A few remarks on the bootstrap

For some moderately difficult statistical problems (a.k.a. in moderate and high dimensions)

N_2 El Karoui (joint with Elizabeth Purdom)

Department of Statistics + Criteo AI Lab
UC, Berkeley + Paris/Palo Alto

N. El Karoui's 3x25 birthday conference
May 2019

What is the bootstrap?

Bootstrap (Efron, '79): care about statistic $\widehat{\theta}_n$; would like to know its law. Can we do this from the data/sample we observe? Example: sample mean; suppose we have data $X_1, \ldots, X_n$, i.i.d, $X_i \in \mathbb{R}$. $\mathbf{E}(X_i) = \mu$, $\mathrm{var}(X_i) = \sigma^2$; interested in

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i \ .$$

What is the bootstrap?

Bootstrap (Efron, '79): care about statistic $\widehat{\theta}_n$; would like to know its law. Can we do this from the data/sample we observe?
Example: sample mean; suppose we have data $X_1, \ldots, X_n$, i.i.d, $X_i \in \mathbb{R}$. $\mathbf{E}(X_i) = \mu$, $\mathrm{var}(X_i) = \sigma^2$; interested in

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i \ .$$

Want to create a confidence interval for μ, the mean of X_i.

What is the bootstrap?

Bootstrap (Efron, '79): care about statistic $\widehat{\theta}_n$; would like to know its law. Can we do this from the data/sample we observe?
Example: sample mean; suppose we have data $X_1, \ldots, X_n$, i.i.d, $X_i \in \mathbb{R}$. $\mathbf{E}(X_i) = \mu$, $\mathrm{var}(X_i) = \sigma^2$; interested in

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \; .$$

Want to create a confidence interval for μ, the mean of X_i.

- Option 1: law of $\widehat{\theta}_n = \bar{X}_n$? Central limit theorem:

$$\sqrt{n}\frac{\bar{X}_n - \mu}{\sigma} \Longrightarrow \mathcal{N}(0, 1) \; .$$

  100 (1-α)%CI: $\bar{X}_n \pm \frac{\sigma}{\sqrt{n}} z_{1-\alpha/2}$; t-distribution variants

- Option 2: bootstrap

Idea: from the original sample, create lots of "new" datasets; this should mimick sampling mechanism which gave us $\bar{X}_n$ from population distribution

In more detail:

- For $b = 1, \ldots, B$, repeat:

Idea: from the original sample, create lots of “new” datasets; this should mimick sampling mechanism which gave us $\bar{X}_n$ from population distribution

In more detail:

- For $b = 1, \ldots, B$, repeat:
- Sample n times with replacement from $\{X_i\}_{i=1}^n$, to get dataset $D_b = \{X^*_{1,b}, \ldots, X^*_{n,b}\}$.

Idea: from the original sample, create lots of "new" datasets; this should mimick sampling mechanism which gave us $\bar{X}_n$ from population distribution

In more detail:

- For $b = 1, \ldots, B$, repeat:
- Sample n times with replacement from $\{X_i\}_{i=1}^n$, to get dataset $D_b = \{X^*_{1,b}, \ldots, X^*_{n,b}\}$.
- Compute $\bar{X}^*_{n,b} = \frac{1}{n} \sum_{i=1}^n X^*_{i,b}$

Idea: from the original sample, create lots of “new” datasets; this should mimick sampling mechanism which gave us $\bar{X}_n$ from population distribution

In more detail:

- For $b = 1, \ldots, B$, repeat:
- Sample n times with replacement from $\{X_i\}_{i=1}^n$, to get dataset $D_b = \{X^*_{1,b}, \ldots, X^*_{n,b}\}$.
- Compute $\bar{X}^*_{n,b} = \frac{1}{n}\sum_{i=1}^n X^*_{i,b}$

Idea: from the original sample, create lots of “new” datasets; this should mimick sampling mechanism which gave us $\bar{X}_n$ from population distribution

In more detail:

- For $b = 1, \ldots, B$, repeat:
- Sample n times with replacement from $\{X_i\}_{i=1}^n$, to get dataset $D_b = \{X^*_{1,b}, \ldots, X^*_{n,b}\}$.
- Compute $\bar{X}^*_{n,b} = \frac{1}{n}\sum_{i=1}^n X^*_{i,b}$

Now use $\{\bar{X}^*_{n,b}\}_{b=1}^B$ as approximation of distribution of $\bar{X}_n$

Idea: from the original sample, create lots of "new" datasets; this should mimick sampling mechanism which gave us $\bar{X}_n$ from population distribution

In more detail:

- For $b = 1, \ldots, B$, repeat:
- Sample n times with replacement from $\{X_i\}_{i=1}^n$, to get dataset $D_b = \{X^*_{1,b}, \ldots, X^*_{n,b}\}$.
- Compute $\bar{X}^*_{n,b} = \frac{1}{n} \sum_{i=1}^n X^*_{i,b}$

Now use $\{\bar{X}^*_{n,b}\}_{b=1}^B$ as approximation of distribution of $\bar{X}_n$

In particular, 95% CI could be, if $\bar{X}_{n,(k)}$ are increasingly ordered values of $\{\bar{X}_{n,b}\}_{b=1}^B$

$$\left(\bar{X}^*_{n,(2.5\%*B)}, \bar{X}^*_{n,(97.5\%*B)}\right) .$$

So called bootstrap percentile interval; simple computation shows asymptotically valid

Of course use it for much more complicated statistics

P: data generating distribution. Empirical distribution:

$$\hat{P}_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i}$$

Let θ be a functional of those distributions: e.g $\theta(P)$: median or trimmed mean of population
Often: use $\theta(\hat{P}_n)$ to get confidence interval/statement about $\theta(P)$.
Question e.g.:

(Asymptotic) law of $\theta(\hat{P}_n) - \theta(P)$?

P: data generating distribution. Empirical distribution:

$$\hat{P}_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i}$$

Let θ be a functional of those distributions: e.g $\theta(P)$: median or trimmed mean of population
Often: use $\theta(\hat{P}_n)$ to get confidence interval/statement about $\theta(P)$.
Question e.g.:

$$\text{(Asymptotic) law of } \theta(\hat{P}_n) - \theta(P)?$$

bootstrap: if $\hat{P}_n^*$ is bootstrapped version of $\hat{P}_n$,

$$\text{Bootstrap law of } [\theta(\hat{P}_n^*) - \theta(\hat{P}_n)] \text{``} \simeq \text{''} \text{ Law of } [\theta(\hat{P}_n) - \theta(P)]?$$

Left-hand side: we can “resample” the data to get this
Righ-hand side: ideally, we would like to know it, but not accessible

Suppose we are interested in random variable

$$\widehat{\theta}(\hat{P}_n, P) \text{ and its law } \mathcal{L}_n(\widehat{\theta}(\hat{P}_n, P))$$

E.g $\widehat{\theta}(\hat{P}_n, P) = \sqrt{n}(\mu(\hat{P}_n) - \mu(P))$
Suppose

$$\mathcal{L}_n(\widehat{\theta}(\hat{P}_n, P)) \Longrightarrow \mathcal{L}$$

Call $\mathcal{L}_{n,boot}(\hat{P}_n)$ the conditional law of $\widehat{\theta}(\hat{P}_n^*, \hat{P}_n)|\hat{P}_n$
Then bootstrap works if, e.g,

$$\lim_{n\to\infty} d(\mathcal{L}_{n,boot}(\hat{P}_n), \mathcal{L}) \to 0 \text{ , } a.s \;\; X_1, \ldots, X_n, \ldots$$

where d: distance between probability measures; alternative: convergence in probability

Suppose we are interested in random variable

$$\widehat{\theta}(\hat{P}_n, P) \text{ and its law } \mathcal{L}_n(\widehat{\theta}(\hat{P}_n, P))$$

E.g $\widehat{\theta}(\hat{P}_n, P) = \sqrt{n}(\mu(\hat{P}_n) - \mu(P))$
Suppose

$$\mathcal{L}_n(\widehat{\theta}(\hat{P}_n, P)) \Longrightarrow \mathcal{L}$$

Call $\mathcal{L}_{n,boot}(\hat{P}_n)$ the conditional law of $\widehat{\theta}(\hat{P}_n^*, \hat{P}_n)|\hat{P}_n$
Then bootstrap works if, e.g,

$$\lim_{n\to\infty} d(\mathcal{L}_{n,boot}(\hat{P}_n), \mathcal{L}) \to 0 \text{ , } a.s \;\; X_1, \ldots, X_n, \ldots$$

where d: distance between probability measures; alternative: convergence in probability
Example: X_i i.i.d mean μ, $\operatorname{cov}(X_i) = \Sigma$, then conditionally on $X_1, \ldots, X_n$

$$\sqrt{n}(\bar{X}_n^* - \bar{X}_n) \Longrightarrow \mathcal{N}(0, \Sigma)$$

for almost every sequence $X_1, \ldots, X_n, \ldots$

Bootstrap: brilliant idea, **huge** impact for applied, methodological and theoretical statistics; probably one of the most widely used tool in applied statistics

Everything seems possible; no need for asymptotics. Now, beside stat practice, very useful in teaching data science and inferential ideas.

Bootstrap: brilliant idea, **huge** impact for applied, methodological and theoretical statistics; probably one of the most widely used tool in applied statistics

Everything seems possible; no need for asymptotics. Now, beside stat practice, very useful in teaching data science and inferential ideas.

Theory started almost immediately: Bickel-Freedman (AoS, '81) first fairly general paper.

Lots of activity both practical and theoretical for 30+ years

Bootstrap: brilliant idea, **huge** impact for applied, methodological and theoretical statistics; probably one of the most widely used tool in applied statistics

Everything seems possible; no need for asymptotics. Now, beside stat practice, very useful in teaching data science and inferential ideas.

Theory started almost immediately: Bickel-Freedman (AoS, '81) first fairly general paper.

Lots of activity both practical and theoretical for 30+ years

Standard books: Davison-Hinkley (applied/theory), Hall (mostly theory), Politis-Romano-Wolf (subsampling)

And lots of variants of bootstrap (e.g m-out-of-n bootstrap (Bickel et al.), various other subsampling methods...)

Other old techniques discussed later

- Bootstrap: brilliant idea, **huge** impact for applied, methodological and theoretical statistics; probably one of the most widely used tool in applied statistics
- Everything seems possible; no need for asymptotics. Now, beside stat practice, very useful in teaching data science and inferential ideas.
- Theory started almost immediately: Bickel-Freedman (AoS, '81) first fairly general paper.
- Lots of activity both practical and theoretical for 30+ years
- Standard books: Davison-Hinkley (applied/theory), Hall (mostly theory), Politis-Romano-Wolf (subsampling)
- And lots of variants of bootstrap (e.g m-out-of-n bootstrap (Bickel et al.), various other subsampling methods...)
- Other old techniques discussed later
- One big question: when does it work?

Example where it does not work: $X_i \overset{iid}{\backsim} Unif[0, a]$, distribution of the $(a - \max X_i)$

Example where it does not work: $X_i \overset{iid}{\backsim} Unif[0, a]$, distribution of the $(a - \max X_i)$

Essentially, need the function θ to be "smooth" enough. Formal results on next slide. Informally: von Mises calculus: θ differentiable implies: if $\theta'(\cdot; P)$ is linear

$$\theta(\hat{P}_n) - \theta(P) \simeq \frac{1}{\sqrt{n}}\theta'(G_n; P) ,$$

where $G_n = \sqrt{n}(\hat{P}_n - P)$ (Donsker thm: limit of G_n is (P-)Brownian bridge)

Bootstrap

When does it work? 1-dimensional case

Example where it does not work: $X_i \overset{iid}{\backsim} Unif[0, a]$, distribution of the $(a - \max X_i)$

Essentially, need the function θ to be "smooth" enough. Formal results on next slide. Informally: von Mises calculus:

θ differentiable implies: if $\theta'(\cdot; P)$ is linear

$$\theta(\hat{P}_n) - \theta(P) \simeq \frac{1}{\sqrt{n}}\theta'(G_n; P) \ ,$$

where $G_n = \sqrt{n}(\hat{P}_n - P)$ (Donsker thm: limit of G_n is (P-)Brownian bridge)

Bootstrap: expand $\theta(\hat{P}_n^*)$ around $\theta(P)$ + linearity to get:

$$\theta(\hat{P}_n^*) - \theta(\hat{P}_n) \simeq \frac{1}{\sqrt{n}}\theta'(G_n^*; P) \ ,$$

$G_n^* = \sqrt{n}(\hat{P}_n^* - \hat{P}_n)$; G_n^* also has P-Brownian bridge as limit

Look at θ as mapping from $(D[-\infty, \infty], \|\cdot\|_\infty) \mapsto \mathbb{R}$, where D càdlàg/rcll functions. If θ Hadamard differentiable, i.e

$$\left|\frac{\theta(F + th_t) - \theta(F)}{t} - \theta'(h; F)\right| \to 0,$$

$$\text{as } t \to 0^+, \forall h_t : \sup_{x \in \mathbb{R}} |h_t(x) - h(x)| \to 0 .$$

$\theta'(\cdot; F)$: continuous linear map, $(D, \|\cdot\|_\infty) \mapsto \mathbb{R}$.
Then bootstrap works.
Then not much need to understand fluctuation properties of $\theta(\hat{P}_n)$: resampling does it for us.
Often summarized as : “bootstrap works for smooth statistics”

Plan for rest of talk

Work in the high-dimensional case: data vectors $\{X_i\}_{i=1}^n \in \mathbb{R}^p$, $p/n \to \kappa \in (0,1)$

Arguments above (proximity of empirical and population distribution) fail; but what about bootstrap?

1. Bootstrapping (robust) regression: review
2. Bootstrapping regression in high-dimension: results
3. RM issues in bootstrap

Why p/n not close to 0?

Plan for rest of talk

Work in the high-dimensional case: data vectors $\{X_i\}_{i=1}^n \in \mathbb{R}^p$, $p/n \to \kappa \in (0,1)$
Arguments above (proximity of empirical and population distribution) fail; but what about bootstrap?

1. Bootstrapping (robust) regression: review
2. Bootstrapping regression in high-dimension: results
3. RM issues in bootstrap

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

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

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

For ρ loss function, consider

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

Simplest question: can get CI for $\beta_0(1)$ based on $\widehat{\beta}_\rho(1)$?

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

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

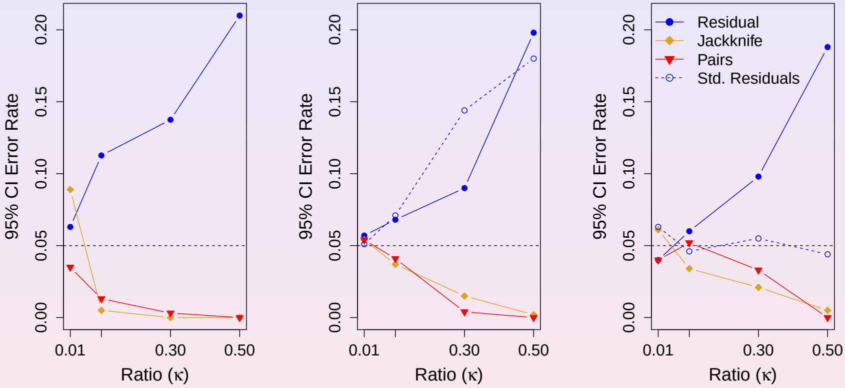
What's random? ϵ_i in this context; they are i.i.d.
X_i assumed "fixed" in this example.
So **bootstrap from the residuals:**

1. estimate β_0 by $\widehat{\beta}_\rho$
2. estimate ϵ_i by e_i's; typically $e_i = Y_i - X_i^T \widehat{\beta}$
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 $Y^*_{i,b} = X_i^T \widehat{\beta} + e^*_{i,b}$
    3. Fit this new dataset to get $\widehat{\beta}^*_b$

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

Bootstrapping from the residuals

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



(a) L_1 loss (b) Huber loss (c) L_2 loss

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

Note: Bickel and Freedman ('83) studied high-dimensional residual bootstrap for least-squares; showed that residuals did not have the right distribution. Mammen ('89) for robust regression when $p^2/n \to 0$

Note: Bickel and Freedman ('83) studied high-dimensional residual bootstrap for least-squares; showed that residuals did not have the right distribution. Mammen ('89) for robust regression when $p^2/n \to 0$
Of course, if $e = \{e_i\}_{i=1}^n$ are residuals,

$$e = (\mathrm{Id} - X(X^TX)^{-1}X^T)\epsilon \triangleq (\mathrm{Id} - H)\epsilon \; .$$

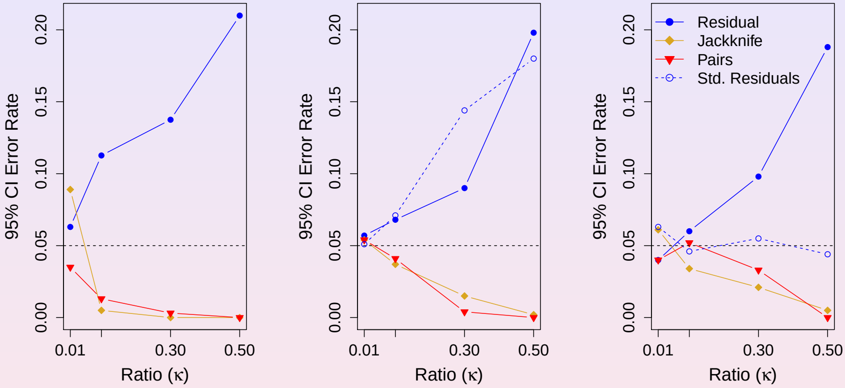
So suggestion for resampling (see e.g Davison-Hinkley '97, many others): use

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

In low-dimension, this correction is minimal; in high-d, Gaussian case, $H_{i,i} \simeq 1 - \frac{p}{n}$: non-negligible correction

Bootstrapping from the standardized residuals

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



(a) L_1 loss (b) Huber loss (c) L_2 loss

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.

Recall M-estimation problem above. Suppose $p/n \to \kappa \in (0, 1)$. For simplicity of statement, X_i i.i.d with mean-0 i.i.d entries with many moments.

Recall M-estimation problem above. Suppose $p/n \to \kappa \in (0,1)$. For simplicity of statement, X_i i.i.d with mean-0 i.i.d entries with many moments.

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,

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

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

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

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

1. if $\rho(x) = x^2/2$, $\operatorname{prox}(c\rho)[x] = \frac{x}{1+c}$; hence, here $\frac{1}{1+c} = 1 - \kappa$
2. if $\rho(x) = |x|$, $\operatorname{prox}(c\rho)[x] = sgn(x)(|x| - c)_+$

Comments:

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

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

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

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

1. if $\rho(x) = x^2/2$, $\mathrm{prox}(c\rho)[x] = \frac{x}{1+c}$; hence, here $\frac{1}{1+c} = 1 - \kappa$
2. if $\rho(x) = |x|$, $\mathrm{prox}(c\rho)[x] = sgn(x)(|x| - c)_+$

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
2. For other loss functions, clear that performance depends on more than a few moments, hence problems

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

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

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

1. if $\rho(x) = x^2/2$, $\mathrm{prox}(c\rho)[x] = \frac{x}{1+c}$; hence, here $\frac{1}{1+c} = 1 - \kappa$
2. if $\rho(x) = |x|$, $\mathrm{prox}(c\rho)[x] = sgn(x)(|x| - c)_+$

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
2. For other loss functions, clear that performance depends on more than a few moments, hence problems
3. Bickel-Freedman, '83, for OLS - answered a slightly different question

1. Advocated for a long-time even in robust regression (e.g Shorack '81): 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?

Bootstrapping from residuals
A couple ideas

Recall that in robust regression, asymptotically, in setting considered here:

$$Y_i - X_i^T \widehat{\beta} = \mathbf{e_i} \stackrel{\mathcal{L}}{=} \mathrm{prox}(c\rho)(\epsilon_i + r_\rho(\kappa) Z_i) \ , Z_i \sim \mathcal{N}(0,1) \perp\!\!\!\perp \epsilon_i$$

$\mathrm{prox}(c\rho)$ problematic: so instead, use as basis of work

$$\tilde{e}_{i,(i)} = Y_i - X_i^T \widehat{\beta}_{(i)} = \epsilon_i + X_i^T(\beta_0 - \widehat{\beta}_{(i)}) \ , \text{ because}$$
$$e_i = \mathrm{prox}(c\rho)(\tilde{e}_{i,(i)}) \ .$$

where $\widehat{\beta}_{(i)}$ is leave-i-th-observation out estimate. Remarks:

- Stochastic structure of $\tilde{e}_{i,(i)}$ comparatively simpler than that of e_i

Bootstrapping from residuals

A couple ideas

Recall that in robust regression, asymptotically, in setting considered here:

$$Y_i - X_i^T \widehat{\beta} = \mathbf{e_i} \stackrel{\mathcal{L}}{=} \mathrm{prox}(c\rho)(\epsilon_i + r_\rho(\kappa) Z_i) \;, Z_i \sim \mathcal{N}(0,1) \perp\!\!\!\perp \epsilon_i$$

$\mathrm{prox}(c\rho)$ problematic: so instead, use as basis of work

$$\tilde{e}_{i,(i)} = Y_i - X_i^T \widehat{\beta}_{(i)} = \epsilon_i + X_i^T(\beta_0 - \widehat{\beta}_{(i)}) \;, \text{ because}$$
$$e_i = \mathrm{prox}(c\rho)(\tilde{e}_{i,(i)}) \;.$$

where $\widehat{\beta}_{(i)}$ is leave-i-th-observation out estimate. Remarks:

- Stochastic structure of $\tilde{e}_{i,(i)}$ comparatively simpler than that of e_i
- Problem 1 with $\tilde{e}_{i,(i)}$: excess variance compared to ϵ_i

Recall that in robust regression, asymptotically, in setting considered here:

$$Y_i - X_i^T \widehat{\beta} = \mathbf{e_i} \stackrel{\mathcal{L}}{=} \mathrm{prox}(c\rho)(\epsilon_i + r_\rho(\kappa) Z_i) \ , Z_i \sim \mathcal{N}(0,1) \perp\!\!\!\perp \epsilon_i$$

$\mathrm{prox}(c\rho)$ problematic: so instead, use as basis of work

$$\tilde{e}_{i,(i)} = Y_i - X_i^T \widehat{\beta}_{(i)} = \epsilon_i + X_i^T (\beta_0 - \widehat{\beta}_{(i)}) \ , \text{ because}$$
$$e_i = \mathrm{prox}(c\rho)(\tilde{e}_{i,(i)}) \ .$$

where $\widehat{\beta}_{(i)}$ is leave-i-th-observation out estimate. Remarks:

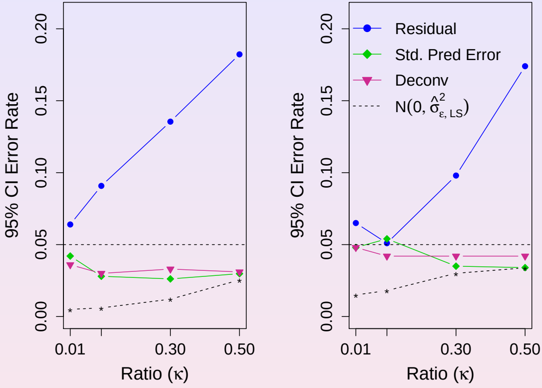
- Stochastic structure of $\tilde{e}_{i,(i)}$ comparatively simpler than that of e_i
- Problem 1 with $\tilde{e}_{i,(i)}$: excess variance compared to ϵ_i
- Problem 2 with $\tilde{e}_{i,(i)}$: extra "Gaussian" component

Idea: resample from $\tilde{e}_{i,(i)}$ but properly scale them. Need at least right variance...
How to do so?

1. Estimate $\sigma^2(\epsilon)$ using least squares: easy to get consistent estimator in high-dimension for that
2. Easy to get estimate of $\|(\beta_0 - \widehat{\beta}_{(i)})\|$ then.
3. Normalize $e_{i,(i)}$ to $\tilde{e}_{i,(i)}$ so variance of the latter is $\widehat{\sigma}(\epsilon)$.
4. Use $\tilde{e}_{i,(i)}$ in bootstrap resampling

Bootstrapping the residuals

Approach 1: scaling predicted errors; $\epsilon_i \overset{iid}{\backsim}$ double exponential



(a) L_1 loss (b) Huber loss

Figure: **Bootstrap based on predicted errors:** We plotted the error rate of 95% confidence intervals for alternative bootstrap methods: bootstrapping from standardized predicted errors (blue) and from deconvolution of predicted error (magenta).

Further bootstraps

Conclusion about bootstrapping residuals:

1. Need to be careful - in general not accurate/can fail
2. Anti-conservative in general: CI do not cover the true value with the probability we want
3. Appears possible to fix to a certain/large extent the problems

Will now discuss another type of bootstrap: **pairs-resampling**

Will now discuss another type of bootstrap: **pairs-resampling**
In standard books, this is the technique that is favored in general.
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 to get $\{\widehat{\beta}^*_b\}_{b=1}^B$

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

Pairs bootstrap
More details

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

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

Pairs bootstrap
More details

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 $p > (1-1/e)n$

Pairs bootstrap
More details

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. Understood in NEK et al. '11 that weighted robust regression has very different statistical properties than unweighted; measure concentration

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. Understood in NEK et al. '11 that weighted robust regression has very different statistical properties than unweighted; measure concentration
3. RM point of view: least squares: $X_i \to \sqrt{w^*_{i,b}} X_i$: move from "Gaussian to elliptical".

Pairs bootstrap
More details

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. Understood in NEK et al. '11 that weighted robust regression has very different statistical properties than unweighted; measure concentration
3. RM point of view: least squares: $X_i \to \sqrt{w^*_{i,b}} X_i$: move from "Gaussian to elliptical".
4. "Reweighting changes the effective geometry of the dataset": so potentially problematic here

Pairs bootstrap
More details

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. Understood in NEK et al. '11 that weighted robust regression has very different statistical properties than unweighted; measure concentration
3. RM point of view: least squares: $X_i \to \sqrt{w_{i,b}^*} X_i$: move from "Gaussian to elliptical".
4. "Reweighting changes the effective geometry of the dataset": so potentially problematic here
5. Note however that reweighting also affects ϵ_i's

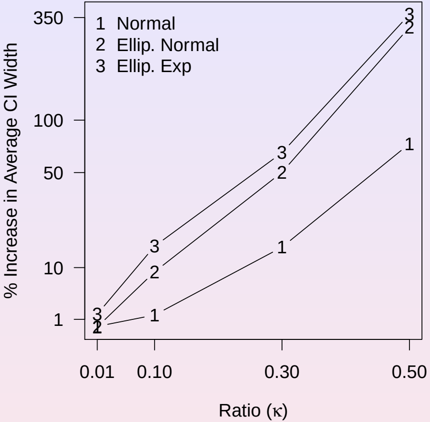


Figure: **Comparison of width of 95% confidence intervals of β_1 for L_2 loss:** y-axis is the percent increase of the average confidence interval width based on simulation ($n = 500$), as compared to the average for the standard confidence interval based on normal theory in L_2; the percent increase is plotted against the ratio $\kappa = p/n$ (x-axis)

Theorem

Weights $(w_i)_{i=1}^n$ *be 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 unit vector.*

Suppose $\widehat{\beta}$ *is obtained by solving a least-squares problem - linear model holds;* $\mathrm{var}(\epsilon_i) = \sigma_\epsilon^2$

If $\lim p/n = \kappa < 1$ *then asymptotically as* $n \to \infty$

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

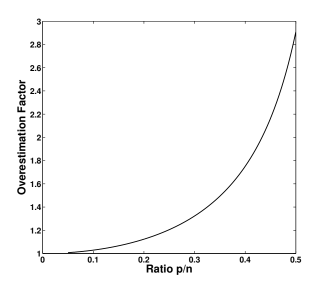
c : *unique solution of*

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

Note that of course in setup above,

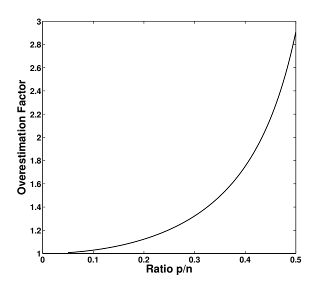
$$p \operatorname{var}\left(v^T \widehat{\beta}\right) \to \sigma_\epsilon^2 \frac{\kappa}{1-\kappa}$$

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. Ratio $\mathbf{E}\left(\operatorname{var}\left(v^T \widehat{\beta}_w^*\right)\right) / \operatorname{var}\left(v^T \widehat{\beta}\right)$ does not depend on $\operatorname{cov}(X_i) = \Sigma$ - results true for any Σ
4. Suggest weight corrections (not discussed because of time constraints)

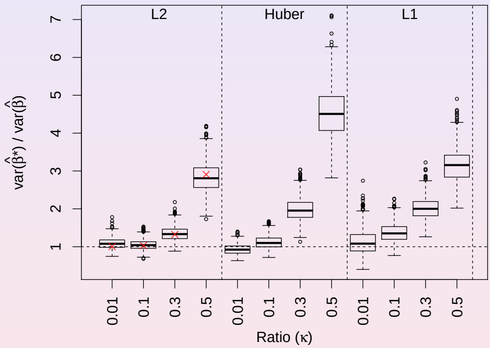


(a) L_2 (Theoretical)

Figure: **Factor by which standard pairs bootstrap over-estimates the variance:** Gaussian design, Gaussian errors



(a) L_2 (Theoretical)



(b) All (Simulated)

Figure: **Factor by which standard pairs bootstrap over-estimates the variance:** Gaussian design, Gaussian errors

Beyond regression problems

Are these issues limited to the simple setting of regression?

Another type of statistics: eigenvalues of covariance matrices

Sample covariance matrices and their eigenvalues

Recall if data is X_i,

$$\widehat{\Sigma} = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})(X_i - \bar{X})^T \; .$$

Bootstrap quite widely used to assess fluctuation behavior of eigenvalues of sample covariance matrices. See Beran and Srivastava ('85), Eaton and Tyler ('91)

Sample covariance matrices and their eigenvalues

Recall if data is X_i,

$$\widehat{\Sigma} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})(X_i - \bar{X})^T .$$

Bootstrap quite widely used to assess fluctuation behavior of eigenvalues of sample covariance matrices. See Beran and Srivastava ('85), Eaton and Tyler ('91)

In context of p fixed and $n \to \infty$, showed that when Σ has eigenvalues of multiplicity 1, bootstrap works.

Sample covariance matrices and their eigenvalues

Recall if data is X_i,

$$\widehat{\Sigma} = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})(X_i - \bar{X})^T \; .$$

Bootstrap quite widely used to assess fluctuation behavior of eigenvalues of sample covariance matrices. See Beran and Srivastava ('85), Eaton and Tyler ('91)

In context of p fixed and $n \to \infty$, showed that when Σ has eigenvalues of multiplicity 1, bootstrap works. Fails when eigenvalues have multiplicity higher than 1.

Sample covariance matrices and their eigenvalues

Recall if data is X_i,

$$\widehat{\Sigma} = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})(X_i - \bar{X})^T .$$

Bootstrap quite widely used to assess fluctuation behavior of eigenvalues of sample covariance matrices. See Beran and Srivastava ('85), Eaton and Tyler ('91)
In context of p fixed and $n \to \infty$, showed that when Σ has eigenvalues of multiplicity 1, bootstrap works. Fails when eigenvalues have multiplicity higher than 1. Can use subsampling to fix the problem.

Sample covariance matrices and their eigenvalues

Recall if data is X_i,

$$\widehat{\Sigma} = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})(X_i - \bar{X})^T \, .$$

Bootstrap quite widely used to assess fluctuation behavior of eigenvalues of sample covariance matrices. See Beran and Srivastava ('85), Eaton and Tyler ('91)

In context of p fixed and $n \to \infty$, showed that when Σ has eigenvalues of multiplicity 1, bootstrap works. Fails when eigenvalues have multiplicity higher than 1. Can use subsampling to fix the problem.

Bootstrapping eigenvalues currently used in a number of fields (see e.g several papers in British Journal of Psychology '07)

Now question: is that true if $p/n \to c \neq 0$?

Classic results

Recall

Theorem (Johnstone ('01))

If X_i are i.i.d $\mathcal{N}(0, \mathrm{Id}_p)$, then as $p/n \to \gamma \in (0, \infty)$

$$n^{2/3} \frac{\lambda_{max}(\widehat{\Sigma}) - (1 + \sqrt{p/n})^2}{\sigma_{n,p}} \Rightarrow TW_1 \ .$$

Further results: phase transition at $\lambda_1(\Sigma) = 1 + \sqrt{p/n}$ (BBP, '04); general Σ case (N_2EK, '05; Lee and Schnelli '13). Much work since then.

Classic results

Recall

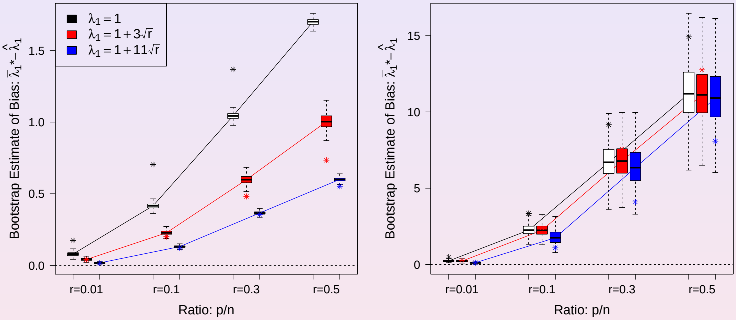
Theorem (Johnstone ('01))

If X_i are i.i.d $\mathcal{N}(0, \mathrm{Id}_p)$, then as $p/n \to \gamma \in (0, \infty)$

$$n^{2/3} \frac{\lambda_{max}(\widehat{\Sigma}) - (1 + \sqrt{p/n})^2}{\sigma_{n,p}} \Rightarrow TW_1 \ .$$

Further results: phase transition at $\lambda_1(\Sigma) = 1 + \sqrt{p/n}$ (BBP, '04); general Σ case (N_2EK, '05; Lee and Schnelli '13). Much work since then.
Also classic work (Marcenko-Pastur ('67), Wachter ('78)) about empirical spectral distribution of eigenvalues

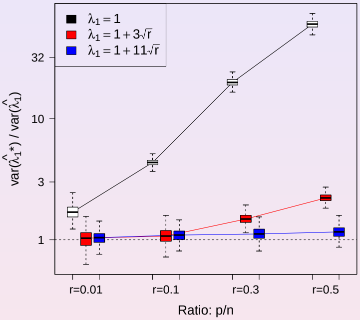
Eigenvalues
Numerics: bias



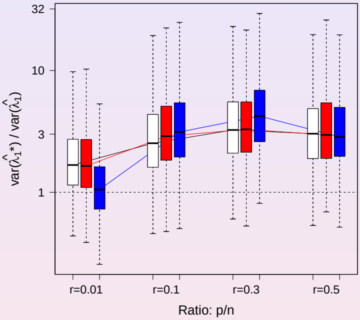
(a) $Z \sim$ Normal (b) $Z \sim$ Ellip. Exp

Figure: **Bias of Largest Bootstrap Eigenvalue, n=1,000:** Plotted are boxplots of the difference of the average bootstrap value of λ_1 over 999 bootstrap samples, minus the estimate $\hat{\lambda}_1$ over 1000 simulations; $\bar{\lambda}_1^* - \hat{\lambda}_1$ is also the standard bootstrap estimate of bias.

Eigenvalues
Numerics: variance



(a) $Z \sim$ Normal

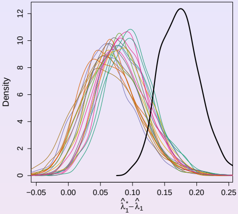


(b) $Z \sim$ Ellip. Exp

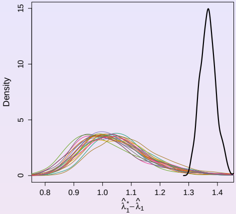
Figure: **Ratio of Bootstrap Estimate of Variance to True Variance for Largest Eigenvalue, n=1,000:** Plotted are boxplots of the bootstrap estimate of variance ($B = 999$) as a ratio of the true variance of $\hat{\lambda}_1$; boxplots represent the bootstrap estimate of variance

Eigenvalues

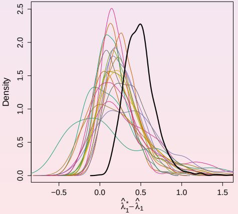
Numerics: distribution in null case



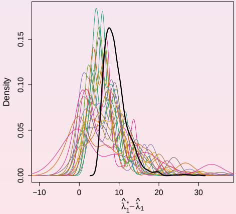
(a) $Z \sim$ Normal, r=0.01



(b) $Z \sim$ Normal, r=0.3



(c) $Z \sim$ Ellip. Exp, r=0.01



(d) $Z \sim$ Ellip. Exp, r=0.3

Eigenvalues

Theory: skipped in interest of time

- Simple theory for well separated eigenvalues
- Possible to do theory of spectral distribution of eigenvalues: Results are negative: bootstrapped Stieltjes transform concentrates but around the “wrong” Stieltjes transform.
- Can be used (with a few more refined tools) to understand bootstrap bias

Conclusions

Bootstrap :

- Standard techniques/intuition do not perform well (jackknife)

Conclusions

Bootstrap :

- Standard techniques/intuition do not perform well (jackknife)
- Caveat: resampling from scaled residuals work for least-squares (but do we need it in our context?)

Conclusions

Bootstrap :

- Standard techniques/intuition do not perform well (jackknife)
- Caveat: resampling from scaled residuals work for least-squares (but do we need it in our context?)
- Hinted at possible fixes for in robust-regression setups

Conclusions

Bootstrap :

- Standard techniques/intuition do not perform well (jackknife)
- Caveat: resampling from scaled residuals work for least-squares (but do we need it in our context?)
- Hinted at possible fixes for in robust-regression setups
- Main Problem: we do not know in what direction standard bootstrap has issues... Beyond our simple examples, what about truly complicated applied setups?

Conclusions

Bootstrap :

- Standard techniques/intuition do not perform well (jackknife)
- Caveat: resampling from scaled residuals work for least-squares (but do we need it in our context?)
- Hinted at possible fixes for in robust-regression setups
- Main Problem: we do not know in what direction standard bootstrap has issues... Beyond our simple examples, what about truly complicated applied setups?
- Slightly more complicated problem of eigenvalues results in severe problems... unless the problem is effectively low-d and trivial

Conclusions

Bootstrap :

- Standard techniques/intuition do not perform well (jackknife)
- Caveat: resampling from scaled residuals work for least-squares (but do we need it in our context?)
- Hinted at possible fixes for in robust-regression setups
- Main Problem: we do not know in what direction standard bootstrap has issues... Beyond our simple examples, what about truly complicated applied setups?
- Slightly more complicated problem of eigenvalues results in severe problems... unless the problem is effectively low-d and trivial
- Seems bootstrap genuinely perturbation-analytic method

Conclusions

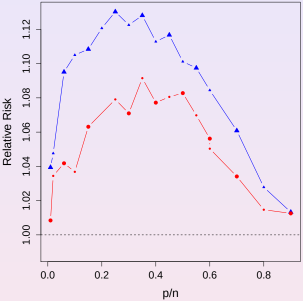
Bootstrap :

- Standard techniques/intuition do not perform well (jackknife)
- Caveat: resampling from scaled residuals work for least-squares (but do we need it in our context?)
- Hinted at possible fixes for in robust-regression setups
- Main Problem: we do not know in what direction standard bootstrap has issues... Beyond our simple examples, what about truly complicated applied setups?
- Slightly more complicated problem of eigenvalues results in severe problems... unless the problem is effectively low-d and trivial
- Seems bootstrap genuinely perturbation-analytic method
- Large n, p theory seems to capture some phenomena observed in practice - may lead to a practically informative theory.

Bon anniversaire!

Bon anniversaire!

Bon anniversaire!



(a)

Figure: **Solid line: Relative Risk of $\widehat{\beta}$ for scaled predicted errors vs original errors - population version**

Robust regression estimator

Impact of error distribution

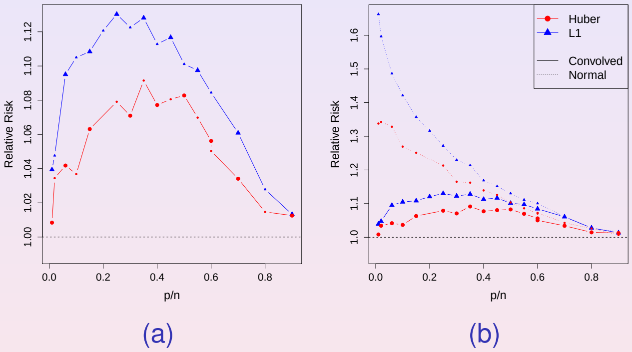


Figure: **Solid line: Relative Risk of $\widehat{\beta}$ for scaled predicted errors vs original errors - population version** Dotted line: using $\eta_i \overset{iid}{\backsim} \mathcal{N}(0, \sigma_\epsilon^2)$