Variable selection in Ultra-high dimensional statistical problems

Jamshid Namdari

University of California Davis
RTG presentation series
Modern Tools in Statistics and Application

jamnamdari@ucdavis.edu

June 1, 2018

Overview

Introduction

- Development in technology and dependence of many scientific investigations has led to rapidly increasing volume of data.
- Characteristic of the data: high in dimension and sample size. Applications in genomics, health sciences, economics, finance, climatology, ...
- In this talk I will consider the case that dimension(p) can grow exponentially in the sample size(n) and I will present the methodologies proposed in Fan and Lv (2008) and Fan, Samworth, and We (2009).

Examples

- Disease classification using microarray gen expression.
  - # of arrays on the order of tens
  - # of gen expression profile on the order of tens of thousands.
- When interactions are considered for instance in portfolio allocation among two thousand stocks, the covariance matrix involves over two million parameters.
- Analysis of high resolution images.

Model

- Suppose $X_1, ..., X_n \in \mathbb{R}^p$ are i.i.d. predictors.
- $Y \in \mathbb{R}^n$ is the response.
- Consider linear model $Y = \mathbf{X}\beta + \epsilon$ where:
  $\mathbf{X} = [X_1, .., X_n]$, $\epsilon \in \mathbb{R}^n$ is an n-vector of i.i.d random error.
- Goal: estimate β.

Difficulties

- $\mathbf{X}^T\mathbf{X}$ is huge and singular.
- Maximum sample correlation between predictors can be large despite that predictors are independent.
- Unimportant predictor may be highly correlated with important predictors which usually increases with dimensionality.
- Unimportant predictors can be highly correlated with the response due to correlation with an important predictor.
- Population covariance matrix may be ill-conditioned as n grows.
- Minimum non-zero absolute coefficient $|\beta_j|$ may decay with n and fall close to the noise level.
- Noise accumulation in high dimensional prediction.

One way to overcome the problems

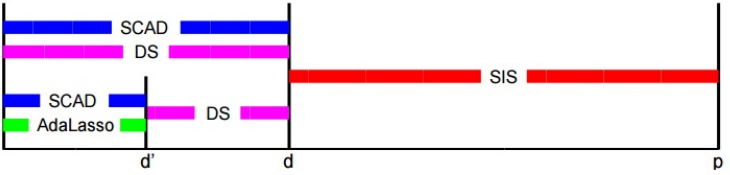
Assume the p-dimensional regression parameters are sparse(with many components being zero).
E.g. in Genomic studies in general it is believed that only a fraction of molecules are related to biological outcomes.
This suggests to use variable selection procedures such as: LASSO or SCAD, ... Or in other words, use penalized least squares with suitable choice of penalty, i.e.

$$\hat{\beta}_{PLS} = argmin_{\beta_0,\beta} n^{-1} \sum_{i=1}^{n} (Y_i - \beta_0 - X_i^T \beta)^2 + \sum_{j=1}^{p} p_\lambda(|\beta_j|).$$

A remedy for Ultra-high dimension

But commonly used variable selection procedures work when dimension is in the same order as sample size.

- idea:
  - First, reduce the dimensionality from ultra high ($log(p) = O(n^a)$ for some $a > 0$) to moderate scale ($d = n - 1$ or $d = \lfloor n/\log(n) \rfloor$).
  - Second, use a well developed variable selection technique.



Methods of model selection with ultra high dimensionality.

Source: Fan and Lv (2008)

Large Scale Screening

Independence Screening: Ranking features according to marginal utility. Each covariate is used independently as a predictor to decide its usefulness for predicting the response.

- Correlation ranking(Fan and Lv (2008))
- Two sample test(Story and Tibshirani(2003))
- Feature ranking using generalized correlation(Hall and Miller(2009))
- Using marginal bridge estimators(Huang, Horowitz and Ma(2008))
- Using tilting methods and empirical likelihood(Hall, Titterington and Xue(2009)).

Sure Independence Screening: All the important variables survive after applying variable screening procedure with probability tending to 1.

Correlation Ranking

- **Ranking features according to the magnitude of its sample correlation with the response variable**. I.e. Consider component wise regression coefficient $\hat{\beta} = (\hat{\beta}_1, \dots, \hat{\beta}_p)^T = \mathbf{X}^T Y$, where each column of $\mathbf{X}$ has been centered and standardized.
  Take the submodel to be :

  $$\mathbb{M}_d = \{1 \leq j \leq p : |\hat{\beta}_j| \text{ is among the first } d \text{ largest of all}\}$$

  $d \leq n$ e.g. $d = n - 1$ or $d = \frac{n}{log(n)}$.
- **Rank features according to the marginal loss**

  $$L_j = \min_{\beta_0, \beta_j} n^{-1} \sum_{i=1}^{n} (Y_i - \beta_0 - X_{ij}^T \beta_j)^2$$

  and choose features corresponding to the first d smallest of the marginal losses L_j.

Drawbacks

It is possible that:

- some unimportant predictors that are highly correlated with an important predictor can have higher priority for being selected.
- an important predictor that is marginally uncorrelated but jointly correlated with the response cannot be picked up.

In general collinearity between predictors add difficulty to variable selection.

Iterative correlation learning

To overcome those problems apply iteratively correlation learning as follow:

- Select a subset of k_1 variables $\mathcal{A}_1 = \{X_{i_1}, \ldots, X_{i_{k_1}}\}$ using an SIS-based model selection such as SIS-Lasso.
- Let r_1 be the residual after regressing Y on $\{X_{i_1}, \ldots, X_{i_{k_1}}\}$.
- Treat r_1 as the new responses and apply the same procedure to the remaining $p - k_1$ variables to obtain $\mathcal{A}_2 = \{X_{j_1}, \ldots, X_{j_{k_2}}\}$.
- Continue until $\mathcal{A} = \cup_{s=1}^{\ell} \mathcal{A}_s$ has size $d < n$, then apply a moderate scale method such as Lasso or SCAD.

Does it address our concerns?

- Weakens the priority of unimportant variables that are highly correlated with the response through $X_{i_1}, \ldots, X_{i_{k_1}}$. (Since the remaining covariates in each step have lower correlation with the residuals than with the original response.)
- Gives a chance to those important predictors that are missed in the previous step to be selected.

Beyond linear model

In the more general pseudo-likelihood framework we can perform screening by choosing covariates with lowest marginal loss

$$L_j = \min_{\beta_0,\beta_j} n^{-1} \sum_{i=1}^{n} L(Y_i, \beta_0 + X_{ij}^T \beta_j), \qquad j = 1, \ldots, p,$$

for suitable loss function L, followed by moderated scale variable selection procedure s.t. k_1 of them are retained.

- In logistic regression model

$$L(Y_i, \beta_0, X_{ij}\beta_j) = \sum_{i=1}^{n} \{log(1 + e^{\beta_0 + X_{ij}\beta_j}) - Y_i(\beta_0 + X_{ij}\beta_j)\}.$$

- In classification using support vector machine

$$L(Y_i, \beta_0 + X_{ij}\beta) = \{1 - Y_i(\beta_0 + X_{ij}\beta)\}_+.$$

Iterative feature selection

- Apply SIS followed by a penalized (pseudo)-likelihood method to select a subset $\widehat{\mathcal{M}}_1 \subset \{1, \dots, p\}$.
- order $\{L_j^{(2)} : j \in \widehat{\mathcal{M}}_1^c\}$ where

$$L_j^{(2)} = \min_{\beta_0, \beta_{\widehat{\mathcal{M}}_1}, \beta_j} n^{-1} L(Y_i, \beta_0 + X_{i,\widehat{\mathcal{M}}_1}^T \beta_{\widehat{\mathcal{M}}_1} + X_{ij}\beta_j),$$

  and add indexes of k_2 smallest of them to $\widehat{\mathcal{M}}_1$. Then apply a (pseudo)-likelihood method to select a subset $\widehat{\mathcal{M}}_2 \subset \{1, \dots, p\}$.
- Repeat until $\widehat{\mathcal{M}}_\ell$ either reaches a prescribed size d or satisfies $\widehat{\mathcal{M}}_\ell = \widehat{\mathcal{M}}_{\ell-1}$.

Reduction of False Selection Rate

- Let $A = \{j : \beta_j \neq 0\}$.
- Split the sample into two halves at random.
- Apply SIS or (I)SIS separately to the data in each partition to select two sets of active indices $\widehat{A}_1$ and $\widehat{A}_2$.
- **First Variant:** Consider $\widehat{A}_1 \cap \widehat{A}_2$ as an estimate of A.
- **Second Variant:** Recruit as many features into equal-sized sets of active indices $\tilde{A}_1$ and $\tilde{A}_2$ as are required to ensure that the intersection $\tilde{A}_1 \cap \tilde{A}_1$ has d elements.

Software

This two stage variable selection has been implemented in R by Jianqing Fan, Yang Feng, Diego Franco Saldana, Richard Samworth, Yichao Wu.
To install and call the library do the following:

- install.packages("SIS")
- library("SIS")

Then simply call the function SIS.
SIS(x, y, family="binomial", penalty="lasso", tune="bic")

Neuroblastoma data

- Study consists of 251 patients of the German Neuroblastoma Trials diagnosed between 1989 and 2004.
- At diagnosis, patients‘ ages range from 0 to 296 months with a median age of 15 months.
- Interested in predicting if each patient survived 3 years after the diagnosis of Neuroblastoma.
- Goal: develop a gen expression based classification rule for Neuroblastoma patients and obtain a view on which set of genes is responsible for Neuroblastoma.
- $p = 10707$ genes.
- 125 Randomly selected subjects in the training set and remaining subjects in the testing set.

Analysis (Fan, Samworth, and We (2009))

- For the initial screening 50 variables are selected for (I)SIS methods.
- From table below we can see that (I)SIS methods outperforms LASSO in the sense that they used fewer predictors while giving smaller or equal testing error.

Method	SIS	(I)SIS	var2-SIS	var2-(I)SIS	LASSO
No. of predictors	5	23	10	12	57
Testing error	19/114	22/114	22/114	21/114	22/114

In var2-(I)SIS, data set is randomly partitioned into two groups. The (I)SIS is applied to both and equal sized sets of active indices $\mathcal{A}^{(1)}$ and $\mathcal{A}^{(2)}$ are selected to ensure that $\mathcal{A}^{(1)} \cap \mathcal{A}^{(2)}$ has a pre-specified number of elements.

References

[1] Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. Journal of Royal Statistical Society, B, Vol. 70, 849–911.

[2] Fan, J., Samworth, R., and We, Y. (2009). Ultrahigh Dimensional Feature Selection: Beyond The Linear Model. Journal of Machine Learning Research, Vol. 10, 2013–2038.

[3] Hall, P., Titterington, D. M., and Xue, J.(2009) Tiling methods for assessing the influence of components in a classifier. J. Roy. Statist. Soc., Ser. B, Vol. 71, 783–803.

[4] Hastie, T., Tibshirani, R., and Friedman, J. The elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer, New York, 2001.

[5] Storey, J. D. and Tibshirani R. (2003). Statistical significance for genome-wide studies. Proc. Natl. Aca. Sci. 100, 94409445.

[6] Huang, J., Horowitz, J. and Ma, S. (2007). Asymptotic properties of bridge estimators in sparse high-dimensional regression models. Ann. Statist.

Thank You!