ELSEVIER

Statistics & Probability Letters 59 (2002) 135–144

STATISTICS & PROBABILITY LETTERS

www.elsevier.com/locate/stapro

Data analysis in supersaturated designs

Runze Li[a,b,*], Dennis K.J. Lin[a,b]

[a] *Department of Statistics, The Pennsylvania State University, University Park, PA 16802-2111, USA*
[b] *Department of Management Science and Information Systems, The Pennsylvania State University, University Park, PA 16802-2111, USA*

Received November 2001; received in revised form March 2002

Abstract

Supersaturated designs (SSDs) can save considerable cost in industrial experimentation when many potential factors are introduced in preliminary studies. Analyzing data in SSDs is challenging because the number of experiments is less than the number of candidate factors. In this paper, we introduce a variable selection approach to identifying the active effects in SSD via nonconvex penalized least squares. An iterative ridge regression is employed to find the solution of the penalized least squares. We provide both theoretical and empirical justifications for the proposed approach. Some related issues are also discussed.

Keywords: AIC; BIC; Penalized least squares; SCAD; Stepwise regression

1. Introduction

Many preliminary studies in industrial experimentation contain a large number of potentially relevant factors, but actual active effects are believed to be sparse. To save experimental cost, experimenters have tried to reduce the number of experiments. An efficient use of experimental units is the employment of supersaturated design (SSD), in which the number of experiments is less than the number of candidate factors (see, for example, Lin, 1991). Construction of SSDs has received increasing attention recently. For example, Lin (1995) introduced an approach to generating system SSDs, Fang et al. (2000) proposed an approach for constructing multi-level SSD via a quasi-Monte Carlo approach. Also, see Lin (2000) and references therein. Since the number of candidate factors is more than the number of experiments in SSD, variable selection is fundamental in analyzing SSD for identifying sparse active effects. Some traditional approaches, such as best subset variable selection, are not feasible, while stepwise variable selection may not be appropriate (see, for example,

* Corresponding author.

0167-7152/02/$ - see front matter
PII: S0167-7152(02)00140-2

Westfall et al., 1998). In this paper, we introduce an approach from the viewpoint of frequentist analysis.

Fan and Li (2001) proposed a class of variable selection procedures via nonconcave penalized likelihood. They showed that with the proper choice of penalty function and regularization parameter, their approaches possess an oracle property. Namely, the true coefficients that are zero are automatically estimated as zero, and the other coefficients are estimated as if the true submodel were known in advance. However, their approach cannot be directly applied for analyzing SSD because regularity conditions imposed in their paper require the design matrix to be full rank. This cannot be satisfied by an SSD. In this paper, we extend the nonconcave penalized likelihood approaches to least squares, namely nonconvex penalized least squares, and focus on the situation in which the design matrix is not full rank. Theoretic properties of the proposed approach are investigated, and empirical comparisons via Monte Carlo simulation are conducted.

This paper is organized as follows. In Section 2, we briefly discuss nonconvex penalized least squares and introduce a variable selection procedure for SSD. Root n consistency of the resulting estimator via penalized least squares is established. We also show that the introduced procedure possesses an oracle property. An iterative ridge regression algorithm is employed to find the solution of the penalized least squares. A choice of initial value of unknown coefficients for the iterative ridge regression is suggested. In Section 3, some empirical comparisons via Monte Carlo simulation are conducted. A real data example is used for illustration. Section 4 gives the proof of the main theorem. Some conclusions are given in Section 5.

2. Variable selection for screening active effects

2.1. Preliminary

Assume that observations $Y_1, \ldots, Y_n$ are independent samples from the linear regression model

$$Y_i = \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} + \varepsilon_i \tag{2.1}$$

with $\mathrm{E}(\varepsilon_i) = 0$ and $\mathrm{var}(\varepsilon_i) = \sigma^2$, where $\mathbf{x}_i$ is the vector of input variables. Following Fan and Li (2001), a form of penalized least squares is defined as

$$Q(\boldsymbol{\beta}) \equiv \frac{1}{2n}\sum_{i=1}^{n}(Y_i - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta})^2 + \sum_{j=1}^{d} p_{\lambda_n}(|\beta_j|), \tag{2.2}$$

where $p_{\lambda_n}(\cdot)$ is a penalty function, and λ_n is a tuning parameter, which can be chosen by a data-driven approach, such as cross-validation (CV) and generalized cross-validation (GCV, Craven and Wahba, 1979). Many variable selection criteria are closely related to this penalized least squares. Take the penalty function to be the entropy penalty, namely, $p_{\lambda_n}(|\beta|) = \frac{1}{2}\lambda_n^2 I(|\beta| \neq 0)$, where $I(\cdot)$ is an indicator function. Note that the dimension or the size of a model equals the number of nonzero regression coefficients in the model. This actually equals $\sum_j I(|\beta_j| \neq 0)$. In other words, the penalized least squares (2.2) with the entropy penalty can be rewritten as

$$\frac{1}{2n}\sum_{i=1}^{n}(Y_i - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta})^2 + \frac{1}{2}\lambda_n^2|M|, \tag{2.3}$$

where $|M|=\sum_j I(|\beta_j|\neq 0)$, the size of the underlying candidate model. Hence, many popular variable selection criteria can be derived from the penalized least squares (2.3) by taking different values of λ_n. For instance, the AIC and BIC correspond to $\lambda_n=\sqrt{2}(\sigma/\sqrt{n})$ and $\sqrt{\log n}(\sigma/\sqrt{n})$, respectively, although these two criteria were motivated from different principles. The entropy penalty is not continuous, and can be improved by its smooth version, namely a hard thresholding penalty function,

$$p_\lambda(|\beta|)=\lambda^2-(|\beta|-\lambda)^2 I(|\beta|<\lambda),$$

proposed by Fan (1997).

Other penalty functions have been used in the literature. The L_2 penalty $p_{\lambda_n}(|\beta|)=2^{-1}\lambda_n|\beta|^2$ yields a ridge regression, and the L_1 penalty $p_{\lambda_n}(|\beta|)=\lambda_n|\beta|$ results in LASSO (Tibshirani, 1996). More generally, the L_q penalty leads to a bridge regression (see Frank and Friedman, 1993; Fu, 1998; Knight and Fu, 2000).

All aforementioned penalties do not satisfy the conditions for desired properties in terms of continuity, sparsity and unbiasedness, advocated by Fan and Li (2001). They suggested using the smoothly clipped absolute deviation (SCAD) penalty, proposed by Fan (1997). The first-order derivative of SCAD is defined by

$$p'_\lambda(\beta)=\lambda\left\{I(\beta\leqslant\lambda)+\frac{(a\lambda-\beta)_+}{(a-1)\lambda}I(\beta>\lambda)\right\}$$

for $\beta>0$ with $a=3.7$, and $p_\lambda(0)=0$. For simplicity of presentation, we will use SCAD for all procedures using the SCAD penalty. As recommended by Fan and Li, we employ penalized least squares with the SCAD penalty to identify the sparse active effects in the analysis stage of SSD in this paper.

2.2. Asymptotic properties

When $(\mathbf{x}_i, Y_i)$, $i=1,\ldots,n$, are independent and identically distributed, Fan and Li have established an oracle property for their penalized likelihood estimator. In what follows, we will show that the oracle property still holds for the penalized least-squares estimator when the predictor variable $\mathbf{x}$ is a fixed design even if design matrix is singular. We next introduce some related notations.

Denote by $\boldsymbol{\beta}_0$ the true value of $\boldsymbol{\beta}$, and let $\boldsymbol{\beta}_0=(\beta_{10},\ldots,\beta_{d0})^{\mathrm{T}}=(\boldsymbol{\beta}_{10}^{\mathrm{T}},\boldsymbol{\beta}_{20}^{\mathrm{T}})^{\mathrm{T}}$. Without loss of generality, assume that $\boldsymbol{\beta}_{20}=\mathbf{0}$, and all components of $\boldsymbol{\beta}_{10}$ are not equal to 0. Denote by s the number of nonzero components of $\boldsymbol{\beta}$. Let $\mathbf{X}_1$ consist of the first s columns of $\mathbf{X}$, the design matrix of the linear regression model (2.1), and $\mathbf{x}_{1k}$ consist of the first s components of $\mathbf{x}_k$, $k=1,\ldots,n$. Define $\mathbf{V}=\lim_{n\to\infty}(1/n)\mathbf{X}^{\mathrm{T}}\mathbf{X}$ and let $\mathbf{V}_{11}$ consist of the first s columns and rows of $\mathbf{V}$.

To establish the oracle property for the penalized least-squares estimate, we need the following two conditions:

(C1) $\max_{1\leqslant k\leqslant n}\mathbf{x}_{1k}^{\mathrm{T}}(\mathbf{X}_1^{\mathrm{T}}\mathbf{X}_1)^{-1}\mathbf{x}_{1k}\to 0$, as $n\to\infty$.
(C2) $\mathbf{V}$ is finite and $\mathbf{V}_{11}>0$.

Conditions (C1) and (C2) guarantee that the asymptotic normality holds for the least-squares estimator of $\hat{\boldsymbol{\beta}}_1$. We show in Theorem 1 that the penalized least-squares estimate is root n consistent and possesses the oracle property for the penalized least-squares estimator with the SCAD penalty.

Theorem 1. *Consider model* (2.1) *and suppose that the random errors are independent and identically distributed with zero mean and finite positive variance* σ^2. *Suppose that conditions* (C1) *and* (C2) *hold. For the SCAD penalty, if* $\lambda_n \to 0$ *and* $\sqrt{n}\lambda_n \to \infty$, *then*

(a) (Root n consistency). *With probability tending to one, there exists a local minimizer* $\hat{\boldsymbol{\beta}}$ *of* $Q(\boldsymbol{\beta})$, *defined in* (2.2), *such that* $\hat{\boldsymbol{\beta}}$ *is a root n consistent estimator of* $\boldsymbol{\beta}$;
(b) (Oracle property). *With probability tending to one, the root n consistent estimator in Part* (a) *satisfies* $\hat{\boldsymbol{\beta}}_2 = \mathbf{0}$ *and*

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{10}) \to \mathrm{N}(0, \sigma^2 \mathbf{V}_{11}^{-1}).$$

The proof of Theorem 1 is given in Section 4. From Theorem 1, with proper rate of λ_n, the SCAD results in a root n consistency estimator. Moreover, the resulting estimator correctly identifies the inactive effects and estimates the active effects as if we knew the true submodel in advance.

2.3. Iterative ridge regression

2.3.1. Local quadratic approximation

Note that the SCAD penalty function is singular at the origin and may not have the second derivative at some points. In order to apply the Newton Raphson algorithm to the penalized least squares, Fan and Li (2001) locally approximate the SCAD penalty function by a quadratic function as follows. Given an initial value $\boldsymbol{\beta}^{(0)}$ that is close to the true value of $\boldsymbol{\beta}$, when $\beta_j^{(0)}$ is not very close to 0, the penalty $p_\lambda(|\beta_j|)$ can be locally approximated by the quadratic function as

$$[p_\lambda(|\beta_j|)]' = p'_\lambda(|\beta_j|)\,\mathrm{sgn}(\beta_j) \approx \{p'_\lambda(|\beta_j^{(0)}|)/|\beta_j^{(0)}|\}\beta_j$$

and set $\hat{\beta}_j = 0$ if $\beta_j^{(0)}$ is very close to 0. With the local quadratic approximation, the solution for the penalized least squares can be found by iteratively computing the following ridge regression with an initial value $\boldsymbol{\beta}^{(0)}$:

$$\boldsymbol{\beta}^{(1)} = \{\mathbf{X}^{\mathrm{T}}\mathbf{X} + n\Sigma_\lambda(\boldsymbol{\beta}^{(0)})\}^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y},$$

where

$$\Sigma_\lambda(\boldsymbol{\beta}^{(0)}) = \mathrm{diag}\{p'_\lambda(|\beta_1^{(0)}|)/|\beta_1^{(0)}|, \ldots, p'_\lambda(|\beta_d^{(0)}|)/|\beta_d^{(0)}|\}.$$

This can be easily implemented in many statistical packages.

2.3.2. Initial value

The local quadratic approximation requires a good initial value which is close to the true value $\boldsymbol{\beta}_0$. When the design matrix is full rank, the least-squares estimate can serve as the initial value of $\boldsymbol{\beta}$. When the sample size is relatively large, the least-squares estimate possesses root n consistency, and hence it is very close to the true value of $\boldsymbol{\beta}$. In the SSD settings, however, the number of experiments may be less than the number of potential candidates, and therefore the design matrix is not full rank. In fact, the regression coefficients are not identifiable without further assumptions, for example, the effect sparsity assumption. Here we use stepwise variable selection to find an initial

value of $\boldsymbol{\beta}$. In other words, we first apply stepwise variable selection to the full model with small thresholding values (i.e. large value of significance level α) such that all active factors are included in the selected model. Of course, some insignificant factors may still stay in the resulting model at this step.

3. Simulation study and example

In this section we compare the performance of the SCAD and stepwise variable selection procedure for analyzing SSD. All simulations are conducted using MATLAB codes. GCV (Craven and Wahba, 1979) was used for determining the regularization parameter λ in the SCAD. In the following examples, the level of significance is set to be 0.1 for both F-enter and F-remove in the stepwise variable selection procedure for choosing an initial value for the SCAD. The comparison is made in terms of the ability of identifying true model and the size of resulting model.

Example 1. Given the design matrix constructed in Lin (1993), we simulated 1000 data sets from the linear model

$$Y = \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta} + \varepsilon,$$

where the random error ε is $\mathrm{N}(0,1)$. We generate data from the following three models:

Model I: $\beta_1 = 8,\ \beta_{12} = 5$ and all other components of $\boldsymbol{\beta}$ are equal to 0.
Model II: $\beta_1 = 10,\ \beta_2 = 9,\ \beta_3 = 2$ and all other components of $\boldsymbol{\beta}$ are equal to 0.
Model III: $\beta_1 = -20,\ \beta_3 = 12,\ \beta_5 = 10,\ \beta_7 = 5,\ \beta_{16} = 2$ and all other components of $\boldsymbol{\beta}$ are equal to 0.

In Model I, there are two large active effects; while in Models II and III, there are some large effects, some moderate effects and a small effect. In our simulations, we also employ stepwise regression with three different criteria. One selects significant effects by setting the level of significance to be 0.05, the other two select a subset with the best AIC and BIC score in the stepwise regression, respectively. Table 1 summarizes the simulation results. From Table 1, it is clear that SCAD outperforms all three stepwise regression methods in terms of rate of identifying true models and size of selected models.

Example 2 (Williams Rubber Experiment). The proposed procedure is applied to the SSD constructed by Lin (1993). Stepwise variable selection selects the following factors: X_{15}, X_{12}, X_{20}, X_4, X_{10}, X_{11}, X_7, X_1, X_{14}, X_{17} and X_{22}. Then the SCAD procedure is applied. The selected tuning parameter is $\hat{\lambda} = 6.5673$. The final model selected by the SCAD identifies X_4, X_{12}, X_{15}, X_{20} as the active effects. The result is consistent with the conclusion of Williams (1968).

4. Proofs

To prove Theorem 1, we prove the following two lemmas first.

Table 1
Percent of 1000 simulations in Example 1

Model	Method	Rate of the true model being identified	Avg. size of fitted model	
			Median	Mean
I	*True model*: $Y = 8x_1 + 5x_{12} + \varepsilon$			
	Stepwise ($p = 0.05$)	27.00%	3	3.7
	Stepwise (AIC)	2.20%	5	5.0
	Stepwise (BIC)	11.10%	4	4.2
	SCAD	82.70%	2	2.2
II	*True model*: $Y = 10x_1 + 9x_2 + 2x_3 + \varepsilon$			
	Stepwise ($p = 0.05$)	32.30%	4	4.5
	Stepwise (AIC)	3.00%	6	5.74
	Stepwise (BIC)	13.50%	5	5.00
	SCAD	74.70%	3	3.34
III	*True model*: $Y = -20x_1 + 12x_3 + 10x_5 + 5x_7 + 2x_{16} + \varepsilon$			
	Stepwise ($p = 0.05$)	38.90%	6	6.17
	Stepwise (AIC)	7.00%	7	7.01
	Stepwise (BIC)	24.70%	6	6.43
	SCAD	71.90%	5	5.39

Lemma 1. *Under the conditions of Theorem* 1, *there exists a large constant C such that*

$$P\left\{\inf_{\|\mathbf{u}\|=C} Q(\boldsymbol{\beta}_0 + n^{-1/2}\mathbf{u}) > Q(\boldsymbol{\beta}_0)\right\} \geqslant 1 - \varepsilon, \tag{4.1}$$

where $\mathbf{u} = (\mathbf{u}_1^{\mathrm{T}}, \mathbf{0})$ *and the dimension of* $\mathbf{u}_1$ *is the same as that of* $\boldsymbol{\beta}_{10}$.

Proof. Define

$$D_n(\mathbf{u}) \equiv Q(\boldsymbol{\beta}_0 + \alpha_n\mathbf{u}) - Q(\boldsymbol{\beta}_0).$$

Note that $p_{\lambda_n}(0) = 0$ and $\boldsymbol{\beta}_{20} = 0$,

$$D_n(\mathbf{u}) \geqslant \frac{1}{2n}\sum_{i=1}^{n}\{\|\mathbf{y} - \mathbf{X}_1(\boldsymbol{\beta}_{10} + \mathbf{u}_1)\|^2 - \|\mathbf{y} - \mathbf{X}_1\boldsymbol{\beta}_{10}\|^2\}$$
$$+\sum_{j=1}^{s}\{p_{\lambda_n}(|\beta_{j0} + \alpha_n u_j|) - p_{\lambda_n}(|\beta_{j0}|)\}, \tag{4.2}$$

where $\mathbf{y} = (Y_1, \ldots, Y_n)^{\mathrm{T}}$. The first term in (4.2) can be simplified

$$\frac{1}{2}n^{-1}\mathbf{u}_1^{\mathrm{T}}\left(\frac{1}{n}\mathbf{X}_1^{\mathrm{T}}\mathbf{X}_1\right)\mathbf{u} - n^{-1/2}\sum_{i=1}^{n}\mathbf{u}^{\mathrm{T}}\mathbf{x}_{1i}(Y_i - \mathbf{x}_{1i}^{\mathrm{T}}\boldsymbol{\beta}_{10}). \tag{4.3}$$

By the assumption of (C2), the first term in (4.3) equals $n^{-1}\frac{1}{2}\mathbf{u}_1^{\mathrm{T}}\mathbf{V}_{11}\mathbf{u}_1\{1+\mathrm{o}(1)\}$. Using $R=E(R)+\mathrm{O_P}\{\sqrt{\mathrm{var}(R)}\}$ for any random variable with finite second moment, it follows that the second term in (4.3) equals to $n^{-1}\sqrt{\mathbf{u}_1^{\mathrm{T}}\mathbf{V}_{11}\mathbf{u}_1}\mathrm{O_P}(1)$. Note that $\mathbf{V}_{11}$ is finite and positive definite. By choosing a sufficiently large C, the first term will dominate the second term, uniformly in $\|\mathbf{u}_1\|=C$.

By Taylor's expansion, the second term on the right-hand side of (4.2) becomes

$$\sum_{j=1}^{s}[n^{-1/2}p'_{\lambda_n}(|\beta_{j0}|)\,\mathrm{sgn}(\beta_{j0})u_j+n^{-1}p''_{\lambda_n}(|\beta_{j0}|)u_j^2(1+\mathrm{o}(1))]$$

which is bounded by

$$\sqrt{s}n^{-1/2}a_n\|\mathbf{u}_1\|+n^{-1}b_n\|\mathbf{u}_1\|^2,$$

where

$$a_n=\max\{|p'_{\lambda_n}(|\beta_{j0}|)|:\ \beta_{j0}\neq 0\}\quad\text{and}\quad b_n=\max\{|p''_{\lambda_n}(|\beta_{j0}|)|:\ \beta_{j0}\neq 0\}.$$

For the SCAD penalty, if $\lambda_n\to 0$, then $a_n=0$ and $b_n=0$ when n is large enough. Therefore, the second term in (4.2) is dominated by the first term of (4.3). Hence, by choosing sufficiently large C, (4.1) holds. This completes the proof. □

Lemma 2. *Under the conditions of Theorem* 1, *for any given* $\boldsymbol{\beta}_1$ *satisfying that* $\|\boldsymbol{\beta}_1-\boldsymbol{\beta}_{10}\|=\mathrm{O_P}(n^{-1/2})$ *and any constant* C, *the following equation holds with probability tending to one*,

$$Q\left\{\begin{pmatrix}\boldsymbol{\beta}_1\\ \mathbf{0}\end{pmatrix}\right\}=\min_{\|\boldsymbol{\beta}_2\|\leqslant Cn^{-1/2}}Q\left\{\begin{pmatrix}\boldsymbol{\beta}_1\\ \boldsymbol{\beta}_2\end{pmatrix}\right\}.$$

Proof. It is sufficient to show that, with probability tending to 1 as $n\to\infty$, for any $\boldsymbol{\beta}_1$ satisfying $\|\boldsymbol{\beta}_1-\boldsymbol{\beta}_{10}\|=\mathrm{O_P}(n^{-1/2})$ and for some small $\varepsilon_n=Cn^{-1/2}$ and $j=s+1,\ldots,d$

$$\frac{\partial Q(\boldsymbol{\beta})}{\partial\beta_j}>0\quad\text{for } 0<\beta_j<\varepsilon_n \tag{4.4}$$

and

$$\frac{\partial Q(\boldsymbol{\beta})}{\partial\beta_j}<0\quad\text{for } -\varepsilon_n<\beta_j<0. \tag{4.5}$$

By some straightforward computations, it follows that

$$\frac{\partial Q(\boldsymbol{\beta})}{\partial\beta_j}=-\frac{1}{n}\mathbf{x}_{(j)}^{\mathrm{T}}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}_0)+\frac{1}{n}\mathbf{x}_{(j)}^{\mathrm{T}}\mathbf{X}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)+p'_{\lambda_n}(|\beta_j|)\,\mathrm{sgn}(\beta_j),$$

where $\mathbf{x}_{(j)}$ is the jth column of $\mathbf{X}$.

Since $\mathbf{V}$ is finite, $\mathrm{var}\{(1/n)\mathbf{x}_{(j)}^{\mathrm{T}}(\mathbf{y}-\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0)\}=\mathrm{O}(n^{-1})$, and it follows that

$$\frac{1}{n}\mathbf{x}_{(j)}^{\mathrm{T}}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}_0)=\mathrm{O_P}(n^{-1/2}).$$

By condition (C2)

$$\frac{1}{n}\mathbf{x}_{(j)}^{\mathrm{T}}\mathbf{X} = \mathbf{V}_j(1 + \mathrm{o}(1)),$$

where $\mathbf{V}_j$ is the jth column of $\mathbf{V}$.

By the assumption that $\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\| = \mathrm{O_P}(n^{-1/2})$, we have

$$\frac{\partial Q(\boldsymbol{\beta})}{\partial \beta_j} = \lambda_n\{\lambda_n^{-1} p'_{\lambda_n}(|\beta_j|)\,\mathrm{sgn}(\beta_j) + \mathrm{O_P}(n^{-1/2}/\lambda_n)\}.$$

Since $\liminf_{\beta\to 0^+} \lambda_n^{-1} p'_{\lambda_n}(\beta) = 1$ and $n^{-1/2}/\lambda_n \to 0$ as $n \to \infty$, the sign of the derivative is completely determined by that of β_j. Hence, (4.4) and (4.5) follows. This completes the proof. □

Proof of Theorem 1. . To show Part (a), it suffices to show that there exists a large constant C such that

$$P\left\{\inf_{\|\mathbf{u}\|=C} Q(\boldsymbol{\beta}_0 + n^{-1/2}\mathbf{u}) > Q(\boldsymbol{\beta}_0)\right\} \geqslant 1 - \varepsilon. \tag{4.6}$$

This implies that with probability tending to one there exists a local minimum in the ball $\{\boldsymbol{\beta}_0 + n^{-1/2}\mathbf{u} : \|\mathbf{u}\| \leqslant C\}$. Hence, there exists a local minimizer $\hat{\boldsymbol{\beta}}$ such that $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\| = \mathrm{O_P}(n^{-1/2})$.

Since

$$\begin{aligned} &Q(\boldsymbol{\beta}_0 + n^{-1/2}\mathbf{u}) - Q(\boldsymbol{\beta}_0) \\ &\quad = \left[Q\left\{\begin{pmatrix} \boldsymbol{\beta}_{10} + n^{-1/2}\mathbf{u}_1 \\ \boldsymbol{\beta}_{20} + n^{-1/2}\mathbf{u}_2 \end{pmatrix}\right\} - Q\left\{\begin{pmatrix} \boldsymbol{\beta}_{10} + n^{-1/2}\mathbf{u}_1 \\ \boldsymbol{\beta}_{20} \end{pmatrix}\right\}\right] + \left[Q\left\{\begin{pmatrix} \boldsymbol{\beta}_{10} + n^{-1/2}\mathbf{u}_1 \\ \boldsymbol{\beta}_{20} \end{pmatrix}\right\} - Q(\boldsymbol{\beta}_0)\right]. \end{aligned}$$

By Lemma 2, with probability tending to 1 the first term is positive as $\boldsymbol{\beta}_{20} = 0$. By Lemma 1, (4.6) holds. This completes the proof of Part (a).

Now we show Part (b). It follows by Lemma 2 that with probability tending to one, $\hat{\boldsymbol{\beta}}_2 = \mathbf{0}$.

It can be easily shown that there exists a $\hat{\boldsymbol{\beta}}_1$ in Part (a) that is a root n consistent local minimizer of

$$Q\left\{\begin{pmatrix} \boldsymbol{\beta}_1 \\ \mathbf{0} \end{pmatrix}\right\},$$

regarded as a function of $\boldsymbol{\beta}_1$, and satisfying the following equations:

$$\left.\frac{\partial Q(\boldsymbol{\beta})}{\partial \beta_j}\right|_{\boldsymbol{\beta}=\begin{pmatrix} \hat{\boldsymbol{\beta}}_1 \\ \mathbf{0} \end{pmatrix}} = 0 \quad \text{for } j = 1, \ldots, s.$$

Note that $\hat{\boldsymbol{\beta}}_1$ is a consistent estimator and $\boldsymbol{\beta}_{20} = 0$,

$$\begin{aligned} \frac{\partial Q(\hat{\boldsymbol{\beta}})}{\partial \beta_j} &= -\frac{1}{n}\mathbf{x}_{(j)}^{\mathrm{T}}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}_0) + \frac{1}{n}\mathbf{x}_{(j)}^{\mathrm{T}}\mathbf{X}_{(1)}(\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{10}) - p'_{\lambda_n}(|\hat{\beta}_j|) \\ &= -\frac{1}{n}\mathbf{x}_{(j)}^{\mathrm{T}}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}_0) + \mathbf{V}_{(j)}^{\mathrm{T}}(\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{10})(1 + \mathrm{o_P}(1)) \\ &\quad -(p'_{\lambda_n}(|\beta_j^0|)\,\mathrm{sgn}(\beta_{j0}) + \{p''_{\lambda_n}(|\beta_{j0}|) + \mathrm{o_P}(1)\}(\hat{\beta}_j - \beta_{j0})), \end{aligned}$$

where $X_{(1)}$ consists of the first s columns of $\mathbf{X}$ and $\mathbf{V}_{(j)}$ is the jth-column of V_{11}. It follows by Slutsky's Theorem and the Hájek–Šidák Central Limit Theorem (CLT) that

$$\sqrt{n}(\mathbf{V}_{11}+\Sigma)\{\hat{\boldsymbol{\beta}}_1-\boldsymbol{\beta}_{10}+(\mathbf{V}_{11}+\Sigma)^{-1}\mathbf{b}\}\to \mathrm{N}(\mathbf{0},\sigma^2\mathbf{V}_{11})$$

in distribution, where

$$\Sigma=\mathrm{diag}\{p''_{\lambda_n}(|\beta_{10}|),\ldots,p''_{\lambda_n}(|\beta_{s0}|)\},$$

$$\mathbf{b}=(p'_{\lambda_n}(|\beta_{10}|)\,\mathrm{sgn}(\beta_{10}),\ldots,p'_{\lambda_n}(|\beta_{s0}|)\,\mathrm{sgn}(\beta_{s0}))^{\mathrm{T}}.$$

Note that for the SCAD penalty, $\Sigma=0$ and $\mathbf{b}=\mathbf{0}$ as $\lambda_n\to 0$. Thus,

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_1-\boldsymbol{\beta}_{10})\to \mathrm{N}(\mathbf{0},\sigma^2\mathbf{V}_{11}^{-1})$$

in distribution. This completes the proof of Part (b). □

5. Conclusion

While the construction of SSDs has been paid increasing attention, the analysis of SSDs deserves special attention. In this paper, we proposed a variable selection approach for analyzing experiments when the full model contains many potential candidate effects. It has been shown that the proposed approach possesses an oracle property. The proposed approach has been empirically tested. A real data example was used to illustrate the effectiveness of the proposed approach.

Acknowledgements

The authors thank the referee and Dr. H. McGrath for their valuable comments and suggestions which led to a great improvement in the presentation of this paper. Li's research was supported by an NSF Grant DMS-0102505.

References

Craven, P., Wahba, G., 1979. Smoothing noisy data with spline functions: estimating the correct degree of smoothing by the method of generalized cross-validation Numer. Math. 31, 377–403.

Fan, J., 1997. Comments on "Wavelets in statistics: a review" by A. Antoniadis. J. Italian Statist. Assoc. 6, 131–138.

Fan, J., Li, R., 2001. Variable selection via nonconcave penalized likelihood and its oracle properties. J. Amer. Statist. Assos. 96, 1348–1360.

Fang, K.T., Lin, D.K.J., Ma, C.X., 2000. On the construction of multi-level supersaturated designs. J. Statist. Plan. Inference 86, 239–252.

Frank, I.E., Friedman, J.H., 1993. A statistical view of some chemometrics regression tools. Technometrics 35, 109–148.

Fu, W.J., 1998. Penalized regression: the bridge versus the LASSO. J. Comput. Graphical Statist. 7, 397–416.

Knight, K., Fu, W., 2000. Asymptotic for Lasso-type estimators. Ann. Statist. 28, 1356–1378.

Lin, D.K.J., 1991. Systematic supersaturated designs. Working Paper, No. 264, Department of Statistics, University of Tennessee.

Lin, D.K.J., 1993. A new class of supersaturated designs. Technometrics 35, 28–31.

Lin, D.K.J., 1995. Generating system supersaturated designs. Technometrics 37, 213–225.

Lin, D.K.J., 2000. Supersaturated design: theory and application. In: Park, S.H., Vining, G.G. (Eds.), Statistical Process Monitoring and Optimization. Marcel Dekker, New York (Chapter 18).

Tibshirani, R.J., 1996. Regression shrinkage and selection via the lasso. J. Roy. Statist. Soc. B 58, 267–288.

Westfall, P.H., Young, S.S., Lin, D.K.J., 1998. Forward selection error control in analysis of supersaturated designs. Statist. Sinica 8, 101–117.

Williams, K.R., 1968. Designed experiments. Rubber Age 100, 65–71.