

Statistica Sinica Preprint No: SS-2026-0193	
Title	Robust Classification with Feature Selection Using Non-Gaussian Data
Manuscript ID	SS-2026-0193
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202026.0193
Complete List of Authors	Xiuchuan Liu and Xianzheng Huang
Corresponding Authors	Xiuchuan Liu
E-mails	xiuchuan@email.sc.edu
Notice: Accepted author version.	

ROBUST CLASSIFICATION WITH FEATURE SELECTION USING NON-GAUSSIAN DATA

Xiuchuan Liu and Xianzheng Huang

University of South Carolina

Abstract: When confronted with features subject to outliers or asymmetry that cause deviations from Gaussianity, classification remains a challenging task in statistical learning. We develop novel classification methods robust to Gaussian assumptions by adopting a flexible feature distribution, which is further exploited to construct transformed features used in regularized logistic regression to achieve reliable binary classification and perform feature selection. We investigate theoretical properties of the proposed classifier under realistic conditions that allow model misspecification, and establish consistency in separating signal and noise features. Superior robustness and accuracy in classification of the proposed methods compared to many existing methods are demonstrated in simulation experiments, where we generate synthetic data representing a wide range of distributional characteristics of features and data patterns. We showcase the rich interpretable information beyond class labels that the proposed method extracts from data in three real-life case studies.

Key words and phrases: heavy-tailed distribution, LASSO, likelihood ratio, logistic regression, mode.

1. Introduction

Classification is a classic and fundamental task, and an effective classifier is expected to achieve more than an accurate class label. In high-stakes domains such as medicine and finance, trustworthy decision-making and scientific inquiry demand a classifier that embodies three key virtues: (i) robustness, to handle heavy-tailed and skewed features ubiquitous in real-world data; (ii) interpretability, providing a transparent mechanism to explain and validate both the model and its predictions (Rudin, 2019); and (iii) parsimony, the capacity to capture informative features among potentially many within a simple and generalizable model. Existing classification methods, reviewed comprehensively in James et al. (2023), mostly fall into two mainstreams that typically excel in at most two of these objectives.

The first mainstream is based on penalized generalized linear models, such as L_1 -regularized logistic regression. These methods are widely employed for their excellent interpretability and parsimony (Tibshirani, 1996; Friedman et al., 2010). However, maximum likelihood estimators in this setting suffer from unbounded influence functions (Hampel, 1986). Consequently, the framework can be overly sensitive to outliers. The linear predictor inherently restricts these models from learning nonlinear decision boundaries, which are ubiquitous in real-world applications. Although

outlier-robust loss functions can mitigate the effect of anomalies, such as the weighted least absolute deviation loss, the adaptive Huber loss, and the density power divergence (Arslan, 2012; Sun et al., 2020; Basu et al., 2024), they still fail to address the nonlinearity of the underlying decision boundaries.

The second mainstream emerges in pursuit of flexibility. Methods along this vein include ensemble methods such as random forest (Breiman, 2001) and the light gradient boosting machine (Ke et al., 2017), parametric methods based on mixture models such as Gaussian mixture models (Scrucca et al., 2023), and nonparametric approaches such as the mode-based kernel density classifier (Xiong et al., 2025). While these methods demonstrate robustness in handling non-Gaussian features and nonlinear classification boundaries, they typically sacrifice interpretability. Ensemble methods are inherently black-box and rely on post-hoc attributions (Lundberg and Lee, 2017; Rudin, 2019), whose feature importance scores carry no information on directionality or magnitude, making it impossible to reconstruct a classifier’s decision process from these scores alone. Methods based on nonparametric or mixture models generally lack explicit mechanisms for achieving parsimony. Consequently, these methods still fall short of simultaneously delivering all three virtues: robustness, interpretability, and parsimony.

In this paper, we propose a framework of classification built upon ro-

bust likelihood ratios that addresses this trilemma in two steps. In the first generative step, we flexibly model each feature and capture its information via the ratio of its class-specific likelihood functions. In the second discriminative step, we exploit these log-likelihood ratios as transformed features and feed them as covariates into an L_1 -penalized logistic regression model for the class label. These are elaborated in Section 2, and Section 3 presents asymptotic analyses providing theoretical guarantees for the new classifier. Section 4 reports a comparative study on data simulated from a variety of data-generating mechanisms, and Section 5 analyzes three datasets from biomedical studies and astrophysics research. Section 6 considers extensions of the proposed method to accommodate various types of features modeled by different distributions, and Section 7 summarizes our study and discusses the future research agenda. The source R code and reproducible analysis are available on the first author’s GitHub site, <https://github.com/xiuchuan18/robust-classification>.

2. Methodology

2.1 A flexible feature distribution

Consider n independent observations $\{(\mathbf{X}_i, Y_i)\}_{i=1}^n$, where $Y_i \in \{0, 1\}$ is the class label of the i -th subject and $\mathbf{X}_i = (X_{i1}, \dots, X_{ip})^\top \in \mathbb{R}^p$ collects its p

2.1 A flexible feature distribution

continuous features. We suppress the index i when referring to a subject's class label and features generically.

For a generative model of a feature robust to violations of Gaussianity, we adopt the two-piece scale distribution (TPSC, Fernández and Steel, 1998; Liu et al., 2024) indexed by four parameters in $\Psi = (\tau, w, \sigma, \delta)^\top$ and specified by the probability density function (pdf) below,

$$f_{\text{TPSC}}(x|\Psi) = w \frac{2}{\sigma_1} g\left(\frac{x - \tau}{\sigma_1}; \delta\right) \mathbb{1}(x < \tau) + (1 - w) \frac{2}{\sigma_2} g\left(\frac{x - \tau}{\sigma_2}; \delta\right) \mathbb{1}(x \geq \tau), \quad (2.1)$$

where $g(t; \delta) = (1 + t^2/\delta)^{-(\delta+1)/2} \Gamma((\delta+1)/2) / \{\sqrt{\pi\delta} \Gamma(\delta/2)\}$ is the pdf of Student's t with δ degrees of freedom, $\mathbb{1}(\cdot)$ is an indicator function, $\sigma_1 = \sigma \sqrt{w/(1-w)}$, $\sigma_2 = \sigma \sqrt{(1-w)/w}$, $\tau \in \mathbb{R}$ is the unique mode providing a robust anchor for central tendency, $w \in (0, 1)$ is the weight parameter modeling skewness, and $\sigma > 0$ is the scale parameter. Setting $w = 0.5$ gives a symmetric distribution centered at τ , and letting $\delta \rightarrow \infty$ gives Gaussian-like light tails. Denoting by $f_{jk}(\cdot)$ the pdf of feature X_j under class k , we assume $f_{jk}(x) = f_{\text{TPSC}}(x|\Psi_{jk})$, which allows the j -th feature to be light-tailed or heavy-tailed, symmetric or asymmetric within class k , for $j \in [p] = \{1, \dots, p\}$ and $k \in \{0, 1\}$.

Density (2.1) joins the left and the right half of two shifted-and-scaled t densities at their common mode. The more general TPSC in Fernández and

2.2 The log-likelihood ratio transformation

Steel (1998) allows any unimodal symmetric densities, not limited to the t densities, in its formulation. The construction models the skewness and tail behaviors of non-Gaussian features flexibly. The location (mode) parameter τ provides a more faithful representation of the typical data region compared with the mean when a distribution is asymmetric or heavy-tailed. Next, we use this generative model to capture features' characteristics such as mode, skewness, and tail behavior. Incorporating this information allows downstream classifiers to establish more informative decision boundaries.

2.2 The log-likelihood ratio transformation

By Bayesian decision theory, the optimal classification boundary is specified by $P(Y = 1|\mathbf{X}) = P(Y = 0|\mathbf{X})$ (Duda et al., 2012), or equivalently $f(\mathbf{X}|Y = 1)/f(\mathbf{X}|Y = 0) = P(Y = 0)/P(Y = 1)$, with the left-hand side being a ratio typically nonlinear in $\mathbf{X}$. Under the naive Bayes assumption stating that features are independent given the class label, the logarithm of this ratio decomposes into $\sum_{j=1}^p \log(f_{j1}(X_j)/f_{j0}(X_j)) = \sum_{j=1}^p T_j(X_j)$, with $T_j(X_j) = \log(f_{j1}(X_j)/f_{j0}(X_j))$. Therefore, the optimal Bayes classification boundary under this assumption is $\sum_{j=1}^p T_j(X_j) = \log(P(Y = 0)/P(Y = 1))$, which is linear in the transformed features in $\mathbf{T} = (T_1(X_1), \dots, T_p(X_p))^T$.

Even without the naive Bayes assumption that supports the optimality

2.3 Classification based on transformed features

of this decision boundary, the transformed features are inherently interpretable. Firstly, unlike features in $\mathbf{X}$, which can be in different units, each transformed feature is unitless, with a larger $T_j(X_j)$ indicating more evidence from the j -th feature supporting $Y = 1$. Secondly, $\mathbf{T}$ carries a natural information-theoretic interpretation. Using the definition of $T_j(X_j)$, for $j \in [p]$, one can show that $E\{T_j(X_j)|Y = 1\} - E\{T_j(X_j)|Y = 0\} = D_{\text{KL}}(f_{j1} \parallel f_{j0}) + D_{\text{KL}}(f_{j0} \parallel f_{j1})$, where $D_{\text{KL}}(a \parallel b)$ denotes the Kullback-Leibler divergence of a from b . With the right-hand side of the above identity non-negative and symmetric in (f_{j1}, f_{j0}) , the expected difference on the left-hand side encapsulates the overall discriminative information between the two classes. This non-negative expected difference is zero if and only if the j -th feature is non-informative, i.e., $f_{j1} \equiv f_{j0}$. In essence, the mapping from $\mathbf{X}$ to $\mathbf{T}$ converts implicit distributional differences into explicit mean shifts in the transformed feature space.

2.3 Classification based on transformed features

Because $\mathbf{T}$ depends on unknown parameters in $\{(\Psi_{j0}, \Psi_{j1}), j \in [p]\}$, the features we ultimately use are $\hat{\mathbf{T}} = (\hat{T}_1(X_1), \dots, \hat{T}_p(X_p))^\top$, in which $\hat{T}_j(X_j) = \log(f_{\text{TPSC}}(X_j|\hat{\Psi}_{j1})/f_{\text{TPSC}}(X_j|\hat{\Psi}_{j0}))$, for $j \in [p]$, and $\{(\hat{\Psi}_{j0}, \hat{\Psi}_{j1}), j \in [p]\}$ are the maximum likelihood estimators (MLE) of the unknown parameters.

2.3 Classification based on transformed features

Adapting the optimal Bayes classifier under the naive Bayes assumption gives a simple classifier that assigns a new observation $\mathbf{x} = (x_1, \dots, x_p)^\top$ to class 1 if $\sum_{j=1}^p \hat{T}_j(x_j) > 0$. To highlight its robustness to non-Gaussian features, we name this method the robust log likelihood ratio classifier, or RoLLR for short. By signifying the mode, the use of TPSC makes RoLLR a parametric analogue to the mode-based kernel classifier (MOKE) in Xiong et al. (2025), which relies on nonparametric mode-based density estimates for robustness against skewness and heavy tails.

Like MOKE in its original form, RoLLR is a component-wise classifier that puts equal weight on all transformed features, which is well justified under the naive Bayes assumption. To let features contribute differently, with the contributions learnt from data, we regress Y on $\hat{\mathbf{T}}$ according to the logistic regression model,

$$P(Y = 1 | \hat{\mathbf{T}}) = \frac{1}{1 + \exp(-\beta_0 - \sum_{j=1}^p \beta_j \hat{T}_j(X_j))}, \quad (2.2)$$

where entries of $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)^\top$ reflect different contributions of the p features. To extract informative features, we estimate $\boldsymbol{\theta} = (\beta_0, \boldsymbol{\beta})$ via an L_1 -penalized MLE, $\hat{\boldsymbol{\theta}} = (\hat{\beta}_0, \hat{\boldsymbol{\beta}})$, that minimizes the negative log-likelihood,

$$-\frac{1}{n} \sum_{i=1}^n \left\{ Y_i(\beta_0 + \hat{\mathbf{T}}_i^\top \boldsymbol{\beta}) - \log(1 + \exp(\beta_0 + \hat{\mathbf{T}}_i^\top \boldsymbol{\beta})) \right\} + \lambda_n \|\boldsymbol{\beta}\|_1, \quad (2.3)$$

where $\hat{\mathbf{T}}_i = (\hat{T}_1(X_{i1}), \dots, \hat{T}_p(X_{ip}))^\top$, and $\lambda_n > 0$ controls the penalty, which

tends to zero as $n \rightarrow \infty$.

In contrast to the hard classifier RoLLR, we now have a soft classifier assigning $\mathbf{x}$ the probability $1/\{1+\exp(-\hat{\beta}_0-\sum_{j=1}^p\hat{\beta}_j\hat{T}_j(x_j))\}$ of belonging to class 1, and conventionally assigns the label 1 if this value exceeds 0.5. The L_1 penalty facilitates feature selection by pushing regression coefficients associated with non-informative features to zero, as shown in the next section. We call this classification method the sparse (or soft version of) RoLLR, or S-RoLLR for short.

3. Theoretical Properties

We present in this section properties of S-RoLLR as the sample size $n \rightarrow \infty$ while p is fixed, focusing on the risk of the classifier and its ability to filter out non-informative/noise features from a fixed number of candidate features. Henceforth, denote by $\mathcal{I} = \{j \in [p] : X_j \not\perp Y\}$ the signal set and by $\mathcal{I}_N = [p] \setminus \mathcal{I}$ the noise set, and define $s = |\mathcal{I}|$ as the cardinality of $\mathcal{I}$.

3.1 Risk analysis under correct model specification

Let $\pi^*(\mathbf{X}) = P(Y = 1|\mathbf{X})$, then $\eta^*(\mathbf{X}) = \log(\pi^*(\mathbf{X})/\{1 - \pi^*(\mathbf{X})\})$ is the true logarithm of odds, and $g^*(\mathbf{X}) = \mathbb{1}(\eta^*(\mathbf{X}) > 0)$ is the Bayes classifier achieving the minimal Bayes risk $R^* = E\{\min(\pi^*(\mathbf{X}), 1 - \pi^*(\mathbf{X}))\}$.

3.1 Risk analysis under correct model specification

S-RoLLR estimates $\eta^*(\mathbf{X})$ via $\hat{\eta}(\mathbf{X}) = \hat{\beta}_0 + \sum_{j=1}^p \hat{\beta}_j \hat{T}_j(X_j)$, and uses the classifier $\hat{g}_n(\mathbf{X}) = \mathbb{1}(\hat{\eta}(\mathbf{X}) > 0)$ with risk $R(\hat{g}_n) = P(\hat{g}_n(\mathbf{X}) \neq Y)$. Theorem 1 suggests that S-RoLLR is asymptotically optimal under the following conditions.

- (A1) The class-specific distribution of X_j is a TPSC, with the true parameter vector Ψ_{jk}^* in (2.1) lying in the interior of a compact parameter space, for $j \in [p]$ and $k \in \{0, 1\}$.
- (A2) $X_1, \dots, X_p$ are independent given the class label.
- (A3) The regression coefficients θ in (2.2) belong to the compact parameter space $\Theta = \{\theta = (\beta_0, \beta^\top)^\top : \|\theta\|_1 \leq M\}$, where M is a constant satisfying $M > |\beta_0^\circ| + s$ with $\beta_0^\circ = \log(P(Y = 1)/P(Y = 0))$.

Theorem 1. *Under Conditions (A1)–(A3), the classification risk of S-RoLLR converges to the minimal Bayes risk, that is, $\lim_{n \rightarrow \infty} R(\hat{g}_n) = R^*$.*

Appendix A in the Supplementary Materials gives the proof. Condition (A1) enforces correct specification of class-specific feature distributions and guarantees consistency of MLEs in $\{(\hat{\Psi}_{j0}, \hat{\Psi}_{j1}), j \in [p]\}$. Condition (A2) is the naive Bayes assumption. Condition (A3) is mostly standard in regularized regression, with the lower bound for M imposed to ensure that the true log-odds $\eta^*(\mathbf{X})$ is attained at some value of θ in Θ . Simply put, Theorem 1 states that, when the data-generating mechanism is consistent with

3.2 Risk analysis under model misspecification

the assumed feature distributions and satisfies the naive Bayes assumption, S-RoLLR achieves the theoretical minimum classification risk.

3.2 Risk analysis under model misspecification

Relaxing (A1) and (A2), we now allow the true feature distributions to fall outside the TPSC family and (2.2) to be an incorrect model for the class label. We derive in the Supplementary Materials (see Appendix B) an upper bound for the increment in risk of S-RoLLR presented in Theorem 2. In addition to (A3), the following conditions are imposed in the derivations:

- (A4) The true class-specific density of X_j , denoted by $\tilde{f}_j(\cdot | Y = k)$, does not vanish, for $j \in [p]$, $k \in \{0, 1\}$.
- (A5) Define $\Psi_{jk}^\dagger = \arg \min_{\Psi} D_{\text{KL}}\{\tilde{f}_j(\cdot | Y = k) \| f_{\text{TPSC}}(\cdot | \Psi)\}$ as the model parameters in the Kullback–Leibler projection of the true class-specific density of X_j onto the TPSC family, and define $e_{jk}(t) = \log \tilde{f}_j(t | Y = k) - \log f_{\text{TPSC}}(t | \Psi_{jk}^\dagger)$, for $j \in [p]$, $k \in \{0, 1\}$. For every $j \in \mathcal{I}$ and $k \in \{0, 1\}$, $E\{e_{jk}^2(X_j)\} \leq \delta_f$ for some positive constant δ_f .
- (A6) The true class-specific joint density of p features, denoted by $\tilde{f}(\mathbf{X}|Y)$, satisfies that $E[\{\log(\tilde{f}(\mathbf{X}|Y = 1)/\tilde{f}(\mathbf{X}|Y = 0))\}^2] < \infty$.
- (A7) For $k \in \{0, 1\}$, $E[\{\log(\tilde{f}(\mathbf{X}|Y = k)/\prod_{j=1}^p \tilde{f}_j(X_j|Y = k))\}^2] < \infty$.

Conditions (A4) and (A5) require the discrepancy between the true and

3.2 Risk analysis under model misspecification

the assumed feature distribution to be well controlled, which is realistic given the flexibility of the TPSC family. The restriction is needed only on the informative features, since for $j \in \mathcal{I}_N$ the two Kullback–Leibler projections coincide and the two class-specific fitting errors cancel. Conditions (A6) and (A7) make the linear predictor a reasonable approximation for $\eta^*(\mathbf{X})$: writing $T_j^*(X_j) = \log(\tilde{f}_j(X_j|Y=1)/\tilde{f}_j(X_j|Y=0))$, they imply that $\eta^*(\mathbf{X}) - \{a + \sum_{j=1}^p b_j T_j^*(X_j)\}$ has a finite second moment for some $(a, b_1, \dots, b_p)^\top \in \mathbb{R}^{p+1}$, whose tight upper bound we denote by δ_{lo} .

Theorem 2. *Under Conditions (A3)–(A7), the S-RoLLR classification risk $R(\hat{g}_n)$ converges to $R^* + \Delta R$ as $n \rightarrow \infty$, where the asymptotic excess risk ΔR compared with the optimal Bayes classifier satisfies*

$$\Delta R \leq \sqrt{\frac{1}{2} [\delta_{lo} + \{2|J_1|(1 - \bar{\rho}_1) + |J_2|\} M^2 \delta_f]}. \quad (3.4)$$

Here, M and δ_f are given in (A3) and (A5), respectively; $J_1 = \{j \in \mathcal{I} : -1 \leq \rho_j \leq 0.5\}$, $J_2 = \{j \in \mathcal{I} : 0.5 < \rho_j \leq 1\}$, $\bar{\rho}_1 = |J_1|^{-1} \sum_{j \in J_1} \rho_j$, with $\bar{\rho}_1 = 0$ when $J_1 = \emptyset$, and

$$\rho_j = \frac{E\{e_{j1}(X_j)e_{j0}(X_j)\}}{\sqrt{E\{e_{j1}^2(X_j)\}E\{e_{j0}^2(X_j)\}}}, \quad j \in \mathcal{I}. \quad (3.5)$$

If either second moment in the denominator of (3.5) is zero, we set $\rho_j = 1/2$ by convention.

3.2 Risk analysis under model misspecification

Theorem 2 reveals three insights. First, as shown in (3.4), the excess risk arises from two distinct sources: the generative stage, reflected in δ_f that arises from modeling non-TPSC features via the TPSC family; and the discriminative stage, captured by δ_{l_o} that stems from assuming the log-odds are linear in the transformed features. Both quantities vanish in the absence of misspecification, reducing the bound to zero as in Theorem 1. Second, when both second moments in (3.5) are positive, ρ_j measures the normalized alignment between the fitting errors of the j -th feature's two class-specific distributions. The bound remains small when these errors are positively, but moderately, aligned. This behavior highlights an intrinsic robustness in the design of the transformed features: the two estimation errors partially offset each other within the likelihood ratio $\hat{T}_j(X_j)$, thereby preserving the discriminative signal even under imprecise marginal fits. Conversely, strong negative alignment between the errors tends to distort the signal, elevating the overall classification risk. Third, the generative contribution in (3.4) is governed by the number of informative features, $s = |\mathcal{I}|$, rather than the ambient dimension p , as the two class-specific fitting errors of any non-informative feature cancel out in the log-likelihood ratio. As we establish in the next subsection, this property is crucial for screening out noise features via L_1 -penalized logistic regression. Section 6.2 further elaborates on the im-

3.3 Variable selection analysis

plications of this signal-adaptive structure for high-dimensional extensions.

3.3 Variable selection analysis

Putting back (A1)–(A3), we now investigate the feature-selection aspect of S-RoLLR, with the finding summarized in Theorem 3. Let $Q_n(\boldsymbol{\theta})$ be (2.3) without the LASSO penalty, so that $-Q_n(\boldsymbol{\theta})$ is n^{-1} times the logistic regression log-likelihood evaluated at $\{(Y_i, \hat{\mathbf{T}}_i)\}_{i=1}^n$. Under (A1)–(A2), the true log-odds $\eta^*(\mathbf{X})$ can be attained by setting $\boldsymbol{\theta}$ at $\boldsymbol{\theta}^\circ = (\beta_0^\circ, \boldsymbol{\beta}^{\circ\top})^\top$, yielding $\eta^*(\mathbf{X}) = \beta_0^\circ + \mathbf{T}^{*\top} \boldsymbol{\beta}^\circ$, where $\mathbf{T}^* = (T_1^*(X_1), \dots, T_p^*(X_p))^\top$ contains the oracle transformed features resulting from evaluating $\mathbf{T}$ at the truth of TPSC parameters, $\{(\boldsymbol{\Psi}_{j0}^*, \boldsymbol{\Psi}_{j1}^*), j \in [p]\}$, and

$$\beta_0^\circ = \log(P(Y = 1)/P(Y = 0)), \quad \beta_j^\circ = \mathbb{1}(j \in \mathcal{I}), \text{ for } j \in [p]. \quad (3.6)$$

Because $\mathcal{I}_N = \{j \in [p] : f_{j0} \equiv f_{j1}\}$, $T_j^*(\cdot) \equiv 0$ for $j \in \mathcal{I}_N$, and thus any $\boldsymbol{\theta} \in \boldsymbol{\Theta}^\circ = \{\boldsymbol{\theta} \in \boldsymbol{\Theta} : \beta_0 = \log(P(Y = 1)/P(Y = 0)), \beta_j = 1 \text{ for } j \in \mathcal{I}\}$ gives the same $\eta^*(\mathbf{X})$ as the oracle linear predictor. We take $\boldsymbol{\theta}^\circ \in \boldsymbol{\Theta}^\circ$ defined in (3.6) as the target of the regularized logistic regression analysis, which is the canonical sparse representative of the set of regression coefficients $\boldsymbol{\Theta}^\circ$. This canonical sparse representative is an identifiable target because $\boldsymbol{\theta}^\circ$ is the unique minimizer of $\|\boldsymbol{\beta}\|_1$ over $\boldsymbol{\Theta}^\circ$.

3.3 Variable selection analysis

Theorem 3. *Under Conditions (A1)–(A3),*

$$\max_{j \in \mathcal{I}} \left| \frac{\partial Q_n(\boldsymbol{\theta})}{\partial \beta_j} \right|_{\boldsymbol{\theta}=\boldsymbol{\theta}^\circ} = O_p(n^{-1/2}), \quad \max_{j \in \mathcal{I}_N} \left| \frac{\partial Q_n(\boldsymbol{\theta})}{\partial \beta_j} \right|_{\boldsymbol{\theta}=\boldsymbol{\theta}^\circ} = O_p(n^{-1}). \quad (3.7)$$

Theorem 3 suggests that, at $\boldsymbol{\theta}^\circ$, the gradient functions in (3.7) fall on two scales, those associated with noise features being smaller by a factor of order $n^{-1/2}$. As elaborated in the proof of Theorem 3 in Appendix C, having $T_j^*(\cdot) \equiv 0$ for $j \in \mathcal{I}_N$ contributes to this separation in scale, causing $\hat{T}_j(X_j)$ to reflect only the gap between two fits of the same distribution ($f_{j0} \equiv f_{j1}$). We show in Appendix D that the penalty can exploit the separation to achieve consistent feature selection. In particular, setting $\lambda_n \asymp n^{-1/2}$ makes the gradients for noise features fall below λ_n , which in turn push their regression coefficients to zero by the Karush–Kuhn–Tucker conditions (Park and Hastie, 2007). This is formally stated in the following theorem.

(A8) The principal submatrix of the oracle Fisher information $\mathbf{I}^* = E\{\nabla^2 Q_n^*(\boldsymbol{\theta}^\circ)\}$,

denoted by $\mathbf{I}_{\mathcal{S}\mathcal{S}}^*$, has its minimum eigenvalue bounded from below by 2κ , where κ is a positive constant, $\mathcal{S} = \{0\} \cup \mathcal{I}$ indexes the intercept together with the informative features, and $Q_n^*(\boldsymbol{\theta}^\circ)$ is (2.3) with $\lambda_n = 0$, $\boldsymbol{\theta}$ evaluated at $\boldsymbol{\theta}^\circ$ in (3.6), and $\{\hat{\mathbf{T}}_i\}_{i=1}^n$ replaced by $\{\mathbf{T}_i^*\}_{i=1}^n$.

Theorem 4. *If $\lambda_n \asymp n^{-1/2}$, then the penalized MLE $\hat{\boldsymbol{\theta}}$ satisfies*

$\lim_{n \rightarrow \infty} P(\hat{\beta}_j = 0 \text{ for all } j \in \mathcal{I}_N) = 1$ and $\lim_{n \rightarrow \infty} P(\hat{\beta}_j \neq 0 \text{ for all } j \in \mathcal{I}) = 1$ under Conditions (A1)–(A3) and (A8).

Condition (A8) constrains only the signal block $\mathbf{I}_{\mathcal{S}}^*$ of the information matrix $\mathbf{I}^*$. The noise block of $\mathbf{I}^*$ is irrelevant because the noise features are controlled through the rate separation indicated in Theorem 3 and the L_1 penalty, as elaborated in Appendix D. Theorem 4 indicates that, on one hand, S-RoLLR achieves consistent noise elimination without requiring the irrepresentable condition (Zhao and Yu, 2006), as long as λ_n matches the order of the signal rates in (3.7). On the other hand, S-RoLLR asymptotically recovers informative features because $\beta_j^\circ = 1$ for such features, so that the $\sqrt{n}$ -consistency of $\hat{\boldsymbol{\theta}}$ established in Appendix D rules out false negatives without a separate minimum-signal condition. Together, the two give consistent support recovery, that is, $\lim_{n \rightarrow \infty} P(\{j \in [p] : \hat{\beta}_j \neq 0\} = \mathcal{I}) = 1$.

4. Simulation Study

4.1 Design of the simulation experiment

We now compare RoLLR and S-RoLLR with six widely accepted classifiers in their finite-sample performance: MOKE, the mode-based nonparametric classifier of Xiong et al. (2025); the Gaussian mixture model under the naive Bayes assumption (GMM-NB, Scrucca et al., 2023); L₁-regularized logistic regression (L1-RL), the classical classifier that implements feature selection on $\mathbf{X}$; the support vector machine (SVM) with radial basis function kernel

4.1 Design of the simulation experiment

(Schölkopf and Smola, 2002), a nonparametric black-box method; random forest (RF, Breiman, 2001), well-suited for nonlinear and high-dimensional data; and the light gradient boosting machine (LightGBM, Ke et al., 2017), designed specifically for high-efficiency learning with large n and p .

We simulate data $\{(Y_i, \mathbf{X}_i)\}_{i=1}^n$ according to twelve data-generating mechanisms referred to as Cases 1–12, tabulated in Table 1, with detailed formulation of each case provided in Appendix E. The first $|\mathcal{I}|$ features are informative, and the remaining $p - |\mathcal{I}|$ are non-informative drawn from $\mathcal{N}(\mathbf{0}, \mathbf{I})$. We set $n = 800$, $p = 30$ and $|\mathcal{I}| = 20$ throughout, except in Case 9 where we reduce n to 300 while raising p to exceed n at 700. Class labels are drawn from Bernoulli(0.5) in all cases but Case 8, where we first generate features, conditioning on which labels are simulated from a probit regression model with a nonlinear predictor.

The twelve simulation settings cover diverse classification challenges. Case 1 serves as a Gaussian baseline where classes differ solely in their means. Cases 2–5 and 10 introduce various departures from Gaussianity: nonexistence of moments (Case 2), high skewness (Case 3), asymmetric tail behavior (one heavy-tailed versus one light-tailed class, Case 4), severe outliers (Case 5), and mean-matched skewness that yields zero first-moment separation (Case 10). While the naive Bayes assumption holds in Cases 1–5,

4.1 Design of the simulation experiment

Table 1: Generative models of the informative features $\mathbf{X}_{\mathcal{I}}$ in Section 4

Case	$X_j \mid Y = 0$	$X_j \mid Y = 1$	$X_j \perp X_{j'} \mid Y$
1	$\mathcal{N}(-0.5, 4)$	$\mathcal{N}(0.5, 4)$	Yes
2	Cauchy(0, 1)	Cauchy(0.5, 1)	Yes
3	Lognormal(0, 1)	Lognormal(0.5, 1)	Yes
4	$\mathcal{N}(0, 1)$	$t_2 + 0.5$	Yes
5	$0.9\mathcal{N}(\mu_{0j}, \sigma_j^2) +$ $0.1\mathcal{N}(\mu_{0j}, 100\sigma_j^2)$	$0.9\mathcal{N}(\mu_{1j}, \sigma_j^2) +$ $0.1\mathcal{N}(\mu_{1j}, 100\sigma_j^2)$	Yes
6	$\mathcal{N}(\mathbf{0}_{ \mathcal{I} }, \Sigma)$	$\mathcal{N}(0.3\mathbf{1}_{ \mathcal{I} }, 2\Sigma)$	No
7	$\mathcal{N}(0.5\mathbf{1}_{ \mathcal{I} }, \tilde{\Sigma})$	$\mathcal{N}(-0.5\mathbf{1}_{ \mathcal{I} }, \tilde{\Sigma})$	No
8	$\mathbf{X}_{\mathcal{I}} \sim \mathcal{N}(\mathbf{0}_{ \mathcal{I} }, \mathbf{I}_{ \mathcal{I} }), Y \mid \mathbf{X}_{\mathcal{I}} \sim \text{Bernoulli}\{\Phi(h(\mathbf{X}_{\mathcal{I}}))\}$		No
9	same as Case 4 but with $n = 300$ and $p = 700$		Yes
10	mean-zero TPSC	mirror image of f_{j0}	Yes
11	bimodal mixture normal	mirror image of f_{j0}	No
12	bimodal mixture t	mirror image of f_{j0}	No

4.2 Simulation results

9, and 10, the remaining settings have dependent features given Y , with Case 7 exhibiting stronger dependence than Cases 6, 11, and 12. Case 6 further introduces unequal covariance matrices for informative features between two classes. Cases 10–12 represent mean-matched scenarios: each informative feature shares the same mean across classes, with class-specific distributions that are mirror images of each other. Consequently, the classes differ only in the sign of their skewness while sharing all even-order moments, making them distinguishable solely through higher-order distributional shapes. Among these, Case 10 uses TPSC-distributed features to match the working model of (S-)RoLLR; Cases 11 and 12 use bimodal features to make the working model a severely misspecified one.

4.2 Simulation results

We train the eight classifiers on 70% of each simulated dataset and test them on the remaining 30%. We choose the penalty term λ_n in S-RoLLR by five-fold cross-validation, minimizing the binomial deviance over the data-dependent grid constructed via the R package `glmnet`, and in L1-RL by ten-fold cross-validation as the default in `cv.glmnet`. Appendix F gives the numerical protocol for the exercise of testing following training, including the construction and the realized range of λ_n . This exercise is repeated for

4.2 Simulation results

300 Monte Carlo replicates under each case, and the area under the receiver operating characteristic curve (AUC) on the testing data is presented in Figure 1. A larger AUC suggests a better classification performance.

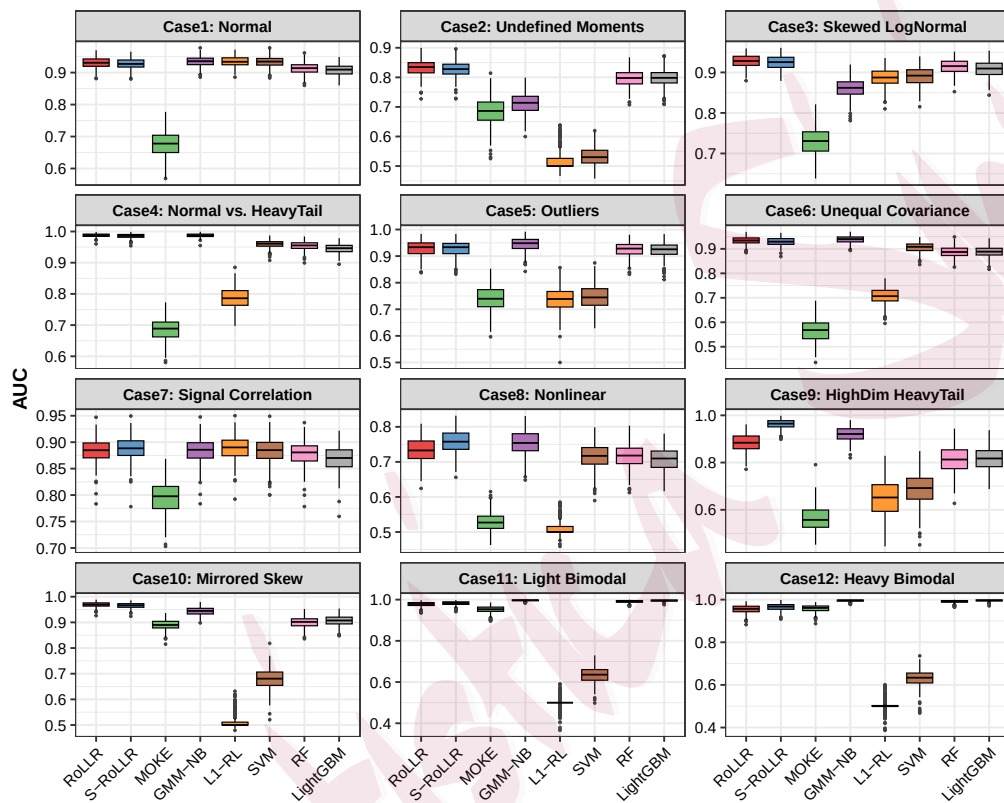


Figure 1: AUC boxplots for eight classifiers across twelve simulation settings

RoLLR and S-RoLLR show strong and competitive performance across all twelve cases. They remain effective even when the naive Bayes assumption is violated, as in Cases 6–8, 11, and 12. The advantage of the flexible TPSC distribution becomes clear when features are far from Gaussian (Cases

4.2 Simulation results

2–5). GMM-NB is a strong competitor, especially with Gaussian or contaminated Gaussian features in Cases 1 and 5. However, it falls behind when features have heavy tails or high skewness (Cases 2 and 3). While both RoLLR and MOKE model the mode of a feature, RoLLR clearly outperforms MOKE in Cases 1–11. This shows the advantage of flexible parametric modeling over nonparametric modeling. Although S-RoLLR assumes a linear predictor with the logit link in the binary regression, its performance remains robust to violations of these assumptions (Cases 6–8), or is less affected than other methods, such as L1-RL, a method also assuming logistic regression. When $p > n$ (Case 9), S-RoLLR clearly improves upon RoLLR, thanks to the feature screening function of S-RoLLR. Even so, with informative features that differ in tail heaviness between two classes, RoLLR still beats most other competitors.

Cases 10–12 illustrate the benefit of extracting richer distributional characteristics via flexibly modeling features within each class, as opposed to classification mostly relying on mean separation of features between classes. Without such separation, L1-RL finds no useful signal for classification, resulting in an AUC around 0.5 in these cases; SVM is another method suffering a noticeable decline in performance. When the feature distributions are actually TPSC (Case 10), RoLLR and S-RoLLR achieve the highest

4.2 Simulation results

accuracy. In the presence of bimodal features (Cases 11 and 12), GMM-NB and tree-based methods slightly outperform our methods.

For the two sparse classifiers, S-RoLLR and L1-RL, Table 2 summarizes their performance in feature selection, including the number of selected features denoted by $|\hat{\mathcal{I}}|$, true positive rate (TPR), false positive rate (FPR), and the F_1 score (Christen et al., 2023). In all twelve cases, S-RoLLR achieves higher F_1 scores. Notably, L1-RL fails (with $F_1 < 0.1$) in Cases 2, 8, and 10–12, where L1-RL selects so few features that its TPR drops to near zero. S-RoLLR maintains high TPR and low FPR in these challenging settings, avoiding the selection collapse seen in L1-RL.

Table 2: Averages of $|\hat{\mathcal{I}}|$, TPR, FPR, and F_1 score associated with S-RoLLR/L1-RL across 300 Monte Carlo replicates in each of twelve simulation settings

#	$ \hat{\mathcal{I}} $	TPR	FPR	F_1	#	$ \hat{\mathcal{I}} $	TPR	FPR	F_1
1	23.8/26.0	1.00/1.00	0.38/0.60	0.91/0.87	7	15.4/15.9	0.69/0.63	0.17/0.32	0.77/0.71
2	25.9/1.9	1.00/0.07	0.59/0.06	0.87/0.09	8	23.6/1.6	0.83/0.05	0.70/0.05	0.76/0.08
3	23.9/25.8	1.00/1.00	0.39/0.59	0.91/0.87	9	34.1/24.6	0.95/0.32	0.02/0.03	0.71/0.27
4	21.3/23.5	1.00/0.94	0.13/0.47	0.97/0.86	10	21.9/1.5	1.00/0.05	0.19/0.05	0.96/0.08
5	22.1/19.2	0.92/0.76	0.36/0.39	0.88/0.77	11	16.9/1.7	0.83/0.05	0.04/0.06	0.89/0.08
6	23.4/23.2	1.00/0.91	0.35/0.50	0.92/0.84	12	16.4/1.6	0.80/0.05	0.03/0.06	0.88/0.08

In summary, S-RoLLR outperforms L1-RL in both classification and fea-

ture selection. As for computational cost, RoLLR and S-RoLLR are mostly comparable with SVM and LightGBM. These methods are far less cumbersome than GMM-NB, whereas MOKE and L1-RL are the fastest, and LightGBM scales best for large p . Runtime comparison of these eight classifiers in the simulation experiment is reported in Table E.2 of Appendix E.

5. Three Case Studies in Real-life Application

We now apply the classifiers considered in Section 4 to three datasets from biomedical studies and astrophysics.

5.1 Prostate cancer dataset

We first consider expression profiles of $p = 200$ genes from $n = 102$ radical prostatectomy specimens, 50 nontumor and 52 tumor. The 200 genes result from a pre-screening step, as in Fan and Lv (2008) and Xiong et al. (2025), applied to the 6,033 genes of the prostate cancer study of Singh et al. (2002). Even after pre-screening, p exceeds n , and a common goal in such studies is to identify genes that correlate with the clinical behavior of a disease, which calls for a classifier that achieves high accuracy and effective feature selection. We compare the eight classifiers over 300 random 70/30 training-testing splits. According to the AUC in the left panel of Figure 2, S-RoLLR

5.1 Prostate cancer dataset

and L1-RL yield the most accurate classification of specimens, followed by LightGBM and RF, whose post-hoc importance scores can facilitate feature selection via appropriate thresholding.

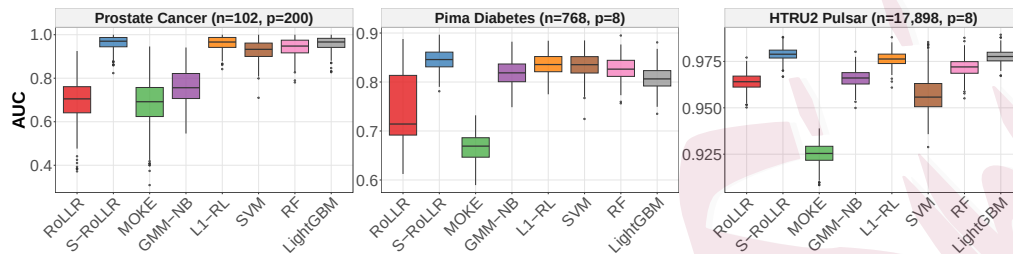


Figure 2: AUC boxplots of eight classifiers across three datasets: prostate cancer (left), Pima diabetes (middle), and HTRU pulsar (right)

Both leading methods incorporate intrinsic feature selection capabilities. Across the 500 splits, L1-RL selects an average of 10.1 genes, while S-RoLLR selects only 7.1. Thus, at comparable classification accuracy, S-RoLLR yields a more parsimonious model. It is also more consistent in gene selection. As shown in Figure 3, only a few of its top 15 genes are frequently selected while most are rarely chosen—a contrast that is far less distinct in L1-RL. Moreover, S-RoLLR selects 5/4 of the top 7 genes ranked by the importance scores of RF/LightGBM, against 4/3 of the top 10 for L1-RL, indicating closer agreement with the two ensemble classifiers on the most influential features. This is because the linear predictor on transformed features ($\hat{\mathbf{T}}$) in S-RoLLR is more powerful than that on raw features ($\mathbf{X}$) in

5.1 Prostate cancer dataset

L1-RL. As a result, S-RoLLR selects informative features more effectively, even when the data are nonlinear.

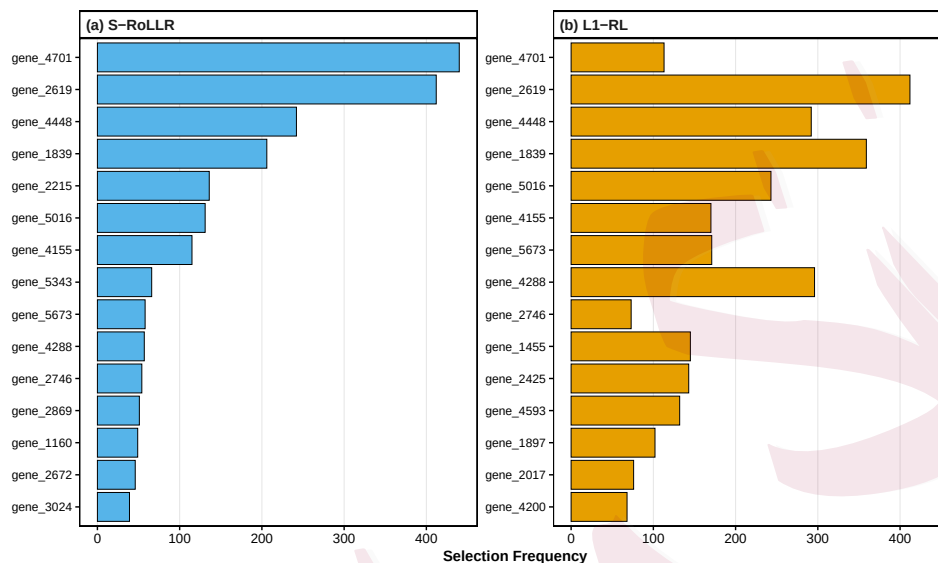


Figure 3: Selection frequencies of the fifteen most frequently selected genes in the prostate cancer data, for (a) S-RoLLR (left panel) and (b) L1-RL (right panel)

A closer look at the selected genes highlights gene 2215. S-RoLLR selects it 136 times across the 500 splits, as its fifth highest frequency. In contrast, L1-RL selects it in only 20 splits (4.0%), and it does not appear in the top ten importance scores for LightGBM or RF. Figure 4 (see the left panel) explains this difference: the discriminative signal of gene 2215 is encoded in its variance and distributional shape rather than its location. While the between-class difference in sample means is modest (a gap of 0.41),

5.1 Prostate cancer dataset

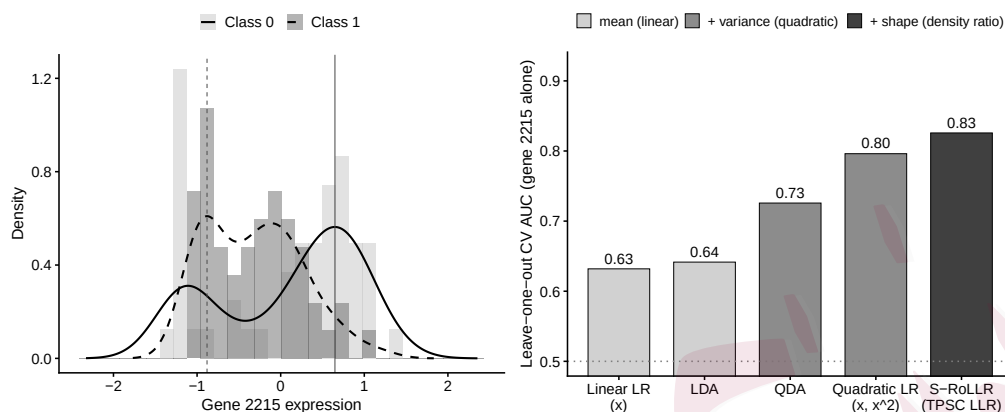


Figure 4: Histograms and kernel density estimates (left panel) of the expression level of gene 2215 for the nontumor group (Class 0, solid line) and the tumor group (Class 1, dashed line), with vertical lines marking the modes; and AUC (right panel) of five classifiers using gene 2215 alone.

the shape discrepancy is substantial. Specifically, the estimated modes from kernel density estimates differ by 1.53, and the tumor class exhibits a noticeably more concentrated spread (variance ratio 0.46, with p -value of 0.006 from an F -test for equal variance). To further confirm the hidden information in this gene is useful for classification, we compare the leave-one-out cross-validated AUC of five classifiers using gene 2215 alone in Figure 4 (see the right panel). The traditional mean-based classifiers, including logistic regression (linear LR) and linear discriminant analysis (LDA), yield the lowest AUC. By accommodating higher moments of the feature, quadratic discriminant analysis (QDA) and logistic regression with a quadratic term

5.2 Pima Indians diabetes dataset

(quadratic LR) improve over the first two. S-RoLLR achieves the highest classification accuracy by using this gene alone and more fully accounting for its distinct non-Gaussian characteristics across the two classes, which are absorbed into the transformed feature.

This case study demonstrates that S-RoLLR can capture signals in features reflecting their shapes and higher-order moments, which are utilized to achieve more accurate classification and feature selection when compared with mean-based linear selectors and tree ensembles. This contributes to S-RoLLR's ability to yield more parsimonious models with greater stability in identifying informative genes, corroborating Theorems 3–4.

5.2 Pima Indians diabetes dataset

Model interpretability and decision transparency are critical in precision medicine, an area where S-RoLLR demonstrates clear advantages when analyzing the Pima Indians diabetes dataset (Smith et al., 1988). Available in the R package `mlbench`, this dataset comprises $p = 8$ routine clinical measurements from $n = 768$ females of Pima Indian heritage (500 non-diabetic and 268 diabetic). The eight features are number of pregnancies, 2-hour post-load plasma glucose from an oral glucose tolerance test, diastolic blood pressure, triceps skinfold thickness, 2-hour serum insulin, body mass

5.2 Pima Indians diabetes dataset

index (BMI), a diabetes pedigree function reflecting genetic predisposition, and age. Physiologically implausible zero entries are imputed using the respective feature medians.

The class-imbalanced data here present challenges in classification. Nevertheless, S-RoLLR outperforms all competing methods in terms of AUC based on 300 random splits (see Figure 2). While L1-RL remains a strong competitor, S-RoLLR offers deeper interpretability. Figure 5-(a1) presents the class-specific histograms of BMI, a well-established risk factor for type 2 diabetes (Kahn et al., 2006), with the fitted TPSC densities superimposed. These capture the right-skewed empirical distributions well and show a distinct upward shift in the mode under the diabetes group. Figure 5-(a2) plots the transformed feature against BMI. The curve is clearly nonlinear, switching from negative to positive at around $\text{BMI} = 30$ and steepening beyond 40. These are precisely the clinical thresholds for obesity and for extreme obesity (World Health Organization, 2000). The transformed feature successfully captures the critical risk threshold at $\text{BMI} = 30$ and reveals an elevated risk above 40. This demonstrates how feature transformations can uncover domain insights hidden in the raw data.

S-RoLLR also enables transparent examination of individual predictions. As shown in Figure 5-(b1), S-RoLLR correctly and confidently identifies a

5.2 Pima Indians diabetes dataset

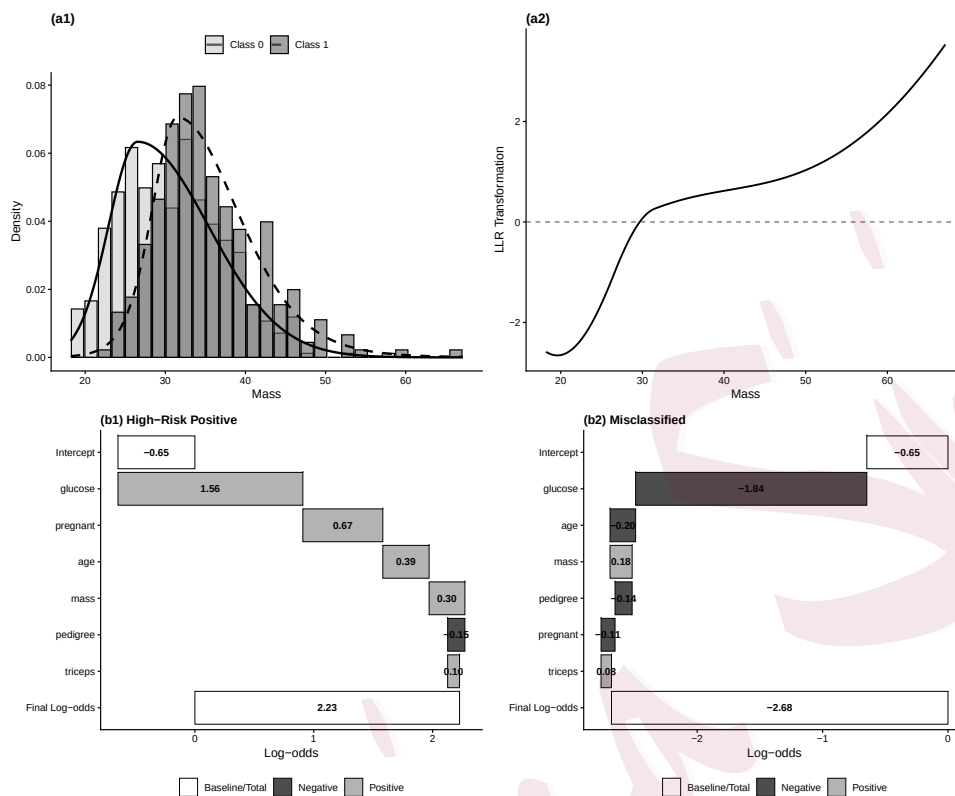


Figure 5: Pima diabetes analysis: (a1) class-wise BMI histograms with fitted TPSC densities; (a2) transformed BMI vs. BMI; (b1) waterfall plot for a true positive instance; and (b2) waterfall plot for a false negative instance.

diabetic patient, presenting her six selected transformed features alongside the estimated log-odds. The model's prediction is primarily driven by high plasma glucose combined with frequent pregnancies at a relatively young age, characteristic of gestational diabetes. Conversely, Figure 5-(b2) illustrates a misclassified diabetic patient. Her unexpectedly low plasma glucose and

5.3 HTRU pulsar dataset

minimal contributions from other features yield an atypical clinical profile. Such an anomalous profile presents a diagnostic challenge in clinical practice, suggesting possible data quality issues. The negative intercept indicates that the model tends to assume a baseline diagnosis of non-diabetic when all transformed features are null.

This case study showcases that S-RoLLR can provide stage-wise interpretability, explaining each step of the classification process. Such transparency is crucial for both scientific insight and high-stakes decisions. The weak feature correlations and accurate marginal fits observed here satisfy the conditions of Theorem 1, validating S-RoLLR’s theoretical strengths.

5.3 HTRU pulsar dataset

In certain applications, feature construction inherently introduces within-class correlations, as observed in the present dataset. Lyon et al. (2016) used this dataset to develop pulsar candidate filters, identifying genuine pulsar signals from sky surveys while filtering out radio frequency interference and other celestial phenomena. The data are available at the UCI Machine Learning Repository (<https://archive.ics.uci.edu/dataset/372/htru2>).

Measured by the High Time Resolution Universe (HTRU) survey, the dataset contains $n = 17,898$ candidate signals: 1,639 pulsars and 16,259

non-pulsars (e.g., radio frequency interference and artifacts). The $p = 8$ features comprise four summary statistics (mean, standard deviation, excess kurtosis, and skewness) of the integrated pulse profile, and the same four statistics for the dispersion-measure signal-to-noise ratio curve, inherently violating the naive Bayes assumption of Theorem 1. Regardless, S-RoLLR still outperforms the others in classification (see Figure 2), demonstrating robustness to strong correlations among features. This resilience stems from the L_1 penalty, which selects a sparse subset from correlated feature groups, ensuring that the retained features remain weakly correlated.

6. Extension of S-RoLLR

6.1 Methodological extension

An immediate extension of S-RoLLR is to adopt different distribution families, not limited to TPSC, at the generative step to model features that can be continuous or discrete. After all, the ratio of two class-specific feature likelihoods is well-defined for most distributions under mild conditions. These likelihood ratios of different features can then be concatenated to form the transformed features $\mathbf{T}$. The discriminative step of S-RoLLR based on (2.3) operates on $\hat{\mathbf{T}}$ regardless of the specific distributions used to construct $\mathbf{T}$. For this extension, let $\mathcal{D}$ be the set of (working) distributions adopted

6.1 Methodological extension

at the generative step, and denote the resulting classifier as S-RoLLR($\mathcal{D}$).

When the two class-conditional Gaussian models are homoscedastic, T_j is affine in X_j . Hence, S-RoLLR($\mathcal{D}$) and L1-RL have the same linear predictor space, although their penalized finite-sample fits need not coincide because the L_1 penalty is scale dependent. More generically, when a working distribution for X_j is a one-parameter exponential family with a complete sufficient statistic linear in X_j , such as Poisson and Bernoulli distributions, $T_j(X_j)$ is affine in X_j , leaving $\hat{T}_j$ in the same linear span as X_j . The transformed feature thus provides little information beyond a location shift of the raw feature between two classes. This in turn suggests the potential gain in incorporating flexible working distributions in $\mathcal{D}$ to extract richer feature information that can be used to distinguish two classes.

For illustration purposes, we consider six choices of $\mathcal{D}$: Gaussian, t , skew-normal, skew- t (Azzalini and Capitanio, 2003), TPSC, and a finite location-scale mixture (originating from t 's) whose number of components $K \in \{1, 2, 3\}$ and degrees of freedom $\nu \in \{5, \infty\}$ are selected via the Bayesian information criterion. We compare the resultant six variants of S-RoLLR($\mathcal{D}$) with L1-RL using data generated from the six configurations in Table 3. Table 4 summarizes their performance in classification and feature selection. More elaborate experiments on these and other settings

6.2 Theoretical extension

with discrete features are reported in Appendix G. Results in Table 4 and Appendix G suggest that S-RoLLR(TPSC) strikes a better balance between flexibility and computational complexity than other variants considered, while remaining competitive in classification and feature selection.

Table 3: Generative models of $\mathbf{X}_{\mathcal{I}}$ in the methodology extension experiment

Setting	$X_j Y = 0$	$X_j Y = 1$
C1	$\mathcal{N}(0, 1)$	$\mathcal{N}(0.55, 1)$
C2 [†]	$\text{SN}(0, 1.5, 5)$	$\text{SN}(0.4, 1.5, 5)$
C3 [†]	$\text{ST}(0, 1.5, 3, 3)$	$\text{ST}(0.75, 1.5, 3, 3)$
C4	mean-zero TPSC	mirror image of f_{j0}
C5	bimodal mixture normal	mirror image of f_{j0}
C6	$0.9\mathcal{N}(0, 1) + 0.1\mathcal{N}(0, 100)$	$0.55 + \{0.9\mathcal{N}(0, 1) + 0.1\mathcal{N}(0, 100)\}$

[†] $\text{SN}(\xi, \omega, \alpha)$ and $\text{ST}(\xi, \omega, \alpha, \nu)$ denote the skew-normal and the skew- t distributions with location ξ , scale ω , slant α , and ν degrees of freedom.

6.2 Theoretical extension

Practically speaking, the choice of $\mathcal{D}$ depends on the features' distributional characteristics one wishes to capture and use for classification, as well as the data needed to infer the chosen distributions reliably. On the other hand, if

6.2 Theoretical extension

$\mathcal{D}$ encompasses correct models for the features, a scenario parallel to that under which Theorems 1, 3, and 4 are established, most of these theoretical results hold for S-RoLLR($\mathcal{D}$) with a revised Condition (A1) to account for the more general $\mathcal{D}$. In particular, under a correct model specification, because $T_j^*(\cdot) \equiv 0$ for a noise feature, the canonical sparse representation of regression coefficients in (3.6) remains intact, which is the key to proving Theorem 3. Consequently, the gradient separation between signal and noise features stated in this theorem is preserved for S-RoLLR($\mathcal{D}$).

A further theoretical extension of S-RoLLR is to allow p to grow as n increases. Although the asymptotic analyses in Section 3 assume a fixed p , many of the underlying mechanisms are not tied to a fixed p . For example, the excess risk bound in Theorem 2 depends on p only through the number of signal features s , which can be fixed even when p grows. Without diving into technical details, we point out four modifications or challenges in the extension of S-RoLLR($\mathcal{D}$) to a diverging- p regime with theoretical guarantees similar to those stated in Theorems 1–4. First, the error incurred in estimating model parameters in $\mathcal{D}$ that are passed into $\hat{\mathbf{T}}$ would have to be controlled uniformly over $j \in [p]$. Second, in (A3), which bounds $\|\boldsymbol{\theta}^\circ\|_1$ by a constant exceeding $|\beta_0^\circ| + s$, now with p diverging, this bound would be replaced by an explicit sparsity condition that ties s

and $\log p$ to n . Third, the assumption on the Fisher information in (A8) and the proof of Theorem 4 need to be restructured. If $p > n$, a restricted eigenvalue condition on $\hat{\mathbf{T}}$ would be needed, which is harder to establish than in standard regression with an observed design matrix. Fourth, while Theorem 4 scales λ_n at order $n^{-1/2}$ to match an individual signal gradient in (3.7), a diverging p would require λ_n to dominate the maximum gradient across all p features, shifting its leading scale to $\sqrt{\log p/n}$.

In summary, extending S-RoLLR to S-RoLLR($\mathcal{D}$) is operationally straightforward; asymptotic analyses of S-RoLLR($\mathcal{D}$) with p growing call for a separate study to follow through.

7. Discussion

In this study, we propose a new classifier, S-RoLLR, for classification using continuous features that can be non-Gaussian. S-RoLLR integrates a flexible log-likelihood ratio transformation into regularized logistic regression, resolving the trilemma of robustness, interpretability, and parsimony. We provide theoretical insights on its properties in the absence or presence of model misspecification. Its feature selection capability is driven by asymptotic rate separation in regression coefficient estimation between signal and noise features, which is further supported by empirical evidence from studies

covering Gaussian and non-Gaussian features in low or high dimensions.

Even with the extension of S-RoLLR in Section 6, a source of information S-RoLLR overlooks is cross-feature dependencies since, here, one models one feature at a time at the generative step. Modeling features jointly can lead to classifiers outperforming S-RoLLR. One possible solution is to use copula functions to model the joint distribution of the features while preserving the flexibility of the marginals. Another methodological extension is to go beyond binary classification. Extending S-RoLLR to multiclass settings is nontrivial, and a careful redesign of the feature transformation is needed to balance theoretical guarantees with computational efficiency.

Supplementary Materials

Supplementary Materials include Appendices A–D for proofs referenced in Section 3, Appendices E and F for implementation and simulation referenced in Section 4, and Appendix G complementing discussions in Section 6.1.

Table 4: Average classification AUC and average feature selection F_1 score across 300 Monte Carlo replicates for L1-RL and six variants of S-RoLLR in the six simulation settings with continuous features, along with maximum AUC shortfall and median runtime across the six settings

Setting	L1-RL	Gaussian	Student- t	skew-normal	skew- t	TPSC	mixture
<i>AUC for classification (with the highest in each setting highlighted in boldface)</i>							
C1	0.949	0.948	0.947	0.942	0.941	0.941	0.947
C2	0.894	0.892	0.883	0.942	0.940	0.940	0.913
C3	0.875	0.792	0.961	0.910	0.982	0.982	0.938
C4	0.511	0.527	0.780	0.914	0.959	0.961	0.923
C5	0.507	0.526	0.526	0.992	0.992	0.994	1.000
C6	0.652	0.530	0.889	0.532	0.871	0.877	0.888
<i>F_1 score for feature selection (with the highest in each setting highlighted in boldface)</i>							
C1	0.655	0.785	0.743	0.694	0.672	0.672	0.765
C2	0.655	0.700	0.604	0.691	0.670	0.659	0.550
C3	0.682	0.541	0.757	0.591	0.810	0.805	0.573
C4	0.032	0.198	0.521	0.664	0.695	0.690	0.537
C5	0.025	0.148	0.146	0.944	0.944	0.952	0.992
C6	0.565	0.199	0.676	0.324	0.578	0.581	0.528
<i>Summary across the six settings (with the lowest value highlighted in italics)</i>							
AUC shortfall [†]	0.494	0.474	0.474	0.357	0.018	<i>0.013</i>	0.043
Runtime (second)	<i>0.4</i>	0.6	7.8	7.7	57.8	8.8	81.4

[†] AUC shortfall of a classifier is the gap between its average AUC across 300 Monte Carlo replicates and that of the best-performing classifier.

REFERENCES

References

- Arslan, O. (2012). Weighted LAD-LASSO method for robust parameter estimation and variable selection in regression. *Computational Statistics & Data Analysis* 56(6), 1952–1965.
- Azzalini, A. and A. Capitanio (2003). Distributions generated by perturbation of symmetry with emphasis on a multivariate skew t-distribution. *Journal of the Royal Statistical Society: Series B* 65(2), 367–389.
- Basu, A., A. Ghosh, M. Jaenada, and L. Pardo (2024). Robust adaptive LASSO in high-dimensional logistic regression. *Statistical Methods & Applications* 33(5), 1217–1249.
- Breiman, L. (2001). Random forests. *Machine learning* 45(1), 5–32.
- Christen, P., D. J. Hand, and N. Kirielle (2023). A review of the f-measure: its history, properties, criticism, and alternatives. *ACM Computing Surveys* 56(3), 1–24.
- Duda, R., P. Hart, and D. Stork (2012). *Pattern Classification*. Wiley.
- 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.
- Fernández, C. and M. F. Steel (1998). On Bayesian modeling of fat tails and skewness. *Journal of the American Statistical Association* 93(441), 359–371.
- Friedman, J. H., T. Hastie, and R. Tibshirani (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software* 33, 1–22.
- Hampel, F. (1986). *Robust Statistics: The Approach Based on Influence Functions*. Probability and Statistics Series. Wiley.
- James, G., D. Witten, T. Hastie, R. Tibshirani, and J. Taylor (2023). *An Introduction to Statistical Learning: with Applications in Python*. Springer.
- Kahn, S. E., R. L. Hull, and K. M. Utzschneider (2006). Mechanisms linking obesity to insulin

REFERENCES

- resistance and type 2 diabetes. *Nature* 444(7121), 840–846.
- Ke, G., Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu (2017). Light-GBM: A highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems* 30.
- Liu, Q., X. Huang, and R. Bai (2024). Bayesian modal regression based on mixture distributions. *Computational Statistics & Data Analysis* 199, 108012.
- Lundberg, S. M. and S.-I. Lee (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems* 30.
- Lyon, R. J., B. W. Stappers, S. Cooper, J. M. Brooke, and J. D. Knowles (2016). Fifty years of pulsar candidate selection: from simple filters to a new principled real-time classification approach. *Monthly Notices of the Royal Astronomical Society* 459(1), 1104–1123.
- Park, M. Y. and T. Hastie (2007). L1-regularization path algorithm for generalized linear models. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* 69(4), 659–677.
- Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence* 1(5), 206–215.
- Schölkopf, B. and A. J. Smola (2002). *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT press.
- Scrucca, L., C. Fraley, T. Murphy, and A. Raftery (2023). *Model-Based Clustering, Classification, and Density Estimation Using mclust in R*. Chapman & Hall/CRC The R Series. CRC Press.
- Singh, D., P. G. Febbo, K. Ross, D. G. Jackson, J. Manola, C. Ladd, P. Tamayo, A. A. Renshaw, A. V. D’Amico, J. P. Richie, et al. (2002). Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell* 1(2), 203–209.

REFERENCES

- Smith, J. W., J. E. Everhart, W. C. Dickson, W. C. Knowler, and R. S. Johannes (1988). Using the adap learning algorithm to forecast the onset of diabetes mellitus. In *Proceedings of the Annual Symposium on Computer Application in Medical Care*, pp. 261.
- Sun, Q., W.-X. Zhou, and J. Fan (2020). Adaptive Huber regression. *Journal of the American Statistical Association* 115(529), 254–265.
- Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology* 58(1), 267–288.
- World Health Organization (2000). Obesity: preventing and managing the global epidemic: report of a WHO consultation.
- Xiong, W., W. K. Härdle, J. Wang, K. Yu, and M. Tian (2025). Mode-based classifier: A robust and flexible discriminant analysis for high-dimensional data. *Statistica Sinica* 35, 1391–1422.
- Zhao, P. and B. Yu (2006). On model selection consistency of Lasso. *The Journal of Machine Learning Research* 7, 2541–2563.
- Department of Statistics, University of South Carolina, Columbia, South Carolina 29208, USA.
E-mail: xiuchuan@email.sc.edu
ORCID: 0009-0007-9283-2814
- Department of Statistics, University of South Carolina, Columbia, South Carolina 29208, USA.
E-mail: huang@stat.sc.edu
ORCID: 0000-0001-7077-0869