Sure Independence Screening for Ultra-High Dimensional Feature Space

Rong Wei (Carter) Man, Chen Yan Michael Chou

STA315
University of Toronto Mississauga

April 2, 2019

Overview

Background Information

- Variable selection plays an important role in high dimensional statistical modeling
- Computational cost and estimation accuracy are top two concerns
- The results from existing methods are challenged when dimensionality is ultra high
- Frequent in genomics and finance

Insight on Ultra High Dimensionality

- 4 facts for real difficulty when p$>>$n:
  - Design matrix X is rectangular
  - Population covariance matrix Σ may become ill-conditioned as n grows
  - Minimum non-zero absolute coefficient $|\beta_i|$ may decay with n to noise level
  - Distribution of z may have heavy tails

Goal of Paper

Reduce dimensionality p from a large or huge scale (say, $exp(O(n^{\xi}))$ for some $\xi > 0$) to a relatively large scale d (e.g., $o(n)$) by a fast and efficient method.

Existing Methods

- SCAD
  - coincides with lasso until $|x| = \lambda$, then smoothly transitions to a quadratic penalization
  - retains the penalization rate (and bias) of the LASSO for small coefficients, but continuously relaxes the rate of penalization as the absolute value of the coefficient increases
- Adaptive Lasso (AdaLasso)
  - use weighted penalty approach (bias of estimate comes from λ)
  - choose weights such that variable with large coefficients have smaller weights
- Lasso

Existing Methods

Dantzig Selector (DS)

Dantzig Selector (DS)

The Dantzig selector was proposed in Candes and Tao (2007) to recover a sparse high dimensional parameter vector in the linear model. It is the solution $\hat{\beta_{DS}}$ to the following ℓ_1-regularization problem

$$\min_{\zeta \in R^d} ||\zeta||_1 \text{ subject to } ||(X_{\mathcal{M}})^T r||_\infty \leq \lambda_d \sigma,$$

The authors pointed out that the above convex optimization problem can be easily recast as a linear program:

$$\min \sum_{i=1}^{d} u_i \text{ subject to } -u \leq \zeta \leq u \text{ and } -\lambda_d \sigma 1 \leq (X_{\mathcal{M}})^T (y - X_{\mathcal{M}} \zeta) \leq \lambda_d \sigma 1.$$

Dantzig Selector (DS)

- Well-developed lower dimensional technique that can be applied to estimate the d-vector β at d $<$ n
- Problems:
  - Computational cost high for huge scale problems
  - Log p grows large and not negligible
  - Large dimensionality causes Uniform Uncertainty Principle (UUP) to fail
  - No guarantee of model selection

Sure Independence Screening (SIS)

Let $\omega = (\omega_1, ..., \omega_p)^T$ be a p-vector obtained by the componentwise regression, that is,

$$\omega = X^T y,$$

where the $n \times p$ data matrix X is first standardized columnwise.
For any given $\gamma \in (0, 1)$, we sort the p componentwise magnitudes of the vector ω in a decreasing order and define a submodel

$$\mathcal{M}_\gamma = \{1 \le i \le p : |\omega_i| \text{ is among the first } [\gamma n] \text{ largest of all}\},$$

where $[\gamma n]$ denotes the integer part of γn. This is a straightforward way to shrink the full model $\{1, ..., p\}$ down to a submodel $\mathcal{M}_\gamma$ with size $d = [\gamma n] < n$.

Sure Independence Screening (SIS)

New Definition:
SIS ranks all the p features using the marginal utilities based on the marginal correlations $\hat{corr}(x_j, y)$ of x_j's with the response y and retains the top d covariates with the largest absolute correlations collected in the set $\hat{\mathcal{M}}$; that is,

$$\hat{\mathcal{M}} = \{1 \leq j \leq p : |\hat{corr}(x_j, y)| \text{ is among the top } d \text{ largest ones}\},$$

where $\hat{corr}(x_j, y)$ denotes the sample correlation. This achieves the goal of variable screening.

Sure Independence Screening (SIS)

Given $\mathcal{M}_*$ the true model and $\mathcal{M}_\gamma$ the model selected by SIS:

Theorem

$P(\mathcal{M}_* \subset \mathcal{M}_\gamma) \to 1$ as $n \to \infty$

Sure Independence Screening (SIS)

Why pick SIS over other existing methods?

- Narrows down the search for important predictors, speeds up Dantzig selector
- Reduces logarithmic factor: $Log(p) \rightarrow Log(d) < log(n)$
- Oracle Property: Selecting right model; estimating parameters efficiently

Simulation Procedure

- Apply SIS to reduce dimensionality from p to large scale d ($d < n$)
- Use lower dimensional mode selection method (SCAD, DS, AdaLasso)

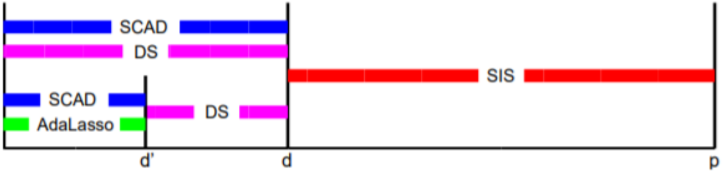

Figure 2: Methods of model selection with ultra high dimensionality.

Simulation I

(n, p, s) = (200, 1000, 8) and (800, 20000, 18)

Table 1: Results of simulation I

	Medians of the selected model sizes (upper entry) and the estimation errors (lower entry)					
p	**DS**	**Lasso**	**SIS-SCAD**	**SIS-DS**	**SIS-DS-SCAD**	**SIS-DS-AdaLasso**
1000	10^3	62.5	15	37	27	34
	1.381	0.895	0.374	0.795	0.614	1.269
20000	—	—	37	119	60.5	99
	—	—	0.288	0.732	0.372	1.014

Simulation II

$(n, p, s) = (200, 1000, 5), (200, 1000, 8)$ and $(800, 20000, 14)$

Table 2: Results of simulation II

	Medians of the selected model sizes (upper entry) and the estimation errors (lower entry)					
p	**DS**	**Lasso**	**SIS-SCAD**	**SIS-DS**	**SIS-DS-SCAD**	**SIS-DS-AdaLasso**
1000	10^3	91	21	56	27	52
$(s = 5)$	1.256	1.257	0.331	0.727	0.476	1.204
	10^3	74	18	56	31.5	51
$(s = 8)$	1.465	1.257	0.458	1.014	0.787	1.824
20000	—	—	36	119	54	86
	—	—	0.367	0.986	0.743	1.762

Real Data Analysis

Table 3: Classification errors on the Leukemia data set

Method	Training error	Test error	Number of genes
SIS-SCAD-LD	0/38	1/34	16
SIS-SCAD-NB	4/38	1/34	16
Nearest shrunken centroids	1/38	2/34	21

Problems of SIS

- Unimportant predictors correlated with important predictors can have higher priority to be selected
- Important predictors that are marginally uncorrelated but jointly correlated with response wont be picked
  - $cov(x_i, x_j)$ is low, but $cov(x_i, x_j, y)$ is high
- Issue of collinearity

Iterative Sure Independence Screening (ISIS)

- Select subset of k_1 variables $A_1 = \left\{X_{i_1}, X_{i_2}, ..., X_{i_{k_1}}\right\}$
- Use n-vector of residuals as new responses and reapply SIS to remaining $p - k_1$ variables $A_2 = \left\{X_{j_1}, X_{j_2}, ..., X_{j_{k_2}}\right\}$
- Weaken priority of unimportant variables
- Variables missed in first screening will survive
- Stop until we get ℓ disjoint subsets of A_1, ..., A_ℓ whose union $A = \cup_{i=1}^{\ell} A_i$ has a size d, which is less than n

Simulation III

Table 7: Simulations I and II in Section 3.3 revisited: Medians of the selected model sizes (upper entry) and the estimation errors (lower entry)

	Simulation I		Simulation II
p	ISIS-SCAD		ISIS-SCAD
1000	13	$(s = 5)$	11
	0.329		0.223
		$(s = 8)$	13.5
			0.366
20000	31		27
	0.246		0.315

SIS-SCAD
21
0.331
18
0.458
36
0.367

(a) SIS Simulation I

SIS-SCAD
15
0.374
37
0.288

(b) SIS Simulation II

Conclusions from Paper

- Reduce dimensionality from high up
- Speed up variable selection and improve estimation accuracy
- Can be combined with lower dimensional techniques

Thank You for Listening!