Contemporary Mathematics
https://ojs.wiserpub.com/index.php/CM/

UNIVERSAL WISER PUBLISHER

Research Article

Generalized Ridge Estimators in High-Dimensional Generalized Linear Models

Jianglin Fang, **Cunyun Nie***

School of Computational Science and Electronics, Hunan Institute of Engineering, Xiangtan, Hunan, 411104, China
E-mail: 05127@hnie.edu.cn

Received: 10 April 2025; **Revised:** 26 May 2025; **Accepted:** 10 June 2025

Abstract: The paper presents a generalized ridge approach for high dimensional generalized linear models where dimensionality exceeds sample size. When the vector of covariate coefficient paraters β is sparse, a thresholding method for high dimensional generalized linear models is presented, enabling simultaneous variable selection and parameter estimation without the uniform signal strength assumption. Theoretical guarantees for variable selection and estimation, including consistency and convergence rates, are derived under some regularity conditions. Simulation studies and a real world data analysis are conducted to examine the performance of the proposed approach.

***Keywords*:** generalized linear model, ridge estimator, high dimensional data, consistency, asymptotic normality

MSC: 62H15, 62G10, 62G20

1. Introduction

Generalized linear models are a large class of models including, for example, linear normal models, logistic regression models, log-linear models for contingency tables. Therefore, since it was introduced by [1], this has been extensively studied, such as [2–5], and so on. Generally speaking, most of the existing methods for generalized linear models have the same assumption that the covariates are known constants or independent identically distributed random variables. When the covariates are multi-collinear, the estimation will become unstable, and the mean square error of it will be very large, e.g., maximum likelihood estimation. [6] proposed the method of ridge regression that is an alternative to the method of the ordinary least squares that can address multicollinearity. Extensive research has been done on this method since then, for example, [7–9]. The ridge technique in generalized linear models has also been studied. [10] proposed ridge estimators for logistic regression models to diminish the error made by predictions; [11] introduced ridge technique for the Cox model; [12, 13] suggested using the ridge method in generalized linear models to alleviate effects of the ill-conditioned data matrix; [14] proposed the ridge estimation in linear mixed measurement error models; [15] developed the ridge regularized estimation of Visual Autoregressive (VAR) models.

Owing to the rapid development of science and information technology, high dimensional data becomes more and more popular in many fields, such as biometrical engineering, mechanical systems, genetic engineering. A large number of statistical methods, algorithms and theories have been developed for high dimensional data; see, e.g., [5, 16–19]. Recently

DOI: https://doi.org/10.37256/cm.6420257162

a few papers have been published about statistical inference for high dimensional generalized linear models. [20] proposed the variable selection method for high dimensional generalized linear models based on lasso. [21] introduced a feature screening approach. [5] discussed tests for regression coefficients in this case. Because of the computational and statistical complexity of high dimensional data analysis, especially the so-called large p small n problem, statistical inference for it is still largely untouched. This paper is focused on estimating the parameter β in generalized linear models under high dimensional settings.

In the high dimensional setting, when the vector of covariate coefficient parameters β is sparse, variable selection consistency is very important to statistical inference methods of variable selection including the Lasso [16], nonconcave penalized likelihood method [17], nonconvex method [22], and so on. However, most of the existing results, selection consistency theory usually requires a uniform signal strength assumption, which means that nonzero coefficients should be larger than an inflated level of noise. Unfortunately, when the existence of weak signals cannot be excluded, the uniform signal strength assumption is scarcely supported in application. The other goal of the paper is to propose a thresholding method for variable selection in high dimensional generalized linear models without the uniform signal strength assumption.

The main contributions of this work includes two aspects. On one hand, for generalized linear models with high dimensional data, we propose an approach based on ridge technique to estimate the parameter β and establish the asymptotic properties of it. On the other hand, under the large p small n settings, when the vector of regression coefficient parameters β is sparse, a thresholding method for generalized linear models is presented, enabling simultaneous variable selection and parameter estimation without the uniform signal strength assumption.

The rest of the paper organized as follows. In Section 2, we present the methodology and theoretical results. The generalized ridge estimators for generalized linear models and the asymptotic properties of it are shown in Section 2.1. We describe the thresholding method for variable selection when the parameter β is sparse in Section 2.2. Simulation studies are presented in Section 3 and a real data example is presented in Sections 4. All technical details and proofs are given in the Appendix.

2. Methodology and theoretical results

Assume $\{Y_i,\ i=1,\cdots,n\}$ are independent and identically distributed with exponential type densities

$$f(y_i;\ \theta_i)=\exp\left(\theta_i y_i-b(\theta_i+c(y_i))\right),\ \ i=1,\cdots,n, \tag{1}$$

where $c(\cdot)$ and $b(\cdot)$ are specific functions, $\theta\in\Theta^0$ $(\Theta^0\subset\Theta)$ is the nature parameter and Θ is the convex nature parameter space. $\{X_i=(X_{1i},\ \cdots,\ X_{pi})^\top,\ i=1,\ \cdots,\ n\}$ are independent and identically distributed p-dimensional covariates. X influences Y via a linear combination $\eta=X^\top\beta$ where β is a p-dimensional parameter. Define $E(Y|X)=\mu(\theta(x))=g(\eta)$, $V=\mathrm{Var}(Y)$ and $\mathrm{cov}(Y)=\Sigma(\theta)$, where $g(\cdot)$ is a monotonic differentiable function and $\Sigma(\theta)$ is positive in Θ^0. Let $h(\mu)=\eta=X^\top\beta$ be the link function. If h is the canonical link, then $\theta(x)=\sum_{j=1}^{n}\beta_j x_j$. The log-likelihood function is given by

$$l(\beta_1,\cdots,\beta_n;\ x_1,\cdots,x_n;\ y_1,\cdots,y_n)=\frac{1}{n}\sum_{i=1}^{n}\log f\left(y_i;\ \sum_{j=1}^{p}\beta_j x_{ji}\right). \tag{2}$$

The maximum likelihood estimator $\hat{\beta}$ is given by

$$\hat{\beta} = \arg\min_{\beta} \left\{ -\frac{1}{n} \sum_{i=1}^{n} \log f\left(y_i;\ \sum_{j=1}^{p} \beta_j x_{ji} \right) \right\}.$$

Based on [23], we can obtain the iterative equation

$$X^{\top} W X \beta^{k+1} = X^{\top} W Z, \tag{3}$$

where Z is a $n \times 1$ vector with elements

$$Z_i = \sum_{j=1}^{p} X_{ij} \beta_j^k + (Y_i - \mu_i) \frac{d\eta_i}{d\mu_i}$$

and

$$W = \text{diag} \left\{ \frac{1}{V_i} \left(\frac{d\mu_i}{d\eta_i} \right)^2;\ i = 1, \cdots, n \right\}.$$

The maximum likelihood estimator $\hat{\beta}$ of β can be obtained by repeating (3) until convergence when V_i, μ_i and $d\mu_i/d\eta_i$ are evaluated. [2] proved, under mild regularity assumptions, that the maximum likelihood estimator $\hat{\beta}$ is consistent and asymptotically normal.

2.1 *Generalized ridge estimators for generalized linear models*

Among of the regularity assumptions, one of them is that $\sum_{i=1}^{n} X_i X_i^{\top}$ has full rank, which ensure that the information matrix is positive define. If the covariates are highly corrected, the above assumption may be no longer true. This can make the consistency and asymptotic normality properties of $\hat{\beta}$ invalid. In order to solve this problem, [12, 13] proposed to use ridge estimators combining with generalized linear models to improve the estimates in such cases. Define

$$l^{\lambda}(\beta_1, \cdots, \beta_n;\ x_1, \cdots, x_n;\ y_1, \cdots, y_n) = \frac{1}{n} \sum_{i=1}^{n} \log \left(f\left(y_i;\ \sum_{j=1}^{p} \beta_j x_{ji} \right) \right) - \lambda \sum_{j=1}^{p} \beta_j^2, \tag{4}$$

where $\lambda > 0$ is the ridge parameter. The maximizer of expression (4) is the ridge estimator of β, denoted $\hat{\beta}^{\lambda}$,

$$\hat{\beta}^{\lambda} = \arg\min_{\beta} \left\{ -\frac{1}{n} \sum_{i=1}^{n} \log f\left(y_i;\ \sum_{j=1}^{p} \beta_j x_{ji} \right) + \lambda \sum_{j=1}^{p} \beta_j^2 \right\}.$$

Similar to the unrestricted maximum likelihood estimators, $\hat{\beta}^{\lambda}$ can be obtained by the Newton-Raphson maximization procedure as follows:

$$\hat{\beta}^{\lambda} = \left(X^{\top}WX + 2\lambda I\right)^{-1} X^{\top}WX\hat{\beta}, \tag{5}$$

where I is a $p \times p$ identity matrix. Here we can see that $\hat{\beta}^{\lambda}$ shrinks towards 0 if the value of the ridge parameter increases, and the mean square error of $\hat{\beta}^{\lambda}$ should be smaller than $\hat{\beta}$ by selecting an appropriate ridge parameter. Because ridge techniques shrink the estimator $\hat{\beta}$ by adding the same positive quantity to each diagonal element the system $X^{\top}WX$, it is easy to cause too much shrinkage. In order to solve this problem, we propose to use generalized ridge techniques to estimate parameters of generalized linear models. Define

$$l^{\Lambda}(\beta_1, \cdots, \beta_n;\ x_1, \cdots, x_n;\ y_1, \cdots, y_n) = \frac{1}{n}\sum_{i=1}^{n} \log\left(f\left(y_i;\ \sum_{j=1}^{p} \beta_j x_{ji}\right)\right) - \sum_{j=1}^{p} \lambda_j \beta_j^2, \tag{6}$$

where $\lambda_j > 0$ for $j = 1, \cdots, p$. The maximizer of expression (6) is, denoted $\hat{\beta}(\Lambda)$,

$$\hat{\beta}(\Lambda) = \arg\min_{\beta}\left\{-\frac{1}{n}\sum_{i=1}^{n} \log f\left(y_i;\ \sum_{j=1}^{p} \beta_j x_{ji}\right) + \sum_{j=1}^{p} \lambda_j \beta_j^2\right\}.$$

Similar to maximizing (5), we can obtain that

$$\hat{\beta}(\Lambda) = \left(X^{\top}WX + 2\Lambda\right)^{-1} X^{\top}WX\hat{\beta}, \tag{7}$$

where $\Lambda = \mathrm{diag}(\lambda_1, \cdots, \lambda_p)$. Without loss of generality, we define the generalized ridge estimator for β as

$$\hat{\beta}^{\Lambda} = \left(X^{\top}WX + Q\Lambda Q^{\top}\right)^{-1}\left(X^{\top}WX + 2\Lambda\right)\hat{\beta}(\Lambda)$$

$$= \left(X^{\top}WX + Q\Lambda Q^{\top}\right)^{-1} X^{\top}WX\hat{\beta},$$

where Q is a orthogonal matrix, $Q^{\top}(X^{\top}WX)Q = D$, $D = \mathrm{diag}(d_1, \cdots, d_p)$, $d_1, \cdots, d_p$ are eigenvalues of $X^{\top}WX$.

When p is fixed, it is easy to show that $\hat{\beta}^{\Lambda}$ is consistent and asymptotically normal under some mild assumptions. However, when p grows with n, the existence and asymptotic behavior of $\hat{\beta}^{\Lambda}$ are unknown, and the related questions on $\hat{\beta}^{\Lambda}$ are more complicated under the high dimensional setting $p > n$. We first study theoretical properties for generalized ridge estimation in generalized linear models with high dimensional data.

Let $B_p(\tau) = \{\beta \in \mathbb{R}^p\colon \|\beta\| \leq \tau\}$ and β_0 be the true value of the parameter β, where $B_p(\tau) \in \mathbb{R}^p$ is a compact and convex set and τ is a sufficiently large positive constant which can ensure that β_0 is an interior point. The following assumptions are necessary:

Assumption 1. Let $l^{\Lambda}(X^{\top}\beta, Y) = l^{\Lambda}(\beta_1, \cdots, \beta_n; x_1, \cdots, x_n; y_1, \cdots, y_n)$ and

$$I^{\Lambda}(\beta) = \mathrm{E}\left\{\left[\frac{\partial}{\partial\beta} l^{\Lambda}(X^{\top}\beta, Y)\right]\left[\frac{\partial}{\partial\beta} l^{\Lambda}(X^{\top}\beta, Y)\right]^{\top}\right\}.$$

The information matrix $I^{\Lambda}(\beta)$ is finite and positive definite at $\beta = \beta_0$. Furthermore, $\sup_{\beta\in B,\, \|X\|=1} \|I^{\Lambda}(\beta)^{1/2}X\|$ exists, where $\|\cdot\|$ denotes the l_2-norm.

Assumption 2. $l^{\Lambda}(X^{\top}\beta, Y)$ meets the Lipschitz condition with a positive constant k_n

$$|l^{\Lambda}(X^{\top}\beta_1, Y) - l^{\Lambda}(X^{\top}\beta_2, Y)|I_n(X, y) \le k_n|X^{\top}\beta_1 - X^{\top}\beta_2|I_n(X, y)$$

for $\beta_1, \beta_2 \in B$, where $I_n(X, y) = I((X, y) \in \Omega_n)$ and

$$\Omega_n = \{(X, y) : \|X\|_\infty \le K_n,\ |y| \le K_n^*\},$$

where $\|\cdot\|_\infty$ denotes the supremum norm, and K_n and K_n^* are positive constants.

Assumption 3.

$$\sup_{\beta\in B,\, \|\beta-\beta_0\|\le b_n} |\mathrm{E}(l^{\Lambda}(X^{\top}\beta, Y) - l^{\Lambda}(X^{\top}\beta_0, Y))|(1 - I_n(X, y)) \le o\left(\frac{p}{n}\right),$$

where C is a positive constant, $b_n = Ck_nV_n^{-1}(p/n)^{1/2}$ and V_n is the a constant found in Assumption 4.

Assumption 4.

$$\mathrm{E}(l^{\Lambda}(X^{\top}\beta, Y) - l^{\Lambda}(X^{\top}\beta_0, Y)) \ge V_n\|\beta - \beta_0\|^2,$$

where the function $l^{\Lambda}(X^{\top}\beta, Y)$ is convex, $\|\beta - \beta_0\| \le b_n$ and V_n is a positive constant.

Assumption 5. Let $d = \min\{d_1, \cdots, d_p\}$, where d_j is the eigenvalue of $X^{\top}WX$ for $j = 1, \cdots, p$. We assume that d satisfies

$$d^{-1} = O(n^{-\eta}),\ \text{where}\ \ \eta \le 1 \text{ and } \eta \text{ does not depend on } n.$$

Assumption 1 guarantees identifiability and existence for the generalized estimation $\hat{\beta}^{\Lambda}$. Assumptions 2-4 are similar to conditions 1-3 of [21], which are necessary to establish the bound for the tail probability of $\hat{\beta}^{\Lambda}$. Assumption 5 controls the asymptotic convergence rate of the linear combination of the generalized ridge estimation $\hat{\beta}^{\Lambda}$.

Theorem 1 Under Assumptions 1-5, for any $\varepsilon > 0$, there exists a matrix Λ such that

$$\text{(i) } P\left(\|\hat{\beta}^{\Lambda}-\beta_0\| \geq 16k_n(1+\varepsilon)/(V_n\sqrt{n})\right) \leq \exp(-2\varepsilon^2/K_n^2)+nP(\Omega_n^c);$$

$$\text{(ii) } E(\mathrm{I}^{\top}\hat{\beta}^{\Lambda}-\mathrm{I}^{\top}\beta)^2 = O(\lambda_{\min}^{-1})+O(\lambda_{\max}^2 n^{-2\eta}),\ \text{ where } \lambda_{\min}=\min\{\lambda_1,\cdots,\lambda_p\},$$

$\lambda_{\max}=\max\{\lambda_1,\cdots,\lambda_p\}$ and I is a p-dimensional deterministic vector with $\|\mathrm{I}\|=1$.

Theorem 1 (i) shows an upper bound for the tail probability of the generalized ridge estimation $\hat{\beta}^{\Lambda}$ under the high dimensional setting, and Theorem 1 (ii) shows the asymptotic convergence rate of the mean squared error of $\mathrm{I}^{\top}\hat{\beta}^{\Lambda}$. If $\lambda_{\min}\to\infty$ and $\lambda_{\max}^2 n^{-2\eta}\to 0$, the mean squared error of any linear combination of the generalized ridge estimation $\mathrm{I}^{\top}\hat{\beta}^{\Lambda}$ convergence to 0.

Theorem 2 Under Assumptions 1-4, there exists λ and Λ such that

$$\mathrm{MSE}(\hat{\beta}^{\Lambda}) \leq \mathrm{MSE}(\hat{\beta}^{\lambda}) \leq \mathrm{MSE}(\hat{\beta}),$$

where $\mathrm{MSE}(\hat{\beta})=\mathrm{E}\left\{(\hat{\beta}-\beta)^{\top}(\hat{\beta}-\beta)\right\}$.

Theorem 2 shows that, for a good choice of Λ, the estimate $\hat{\beta}^{\Lambda}$ is expected to be on average closer to the real value of β than $\hat{\beta}^{\lambda}$ and $\hat{\beta}$ under the high dimensional setting.

2.2 *Thresholding method for variable selection*

According to Theorem 1, $\hat{\beta}^{\Lambda}$ is not consistent, that is, $\|\hat{\beta}^{\Lambda}-\beta\|$ may not converge to 0 when $p>n$. However, when the vector of regression coefficient parameters β is sparse, existing regularized estimators for β, such as the Lasso and Smoothly Clipped Absolute Deviation (SCAD) penalty methods (see Avella-Medina and Ronchetti, 2018), have the consistency and sparsity. Thus, the generalized ridge estimator is not optimal for variable selection or for estimation of the entire β, although arbitrary linear combination of the estimator is consistent if $\lambda_{\min}\to\infty$ and $\lambda_{\max}^2 n^{-2\eta}\to 0$. When the vector of regression coefficient parameters β is sparse, β contains many zero components. Though the generalized ridge estimator $\hat{\beta}^{\Lambda}$ lacks structural sparsity, many coefficients are near zero. These negligible yet non-zero components of $\hat{\beta}^{\Lambda}$ contribute little to inference while inflating variance. This motivates variable selection via small-component truncation of $\hat{\beta}^{\Lambda}$.

Our recommendation is instead to use a thresholding method to improve the generalized ridge estimator. Let $\hat{\beta}_j^{\Lambda}$, $j=1,\cdots,p$, be the j-th components of $\hat{\beta}^{\Lambda}$. We define the thresholded generalized ridge estimator $\tilde{\beta}^{\Lambda}$ as

$$\tilde{\beta}_j^{\Lambda}=\begin{cases}\hat{\beta}_j^{\Lambda} & \text{if } |\hat{\beta}_j^{\Lambda}|>a_n,\\ 0 & \text{if } |\hat{\beta}_j^{\Lambda}|\leq a_n,\end{cases} \tag{8}$$

where $a_n=C_1n^{-\alpha}$ is the thresholding value, $0<\alpha<1/2$ and $C_1>0$ is not depending on n, $j=1,\cdots,p$. Variable selection is achievable via the thresholding method. Thresholding-based selection retains components indexed by $\mathscr{A}_{\hat{\beta}^{\Lambda},\,a_n}$. Define $\mathscr{A}_{\beta,\,c_n}$ as the index set of $|\beta_j|>c_n$. Next, we will establish asymptotic properties of $\mathscr{A}_{\hat{\beta}^{\Lambda},\,a_n}$ under regular assumptions.

In order to facilitate the technical derivations of the proof, the following conditions that are imposed on the likelihood functions are assumed.

Assumption 6. Let $B(\beta) = \mathrm{diag}\{b''(X_1^\top \beta), \cdots, b''(X_n^\top \beta)\}$, $\mathbf{X} = (X_1, \cdots, X_n)^\top$ and $G = -1/(np^{(1+\delta)})\mathbf{X}^\top B(\beta_0)\mathbf{X}$, where δ is a positive constant. There exists a $\delta > 0$ such that the eigenvalues of G and $1/n\mathbf{X}^\top\mathbf{X}$ are bounded away from zero and infinity.

Assumption 7. There is a positive constant C_{00} such that $|b^{(3)(X_i^\top \beta)}| < C_{00}$.

Assumption 8. Let

$$I_n(\beta) = \mathrm{E}\left[\left\{\frac{\partial \log f(y_n; \sum_{j=1}^{p} \beta_j x_{jn})}{\partial \beta}\right\}\left\{\frac{\partial \log f(y_n; \sum_{j=1}^{p} \beta_j x_{jn})}{\partial \beta}\right\}^\top\right].$$

The information matrix $I_n(\beta)$ is positive definite. Furthermore, there exists a positive constant C_{01} such that

$$\mathrm{E}\left\{\frac{\partial^2 \log f(y_n; \sum_{j=1}^{p} \beta_j x_{jn})}{\partial \beta_j \partial \beta_k}\right\}^2 < C_{01} < \infty.$$

Assumption 9. Denote $\beta = (\beta_1^\top, \beta_2^\top)^\top$, where $\beta_1 \in \mathbb{R}^q$ and $\beta_2 \in \mathbb{R}^{p-q}$ correspond to the nonzero component and zero components, respectively, i.e. $\beta_0 = (\beta_{01}^\top, 0^\top)^\top$. We assume that $q \le R(X)$, where $R(X)$ is the rank of X.

As far as we know, the maximum likelihood estimator $\hat{\beta}$ is asymptotically normal when p is fixed, and the component $\hat{\beta}_j$ of $\hat{\beta}$ is also asymptotically normal for $j = 1, \cdots, p$ when $p^2/n \to 0$ (see Prontnoy [24]). Assumptions 6-8 ensure the asymptotic normality of the component $\hat{\beta}_j$ of $\hat{\beta}$ as $p/n \in (0, \infty)$. Assumptions 6-8 are stronger than the asymptotic likelihood theory in [24], but they facilitate the technical derivations of the proof. Assumption 9 is a sparse assumption, which ensures β is identifiable. According to Assumption 9, although β needs to has many zero components, it allows the number nonzero parameters to diverge to infinity when $n \to \infty$ and $R(X) \to \infty$.

Theorem 3 Let a_n be given by (8) with $\alpha < (\eta - \nu)/3$, $u_n = 1 + (\log\log n)^{-1}$ and $\lambda_{min} = C_2 a_n^{-2}(\log\log n)^3 \log(n \vee p)$, where $\nu > 0$ and $C_2 > 0$ are constants, and $(n \vee p) = \max\{n, p\}$. Under Assumptions 1-9 and $0 < p/n < \infty$, for any constant $t > 0$, we have that

$$P(\mathscr{A}_{\beta, a_n u_n} \subset \mathscr{A}_{\hat{\beta}^\Lambda, a_n} \subset \mathscr{A}_{\beta, a_n/u_n}) = 1 - O(n^{-t}). \tag{9}$$

According to above Theorem, we can keep components where $|\beta_j| > a_n u_n$ and remove those with $|\beta_j| < a_n/u_n$ by thresholding $\hat{\beta}^\Lambda$. Furthermore, Theorem 3 implies that

$$\lim_{n\to\infty} P(\mathscr{A}_{\hat{\beta}^\Lambda, a_n} = \mathscr{A}_{\beta, a_n}) = 1,$$

which means that the thresholding method is consistent for variable selection.

Remark 1 The major difference between the existing variable selection consistency method for generalized linear models with high-dimensional data and expression (9) is the signal strength assumptions. the existing variable selection consistency method, such as Lasso, adaptive lasso, SCAD, etc., relies on uniform signal strength assumptions on all nonzero parameters as in [25], which can be written as

$$\min_{\beta_j \neq 0} |\beta_j| \ge 2n^{-\gamma}\log n, \quad \gamma \in (0, 1/2].$$

Furthermore, the existing method can not guarantee to select correctly variables with large $|\beta_j|$ or $\beta_j = 0$ in the presence of small $|\beta_j| \neq 0$. By contract, Theorem 3 does not require minimal signal assumptions on the parameters. On the other hand, large $|\beta_j|$ can be selected and $\beta_j = 0$ can not be selected by the threholding method in the presence of possibly many small nonzero $|\beta_j|$.

Without loss of generality, denote $\hat{\beta}^{\Lambda} = ((\hat{\beta}_{n1}^{\Lambda})^{\top}, (\hat{\beta}_{n2}^{\Lambda})^{\top})^{\top}$, where $\hat{\beta}_{n1}^{\Lambda} \in \mathbb{R}^{s}$ and $\hat{\beta}_{n2}^{\Lambda} \in \mathbb{R}^{p-s}$, $\hat{\beta}_{n1i}^{\Lambda} > a_n$, $\hat{\beta}_{n2j}^{\Lambda} \leq a_n$ $i = 1, \cdots, s$, $j = s+1, \cdots, p$. Similarly, we let $\beta_0 = (\beta_{n10}^{\top}, \beta_{n20}^{\top})^{\top}$ and $\tilde{\beta}^{\Lambda} = ((\tilde{\beta}_{n1}^{\Lambda})^{\top}, (\tilde{\beta}_{n2}^{\Lambda})^{\top})^{\top}$, where $\tilde{\beta}_{n2}^{\Lambda} = 0$. An estimator is said to be consistent, if the estimator converges in probability to the true value of the parameter. Although Theorem 1 (i) shows an exponential bound for the tail probability of the generalized ridge estimation $\hat{\beta}^{\Lambda}$, $\hat{\beta}^{\Lambda}$ may not be consistent under the high dimensional setting. The following Theorem shows that the thresholded ridge estimator $\tilde{\beta}^{\Lambda}$ is asymptotically consistent under the sparsity assumption on β.

Theorem 4 Assume $qn^{-2\alpha} \to 0$ as $n \to \infty$ and Assumptions in Theorem 3, for any $\varepsilon > 0$, we have

$$P(\|\tilde{\beta}^{\Lambda} - \beta_0\| > \varepsilon) \to 0.$$

3. Simulation studies

For the purpose of comparison, we examined the L_2-errors of generalized ridge estimators, thresholded generalized ridge estimators, and Lasso, Elastic Net and SCAD penalized estimators in the following simulations. The selection of the constant C_1 and the thresholding value a_n are similar to that of [18].

Simulation study 1. In this example, we generate observations independently from the mean model $E(Y_i|X_i) = g(X_i^{\top}\beta)$, $i = 1, \cdots, n$, where $g(\cdot)$ is the identical transformation, ε_i is a random error with mean 0 and $\sigma = 1$, and X_i is generated from $N(0, \Sigma)$, where Σ is a $p \times p$ compound symmetry covariance matrix with correlation $\rho = 0.5$. We consider $\beta = (1, 0.5, 1, 0.5, 1, 0.5, 1, 0.5, 1, 0.5, 0, \cdots, 0)^{\top}$, $p = 100$, 500 or 1,000 and $p = 50$, 100 or 150, respectively. Such simulation is repeated 500 rounds, and we summarize the simulation results in Table 1 and 2. From Table 1 and 2, we can find that the proposed estimator is better than other three estimators. Lasso and Elastic Net perform worse than SCAD and our proposed estimators, but better than generalized ridge estimators, and SCAD performs worse than our proposed estimators.

Table 1. L_2-errors of estimators for parameters

		Method				
p	n	Ridge	Thres. Ridge	Lasso	Elastic Net	SCAD
100	50	43.17	24.29	36.48	35.64	31.98
500	100	39.52	22.76	34.15	33.21	27.47
1,000	150	38.38	21.34	30.47	31.75	25.63

Table 2. Results of variable selection for four approaches

p	n	Approach	Average number of zeros estimators	
			Correct	Incorrect
100	50	Thres. Ridge	84.17 [93%]	4.89
		Lasso	76.43 [85%]	8.51
		Elastic Net	74.82 [83%]	9.17
		SCAD	78.24 [87%]	7.12
500	100	Thres. Ridge	467.53 [95%]	16.93
		Lasso	438.67 [89%]	32.47
		Elastic Net	431.35 [88%]	31.28
		SCAD	446.21 [91%]	25.82
1,000	150	Thres. Ridge	931.16 [94%]	39.64
		Lasso	885.38 [89%]	76.25
		Elastic Net	879.52 [89%]	74.18
		SCAD	894.78 [90%]	68.36

Simulation study 2. We consider the Logistic regression model. The generated data are independent and identically distributed, and the conditional distribution of the response Y are given by binomial distribution with probability of success $p(X) = \exp(X^\top\beta)/[1+\exp(X^\top\beta)]$, ε_i is a random error with mean 0 and $\sigma = 1$, and X_i is generated from $N(0, \Sigma)$, where Σ is a $p \times p$ compound symmetry covariance matrix with correlation $\rho = 0.5$. We consider $p = 100$, 500 or 1,000 and $p = 50$, 100 or 150, respectively, and $\beta = (1, 2, -1, 1, 2, 1, 2, -1, 1, 2, 0, \cdots, 0)^\top$. Such simulation is repeated 500 rounds, and we summarize the simulation results in Table 3 and 4. The results in the simulation study 2 are similar to the first one, which support our asymptotic theory.

Table 3. L_2-errors of estimators for parameters

		Method				
p	n	Ridge	Thres. Ridge	Lasso	Elastic Net	SCAD
100	50	48.32	26.38	39.25	40.17	37.42
500	100	46.75	24.13	33.46	32.63	31.58
1,000	150	45.59	22.87	30.61	31.82	28.93

Table 4. Results of variable selection for four approaches

p	n	Approach	Average number of zeros estimators	
			Correct	Incorrect
100	50	Thres. Ridge	83.51 [92%]	5.41
		Lasso	75.24 [84%]	10.26
		Elastic Net	77.18 [85%]	9.39
		SCAD	78.63 [87%]	8.95
500	100	Thres. Ridge	463.32 [94%]	19.37
		Lasso	416.18 [85%]	38.52
		Elastic Net	408.73 [83%]	36.45
		SCAD	428.09 [87%]	29.13
1,000	150	Thres. Ridge	919.25 [92%]	46.71
		Lasso	873.42 [88%]	85.36
		Elastic Net	886.17 [89%]	83.92
		SCAD	883.39 [89%]	79.25

To consider scenarios involving missing data and categorical covariates, we generate data a $(p+1)$ vector X^* form $X^* \sim N(0, \Sigma)$, where $\Sigma = (\rho^{|k-l|})$ for $k, l = 1, \cdots, (p+1)$. The first component of X^* is replaced by a categorical covariate, each taking the value 0 or 1 with 50% probability. Set $(n, p) = (50, 100)$ and $(100, 500)$, and $\beta = (1, -2, 1, 1, -2, 1, 1, -2, 1, 0, \cdots, 0)^\top$. We assume that the missing rate is 5% and 10% respectively, and the simulation results are summarized in Table 5 and 6. The results in this simulation are similar to the above.

Table 5. L_2-errors of estimators for parameters

			Method				
Missing rate	p	n	Ridge	Thres. Ridge	Lasso	Elastic Net	SCAD
5%	100	50	48.69	27.94	39.62	37.56	39.15
	500	100	48.58	26.37	34.51	36.28	32.96
10%	100	50	50.49	29.63	42.38	45.72	41.17
	500	100	51.36	28.52	38.29	39.36	36.45

Table 6. Results of variable selection for four approaches

Missing rate	p	n	Approach	Average number of zeros estimators	
				Correct	Incorrect
5%	100	50	Thres. Ridge	81.36 [89%]	6.87
			Lasso	72.51 [80%]	11.43
			Elastic Net	71.84 [79%]	12.08
			SCAD	74.29 [82%]	10.62
10%	500	100	Thres. Ridge	425.43 [86%]	28.95
			Lasso	382.64 [79%]	47.38
			Elastic Net	390.75 [80%]	45.61
			SCAD	407.82 [83%]	36.54

4. Real data application

We illustrate the proposed approach by using the logistic regression model to the Acute Lymphoblastic Leukemia (ALL) data, which is available from http://www.bioconductor.org/. The ALL data, analyzed by [26], contains gene expression data (from microarrays) for 128 patients with specific leukemia types (T-cell or B-cell).

This study utilizes a curated subset of the ALL dataset comprising 79 B-cell Acute Lymphoblastic Leukemia (B-ALL) patient specimens. The cohort includes 37 cases with BCR/ABL fusion and 42 NEG-classified samples. Given the high proportion of non-expressed genes among the 12,625 initial probesets, we implemented a two-stage gene filtering protocol: (i) Retention required at least 75% of samples to exhibit intensity values no less than 100; (ii) Genes were excluded if their cross-sample intensity coefficient of variation fell outside the range of [0.7, 10]. This process yielded 2,396 functionally relevant genes. Computational constraints preclude direct analysis of all 2,396 genes, therefore, we applied the method of sure independence screening for dimensionality reduction, retaining the 30 probesets demonstrating strongest marginal associations with outcome variables. We adopt a logistic regression framework to cancer classification. By using the proposed method with the thresholding value $a_n = 0.05$, 6 probesets are demonstrated statistically significant nonzero coefficients in the final model. Table 7 summarizes the final feature selection results, providing Affymetrix probe identifiers and regression coefficients for the 6 predictors with nonzero effects.

Table 7. Affymetrix probe identifiers and the proposed estimators

Variables	Thres. Ridge estimators
Intercept	0.2173
31536_at	-0.0875
36131_at	-0.0521
37761_at	0.1264
39837_s_at	-0.2176
40718_at	-0.0839
754_s_at	0.1152

Acknowledgement

We are also grateful to Editor-in-Chief, Associate Editor and two anonymous referees for their insightful comments and suggestions, which led to substantial improvements. This work was supported by Provincial Natural Science Foundation of Hunan (Grant No. 2023JJ30187).

Conflict of interest

The authors declare no competing financial interest.

References

[1] Nelder JA, Wedderburn RWM. Generalized linear models. *Journal of the Royal Statistical Society, Series A*. 1972; 135(3): 370-384.

[2] Fahrmeir L, Kaufmann H. Consistency and asymptotic normality of the maximum likelihood estimator in generalized linear models. *The Annals of Statistics*. 1985; 13(1): 342-368. Available from: https://doi.org/10.1214/aos/1176346597.

[3] Liang KY, Zeger SL. Longitudinal data analysis using generalized linear models. *Biometrika*. 1986; 73(1): 13-22. Available from: https://doi.org/10.1093/biomet/73.1.13.

[4] Lee Y, Nelder JA. Hierarchical generalized linear models. *Journal of the Royal Statistical Society, Series B*. 1996; 58(4): 619-656. Available from: https://doi.org/10.1111/j.2517-6161.1996.tb02105.x.

[5] Guo B, Chen SX. Tests for high dimensional generalized linear models. *Journal of the Royal Statistical Society, Series B*. 2016; 78(5): 1079-1102. Available from: https://doi.org/10.1111/rssb.12152.

[6] Hoerl AE, Kennard RW. Ridge regression: Biased estimation for nonorthogonal problems. *Technometric*. 1970; 12(1): 55-67. Available from: https://doi.org/10.1080/00401706.1970.10488634.

[7] Hoerl E, Kennard RW, Balawin KF. Ridge regression–some simulations. *Communications in Statistics-Theory and Methods*. 1975; 4(2): 105-123. Available from: https://doi.org/10.1080/03610927508827232.

[8] Smith G, Campbell F. A critique of some ridge regression mothods. *Journal of the American Statistical Association*. 1980; 75(369): 74-81. Available from: https://doi.org/10.1080/01621459.1980.10477428.

[9] Lecessie S, Vanhouwelingen JC. Ridge estimators in logistic regression. *Journal of the Royal Statistical Society, Series C*. 1992; 41(1): 191-201. Available from: https://doi.org/10.2307/2347628.

[10] Schaefer RL, Roi LD, Wolfe RA. A ridge logistic estimator. *Communication in Statistics-Theory and Methods*. 1984; 13(1): 99-113. Available from: https://doi.org/10.1080/03610928408828664.

[11] Hearne EM, Mason RL, Clark GM. Ridge regression for Cox model. *Biometrics*. 1985; 41(1): 329-329.

[12] Nyquist H. Restricted estimation of generalized linear models. *Journal of the Royal Statistical Society, Series C*. 1991; 40(1): 133-141. Available from: https://doi.org/10.2307/2347912.

[13] Segerstedt B. On ordinary ridge regression in generalized linear models. *Communication in Statistics-Theory and Methods*. 1992; 21(8): 2227-2246. Available from: https://doi.org/10.1080/03610929208830909.

[14] Janamiri F, Rasekh A, Chaji A, Babadi B. Ridge estimation in linear mixed measurement error models using generalized maximum entropy. *Statistics*. 2022; 56(5): 1095-1112. Available from: https://doi.org/10.1080/02331888.2022.2126474.

[15] Janamiri G. Ridge regularized estimation of VAR models for inference. *Journal of Time Series Analysis*. 2024; 46(2): 235-257. Available from: https://doi.org/10.1111/jtsa.12737.

[16] Tibshirani R. Regression shinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*. 1997; 58(1): 267-288. Available from: https://doi.org/10.1111/j.2517-6161.1996.tb02080.x.

[17] Fan JQ, Li RZ. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Associatoion*. 2001; 96(456): 1348-1360. Available from: https://doi.org/10.1198/016214501753382273.

[18] Shao J, Deng XW. Estimation in high-dimensional linear models with deterministic design matrices. *The Annals of Statistics*. 2012; 40(2): 812-831. Available from: https://doi.org/10.1214/12-AOS982.

[19] Zhang CH, Zhang SS. Confidence interval for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society, Series B*. 2014; 76(1): 217-242. Available from: https://doi.org/10.1111/rssb.12026.

[20] Van de Geer SA. High dimensional generalized linear models and the lasso. *The Annals of Statistics*. 2008; 36(2): 614-645. Available from: https://doi.org/10.1214/009053607000000929.

[21] Fan JQ, Song R. Sure independent screening in generalized linear models with NP-dimensionality. *The Annals of Statistics*. 2010; 38(6): 3567-3604. Available from: https://doi.org/10.1214/10-AOS798.

[22] Zhang CH. Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*. 2010; 38(2): 894-942. Available from: https://doi.org/10.1214/09-AOS729.

[23] Greenm PJ. Iteratively reweighted least squares for maximum likelihood estimation, and some robust and resistant alternatives. *Journal of the Royal Statistical Society, Series B*. 1984; 46(2): 149-192. Available from: https://doi.org/10.1111/j.2517-6161.1984.tb01288.x.

[24] Prontnoy S. Asymptotic behavior of likelihood methods for exponential families when the number of parameters tends to infinity. *The Annals of Statistics*. 1988; 16(1): 356-366. Available from: https://doi.org/10.1214/aos/1176350710.

[25] Avella-Medina M, Ronchetti E. Robust and consistent variable selection in high-dimensional generalized linear models. *Biometrika*. 2018; 105(1): 31-44. Available from: https://doi.org/10.1093/biomet/asx070.

[26] Dudoit S, Keles S, Van der Laan M. Multiple tests of association with biological annotation metadata. *Institute of Mathematical Statistics*. 2008; 2(5): 153-218. Available from: https://doi.org/10.1214/193940307000000446.

[27] Fan JQ, Peng H. Nonconcave penalized likelihood with a diverging number of parameters. *The Annals of Statistics*. 2004; 32(3): 928-961. Available from: https://doi.org/10.1214/009053604000000256.

Appendix

Proof of Theorem 1 (i) the proof of Theorem 1 (i) is similar to Theorem 1 of [21], hence is omitted.

(ii) Let $\tilde{X} = W^{1/2}X$ and r be the rank of $\tilde{X}$. It is easy to show that

$$\mathrm{E}(\hat{\beta}) = \beta \quad \text{and} \quad \mathrm{Cov}(\hat{\beta}) = (X^\top WX)^{-1} = (\tilde{X}\tilde{X})^{-1}. \tag{10}$$

Then, we have

$$\begin{aligned}
\mathrm{bias}(\hat{\beta}^\Lambda) &= \mathrm{E}(\hat{\beta}^\Lambda) - \beta \\
&= (X^\top WX + Q\Lambda Q^\top)^{-1}X^\top WX\beta - \beta \\
&= -(\tilde{X}^\top\tilde{X} + Q\Lambda Q^\top)^{-1}Q\Lambda Q^\top\beta \\
&= -Q(Q^\top\tilde{X}^\top\tilde{X}Q + Q^\top Q\Lambda Q^\top Q)^{-1}\Lambda Q^\top\beta \\
&= -Q(D+\Lambda)^{-1}\Lambda Q^\top\beta,
\end{aligned}$$

and

$$\begin{aligned}
\mathrm{Cov}(\hat{\beta}^\Lambda) &= (X^\top WX + Q\Lambda Q^\top)^{-1}X^\top WX(X^\top WX + Q\Lambda Q^\top)^{-1} \\
&\leq (X^\top WX + Q\Lambda Q^\top)^{-1} \\
&\leq (Q\Lambda Q^\top)^{-1} = Q\Lambda^{-1}Q^\top,
\end{aligned}$$

where $\mathrm{A} \geq \mathrm{B}$ means that $(\mathrm{A}-\mathrm{B})$ is nonnegative definite for nonnegative definite matrices A and B. Because Q is a orthogonal matrix, we can obtain

$$\|Q(D+\Lambda)^{-1}\Lambda\| = \|(D+\Lambda)^{-1}\Lambda\| \quad \text{and} \quad \|Q^\top\beta\| = \|\beta\|,$$

where $\|\cdot\|$ denotes the L_2-norm. Thus

$$\|\mathrm{bias}(\hat{\beta}^\Lambda)\| \leq \|(D+\Lambda)^{-1}\Lambda\|\|\beta\| \leq \rho\|\beta\| \leq \frac{\lambda_{\max}}{d}\|\beta\|,$$

where $\rho = \max\{\lambda_1/(d_1+\lambda_1), \cdots, \lambda_p/(d_p+\lambda_p)\}$. By Assumption 5, we have

$$\left[\mathrm{I}^\top \mathrm{bias}(\hat{\beta}^\Lambda)\right]^2 \leq \|\mathrm{I}\| \|\mathrm{bias}(\hat{\beta}^\Lambda)\|^2 = O(\lambda_{\max}^2 n^{-2\eta}).$$

Similarly, we can obtain that

$$\|Q\Lambda^{-1}Q^\top\| = \|\Lambda^{-1}\| \quad \text{and} \quad \mathrm{I}^\top \mathrm{Cov}(\hat{\beta}^\Lambda)\mathrm{I} = O\left(\frac{1}{\lambda_{\min}}\right).$$

Therefore, the result of Theorem 1 (ii) follows from

$$\mathrm{E}(\mathrm{I}^\top \hat{\beta}^\Lambda - \mathrm{I}^\top \beta)^2 = [\mathrm{I}^\top \mathrm{bias}(\hat{\beta}^\Lambda)]^2 + \mathrm{I}^\top \mathrm{Cov}(\hat{\beta}^\Lambda)\mathrm{I}.$$

□

Proof of Theorem 2 We first expand

$$\begin{aligned}
\mathrm{MSE}(\hat{\beta}^\Lambda) &= \mathrm{E}\left\{(\hat{\beta}^\Lambda - \beta)^\top(\hat{\beta}^\Lambda - \beta)\right\} \\
&= \mathrm{E}\left\{(\tilde{X}^\top\tilde{X} + Q\Lambda Q^\top)^{-1}\tilde{X}^\top\tilde{X}\hat{\beta} - \beta)^\top((\tilde{X}^\top\tilde{X} + Q\Lambda Q^\top)^{-1}\tilde{X}^\top\tilde{X}\hat{\beta} - \beta)\right\} \\
&= \mathrm{E}\left\{(\tilde{X}^\top\tilde{X} + Q\Lambda Q^\top)^{-1}\tilde{X}^\top\tilde{X}(\hat{\beta} - \beta)\tilde{X}^\top\tilde{X}(\tilde{X}^\top\tilde{X} + Q\Lambda Q^\top)^{-1}\right\} \\
&\quad + \left((\tilde{X}^\top\tilde{X} + Q\Lambda Q^\top)^{-1}\tilde{X}^\top\tilde{X}\hat{\beta} - \beta\right)^\top \left((\tilde{X}^\top\tilde{X} + Q\Lambda Q^\top)^{-1}\tilde{X}^\top\tilde{X}\hat{\beta} - \beta\right) \\
&= \mathrm{trace}\left\{(\tilde{X}^\top\tilde{X} + Q\Lambda Q^\top)^{-1}\tilde{X}^\top\tilde{X}(\tilde{X}^\top\tilde{X} + Q\Lambda Q^\top)^{-1}\right\} \\
&\quad + \beta^\top \left((\tilde{X}^\top\tilde{X} + Q\Lambda Q^\top)^{-1}\tilde{X}^\top\tilde{X} - \mathrm{I}_p\right)^\top \left((\tilde{X}^\top\tilde{X} + Q\Lambda Q^\top)^{-1}\tilde{X}^\top\tilde{X} - \mathrm{I}_p\right)\beta \\
&= A_1 + A_2.
\end{aligned}$$

Because $\tilde{X}^\top\tilde{X} = QDQ^\top$ and $(\tilde{X}^\top\tilde{X} + Q\Lambda Q^\top)^{-1}\tilde{X}^\top\tilde{X} = Q(D+\Lambda)^{-1}DQ^\top$, we can show that

$$A_1 = \mathrm{trace}\left\{Q(D+\Lambda)^{-1}D(D+\Lambda)^{-1}Q^\top\right\}$$

Note that $D = \mathrm{diag}\{d_1, \cdots, d_p\}$, $\Lambda = \mathrm{diag}\{\lambda_1, \cdots, \lambda_p\}$ and Q is a orthogonal matrix, we have

$$
\begin{aligned}
A_1 &= \text{trace}\left\{QD(D+\Lambda)^{-2}Q^\top\right\}\\
&= \text{trace}\left\{D(D+\Lambda)^{-2}Q^\top Q\right\}\\
&= \text{trace}\left\{D(D+\Lambda)^{-2}\right\}\\
&= \sum_{j=1}^{p}\frac{d_j}{(d_j+\lambda_j)^2}.
\end{aligned}
$$

Let $\beta^\top Q = (a_1, \cdots, a_p)$. From the facts that

$$
(\tilde{X}^\top\tilde{X}+Q\Lambda Q^\top)^{-1}\tilde{X}^\top\tilde{X}-\mathrm{I}_p = -(\tilde{X}^\top\tilde{X}+Q\Lambda Q^\top)^{-1}Q\Lambda Q^\top,
$$

we can obtain that

$$
\begin{aligned}
A_2 &= \beta^\top\left[(\tilde{X}^\top\tilde{X}+Q\Lambda Q^\top)^{-1}Q\Lambda Q^\top\right]^\top\left[(\tilde{X}^\top\tilde{X}+Q\Lambda Q^\top)^{-1}Q\Lambda Q^\top\right]\beta\\
&= \beta^\top\left[(Q(D+\Lambda)Q^\top)^{-1}Q\Lambda Q^\top\right]^\top\left[(Q(D+\Lambda)Q^\top)^{-1}Q\Lambda Q^\top\right]\beta\\
&= \beta^\top\left[(Q(D+\Lambda)^{-1}\Lambda Q^\top\right]^\top\left[(Q(D+\Lambda)^{-1}\Lambda Q^\top\right]\beta\\
&= \beta^\top Q\Lambda(D+\Lambda)^{-2}\Lambda Q^\top\beta\\
&= \sum_{j=1}^{p}a_j^2\frac{\lambda_j^2}{(d_j+\lambda_j)^2}.
\end{aligned}
$$

Hence,

$$
\mathrm{MSE}(\hat{\beta}^\Lambda) = A_1 + A_2 = \sum_{j=1}^{p}\frac{d_j+a_j^2\lambda_j^2}{(d_j+\lambda_j)^2},
$$

$$
\mathrm{MSE}(\hat{\beta}^\lambda) = \sum_{j=1}^{p}\frac{d_j+a_j^2\lambda}{(d_j+\lambda)^2} \quad \text{and} \quad \mathrm{MSE}(\hat{\beta}) = \sum_{j=1}^{n}\frac{1}{d_j}.
$$

Let $\Phi = \min\left\{\frac{2d_j}{a_j^2 d_j - 1}, \cdots, \frac{2d_j}{a_j^2 d_j - 1}\right\}$. For $0 < \lambda < \Phi$,

$$
\begin{aligned}
\mathrm{MSE}(\hat{\beta}) - \mathrm{MSE}(\hat{\beta}^{\lambda}) &= \sum_{j=1}^{n} \frac{1}{d_j} - \sum_{j=1}^{p} \frac{d_j + a_j^2 \lambda}{(d_j + \lambda)^2} \\
&= \sum_{j=1}^{n} \frac{a_j^2 d_j \lambda^2 - 2\lambda d_j - \lambda^2}{d_j (d_j + \lambda)^2} \\
&\geq 0.
\end{aligned}
\tag{11}
$$

Define $f_j(t) = (d_j + a_j^2 t^2)/((d_j + t)^2)$, $j = 1, \cdots, p$, which is a continuous function. For sufficiently large constant T_0, there exists a t_{0j} ($t_{0j} \in [0, T_0]$) such that

$$
f_j(t_{0j}) \leq f_j(t) = \frac{d_j + a_j^2 t^2}{(d_j + t)^2} \quad \text{for all} \quad t \in [0, T_0], \quad j = 1, \cdots, p.
$$

Let $\Lambda = \mathrm{diag}\{\lambda_{01}, \cdots, \lambda_{0p}\}$ and

$$
\lambda_{0j} = \begin{cases} \lambda & \text{if } \dfrac{d_j + a_j^2 \lambda}{(d_j + \lambda)^2} = f_j(t_{0j}), \\[2ex] t_{0j} & \text{if } \dfrac{d_j + a_j^2 \lambda}{(d_j + \lambda)^2} > f_j(t_{0j}). \end{cases}
$$

Hence,

$$
\mathrm{MSE}(\hat{\beta}^{\lambda}) - \mathrm{MSE}(\hat{\beta}^{\Lambda}) = \sum_{j=1}^{p} \frac{d_j + a_j^2 \lambda}{(d_j + \lambda)^2} - \sum_{j=1}^{p} \frac{d_j + a_j^2 \lambda_{oj}}{(d_j + \lambda_{oj})^2} \geq 0. \tag{12}
$$

By combining (11) and (12), we obtain that $\mathrm{MSE}(\hat{\beta}^{\Lambda}) \leq \mathrm{MSE}(\hat{\beta}^{\lambda}) \leq \mathrm{MSE}(\hat{\beta})$. This completes the proof. □

Proof of Theorem 3 Because $\|\mathrm{bias}(\hat{\beta}^{\Lambda})\| \leq (\lambda_{\max}/d)\|\beta\|$ and $\|\mathrm{Cov}(\hat{\beta}^{\Lambda})\| = O(1/\lambda_{\min})$, we can obtain that $\mathrm{var}(\hat{\beta}_j^{\Lambda})$ $= O(1/\lambda_{\min})$ and

$$
\mathrm{bias}(\hat{\beta}_j^{\Lambda}) = O\left(\frac{\lambda_{\max}}{d}\|\beta\|\right) = O(\lambda_{\max} n^{-\eta}), \quad \text{for} \quad j = 1, \cdots, n.
$$

For some small constant $\nu > 0$,

$$
\begin{aligned}
\frac{\text{bias}(\hat{\beta}_j^{\Lambda})}{(u_n-1)a_n} &= O\left(\frac{\lambda_{\max}}{n^{\eta}(u_n-1)a_n}\right)\\
&= O\left(\frac{(\log\log n)^4\log(n\vee p)}{n^{\eta}a_n^3}\right)\\
&= O\left(\frac{(\log\log n)^4}{n^{\eta-\nu-3\alpha}}\right).
\end{aligned}
$$

For sufficiently large n, $\log\log n > 0$. When $\alpha < (\eta-\nu)/3$, we have $\text{bias}(\hat{\beta}_j^{\Lambda})/(u_n-1)a_n \to 0$ for $j = 1, \cdots, p$. Hence, there exists constants $c_1 > 0$ and $c_2 > 0$ such that

$$
\frac{\text{bias}(\hat{\beta}_j^{\Lambda})-(u_n-1)a_n}{[\text{var}\hat{\beta}_j^{\Lambda}]^{1/2}} \leq -\frac{\sqrt{2}c_1\sqrt{\lambda_{\min}}a_n}{(\log\log n)}
$$

and

$$
\frac{\text{bias}(\hat{\beta}_j^{\Lambda})-(1-u_n^{-1})a_n}{[\text{var}\hat{\beta}_j^{\Lambda}]^{1/2}} \leq -\frac{\sqrt{2}c_2\sqrt{\lambda_{\min}}a_n}{(\log\log n)}.
$$

Let $\hat{\beta}_j^{\Lambda}$ be the j-th component of $\hat{\beta}^{\Lambda}$ for $j = 1, \cdots, p$. If we can show that $\hat{\beta}_j^{\Lambda}$ is normally distributed, we have

$$
\begin{aligned}
P(|\hat{\beta}_j^{\Lambda}-\beta_j| > (u_n-1)a_n) &\leq 2\Phi\left(\frac{|\text{bias}(\hat{\beta}_j^{\Lambda})|-(u_n-1)a_n}{[\text{var}\hat{\beta}_j^{\Lambda}]^{1/2}}\right)\\
&\leq \exp\left\{-\frac{c_1^2\lambda_{\min}a_n^2}{(\log\log n)^2}\right\}
\end{aligned}
$$

and

$$
\begin{aligned}
P(|\hat{\beta}_j^{\Lambda}-\beta_j| > (1-u_n^{-1})a_n) &\leq 2\Phi\left(\frac{|\text{bias}(\hat{\beta}_j^{\Lambda})|-(1-u_n^{-1})a_n}{[\text{var}\hat{\beta}_j^{\Lambda}]^{1/2}}\right)\\
&\leq \exp\left\{-\frac{c_2^2\lambda_{\min}a_n^2}{(\log\log n)^2}\right\},
\end{aligned}
$$

where Φ is the standard normal distribution function. Analogous to the proof process of Theorem 2 in [18], for any $t > 0$, we have

$$P(\mathscr{A}_{\beta,\, a_n u_n} \subset \mathscr{A}_{\hat{\beta}^{\Lambda},\, a_n}) \geq 1 - P\left(\bigcup_{j\,:\,|\beta_j| > a_n u_n} \{(u_n - 1)a_n\} \right)$$

$$\geq 1 - (n \vee p)^{-t} \tag{13}$$

and

$$P(\mathscr{A}_{\hat{\beta}^{\Lambda},\, a_n} \subset \mathscr{A}_{\beta,\, a_n/u_n}) \geq 1 - P\left(\bigcup_{j\,:\,|\beta_j| \leq a_n/u_n} \{(1 - u_n^{-1})a_n\} \right)$$

$$\geq 1 - (n \vee p)^{-t}. \tag{14}$$

According to (13)-(14) and the assumption of $0 < p/n < +\infty$, we can also obtain that

$$P(\mathscr{A}_{\beta,\, a_n u_n} \subset \mathscr{A}_{\hat{\beta}^{\Lambda},\, a_n} \subset \mathscr{A}_{\beta,\, a_n/u_n}) = 1 - O(n^{-t}). \tag{15}$$

Next we verify that $\hat{\beta}_j^{\Lambda}$ is normally distributed for $j = 1, \cdots, p$. Define

$$l(X^{\top}\beta, Y) = l(\beta_1, \cdots, \beta_n;\ x_1, \cdots, x_n;\ y_1, \cdots, y_n)$$

and

$$\Re = \frac{1}{2}(\hat{\beta} - \beta_0)^{\top} \nabla^2 \left(\frac{\partial l(X^{\top}\beta^*, Y)}{\partial \beta} \right)^{\top} (\hat{\beta} - \beta_0),$$

where β^* is between $\hat{\beta}$ and β_0. By some simple calculation, we have

$$\frac{\partial^2 l(X^{\top}\beta_0, Y)}{p^{1+\delta}\partial\beta\partial\beta^{\top}} = G \quad \text{and} \quad \frac{\partial^3 l(X^{\top}\beta^*, Y)}{\partial\beta_j\partial\beta_k\partial\beta_l} = -\frac{1}{n}\sum_{i=1}^{n} b^{(3)}(X_i^{\top}\beta^*)X_{ij}X_{ik}X_{il}.$$

As $\hat{\beta}(\Lambda)$ satisfy the equation $\partial l^{\Lambda}(X^{\top}\beta, Y)/\partial\beta = 0$, based on the Taylor expansion on $\partial l^{\Lambda}(X^{\top}\beta, Y)/\partial\beta$ at point β, we can show that

$$\hat{\beta} - \beta_0 = -G^{-1}\frac{\partial l(X^{\top}\beta_0, Y)}{p^{1+\delta}\partial\beta} - \frac{1}{p^{1+\delta}}G^{-1}\Re. \tag{16}$$

By Assumptions 3, 6 and 7,

$$\left|\left|\frac{1}{p^{1+\delta}}G^{-1}\right|\right|^2 \leq \frac{1}{2p^2}\gamma_{max}^2(G^{-1})\|\Re\|^2$$

$$= \frac{1}{2p^{2+2\delta}}\gamma_{max}^2(G^{-1})\sum_{j=1}^{p^{2+2\delta}}\left[(\beta^*-\beta_0)^\top\frac{1}{n}\sum_{i=1}^{n}X_{ij}b^{(3)}(X_i^\top\beta^*)X_iX_i^\top(\beta^*-\beta_0)\right]^2$$

$$\leq \frac{p}{2p^{2+2\delta}}C_{00}^2\gamma_{max}^2(G^{-1})\gamma_{max}^2\left(\frac{1}{n}\mathbf{X}^\top\mathbf{X}\right)\|\beta^*-\beta_0\|^4$$

$$= o_p\left(\frac{1}{p}\right), \tag{17}$$

where $\gamma_{max}(G^{-1})$ and $\gamma_{max}(1/n\mathbf{X}^\top\mathbf{X})$ are the largest eigenvalue of G^{-1} and $1/n\mathbf{X}^\top\mathbf{X}$ respectively. Let $e_j = (0, \cdots, 1, \cdots, 0)^\top$ be a unit vector in $\mathbb{R}^p$ with the j-th entry 1 and 0 elsewhere, $j = 1, \cdots, p$. Define

$$Z_{nij} = \frac{1}{\sqrt{n}}e_jG^{-1}\frac{\partial \log f(y_i; \sum_{k=1}^{p}\beta_{0k}x_{ik})}{p^{1+\delta}\partial\beta}, \quad i = 1, \cdots, n, \quad j = 1, \cdots, p.$$

Similar to the proof of Theorem 2 in [27], we can obtain that Z_{nij} satisfies the Lindeberg-Feller conditions and it has an asymptotic normal distribution. Therefore, by combining (16) and (17), $\hat{\beta}_j$ is normally distributed for $j = 1, \cdots, p$. Finally, according to (5) and the asymptotic normality of $\hat{\beta}_j$, we can show that $\hat{\beta}_j^\Lambda$, $j = 1, \cdots, p$, is normally distributed. The proof of Theorem 3 is completed. □

Proof of Theorem 4 Let $A_n = \{\mathscr{A}_{\hat{\beta}^\Lambda, a_n} = \mathscr{A}_{\beta, a_n}\}$ and A_n^c be its complement. For any $\varepsilon > 0$,

$$P(\|\tilde{\beta}^\Lambda - \beta_0\| > \varepsilon) = p(\|\tilde{\beta}^\Lambda - \beta_0\| > \varepsilon, A_n) + p(\|\tilde{\beta}^\Lambda - \beta_0\| > \varepsilon, A_n^c)$$

$$\leq p(\|\tilde{\beta}^\Lambda - \beta_0\| > \varepsilon, A_n) + p(A_n^c). \tag{18}$$

$$P(\|\tilde{\beta}^\Lambda - \beta_0\| > \varepsilon, A_n) \leq P(\|\tilde{\beta}_{n1}^\Lambda - \beta_{n10}\| > \varepsilon, A_n) + P(\|\tilde{\beta}_{n2}^\Lambda - \beta_{n20}\| > \varepsilon, A_n) \tag{19}$$

By Chebyshev's inequality, we have

$$P(\|\tilde{\beta}_{n2}^\Lambda - \beta_{n20}\| > \varepsilon, A_n) \leq \frac{\mathrm{E}\|\tilde{\beta}_{n2}^\Lambda - \beta_{n20}\|^2}{\varepsilon^2}I_{A_n} \tag{20}$$

where I_{A_n} is the indicator of the set. Because $\tilde{\beta}_{n2}^\Lambda = 0$, we can obtain

$$\frac{\mathrm{E}\|\tilde{\beta}_{n2}^{\Lambda}-\beta_{n20}\|^2}{\varepsilon^2}I_{A_n}=\left\{\frac{E\left((\tilde{\beta}_{n2}^{\Lambda}-\beta_{n20})^{\top}(\tilde{\beta}_{n2}^{\Lambda}-\beta_{n20})\right)}{\varepsilon^2}\right\}I_{A_n}$$

$$\leq\frac{qa_n^2}{\varepsilon^2}. \tag{21}$$

From the proof of Theorem 3, we have

$$P(|\tilde{\beta}_{n1i}^{\Lambda}-\beta_{n10i}|>\varepsilon,A_n)\leq\exp\left\{-c_1^2\lambda_{\min}\varepsilon^2\right\}\quad i=1,\cdots,s,$$

where c_1 is given in the proof of Theorem 3. Therefore,

$$P(\|\tilde{\beta}_{n1}^{\Lambda}-\beta_{n10}\|>\varepsilon,A_n)\leq\sum_{i=1}^{q}p(|\tilde{\beta}_{n1i}^{\Lambda}-\beta_{n10i}|>\varepsilon,A_n)$$

$$\leq q\exp\left\{-c_1^2\lambda_{\min}\varepsilon^2\right\}. \tag{22}$$

Theorem 3 implies that

$$P\{\mathscr{A}_{\hat{\beta}^{\Lambda},\,a_n}=\mathscr{A}_{\beta,\,a_n}\}=1-O(n^{-t}),$$

so we have, for $\forall\, t>0$,

$$P(A_n^c)=O(n^{-t}). \tag{23}$$

By combining (18)-(23), we can establish that $P(\|\tilde{\beta}^{\Lambda}-\beta_0\|>\varepsilon)\to 0$ as $n\to\infty$, and the proof is completed. □