

Incentive sparse regression

Yongdai Kim

Department of Statistics, Seoul National University, Korea

1. Introduction

- Variable selection is important for high dimensional models.
- Traditional approaches such as stepwise selections are
 - computationally intensive
 - hard to draw sampling properties
 - unstable
- Alternative approach is sparse, which means some coefficients are exactly zero, penalized approaches including
 - bridge regression (Frank and Friedman, 1993)
 - Lasso (Least Absolute Shrinkage and Selection Operator, Tibshirani, 1997)
 - SCAD (Smoothly Clipped Absolute Deviation, Fan and Li 2001)

Sparse penalized approaches

- Data: $(y_1, \mathbf{x}_1), \dots, (y_n, \mathbf{x}_n)$ where $y_i \in R$ and $\mathbf{x}_i \in R^p$.
- Let $C_n(\beta) = \sum_{i=1}^n l(y_i, \mathbf{x}_i' \beta) / n$.
- General form of sparse penalized estimators

$$\hat{\beta} = \operatorname{argmin}_{\beta} C_n(\beta) + \lambda \sum_{j=1}^p J_{\lambda}(|\beta_j|)$$

for some penalty function J .

- Various penalty functions

- Bridge: $J(\beta) = \beta^q, q > 0$

- Lasso: $J(\beta) = \beta$

- SCAD:

$$\begin{aligned} J_\lambda(\beta) &= \lambda\beta I(0 \leq \beta < \lambda) \\ &+ \left(\frac{a\lambda(\beta - \lambda) - (\beta^2 - \lambda^2)/2}{(a - 1)} + \lambda^2 \right) I(\lambda \leq \beta \leq a\lambda) \\ &+ \left(\frac{(a - 1)\lambda^2}{2} + \lambda^2 \right) I(\beta \geq a\lambda). \end{aligned}$$

- * Ridge: $J(\beta) = \beta^2$ (non-sparse)

Less sparse penalized approaches

- Sparse methods are good only when the true model is sparse.
- When there are highly correlated covariates, in some cases, the average of the covariates would be better than the selection of a covariate.

- An example is a measurement error model.
 - True model: $Y = F + \epsilon$ where $F \sim N(0, 1) \perp \epsilon \sim N(0, \sigma^2)$.
 - Data: (Y, X_1, X_2) where $X_j = F + \xi_j$ and $\xi_j \sim N(0, \tau^2) \perp F$.
 - Then

$$\operatorname{argmin}_{\beta_1, \beta_2} \mathbf{E}(Y - \beta_1 X_1 + \beta_2 X_2)^2 = (1/2, 1/2).$$

- Highly correlated covariates are frequently met in high dimensional problems.
- We need a method which provides a less sparse solution.
- There are two such approaches
 - Elastic net: Zou and Hastie (2005)
 - Gradient Directed regularization: Friedman and Popescu (2004)

The objective of the talk

- Propose a new penalty called *the incentive sparse penalty*, which compromises advantages of the aforementioned less sparse approaches.
- Develop an efficient computational algorithm
- Perform numerical studies

2. Less sparse approaches: Review

Elastic net

- The main idea of the elastic net is to combine the ridge and Lasso.
- That is, we minimize the following objective function:

$$\sum_{i=1}^n (y_i - \boldsymbol{\beta}' \mathbf{x}_i)^2 + \lambda_1 \|\boldsymbol{\beta}\|_1 + \lambda_2 \|\boldsymbol{\beta}_2\|_2^2.$$

- A problem of the above naive elastic net is that the estimator is doubly regularized.
- In fact, we want to control the sparsity by controlling the ratio of λ_1/λ_2 .
- That is, we want to have an estimator in between the lasso and univariate estimator by controlling λ_1/λ_2 .

- Here, the univariate estimator is defined by

$$\hat{\beta}_j = \sum_{i=1}^n (y_i - \beta_j x_{ij})^2$$

for $j = 1, \dots, p$.

- The univariate estimator can be thought to be the least sparse solution.
- And, we wish that the univariate estimator could be obtained by letting $\lambda_1 \rightarrow 0$ and $\lambda_2 \rightarrow \infty$.
- In this case, however, the norm of the naive elastic net estimator converges to 0.
- To overcome such deficiency, Zou and Hastie (2005) proposed an ad-hoc modification:

$$\hat{\beta}(\text{elastic net}) = (1 + \lambda_2) \hat{\beta}(\text{naive elastic net}).$$

- It can be shown that the elastic net estimator converges to the univariate estimator if we let $\lambda_1 \rightarrow 0$ and $\lambda_2 \rightarrow \infty$.
- Also, Zou and Hastie (2005) showed numerically that the elastic net outperforms the naive elastic net as well as the lasso when predictive variables are highly correlated.
- Problem: It is not easy to find an appropriate scaling factor (i.e. $1 + \lambda_2$ for the squared error loss) for general loss functions.

Gradient directed regularization

- Let $\hat{\beta}(\lambda)$ be the minimizer of the penalized empirical risk

$$C_n(\beta) + \lambda \sum_{j=1}^p J_\lambda(|\beta_j|).$$

- The gradient descent method update $\hat{\beta}(\lambda)$ by

$$\hat{\beta}(\lambda + \Delta\lambda) = \hat{\beta}(\lambda) + \Delta\lambda \cdot g(\lambda) \quad (1)$$

where $g(\lambda)$ is the gradient of the empirical risk evaluated at $\hat{\beta}(\lambda)$

$$g(\lambda) = \left. \frac{\partial C_n(\beta)}{\partial \beta} \right|_{\beta=\hat{\beta}(\lambda)}.$$

- Gradient directed approach replaces $g(\lambda)$ with other vector $f(\lambda)$.

- For example, we can set

$$f_j(\lambda) = g_j(\lambda) I \left(|g_j(\lambda)| \geq \tau \max_{k=1, \dots, p} |g_k(\lambda)| \right)$$

for some value $\tau \in [0, 1]$.

- If $\tau = 1$, (1) is related to the LARS algorithm and hence the Lasso estimator.
- If $\tau = 0$, (1) is related to the PLS (partial least square) algorithm.
- We can control the sparsity of the solution by controlling τ .
- Problem: It is not easy to study properties of the estimator since it is not defined as the minimizer of a penalized empirical risk.

3. Incentive sparse regression

Motivation

- Two required properties
 - Selection (i.e. sparsity)
 - Average among highly correlated predictive variables
- We use the idea of LARS for selection.
- For average effect, we use a modified gradient $\tilde{g}(\lambda)$ such that $|\tilde{g}_j(\lambda)|$ is larger than $|g_j(\lambda)|$ when the j th predictive variable is highly correlated with other **signal** predictive variables.

- For this purpose, we let

$$\tilde{g}_j(\lambda) = g_j(\lambda) - \lambda J_j(\hat{\beta}(\lambda))$$

where

$$J_j(\beta) = \sum_{k \neq j} \beta_k \text{corr}(\mathbf{z}_k, \mathbf{z}_j).$$

- Here, $\mathbf{z}_j = (x_{1j}, \dots, x_{nj})'$.
- Note that
 - $g_j(\lambda)$ is usually negative since we minimize the empirical risk.
 - When β_k is large and \mathbf{z}_j is highly correlated with \mathbf{z}_k , $\tilde{g}_j(\lambda) < g_j(\lambda)$, and hence the chance of selecting \mathbf{z}_j is larger than that in the LARS.
- To sum up, at each stage, we select a predictive variable which has the largest value of $|\tilde{g}_j(\lambda)|$.

Incentive Sparse regression

- Note that the modified gradient is the gradient of

$$C_n(\beta) - \lambda \beta' \text{corr}(\mathbf{Z})\beta + \lambda \sum_{k=1}^p \beta_k^2.$$

- We propose an incentive sparse regression estimator as a minimizer of the following penalized empirical risk

$$C_n(\beta) - \lambda_2 \beta' \text{corr}(\mathbf{Z})\beta + \lambda_2 \sum_{k=1}^p \beta_k^2 + \lambda_1 \sum_{k=1}^p |\beta_k|.$$

- The term $-\lambda_2 \beta' \text{corr}(\mathbf{Z})\beta$ can be thought of as an incentive (an antonym of the penalty).

Relation to the Elastic net

- Suppose $l(y, a) = (y - a)^2$ and $\mathbf{X}'\mathbf{X} = \text{corr}(\mathbf{Z})$.
- Let $\beta^E(\lambda_1^E, \lambda_2^E)$ be the elastic net estimator with the regularization parameter $(\lambda_1^E, \lambda_2^E)$.
- Then, $\beta^E(\lambda_1^E, \lambda_2^E)$ is an incentive sparse regression estimator with
 - $\lambda_1^E = \lambda_1 / (1 - \lambda_2)$
 - $\lambda_2^E = \lambda_2 / (1 - \lambda_2)$.
- A surprising result is that we don't need an ad-hoc rescaling.
- Hence, the incentive sparse regression can be applied to general loss functions.

Remark

- About $\text{corr}(\mathbf{Z})$.
 - We can use any (positive definite) matrix \mathbf{A} instead of $\text{corr}(\mathbf{Z})$.
 - For example, we can let $\mathbf{A} = [K(\mathbf{z}_j, \mathbf{z}_k)]$ where K is a some Mercer kernel popularly used in SVM.
 - This would be helpful when the correlation is not the best similarity measure (e.g. categorical predictive variables).
- About the choice of λ_2
 - When λ_2 is too large, the objective function may not be convex.
 - Sometimes, it can be almost concave.
 - We should be careful to choose λ_2 .

4. Computation

Convex-Concave Procedure (CCCP, Yille and Rangarajan 2003)

- Suppose we are to minimize a non-convex function $C(\beta)$.
- Suppose $C(\beta)$ is a sum of convex and concave functions $C_{vex}(\beta)$ and $C_{cav}(\beta)$ such as

$$C(\beta) = C_{vex}(\beta) + C_{cav}(\beta).$$

- For a given current solution β^c , the tight convex upper bound is defined by $Q(\beta) = C_{vex}(\beta) + \nabla C_{cav}(\beta^c)' \beta$ where $\nabla C_{cav}(\beta) = \partial C_{cav}(\beta) / \partial \beta$.
- We update the solution by the minimizer of $Q(\beta)$ which is convex.
- Repeat this until convergence. It always converges to a local minimum.

CCCP for the incentive sparse regression

- $C_{vex}(\beta) = C_n(\beta) + \lambda_2 \sum_{k=1}^p \beta_k^2 + \lambda_1 \sum_{k=1}^p |\beta_k|$
- $C_{cav}(\beta) = -\lambda \beta' \text{corr}(\mathbf{Z}) \beta$
- Hence

$$Q(\beta) = C_n(\beta) + \lambda_2 \sum_{k=1}^p \beta_k^2 + \lambda_1 \sum_{k=1}^p |\beta_k| - 2\lambda_2 \beta^{c'} \text{corr}(\mathbf{Z}) \beta.$$

- There are many algorithms for minimizing $Q(\beta)$ (Osborne 2000, Rosset and Zhu 2007, Park and Hastie 2007, Kim et al. 2008).

5. Numerical studies

Simulation 1: Highly correlated case

- Logistic regression
- $n = 20, p = 15$
- 15 predictive variables generated from multivariate Gaussian distribution are divided into the five blocks of equal size where
 - Within block: highly correlated (correlation is close 1)
 - Between block: independent
- The first three blocks are signal and the other two blocks are noise.
- The true regression coefficients are set to be

$$\beta = (5, 4, 1, 0.5, 0.4, 0.1, 0.3, 0.3, 0.3, 0, 0, 0, 0, 0, 0)$$

- The regularization parameter is selected using a validation data set of size 20.
- Test error is calculated using a test data set of size 1000.
- We repeat the simulation 50 times.
- The results (average error rates) are

Table 1: Error rates (std. errors)

	lasso	naive EN	ISR
0-1 loss	0.0969(0.00554)	0.1009(0.00572)	0.0857(0.00417)
log-likelihood	0.1018(0.00569)	0.1011(0.00551)	0.0820(0.00376)

Simulation 2: Measurement error model

- True model

$$\text{logit Pr}(Y = 1|Z) = Z' \eta$$

where Z is generated from a 9 dimensional multivariate Gaussian distribution with mean 0 and identity covariance matrix.

- $\eta = (4.7, -4.3, 0.3, 0.4, -0.3, 0, 0, 0, 0)$
- $X_{3i+j} = (-1)^{j-1} Z_i + \epsilon_{ij}$, $\epsilon_{ij} \sim N(0, 1/16)$ for $j = 1, 2, 3$ and $i = 0, \dots, 8$.
- That is, X is a 27 dimensional vector whose entries are divided into 9 groups of size 3.
- Construct a predictive model using $(Y_1, X_1), \dots, (Y_n, X_n)$.
- The other set-ups are the same as those for Case 1.
- Results (average error rates) are

Table 2: Error rates (std. errors)

	lasso	naive EN	ISR
0-1 loss	0.0767	0.0817	0.0717
log-likelihood	0.2787	0.3791	0.2528

Real data analysis

- Analyze 4 micorarray data sets
- 100 covariates are selected in advance using marginal correlation
- Test errors are measured by 100 random partition (70% traing and 30% test)
- The regularization parameters are selected by 5 fold cross-validation

- Results (average error rates) are

Table 3: Missclassification error

	lasso	naive EN	ISR
prostate	0.366(0.00859)	0.360(0.00808)	0.324 (0.00662)
lung	0.0393(0.00248)	0.0410(0.00245)	0.0295 (0.00229)
colon	0.182(0.0100)	0.177 (0.00948)	0.180(0.00891)
Singapore	0.349(0.0090)	0.317 (0.0101)	0.326 (0.00905)

Table 4: Log-likelihood errors

	lasso	naive	ISR
prostate	1.947(0.1221)	1.908(0.100)	1.529 (0.0577)
lung	0.0967(0.00563)	0.0966(0.00564)	0.0951 (0.00549)
colon	3.868(0.1053)	4.072(0.1914)	2.764 (0.0860)
Singapore	0.147(0.00763)	0.140 (0.00754)	0.142(0.00755)

- Comments

- Incentive sparse regression is always better than the lasso estimator.
- Incentive sparse regression is competitive to the naive elastic net for the missclassification error rates.
- This is not surprising since the naive elastic net and the incentive sparse regression give the same missclassification error rates when we use the square error loss (the signs of the predictive models are the same).
- For the log-likelihood loss, the incentive sparse regression seems to perform better than the naive elastic net. For ‘colon’ data set, the log-likelihood error of the incentive sparse regression is 33% less than that of the naive elastic net while the missclassification error rates are close.

- That is, the naive elastic net still works well for classification, but not for estimating the probability. In contrast, the incentive sparse regression works well for both classification and probability estimation.

6. Concluding Remark

- What I have done
 - proposed a new penalty for less sparse solutions
 - developed an efficient computational algorithm
 - did simulation and real data analysis
- What I have to do
 - More simulation (choice of initial value, choice of \mathbf{A} , high dimensional problem etc)
 - More real data analysis (more data sets, correlation structure vs performance etc)
 - Enjoy Beijing!!!