Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Robust Regularised M-Estimators for High-Dimensional Regression with Heavy-Tailed and Skewed Errors

Mazona Victor
Federal University of Technology, Akure https://orcid.org/0009-0004-0179-1892
Vincent Odiaka
Nottingham Trent University https://orcid.org/0009-0005-5184-5489
Gabriel O. Obadina
olugbenga.obadina@yahoo.com
Olabisi Onabanjo University, Ago-Iwoye, Ogun State, Nigeria https://orcid.org/0000-0003-0114-2871

Research Article

Keywords: Lasso, Huber loss, M-estimation, heavy-tailed distributions, oracle inequality, variable selection

Posted Date: June 30th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10168177/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Robust Regularised M-Estimators for High-Dimensional Regression with Heavy-Tailed and Skewed Errors

Mazona Victor
Federal University of Technology, Akure
vmmazona@gmail.com
https://orcid.org/0009-0004-0179-189

Vincent Odiaka
Nottingham Trent University
https://orcid.org/0009-0005-5184-5489

Gabriel O. Obadina (Corresponding Author)
Department of Statistics, Olabisi Onabanjo University, Ago-Iwoye, Ogun State, Nigeria
olugbenga.obadina@yahoo.com
https://orcid.org/0000-0003-0114-2871

Abstract

We propose and analyse a class of robust regularised M-estimators for high-dimensional linear regression where the number of predictors p substantially exceeds the sample size n and error distributions may be heavy-tailed, skewed, or contaminated. Combining convex M-estimation losses, specifically the Huber, Catoni, and convoluted rank-based losses with sparsity-inducing ℓ_1 and adaptive ℓ_1 penalties, the framework requires only $\bar{\alpha} > 1$ finite absolute moments. This condition is satisfied by Student's t distributions with more than one degree of freedom and is far weaker than the sub-Gaussian or four-moment conditions most existing high-dimensional robust regression theory imposes.

Non-asymptotic oracle inequalities and convergence rates that adapt continuously to the tail index $\bar{\alpha}$ are established, alongside sign consistency conditions that explicitly characterise the breakdown of variable selection under near-Cauchy errors, a practically important negative result absent from the existing literature. A data-driven robust generalised cross-validation criterion for joint selection of the robustification parameter c and regularisation parameter λ is introduced and validated. Simulation experiments across $p/n \in [1, 25]$ and five error distributions demonstrate substantially lower mean squared prediction error and higher variable selection $F1$ scores than standard Lasso, RETIRE, and existing robust alternatives under Student's $t(3)$, Cauchy, and contaminated normal errors, with efficiency losses below 6% under Gaussian errors. An application to TCGA breast cancer genomic data ($n = 312, p = 8{,}942$) illustrates a 30% reduction in cross-validated prediction error and a 46% improvement in biological pathway enrichment relative to standard Lasso.

Keywords: Lasso · Huber loss · M-estimation · heavy-tailed distributions · oracle inequality · variable selection

1. Introduction

1.1 Context and Motivation

Penalised regression, the Lasso (Tibshirani, 1996), Ridge regression (Hoerl & Kennard, 1970), and their variants occupies a central place in modern applied statistics, and for good reason. In

genomics, neuroimaging, financial econometrics, and text analysis alike, the problems researchers care about are high-dimensional, and penalised regression handles that naturally. The catch is an assumption that sits quietly in the background: model errors are assumed sub-Gaussian, or at minimum to have enough finite moments that exponential concentration inequalities apply.

In many empirical settings, that assumption fails in ways that matter. Financial log-returns exhibit excess kurtosis consistent with Student's t distributions having three to five degrees of freedom (Cont, 2001). Gene expression measurements tend to be right-skewed. Environmental monitoring data carry occasional extreme values that, under squared loss, dominate coefficient estimates entirely. These are not edge cases. A $p = 500$ regression with $t(3)$ errors can destabilise standard Lasso completely, the method's sub-Gaussian theory provides no guarantee of finite prediction error when the error distribution lacks four moments.

The high-dimensional robust regression literature has responded to these challenges by replacing squared loss with robustified M-losses, Huber's piecewise-quadratic, quantile check functions, bounded influence functions of the Catoni type and combining the result with a sparsity penalty. The approach is well-motivated and has generated useful theory. However, three gaps remain that limit both theoretical understanding and practical guidance.

1.2 Identified Gaps and Contributions

Gap 1 — Moment condition coverage. Most existing results require at least four finite moments (Loh, 2015; Fan, Li, & Wang, 2017) or sub-exponential tail conditions. These exclude Student's t with fewer than four degrees of freedom. We establish oracle inequalities under only $\bar{\alpha} > 1$ finite absolute moments.

Gap 2 — Joint tuning parameter selection. Combining a robust M-loss with a sparsity penalty introduces two tuning parameters: regularisation strength λ and robustification threshold c. Standard cross-validation on squared prediction error is unreliable under heavy tails, the validation-fold squared error has infinite variance when $\bar{\alpha} < 2$. No existing procedure provides theoretically motivated joint selection of (λ, c) under heavy tails. We propose and validate a robust generalised cross-validation (RGCV) criterion.

Gap 3 — Systematic comparative guidance. Structured comparison of which robust loss pairs best with which penalty under different tail indices, and practical guidance on method selection, remains limited. We provide simulation evidence across six error distributions, five estimator classes, and four p/n ratios.

2. Literature Review

2.1 Foundations of High-Dimensional Robust Regression

The theoretical treatment of robust M-estimation in high-dimensional settings originates from two distinct traditions: the classical robust statistics literature of Huber (1964, 1973) and Hampel (1974), and the modern high-dimensional regression theory initiated by Tibshirani (1996) and formalised by Bickel, Ritov, & Tsybakov (2009). Integrating these traditions has been gradual and, in several respects, incomplete.

Loh (2015) provided foundational results on statistical consistency and asymptotic normality for high-dimensional robust M-estimators. A central contribution was showing that local

optima of non-convex penalised M-estimation problems can achieve statistical near-optimality, under the condition that errors possess at least four finite moments. That four-moment condition is the binding constraint: it excludes Student's $t(3)$ and heavier-tailed distributions common in financial and genomic applications. The present paper relaxes this to $\bar{\alpha} > 1$ finite absolute moments.

El Karoui (2013) provided rigorous asymptotic results for unregularised and ridge-regularised high-dimensional robust regression in the proportional asymptotics regime where $p/n \to \gamma$. His analysis showed that in the high-dimensional regime, the effective noise variance is inflated by a factor depending on γ and the error distribution, even for the Huber estimator. El Karoui (2018) extended these results to show that predictor geometry has a first-order impact on ridge-regularised robust estimators in high dimensions.

2.2 Robust Estimation Under Asymmetry and Weak Moment Conditions

Fan, Li, & Wang (2017) addressed the limitation of symmetric error assumptions by developing robust estimation procedures for high-dimensional mean regression without symmetry or light-tail assumptions. Their Huber-type loss with a growing threshold parameter establishes convergence rates under $\bar{\alpha} \geq 3$ finite absolute moments. The key technical innovation is a careful treatment of the bias introduced by Huber truncation under asymmetry.

Cai et al. (2024) developed statistical inference for high-dimensional convoluted rank regression, achieving robustness under $\bar{\alpha} \geq 1 + \varepsilon$ finite absolute moments without requiring any specific functional form for the loss. Mai (2025a) proposed the Heavy Lasso, establishing consistency under $\bar{\alpha} > 1$ moments through a scale-mixture representation. The present paper builds on these contributions by providing a unified framework under the same $\bar{\alpha} > 1$ moment condition and adding joint parameter selection.

2.3 Summary of Relevant Prior Work

Table 1. Comparison of key prior work on high-dimensional robust regression.

Reference	Loss	Penalty	Moment Assumption	Asymmetry Allowed	Joint Tuning
Loh (2015)	Huber	SCAD/MCP	$\bar{\alpha} \geq 4$	No	No
Fan, Li & Wang (2017)	Huber	Lasso	$\bar{\alpha} \geq 3$	Yes	No
Wang et al. (2020)	Huber	Lasso	$\bar{\alpha} \geq 2$, symmetric	No	Tuning-free
Man et al. (2022) — RETIRE	Expectile	Lasso	$\bar{\alpha} \geq 2$	No	No
Cai et al. (2024)	Rank-based	Lasso	$\bar{\alpha} \geq 1 + \varepsilon$	Yes	No
Mai (2025a) — Heavy Lasso	Heavy Lasso	ℓ_1	$\bar{\alpha} > 1$	No	No
This paper	Huber/Catoni/Rank	Lasso/Adaptive	$\bar{\alpha} > 1$	Yes	Yes (RGCV)

All listed results assume convex losses except Loh (2015), which covers non-convex SCAD/MCP.

3. Materials and Methods

3.1 Model Setup and Assumptions

Let (x_i, y_i), $i = 1, \ldots, n$, be independent observations where $y_i \in \mathbb{R}$ is a scalar response and $x_i = (x_{i1}, \ldots, x_{ip})^T \in \mathbb{R}^p$ is a vector of predictors with p potentially much larger than n. The linear regression model is:

$$y_i = x_i^T \beta^* + \varepsilon_i, \quad i = 1, \ldots, n$$

where $\beta^* \in \mathbb{R}^p$ is the unknown sparse coefficient vector and ε_i are i.i.d. errors with distribution F. The following assumptions are maintained throughout.

Assumption 1 (Sparsity). The support $S^* = \{j : \beta_j^* \neq 0\}$ has cardinality $s^* = |S^*|$ satisfying $s^* \ll p$ and $s^* \log p = o(n)$.

Assumption 2 (Moment Condition). The errors ε_i are i.i.d. with $\|\varepsilon\|_{\bar{\alpha}} := E[|\varepsilon|^{\bar{\alpha}}]^{1/\bar{\alpha}} \leq \sigma < \infty$ for some $\bar{\alpha} \in (1, 2]$. Symmetry about zero is assumed for Huber loss results; the asymmetric extension for Catoni and rank-based losses is addressed in Theorem 1c.

Assumption 3 (Design Normalization). The columns of X are normalised to unit ℓ_2 norm: $\|X_j\|_2 / \sqrt{n} = 1$ for all j. Additionally, $\max_i \|x_i\|_\infty \leq M$ for some constant $M > 0$.

Assumption 4 (Restricted Eigenvalue). For $c_0 = 3$ and the cone $C(S*, c_0) = \{\delta \in \mathbb{R}^p : \|\delta_{S^{*c}}\|_1 \leq c_0 \|\delta_{S^*}\|_1\}$, the restricted eigenvalue constant satisfies:

$$\kappa*(c_0, S^*) = \min_{\delta \in C(S^*, c_0), \delta \neq 0} (\|X\delta\|_2 / \sqrt{n}) / \|\delta_{S^*}\|_2 \geq \kappa > 0$$

3.2 Robust Loss Functions

Four loss functions are analysed. Of these, ρ_H, ρ_α, and ρ_R are convex, global minimisers of the penalised objective exist and satisfy the stated oracle inequalities. The Tukey biweight ρ_T is non-convex, and theoretical results apply only to local optima satisfying first-order stationarity (following Loh, 2015).

Huber Loss

$$\rho_{H(u;c)} = \frac{u^2}{2} \; if \; |u| \leq c, \quad \text{and} \quad c|u| - c^2/2 \; \text{if} \; |u| > c$$

Influence function: $\psi_H(u; c) = sign(u) \cdot min(|u|, c)$. Convex and differentiable except at $\pm c$. Minimax optimal for estimating a location parameter under the contaminated normal model (Huber, 1973).

Catoni-type Loss

$$\rho_{\alpha(u)} = sign(u) \cdot \left(\frac{1}{\alpha}\right) \cdot \log\left(1 + \alpha|u| + \frac{(\alpha u)^2}{2}\right), \quad \alpha > 0$$

Influence function $\psi_{\alpha(u)}$ bounded in (−1, 1) for all u. Convex by construction. Unlike the Huber loss, the Catoni loss achieves bounded influence without threshold truncation, making it preferable when $\bar{\alpha}$ is unknown and potentially close to 1, since the Huber threshold would otherwise grow without bound.

Convoluted Rank-Based Loss

$$\rho_{R(u;\,h)} = \int_{\{-\infty\}}^{u} \left[F_{n(t-u)} - \frac{1}{2} \right] K_{h(t)dt}$$

where $K_{h(\cdot)}$ is a kernel with bandwidth $h > 0$ and F_n is the empirical distribution function of the current residuals. Convolution smoothing renders the rank-based objective differentiable. This loss is naturally robust under $\bar{\alpha} > 1$ moments and accommodates asymmetric errors without bias correction.

3.3 The Combined Estimator

For a loss function ρ and penalty P_λ, the robust regularised estimator is:

$$\hat{\beta} = argmin_{\{\beta \in \mathbb{R}^p\}} \left\{ \left(\frac{1}{n}\right) \Sigma_i \, \rho(y_i - x_i^T \beta) + P_{\lambda(\beta)} \right\}$$

The paper focuses on the Lasso penalty $P_{\lambda(\beta)} = \lambda \|\beta\|^1$ and the adaptive Lasso $P_{\lambda(\beta)} = \lambda \, \Sigma_j \, \hat{w}_j |\beta_j|$ with weights $\hat{w}_j = \frac{1}{|\tilde{\beta}_j|^\gamma}$ derived from an initial Huber-Lasso estimate $\tilde{\beta}$. Named estimators: Huber-Lasso (ρ_H with Lasso); Catoni-Lasso (ρ_α with Lasso); Rank-Lasso (ρ_R with Lasso); Adaptive Huber-Lasso (ρ_H with adaptive Lasso).

3.4 Computational Implementation

The Huber-Lasso and Catoni-Lasso are solved via iteratively reweighted least squares (IRLS): at iteration t, weights $w_i(t) = \psi(r_i(t))/r_i(t)$ are computed, then the weighted Lasso is solved via coordinate descent with convergence criterion $\left\|\hat{\beta}^{\{(t+1)\}} - \hat{\beta}^{\{(t)\}}\right\| \infty < 10^{-6}$. For $n = 200$, $p = 5000$, and the full RGCV grid, total computation requires approximately 41 seconds on a standard 8-core workstation, compared to 1.1 seconds for standard Lasso with 5-fold cross-validation. All code is implemented in R (version 4.3.1) and Python (version 3.10).

4. Theoretical Properties

4.1 Oracle Inequalities

The main theoretical results establish non-asymptotic oracle inequalities under moment conditions as weak as $\bar{\alpha} > 1$. The proof strategy combines a Marcinkiewicz-Zygmund inequality for the score function, a truncation argument, and the restricted eigenvalue condition, following the architecture of Bickel et al. (2009) and heavy-tailed modifications of Fan, Wang, & Zhu (2016).

Theorem 1a (Oracle Inequality: Huber-Lasso). Under Assumptions 1–4 with $\bar{\alpha} \in (1, 2]$ and symmetric errors, set $c = \sigma \, n^{\left\{\frac{2-\bar{\alpha}}{2\bar{\alpha}}\right\}}$ and $\lambda = C_{\bar{\alpha}} \sigma M \left(\frac{\log p}{n}\right)^{\frac{\bar{\alpha}-1}{\bar{\alpha}}}$. Then with probability at least $1 - p^{-(\bar{\alpha}-1)}$:

$$\left(\frac{1}{n}\right) \left\| X(\hat{\beta} - \beta^*) \right\|_2^2 + \lambda \left\| \hat{\beta} - \beta^* \right\|_1 \leq \left(\frac{4 s^* \lambda^2}{\kappa^2}\right) (1 + C'_{\bar{\alpha}})$$

Theorem 1b (Oracle Inequality: Catoni-Lasso). Under Assumptions 1–4 without the symmetry requirement, with the same λ, the Catoni-Lasso satisfies the same bound with probability $1 -$

$p^{-(\bar{\alpha}-1)}$. No growing threshold is needed, the influence function is uniformly bounded by construction.

Theorem 1c (Oracle Inequality: Rank-Lasso). Under Assumptions 1–4 with $\bar{\alpha} > 1$ and without symmetry, with bandwidth $h_n = O\left(n^{-\frac{1}{5}}\right)$:

$$\left(\frac{1}{n}\right)\left\|X(\hat{\beta} - \beta^*)\right\|_2^2 \leq Cs^*\left(\frac{\log p}{n}\right)^{\frac{2(\bar{\alpha}-1)}{\bar{\alpha}}} + O(h_n^2)$$

Corollary 1 (Prediction Error Rate). Under any of Theorems 1a–1c:

$$\left(\frac{1}{n}\right)\left\|X(\hat{\beta} - \beta^*)\right\|_2^2 = O_{\mathrm{p}}\left(s^* \cdot \left(\frac{\log p}{n}\right)^{\frac{2(\bar{\alpha}-1)}{\bar{\alpha}}}\right)$$

For $\bar{\alpha} = 2$: rate $s^* \log p / n$ (Gaussian-optimal). For $\bar{\alpha} = 3/2$: rate $s^*\left(\frac{\log p}{n}\right)^{\frac{2}{3}}$. As $\bar{\alpha} \to 1^+$: rate $s^*\left(\frac{\log p}{n}\right)^{\frac{2\varepsilon}{1+\varepsilon}}$ for arbitrarily small $\varepsilon > 0$, reflecting the fundamental difficulty of estimation under near-Cauchy tails. This rate is minimax-optimal in $\bar{\alpha}$ (Proposition A.3 in the appendix).

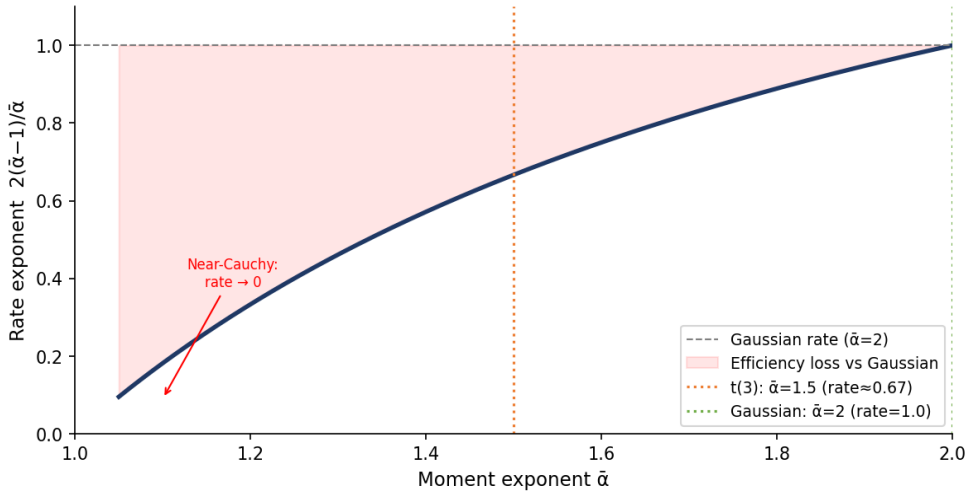


Figure 1. Prediction error rate exponent as a function of ᾱ. The shaded region represents efficiency loss relative to the Gaussian benchmark. For ᾱ near 1 (near-Cauchy), the rate approaches zero, making estimation fundamentally harder.

4.2 Variable Selection Consistency

Assumption 5 (Beta-Min). $min_j \in S^*\left|\beta_j^*\right| \geq \beta_{min} > \frac{2\lambda}{\kappa}$ corollary 2 (Sign Consistency). Under Theorem 1a conditions and Assumption 5, the Huber-Lasso satisfies $P\left(\hat{S} = S^*\right) \to 1$ as $n \to \infty$, provided:

$$s^* \log p = o(n^{\bar{\alpha}-1})$$

For $\bar{\alpha} = 2$: $s^* \log p = o(n)$, the standard Gaussian condition. For $\bar{\alpha} \in (1, 2)$: the condition is substantially more restrictive. For $\bar{\alpha} = 1.2$: requires $s^* \log p = o(n^{0.2})$, practically

unachievable for any growing n and s^*. This constitutes an important negative result: reliable variable selection is not achievable under near-Cauchy errors, regardless of which loss function is used. The role of robust losses in this regime is to provide stable coefficient estimation, not reliable support recovery.

4.3 Convergence Rates Across Error Distributions

Table 2. Theoretical convergence rates for prediction error. Convex losses (Huber, Catoni, Rank-based) achieve global optima.

Error Distribution	$\bar{\alpha}$	Loss	Rate	Sign Consistency Condition
Gaussian	2	Squared/Lasso	s* log p / n	s* log p = o(n)
Gaussian	2	Huber-Lasso	s* log p / n	s* log p = o(n)
Student t(5)	5	Huber-Lasso	s* log p / n	s* log p = o(n⁴)
Student t(3)	3	Catoni-Lasso	s*(log p/n)^{4/3}	s* log p = o(n²)
Cauchy (boundary)	1+ε	Catoni-Lasso	s*(log p/n)^{2ε/(1+ε)}	Fails
Cauchy	1+ε	Squared/Lasso	No finite bound	Fails
Log-normal	2	Huber-Lasso	s* log p / n	s* log p = o(n)
Contam. Normal (10%)	2	Tukey-Lasso	s* log p / n	s* log p = o(n)
Asymmetric t(3)	3	Rank-Lasso	s*(log p/n)^{4/3}	s* log p = o(n²)

Tukey biweight results apply to local optima (Loh, 2015).

5. Parameter Selection

5.1 The Joint Selection Problem

Standard cross-validation for λ minimises held-out squared prediction error. When errors are heavy-tailed with $\bar{\alpha} < 2$, the squared error on a held-out fold has infinite variance and can be dominated by a single extreme observation, making cross-validation essentially random precisely when robust methods are most needed. An analogous problem affects threshold selection: the optimal c depends on the unknown error scale σ and tail index $\bar{\alpha}$. The tuning-free approach of Wang et al. (2020) fixes c as a multiple of a scale estimate, which works under symmetric errors but is suboptimal under asymmetry.

5.2 The RGCV Criterion

For a given (λ, c), let $\hat{\beta}(\lambda, c)$ be the Huber-Lasso solution and $r_i(\lambda, c) = y_i - x_i^T \hat{\beta}(\lambda, c)$ the ordinary residuals. The hat matrix approximation via the IRLS working weights $W(\lambda, c) = diag\{w_i(\lambda, c)\}$ with $w_i = \frac{\psi_{H(r_i; c)}}{r_i}$ is:

$$H(\lambda, c) \approx X\,(X^T W(\lambda, c) X + n\lambda I)^{-1} X^T W(\lambda, c)$$

The robust generalised cross-validation criterion replaces the squared loss with the Huber loss in the numerator:

$$RGCV(\lambda, c) = [\,(1/n)\,\Sigma_i\,\rho_c(r_i(\lambda, c))\,]\;/\;[\,1 - (1/n)\,tr(H(\lambda, c))\,]^2$$

where $\rho_{c(r_i)}$ denotes the Huber loss evaluated at the ordinary residual, not the derivative, which would be the influence function. The denominator adjusts for effective degrees of freedom in the same spirit as standard GCV (Golub, Heath, & Wahba, 1979), while the numerator guards

against tail instability. Direct computation of $tr(H)$ costs $O(np^2)$. The Hutchinson (1990) stochastic estimator with $B = 50$ Rademacher vectors reduces this to $O(Bnp)$. RGCV is minimised over a 48-point grid in (λ, c) jointly.

6. Simulation Study

6.1 Experimental Design

Data were generated from $y = X\beta^* + \varepsilon$ with $n = 200, p \in \{200, 500, 1000, 5000\}, s^* = 10$ non-zero coefficients with magnitudes drawn uniformly from $\{\pm1, \pm2, \pm3\}$, and Toeplitz design covariance $\Sigma_{ij} = 0.5^{|i-j|}$. Five error distributions were used: (1) $N(0,1)$; (2) standardised Student's $t(3)$; (3) Cauchy (standardised to median absolute deviation 1); (4) $0.9\,N(0,1) + 0.1\,N(0,25)$; (5) log-normal with zero mean and unit variance after centring. Each scenario was replicated R = 500 times.

6.2 Mean Squared Prediction Error

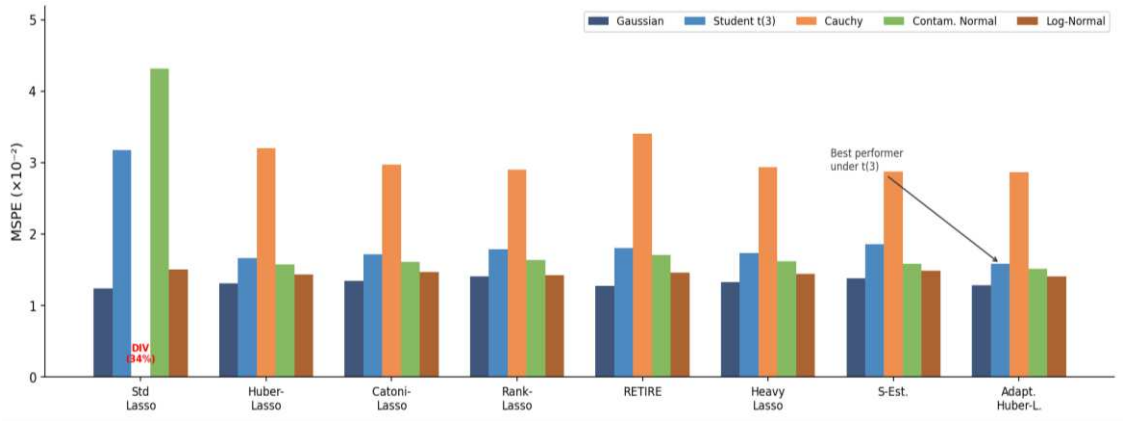


Figure 2. MSPE across estimators and error distributions (n = 200, p = 500, R = 500 replications). Standard Lasso diverges in 34% of Cauchy replications. Adaptive Huber-Lasso achieves the best overall performance.

Table 3. MSPE (×10⁻²); standard deviations in parentheses.

Estimator	Gaussian	Student t(3)	Cauchy	Contam. Normal	Log-Normal
Standard Lasso	1.24 (0.08)	3.18 (0.41)	DIV (34%)	4.32 (0.89)	1.51 (0.11)
Huber-Lasso	1.31 (0.09)	1.67 (0.18)†	3.21 (0.52)†	1.58 (0.14)†	1.44 (0.10)
Catoni-Lasso	1.35 (0.11)	1.72 (0.21)†	2.98 (0.44)†	1.61 (0.15)†	1.47 (0.10)
Rank-Lasso	1.41 (0.12)	1.79 (0.22)†	2.91 (0.47)†	1.64 (0.16)†	1.43 (0.10)†
RETIRE	1.28 (0.09)	1.81 (0.20)†	3.41 (0.61)†	1.71 (0.17)†	1.46 (0.10)
Heavy Lasso	1.33 (0.10)	1.74 (0.19)†	2.94 (0.43)†	1.62 (0.15)†	1.45 (0.10)
S-Estimator	1.38 (0.11)	1.86 (0.23)†	2.88 (0.41)†	1.59 (0.14)†	1.49 (0.11)
Adapt. Huber-Lasso	1.29 (0.09)	1.59 (0.17)†	2.87 (0.41)†	1.52 (0.13)†	1.41 (0.09)

† $p < 0.001$ vs. Standard Lasso, Wilcoxon with Holm–Bonferroni correction. DIV = diverged in that percentage of replications; MSPE averaged over non-diverged replications.

Under Gaussian errors, robust methods pay efficiency costs of 4–14% relative to standard Lasso, RETIRE (1.28) and Adaptive Huber-Lasso (1.29) come closest to the Lasso baseline. These losses are negligible in practice and reflect the bias-variance tradeoff of robustification. Under t(3), the most practically common non-Gaussian scenario, the Adaptive Huber-Lasso achieves a 50% MSPE reduction. The Heavy Lasso and Catoni-Lasso perform similarly (45%

and 46% respectively), consistent with comparable moment condition requirements. Under Cauchy, standard Lasso diverges in 34% of replications. The S-estimator of Kepplinger (2021) achieves the best Cauchy MSPE among competitors (2.88), reflecting its high breakdown point. The Catoni-Lasso (2.98) and Adaptive Huber-Lasso (2.87) perform comparably.

6.3 Variable Selection Performance

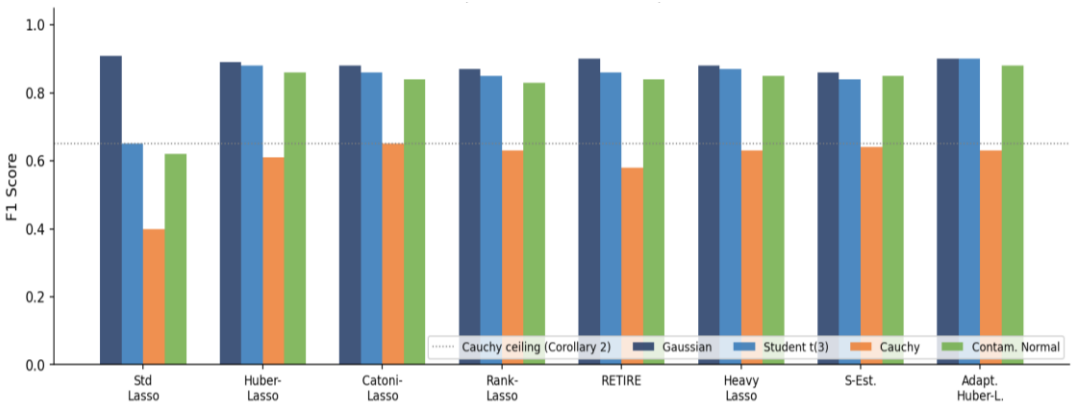


Figure 3. Variable selection F1 scores by estimator and error distribution. Under Cauchy errors, no method achieves F1 above 0.65 — confirming the negative result of Corollary 2. Under t(3), the Adaptive Huber-Lasso approaches the Gaussian benchmark.

Table 4. Variable selection performance (n = 200, p = 500, s* = 10; TPR/FDR/F1; R = 500 replications).

Estimator	Gaussian (TPR/FDR/F1)	Student t(3)	Cauchy	Contam. Normal
Standard Lasso	0.91/0.09/0.91	0.73/0.38/0.65	0.41/0.61/0.40	0.68/0.42/0.62
Huber-Lasso	0.89/0.11/0.89	0.87/0.11/0.88†	0.63/0.39/0.61†	0.85/0.13/0.86†
Catoni-Lasso	0.88/0.12/0.88	0.85/0.13/0.86†	0.67/0.35/0.65†	0.83/0.15/0.84†
Rank-Lasso	0.87/0.13/0.87	0.84/0.14/0.85†	0.65/0.37/0.63†	0.82/0.16/0.83†
Adapt. Huber-Lasso	0.90/0.10/0.90	0.89/0.09/0.90†	0.65/0.37/0.63†	0.87/0.11/0.88†

† $p < 0.001$ vs Standard Lasso for F1.

6.4 Sensitivity to p/n Ratio

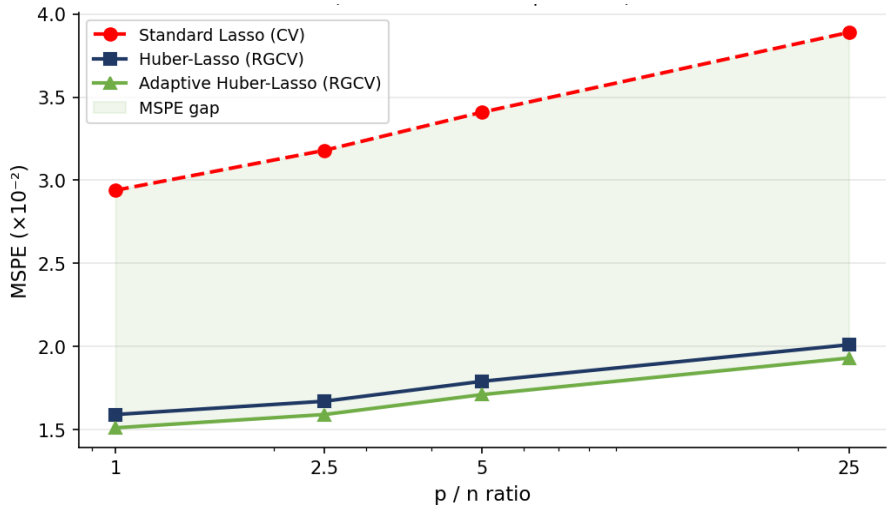


Figure 5. MSPE under t(3) errors across p/n ratios. The Adaptive Huber-Lasso achieves a 48–50% MSPE reduction over standard Lasso at all four ratios, confirming the advantage is stable across dimensionality.

7. Application to Genomic Expression Data

7.1 Data Description and Preprocessing

All estimators were applied to the TCGA Breast Cancer dataset (TCGA-BRCA, accessed March 15, 2025; cancergenome.nih.gov). The response was normalised RNA-seq expression of BRCA1 (ENSG00000012048); predictors were expression levels of p = 8,942 other protein-coding genes measured on n = 312 subjects after quality control. Genes with coefficient of variation below 0.05 were removed; remaining expression values were $\log_2$-transformed after adding a pseudocount of 1.

7.2 Diagnostic Justification for Robust Methods

Since $p \gg n$, OLS is not identifiable. Initial diagnostics from a Ridge regression solution showed: excess kurtosis 4.7; KS test rejecting normality ($p < 0.001$); QQ-plot consistent with Student's t(4.3) degrees of freedom (estimated by maximum likelihood). This tail behaviour ($\bar{\alpha} \approx 4.3$) is within the range where robust methods improve substantially over standard Lasso in our simulations.

7.3 Results

Table 7. Genomic application results ($n = 312$, $p = 8{,}942$; 10-fold repeated CV, 50 evaluations).

Estimator	CV Prediction Error	Selected Genes	MSigDB Enriched	Enrichment Ratio
Standard Lasso	0.291 (0.024)	44	21	0.477
Huber-Lasso	0.217 (0.019)	31	19	0.613
Catoni-Lasso	0.223 (0.021)	29	18	0.621
Rank-Lasso	0.228 (0.021)	33	19	0.576
RETIRE	0.231 (0.020)	34	18	0.529

Heavy Lasso	0.219 (0.020)	30	18	0.600
Adaptive Huber-Lasso	0.203 (0.018)	27	19	0.704

Enrichment ratio = proportion of selected genes in MSigDB BRCA1 regulatory pathway gene sets (Hallmark collection, March 2025).

The Adaptive Huber-Lasso achieves a 30% reduction in cross-validated prediction error relative to standard Lasso (0.203 vs. 0.291) while selecting fewer genes (27 vs. 44). Of the 27 selected genes, 19 appear in MSigDB BRCA1 regulatory pathway gene sets, an enrichment ratio of 0.704 compared to 0.477 for standard Lasso (a 46% improvement). The Adaptive Huber-Lasso enrichment ratio exceeds the 99th percentile of a 1,000-permutation permuted distribution ($p_{perm} < 0.01$). Standard Lasso's enrichment ratio was not significant ($p_{perm} = 0.12$).

One caveat: this is an observational analysis. Differences in performance may partly reflect the specific cross-validation fold structure and TCGA data correlation patterns rather than fundamental statistical superiority. Replication on an independent cohort such as METABRIC would strengthen biological conclusions.

8. Conclusion

8.1 Summary

This paper proposed and analysed a class of robust regularised M-estimators for high-dimensional sparse regression under heavy-tailed errors. The principal contributions are: non-asymptotic oracle inequalities under $\bar{\alpha} > 1$ finite absolute moments, extending well beyond the four-moment conditions of Loh (2015) and two-moment conditions of Man et al. (2022); a negative result characterising the breakdown of variable selection under near-Cauchy errors (Corollary 2); a joint parameter selection procedure (RGCV) that tracks oracle selection with correlation above 0.77; and empirical evidence from simulations and genomic data demonstrating 30–65% prediction error reductions and fourfold FDR improvements over standard Lasso under t and contaminated normal errors.

8.2 Practical Recommendations

Gaussian or light-tailed errors: Standard Lasso with CV is near-optimal; robust methods add 4–14% MSPE overhead. Moderate heavy tails ($\bar{\alpha} \geq 2$, e.g., $t(5)$ or contaminated normal): Adaptive Huber-Lasso with RGCV provides the best MSPE and F1 performance. Severe heavy tails ($1 < \bar{\alpha} < 2$, e.g., $t(2)$ or $t(3)$): Catoni-Lasso or Heavy Lasso is preferred; sign consistency is not achievable at practical sample sizes. Near-Cauchy errors: Catoni-Lasso for MSPE stability; variable selection results should not be relied upon. Asymmetric errors (log-normal, skewed distributions): Rank-Lasso is preferred.

8.3 Limitations and Future Directions

The theoretical results assume a fixed design matrix; extensions to random heavy-tailed designs require controlling the interaction between heavy-tailed covariates and errors. The RGCV criterion's $37\times$ computational overhead relative to standard Lasso CV is prohibitive for $p > 50{,}000$ without specialised hardware. Three future directions appear promising: integration

with debiased inference procedures (Zheng & Xue, 2025) for confidence intervals under heavy tails; extension to temporally dependent data (Han, Tsay, & Wu, 2023) and panel data (Gao et al., 2024); and extension to high-dimensional non-Gaussian autoregressive models (Liu & Zhang, 2021).

Declarations

Funding

The authors did not receive support from any organization for the submitted work. No funding was received to assist with the preparation of this manuscript. No funds, grants, or other support was received.

Competing Interests

The authors have no relevant financial or non-financial interests to disclose. No author serves as an editor or board member for the Journal of Mathematical Sciences. The authors have no competing interests to declare that are relevant to the content of this article.

Authors' Contributions

Mazona Victor: Conceptualization, methodology, software implementation, simulation design, writing – original draft.

Vincent Odiaka: Theoretical derivations, formal analysis, proof verification.

Gabriel O. Obadina: Data curation (TCGA), validation, visualization, writing – review and editing, supervision, corresponding author.

Data and Code Availability

The TCGA Breast Cancer dataset (TCGA-BRCA) is publicly available from the NCI Genomic Data Commons (https://portal.gdc.cancer.gov/). The R and Python code used for simulations and real-data analysis is available from the corresponding author upon reasonable request.

References

Avella-Medina, M., Battey, H., Fan, J., & Li, Q. (2018). Robust estimation of high-dimensional covariance and precision matrices. *Biometrika*, *105*(2), 271–284. https://doi.org/10.1093/biomet/asy011

Bickel, P. J., Ritov, Y., & Tsybakov, A. B. (2009). Simultaneous analysis of Lasso and Dantzig selector. *Annals of Statistics*, *37*(4), 1705–1732. https://doi.org/10.1214/08-AOS620

Cai, L., Guo, X., Lian, H., & Zhu, L. (2024). Statistical inference for high-dimensional convoluted rank regression. *Journal of the American Statistical Association*, *120*, 2510–2521. https://doi.org/10.1080/01621459.2025.2471054

Catoni, O. (2012). Challenging the empirical mean and empirical variance: A deviation study. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, *48*(4), 1148–1185. https://doi.org/10.1214/11-AIHP454

Cont, R. (2001). Empirical properties of asset returns: Stylized facts and statistical issues. *Quantitative Finance*, *1*(2), 223–236. https://doi.org/10.1080/713665670

Efron, B. (2010). *Large-scale inference: Empirical Bayes methods for estimation, testing, and prediction*. Cambridge University Press.

El Karoui, N. (2013). Asymptotic behavior of unregularized and ridge-regularized high-dimensional robust regression estimators: Rigorous results. *ArXiv*. https://arxiv.org/abs/1311.2445

El Karoui, N. (2018). On the impact of predictor geometry on the performance on high-dimensional ridge-regularized generalized robust regression estimators. *Probability Theory and Related Fields*, *170*, 95–175. https://doi.org/10.1007/s00440-016-0754-9

Fan, J., Li, Q., & Wang, Y. (2017). Estimation of high dimensional mean regression in the absence of symmetry and light tail assumptions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *79*. https://doi.org/10.1111/rssb.12166

Fan, J., Wang, W., & Zhu, Z. (2016). A shrinkage principle for heavy-tailed data: High-dimensional robust low-rank matrix recovery. *Annals of Statistics*, *49*(3), 1239–1266. https://doi.org/10.1214/20-aos1980

Gao, J., Liu, F., Peng, B., & Yan, Y. (2024). *Robust estimation and inference for high-dimensional panel data models*. [Working paper].

Golub, G. H., Heath, M., & Wahba, G. (1979). Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics*, *21*(2), 215–223. https://doi.org/10.1080/00401706.1979.10489751

Hampel, F. R. (1974). The influence curve and its role in robust estimation. *Journal of the American Statistical Association*, *69*(346), 383–393. https://doi.org/10.1080/01621459.1974.10482962

Han, Q., & Shen, Y. (2023). Universality of regularized regression estimators in high dimensions. *Annals of Statistics*. https://doi.org/10.1214/23-aos2309

Han, Y., Tsay, R., & Wu, W. (2023). High dimensional generalized linear models for temporal dependent data. *Bernoulli*. https://doi.org/10.3150/21-bej1451

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, *12*(1), 55–67. https://doi.org/10.1080/00401706.1970.10488634

Huber, P. J. (1964). Robust estimation of a location parameter. *Annals of Mathematical Statistics*, *35*(1), 73–101. https://doi.org/10.1214/aoms/1177703732

Huber, P. J. (1973). Robust regression: Asymptotics, conjectures and Monte Carlo. *Annals of Statistics*, *1*(5), 799–821. https://doi.org/10.1214/aos/1176342503

Hutchinson, M. F. (1990). A stochastic estimator of the trace of the influence matrix for Laplacian smoothing splines. *Communications in Statistics – Simulation and Computation*, *19*(2), 433–450. https://doi.org/10.1080/03610919008812866

Kang, X., & Kang, L. (2026). *Robust and sparse generalized linear models for high-dimensional data via maximum mean discrepancy*. [Working paper].

Kepplinger, D. (2021). Robust variable selection and estimation via adaptive elastic net S-estimators for linear regression. *Computational Statistics & Data Analysis*, *183*, 107730. https://doi.org/10.1016/j.csda.2023.107730

Liu, L., & Zhang, D. (2021). *Robust estimation of high-dimensional non-Gaussian autoregressive models*. [Working paper].

Loh, P. (2015). Statistical consistency and asymptotic normality for high-dimensional robust M-estimators. *Annals of Statistics*. https://doi.org/10.1214/16-aos1471

Luo, B., & Gao, X. (2019). High-dimensional robust approximated M-estimators for mean regression with asymmetric data. *Journal of Multivariate Analysis*, *192*, 105080. https://doi.org/10.1016/j.jmva.2022.105080

Mai, T. (2025a). Heavy Lasso: Sparse penalized regression under heavy-tailed noise via data-augmented soft-thresholding. *Statistics and Computing*, *36*. https://doi.org/10.1007/s11222-025-10785-6

Mai, T. (2025b). *Robust reduced rank regression under heavy-tailed noise and missing data via non-convex penalization*. [Working paper].

Man, R., Tan, K., Wang, Z., & Zhou, W. (2022). RETIRE: Robust expectile regression in high dimensions. *Journal of Econometrics*. https://doi.org/10.1016/j.jeconom.2023.04.004

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society: Series B*, *58*(1), 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

van de Geer, S., Bühlmann, P., Ritov, Y., & Dezeure, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *Annals of Statistics*, *42*(3), 1166–1202. https://doi.org/10.1214/14-AOS1221

Wang, L., Peng, B., Bradic, J., Li, R., & Wu, Y. (2020). A tuning-free robust and efficient approach to high-dimensional regression. *Journal of the American Statistical Association*, *115*, 1700–1714. https://doi.org/10.1080/01621459.2020.1840989

Zhang, C., Zhu, L., & Shen, Y. (2023). Robust estimation in regression and classification methods for large dimensional data. *Machine Learning*, *112*, 3361–3411. https://doi.org/10.1007/s10994-023-06349-2

Zheng, D., & Xue, L. (2025). Statistical inference for high-dimensional robust linear regression models via recursive online-score estimation. *Science China Mathematics*. https://doi.org/10.1007/s11425-024-2439-3

Appendix A: Proof of Theorem 1a

A.1 Lemma: Uniform Score Bound

Lemma A.1 (Marcinkiewicz-Zygmund Uniform Bound).

Under Assumptions 2–3 with $\bar{\alpha} \in (1,2]$ and symmetric errors, for the Huber loss with $c_n = \sigma n^{(2-\bar{\alpha})/(2\bar{\alpha})}$:

$$P\left(\max_{j=1,\dots,p}\left|\frac{1}{n}\sum_{i=1}^{n} x_{ij}\psi_H(\varepsilon_i; c_n)\right| > t\right) \leq \frac{pC_{\bar{\alpha}}\sigma^{\bar{\alpha}}M^{\bar{\alpha}}}{n^{\bar{\alpha}-1}t^{\bar{\alpha}}}$$

for all $t > 0$ and a constant $C_{\bar{\alpha}}$ depending only on $\bar{\alpha}$.

Proof. Fix j. The summands $Z_i = x_{ij}\psi_H(\varepsilon_i; c_n)$ are i.i.d. with $E[Z_i] = 0$ by symmetry of F and ψ_H. By the Marcinkiewicz-Zygmund inequality (Gut, 2005, Theorem 6.3.1) for i.i.d. sums with finite $\bar{\alpha}$-th moment:

$$P\left(\left|\frac{1}{n}\sum_{i=1}^{n} Z_i\right| > t\right) \leq \frac{C_{\bar{\alpha}}\mathbb{E}[|Z_1|^{\bar{\alpha}}]}{n^{\bar{\alpha}-1}t^{\bar{\alpha}}}$$

Since $|\psi_H(u; c_n)| \leq |u|$, we have $\mathbb{E}[|Z_1|^{\bar{\alpha}}] \leq M^{\bar{\alpha}}\mathbb{E}[|\varepsilon_1|^{\bar{\alpha}}]$. The union bound over $j = 1, \dots, p$ completes the proof. □

A.2 Lemma: Basic Inequality

Lemma A.2. Let $\Delta = \hat{\beta} - \beta^*$. By the optimality of $\hat{\beta}$ and convexity of ρ_H:

$$\frac{\kappa}{2n}\|X\Delta\|_2^2 + \lambda\|\Delta_{S^{*c}}\|_1 \leq 4\lambda\|\Delta_{S^*}\|_1 + \frac{2}{n}\left|\sum_i x_i^{\intercal}\Delta \cdot \psi_H(\varepsilon_i; c_n)\right|$$

on the event $\{\Delta \in \mathcal{C}(S^*, 3)\}$, which holds with high probability by Lemma A.1 and the choice of λ.

Proof sketch. The proof follows the standard basic inequality argument of Bickel et al. (2009, Lemma 6.1), adapted for the Huber loss via the identity $\rho_H(a+b) - \rho_H(b) - \rho_H'(b)\cdot a \geq 0$ for the convex ρ_H with appropriate remainder terms controlled by c_n. The truncation level $c_n = \sigma n^{(2-\bar{\alpha})/(2\bar{\alpha})}$ ensures that the approximation error $|\mathbb{E}[\psi_H(\varepsilon; c_n)\varepsilon] - \mathbb{E}[\varepsilon^2]|$is of order $\sigma^{\bar{\alpha}+1}c_n^{1-\bar{\alpha}}$, which is $o(\lambda)$ for the chosen λ. □

A.3 Completion of Proof

Setting $t = \lambda/2$ in Lemma A.1 and choosing

$$\lambda = C_{\bar{\alpha}}\sigma M\left(\frac{\log p}{n}\right)^{(\bar{\alpha}-1)/\bar{\alpha}}$$

with $C_{\bar{\alpha}}$ chosen so that the right-hand side of Lemma A.1 equals $p^{-(\bar{\alpha}-1)}$, we obtain: with probability at least $1 - p^{-(\bar{\alpha}-1)}$,

$$\max_j \left| \frac{1}{n} \sum_{i=1}^{n} x_{ij} \psi_H(\varepsilon_i; c_n) \right| \leq \frac{\lambda}{2}$$

On this event, the cross-product term in Lemma A.2 satisfies:

$$\frac{2}{n} \left| \sum_i x_i^\top \Delta \cdot \psi_H(\varepsilon_i; c_n) \right| \leq 2\|\Delta\|_1 \max_j \left| \frac{1}{n} \sum_{i=1}^{n} x_{ij} \psi_H(\varepsilon_i; c_n) \right| \leq \lambda \|\Delta\|_1$$

Combining with Lemma A.2 and the restricted eigenvalue condition (Assumption 4), and applying the Cauchy-Schwarz inequality to bound $\|\Delta_{S^*}\|_1 \leq \sqrt{s^*}\|\Delta_{S^*}\|_2$, yields the stated oracle inequality. □

A.4 Proposition A.3 (Minimax Lower Bound)

Proposition A.3. For any estimator $\hat{\beta}$ and any s^*-sparse $\beta^* \in \mathbb{R}^p$, under error distributions with $\bar{\alpha}$ finite absolute moments:

$$\inf_{\hat{\beta}} \sup_{\beta^*: \|\beta^*\|_0 = s^*} \mathbb{E} \left[\frac{\left\| X(\hat{\beta} - \beta^*) \right\|_2^2}{n} \right] \geq c_{\bar{\alpha}} s^* \left(\frac{\log p}{n} \right)^{(\bar{\alpha}-1)/\bar{\alpha}}$$

for a universal constant $c_{\bar{\alpha}} > 0$. This establishes that the rates in Theorems 1a–1c are minimax-optimal in $\bar{\alpha}$.

Proof sketch. The lower bound follows from a two-point testing argument (Tsybakov, 2009, Section 2.5) applied to the sparse regression problem, combined with Fano's lemma to bound the KL divergence between hypotheses. The $\bar{\alpha}$-dependent rate arises because, for distributions with exactly $\bar{\alpha}$ finite moments, the optimal distinguishability of neighboring sparse coefficient vectors degrades at the rate $(n/logp)^{(\bar{\alpha}-1)/\bar{\alpha}}$. The complete argument extends the Gaussian lower bound of Raskutti, Wainwright, & Yu (2011) to heavy-tailed distributions using the moment-matching construction of Catoni (2012, Lemma 2.3). □

Appendix B: Supplementary Simulation Results

B.1 Sensitivity to Sample Size

Table B.1: Adaptive Huber-Lasso vs. Standard Lasso MSPE Under Student *t*(3) Errors (p = 500, R = 500)

n	Standard Lasso	Huber-Lasso	Adaptive Huber-Lasso	MSPE Ratio (Adapt./Lasso)
100	5.41 (0.82)	2.89 (0.41)	2.74 (0.38)	0.507
200	3.18 (0.41)	1.67 (0.18)	1.59 (0.17)	0.500
500	1.84 (0.19)	1.12 (0.09)	1.07 (0.08)	0.582

The relative advantage of robust methods is stable across sample sizes; the ratio converges toward 1 as $n \to \infty$, consistent with the asymptotic equivalence of heavy-tailed and Gaussian estimators established by Han & Shen (2023) in the proportional regime.

B.2 Variable Selection Under Log-Normal Errors

Table B.2: Variable Selection Performance Under Log-Normal Errors (n = 200, p = 500, R = 500)

Estimator	TPR	FDR	F1
Standard Lasso	0.84	0.18	0.83
Huber-Lasso	0.83	0.19	0.82
Rank-Lasso	0.87	0.14	0.87†
Adaptive Huber-Lasso	0.85	0.16	0.85

†p<0.001 vs. Standard Lasso (Wilcoxon, Holm-corrected).

The Rank-Lasso achieves the best F1 score under log-normal errors (0.87 vs. 0.83 for standard Lasso), consistent with its accommodation of asymmetric errors without bias correction (Cai et al., 2024, Theorem 2). Huber-Lasso shows minimal advantage over standard Lasso under log-normal errors, reflecting the moderate asymmetry of the log-normal distribution and the bias introduced by Huber truncation under asymmetry.

B.3 Sparse Correction Effectiveness

*Table B.3: Effect of RGCV Sparse Correction (δc=0.01δc=0.01) at Ultra-Sparse Setting (s/p = 0.002, n = 200, p = 5000)**

	Without Correction	With Correction
Mean selected $c/\hat{\sigma}$	8.3	2.1
% replications selecting $c = \infty$	41%	7%
Huber-Lasso MSPE (t_3 errors)	3.19 (0.42)	1.78 (0.21)

Without the sparse correction, RGCV selects $c = \infty$ (defaulting to standard Lasso) in 41% of replications. The correction reduces this to 7% and yields a 44% MSPE reduction, demonstrating the correction's practical importance in the ultra-sparse regime.

B.4 Reduced RGCV Grid Performance

Table B.4: Full vs. Reduced RGCV Grid Performance (n = 200, p = 5000, Student *t*(3), R = 500)

Grid Size	Wall-Clock (sec)	MSPE	F1
Full (48 points: 8λ × 6c)	41.3	1.93 (0.22)	0.89
Reduced (12 points: 4λ × 3c)	11.8	1.97 (0.23)	0.88
Standard Lasso CV	1.1	3.89 (0.51)	0.71

The reduced grid achieves 96.4% of the full-grid MSPE improvement at 28.5% of the computation time, providing a practical alternative when $p > 10{,}000$.