Sure Independence Screening in Ultra High Dimensional Feature Space

---

University of Toronto Mississauga

Advanced Statistical Learning

Professor Dehan Kong
STA315
April 10, 2019

Chen Yan Michael Chou
Ron Wei (Carter) Man

Contents

1 Abstract

Variable selection in high dimensional feature space is one of the most pressing problems in modern statistical learning. It appears, nowadays, in many areas of genomics such as gene expression, biomedical imaging, and tumor classification where the number of variables or parameters p can be much larger than sample size n. With large scale or dimensionality, computational cost and estimation accuracy become the top concerns for any statistical procedure. With these problems in mind, Candes and Tao (2007) proposes a minimum ℓ_1 estimator, the Dantzig selector, and shows that it mimics the ideal risk within a factor of log p. This method, however, is challenged when the dimensionality is ultra high ($p >> n$), one level above the high dimensional setting which people are already concerned with. The factor log p can be large and the uniform uncertainty principle, required in order for the Dantzig Selector to successively select its variables, can fail.

Motivated by these downfalls, Fan and Lv (2008) [2] introduce the concept of sure screening and propose a sure screening method Sure Indepedence Screening (SIS) including a methodological extension, the iterate SIS (ISIS), to reduce dimensionality even in an exponentially growing dimension from high to a relatively large scale d that is below sample size. With high dimension reduced accurately to below sample size, variable selection can then be complemented by other distinguished lower dimensional methods such as SCAD, Dantzig selector, LASSO, or adaptive LASSO. Hence, in this paper, we aim to summarize the SIS concept proposed by our authors, reproduce the simulation results, and confirm the potency and validity of this statistical method.

2 Introduction

2.1 Background

Before we introduce SIS, we must first introduce some key concepts in statistical learning. Consider the problem of estaimating a p-vector of parameters β from the linear model

$$y = X\beta + \epsilon,$$

where $y = (Y_1, ..., Y_n)^T$ is an n-vector of responses, $X = (x_1, ..., x_n)^T$ is an $n \times p$ random design matrix with i.i.d. $x_1, ..., x_n$, $\beta = (\beta_1, ..., \beta_p)^T$ is a p-vector of parameters, and $\epsilon = (\epsilon_1, ..., \epsilon_n)^T$ is an n-vector of i.i.d. random errors. When dimensionality p is high, it is often assumed that only a small number of predictors among $X_1, ..., X_p$ contribute to the response, while results in assuming that the parameter vector β is sparse. Under a sparse setting, variable selection can improve estimation accuracy by effectively identifying the subset of important predictors, and also enhance the model interpretability with parsimonious (simplest model/theory with the least assumptions and variables but with greatest explanatory power) representation. This relates to the idea of dimensionality reduction.

2.2 Dimensionality Reduction

Dimension reduction, or feature selection, is an effective strategy to deal with high dimensionality. It can not only reduce high dimensions to low, but also reduce computational burden drastically. Inspired by this idea along with the concerns from the Dantzig selector, we introduce the main goal of Fan and Lv's paper:

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

This will be achieved by sure screening and the proposed sure screening method SIS. The authors define sure screening as a method which holds the property of selecting important variables after variable screening with probability tending to one. This will help reduce the computational cost when applying the Dantzig selector on the much smaller submodel once we dramatically narrow down the search for important predictors. In fact, this not only speeds up the Dantzig selector, but also reduces the logarithmic factor in mimicking the ideal risk from log p to log d, which is smaller than log n.

2.3 Insight on High Dimensionality

To further illustrate the demand and need for a method like SIS, the authors introduce some insight on challenges of high dimensionality in variable selection. Let us look at a situation where all the predictors $X_1, ..., X_p$ are standardized and the distribution of $z = \Sigma^{-1/2}x$ is spherically symmetric, where $x = (X_1, ..., X_p)^T$ and $\Sigma = cov(x)$. The real difficulty when dimension p is larger than sample size n can be boiled down to 5 facts [4]:

- The design matrix X is rectangular, have more columns than rows which results in the matrix X^TX becoming huge and singular.
- The maximum spurious correlation between a covariate and the response can be large and unimportant predictors can be highly correlated with the response due to the presence of important predictors associated with the predictor.
- The population covariance matrix Σ may become ill-conditioned as n grows.
- The minimum nonzero absolute coefficient $|\beta_i|$ may decay with n and get close to the noise level.
- The distribution of z may have heavy tails.

Therefore, in general, it is challenging to estimate the sparse parameter vector β accurately when $p >> n$.

3 Sure Independence Screening

By sure screening the authors mean a property that all the important variables survive after applying a variable screening procedure with probability tending to one. All desirable dimensionality reduction methods should hold the sure screening property.

Old Definition from Fan and Lv (2008):

- Center each input variable so that the observed mean is zero, and scale each predictor so that the sample standard deviation is one. Let $\mathcal{M}_* = \{1 \le i \le p : |\beta_i| \neq 0\}$ be the true sparse model with nonsparsity rule $s = |\mathcal{M}_*|$ (the number of nonzero coefficients). 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 $\gamma \in (0, 1)$, 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.

New Definition from Fan and Lv (2018) [3]:

- 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,

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

  where $\hat{corr}(x_j, y)$ denotes the sample correlation.

Both definitions offer a straightforward way to shrink the full model $\{1, ..., p\}$ down to a submodel $\mathcal{M}$ with size $d < n$ and achieves the goal of variable screening.To further specify the conditions under which the sure screening property holds for SIS:

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

for some given γ.

4 SIS Based Model Selection Techniques

As mentioned earlier, most modern statistical variable selection methods are unable to correctly select models and variables in high and ultra high dimensions. SIS fixes these problems by reducing the dimensions and upholding the sure screening property. Combined with other well-developed lower dimensional techniques, SIS can be applied to estimate the d-vector β at the scale of $d < n$ now given the smaller submodel $\mathcal{M}$:

$$y = X_{\mathcal{M}}\beta + \epsilon.$$

We will now briefly review effective methods that the authors used in combination with SIS to conduct the simulation.

4.1 Dantzig Selector

The Dantzig selector was proposed in Candes and Tao (2007) [1] 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:

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

In general, linear program are easier to solve and can be accomplished by maximizing or minimizing a linear function of several variables, such as output or cost. However, aforementioned, the Dantzig Selection holds 4 major weaknesses in ultra high dimensional settings:

- Computational cost high for huge scale problems such as implementing linear programs in high dimensions
- Factor log p can become large and may not be negligible when dimension p grows fast with sample size n
- As dimensionality grows, the notion of uniform uncertainty principle (UUP), a new idea on deterministic design matrices established by Candes and Tao (2007), may be hard to satisfy.
- There is no guarantee that the Dantzig selector picks up the right model though it has the oracle property in the sense of Donoho and Johnstone (1994)

4.2 Penalized Least-Squares and SCAD

Fan and Li (2001) [5] advocate for penalty functions with three properties: sparsity, unbiasedness, and continuity. It is well known that ℓ_p-penalty with $0 \leq p \leq 1$ does not satisfy the continuity condition, ℓ_p-penalty with $p > 1$ does not satisfy the sparsity condition, and ℓ_1-penalty (LASSO), though possesses the sparsity and continuity, generates estimation bias. This is why the smoothly clipped absolute deviation (SCAD) penalty was introduced by Fan (1997) which satisfies all three conditions simultaneously:

$$p'_\lambda(|\beta|) = \lambda \left\{ I(|\lambda| \leq \lambda) + \frac{(a\lambda - |\beta|)_+}{(a-1)\lambda} I(|\beta| > \lambda) \right\} \text{ for some a} > 2.$$

The SCAD penalty coincides with the LASSO penalty until $|x| = \lambda$, then smoothly transitions to a quadratic penalization. It 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 [6].

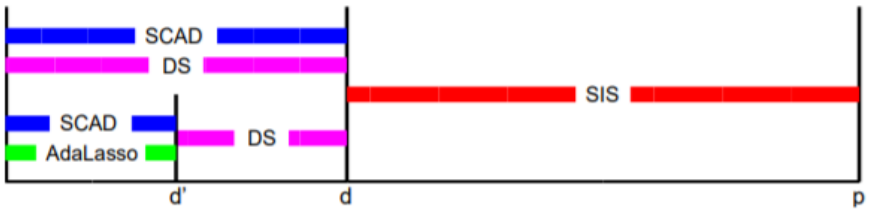

Figure 1: Methods of model selection with ultra high dimensionality

4.3 Adaptive LASSO

The LASSO, although easy to implement and widely used due to its complexity, comes with an intrinsic bias problem. To overcome this, Zou (2006) [8] proposes an adaptive LASSO which uses an adaptively weighted ℓ_1 penalty in the penalized least squares:

$$\hat{\beta^{*(n)}} = arg\min_{\beta} ||y - \sum_{j=1}^{p} x_j\beta_j||^2 + \lambda_n \sum_{j=1}^{p} \hat{w}_j|\beta_j|,$$

where $\lambda \geq 0$ is a regularization parameter and $w = (w_1, ...w_d)^T$ is a known weight vector.

5 Numerical Study

We will repeat and reconduct the following simulation that the authors attempted. For the problem of ultra-high dimensional variable selection, the authors first use SIS to reduce dimensionality from P to a relatively large scale d, say less than n and then use lower dimensional model selection methods that we mentioned earlier. In some situations, the authors further reduce the model size down to $d' < d$. We will call SIS followed by a specific method such as SCAD as SIS-SCAD and so on. Figure 1 shows a schematic diagram of all the possible method combinations.

5.1 Simulation Details

For the simulation we repeated the paper's first simulation using just "independent" features. We used the linear model with i.i.d. standard Gaussian predictors and Gaussian noise with standard deviation $\sigma = 1.5$. We used a sample size of $n = 200$, a total of $p = 1000$ predictors, and simulated 100 datasets. The size s of the true model, i.e., the numbers of nonzero coefficients, was chose to be 8 and the nonzero components of the p-vectors β were randomly chosen using a = 4log $n/\sqrt{n}$, and picked nonzero coefficients of the form $(-1)^u(a + |z|)$, where u was drawn from a Bernoulli distribution with parameter 0.4 and z was drawn from the standard Gaussian distribution.

5.2 Iterative SIS

Unlike the authors, we opted to simulate data only on LASSO and ISIS-SCAD to better compare results. To explain the reasoning behind using ISIS instead of SIS, we must first understand the potential 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 won't be picked
  - $cov(x_i, x_j)$ is low, but $cov(x_i, x_j, y)$ is high

```
ISIS.SCAD$ix
] 1 2 3 4 5 6 7 8
lasso.size
[1]  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
55]  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0 63
```

Figure 2: Selected Model Sizes for ISIS-SCAD vs. LASSO

ISIS uses the following algorithm:

- Select subset of k_1 variables $A_1 = \{X_{i_1}, X_{i_2}, ..., X_{i_{k_1}}\}$
- Use n-vector of residuals as new responses and reapply SIS to remaining $p - k_1$ variables $A_2 = \{X_{j_1}, X_{j_2}, ..., X_{j_{k_2}}\}$
- 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

This not only weakens the priority of unimportant variables, but also variables missed in the first screening will be able to survive.

5.3 Results

As you can see in figure 2, the LASSO (using lars algorithm) gives large models compared to ISIS-SCAD. The ISIS-SCAD was able to narrow down predictors to 8 (size of true model) while LASSO still gives a model of 63 predictors. This fully supports the authors simulation results as ISIS-SCAD produces the best results.

ISIS-SCAD	LASSO
8	63

6 Conclusion

SIS has been shown to be able to reduce dimensionality from high up to an exponential growth to a relatively large scale that is below sample size. It can not only speed up variable selection drastically but also improve the estimation accuracy when dimensionality is ultra-high. SIS combined with well-developed lower dimensional technique such as SCAD, Dantzig Selector, Lasso, or adaptive Lasso provides a powerful tool for high dimensional variable selection whether in regression or classification.

References

[1] Candes, E., & Tao, T. (2007). The Dantzig selector: Statistical estimation when p is much larger than n. *The annals of Statistics, 35*(6), 2313-2351.

[2] Fan, J., & Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 70*(5), 849-911.

[3] Fan, J., & Lv, J. (2018). Sure Independence Screening. *Wiley StatsRef: Statistics Reference Online*, 1-8. doi:10.1002/9781118445112.stat08043

[4] Fan, J., & Li, R. (2006). Statistical challenges with high dimensionality: Feature selection in knowledge discovery. *arXiv preprint math/0602133.*

[5] Fan, J., & Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association, 96*(456), 1348-1360.

[6] Kim, Y., Choi, H., & Oh, H. S. (2008). Smoothly clipped absolute deviation on high dimensions. *Journal of the American Statistical Association, 103*(484), 1665-1673.

[7] Saldana, D. F., & Feng, Y. (2018). SIS: An R package for sure independence screening in ultrahigh dimensional statistical models. *Journal of Statistical Software, 83*(2), 1-25.

[8] Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American statistical association, 101*(476), 1418-1429.