STAT 992 Paper Review: Sure Independence Screening in Generalized Linear Models with NP-Dimensionality

J.Fan and R.Song

Presenter: Jiwei Zhao
Department of Statistics
University of Wisconsin–Madison

April 9, 2010

Outline

Outline

Introduction

- high dimensional data analysis
  - $p \to p_n$
  - $p = O(n^a)$ for some $a > 0$
  - $\log p = O(n^a)$ for some $a > 0$, which is NP-dimensionality, or loosely ultra high dimensionality
- variable/model selection: approach 1 (penalized pseudo-likelihood)
  - bridge regression (1993), LASSO (1996), SCAD (2001), Dantzig selector (2007), and their variants.
  - theoretical studies on persistency, consistency and oracle properties.
  - limitations: may not perform well in ultra high dimensional setting due to the simultaneous challenges of computational expediency, statistical accuracy and algorithmic stability (Fan et al. 2009)

Introduction

- variable/model selection: approach 2
  - Fan and Lv (2008): SIS method
  - limitations: only restricts to the ordinary linear model, and technical arguments can not be easily extended
  - Huang et al. (2008): based on marginal bridge regression
  - similar limitations

  - SIS in GLMs: Fan et al. (2009), Fan and Song (2010)

  - Hall et al. (2009): used a different marginal utility, derived from an empirical likelihood point of view
  - Hall and Miller (2009): proposed a generalized correlation ranking, which allows nonlinear regression
  - Wang (2009): Forward Regression
  - Fan and his colleagues (2010+): nonparametric IS for additive model, penalized composite quasi-likelihood, SIS for Cox's proportional hazards model, ...

Introduction

- In this paper, rank the maximum marginal likelihood estimator (MMLE) or maximum marginal likelihood, for GLMs
- An important extension of SIS in Fan and Lv (2008)
- Advantages: a new framework, which does not depend on the normality assumption even in the linear model setting; can be applied to possibly other models, like Cox model; can easily be applied to the generalized correlation ranking and other rankings based on a group of marginal variables.
- The two methods, ranking MMLE or maximum marginal likelihood, are equivalent in terms of sure screening properties.

Outline

- assume the pdf of Y is

$$f_Y(y;\theta) = \exp\{y\theta - b(\theta) + c(y)\}$$

- only model the mean regression, do not consider the dispersion parameter
- the model to be considered is

$$E(Y|\boldsymbol{X} = \boldsymbol{x}) = b'(\theta(\boldsymbol{x})) = g^{-1}(\sum_{j=0}^{p_n} \beta_j x_j)$$

- focus on the canonical link function for simplicity of presentation
- assume $EX_j = 0, EX_j^2 = 1, j = 1, ..., p_n$.

Outline

- assume the true sparse model is $\mathcal{M}_\star = \{1 \le j \le p_n : \beta_j^\star \neq 0\}$, where $\boldsymbol{\beta}^\star = (\beta_0^\star, \beta_1^\star, ..., \beta_{p_n}^\star)$ denotes the true value, and $s_n = |\mathcal{M}_\star|$
- the maximum marginal likelihood estimator is defined as

$$\hat{\boldsymbol{\beta}}_j^M = (\hat{\beta}_{j,0}^M, \hat{\beta}_j^M) = argmin_{\beta_0,\beta_j} \mathbb{P}_n l(\beta_0 + \beta_j X_j, Y)$$

- similarly, define the population version of the minimizer of the componentwise regression

$$\boldsymbol{\beta}_j^M = (\beta_{j,0}^M, \beta_j^M) = argmin_{\beta_0,\beta_j} El(\beta_0 + \beta_j X_j, Y),$$

  where E denotes the expectation under the true model
- variables selected are $\hat{\mathcal{M}}_{\gamma_n} = \{1 \le j \le p_n : |\hat{\beta}_j^M| \ge \gamma_n\}$, where γ_n is a predefined threshold value

Outline

- consider i.i.d. data $\{\boldsymbol{X}_i, Y_i\}$
- a regression model for $\boldsymbol{X}$ and Y is assumed with quasi-likelihood function $-l(\boldsymbol{X}^T\boldsymbol{\beta}, Y)$
- define $\boldsymbol{\beta}_0 = argmin_{\boldsymbol{\beta}} El(\boldsymbol{X}^T\boldsymbol{\beta}, Y)$ to be the population parameter
- define $\hat{\boldsymbol{\beta}} = argmin_{\boldsymbol{\beta}} \mathbb{P}_n l(\boldsymbol{X}^T\boldsymbol{\beta}, Y)$, which is QMLE
- assume that $\boldsymbol{\beta}_0$ is an interior point of a sufficient large, compact and convex set $\boldsymbol{B} \in \boldsymbol{R}^q$

Regularity Conditions

A. The Fisher information

$$I(\boldsymbol{\beta}) = E\{[\frac{\partial}{\partial \boldsymbol{\beta}} l(\boldsymbol{X}^T \boldsymbol{\beta}, Y)][\frac{\partial}{\partial \boldsymbol{\beta}} l(\boldsymbol{X}^T \boldsymbol{\beta}, Y)]^T\}$$

is finite and positive definite at $\boldsymbol{\beta} = \boldsymbol{\beta}_0$. Moreover, $\|I(\boldsymbol{\beta})\|_{\boldsymbol{B}} = \sup_{\boldsymbol{\beta} \in \boldsymbol{B}, \|\boldsymbol{x}\|=1} \|I(\boldsymbol{\beta})^{1/2} \boldsymbol{x}\|$ exists.

Regularity Conditions

B. The function $l(\boldsymbol{x}^T\boldsymbol{\beta}, y)$ satisfies the Lipschitz property with positive constant k_n:

$$|l(\boldsymbol{x}^T\boldsymbol{\beta}, y) - l(\boldsymbol{x}^T\boldsymbol{\beta}', y)|I_n(\boldsymbol{x}, y) \le k_n|\boldsymbol{x}^T\boldsymbol{\beta} - \boldsymbol{x}^T\boldsymbol{\beta}'|I_n(\boldsymbol{x}, y),$$

for $\boldsymbol{\beta}, \boldsymbol{\beta}' \in \boldsymbol{B}$, where $I_n(\boldsymbol{x}, y) = I((\boldsymbol{x}, y) \in \Omega_n)$ with $\Omega_n = \{(\boldsymbol{x}, y) : \|\boldsymbol{x}\|_\infty \le K_n, |y| \le K_n^\star\}$, for some sufficiently large positive constants K_n and $K_n^\star$. In addition, there exists a sufficiently large constant C such that with $b_n = Ck_nV_n^{-1}(q/n)^{1/2}$ and V_n given in condition C,

$$\sup_{\boldsymbol{\beta}\in\boldsymbol{B}, \|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|\le b_n} |E[l(\boldsymbol{X}^T\boldsymbol{\beta}, Y) - l(\boldsymbol{X}^T\boldsymbol{\beta}_0, Y)](1 - I_n(\boldsymbol{X}, Y))| \le o(q/n).$$

where V_n is the constant given in condition C.

Regularity Conditions

C. The function $l(\boldsymbol{X}^T\beta, Y)$ is convex in β, satisfying

$$E[l(\boldsymbol{X}^T\beta, Y) - l(\boldsymbol{X}^T\beta_0, Y)] \geq V_n\|\beta - \beta_0\|^2,$$

for all $\|\beta - \beta_0\| \leq b_n$ and some positive constants V_n.

Theorem 1

Theorem

Theorem 1. Under conditions A-C, it holds that for any $t > 0$,

$$P(\sqrt{n}\|\hat{\beta} - \beta_0\| \geq 16k_n(1+t)/V_n) \leq \exp(-2t^2/K_n^2) + nP(\Omega_n^c).$$

Proof of Theorem 1

Lemma

Lemma 2. (Symmetrization,Lemma 2.3.1,van der Vaart and Wellner,1996) Let $Z_1, \dots, Z_n$ be independent random variables with values in $\mathcal{Z}$ and $\mathcal{F}$ is a class of real valued functions on $\mathcal{Z}$. Then

$$E\{\sup_{f\in\mathcal{F}} |(\mathbb{P}_n - P)f(Z)|\} \leq 2E\{\sup_{f\in\mathcal{F}} |\mathbb{P}_n \varepsilon f(Z)|\},$$

where $\varepsilon_1, \dots, \varepsilon_n$ be a Rademacher sequence (i.e., i.i.d. sequence taking values ± 1 with probability $1/2$) independent of $Z_1, \dots, Z_n$ and $Pf(Z) = Ef(Z)$.

Proof of Theorem 1

Lemma

Lemma 3. (Contraction theorem,Ledoux and Talagrand,1991) Let $z_1, ..., z_n$ be nonrandom elements of some space $\mathcal{Z}$ and let $\mathcal{F}$ be a class of real valued functions on $\mathcal{Z}$. Let $\varepsilon_1, ..., \varepsilon_n$ be a Rademacher sequence. Consider Lipschitz functions $\gamma_i : \boldsymbol{R} \mapsto \boldsymbol{R}$, that is,

$$|\gamma_i(s) - \gamma_i(\tilde{s})| \leq |s - \tilde{s}|, \forall s, \tilde{s} \in \boldsymbol{R}.$$

Then for any function $\tilde{f} : \mathcal{Z} \mapsto \boldsymbol{R}$, we have

$$E\{\sup_{f \in \mathcal{F}} |\mathbb{P}_n \varepsilon(\gamma(f) - \gamma(\tilde{f}))|\} \leq 2E\{\sup_{f \in \mathcal{F}} |\mathbb{P}_n \varepsilon(f - \tilde{f})|\}.$$

Proof of Theorem 1

Lemma

Lemma 4. (Concentration theorem,Massart,2000) Let $Z_1, ..., Z_n$ *be independent random variables with values in some space* $\mathcal{Z}$ *and let* $\gamma \in \Gamma$*, a class of real valued functions on* $\mathcal{Z}$*. We assume that for some positive constants* $l_{i,\gamma}$ *and* $u_{i,\gamma}$*,* $l_{i,\gamma} \leq \gamma(Z_i) \leq u_{i,\gamma} \forall \gamma \in \Gamma$*. Define*

$$L^2 = \sup_{\gamma \in \Gamma} \sum_{i=1}^{n} (u_{i,\gamma} - l_{i,\gamma})^2 / n, \boldsymbol{Z} = \sup_{\gamma \in \Gamma} |(\mathbb{P}_n - P)\gamma(Z)|,$$

then for any $t > 0$*,*

$$P(\boldsymbol{Z} \geq E\boldsymbol{Z} + t) \leq \exp(-\frac{nt^2}{2L^2}).$$

Proof of Theorem 1

- Let $N > 0$, define $\mathcal{B}(N) = \{\beta \in \mathcal{B} : \|\beta - \beta_0\| \leq N\}$, and

$$\mathbb{G}_1(N) = \sup_{\beta \in \mathcal{B}(N)} |(\mathbb{P}_n - P)\{l(\boldsymbol{X}^T\beta, Y) - l(\boldsymbol{X}^T\beta_0, Y)\}I_n(\boldsymbol{X}, Y)|.$$

Lemma

Lemma 5. For all $t > 0$, it holds that

$$P(\mathbb{G}_1(N) \geq 4Nk_n(q/n)^{1/2}(1+t)) \leq \exp(-2t^2/K_n^2).$$

Proof of Theorem 1

- Proof of Lemma 5:
- On the set Ω_n,

$$\begin{aligned} |l(\boldsymbol{X}^T\beta, Y) - l(\boldsymbol{X}^T\beta_0, Y)| &\leq k_n|\boldsymbol{X}^T(\beta - \beta_0)| \\ &\leq k_n\|\boldsymbol{X}\|\|\beta - \beta_0\| \\ &\leq k_n \cdot q^{1/2}K_n \cdot N, \end{aligned}$$

  by Lipschitz and Cauchy-Schwartz respectively. Hence, $L^2 = 4k_n^2 qK_n^2N^2$.
- On the other hand,

Proof of Theorem 1

- $$\begin{aligned} E\mathbb{G}_1(N) &\leq 2E[\sup_{\boldsymbol{\beta}\in\mathcal{B}(N)} |\mathbb{P}_n\varepsilon\{l(\boldsymbol{X}^T\boldsymbol{\beta}, Y) - l(\boldsymbol{X}^T\boldsymbol{\beta}_0, Y)\}I_n(\boldsymbol{X}, Y)|] \\ &\leq 4k_nE[\sup_{\boldsymbol{\beta}\in\mathcal{B}(N)} |\mathbb{P}_n\varepsilon\boldsymbol{X}^T(\boldsymbol{\beta} - \boldsymbol{\beta}_0)I_n(\boldsymbol{X}, Y)|] \\ &\leq 4k_nE\|\mathbb{P}_n\varepsilon\boldsymbol{X}I_n(\boldsymbol{X}, Y)\| \sup_{\boldsymbol{\beta}\in\mathcal{B}(N)} \|\boldsymbol{\beta} - \boldsymbol{\beta}_0\| \\ &\leq 4k_nN(E\|\mathbb{P}_n\varepsilon\boldsymbol{X}I_n(\boldsymbol{X}, Y)\|^2)^{1/2} \\ &= 4k_nN(E\|\boldsymbol{X}\|^2I_n(\boldsymbol{X}, Y)/n)^{1/2} \\ &\leq 4k_nN(E\|\boldsymbol{X}\|^2/n)^{1/2} = 4k_nN(q/n)^{1/2}. \end{aligned}$$

- Then, use lemma 4,

$$P(\mathbb{G}_1(N) \geq 4Nk_n(q/n)^{1/2}(1 + t)) \leq \exp -2t^2/K_n^2.$$

Proof of Theorem 1

- Proof of Theorem 1. (See Appendix)

Outline

- First, for the sure screening purpose, if a variable X_j is jointly important ($\beta_j^\star \neq 0$), will it still be marginally important ($\beta_j^M \neq 0$)?
- Second, for the model selection consistency purpose, if a variable X_j is jointly unimportant ($\beta_j^\star = 0$), will it still be marginally unimportant ($\beta_j^M = 0$)?

Theorem 2

Theorem

Theorem 2. For $j = 1, ..., p_n$, the marginal regression parameters $\beta_j^M = 0$ iff $cov(b'(\boldsymbol{X}^T \boldsymbol{\beta}^\star), X_j) = 0$.

Corollary

Corollary 1. If the partial orthogonality condition holds, i.e., $\{X_j, j \notin \mathcal{M}_\star\}$ is independent of $\{X_i, i \in \mathcal{M}_\star\}$, then $\beta_j^M = 0$, for $j \notin \mathcal{M}_\star$.

Proof of Theorem 2

- Proof of Theorem 2. (See Appendix)

Theorem 3

Theorem

Theorem 3. If $|cov(b'(\mathbf{X}^T\beta^\star), X_j)| \geq c_1 n^{-\kappa}$ *for* $j \in \mathcal{M}_\star$ *and a positive constant* $c_1 > 0$, *then there exists a positive constant* c_2 *such that* $\min_{j\in\mathcal{M}_\star}|\beta_j^M| \geq c_2 n^{-\kappa}$, *provided that* $b''(.)$ *is bounded or*

$$EG(a|X_j|)|X_j|I(|X_j| \geq n^\eta) \leq dn^{-\kappa}, \text{for some } 0 < \eta < \kappa,$$

and some sufficiently small positive constants a *and* d, *where* $G(|x|) = \sup_{|u|\leq|x|}|b'(u)|$.

Proof of Theorem 3

- Proof of Theorem 3. (See Appendix)

- Note that for the normal and Bernoulli distributions, $b''(.)$ is bounded, whereas for the Poisson distribution, $G(|x|) = \exp(|x|)$ and Theorem 3 requires the tails of X_j to be light.
- In the proof of Theorem 5, it can be shown that

$$\sum_{j=1}^{p_n} |\beta_j^M|^2 = O(\|\Sigma\beta^\star\|^2) = O(\lambda_{\max}(\Sigma)),$$

  which means there can not be too many variables that have marginal coefficient $|\beta_j^M|$ that exceeds certain thresholding level. That achieves the sparsity in final selected model.

Gaussian Covariates

- Proposition 1. Suppose that Z and X are jointly normal with mean zero and standard deviation 1. For a strictly monotonic function f, $cov(X, Z) = 0$ iff $cov(X, f(Z)) = 0$, provided the latter covariance exists. In addition,

$$|cov(X, f(Z))| \geq |\rho| \inf_{|x| \leq c|\rho|} |g'(x)| EX^2 I(|X| \leq c),$$

  for any $c > 0$, where $\rho = EXZ$, $g(x) = Ef(x + \varepsilon)$ with $\varepsilon \sim N(0, 1 - \rho^2)$.
- Using Proposition 1, for Gaussian covariates, $\beta_j^M = 0$ iff $cov(\boldsymbol{X}^T \beta^\star, X_j) = 0$. Also, condition for Theorem 3 is that $|cov(\boldsymbol{X}^T \beta^\star, X_j)| \geq c_1 n^{-\kappa}$, which is a minimum condition required even for the least squares model (Fan and Lv, 2008).

Regularity Conditions

- A'. The marginal Fisher information: $I_j(\beta_j) = E\{b''(\boldsymbol{X}_j^T\beta_j)\boldsymbol{X}_j\boldsymbol{X}_j^T\}$ is finite and positive definite at $\beta_j = \beta_j^M$, for $j = 1, ..., p_n$. Moreover, $\|I_j(\beta_j)\|_{\mathcal{B}}$ is bounded from above.
- B'. The second derivative of $b(\theta)$ is continuous and positive. There exists an $\varepsilon_1 > 0$ such that for all $j = 1, ..., p_n$,

$$\sup_{\boldsymbol{\beta}\in\mathcal{B}, \|\boldsymbol{\beta}-\boldsymbol{\beta}_j^M\|\le\varepsilon_1} |Eb(\boldsymbol{X}_j^T\boldsymbol{\beta})I(|X_j| > K_n)| \le o(n^{-1}).$$

- C'. For all $\beta_j \in \mathcal{B}$, we have $E[l(\boldsymbol{X}_j^T\beta_j, Y) - l(\boldsymbol{X}_j^T\beta_j^M, Y)] \ge V\|\beta_j - \beta_j^M\|^2$, for some positive V, bounded from below uniformly over $j = 1, ..., p_n$.

Regularity Conditions

- D. There exists some positive constants m_0, m_1, s_0, s_1 and α, such that for sufficiently large t,

$$P(|X_j| > t) \leq (m_1 - s_1)\exp\{-m_0 t^{\alpha}\}, j = 1, ..., p_n,$$

and that

$$E\exp(b(\boldsymbol{X}^T\beta^\star + s_0) - b(\boldsymbol{X}^T\beta^\star)) + E\exp(b(\boldsymbol{X}^T\beta^\star - s_0) - b(\boldsymbol{X}^T\beta^\star)) \leq s_1.$$

- E. The conditions in Theorem 3 hold.

- Conditions A'-C' are satisfied in a lot of examples of GLMs, such as linear regression, logistic regression and Poisson regression. Note that the second part of condition D ensures the tail of the response variable Y to be exponentially light, as shown in the following.

Lemma

Lemma 1. If condition D holds, for any $t > 0$,

$$P(|Y| \geq m_0 t^{\alpha}/s_0) \leq s_1 \exp(-m_0 t^{\alpha}).$$

Proof of Lemma 1

- Proof of Lemma 1. (See Appendix)

Theorem 4

Theorem

Theorem 4. Suppose that conditions A', B', C' and D hold.
(1) If $n^{1-2\kappa}/(k_n^2 K_n^2) \to \infty$*, then for any* $c_3 > 0$*, there exists a positive constant* c_4 *such that*

$$\begin{aligned} & P(\max_{1 \le j \le p_n} |\hat{\beta}_j^M - \beta_j^M| \ge c_3 n^{-\kappa}) \\ \le \quad & p_n\{\exp(-c_4 n^{1-2\kappa}/(k_n^2 K_n^2)) + n m_1 \exp(-m_0 K_n^\alpha)\}. \end{aligned}$$

(2) If, in addition, condition E holds, then by taking $\gamma_n = c_5 n^{-\kappa}$ *with* $c_5 \le c_2/2$*, we have*

$$\begin{aligned} & P(\mathcal{M}_\star \subset \hat{\mathcal{M}}_{\gamma_n}) \\ \ge \quad & 1 - s_n\{\exp(-c_4 n^{1-2\kappa}/(k_n^2 K_n^2)) + n m_1 \exp(-m_0 K_n^\alpha)\}. \end{aligned}$$

Proof of Theorem 4

- Proof of Theorem 4. (See Appendix)

- No conditions on covariance matrix!
- If we assume that $\min_{j\in\mathcal{M}_\star} |cov(b'(\boldsymbol{X}^T\boldsymbol{\beta}^\star), X_j)| \geq c_1 n^{-\kappa+\delta}$ for any $\delta > 0$, then one can take $\gamma_n = cn^{-\kappa+\delta/2}$ for any $c > 0$ in Theorem 4. This is essentially the thresholding used in Fan and Lv (2008).

Regularity Conditions

- F. The variance $var(\boldsymbol{X}^T\beta^\star)$ is bounded from above and below.
- G. Either $b''(.)$ is bounded or $\boldsymbol{X}_M = (X_1, ..., X_{p_n})^T$ follows an elliptically contoured distribution, i.e., $\boldsymbol{X}_M = \Sigma_1^{1/2} RU$, and $|Eb'(\boldsymbol{X}^T\beta^\star)(\boldsymbol{X}^T\beta^\star - \beta_0^\star)|$ is bounded, where U is uniformly distributed on the unit sphere in p-dimensional Euclidean space, independent of the nonnegative random variable R, and $\Sigma_1 = var(\boldsymbol{X}_M)$.

Theorem 5

Theorem

Theorem 5. Under conditions A', B', C', D, F and G, we have for any $\gamma_n = c_5 n^{-\kappa}$, there exists a c_4 such that

$$\begin{aligned} & P[|\hat{\mathcal{M}}_{\gamma_n}| \leq O\{n^{2\kappa}\lambda_{\max}(\Sigma)\}] \\ \geq \; & 1 - p_n\{\exp(-c_4 n^{1-2\kappa}/(k_n^2 K_n^2)) + n m_1 \exp(-m_0 K_n^{\alpha})\}. \end{aligned}$$

Proof of Theorem 5

- Proof of Theorem 5. (See Appendix)

Outline

- The likelihood ratio screening is equivalent to the MMLE screening in the sense that they both possess the sure screening property and that the number of selected variables of the two methods are of the same order of magnitude.
- Marginal utility: letting $\hat{L}_0 = \min_{\beta_0} \mathbb{P}_n l(Y, \beta_0)$, define $\hat{L}_j = \hat{L}_0 - \min_{\beta_0,\beta_j} \mathbb{P}_n l(Y_i, \beta_0 + X_j \beta_j)$.
- Feature ranking: select features with largest marginal utilities: $\hat{\mathcal{M}}_{\nu_n} = \{1 \leq j \leq p_n : \hat{L}_j \geq \nu_n\}$.
- The main results are Theorems 6-9, which are analogous to Theorem 2-5.

Outline

- Consider 3 settings of how to generate the covariates in logistic regressions and linear models
- Compare two SIS procedures with LASSO, and SCAD
- The minimum model size is used as a measure of the effectiveness of a screening method
- The initial intension is to demonstrate that the simple SIS does not perform much worse than the far more complicated procedures like the LASSO and the SCAD
- The SIS can even outperform those more complicated methods in terms of variable screening

Outline

- Any surrogates screening, as long as which can preserve the non-sparsity structure of the true model and is feasible in computation, can be a good option for population variable screening.
- The proposed procedure does not cover all GLMs, such as some non-canonical link cases.
- The main idea of the technical proofs is broadly applicable.
- Another important extension is to generalize the concept of marginal regression to the marginal group regression, where the number of covariates m in each marginal regression is greater than or equal to one. This leads to a new procedure called group variables screening.
- How to choose the tuning parameter γ_n is another interesting and important problem.

THANK YOU!!