Penalized Principal Logistic Regression for Sparse Sufficient Dimension Reduction

Seung Jun Shin and Andreas Artemiou
Korea University and Cardiff University

Abstract

Sufficient dimension reduction (SDR) is a successive tool for reducing the dimensionality of predictors by finding the central subspace, a minimal subspace of predictors that preserves all the regression information. When predictor dimension is large, it is often assumed that only a small number of predictors is informative. In this regard, sparse SDR is desired to achieve variable selection and dimension reduction simultaneously. We propose a principal logistic regression (PLR) as a new SDR tool and extend it to a penalized version for sparse SDR. Asymptotic analysis shows that the penalized PLR enjoys the oracle property. Numerical investigation supports the advantageous performance of the proposed methods.

Keywords: max-SCAD penalty, principal logistic regression, sparse sufficient dimension reduction, sufficient dimension reduction

1. Introduction

It is often of primary interest to identify the relationship between the univariate response Y and the p-dimensional predictor $\mathbf{X} \in \mathbb{R}^p$. Sufficient dimension reduction (SDR) efficiently reduces the dimensionality of $\mathbf{X}$ by finding a lower dimensional subspace of span($\mathbf{X}$) while preserving regression information in $\mathbf{X}$. Specifically, SDR seeks a matrix $\mathbf{B} = (\mathbf{b}_1, \cdots, \mathbf{b}_d) \in \mathbb{R}^{p\times d}$ that satisfies

$$Y \perp \mathbf{X} | \mathbf{B}^\top \mathbf{X}, \tag{1}$$

where $\perp$ denotes statistical independence. Compared to conventional parametric models, (1) is less stringent since it does not assume any specific link functions between Y and $\mathbf{X}$. The space spanned by $\mathbf{B}$ satisfying (1) is called the dimension reduction subspace (DRS). The central subspace, denoted by $\mathcal{S}_{Y|\mathbf{X}}$ is defined as the intersection of all DRSes, and hence it is the lowest dimensional DRS. **(author?)** [1] showed that $\mathcal{S}_{Y|\mathbf{X}}$ uniquely exists under mild conditions. In SDR, it is assumed that $\mathcal{S}_{Y|\mathbf{X}} = \text{span}(\mathbf{B})$ to make $\mathbf{B}$ an identifiable target. The dimension d of $\mathcal{S}_{Y|\mathbf{X}}$ is referred to as structural dimension, another important quantity to be inferred from the data.

Since the seminal paper on sliced inverse regression [SIR, 2], there have been various methods developed to estimate $\mathcal{S}_{Y|\mathbf{X}}$, which include but are not limited to sliced averaged variance estimation [SAVE, 3], directional regression [DR, 4], sliced regression [5], contour regression [6], and principal support vector machine [PSVM, 7].

Among many others, PSVM is a recently developed SDR method and brings new insight by connecting SDR to penalized machine learning methods such as the support vector machine (SVM). The idea of PSVM is simple as follows. First, dichotomize the continuous response Y by introducing a pseudo response $\tilde{Y} = 1$ if Y is greater than a given cutoff value r, and -1 otherwise. A sequence of linear SVMs are then repeatedly trained for $(\tilde{Y}_r, \mathbf{X})$ as varying the cutoff value r. **(author?)** [7] showed that normals of the optimal hyperplanes from the linear SVMs lie on $\mathcal{S}_{Y|\mathbf{X}}$ regardless of the value of r. Finally, $\mathcal{S}_{Y|\mathbf{X}}$ can be recovered by the spectral decomposition of these normals. PSVM is known to perform better than classical SDR methods such as SIR, and it tackles both linear and nonlinear SDR in a unified framework via kernel trick, as SVM does.

In this article, we propose a principal logistic regression (PLR) as an alternative to PSVM. Namely, we apply the logistic regression to $(\tilde{Y}, \mathbf{X})$ instead of SVM. The advantages of the logistic regression over SVM are obvious since its loss function is smooth and strictly convex (see Figure 1). PLR not only entails simpler asymptotic results under less stringent conditions but also is computationally stable. It is important to note that PLR is not a parametric method

for SDR since we replace the loss in population level and a target of estimation changes.

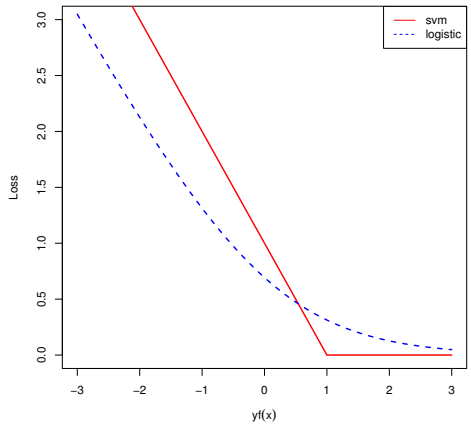


Figure 1: SVM (hinge) loss versus logistic (binomial log-likelihood) loss: The (dashed blue) logistic loss is a smooth, convex, and continuously differentiable function while the (solid red) SVM hinge loss is not.

Sparse SDR that seeks a sparse representation of the basis of $\mathcal{S}_{Y|\mathbf{X}}$ is often desired to achieve the dimension reduction and variable selection simultaneously. Sparse SDR facilitates the interpretation of the results and improves the estimation accuracy by eliminating negligible uncertainties from the predictors with weak signals [8]. Toward sparse SDR, several methods have been proposed. See, for example, **(author?)** [8], **(author?)** [9], **(author?)** [10] and **(author?)** [11].

Sparse SDR assumes a unique partition $\mathbf{X} = (\mathbf{X}_+^\top, \mathbf{X}_-^\top)$ that satisfies

$$Y \perp \mathbf{X}_- | \mathbf{X}_+, \tag{2}$$

where $\mathbf{X}_+ \in \mathbb{R}^q$ and $\mathbf{X}_- \in \mathbb{R}^{p-q}$ for some $q(\ll p)$ [12, 9]. We call $\mathbf{X}_+$ and $\mathbf{X}_-$ relevant and irrelevant variables, respectively. Without loss of generality we assume that the first q predictors are the relevant ones throughout this article. Under (1) and (2), the last $p-q$ rows of $\mathbf{B}$ are all zeros, which makes $\mathbf{B}$ sparse and has an identical sparsity structure across different columns. That is, the last $p-q$ elements of $\mathbf{B}$ are zeros regardless of the cutoff values. In order to

preserve such sparsity structure, we employ a max-SCAD penalty. The SCAD penalty [13] is known to enjoy the oracle property, but its computation is more challenging due to its nonconvexity. However, the logistic loss can minimize the additional computational burden thanks to its smoothness. As a result, we establish the oracle property of the max-SCAD penalized PLR and develop an efficient algorithm for its sample estimation.

The rest of the article is organized as follows. In Section 2 we propose PLR and describe related details including its sample estimation, asymptotic properties, and structural dimension estimation. The penalized PLR is developed in Section 3 in which we establish its oracle property and develop an efficient algorithm for the sample estimation. In Section 4, simulation studies are carried out to investigate finite sample performances of both PLR and the penalized PLR, and real data analysis results are given in Section 5. Final discussions follow in Section 6. All the technical proofs are relegated to Appendix.

2. Principal Logistic Regression For SDR

2.1. Principal Logistic Regression

We start by briefly introducing PSVM which motivates PLR. For a pair of random variables $(Y, \mathbf{X})$, **(author?)** [7] proposed PSVM by solving the following optimization problem:

$$(a_{0,r}, \mathbf{b}_{0,r}) = \underset{a,\mathbf{b}}{\operatorname{argmin}}\, \mathbf{b}^\top \boldsymbol{\Sigma} \mathbf{b} + CE\left[\left|1 - \tilde{Y}_r\{a + \mathbf{b}^\top(\mathbf{X} - E(\mathbf{X}))\}\right|_+\right], \qquad (3)$$

where $|u|_+ = \max\{0, u\}$, $\boldsymbol{\Sigma} = \mathrm{Var}(\mathbf{X})$, and $\tilde{Y}_r$ denotes an artificially dichotomized response having 1 if $Y < r$ and -1 otherwise for a given cutoff value r. A fixed positive constant C is a cost parameter. Notice that (3) is akin to the linear SVM for $(\tilde{Y}_r, \mathbf{X})$. **(author?)** [7] showed that $\mathbf{b}_{0,r} \in \mathcal{S}_{Y|\mathbf{X}}$ for any cutoff r, and thus $\mathrm{span}\{\mathbf{b}_{0,1}, \cdots, \mathbf{b}_{0,h}\} \subseteq \mathcal{S}_{Y|\mathbf{X}}$ where $\mathbf{b}_{0,k}$ denote the minimizer of (3) when $r = r_k, k = 1, \cdots, h$ with $r_1 < \cdots < r_h$ being an arbitrarily given grid of r. **(author?)** [7] assumed the coverage condition that $\mathrm{span}\{\mathbf{b}_{0,1}, \cdots, \mathbf{b}_{0,h}\} = \mathcal{S}_{Y|\mathbf{X}}$

whenever $\text{span}\{\mathbf{b}_{0,1}, \cdots, \mathbf{b}_{0,h}\} \subseteq \mathcal{S}_{Y|\mathbf{X}}$. The coverage condition is known to be held in practice [14].

Motivated by PSVM, we propose PLR by replacing the hinge loss in (3) with the logistic one. As shown in Figure 1 the logistic loss can be regarded as a smooth approximation of the non-differentiable hinge loss function of SVM. Now, the PLR objective function is given by

$$\Lambda_r(\boldsymbol{\theta}) = \boldsymbol{\beta}^\top \boldsymbol{\Sigma} \boldsymbol{\beta} + CE\left[\ln\left(1 + e^{-\tilde{Y}_r\{\alpha + \boldsymbol{\beta}^\top(\mathbf{X} - E(\mathbf{X}))\}}\right)\right]. \tag{4}$$

where $\boldsymbol{\theta}^\top = (\alpha, \boldsymbol{\beta}^\top)$. The PLR objective function (4) is akin to that of the linear kernel logistic regression [15] where its name comes from. Let $\boldsymbol{\theta}_{0,r}^\top = (\alpha_{0,r}, \boldsymbol{\beta}_{0,r}^\top)$ denote the minimizer of (4), which we call PLR solution. Theorem 1 provides a theoretical foundation of PLR for SDR.

Theorem 1. *Under the linearity condition, $\boldsymbol{\beta}_{0,r} \in \mathcal{S}_{Y|\mathbf{X}}$ for an arbitrary given cutoff r.*

Theorem 1 establishes the unbiasedness of $\boldsymbol{\beta}_{0,r}$ for SDR defined in (1). Given a grid $r_1 < \cdots < r_h$ where $h > d$, let $\boldsymbol{\theta}_{0,k}^\top = (\alpha_{0,k}, \boldsymbol{\beta}_{0,k}^\top) = \text{argmin}_{\boldsymbol{\theta}} \Lambda_k(\boldsymbol{\theta})$ where $\Lambda_k(\boldsymbol{\theta}) = \Lambda_{r_k}(\boldsymbol{\theta}), k = 1, \cdots, h$. By Theorem 1, we have $\text{span}\{\boldsymbol{\beta}_{0,1}, \cdots, \boldsymbol{\beta}_{0,h}\} = \mathcal{S}_{Y|\mathbf{X}}$ under the coverage condition.

The linearity condition states that $\text{E}(\mathbf{X}|\mathbf{B}^\top\mathbf{X})$ is a linear function of $\mathbf{B}^\top\mathbf{X}$ where $\mathbf{B}$ is defined in (1), and it implies $\text{E}(\boldsymbol{\beta}^\top\mathbf{X}|\mathbf{B}^\top\mathbf{X}) = \boldsymbol{\beta}^\top\mathbf{P}_{\mathbf{B}}(\boldsymbol{\Sigma})\mathbf{X}$ where $\mathbf{P}_{\mathbf{B}}(\boldsymbol{\Sigma}) = \mathbf{B}(\mathbf{B}^\top\boldsymbol{\Sigma}\mathbf{B})^{-1}\mathbf{B}^\top\boldsymbol{\Sigma}$. The linearity condition plays an essential role and is routinely assumed in many SDR methods. We remark that the linearity condition is not testable but is known to be held when $\mathbf{X}$ is elliptically symmetric [16, 17] or p is large [18]. We remark that PLR is still a model-free approach since the linearity condition restricts the marginal distribution of $\mathbf{X}$ only.

In the classification context, SVM is often preferred to the logistic regression since the logistic regression is fully parametric and fails to recover true classification rule if the model assumption is violated. However, PLR replace the loss function in the population level (4) and is free from the model misspecification.

Robustness is another reason for popularity of SVM. We note, however, that the relation between the logistic loss and the hinge loss is essentially different from that between the squared loss for mean regression and the absolute deviance loss for median regression in the sense that the magnitude of their difference is nearly fixed as the margin gets away from the origin. Therefore PLR can perform comparably well under the presence of outliers. Consequently, PLR and PSVM show nearly identical performance while PLR enjoys additional benefits from smoothness of its loss function, which motivates PLR.

2.2. Sample Estimation

Given a set of data $(y_i, \mathbf{x}_i), i = 1, \cdots, n$, a sample version of $\Lambda_k(\boldsymbol{\theta})$ for $r_k, k = 1, \cdots, h$ is

$$\hat{\Lambda}_k(\boldsymbol{\theta}) = \boldsymbol{\beta}^\top \widehat{\boldsymbol{\Sigma}} \boldsymbol{\beta} + \frac{C}{n} \sum_{i=1}^{n} \ln\left(1 + e^{-\tilde{y}_{ik}\{\alpha + \boldsymbol{\beta}^\top (\mathbf{x}_i - \bar{\mathbf{x}})\}}\right), \qquad k = 1, \cdots, h, \quad (5)$$

where $\bar{\mathbf{x}}$ and $\widehat{\boldsymbol{\Sigma}}$ denote the sample covariance matrix of the predictors, respectively; and $\tilde{y}_{ik} = 1$ if $y_i > r_k$ and -1 otherwise. Let the minimizer of (5) denote $\hat{\boldsymbol{\theta}}_k^\top = (\hat{\alpha}_k, \hat{\boldsymbol{\beta}}_k^\top)$. To obtain $\hat{\boldsymbol{\theta}}_k^\top$, we consider linear transformations $\boldsymbol{\eta} = \widehat{\boldsymbol{\Sigma}}^{1/2} \boldsymbol{\beta}$ and $\tilde{\mathbf{x}}_i = \widehat{\boldsymbol{\Sigma}}^{-1/2} (\mathbf{x}_i - \bar{\mathbf{x}})$, then (5) becomes

$$\boldsymbol{\eta}^\top \boldsymbol{\eta} + \frac{C}{n} \sum_{i=1}^{n} \ln\left(1 + e^{-\tilde{y}_{ik}(\alpha + \boldsymbol{\eta}^\top \tilde{\mathbf{x}}_i)}\right) \quad (6)$$

which is equivalent to the objective function of the linear kernel logistic regression [15] with respect to $(\alpha, \boldsymbol{\eta})$. Now, we have $\hat{\boldsymbol{\beta}}_k = \widehat{\boldsymbol{\Sigma}}^{-1/2} \hat{\boldsymbol{\eta}}_k$ where $(\hat{\alpha}_k, \hat{\boldsymbol{\eta}}_k)$ denotes the minimizer of (6). Finally, $\mathbf{B} = \text{span}(\widehat{\mathbf{V}})$ where $\widehat{\mathbf{V}} = (\hat{\mathbf{v}}_1, \cdots, \hat{\mathbf{v}}_d)$ is a $(p \times d)$ matrix whose jth column, $\hat{\mathbf{v}}_j$ is the jth leading eigenvector of

$$\widehat{\mathbf{M}} = \sum_{k=1}^{h} \hat{\boldsymbol{\beta}}_k \hat{\boldsymbol{\beta}}_k^\top,$$

which we call the PLR working matrix.

2.3. Large Sample Properties

For the sake of simplicity, the subscript k is omitted when a result holds for an arbitrary chosen r_k. Let $\mathbf{X}^* = (1, \mathbf{X}^\top)^\top$, $\mathbf{Z}^\top = (\tilde{Y}, \mathbf{X}^\top)$, $\mathbf{\Sigma}^* = \text{Diag}\{0, \mathbf{\Sigma}\}$, and

$$m_{\boldsymbol{\theta}}(\mathbf{Z}) = \boldsymbol{\theta}^\top \mathbf{\Sigma}^* \boldsymbol{\theta} + C \log\left(1 + \exp\{-\tilde{Y} \cdot \boldsymbol{\theta}^\top \mathbf{X}^*\}\right),$$

and hence $\Lambda(\boldsymbol{\theta}) = E[m_{\boldsymbol{\theta}}(\mathbf{Z})]$.

Theorem 2 states the consistency and asymptotic normality of $\hat{\boldsymbol{\theta}}$.

Theorem 2. *Under the regularity conditions in Appendix A.1,*

a) $\hat{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}_0$ *and,*

b) $\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{d} N(\mathbf{0}, \mathbf{H}_{\boldsymbol{\theta}_0}^{-1} \mathbf{A}_{\boldsymbol{\theta}_0} \mathbf{H}_{\boldsymbol{\theta}_0}^{-1})$, *where* $\mathbf{H} = E[m''_{\boldsymbol{\theta}_0}(\mathbf{Z})]$ *and* $\mathbf{A} = E[m'_{\boldsymbol{\theta}_0}(\mathbf{Z})\{m'_{\boldsymbol{\theta}_0}(\mathbf{Z})\}^\top]$

where $m'_{\boldsymbol{\theta}}$ *and* $m''_{\boldsymbol{\theta}}$ *denote the first and second order derivatives of* $m_{\boldsymbol{\theta}}$ *with respect to* $\boldsymbol{\theta}$*.*

We remark that the asymptotic results for PLR are straightforward from the standard M-estimation theory [19] and do not rely on any stringent technical conditions while PSVM does.

Let $\mathbf{M}_0 = \sum_{h=1}^{H} \boldsymbol{\beta}_{0,k} \boldsymbol{\beta}_{0,k}^\top$ and $\mathbf{V}_0 = (\mathbf{v}_{0,1}, \cdots, \mathbf{v}_{0,d})$ where $\mathbf{v}_{0,j}$ is the jth leading eigenvector of $\mathbf{M}_0$. Theorem 3 establishes asymptotic normalities of $\widehat{\mathbf{M}}$ and $\widehat{\mathbf{V}}$.

Theorem 3. *Under the regularity conditions in Appendix A.1 and rank*$(\mathbf{M}_0) = d$*,*

a) $\sqrt{n}\ \text{vec}(\widehat{\mathbf{M}} - \mathbf{M}_0) \xrightarrow{d} N(\mathbf{0}, \mathbf{\Sigma}_{\mathbf{M}})$

b) $\sqrt{n}\ \text{vec}(\widehat{\mathbf{V}} - \mathbf{V}_0) \xrightarrow{d} N(\mathbf{0}, \mathbf{\Sigma}_{\mathbf{V}})$

for some variance matrices $\mathbf{\Sigma}_{\mathbf{M}}$ *and* $\mathbf{\Sigma}_{\mathbf{V}}$ *explicitly given in Appendix A.4.*

2.4. Structure Dimensionality Determination

To estimate the structural dimension d, we consider the following estimate based on BIC-type criterion proposed by **(author?)** [7].

$$\hat{d} = \underset{d}{\operatorname{argmax}}\, G(d; \rho, \widehat{\mathbf{M}}) = \sum_{j=1}^{d} \upsilon_j - \rho \frac{d \log n}{\sqrt{n}} \upsilon_1, \tag{7}$$

where υ_j is the jth leading eigenvalue of $\widehat{\mathbf{M}}$ and ρ is a tuning parameter. Consistency of $\hat{d}$, i.e., $\lim_{n\to\infty} P(\hat{d} = d) = 1$ directly follows from the asymptotic property of $\widehat{\mathbf{M}}$ in Theorem 3.

In order to select ρ, we propose the following algorithm: First, randomly split the data into the training and testing sets, respectively denoted by $\{(\mathbf{x}_j^{\text{tr}}, y_j^{\text{tr}}) : j = 1, \cdots, n_{\text{tr}}\}$ and $\{(\mathbf{x}_{j'}^{\text{ts}}, y_{j'}^{\text{ts}}) : j' = 1, \cdots, n_{\text{ts}}\}$ where $n_{\text{ts}} + n_{\text{tr}} = n$. We then apply PLR to the training set $\{(\mathbf{x}_j^{\text{tr}}, y_j^{\text{tr}}) : j = 1, \cdots, n_{\text{tr}}\}$ and let $\widehat{\mathbf{M}}^{\text{tr}}$ be the corresponding working matrix. For a given appropriate grid of ρ, repeat the following steps 1 to 4 and select ρ^* that minimizes $TC(\rho)$ defined below.

1. Compute $\hat{d}_{\text{tr}} = \operatorname{argmax}_{d\in\{1,\cdots,p\}} G(d; \rho, \widehat{\mathbf{M}}^{\text{tr}})$.

2. Transform the test predictors by $\tilde{\mathbf{x}}_{j'}^{\text{ts}} = (\widehat{\mathbf{V}}^{\text{tr}})^\top \mathbf{x}_{j'}^{\text{tr}}$, where $\widehat{\mathbf{V}}^{\text{tr}} = (\widehat{\mathbf{v}}_1^{\text{tr}}, \cdots, \widehat{\mathbf{v}}_{\hat{d}_{\text{tr}}}^{\text{tr}})$ denotes the $(p \times \hat{d}_{\text{tr}})$ eigenvector matrix of $\widehat{\mathbf{M}}^{\text{tr}}$.

3. For each $r_k, k = 1, \cdots, h$, apply the logistic regression to $\{(\tilde{\mathbf{x}}_{j'}^{\text{ts}}, \tilde{y}_{j'k}^{\text{ts}}) : j = 1, \cdots, n_{\text{tr}}\}$ where $\tilde{y}_{j'k}^{\text{ts}} = \mathbb{1}\{y_{j'}^{\text{ts}} > r_k\}$.

4. Compute $TC(\rho) = \sum_{k=1}^{h} \sum_{j'=1}^{n_{\text{ts}}} \mathbb{1}\{\tilde{y}_{j',k}^{\text{ts}} \neq \hat{y}_{j',k}^{\text{ts}}\}$ where $\hat{y}_{j',k}^{\text{ts}}$ denotes a predicted value of $\tilde{y}_{j',k}^{\text{ts}}$ from the logistic model obtained from Step 3 above.

3. Penalized PLR for Sparse SDR

3.1. Penalized Principal Logistic Regression

Under the sparsity assumption (2), we have $\boldsymbol{\theta}_0^\top = (\boldsymbol{\theta}_{0,+}^\top, \mathbf{0}^\top)$ where

$$\boldsymbol{\theta}_{0,+}^\top = (\alpha_{0,+}, \boldsymbol{\beta}_{0,+}^\top) = \underset{\alpha,\boldsymbol{\beta}}{\operatorname{argmin}}\, \boldsymbol{\beta}^\top \boldsymbol{\Sigma}_+ \boldsymbol{\beta} + C\mathrm{E}\left[\ln\left(1 + e^{-\tilde{Y}\{\alpha + \boldsymbol{\beta}^\top(\mathbf{X}_+ - \mathrm{E}(\mathbf{X}_+))\}}\right)\right], \tag{8}$$

with $\boldsymbol{\Sigma}_+ = \mathrm{Var}(\mathbf{X}_+)$. This is because $\boldsymbol{\beta}_0 \in \mathcal{S}_{Y|\mathbf{X}}$ by Theorem 1 and $\boldsymbol{\beta}_0$ can be written as a linear combination of the columns of $\mathbf{B}$ whose q rows associated with $\mathbf{X}_-$ are all zeros, i.e. $\boldsymbol{\beta}_0^\top = (\boldsymbol{\beta}_{0,+}^\top, \mathbf{0}_{p-q}^\top)$. Therefore, $\boldsymbol{\beta}_{0,k}, k = 1, \cdots, h$ should be sparse and share a common sparsity structure across k.

To impose such a sparsity structure, we propose a penalized PLR that minimizes the following objective function:

$$Q(\boldsymbol{\Omega}) = \sum_{k=1}^{h} \hat{\Lambda}_k(\boldsymbol{\theta}_k) + \sum_{j=1}^{p} p_\lambda\big(\max_{1\le k\le h} |\beta_{jk}|\big), \tag{9}$$

where $\boldsymbol{\Omega}$ is a $(p+1)\times h$ dimensional matrix whose kth column is $\boldsymbol{\theta}_k, k = 1, \cdots, h$. p_λ denotes a nonconvex penalty function and depends on a tuning parameter λ that controls the sparsity of the solution. Because (9) penalizes the maximum of $|\beta_{jk}|$ over $k = 1, \cdots, h$, the entire elements in the same row of $\boldsymbol{\Omega}$ simultaneously shrink toward zero so that the desired sparsity structure is naturally attained. It is crucial to tune λ in practice and we discuss this issue in Section 3.3.

For the penalty function, we exploit the SCAD penalty of **(author?)** [13] which is defined through its derivative as

$$p'_\lambda(\theta) = \lambda\left[I(\theta \le \lambda) + \frac{(a\lambda - \theta)_+}{(a-1)\lambda} I(\theta > \lambda)\right], \qquad \theta > 0$$

for some $a > 2$. We set $a = 3.7$ as recommended by **(author?)** [13]. The SCAD penalty has been popular in the context of variable selection due to the oracle property [13]. In order to establish the oracle property of the penalized PLR, consider a partition of $\boldsymbol{\Omega}^\top = (\boldsymbol{\Omega}_+^\top, \boldsymbol{\Omega}_-^\top)$, where $\boldsymbol{\Omega}_+ = (\boldsymbol{\theta}_{1+}, \cdots, \boldsymbol{\theta}_{h+})$ and $\boldsymbol{\Omega}_- =$

$(\boldsymbol{\theta}_{1-}, \cdots, \boldsymbol{\theta}_{h-})$ with $\boldsymbol{\theta}_{k+} = (\alpha_k, \beta_{1k}, \cdots, \beta_{qk})^\top$ and $\boldsymbol{\theta}_{k-} = (\beta_{jq+1} \cdots, \beta_{jp})^\top$ for $k = 1, \cdots, h$, respectively. Theorem 4 states oracle property of the solution of the penalized PLR, $\widehat{\boldsymbol{\Omega}}^\top = (\widehat{\boldsymbol{\Omega}}_+^\top, \widehat{\boldsymbol{\Omega}}_-^\top) = \text{argmin}_{\boldsymbol{\Omega}} Q(\boldsymbol{\Omega})$. Namely, $\widehat{\boldsymbol{\Omega}}^\top$ behaves asymptotically as if we know which variables are relevant.

Theorem 4. *(Oracle property) Let $\lambda = \lambda_n$ to emphasize that λ is a function of n. Suppose that $\lambda_n \to 0$ and $\sqrt{n}\lambda_n \to \infty$ as $n \to \infty$. In addition to the regularity conditions in Appendix A.1, we further assume that p_{λ_n} satisfies*

$$\liminf_{n\to\infty} \liminf_{\theta\to 0+} p'_{\lambda_n}(\theta)/\lambda_n > 0.$$

If $\lambda_n \to 0$ and $\sqrt{n}\lambda_n \to \infty$ as $n \to \infty$, then with probability tending to one, the $\sqrt{n}$-consistent minimizer of $Q(\boldsymbol{\Omega})$ denoted by $\widehat{\boldsymbol{\Omega}}^\top = (\widehat{\boldsymbol{\Omega}}_+^\top, \widehat{\boldsymbol{\Omega}}_-^\top)$ must satisfy

(a) $\widehat{\boldsymbol{\Omega}}_- = \mathbf{0}$.

(b) For $k = 1, \cdots, h$,

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_{k+} - \boldsymbol{\theta}_{0,k+}) \quad \to \quad N(\mathbf{0}, \mathbf{H}_{k+}^{-1}\mathbf{A}_{k+}\mathbf{H}_{k+}^{-1}),$$

where $\mathbf{H}_+ = \mathrm{E}[m''_{\boldsymbol{\theta}_0}(\mathbf{Z}_{k+})]$ and $\mathbf{A}_+ = \mathrm{E}[m'_{\boldsymbol{\theta}_0}(\mathbf{Z}_+)\{m'_{\boldsymbol{\theta}_0}(\mathbf{Z}_{k+})\}^\top]$ with $\mathbf{Z}_{k+}^\top = (\tilde{Y}_k, \mathbf{X}_+^\top)$.

3.2. Computation

For ease of representation, we assume that predictors are centered without loss of generality (i.e., $\sum_{i=1}^n \mathbf{x}_i = \mathbf{0}$). Now, the objective function (9) is equivalently rewritten in a vector format as follows:

$$\underset{\boldsymbol{\theta}}{\text{argmin}}\, \boldsymbol{\theta}^\top \mathbf{S}\boldsymbol{\theta} + \frac{C}{n}\left[\mathbf{1}^\top \ln\left(1 + e^{-\tilde{\mathbf{y}}\odot(\mathbf{W}\boldsymbol{\theta})}\right)\right] + \sum_{j=1}^{p} p_\lambda\big(\max_{1\le k\le h} |\beta_{jh}|\big), \qquad (10)$$

where

$$
\begin{aligned}
\boldsymbol{\theta}^\top &= (\boldsymbol{\theta}_1^\top, \cdots, \boldsymbol{\theta}_h^\top); \\
\tilde{\mathbf{Y}}^\top &= (\tilde{\mathbf{y}}_1^\top, \cdots, \tilde{\mathbf{y}}_h^\top) \text{ with } \tilde{\mathbf{y}}_k^\top = (\tilde{y}_{1k}, \cdots, \tilde{y}_{nk}); \\
\mathbf{S} &= \text{Diag}\{\widehat{\boldsymbol{\Sigma}}^*, \cdots, \widehat{\boldsymbol{\Sigma}}^*\} \text{ with } \widehat{\boldsymbol{\Sigma}}^* = \text{Diag}\{0, \widehat{\boldsymbol{\Sigma}}\}; \\
\mathbf{W} &= \text{Diag}\{\mathbf{X}^*, \cdots, \mathbf{X}^*\} \text{ with } \mathbf{X}^* = (\mathbf{x}_1^*, \cdots, \mathbf{x}_n^*)^\top \text{ and } \mathbf{x}_i^* = (1, \mathbf{x}_i^\top)^\top,
\end{aligned}
$$

and $\odot$ denotes the Hadamard product. We slightly abuse the notation in (10) by elementwisely applying the exponential function to the power in a vector form, i.e., $e^{\mathbf{a}} = (e^{a_1}, \cdots, e^{a_p})^\top$ for a vector $\mathbf{a} = (a_1, \cdots, a_p)^\top$.

It is not trivial to solve (10) with respect to $\boldsymbol{\theta}$ due to the nonconvexity of the SCAD penalty. There are several existing algorithms for solving the SCAD-penalized problems that include, for example, local quadratic approximation [13], minorize-maximize algorithm [20] and local linear approximation [21] among many others. In this article, we employ the difference convex algorithm [DC, 22, 23, 11] as described in the following paragraph.

First, we approximate the logistic loss to its second order Taylor expansion at the value of the t^{th} iteration denoted by $\boldsymbol{\theta}^{(t)}$. We then have a familiar form of the iteratively reweighted least squares algorithm commonly used to fit the logistic regression [24, 25]. In particular, $\boldsymbol{\theta}$ can be updated at the t^{th} iteration as follows.

$$
\boldsymbol{\theta}^{(t+1)} = \underset{\boldsymbol{\theta}}{\text{argmin}}\, \boldsymbol{\theta}^\top \mathbf{S}\boldsymbol{\theta} + \frac{C}{2n}(\tilde{\mathbf{u}}^{(t)} - \mathbf{W}\boldsymbol{\theta})^\top \mathbf{G}^{(t)}(\tilde{\mathbf{u}}^{(t)} - \mathbf{W}\boldsymbol{\theta}) + \sum_{j=1}^{p} p_\lambda\left(\max_{1\le k\le h}\{|\beta_{jk}|\}\right), \tag{11}
$$

where $\mathbf{G}^{(t)} = \text{Diag}\{\mathbf{p}_1^{(t)} \odot (1-\mathbf{p}_1^{(t)}), \cdots, \mathbf{p}_h^{(t)} \odot (1-\mathbf{p}_h^{(t)})\}$ with $\mathbf{p}_k^{(t)} = (\pi_{1k}^{(t)}, \cdots, \pi_{nk}^{(t)})^\top$ and $\pi_{ik}^{(t)} = \left(1 + \exp\{\tilde{y}_{ik} \cdot \{\boldsymbol{\theta}_k^{(t)}\}^\top \mathbf{x}_i^*\}\right)^{-1}$, and $\tilde{\mathbf{u}}^{(t)} = \mathbf{W}\boldsymbol{\theta}^{(t)} + \{\mathbf{G}^{(t)}\}^{-1}(\tilde{\mathbf{y}} \odot \mathbf{p}^{(t)})$ for $\mathbf{p}^{(t)} = (\{\mathbf{p}_1^{(t)}\}^\top, \cdots, \{\mathbf{p}_h^{(t)}\}^\top)^\top$. The superscript is used to denote the quantities obtained at the t^{th} iteration.

The DC algorithm decomposes the SCAD penalty as the difference of two

convex functions as $p_\lambda(\theta) = p_{\lambda,1}(\theta) - p_{\lambda,2}(\theta)$ where $p'_{\lambda,1}(\theta) = \lambda$, and $p'_{\lambda,2}(\theta) = \lambda\left[1 - \frac{(a\lambda-\theta)_+}{(a-1)\lambda}\right] I(\theta > \lambda)$.

Letting $\xi_j = \max_{1\le k\le h} |\beta_{jk}|$ and $\xi_j^{(t)} = \max_{1\le k\le h} |\beta_{jk}^{(t)}|$, we have a linear approximation of $p_{\lambda,2}(\xi_j) \approx p'_{\lambda,2}(\xi_j^{(t)})(\xi_j - \xi_j^{(t)})$ and, thus,

$$p_\lambda\big(\max_{1\le k\le h}\{|\beta_{jk}|\}\big) = p_\lambda(\xi_j) \approx \{\lambda - p'_{\lambda,2}(\xi_j^{(t)})\}\xi_j + Constant \tag{12}$$

Plugging (12) into (11), we have a standard quadratic programming (QP) problem for updating $\boldsymbol{\theta}$ and $\boldsymbol{\xi} = (\xi_1, \cdots, \xi_p)^\top$:

$$(\boldsymbol{\theta}^{(t+1)}, \boldsymbol{\xi}^{(t+1)}) = \underset{\boldsymbol{\theta},\boldsymbol{\xi}}{\operatorname{argmin}}\, \boldsymbol{\theta}^\top \mathbf{S}\boldsymbol{\theta} + \frac{C}{2n}(\tilde{\mathbf{u}}^{(t)} - \mathbf{W}\boldsymbol{\theta})^\top \mathbf{G}^{(t)}(\tilde{\mathbf{u}}^{(t)} - \mathbf{W}\boldsymbol{\theta}) + \sum_{i=1}^{p}\left(\lambda - p'_{\lambda,2}(\xi_j^{(t)})\right)\xi_j \tag{13}$$

subject to $\xi_j \ge \beta_{jk}$ and $\xi_j \ge -\beta_{jk}, j = 1, \cdots, p, k = 1, \cdots, h$. Existing software can be readily applied to solve (13).

Finally, the algorithm to solve the max-SCAD penalized PLR is summarized as follows.

1. Initialize $\boldsymbol{\theta}^{(0)}$ (e.g., unpenalized PLR solution) and $\boldsymbol{\xi}^{(0)}$
2. Update $(\boldsymbol{\theta}^{(t)}, \boldsymbol{\xi}^{(t)}) \to (\boldsymbol{\theta}^{(t+1)}, \boldsymbol{\xi}^{(t+1)})$ from (13).
3. Stop updating if $\|\boldsymbol{\theta}^{(t+1)} - \boldsymbol{\theta}^{(t)}\|/\|\boldsymbol{\theta}^{(t)}\|$ is sufficiently small, for example, less then 10^{-4}.

3.3. Tuning λ

It is important to tune λ that controls the degree of sparsity. To this end, an L-fold cross-validation procedure is proposed as follows. First, we randomly split the data into a training set $(\mathbf{x}_{j,\mathrm{tr}}^{[\ell]}, y_{j,\mathrm{tr}}^{[\ell]}), j = 1, \cdots, n_{\mathrm{tr}}$ and test set $(\mathbf{x}_{j',\mathrm{ts}}^{[\ell]}, y_{j',\mathrm{ts}}^{[\ell]}), j' = 1, \cdots, n_{\mathrm{ts}}$, where the superscript $\ell = 1, \cdots, L$ denotes the ℓth fold. Next, we apply the penalized PLR to the training set for a given λ and obtain $\widehat{\mathbf{V}}_{\mathrm{tr}}^{[\ell]}$. Third, we project test predictors onto the estimated $\mathcal{S}_{Y|\mathbf{X}}$ from the training set, i.e., $\tilde{\mathbf{x}}_{j',\mathrm{ts}}^{[\ell]} = \{\widehat{\mathbf{V}}_{\mathrm{tr}}^{[\ell]}\}^\top \mathbf{x}_{j',\mathrm{ts}}^{[\ell]}$. Fourth, we compute

$\hat{\gamma}^{[\ell]} = \widehat{\mathrm{dcor}}(y^{[\ell]}_{j',\mathrm{ts}}, \tilde{\mathbf{x}}^{[\ell]}_{j',\mathrm{ts}})$ where $\widehat{\mathrm{dcor}}(y, \mathbf{x})$ denotes the sample distance correlation between y and $\mathbf{x}$ [DC, 26], which is defined by $\widehat{\mathrm{dcor}}(y, \mathbf{x}) = \frac{\widehat{\mathrm{dcov}}(y,\mathbf{x})}{\sqrt{\widehat{\mathrm{dcov}}(y,y)\widehat{\mathrm{dcov}}(\mathbf{x},\mathbf{x})}}$. Here $\widehat{\mathrm{dcov}}(y, \mathbf{x})$ is the sample distance covariance between y and $\mathbf{x}$ and obtained by

$$\widehat{\mathrm{dcov}}(y, \mathbf{x}) = \hat{S}_1 + \hat{S}_2 - 2\hat{S}_3.$$

where

$$\begin{aligned}
\hat{S}_1 &= \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \|y_i - y_j\| \|\mathbf{x}_i - \mathbf{x}_j\|, \\
\hat{S}_2 &= \frac{1}{n^4} \left(\sum_{i=1}^{n} \sum_{j=1}^{n} \|y_i - y_j\| \right) \left(\sum_{i=1}^{n} \sum_{j=1}^{n} \|\mathbf{x}_i - \mathbf{x}_j\| \right), \\
\hat{S}_3 &= \frac{1}{n^3} \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{l=1}^{n} \|y_i - y_l\| \|\mathbf{x}_j - \mathbf{x}_l\|,
\end{aligned}$$

with $\|\cdot\|$ denoting the Euclidian norm.

Finally, we repeat these steps and find an optimal λ that minimizes $R(\lambda) = \frac{1}{K} \sum_{\ell=1}^{L} \hat{\gamma}^{[\ell]}$.

4. Simulation

We conducted a number of simulation studies to evaluate the finite sample performance of the proposed methods under various scenarios. The data are generated from the following nonlinear regression model:

$$y_i = f(\mathbf{x}_i) + 0.2\epsilon,$$

where $\mathbf{x}_i \overset{iid}{\sim} N(\mathbf{0}_p, \mathbf{I}_p)$ and $\epsilon \sim N(0,1)$ with $n = 100$ and $p = 10, 20, 30$. Three different regression functions are considered as follows:

$$f_1(\mathbf{x}) = \frac{\mathbf{b}_1^\top \mathbf{x} - 1}{0.5 + (\mathbf{b}_2^\top \mathbf{x} + 1)^2},$$
$$f_2(\mathbf{x}) = \sin(\mathbf{b}_1^\top \mathbf{x}) + (\mathbf{b}_2^\top \mathbf{x} + 1)^2,$$
$$f_3(\mathbf{x}) = \cos(\mathbf{b}_1^\top \mathbf{x} + 1)/\exp(\mathbf{b}_2^\top \mathbf{x})$$

with three basis matrices $\mathbf{B} = (\mathbf{b}_1, \mathbf{b}_2)$ to represent different sparsity structures:

$$\text{(Case 1)} \quad \mathbf{b}_1 = \mathbf{b}_2 = \mathbf{1}_p/\sqrt{p},$$
$$\text{(Case 2)} \quad \mathbf{b}_1 = \sum_{j=1}^{p/2} \mathbf{e}_j/\sqrt{p/2}, \text{ and } \mathbf{b}_2 = \sum_{j=p/2+1}^{p} \mathbf{e}_j/\sqrt{p/2},$$
$$\text{(Case 3)} \quad \mathbf{b}_1 = \mathbf{e}_1, \text{ and } \mathbf{b}_2 = \mathbf{e}_2.$$

Here $\mathbf{e}_j$ denotes the p-dimensional vector whose jth element is 1 and 0 for all others. Case 1 represents a nonsparse structure with $q = p$, and Case 3 is the most sparse with $q = 2$. Case 2 can be regarded as intermediate between Case 1 and 3 since $q = p/2$. The true structural dimension d is 1 for Case 1 and $d = 2$ for Case 2 and 3.

4.1. Dimension Reduction Performance

We compare two versions of PLR to the existing methods. For the conventional SDR without pursuing sparsity, SIR and PSVM are considered as competing methods against PLR. For the sparse SDR, penalized SIR [PSIR, 11] and sparse partial least square regression [SPLS, 10] are compared to the penalized PLR which we denote PPLR for short. For the two PLR methods as well as PSVM, we set $r_k, k = 1, \cdots, 9$ as the $(100 \times k/10)$th sample percentile of y_i, and $C = 1$. It is empirically shown that the performance of PSVM is not overly sensitive to the choices of either h or C [7], and hence PLR will not as well. We set the number of slices to be 10 for both SIR and PSIR. PPLR is tuned as described in Section 3.3. PSIR is tuned based on a BIC criterion

as suggested by **(author?)** [11]. For SPLS, we tried several different values of tuning parameters and reported the best result in each case.

As a performance measure, we compute the distance between the true and estimated $\mathcal{S}_{Y|\mathbf{X}}$ in terms of the following criterion:

$$\|\mathbf{P}_{\widehat{\mathbf{B}}} - \mathbf{P}_{\mathbf{B}}\|_F, \tag{14}$$

where $\mathbf{P}_{\mathbf{B}} = \mathbf{B}(\mathbf{B}^\top\mathbf{B})^{-1}\mathbf{B}^\top$ is a projection matrix to span($\mathbf{B}$) and $\|\mathbf{A}\|_F$ denotes the Frobenius norm of a matrix $\mathbf{A}$.

case	f	p	SDR			sparse SDR		
			SIR	PSVM	PLR	PSIR	SPLS	PPLR
1	f_1	10	0.296 (.091)	**0.213** (.058)	0.220 (.064)	0.309 (.088)	0.403 (.143)	0.290 (.082)
		20	0.457 (.103)	0.358 (.091)	**0.334** (.079)	0.439 (.113)	0.684 (.148)	0.392 (.084)
		30	0.698 (.177)	0.544 (.113)	**0.471** (.105)	0.664 (.152)	0.854 (.102)	0.526 (.095)
	f_2	10	0.271 (.120)	**0.172** (.064)	0.210 (.105)	0.296 (.120)	0.305 (.116)	0.279 (.112)
		20	0.411 (.141)	**0.313** (.088)	0.344 (.126)	0.421 (.133)	0.601 (.139)	0.393 (.111)
		30	0.593 (.244)	0.557 (.166)	**0.506** (.182)	0.667 (.203)	0.807 (.109)	0.526 (.147)
	f_3	10	0.290 (.112)	**0.202** (.069)	0.236 (.106)	0.307 (.110)	0.498 (.344)	0.284 (.099)
		20	0.459 (.161)	**0.363** (.115)	0.383 (.152)	0.464 (.147)	0.713 (.259)	0.415 (.130)
		30	0.658 (.211)	0.590 (.183)	**0.549** (.214)	0.642 (.210)	0.916 (.208)	0.549 (.179)
2	f_1	10	1.054 (.240)	**0.985** (.244)	0.996 (.251)	1.146 (.231)	1.093 (.041)	1.057 (.240)
		20	1.370 (.128)	1.269 (.148)	**1.256** (.156)	1.356 (.129)	1.208 (.064)	1.289 (.151)
		30	1.529 (.097)	1.451 (.095)	**1.419** (.096)	1.489 (.076)	1.308 (.073)	1.426 (.100)
	f_2	10	1.145 (.240)	**0.852** (.199)	0.852 (.208)	1.082 (.237)	1.071 (.043)	1.004 (.230)
		20	1.389 (.120)	1.180 (.157)	**1.170** (.169)	1.318 (.154)	1.333 (.073)	1.219 (.159)
		30	1.523 (.115)	1.383 (.124)	**1.342** (.135)	1.446 (.118)	1.434 (.089)	1.362 (.146)
	f_3	10	0.659 (.175)	0.633 (.149)	**0.627** (.159)	0.709 (.161)	1.171 (.093)	0.686 (.151)
		20	1.133 (.201)	0.985 (.182)	**0.955** (.179)	1.071 (.196)	1.315 (.093)	0.992 (.162)
		30	1.530 (.174)	1.288 (.172)	**1.220** (.188)	1.362 (.179)	1.413 (.079)	1.244 (.181)
3	f_1	10	1.049 (.252)	0.923 (.227)	0.965 (.232)	0.240 (.394)	1.029 (.026)	**0.199** (.403)
		20	1.339 (.176)	1.278 (.162)	1.248 (.168)	0.338 (.436)	1.057 (.040)	**0.230** (.385)
		30	1.521 (.106)	1.438 (.113)	1.399 (.120)	0.394 (.391)	1.078 (.057)	**0.246** (.422)
	f_2	10	1.087 (.233)	0.850 (.207)	0.842 (.198)	0.410 (.515)	1.014 (.023)	**0.263** (.408)
		20	1.383 (.135)	1.182 (.147)	1.174 (.162)	0.489 (.534)	1.025 (.030)	**0.316** (.435)
		30	1.534 (.094)	1.382 (.105)	1.343 (.124)	0.507 (.519)	1.034 (.037)	**0.386** (.484)
	f_3	10	0.610 (.147)	0.599 (.132)	0.595 (.140)	0.061 (.121)	1.080 (.056)	**0.050** (.124)
		20	1.104 (.200)	0.978 (.155)	0.936 (.148)	0.061 (.071)	1.154 (.083)	**0.056** (.101)
		30	1.452 (.179)	1.267 (.150)	1.180 (.170)	0.090 (.092)	1.202 (.095)	**0.043** (.069)

Table 1: Averaged distance measures (14) over 100 independent repetitions. Bold cases represent a winning method for each scenario. Corresponding standard deviations are given in parentheses.

Table 1 contains averaged distance measure (14) over 100 independent repetitions. Under Case 1 representing a nonsparse scenario, PSVM and PLR show comparable performance, and outperforms all others including the three methods for sparse SDR, which we believe natural. Similar patterns are observed in Case 2 whose bases are not very sparse. Under Case 3 representing a sparse scenario, both PSIR and PPLR show nearly perfect results with PPLR being

slightly and consistently better under all scenarios under consideration. We remark that the performance of SPLS is not very satisfactory since it can estimate at most one basis of $\mathcal{S}_{Y|\mathbf{X}}$, which is one drawback of SPLS. We also note that PPLR shows comparable performance to PLR even in Cases 1 and 2 with nonsparse scenarios. This is because if the true signal is not sparse, then the λ adaptively selected from the data would be sufficiently small and the effect of penalization becomes negligible. In practice, we are not aware of the true sparsity of $\mathbf{B}$ and, hence, PPLR would be a safer choice because the tuning procedure automatically takes into account the unknown sparsity structure.

4.2. Structural Dimension Estimation

We also check the performance of the proposed procedure for structural dimension estimation developed in 2.4. As a comparison, the sequential χ^2-test [2] is applied for SIR. Table 2 reports empirical probabilities (in percentage) of correctly estimating d over 100 independent repetitions. It is clearly observed that the proposed procedure provides quite promising results for both PLR and PPLR in estimating d.

f	p	Case1			Case2			Case3		
		SIR	PLR	PPLR	SIR	PLR	PPLR	SIR	PLR	PPLR
f_1	10	95%	95%	90%	20%	85%	83%	31%	78%	83%
	20	40%	83%	77%	5%	83%	81%	6%	73%	74%
	30	18%	74%	73%	4%	73%	77%	7%	70%	43%
f_2	10	94%	93%	91%	17%	86%	87%	21%	82%	85%
	20	50%	81%	79%	6%	79%	77%	11%	74%	77%
	30	22%	79%	76%	2%	71%	71%	3%	70%	67%
f_3	10	95%	90%	82%	68%	98%	97%	81%	100%	100%
	20	52%	84%	75%	21%	93%	92%	21%	96%	97%
	30	20%	70%	74%	7%	89%	77%	3%	89%	83%

Table 2: Empirical probabilities (in percentage) of correctly estimating true d based on 100 independent repetitions: The proposed procedure shows promising performance in estimating structural dimension.

4.3. Variable Selection Performance in Sparse SDR

In order to evaluate the variable selection performance of the three methods for sparse SDR, we consider three measures as follows: the number of nonzero elements in the basis which are correctly estimated as nonzero (denoted by

"CNZ"), the number of zero elements of the basis which are incorrectly set to nonzero (denoted by "INZ"), and the frequency of recovering the correct sparsity structure of basis (denoted by "C"). Table 3 contains the three measures averaged over 100 repetitions under Case 3 where only two elements in $\mathbf{B}$ are nonzero and hence (CNZ, INZ, C) $= (2, 0, 100)$ indicates perfect performance. In terms of CNZ, all the methods performs reasonably well. However, PPLR outperforms others in terms of INZ and C. That is, PPLR is less likely to over-select nonzero elements in the basis than PSIR and SPLS.

f	p	CNZ			INZ			C		
		PSIR	SPLS	PPLR	PSIR	SPLS	PPLR	PSIR	SPLS	PPLR
f_1	10	2.00	2.00	1.99	1.63	2.15	1.45	41	28	62
	20	2.00	2.00	2.00	2.19	3.89	2.14	25	21	51
	30	2.00	2.00	2.00	3.54	5.08	2.54	6	20	35
f_2	10	1.99	2.00	2.00	2.89	0.88	1.93	40	57	51
	20	1.98	2.00	2.00	5.22	1.55	3.13	19	42	39
	30	1.96	2.00	1.98	5.46	2.33	4.45	13	33	29
f_3	10	2.00	1.98	2.00	0.86	2.75	0.63	55	9	74
	20	2.00	1.98	2.00	1.10	5.30	0.98	32	4	62
	30	2.00	1.98	2.00	1.76	6.99	0.72	23	2	58

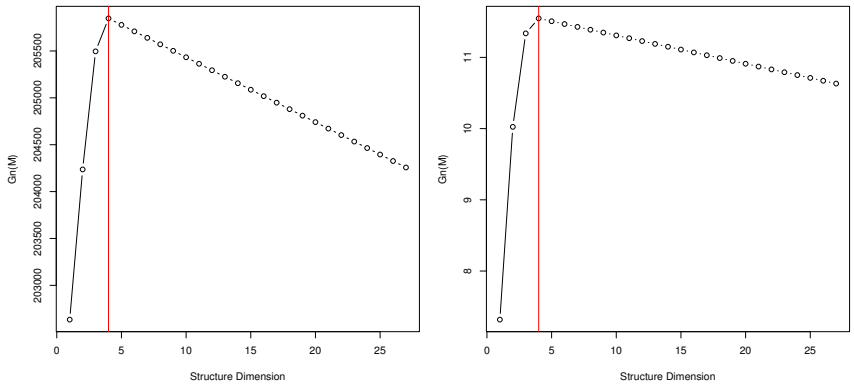
Table 3: In Case3, variable selection performance of the methods for sparse SDR are compared over 100 independent repetitions. 'CNZ' denotes the number of nonzero elements of the basis which are correctly estimated as nonzero; 'INZ' denotes the number of zero elements of the basis which are incorrectly set to nonzero; and 'C' denotes the frequency of recovering the correct sparsity structure of basis. PPLR outperforms all others.

5. Real Data Analysis

In order to carefully evaluate the proposed method on real data, we apply our method to *Pyrimidines* dataset which was collected to understand the quantitative structure-activity relationship (QSAR) in drug design. The data are available at `http://www.dcc.fc.up.pt/~ltorgo/Regression/DataSets.html`. The *Pyrimidines* data set contains three versions of predicates of nine properties (hence 27 predictors in total) for 74 nonhydrogen substituents. The nine properties include polarity, size, flexibility, hydrogen-bond donor, hydrogenbond acceptor, π-donor, π-acceptor, polarizability, and σ-effect of the substituent. For each substituents, their biological activities are recored as response and the goal of the study is to find relation between the biological activity and

the aforementioned properties. Details of the *Pyrimidines* data can be found in **(author?)** [27].

We apply both PLR and PPLR to the *Pyrimidines* data. All tuning parameters including λ, C, and H are set in the same manner as described in Section 4. Structural dimension d is estimated as described in Section 2.4. To be more precise, we first tune ρ in (7) which is selected as .00068 for PLR and .01081 for PPLR. Figure 2 depicts the BIC-type values in (7) as a function of d which results d estimated as 4 for both PLR and PPLR. It turns out that the sequential χ^2 test for SIR also gives 4 for d estimate. Therefore, we conclude that SDR reduces predictor dimension from 27 to 4 without much loss of regression information.



(a) PLR (b) Penalized-PLR

Figure 2: CVBIC for determining d for (a) PLR and (b) the penalized PLR. The tuning parameters ρ in (7) are selected as .00068 for the PLR and .01081 for the penalized PLR. Estimated d for the two methods are the same, 4, which is concordant with the sequential χ^2 test for SIR under the significance level $\alpha = .05$.

The final goal of SDR is to find a regression model on the central subspace, $\mathcal{S}_{Y|\mathbf{X}}$ where complete regression information is contained. To this end, we train both linear and nonlinear regression models for y and predictors projected on the estimated $\mathcal{S}_{Y|\mathbf{X}}$, $\mathbf{x}^\top \widehat{\mathbf{B}}$. We employ local polynomial regression for fitting nonlinear regression. We note that SDR is model-free and a final regression model on $\mathcal{S}_{Y|\mathbf{X}}$ does not need to be linear. Figure 3 depicts scatter plots of observed and fitted responses from both linear and nonlinear models built on

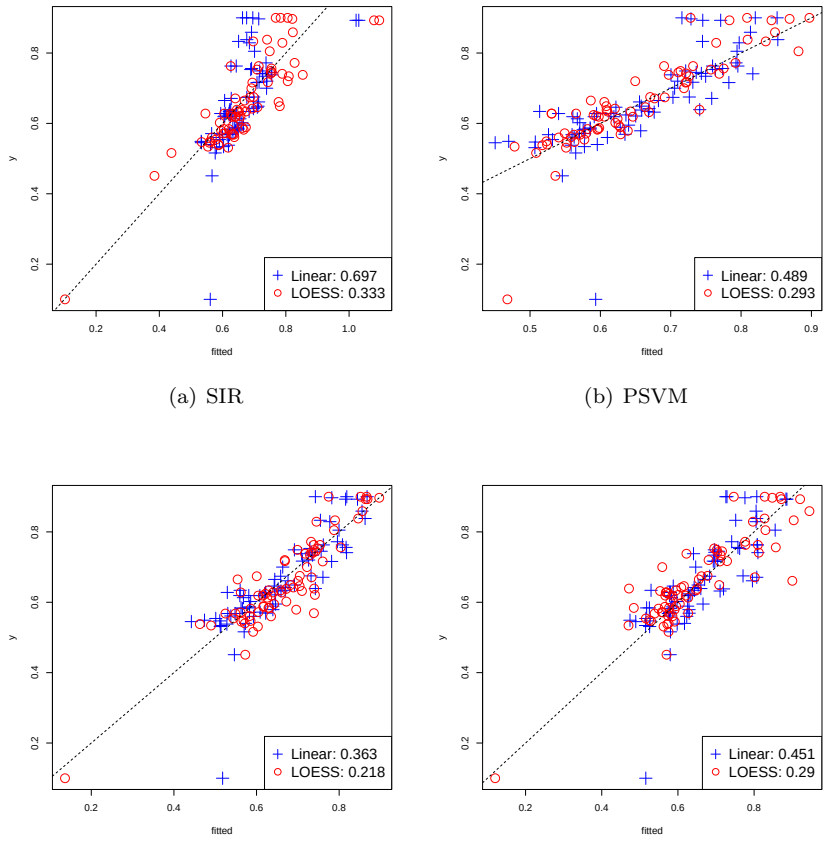
the estimated $\mathcal{S}_{Y|\mathbf{X}}$ by different SDR methods. If a SDR method performs well, prediction accuracy of the final model on the estimated $\mathcal{S}_{Y|\mathbf{X}}$ is to be improved. In this context, PLR outperforms others in terms of prediction. SIR is particularly unsatisfactory for this *Pyrimidines* data

Finally, we conducted cross-validation to evaluate more precisely the performance of different SDR methods for *Pyrimidines* data as follows. First, we randomly split the data into training and testing data with approximately equal sizes and apply different SDR methods to the training data to estimate $\mathcal{S}_{Y|\mathbf{X}}$. Second, we project the test predictors to the central spaces estimated by the different SDR methods. Third, we compute the distance correlation between the test response and the test predictors projected on the estimated $\mathcal{S}_{Y|\mathbf{X}}$. This procedure is repeated 100 times for different random partition of training and test data. Box plots of the cross-validated distance correlations for different SDR methods are depicted in Figure 4. As a reference, we also include distance correlation between the test response and test predictors without applying SDR. All methods except SIR successively reduce dimensionality of the predictors without losing regression information contained in the *Pyrimidines* data. We note that PPLR shows the best performance in terms of preserving dependency structure between the response and predictors projected on the estimated central subspace by PPLR.

6. Discussion

In this paper, we propose PLR as an efficient tool for SDR. Its estimation as well as asymptotic analysis are straightforward due to the similarity to the conventional logistic regression. We then further develop its penalized version for sparse SDR. The max-SCAD penalized PLR adaptively takes into account the unknown sparsity structure of the basis of the central subspace and presents dramatic improvement when the true signal is indeed sparse.

A distinguished feature of PSVM which motivates PLR is that it can be readily extended to the nonlinear SDR by employing the kernel trick. This



(a) SIR (b) PSVM

(c) PLR (d) PPLR

Figure 3: Scatter plots of y versus fitted $\hat{y}$ from linear (plus) and local polynomial (circle) regression models on the estimated central subspaces by different SDR methods. Values in the legend are sum of squared residuals.

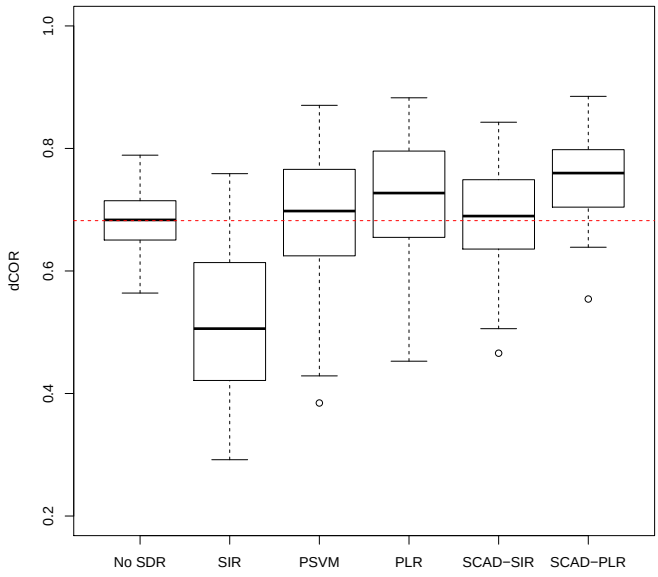


Figure 4: Boxplots of 100 cross-validated distance correlations between test response and test predictors projected on $\mathcal{S}_{Y|\mathbf{X}}$ estimated by different SDR methods: As a reference, both boxplot and average (dashed horizontal line) of distance correlations between test response and test predictors before projection are depicted.

leads us to develop the kernel PLR by simply replacing loss function. However, the penalized version is not straightforward due to the use of kernels that map predictors to an infinite feature space.

In this article, we assume that p can be large but fixed. When p is diverging, its asymptotic analysis becomes challenging because of the covariance matrix of the predictors in the objective function. For diverging p, we may need much stronger conditions to guarantee that the covariance matrix estimators behave nicely as p increases. We leave this as a further research topic.

Acknowldegement

S. J. Shin was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (No.R1C1A1A01054913).

References

[1] R. D. Cook, Graphics for regressions with a binary response, Journal of the American Statistical Association 91 (435) (1996) 983–992.

[2] K.-C. Li, Sliced inverse regression for dimension reduction (with discussion), Journal of the American Statistical Association 86 (414) (1991) 316–342.

[3] R. D. Cook, S. Weisberg, Discussion of "sliced inverse regression for dimension reduction", Journal of the American Statistical Association 86 (414) (1991) 28–33.

[4] B. Li, S. Wang, On directional regression for dimension reduction, Journal of the American Statistical Association 102 (479) (2007) 997–1008.

[5] H. Wang, Y. Xia, Sliced regression for dimension reduction, Journal of the American Statistical Association 103 (482) (2008) 811–821.

[6] B. Li, H. Zha, F. Chiaromonte, Contour regression: a general approach to dimension reduction, The Annals of Statistics 33 (4) (2005) 1580–1616.

[7] B. Li, A. Artemiou, L. Li, Principal support vector machines for linear and nonlinear sufficient dimension reduction, The Annals of Statistics 39 (6) (2011) 3182–3210.

[8] L. Li, Sparse sufficient dimension reduction, Biometrika 94 (3) (2007) 603–613.

[9] H. D. Bondell, L. Li, Shrinkage inverse regression estimation for model-free variable selection, Journal of the Royal Statistical Society: Series B (Statistical Methodology) 71 (1) (2009) 287–299.

[10] H. Chun, S. Keleş, Sparse partial least squares regression for simultaneous dimension reduction and variable selection, Journal of the Royal Statistical Society: Series B (Statistical Methodology) 72 (1) (2010) 3–25.

[11] Y. Wu, L. Li, Asymptotic properties of sufficient dimension reduction with a diverging number of predictors, Statistica Sinica 2011 (21) (2011) 707.

[12] R. D. Cook, et al., Testing predictor contributions in sufficient dimension reduction, The Annals of Statistics 32 (3) (2004) 1062–1092.

[13] J. Fan, R. Li, Variable section via nonconcave penalized likelihood and its oracle properties, Journal of the American Statistical Association 96 (456) (2001) 1348–1360.

[14] R. D. Cook, L. Ni, Using intraslice covariances for improved estimation of the central subspace in regression, Biometrika 93 (1) (2006) 65–74.

[15] J. Zhu, T. Hastie, Kernel logistic regression and the import vector machine, Journal of Computational and Graphical Statistics 14 (1).

[16] K.-C. Li, N. Duan, Regression analysis under link violation, The Annals of Statistics 17 (3) (1989) 1009–1052.

[17] B. Li, Y. Dong, Dimension reduction for nonelliptically distributed predictors, The Annals of Statistics 37 (2009) 1272–1298.

[18] P. Hall, K.-C. Li, On almost linearity of low-dimensional projections from high-dimensional data, The Annals of Statistics 21 (2) (1993) 867–889.

[19] A. W. van der Vaart, Asymptotic Statistics, Vol. 3, Cambridge university press, 2000.

[20] D. R. Hunter, R. Li, Variable selection using mm algorithms, The Annals of Statistics 33 (4) (2005) 1617.

[21] H. Zou, R. Li, One-step sparse estimates in nonconcave penalized likelihood models, The Annals of statistics 36 (4) (2008) 1509.

[22] L. T. H. An, P. D. Tao, Solving a class of linearly constrained indefinite quadratic problems by dc algorithms, Journal of Global Optimization 11 (3) (1997) 253–285.

[23] Y. Wu, Y. Liu, Variable selection in quantile regression, Statistica Sinica 19 (2) (2009) 801.

[24] P. McCullagh, J. A. Nelder, Generalized linear models, Vol. 37, CRC press, 1989.

[25] P. Breheny, J. Huang, Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection, The Annals of Applied Statistics 5 (1) (2011) 232.

[26] G. J. Székely, M. L. Rizzo, N. K. Bakirov, et al., Measuring and testing dependence by correlation of distances, The Annals of Statistics 35 (6) (2007) 2769–2794.

[27] R. D. King, S. Muggleton, R. A. Lewis, M. Sternberg, Drug design by machine learning: The use of inductive logic programming to model the

structure-activity relationships of trimethoprim analogues binding to dihydrofolate reductase, Proceedings of the National Academy of Sciences 89 (23) (1992) 11322–11326.

[28] D. Pollard, Asymptotics for least absolute deviation regression estimator, Econometric Theory 7 (02) (1991) 186–199.

[29] W. K. Newey, D. McFadden, Large sample estimation and hypothesis testing, in: Hnadbook of Economietrics IV, Amsterdam: North Holland, 1994, pp. 2113–2245.

[30] E. Bura, C. Pfeiffer, On the distribution of the left singular vectors of a random matrix and its applications, Statistics and Probability Letters 78 (15) (2008) 2275–2280.

Appendix A. Appendix

Appendix A.1. Regularity Conditions

We assume the following regularity conditions.

(A) $\mathbf{\Sigma} = var(\mathbf{X})$ is nonsingular.

(B) $\mathbf{Z}_i$ are independent and identically distributed with probability density $f_{\mathbf{Z}}$. $f_{\mathbf{Z}}$ is identifiable and has a common support, and a unique solution $\boldsymbol{\theta}_0$ exists that satisfies

$$E\left[m'_{\boldsymbol{\theta}}(\mathbf{Z})\right] = \mathbf{0}$$

(C) $E\left[\frac{\partial^2}{\partial \boldsymbol{\theta}\boldsymbol{\theta}^\top} m_{\boldsymbol{\theta}}(\mathbf{Z})\right]$ is nonsingular at $\boldsymbol{\theta} = \boldsymbol{\theta}_0$.

(D) $E(X_j^4) < \infty$ for $j = 1, \cdots, p$

The regularity conditions are rather standard in the context of M-estimation. See, for example, **(author?)** [19].

Appendix A.2. Proof of Theorem 1

Assume that $E(\mathbf{X}) = \mathbf{0}$ without loss of generality, then, the objective function (4) is

$$\Lambda(\alpha, \boldsymbol{\beta}) = \mathrm{Var}(\boldsymbol{\beta}^\top \mathbf{X}) + C\mathrm{E}\left[\ln\left(1 + e^{-\tilde{Y}(\alpha + \boldsymbol{\beta}^\top \mathbf{X})}\right)\right].$$

We have that

$$\mathrm{Var}(\boldsymbol{\beta}^\top \mathbf{X}) = \mathrm{Var}[\mathrm{E}(\boldsymbol{\beta}^\top \mathbf{X}|\mathbf{B}^\top \mathbf{X})] + \mathrm{E}[\mathrm{Var}(\boldsymbol{\beta}^\top \mathbf{X}|\mathbf{B}^\top \mathbf{X})] \geq \mathrm{Var}[\mathrm{E}(\boldsymbol{\beta}^\top \mathbf{X}|\mathbf{B}^\top \mathbf{X})], \tag{Appendix A.1}$$

and

$$\begin{aligned} \mathrm{E}\left[\ln\left(1 + e^{-\tilde{Y}(\alpha + \boldsymbol{\beta}^\top \mathbf{X})}\right)\right] &= \mathrm{E}\left[\mathrm{E}\left[\ln\left(1 + e^{-\tilde{Y}(\alpha + \boldsymbol{\beta}^\top \mathbf{X})}\right)|\tilde{Y}, \mathbf{B}^\top \mathbf{X}\right]\right] \qquad \text{(Appendix A.2)} \\ &\geq \mathrm{E}\left[\ln\left(1 + e^{-\tilde{Y}(\alpha + \mathrm{E}[\boldsymbol{\beta}^\top \mathbf{X}|\tilde{Y}, \mathbf{B}^\top \mathbf{X}])}\right)\right] \\ &= \mathrm{E}\left[\ln\left(1 + e^{-\tilde{Y}(\alpha + \mathrm{E}[\boldsymbol{\beta}^\top \mathbf{X}|\mathbf{B}^\top \mathbf{X}])}\right)\right] \end{aligned}$$

The inequality holds since the logistic loss function is convex, and the last equality is true under (1). Thus, the (possibly non-unique) minimum of (Appendix A.2) is achieved at $\mathrm{E}[\boldsymbol{\beta}^\top \mathbf{X}|\mathbf{B}^\top \mathbf{X}] = P_{\mathbf{B}}(\boldsymbol{\Sigma})\boldsymbol{\beta} \in \mathcal{S}_{Y|\mathbf{X}}$ for any $\alpha \in \mathbb{R}$.

Suppose $\tilde{\boldsymbol{\beta}} \notin \mathcal{S}_{Y|\mathbf{X}}$ is a minimizer of (Appendix A.2), then, $\mathrm{Var}(\tilde{\boldsymbol{\beta}}^\top \mathbf{X}|\mathbf{B}^\top \mathbf{X}) > 0$ by (Appendix A.1) and

$$\Lambda(\alpha, \tilde{\boldsymbol{\beta}}) > \Lambda(\alpha, P_{\mathbf{B}}(\boldsymbol{\Sigma})\tilde{\boldsymbol{\beta}}).$$

Therefore, $\tilde{\boldsymbol{\beta}}$ cannot be the minimizer of $\Lambda(\alpha, \boldsymbol{\beta})$. □

Appendix A.3. Proof of Theorem 2

(a) - consistency

Toward (a), we have $\hat{\Lambda}(\boldsymbol{\theta}) \xrightarrow{p} \Lambda(\boldsymbol{\theta})$ by weak law of large numbers and the consistency of the sample covariance matrix, $\widehat{\boldsymbol{\Sigma}}$. Notice that $\Lambda(\boldsymbol{\theta})$ is a strictly

convex function of $\boldsymbol{\theta}$. By *Convexity Lemma* [28], pointwise convergence of $\hat{\Lambda}(\boldsymbol{\theta})$ implies uniform convergence, that is, $\sup_{\boldsymbol{\theta}} |\hat{\Lambda}(\boldsymbol{\theta}) - \Lambda(\boldsymbol{\theta})| \xrightarrow{p} 0$. Finally, we have $\hat{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}_0$ by Theorem 2.1 of [29].

(b) - asymptotic normality

Notice that $\boldsymbol{\theta}_0$ is a unique solution of

$$\mathrm{E}\left[m'_{\boldsymbol{\theta}}(\mathbf{Z})\right] = E\left[2\boldsymbol{\Sigma}\boldsymbol{\theta} - C\tilde{Y}(1-\pi)\cdot\mathbf{X}^*\right] = \mathbf{0},$$

where $\pi = \left\{1 + \exp(-\tilde{Y}\cdot\boldsymbol{\theta}^\top\mathbf{X}^*)\right\}^{-1}$.

We remark that

i) A mapping $\boldsymbol{\theta} \to m_{\boldsymbol{\theta}}(\mathbf{Z})$ is continuously differentiable for every $\mathbf{Z}$.

ii) $E\|m'_{\boldsymbol{\theta}_0}(\mathbf{Z})\|^2 < \infty$

iii) there exists a constant $A > 0$ such that

$$\left|\frac{\partial^3 m(\boldsymbol{\theta},\mathbf{Z})}{\partial\theta_i\partial\theta_j\partial\theta_k}\right| = \left|C\pi(1-\pi)(1-2\pi)\cdot X_i^*X_j^*X_k^*\tilde{Y}\right| \le A\|\mathbf{X}\|^3.$$

by the condition (D).

Finally, under i) – iii) combined with the regularity condition (C), the desired result follows from Theorem 5.41 of **(author?)** [19]. □

Appendix A.4. Proof of Theorem 3

(a) - Asymptotic normality of $\widehat{\mathbf{M}}$

Let $\mathbf{S}(\boldsymbol{\theta}_{0,k};\mathbf{Z}) = \mathbf{F}_{\boldsymbol{\theta}_{0,k}} m'_{\boldsymbol{\theta}_0}(\mathbf{Z})$ where $F_{\boldsymbol{\theta}_0}$ is the last p rows of $\mathbf{H}^{-1}_{\boldsymbol{\theta}_0}$, then

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_{0,k}) = -n^{-1/2}\sum_{i=1}^{n}\mathbf{S}(\boldsymbol{\theta}_{0,k};\mathbf{Z}_i) + o_p(1).$$

Let $\bar{\mathbf{S}}(\boldsymbol{\theta}_{0,k}, \mathbf{Z}) = n^{-1}\sum_{i=1}^{n}\mathbf{S}(\boldsymbol{\theta}_{0,k}; \mathbf{Z}_i)$ and $\bar{\mathbf{S}}_n(\boldsymbol{\theta}_{0,k}; \mathbf{Z}) = O_p(n^{-1/2})$ because $\mathrm{E}[\mathbf{S}(\boldsymbol{\theta}_{0,k}, \mathbf{Z})] = 0$. We then have

$$
\begin{aligned}
&\mathrm{vec}(\widehat{\mathbf{M}}_n) - \mathrm{vec}(\mathbf{M}_0) \\
=&\sum_{h=1}^{H}\hat{\boldsymbol{\beta}}_{n,k}\otimes\hat{\boldsymbol{\beta}}_{n,k} - \sum_{h=1}^{H}\boldsymbol{\beta}_{0,k}\otimes\boldsymbol{\beta}_{0,k} \\
=&\sum_{h=1}^{H}\left\{\boldsymbol{\beta}_{0,k} - \bar{\mathbf{S}}_n(\boldsymbol{\theta}_{0,k}; \mathbf{Z}) + o_p(n^{-1/2})\right\}\otimes\left\{\boldsymbol{\beta}_{0,k} - \bar{\mathbf{S}}_n(\boldsymbol{\theta}_{0,k}; \mathbf{Z}) + o_p(n^{-1/2})\right\} - \sum_{h=1}^{H}\boldsymbol{\beta}_{0,k}\otimes\boldsymbol{\beta}_{0,k} \\
=&-\sum_{h=1}^{H}\left\{\boldsymbol{\beta}_{0,k}\otimes\bar{\mathbf{S}}_n(\boldsymbol{\theta}_{0,k}; \mathbf{Z}) + \bar{\mathbf{S}}_n(\boldsymbol{\theta}_{0,k}; \mathbf{Z})\otimes\boldsymbol{\beta}_{0,k}\right\} + \sum_{h=1}^{H}\bar{\mathbf{S}}_n(\boldsymbol{\theta}_{0,k}; \mathbf{Z})\otimes\bar{\mathbf{S}}_n(\boldsymbol{\theta}_{0,k}; \mathbf{Z}) + o_p(n^{-1/2}) \\
=&-\sum_{h=1}^{H}\left\{\boldsymbol{\beta}_{0,k}\otimes\bar{\mathbf{S}}_n(\boldsymbol{\theta}_{0,k}; \mathbf{Z}) + \bar{\mathbf{S}}_n(\boldsymbol{\theta}_{0,k}; \mathbf{Z})\otimes\boldsymbol{\beta}_{0,k}\right\} + o_p(n^{-1/2}).
\end{aligned}
$$

where $\otimes$ denotes the Kronecker product operator. Let $\mathbf{T}_{u,v} \in \mathbb{R}^{uv\times uv}$ denote a commutation matrix that satisfies $\mathbf{T}_{u,v}vec(\mathbf{A}) = vec(\mathbf{A}^\top)$ for a matrix $\mathbf{A} \in \mathbb{R}^{u\times v}$. It is known that the commutation matrix $\mathbf{T}$ has the following properties:

- $\mathbf{T}_{i_1,i_2} = \mathbf{T}_{i_2,i_1}^\top$.
- $\mathbf{A}\otimes\mathbf{B} = \mathbf{T}_{i_1,i_3}(\mathbf{B}\otimes\mathbf{A})\mathbf{T}_{i_4,i_2}$ for $\mathbf{A} \in \mathbb{R}^{i_1\times i_2}$ and $\mathbf{B} \in \mathbb{R}^{i_3\times i_4}$.

Therefore

$$
\sqrt{n}\{\mathrm{vec}(\widehat{\mathbf{M}}_n) - \mathrm{vec}(\mathbf{M}_0)\} = -\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left\{(\mathbf{I}_{p^2} + \mathbf{T}_{p,p})\sum_{k=1}^{h}\boldsymbol{\beta}_{0,k}\otimes\mathbf{S}(\boldsymbol{\theta}_{0,k}; \mathbf{Z}_i)\right\} + o_p(1),
$$

Finally the desired result is then followed by Central Limit Theorem with covariance matrix $\boldsymbol{\Sigma}_{\mathbf{M}}$ as follows:

$$
\boldsymbol{\Sigma}_{\mathbf{M}} = (\mathbf{I}_{p^2} + \mathbf{T}_{p,p})\sum_{k=1}^{h}\sum_{k'=1}^{h}\left(\boldsymbol{\beta}_{0,k}\boldsymbol{\beta}_{0,k'}^\top\otimes\mathrm{E}\left[\mathbf{S}(\boldsymbol{\theta}_{0,k}, \mathbf{Z})\mathbf{S}^\top(\boldsymbol{\theta}_{0,k'}, \mathbf{Z})\right]\right)(\mathbf{I}_{p^2} + \mathbf{T}_{p,p}),
$$

where $\mathbf{I}_p$ denotes the p-dimensional identity matrix.

(b) - Asymptotic normality of $\widehat{\mathbf{V}}$

Finally, asymptotic normality of $\widehat{\mathbf{V}}$ follows as a direct consequence of that of $\widehat{\mathbf{M}}$ and **(author?)** [30]. The asymptotic variance is given by

$$\mathbf{\Sigma_V} = (\mathbf{D}^{-1}\mathbf{U}^\top \otimes \mathbf{I}_p)\mathbf{\Sigma_M}(\mathbf{U}\mathbf{D}^{-1} \otimes \mathbf{I}_p)$$

where $\mathbf{U}$ is a $p \times d$ matrix with columns that are eigenvectors of $\mathbf{M}_0$ corresponding to its nonzero eigenvalues, and $\mathbf{D}$ is a $d \times d$ diagonal matrix with the nonzero eigenvalues as diagonal elements.

□

Appendix A.5. Proof of Theorem 4

In order to prove Theorem 4, we first introduce two lemmas.

Lemma 1. *Under the regularity conditions, if* $b_n = \max_{1\leq j\leq p} p''_{\lambda_n}(\max_{1\leq k\leq h} |\beta^0_{jk}|)$ *converges to* 0*, then there exists a local minimizer* $\widehat{\mathbf{\Omega}}$ *of* $Q(\mathbf{\Omega})$ *such that* $\|\hat{\boldsymbol{\theta}}_k - \boldsymbol{\theta}_{0,k}\| = O_p(n^{-1/2}+a_n)$ *for* $k = 1, \cdots, h$ *where* $a_n = \max_{1\leq j\leq p} p'_{\lambda_n}(\max_{1\leq k\leq h} |\beta^0_{jk}|)$ *with* β^0_{jk} *being the* j*th element of* $\boldsymbol{\beta}_{0,k}$*.*

Proof of Lemma 1. Let $\delta_n = n^{-1/2} + a_n$. We show that for arbitrary given $\epsilon > 0$ a constant $C_1 > 0$ exists such that

$$P\left\{\inf_{\|\mathbf{E}\|=C_1} Q(\mathbf{\Omega} + \delta_n\mathbf{E}) > Q(\mathbf{\Omega})\right\} \geq 1 - \epsilon$$

Now we have that

$$\begin{aligned}
&Q(\mathbf{\Omega}_0 + \delta_n\mathbf{E}) - Q(\mathbf{\Omega}_0)\\
&\geq \sum_{k=1}^{h} \hat{\Lambda}_k(\boldsymbol{\theta}_{0,k} + \delta_n\mathbf{e}_k) - \hat{\Lambda}_k(\boldsymbol{\theta}_{0,k}) + \sum_{j=1}^{q}\left(p_{\lambda_n}(\max_{1\leq k\leq h}|\beta^0_{jk} + \delta_n e_{jk}|) - p_{\lambda_n}(\max_{1\leq k\leq h}|\beta^0_{jk}|)\right)\\
&=: D_1 + D_2
\end{aligned}$$

By Taylor expansion,

$$D_1 = \delta_n \sum_{k=1}^{h} \mathbf{e}_k^\top \frac{\partial}{\partial \boldsymbol{\theta}} \hat{\Lambda}_k(\boldsymbol{\theta}_{0,k}) + \frac{1}{2}\delta_n^2 \sum_{k=1}^{h} \mathbf{e}_k^\top \left\{ \frac{\partial^2}{\partial \boldsymbol{\theta}^2} \hat{\Lambda}_k(\boldsymbol{\theta}_{0,k}) \right\} \mathbf{e}_k (1+o_p(1)) := D_{11} + D_{12}.$$

Then,

$$|D_{11}| \le \delta_n \sum_{k=1}^{h} \left| \mathbf{e}_k^\top \frac{\partial}{\partial \boldsymbol{\theta}} \hat{\Lambda}_k(\boldsymbol{\theta}_{0,k}) \right| \le \delta_n \sum_{k=1}^{h} \|\mathbf{e}_k\| \cdot \left\| \frac{\partial}{\partial \boldsymbol{\theta}} \hat{\Lambda}_k(\boldsymbol{\theta}_{0,k}) \right\| = O_p(\delta_n n^{-1/2}) \sum_{k=1}^{h} \|\mathbf{e}_k\|$$

and

$$D_{12} = \frac{1}{2}\delta_n^2 \sum_{k=1}^{h} \mathbf{e}_k^\top \mathbf{H}_{\boldsymbol{\theta}_{0,k}} \mathbf{e}_k (1 + o_p(1)).$$

Now,

$$D_2 = \sum_{j=1}^{q} p'_{\lambda_n}(\max_{1\le k\le h} |\beta_{jk}^0|)(\max_{1\le k\le h} |\beta_{jk}^0 + \delta_n e_{jk}| - \max_{1\le k\le h} |\beta_{jk}^0|) +$$
$$\sum_{j=1}^{q} \frac{1}{2} p''_{\lambda_n}(\max_{1\le k\le h} |\beta_{jk}^0|)(\max_{1\le k\le h} |\beta_{jk}^0 + \delta_n e_{jk}| - \max_{1\le k\le h} |\beta_{jk}^0|)^2 (1 + o(1)),$$

and hence

$$|D_2| \le \sqrt{q}\delta_n a_n \sum_{k=1}^{h} \|\mathbf{e}_k\| + \frac{1}{2}\delta_n^2 \max_{1\le j\le q} p''_{\lambda_n}(|\beta_{jk}^0|) h \sum_{k=1}^{h} \|\mathbf{e}_k\|^2.$$

Note that D_{12}, which is always positive, dominates all other terms, hence, the desired result follows by letting $C_1 = \|\mathbf{E}\| = (\sum_{k=1}^{h} \|\mathbf{e}_k\|^2)^{1/2}$ sufficiently large.

□

Lemma 2. *Under the conditions of Theorem 4, for any given* $(q+1)\times h$ *submatrix* $\boldsymbol{\Omega}_+ = (\boldsymbol{\theta}_{1+}, \cdots, \boldsymbol{\theta}_{h+})$ *satisfying* $\|\boldsymbol{\theta}_{k+} - \boldsymbol{\theta}_{0,k+}\| = O_p(n^{-1/2}), k = 1, \cdots, h$ *and any* $(p-q)\times h$ *submatrix* $\boldsymbol{\Omega}_- = (\boldsymbol{\theta}_{1-}, \cdots, \boldsymbol{\theta}_{h_-})$ *satisfying* $\|\boldsymbol{\theta}_{k-} - \boldsymbol{\theta}_{0,k-}\| \le C_2 n^{-1/2}$ *for a constant* C_2, $k = 1, \cdots, h$, *we have, with probability tending to*

one,

$$Q((\boldsymbol{\Omega}_+^\top, \mathbf{0}_{(p-q)\times h}^\top)^\top) = \min_{\|\boldsymbol{\beta}_{k-}\| \le C_2 n^{-1/2}} Q((\boldsymbol{\Omega}_+^\top, \boldsymbol{\Omega}_-^\top)^\top)$$

Proof of Lemma 2. It is sufficient to show that with probability tending to one as $n \to \infty$ for any given $\boldsymbol{\theta}_{k+}$ satisfying $\|\boldsymbol{\theta}_{k+} - \boldsymbol{\theta}_{0,k+}\| = O_p(n^{-1/2})$ and any constant $C_2 > 0$, $j = q+1, \cdots, p$,

$$\frac{\partial}{\partial \beta_{jk}^r} Q(\boldsymbol{\Omega}) > 0 \text{ for } 0 < \beta_{jk} < C_2 n^{-1/2} \text{ and } \beta_{jk} = \max_{1 \le m \le h} |\beta_{jm}|$$
$$\frac{\partial}{\partial \beta_{jk}^l} Q(\boldsymbol{\Omega}) < 0 \text{ for } 0 < -\beta_{jk} < C_2 n^{-1/2} \text{ and } \beta_{jk} = -\max_{1 \le m \le h} |\beta_{jm}|$$

where $\partial/\partial\beta_{jk}^l$ and $\partial/\partial\beta_{jk}^r$ denote the left and right hand partial derivative, respectively.

Applying Taylor expansion, we have

$$\frac{\partial}{\partial \beta_{jk}} \hat{\Lambda}_k(\boldsymbol{\theta}_k) = \frac{\partial}{\partial \beta_{jk}} \hat{\Lambda}_k(\boldsymbol{\theta}_{0,k}) + \sum_{l=1}^{p} \frac{\partial^2}{\partial \beta_{jk} \partial \beta_{lk}} \hat{\Lambda}_k(\boldsymbol{\theta}_{0,k})(\beta_{lk} - \beta_{lk}^0) := E_1 + E_2.$$

Note that $E_1 = O_p(n^{-1/2})$ and E_2 can be decomposed as

$$E_2 = \sum_{l=1}^{p} \left[\frac{\partial^2}{\partial \beta_{jk} \partial \beta_{lk}} \hat{\Lambda}_k(\boldsymbol{\theta}_{0,k}) - h_{\boldsymbol{\theta}_{0,k}}^{lj} \right] (\beta_{lk} - \beta_{lk}^0) + \sum_{l=1}^{p} h_{\boldsymbol{\theta}_{0,k}}^{lj} (\beta_{lk} - \beta_{lk}^0) := E_{21} + E_{22}$$

where $h_{\boldsymbol{\theta}_{0,k}}^{lj}$ denotes the (l,j)th element of $\mathbf{H}_{\boldsymbol{\theta}_{0,k}}$. Now, it is clear that $E_{21} = O_p(n^{-1})$ and $E_{21} = O_p(n^{-1/2})$ by applying Cauchy-Schwarz inequalities, respectively. As a consequence, we have

$$\frac{\partial}{\partial \beta_{jk}} \hat{\Lambda}_k(\boldsymbol{\theta}_k) = O_p(n^{-1/2}).$$

Finally, we have for $\beta_{jk} = \max_{1 \le m \le h} |\beta_{jm}|$,

$$\frac{\partial}{\partial \beta_{jk}} Q(\boldsymbol{\Omega}) = \lambda_n \left\{ \lambda_n^{-1} p'_{\lambda_n}(|\beta_{jk}|) \text{sign}(\beta_{jk}) + O_p(n^{-1/2}/\lambda_n) \right\}.$$

By the assumptions that $n^{-1/2}\lambda_n \to 0$ and $\liminf_{n\to\infty} \liminf_{\theta\to 0+} p'_{\lambda_n}(\theta)/\lambda_n > 0$, the first term dominates the second term and the desired result follows. $\square$

Proof of Theorem 4

Notice for the SCAD penalty that $a_n = 0$ and $b_n = 0$ when $\lambda_n < a^{-1} \max_{1\le j\le p} \max_{1\le k\le h} |\beta^0_{jk}|$. By Lemma 1, a $\sqrt{n}$-consistent local minimizer $\hat{\mathbf{\Omega}}$ of $Q(\mathbf{\Omega})$ exists. By Lemma 2, $\hat{\mathbf{\Omega}} = (\mathbf{\Omega}_+^\top, \mathbf{0}_{(p-q)\times h})^\top$ with probability tending to one, which proves part (a). As a consequence, we are in effect minimizing

$$\tilde{Q}(\mathbf{\Omega}_+) = \boldsymbol{\beta}_{k+}^\top \widehat{\mathbf{\Sigma}} \boldsymbol{\beta}_{k+} + \frac{C}{n} \sum_{i=1}^{n} \ln(1 + e^{-\tilde{y}_{ik}\{\alpha_k + \boldsymbol{\beta}_{k+}^\top \mathbf{x}_{i+}\}})$$

over $\mathbf{\Omega}_+$ with probability tending to one. This completes the proof of part (b). $\square$