High-dimensional Variable Selection

Larry Wasserman, Kathryn Roeder

Xinxin Yu

Department of Statistics,
University of Wisconsin Madison

May 4, 2010

Outline

1. Introduction
   - Goal of the Paper
   - Notation and Assumptions
2. Error Control
3. Loss and Cross-Validation
4. Multi-Stage Methods
   - The Lasso
   - Stepwise Regression
   - Marginal Regression
   - Modifications

Outline

1. Introduction
   - Goal of the Paper
   - Notation and Assumptions
2. Error Control
3. Loss and Cross-Validation
4. Multi-Stage Methods
   - The Lasso
   - Stepwise Regression
   - Marginal Regression
   - Modifications

Goals of high dimensional problems

1. Find models with good prediction error.
2. Estimate the true ‘sparsity pattern’, the set of covariates with nonzero regression coefficients.

This paper will deal with the second goal and builds on ideas in Meinshausen and Yu (2008) and Meinshausen (2007).

Regression Model

Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be iid observations from the regression model

$$Y_i = X_i^T \beta + \epsilon_i$$

where $\epsilon_i \sim N(0, \sigma^2)$, $X_i = (X_{i1}, \ldots, X_{ip})^T \in \mathbb{R}^p$ and $p = p_n > n$.Let

$$D = \{j : \beta_j \neq 0\}$$

be the set of the covariates with nonzero regression coefficients. Assume that $|D| = s$. A variable selection procedure $\hat{D}_n$ maps the data into subsets of $\{1, \ldots, p\}$.

Goal of the paper

The main goal of this paper is to derive a procedure $\hat{D}_n$ such that

$$\limsup_{n\to\infty} \mathbb{P}(\hat{D}_n \subset D) \geq 1 - \alpha,$$

that is, the asymptotic type I error is no more than α.
The procedure involves three stages

1. Fit a suite of candidate models, each model depending on a tuning parameter λ

$$\mathcal{S} = \{\hat{S}_n(\lambda) : \lambda \in \Lambda\}.$$

2. Select one of those models $\hat{S}_n(\lambda)$ using cross-validation to select $\hat{\lambda}$.
3. Eliminate some variables by hypothesis tesing, to get $\hat{D}_n$.

Methods in Stage I

1. LASSO
$$\hat{S}_n(\lambda) = \{j : \tilde{\beta}_j(\lambda) \neq 0\},$$
where $\tilde{\beta}_j(\lambda)$ is the lasso estimator, the value of β that minimizes
$$\sum_{i=1}^{n}(Y_i - X_i^T\beta)^2 + \lambda \sum_{j=1}^{p}|\beta_j|.$$
2. Take $\hat{S}_n(\lambda)$ to be the set of variables chosen by forward stepwise regression after λ steps.
3. Marginal regression taking
$$\hat{S}_n = \{j : |\hat{\mu}_j| > \lambda\}$$
where $\hat{\mu}_j$ is the marginal regression coefficient from regressing Y on $X_{.j}$.

Outline

1 Introduction
- Goal of the Paper
- Notation and Assumptions

2 Error Control

3 Loss and Cross-Validation

4 Multi-Stage Methods
- The Lasso
- Stepwise Regression
- Marginal Regression
- Modifications

Notation

- Define the loss of any estimator $\hat{\beta}$ by

  $$L(\hat{\beta}) = \frac{1}{n}(\hat{\beta} - \beta)^T X^T X(\hat{\beta} - \beta) = (\hat{\beta} - \beta)^T \hat{\Sigma}_n (\hat{\beta} - \beta)$$

  where $\hat{\Sigma}_n = n^{-1} X^T X$. For convenience, when $\hat{\beta} = \hat{\beta}(\lambda)$ depends on λ we write $L(\lambda)$ instead of $L(\hat{\beta}(\lambda))$.
- Let X_M be the design matrix with columns $(X_{\cdot j} : j \in M)$ and let $\hat{\beta}_M = (X_M^T X_M)^{-1} X_M^T Y$ denote the least squares estimator.

Notation (Continue)

- If C is any square matrix, let $\phi(C)$ and $\Phi(C)$ denote the smallest and largest eigenvalues of C. Also if k is an integer define

$$\phi_n(k) = \min_{M:|M|=k} \phi\left(\frac{1}{n} X_M^T X_M\right)$$
$$\Phi_n(k) = \max_{M:|M|=k} \Phi\left(\frac{1}{n} X_M^T X_M\right).$$

- Define the type I error rate $q(\hat{D}_n) = \mathbb{P}(\hat{D}_n \cap D^c \neq \emptyset)$ and the asymptotic error rate $\limsup_{n\to\infty} q(\hat{D}_n)$. Also define the power $\pi(\hat{D}_n) = \mathbb{P}(D \subset \hat{D}_n)$.

Assumptions

The following assumptions will be used throughout the talk:

- A1 $Y_i = X_i^T \beta + \epsilon_i$ where $\epsilon \sim N(0, \sigma^2)$, for $i = 1, \ldots, n$.
- A2 The dimension p_n of X satisfies $p_n \to \infty$ and $p_n \leq c_1 e^{n^{c_2}}$ for some $c_1 > 0$ and $0 \leq c_2 < 1$.
- A3 $s = |\{j : \beta_j \neq 0\}| = O(1)$ and $\psi = \min\{|\beta_j| : \beta_j \neq 0\} > 0$.
- A4 There exist positive constants C_0, C_1 and κ such that $\mathbb{P}(\limsup_{n\to\infty} \Phi_n(n) \leq C_0) = 1$ and $\mathbb{P}(\liminf_{n\to\infty} \phi_n(C_1 \log n) \geq \kappa) = 1$. Also, $\mathbb{P}(\phi_n(n) > 0) = 1$ for all n.
- A5 The covariates are standardized: $\mathbb{E}(X_{ij}) = 0$ and $\mathbb{E}(X_{ij}^2) = 1$. Also, there exists $0 < B < \infty$ such that $\mathbb{P}(|X_{jk}| \leq B) = 1$.

Error Control

The error rate is difficult to control for three reasons

1. Correlation of covariates.
   An example where $\pi(\hat{D}_n) \approx \alpha$ if $q(\hat{D}_n) \leq \alpha$
2. High-dimensionality of the covariates.
   restrictions on the number s of nonzero β_j's.
3. Unfaithfullness (cancellations of correlations).
   Let $\hat{\mu}_j$ denote the regression coefficient from regressing Y on X_j. Fix $j < s$ and note that

$$\mu_j = \mathbb{E}(\hat{\mu}_j) = \beta_j + \sum_{k \neq j, 1 \leq k \leq s} \beta_k \rho_{kj}$$

   If $\sum_{k \neq j, 1 \leq k \leq s} \beta_k \rho_{kj} \approx -\beta_j$, then $\mu_j \approx 0$ no matter how large β_j is.

Loss

Now we record some properties of the loss function. The first part of the following lemma is essentially Lemma 3 of Meinshausen and Yu (2008).

Lemma 3.1

Let $\mathcal{M}_m^+ = \{M \subset S : |M| \leq m, D \subset M\}$. Then

$$\mathbb{P}\left(\sup_{M \in \mathcal{M}_m^+} L(\hat{\beta}_M) \leq \frac{4m \log p}{n\phi_n(m)}\right) \to 1. \tag{1}$$

Let $\mathcal{M}_m^- = \{M \subset S : |M| \leq m, D \not\subset M\}$. Then

$$\mathbb{P}\left(\inf_{M \in \mathcal{M}_m^-} L(\hat{\beta}_M) \geq \psi^2 \phi_n(m+s)\right) \to 1. \tag{2}$$

Cross-Validation

The data are split into groups $\mathcal{D}_1, \mathcal{D}_2, \mathcal{D}_3$ for each stage of size n. Construct $\hat{\beta}(\lambda)$ from $\mathcal{D}_1$ and let

$$\hat{L}(\lambda) = \frac{1}{n} \sum_{X_i \in \mathcal{D}_2} (Y_i - X_i^T \hat{\beta}(\lambda))^2.$$

We would like $\hat{L}(\lambda)$ to order the models the same way as the true loss $L(\lambda)$. This requires that, asymptotically, $\hat{L}(\lambda) - L(\lambda) \approx \delta_n$ where δ_n does not involve λ.

Theorem 3.2

Suppose that $\max_{\lambda \in \Lambda_n} |\hat{S}_n(\lambda)| \leq k_n$. Then there exists a sequence of random variables $\delta_n = O_P(1)$ that do not depend on λ or X, such that with probability tending to 1,

$$\sup_{\lambda \in \Lambda_n} |L(\lambda) - \hat{L}(\lambda) - \delta_n| = O_P\left(\frac{k_n}{n^{1-c_2}}\right) + O_P\left(\frac{k_n}{\sqrt{n}}\right)$$

Multi-Stage Procedure

The multi-stage methods use the following steps. As mentioned earlier, the data is randomly split into three parts $\mathcal{D}_1, \mathcal{D}_2, \mathcal{D}_3$ of equal size.

1. Use $\mathcal{D}_1$ to find $\hat{S}_n(\lambda)$ for each λ
2. Use $\mathcal{D}_2$ to find $\hat{\lambda}$ by cross-validation and let $\hat{S}_n = \hat{S}_n(\hat{\lambda})$
3. Use $\mathcal{D}_3$ to find the least square estimate $\hat{\beta}$ for the model $\hat{S}_n$. Let

$$\hat{D}_n = \{j \in \hat{S}_n : |T_j| > c_n\}$$

where T_j is the usual t-statistic, $c_n = z_{\alpha/2m}$ and $m = |\hat{S}_n|$.

Outline

1. Introduction
   - Goal of the Paper
   - Notation and Assumptions
2. Error Control
3. Loss and Cross-Validation
4. Multi-Stage Methods
   - The Lasso
   - Stepwise Regression
   - Marginal Regression
   - Modifications

Theorems

Let $k_n = A \log n$ where $A > 0$ is a positive constant.

Theorem 4.1

Assume that (A1)-(A5) hold, let $\Lambda_n = \{\lambda : |\hat{S}_n(\lambda)| \leq k_n\}$. Then

1. The true loss overfits: $\mathbb{P}(D \subset \hat{S}_n(\lambda_*)) \to 1$ where $\lambda_* = \underset{\lambda \in \Lambda_n}{\operatorname{argmin}} \hat{L}(\lambda)$.
2. Cross-Validation also overfits: $\mathbb{P}(D \subset \hat{S}_n(\hat{\lambda})) \to 1$ where $\hat{\lambda} = \underset{\lambda \in \Lambda_n}{\operatorname{argmin}} \hat{L}(\lambda)$.
3. Type I error is controlled: $\limsup_{n \to \infty} \mathbb{P}(D^c \cap \hat{D}_n \neq \emptyset) \leq \alpha$.

If let $\alpha = \alpha_n \to 0$ then $\hat{D}_n$ is consistent for variable selection.

Theorems (cont'd)

Theorem 4.2

Assume that (A1)-(A5) hold. Let $\alpha_n \to 0$ and $\sqrt{n}\alpha_n \to \infty$. Then the multi-stage lasso is consistent,

$$\mathbb{P}(\hat{D}_n = D) \to 1 \tag{3}$$

Theorem 4.3

Assume that (A1)-(A5) hold. Let α be fixed. Then $(\hat{D}_n, \hat{S}_n)$ forms a confidence sandwich:

$$\liminf_{n\to\infty} \mathbb{P}(\hat{D}_n \subset D \subset \hat{S}_n) \geq 1 - \alpha$$

Remark: This confidence sandwich is expected to be conservative in the sense that the coverage can be much larger than $1 - \alpha$.

Outline

1. Introduction
   - Goal of the Paper
   - Notation and Assumptions
2. Error Control
3. Loss and Cross-Validation
4. Multi-Stage Methods
   - The Lasso
   - Stepwise Regression
   - Marginal Regression
   - Modifications

Stepwise Regression

The version of stepwise regression considered is as follows. Let $k_n = A \log n$ for some $A > 0$.

1. Initialize: $\text{Res} = Y$, $\lambda = 0$, $\hat{Y} = 0$ and $\hat{S}_n(\lambda) = 0$.
2. Let $\lambda \leftarrow \lambda + 1$. Compute $\hat{\mu}_j = n^{-1}\langle X_j, \text{Res}\rangle$ for $j = 1, \ldots, p$.
3. Let $J = \underset{j}{\text{argmax}}|\hat{\mu}_j|$. Set $\hat{S}_n(\lambda) = \{\hat{S}_n(\lambda - 1), J\}$. Set $\hat{Y} = X_\lambda \hat{\beta}(\lambda)$ where $\hat{\beta}_\lambda = (X_\lambda^T X_\lambda)^{-1} X_\lambda^T Y$ and let $\text{Res} = Y - \hat{Y}$.
4. If $\lambda = k_n$ stop. Otherwise, go to step 2.

Theorem 4.5

With $\hat{S}_n(\lambda)$ defined as above, the statements of Theorems 4.1, 4.2, 4.3 hold.

Outline

1. Introduction
   - Goal of the Paper
   - Notation and Assumptions
2. Error Control
3. Loss and Cross-Validation
4. Multi-Stage Methods
   - The Lasso
   - Stepwise Regression
   - Marginal Regression
   - Modifications

Marginal Regression

A version appears in a recent paper by Fan and LV(2008). Let $\hat{S}_n(\lambda) = \{j : |\hat{\mu}_j| \geq \lambda\}$ where $\hat{\mu}_j = n^{-1}\langle Y, X_{.j}\rangle$. Let $\mu_j = \mathbb{E}(\hat{\mu}_j)$ and let $\mu_{(j)}$ denote the value of μ ordered by their absolute values:

$$|\mu_{(1)}| \geq |\mu_{(2)}| \geq \cdots$$

Theorem 4.6

Let $k_n \to \infty$ with $k_n = o(\sqrt{n})$. Let $\Lambda_n = \{\lambda : |\hat{S}_n(\lambda)| \leq k_n\}$. Assume that

$$\min_{j \in D} |\mu_j| > |\mu_(k_n)|.$$

Then, the statements of Theorems 4.1, 4.2, 4.3 hold.

Comments on the assumption in Theorem 4.6:

- The assumption on μ_j's limits the degree of unfaithfulness.
- Fan and Lv make similar assumptions. They assume that there is a $C > 0$ such that $|\mu_j| \geq C|\beta_j|$ for all j, which also rules out unfaithfulness.
- They also assume $Z = \Sigma^{-1/2}X$ has a spherically symmetric distribution. Under this assumption, they deduce that the μ_j's outside D cannot dominate the μ_j's within D, which is the same assumption as in Theorem 4.3.
- Any method that start from marginal regression must take some sort of faithfulness assumptions to succeed.

Outline

1. Introduction
   - Goal of the Paper
   - Notation and Assumptions
2. Error Control
3. Loss and Cross-Validation
4. Multi-Stage Methods
   - The Lasso
   - Stepwise Regression
   - Marginal Regression
   - Modifications

Let us now discuss a few modifications of the basic method. First, consider splitting the data only into two groups $\mathcal{D}_1$ and $\mathcal{D}_2$. The do these steps:

1. Find $\hat{S}_n(\lambda)$ for $\lambda \in \Lambda$, where $|\hat{S}_n(\lambda)| \leq k_n$ for each $\lambda \in \Lambda$ using $\mathcal{D}_1$.
2. Find $\hat{\lambda}$ by cross-validation and let $\hat{S}_n = \hat{S}_n(\hat{\lambda})$ using $\mathcal{D}_2$.
3. Find the least squares estimate $\hat{\beta}_{\hat{S}_n}$ using $\mathcal{D}_2$. Let $\hat{D}_n = \{j \in \hat{S}_n : |T_j| > c_n\}$ where T_j is the usual t-statistic.

Theorem 4.7

Choosing

$$c_n = \frac{\log \log n \sqrt{2k_n \log(2p_n)}}{\alpha}$$

controls asymptotic type I error.

The critical value c_n is hopelessly large. This part is mainly to show the value of extra data-splitting step.