A Resized Bootstrap Method for Inference of a High-Dimensional Logistic Regression

Qian Zhao

University of Massachusetts, Amherst

EcoSta 2024

Table of Contents

High-dimensional logistic regression

- We model Y by a logistic model:

$$\mathrm{P}(Y = 1 \,|\, X) = \frac{1}{1 + e^{-\beta_0 - \sum_{j=1}^{p} \beta_j X_j}},$$

- Given n i.i.d. pairs (X_i, Y_i), the MLE $\hat{\beta}$ exists.
- Large number of covariates: $p/n \approx \kappa > 0$, e.g., $p = 1,000$ and $n = 5,000$.

High-dimensional logistic regression

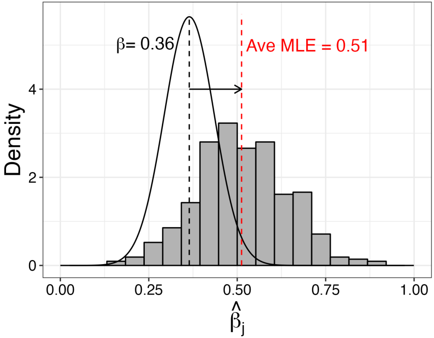
- We model Y by a logistic model:

$$\mathrm{P}(Y = 1 \mid X) = \frac{1}{1 + e^{-\beta_0 - \sum_{j=1}^{p} \beta_j X_j}},$$

- Given n i.i.d. pairs (X_i, Y_i), the MLE $\hat{\beta}$ exists.
- Large number of covariates: $p/n \approx \kappa > 0$, e.g., $p = 1,000$ and $n = 5,000$.

Question: How to construct a 95% confidence interval for β_j?

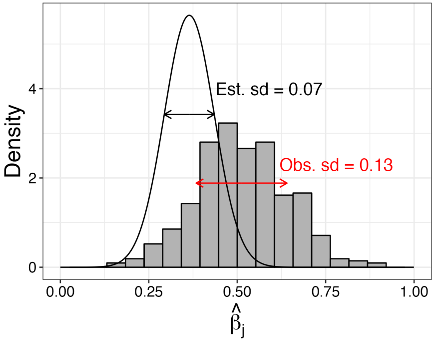
MLE is inflated when p is large



- Classical theory: $\hat{\beta}_j - \beta_j \sim \mathcal{N}(0, \mathcal{I}_{jj}^{-1})$.
- MLE is inflated: $\alpha = \hat{\beta}_j / \beta_j \approx 1.4$.

Simulation setting: Sample $X \sim \mathcal{N}(0, \Sigma)$, Σ is a circular matrix. Simulate Y from a logistic model (100 non-null $\beta_j \sim \mathcal{N}(0, 0.2)$, $\beta_0 = 0$). $n = 2,000$ and $p = 400$.

Classical theory underestimates SD of the MLE



- Classical theory: $\hat{\beta}_j - \beta_j \approx \mathcal{N}(0, \mathcal{I}_{jj}^{-1})$.
- Inverse Fisher information under-estimates the SD.

Simulation setting: Sample $X \sim \mathcal{N}(0, \Sigma)$, Σ is a circular matrix. Simulate Y from a logistic model (100 non-null $\beta_j \sim \mathcal{N}(0, 0.2)$, $\beta_0 = 0$). $n = 2,000$ and $p = 400$.

Distribution of the high-dimensional logistic MLE (**Zhao**, Sur & Candès, 2022)

- n i.i.d. obs (X_i, Y_i),
- $X_i \in \mathbb{R}^p$ are multivariate Gaussian $X_i \sim \mathcal{N}(0, \Sigma)$.
- $Y_i \in \{0, 1\}$, $Y_i \,|\, X_i$ is from a logistic model with parameters β.

As $p, n \to \infty$ at a constant ratio $p/n \to \kappa$, if $\sqrt{n}\tau_j\beta_j = O(1)$, then the MLE $\hat{\beta}_j$ satisfy

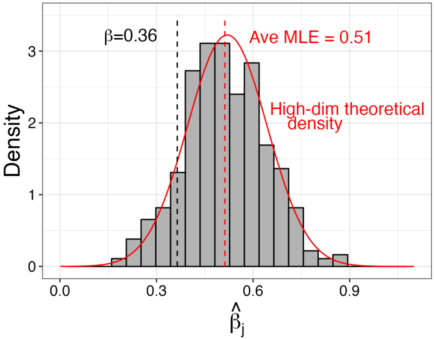
$$\frac{\sqrt{n}(\hat{\beta}_j - \alpha_\star \beta_j)}{\sigma_\star/\tau_j} \xrightarrow{d} \mathcal{N}(0, 1),$$

where $\tau_j^2 = \mathrm{Var}(X_j | X_{-j})$, $(\alpha_\star, \sigma_\star)$ depend on two parameters:

- Problem dimension $\kappa = p/n$,
- Signal strength $\gamma = \mathrm{Var}(X^\top \beta)^{1/2}$.

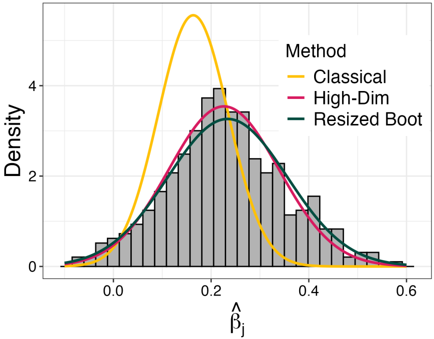
Applying the high-dimensional theory (HDT) to the simulated example

$$\frac{\sqrt{n}(\hat{\beta}_j - \alpha_\star \beta_j)}{\sigma_\star / \tau_j} \xrightarrow{d} \mathcal{N}(0,1),$$



- Estimated inflation from one sample: $\hat{\alpha} = 1.43$ (observed inflation $= 1.41$).
- Estimated SD: $\hat{\sigma} = 0.124$ (observed SD $= 0.128$).

HDT underestimates bias and SD when covariates are heavy-tailed



- Inflation:
  - Obs. inflation = 1.41
  - Est. HDT= 1.38
  - Est. Resized boot = 1.43
- SD
  - Obs. SD = 0.12
  - Est. SD (HDT) = 0.11
  - Est. SD (Resized boot) = 0.12

Simulation setting: Sample $X \sim t_8(0, \Sigma)$ (standardized), Σ is a circular matrix. Simulate Y from a logistic model (100 non-null $\beta_j \sim \mathcal{N}(0, 0.2)$, $\beta_0 = 0$). $n = 2,000$ and $p = 400$.

The bootstrap method

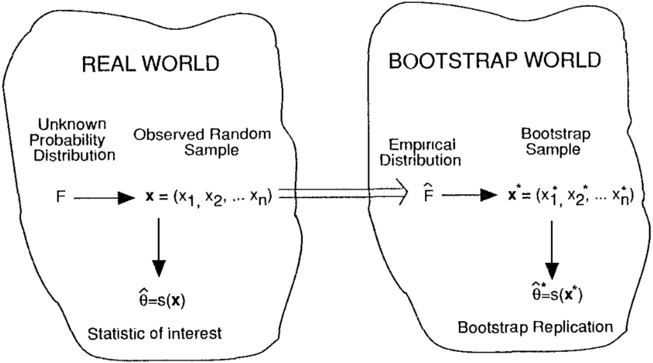


Figure: Figure 8.1 in *An Introduction to the Bootstrap*, by Efron and Tibshirani

What does HDT tell us about pairs bootstrap?

Real world		Bootstrap world	
Unknown distribution	Observed sample	Empirical distribution	Bootstrap sample
$F \longrightarrow$	$(X_i, Y_i) \overset{iid}{\sim} F$, $1 \leq i \leq n$ $\Longrightarrow$	$\hat{F} \longrightarrow$	$(X_i^\star, Y_i^\star) \overset{iid}{\sim} \hat{F}$, $1 \leq i \leq n$
	$\downarrow$		$\downarrow$
	MLE $\hat{\beta} = h(X, Y)$		Bootstrap MLE $\hat{\beta}^\star = h(X^\star, Y^\star)$

- Linear regression:
  - If $p, n \to \infty$, $p^{1+\delta}/n \to 0$ for $\delta > 0$, then bootstrap is weakly consistent (Mammen, 1993).
  - Conservative as $p/n \to \kappa > 0$ (El Karoui & Purdom, 2015).

What does HDT tell us about pairs bootstrap?

Real world

Unknown distribution | Observed sample

$$F \longrightarrow (X_i, Y_i) \overset{iid}{\sim} F, \quad 1 \le i \le n \;\Longrightarrow\; \hat{F} \longrightarrow (X_i^\star, Y_i^\star) \overset{iid}{\sim} \hat{F}, \quad 1 \le i \le n$$

$$\downarrow$$

MLE

$$\hat{\beta} = h(X, Y)$$

Bootstrap world

Empirical distribution | Bootstrap sample

$$\downarrow$$

Bootstrap MLE

$$\hat{\beta}^\star = h(X^\star, Y^\star)$$

- Linear regression:
  - If $p, n \to \infty$, $p^{1+\delta}/n \to 0$ for $\delta > 0$, then bootstrap is weakly consistent (Mammen, 1993).
  - Conservative as $p/n \to \kappa > 0$ (El Karoui & Purdom, 2015).
- Logistic regression: $\#$ distinct obs in a bootstrap sample is $< n$ $\implies$ $\kappa^\star > \kappa$, i.e., we tend to overestimate α and σ.

What about the parametric bootstrap?

Covariates

$(X_1, \dots, X_n)$ $\rightarrow$

Real world

Unknown parameter

$$\beta$$

$$\downarrow$$

Observed sample

$$Y_i \mid X_i \sim F_\beta(X_i)$$

$$\downarrow$$

MLE

$$\hat{\beta} = h(X, Y)$$

Bootstrap world

Estimated parameter

$$\hat{\beta}$$

$$\downarrow$$

Bootstrap sample

$$Y_i^\star \mid X_i \sim F_{\hat{\beta}}(X_i)$$

$$\downarrow$$

Bootstrap MLE

$$\hat{\beta}^\star = h(X^\star, Y^\star)$$

What about the parametric bootstrap?

	Real world		Bootstrap world
	Unknown parameter		Estimated parameter
	β		$\hat{\beta}$
	$\downarrow$		$\downarrow$
Covariates $(X_1, \ldots, X_n)$	$\rightarrow$ Observed sample $Y_i \mid X_i \sim F_\beta(X_i)$		Bootstrap sample $Y_i^\star \mid X_i \sim F_{\hat{\beta}}(X_i)$
	$\downarrow$		$\downarrow$
	MLE $\hat{\beta} = h(X, Y)$		Bootstrap MLE $\hat{\beta}^\star = h(X^\star, Y^\star)$

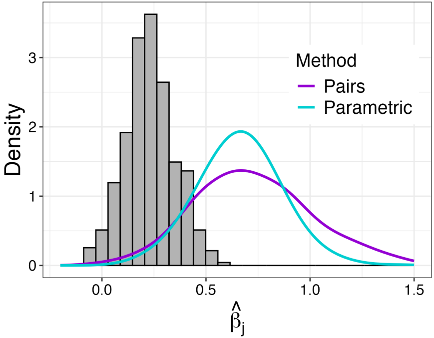
- Linear regression (residual bootstrap):
  - If p fixed, $n \to \infty$, then bootstrap is weakly consistent (Freedman, 1981).
  - $\mathrm{Var}(r_i) \approx (1 - p/n)\sigma_\varepsilon^2$ for i.i.d. Gaussian covariates.

What about the parametric bootstrap?

	Real world	Bootstrap world
	Unknown parameter β	Estimated parameter $\hat{\beta}$
	$\downarrow$	$\downarrow$
Covariates $(X_1, \ldots, X_n)$ $\rightarrow$	Observed sample $Y_i \mid X_i \sim F_\beta(X_i)$	Bootstrap sample $Y_i^\star \mid X_i \sim F_{\hat{\beta}}(X_i)$
	$\downarrow$	$\downarrow$
	MLE $\hat{\beta} = h(X, Y)$	Bootstrap MLE $\hat{\beta}^\star = h(X^\star, Y^\star)$

- Linear regression (residual bootstrap):
  - If p fixed, $n \to \infty$, then bootstrap is weakly consistent (Freedman, 1981).
  - $\mathrm{Var}(r_i) \approx (1 - p/n)\sigma_\varepsilon^2$ for i.i.d. Gaussian covariates.
- Logistic regression: $\hat{\beta}$ is far from $\beta \implies \gamma^\star > \gamma \implies$ we tend to over-estimates α and σ.

Classical pairs & parametric bootstrap fails in the high-dimensional setting!



- Pairs bootstrap:
  - $n_{\text{eff}} < n \implies \kappa^\star > \kappa$.
- Parametric bootstrap:
  - $\hat{\beta} \approx \alpha_\star \beta + \sigma_\star / \tau_j Z$.
  - $\gamma^\star > \gamma$.
- Both overestimate inflation and SD.

The resized parametric bootstrap (**Zhao** & Candès, 2023)

Covariates
$(X_1, \ldots, X_n)$ $\rightarrow$

Real world

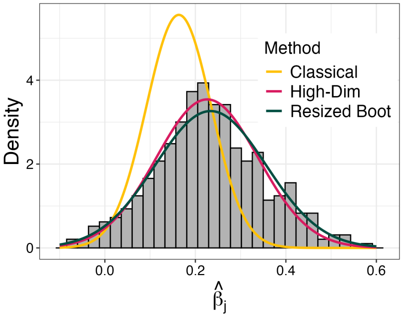
Unknown parameter
β
$\downarrow$
Observed sample
$Y_i \mid X_i \sim F_\beta(X_i)$
$\downarrow$
MLE
$\hat{\beta} = h(X, Y)$

Bootstrap world

Resized MLE
$\beta_\star$
$\downarrow$
Bootstrap sample
$Y_i^\star \mid X_i \sim F_{\hat{\beta}}(X_i)$
$\downarrow$
Bootstrap MLE
$\hat{\beta}^\star = h(X^\star, Y^\star)$

> Resized bootstrap: Resample Y using a rezied MLE $\beta^\star$ s.t. $\mathrm{Var}(X^\top \beta^\star) \approx \gamma^2$.

Applying the resized bootstrap to the previous example



A $(1-q)$ CI is given by

$$\mathrm{CI}_{\mathrm{low}} = \frac{1}{\hat{\alpha}}\left(\hat{\beta}_j - t_j^b[1-q/2]\hat{\sigma}_j\right)$$
$$\mathrm{CI}_{\mathrm{up}} = \frac{1}{\hat{\alpha}}\left(\hat{\beta}_j - t_j^b[q/2]\hat{\sigma}_j\right),$$

- $\hat{\alpha}$ and $\hat{\sigma}_j$ are the estimated inflation and SD using bootstrap samples.
- $t_j^b[q/2]$ and $t_j^b[1-q/2]$ are the q and $(1-q/2)$ quantiles of $\frac{\hat{\beta}_j^b - \hat{\alpha}\beta_{\star,j}}{\hat{\sigma}_j}$.

Implemented in the `R` package `glmhd`.

Summary

- In a high-dimensional logistic regression ($n, p \to \infty$, $p/n \to \kappa > 0$), the MLE is inflated and the SD is larger than given by the inverse Fisher information.
- High-dimensional theory (HDT): When $X \sim \mathcal{N}(0, \Sigma)$, $\tau_j^2 = \mathrm{Var}(X_j \,|\, X_{-j})$ and $\sqrt{n}\tau_j\beta_j = O(1)$, the logistic MLE satisfy

$$\sqrt{n}(\hat{\beta}_j - \alpha_\star\beta_j) \xrightarrow{d} \mathcal{N}(0, \sigma_\star^2/\tau_j^2),$$

- HDT underestimates $\alpha_\star$ and $\sigma_\star$ when X is heavy-tailed.
- The resized parametric bootstrap applies the parametric bootstrap at the shrinked MLE guided by HDT. Distribution of the bootstrap MLE approximates the MLE distribution well when X from a general distribution.

Future studies

- HDT requires $\sqrt{n}\tau_j\beta_j = O(1)$. Empirically we observed SD increases with $|\beta_j|$.
  - Can we characterize how SD changes with β_j?
- Can we apply the high-dimensional theory to genetic studies? For example,
  - Model includes a random effect to account for all the other SNPs.
  - Model multiple phenotypes at the same time.

Thank you!

- Paper: *An Adaptively Resized Parametric Bootstrap for Inference in High-dimensional Generalized Linear Models*, Zhao, Q., and Candès, E., *Statistica Sinica* 2022
- `R` package: zq00/glmhd

References I

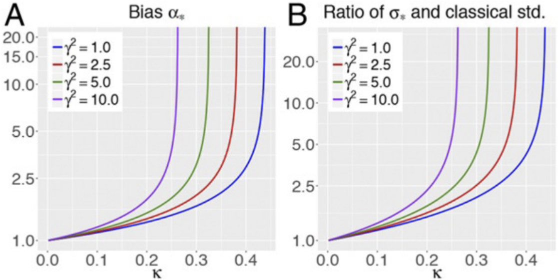
Freedman, D.A.
Bootstrapping regression models *Ann. Stat.*, 1981

Mammen, E.
Bootstrap and wild bootstrap for high-dimensional linear models *Ann.Stat.*, 1993

El Karoui, N. and Purdom, E.
Can We Trust the Bootstrap in High-Dimensions? The Case of Linear Models, *J. Mach. Learn. Res.*, 2018

Zhao, Q., Sur, P. and Candès, E.
The asymptotic distribution of the MLE in high-dimensional logistic models: Arbitrary covariance *Bernoulli*, 2022

How model parameters affect inflation and SD

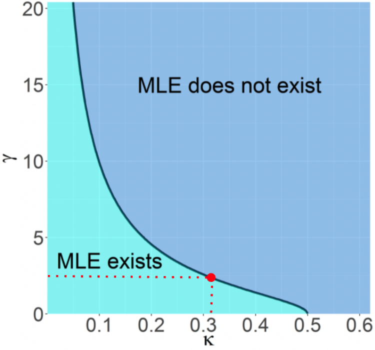


A modern maximum-likelihood theory forhigh-dimensional logistic regression, Sur and Candès, *PNAS*, 2019

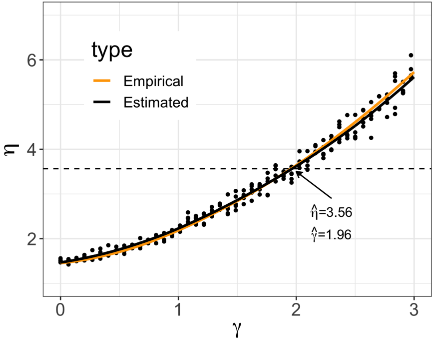
Estimating γ (Method 1)

Using the ProbeFrontier method.

- If $\gamma = 2.32$, then the MLE does not exist if $\kappa > 0.32$
- Sub-sample data to find the threshold in κ
- Estimate γ by the point on the phase transition curve.



Estimating γ (Method 2)



- Use the one-to-one correspondence between $\text{Var}(X_{\text{new}}^{\top}\hat{\beta})$ with γ.
- Use the SLOE estimator to estimate $\text{Var}(X_{\text{new}}^{\top}\hat{\beta})$ from the MLE.
- Apply the parametric bootstrap to compute $\text{Var}(X_{\text{new}}^{\top}\hat{\beta})$ when $\beta = s \times \hat{\beta}$.

Resized bootstrap confidence intervals

- Compute a resized MLE $\beta_\star$
- Generate B bootstrap samples using $\beta_\star$ as the true coefficient, and compute the bootstrap MLE $\hat{\beta}^b$.
- Estimate the inflation and SD of the MLE:
  - $\hat{\sigma}_j^2 = \frac{1}{B-1}\sum_{b=1}^{B}(\hat{\beta}_j^b - \bar{\beta}_j)^2$, where $\bar{\beta}_j = \frac{1}{B}\sum_{b=1}^{B}\bar{\beta}_j^b$.
  - Compute $\hat{\alpha}$ by regressing $\hat{\beta}^b$ onto $\beta_\star$, with weights inversely proportional to $\hat{\sigma}_j^2$.
- Compute the q and $1 - q/2$ quantile of $\frac{\hat{\beta}_j^b - \hat{\alpha}\beta_{\star,j}}{\hat{\sigma}_j}$, denote them as $t_j^b[q/2]$ and $t_j^b[1 - q/2]$.
- Compute a $(1 - q)$ CI as

$$\left[\frac{1}{\hat{\alpha}}\left(\hat{\beta}_j - t_j^b[1 - q/2]\hat{\sigma}_j\right), \frac{1}{\hat{\alpha}}\left(\hat{\beta}_j - t_j^b[q/2]\hat{\sigma}_j\right)\right] \tag{1}$$