Pictures From My Backyard of Niagara Falls, Canada

IMPLICIT BIAS in BIG DATA ANALYTICS

S. Ejaz Ahmed
Brock University

Workshop on "New and Evolving Roles of Shrinkage in Large-Scale Prediction and Inference"

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

BIRS ---April 7-12, 2019

Outline of Presentation

- Estimation and Prediction Strategies in Sparse Regression Model
- Supervised Learning
- Submodel Selection
- Asymptotics, Numerical Studies and Applications
- Envoi

Outline of Presentation

- Estimation and Prediction Strategies in Sparse Regression Model
- Supervised Learning
- Submodel Selection
- Asymptotics, Numerical Studies and Applications
- Envoi

Outline of Presentation

- Estimation and Prediction Strategies in Sparse Regression Model
- Supervised Learning
- Submodel Selection
- Asymptotics, Numerical Studies and Applications
- Envoi

Outline of Presentation

- Estimation and Prediction Strategies in Sparse Regression Model
- Supervised Learning
- Submodel Selection
- Asymptotics, Numerical Studies and Applications
- Envoi

Outline of Presentation

- Estimation and Prediction Strategies in Sparse Regression Model
- Supervised Learning
- Submodel Selection
- Asymptotics, Numerical Studies and Applications
- Envoi

Big Data and Data Science

What is Data Science?

What is Big Data?

What is High Dimensional Data?

Data Science recognized as a Science at the beginning of century

Volume of data (Big Data) due to Technological advancement can be stored

Big Data and Data Science

What is Data Science?

What is Big Data?

What is High Dimensional Data?

Data Science recognized as a Science at the beginning of century

Volume of data (Big Data) due to Technological advancement can be stored

Big Data and Data Science

What is Data Science?

What is Big Data?

What is High Dimensional Data?

Data Science recognized as a Science at the beginning of century

Volume of data (Big Data) due to Technological advancement can be stored

Big Data and Data Science

What is Data Science?

What is Big Data?

What is High Dimensional Data?

Data Science recognized as a Science at the beginning of century

Volume of data (Big Data) due to Technological advancement can be stored

Big Data and Data Science

What is Data Science?

What is Big Data?

What is High Dimensional Data?

Data Science recognized as a Science at the beginning of century

Volume of data (Big Data) due to Technological advancement can be stored

Big Data and Data Science

Big Data – Big Problem: May not be Valuable? A Paradox!

Data Needs to be extracted to make the data valuable for further study

Data Mining/Cleaning/Massaging/Introgating/Touturing?

Data Reduction and/or Variables Reduction

Data Reduction (extracting valuable information) from a Big Data is called "Data Science"!

Big Data and Data Science

Big Data – Big Problem: May not be Valuable? A Paradox!

Data Needs to be extracted to make the data valuable for further study

Data Mining/Cleaning/Massaging/Introgating/Touturing?

Data Reduction and/or Variables Reduction

Data Reduction (extracting valuable information) from a Big Data is called "Data Science"!

Big Data and Data Science

Big Data – Big Problem: May not be Valuable? A Paradox!

Data Needs to be extracted to make the data valuable for further study

Data Mining/Cleaning/Massaging/Introgating/Touturing?

Data Reduction and/or Variables Reduction

Data Reduction (extracting valuable information) from a Big Data is called "Data Science"!

Big Data and Data Science

Big Data – Big Problem: May not be Valuable? A Paradox!

Data Needs to be extracted to make the data valuable for further study

Data Mining/Cleaning/Massaging/Introgating/Touturing?

Data Reduction and/or Variables Reduction

Data Reduction (extracting valuable information) from a Big Data is called "Data Science"!

Big Data and Data Science

Big Data – Big Problem: May not be Valuable? A Paradox!

Data Needs to be extracted to make the data valuable for further study

Data Mining/Cleaning/Massaging/Introgating/Touturing?

Data Reduction and/or Variables Reduction

Data Reduction (extracting valuable information) from a Big Data is called "Data Science"!

BIG DATA / HIGH DIMENSIONAL DATA

BIG DATA MATRIX (*nxp*)

n is large and *p* is also large, $n > p$; *p* is fixed

Big Data with a Large number of Variables

$n <<< p$, HIGH DIMENSIONAL DATA (HDD)

BIG DATA / HIGH DIMENSIONAL DATA

BIG DATA MATRIX (*nxp*)

n is large and p is also large, $n > p$; p is fixed

Big Data with a Large number of Variables

$n <<< p$, HIGH DIMENSIONAL DATA (HDD)

BIG DATA / HIGH DIMENSIONAL DATA

BIG DATA MATRIX (*nxp*)

n is large and p is also large, $n > p$; p is fixed

Big Data with a Large number of Variables

$n <<< p$, HIGH DIMENSIONAL DATA (HDD)

BIG DATA / HIGH DIMENSIONAL DATA

BIG DATA MATRIX (*nxp*)

n is large and *p* is also large, $n > p$; *p* is fixed

Big Data with a Large number of Variables

$n <<< p$, HIGH DIMENSIONAL DATA (HDD)

BIG DATA / HIGH DIMENSIONAL DATA

BIG DATA MATRIX (*nxp*)

n is large and *p* is also large, $n > p$; *p* is fixed

Big Data with a Large number of Variables

$n <<< p$, HIGH DIMENSIONAL DATA (HDD)

Some of Important Features Related to Big Data

Volume

Variety

Velocity

Viability/Sparsity

Some of Important Features Related to Big Data

Volume

Variety

Velocity

Viability/Sparsity

Some of Important Features Related to Big Data

Volume

Variety

Velocity

Viability/Sparsity

Some of Important Features Related to Big Data

Volume

Variety

Velocity

Viability/Sparsity

Big Data

Describes the situation where the volume, variety and velocity of data may exceed an organization's storage or compute capacity?

Volume: Terabytes to Petabytes or up

Variety: Pictures, audios, videos, texts, emails, etc.

Velocity: The speed of data processing

Big Data

Describes the situation where the volume, variety and velocity of data may exceed an organization's storage or compute capacity?

Volume: Terabytes to Petabytes or up

Variety: Pictures, audios, videos, texts, emails, etc.

Velocity: The speed of data processing

BIG DATA

- In genetic micro-array studies, *n* is measured in hundreds, the number of features *d* per sample can exceed millions!!!
- 22,500 unique proteins, implies about 253,000,000 possible protein-protein interactions. So far 42,000 are identified.
- Facebook: More than 950 millions user spends approximately 7 hours on average for a given month, a huge task to store and analyze such Big and Complex Data.
- Combining different Health Data banks with a great heterogeneity collected over the time.
- The developments in the arena of Big Data is still infancy.

BIG DATA

- In genetic micro-array studies, *n* is measured in hundreds, the number of features *d* per sample can exceed millions!!!
- 22,500 unique proteins, implies about 253,000,000 possible protein-protein interactions. So far 42,000 are identified.
- Facebook: More than 950 millions user spends approximately 7 hours on average for a given month, a huge task to store and analyze such Big and Complex Data.
- Combining different Health Data banks with a great heterogeneity collected over the time.
- The developments in the arena of Big Data is still infancy.

BIG DATA

- In genetic micro-array studies, n is measured in hundreds, the number of features d per sample can exceed millions!!!
- 22,500 unique proteins, implies about 253,000,000 possible protein-protein interactions. So far 42,000 are identified.
- Facebook: More than 950 millions user spends approximately 7 hours on average for a given month, a huge task to store and analyze such Big and Complex Data.
- Combining different Health Data banks with a great heterogeneity collected over the time.
- The developments in the arena of Big Data is still infancy.

BIG DATA

- In genetic micro-array studies, n is measured in hundreds, the number of features d per sample can exceed millions!!!
- 22,500 unique proteins, implies about 253,000,000 possible protein-protein interactions. So far 42,000 are identified.
- Facebook: More than 950 millions user spends approximately 7 hours on average for a given month, a huge task to store and analyze such Big and Complex Data.
- Combining different Health Data banks with a great heterogeneity collected over the time.
- The developments in the arena of Big Data is still infancy.

BIG DATA

- In genetic micro-array studies, *n* is measured in hundreds, the number of features *d* per sample can exceed millions!!!
- 22,500 unique proteins, implies about 253,000,000 possible protein-protein interactions. So far 42,000 are identified.
- Facebook: More than 950 millions user spends approximately 7 hours on average for a given month, a huge task to store and analyze such Big and Complex Data.
- Combining different Health Data banks with a great heterogeneity collected over the time.
- The developments in the arena of Big Data is still infancy.

BIG DATA

- In genetic micro-array studies, n is measured in hundreds, the number of features d per sample can exceed millions!!!
- 22,500 unique proteins, implies about 253,000,000 possible protein-protein interactions. So far 42,000 are identified.
- Facebook: More than 950 millions user spends approximately 7 hours on average for a given month, a huge task to store and analyze such Big and Complex Data.
- Combining different Health Data banks with a great heterogeneity collected over the time.
- The developments in the arena of Big Data is still infancy.

Regression Model

Supervised Learning

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

y_i: observed response variable

$\mathbf{x}_i = (x_{i1}, \cdots, x_{ip_n})'$

$\boldsymbol{\beta}_n = (\beta_1, \cdots, \beta_{p_n})'$ is a p_n-dimensional vector of the unknown parameters

ϵ_i's are independent and identically distributed with center 0 and variance σ^2.

Sparsity

Suppose $\beta = (\beta_1', \beta_2')'$

In many applications it is assumed that model is sparse, i.e. $\beta_2 = \mathbf{0}$, $p_1 < n$, $p_2 > n$.

Variable Selection: Submodel

The parameter of interest is β_1

.

Post Estimation of β_1

.

Submodel Selection: Penalized Likelihood

- 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

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

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

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

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

Penalty Estimation Strategy

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 with the absolute penalty $||\beta_j||_1$ in the LASSO–

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

It shrinks the coefficient towards zero, and depending on the value of λ, it sets some of the coefficients to exactly zero.

Other: The smoothly clipped absolute deviation(SCAD), Adaptive LASSO, MCP and others

Bayesian Perspectives

- The penalized likelihood methods have a close connection to Bayesian procedures.
- The LASSO estimate corresponds to a Bayes method that puts a Laplacian (double exponential) prior on the regression coefficients.
- Recent results Armagan (2013), Bhattacharya (2012), and Carvalho (2010) have shown that better results can be obtained in small *n*, large *p* case by using priors with heavier tails than the double exponential prior, in particular, priors with polynomial tails.

Penalty Estimation Strategy

Good Strategy if Model is **Truly** Sparse

Unrevealing Tale of Under-fitted Model

Submodel Estimators are BIASED!!!

BIAS is the **BIG** Problem in Big Data

NEED TO CONTROL BIAS

Unbearable Truth about Submodel Estimation

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

- **Condition 1**: the regression coefficients corresponding to deleted variables (β_2) are zero
- **Condition 2**: the retained variables are orthogonal to the deleted variables,$\mathbf{X}_1'\mathbf{X}_2 = \mathbf{0}$
- **Dominance Condition**: 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.

All Submodels are BIASED

Bias will increase without a bound! Consequentially MSE will explode!!

Not a good Tradeoff!!!

"All Submodels are Biased, but Some are Useful!!!"

- A naive data analyst/data scientist/data miner and others may not comprehend that by dropping $\mathbf{X}_2$ from the model the impact will be drastic.
- $\mathbf{X}_2\boldsymbol{\beta}_2$ will **covertly influence** the estimation and testing of $\boldsymbol{\beta}_1$.
- The law of **Unconsciousness** may not work!

Strong Signal, Weak signal and 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.**
- The assumptions on designed covariates include adaptive irrepresentable condition, or the restricted eigenvalue condition.

Strong Signal, Weak signal and 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.**
- The assumptions on designed covariates include adaptive irrepresentable condition, or the restricted eigenvalue condition.

Strong Signal, Weak signal and 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.**
- The assumptions on designed covariates include adaptive irrepresentable condition, or the restricted eigenvalue condition.

Strong Signal, Weak signal and 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.**
- The assumptions on designed covariates include adaptive irrepresentable condition, or the restricted eigenvalue condition.

Strong Signal, Weak signal and 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.**
- The assumptions on designed covariates include adaptive irrepresentable condition, or the restricted eigenvalue condition.

Strong Signal, Weak signal and 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.**
- The assumptions on designed covariates include adaptive irrepresentable condition, or the restricted eigenvalue condition.

Strong Signal, Weak signal and 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.**
- The assumptions on designed covariates include adaptive irrepresentable condition, or the restricted eigenvalue condition.

Strong Signal, Weak signal and Noise

- Split the signals into Strong and Weak ones.
- The Usual one-step method such as LASSO may be very effective in detecting strong signals while failing to identify some weak ones.
- Resulting in significant negative impact on the model fitting and prediction due to inherited Bias.

Integrating Penalty Strategies

1. We propose a post selection shrinkage strategy to improve the performance of the Lasso-type estimators in high-dimensional settings by combining two model estimators.
2. This post selection shrinkage strategy is data-adaptive and efficient when selected submodel ignores predictors with joint weak effects.

Model Sparsity and Signal Strength

The effect of all p_n predictors are characterized into three categories based upon their signal strength:

- important predictors with strong effects in S_1
- predictors with no effect in S_3,
- intermediate group in S_2 with joint weak effects.
- Most existing high-dimensional sparse models investigate the variable selection consistency by only considering the existence of the strong signals in and sparse signals

If we let $\widehat{S}_1 \subset \{1, \cdots, p_n\}$ index an active subset, then a data-adaptive candidate subset model is produced such that

$$\hat{\beta}_j^{\mathrm{PLS}} = 0 \quad \text{if and only if } j \notin \widehat{S}_1 .$$

Model Sparsity and Signal Strength

- Denote design matrix $\mathbf{X} = (\mathbf{x_1} \cdots \mathbf{x}_{p_n})$.
- Without loss of generality, we rearrange the designed vectors such that $\mathbf{X} = (\mathbf{X}_{S_1} | \mathbf{X}_{S_2} | \mathbf{X}_{S_3})$, where $\mathbf{X}_S$ is the submatrix consists of vectors in $S \subset \{1, \cdots, p_n\}$.
- Chen, Donoho and Saunders (1998) and Wainwright (2009) argued that S_2 cannot be separated from S_3,

Under some regularity conditions, Belloni, Chernozhukov (2009) and Liu and Yu (2013) studied post selection least square estimators, to improve the prediction performance of the PLS estimator.

$$\hat{\beta}_{\widehat{S}_1}^{RE} = (\mathbf{X}'_{\widehat{S}_1} \mathbf{X}_{\widehat{S}_1})^{-1} \mathbf{X}'_{\widehat{S}_1}$$

Our Contribution: Discussion Paper

Gao X., Ahmed, S. Ejaz and Yang, F. (2017). Post Selection Shrinkage Estimation for High Dimensional Data Analysis. *Applied Stochastic Models in Business and Industry*, 33, 97–120.

Discussants

- Kjell Doksum and Joan Fujimura, U of Wisconsin, Madison
- Jianqing Fan, Princton
- Peihua Qiu, Kai Yang, and Lu You, U of Florida
- Yanming Li, Hyokyoung Grace Hong, and Yi Li, U of Michigan

Rejoinder:Applied Stochastic Models in Business and Industry, 33, 131–135
We thank them for their thought-provoking and insightful discussions on our paper.

Weighted Ridge Estimation

We construct a post selection weighted ridge estimation based upon $\widehat{S}_1$ in two steps.
Once $\widehat{S}_1$ is obtained from Step 1, we seek to minimize a penalized objective function with a ridge penalty on coefficients in $\widehat{S}_1^c$,

$$\tilde{\beta}^{WR}(r_n) = argmin\{L(\beta; \widehat{S}_1)\} = argmin\left\{\|\mathbf{y} - \mathbf{X}_n\beta_n\|^2 + r_n\|\beta_{\widehat{S}_1^c}\|^2\right\}$$

where $r_n > 0$ is a tuning parameter controlling the penalty effect on $\beta_{\widehat{S}_1}$. Then a post selection weighted ridge (WR) estimator $\hat{\beta}^{WR}(r_n, a_n; \widehat{S}_1)$ is obtained from,

$$\hat{\beta}_j^{WR}(r_n, a_n) = \begin{cases} \tilde{\beta}_j(r_n), & j \in \widehat{S}_1; \\ \tilde{\beta}_j(r_n) I(\tilde{\beta}_j(r_n) > a_n), & j \in \widehat{S}_1^c, \end{cases}$$

Shrinkage Estimation: Integrating Penalty Estimation

$$\begin{aligned}\hat{\beta}^{S}_{\widehat{S}_1} &= \hat{\beta}^{RE}_{\widehat{S}_1} + (\hat{\beta}^{WR}_{\widehat{S}_1} - \hat{\beta}^{RE}_{\widehat{S}_1})(1 - (\widehat{s}_2 - 2)/\widehat{T}_n) \\ &= \hat{\beta}^{WR}_{\widehat{S}_1} - ((\widehat{s}_2 - 2)/\widehat{T}_n)(\hat{\beta}^{WR}_{\widehat{S}_1} - \hat{\beta}^{RE}_{\widehat{S}_1}),\end{aligned}$$

where $\widehat{s}_2 = |\widehat{S}_2|$ and $\widehat{T}_n$ is given by

$$\widehat{T}_n = (\hat{\beta}^{WR}_{\widehat{S}_2})'(\mathbf{X}'_{\widehat{S}_2}\mathbf{M}_{\widehat{S}_1}\mathbf{X}_{\widehat{S}_2})\hat{\beta}^{WR}_{\widehat{S}_2}/\sigma^2, \tag{4}$$

where $\mathbf{M}_{\widehat{S}_1} = \mathbf{I}_n - \mathbf{X}_{\widehat{S}_1}(\mathbf{X}'_{\widehat{S}_1}\mathbf{X}_{\widehat{S}_1})^{-1}\mathbf{X}'_{\widehat{S}_1}$.

Positive-part Shrinkage Estimation: Integrating Penalty Estimation

$$\hat{\beta}^{S+}_{\widehat{S}_1} = \hat{\beta}^{WR}_{\widehat{S}_1} - ([(\widehat{s}_2 - 2)/\widehat{T}_n] \wedge 1)(ug_{\widehat{S}_1} - \hat{\beta}^{RE}_{\widehat{S}_1}).$$

$\widehat{S}_1^c$

Consistency and Asymptotic Normality

Weighted Ridge Estimation

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

$$n^{1/2} m_n^{-1} \mathbf{d}_n' (\hat{\beta}_{S_3^c}^{WR} - \beta_{S_3^c}^{*}) \xrightarrow{\mathrm{d}} N(0,1).$$

Asymptotic Distributional Risk

Define

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

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

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

Asymptotic Distributional Risk

Mathematical/Creative Proof

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

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

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

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

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

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

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

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

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

Asymptotic Distributional Risk

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

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

Creativity and Truth

"Creativity is Subjective"

"The truth is not"

Simulation

Engineering/Artistic 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 MSE (RMSE) of the estimator $\beta^{\diamond}$ to another estimator $\hat{\beta}^{WR}$ is denoted by

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

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

Simulation

Engineering/Artistic Proof

Relative Performance

- We let $\beta_{10} = (1.5, 3, 2)'$ be fixed for every design.
- Let $\Delta^* = \|\beta_2 - \mathbf{0}\|^2$ varying between 0 and 4.
- We choose $n = 30$ or 100.

(n,p)	Δ^*	$\hat{\beta}_{1n}^{RE}$	$\hat{\beta}_{1n}^{PSE}$	(n,p)	Δ^*	$\hat{\beta}_{1n}^{RE}$	$\hat{\beta}_{1n}^{PSE}$
$(30,30)$	0.00	16.654	4.101	$(30,59)$	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
	0.35	1.643	1.949		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
$(100,158)$	0.00	12.672	4.260	$(100,398)$	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
	0.40	0.311	2.272		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

Shrinkage Versus Penalty Estimators

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 $p_{1n} = 3, 4, 10, 20$.
- All tuning parameters are chosen using the generalized cross validation.

p_1	p_n	$\hat{\beta}_{1n}^{RE}$	$\hat{\beta}_{1n}^{PSE}$	$\hat{\beta}_{1n}^{\text{SCAD}}$	$\hat{\beta}_{1n}^{\text{MCP}}$	$\hat{\beta}_{1n}^{\text{ALasso}}$	$\hat{\beta}_{1n}^{\text{Lasso}}$	$\hat{\beta}_{1n}^{\text{TR}}$
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

Application: Genomics Data

- 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

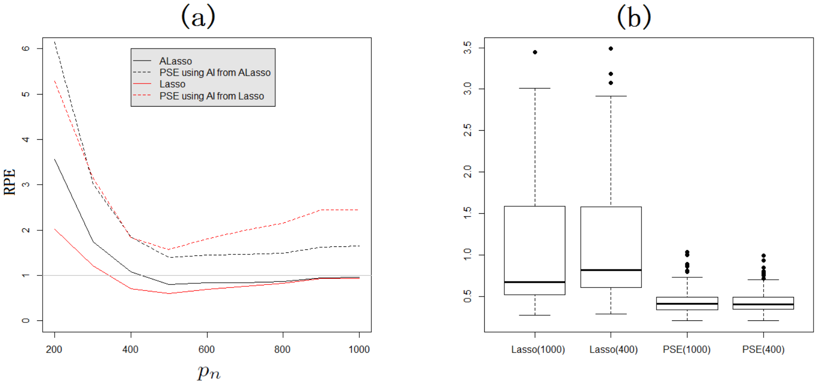
- Huang, Ma and Zhang (2008) studied a total of 18,976 probes including gene TRIM32, which was recently found to cause Bardet-Biedl syndrome (Chiang et al. (2006)), a genetically heterogeneous disease of multiple organ systems including the retina.
- A regression analysis was conducted to find the probes 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

- 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)
RPE
0
1
2
3
4
5
6
ALasso
PSE using AI from ALasso
Lasso
PSE using AI from Lasso
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)

Executive Summary

- Bancroft (1944) suggested two problems on pretest 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 pretest 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 pretest 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 pretest 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 pretest 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 pretest 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.

Envoi

- We generalized the classical Stein’s shrinkage estimation to a high-dimensional sparse model with some predictors with weak signals.
- 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 predictors and has dominant prediction performances over submodel estimates generated from Lasso-type methods.

References

Pretest, Penalty and Shrinkage Strategies

- 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.
- Hossien, Ahmed and Doksum(2015) for Generalized Linear Models.

References

Pretest, Penalty and Shrinkage Strategies

Li, Hong, Ahmed, Li (2018) -- CIS

- 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.
- Hossien, Ahmed and Doksum(2015) for Generalized Linear Models.

References

Pretest, Penalty and Shrinkage Strategies

- 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.
- Hossien, Ahmed and Doksum(2015) for Generalized Linear Models.

References

Pretest, Penalty and Shrinkage Strategies

- 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.
- Hossien, Ahmed and Doksum(2015) for Generalized Linear Models.

References

Pretest, Penalty and Shrinkage Strategies

- 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.
- Hossien, Ahmed and Doksum(2015) for Generalized Linear Models.

References

Pretest, Penalty and Shrinkage Strategies

- 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.
- Hossien, Ahmed and Doksum(2015) for Generalized Linear Models.

References

Pretest, Penalty and Shrinkage Strategies

- 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. Contemporary Mathematics, a co-publication of American Mathematical Society and CRM, 2014*.
- S. E. Ahmed (Editor). *Big and Complex Data Analysis: Statistical Methodologies and Applications. Springer, 2017*.

References

Pretest, Penalty and Shrinkage Strategies

- 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. Contemporary Mathematics, a co-publication of American Mathematical Society and CRM, 2014*.

- S. E. Ahmed (Editor). *Big and Complex Data Analysis: Statistical Methodologies and Applications. Springer, 2017*.

References

Pretest, Penalty and Shrinkage Strategies

- 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. Contemporary Mathematics, a co-publication of American Mathematical Society and CRM, 2014*.
- S. E. Ahmed (Editor). *Big and Complex Data Analysis: Statistical Methodologies and Applications. Springer, 2017*.

References

Pretest, Penalty and Shrinkage Strategies

- 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. Contemporary Mathematics, a co-publication of American Mathematical Society and CRM, 2014*.

- S. E. Ahmed (Editor). *Big and Complex Data Analysis: Statistical Methodologies and Applications. Springer, 2017*.

World’s Data is Growing Exponentially!

- How to Acquire, Manage, Process, Analyze and Make Sense of Big Data?
- Big data is not going away and Trans-disciplinary research in Statistical and Data 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.

Clash of Cultures

Statistical Science and Data Science

- Pre-processed Sampled Data
- Exact/Analytic Solutions
- Low-dimensional Data Analysis
- SAS, SPSS and other; R (open source program)
- Idealistic
- Work Alone or in Small Team
- Glory of the Individual

Clash of Cultures

Statistical Science and Data Science

- Pre-processed Sampled Data
- Exact/Analytic Solutions
- Low-dimensional Data Analysis
- SAS, SPSS and other; R (open source program)
- Idealistic
- Work Alone or in Small Team
- Glory of the Individual

Clash of Cultures

Statistical Science and Data Science

- Pre-processed Sampled Data
- Exact/Analytic Solutions
- Low-dimensional Data Analysis
- SAS, SPSS and other; R (open source program)
- Idealistic
- Work Alone or in Small Team
- Glory of the Individual

Clash of Cultures

Statistical Science and Data Science

- Pre-processed Sampled Data
- Exact/Analytic Solutions
- Low-dimensional Data Analysis
- SAS, SPSS and other; R (open source program)
- Idealistic
- Work Alone or in Small Team
- Glory of the Individual

Clash of Cultures

Statistical Science and Data Science

- Pre-processed Sampled Data
- Exact/Analytic Solutions
- Low-dimensional Data Analysis
- SAS, SPSS and other; R (open source program)
- Idealistic
- Work Alone or in Small Team
- Glory of the Individual

Clash of Cultures

Statistical Science and Data Science

- Pre-processed Sampled Data
- Exact/Analytic Solutions
- Low-dimensional Data Analysis
- SAS, SPSS and other; R (open source program)
- Idealistic
- Work Alone or in Small Team
- Glory of the Individual

Clash of Cultures

Statistical Science and Data Science

- Pre-processed Sampled Data
- Exact/Analytic Solutions
- Low-dimensional Data Analysis
- SAS, SPSS and other; R (open source program)
- Idealistic
- Work Alone or in Small Team
- Glory of the Individual

Clash of Cultures

Statistical Science and Data Science

- Pre-processed Sampled Data
- Exact/Analytic Solutions
- Low-dimensional Data Analysis
- SAS, SPSS and other; R (open source program)
- Idealistic
- Work Alone or in Small Team
- Glory of the Individual

Clash of Cultures

Statistical Science and Data Science

- Big Data (available from different sources)
- Frame the Problem
- Identify the Data for Analysis (Population or Sample)
- **New tools for New data**: Python and R (open source programs)
- **Parallel Computing Systems**: Hadoop, Spark and other
- Combining Data (Data Fusion, Record Linkage): Tax returns, Court records, and Health agencies, etc.
- High-Dimensional Statistical Inference
- Pragmatic
- Think Tanks - Trans-disciplinary Research
- Glory of the Research Team

Statistical Science and Data Science

- Big Data (available from different sources)
- Frame the Problem
- Identify the Data for Analysis (Population or Sample)
- **New tools for New data**: Python and R (open source programs)
- **Parallel Computing Systems**: Hadoop, Spark and other
- Combining Data (Data Fusion, Record Linkage): Tax returns, Court records, and Health agencies, etc.
- High-Dimensional Statistical Inference
- Pragmatic
- Think Tanks - Trans-disciplinary Research
- Glory of the Research Team

Clash of Cultures

Statistical Science and Data Science

- Big Data (available from different sources)
- Frame the Problem
- Identify the Data for Analysis (Population or Sample)
- **New tools for New data**: Python and R (open source programs)
- **Parallel Computing Systems**: Hadoop, Spark and other
- Combining Data (Data Fusion, Record Linkage): Tax returns, Court records, and Health agencies, etc.
- High-Dimensional Statistical Inference
- Pragmatic
- Think Tanks - Trans-disciplinary Research
- Glory of the Research Team

Clash of Cultures

Statistical Science and Data Science

- Big Data (available from different sources)
- Frame the Problem
- Identify the Data for Analysis (Population or Sample)
- **New tools for New data**: Python and R (open source programs)
- **Parallel Computing Systems**: Hadoop, Spark and other
- Combining Data (Data Fusion, Record Linkage): Tax returns, Court records, and Health agencies, etc.
- High-Dimensional Statistical Inference
- Pragmatic
- Think Tanks - Trans-disciplinary Research
- Glory of the Research Team

Clash of Cultures

Statistical Science and Data Science

- Big Data (available from different sources)
- Frame the Problem
- Identify the Data for Analysis (Population or Sample)
- **New tools for New data**: Python and R (open source programs)
- **Parallel Computing Systems**: Hadoop, Spark and other
- Combining Data (Data Fusion, Record Linkage): Tax returns, Court records, and Health agencies, etc.
- High-Dimensional Statistical Inference
- Pragmatic
- Think Tanks - Trans-disciplinary Research
- Glory of the Research Team

Clash of Cultures

Statistical Science and Data Science

- Big Data (available from different sources)
- Frame the Problem
- Identify the Data for Analysis (Population or Sample)
- **New tools for New data**: Python and R (open source programs)
- **Parallel Computing Systems**: Hadoop, Spark and other
- Combining Data (Data Fusion, Record Linkage): Tax returns, Court records, and Health agencies, etc.
- High-Dimensional Statistical Inference
- Pragmatic
- Think Tanks - Trans-disciplinary Research
- Glory of the Research Team

Clash of Cultures

Statistical Science and Data Science

- Big Data (available from different sources)
- Frame the Problem
- Identify the Data for Analysis (Population or Sample)
- **New tools for New data**: Python and R (open source programs)
- **Parallel Computing Systems**: Hadoop, Spark and other
- Combining Data (Data Fusion, Record Linkage): Tax returns, Court records, and Health agencies, etc.
- High-Dimensional Statistical Inference
- Pragmatic
- Think Tanks - Trans-disciplinary Research
- Glory of the Research Team

Clash of Cultures

Statistical Science and Data Science

- Big Data (available from different sources)
- Frame the Problem
- Identify the Data for Analysis (Population or Sample)
- **New tools for New data**: Python and R (open source programs)
- **Parallel Computing Systems**: Hadoop, Spark and other
- Combining Data (Data Fusion, Record Linkage): Tax returns, Court records, and Health agencies, etc.
- High-Dimensional Statistical Inference
- Pragmatic
- Think Tanks - Trans-disciplinary Research
- Glory of the Research Team

Clash of Cultures

Statistical Science and Data Science

- Big Data (available from different sources)
- Frame the Problem
- Identify the Data for Analysis (Population or Sample)
- **New tools for New data**: Python and R (open source programs)
- **Parallel Computing Systems**: Hadoop, Spark and other
- Combining Data (Data Fusion, Record Linkage): Tax returns, Court records, and Health agencies, etc.
- High-Dimensional Statistical Inference
- Pragmatic
- Think Tanks - Trans-disciplinary Research
- Glory of the Research Team

Clash of Cultures

Statistical Science and Data Science

- Big Data (available from different sources)
- Frame the Problem
- Identify the Data for Analysis (Population or Sample)
- **New tools for New data**: Python and R (open source programs)
- **Parallel Computing Systems**: Hadoop, Spark and other
- Combining Data (Data Fusion, Record Linkage): Tax returns, Court records, and Health agencies, etc.
- High-Dimensional Statistical Inference
- Pragmatic
- Think Tanks - Trans-disciplinary Research
- Glory of the Research Team

Clash of Cultures

Statistical Science and Data Science

- Big Data (available from different sources)
- Frame the Problem
- Identify the Data for Analysis (Population or Sample)
- **New tools for New data**: Python and R (open source programs)
- **Parallel Computing Systems**: Hadoop, Spark and other
- Combining Data (Data Fusion, Record Linkage): Tax returns, Court records, and Health agencies, etc.
- High-Dimensional Statistical Inference
- Pragmatic
- Think Tanks - Trans-disciplinary Research
- Glory of the Research Team

Thanks a bundle!

Thanks to organizers (BIRS) for the kind invitation!

Thank you for sharing time with me!

The research is supported by the Natural Sciences and the Engineering Research Council of Canada -- Since 1988