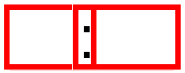
Statistical inference based on optimal subdata

HaiYing Wang

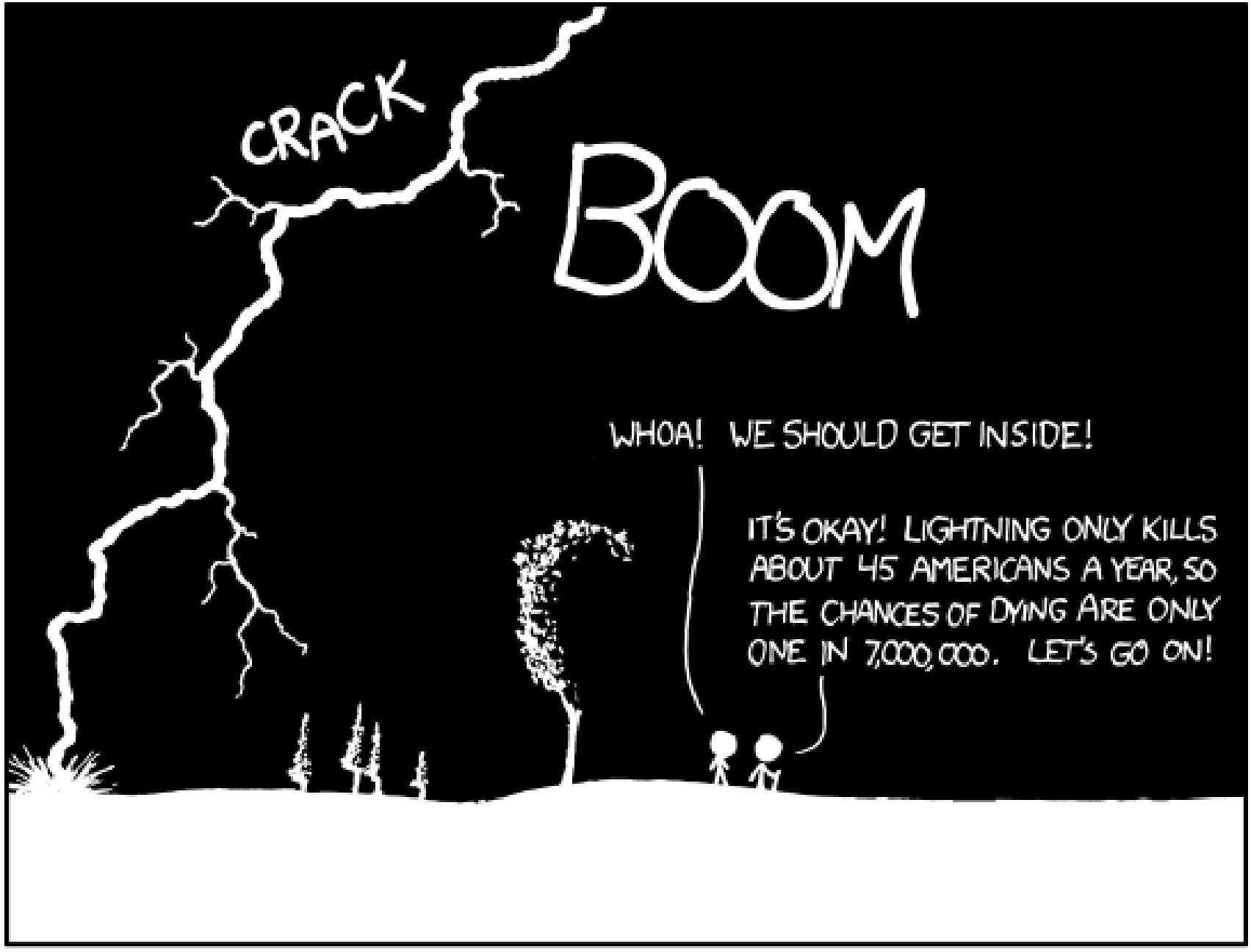
University of Connecticut

CIRM, April 30, 2018



Outline

Outline



THE ANNUAL DEATH RATE AMONG PEOPLE WHO KNOW THAT STATISTIC IS ONE IN SIX.

- Subsampling is a practical technique to extract useful information with limited computational resources.
- The key to success of subsampling methods is to specify nonuniform sampling probabilities to include the "most informative" data points.
- For this purpose, numerous approaches have been developed.
  - Leveraging sampling: Ma, *et al* (2014, ICML; 2015 JMLR) [1, 2].
  - OSMAC: Wang, *et al* (2017, JASA) [3]; Wang (2018) [4].
  - IBOSS: Wang, *et al* (2018, JASA) [5].
- Optimal subsample are often biased
  - The original data $(\mathbf{x}_1, y_1), ..., (\mathbf{x}_N, y_N) \sim f(\mathbf{x}, y; \boldsymbol{\beta})$.
  - An optimal subsample $(\mathbf{x}_1^*, y_1^*), ..., (\mathbf{x}_n^*, y_n^*) \nsim f(\mathbf{x}, y; \boldsymbol{\beta})$.
- Weighted estimator is often used to remove bias, but it may reduce estimation efficiency.

Logistic regression

- Given the covariate $\mathbf{x} \in \mathbb{R}^d$, logistic regression models are of the form

$$P(y=1|\mathbf{x}) = p(\boldsymbol{\beta}) = \frac{\exp(\mathbf{x}^T\boldsymbol{\beta})}{1+\exp(\mathbf{x}^T\boldsymbol{\beta})}.$$

  where $y \in \{0,1\}$ is the response variable
  $\boldsymbol{\beta} \in \Theta$ is a $d \times 1$ vector of unknown regression parameters.

- Let $\mathcal{D}_N = \{(x_1,y_1),...,(x_N,y_N)\}$ be the full data. The MLE is

$$\hat{\boldsymbol{\beta}}_f = \arg\max_{\boldsymbol{\beta}} \sum_{i=1}^{N} \left\{ y_i \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta} - \log\left(1+e^{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i}\right)\right\}. \quad (1)$$

- There is no general closed-form solution to $\hat{\boldsymbol{\beta}}_f$, and for massive data ($N$ is large), iterative calculations may take too long.

General Subsampling Procedure

- Let π_1, ..., π_N be subsampling probabilities such that $\sum \pi_i = 1$, where π_i can be dependent on the full data $\mathcal{D}_N$.
- **Sample with replacement** for a subsample of size $n \ll N$ from the full sample according to the probability $\{\pi_i\}_{i=1}^N$, denoted as (x_i^*, y_i^*), $i = 1, ..., n$.
- Let the subsample estimator $\hat{\boldsymbol{\beta}}_w^{\pi}$ be the weighted MLE, i.e.,

$$\hat{\boldsymbol{\beta}}_w^{\pi} = \arg\max_{\boldsymbol{\beta}} \sum_{i=1}^{n} \frac{y_i^* \boldsymbol{\beta}^{\mathrm{T}} \mathbf{x}_i^* - \log\left(1 + e^{\boldsymbol{\beta}^{\mathrm{T}} \mathbf{x}_i^*}\right)}{\pi_i^*}.$$

- The inverse probability weighting is used to remove bias.

Outline

OSMAC

- The $\hat{\boldsymbol{\beta}}_w^{\pi}$ is asymptotically normal, namely, conditional on $\mathcal{D}_N$

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_w^{\pi} - \hat{\boldsymbol{\beta}}_f) \overset{a}{\sim} \mathbb{N}(\mathbf{0}, \mathbf{V}_N), \text{ where } \mathbf{V}_N = \mathbf{M}_N^{-1}\mathbf{V}_{Nc}\mathbf{M}_N^{-1}$$

- The **O**ptimal **S**ubsampling **M**ethod under the **A**-optimality **C**riterion specifies

$$\pi_i^{\mathrm{OS}}(\boldsymbol{\beta}) = \frac{|y_i - p_i(\boldsymbol{\beta})|h(\mathbf{x}_i)}{\sum_{j=1}^{N}|y_j - p_j(\boldsymbol{\beta})|h(\mathbf{x}_j)}, \quad i = 1, ..., N. \tag{2}$$

- **A-optimality**: If $h(\mathbf{x}) = \|\mathbf{M}_N^{-1}\mathbf{x}\|$, then π_i^{OS} minimize $\mathrm{tr}(\mathbf{V}_N)$. Denote it as π_i^{mMSE}.

- **L-optimality**: If $h(\mathbf{x}) = \|\mathbf{x}\|$, then π_i^{OS} minimize $\mathrm{tr}(\mathbf{V}_{Nc})$. Denote it as π_i^{mVc}.

- **Local Case-control**: $h(\mathbf{x}) = 1$.

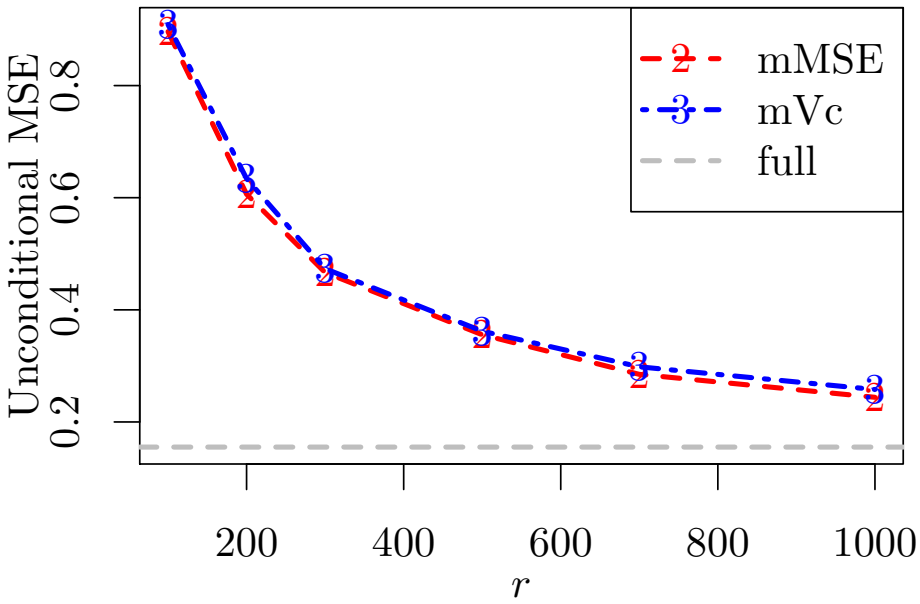
The OSMAC estimator

- Since $\pi_i^{\mathrm{OS}}(\boldsymbol{\beta})$ depend $\boldsymbol{\beta}$, a pilot estimate is required.
- Let $\hat{\boldsymbol{\beta}}_1$ be a data-dependent pilot estimator.
- The original weighted OSMAC estimator is

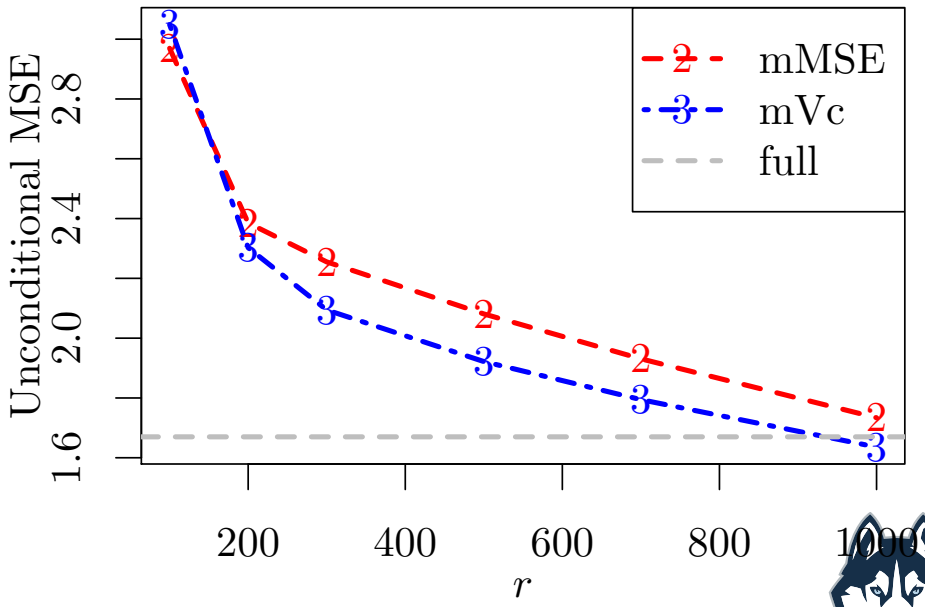
$$\hat{\boldsymbol{\beta}}_w = \arg\max_{\boldsymbol{\beta}} \sum_{i=1}^{n} \frac{y_i^* \boldsymbol{\beta}^{\mathrm{T}} \mathbf{x}_i^* - \log\left(1 + e^{\boldsymbol{\beta}^{\mathrm{T}} \mathbf{x}_i^*}\right)}{\pi_i^{\mathrm{OS}}(\hat{\boldsymbol{\beta}}_1)^*}.$$

Numerical results on rare events data

- Full data size $N = 10,000$, $d = 7$, $\boldsymbol{\beta}_t$ is a 7×1 vector of 0.5.
- $\mathbf{x} \sim \mathbb{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}_{ij} = 0.5^{I(i \neq j)}$.
- Consider two values of $\boldsymbol{\mu}$:
  1) $\boldsymbol{\mu} = \mathbf{-2.14}$ so that 1.01% of the responses are 1's.
  2) $\boldsymbol{\mu} = \mathbf{-2.9}$ so that 0.14% of the responses are 1's.



(a) 1.01% of y_i's are 1



(b) 0.14% of y_i's are 1

Comparisons with deep learning (DL)

Consider a supersymmetric (SUSY) benchmark data set. [1]

- $N = 5,000,000$, $d = 18$.
- The goal is to distinguish determine whether new SUSY particles are produced using 18 kinematic properties.
- The DL method (Baldi *et al.* 2014) produced an AUC of **0.88**.
- Our method gives an AUC of about **0.85**.
- Baldi *et al.*'s (2014) method requires special computing resources and coding skills.
- Anyone with basic programming ability are able to implement our method.

[1] P. Baldi, P. Sadowski, and D. Whiteson. Searching for exotic particles in high-energy physics with deep learning. *Nature Communications*, 5(4308), 2014.

Our method	DL
$n_1 + n = 1000$	$N = 5,000,000$
Logistic model	A five-layer neural nets with 300 hidden units in each layer
Newton's Method	Combinations of pre-training methods, network architectures, initial learning rates, and regularization methods
A normal PC with an Intel I7 processor and 8GB memory	Machines with 16 Intel Xeon cores, an NVIDIA Tesla C2070 graphics processor, and 64 GB memory. All neural networks were trained using the GPU-accelerated Theano and Pylearn2 software libraries

Outline

Drawbacks of the weighting scheme

$$\hat{\boldsymbol{\beta}}_w = \arg\max_{\boldsymbol{\beta}} \sum_{i=1}^{n} \frac{y_i^* \boldsymbol{\beta}^{\mathrm{T}} \mathbf{x}_i^* - \log\left(1 + e^{\boldsymbol{\beta}^{\mathrm{T}} \mathbf{x}_i^*}\right)}{\pi_i^{\mathrm{OS}}(\hat{\boldsymbol{\beta}}_1)^*}.$$

- Intuitively, a larger $\pi_i^{\mathrm{OS}}(\hat{\boldsymbol{\beta}}_1)$ means that the data point $(\mathbf{x}_i, y_i)$ contains more information about $\boldsymbol{\beta}$, but it has a smaller weight in the weighted log-likelihood.
- The weighting reduce the contributions of more informative data points to the likelihood function.
- An unweighted estimator may have better performance, but we have to correct the bias.

Unweighted estimator

Let the pilot estimate be $\hat{\boldsymbol{\beta}}_1$, and subsampling probabilities be

$$\pi_i^{\mathrm{OS}}(\hat{\boldsymbol{\beta}}_1) = \frac{|y_i - p_i(\hat{\boldsymbol{\beta}}_1)|h(\mathbf{x}_i)}{\sum_{j=1}^{N}|y_j - p_j(\hat{\boldsymbol{\beta}}_1)|h(\mathbf{x}_j)} \quad i = 1, ..., N, \tag{3}$$

where $h(\mathbf{x})$ is a univariate function of $\mathbf{x}$.
Use the subsample, calculate

$$\tilde{\boldsymbol{\beta}}_{uw} = \arg\max_{\boldsymbol{\beta}} \ell_r^*(\boldsymbol{\beta}) = \arg\max_{\boldsymbol{\beta}} \sum_{i=1}^{n} \left\{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i^* y_i^* - \log\left(1 + e^{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i^*}\right)\right\}; \tag{4}$$

the unweighted estimator is

$$\hat{\boldsymbol{\beta}}_{uw} = \tilde{\boldsymbol{\beta}}_{uw} + \hat{\boldsymbol{\beta}}_1. \tag{5}$$

Asymptotic normality

Theorem

Under mild assumptions, conditional on $\mathcal{D}_N$ *and* $\hat{\boldsymbol{\beta}}_1$, *as* $n, N \to \infty$,

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{uw} - \hat{\boldsymbol{\beta}}_{wf}) \xrightarrow{d} \mathbb{N}\big(\mathbf{0},\ \boldsymbol{\Sigma}_{\boldsymbol{\beta}_t}\big), \tag{6}$$

where $\hat{\boldsymbol{\beta}}_{wf}$ *is a weighted MLE based on the full data satisfying that*

$$\sqrt{N}(\hat{\boldsymbol{\beta}}_{wf} - \boldsymbol{\beta}_t) \xrightarrow{d} \mathbb{N}\big(\mathbf{0},\ \boldsymbol{\Sigma}_{wf}\big). \tag{7}$$

If $n/N \to 0$, *then*

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{uw} - \boldsymbol{\beta}_t) \xrightarrow{d} \mathbb{N}\big(\mathbf{0},\ \boldsymbol{\Sigma}_{\boldsymbol{\beta}_t}\big). \tag{8}$$

Asymptotic normality

$$\hat{\boldsymbol{\beta}}_{wf} = \arg\max_{\boldsymbol{\beta}} \sum_{i=1}^{N} |y_i - p_i(\hat{\boldsymbol{\beta}}_1)| h(\mathbf{x}_i) \left[y_i \mathbf{x}_i^{\mathrm{T}} (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_1) - \log\{1 + e^{\mathbf{x}_i^{\mathrm{T}}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_1)}\} \right]$$

$$\boldsymbol{\Sigma}_{\boldsymbol{\beta}} = \left[\frac{\mathbb{E}\{\phi(\boldsymbol{\beta}) h(\mathbf{x}) \mathbf{x}\mathbf{x}^{\mathrm{T}}\}}{4\Phi(\boldsymbol{\beta})} \right]^{-1}$$

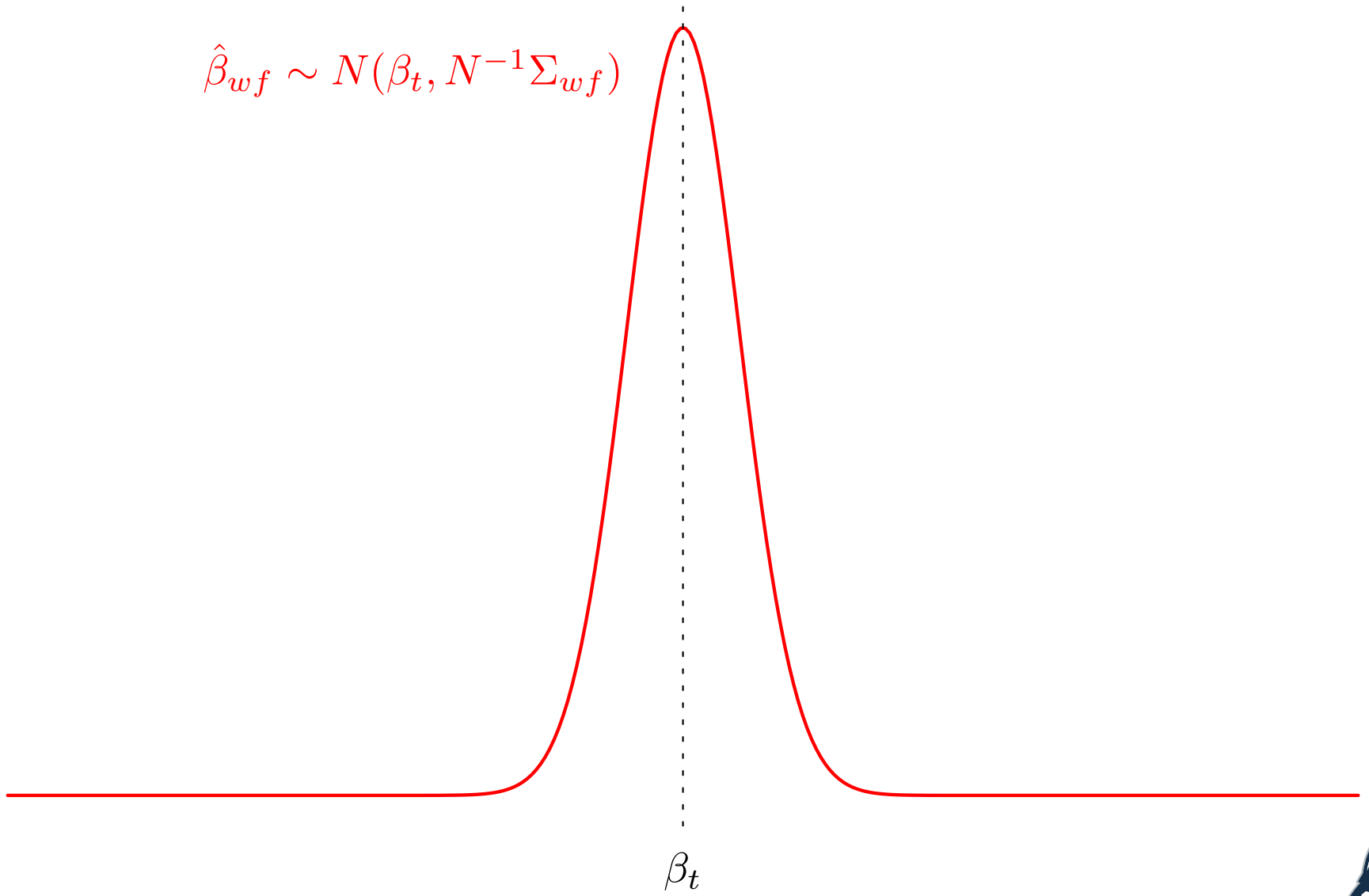
$$\Phi(\boldsymbol{\beta}) = \mathbb{E}\{\phi(\boldsymbol{\beta}) h(\mathbf{x})\}$$

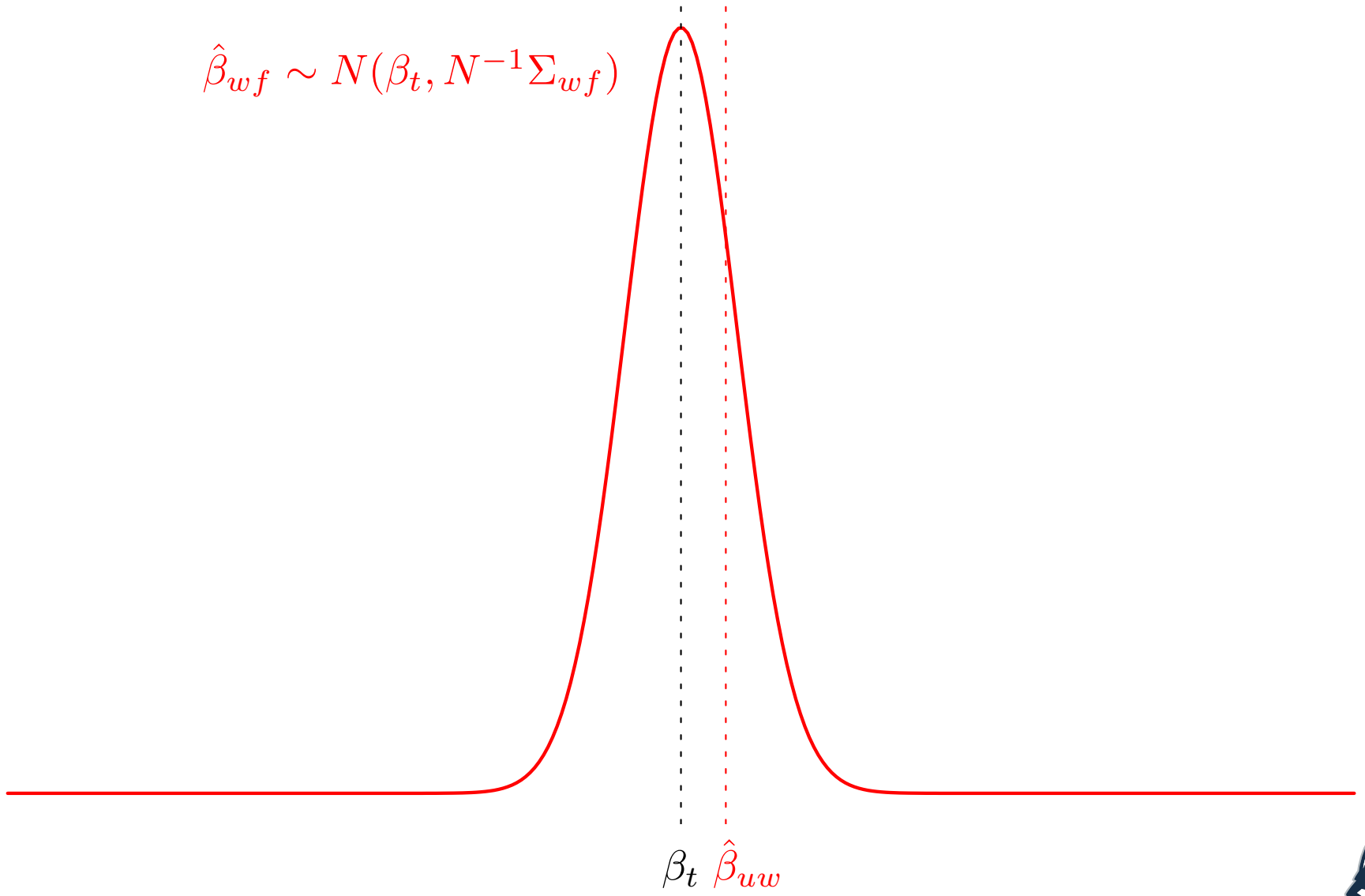
$$\phi(\boldsymbol{\beta}) = p(\boldsymbol{\beta})\{1 - p(\boldsymbol{\beta})\}$$

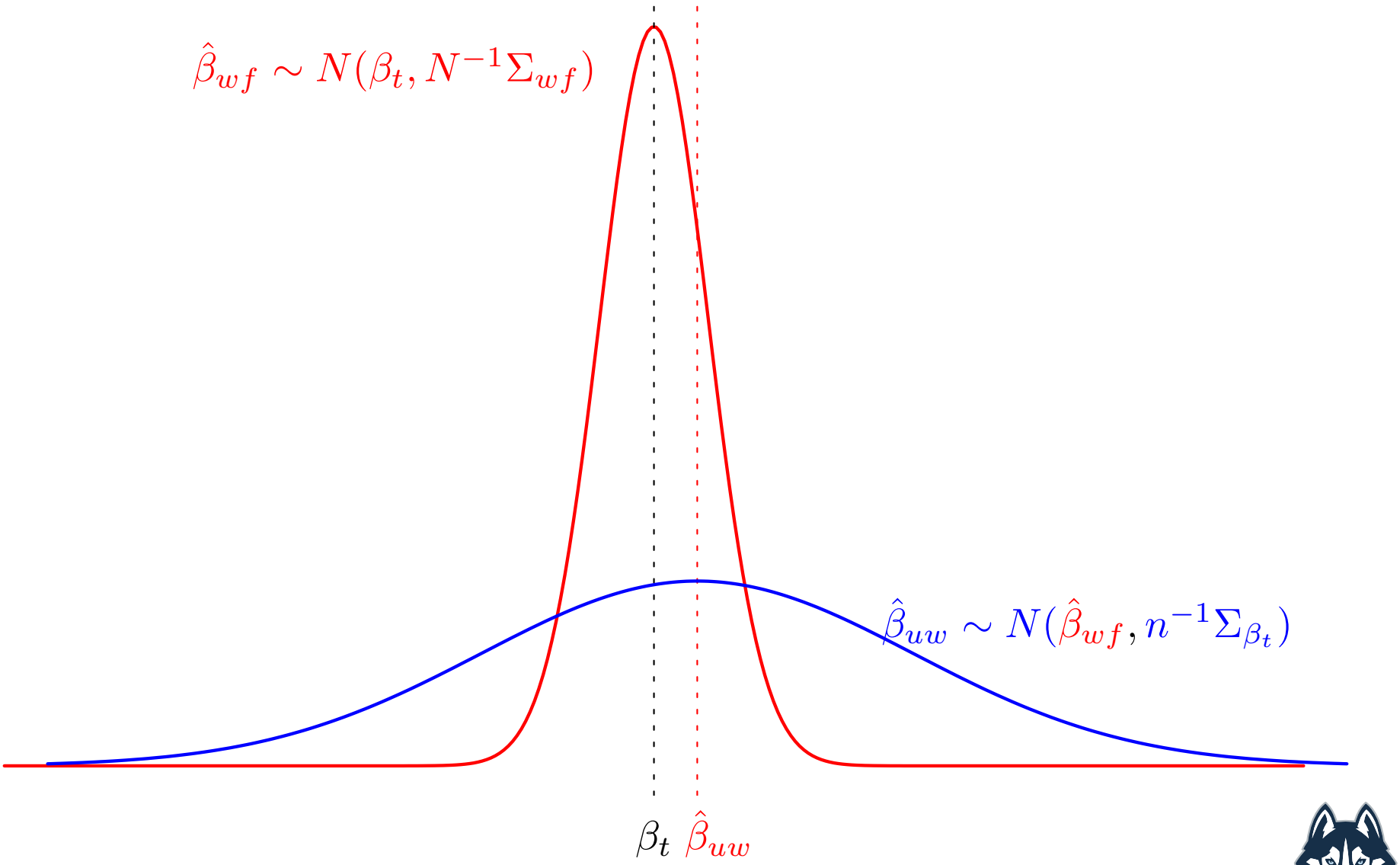
$$\boldsymbol{\Sigma}_{wf} = \frac{\boldsymbol{\Sigma}_{\boldsymbol{\beta}} \mathbb{E}\{\phi(\boldsymbol{\beta}_t) h^2(\mathbf{x}) \mathbf{x}\mathbf{x}\} \boldsymbol{\Sigma}_{\boldsymbol{\beta}}}{16\Phi^2(\boldsymbol{\beta})}$$

Remarks

- The naive unweighted estimator $\tilde{\boldsymbol{\beta}}_{uw}$ is biased, and $\tilde{\boldsymbol{\beta}}_{uw} \xrightarrow{p} 0$.
- $\hat{\boldsymbol{\beta}}_{uw}$ varies around $\hat{\boldsymbol{\beta}}_{wf}$ with variance-covariance matrix $n^{-1}\boldsymbol{\Sigma}_{\boldsymbol{\beta}_t}$,
- $\hat{\boldsymbol{\beta}}_{wf}$ varies around $\boldsymbol{\beta}_t$ with variance-covariance matrix $N^{-1}\boldsymbol{\Sigma}_{wf}$.
- Thus, both $n^{-1}\boldsymbol{\Sigma}_{\boldsymbol{\beta}_t}$ and $N^{-1}\boldsymbol{\Sigma}_{wf}$ should be considered in accessing the quality of $\hat{\boldsymbol{\beta}}_{uw}$ in estimating the true parameter $\boldsymbol{\beta}_t$.
- It is expected that $n \ll N$ for computational benefit. Thus, $n^{-1}\boldsymbol{\Sigma}_{\boldsymbol{\beta}_t}$ is the dominating term in quantifying the variation of $\hat{\boldsymbol{\beta}}_{uw}$.
- If $n/N \to 0$, then the variation of $\hat{\boldsymbol{\beta}}_{wf}$ can be ignored.


$\hat{\beta}_{wf} \sim N(\beta_t, N^{-1}\Sigma_{wf})$
β_t


$\hat{\beta}_{wf} \sim N(\beta_t, N^{-1}\Sigma_{wf})$
β_t
$\hat{\beta}_{uw}$


$\hat{\beta}_{wf} \sim N(\beta_t, N^{-1}\Sigma_{wf})$
$\hat{\beta}_{uw} \sim N(\hat{\beta}_{wf}, n^{-1}\Sigma_{\beta_t})$
β_t
$\hat{\beta}_{uw}$

Weighted OSMAC estimator

For the weighted estimator, conditional on $\mathcal{D}_N$ and $\hat{\boldsymbol{\beta}}_1$, as $n, N \to \infty$,

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_w - \hat{\boldsymbol{\beta}}_f) \xrightarrow{d} \mathbb{N}(\mathbf{0}, \mathbf{V}^{\mathrm{OS}})$$
$$\sqrt{N}(\hat{\boldsymbol{\beta}}_f - \boldsymbol{\beta}_t) \xrightarrow{d} \mathbb{N}\big(\mathbf{0}, \mathbf{M}^{-1}\big),$$

where $\mathbf{V}^{\mathrm{OS}} = \mathbf{M}^{-1}\mathbf{V}_c^{\mathrm{OS}}\mathbf{M}^{-1}$,

$$\mathbf{M} = \mathbb{E}\{\phi(\boldsymbol{\beta}_t)\mathbf{x}\mathbf{x}^{\mathrm{T}}\} \quad \text{and} \quad \mathbf{V}_c^{\mathrm{OS}} = 4\Phi(\boldsymbol{\beta}_t)\mathbb{E}\left\{\frac{\phi(\boldsymbol{\beta}_t)\mathbf{x}\mathbf{x}^{\mathrm{T}}}{h(\mathbf{x})}\right\}.$$

- Both $n^{-1}\mathbf{V}^{\mathrm{OS}}$ and $N^{-1}\mathbf{M}^{-1}$ should be considered in accessing the quality of $\hat{\boldsymbol{\beta}}_w$ in estimating $\boldsymbol{\beta}_t$.
- Since typically $n \ll N$, $n^{-1}\mathbf{V}^{\mathrm{OS}}$ is the dominating term.
- Therefore, the relative performance between $\hat{\boldsymbol{\beta}}_{uw}$ and $\hat{\boldsymbol{\beta}}_w$ are mainly determined by the relative magnitude between $\mathbf{V}^{\mathrm{OS}}$ and $\boldsymbol{\Sigma}_{\boldsymbol{\beta}_t}$.

Comparison with weighted estimator

Proposition

Suppose that $\mathbf{M}$, $\mathbf{V}_c^{\mathrm{OS}}$, *and* $\boldsymbol{\Sigma}_{\boldsymbol{\beta}_t}$ *are finite and positive definite matrices. We have that*

$$\boldsymbol{\Sigma}_{\boldsymbol{\beta}_t} \leq \mathbf{V}^{\mathrm{OS}}, \tag{9}$$

where the inequality is in the Loewner ordering, i.e., for positive semi-definite matrices $\mathbf{A}$ *and* $\mathbf{B}$, $\mathbf{A} \geq \mathbf{B}$ *if and only if* $\mathbf{A} - \mathbf{B}$ *is positive semi-definite. If* $h(\mathbf{x}) = 1$, *then the equality in* (9) *holds.*

- This means that the unweighted estimator has a smaller variance-covariance matrix.

Combine the two stage estimates

$$\check{\boldsymbol{\beta}}_{uw} = \{\ddot{\ell}_r^{*1}(\tilde{\boldsymbol{\beta}}_1) + \ddot{\ell}_{uw}^{*}(\tilde{\boldsymbol{\beta}}_{uw})\}^{-1}\{\ddot{\ell}_r^{*1}(\tilde{\boldsymbol{\beta}}_1)\hat{\boldsymbol{\beta}}_1 + \ddot{\ell}_{uw}^{*}(\tilde{\boldsymbol{\beta}}_{uw})\hat{\boldsymbol{\beta}}_{uw}\}$$

where $\ddot{\ell}_r^{*1}(\tilde{\boldsymbol{\beta}}_1) = \sum_{i=1}^{n_1} \phi^{*1}(\tilde{\boldsymbol{\beta}}_1)\mathbf{x}_i^{*1}(\mathbf{x}_i^{*1})^{\mathrm{T}}$ and

$\ddot{\ell}_r^{*}(\tilde{\boldsymbol{\beta}}_{uw}) = \sum_{i=1}^{n} \phi^{*}(\tilde{\boldsymbol{\beta}}_{uw})\mathbf{x}_i^{*}(\mathbf{x}_i^{*})^{\mathrm{T}}$.

The variance-covariance matrix of $\check{\boldsymbol{\beta}}_{uw}$ can be estimated by

$$\hat{\mathbb{V}}(\check{\boldsymbol{\beta}}_{uw}) = \{\ddot{\ell}_r^{*1}(\tilde{\boldsymbol{\beta}}_1) + \ddot{\ell}_{uw}^{*}(\tilde{\boldsymbol{\beta}}_{uw})\}^{-1} \tag{10}$$

- To combine $\hat{\boldsymbol{\beta}}_1$ and $\hat{\boldsymbol{\beta}}_{uw}$ using $\ddot{\ell}_r^{*1}(\tilde{\boldsymbol{\beta}}_1)$ and $\ddot{\ell}_r^{*}(\tilde{\boldsymbol{\beta}}_{uw})$, the inconsistent estimators $\tilde{\boldsymbol{\beta}}_1$ and $\tilde{\boldsymbol{\beta}}_{uw}$ should be used.
- This is an advantage for implementation using existing software to fit logistic regression.
- Just use the inverse of the estimated variance-covariance matrix from the software to replace the second derivative of the likelihood.

Sampling with replacement vs Poisson sa

$$\pi_i^{\mathrm{OS}}(\hat{\boldsymbol{\beta}}_1) = \frac{N^{-1}|y_i - p_i(\hat{\boldsymbol{\beta}}_1)|h(\mathbf{x}_i)}{N^{-1}\sum_{j=1}^{N}|y_j - p_j(\hat{\boldsymbol{\beta}}_1)|h(\mathbf{x}_j)} \quad i = 1, ..., N,$$

- Sampling with replacement
  - Advantages
    - The subsample are i.i.d conditional on the full data.
    - It is faster to compute than sampling without replacement.
  - Disadvantages
    - **Need to calculate sampling probabilities all at once.**
    - **The subsample are not independent unconditionally.**
- Poisson sampling
  - Advantages
    - **The subsample are independent unconditionally.**
    - **No need to calculate sampling probabilities all at once.**
    - It is fast to compute.
  - Disadvantage
    - Subsample size is random.

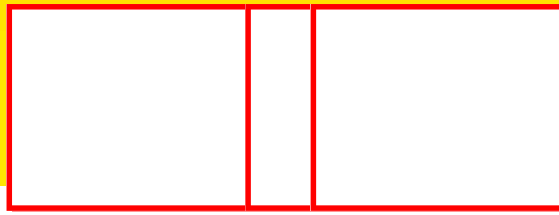
Algorithms based on Poisson sampling

- Let $\hat{\boldsymbol{\beta}}_1$ be pilot estimates of $\boldsymbol{\beta}$;
  $\hat{\Psi}_1$ be pilot estimates of $\Psi(\boldsymbol{\beta}) = \mathbb{E}\{|y - p(\boldsymbol{\beta})|h(\mathbf{x})\}$.
- For $i = 1, ..., N$,
  - calculate $\pi_i^p(\hat{\boldsymbol{\beta}}_1) = \dfrac{|y_i - p_i(\hat{\boldsymbol{\beta}}_1)|h(\mathbf{x}_i)}{\hat{\Psi}_1}$;
  - generate $u_i \sim U(0,1)$;
  - if $u_i \leq n\pi_i^p(\hat{\boldsymbol{\beta}}_1)$, include $\{\mathbf{x}_i, y_i, \pi_i^p\}$ in the subsample;
- For the obtained subsample, calculate

$$\tilde{\boldsymbol{\beta}}_{poi} = \arg\max_{\boldsymbol{\beta}} \sum_{i=1}^{n^*} (n\pi_i^{p*} \vee 1)\{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i^* y_i^* + \log(1 + e^{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i^*})\},$$

and let $\hat{\boldsymbol{\beta}}_{poi} = \tilde{\boldsymbol{\beta}}_{poi} + \hat{\boldsymbol{\beta}}_1$.

Asymptotic normality



Theorem

Under mild assumptions, conditional on $\mathcal{D}_N$ *and* $\hat{\boldsymbol{\beta}}_1$, *as* $n \to \infty$ *and* $N \to \infty$, *if* $n/N \to 0$

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{poi} - \boldsymbol{\beta}_t) \xrightarrow{d} \mathbb{N}(0,\ \boldsymbol{\Sigma}_{\boldsymbol{\beta}_t}); \tag{11}$$

if $n/N \to \rho \in (0,1)$, *then*

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{poi} - \hat{\boldsymbol{\beta}}_{wf}) \xrightarrow{d} \mathbb{N}(0,\ \boldsymbol{\Sigma}_{\boldsymbol{\beta}_t}\boldsymbol{\Lambda}_\rho\boldsymbol{\Sigma}_{\boldsymbol{\beta}_t}). \tag{12}$$

Furthermore,

$$\boldsymbol{\Sigma}_{\boldsymbol{\beta}_t}\boldsymbol{\Lambda}_\rho\boldsymbol{\Sigma}_{\boldsymbol{\beta}_t} < \boldsymbol{\Sigma}_{\boldsymbol{\beta}_t}, \tag{13}$$

under the Loewner ordering.

Unconditional asymptotic normality

Poisson subsampling produce unconditionally independent sample, so it is possible to obtain unconditional asymptotic distribution.

Theorem

Under some mild assumptions, as $n \to \infty$ *and* $N \to \infty$, *if* $n/N \to 0$

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{poi} - \boldsymbol{\beta}_t) \xrightarrow{d} \mathbb{N}(0,\ \boldsymbol{\Sigma}_{\boldsymbol{\beta}_t}); \tag{14}$$

if $n/N \to \rho \in (0, 1)$, *then*

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{poi} - \boldsymbol{\beta}_t) \longrightarrow \mathbb{N}(0,\ \boldsymbol{\Sigma}_{\boldsymbol{\beta}_t}\boldsymbol{\Lambda}_u\boldsymbol{\Sigma}_{\boldsymbol{\beta}_t}). \tag{15}$$

Furthermore,

$$\boldsymbol{\Sigma}_{\boldsymbol{\beta}_t}\boldsymbol{\Lambda}_u\boldsymbol{\Sigma}_{\boldsymbol{\beta}_t} \geq \boldsymbol{\Sigma}_{\boldsymbol{\beta}_t} > \boldsymbol{\Sigma}_{\boldsymbol{\beta}_t}\boldsymbol{\Lambda}_\rho\boldsymbol{\Sigma}_{\boldsymbol{\beta}_t}, \tag{16}$$

Outline

Simulation setup

Here we used the same simulation setup as in Wang *et al* (2017)

- Full data of size $N = 10,000$ are generated from a logistic model with $\boldsymbol{\beta}$ being a 7×1 vector of 0.5.
- Consider the following distributions for $\mathbf{x}$
  1) **mzNormal**. $\mathbf{x} \sim \mathbb{N}(\mathbf{0}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}_{ij} = 0.5^{I(i \neq j)}$.
  2) **nzNormal**. $\mathbf{x} \sim \mathbb{N}(\mathbf{1.5}, \boldsymbol{\Sigma})$. About 95% of responses are 1.
  3) **ueNormal**. $\mathbf{x} \sim \mathbb{N}(\mathbf{0}, \boldsymbol{\Sigma}_{ue})$; diagonal elements of $\boldsymbol{\Sigma}_{ue}$ are unequal.
  4) **mixNormal**. $\mathbf{x} \sim 0.5\mathbb{N}(\mathbf{1}, \boldsymbol{\Sigma}) + 0.5\mathbb{N}(-\mathbf{1}, \boldsymbol{\Sigma})$.
  5) $\mathbf{T_3}$. $\mathbf{x} \sim \mathbb{T}_3(\mathbf{0}, \boldsymbol{\Sigma})/10$.
  6) **EXP**. Components of $\mathbf{x}$ are independent and $x_j \sim \mathbb{EXP}(2)$. About 84% of responses are 1.

MSE: performance on parameter estimati

We calculate the estimation efficiency of $\check{\boldsymbol{\beta}}_{\text{uw}}$ relative to $\check{\boldsymbol{\beta}}_{w}$ as

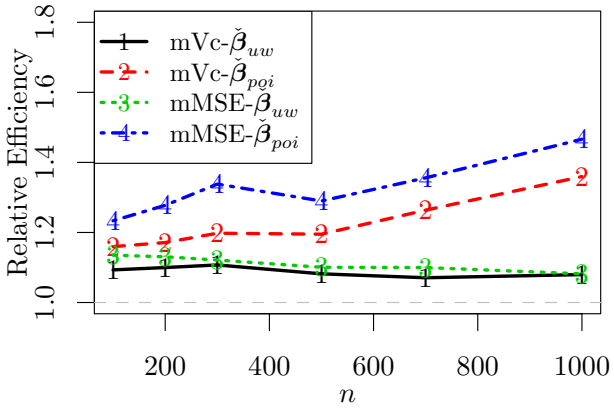
$$\text{Relative Efficiency} = \frac{\text{MSE}(\check{\boldsymbol{\beta}}_{w})}{\text{MSE}(\check{\boldsymbol{\beta}}_{\text{uw}})}. \tag{17}$$

We consider

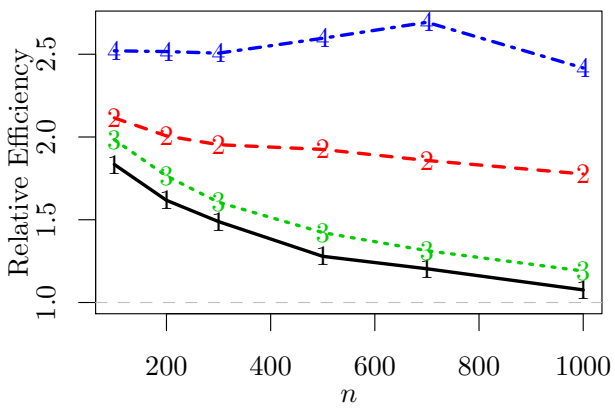
- $\check{\boldsymbol{\beta}}_{\text{uw}} = \check{\boldsymbol{\beta}}_{uw}$ for subsampling with replacement
- $\check{\boldsymbol{\beta}}_{\text{uw}} = \check{\boldsymbol{\beta}}_{poi}$ for Poisson subsampling
- $\pi_i^{\text{OS}} = \pi_i^{\text{mMSE}}$ (A-optimality)
- $\pi_i^{\text{OS}} = \pi_i^{\text{mVc}}$ (L-optimality)
- We calculate empirical MSEs from $S = 1000$ subsamples using

$$\text{MSE}(\check{\boldsymbol{\beta}}) = S^{-1} \sum_{s=1}^{S} \|\check{\boldsymbol{\beta}}^{(s)} - \boldsymbol{\beta}_t\|^2.$$

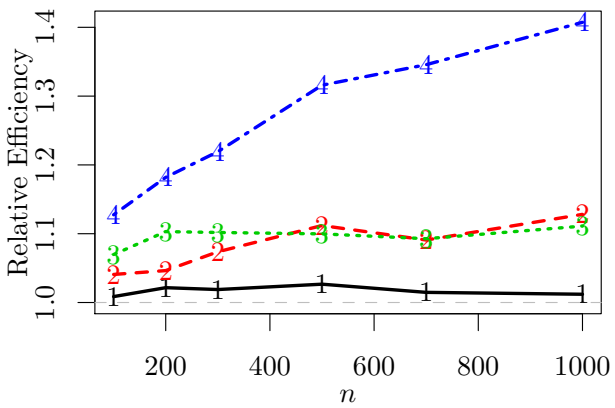
Relative Efficiency: $n_1 = 200$, n varies.



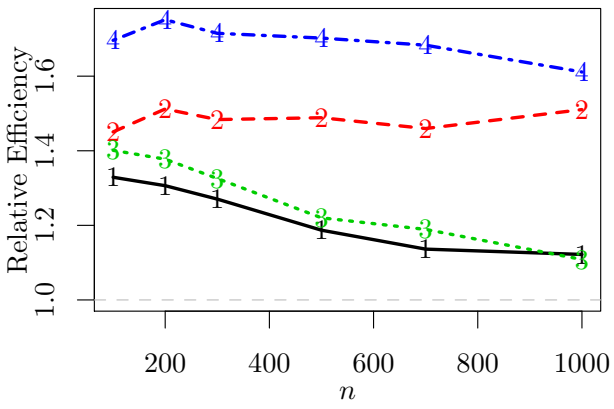
(a) mzNormal



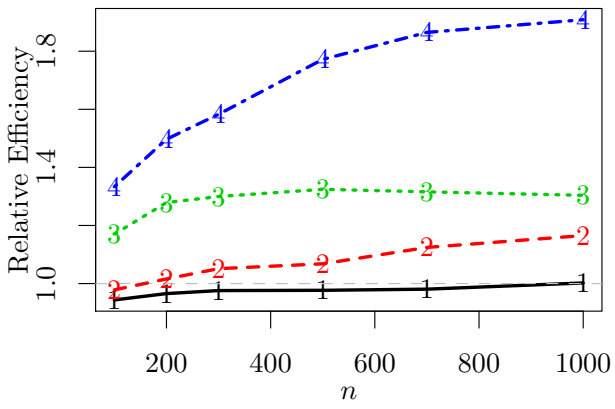
(b) nzNormal



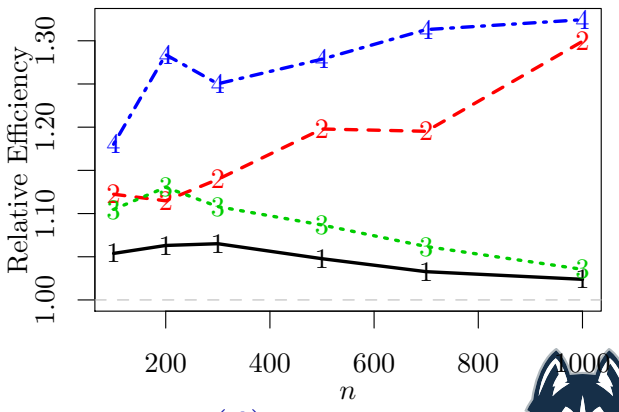
(c) ueNormal



(d) mixNormal

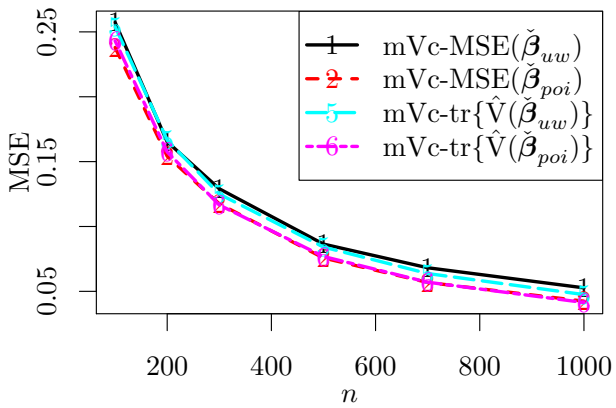


(e) T_3

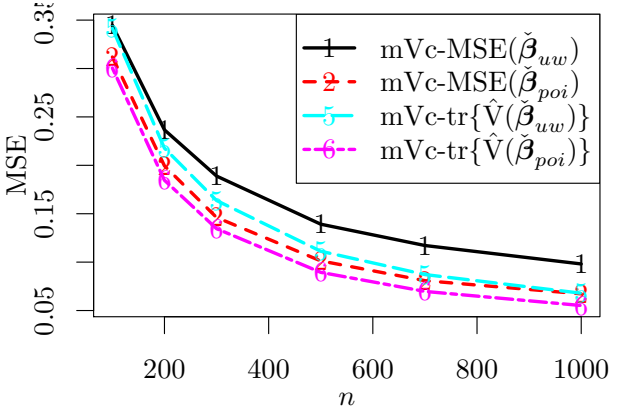


(f) EXP

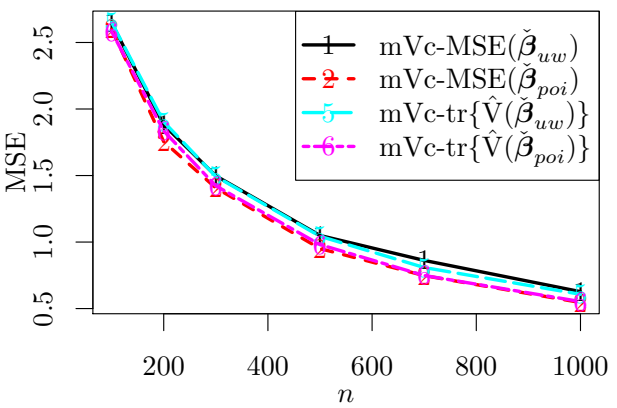
MSE($\check{\boldsymbol{\beta}}$) and estimated MSE, tr$\{\hat{\mathbb{V}}(\check{\boldsymbol{\beta}})\}$.



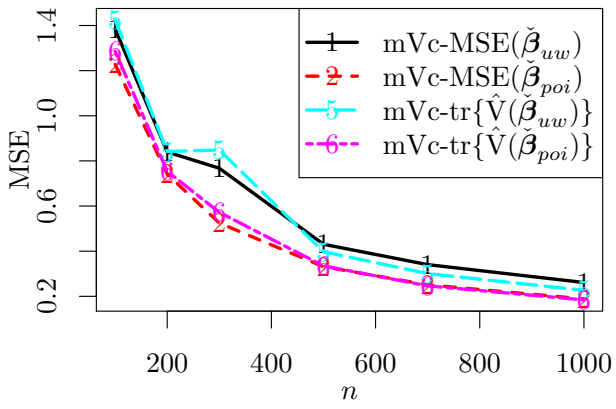
(a) mzNormal



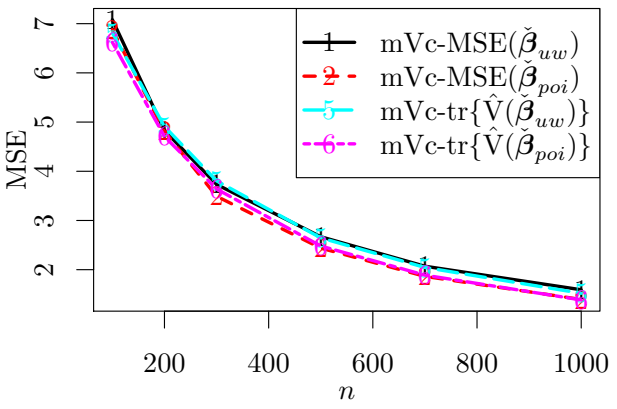
(b) nzNormal



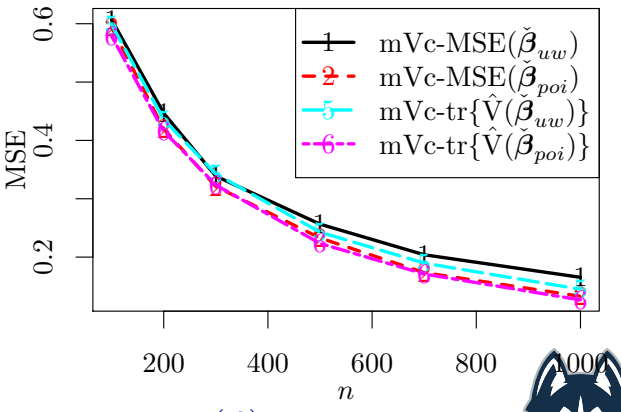
(c) ueNormal



(d) mixNormal



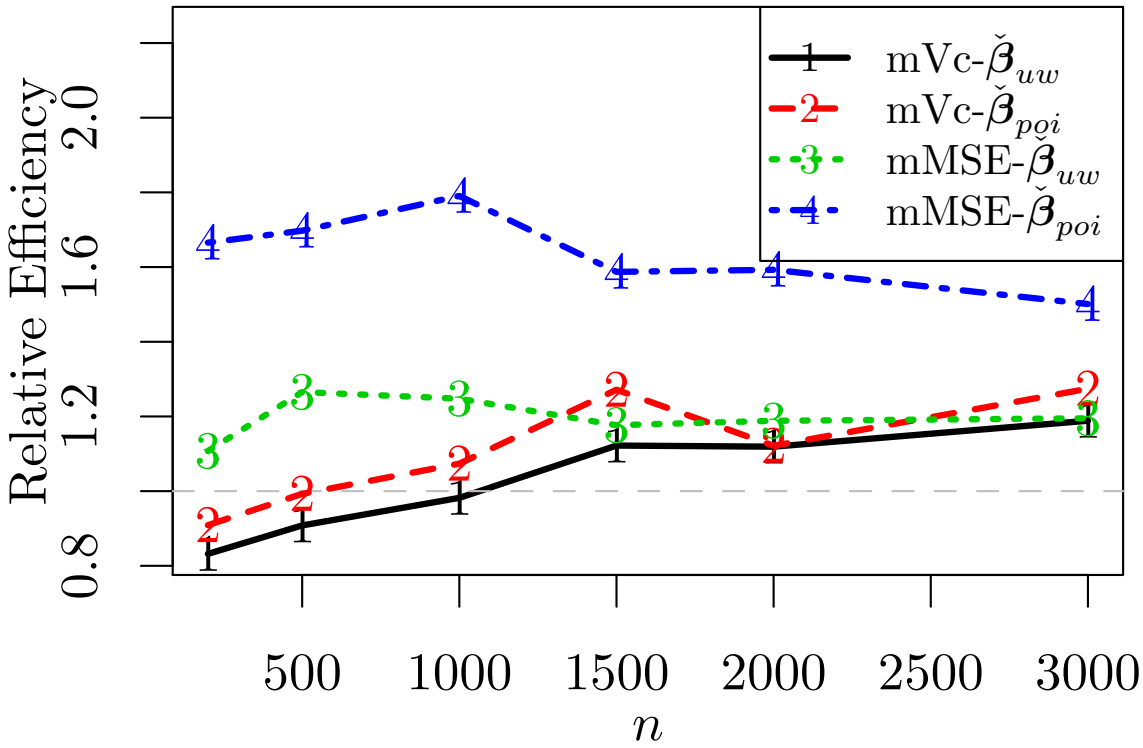
(e) T_3



(f) EXP

SUSY benchmark dataset

- Consider a SUSY benchmark data set in Baldi *et al.* (2014).
- The sample size $N = 5,000,000$ and the data file is 2.4GB.
- The goal is to distinguish between a process where new supersymmetric particles are produced and a background process.
- There 18 features that are kinematic properties in the data set.



Computing time

Table: CPU seconds when the full data are generate and kept in the RAM. Here $n_1 = 200$, $n = 1000$, and the full data size N varies; the covariates are from a $d = 50$ dimensional multivariate normal distribution.

Method	N			
	10^4	10^5	10^6	10^7
mVc, $\check{\boldsymbol{\beta}}_w$	0.14	0.13	0.45	5.24
mVc, $\check{\boldsymbol{\beta}}_{uw}$	0.08	0.11	0.41	3.71
mVc, $\check{\boldsymbol{\beta}}_{poi}$	0.08	0.11	0.43	3.88
mMSE, $\check{\boldsymbol{\beta}}_w$	0.13	0.32	3.31	35.15
mMSE, $\check{\boldsymbol{\beta}}_{uw}$	0.12	0.31	3.29	34.98
mMSE, $\check{\boldsymbol{\beta}}_{poi}$	0.12	0.31	3.29	35.06
Full	0.15	1.62	15.05	247.89

Computing time

Table: CPU seconds when the full data are scanned from hard drive. Here $n_1 = 200$, $n = 1000$, and the full data size N varies; the covariates are from a $d = 50$ dimensional multivariate normal distribution.

Method	N			
	10^4	10^5	10^6	10^7
mVc, $\check{\boldsymbol{\beta}}_w$	4.26	41.60	441.46	4374.94
mVc, $\check{\boldsymbol{\beta}}_{uw}$	4.13	41.42	413.09	4384.99
mVc, $\check{\boldsymbol{\beta}}_{poi}$	2.77	27.58	272.32	2699.13
mMSE, $\check{\boldsymbol{\beta}}_w$	4.43	41.75	434.96	4393.38
mMSE, $\check{\boldsymbol{\beta}}_{uw}$	4.10	41.83	417.55	4369.04
mMSE, $\check{\boldsymbol{\beta}}_{poi}$	2.88	27.93	273.24	2719.51
Full	139.46	1411.78	14829.63	138134.69

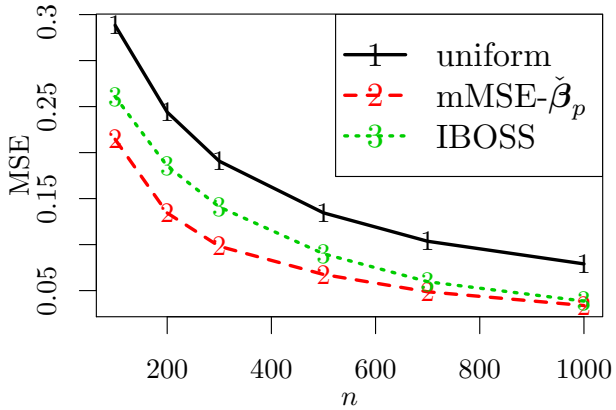
Outline

D-OPT IBOSS algorithm

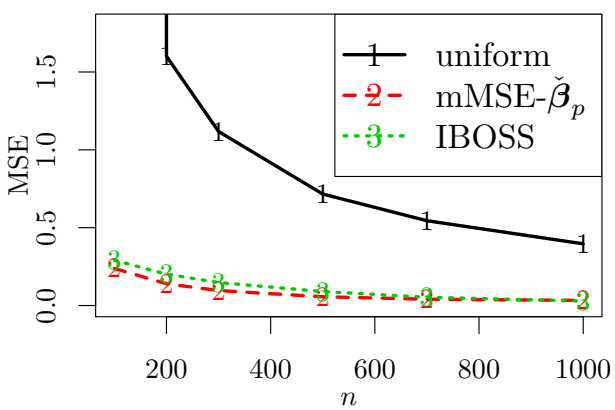
Let $\mathbf{w}_i(\boldsymbol{\beta}) = \phi_i^{1/2}(\boldsymbol{\beta})\mathbf{x}_i$, where $\phi_i(\boldsymbol{\beta}) = p_i(\boldsymbol{\beta})\{1 - p_i(\boldsymbol{\beta})\}$.

1. Use pilot estimate $\hat{\boldsymbol{\beta}}_1$ to calculate $\hat{\mathbf{w}}_i = \mathbf{w}_i(\hat{\boldsymbol{\beta}}_1)$, for $i = 1, ..., N$.
2. Let $s = [r/(2d)]$. Using a partition-based selection algorithm, perform the following steps:
   1. For $\hat{w}_{i1}$, $1 \le i \le n$, include s data points with the s smallest $\hat{w}_{i1}$ values and s data points with the s largest $\hat{w}_{i1}$ values;
   2. For $j = 2, ..., d$, exclude data points that were previously selected, and from the remainder select s data points with the smallest $\hat{w}_{ij}$ values and s data points with the largest $\hat{w}_{ij}$ values.
   3. Calculate the parameter estimator, $\hat{\boldsymbol{\beta}}_{iboss}$, and associated statistics using the selected subdata.
3. Combine $\hat{\boldsymbol{\beta}}_1$ and $\hat{\boldsymbol{\beta}}_{iboss}$.

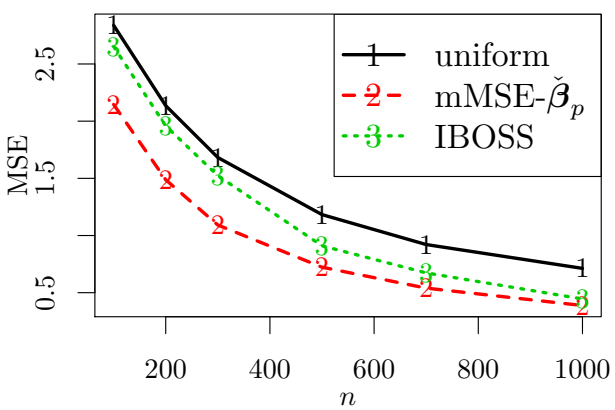
MSE: $n_1 = 200$, n varies.



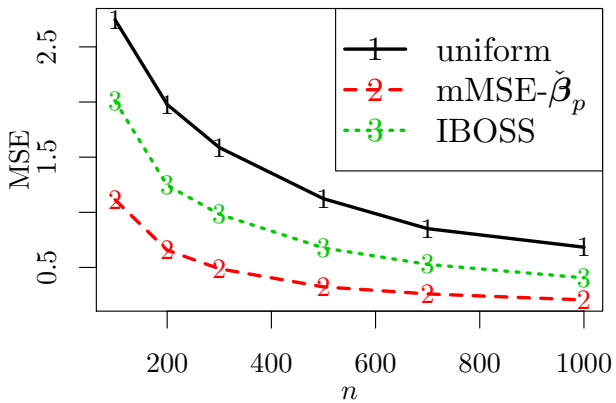
(a) mzNormal



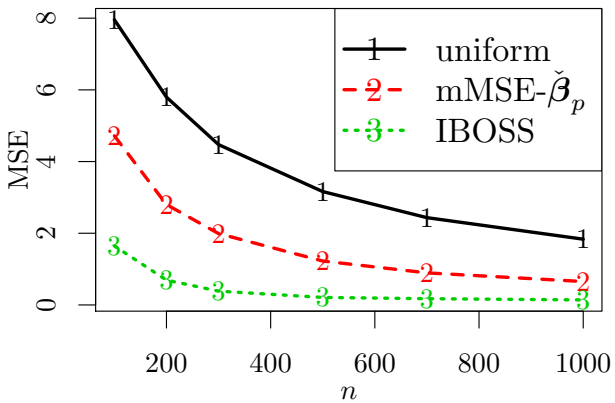
(b) nzNormal



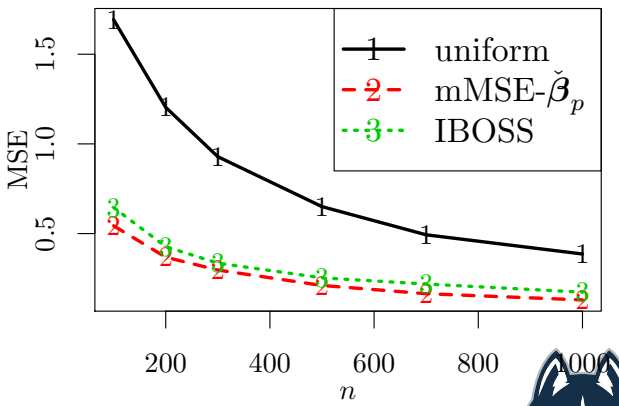
(c) ueNormal



(d) mixNormal



(e) T_3



(f) EXP

Challenge of statistical inference

- For light-tailed covariates, OSMAC can be better than IBOSS because it use information in responses.
- How to remove bias if we let IBOSS to be dependent on the responses.
- Asymptotic properties of IBOSS estimators are complicated.
- The data dependent pilot estimator $\hat{\boldsymbol{\beta}}_1$ makes it harder.

References

[1] Ping Ma, Michael Mahoney, and Bin Yu.
A statistical perspective on algorithmic leveraging.
In *Proceedings of the 31st International Conference on Machine Learning*, pages 91–99, 2014.

[2] Ping Ma, Michael Mahoney, and Bin Yu.
A statistical perspective on algorithmic leveraging.
Journal of Machine Learning Research, 16:861–911, 2015.

[3] Haiying Wang, Rong Zhu, and Ping Ma.
Optimal subsampling for large sample logistic regression.
Journal of the American Statistical Association, page http://dx.doi.org/10.1080/01621459.2017.1292914, 2017.

[4] HaiYing Wang.
More efficient estimation for logistic regression with optimal subsample.
arXiv preprint, page https://arxiv.org/abs/1802.02698, 2018.

[5] Haiying Wang, Min Yang, and John Stufken.
Information-based optimal subdata selection for big data linear regression.
Journal of the American Statistical Association, page https://doi.org/10.1080/01621459.2017.1408468, 2018.

Thank you!

April 30, 2018