Factor-Adjusted Regularized Model Selection

Yuan Ke

University of Georgia

Joint work with Jianqing Fan and Kaizheng Wang

GSD 2018

October 26, 2018

Outline

Background and Motivation

High dimensional sparse regression

Model selection has become a fundamental approach in high dimensional regression problems

- LASSO Tibshirani,1996
- SCAD Fan and Li, 2001
- Elastic net Zou and Hastie, 2005
- Dantzig selector Candes and Tao, 2007 and more

- Computational biology
- Health studies
- Financial engineering and risk management
- Machine learning and data mining

...

Fan and Li, 2006; Johnston and Titterington, 2009; Bühlmann and Van De Geer, 2011.

Selection consistency

How close between the estimator and true parameter?

Estimation consistency $\|\widehat{\beta} - \beta^*\| \to 0$

How well the sparse solution associates with the true model?

Selection consistency $P(\text{supp}(\widehat{\beta}) = \text{supp}(\beta^*)) \to 1$

- Fan and Li (2001) studied the oracle property for folded concave penalty functions.
- Zhao and Yu (2006) studied sign consistency and derived the *irrepresentable condition*.
- Bunea (08) and Ravikumar *etal.* (2010) regularized logistic regression.
- Van De Geer and Müller (2012) θ-*irrepresentable condition*.

and references therein.

Selection consistency

How close between the estimator and true parameter?

Estimation consistency $\|\widehat{\beta} - \beta^*\| \to 0$

How well the sparse solution associates with the true model?

Selection consistency $P(\mathrm{supp}(\widehat{\beta}) = \mathrm{supp}(\beta^*)) \to 1$

- Fan and Li (2001) studied the oracle property for folded concave penalty functions.
- Zhao and Yu (2006) studied sign consistency and derived the *irrepresentable condition*.
- Bunea (08) and Ravikumar *etal.* (2010) regularized logistic regression.
- Van De Geer and Müller (2012) θ-*irrepresentable condition*.

and references therein.

Irrepresentable condition

LASSO estimator $\widehat{\beta} \in \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \left\{ \frac{1}{2n} \|\mathbf{Y} - \mathbf{X}\beta\|_2^2 + \lambda \|\beta\|_1 \right\}$.

- $\operatorname{supp}(\beta^*) = [S] = s$
- $\mathbf{X}_S$ and $\mathbf{X}_{S^c}$ the first s columns and the rest $p - s$ columns of $\mathbf{X}$

Irrepresentable condition (Zhao and Yu, 06)

$$\|\mathbf{X}_{S^c}^T \mathbf{X}_S (\mathbf{X}_S^T \mathbf{X}_S)^{-1}\|_\infty < 1 - \tau, \quad \tau \in (0, 1)$$

Problems:

★ Hard to verify!

★ Correlated datasets!

★ Superious correlation in High-D data!

Irrepresentable condition

LASSO estimator $\widehat{\beta} \in \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \left\{ \frac{1}{2n} \|\mathbf{Y} - \mathbf{X}\beta\|_2^2 + \lambda \|\beta\|_1 \right\}.$

- $\operatorname{supp}(\beta^*) = [S] = s$
- $\mathbf{X}_S$ and $\mathbf{X}_{S^c}$ the first s columns and the rest $p - s$ columns of $\mathbf{X}$

Irrepresentable condition (Zhao and Yu, 06)

$$\|\mathbf{X}_{S^c}^T \mathbf{X}_S (\mathbf{X}_S^T \mathbf{X}_S)^{-1}\|_\infty < 1 - \tau, \quad \tau \in (0, 1)$$

Problems:

★ Hard to verify!

★ Correlated datasets!

★ Superious correlation in High-D data!

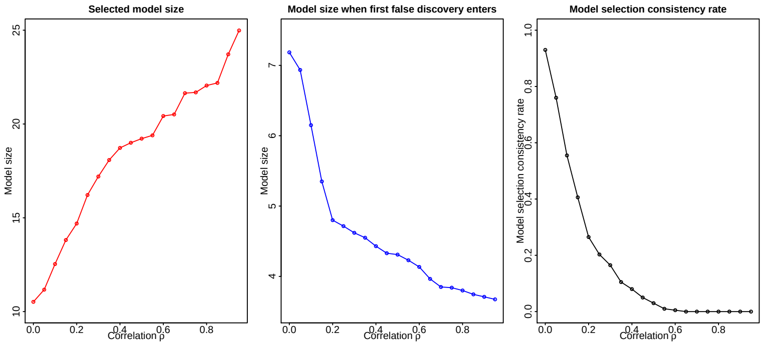
A motivative example

Sparse linear model $\mathbf{Y} = \mathbf{X}\beta^* + \varepsilon$ with $n = 100$ and $p = 200$

- $\beta^* = (\beta_1, \cdots, \beta_{10}, \mathbf{0}^T_{(p-10)})^T$, Nonzero $\beta \sim$ i.i.d. Uniform $[2, 5]$
- $\varepsilon \sim N_n(\mathbf{0}, \mathbf{I})$
- $\mathbf{X} = (x_1, \cdots, x_p)^T \sim N_p(\mathbf{0}, \Sigma)$
- $\Sigma =$ with diag. 1 and off-diag. some $\rho \in [0, 1)$.

★ Model selection with LASSO when ρ increase from 0 to 0.95 by a step size 0.05. For each given ρ, we simulate 200 replications.

A motivative example



★ X axis: correlation level ρ increase from 0 to 0.95

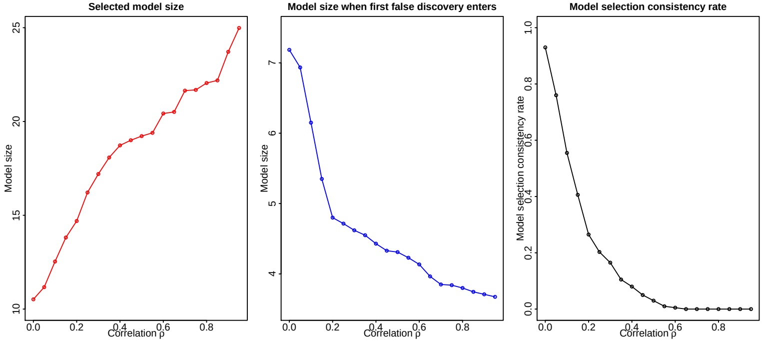
★ Y axis from left to right:

L: Average model size selected by LASSO

M: Average model size when the first false discovery $(x_j, j > 10)$ enters the solution path

R: Average model selection consistency rate (ratio of exactly model recovery)

A motivative example



When covariates are strongly correlated:

★Inflated model size ★Early selection of false variables ★Selection inconsistency

Beyond weakly correlated assumption

Weakly correlated → Conditional weakly correlated

Approximate factor model

$$\mathbf{X} = \mathbf{F}\mathbf{B}^T + \mathbf{U}.$$

- Strongly dependent K latent common factors $\mathbf{F} \in \mathbb{R}^{n \times K}$
- Weakly dependent idiosyncratic components $\mathbf{U} =\in \mathbb{R}^{n \times p}$

Beyond weakly correlated assumption

Weakly correlated → Conditional weakly correlated

Approximate factor model

$$\mathbf{X} = \mathbf{F}\mathbf{B}^T + \mathbf{U}.$$

- Strongly dependent K latent common factors $\mathbf{F} \in \mathbb{R}^{n \times K}$
- Weakly dependent idiosyncratic components $\mathbf{U} = \in \mathbb{R}^{n \times p}$

Factor-Adjusted Regularized Model Selection (FarmSelect)

Problem setup

Regularized M-estimator

$$\widehat{\boldsymbol{\beta}} \in \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\operatorname{argmin}} \{L_n(\mathbf{Y},\, \mathbf{X}\boldsymbol{\beta}) + \lambda R_n(\boldsymbol{\beta})\},$$

- $\mathbf{Y} = (y_1,\, \cdots y_n)^T \in \mathbb{R}^n$ and $\mathbf{X} = (x_1,\, \cdots x_n)^T \in \mathbb{R}^{n\times p}$
- $L_n(\mathbf{Y},\, \mathbf{X}\boldsymbol{\beta})$ convex and differentiable loss function
- $\boldsymbol{\beta}^* \in \mathbb{R}^p$ unique minimizer $\mathbb{E}L_n(\mathbf{Y},\, \mathbf{X}\beta)$, sparse with s non-zero elements
- $R_n : \mathbb{R}^p \to \mathbb{R}_+$ penalty and $\lambda > 0$ is a tuning parameter

Intuition

By the approximate factor model

$$\mathbf{X}\beta = \mathbf{F}\mathbf{B}^T\beta + \mathbf{U}\beta := \mathbf{F}\gamma + \mathbf{U}\beta,$$

The regularized M-estimator can be rewritten as

$$\widehat{\beta} \in \underset{\gamma\in\mathbb{R}^K,\ \beta\in\mathbb{R}^p}{\operatorname{argmin}} \left\{ L_n(\mathbf{Y}, \mathbf{F}\gamma + \mathbf{U}\beta) + \lambda R_n(\beta) \right\}.$$

Our goal:

(1) Identifying the highly correlated latent factors $\mathbf{F}$.

(2) Transform to model selection with weakly correlated $\mathbf{U}$.

FarmSelect procedure

Step 1: Factor estimation

Fit the approximate factor model and denote $\widehat{\mathbf{B}}$, $\widehat{\mathbf{F}}$ and $\widehat{\mathbf{U}} = \mathbf{X} - \widehat{\mathbf{F}}\widehat{\mathbf{B}}^T$ the obtained estimates of $\mathbf{B}$, $\mathbf{F}$ and $\mathbf{U}$ respectively.

Step 2: Augmented M-estimation

Define $\widehat{\mathbf{W}} = (\widehat{\mathbf{F}}, \widehat{\mathbf{U}})$ and $\theta = (\gamma^T, \beta^T)^T$. Then $\widehat{\beta}$ can be obtained by solving the following augmented problem

$$\widehat{\theta} \in \underset{\theta \in \mathbb{R}^{K+p}}{\operatorname{argmin}} \left\{ L_n(\mathbf{Y}, \widehat{\mathbf{W}}\theta) + \lambda R_n(\theta_{[K^c]}) \right\}.$$

- Convex opt. algorithms: coordinate descent and ADMM.

FarmSelect procedure

Step 1: Factor estimation

Fit the approximate factor model and denote $\widehat{\mathbf{B}}$, $\widehat{\mathbf{F}}$ and $\widehat{\mathbf{U}} = \mathbf{X} - \widehat{\mathbf{F}}\widehat{\mathbf{B}}^T$ the obtained estimates of $\mathbf{B}$, $\mathbf{F}$ and $\mathbf{U}$ respectively.

Step 2: Augmented M-estimation

Define $\widehat{\mathbf{W}} = (\widehat{\mathbf{F}}, \ \widehat{\mathbf{U}})$ and $\theta = (\gamma^T, \ \beta^T)^T$. Then $\widehat{\beta}$ can be obtained by solving the following augmented problem

$$\widehat{\theta} \in \underset{\theta \in \mathbb{R}^{K+p}}{\operatorname{argmin}} \left\{ L_n(\mathbf{Y}, \widehat{\mathbf{W}}\theta) + \lambda R_n(\theta_{[K^c]}) \right\}.$$

- Convex opt. algorithms: coordinate descent and ADMM.

Estimating approximate factor model

Estimation of factors

- Applying PCA on the the $n \times n$ matrix $\mathbf{X}\mathbf{X}^T$
- $\widehat{\mathbf{F}}/\sqrt{n}$ is estimated as top K eigenvectors
- Normalization $\mathbf{F}^T\mathbf{F}/n = \mathbf{I}_K$ yields $\widehat{\mathbf{B}} = \mathbf{X}^T\widehat{\mathbf{F}}/n$.

Estimation of the number of factors

Eigen-ratio method (Lam and Yao, 2013; Ahn and Horenstein, 2013)

$$\widehat{K} = \underset{k \le K_{max}}{\operatorname{argmax}} \frac{\lambda_k(\mathbf{X}\mathbf{X}^T)}{\lambda_{k+1}(\mathbf{X}\mathbf{X}^T)}.$$

- K_{max} a prescribed upper bound
- $\lambda_k(\mathbf{X}\mathbf{X}^T)$ the kth largest eigenvalue of $\mathbf{X}\mathbf{X}^T$

Example: sparse linear model

Penalized profile least-squares solution

$$\widehat{\beta} \in \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \left\{ \frac{1}{2n} \|(\mathbf{I}_n - \widehat{\mathbf{P}})(\mathbf{Y} - \widehat{\mathbf{U}}\beta)\|_2^2 + \lambda \|\beta\|_1 \right\},$$

- $\widehat{\mathbf{P}} = \widehat{\mathbf{F}}(\widehat{\mathbf{F}}^T\widehat{\mathbf{F}})^{-1}\widehat{\mathbf{F}}^T$ is the $n \times n$ projection matrix onto the column space of $\widehat{\mathbf{F}}$.

Projection representation

$$\begin{aligned}(\mathbf{I}_n - \widehat{\mathbf{P}})\mathbf{Y} &= (\mathbf{I}_n - \widehat{\mathbf{P}})\widehat{\mathbf{U}}\beta^* + (\mathbf{I}_n - \widehat{\mathbf{P}})\varepsilon \\ &\approx \widehat{\mathbf{U}}\beta^* + \varepsilon\end{aligned}$$

★Model selection with decorrelated design matrix $(\mathbf{I}_n - \widehat{\mathbf{P}})\widehat{\mathbf{U}}$ (Kneip and Sarda, 2011)

Numerical Results

Simulated example: linear regression

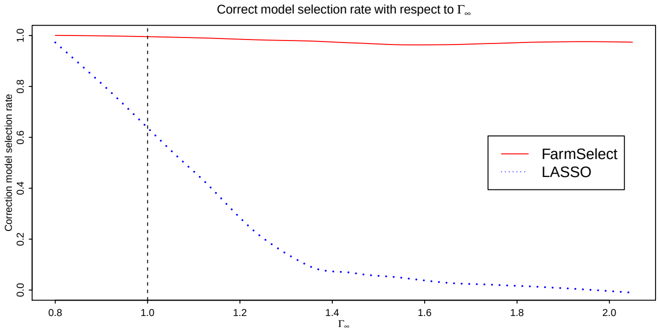
Sparse linear regression

$$\mathbf{Y} = \mathbf{X}\beta^* + \varepsilon$$

- The correlation structure is calibrated from S&P 500 monthly excess returns between 1980 and 2012.
- $\beta^* = (\beta_1, \cdots, \beta_{10}, \mathbf{0}^T_{(p-10)})^T$, with nonzero coefficients drawn from i.i.d. Uniform $[2, 5]$.
- ε drawn from i.i.d. Normal distribution $N(0,\ 1)$
- Tuning parameter λ is selected by the 10-fold cross validation

Impacts of correlations level

Selection Consistency rate with respect to correlation level

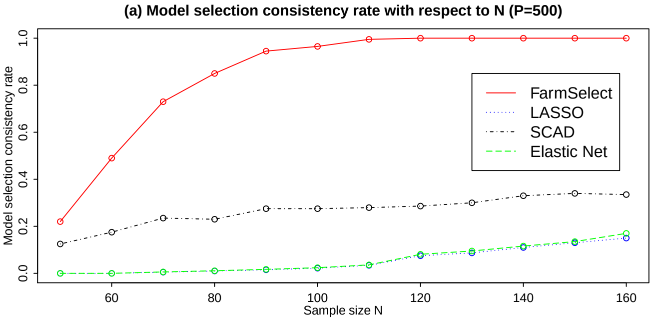


★ $n = 100$ $p = 500$ and 10,000 replications

★ $\Gamma_\infty = \|\mathbf{X}_{S^c}^T \mathbf{X}_S (\mathbf{X}_S^T \mathbf{X}_S)^{-1}\|_\infty$

Impacts on sample size

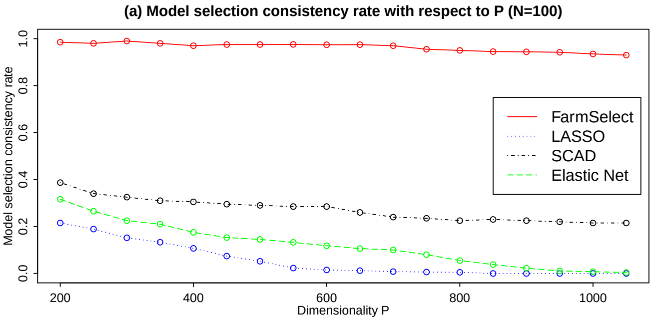
Selection Consistency rate with fixed dim. and an increasing sample size



★ Fix $p = 500$, n increase from 50 to 150, and 200 replications

Impacts on dimensionality

Comparison of MSC rate with fixed sample size and an increasing dim.



★ Fix $n = 100$, p increase from 200 to 1000, and 200 replications

Empirical Application

Gene expression based classifier for Neuroblastoma trials

- German Neuroblastoma Trials NB90-NB2004 diagnosed between 1989 and 2004 Oberthuer *et al.*(06)
- 3-year event-free survival information of 246 neuroblastoma patients (56 positive and 190 negative)
- Gene expressions over 10,707 probe sites

Challenges

- High dimensionality
- Strong correlation caused by gene-gene interaction

Strong correlation among genes

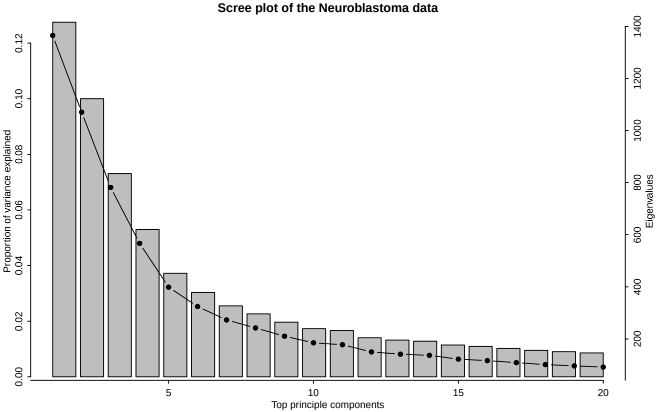


Figure: Eigenvalues (dotted line) and proportion of variance explained (bar) by the top 20 principal components

★ Top ten PC explain more than 50% of the total variance!

Setting

Model selection with FarmSeelct

- High dimensional sparse logistic regression model
- The correlation structure is estimated by a factor model
- The ratio method (Lam and Yao, 2012) suggests $\widehat{K} = 4$

Competing model selection methods

★LASSO ★SCAD ★Elastic net ($\lambda_1 = \lambda_2$)

Performance measure

Bootstrap based out-of-sample prediction

- Select and fit a model with random 200 observations
- Prediction with the remaining 46 observations
- Classified the patient into the group with higher estimated conditional probability

Performance measure

- Selected model size
- Correct prediction rate (# of correct predictions/46).

Selection and classification results

Bootstrap sample ave.	Model selection methods			
	FarmSelect	Lasso	SCAD	elastic net
Model size	**17.6**	46.2	34.0	90.0
Correct prediction rate	**0.813**	0.807	0.809	0.790
Prediction performance with first 17 variables enter the solution path				
	FarmSelect	Lasso	SCAD	elastic net
Correct prediction rate	**0.813**	0.733	0.764	0.705

★ FarmSelect selects smallest model with highest prediction rate.

★ False discovery enters solutions path early for other methods.

Theoretical Results

Questions

Regularized M-estimator

$$\widehat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}\in\mathbb{R}^{K+p}}{\operatorname{argmin}}\{L_n(\boldsymbol{\theta}) + \lambda\|\boldsymbol{\theta}_{[K]^c}\|_1\} \quad \text{and} \quad \widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\theta}}_{[K]^c},$$

★ $S = \text{supp}(\boldsymbol{\theta}^*), \quad S_1 = \text{supp}(\boldsymbol{\beta}^*), \quad S_2 = [p+K]\backslash S$

How the correlation level among covariates will affect:

(1) Estimation consistency

(2) Selection consistency

Questions

Regularized M-estimator

$$\widehat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}\in\mathbb{R}^{K+p}}{\operatorname{argmin}}\{L_n(\boldsymbol{\theta}) + \lambda\|\boldsymbol{\theta}_{[K]^c}\|_1\} \quad \text{and} \quad \widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\theta}}_{[K]^c},$$

★ $S = \text{supp}(\boldsymbol{\theta}^*), \quad S_1 = \text{supp}(\boldsymbol{\beta}^*), \quad S_2 = [p+K]\backslash S$

How the correlation level among covariates will affect:

(1) Estimation consistency

(2) Selection consistency

Estimation consistency: error bounds in norms

Error bounds : Under some Assumptions, if

$$\frac{7}{\tau}\|\nabla L_n(\boldsymbol{\theta}^*)\|_\infty < \lambda < \frac{\kappa_2}{4\sqrt{|S|}} \min\left\{A, \frac{\kappa_\infty \tau}{3M}\right\},$$

then $\mathrm{supp}(\widehat{\boldsymbol{\theta}}) \subseteq S$ and

$$\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_\infty \leq \frac{3}{5\kappa_\infty}(\|\nabla_S L_n(\boldsymbol{\theta}^*)\|_\infty + \lambda),$$

$$\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_2 \leq \frac{2}{\kappa_2}(\|\nabla_S L_n(\boldsymbol{\theta}^*)\|_2 + \lambda\sqrt{|S_1|}),$$

$$\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_1 \leq \min\left\{\frac{3}{5\kappa_\infty}(\|\nabla_S L_n(\boldsymbol{\theta}^*)\|_1 + \lambda|S_1|), \frac{2\sqrt{|S|}}{\kappa_2}(\|\nabla_S L_n(\boldsymbol{\theta}^*)\|_2 + \lambda\sqrt{|S_1|})\right\}.$$

★ τ denotes the correlation level between active and in-active sets

★ κ_∞ and κ_2 are two positive constants.

Selection consistency

Sign consistency : In addition, if the following two conditions

$$\min\{|\beta_j^*| : \beta_j^* \neq 0,\ j \in [p]\} > \frac{C}{\kappa_\infty \tau} \|\nabla L_n(\theta^*)\|_\infty,$$

$$\|\nabla L_n(\theta^*)\|_\infty < \frac{\kappa_2 \tau}{7C\sqrt{|S|}} \min\left\{A, \frac{\kappa_\infty \tau}{3M}\right\}$$

hold for some $C \geq 5$, then by taking

$\lambda \in (\frac{7}{\tau}\|\nabla L_n(\theta^*)\|_\infty, \frac{1}{\tau}(\frac{5C}{3} - 1)\|\nabla L_n(\theta^*)\|_\infty)$, the estimator achieves the sign consistency $\mathrm{sign}(\widehat{\beta}) = \mathrm{sign}(\beta^*)$.

Highlights of theoretical results

Effects of correlated covariates

- L^∞ and L^2 errors will scale with $(\kappa_\infty \tau)^{-1}$ and $(\kappa_2 \tau)^{-1}$
- Sign consistency will fail under strong correlation
- Optimal error bounds $\to$ small λ $\to$ overfitted model

★ Trade-off between model selection and parameter estimation due to the existence of strong correlation!

Summary

Highlights of our method

- Identify strong correlation structure among covariates
- Transform to model selection with weak correlated components
- No price paid under weak correlation case
- Applicable to general regularized M-estimators (loss function, penalty, correlations)

★ FarmSelect method achieves both selection consistency and estimation consistency under strong correlation!

★ R-package named FarmSelect available on CRAN.

The End

Thank You