Big Data Big Bias Small Surprise!

S. Ejaz Ahmed

Faculty of Math and Science
Brock University, ON, Canada

sahmed5@brocku.ca
www.brocku.ca/sahmed

HDDA-IV 2014
August 24-28, 2014
Joint Work with X. Gao

Outline of Presentation

Proposed Estimation Strategies

Asymptotic and Simulation Study

Applications

Envoi

Outline of Presentation

- Proposed Estimation Strategies
- Asymptotic and Simulation Study
- Applications
- Envoi

Outline of Presentation

Proposed Estimation Strategies

Asymptotic and Simulation Study

Applications

Envoi

Outline of Presentation

- Proposed Estimation Strategies
- Asymptotic and Simulation Study
- Applications
- Envoi

Outline of Presentation

Outline of Presentation

Classical Linear Model

Consider a classical linear model with observed response variable y_i and covariates $\mathbf{x}_i = (x_{i1}, \cdots, x_{ip_n})'$ as follows,

$$y_i = \mathbf{x}_i' \boldsymbol{\beta}_n + \epsilon_i, \quad 1 \leq i \leq n,$$

where $\boldsymbol{\beta}_n = (\beta_1, \cdots, \beta_{p_n})'$ is a p_n-dimensional vector of the unknown parameters, and ϵ_i's are independent and identically distributed with center 0 and variance σ^2.

Subscript n in p_n indicates that the number of coefficients may increase with the sample size n.

Model Selection & Estimation Problem

Candidate Full Model Estimation

A Great Deal of Redundancy in the Candidate Full Model

Too Many Nuisance Regression Parameters

Candidate Full Model is Sparse

Candidate Subspace – Candidate Submodel

Model Selection & Estimation Problem

Candidate Full Model Estimation

A Great Deal of Redundancy in the Candidate Full Model

Too Many Nuisance Regression Parameters

Candidate Full Model is Sparse

Candidate Subspace – Candidate Submodel

Model Selection & Estimation Problem

Candidate Full Model Estimation

A Great Deal of Redundancy in the Candidate Full Model

Too Many Nuisance Regression Parameters

Candidate Full Model is Sparse

Candidate Subspace – Candidate Submodel

Model Selection & Estimation Problem

Candidate Full Model Estimation

A Great Deal of Redundancy in the Candidate Full Model

Too Many Nuisance Regression Parameters

Candidate Full Model is Sparse

Candidate Subspace – Candidate Submodel

Model Selection & Estimation Problem

Candidate Full Model Estimation

A Great Deal of Redundancy in the Candidate Full Model

Too Many Nuisance Regression Parameters

Candidate Full Model is Sparse

Candidate Subspace – Candidate Submodel

Model Selection & Estimation Problem

Candidate Full Model Estimation

A Great Deal of Redundancy in the Candidate Full Model

Too Many Nuisance Regression Parameters

Candidate Full Model is Sparse

Candidate Subspace – Candidate Submodel

Model Selection & Estimation Problem

Candidate Full Model Estimation

A Great Deal of Redundancy in the Candidate Full Model

Too Many Nuisance Regression Parameters

Candidate Full Model is Sparse

Candidate Subspace – Candidate Submodel

Model Selection & Estimation Problem

We want to estimate β when it is plausible that β lie in the subspace

$$\mathbf{H}\beta = \mathbf{h}$$

- Human Eye: Uncertain Prior Information (UPI)
- Machine Eye: Auxiliary Information (AE)

$$\textit{UPI or AI}: \ \mathbf{H}\beta = \mathbf{h}$$

In many applications it is assumed that model is sparse, i.e. $\beta = (\beta_1', \beta_2')', \quad \beta_2 = \mathbf{0}$.

Model Selection & Estimation Problem

We want to estimate β when it is plausible that β lie in the subspace

$$\mathbf{H}\beta = \mathbf{h}$$

- Human Eye: Uncertain Prior Information (UPI)
- Machine Eye: Auxiliary Information (AE)

$$\textit{UPI or AI}: \ \mathbf{H}\beta = \mathbf{h}$$

In many applications it is assumed that model is sparse, i.e. $\beta = (\beta_1', \beta_2')', \quad \beta_2 = \mathbf{0}$.

Model Selection & Estimation Problem

We want to estimate β when it is plausible that β lie in the subspace

$$\mathbf{H}\beta = \mathbf{h}$$

- Human Eye: Uncertain Prior Information (UPI)
- Machine Eye: Auxiliary Information (AE)

$$\textit{UPI or AI}: \ \mathbf{H}\beta = \mathbf{h}$$

In many applications it is assumed that model is sparse, i.e. $\beta = (\beta_1', \beta_2')', \quad \beta_2 = \mathbf{0}.$

Model Selection & Estimation Problem

We want to estimate β when it is plausible that β lie in the subspace

$$\mathbf{H}\beta = \mathbf{h}$$

- Human Eye: Uncertain Prior Information (UPI)
- Machine Eye: Auxiliary Information (AE)

$$\textit{UPI or AI}: \ \mathbf{H}\beta = \mathbf{h}$$

In many applications it is assumed that model is sparse, i.e. $\beta = (\beta_1', \beta_2')', \quad \beta_2 = \mathbf{0}.$

Model Selection & Estimation Problem

We want to estimate β when it is plausible that β lie in the subspace

$$\mathbf{H}\beta = \mathbf{h}$$

- Human Eye: Uncertain Prior Information (UPI)
- Machine Eye: Auxiliary Information (AE)

$$UPI \ or \ AI: \ \mathbf{H}\beta = \mathbf{h}$$

In many applications it is assumed that model is sparse, i.e. $\beta = (\beta_1', \beta_2')', \quad \beta_2 = \mathbf{0}$.

Candidate Full Model Estimation

- Maximum Likelihood
- Least Square
- Ridge regression Or any other

A Revealing Tale of Overfitted Model

- Gauss offered two justifications for least squares: First, what we now call the maximum likelihood argument in the Gaussian error model. Second, the concept of risk and the start of what we now call the Gauss-Markov theorem.

- Stein's 1956 paper revealed that neither maximum likelihood estimators nor unbiased estimators have desirable risk functions when the dimension of the parameter space is not small.

Candidate Full Model Estimation

- Maximum Likelihood
- Least Square
- Ridge regression Or any other

A Revealing Tale of Overfitted Model

- Gauss offered two justifications for least squares: First, what we now call the maximum likelihood argument in the Gaussian error model. Second, the concept of risk and the start of what we now call the Gauss-Markov theorem.
- Stein's 1956 paper revealed that neither maximum likelihood estimators nor unbiased estimators have desirable risk functions when the dimension of the parameter space is not small.

Candidate Full Model Estimation

- Maximum Likelihood
- Least Square
- Ridge regression Or any other

A Revealing Tale of Overfitted Model

- Gauss offered two justifications for least squares: First, what we now call the maximum likelihood argument in the Gaussian error model. Second, the concept of risk and the start of what we now call the Gauss-Markov theorem.
- Stein's 1956 paper revealed that neither maximum likelihood estimators nor unbiased estimators have desirable risk functions when the dimension of the parameter space is not small.

Candidate Full Model Estimation

- Maximum Likelihood
- Least Square
- Ridge regression Or any other

A Revealing Tale of Overfitted Model

- Gauss offered two justifications for least squares: First, what we now call the maximum likelihood argument in the Gaussian error model. Second, the concept of risk and the start of what we now call the Gauss-Markov theorem.
- Stein's 1956 paper revealed that neither maximum likelihood estimators nor unbiased estimators have desirable risk functions when the dimension of the parameter space is not small.

Candidate Full Model Estimation

- Maximum Likelihood
- Least Square
- Ridge regression Or any other

A Revealing Tale of Overfitted Model

- Gauss offered two justifications for least squares: First, what we now call the maximum likelihood argument in the Gaussian error model. Second, the concept of risk and the start of what we now call the Gauss-Markov theorem.
- Stein's 1956 paper revealed that neither maximum likelihood estimators nor unbiased estimators have desirable risk functions when the dimension of the parameter space is not small.

Candidate Full Model Estimation

- Maximum Likelihood
- Least Square
- Ridge regression Or any other

A Revealing Tale of Overfitted Model

- Gauss offered two justifications for least squares: First, what we now call the maximum likelihood argument in the Gaussian error model. Second, the concept of risk and the start of what we now call the Gauss-Markov theorem.
- Stein's 1956 paper revealed that neither maximum likelihood estimators nor unbiased estimators have desirable risk functions when the dimension of the parameter space is not small.

Candidate Submodel Estimation

$$\hat{\beta}^{SM} = \hat{\beta}^{FM} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{H}'(\mathbf{H}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{H}')^{-1}(\mathbf{H}\hat{\beta}^{FM} - \mathbf{h}).$$

A Unrevealing Tale of Underfitted Model

Submodel Estimators are BIASED!!!

An interesting application of the restriction is that β can be partitioned as $\beta = (\beta_1', \beta_2')'$, if model is sparse, then $\beta_2 = \mathbf{0}$

Unbearable Truth about Submodel Estimation

$$E(\hat{\beta}_1) = \beta_1 - (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{X}_2\beta_2.$$

Clearly $\hat{\beta}_1$ is a biased estimator.

- unless the regression coefficients corresponding to deleted variables (β_2) are zero
- or the retained variables are orthogonal to the deleted variables,$\mathbf{X}_1'\mathbf{X}_2 = \mathbf{0}$
- Submodel estimates have smaller MSE than Full model estimates when the deleted regression variables have regression coefficients that are smaller than the standard errors of their estimates in full model.
- A naive data analyst may not comprehend that by dropping $\mathbf{X}_2$ from the model,S/he risk letting $\mathbf{X}_2\beta_2$ covertly influence the estimation and testing of β_1.

Pretest Estimation Strategy

The pretest estimator (PTE) of β based on $\hat{\beta}^{FM}$ and $\hat{\beta}^{SM}$ is defined as

$$\hat{\beta}^{PT} = \hat{\beta}^{FM} - (\hat{\beta}^{FM} - \hat{\beta}^{SM}) I(T_n \leq \chi^2_{p_2,\alpha}), \quad p_2 \geq 1,$$

$I(A)$ is an indicator function of a set A and $\chi^2_{p_2,\alpha}$ is the α-level

critical value of the distribution of T_n under H_0.

Shrinkage Estimation Strategy

$$\hat{\beta}^{S} = \hat{\beta}^{SM} + \left(1 - (p_2 - 2)\, T_n^{-1}\right)(\hat{\beta}^{FM} - \hat{\beta}^{SM}), \quad p_2 \geq 3,$$

Possible over-shrinking problem is defined as

$$\hat{\beta}^{S+} = \hat{\beta}^{SM} + \left(1 - (p_2 - 2)\, T_n^{-1}\right)^{+}(\hat{\beta}^{FM} - \hat{\beta}^{SM}),$$

where $z^{+} = max(0, z)$.

Shrinkage Estimation Strategy

$$\hat{\beta}^{S} = \hat{\beta}^{SM} + \left(1 - (p_2 - 2)\, T_n^{-1}\right)(\hat{\beta}^{FM} - \hat{\beta}^{SM}), \quad p_2 \geq 3,$$

Possible over-shrinking problem is defined as

$$\hat{\beta}^{S+} = \hat{\beta}^{SM} + \left(1 - (p_2 - 2)\, T_n^{-1}\right)^{+}(\hat{\beta}^{FM} - \hat{\beta}^{SM}),$$

where $z^{+} = max(0, z)$.

Shrinkage Estimation Strategy

$$\hat{\beta}^{S} = \hat{\beta}^{SM} + \left(1 - (p_2 - 2)T_n^{-1}\right)(\hat{\beta}^{FM} - \hat{\beta}^{SM}), \quad p_2 \geq 3,$$

Possible over-shrinking problem is defined as

$$\hat{\beta}^{S+} = \hat{\beta}^{SM} + \left(1 - (p_2 - 2)T_n^{-1}\right)^{+}(\hat{\beta}^{FM} - \hat{\beta}^{SM}),$$

where $z^{+} = max(0, z)$.

Executive Summary

- Bancroft (1944) suggested two problems on preliminary test strategy.
  - Data pooling problem based on a pretest. This stream followed by a host of researchers.
  - Model misspecification problem in linear regression model based on a pretest.
- Stein (1956, 1961) developed highly efficient shrinkage estimators in balanced designs. Most statisticians have ignored these (perhaps due to lack of understanding)
- Modern regularization estimation strategies based on penalized least squares with penalties extend Stein's procedures powerfully.

Executive Summary

- Bancroft (1944) suggested two problems on preliminary test strategy.
    - Data pooling problem based on a pretest. This stream followed by a host of researchers.
    - Model misspecification problem in linear regression model based on a pretest.
- Stein (1956, 1961) developed highly efficient shrinkage estimators in balanced designs. Most statisticians have ignored these (perhaps due to lack of understanding)
- Modern regularization estimation strategies based on penalized least squares with penalties extend Stein's procedures powerfully.

Executive Summary

- Bancroft (1944) suggested two problems on preliminary test strategy.
  - Data pooling problem based on a pretest. This stream followed by a host of researchers.
  - Model misspecification problem in linear regression model based on a pretest.
- Stein (1956, 1961) developed highly efficient shrinkage estimators in balanced designs. Most statisticians have ignored these (perhaps due to lack of understanding)
- Modern regularization estimation strategies based on penalized least squares with penalties extend Stein's procedures powerfully.

Executive Summary

- Bancroft (1944) suggested two problems on preliminary test strategy.
  - Data pooling problem based on a pretest. This stream followed by a host of researchers.
  - Model misspecification problem in linear regression model based on a pretest.
- Stein (1956, 1961) developed highly efficient shrinkage estimators in balanced designs. Most statisticians have ignored these (perhaps due to lack of understanding)
- Modern regularization estimation strategies based on penalized least squares with penalties extend Stein's procedures powerfully.

Executive Summary

- Bancroft (1944) suggested two problems on preliminary test strategy.
  - Data pooling problem based on a pretest. This stream followed by a host of researchers.
  - Model misspecification problem in linear regression model based on a pretest.
- Stein (1956, 1961) developed highly efficient shrinkage estimators in balanced designs. Most statisticians have ignored these (perhaps due to lack of understanding)
- Modern regularization estimation strategies based on penalized least squares with penalties extend Stein's procedures powerfully.

Executive Summary

- Bancroft (1944) suggested two problems on preliminary test strategy.
  - Data pooling problem based on a pretest. This stream followed by a host of researchers.
  - Model misspecification problem in linear regression model based on a pretest.
- Stein (1956, 1961) developed highly efficient shrinkage estimators in balanced designs. Most statisticians have ignored these (perhaps due to lack of understanding)
- Modern regularization estimation strategies based on penalized least squares with penalties extend Stein's procedures powerfully.

Penalty Estimation Strategy

- The penalty estimators are members of the penalized least squares (PLS) family and they are obtained by optimizing a quadratic function subject to a penalty.
- A popular version of the PLS is given by Tikhonov (1963) regularization.
- A generalized version of penalty estimator is the bridge regression (Frank and Friedman,1993).

Penalty Estimation Strategy

- The penalty estimators are members of the penalized least squares (PLS) family and they are obtained by optimizing a quadratic function subject to a penalty.
- A popular version of the PLS is given by Tikhonov (1963) regularization.
- A generalized version of penalty estimator is the bridge regression (Frank and Friedman,1993).

Penalty Estimation Strategy

- The penalty estimators are members of the penalized least squares (PLS) family and they are obtained by optimizing a quadratic function subject to a penalty.
- A popular version of the PLS is given by Tikhonov (1963) regularization.
- A generalized version of penalty estimator is the bridge regression (Frank and Friedman,1993).

Penalty Estimation Strategy

- The penalty estimators are members of the penalized least squares (PLS) family and they are obtained by optimizing a quadratic function subject to a penalty.
- A popular version of the PLS is given by Tikhonov (1963) regularization.
- A generalized version of penalty estimator is the bridge regression (Frank and Friedman,1993).

Penalty Estimation Strategy

- For a given penalty function $\pi(\cdot)$ and regularization parameter λ, the general form of the objective function can be written as

$$\phi(\beta) = (\boldsymbol{y} - \boldsymbol{X}\beta)^T(\boldsymbol{y} - \boldsymbol{X}\beta) + \lambda\pi(\beta),$$

- Penalty function is of the form

$$\pi(\beta) = \sum_{j=1}^{p} |\beta_j|^\gamma, \ \gamma > 0. \tag{1}$$

Penalty Estimation Strategy

- For a given penalty function $\pi(\cdot)$ and regularization parameter λ, the general form of the objective function can be written as

$$\phi(\boldsymbol{\beta}) = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) + \lambda\pi(\boldsymbol{\beta}),$$

- Penalty function is of the form

$$\pi(\boldsymbol{\beta}) = \sum_{j=1}^{p} |\beta_j|^\gamma, \ \gamma > 0. \tag{1}$$

Penalty Estimation Strategy

- For a given penalty function $\pi(\cdot)$ and regularization parameter λ, the general form of the objective function can be written as

$$\phi(\boldsymbol{\beta}) = (\boldsymbol{y} - \boldsymbol{X}\beta)^T(\boldsymbol{y} - \boldsymbol{X}\beta) + \lambda\pi(\boldsymbol{\beta}),$$

- Penalty function is of the form

$$\pi(\boldsymbol{\beta}) = \sum_{j=1}^{p} |\beta_j|^\gamma, \ \gamma > 0. \tag{1}$$

Penalty Estimation Strategy

For $\gamma = 2$, we have ridge estimates which are obtained by minimizing the penalized residual sum of squares

$$\hat{\beta}^{\text{ridge}} = \arg\min_{\beta} \left|\left| \boldsymbol{y} - \sum_{j=1}^{p} \boldsymbol{X}_j \beta_j \right|\right|^2 + \lambda \sum_{j=1}^{p} ||\beta_j||^2, \qquad (2)$$

λ is the tuning parameter which controls the amount of shrinkage and $||\cdot|| = ||\cdot||_2$ is the L_2 norm.

Penalty Estimation Strategy

- For $\gamma < 2$, it shrinks the coefficient towards zero, and depending on the value of λ, it sets some of the coefficients to exactly zero.
- The procedure combines variable selection and shrinking of the coefficients of a penalized regression.
- An important member of the penalized least squares family is the L_1 penalized least squares estimator, which is obtained when $\gamma = 1$.
- This is known as the Least Absolute Shrinkage and Selection Operator (LASSO): Tibshirani(1996)

Penalty Estimation Strategy

- For $\gamma < 2$, it shrinks the coefficient towards zero, and depending on the value of λ, it sets some of the coefficients to exactly zero.
- The procedure combines variable selection and shrinking of the coefficients of a penalized regression.
- An important member of the penalized least squares family is the L_1 penalized least squares estimator, which is obtained when $\gamma = 1$.
- This is known as the Least Absolute Shrinkage and Selection Operator (LASSO): Tibshirani(1996)

Penalty Estimation Strategy

- For $\gamma < 2$, it shrinks the coefficient towards zero, and depending on the value of λ, it sets some of the coefficients to exactly zero.
- The procedure combines variable selection and shrinking of the coefficients of a penalized regression.
- An important member of the penalized least squares family is the L_1 penalized least squares estimator, which is obtained when $\gamma = 1$.
- This is known as the Least Absolute Shrinkage and Selection Operator (LASSO): Tibshirani(1996)

Penalty Estimation Strategy

- For $\gamma < 2$, it shrinks the coefficient towards zero, and depending on the value of λ, it sets some of the coefficients to exactly zero.
- The procedure combines variable selection and shrinking of the coefficients of a penalized regression.
- An important member of the penalized least squares family is the L_1 penalized least squares estimator, which is obtained when $\gamma = 1$.
- This is known as the Least Absolute Shrinkage and Selection Operator (LASSO): Tibshirani(1996)

Penalty Estimation Strategy

- For $\gamma < 2$, it shrinks the coefficient towards zero, and depending on the value of λ, it sets some of the coefficients to exactly zero.
- The procedure combines variable selection and shrinking of the coefficients of a penalized regression.
- An important member of the penalized least squares family is the L_1 penalized least squares estimator, which is obtained when $\gamma = 1$.
- This is known as the Least Absolute Shrinkage and Selection Operator (LASSO): Tibshirani(1996)

Penalty Estimation Strategy

- LASSO is closely related to the ridge regression and its solutions are similarly obtained by replacing the squared penalty $||\beta_j||^2$ in the ridge solution (3) with the absolute penalty $||\beta_j||_1$ in the LASSO–

$$\hat{\beta}^{\text{LASSO}} = \arg\min_{\beta} \left|\left| \boldsymbol{y} - \sum_{j=1}^{p} \boldsymbol{X}_j \beta_j \right|\right|^2 + \lambda \sum_{j=1}^{p} ||\beta_j||_1. \qquad (3)$$

Good Strategy if Model is **Truly** Sparse

Penalty Estimation Strategy

- LASSO is closely related to the ridge regression and its solutions are similarly obtained by replacing the squared penalty $||\beta_j||^2$ in the ridge solution (3) with the absolute penalty $||\beta_j||_1$ in the LASSO–

$$\hat{\beta}^{\text{LASSO}} = \arg\min_{\beta} \left|\left| \boldsymbol{y} - \sum_{j=1}^{p} \boldsymbol{X}_j \beta_j \right|\right|^2 + \lambda \sum_{j=1}^{p} ||\beta_j||_1. \quad (3)$$

Good Strategy if Model is **Truly** Sparse

Penalty Estimation Family Ever Growing!!

Adaptive LASSO (aLASSO)

Elastic Net Penalty

Minimax Concave Penalty (MCP)

SCAD

Innate Difficulties: Can Signals be Septated from Noise?

- All penalty estimators may not provide an estimator with both estimation consistency and variable selection consistency simultaneously.
- aLASSO, SCAD, and MCP are Oracle (asymptoticaly).
- Asymptotic properties are based on assumptions on both true model and designed covariates.
- Sparsity in the model (most coefficients are exactly 0), few are not
- Nonzero coefficients are big enough to to be separated from zero ones.

Innate Difficulties: Can Signals be Septated from Noise?

- All penalty estimators may not provide an estimator with both estimation consistency and variable selection consistency simultaneously.
- aLASSO, SCAD, and MCP are Oracle (asymptoticaly).
- Asymptotic properties are based on assumptions on both true model and designed covariates.
- Sparsity in the model (most coefficients are exactly 0), few are not
- Nonzero coefficients are big enough to to be separated from zero ones.

Innate Difficulties: Can Signals be Septated from Noise?

- All penalty estimators may not provide an estimator with both estimation consistency and variable selection consistency simultaneously.
- aLASSO, SCAD, and MCP are Oracle (asymptoticaly).
- Asymptotic properties are based on assumptions on both true model and designed covariates.
- Sparsity in the model (most coefficients are exactly 0), few are not
- Nonzero coefficients are big enough to to be separated from zero ones.

Innate Difficulties: Can Signals be Septated from Noise?

- All penalty estimators may not provide an estimator with both estimation consistency and variable selection consistency simultaneously.
- aLASSO, SCAD, and MCP are Oracle (asymptoticaly).
- Asymptotic properties are based on assumptions on both true model and designed covariates.
- Sparsity in the model (most coefficients are exactly 0), few are not
- Nonzero coefficients are big enough to to be separated from zero ones.

Innate Difficulties: Can Signals be Septated from Noise?

- All penalty estimators may not provide an estimator with both estimation consistency and variable selection consistency simultaneously.
- aLASSO, SCAD, and MCP are Oracle (asymptoticaly).
- Asymptotic properties are based on assumptions on both true model and designed covariates.
- Sparsity in the model (most coefficients are exactly 0), few are not
- Nonzero coefficients are big enough to to be separated from zero ones.

Innate Difficulties: Can Signals be Septated from Noise?

- All penalty estimators may not provide an estimator with both estimation consistency and variable selection consistency simultaneously.
- aLASSO, SCAD, and MCP are Oracle (asymptoticaly).
- Asymptotic properties are based on assumptions on both true model and designed covariates.
- Sparsity in the model (most coefficients are exactly 0), few are not
- Nonzero coefficients are big enough to to be separated from zero ones.

Innate Difficulties: Ultrahigh Dimensional Features

- In genetic micro-array studies, n is measured in hundreds, the number of features p per sample can exceed millions!!!
- 22,500 unique proteins, implies about 253,000,000 possible protein-protein interactions. So far 42,000 are identified.
- penalty estimators may not be efficient when the dimension p becomes extremely large compared with sample size n.
- There are still challenging problems when p grows at a non-polynomial rate with n.
- Non-polynomial dimensionality poses substantial computational challenges.
- The developments in the arena of penalty estimation is still infancy.

Innate Difficulties: Ultrahigh Dimensional Features

- In genetic micro-array studies, n is measured in hundreds, the number of features p per sample can exceed millions!!!
- 22,500 unique proteins, implies about 253,000,000 possible protein-protein interactions. So far 42,000 are identified.
- penalty estimators may not be efficient when the dimension p becomes extremely large compared with sample size n.
- There are still challenging problems when p grows at a non-polynomial rate with n.
- Non-polynomial dimensionality poses substantial computational challenges.
- The developments in the arena of penalty estimation is still infancy.

Innate Difficulties: Ultrahigh Dimensional Features

- In genetic micro-array studies, n is measured in hundreds, the number of features p per sample can exceed millions!!!
- 22,500 unique proteins, implies about 253,000,000 possible protein-protein interactions. So far 42,000 are identified.
- penalty estimators may not be efficient when the dimension p becomes extremely large compared with sample size n.
- There are still challenging problems when p grows at a non-polynomial rate with n.
- Non-polynomial dimensionality poses substantial computational challenges.
- The developments in the arena of penalty estimation is still infancy.

Innate Difficulties: Ultrahigh Dimensional Features

- In genetic micro-array studies, n is measured in hundreds, the number of features p per sample can exceed millions!!!
- 22,500 unique proteins, implies about 253,000,000 possible protein-protein interactions. So far 42,000 are identified.
- penalty estimators may not be efficient when the dimension p becomes extremely large compared with sample size n.
- There are still challenging problems when p grows at a non-polynomial rate with n.
- Non-polynomial dimensionality poses substantial computational challenges.
- The developments in the arena of penalty estimation is still infancy.

Innate Difficulties: Ultrahigh Dimensional Features

- In genetic micro-array studies, n is measured in hundreds, the number of features p per sample can exceed millions!!!
- 22,500 unique proteins, implies about 253,000,000 possible protein-protein interactions. So far 42,000 are identified.
- penalty estimators may not be efficient when the dimension p becomes extremely large compared with sample size n.
- There are still challenging problems when p grows at a non-polynomial rate with n.
- Non-polynomial dimensionality poses substantial computational challenges.
- The developments in the arena of penalty estimation is still infancy.

Innate Difficulties: Ultrahigh Dimensional Features

- In genetic micro-array studies, n is measured in hundreds, the number of features p per sample can exceed millions!!!
- 22,500 unique proteins, implies about 253,000,000 possible protein-protein interactions. So far 42,000 are identified.
- penalty estimators may not be efficient when the dimension p becomes extremely large compared with sample size n.
- There are still challenging problems when p grows at a non-polynomial rate with n.
- Non-polynomial dimensionality poses substantial computational challenges.
- The developments in the arena of penalty estimation is still infancy.

Innate Difficulties: Ultrahigh Dimensional Features

- In genetic micro-array studies, n is measured in hundreds, the number of features p per sample can exceed millions!!!
- 22,500 unique proteins, implies about 253,000,000 possible protein-protein interactions. So far 42,000 are identified.
- penalty estimators may not be efficient when the dimension p becomes extremely large compared with sample size n.
- There are still challenging problems when p grows at a non-polynomial rate with n.
- Non-polynomial dimensionality poses substantial computational challenges.
- The developments in the arena of penalty estimation is still infancy.

Can Noise be Septated from Signals?

Pretest and Shrinkage Strategies are Useful in this Situation

Extension and Comparison with non-penalty Estimators

- Ahmed et al. (2008, 2009) for partially linear models.
- Fallahpour, Ahmed and Doksum (2010) and Ahmed and Fallahpour (2014)for partially linear models with Random Coefficient autoregressive Errors.
- Ahmed and Fallahpour (2012) for Quasi-likelihood models.
- Ahmed et al. (2012) for Weibull censored regression models.

Can Noise be Septated from Signals?

Pretest and Shrinkage Strategies are Useful in this Situation

Extension and Comparison with non-penalty Estimators

- Ahmed et al. (2008, 2009) for partially linear models.
- Fallahpour, Ahmed and Doksum (2010) and Ahmed and Fallahpour (2014)for partially linear models with Random Coefficient autoregressive Errors.
- Ahmed and Fallahpour (2012) for Quasi-likelihood models.
- Ahmed et al. (2012) for Weibull censored regression models.

Can Noise be Septated from Signals?

Pretest and Shrinkage Strategies are Useful in this Situation

Extension and Comparison with non-penalty Estimators

- Ahmed et al. (2008, 2009) for partially linear models.
- Fallahpour, Ahmed and Doksum (2010) and Ahmed and Fallahpour (2014)for partially linear models with Random Coefficient autoregressive Errors.
- Ahmed and Fallahpour (2012) for Quasi-likelihood models.
- Ahmed et al. (2012) for Weibull censored regression models.

Can Noise be Septated from Signals?

Pretest and Shrinkage Strategies are Useful in this Situation

Extension and Comparison with non-penalty Estimators

- Ahmed et al. (2008, 2009) for partially linear models.
- Fallahpour, Ahmed and Doksum (2010) and Ahmed and Fallahpour (2014)for partially linear models with Random Coefficient autoregressive Errors.
- Ahmed and Fallahpour (2012) for Quasi-likelihood models.
- Ahmed et al. (2012) for Weibull censored regression models.

Can Noise be Septated from Signals?

Pretest and Shrinkage Strategies are Useful in this Situation

Extension and Comparison with non-penalty Estimators

- Ahmed et al. (2008, 2009) for partially linear models.
- Fallahpour, Ahmed and Doksum (2010) and Ahmed and Fallahpour (2014)for partially linear models with Random Coefficient autoregressive Errors.
- Ahmed and Fallahpour (2012) for Quasi-likelihood models.
- Ahmed et al. (2012) for Weibull censored regression models.

Can Noise be Septated from Signals?

Pretest and Shrinkage Strategies are Useful in this Situation

Extension and Comparison with non-penalty Estimators

- Ahmed et al. (2008, 2009) for partially linear models.
- Fallahpour, Ahmed and Doksum (2010) and Ahmed and Fallahpour (2014)for partially linear models with Random Coefficient autoregressive Errors.
- Ahmed and Fallahpour (2012) for Quasi-likelihood models.
- Ahmed et al. (2012) for Weibull censored regression models.

Extension and Comparison with non-penalty Estimators

- S. E. Ahmed (2014). *Penalty, Pretest and Shrinkage Estimation: Variable Selection and Estimation*. Springer.
- S. E. Ahmed (Editor). *Perspectives on Big Data Analysis: Methodologies and Applications. To be published by Contemporary Mathematics, a co-publication of American Mathematical Society and CRM, 2014*.

Extension and Comparison with non-penalty Estimators

- S. E. Ahmed (2014). *Penalty, Pretest and Shrinkage Estimation: Variable Selection and Estimation*. Springer.
- S. E. Ahmed (Editor). *Perspectives on Big Data Analysis: Methodologies and Applications. To be published by Contemporary Mathematics, a co-publication of American Mathematical Society and CRM, 2014*.

Extension and Comparison with non-penalty Estimators

- S. E. Ahmed (2014). *Penalty, Pretest and Shrinkage Estimation: Variable Selection and Estimation*. Springer.
- S. E. Ahmed (Editor). *Perspectives on Big Data Analysis: Methodologies and Applications. To be published by Contemporary Mathematics, a co-publication of American Mathematical Society and CRM, 2014*.

Extension and Comparison with non-penalty Estimators

- S. E. Ahmed (2014). *Penalty, Pretest and Shrinkage Estimation: Variable Selection and Estimation*. Springer.
- S. E. Ahmed (Editor). *Perspectives on Big Data Analysis: Methodologies and Applications. To be published by Contemporary Mathematics, a co-publication of American Mathematical Society and CRM, 2014*.

Shrinkage Estimation for Big Data

- The classical shrinkage estimation methods are limited to fixed p.
- The asymptotic results depend heavily on a maximum likelihood full estimation with component-wise consistency at rate of $\sqrt{n}$.
- When $p_n > n$, a component-wise consistent estimator of β_n is not available since β_n is not identifiable.
- Here β_n is not identifiable in the sense that there always exist two different estimations of β_n, $\beta_n^{(1)}$ and $\beta_n^{(2)}$, such that $\mathbf{x}_i'\beta_n^{(1)} = \mathbf{x}_i'\beta_n^{(2)}$ for $1 \leq i \leq n$.

Shrinkage Estimation for Big Data

- The classical shrinkage estimation methods are limited to fixed p.
- The asymptotic results depend heavily on a maximum likelihood full estimation with component-wise consistency at rate of $\sqrt{n}$.
- When $p_n > n$, a component-wise consistent estimator of β_n is not available since β_n is not identifiable.
- Here β_n is not identifiable in the sense that there always exist two different estimations of β_n, $\beta_n^{(1)}$ and $\beta_n^{(2)}$, such that $\mathbf{x}_i'\beta_n^{(1)} = \mathbf{x}_i'\beta_n^{(2)}$ for $1 \leq i \leq n$.

Shrinkage Estimation for Big Data

- The classical shrinkage estimation methods are limited to fixed p.
- The asymptotic results depend heavily on a maximum likelihood full estimation with component-wise consistency at rate of $\sqrt{n}$.
- When $p_n > n$, a component-wise consistent estimator of β_n is not available since β_n is not identifiable.
- Here β_n is not identifiable in the sense that there always exist two different estimations of β_n, $\beta_n^{(1)}$ and $\beta_n^{(2)}$, such that $\mathbf{x}_i'\beta_n^{(1)} = \mathbf{x}_i'\beta_n^{(2)}$ for $1 \leq i \leq n$.

Shrinkage Estimation for Big Data

- The classical shrinkage estimation methods are limited to fixed p.
- The asymptotic results depend heavily on a maximum likelihood full estimation with component-wise consistency at rate of $\sqrt{n}$.
- When $p_n > n$, a component-wise consistent estimator of β_n is not available since β_n is not identifiable.
- Here β_n is not identifiable in the sense that there always exist two different estimations of β_n, $\beta_n^{(1)}$ and $\beta_n^{(2)}$, such that $\mathbf{x}_i'\beta_n^{(1)} = \mathbf{x}_i'\beta_n^{(2)}$ for $1 \leq i \leq n$.

Shrinkage Estimation for Big Data

- The classical shrinkage estimation methods are limited to fixed p.
- The asymptotic results depend heavily on a maximum likelihood full estimation with component-wise consistency at rate of $\sqrt{n}$.
- When $p_n > n$, a component-wise consistent estimator of $\boldsymbol{\beta}_n$ is not available since $\boldsymbol{\beta}_n$ is not identifiable.
- Here $\boldsymbol{\beta}_n$ is not identifiable in the sense that there always exist two different estimations of $\boldsymbol{\beta}_n$, $\boldsymbol{\beta}_n^{(1)}$ and $\boldsymbol{\beta}_n^{(2)}$, such that $\mathbf{x}_i'\boldsymbol{\beta}_n^{(1)} = \mathbf{x}_i'\boldsymbol{\beta}_n^{(2)}$ for $1 \leq i \leq n$.

Shrinkage Estimation for Big Data

- we write the p_n-dimensional coefficients vector $\boldsymbol{\beta}_n = (\boldsymbol{\beta}_{1n}', \boldsymbol{\beta}_{2n}')'$,, where $\boldsymbol{\beta}_{1n}$ is the coefficient vector for main covariates, $\boldsymbol{\beta}_{2n}$ include all nuisance parameters.
- Sub-vectors $\boldsymbol{\beta}_{1n}$, $\boldsymbol{\beta}_{2n}$, have dimensions p_{1n}, p_{2n}, respectively, where $p_{1n} \leq n$ and $p_{1n} + p_{2n} = p_n$.
- Let $\mathbf{X}_{1n}$ and $\mathbf{X}_{2n}$ be the sub-matrices of $\mathbf{X}_n$ corresponding to $\boldsymbol{\beta}_{1n}$ and $\boldsymbol{\beta}_{2n}$, respectively.
- Let us assume true parameter vector

$$\boldsymbol{\beta}_0 = (\beta_{01}, \cdots, \beta_{0p_n})' = (\boldsymbol{\beta}_{10}', \boldsymbol{\beta}_{20}')'.$$

- Let S_{10} and S_{20} represent the corresponding index sets for $\boldsymbol{\beta}_{10}$ and $\boldsymbol{\beta}_{20}$, respectively.
- Specifically, S_{10} includes important predictors and S_{20} includes sparse and weak signals satisfying the following assumption.

(A0) $|\beta_{0j}| = O(n^{-\varsigma})$, for $\forall j \in S_{20}$, where $\varsigma > 1/2$ does not change with n.

- Condition (A0) is considered to be the sparsity of the model. A simpler representation for the finite sample is that $\beta_{0j} = 0 \; \forall j \in S_{20}$, that is, most coefficients are 0 exactly.

A Class of Submodels

- Predictors indexed by S_{10} are used to construct a submodel.
- However, other predictors, especially ones in S_{20} may also make some contributions to the response and cannot be ignored.

Consider

$$\text{UPI or AI}: \quad (\beta'_{20})' = \mathbf{0}_{p_{2n}}.$$

A Candidate Submodel Estimator

We make the following assumptions on the random error and design matrix of the true model:

(A1) The random error ϵ_i's are independent and identically distributed with mean 0 and variance $0 < \sigma^2 < \infty$. Further, $E(\epsilon_i^m) < \infty$, for an even integer m not depending on n.

(A2) $\rho_{1n} > 0$, for all n, the smallest eigenvalue of $\mathbf{C}_{12n}$

Under (A1-A2) and UPI/AE, the submodel estimator (SME) of $\boldsymbol{\beta}_{1n}$ is defined as

$$\hat{\boldsymbol{\beta}}_{1n}^{SM} = (\mathbf{X}_{1n}'\mathbf{X}_{1n})^{-1}\mathbf{X}_{1n}'\mathbf{y}.$$

A Candidate Full Model Estimator

Weighted Ridge Estimation

We estimate an estimator of β_n by minimizing a partial penalized objective function,

$$\hat{\beta}(r_n) = argmin\{\|\mathbf{y} - \mathbf{X}_{1n}\beta_{1n} - \mathbf{X}_{2n}\beta_{2n}\|^2 + r_n\|\beta_{2n}\|^2\}$$

where "$\| \cdot \|$" is the ℓ_2 norm and $r_n > 0$ is a tuning parameter.

Weighted Ridge Estimation

Since $p_n >> n$ and under the sparsity assumption
Define

$$a_n = c_1 n^{-\omega}, \quad 0 < \omega \leq 1/2,\ c_1 > 0.$$

We define a weighted ridge estimator of β_n is denoted as

$$\hat{\beta}_n^{WR}(r_n, a_n) = \begin{pmatrix} \hat{\beta}_{1n}^{WR}(r_n) \\ \hat{\beta}_{2n}^{WR}(r_n, a_n) \end{pmatrix}, \text{ where}$$

$$\hat{\beta}_{1n}^{WR}(r_n) = \hat{\beta}_{1n}(r_n)$$

and for $j \notin S_{10}$,

$$\hat{\beta}_j^{\mathrm{WR}}(r_n, a_n) = \begin{cases} \hat{\beta}_j(r_n, a_n), & \hat{\beta}_j(r_n, a_n) > a_n; \\ 0, & \text{otherwise.} \end{cases}$$

Weighted Ridge Estimation

- We call $\hat{\beta}(r_n, a_n)$ as a weighted ridge estimator from two aspects.
- We use a weighted ridge instead of ridge penalty for the HD shrinkage estimation strategy since we do not want to generate some additional biases caused by an additional penalty on β_{1n} if we already have a candidate subset model.
- Here $\hat{\beta}_{1n}^{WR}(r_n)$ changes with r_n and $\hat{\beta}_{2n}^{WR}(r_n, a_n)$ changes with both r_n and a_n.
- For the notation's convenience, we denote the weighted ridge estimators as $\hat{\beta}_{1n}^{WR}$ and $\hat{\beta}_{2n}^{WR}$.

A Candidate HD Shrinkage Estimator

A HD shrinkage estimators (HD-SE) $\hat{\beta}_{1n}^{S}$ is

$$\hat{\beta}_{1n}^{S} = \hat{\beta}_{1n}^{WR} - (h-2)T_n^{-1}(\hat{\beta}_{1n}^{WR} - \hat{\beta}_{1n}^{SM}),$$

$h > 2$ is the number of nonzero elements in $\hat{\beta}_{2n}^{WR}$

$$T_n = (\hat{\beta}_2^{WR})'(\mathbf{X}_2'\mathbf{M}_1\mathbf{X}_2)\hat{\beta}_2^{WR}/\hat{\sigma}^2, \tag{4}$$

$$\mathbf{M}_1 = \mathbf{I}_n - \mathbf{X}_{1n}(\mathbf{X}_{1n}'\mathbf{X}_{1n})^{-1}\mathbf{X}_{1n}'$$

- $\hat{\sigma}^2$ is a consistent estimator of σ^2.
- For example, we can choose $\hat{\sigma}^2 = \sum_{i=1}^{n}(y_i - \mathbf{x}_i'\hat{\beta}^{SM})^2/(n-1)$ under UPI or AI.

A Candidate HD Positive Shrinkage Estimator

A HD positive shrinkage estimator (HD-PSE),

$$\hat{\beta}_{1n}^{PSE} = \hat{\beta}_{1n}^{WR} - ((h-2)T_n^{-1})_1(\hat{\beta}_{1n}^{WR} - \hat{\beta}_{1n}^{SM}),$$

where $(a)_1 = 1$ and a for $a > 1$ and $a \leq 1$, respectively.

Consistency and Asymptotic Normality

Weighted Ridge Estimation

Let $s_n^2 = \sigma^2 \mathbf{d}_n' \Sigma_n^{-1} \mathbf{d}_n$ for any $p_{12n} \times 1$ vector $\mathbf{d}_n$ satisfying $\|\mathbf{d}_n\| \leq 1$.

$$n^{1/2} s_n^{-1} \mathbf{d}_n' (\hat{\beta}_{12n}^{WR} - \beta_{120}) = n^{-1/2} s_n^{-1} \sum_{i=1}^{n} \epsilon_i \mathbf{d}_n' \Sigma_n^{-1} \mathbf{z}_i + o_P(1)$$

$$\xrightarrow{\mathrm{d}} N(0, 1).$$

Define

$$\Sigma_{n11} = \lim_{n\to\infty} \mathbf{X}_{1n}'\mathbf{X}_{1n}/n, \quad \Sigma_{n22} = \lim_{n\to\infty} \mathbf{X}_{2n}'\mathbf{X}_{2n}/n,$$
$$\Sigma_{n12} = \lim_{n\to\infty} \mathbf{X}_{1n}'\mathbf{X}_{2n}/n, \quad \Sigma_{n21} = \lim_{n\to\infty} \mathbf{X}_{2n}'\mathbf{X}_{1n}/n,$$
$$\Sigma_{n22.1} = \lim_{n\to\infty} n^{-1}\mathbf{X}_{2n}'\mathbf{X}_{2n} - \mathbf{X}_{2n}'\mathbf{X}_{1n}(\mathbf{X}_{1n}'\mathbf{X}_{1n})^{-1}\mathbf{X}_{1n}'\mathbf{X}_{2n}$$
$$\Sigma_{n11.2} = \lim_{n\to\infty} n^{-1}\mathbf{X}_{1n}'\mathbf{X}_{1n} - \mathbf{X}_{1n}'\mathbf{X}_{2n}(\mathbf{X}_{2n}'\mathbf{X}_{2n})^{-1}\mathbf{X}_{2n}'\mathbf{X}_{1n}$$

Asymptotic Distributional Risk (ADR)

$$K_n : \boldsymbol{\beta}_{20} = n^{-1/2}\boldsymbol{\delta} \quad \text{and} \quad \boldsymbol{\beta}_{30} = \mathbf{0}_{p_{3n}},$$

$$\boldsymbol{\delta} = (\delta_1, \delta_2, \cdots, \delta_{p_{2n}})' \in \Re^{p_{2n}}, \delta_j \quad \text{is fixed.}$$

- Define $\Delta_n = \boldsymbol{\delta}'\boldsymbol{\Sigma}_{n22.1}\boldsymbol{\delta}$,
- $n^{1/2}\mathbf{d}'_{1n}s^{-1}_{1n}(\boldsymbol{\beta}^*_{1n} - \boldsymbol{\beta}_{10})$ is asymptotically normal under $\{K_n\}$, where $s^2_{1n} = \sigma^2\mathbf{d}'_{1n}\boldsymbol{\Sigma}^{-1}_{n11.2}\mathbf{d}_{1n}$.
- The asymptotic distributional risk (ADR) of $\mathbf{d}'_{1n}\boldsymbol{\beta}^*_{1n}$ is

$$\mathrm{ADR}(\mathbf{d}'_{1n}\boldsymbol{\beta}^*_{1n}) = \lim_{n\to\infty} E\{[n^{1/2}s^{-1}_{1n}\mathbf{d}'_{1n}(\boldsymbol{\beta}^*_{1n} - \boldsymbol{\beta}_{10})]^2\}.$$

Asymptotic Distributional Risk

Mathematical Proof

Under regularity conditions and K_n, and suppose there exists $0 \le c \le 1$ such that $c = \lim_{n\to\infty} s_{1n}^{-2}\mathbf{d}'_{1n}\Sigma_{n11}^{-1}\mathbf{d}_{1n}$, we have

$$\mathrm{ADR}(\mathbf{d}'_{1n}\hat{\beta}_{1n}^{WR}) = 1, \tag{5a}$$

$$\mathrm{ADR}(\mathbf{d}'_{1n}\hat{\beta}_{1n}^{SM}) = 1 - (1-c)(1-\Delta_{\mathbf{d}_{1n}}), \tag{5b}$$

$$\mathrm{ADR}(\mathbf{d}'_{1n}\hat{\beta}_{1n}^{S}) = 1 - E[g_1(\mathbf{z}_2 + \delta)], \tag{5c}$$

$$\mathrm{ADR}(\mathbf{d}'_{1n}\hat{\beta}_{1n}^{PSE}) = 1 - E[g_2(\mathbf{z}_2 + \delta)], \tag{5d}$$

$$\Delta_{\mathbf{d}_{1n}} = \frac{\mathbf{d}'_{1n}(\Sigma_{n11}^{-1}\Sigma_{n12}\delta\delta'\Sigma_{n21}\Sigma_{n11}^{-1})\mathbf{d}_{1n}}{\mathbf{d}'_{1n}(\Sigma_{n11}^{-1}\Sigma_{n12}\Sigma_{n22.1}^{-1}\Sigma_{n21}\Sigma_{n11}^{-1})\mathbf{d}_{1n}}.$$

$$s_{2n}^{-1}\mathbf{d}'_{2n}\mathbf{z}_2 \to N(0,1)$$

$$\mathbf{d}_{2n} = \Sigma_{n21}\Sigma_{n11}^{-1}\mathbf{d}_{1n}$$

$$s_{2n}^{2} = \mathbf{d}'_{2n}\Sigma_{n22.1}^{-1}\mathbf{d}_{2n}$$

Mathematical Proof

$$g_1(\mathbf{x}) = \lim_{n\to\infty}(1-c)\frac{p_{2n}-2}{\mathbf{x}'\Sigma_{n22.1}\mathbf{x}}\left[2-\frac{\mathbf{x}'((p_{2n}+2)\mathbf{d}_{2n}\mathbf{d}_{2n}')\mathbf{x}}{s_{2n}^2\mathbf{x}'\Sigma_{n22.1}\mathbf{x}}\right],$$

$$\begin{aligned} g_2(\mathbf{x}) = \quad & \lim\nolimits_{n\to\infty}\frac{p_{2n}-2}{\mathbf{x}'\Sigma_{n22.1}\mathbf{x}}\left[(1-c)\left(2-\frac{\mathbf{x}'((p_{2n}+2)\mathbf{d}_{2n}\mathbf{d}_{2n}')\mathbf{x}}{s_{2n}^2\mathbf{x}'\Sigma_{n22.1}\mathbf{x}}\right)\right] \\ & I(\mathbf{x}'\Sigma_{n22.1}\mathbf{x} \geq p_{2n}-2) \\ & +\lim\nolimits_{n\to\infty}[(2-s_{2n}^{-2}\mathbf{x}'\delta_{2n}\delta_{2n}'\mathbf{x})(1-c)]I(\mathbf{x}'\Sigma_{n22.1}\mathbf{x} \leq p_{2n}-2) \end{aligned}$$

Moral of the Story

By Ignoring the Bias, it will Not go away!

- Submodel estimator provided by some existing variable selection techniques when $p_n \gg n$ are subject to BIAS.

- The prediction performance can be improved by the shrinkage strategy.

- Particulary when an under-fitted submodel is selected by an aggressive penalty parameter.

Moral of the Story

By Ignoring the Bias, it will Not go away!

- When $p \gg n$, we assume the true model is sparse in the sense that most coefficients goes to 0 when $n \to \infty$.
- However, it is realistic to assume that some β_j may be small, but not exactly 0.
- Such predictors with small amount of influence on the response variable are often ignored incorrectly in HD variable selection methods.
- We borrow (re-gain) some information from those predictors using the shrinkage strategy to improve the prediction performance.

Engineering Proof: Simulation

- In all experiments, ϵ_i's are simulated from i.i.d standard normal random variables, $x_{is} = (\xi^1_{(is)})^2 + \xi^2_{(is)}$, where $\xi^1_{(is)}$ and $\xi^2_{(is)}$, $i = 1, \cdots, n$, $s = 1, \cdots, p_n$ are also independent copies of standard normal distribution.

- In all sampling experiments, we let $p_n = n^\alpha$ for different sample size n, where α changes from 1 to 1.8 with an increment of 0.2. The HD-PSE is computed for $r_n = p_n^{1/8}$ and $a_n = 0.1n^{-1/3}$.

Simulation Results

Engineering Proof

- The performance of an estimator of β will be appraised using the mean squared error (MSE) criterion.
- All computations were conducted using the **R** statistical software.
- We have numerically calculated the relative MSE of the estimators with respect to $\hat{\beta}^{WR}$ by simulation.
- The simulated relative efficiency (SRE) of the estimator $\beta^{\diamond}$ to the maximum likelihood estimator $\hat{\beta}^{FM}$ is denoted by

$$\text{SRE}(\hat{\beta}^{FM} : \beta^{\diamond}) = \frac{\text{MSE}(\hat{\beta}^{WR})}{\text{MSE}(\beta^{\diamond})}.$$

- A SRE larger than one indicates the degree of superiority of the estimator $\beta^{\diamond}$ over $\hat{\beta}^{WR}$.

Engineering Proof

Relative Performance

- We let $\beta_{10} = (1.5, 3, 2)'$ be fixed for every design.
- Let $\Delta^* = \|\beta_{20} - \mathbf{0}\|^2$ varying between 0 and 4.
- We choose $n = 30$ or 100.

Table: Simulated RMSEs

.

(n,p)	Δ^*	$\hat{\beta}_{1n}^{SM}$	$\hat{\beta}_{1n}^{PSE}$	(n,p)	Δ^*	$\hat{\beta}_{1n}^{SM}$	$\hat{\beta}_{1n}^{PSE}$
	0.00	16.654	4.101		0.00	8.953	5.385
	0.05	8.202	3.446		0.05	4.456	3.794
	0.20	2.855	2.610		0.20	1.551	3.216
	0.25	2.074	2.437		0.25	1.422	2.833
	0.30	1.857	2.180		0.30	1.091	2.459
$(30,30)$	0.35	1.643	1.949	$(30,59)$	0.35	0.986	2.447
	0.80	0.649	1.506		0.80	0.542	1.601
	2.50	0.232	1.160		2.50	0.234	1.171
	3.30	0.170	1.095		3.30	0.210	1.108
	0.00	12.672	4.260		0.00	5.546	5.388
	0.05	2.546	3.538		0.05	1.255	1.900
	0.10	1.129	3.256		0.15	0.441	1.322
	0.20	0.628	2.948		0.20	0.361	1.382
	0.25	0.481	3.366		0.25	0.316	1.358
$(100,158)$	0.40	0.311	2.272	$(100,398)$	0.40	0.198	1.543
	1.40	0.110	1.500		1.40	0.096	1.826
	3.10	0.066	1.181		3.10	0.079	1.304
	3.50	0.060	1.217		3.50	0.075	1.297

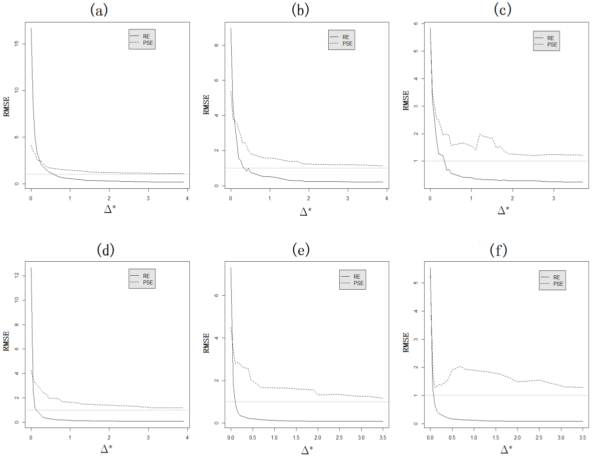


Figure: The top three panels (a-c) are for $n = 30$ and $p_n = 30, 59, 117$ from the left to the right. The bottom panels (d-f) are for $n = 100$ and $p_n = 158, 251, 398$ from the left to the right. Solid curves: $\mathrm{RMSE}(\hat{\beta}_{1n}^{SM})$; Dashed curves: $\mathrm{RMSE}(\hat{\beta}_{1n}^{PSE})$.

Shrinkage Versus Penalty Estimators

Engineering Solution: Simulation Results

- Performance of HD-PSE relative to penalty estimators including Lasso, ALasso, SCAD, MCP and Threshold Ridge (TR).
- We let $\beta_{10} = (1.5, 3, 2, \underbrace{0.1, \cdots, 0.1}_{p_{1n}-3})'$, $\beta_{20} = \mathbf{0}'_{p_{2n}}$.
- The model includes some predictors with weak signals. We consider $n = 30$ and $p_{1n} = 3, 4, 10, 20$.
- We choose $a = 3.7$ and $\gamma = 3$ for SCAD and MCP, respectively.
- For TR, we choose $\alpha_n = c_6 n^{-1/3}$ and $\lambda = c_7(\log\log n)^3/\alpha_n^2$, where c_6 and c_7 are two tuning parameters.
- All tuning parameters are chosen using the generalized cross validation.

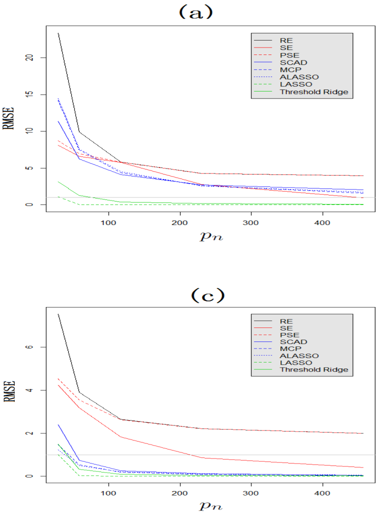
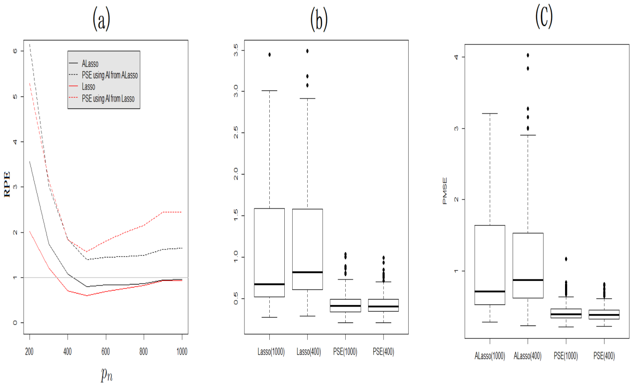


Figure: RMSEs for $n = 30$. Plots (a-d) are for $p_1 = 3, 4, 10, 20$, respectively.


(a)
ALasso
PSE using AI from ALasso
Lasso
PSE using AI from Lasso
RPE
p_n
(b)
Lasso(1000)
Lasso(400)
PSE(1000)
PSE(400)
(C)
PMSE
ALasso(1000)
ALasso(400)
PSE(1000)
PSE(400)

p_1	p_n	$\hat{\beta}_{1n}^{SM}$	$\hat{\beta}_{1n}^{PSE}$	$\hat{\beta}_{1n}^{\mathrm{SCAD}}$	$\hat{\beta}_{1n}^{\mathrm{MCP}}$	$\hat{\beta}_{1n}^{\mathrm{ALasso}}$	$\hat{\beta}_{1n}^{\mathrm{Lasso}}$	$\hat{\beta}_{1n}^{\mathrm{TR}}$
3	30	23.420	8.740	14.486	14.247	11.399	3.130	1.097
	59	9.900	6.951	7.588	7.499	6.244	1.257	0.015
	231	4.292	4.291	2.568	2.622	2.714	0.166	0.003
	456	3.977	3.977	1.739	1.576	2.059	0.099	0.002
4	30	15.055	6.882	11.809	11.291	9.528	2.830	0.993
	59	6.954	4.933	5.260	5.204	4.469	0.966	0.019
	231	3.605	3.605	2.222	2.154	2.045	0.167	0.004
	456	3.184	3.184	1.648	1.436	1.703	0.102	0.003
10	30	7.528	4.526	1.232	1.469	2.391	1.497	1.001
	59	3.899	3.534	0.493	0.538	0.746	0.321	0.032
	231	2.212	2.212	0.104	0.083	0.117	0.034	0.005
	456	1.997	1.997	0.052	0.032	0.050	0.017	0.003
20	30	4.603	3.139	0.099	0.128	0.892	0.599	0.981
	59	2.231	2.194	0.016	0.018	0.067	0.031	0.013
	231	1.489	1.489	0.002	0.002	0.003	0.002	0.002
	456	1.392	1.392	0.001	0.001	0.002	0.001	0.001

Threshold Ridge Regression

A Threshold ridge (TR) for $1 \leq j \leq p_n$ of β_j is given by (Shao and Deng (2008))

$$\hat{\beta}_j^{\mathrm{TR}} = \begin{cases} \widetilde{\beta}_j, & |\widetilde{\beta}_j| > a_n, \\ 0, & |\widetilde{\beta}_j| \leq a_n, \end{cases}$$

where

$$\widetilde{\beta}_n = \arg\min_{\beta} \left\{ \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p_n} x_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{p_n} \beta_j^2 \right\}$$

and $a_n = cn^{-\omega}$ for $0 < \omega < 1/2$ and $c > 0$.

Shrinkage Versus Penalty Estimators

- The submodel estimator dominates all other estimators in the class, since $\hat{\beta}^{SM}$ is computed based on the true submodel.
- SCAD and MCP work better than the HD-PSE for smaller p_n.
- HD-PSE performs better than penalty estimators for larger p_n.
- Penalty estimators are even less efficient than the weighted ridge estimate.
- This phenomenon can be explained by the existence of predictors with weak effects, which cannot be separated from zero effects using Lasso-type methods.
- The predictors are designed to be correlated, the weighted ridge estimator can generate a better estimation at the starting point.

Shrinkage Versus Penalty Estimators

- The submodel estimator dominates all other estimators in the class, since $\hat{\beta}^{SM}$ is computed based on the true submodel.
- SCAD and MCP work better than the HD-PSE for smaller p_n.
- HD-PSE performs better than penalty estimators for larger p_n.
- Penalty estimators are even less efficient than the weighted ridge estimate.
- This phenomenon can be explained by the existence of predictors with weak effects, which cannot be separated from zero effects using Lasso-type methods.
- The predictors are designed to be correlated, the weighted ridge estimator can generate a better estimation at the starting point.

Shrinkage Versus Penalty Estimators

- The submodel estimator dominates all other estimators in the class, since $\hat{\beta}^{SM}$ is computed based on the true submodel.
- SCAD and MCP work better than the HD-PSE for smaller p_n.
- HD-PSE performs better than penalty estimators for larger p_n.
- Penalty estimators are even less efficient than the weighted ridge estimate.
- This phenomenon can be explained by the existence of predictors with weak effects, which cannot be separated from zero effects using Lasso-type methods.
- The predictors are designed to be correlated, the weighted ridge estimator can generate a better estimation at the starting point.

Shrinkage Versus Penalty Estimators

- The submodel estimator dominates all other estimators in the class, since $\hat{\beta}^{SM}$ is computed based on the true submodel.
- SCAD and MCP work better than the HD-PSE for smaller p_n.
- HD-PSE performs better than penalty estimators for larger p_n.
- Penalty estimators are even less efficient than the weighted ridge estimate.
- This phenomenon can be explained by the existence of predictors with weak effects, which cannot be separated from zero effects using Lasso-type methods.
- The predictors are designed to be correlated, the weighted ridge estimator can generate a better estimation at the starting point.

- The submodel estimator dominates all other estimators in the class, since $\hat{\beta}^{SM}$ is computed based on the true submodel.
- SCAD and MCP work better than the HD-PSE for smaller p_n.
- HD-PSE performs better than penalty estimators for larger p_n.
- Penalty estimators are even less efficient than the weighted ridge estimate.
- This phenomenon can be explained by the existence of predictors with weak effects, which cannot be separated from zero effects using Lasso-type methods.
- The predictors are designed to be correlated, the weighted ridge estimator can generate a better estimation at the starting point.

Shrinkage Versus Penalty Estimators

- The submodel estimator dominates all other estimators in the class, since $\hat{\beta}^{SM}$ is computed based on the true submodel.
- SCAD and MCP work better than the HD-PSE for smaller p_n.
- HD-PSE performs better than penalty estimators for larger p_n.
- Penalty estimators are even less efficient than the weighted ridge estimate.
- This phenomenon can be explained by the existence of predictors with weak effects, which cannot be separated from zero effects using Lasso-type methods.
- The predictors are designed to be correlated, the weighted ridge estimator can generate a better estimation at the starting point.

Shrinkage Versus Penalty Estimators

- The submodel estimator dominates all other estimators in the class, since $\hat{\beta}^{SM}$ is computed based on the true submodel.
- SCAD and MCP work better than the HD-PSE for smaller p_n.
- HD-PSE performs better than penalty estimators for larger p_n.
- Penalty estimators are even less efficient than the weighted ridge estimate.
- This phenomenon can be explained by the existence of predictors with weak effects, which cannot be separated from zero effects using Lasso-type methods.
- The predictors are designed to be correlated, the weighted ridge estimator can generate a better estimation at the starting point.

Microarray Data Example

- We apply the proposed HD-PSE strategy to the data set reported in Scheetz et al. (2006) and also analyzed by Huang, Ma and Zhang (2008).
- In this dataset, 120 twelve-week-old male offsprings of F1 animals were selected for tissue harvesting from the eyes for microarray analysis.
- The microarrays used to analyze the RNA from the eyes of these F2 animals contain over $31,042$ different probe sets (Affymetric GeneChip Rat Genome 230 2.0 Array).

Microarray Data Example

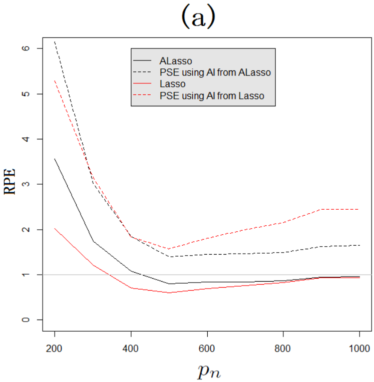
- Huang, Ma and Zhang (2008) studied a total of 18,976 probes including gene TRIM32, which was recently found to cause Bardet-Biedl syndrome (Chiang et al. (2006)), a genetically heterogeneous disease of multiple organ systems including the retina.
- A regression analysis was conducted to find the probes among the remaining $18,975$ probes that are most related to TRIM32 (Probe ID: 1389163_at). Huang et al (2008) found 19 and 24 probes based on Lasso and adaptive Lasso methods, respectively.
- We compute HD-PSEs based on two different candidate subset models consisting of 24 and 19 probes selected from Lasso and adaptive Lasso, respectively.

Microarray Data Example

- In the largest full set model, we consider at most $1,000$ probes with the largest variances. Other smaller full set model with top p_n probes are also considered.
- Here we choose different p_n's between 200 and $1,000$.
- The relative prediction error (RPE) of the estimator $\beta^*_{\mathcal{J}}$ relative to weighted ridge estimator $\hat{\beta}^{WR}_{\mathcal{J}}$ is computed as follows

$$\mathrm{RPE}(\beta^*_{\mathcal{J}}) = \frac{\sum_{i=1}^{n} \|\mathbf{y} - \sum_{j\in\mathcal{J}} \mathbf{X}_{\mathcal{J}} \hat{\beta}^{WR}_{\mathcal{J}}\|^2}{\sum_{i=1}^{n} \|\mathbf{y} - \sum_{j\in\mathcal{J}} \mathbf{X}_{\mathcal{J}} \beta^*_{\mathcal{J}}\|^2},$$

where $\mathcal{J}$ is the index of the submodel including either 24 or 19 elements.


(a)
ALasso
PSE using AI from ALasso
Lasso
PSE using AI from Lasso
RPE
0
1
2
3
4
5
6
200
400
600
800
1000
p_n
(b)
0.5
1.0
1.5
2.0
2.5
3.0
3.5
Lasso(1000)
Lasso(400)
PSE(1000)
PSE(400)

Envoi

- We generalized the classical Stein's shrinkage estimation to a high-dimensional sparse model with some predictors with weak signals.
- When p_n grows with n quickly, it is reasonable to suspect that most predictors do not contribute, that is model is sparse.
- We proposed a HD shrinkage estimation strategy by shrinking a weighted ridge estimator in the direction of a candidate submodel.

Envoi

- Existing penalized regularization approaches have some advantages of generating a parsimony sparse model, but tends to ignore the possible small contributions from some predictors.
- Lasso-type methods provide estimation and prediction only based on the selected candidate submodel, which is often inefficient with the existence of mild or weak signals.
- Our proposed HD shrinkage strategy takes into account possible contributions of all other possible nuisance parameters and has dominant prediction performances over submodel estimates generated from Lasso-type methods, which depend strongly on the sparsity assumption of the true model.

Is Classical Shrinkage Estimation Dead?

Long Live L_2 Shrinkage!

Long Live L_2 Shrinkage!

Long Live L_2 Shrinkage!

Is Classical Shrinkage Estimation Dead?

Long Live L_2 Shrinkage!

Long Live L_2 Shrinkage!

Long Live L_2 Shrinkage!

Is Classical Shrinkage Estimation Dead?

Long Live L_2 Shrinkage!

Long Live L_2 Shrinkage!

Long Live L_2 Shrinkage!

Is Classical Shrinkage Estimation Dead?

Long Live L_2 Shrinkage!

Long Live L_2 Shrinkage!

Long Live L_2 Shrinkage!

World's Data is Growing Exponentially!

- How to Acquire, Manage, Process, Analyze and Make Sense of Big Data?
- Big data is the future of Science and Trans-disciplinary research in Statistical Sciences is a must.
- "Think of big data as an epic wave gathering now, starting to crest," says the Harvard Business Review. "If you want to catch it, you need people who can surf"
- By 2015 there will be 4.4 M jobs available globally for Big Data analysis.
- Are we training "Wave Jockeys"?

World's Data is Growing Exponentially!

- A greater collaboration between statisticians, computer scientists and social scientists (Facebook clicks, Netflix queues, and GPS data, a few to mention, 12 billions devices are connected to internet).
- Data is never neutral and unbiased, we must pull expertise across a host of fields to combat the biases in the estimation.
- Need to be careful with algorithmic based predictions. For example, protein interaction prediction.
- "The purpose of computing is insight, not numbers." R.W. Hamming, 1962.
- "Big Data can't tell us why easily – it can only tell us the what, but most often that's enough." Mayer-Schonberger, CBC Radio.

Culture in Statistical Sciences

- Study classical problems - Classical assumptions
- Exact/Analytic Solutions
- Low-dimensional Data Analysis
- Work Alone or in Small Teams
- Glory of the Individual

Culture in Statistical Sciences

- Study classical problems - Classical assumptions
- Exact/Analytic Solutions
- Low-dimensional Data Analysis
- Work Alone or in Small Teams
- Glory of the Individual

Culture in Statistical Sciences

- Study classical problems - Classical assumptions
- Exact/Analytic Solutions
- Low-dimensional Data Analysis
- Work Alone or in Small Teams
- Glory of the Individual

Culture in Statistical Sciences

- Study classical problems - Classical assumptions
- Exact/Analytic Solutions
- Low-dimensional Data Analysis
- Work Alone or in Small Teams
- Glory of the Individual

Culture in Statistical Sciences

- Study classical problems - Classical assumptions
- Exact/Analytic Solutions
- Low-dimensional Data Analysis
- Work Alone or in Small Teams
- Glory of the Individual

Culture in Statistical Sciences

- Study classical problems - Classical assumptions
- Exact/Analytic Solutions
- Low-dimensional Data Analysis
- Work Alone or in Small Teams
- Glory of the Individual

World is Changing

- Complex Problems, Approximate Solutions
- Visualizing Complex Data - Use of Technology
- High-Dimensional Statistical Inference
- Think Tanks - Trans-disciplinary Research
- Glory of the Research Team

World is Changing

- Complex Problems, Approximate Solutions
- Visualizing Complex Data - Use of Technology
- High-Dimensional Statistical Inference
- Think Tanks - Trans-disciplinary Research
- Glory of the Research Team

World is Changing

- Complex Problems, Approximate Solutions
- Visualizing Complex Data - Use of Technology
- High-Dimensional Statistical Inference
- Think Tanks - Trans-disciplinary Research
- Glory of the Research Team

World is Changing

- Complex Problems, Approximate Solutions
- Visualizing Complex Data - Use of Technology
- High-Dimensional Statistical Inference
- Think Tanks - Trans-disciplinary Research
- Glory of the Research Team

World is Changing

- Complex Problems, Approximate Solutions
- Visualizing Complex Data - Use of Technology
- High-Dimensional Statistical Inference
- Think Tanks - Trans-disciplinary Research
- Glory of the Research Team

Clash of Cultures

World is Changing

- Complex Problems, Approximate Solutions
- Visualizing Complex Data - Use of Technology
- High-Dimensional Statistical Inference
- Think Tanks - Trans-disciplinary Research
- Glory of the Research Team

Thanks a bundle!

Thank you and thanks to organizers!