Random matrices and statistics: beyond covariance matrices

N. El Karoui

Department of Statistics
UC, Berkeley

ICM, Rio de Janeiro, August 2018

Linear regression: fundamental statistical problem

Consider following situations: n observations,
responses: $\{Y_i\}_{i=1}^n$, $Y_i \in \mathbb{R}$
predictors $\{X_i\}_{i=1}^n$, $X_i \in \mathbb{R}^p$.

Linear regression: fundamental statistical problem

Consider following situations: n observations,

responses: $\{Y_i\}_{i=1}^n$, $Y_i \in \mathbb{R}$
predictors $\{X_i\}_{i=1}^n$, $X_i \in \mathbb{R}^p$.

Simple examples:

- Y_i : price of house i, X_i: characteristics of house
- Y_i : disease indicator (0/1), X_i: characteristics of patient

Question: want to predict Y_i from X_i

Linear regression: fundamental statistical problem

Consider following situations: n observations,

responses: $\{Y_i\}_{i=1}^n$, $Y_i \in \mathbb{R}$
predictors $\{X_i\}_{i=1}^n$, $X_i \in \mathbb{R}^p$.

Simple examples:

- Y_i : price of house i, X_i: characteristics of house
- Y_i : disease indicator (0/1), X_i: characteristics of patient

Question: want to predict Y_i from X_i Simplest method: linear regression;

$$\text{predict } Y_i \text{ by } \widehat{Y}_i(\beta) = X_i^T \beta$$

Linear regression: fundamental statistical problem

Consider following situations: n observations,

responses: $\{Y_i\}_{i=1}^n$, $Y_i \in \mathbb{R}$
predictors $\{X_i\}_{i=1}^n, X_i \in \mathbb{R}^p$.

Simple examples:

- Y_i : price of house i, X_i: characteristics of house
- Y_i : disease indicator (0/1), X_i: characteristics of patient

Question: want to predict Y_i from X_i Simplest method: linear regression;

$$\text{predict } Y_i \text{ by } \widehat{Y}_i(\beta) = X_i^T \beta$$

How to find "good" β:

$$\widehat{\beta}_\rho = \text{argmin}_{\beta \in \mathbb{R}^p} \sum_{i=1}^n \rho(Y_i - X_i^T \beta) \text{ , where } \rho \text{ is a function.}$$

Least-squares: $\rho(x) = x^2$

Uncertainty assessment

Key issue: stability of estimator $\widehat{\beta}_\rho$ or prediction $X_{\text{new}}^T \widehat{\beta}$.
Assume statistical model/true relationship

$$Y_i = X_i^T \beta_0 + \epsilon_i \ , i = 1, \ldots, n \ .$$

- $\widehat{\beta}_\rho$: an estimate of (unknown) β_0.
- Y_i's, X_i's observed.
- ϵ_i's noise (unobserved). $\mathbf{E}(\epsilon_i) = 0$, $\text{var}(\epsilon_i) = \sigma_\epsilon^2$

Uncertainty assessment

Key issue: stability of estimator $\widehat{\beta}_\rho$ or prediction $X_{\text{new}}^T\widehat{\beta}$.
Assume statistical model/true relationship

$$Y_i = X_i^T\beta_0 + \epsilon_i \ , i = 1, \ldots, n \ .$$

- $\widehat{\beta}_\rho$: an estimate of (unknown) β_0.
- Y_i's, X_i's observed.
- ϵ_i's noise (unobserved). $\mathbf{E}(\epsilon_i) = 0$, $\text{var}(\epsilon_i) = \sigma_\epsilon^2$

Questions:

1. Confidence interval (CI) for β_0 based on $\widehat{\beta}_\rho$? Probabilistic fluctuation behavior of $\widehat{\beta}_\rho$?
2. Prediction performance: Y_{new},X_{new} new observations:

   Expected Prediction error $= EPE = \mathbf{E}\left((Y_{\text{new}} - X_{\text{new}}^T\widehat{\beta}_\rho)^2\right)$?
3. Risk of estimator $\mathbf{E}\left(\|\widehat{\beta}_\rho - \beta_0\|_2^2\right)$.

Uncertainty assessment

Key issue: stability of estimator $\widehat{\beta}_\rho$ or prediction $X_{\text{new}}^T\widehat{\beta}$.
Assume statistical model/true relationship

$$Y_i = X_i^T\beta_0 + \epsilon_i \ , i = 1, \ldots, n \ .$$

- $\widehat{\beta}_\rho$: an estimate of (unknown) β_0.
- Y_i's, X_i's observed.
- ϵ_i's noise (unobserved). $\mathbf{E}\left(\epsilon_i\right) = 0$, $\text{var}\left(\epsilon_i\right) = \sigma_\epsilon^2$

Questions:

1. Confidence interval (CI) for β_0 based on $\widehat{\beta}_\rho$? Probabilistic fluctuation behavior of $\widehat{\beta}_\rho$?
2. Prediction performance: Y_{new},X_{new} new observations:

   Expected Prediction error $= EPE = \mathbf{E}\left((Y_{\text{new}} - X_{\text{new}}^T\widehat{\beta}_\rho)^2\right)$?
3. Risk of estimator $\mathbf{E}\left(\|\widehat{\beta}_\rho - \beta_0\|_2^2\right)$.

ℓ_2 errors used above; could do it with more general metrics

Main question of talk: **influence of dimensionality on properties of estimator**. Will assume that

$$\frac{p}{n} \to \kappa \neq 0 \ .$$

Main question of talk: **influence of dimensionality on properties of estimator**. Will assume that

$$\frac{p}{n} \to \kappa \neq 0 \ .$$

Connection with random matrices: sample covariance matrix

$$\widehat{\Sigma} = \frac{1}{n} \sum_{i=1}^{n} X_i X_i^T \, , \text{e.g.} X_i \sim \mathcal{N}(0, \mathrm{Id}_p)$$

Workhorse of statistical methods such as Principal Components Analysis (Pearson, 1900's, Hotelling '33). Still very widely used. See Johnstone ICM '06
Main message: **widely different behavior of sample covariance matrix in low ($p/n \to 0$) and high dimension ($p/n \to \kappa > 0$).**

RM illustration; Marcenko-Pastur law

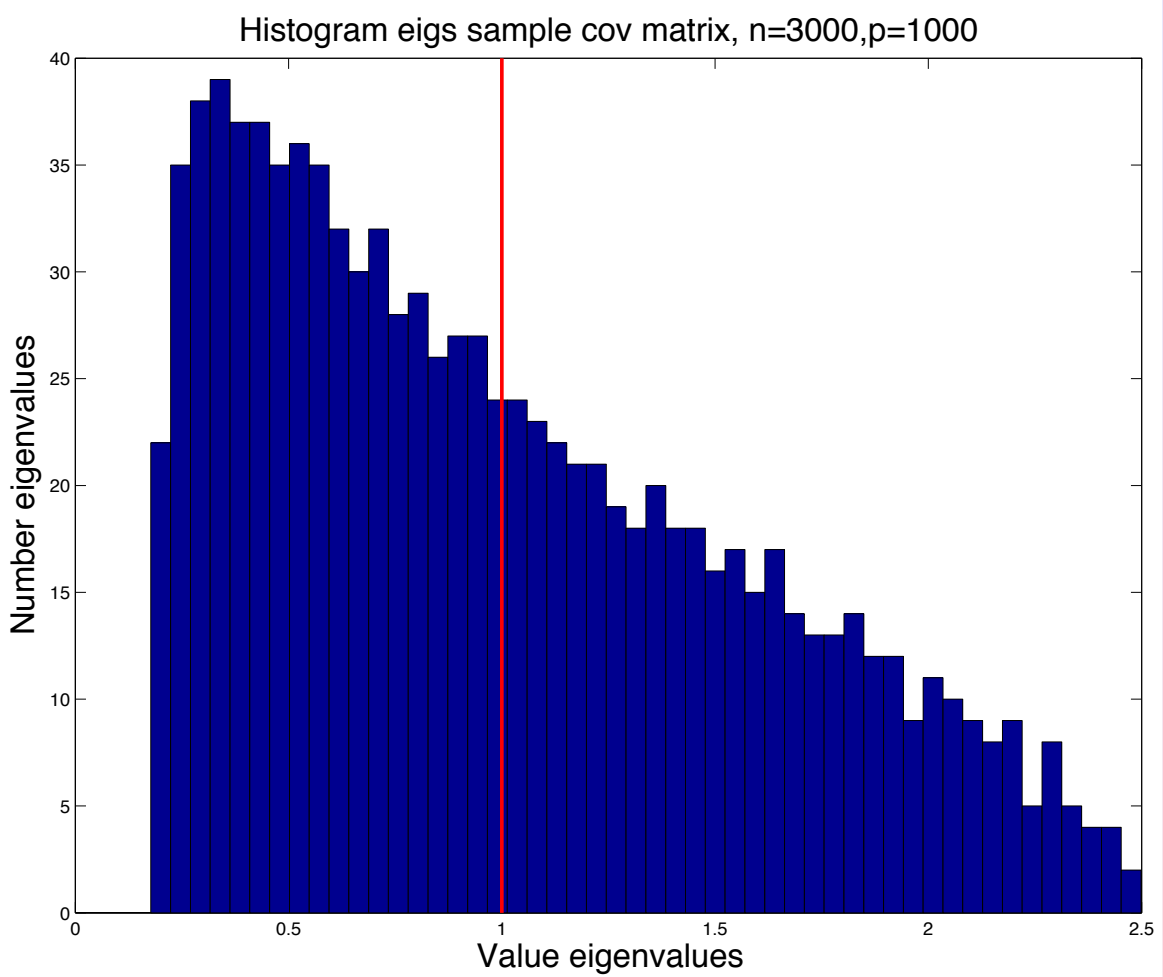


Figure: Histogram of eigenvalues of $\widehat{\Sigma}$, sample covariance matrix, population/true covariance = Id_p; $p/n = .3$ $\widehat{\Sigma}$ "bad" spectral estimate of Σ in high dimensions

RMs and sample covariance matrices

Huge renewed interest in random matrices questions in last 25 years.
To name just a few contributors (in no specific order; list is partial and subjective): Wishart, Wigner, Marcenko, Pastur, Voiculescu, Wachter, Bai, Silverstein, Tracy, Widom, Baik, Deift, Johannson, Soshnikov, Guionnet, Zeitouni, Ledoux, Ben Arous, Péché, Götze, Tikhomirov, (H-T) Yau, Erdös, Tao, Vu, Paul, Pajor, Scherbina, Chatterjee, Diaconis, Evans, Capitaine, Knowles, Yin, Bloemendal, etc...

Questions mostly around behavior of eigenvalues of large random matrices. Ideas developed there informative for understanding high-dimensional linear model?

Case of least-squares: $\rho(x) = x^2$

Well-known that

$$\begin{aligned}\widehat{\beta}_{LS} &= (X^T X)^{-1} X^T Y \\ &= \beta_0 + (X^T X)^{-1} X^T \epsilon \quad \text{(in linear model case)}\end{aligned}$$

Case of least-squares: $\rho(x) = x^2$

Well-known that

$$\begin{aligned}\widehat{\beta}_{LS} &= (X^T X)^{-1} X^T Y \\ &= \beta_0 + (X^T X)^{-1} X^T \epsilon \quad \text{(in linear model case)}\end{aligned}$$

So risk of estimator

$$\begin{aligned}\mathbf{E}\left(\|\widehat{\beta}_{LS} - \beta_0\|_2^2\right) &= \operatorname{trace}\left((X^T X)^{-1}\right)\sigma_\epsilon^2 \\ &\simeq \frac{p/n}{1-p/n}\sigma_\epsilon^2 . \quad \text{if e.g. } X_i \sim \mathcal{N}(0, \mathrm{Id}_p)\end{aligned}$$

Influence of dimensionality very clear! Big contrast with low-dimension

Case of least-squares: $\rho(x) = x^2$

Well-known that

$$\begin{aligned}\widehat{\beta}_{LS} &= (X^T X)^{-1} X^T Y \\ &= \beta_0 + (X^T X)^{-1} X^T \epsilon \quad \text{(in linear model case)}\end{aligned}$$

So risk of estimator

$$\begin{aligned}\mathbf{E}\left(\|\widehat{\beta}_{LS} - \beta_0\|_2^2\right) &= \operatorname{trace}\left((X^T X)^{-1}\right)\sigma_\epsilon^2 \\ &\simeq \frac{p/n}{1-p/n}\sigma_\epsilon^2 . \quad \text{if e.g. } X_i \sim \mathcal{N}(0, \mathrm{Id}_p)\end{aligned}$$

Influence of dimensionality very clear! Big contrast with low-dimension Easy to show no statistical "universality"

Case of least-squares: $\rho(x) = x^2$

Well-known that

$$\begin{aligned}\widehat{\beta}_{LS} &= (X^T X)^{-1} X^T Y \\ &= \beta_0 + (X^T X)^{-1} X^T \epsilon \quad \text{(in linear model case)}\end{aligned}$$

So risk of estimator

$$\begin{aligned}\mathbf{E}\left(\|\widehat{\beta}_{LS} - \beta_0\|_2^2\right) &= \text{trace}\left((X^T X)^{-1}\right)\sigma_\epsilon^2 \\ &\simeq \frac{p/n}{1-p/n}\sigma_\epsilon^2 . \quad \text{if e.g. } X_i \sim \mathcal{N}(0, \text{Id}_p)\end{aligned}$$

Influence of dimensionality very clear! Big contrast with low-dimension Easy to show no statistical "universality" Least-squares case easy because all formulae in closed form. (See work with Koesters for many generalizations to penalized estimators)

Aim of talk

RMs and high-dimensional statistics (κ not close to 0)

- Risk results for robust regression estimators (i.e. general ρ above)
- Optimality results: a new class of natural ρ to use in practice
- Problems with data-driven methods: bootstrap issues in this context, 2nd part of talk
- Many interesting statistical phenomena at play

Why p/n not close to 0?

Aim of talk

RMs and high-dimensional statistics (κ not close to 0)

- Risk results for robust regression estimators (i.e. general ρ above)
- Optimality results: a new class of natural ρ to use in practice
- Problems with data-driven methods: bootstrap issues in this context, 2nd part of talk
- Many interesting statistical phenomena at play

Why p/n not close to 0? 1) often better small sample approximations; 2) often allows comparison of methods at 1st order and not second order; so more dramatic differencing of methods - often consistent with practical knowledge 3) power series vs 1st order approximation 4) problems statistically non-trivial; p/n metric of difficulty

Part I: Risk and optimality results

Based on multiple works, including some with Bean, Bickel, Lim, Yu at UC, Berkeley

Linear regression: basic question

Consider linear regression model:

$$Y_i = X_i^T \beta_0 + \epsilon_i \ , i = 1, \ldots, n \ .$$

Here $Y_i \in \mathbb{R}$, $X_i \in \mathbb{R}^p$, $\beta_0 \in \mathbb{R}^p$ and $\epsilon_i \in \mathbb{R}$.

- Aim: estimate (unknown) β_0.
- Setting: X_i's vectors of predictors (observed). ϵ_i's noise (unobserved).
- Standard method: (say $p < n$): estimate β_0 by

$$\widehat{\beta}_\rho = \text{argmin}_{\beta \in \mathbb{R}^p} \sum_{i=1}^{n} \rho(Y_i - X_i^T \beta) \ , \text{ where } \rho \text{ is a function.}$$

Linear regression: basic question

Consider linear regression model:

$$Y_i = X_i^T \beta_0 + \epsilon_i \ , i = 1, \ldots, n \ .$$

Here $Y_i \in \mathbb{R}$, $X_i \in \mathbb{R}^p$, $\beta_0 \in \mathbb{R}^p$ and $\epsilon_i \in \mathbb{R}$.

- Aim: estimate (unknown) β_0.
- Setting: X_i's vectors of predictors (observed). ϵ_i's noise (unobserved).
- Standard method: (say $p < n$): estimate β_0 by

$$\widehat{\beta}_\rho = \text{argmin}_{\beta \in \mathbb{R}^p} \sum_{i=1}^{n} \rho(Y_i - X_i^T \beta) \ , \text{ where } \rho \text{ is a function.}$$

Questions:

- **can we do inference on β_0?, i.e confidence intervals/error bars based on $\widehat{\beta}_\rho$**
- **how to pick ρ? (e.g if know something about errors)**
- **Influence of dimension or ratio p/n on these questions?**

How to pick ρ? Optimality questions

Very classical question.
Much work on this starting with Fisher in 30's.
Very nice work in the late 60's, 70's, 80's on properties of these estimators.
Contributors include: Relles ('68), Huber ('72), Portnoy ('84-85), Mammen ('89), Yohai, Bickel, etc... More recently e.g Duembgen, Samworth and Schuhmacher ('11)

How to pick ρ? Optimality questions

Very classical question.
Much work on this starting with Fisher in 30's.
Very nice work in the late 60's, 70's, 80's on properties of these estimators.
Contributors include: Relles ('68), Huber ('72), Portnoy ('84-85), Mammen ('89), Yohai, Bickel, etc... More recently e.g Duembgen, Samworth and Schuhmacher ('11)
Short answer: in low dimension, if ϵ_i's i.i.d f_ϵ, pick

$$\boxed{\rho = -\log f_\epsilon \ .}$$

How to pick ρ? Optimality questions

Very classical question.
Much work on this starting with Fisher in 30's.
Very nice work in the late 60's, 70's, 80's on properties of these estimators.
Contributors include: Relles ('68), Huber ('72), Portnoy ('84-85), Mammen ('89), Yohai, Bickel, etc... More recently e.g Duembgen, Samworth and Schuhmacher ('11)
Short answer: in low dimension, if ϵ_i's i.i.d f_ϵ, pick

$$\boxed{\rho = -\log f_\epsilon \ .}$$

Remarkable fact: independent of design matrix, X. (X is $n \times p$, i-th row is X_i^T); consistent with **maximum likelihood** ideas

How to pick ρ? Optimality questions

Very classical question.
Much work on this starting with Fisher in 30's.
Very nice work in the late 60's, 70's, 80's on properties of these estimators.
Contributors include: Relles ('68), Huber ('72), Portnoy ('84-85), Mammen ('89), Yohai, Bickel, etc... More recently e.g Duembgen, Samworth and Schuhmacher ('11)
Short answer: in low dimension, if ϵ_i's i.i.d f_ϵ, pick

$$\boxed{\rho = -\log f_\epsilon \ .}$$

Remarkable fact: independent of design matrix, X. (X is $n \times p$, i-th row is X_i^T); consistent with **maximum likelihood** ideas
Classic result of Huber ('70s) is that (under regularity conditions and asymptotically as $n \to \infty$)

$$\boxed{\mathrm{cov}\left(\widehat{\beta}_\rho\right) = (X^T X)^{-1} \frac{\mathbf{E}\left(\psi^2(\epsilon)\right)}{[\mathbf{E}\left(\psi'(\epsilon)\right)]^2}, \psi = \rho' \ .}$$

An example: double exponential errors

ϵ_i's double exponential, i.e $f_\epsilon(x) = \exp(-|x|)/2$.
According to classical results/intuition, i.e. above results, ℓ_1 should be optimal
So natural to think that: need to solve

$$\widehat{\beta}_{\ell_1} = \text{argmin}_{\beta \in \mathbb{R}^p} \sum_{i=1}^{n} |Y_i - X_i^T \beta| \ .$$

A proposal for ρ

ϵ with log-concave errors; $p/n = \kappa$ not close to 0

Let $p_2(x) = x^2/2$. Suppose ϵ has log-concave density, f_ϵ. For sake of argument, assume f_ϵ known. Also, $p/n < 1$.
For reasons explained later, let us try

$$\boxed{\rho_{opt} = (p_2 + r_{opt}^2 \log \phi_{r_{opt}} \star f_\epsilon)^* - p_2 \ .}$$

where $r_{opt} = \min\{r : r^2 I_\epsilon(r) = p/n\}$.

A proposal for ρ

ϵ with log-concave errors; $p/n = \kappa$ not close to 0

Let $p_2(x) = x^2/2$. Suppose ϵ has log-concave density, f_ϵ. For sake of argument, assume f_ϵ known. Also, $p/n < 1$.
For reasons explained later, let us try

$$\boxed{\rho_{opt} = (p_2 + r_{opt}^2 \log \phi_{r_{opt}} \star f_\epsilon)^* - p_2 \ .}$$

$$\text{where } r_{opt} = \min\{r : r^2 I_\epsilon(r) = p/n\} \ .$$

ϕ_r: gaussian density with variance r^2.
$I_\epsilon(r)$: Fisher information of $\phi_r \star f_\epsilon$
$g^*(x) = \sup_y(xy - g(y))$, Fenchel-Legendre dual of g

Note: ρ_{opt} depends on p/n! Also, convex.

Comparison ρ_{opt} to ℓ_1, double exponential errors

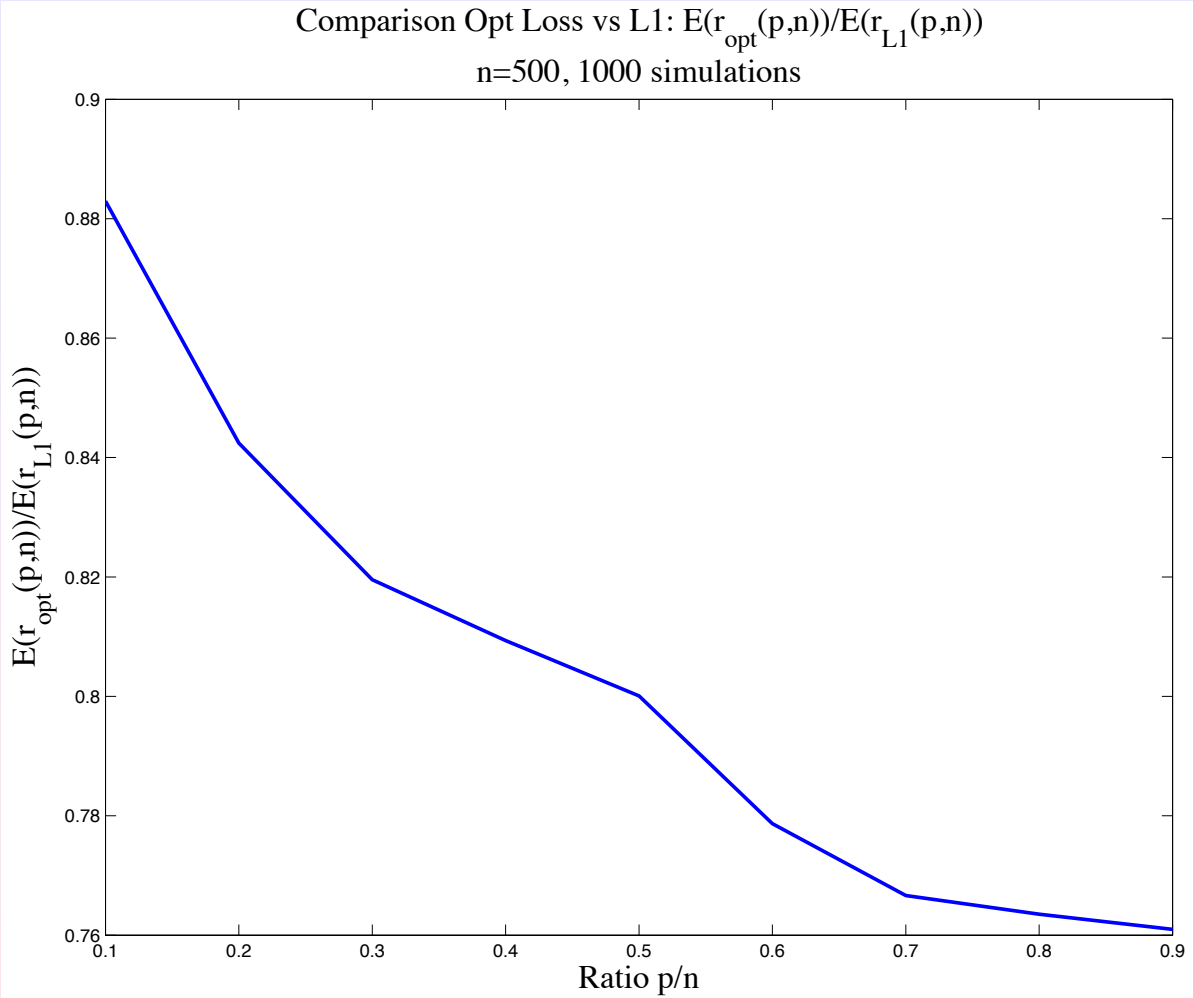


Figure: $\mathbf{E}\left(\|\widehat{\beta}_{opt} - \beta_0\|_2\right) / \mathbf{E}\left(\|\widehat{\beta}_{l_1} - \beta_0\|_2\right)$, double exponential errors.

Ratio always less than 1: ρ_{opt} beats ℓ_1!

Characterization of solution of robust regression problem

Suppose $p/n \to \kappa \in (0,1)$. $X_i \overset{iid}{\backsim} (0, \mathrm{Id}_p)$, with i.i.d entries + moment conditions.

Characterization of solution of robust regression problem

Suppose $p/n \to \kappa \in (0,1)$. $X_i \overset{iid}{\backsim} (0, \mathrm{Id}_p)$, with i.i.d entries + moment conditions.

Theorem

Under regularity conditions on $\{\epsilon_i\}$ *and* ρ *(convex),* $\|\widehat{\beta}_\rho - \beta_0\|_2$ *is asymptotically deterministic. Call* $r_\rho(\kappa)$ *its limit and* $\hat{z}_\epsilon = \epsilon + r_\rho(\kappa)Z$*, where* $Z \sim \mathcal{N}(0,1)$*, independent of* ϵ*. For* c *deterministic, we have*

$$\begin{cases} \mathbf{E}\left([prox(c\rho)]'(\hat{z}_\epsilon)\right) & = 1-\kappa\ , \\ \kappa r_\rho^2(\kappa) & = \mathbf{E}\left([\hat{z}_\epsilon - prox(c\rho)(\hat{z}_\epsilon)]^2\right)\ . \end{cases}$$

Characterization of solution of robust regression problem

Suppose $p/n \to \kappa \in (0,1)$. $X_i \overset{iid}{\backsim} (0, \mathrm{Id}_p)$, with i.i.d entries + moment conditions.

Theorem

Under regularity conditions on $\{\epsilon_i\}$ *and* ρ *(convex),* $\|\widehat{\beta}_\rho - \beta_0\|_2$ *is asymptotically deterministic. Call* $r_\rho(\kappa)$ *its limit and* $\hat{z}_\epsilon = \epsilon + r_\rho(\kappa) Z$*, where* $Z \sim \mathcal{N}(0,1)$*, independent of* ϵ*. For c deterministic, we have*

$$\begin{cases} \mathbf{E}\left([prox(c\rho)]'(\hat{z}_\epsilon)\right) & = 1-\kappa \ , \\ \kappa r_\rho^2(\kappa) & = \mathbf{E}\left([\hat{z}_\epsilon - prox(c\rho)(\hat{z}_\epsilon)]^2\right) \ . \end{cases}$$

By definition, (Moreau '65), for convex function f,

$$\mathrm{prox}(f)(x) = \mathrm{argmin}_y \left(f(y) + \frac{1}{2}(x-y)^2 \right) \ .$$

Characterization of solution of robust regression problem

Suppose $p/n \to \kappa \in (0,1)$. $X_i \overset{iid}{\backsim} (0, \mathrm{Id}_p)$, with i.i.d entries + moment conditions.

Theorem

Under regularity conditions on $\{\epsilon_i\}$ *and* ρ *(convex),* $\|\widehat{\beta}_\rho - \beta_0\|_2$ *is asymptotically deterministic. Call* $r_\rho(\kappa)$ *its limit and* $\hat{z}_\epsilon = \epsilon + r_\rho(\kappa) Z$*, where* $Z \sim \mathcal{N}(0,1)$*, independent of* ϵ*. For c deterministic, we have*

$$\begin{cases} \mathbf{E}\left([prox(c\rho)]'(\hat{z}_\epsilon)\right) & = 1-\kappa \ , \\ \kappa r_\rho^2(\kappa) & = \mathbf{E}\left([\hat{z}_\epsilon - prox(c\rho)(\hat{z}_\epsilon)]^2\right) \ . \end{cases}$$

By definition, (Moreau '65), for convex function f,

$$\mathrm{prox}(f)(x) = \mathrm{argmin}_y \left(f(y) + \frac{1}{2}(x-y)^2 \right) \ .$$

Many generalizations exist... Risk completely different from Huber's characterization!

On the proximal mapping (Moreau '65)

prox operation natural in convex analysis, optimization etc...
Applies to convex functions. Some examples: $c > 0 \in \mathbb{R}_+$:

1. if $f(x) = x^2$, $\text{prox}(cf)[x] = \frac{x}{1+2c}$
2. if $f(x) = |x|$, $\text{prox}(cf)[x] = sgn(x)(|x| - c)_+$, i.e soft-thresholding function

In general, if f has a subdifferential ∂f,

$$\text{prox}(f) = (\text{Id} + \partial f)^{-1} .$$

Remarkable fact: $\text{prox}(f)$ single valued even if ∂f is multivalued.

On the proximal mapping (Moreau '65)

prox operation natural in convex analysis, optimization etc...
Applies to convex functions. Some examples: $c > 0 \in \mathbb{R}_+$:

1. if $f(x) = x^2$, $\text{prox}(cf)[x] = \frac{x}{1+2c}$
2. if $f(x) = |x|$, $\text{prox}(cf)[x] = sgn(x)(|x| - c)_+$, i.e soft-thresholding function

In general, if f has a subdifferential ∂f,

$$\text{prox}(f) = (\text{Id} + \partial f)^{-1} .$$

Remarkable fact: $\text{prox}(f)$ single valued even if ∂f is multivalued.
See also related work of Donoho-Montanari.

Call $e_i = Y_i - X_i^T \widehat{\beta}_\rho$, the i-th residual. In the asymptotic limit,

$$\boxed{e_i \overset{\mathcal{L}}{=} \text{prox}(c\rho)(\epsilon_i + r_\rho(\kappa) Z_i)}$$

where $Z_i \sim \mathcal{N}(0,1)$ independent of ϵ_i.

Call $e_i = Y_i - X_i^T \widehat{\beta}_\rho$, the i-th residual. In the asymptotic limit,

$$\boxed{e_i \overset{\mathcal{L}}{=} \text{prox}(c\rho)(\epsilon_i + r_\rho(\kappa) Z_i)}$$

where $Z_i \sim \mathcal{N}(0,1)$ independent of ϵ_i.

1. Non-trivial relationship between ρ, distribution of ϵ_i and distribution of e_i.
2. Very different from classical setting of p/n close to 0, when $e_i \simeq \epsilon_i$.
3. Bootstrap? See later

Numerical experiments: $\rho(x) = |x|$, $\epsilon \sim \mathcal{N}(0, \sigma_\epsilon^2)$

Can solve the previous system and find:
Let φ be normal density, Φ normal cdf and

$$\zeta(t) = 2\Phi^{-1}(t)\left(\varphi[\Phi^{-1}(t)] - \Phi^{-1}(t)(1-t)\right) \ ,$$

Previous system leads to: if $\kappa = \lim p/n$, $\kappa < 1$,

$$r_{\ell_1}^2(\kappa) = \frac{\kappa - \zeta([1+\kappa]/2)}{\zeta([1+\kappa]/2)}\sigma_\epsilon^2 \ .$$

Checking previous prediction numerically

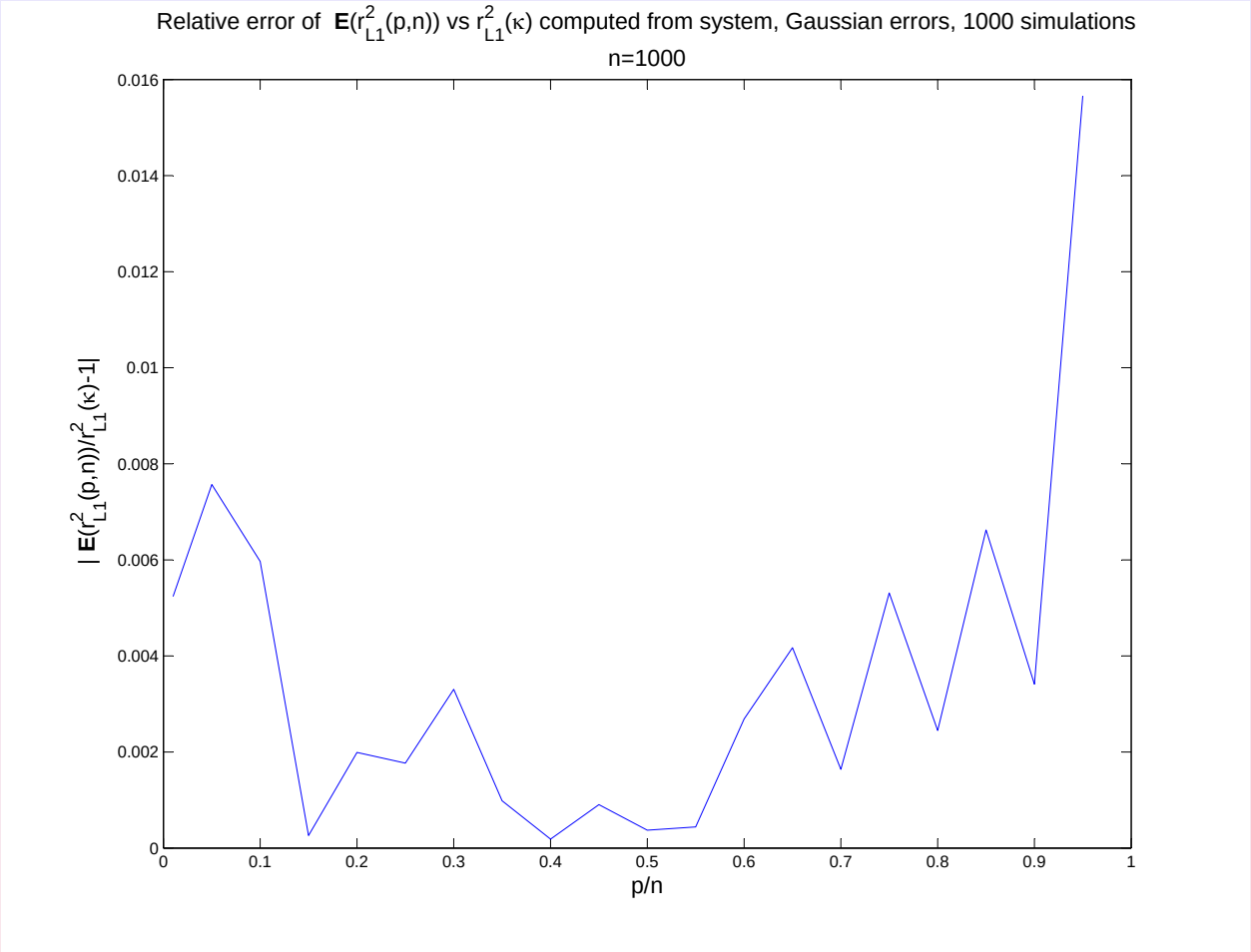


Figure: Relative errors $\left|\frac{\mathbf{E}\left(r_{\ell_1}^2(p,n)\right)}{r_{\ell_1}^2(\kappa)} - 1\right|$, Gaussian errors, 1000 simulations

Extensions on previous result

- Elliptical models: other models for X_i with same covariance but different Euclidean geometry. Risk very different for those models. Hence not only covariance of X_i matters, but indeed geometry of point cloud (similar to RMT). **No statistical universality**
- to handle **weighted regression** (i.e $\rho \to w_i \rho$) replace ρ by $w_i \rho$ everywhere. (Consequence for **non-parametric bootstrap** see later)
- Easy to extend method to penalized problems (add $sP(\beta)$ [$s \in \mathbb{R}_+$, and e.g $P(\beta) = \sum_{i=1}^p f_i(\beta_i)$] to original optimization problem or Generalized Linear Models) Results not presented here for lack of time. Prox of loss and penalty are essential features again.

Key ideas

Analyze, if $\psi = \rho'$ and Y_i's are responses:

$$\widehat{\beta}_\rho : \sum_{i=1}^{n} X_i \, \psi(Y_i - X_i^T \widehat{\beta}_\rho) = 0_p \, .$$

Key elements

- concentration of quadratic forms in X_i; consequence: geometry of dataset influences crucially result. No statistical universality.
- leave-one-out ideas.
- martingale ideas (à la Burkholder/Efron-Stein).
- connection with ideas in random matrix theory and convex analysis

Key ideas

Analyze, if $\psi = \rho'$ and Y_i's are responses:

$$\widehat{\beta}_\rho : \sum_{i=1}^{n} X_i \, \psi(Y_i - X_i^T \widehat{\beta}_\rho) = 0_p \, .$$

Key elements

- concentration of quadratic forms in X_i; consequence: geometry of dataset influences crucially result. No statistical universality.
- leave-one-out ideas.
- martingale ideas (à la Burkholder/Efron-Stein).
- connection with ideas in random matrix theory and convex analysis

Surprise : it is a random matrix problem!
In particular, constant c turns out to be - basically - $\frac{1}{n}\text{trace}\left(S^{-1}\right)$, where $S = \frac{1}{n}\sum_{i=1}^{n} w_i X_i X_i^T$ and $w_i = \psi'(\epsilon_i - X_i^T \widehat{\beta}_\rho)$.

Optimizing $r_\rho(\kappa)$ over ρ

$r_\rho(\kappa) = \lim_{n\to\infty} \|\widehat{\beta}_\rho - \beta_0\|$

Recall system: if $\hat{z}_\epsilon = \epsilon + r_\rho(\kappa) Z$, with $Z \sim \mathcal{N}(0,1)$,

$$\left\{ \begin{array}{rl} \mathbf{E}\left([\mathrm{prox}(c\rho)]'(\hat{z}_\epsilon)\right) & = 1-\kappa \ , \\ \kappa r_\rho^2(\kappa) & = \mathbf{E}\left([\hat{z}_\epsilon - \mathrm{prox}(c\rho)(\hat{z}_\epsilon)]^2\right) \ . \end{array} \right.$$

Possible to optimize $r_\rho(\kappa)$ over ρ!

Strategy

1. Write problem as feasibility problem in $r_\rho(\kappa)$

Strategy

1. Write problem as feasibility problem in $r_\rho(\kappa)$
2. Use Moreau's fundamental prox-identity:

$$x = \text{prox}(\rho)(x) + \text{prox}(\rho^*)(x) \ .$$

to rewrite system. Natural variable for problem: $\text{prox}([c\rho]^*)$; gets rid of c

Strategy

1. Write problem as feasibility problem in $r_\rho(\kappa)$
2. Use Moreau's fundamental prox-identity:

$$x = \text{prox}(\rho)(x) + \text{prox}(\rho^*)(x) \ .$$

   to rewrite system. Natural variable for problem: $\text{prox}([c\rho]^*)$; gets rid of c
3. Work on simplified problem... in space of proximal mappings and not directly on loss functions

Strategy

1. Write problem as feasibility problem in $r_\rho(\kappa)$
2. Use Moreau's fundamental prox-identity:

$$x = \text{prox}(\rho)(x) + \text{prox}(\rho^*)(x) \ .$$

   to rewrite system. Natural variable for problem: $\text{prox}([c\rho]^*)$; gets rid of c
3. Work on simplified problem... in space of proximal mappings and not directly on loss functions
4. Go from optimal $\text{prox}(\rho^*)$ to optimal ρ

Strategy

1. Write problem as feasibility problem in $r_\rho(\kappa)$
2. Use Moreau's fundamental prox-identity:

$$x = \text{prox}(\rho)(x) + \text{prox}(\rho^*)(x) \ .$$

   to rewrite system. Natural variable for problem: $\text{prox}([c\rho]^*)$; gets rid of c
3. Work on simplified problem... in space of proximal mappings and not directly on loss functions
4. Go from optimal $\text{prox}(\rho^*)$ to optimal ρ

Strategy

1. Write problem as feasibility problem in $r_\rho(\kappa)$
2. Use Moreau's fundamental prox-identity:

$$x = \text{prox}(\rho)(x) + \text{prox}(\rho^*)(x) \ .$$

   to rewrite system. Natural variable for problem: $\text{prox}([c\rho]^*)$; gets rid of c
3. Work on simplified problem... in space of proximal mappings and not directly on loss functions
4. Go from optimal $\text{prox}(\rho^*)$ to optimal ρ

Following this strategy, we get that, if $p_2(x) = x^2/2$, if $-\log f_\epsilon$ convex,

$$\boxed{\rho_{opt} = (p_2 + r_{opt}^2 \log \phi_{r_{opt}} \star f_\epsilon)^* - p_2 \ .}$$

where $r_{opt} = \min\{r : r^2 I_\epsilon(r) = p/n\}$.

Further remarks on optimal loss

- In general, ρ_{opt} changes with p/n! Adapts to difficulty of problem.
- Not maximum likelihood. (Contrast with classical theory) Though recovers ML when $p/n \to 0$.
- For Gaussian errors, ℓ_2 still optimal, no matter p/n
- As $p/n \to 1$, performance of ℓ_2 becomes optimal (at least at the level of equivalence b/t two diverging sequences)
- However, limit of optimal loss when $p/n \to 1$ not ℓ_2

Also have data-driven approaches now.

Plot for $p/n = .5$, double exponential errors

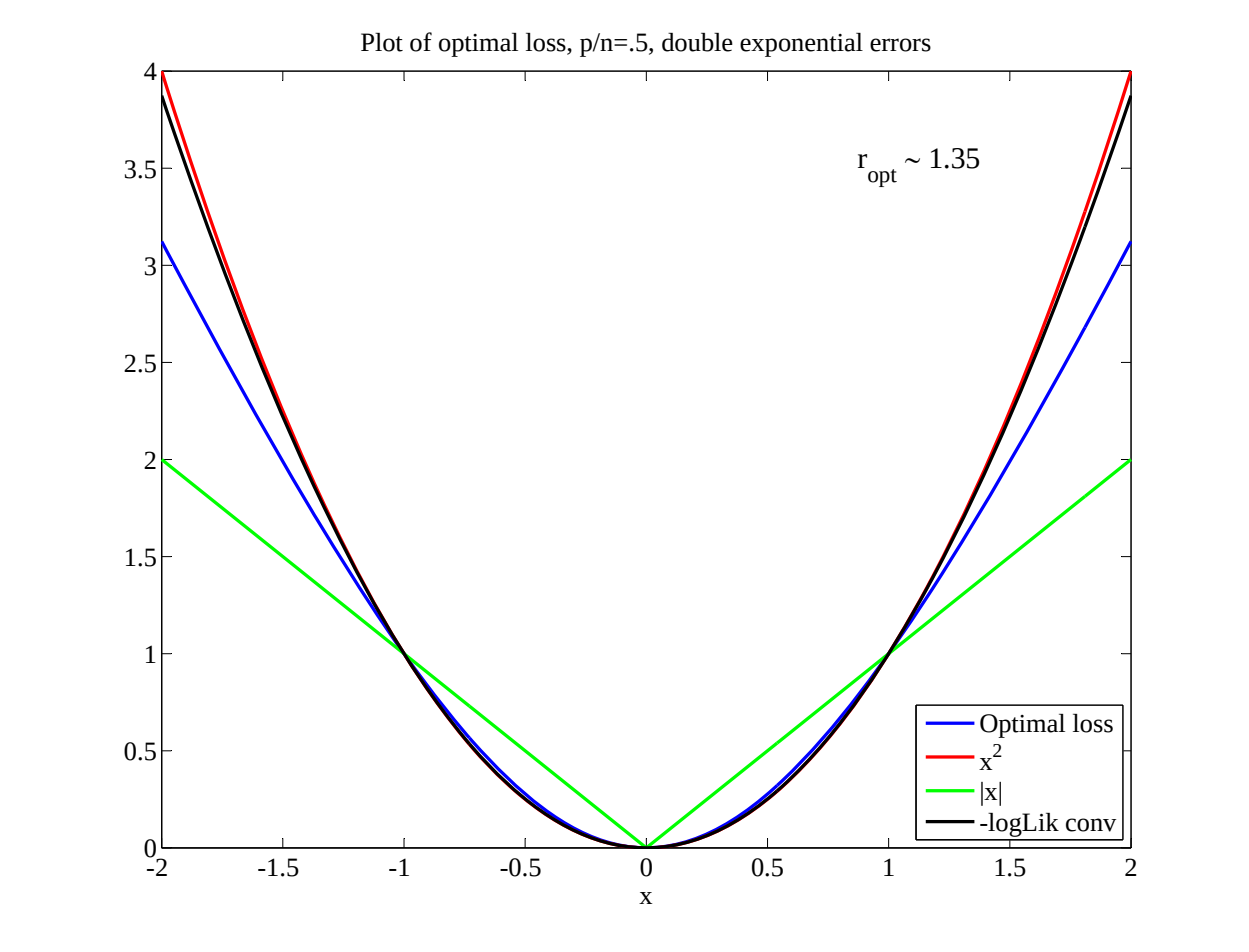


Figure: Some "natural" loss functions; optimal loss is a bit like a Huber. Changes with p/n

Statistical limitations of model

Considerable!

- Model under consideration very special - Euclidean geometry
- Alternative design: X_i has only 1 non-zero coordinate. $X_i(k) = 1$, k picked at random on $\{1, \ldots, p\}$. Then optimality theory and fluctuation theory totally different from above.
- Of course, this model is very far from "classical" RMT...

Many other limitations (inadmissibility of estimators... see also Stein ('60), Baranchik ('73), Klebanov-Jureckova ('97))
But still interesting in furthering our understanding of high-dimensional statistics.

Part II: Bootstrap results

Question: does the bootstrap work for moderately difficult statistical problems?
Joint with Elizabeth Purdom, UC Berkeley

Model: $Y_i \in \mathbb{R}$, $X_i \in \mathbb{R}^p$,

$$Y_i = X_i^T \beta_0 + \epsilon_i \, , 1 \leq i \leq n \, .$$

$$\widehat{\beta}_\rho = \text{argmin}_{\beta \in \mathbb{R}^p} \sum_{i=1}^{n} \rho(Y_i - X_i^T \beta) \, .$$

Simplest question: as before, can we get a CI for $\beta_0(1)$ based on $\widehat{\beta}_\rho(1)$? (Confidence interval: random interval covering true parameter ($\beta_0(1)$ here) with pre-specified probability.)

Model: $Y_i \in \mathbb{R}$, $X_i \in \mathbb{R}^p$,

$$Y_i = X_i^T \beta_0 + \epsilon_i \ , 1 \leq i \leq n \ .$$

$$\widehat{\beta}_\rho = \text{argmin}_{\beta \in \mathbb{R}^p} \sum_{i=1}^{n} \rho(Y_i - X_i^T \beta) \ .$$

Simplest question: as before, can we get a CI for $\beta_0(1)$ based on $\widehat{\beta}_\rho(1)$? (Confidence interval: random interval covering true parameter ($\beta_0(1)$ here) with pre-specified probability.)
I.e. can we understand the (probabilistic) fluctuation behavior/law of

$$\widehat{\beta}_\rho - \beta_0$$

by resampling, i.e. doing simulations from existing data?

Bootstrap

Bootstrap (Efron, '79): Very powerful methodology. Huge impact on methodological, applied and theoretical stats. (100's of papers on theory and practice: early theoretical work of Bickel, Freedman, Beran, Srivastava, Giné, Zinn, Götze, van Zwet, etc...)
Nowadays very useful to introduce inferential ideas, e.g in intro data science class.

Bootstrap

Bootstrap (Efron, '79): Very powerful methodology. Huge impact on methodological, applied and theoretical stats. (100's of papers on theory and practice: early theoretical work of Bickel, Freedman, Beran, Srivastava, Giné, Zinn, Götze, van Zwet, etc...)
Nowadays very useful to introduce inferential ideas, e.g in intro data science class.
Usual take-away message: "bootstrap" works for smooth statistics, i.e. "smooth functionals" of data-generating distribution. See statistics textbooks, e.g. van der Vaart ('99).
For bootstrap problems related to ours in low-d, see Bickel-Freedman, Shorack, Portnoy, Mammen, Jeff Wu in 80s-90s.

Bootstrap

Bootstrap (Efron, '79): Very powerful methodology. Huge impact on methodological, applied and theoretical stats. (100's of papers on theory and practice: early theoretical work of Bickel, Freedman, Beran, Srivastava, Giné, Zinn, Götze, van Zwet, etc...)
Nowadays very useful to introduce inferential ideas, e.g in intro data science class.
Usual take-away message: "bootstrap" works for smooth statistics, i.e. "smooth functionals" of data-generating distribution. See statistics textbooks, e.g. van der Vaart ('99).
For bootstrap problems related to ours in low-d, see Bickel-Freedman, Shorack, Portnoy, Mammen, Jeff Wu in 80s-90s.
As seen above, fluctuation theory not completely trivial in our case. Good test case for bootstrap.

Review: How to bootstrap in regression?

v1: Bootstrapping from residuals

Motto: copy the data-generating process.
Model: $Y_i \in \mathbb{R}$, $X_i \in \mathbb{R}^p$,

$$Y_i = X_i^T \beta_0 + \epsilon_i \ , 1 \leq i \leq n \ .$$

What's random? ϵ_i in this context; they are i.i.d.
X_i assumed "fixed" in this example.
So **bootstrap from the residuals:**

1. estimate β_0 by $\widehat{\beta}_\rho$

Review: How to bootstrap in regression?

v1: Bootstrapping from residuals

Motto: copy the data-generating process.
Model: $Y_i \in \mathbb{R}$, $X_i \in \mathbb{R}^p$,

$$Y_i = X_i^T \beta_0 + \epsilon_i \ , 1 \leq i \leq n \ .$$

What's random? ϵ_i in this context; they are i.i.d.
X_i assumed "fixed" in this example.
So **bootstrap from the residuals:**

1. estimate β_0 by $\widehat{\beta}_\rho$
2. estimate ϵ_i by e_i's; typically $e_i = Y_i - X_i^T \widehat{\beta}_\rho$

Review: How to bootstrap in regression?

v1: Bootstrapping from residuals

Motto: copy the data-generating process.
Model: $Y_i \in \mathbb{R}$, $X_i \in \mathbb{R}^p$,

$$Y_i = X_i^T \beta_0 + \epsilon_i \ , 1 \leq i \leq n \ .$$

What's random? ϵ_i in this context; they are i.i.d.
X_i assumed "fixed" in this example.
So **bootstrap from the residuals:**

1. estimate β_0 by $\widehat{\beta}_\rho$
2. estimate ϵ_i by e_i's; typically $e_i = Y_i - X_i^T \widehat{\beta}_\rho$
3. Repeat for $b = 1, \ldots, B$

Review: How to bootstrap in regression?

v1: Bootstrapping from residuals

Motto: copy the data-generating process.
Model: $Y_i \in \mathbb{R}$, $X_i \in \mathbb{R}^p$,

$$Y_i = X_i^T \beta_0 + \epsilon_i \ , 1 \leq i \leq n \ .$$

What's random? ϵ_i in this context; they are i.i.d.
X_i assumed "fixed" in this example.
So **bootstrap from the residuals:**

1. estimate β_0 by $\widehat{\beta}_\rho$
2. estimate ϵ_i by e_i's; typically $e_i = Y_i - X_i^T \widehat{\beta}_\rho$
3. Repeat for $b = 1, \ldots, B$
   1. Get new errors $e^*_{i,b}$ by sampling i.i.d at random from $\{e_i\}_{i=1}^n$

Review: How to bootstrap in regression?

v1: Bootstrapping from residuals

Motto: copy the data-generating process.
Model: $Y_i \in \mathbb{R}$, $X_i \in \mathbb{R}^p$,

$$Y_i = X_i^T \beta_0 + \epsilon_i \ , 1 \leq i \leq n \ .$$

What's random? ϵ_i in this context; they are i.i.d.
X_i assumed "fixed" in this example.
So **bootstrap from the residuals:**

1. estimate β_0 by $\widehat{\beta}_\rho$
2. estimate ϵ_i by e_i's; typically $e_i = Y_i - X_i^T \widehat{\beta}_\rho$
3. Repeat for $b = 1, \ldots, B$
   1. Get new errors $e^*_{i,b}$ by sampling i.i.d at random from $\{e_i\}_{i=1}^n$
   2. Get new dataset $Y^*_{i,b} = X_i^T \widehat{\beta}_\rho + e^*_{i,b}$

Review: How to bootstrap in regression?

v1: Bootstrapping from residuals

Motto: copy the data-generating process.
Model: $Y_i \in \mathbb{R}$, $X_i \in \mathbb{R}^p$,

$$Y_i = X_i^T \beta_0 + \epsilon_i \ , 1 \leq i \leq n \ .$$

What's random? ϵ_i in this context; they are i.i.d.
X_i assumed "fixed" in this example.
So **bootstrap from the residuals:**

1. estimate β_0 by $\widehat{\beta}_\rho$
2. estimate ϵ_i by e_i's; typically $e_i = Y_i - X_i^T \widehat{\beta}_\rho$
3. Repeat for $b = 1, \ldots, B$
   1. Get new errors $e^*_{i,b}$ by sampling i.i.d at random from $\{e_i\}_{i=1}^n$
   2. Get new dataset $Y^*_{i,b} = X_i^T \widehat{\beta}_\rho + e^*_{i,b}$
   3. Fit this new dataset to get $\widehat{\beta}^*_b$

Review: How to bootstrap in regression?

v1: Bootstrapping from residuals

Motto: copy the data-generating process.
Model: $Y_i \in \mathbb{R}$, $X_i \in \mathbb{R}^p$,

$$Y_i = X_i^T \beta_0 + \epsilon_i \ , 1 \leq i \leq n \ .$$

What's random? ϵ_i in this context; they are i.i.d.
X_i assumed "fixed" in this example.
So **bootstrap from the residuals:**

1. estimate β_0 by $\widehat{\beta}_\rho$
2. estimate ϵ_i by e_i's; typically $e_i = Y_i - X_i^T \widehat{\beta}_\rho$
3. Repeat for $b = 1, \ldots, B$
   1. Get new errors $e^*_{i,b}$ by sampling i.i.d at random from $\{e_i\}_{i=1}^n$
   2. Get new dataset $Y^*_{i,b} = X_i^T \widehat{\beta}_\rho + e^*_{i,b}$
   3. Fit this new dataset to get $\widehat{\beta}^*_b$

Review: How to bootstrap in regression?

v1: Bootstrapping from residuals

Motto: copy the data-generating process.
Model: $Y_i \in \mathbb{R}$, $X_i \in \mathbb{R}^p$,

$$Y_i = X_i^T \beta_0 + \epsilon_i \ , 1 \leq i \leq n \ .$$

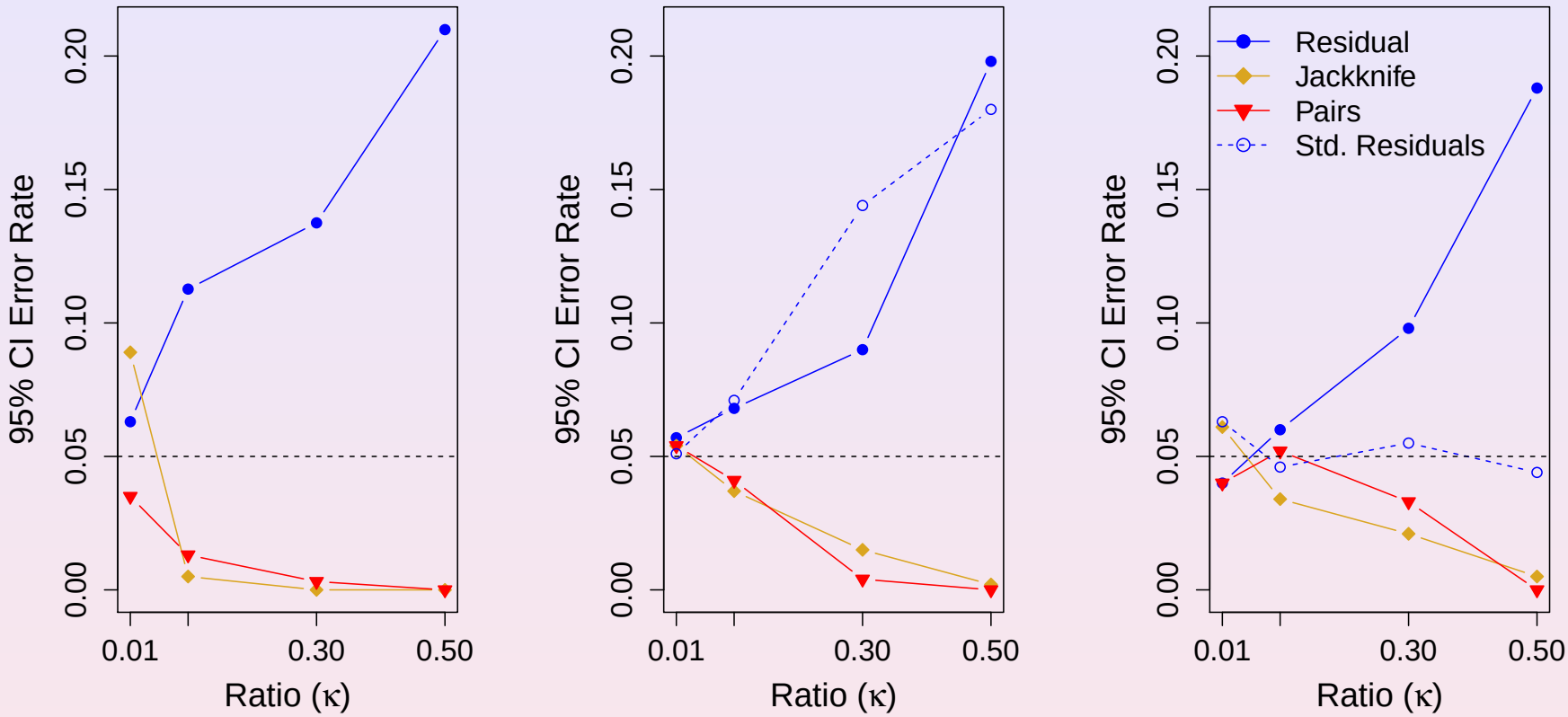
What's random? ϵ_i in this context; they are i.i.d.
X_i assumed "fixed" in this example.
So **bootstrap from the residuals:**

1. estimate β_0 by $\widehat{\beta}_\rho$
2. estimate ϵ_i by e_i's; typically $e_i = Y_i - X_i^T \widehat{\beta}_\rho$
3. Repeat for $b = 1, \ldots, B$
   1. Get new errors $e^*_{i,b}$ by sampling i.i.d at random from $\{e_i\}_{i=1}^n$
   2. Get new dataset $Y^*_{i,b} = X_i^T \widehat{\beta}_\rho + e^*_{i,b}$
   3. Fit this new dataset to get $\widehat{\beta}^*_b$

Do inference using $\{\widehat{\beta}^*_b - \widehat{\beta}_\rho\}_{b=1}^B$ (proxy for distribution of $\widehat{\beta}_\rho - \beta_0$)

Bootstrapping from the residuals

$\epsilon_i \overset{iid}{\sim} \mathcal{N}(0,1)$



(a) L_1 loss: $\rho(x) = |x|$ (b) Huber loss (c) L_2 loss: $\rho(x) = x^2$

Figure: **Performance of 95% confidence intervals of β_1 : $n = 500$, 1,000 simulations** Residuals method is anti-conservative!

In LS case: if residuals $e = \{e_i\}_{i=1}^n$, suggestion for resampling (see e.g Davison-Hinkley '97, based on closed form expressions): use

$$\tilde{e}_i = \frac{e_i}{\sqrt{1 - H_{i,i}}}, H = X(X^T X)^{-1} X^T$$

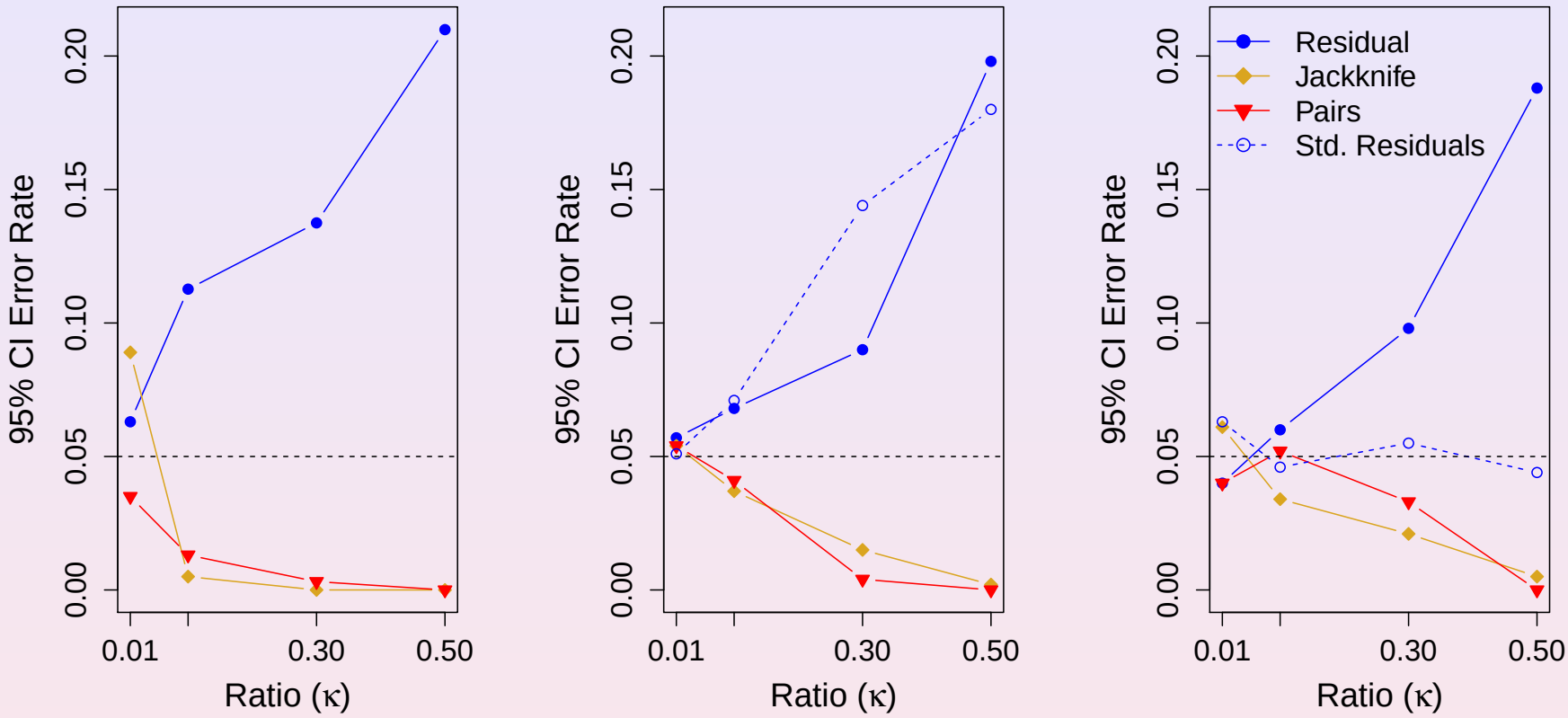
In LS case: if residuals $e = \{e_i\}_{i=1}^n$, suggestion for resampling (see e.g Davison-Hinkley '97, based on closed form expressions): use

$$\tilde{e}_i = \frac{e_i}{\sqrt{1 - H_{i,i}}}, H = X(X^T X)^{-1} X^T$$

In low-dimension, this correction is minimal; in high-d, Gaussian case, $H_{i,i} \simeq \frac{p}{n}$: non-negligible correction

Bootstrapping from the standardized residuals

$\epsilon_i \overset{iid}{\sim} \mathcal{N}(0,1)$; Gaussian design



(a) L_1 loss: $\rho(x) = |x|$ (b) Huber loss (c) L_2 loss: $\rho(x) = x^2$

Figure: **Performance of 95% confidence intervals of β_1 : $n = 500$, 1,000 simulations** Method works for L_2; standardization for Huber (see McKean et al. '93) not effective.

Call $e_i = Y_i - \widehat{\beta}_\rho^T X_i$, the i-th residual. In the asymptotic limit,

$$\boxed{e_i \overset{\mathcal{L}}{=} \mathrm{prox}(c\rho)(\epsilon_i + r_\rho(\kappa) Z_i) \ , Z_i \sim \mathcal{N}(0,1) \perp\!\!\!\perp \epsilon_i}$$

where $Z_i \sim \mathcal{N}(0,1)$ independent of ϵ_i.
Comments:

1. even in LS case, e_i's do not have the right marginal distribution. However, only $\mathrm{var}\,(e_i)$ matters then... Hence, simple scaling works, though usual interpretation misleading/wrong

Bootstrapping from residuals

On the residuals: reminders

Call $e_i = Y_i - \widehat{\beta}_\rho^T X_i$, the i-th residual. In the asymptotic limit,

$$\boxed{e_i \overset{\mathcal{L}}{=} \text{prox}(c\rho)(\epsilon_i + r_\rho(\kappa)Z_i) \ , Z_i \sim \mathcal{N}(0,1) \perp\!\!\!\perp \epsilon_i}$$

where $Z_i \sim \mathcal{N}(0,1)$ independent of ϵ_i.
Comments:

1. even in LS case, e_i's do not have the right marginal distribution. However, only $\text{var}\,(e_i)$ matters then... Hence, simple scaling works, though usual interpretation misleading/wrong
2. For other loss functions, clear from prox system early in talk that performance depends on more than a few moments, hence problems

Bootstrapping from residuals

On the residuals: reminders

Call $e_i = Y_i - \widehat{\beta}_\rho^T X_i$, the i-th residual. In the asymptotic limit,

$$\boxed{e_i \overset{\mathcal{L}}{=} \text{prox}(c\rho)(\epsilon_i + r_\rho(\kappa)Z_i) \ , Z_i \sim \mathcal{N}(0,1) \perp\!\!\!\perp \epsilon_i}$$

where $Z_i \sim \mathcal{N}(0,1)$ independent of ϵ_i.
Comments:

1. even in LS case, e_i's do not have the right marginal distribution. However, only $\text{var}\,(e_i)$ matters then... Hence, simple scaling works, though usual interpretation misleading/wrong
2. For other loss functions, clear from prox system early in talk that performance depends on more than a few moments, hence problems
3. Bickel-Freedman, '83, for OLS - answered a slightly different question

Bootstrapping from the residuals
Further comments

1. Advocated for a long-time even in robust regression (e.g Shorack '81; Mammen '89 when $p^2/n \to 0$): clearly problematic here
2. Many methods suggested in low-dimension to improve second order accuracy: see e.g Koenker ('05), Parzen et al. ('94), De Angelis et al. ('93), McKean et al. ('93); outside of L_2, these methods did not improve our numerical results
3. So question: can we do better? Short answer: yes. Details skipped for lack of time.

Further bootstraps

Conclusion about bootstrapping residuals:

1. Need to be careful - in general will fail
2. Anti-conservative in general: CI do not cover the true value with the probability we want
3. Possible to fix the problems to a large extent

Further bootstraps

Conclusion about bootstrapping residuals:

1. Need to be careful - in general will fail
2. Anti-conservative in general: CI do not cover the true value with the probability we want
3. Possible to fix the problems to a large extent

Will now briefly discuss another type of bootstrap:
pairs-resampling
In standard books, this is the technique that is favored in general.

Another type of bootstrap
Resampling the pairs

Idea:

- For $b = 1, \ldots, B$, sample with replacement from $(X_i, Y_i)_{i=1}^n$.
- Get new dataset $(X_{i,b}^*, Y_{i,b}^*)_{i=1}^n$
- Fit model to this new dataset - using loss function ρ - to get $\{\widehat{\beta}_b^*\}_{b=1}^B$

Do inference using $\{\widehat{\beta}_b^* - \widehat{\beta}_\rho\}_{b=1}^B$

Pairs bootstrap
How does it fare?

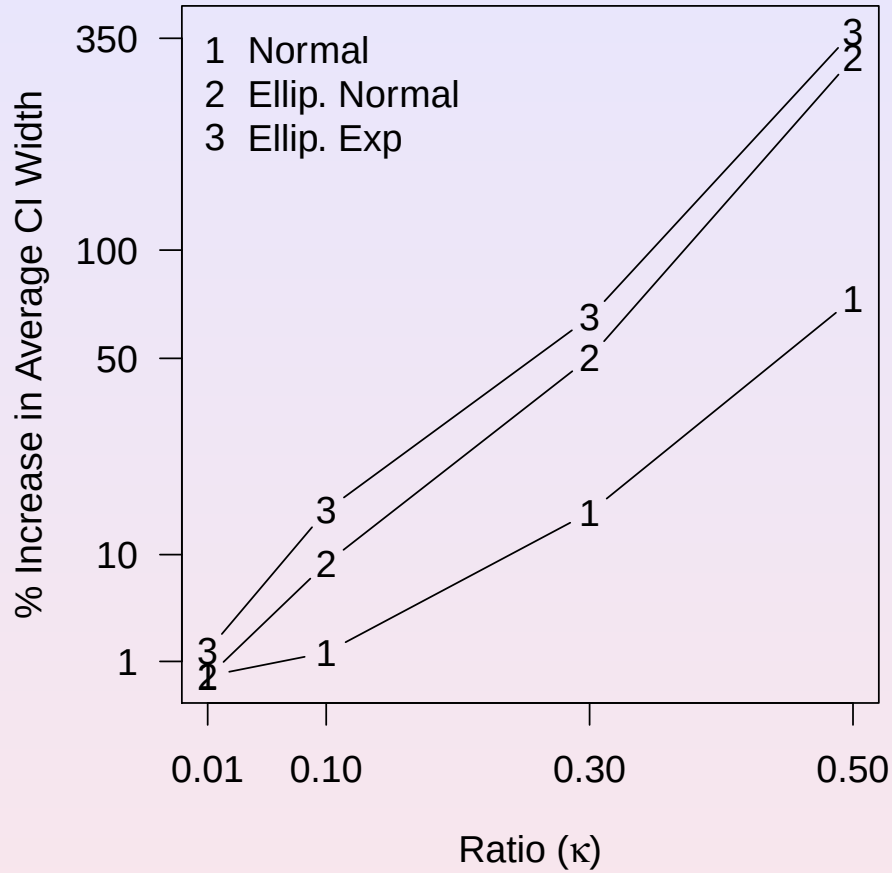
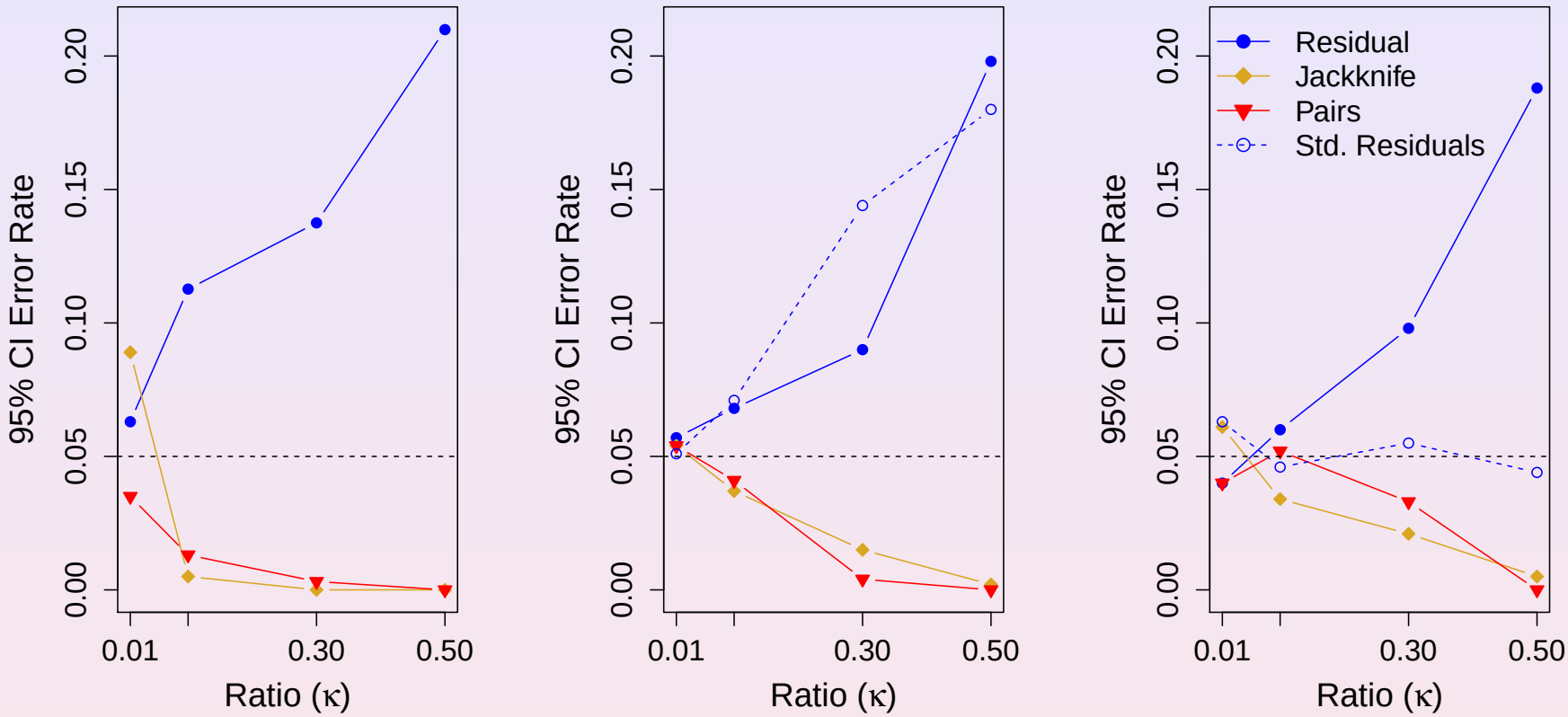


Figure: **Comparison of width of 95% confidence intervals of β_1 for L_2 loss:** on y-axis (log-scale): percent increase of average confidence interval width based on simulations ($n = 500$), vs theory in L_2; percent increase plotted against the ratio $\kappa = p/n$ (x-axis).

Pairs bootstrapping is conservative

$\epsilon_i \overset{iid}{\sim} \mathcal{N}(0,1)$, $X_i \sim \mathcal{N}(0, \mathrm{Id}_p)$; see triangle-curves



(a) L_1 loss: $\rho(x) = |x|$ (b) Huber loss (c) L_2 loss: $\rho(x) = x^2$

Figure: **Performance of 95% confidence intervals of β_1 : $n = 500$, 1,000 simulations** Pairs bootstrapping makes too few mistakes: conservative method.

Note that, if $w^*_{i,b}$ is number of times (X_i, Y_i) appears in b-th boot sample:

$$\widehat{\beta}^*_b = \text{argmin}_{\beta \in \mathbb{R}^p} \sum_{i=1}^{n} w^*_{i,b} \rho(Y_i - X_i^T \beta) \ .$$

Note that, if $w^*_{i,b}$ is number of times (X_i, Y_i) appears in b-th boot sample:

$$\widehat{\beta}^*_b = \text{argmin}_{\beta \in \mathbb{R}^p} \sum_{i=1}^{n} w^*_{i,b} \rho(Y_i - X_i^T \beta) \ .$$

Potential problems :

1. Number of distinct pairs $\{(X_i, Y_i)\}$ in bootstrapped sample is roughly $(1 - 1/e)n$. Problem if $p > (1 - 1/e)n$

Note that, if $w^*_{i,b}$ is number of times (X_i, Y_i) appears in b-th boot sample:

$$\widehat{\beta}^*_b = \text{argmin}_{\beta \in \mathbb{R}^p} \sum_{i=1}^{n} w^*_{i,b} \rho(Y_i - X_i^T \beta) \ .$$

Potential problems :

1. Number of distinct pairs $\{(X_i, Y_i)\}$ in bootstrapped sample is roughly $(1 - 1/e)n$. Problem if $p > (1 - 1/e)n$
2. Weighted robust regression in high-dimension has very different statistical properties than unweighted;

Note that, if $w^*_{i,b}$ is number of times (X_i, Y_i) appears in b-th boot sample:

$$\widehat{\beta}^*_b = \text{argmin}_{\beta \in \mathbb{R}^p} \sum_{i=1}^{n} w^*_{i,b} \rho(Y_i - X_i^T \beta) \ .$$

Potential problems :

1. Number of distinct pairs $\{(X_i, Y_i)\}$ in bootstrapped sample is roughly $(1 - 1/e)n$. Problem if $p > (1 - 1/e)n$
2. Weighted robust regression in high-dimension has very different statistical properties than unweighted;
3. "Reweighting changes the effective geometry of the dataset": so potentially problematic here

Theorem

Weights $(w_i)_{i=1}^n$ *i.i.d.,* $\mathbf{E}(w_i) = 1$*; enough moments and bounded away from 0.* $X_i \overset{iid}{\backsim} \mathcal{N}(0, \mathrm{Id}_p)$*;* v *: deterministic,* $\|v\|_2 = 1$*.*
$\widehat{\beta}$ *obtained by solving a least-squares problem - linear model holds;* $\mathrm{var}(\epsilon_i) = \sigma_\epsilon^2$*.* $\widehat{\beta}_w^*$ *bootstrap version of* $\widehat{\beta}$*, weights* $\{w_i\}_{i=1}^n$
If $\lim p/n = \kappa < 1$ *then asymptotically as* $n \to \infty$

$$p\mathbf{E}\left(\mathrm{var}\left(v^T\widehat{\beta}_w^*\right)\right) \to \sigma_\epsilon^2\left[\kappa\frac{1}{1-\kappa-\mathbf{E}\left(\frac{1}{(1+cw_i)^2}\right)} - \frac{1}{1-\kappa}\right],$$

c *: unique solution of*

$$\mathbf{E}\left(\frac{1}{1+cw_i}\right) = 1-\kappa\,.$$

Note that of course in setup above,

$$p\text{var}\left(v^T\widehat{\beta}\right) \to \sigma_\epsilon^2 \frac{\kappa}{1-\kappa}$$

1. Pairs-bootstrap does not get the right variance
2. Confidence intervals are too wide: method is **conservative** (covers the truth more often than it should)
3. Suggest weight corrections: current work on doing this adaptively

Jackknife and spectra of RMs

Many other resampling techniques exist...
Popular one: **jackknifing** (Quenouille, '50s) Jackknife fails in high-d but more predictably than bootstrap

Jackknife and spectra of RMs

Many other resampling techniques exist...
Popular one: **jackknifing** (Quenouille, '50s) Jackknife fails in high-d but more predictably than bootstrap

Eigenvalues of random matrices: also did work on bootstrapping eigenvalues of random matrices: results are pretty dismal unless the problem is statistically and probabilistically trivial (do you need bootstrap then?) Especially for extreme eigenvalues...

Conclusions

Random matrices provide unifying framework for understanding (at precision level useful to practitioners) large variety of problems in high-dimensional statistics

Conclusions

Random matrices provide unifying framework for understanding (at precision level useful to practitioners) large variety of problems in high-dimensional statistics

- Standard techniques/intuition do not perform well: CI have bad coverage, ML estimators inefficient, bootstrap fails for “new” reasons

Conclusions

Random matrices provide unifying framework for understanding (at precision level useful to practitioners) large variety of problems in high-dimensional statistics

- Standard techniques/intuition do not perform well: CI have bad coverage, ML estimators inefficient, bootstrap fails for "new" reasons
- Standard theoretical techniques based on perturbation analysis; implicitly same for bootstrap. Our problems and non-trivial statistical problems are not.

Conclusions

Random matrices provide unifying framework for understanding (at precision level useful to practitioners) large variety of problems in high-dimensional statistics

- Standard techniques/intuition do not perform well: CI have bad coverage, ML estimators inefficient, bootstrap fails for "new" reasons
- Standard theoretical techniques based on perturbation analysis; implicitly same for bootstrap. Our problems and non-trivial statistical problems are not.
- RM ideas open way to deal with some of those hard problems.

Conclusions

Random matrices provide unifying framework for understanding (at precision level useful to practitioners) large variety of problems in high-dimensional statistics

- Standard techniques/intuition do not perform well: CI have bad coverage, ML estimators inefficient, bootstrap fails for "new" reasons
- Standard theoretical techniques based on perturbation analysis; implicitly same for bootstrap. Our problems and non-trivial statistical problems are not.
- RM ideas open way to deal with some of those hard problems.
- Bootstrap main practical issue: we do not know in what direction bootstrap fails... Beyond our simple examples, what about truly complicated applied setups?

Conclusions

Random matrices provide unifying framework for understanding (at precision level useful to practitioners) large variety of problems in high-dimensional statistics

- Standard techniques/intuition do not perform well: CI have bad coverage, ML estimators inefficient, bootstrap fails for "new" reasons
- Standard theoretical techniques based on perturbation analysis; implicitly same for bootstrap. Our problems and non-trivial statistical problems are not.
- RM ideas open way to deal with some of those hard problems.
- Bootstrap main practical issue: we do not know in what direction bootstrap fails... Beyond our simple examples, what about truly complicated applied setups?
- Large n, p/RM theory seems to capture and explain some phenomena observed in practice