for feature screening. However, the number of features retained by the penalized LTS or MSMOD can not determined in advance. This is due to the final model size in (S1.3) or (S1.4) is determined by the value of λ, which is a continuous tuning parameter generally requiring much additional computational cost to train in practice. Since the retained model size, denoted by K, just needs to be a moderate integer greater than the true model size in feature screening, the selection of K should avoid excessive training costs. In view of the high efficiency of dual sample-feature L_0 fitting, we adopt (2.3) as the objective function in our HDS.

S2 The implementation details of IHT and the final algorithm for HDS

In IHT, u^{-1} can be seen as a step size determining the distance from $\boldsymbol{\gamma}^{(h)}$ to $\boldsymbol{\gamma}^{(h+1)}$. It can be easily proved that $\boldsymbol{\mathcal{L}}(\boldsymbol{v}^{(t+1)}, \boldsymbol{\eta}^{(h)})$ is strictly non-increasing as long as u takes the value not less than $\lambda_{\max}(\mathbf{A}(\boldsymbol{v}^{(t+1)}))$, where $\mathbf{A}(\boldsymbol{v}) = n^{-1}\sum_{i=1}^{n} 2v_i\boldsymbol{x}_i\boldsymbol{x}_i^{\mathrm{T}}$ and $\lambda_{\max}(\cdot)$ denotes the largest eigenvalue of a matrix. Despite the theoretical assurances, an overly small step size may lead IHT to fail to converge to a good solution. In practice, a larger u^{-1} often contributes to boosting the iterations, but an overly large step size makes IHT difficult to converge. Thus, it is suggested to adjust u^{-1} adaptively at each step. A commonly-used strategy is to initialize u^{-1} with a large value and then adaptively decrease its size by multiplying a $\tau \in (0,1)$ until $\mathcal{L}(\boldsymbol{v}^{(t+1)}, \boldsymbol{\gamma}^{(h+1)}) \leq \mathcal{L}(\boldsymbol{v}^{(t+1)}, \boldsymbol{\gamma}^{(h)})$ is satisfied. This seems to be an effective way to balance algorithm convergence and iteration rates.

Next, we provide some discussions regarding the initial value $\boldsymbol{\gamma}^{(0)}$ in (2.7). A natural strategy is to choose $\boldsymbol{\gamma}^{(0)} = \boldsymbol{\beta}_{\mathrm{IHT}}^{(t)}$, based on which u can be determined properly to strictly ensure that $\mathcal{L}(\boldsymbol{v}^{(t+1)}, \boldsymbol{\beta}_{\mathrm{IHT}}^{(t+1)}) \leq \mathcal{L}(\boldsymbol{v}^{(t+1)}, \boldsymbol{\beta}_{\mathrm{IHT}}^{(t)})$. However, some numerical experiences show that IHT with $\boldsymbol{\gamma}^{(0)} = \boldsymbol{\beta}_{\mathrm{IHT}}^{(t)}$ often fails to significantly improve $\mathcal{L}(\boldsymbol{v}, \boldsymbol{\beta})$ in the first few steps (i.e. t is small). The performance of AOS is much restricted by the initial $\boldsymbol{\gamma}^{(0)}$ and the above strictly decreasing property. To be worse, the first $\boldsymbol{\beta}_{\mathrm{IHT}}^{(t)}$ and $\boldsymbol{v}^{(t+1)}$ are often seriously affected by noises; this leads to they are not of high quality. These factors result in that $\boldsymbol{\gamma}^{(h)}$ often converges to a bad local solution near to $\boldsymbol{\beta}_{\mathrm{IHT}}^{(t)}$, and the leading $\boldsymbol{\beta}_{\mathrm{IHT}}^{(t+1)}$ fails to find sufficient relevant features. To overcome the above difficulty, we suggest adopting some more efficient initial for the iterations in IHT, such as lasso-type initial (Tibshirani, 1996). Although the decreasing property can not be guaranteed strictly, the potential significant decline of $\mathcal{L}$ contributes to rapidly searching all relevant features. When $\boldsymbol{v}^{(t)}$ stabilizes after some steps, it is believed that most clean observations are retained. Then, we can set $\boldsymbol{\gamma}^{(0)} = \boldsymbol{\beta}_{\mathrm{IHT}}^{(t)}$ to obtain a more accurate $\hat{\boldsymbol{\beta}}$ with a refined $\hat{\mathcal{M}}$. Based on the above discussion, the initial value $\boldsymbol{\gamma}^{(0)}$ in (2.7) can be set by

$$\boldsymbol{\gamma}^{(0)} = \begin{cases} \gamma_{\mathrm{lasso}}, & \text{if } t < T_1 \text{ and } \boldsymbol{v}^{(t+1)} \neq \boldsymbol{v}^{(t)}, \\ \boldsymbol{\beta}_{\mathrm{IHT}}^{(t)}, & \text{if } t \geq T_1 \text{ or } \boldsymbol{v}^{(t+1)} = \boldsymbol{v}^{(t)}, \end{cases} \tag{S2.5}$$

where $\boldsymbol{\gamma}_{\text{lasso}} = \operatorname{argmin}_{\boldsymbol{\gamma}}\ \mathcal{L}(\boldsymbol{v}^{(t+1)}, \boldsymbol{\gamma}) + \lambda\|\boldsymbol{\gamma}\|_1$ with a λ leading to the sparsity $\|\boldsymbol{\gamma}_{\text{lasso}}\|_0$ closest to $L-1$. Based on our numerical experiments, $T_1 = 20$ demonstrates effective performance across most simulations in this study. Notably, compared to AOS initialized solely with $\boldsymbol{\gamma}^{(0)} = \boldsymbol{\gamma}_{\text{lasso}}$, our proposed initialization scheme ($\boldsymbol{\gamma}^{(0)}$ in (S2.5)) generally exhibits faster convergence and achieves higher screening accuracy.

Last, we show the whole computing procedure of HDS in the following Algorithm 1.

Algorithm 1 The computing process for HDS

```
input: Data $\{y_i, \boldsymbol{x}_i\}_{i=1}^n$.
Initialize $t = 0$, $\tilde{\boldsymbol{v}} = \boldsymbol{1}$, and set $\tilde{\boldsymbol{\beta}}_{\text{IHT}}$ as the IHT solution for the constrained minimization
    $\operatorname{argmin}_{\boldsymbol{\beta}} \sum_{i=1}^n (y_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta})^2$ subject to $\|\boldsymbol{\beta}\|_0 \le K$, where $\boldsymbol{\gamma}^{(0)}$ denotes the lasso solution with sparsity
    level $n-1$. Subsequently, specify two numbers of iterations $T_1$ and $T_2$ $(1 < T_1 < T_2)$ and convergence
    threshold $\epsilon_\gamma$ for the IHT procedure.
while $t \le T_2$ do
    Denoising step: update $\hat{\boldsymbol{v}}$ by
        $\hat{v}_i = \begin{cases} 1, & R(|y_i - \boldsymbol{x}_i^{\mathrm{T}}\tilde{\boldsymbol{\beta}}_{\text{IHT}}|) \le L \\ 0, & R(|y_i - \boldsymbol{x}_i^{\mathrm{T}}\tilde{\boldsymbol{\beta}}_{\text{IHT}}|) > L \end{cases}, \quad i = 1, \ldots, n.$
    If $t < T_1$ & $\hat{\boldsymbol{v}} = \tilde{\boldsymbol{v}}$, set $t = T_1 - 1$.
    If $t \ge T_1$ & $\hat{\boldsymbol{v}} = \tilde{\boldsymbol{v}}$, break.
    Update $\boldsymbol{\gamma}^{(0)}$ by (2.10), where $\boldsymbol{v}^{(t+1)} = \hat{\boldsymbol{v}}$, $\boldsymbol{v}^{(t)} = \tilde{\boldsymbol{v}}$, and $\boldsymbol{\beta}_{\text{IHT}}^{(t)} = \tilde{\boldsymbol{\beta}}_{\text{IHT}}$.
    Iterate $\boldsymbol{\gamma}^{(h+1)}$ by (2.9) until $\|\boldsymbol{\gamma}^{(h+1)} - \boldsymbol{\gamma}^{(h)}\|_2 \le \epsilon_\gamma$.
    Screening step: update $\tilde{\boldsymbol{\beta}}_{\text{IHT}} = \boldsymbol{\gamma}^{(h+1)}$.
    Set $\tilde{\boldsymbol{v}} = \hat{\boldsymbol{v}}$ and $t = t + 1$.
end while
output: $\hat{\boldsymbol{v}}$, $\hat{\boldsymbol{\beta}} = \tilde{\boldsymbol{\beta}}_{\text{IHT}}$, and $\hat{\mathcal{M}} = \{j : \hat{\beta}_j \neq 0, j = 1, \ldots, p\}$.
```

S3 Additional Examples in Simulations

Example 2. Sensitivity Analysis.

In this example, we test the sensitivity of HDS on some factors, such as the number of retained features K, the dimension of features p, the variance of random error in linear model σ^2, and the correlation between features $\rho_{j,k}$. For ease of study, we set all $\rho_{j,k} = \rho$ for $j \neq k$. In addition, we further consider the influence of an important factor, that is, the distance between noises and the clean linear model.

We first assess sensitivity with respect to the choices of K, p, σ^2, and ρ. To facilitate the analysis, we re-conduct a few simulations in Example (1a) with specific setups of noisy distribution $f_0(y_i|\boldsymbol{x}_i) = 0.5 f_{U(-20,-15)}(y_i) + 0.5 f_{U(15,20)}(y_i)$ and NCR $= 20\%$. We give some specific setups as follows.

(2a) **Factor K:** we fix $(n, p, \sigma^2, \rho) = (150, 2000, 1, 0.5)$, and consider $K = 10, 20, \ldots, 60$.

(2b) **Factor p:** we fix $(n, K, \sigma^2, \rho) = (150, 20, 1, 0.5)$, and consider $p = 1000, 2000, \ldots, 8000$.

(2c) **Factor σ^2:** we fix $(n, p, K, \rho) = (200, 3000, 20, 0.3)$, and consider $\sigma^2 = 1, 2, \ldots, 7$.

(2d) **Factor ρ:** we fix $(n, p, K, \sigma^2) = (300, 2000, 20, 1)$, and consider $\rho = 0, 0.1, \ldots, 0.7$.

The rest is in accord with Example (1a). As before, the evaluation criterion is SSR based on $T = 100$ repetitions. The screening performance of HDS on these factors is shown in Figure 1 of supplementary material.

In general, HDS is insensitive to wide ranges of K, p, σ^2, and ρ and shows significant superiority to other methods. Although large values of the above factors may reduce some accuracy of HDS, its accuracy still maintains at a high level. Some detailed analyses are

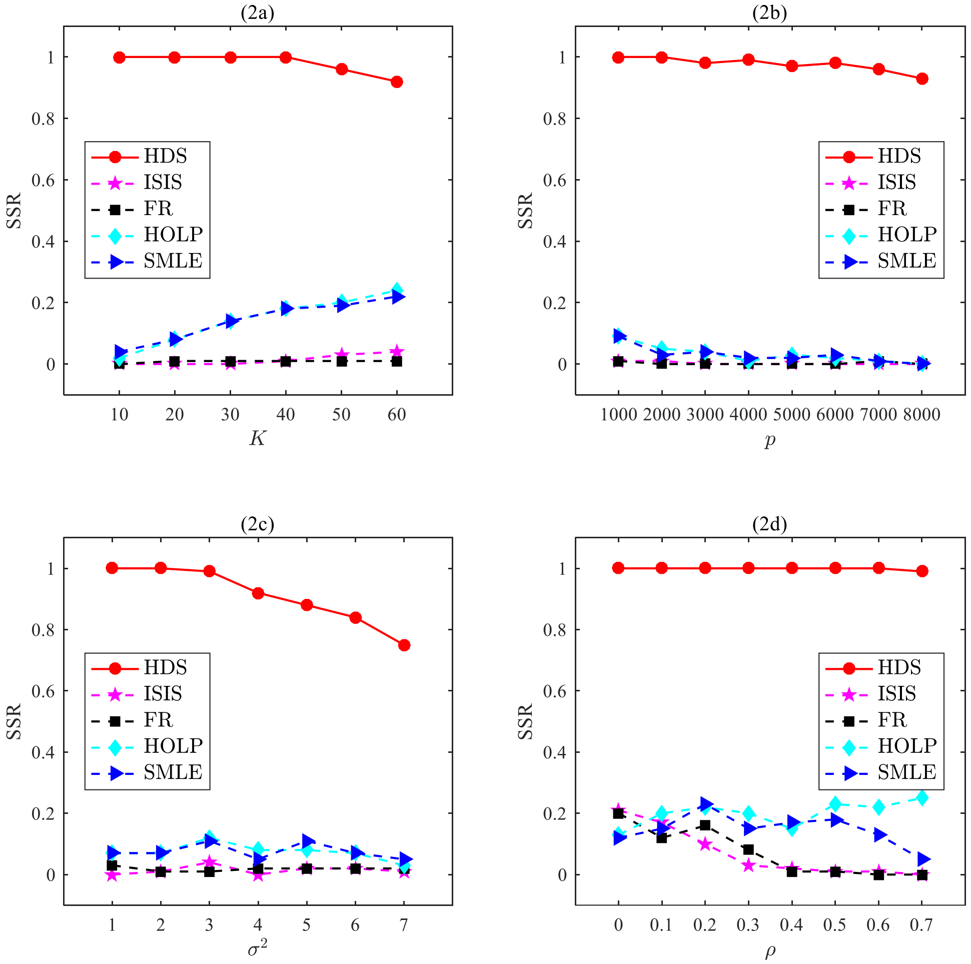

Figure 1: Overall SSRs of the selected screeners over the following setups: (2a) increasing screening size K; (2b) increasing dimensionality p; (2c) increasing variance σ^2; (2d) increasing correlation ρ.

given as follows. When the sample size L is small, a larger K may induce overfitting and hinder the search for good local solutions in HDS, resulting in some accuracy loss. Thus, we recommend against selecting an excessively large K when L is moderate in HDS. The variance σ^2 reflects the signal strength of linear model. As σ^2 increases, the signal of relevant features becomes weaker; this results in some difficulty for HDS. Regarding the correlation, HDS is very robust to the given candidates of ρ.

Then, we measure the influence of the distances between noises and the clean linear model. Assume all $\boldsymbol{x}_i$ and clean y_i follow the same distribution in Example 1, where $\mathcal{M} = \{1, \ldots, 5\}$, $(n, p, K) = (200, 2000, 20)$, and NCR $= 20\%$. Next, assume noisy $y_i = \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta}^* + \epsilon_i$, where the distribution of ϵ_i differs from $N(0, \sigma^2)$. We consider two setups of ϵ_i as follows.

(2e) $\epsilon_i \sim \phi N(0, 64) + \phi N(0, 16) + (1 - 2\phi) N(0, 0.25)$ with $\phi = 0.1, 0.2, \ldots, 0.5$.

(2f) $\epsilon_i \sim 0.5U(-d, -d + 5) + 0.5U(d - 5, d)$ with $d = 5, 10, \ldots, 40$.

In (2e), the noises can be divided into three types with the small, moderate, and large magnitude of influences, and ϕ determines the proportion of three types of noises. In (2f), the distances between noises and the clean linear model are determined by the starting parameter d and the width 5. When ϕ and d are small, clean and noisy responses overlap strongly. As they increase, the differences between clean and noisy responses will be more significant. The screening performance of HDS on ϕ and d are shown in Figure 2 of supplementary material.

It can be seen all classic screening methods are sensitive to the distances between noises and clean linear models. As ϕ and d increase, their SSRs encounter a drastic drop. HDS shows great robustness to the distance. In (2f), when d is small, although the overlap between clean and noisy y_i causes some accuracy loss on HDS, its effectiveness still can be guaranteed. As d increases, the superiority of HDS becomes more significant.

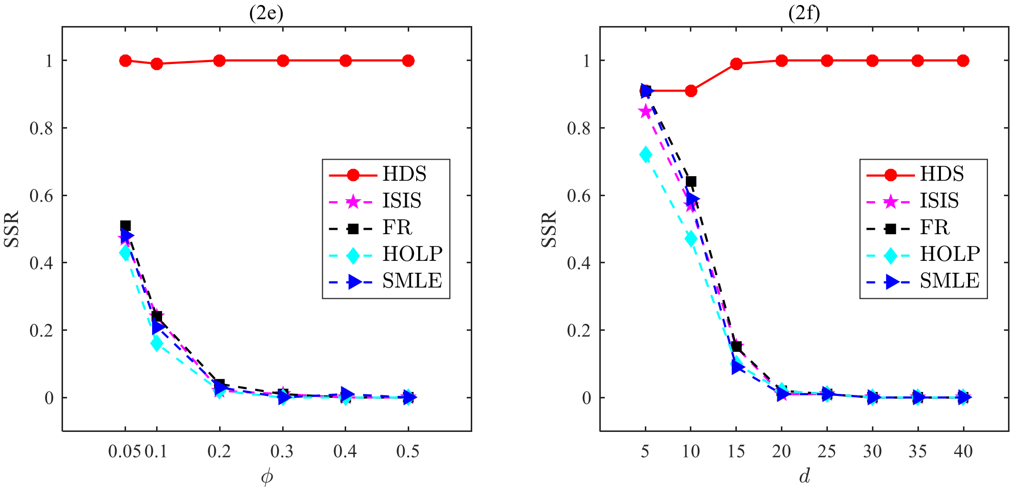

Figure 2: Overall SSRs of the selected screeners over setups (2e) and (2f).

Example 3. Some challenging setups for HDS

In Example 1, HDS shows its anti-noise and joint-screening performance to three different types of contaminated data. In this example, we consider some more challenging setups, such as more complex correlations between features, higher CNRs, a larger number of relevant features, and new setups of noises. The specific setups are given as follows.

(3a) Each $\boldsymbol{x}_i \sim N(\mathbf{0}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma} = [\rho_{j,k}]_{j,k=1,\ldots,p}$, where $\rho_{j,j} = 1$ for $j = 1, \ldots, p$, off-diagonal elements $\rho_{j,k} = 0.5, j \neq k$ except $\rho_{6,j} = \rho_{j,6} = 0.25$ for $j \neq 5$. For $i \in \mathcal{I}_1$,

$$y_i = 3x_{i1} + 3x_{i2} + 3x_{i3} + 3x_{i4} + 3x_{i5} + 1.5x_{i6} + \epsilon_i$$

Then, we consider a two-sided distributed noisy y_i. That is,

$$f_0(y_i|\boldsymbol{x}_i) = 0.5 f_{U(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}^* - 40, \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}^* - 20)}(y_i) + 0.5 f_{U(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}^* + 20, \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}^* + 40)}(y_i).$$

(3b) Each $\boldsymbol{x}_i \sim N(\mathbf{0}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma} = [\rho_{j,k}]_{j,k=1,\ldots,p}$, where $\rho_{j,j} = 1, j = 1, \ldots, p$ for $j = 1, \ldots, p$, off-diagonal elements $\rho_{j,k} = 0.5$ except $\rho_{6,j} = \rho_{j,6} = \sqrt{0.5}$ for $j \neq 5$. For $i \in \mathcal{I}_1$,

$$y_i = 5x_{i1} + 5x_{i2} + 5x_{i3} + 5x_{i4} - 20\sqrt{0.5}x_{i5} + \epsilon_i.$$

Then, we consider one-sided distributed noisy y_i. That is,

$$f_0(y_i|\boldsymbol{x}_i) = f_{U(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}^*-40, \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}^*-20)}(y_i).$$

(3c) Each $\boldsymbol{x}_i$ follows the same setup as example (1a) with $\mathcal{M} = \{1, 2, \ldots, 8\}$. Then, we consider a new setup where clean and noisy responses have overlaps. To be specific, the distribution of y_i is set by

$$f_0(y_i|\boldsymbol{x}_i) = f_{U(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}^*-80, \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}^*)}(y_i).$$

Setups (3a)-(3b) consider the impact of correlations between features on screening. In (3a), most irrelevant features are highly correlated with relevant ones, thus they have a false marginal correlation with uncontaminated response. However, the relevant feature x_6 is weakly correlated with uncontaminated response because of its corresponding small coefficient and lower correlations with other features. In (3b), we consider a special feature x_5, which is jointly dependent but marginally independent on uncontaminated response. These are two difficult setups for feature screening and are commonly seen in many literature, such as Fan and Lv (2008) and Cho and Fryzlewicz (2012). Regarding the distribution of noisy response, we consider the two-sided and one-sided extreme values to the response in (3a)-(3b). In (3c), the overlap between clean and noisy y_i may lead to some challenges in denoising and variable selection. Besides, we increase the screening difficulty by enlarging the NCR up to 90% in (3a)-(3b) and the size of $\mathcal{M}$ up to 8 in (3c).

For comparison purposes, we consider a wide range of methods for feature screening in (3a) and (3b) and variable selection in (3c). To show the superiority of HDS in terms of robustness, we conduct two classic robust screening methods RRCS (Li et al., 2012) and FKF (Mai and Zou, 2015) with 6 slices. To make them better suitable for joint screening, we further consider their iterative versions, named iRRCS and iFKF, where the residual of y_i fitted by the current features are used in the next iteration of screening. Three features are retained in each iteration of iRRCS and iFKF until the total model size is not lower than K. To show the effectiveness of IHT, we consider two other algorithms, primal-dual active set (PDAS) discussed in Wen et al. (2020) and adaptive best-subset selection (ABESS) proposed by Zhu et al. (2020), to update $\boldsymbol{\beta}^{(t+1)}$ in (2.6). The leading screening procedures are denoted by HDS-PDAS and HDS-ABESS, respectively. To demonstrate the necessity of HDS as a preliminary step before in-depth variable selection, we perform EBIC-based post-screening on the data preserved by HDS. The details can be referred to the function "`smle_select`" in R package `SMLE` (Zang et al., 2022). As the competitors of

HDS in variable selection, we consider sparse least trimmed squares (SLTS) proposed by Alfons et al. (2013) and penalized mean-shift-model-based outlier detection (PMSMOD) discussed in Section 7 of She and Owen (2011).

We generate $n = 300$ independent copies from $(y, \boldsymbol{x})$ with $p = 2000$. During screening, we set $L = \lfloor 0.9\pi_1 n \rfloor$ for each setup. Regarding the evaluation criterion in screening and selection, in addition to SSR, positive selection rate (PSR) and false discovery rate (FDR) based on $T = 100$ repetitions are used to assess the screening or selection accuracy. These indices are computed by

$$\text{PSR} = \frac{1}{T}\sum_{t=1}^{T}\frac{\mathbb{N}(\mathcal{M} \cap \hat{\mathcal{M}}(t))}{\mathbb{N}(\mathcal{M})},\ \text{FDR} = \frac{1}{T}\sum_{t=1}^{T}\frac{\mathbb{N}(\hat{\mathcal{M}}(t) - \mathcal{M})}{\mathbb{N}(\hat{\mathcal{M}}(t))},$$

where $\hat{\mathcal{M}}(t)$ is the index set of retained features determined by screening or selection methods. In particular, for the selection procedure in setup (3c), the correct selection rate (CSR), final model size (FMS), and computational time (Time; in seconds) are further exhibited, where CSR and FMS are computed by

$$\text{CSR} = \frac{1}{T}\sum_{t=1}^{T} I(\mathcal{M} = \hat{\mathcal{M}}(t)),\ \text{FMS} = \frac{1}{T}\sum_{t=1}^{T}\mathbb{N}(\hat{\mathcal{M}}(t)).$$

The simulation results of screening on setups (3a) and (3b) are summarized in Table 1, and the results of selection on setups (3c) are summarized in Table 2.

Table 1 shows that HDS outperforms other screening methods in setups (3a) and (3b),

Table 1: The simulation result of setups (3a) and (3b), where $K = \lfloor n^{1/4} \log(n) \rfloor$.

Setup	Methods	NCR = 0%			NCR = 60%			NCR = 90%		
		SSR	PSR	FDR	SSR	PSR	FDR	SSR	PSR	FDR
(3a)	HDS-IHT	1.00	1.00	0.74	0.96	0.98	0.74	0.90	0.95	0.75
	HDS-PDAS	1.00	1.00	0.74	0.24	0.51	0.87	0.10	0.30	0.92
	HDS-ABESS	1.00	1.00	0.74	0.37	0.56	0.85	0.12	0.33	0.91
	RRCS	0.00	0.81	0.79	0.00	0.17	0.96	0.00	0.10	0.98
	iRRCS	1.00	1.00	0.74	0.00	0.11	0.97	0.00	0.08	0.98
	FKF	0.00	0.36	0.91	0.00	0.18	0.95	0.00	0.16	0.96
	iFKF	1.00	1.00	0.74	0.00	0.07	0.98	0.00	0.04	0.99
(3b)	HDS-IHT	1.00	1.00	0.78	0.99	1.00	0.78	0.97	0.98	0.79
	HDS-PDAS	1.00	1.00	0.78	0.84	0.95	0.79	0.49	0.73	0.84
	HDS-ABESS	1.00	1.00	0.78	0.74	0.87	0.81	0.37	0.59	0.87
	RRCS	0.36	0.84	0.82	0.04	0.52	0.89	0.01	0.39	0.91
	iRRCS	1.00	1.00	0.78	0.01	0.47	0.90	0.01	0.45	0.90
	FKF	0.08	0.66	0.86	0.00	0.37	0.92	0.00	0.39	0.92
	iFKF	1.00	1.00	0.78	0.03	0.54	0.88	0.04	0.49	0.89

no matter with NCR=0%, 60%, and 90%; this is reflected by its higher SSRs and PSRs. HDS-PDAS and HDS-ABESS obtain satisfactory accuracy in the setup without noises (NCR = 0%). However, they lose some effectiveness when the number of noises increases; this indicates that they converge to some bad local optimal solution for $\boldsymbol{\beta}^{(t+1)}$. In our HDS-IHT, the adaptive step size and the lasso initials in the first T_1 iterations contribute to avoiding bad local optimal solutions; this leads HDS-IHT to be more accurate in screening. Regarding RRCS, FKF, and their iterative versions, their performances are also sensitive to scenarios with many noises. Moreover, HDS performs promisingly in both setups with two-sided and one-sided noises.

From FMS in Table 2, it can be found that the model size determined by HDS together with EBIC is further reduced and close to $\mathbb{N}(\mathcal{M})$. Together with CSR, it is shown that HDS can determine the true model $\mathcal{M}$ in the majority of simulations. High SSR indicates that only in a very small number of simulations, some relevant features are lost. The results of PSR and FDR further support the effectiveness of the proposed selection procedure. Generally, HDS is robust to setup (3c) with overlap between clean and noisy responses. Regarding SLTS, it tends to output a large model, in which all relevant features are usually can be retained but many irrelevant ones are also selected; this is supported by its high SSR and low CSR. Regarding PMSMOD, its CSR decreases more drastically than that of HDS as NCR increases. To be worse, when NCR=60%, it is hard to retain all relevant features in the final model by PMSMOD. In addition, PMSMOD tends to retain the most features among the three involved methods. Regarding the computational time of all methods, as L decreases, the computational time of HDS is also reduced. When NCR is not lower than 10%, the computational cost of SLTS encounters a drastic increase; this leads it to be computationally expensive. It is worth noting that the computational time of HDS is not least in three methods; this is due to the high computational cost of HDS used for step size adaption and lasso initial; this contributes to improving selection accuracy of HDS.

Table 2: The simulation result of setup (3c), where $K = 2\mathbb{N}(\mathcal{M})$ in HDS.

NCR	Method	SSR	CSR	PSR	FDR	FMS	Time
0%	HDS	1.00	0.94	1.00	0.01	8.06	5.89
	SLTS	0.97	0.08	1.00	0.47	23.46	0.72
	PMSMOD	1.00	0.94	1.00	0.06	18.79	0.80
10%	HDS	0.98	0.89	0.99	0.02	8.03	5.20
	SLTS	0.99	0.01	1.00	0.54	26.25	18.16
	PMSMOD	1.00	0.38	1.00	0.59	97.83	0.81
30%	HDS	0.99	0.83	1.00	0.02	8.19	4.07
	SLTS	1.00	0.02	1.00	0.67	34.95	22.15
	PMSMOD	0.96	0.00	0.99	0.93	117.15	0.90
60%	HDS	0.76	0.64	0.86	0.10	7.52	3.04
	SLTS	1.00	0.00	1.00	0.71	33.62	23.05
	PMSMOD	0.04	0.00	0.45	0.96	93.35	1.09

S4 Proofs

S4.1 Proof of Proposition 1

Denote $(\hat{\boldsymbol{v}}_{\text{HDS}}, \hat{\boldsymbol{\beta}}_{\text{HDS}})$ the optimal solution of problem (2.3) and $\hat{\boldsymbol{\beta}}_{\text{LTS}}$ the optimal solution of problem (S1.1). Based on the definition of $\hat{\boldsymbol{v}}_{\text{HDS}}$, we have

$$\hat{v}_{i,\text{HDS}} = \begin{cases} 1, & R(|y_i - \boldsymbol{x}_i^{\text{T}}\hat{\boldsymbol{\beta}}_{\text{HDS}}|) \leq L \\ 0, & R(|y_i - \boldsymbol{x}_i^{\text{T}}\hat{\boldsymbol{\beta}}_{\text{HDS}}|) > L \end{cases}, \quad i = 1, \ldots, n.$$

Therefore, we have

$$\mathcal{L}(\hat{\boldsymbol{v}}_{\text{HDS}}, \hat{\boldsymbol{\beta}}_{\text{HDS}}) = \sum_{i=1}^{L} \left[r_{(i)}(\hat{\boldsymbol{\beta}}_{\text{HDS}}) \right]^2,$$

where $\|\hat{\boldsymbol{\beta}}_{\text{HDS}}\|_0 \leq K$. Then, with the optimality of $\hat{\boldsymbol{\beta}}_{\text{LTS}}$, it follows that

$$\sum_{i=1}^{L}\left[r_{(i)}(\hat{\boldsymbol{\beta}}_{\text{LTS}})\right]^2 \leq \sum_{i=1}^{L}\left[r_{(i)}(\hat{\boldsymbol{\beta}}_{\text{HDS}})\right]^2 = \mathcal{L}(\hat{\boldsymbol{v}}_{\text{HDS}}, \hat{\boldsymbol{\beta}}_{\text{HDS}}). \tag{S4.6}$$

On the other hand, by the definition of $r_{(i)}(\boldsymbol{\beta})$, we have

$$\sum_{i=1}^{L}\left[r_{(i)}(\hat{\boldsymbol{\beta}}_{\text{LTS}})\right]^2 = \sum_{i=1}^{n}\hat{v}_{i,\text{LTS}}(y_i - \boldsymbol{x}_i^{\text{T}}\hat{\boldsymbol{\beta}}_{\text{LTS}})^2$$

with

$$\hat{v}_{i,\text{LTS}} \quad = \quad \begin{cases} 1, & R(|y_i - \boldsymbol{x}_i^{\text{T}}\hat{\boldsymbol{\beta}}_{\text{LTS}}|) \leq L \\ 0, & R(|y_i - \boldsymbol{x}_i^{\text{T}}\hat{\boldsymbol{\beta}}_{\text{LTS}}|) > L \end{cases}, \quad i = 1, \ldots, n.$$

By letting $\hat{\boldsymbol{v}}_{\text{LTS}} = (\hat{v}_{1,\text{LTS}}, \ldots, \hat{v}_{n,\text{LTS}})$, we have $\sum_{i=1}^{L}\left[r_{(i)}(\hat{\boldsymbol{\beta}}_{\text{LTS}})\right]^2 = \mathcal{L}(\hat{\boldsymbol{v}}_{\text{LTS}}, \hat{\boldsymbol{\beta}}_{\text{LTS}})$. Then, with the optimalty of $(\hat{\boldsymbol{v}}_{\text{HDS}}, \hat{\boldsymbol{\beta}}_{\text{HDS}})$, it follows that

$$\mathcal{L}(\hat{\boldsymbol{v}}_{\text{HDS}}, \hat{\boldsymbol{\beta}}_{\text{HDS}}) \leq \mathcal{L}(\hat{\boldsymbol{v}}_{\text{LTS}}, \hat{\boldsymbol{\beta}}_{\text{LTS}}) = \sum_{i=1}^{L}\left[r_{(i)}(\hat{\boldsymbol{\beta}}_{\text{LTS}})\right]^2. \tag{S4.7}$$

Together with (S4.6) and (S4.7), we have

$$\mathcal{L}(\hat{\boldsymbol{v}}_{\text{HDS}}, \hat{\boldsymbol{\beta}}_{\text{HDS}}) = \sum_{i=1}^{L}\left[r_{(i)}(\hat{\boldsymbol{\beta}}_{\text{HDS}})\right]^2 = \sum_{i=1}^{L}\left[r_{(i)}(\hat{\boldsymbol{\beta}}_{\text{LTS}})\right]^2 = \mathcal{L}(\hat{\boldsymbol{v}}_{\text{LTS}}, \hat{\boldsymbol{\beta}}_{\text{LTS}}). \tag{S4.8}$$

Then, we deal with the relationship between LTS and MSMOD. Denote $(\hat{\boldsymbol{\beta}}_{\text{OD}}, \hat{\boldsymbol{\xi}}_{\text{OD}})$

the optimal solution of problem (S1.2). Therefore, we have

$$\hat{\xi}_{i,\mathrm{OD}} = \begin{cases} 0, & R(|y_i - \boldsymbol{x}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}}_{\mathrm{OD}}|) \leq L \\ y_i - \boldsymbol{x}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}}_{\mathrm{OD}}, & R(|y_i - \boldsymbol{x}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}}_{\mathrm{OD}}|) > L \end{cases}, \quad i = 1, \ldots, n.$$

By letting $\mathcal{L}_{\mathrm{OD}}(\boldsymbol{\beta}, \boldsymbol{\xi}) = \sum_{i=1}^{n}(y_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta} - \xi_i)^2$, it follows that

$$\sum_{i=1}^{L}\left[r_{(i)}(\hat{\boldsymbol{\beta}}_{\mathrm{LTS}})\right]^2 \leq \sum_{i=1}^{L}\left[r_{(i)}(\hat{\boldsymbol{\beta}}_{\mathrm{OD}})\right]^2 = \mathcal{L}_{\mathrm{OD}}(\hat{\boldsymbol{\beta}}_{\mathrm{OD}}, \hat{\boldsymbol{\xi}}_{\mathrm{OD}}). \tag{S4.9}$$

Define

$$\hat{\xi}_{i,\mathrm{LTS}} = \begin{cases} 0, & R(|y_i - \boldsymbol{x}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}}_{\mathrm{LTS}}|) \leq L \\ y_i - \boldsymbol{x}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}}_{\mathrm{LTS}}, & R(|y_i - \boldsymbol{x}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}}_{\mathrm{LTS}}|) > L \end{cases}, \quad i = 1, \ldots, n.$$

Then, it can be obtained that

$$\mathcal{L}_{\mathrm{OD}}(\hat{\boldsymbol{\beta}}_{\mathrm{OD}}, \hat{\boldsymbol{\xi}}_{\mathrm{OD}}) \leq \sum_{i=1}^{n}(y_i - \boldsymbol{x}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}}_{\mathrm{LTS}} - \hat{\xi}_{i,\mathrm{LTS}})^2 = \sum_{i=1}^{L}\left[r_{(i)}(\hat{\boldsymbol{\beta}}_{\mathrm{LTS}})\right]^2. \tag{S4.10}$$

Together with (S4.9) and (S4.10), we have

$$\mathcal{L}_{\mathrm{OD}}(\hat{\boldsymbol{\beta}}_{\mathrm{OD}}, \hat{\boldsymbol{\xi}}_{\mathrm{OD}}) = \sum_{i=1}^{L}\left[r_{(i)}(\hat{\boldsymbol{\beta}}_{\mathrm{LTS}})\right]^2. \tag{S4.11}$$

Based on (S4.8) and (S4.11), Proposition 1 can be proved. ■

S4.2 Proof of Theorem 1

The proof of Theorem 1 is built upon the following technical lemma.

Lemma 1. *Let $e_{max}(\boldsymbol{v}^{(t+1)})$ be the largest eigenvalue of $\mathbf{A}(\boldsymbol{v}^{(t+1)}) = n^{-1}\sum_{i=1}^{n} v_i^{(t+1)}\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{x}_i$ with a given $\boldsymbol{v}^{(t+1)}$. If $u \geq e_{max}(\boldsymbol{v}^{(t+1)})$,*

$$\mathcal{L}(\boldsymbol{v}^{(t+1)}, \boldsymbol{\gamma}^{(h+1)}) \leq \mathcal{L}(\boldsymbol{v}^{(t+1)}, \boldsymbol{\gamma}^{(h)}) - \frac{u - e_{max}(\boldsymbol{v}^{(t+1)})}{2}\|\boldsymbol{\gamma}^{(h+1)} - \boldsymbol{\gamma}^{(h)}\|_2^2. \tag{S4.12}$$

Moreover, if $\mathbf{A}(\boldsymbol{v}^{(t+1)}, s) = n^{-1}\sum_{i=1}^{n} v_i^{(t+1)}\boldsymbol{x}_{i,s}^{\mathrm{T}}\boldsymbol{x}_{i,s}$ is positive defined for any s satisfying $\mathbb{N}(s) \leq K$, $\boldsymbol{\gamma}^{(h)}$ in IHT converges to a local minimum of $\mathcal{L}(\boldsymbol{v}^{(t+1)}, \boldsymbol{\beta})$.

The proof of Lemma 1 can be referred to the proof of Theorem 1 in Xu and Chen (2014). With Lemma 1, we prove Theorem 1 as follows.

Proof of Theorem 1. We first prove that $\mathcal{L}(\boldsymbol{v}^{(t)}, \boldsymbol{\beta}_{\mathrm{IHT}}^{(t)})$ decreases after each iteration. Based on the definition (3.14), it is obvious that $\mathcal{L}(\boldsymbol{v}^{(t+1)}, \boldsymbol{\beta}_{\mathrm{IHT}}^{(t)}) \leq \mathcal{L}(\boldsymbol{v}^{(t)}, \boldsymbol{\beta}_{\mathrm{IHT}}^{(t)})$. Then, we show that $\mathcal{L}(\boldsymbol{v}^{(t+1)}, \boldsymbol{\beta}_{\mathrm{IHT}}^{(t+1)}) \leq \mathcal{L}(\boldsymbol{v}^{(t+1)}, \boldsymbol{\beta}_{\mathrm{IHT}}^{(t)})$. Recall $e_{max} = \max\{\lambda_{\max}(\mathbf{A}(\boldsymbol{v})), \|\boldsymbol{v}\|_0 = L\}$. Given a $\boldsymbol{\gamma}^{(h)}$, together with the condition in Theorem 1 and Lemma 1, we have

$$\mathcal{L}(\boldsymbol{v}^{(t+1)}, \boldsymbol{\gamma}^{(h+1)}) \leq \mathcal{L}(\boldsymbol{v}^{(t+1)}, \boldsymbol{\gamma}^{(h)}). \tag{S4.13}$$

Thus, as $\boldsymbol{\beta}_{\mathrm{IHT}}^{(t+1)}$ is a limiting point of $\boldsymbol{\gamma}^{(h)}$, it follows that $\mathcal{L}(\boldsymbol{v}^{(t+1)}, \boldsymbol{\beta}_{\mathrm{IHT}}^{(t+1)}) \leq \mathcal{L}(\boldsymbol{v}^{(t)}, \boldsymbol{\beta}_{\mathrm{IHT}}^{(t)})$. Since $\mathcal{L}(\boldsymbol{v}, \boldsymbol{\beta})$ is lower bounded by 0, $|\mathcal{L}(\boldsymbol{v}^{(t+1)}, \boldsymbol{\beta}_{\mathrm{IHT}}^{(t+1)}) - \mathcal{L}(\boldsymbol{v}^{(t)}, \boldsymbol{\beta}_{\mathrm{IHT}}^{(t)})| \to 0$ as $t \to \infty$.

Next, we prove $\{\boldsymbol{v}^{(t)}, \boldsymbol{\beta}_{\text{IHT}}^{(t)}\}$ converges by contradiction. If $\{\boldsymbol{v}^{(t)}, \boldsymbol{\beta}^{(t)}\}$ does not converge, there must exist a $\delta > 0$ independent with t and infinite $t_1, t_2, \ldots, t_\infty$ such that $\boldsymbol{v}^{(t_k+1)} \neq \boldsymbol{v}^{(t_k)}$ or $\|\boldsymbol{\beta}_{\text{IHT}}^{(t_k+1)} - \boldsymbol{\beta}_{\text{IHT}}^{(t_k)}\|_2 > \delta$. At least one of the following Cases 1-3 happens, and we prove that each case will result in a contradiction.

Consider **Case 1**: $\boldsymbol{v}^{(t_k+1)} = \boldsymbol{v}^{(t_k)}$, $\|\boldsymbol{\beta}_{\text{IHT}}^{(t_k+1)} - \boldsymbol{\beta}_{\text{IHT}}^{(t_k)}\|_2 > \delta$. Since the intial $\boldsymbol{\beta}^{(t)} = \boldsymbol{\beta}_{\text{IHT}}^{(t_k)}$ and $\boldsymbol{v}^{(t_k+1)} = \boldsymbol{v}^{(t_k)}$, we have $\boldsymbol{\gamma}^{(h)} = \boldsymbol{\gamma}^{(0)}$ for all $h \geq 1$ in IHT update for $\boldsymbol{\beta}_{\text{IHT}}^{(t+1)}$. Thus, $\boldsymbol{\beta}_{\text{IHT}}^{(t_k+1)} = \boldsymbol{\beta}_{\text{IHT}}^{(t_k)}$; this leads to a contradiction.

Consider **Case 2**: $\boldsymbol{v}^{(t_k+1)} \neq \boldsymbol{v}^{(t_k)}$, $\boldsymbol{\beta}_{\text{IHT}}^{(t_k+1)} = \boldsymbol{\beta}_{\text{IHT}}^{(t_k)}$. Since the update of $\boldsymbol{v}^{(t+1)}$ in (3.14) is unique, it follows that $\boldsymbol{v}^{(t_k+2)} = \boldsymbol{v}^{(t_k+1)}$ when $\boldsymbol{\beta}_{\text{IHT}}^{(t_k+1)} = \boldsymbol{\beta}_{\text{IHT}}^{(t_k)}$. By the result of Case 1, for all $t \geq t_k + 1$, we have $\boldsymbol{v}^{(t+1)} = \boldsymbol{v}^{(t)}$ and $\boldsymbol{\beta}_{\text{IHT}}^{(t+1)} = \boldsymbol{\beta}_{\text{IHT}}^{(t)}$. This results in a contradiction with that there must exist a $\delta > 0$ and infinite $t_1, t_2, \ldots, t_\infty$ such that $\boldsymbol{v}^{(t_k+1)} \neq \boldsymbol{v}^{(t_k)}$ or $\|\boldsymbol{\beta}_{\text{IHT}}^{(t_k+1)} - \boldsymbol{\beta}_{\text{IHT}}^{(t_k)}\|_2 > \delta$.

Consider **Case 3**: $\boldsymbol{v}^{(t_k+1)} \neq \boldsymbol{v}^{(t_k)}$, $\|\boldsymbol{\beta}_{\text{IHT}}^{(t_k+1)} - \boldsymbol{\beta}_{\text{IHT}}^{(t_k)}\|_2 > \delta$. By the definition of $\boldsymbol{\beta}_{\text{IHT}}^{(t_k+1)}$ with $\boldsymbol{\gamma}^{(0)} = \boldsymbol{\beta}_{\text{IHT}}^{(t_k)}$, we have $\boldsymbol{\gamma}^{(h)} \to \boldsymbol{\beta}_{\text{IHT}}^{(t_k+1)}$ as $h \to \infty$. Since n and p are finite, the pattern numbers of $(\boldsymbol{v}, s)$ is $\binom{n}{L}\binom{p}{K}$, where $\|\boldsymbol{v}\|_0 = L$ and $\mathbb{N}(s) = K$. Thus, there exists a unified finite h^* such that $\|\boldsymbol{\gamma}^{(h^*)} - \boldsymbol{\beta}_{\text{IHT}}^{(t_k)}\|_2^2 > \delta/2$ for all possible $\boldsymbol{v}^{(t_k+1)}$ and $\boldsymbol{\beta}_{\text{IHT}}^{(t_k)}$. Based on (S4.12) in Lemma 1, we have

$$\mathcal{L}(\boldsymbol{v}^{(t_k+1)}, \boldsymbol{\gamma}^{(h)}) - \mathcal{L}(\boldsymbol{v}^{(t_k+1)}, \boldsymbol{\gamma}^{(h+1)}) \geq 1/2(u - e_{max})\|\boldsymbol{\gamma}^{(h+1)} - \boldsymbol{\gamma}^{(h)}\|_2^2.$$

When $u > e_{max}$, it follows that

$$\begin{aligned}
\mathcal{L}(\boldsymbol{v}^{(t_k+1)}, \boldsymbol{\beta}_{\text{IHT}}^{(t_k)}) - \mathcal{L}(\boldsymbol{v}^{(t_k+1)}, \boldsymbol{\beta}_{\text{IHT}}^{(t_k+1)}) &\geq \mathcal{L}(\boldsymbol{v}^{(t_k+1)}, \boldsymbol{\beta}_{\text{IHT}}^{(t_k)}) - \mathcal{L}(\boldsymbol{v}^{(t_k+1)}, \boldsymbol{\gamma}^{(h^*)}) \\
&= \sum_{h=0}^{h^*-1} [\mathcal{L}(\boldsymbol{v}^{(t_k+1)}, \boldsymbol{\gamma}^{(h)}) - \mathcal{L}(\boldsymbol{v}^{(t_k+1)}, \boldsymbol{\gamma}^{(h+1)})] \\
&\geq \frac{1}{2}(u - e_{max}) \sum_{h=0}^{h^*-1} \|\boldsymbol{\gamma}^{(t+1)} - \boldsymbol{\gamma}^{(t)}\|_2^2 \\
&\geq \frac{1}{2h^*}(u - e_{max}) \|\boldsymbol{\gamma}^{(h^*)} - \boldsymbol{\beta}_{\text{IHT}}^{(t_k)}\|_2^2 \\
&\geq \frac{\delta}{4h^*}(u - e_{max}) > 0.
\end{aligned}$$

There exists a exists a constant $c > 0$ such that

$$\mathcal{L}(\boldsymbol{v}^{(t_k+1)}, \boldsymbol{\beta}_{\text{IHT}}^{(t_k)}) - \mathcal{L}(\boldsymbol{v}^{(t_k+1)}, \boldsymbol{\beta}_{\text{IHT}}^{(t_k+1)}) > c.$$

As $k \to \infty$, $\mathcal{L}(\boldsymbol{v}^{(t_k+1)}, \boldsymbol{\beta}_{\text{IHT}}^{(t_k+1)}) \to -\infty$, this contradicts the fact $\mathcal{L}$ is lower bounded.

Based on the contradictions in Cased 1, 2, and 3, we have $\{\boldsymbol{v}^{(t)}, \boldsymbol{\beta}^{(t)}\}$ converges to a limiting point $\{\widetilde{\boldsymbol{v}}, \widetilde{\boldsymbol{\beta}}\}$. In addition, Lemma 1 shows that given $\boldsymbol{v}^{(t+1)}$, $\boldsymbol{\gamma}^{(h)}$ in IHT converges to a local minimum of $\mathcal{L}(\boldsymbol{v}^{(t+1)}, \boldsymbol{\beta})$ subject to $\|\boldsymbol{\beta}\|_0 \leq K$. This indicates that $\widetilde{\boldsymbol{\beta}}$ is a local minimum of $\mathcal{L}(\widetilde{\boldsymbol{v}}, \boldsymbol{\beta})$ subject to $\|\boldsymbol{\beta}\|_0 \leq K$. The Theorem 1 is therefore proved. ■

S4.3 Proof of Theorem 2

At first, we show that $\nabla_{\boldsymbol{\beta}}\, G_\rho(\boldsymbol{\beta}|\boldsymbol{\beta}^*)\big|_{\boldsymbol{\beta}=\boldsymbol{\beta}^*} = \mathbf{0}$ under the assumption that $f_0(y|\boldsymbol{x}) = 0$ for all $(y, \boldsymbol{x})$ satisfying $|y - \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^*| \leq \nu_\rho^*$. To be specific,

$$
\begin{aligned}
\nabla_{\boldsymbol{\beta}}\, G_\rho(\boldsymbol{\beta}|\boldsymbol{\beta}^*)\big|_{\boldsymbol{\beta}=\boldsymbol{\beta}^*} &= \int_{\mathcal{X}}\int_{\mathcal{Y}} I(|y - \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^*| \leq \nu_\rho^*) 2\boldsymbol{x}(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^* - y) f(\boldsymbol{x}, y)\mathrm{d}y\mathrm{d}\boldsymbol{x} \\
&= 2\int_{\mathcal{X}}\int_{\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^*-\nu_\rho^*}^{\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^*+\nu_\rho^*} \boldsymbol{x}(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^* - y)\frac{\pi_1}{\sqrt{2\pi}\sigma} e^{\frac{-(y-\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^*)^2}{2\sigma^2}} \mathrm{d}y f(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \\
&= 2\int_{\mathcal{X}} \boldsymbol{x}(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^* - \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^*) \int_{\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^*-\nu_\rho^*}^{\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^*+\nu_\rho^*} \frac{\pi_1}{\sqrt{2\pi}\sigma} e^{\frac{-(y-\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^*)^2}{2\sigma^2}} \mathrm{d}y f(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \\
&= \mathbf{0},
\end{aligned}
$$

where $\mathcal{X}$ and $\mathcal{Y}$ are the support of $\boldsymbol{x}$ and y, respectively.

Next, for any $\widetilde{\boldsymbol{\beta}} = (\widetilde{\beta}_1, \ldots, \widetilde{\beta}_p)^{\mathrm{T}}$ satisfying that $\widetilde{\boldsymbol{\beta}} \neq \boldsymbol{\beta}^*$ and $\|\widetilde{\boldsymbol{\beta}}\|_0 \leq K$, we show that $G_\rho(\boldsymbol{\beta}^*|\boldsymbol{\beta}^*) < G_\rho(\widetilde{\boldsymbol{\beta}}|\boldsymbol{\beta}^*)$. Denote $\eta = \int_{\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^*-\nu_\rho^*}^{\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^*+\nu_\rho^*} \frac{\pi_1}{\sqrt{2\pi}\sigma} e^{\frac{-(y-\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^*)^2}{2\sigma^2}}\mathrm{d}y$, and we have

$$
\begin{aligned}
G_\rho(\boldsymbol{\beta}^*|\boldsymbol{\beta}^*) - G_\rho(\widetilde{\boldsymbol{\beta}}|\boldsymbol{\beta}^*) &= \int_{\mathcal{X}}\int_{\mathcal{Y}} I(|y - \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^*| \leq \nu_\rho^*)[(y - \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^*)^2 - (y - \boldsymbol{x}^{\mathrm{T}}\widetilde{\boldsymbol{\beta}})^2] f(\boldsymbol{x}, y)\mathrm{d}y\mathrm{d}\boldsymbol{x} \\
&= \int_{\mathcal{X}}\int_{\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^*-\nu_\rho^*}^{\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^*+\nu_\rho^*} [2y(\boldsymbol{x}^{\mathrm{T}}\widetilde{\boldsymbol{\beta}} - \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^*) + (\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^*)^2 - (\boldsymbol{x}^{\mathrm{T}}\widetilde{\boldsymbol{\beta}})^2] f(y|\boldsymbol{x})\mathrm{d}y f(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \\
&= \eta \int_{\mathcal{X}} [2\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^*(\boldsymbol{x}^{\mathrm{T}}\widetilde{\boldsymbol{\beta}} - \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^*) + (\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^*)^2 - (\boldsymbol{x}^{\mathrm{T}}\widetilde{\boldsymbol{\beta}})^2] f(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \\
&= -\eta \int_{\mathcal{X}} [\boldsymbol{x}(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)]^2 f(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \\
&= -\eta(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)_{\mathcal{M}'}^{\mathrm{T}} \left[\int \boldsymbol{x}_{\mathcal{M}'}\boldsymbol{x}_{\mathcal{M}'}^{\mathrm{T}} f(\boldsymbol{x}_{\mathcal{M}'})\mathrm{d}\boldsymbol{x}_{\mathcal{M}'}\right] (\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)_{\mathcal{M}'}, \qquad \text{(S4.14)}
\end{aligned}
$$

where $\mathcal{M}' = \mathcal{M} \cup \widetilde{\mathcal{M}}$ with $\widetilde{\mathcal{M}} = \{j : \widetilde{\beta}_j \neq 0\}$. Since $\int \boldsymbol{x}_s \boldsymbol{x}_s^{\mathrm{T}} f(\boldsymbol{x}_s) \mathrm{d}\boldsymbol{x}_s$ is assumed to be positive defined for any s satisfying $\mathbb{N}(s) \leq \mathbb{N}(\mathcal{M}) + K$ and $\eta > 0$, we have $G_\rho(\boldsymbol{\beta}^*|\boldsymbol{\beta}^*) - G_\rho(\widetilde{\boldsymbol{\beta}}|\boldsymbol{\beta}^*) < 0$.

Then, we work on $\boldsymbol{\gamma}^{(h+1)} = H(\widetilde{Q}(\boldsymbol{\gamma}^{(h)}); K)$ with $\boldsymbol{\gamma}^{(0)} = \boldsymbol{\beta}^*$. Considering $h = 1$, we have

$$\begin{aligned} \tilde{Q}(\boldsymbol{\gamma}^{(1)}) &= \boldsymbol{\beta}^* - u^{-1} \nabla_{\boldsymbol{\beta}} \, G_\rho(\boldsymbol{\beta}|\boldsymbol{\beta}^*)\big|_{\boldsymbol{\beta}=\boldsymbol{\beta}^*} \\ &= \boldsymbol{\beta}^*. \end{aligned}$$

Since $\|\boldsymbol{\beta}^*\|_0 \leq K$, $\boldsymbol{\gamma}^{(1)} = H(\widetilde{Q}(\boldsymbol{\beta}^*); K) = \boldsymbol{\beta}^*$. As a result, $\boldsymbol{\gamma}^{(h)} = \boldsymbol{\gamma}^{(h-1)} = \ldots = \boldsymbol{\gamma}^{(0)} = \boldsymbol{\beta}^*$ and

$$\boldsymbol{\gamma}^{(h+1)} = H(\widetilde{Q}(\boldsymbol{\gamma}^{(h)}); K) = \boldsymbol{\beta}^*.$$

The proof of Theorem 2 is completed. ■

S4.4 Proof of Theorem 3

Recall that for any $\widetilde{\boldsymbol{\beta}} \in \Theta$, there exists a corresponding $\widetilde{\nu}_\rho$ such that $E[I(|y - \boldsymbol{x}^{\mathrm{T}}\widetilde{\boldsymbol{\beta}}| \leq \widetilde{\nu}_\rho)] = \rho$. In addition, $f(y|\boldsymbol{x}) = \pi_1 f_1(y|\boldsymbol{x}) = \frac{\pi_1}{\sqrt{2\pi}\sigma} e^{\frac{-(y-\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^*)^2}{2\sigma^2}}$ for $(y, \boldsymbol{x}) \in \Lambda$. For any

$\widetilde{\boldsymbol{\beta}} \neq \boldsymbol{\beta}^*$, since $E[I(|y-\boldsymbol{x}^{\mathrm{T}}\widetilde{\boldsymbol{\beta}}| \leq \widetilde{\nu}_\rho)] = E[I(|y-\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^*| \leq \nu_\rho^*)] = \rho$, we have $\widetilde{\nu}_\rho > \nu_\rho^*$ and

$$\begin{aligned}
G_\rho(\widetilde{\boldsymbol{\beta}}|\widetilde{\boldsymbol{\beta}}) &= \int_{\mathcal{X}}\int_{\boldsymbol{x}^{\mathrm{T}}\widetilde{\boldsymbol{\beta}}-\widetilde{\nu}_\rho}^{\boldsymbol{x}^{\mathrm{T}}\widetilde{\boldsymbol{\beta}}+\widetilde{\nu}_\rho}(y-\boldsymbol{x}^{\mathrm{T}}\widetilde{\boldsymbol{\beta}})^2 f(y|\boldsymbol{x})\mathrm{d}yf(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\\
&= \pi_1\int_{\mathcal{X}}\int_{\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^*-\widetilde{\nu}_\rho}^{\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^*+\widetilde{\nu}_\rho}(y-\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^*)^2 f_1(y+\boldsymbol{x}^{\mathrm{T}}\widetilde{\boldsymbol{\beta}}-\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^*|\boldsymbol{x})\mathrm{d}yf(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\\
&= \pi_1\int_{\mathcal{X}}\int_{\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^*-\widetilde{\nu}_\rho}^{\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^*-\nu_\rho^*}(y-\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^*)^2 f_1(y+\boldsymbol{x}^{\mathrm{T}}\widetilde{\boldsymbol{\beta}}-\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^*|\boldsymbol{x})\mathrm{d}yf(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\\
&\quad +\pi_1\int_{\mathcal{X}}\int_{\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^*-\nu_\rho^*}^{\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^*+\nu_\rho^*}(y-\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^*)^2 f_1(y+\boldsymbol{x}^{\mathrm{T}}\widetilde{\boldsymbol{\beta}}-\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^*|\boldsymbol{x})\mathrm{d}yf(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\\
&\quad +\pi_1\int_{\mathcal{X}}\int_{\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^*+\nu_\rho^*}^{\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^*+\widetilde{\nu}_\rho}(y-\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^*)^2 f_1(y+\boldsymbol{x}^{\mathrm{T}}\widetilde{\boldsymbol{\beta}}-\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^*|\boldsymbol{x})\mathrm{d}yf(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\\
&= E_1+E_2+E_3.
\end{aligned}$$

Letting $\Delta(y|\boldsymbol{x}) = f_1(y+\boldsymbol{x}^{\mathrm{T}}\widetilde{\boldsymbol{\beta}}-\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^*|\boldsymbol{x}) - f_1(y|\boldsymbol{x})$, we have

$$\begin{aligned}
& G_\rho(\widetilde{\boldsymbol{\beta}}|\widetilde{\boldsymbol{\beta}}) - G_\rho(\boldsymbol{\beta}^*|\boldsymbol{\beta}^*)\\
= \; & E_1+E_3+\pi_1\int_{\mathcal{X}}\int_{\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^*-\nu_\rho^*}^{\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^*+\nu_\rho^*}(y-\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^*)^2\Delta(y|\boldsymbol{x})\mathrm{d}yf(\boldsymbol{x})\mathrm{d}\boldsymbol{x}. \qquad \text{(S4.15)}
\end{aligned}$$

Obviously, given $\boldsymbol{x}$, there is at most one $y' \in [\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^*-\nu_\rho^*, \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^*+\nu_\rho^*]$ such that $f_1(y'|\boldsymbol{x}) = f_1(y'+\boldsymbol{x}^{\mathrm{T}}\widetilde{\boldsymbol{\beta}}-\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^*|\boldsymbol{x})$. We first discuss the situation that $\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^*-\nu_\rho^* < \boldsymbol{x}^{\mathrm{T}}\widetilde{\boldsymbol{\beta}}-\widetilde{\nu}_\rho$; this means that $\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^* < \boldsymbol{x}^{\mathrm{T}}\widetilde{\boldsymbol{\beta}}$. Then, two potential cases are analyzed as below.

Consider **Case 1**: for any $y \in [\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^*-\nu_\rho^*, \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^*+\nu_\rho^*]$, $f_1(y+\boldsymbol{x}^{\mathrm{T}}\widetilde{\boldsymbol{\beta}}-\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^*|\boldsymbol{x}) < f_1(y|\boldsymbol{x})$.

Given an $\boldsymbol{x}$, we have

$$\int_{\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^*-\widetilde{\nu}_\rho}^{\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^*-\nu_\rho^*}(y-\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^*)^2 f_1(y+\boldsymbol{x}^{\mathrm{T}}\widetilde{\boldsymbol{\beta}}-\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^*|\boldsymbol{x})\mathrm{d}y > \nu_\rho^{*2}\int_{\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^*-\widetilde{\nu}_\rho}^{\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^*-\nu_\rho^*} f_1(y+\boldsymbol{x}^{\mathrm{T}}\widetilde{\boldsymbol{\beta}}-\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^*|\boldsymbol{x})\mathrm{d}y, \tag{S4.16}$$

$$\int_{\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^*+\nu_\rho^*}^{\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^*+\widetilde{\nu}_\rho}(y-\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^*)^2 f_1(y+\boldsymbol{x}^{\mathrm{T}}\widetilde{\boldsymbol{\beta}}-\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^*|\boldsymbol{x})\mathrm{d}y > \nu_\rho^{*2}\int_{\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^*+\nu_\rho^*}^{\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^*+\widetilde{\nu}_\rho} f_1(y+\boldsymbol{x}^{\mathrm{T}}\widetilde{\boldsymbol{\beta}}-\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^*|\boldsymbol{x})\mathrm{d}y, \tag{S4.17}$$

$$\int_{\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^*-\nu_\rho^*}^{\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^*+\nu_\rho^*}(y-\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^*)^2\Delta(y|\boldsymbol{x})\mathrm{d}y > \nu_\rho^{*2}\int_{\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^*-\nu_\rho^*}^{\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^*+\nu_\rho^*}\Delta(y|\boldsymbol{x})\mathrm{d}y. \tag{S4.18}$$

Together with (S4.15)-(S4.18), it follows that

$$\begin{aligned}
& G_\rho(\widetilde{\boldsymbol{\beta}}|\widetilde{\boldsymbol{\beta}}) - G_\rho(\boldsymbol{\beta}^*|\boldsymbol{\beta}^*) \\
> \ & \pi_1\nu_\rho^{*2}\int_{\mathcal{X}}\left[\int_{\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^*-\widetilde{\nu}_\rho}^{\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^*+\widetilde{\nu}_\rho} f_1(y+\boldsymbol{x}^{\mathrm{T}}\widetilde{\boldsymbol{\beta}}-\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^*|\boldsymbol{x})\mathrm{d}y - \int_{\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^*-\nu_\rho^*}^{\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^*+\nu_\rho^*} f_1(y|\boldsymbol{x})\mathrm{d}y\right] f_1(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \\
= \ & \pi_1\nu_\rho^{*2}\int_{\mathcal{X}}\left[\int_{\boldsymbol{x}^{\mathrm{T}}\widetilde{\boldsymbol{\beta}}-\widetilde{\nu}_\rho}^{\boldsymbol{x}^{\mathrm{T}}\widetilde{\boldsymbol{\beta}}+\widetilde{\nu}_\rho} f_1(y|\boldsymbol{x})\mathrm{d}y - \int_{\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^*-\nu_\rho^*}^{\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^*+\nu_\rho^*} f_1(y|\boldsymbol{x})\mathrm{d}y\right] f(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \\
= \ & \nu_\rho^{*2}E[I(|y-\boldsymbol{x}^{\mathrm{T}}\widetilde{\boldsymbol{\beta}}| \le \widetilde{\nu}_\rho)] - \nu_\rho^{*2}E[I(|y-\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^*| \le \nu_\rho^*)] \\
= \ & \nu_\rho^{*2}(\rho-\rho) \ = \ 0.
\end{aligned} \tag{S4.19}$$

Consider **Case 2**: there exists a $y' = (3\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^* - \boldsymbol{x}^{\mathrm{T}}\widetilde{\boldsymbol{\beta}})/2 \in (\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^* - \nu_\rho^*, \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^*)$ such that $f_1(y' + \boldsymbol{x}^{\mathrm{T}}\widetilde{\boldsymbol{\beta}} - \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^*|\boldsymbol{x}) = f_1(y'|\boldsymbol{x})$. Obviously, if $y \in [\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^* - \nu_\rho^*, y')$, we have $f_1(y + \boldsymbol{x}^{\mathrm{T}}\widetilde{\boldsymbol{\beta}} - \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^*|\boldsymbol{x}) - f_1(y|\boldsymbol{x}) = \Delta(y|\boldsymbol{x}) > 0$; if $y \in [y', \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^* + \nu_\rho^*]$, we have $f_1(y +$

$\boldsymbol{x}^{\mathrm{T}}\widetilde{\boldsymbol{\beta}} - \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^*|\boldsymbol{x}) - f_1(y|\boldsymbol{x}) = \Delta(y|\boldsymbol{x}) < 0$. Furthermore,

$$\int_{\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^*-\nu_\rho^*}^{y'} \Delta(y|\boldsymbol{x})\mathrm{d}y + \int_{y'}^{2y'-\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^*+\nu_\rho^*} \Delta(y|\boldsymbol{x})\mathrm{d}y \;=\; 0. \tag{S4.20}$$

Since $f_1(y|\boldsymbol{x})$ is symmetric about $\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^*$, $f_1(y+\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^*|\boldsymbol{x}) = f_1(-y+\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^*|\boldsymbol{x})$ holds. As a result,

$$\begin{aligned}
& \int_{\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^*-\nu_\rho^*}^{y'} (y-\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^*)^2\Delta(y|\boldsymbol{x})\mathrm{d}y + \int_{y'}^{2y'-\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^*+\nu_\rho^*} (y-\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^*)^2\Delta(y|\boldsymbol{x})\mathrm{d}y \\
= & \int_{\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^*-\nu_\rho^*}^{y'} (y-\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^*)^2\Delta(y|\boldsymbol{x})\mathrm{d}y - \int_{\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^*-\nu_\rho^*}^{y'} (-y+2y'-\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^*)^2\Delta(y|\boldsymbol{x})\mathrm{d}y \\
= & \int_{\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^*-\nu_\rho^*}^{y'} \left[(y-\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^*)^2 - (y-2y'+\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^*)^2\right]\Delta(y|\boldsymbol{x})\mathrm{d}y \\
= & \int_{\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^*-\nu_\rho^*}^{y'} 4(y'-\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^*)(y-y')\Delta(y|\boldsymbol{x})\mathrm{d}y > 0.
\end{aligned} \tag{S4.21}$$

Then, based on (S4.20) and (S4.21),

$$\begin{aligned}
& \int_{\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^*-\nu_\rho^*}^{\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^*+\nu_\rho^*} (y-\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^*)^2\Delta(y|\boldsymbol{x})\mathrm{d}y \\
= & \int_{\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^*-\nu_\rho^*}^{2y'-\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^*+\nu_\rho^*} (y-\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^*)^2\Delta(y|\boldsymbol{x})\mathrm{d}y + \int_{2y'-\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^*+\nu_\rho^*}^{\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^*+\nu_\rho^*} (y-\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^*)^2\Delta(y|\boldsymbol{x})\mathrm{d}y \\
> & \nu_\rho^{*2}\left(\int_{\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^*-\nu_\rho^*}^{y'} \Delta(y|\boldsymbol{x})\mathrm{d}y + \int_{y'}^{2y'-\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^*+\nu_\rho^*} \Delta(y|\boldsymbol{x})\mathrm{d}y\right) + \int_{2y'-\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^*+\nu_\rho^*}^{\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^*+\nu_\rho^*} (y-\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^*)^2\Delta(y|\boldsymbol{x})\mathrm{d}y \\
> & \nu_\rho^{*2}\left(\int_{\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^*-\nu_\rho^*}^{y'} \Delta(y|\boldsymbol{x})\mathrm{d}y + \int_{y'}^{2y'-\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^*+\nu_\rho^*} \Delta(y|\boldsymbol{x})\mathrm{d}y + \int_{2y'-\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^*+\nu_\rho^*}^{\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^*+\nu_\rho^*} \Delta(y|\boldsymbol{x})\mathrm{d}y\right) \\
= & \nu_\rho^{*2}\int_{\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^*-\nu_\rho^*}^{\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^*+\nu_\rho^*} \Delta(y|\boldsymbol{x})\mathrm{d}y.
\end{aligned} \tag{S4.22}$$

Similar to (S4.19), together with (S4.15), (S4.16), (S4.17), and (S4.22), we have $G_\rho(\widetilde{\boldsymbol{\beta}}|\widetilde{\boldsymbol{\beta}}) - G_\rho(\boldsymbol{\beta}^*|\boldsymbol{\beta}^*) > 0$.

By using the same techniques, we can also prove $G_\rho(\widetilde{\boldsymbol{\beta}}|\widetilde{\boldsymbol{\beta}}) - G_\rho(\boldsymbol{\beta}^*|\boldsymbol{\beta}^*) > 0$ under the situation that $\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^* + \nu_\rho^* > \boldsymbol{x}^{\mathrm{T}}\widetilde{\boldsymbol{\beta}} + \widetilde{\nu}_\rho$. Therefore, we obtain $G_\rho(\boldsymbol{\beta}^*|\boldsymbol{\beta}^*) < G_\rho(\widetilde{\boldsymbol{\beta}}|\widetilde{\boldsymbol{\beta}})$ under the conditions specified in the Theorem 3. The proof is complete. ■

S4.5 Proof of Theorem 4

The following Lemma 2-6 are provided for the proof of Theorem 4.

Lemma 2. *(Hoeffding's inequality; Hoeffding (1963)) Let $X_1, \ldots, X_n$ be independent random variables. Assume that $P(X_i \in [a_i, b_i]) = 1$ for $1 \le i \le n$, where a_i and b_i are constants. Let $\bar{X} = n^{-1}\sum_{i=1}^n X_i$. Then the following inequality holds*

$$P(|\bar{X} - E(\bar{X})| \ge \epsilon) \le 2\exp\left(-\frac{2n^2\epsilon^2}{\sum_{i=1}^n (b_i - a_i)^2}\right),$$

where ϵ is a positive constant and $E(\bar{X})$ is the expected value of $\bar{X}$.

Lemma 3. *Let $X_1, \ldots, X_n$ is a sample of X with distribution function $F(x) = P(X \le x)$. For $0 < \rho < 1$, suppose that ν_ρ is the unique solution x of $F(-x) \le \rho \le F(x)$. The estimator of ν_ρ is defined by*

$$\hat{\nu}_\rho = \begin{cases} X_{(n\rho)}, & \text{if } n\rho \text{ is an integer,} \\ X_{(\lfloor n\rho \rfloor + 1)}, & \text{if } n\rho \text{ is not an integer,} \end{cases}$$

where $X_{(i)}$ *is the* ith *order statistic. Then, for every* $\epsilon > 0$,

$$P\left(|\hat{\nu}_\rho - \nu_\rho| > \epsilon\right) \leq 2\exp(-2n\delta_\epsilon^2),$$

where $\delta_\epsilon = \min\{F(\nu_\rho + \epsilon) - \rho, \rho - F(\nu_\rho - \epsilon)\}$.

Proof of Lemma 3. We first decompose

$$P\left(|\hat{\nu}_\rho - \nu_\rho| > \epsilon\right) = P\left(\hat{\nu}_\rho > \nu_\rho + \epsilon\right) + P\left(\hat{\nu}_\rho < \nu_\rho - \epsilon\right).$$

By the definition of $\hat{\nu}_\rho$, we have $\hat{\nu}_\rho = \inf\{x : F_n(x) \geq \rho\}$, where $F_n(x)$ is the empirical distribution function given a sample. Thus,

$$\begin{aligned} P\left(\hat{\nu}_\rho > \nu_\rho + \epsilon\right) &= P\left(\rho > F_n(\nu_\rho + \epsilon)\right) \\ &= P\left(\frac{1}{n}\sum_{i=1}^n I(X_i > \nu_\rho + \epsilon) > (1-\rho)\right) \\ &= P\left(\frac{1}{n}\sum_{i=1}^n I(X_i > \nu_\rho + \epsilon) - [1 - F(\nu_\rho + \epsilon)] > \delta_1\right), \end{aligned}$$

where $\delta_1 = F(\nu_\rho + \epsilon) - \rho$. Therefore, utilizing Hoeffding's inequality in Lemma 2, we have

$$P\left(\hat{\nu}_\rho > \nu_\rho + \epsilon\right) \leq \exp(-2n\delta_1^2).$$

Similarly,

$$\begin{aligned} P\left(\hat{\nu}_\rho < \nu_\rho - \epsilon\right) &= P\left(\rho < F_n(\nu_\rho - \epsilon)\right) \\ &= P\left(\frac{1}{n}\sum_{i=1}^{n} I(X_i < \nu_\rho - \epsilon) > \rho\right) \\ &= P\left(\frac{1}{n}\sum_{i=1}^{n} I(X_i < \nu_\rho - \epsilon) - F(\nu_\rho - \epsilon) > \delta_2\right) \\ &\le \exp(-2n\delta_2^2), \end{aligned}$$

where $\delta_2 = \rho - F(\nu_\rho - \epsilon)$. Putting $\delta_\epsilon = \min\{\delta_1, \delta_2\}$, the proof is completed. ■

To better show the following Lemmas 4-6, we introduce some new notations. Let $\boldsymbol{\gamma}^{(0.5)} = Q(\boldsymbol{\gamma}^{(0)})$, $\boldsymbol{\gamma}_\rho^{(0.5)} = Q_\rho(\boldsymbol{\gamma}^{(0)})$, and $\hat{\mathcal{M}}^{(0.5)} = \{j : |\gamma_j^{(0.5)}| \text{ is among the first } K \text{ largest of all}\}$. For an index set $\mathcal{S} \subseteq \{1, \ldots, p\}$, we define the function $T(\boldsymbol{a}; \mathcal{S})$ as

$$[T(\boldsymbol{a}; \mathcal{S})]_j = \begin{cases} a_j, & \text{if } j \in \mathcal{S}, \\ 0, & \text{if } j \notin \mathcal{S}. \end{cases}$$

Then, we write $\hat{\boldsymbol{\gamma}}_\rho^{(1)} = T(\boldsymbol{\gamma}_\rho^{(0.5)}; \hat{\mathcal{M}}^{(0.5)})$.

Lemma 4. *(Lemma 5.1 in Wang et al. (2014)) Suppose that we have* $\|\boldsymbol{\gamma}_\rho^{(0.5)} - \boldsymbol{\beta}^*\|_2 \le \kappa \cdot \|\boldsymbol{\beta}^*\|_2$ *for some* $\kappa \in (0, 1)$. *Assuming that we have*

$$K \ge \frac{4 \cdot (1+\kappa)^2}{(1-\kappa)^2} \cdot \mathcal{M}, \text{ and } \sqrt{K}\|\boldsymbol{\gamma}^{(0.5)} - \boldsymbol{\gamma}_\rho^{(0.5)}\|_\infty \le \frac{(1-\kappa)^2}{2(1+\kappa)} \cdot \|\boldsymbol{\beta}^*\|_2, \tag{S4.23}$$

Then it holds that

$$\|\hat{\boldsymbol{\gamma}}_{\rho}^{(1)} - \boldsymbol{\beta}^*\|_2 \le \frac{C \cdot \sqrt{m}}{\sqrt{1-\kappa}} \cdot \|\boldsymbol{\gamma}^{(0.5)} - \boldsymbol{\gamma}_{\rho}^{(0.5)}\|_\infty + \Big(1 + 4 \cdot \sqrt{m/K}\Big)^{1/2} \cdot \|\boldsymbol{\gamma}_{\rho}^{(0.5)} - \boldsymbol{\beta}^*\|_2, \tag{S4.24}$$

where $m = \mathbb{N}(\mathcal{M})$.

Lemma 5. *Assume that $G_\rho(\boldsymbol{\beta}^*|\boldsymbol{\beta}^*) \le G_\rho(\boldsymbol{\beta}|\boldsymbol{\beta}^*)$ for all $\mathcal{B}(R;\boldsymbol{\beta}^*)$. Define $\boldsymbol{\gamma}_{\rho}^{(0.5)} = \boldsymbol{\beta}^{(t)} - u^{-1}\nabla G_\rho(\boldsymbol{\beta}|\boldsymbol{\beta}^{(t)})|_{\boldsymbol{\beta}=\boldsymbol{\beta}^{(t)}}$ with stepsize $1/\mu = 2/(C_2 + C_3)$. With Conditions C2 and C3, we have*

$$\|\boldsymbol{\gamma}_{\rho}^{(0.5)} - \boldsymbol{\beta}^*\|_2 \le \left(1 - 2 \cdot \frac{C_3 - C_1}{C_3 + C_2}\right) \|\boldsymbol{\beta}^{(t)} - \boldsymbol{\beta}^*\|_2. \tag{S4.25}$$

Proof of Lemma 5. By the definition of $\boldsymbol{\gamma}_{\rho}^{(0.5)}$, we have

$$\begin{aligned}
& \|\boldsymbol{\gamma}_{\rho}^{(0.5)} - \boldsymbol{\beta}^*\|_2 \\
= \ & \|\boldsymbol{\beta}^{(t)} - u^{-1}\nabla_{\boldsymbol{\beta}} G_\rho(\boldsymbol{\beta}|\boldsymbol{\beta}^{(t)})|_{\boldsymbol{\beta}=\boldsymbol{\beta}^{(t)}} - \boldsymbol{\beta}^*\|_2 \\
= \ & \|\boldsymbol{\beta}^{(t)} - u^{-1}\nabla_{\boldsymbol{\beta}} G_\rho(\boldsymbol{\beta}|\boldsymbol{\beta}^*)|_{\boldsymbol{\beta}=\boldsymbol{\beta}^{(t)}} + u^{-1}\nabla_{\boldsymbol{\beta}} G_\rho(\boldsymbol{\beta}|\boldsymbol{\beta}^*)|_{\boldsymbol{\beta}=\boldsymbol{\beta}^{(t)}} - u^{-1}\nabla_{\boldsymbol{\beta}} G_\rho(\boldsymbol{\beta}|\boldsymbol{\beta}^{(t)})|_{\boldsymbol{\beta}=\boldsymbol{\beta}^{(t)}} - \boldsymbol{\beta}^*\|_2 \\
\le \ & \|\boldsymbol{\beta}^{(t)} - u^{-1}\nabla_{\boldsymbol{\beta}} G_\rho(\boldsymbol{\beta}|\boldsymbol{\beta}^*)|_{\boldsymbol{\beta}=\boldsymbol{\beta}^{(t)}} - \boldsymbol{\beta}^*\|_2 \\
& + u^{-1}\|\nabla_{\boldsymbol{\beta}} G_\rho(\boldsymbol{\beta}|\boldsymbol{\beta}^*)|_{\boldsymbol{\beta}=\boldsymbol{\beta}^{(t)}} - \nabla_{\boldsymbol{\beta}} G_\rho(\boldsymbol{\beta}|\boldsymbol{\beta}^{(t)})|_{\boldsymbol{\beta}=\boldsymbol{\beta}^{(t)}}\|_2 \qquad \text{(S4.26)}
\end{aligned}$$

We deal with the first term in (S4.25). Recall $G_\rho(\boldsymbol{\beta}|\boldsymbol{\beta}^*)$ is C_2-Lipschitz-smooth and C_3-

strongly convex. By invoking standard optimization results for minimizing strongly convex and smooth objective functions, e.g., in (Nesterov, 2013), for stepsize $1/\mu = 2/(C_2 + C_3)$, we have

$$\|\boldsymbol{\beta}^{(t)} - u^{-1}\nabla_{\boldsymbol{\beta}} G_\rho(\boldsymbol{\beta}|\boldsymbol{\beta}^*)|_{\boldsymbol{\beta}=\boldsymbol{\beta}^{(t)}} - \boldsymbol{\beta}^*\|_2 \leq \left(1 - \frac{C_2 - C_3}{C_2 + C_3}\right) \cdot \left\|\boldsymbol{\beta}^{(t)} - \boldsymbol{\beta}^*\right\|_2. \tag{S4.27}$$

Next, with Condition C1, the second term can be upper bounded by

$$u^{-1}\|\nabla_{\boldsymbol{\beta}} G_\rho(\boldsymbol{\beta}|\boldsymbol{\beta}^*)|_{\boldsymbol{\beta}=\boldsymbol{\beta}^{(t)}} - \nabla_{\boldsymbol{\beta}} G_\rho(\boldsymbol{\beta}|\boldsymbol{\beta}^{(t)})|_{\boldsymbol{\beta}=\boldsymbol{\beta}^{(t)}}\|_2 = \frac{2C_1}{C_2 + C_3} \cdot \left\|\boldsymbol{\beta}^{(t)} - \boldsymbol{\beta}^*\right\|_2. \tag{S4.28}$$

Taking (S4.27) and (S4.28) into (S4.26), (S4.25) is proved. The proof is completed. ■

Lemma 6. *Assume each feature x_j is bounded, i.e. $\|\boldsymbol{x}\|_\infty \leq d$, where d is a positive constant. Let $\delta_\rho = \min\{P(|y_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}^{(t)}| \leq 2\nu_\rho^{(t)}) - \rho, \rho\}$. For each $\epsilon > 0$ and $n > (8d\nu_\rho^{(t)})/u\epsilon$, we have*

$$P\left(\|\boldsymbol{\gamma}^{(0.5)} - \boldsymbol{\gamma}_\rho^{(0.5)}\|_\infty \leq \epsilon\right) \geq 1 - \delta,$$

where

$$\delta = 2\exp\left(-\frac{nu^2\epsilon^2}{32(d\nu_\rho^{(t)})^2} + \frac{u\epsilon}{2d\nu_\rho^{(t)}} - \frac{2}{n}\right) + 2\exp(-2n\delta_\rho^2) + 2p\cdot\exp\left(-\frac{nu^2\epsilon^2}{(2d\nu_\rho^{(t)})^2}\right).$$

Proof of Lemma 6. Denote

$$G_{n,\rho}(\boldsymbol{\beta}|\boldsymbol{\beta}^{(t)}) = \frac{1}{n}\sum_{i=1}^{n} I(|y_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}^{(t)}| \le \nu_\rho^{(t)})(y_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta})^2,$$

where $\nu_\rho^{(t)}$ satisfies $E[I(|y - \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^{(t)}| \le \nu_\rho^{(t)})] = \rho$. Then, we can bound

$$\begin{aligned} \|\boldsymbol{\gamma}^{(0.5)} - \boldsymbol{\gamma}_\rho^{(0.5)}\|_\infty &= \|Q(\boldsymbol{\gamma}^{(0)}) - Q_\rho(\boldsymbol{\gamma}^{(0)})\|_\infty \\ &= \frac{1}{u}\|\nabla_{\boldsymbol{\gamma}}\mathcal{L}(\boldsymbol{v}^{(t+1)}, \boldsymbol{\gamma})|_{\boldsymbol{\gamma}=\boldsymbol{\gamma}^{(0)}} - \nabla G_\rho(\boldsymbol{\gamma}|\boldsymbol{\beta}^{(t)})|_{\boldsymbol{\gamma}=\boldsymbol{\gamma}^{(0)}}\|_\infty \\ &\le \frac{1}{u}\|\nabla_{\boldsymbol{\gamma}}\mathcal{L}(\boldsymbol{v}^{(t+1)}, \boldsymbol{\gamma})|_{\boldsymbol{\gamma}=\boldsymbol{\gamma}^{(0)}} - \nabla_{\boldsymbol{\gamma}} G_{n,\rho}(\boldsymbol{\gamma}|\boldsymbol{\beta}^{(t)})|_{\boldsymbol{\gamma}=\boldsymbol{\gamma}^{(0)}}\|_\infty \\ &\quad + \frac{1}{u}\|\nabla_{\boldsymbol{\gamma}} G_{n,\rho}(\boldsymbol{\gamma}|\boldsymbol{\beta}^{(t)})|_{\boldsymbol{\gamma}=\boldsymbol{\gamma}^{(0)}} - \nabla_{\boldsymbol{\gamma}} G_\rho(\boldsymbol{\gamma}|\boldsymbol{\beta}^{(t)})|_{\boldsymbol{\gamma}=\boldsymbol{\gamma}^{(0)}}\|_\infty. \end{aligned} \tag{S4.29}$$

Recall that $\boldsymbol{\gamma}^{(0)} = \boldsymbol{\beta}^{(t)}$. By letting $\hat{\nu}_\rho^{(t)} = |y - \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^{(t)}|_{(L)}$, we have

$$\begin{aligned}
& \|\nabla_{\boldsymbol{\gamma}}\mathcal{L}(\boldsymbol{v}^{(t+1)}, \boldsymbol{\gamma})|_{\boldsymbol{\gamma}=\boldsymbol{\beta}^{(t)}} - \nabla_{\boldsymbol{\gamma}} G_{n,\rho}(\boldsymbol{\gamma}|\boldsymbol{\beta}^{(t)})|_{\boldsymbol{\gamma}=\boldsymbol{\beta}^{(t)}}\|_\infty \\
= \quad & \|\frac{2}{n}\sum_{i=1}^n I(|y_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}^{(t)}| \le \hat{\nu}_\rho^{(t)})\boldsymbol{x}_i(y_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}^{(t)}) - \frac{2}{n}\sum_{i=1}^n I(|y_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}^{(t)}| \le \nu_\rho^{(t)})\boldsymbol{x}_i(y_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}^{(t)})\|_\infty \\
= \quad & 2\|\frac{1}{n}\sum_{i=1}^n [I(|y_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}^{(t)}| \le \hat{\nu}_\rho^{(t)}) - I(|y_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}^{(t)}| \le \nu_\rho^{(t)})]\boldsymbol{x}_i(y_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}^{(t)})\|_\infty \\
=: \quad & 2A_1
\end{aligned}$$

Denote $F_{n,r}(x) = n^{-1}\sum_{i=1}^n I(|y_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}^{(t)}| \le x)$ be the empirical distribution function of $|y_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}^{(t)}|$. Then, A_1 can be further bounded in probability as follows.

$$\begin{aligned}
P(A_1 > \epsilon) \quad \le \quad & P\left(\max_j \left|\frac{1}{n}\sum_{i=1}^n I(\nu_\rho^{(t)} \le |y_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}^{(t)}| \le \hat{\nu}_\rho^{(t)})x_{ij}(y_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}^{(t)})\right| > \epsilon, \nu_\rho^{(t)} \le \hat{\nu}_\rho^{(t)}\right) \\
& + P\left(\max_j \left|\frac{1}{n}\sum_{i=1}^n I(\hat{\nu}_\rho^{(t)} \le |y_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}^{(t)}| \le \nu_\rho^{(t)})x_{ij}(y_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}^{(t)})\right| > \epsilon, \hat{\nu}_\rho^{(t)} < \nu_\rho^{(t)}\right) \\
\le \quad & P\left(\left[\frac{1}{n}\sum_{i=1}^n I(\nu_\rho^{(t)} \le |y_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}^{(t)}| \le \hat{\nu}_\rho^{(t)})\right] \cdot d\hat{\nu}_\rho^{(t)} > \epsilon, \nu_\rho^{(t)} \le \hat{\nu}_\rho^{(t)}\right) \\
& + P\left(\left[\frac{1}{n}\sum_{i=1}^n I(\hat{\nu}_\rho^{(t)} \le |y_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}^{(t)}| \le \nu_\rho^{(t)})\right] \cdot d\nu_\rho^{(t)} > \epsilon, \hat{\nu}_\rho^{(t)} < \nu_\rho^{(t)}\right) \\
\le \quad & P\left(F_{n,r}(\hat{\nu}_\rho^{(t)}) - F_{n,r}(\nu_\rho^{(t)}) > \frac{\epsilon}{2d\nu_\rho^{(t)}}, \nu_\rho^{(t)} \le \hat{\nu}_\rho^{(t)}\right) + P\left(\hat{\nu}_\rho^{(t)} > 2\nu_\rho^{(t)}\right) \\
& + P\left(F_{n,r}(\nu_\rho^{(t)}) - F_{n,r}(\hat{\nu}_\rho^{(t)}) > \frac{\epsilon}{d\nu_\rho^{(t)}}, \hat{\nu}_\rho^{(t)} < \nu_\rho^{(t)}\right) \\
\le \quad & P\left(|F_{n,r}(\hat{\nu}_\rho^{(t)}) - F_{n,r}(\nu_\rho^{(t)})| > \frac{\epsilon}{2d\nu_\rho^{(t)}}\right) + P\left(\hat{\nu}_\rho^{(t)} - \nu_\rho^{(t)} > \nu_\rho^{(t)}\right). \qquad \text{(S4.30)}
\end{aligned}$$

We work on the first term in (S4.30). Based on the definition of $\hat{\nu}_\rho^{(t)}$ and $\hat{\nu}_\rho^{(t)}$, for a given $\epsilon > (2d\nu_\rho^{(t)}/n)$, we have

$$\begin{aligned}
& P\left(|F_{n,r}(\hat{\nu}_\rho^{(t)}) - F_{n,r}(\nu_\rho^{(t)})| > \frac{\epsilon}{2d\nu_\rho^{(t)}}\right) \\
\leq \; & P\left(|F_{n,r}(\hat{\nu}_\rho^{(t)}) - \rho| + |F_{n,r}(\nu_\rho^{(t)}) - E[I(|y - \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^{(t)}| \leq \nu_\rho^{(t)})]| > \frac{\epsilon}{2d\nu_\rho^{(t)}}\right) \\
\leq \; & P\left(|F_{n,r}(\nu_\rho^{(t)}) - E[I(|y - \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^{(t)}| \leq \nu_\rho^{(t)})]| > \frac{\epsilon}{2d\nu_\rho^{(t)}} - \frac{1}{n}\right) \\
\leq \; & 2\exp\left(-\frac{n\epsilon^2}{2(d\nu_\rho^{(t)})^2} + \frac{2\epsilon}{d\nu_\rho^{(t)}} - \frac{2}{n}\right),
\end{aligned}$$

where the last inequality is from Hoeffding's inequality in Lemma 2.

Then, we deal with the second term in (S4.30). With Lemma 3, it is easy to show the empirical quantile $\hat{\nu}_\rho^{(t)}$ is consistent. To be specific,

$$P\left(\hat{\nu}_\rho^{(t)} - \nu_\rho^{(t)} > \nu_\rho^{(t)}\right) \leq 2\exp(-2n\delta_\rho^2),$$

where $\delta_\rho = \min\{P(|y_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}^{(t)}| \leq 2\nu_\rho) - \rho, \rho\}$.

Thus, for a given $\epsilon > (4d\nu_\rho^{(t)}/un)$, we have

$$\begin{aligned}
& P\left(\frac{1}{u}\|\nabla_{\boldsymbol{\gamma}}\mathcal{L}(\boldsymbol{v}^{(t+1)},\boldsymbol{\gamma})|_{\boldsymbol{\gamma}=\boldsymbol{\beta}^{(t)}} - \nabla_{\boldsymbol{\gamma}}G_{n,\rho}(\boldsymbol{\gamma}|\boldsymbol{\beta}^{(t)})|_{\boldsymbol{\gamma}=\boldsymbol{\beta}^{(t)}}\|_\infty > \epsilon\right) \\
\leq\ & P\left(A_1 > \frac{u\epsilon}{2}\right) \\
\leq\ & 2\exp\left(-\frac{nu^2\epsilon^2}{8(d\nu_\rho^{(t)})^2} + \frac{u\epsilon}{d\nu_\rho^{(t)}} - \frac{2}{n}\right) + 2\exp(-2n\delta_\rho^2). \qquad \text{(S4.31)}
\end{aligned}$$

Then, we work on $\|\nabla_{\boldsymbol{\gamma}}G_{n,\rho}(\boldsymbol{\gamma}|\boldsymbol{\beta}^{(t)})|_{\boldsymbol{\gamma}=\boldsymbol{\beta}^{(t)}} - \nabla_{\boldsymbol{\gamma}}G_\rho(\boldsymbol{\gamma}|\boldsymbol{\beta}^{(t)})|_{\boldsymbol{\gamma}=\boldsymbol{\beta}^{(t)}}\|_\infty$ in (S4.29). Let $Z_{ij} = I(|y_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}^{(t)}| \leq \nu_\rho^{(t)})x_{ij}(y_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}^{(t)})$. Obviously, $|Z_{ij}|$ can be upper bounded by $d\nu_\rho^{(t)}$. It follows that

$$\begin{aligned}
& P\left(\frac{1}{u}\|\nabla_{\boldsymbol{\gamma}}G_{n,\rho}(\boldsymbol{\gamma}|\boldsymbol{\beta}^{(t)})|_{\boldsymbol{\gamma}=\boldsymbol{\gamma}^{(h)}} - \nabla_{\boldsymbol{\gamma}}G_\rho(\boldsymbol{\gamma}|\boldsymbol{\beta}^{(t)})|_{\boldsymbol{\gamma}=\boldsymbol{\gamma}^{(h)}}\|_\infty > \epsilon\right) \\
=\ & P\left(\max_j \ \Big|\frac{1}{n}\sum_{i=1}^n Z_{ij} - E(Z_{ij})\Big| > u\epsilon\right) \\
\leq\ & \sum_{j=1}^p P\left(\Big|\frac{1}{n}\sum_{i=1}^n Z_{ij} - E(Z_{ij})\Big| > u\epsilon\right) \\
\leq\ & 2p\cdot\exp\left(-\frac{nu^2\epsilon^2}{(d\nu_\rho^{(t)})^2}\right). \qquad \text{(S4.32)}
\end{aligned}$$

Together with (S4.31) and (S4.32),

$$\begin{aligned} P\left(\|\boldsymbol{\gamma}^{(0.5)} - \boldsymbol{\gamma}_{\rho}^{(0.5)}\|_{\infty} \leq \epsilon\right) &\leq P\left(\frac{1}{u}\|\nabla_{\boldsymbol{\gamma}}\mathcal{L}(\boldsymbol{v}^{(t+1)}, \boldsymbol{\gamma})|_{\boldsymbol{\gamma}=\boldsymbol{\beta}^{(t)}} - \nabla_{\boldsymbol{\gamma}} G_{n,\rho}(\boldsymbol{\gamma}|\boldsymbol{\beta}^{(t)})|_{\boldsymbol{\gamma}=\boldsymbol{\beta}^{(t)}}\|_{\infty} > \frac{\epsilon}{2}\right) \\ &\quad + P\left(\frac{1}{u}\|\nabla_{\boldsymbol{\gamma}} G_{n,\rho}(\boldsymbol{\gamma}|\boldsymbol{\beta}^{(t)})|_{\boldsymbol{\gamma}=\boldsymbol{\gamma}^{(h)}} - \nabla_{\boldsymbol{\gamma}} G_{\rho}(\boldsymbol{\gamma}|\boldsymbol{\beta}^{(t)})|_{\boldsymbol{\gamma}=\boldsymbol{\gamma}^{(h)}}\|_{\infty} > \frac{\epsilon}{2}\right) \\ &\leq 1 - 2\exp\left(-\frac{nu^2\epsilon^2}{32(d\nu_{\rho}^{(t)})^2} + \frac{u\epsilon}{2d\nu_{\rho}^{(t)}} - \frac{2}{n}\right) \\ &\quad -2\exp(-2n\delta_{\rho}^2) - 2p \cdot \exp\left(-\frac{nu^2\epsilon^2}{(2d\nu_{\rho}^{(t)})^2}\right). \end{aligned}$$

Since ϵ and p are allowed to be varying with n, the above probability bound can be expressed as $1 - O(\exp\{-cn\epsilon^2 + \log p\})$ with some positive constant c. This lemma is proved. ■

Proof of Theorem 4. Given a $\boldsymbol{\beta}^{(t)}$, we first bound $\|\boldsymbol{\beta}^{(t+1)} - \boldsymbol{\beta}^*\|_2$ as follows. Recall $\boldsymbol{\gamma}^{(0)}$ is set to be $\boldsymbol{\beta}^{(t)}$. When the IHT procedure involves one iteration, we have

$$\begin{aligned} \|\boldsymbol{\beta}^{(t+1)} - \boldsymbol{\beta}^*\|_2 &= \|\boldsymbol{\gamma}^{(1)} - \boldsymbol{\beta}^*\|_2 \\ &= \|H(\boldsymbol{\gamma}^{(0.5)}; K) - \boldsymbol{\beta}^*\|_2 \\ &\leq \|T(\boldsymbol{\gamma}^{(0.5)}; \hat{\mathcal{M}}^{(0.5)}) - T(\boldsymbol{\gamma}_{\rho}^{(0.5)}; \hat{\mathcal{M}}^{(0.5)})\|_2 + \|\hat{\boldsymbol{\gamma}}_{\rho}^{(1)} - \boldsymbol{\beta}^*\|_2. \end{aligned} \tag{S4.33}$$

With Lemma 6, for each $\epsilon > 0$ and $n > (8d\nu_{\rho}^{(t)})/u\epsilon$, the first term in (S4.33) can be further bounded by

$$\begin{aligned}
& \|T(\boldsymbol{\gamma}^{(0.5)}; \hat{\mathcal{M}}^{(t+0.5)}) - T(\boldsymbol{\gamma}_\rho^{(0.5)}; \hat{\mathcal{M}}^{(t+0.5)})\|_2 \\
= \;& \|(\boldsymbol{\gamma}^{(0.5)} - \boldsymbol{\gamma}_\rho^{(0.5)})_{\hat{\mathcal{M}}^{(t+0.5)}}\|_2 \\
\leq \;& \sqrt{K}\|(\boldsymbol{\gamma}^{(0.5)} - \boldsymbol{\gamma}_\rho^{(0.5)})_{\hat{\mathcal{M}}^{(t+0.5)}}\|_\infty \\
\leq \;& \sqrt{K}\|\boldsymbol{\gamma}^{(0.5)} - \boldsymbol{\gamma}_\rho^{(0.5)}\|_\infty \\
\leq \;& \sqrt{K}\epsilon
\end{aligned}$$

with probability at least $1-\delta$. If all assumptions in Lemma 4 hold, the second term $\|\hat{\boldsymbol{\gamma}}_\rho^{(1)} - \boldsymbol{\beta}^*\|_2$ in (S4.33) can be upper bounded by using (S4.24). Denote $\varrho = 1 - 2(C_3 - C_1)/(C_3 + C_2) \in (0,1)$. Then, with Lemma 5, we have

$$\begin{aligned}
\|\boldsymbol{\beta}^{(t+1)} - \boldsymbol{\beta}^*\|_2 &= \|\boldsymbol{\gamma}^{(1)} - \boldsymbol{\beta}^*\|_2 \\
&\leq \left(\sqrt{K} + \frac{C \cdot \sqrt{m}}{\sqrt{1-\kappa}}\right) \cdot \epsilon + \left(1 + 4 \cdot \sqrt{m/K}\right)^{1/2} \cdot \|\boldsymbol{\gamma}_\rho^{(0.5)} - \boldsymbol{\beta}^*\|_2 \\
&\leq \left(\sqrt{K} + \frac{C \cdot \sqrt{m}}{\sqrt{1-\kappa}}\right) \cdot \epsilon + \left(1 + 4 \cdot \sqrt{m/K}\right)^{1/2} \cdot \varrho \cdot \|\boldsymbol{\gamma}^{(0)} - \boldsymbol{\beta}^*\|_2, \\
&\leq \left(\sqrt{K} + \frac{C \cdot \sqrt{m}}{\sqrt{1-\kappa}}\right) \cdot \epsilon + \left(1 + 4 \cdot \sqrt{m/K}\right)^{1/2} \cdot \varrho \cdot \|\boldsymbol{\beta}^{(t)} - \boldsymbol{\beta}^*\|_2
\end{aligned} \tag{S4.34}$$

occurs with probability at least $1-\delta$.

Next, under the event

$$\mathcal{E} = \left\{ \|Q(\boldsymbol{\beta}^{(t)}) - Q_\rho(\boldsymbol{\beta}^{(t)})\|_\infty \le \epsilon, \text{ for all } t = 0, 1, \ldots, T-1 \right\},$$

we prove that

$$\|\boldsymbol{\beta}^{(t)} - \boldsymbol{\beta}^*\|_2 \le \frac{\sqrt{K} + C\sqrt{m/(1-\kappa)}}{1-\sqrt{\varrho}} \cdot \epsilon + \varrho^{t/2} \cdot \|\boldsymbol{\beta}^{(0)} - \boldsymbol{\beta}^*\|_2, \text{ for all } t = 0, 1, \ldots, T-1 \tag{S4.35}$$

by mathematical induction. By invoking Lemma 5 and setting $t = 0$, we have

$$\|\boldsymbol{\gamma}_\rho^{(0.5)} - \boldsymbol{\beta}^*\|_2 \le \varrho\|\boldsymbol{\beta}^{(0)} - \boldsymbol{\beta}^*\|_2 \le \varrho R < R = \kappa\|\boldsymbol{\beta}^*\|_2.$$

Thus, the assumption that $\|\boldsymbol{\gamma}_\rho^{(0.5)} - \boldsymbol{\beta}^*\|_2 \le \kappa \cdot \|\boldsymbol{\beta}^*\|_2$ for some $\kappa \in (0, 1)$ in Lemma 4 holds. In addition, by assuming (3.17) and (3.18) in Theorem 4, we can also verify that the assumption in (S4.23) of Lemma 4 on the event $\mathcal{E}$. Thus, for $t = 0$,

$$\|\boldsymbol{\beta}^{(1)} - \boldsymbol{\beta}^*\|_2 \le \left(\sqrt{K} + \frac{C \cdot \sqrt{m}}{\sqrt{1-\kappa}}\right) \cdot \epsilon + \left(1 + 4 \cdot \sqrt{m/K}\right)^{1/2} \cdot \varrho \cdot \|\boldsymbol{\beta}^{(0)} - \boldsymbol{\beta}^*\|_2. \tag{S4.36}$$

Since it is assumed that $K \geq 16 \cdot (1/\varrho - 1)^{-2} \cdot m$ in (S4.23) of Lemma 4, it follows that

$$(1 + 4 \cdot \sqrt{m/K})^{1/2} \leq 1/\sqrt{\varrho}.$$

Since $1 - \sqrt{\varrho} < 1$ in (S4.35), we can obtain

$$\begin{aligned} \|\boldsymbol{\beta}^{(1)} - \boldsymbol{\beta}^*\|_2 &\leq \left(\sqrt{K} + \frac{C \cdot \sqrt{m}}{\sqrt{1-\kappa}}\right) \cdot \epsilon + \sqrt{\varrho} \cdot \|\boldsymbol{\beta}^{(0)} - \boldsymbol{\beta}^*\|_2 \\ &\leq \frac{\sqrt{K} + C\sqrt{m/(1-\kappa)}}{1 - \sqrt{\varrho}} \cdot \epsilon + \sqrt{\varrho} \cdot \|\boldsymbol{\beta}^{(0)} - \boldsymbol{\beta}^*\|_2. \end{aligned} \tag{S4.37}$$

Obviously, (S4.35) holds for $t = 1$. Next, suppose we have that (S4.35) holds for some $t \geq 1$. Since it is assumed that $\|\boldsymbol{\beta}^{(0)} - \boldsymbol{\beta}^*\|_2 \leq R$ and $(\sqrt{K} + C\sqrt{m/(1-\kappa)}) \cdot \epsilon \leq (1 - \sqrt{\varrho})^2 \cdot R$ in Theorem 4, we have

$$\begin{aligned} \|\boldsymbol{\beta}^{(t)} - \boldsymbol{\beta}^*\|_2 &\leq \frac{\sqrt{K} + C\sqrt{m/(1-\kappa)}}{1 - \sqrt{\varrho}} \cdot \epsilon + \varrho^{t/2} \cdot \|\boldsymbol{\beta}^{(0)} - \boldsymbol{\beta}^*\|_2 \\ &\leq (1 - \sqrt{\varrho}) \cdot R + \sqrt{\varrho} \cdot R = R. \end{aligned} \tag{S4.38}$$

Hence, $\boldsymbol{\beta}^{(t)} \in \mathcal{B}(R; \boldsymbol{\beta}^*)$. Similar to the situation with $t = 0$, by invoking Lemma 5 and setting $\boldsymbol{\gamma}^{(0)} = \boldsymbol{\beta}^{(t)}$, we have

$$\|\boldsymbol{\gamma}_{\varrho}^{(0.5)} - \boldsymbol{\beta}^*\|_2 \leq \varrho \|\boldsymbol{\beta}^{(t)} - \boldsymbol{\beta}^*\|_2 \leq \varrho R < R = \kappa \|\boldsymbol{\beta}^*\|_2.$$

Following the proof of (S4.37), we obtain

$$\begin{aligned}
\|\boldsymbol{\beta}^{(t)} - \boldsymbol{\beta}^*\|_2 &\leq \left(\sqrt{K} + \frac{C \cdot \sqrt{m}}{\sqrt{1-\kappa}}\right) \cdot \epsilon + \sqrt{\varrho} \cdot \|\boldsymbol{\beta}^{(t)} - \boldsymbol{\beta}^*\|_2 \\
&\leq \left(1 + \frac{\sqrt{\varrho}}{1-\sqrt{\varrho}}\right)\left(\sqrt{K} + \frac{C \cdot \sqrt{m}}{\sqrt{1-\kappa}}\right) \cdot \epsilon + \sqrt{\varrho} \cdot \varrho^{t/2} \cdot \|\boldsymbol{\beta}^{(t)} - \boldsymbol{\beta}^*\|_2 \\
&\leq \frac{\sqrt{K} + C\sqrt{m/(1-\kappa)}}{1-\sqrt{\varrho}} \cdot \epsilon + \varrho^{(t+1)/2} \cdot \|\boldsymbol{\beta}^{(0)} - \boldsymbol{\beta}^*\|_2, \qquad \text{(S4.39)}
\end{aligned}$$

where the second inequality is obtained by plugging in (S4.35) for t. Hence, we have (S4.35) holds for $t+1$. By induction, we conclude that (S4.35) holds conditioning on the event $\mathcal{E}$, which occurs with probability at least $1 - t \cdot \delta$. By taking $\varrho = 1 - 2(C_3 - C_1)/(C_3 + C_2) \in (0, 1)$ into (S4.39), Theorem 4 is proved. ■

S4.6 Proof of Theorem 5

By the results in Theorem 4, we have

$$\|\boldsymbol{\beta}^{(t)} - \boldsymbol{\beta}^*\|_2 \leq \varrho^{t/2} \cdot R + \frac{(\sqrt{K} + C'\sqrt{m/(1-\kappa)}) \cdot \epsilon}{1-\sqrt{\varrho}} \qquad \text{(S4.40)}$$

holds with probability at least $1 - t \cdot \delta$ and $\log \delta = O(-cn\epsilon^2 + \log p)$. Together with Condition C4 and the fact $R = \kappa \cdot \|\boldsymbol{\beta}^*\|_2$, (S4.40) can be upper bounded by

$$\|\boldsymbol{\beta}^{(t)} - \boldsymbol{\beta}^*\|_2 \leq \varrho^{t/2} \cdot \kappa \cdot \omega_2 n^{\tau_2} + c_1 n^{\tau_3/2} \cdot \epsilon, \qquad \text{(S4.41)}$$

where c_1 is a positive constant. Recall that it is assumed $\min_{j\in\mathcal{M}}|\beta_j^*| \geq \omega_1 n^{-\tau_1}$ in Condition C4. If the upper bound in (S4.41) is lower than $\omega_1 n^{-\tau_1}$ with an overwhelming probability, the sure screening property ($\lim_{n\to\infty}\mathcal{M}\subset\mathcal{M}^{(t)}\to 1$) is established. To this end, we first set $\epsilon = c_2 n^{-\zeta}$ with $\zeta > \tau_3/2+\tau_1$. Thus, when n is sufficiently large, we have

$$c_1 n^{\tau_3/2}\cdot\epsilon = c_1 c_2 n^{-\zeta+\tau_3/2} < 0.5\cdot\omega_1 n^{-\tau_1}. \tag{S4.42}$$

When $t > 2\log_\varrho[\omega_1/(2\kappa\omega_2)^{-1}\cdot n^{-\tau_1-\tau_2}]$, it can be verified

$$\varrho^{t/2}\cdot\kappa\cdot\omega_2 n^{\tau_2} < 0.5\cdot\omega_1 n^{-\tau_1}. \tag{S4.43}$$

Since $\log p = O(n^a)$ for some $0\leq a<1$ in Condition C4, there exists a positive constant c_3 such that $p\leq c_3 n^a$. Together with (S4.42) and (S4.43), when by setting $\epsilon = cn^{-\zeta}$ and $t = t_0, t_0+1, \ldots t_0+T$, we have that

$$\begin{aligned} P\left(\mathcal{M}\subset\mathcal{M}^{(t)}\right) &\leq P\left(\|\boldsymbol{\beta}^{(t)}-\boldsymbol{\beta}^*\|_2 < \omega_1 n^{-\tau_1}\right)\\ &\leq 1-t\cdot O(\exp\{-cn\epsilon^2+c_3n^a\})\\ &\leq 1-t\cdot O(\exp\{-c\cdot c_2 n^{1-2\zeta}+c_3n^a\}). \end{aligned} \tag{S4.44}$$

Since it is assumed in Condition C4 that there exists a $\zeta\in(\tau_3/2+\tau_1,(1-a)/2)$, we have $1-2\zeta > a$. By the fact that $t\leq t_0+T = o(\exp\{c\cdot c_2 n^{1-2\zeta}-c_3n^a\})$, the above probability bound in (S4.44) goes to 1 as $n\to\infty$. The Theorem 5 is proved. ■

Bibliography

Alfons, A., C. Croux, and S. Gelper (2013). Sparse least trimmed squares regression for analyzing high-dimensional large data sets. *The Annals of Applied Statistics*, 226–248.

Cho, H. and P. Fryzlewicz (2012). High dimensional variable selection via tilting. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 74*(3), 593–622.

Fan, J. and J. Lv (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 70*(5), 849–911.

Hawkins, D. M. (1994). The feasible solution algorithm for least trimmed squares regression. *Computational statistics & data analysis 17*(2), 185–196.

Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association 58*(301), 13–30.

Kurnaz, F. S., I. Hoffmann, and P. Filzmoser (2018). Robust and sparse estimation methods for high-dimensional linear and logistic regression. *Chemometrics and Intelligent Laboratory Systems 172*, 211–222.

Li, G., H. Peng, J. Zhang, and L. Zhu (2012). Robust rank correlation based screening. *The Annals of Statistics 40*(3), 1846–1877.

Mai, Q. and H. Zou (2015). The fused kolmogorov filter: A nonparametric model-free screening method. *The Annals of Statistics 43*(4), 1471–1497.

Nesterov, Y. (2013). *Introductory lectures on convex optimization: A basic course*, Volume 87. Springer Science & Business Media.

Rousseeuw, P. J. and A. M. Leroy (1987). *Robust regression and outlier detection*, Volume 589. John wiley & sons.

She, Y. and A. B. Owen (2011). Outlier detection using nonconvex penalized regression. *Journal of the American Statistical Association 106*(494), 626–639.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological) 58*, 267–288.

Wang, Z., Q. Gu, Y. Ning, and H. Liu (2014). High dimensional expectation-maximization algorithm: Statistical optimization and asymptotic normality. *arXiv preprint arXiv:1412.8729*.

Wen, C., A. Zhang, S. Quan, and X. Wang (2020). Bess: Best subset selection for sparse generalized linear model and cox model. *J Stat Softw 94*, 1–24.

Xu, C. and J. Chen (2014). The sparse mle for ultrahigh-dimensional feature screening. *Journal of the American Statistical Association 109*(507), 1257–1269.

Zang, Q., C. Xu, and K. Burkett (2022). Smle: An r package for joint feature screening in ultrahigh-dimensional glms. *arXiv preprint arXiv:2201.03512*.

Zhu, J., C. Wen, J. Zhu, H. Zhang, and X. Wang (2020). A polynomial algorithm for best-subset selection problem. *Proceedings of the National Academy of Sciences 117*(52), 33117–33123.