Optimal Subsampling for Large Sample Logistic Regression

Wang,Zhu and Ma

June 3, 2020

Introduction

If data is too big, there are several things you can do:

- Operate on a super computer
- Split the data for distributed analysis
- Downsize the data : Subsampling

Review:Logistic Regression

Given covariates $\mathbf{x}_i \in \mathbb{R}$, logistic regression models are of the form

$$P\left(y_i = 1|\mathbf{x}_i\right) = p_i(\boldsymbol{\beta}) = \frac{\exp\left(\mathbf{x}_i^T \boldsymbol{\beta}\right)}{1 + \exp\left(\mathbf{x}_i^T \boldsymbol{\beta}\right)}, \quad i = 1, 2, \ldots, n,$$

where $y_i \in \{0, 1\}$ are the responses and β is a $d \times 1$ vector of unknown parameters.

Review:Logistic Regression

The unknown parameter β is often estimated by the maximum likelihood estimator (MLE) through maximizing the log-likelihood function with respect to β, namely,

$$\begin{aligned}\hat{\boldsymbol{\beta}}_{\mathrm{MLE}} &= \arg\max_{\beta} \ell(\boldsymbol{\beta}) \\ &= \arg\max_{\beta} \sum_{i=1}^{n} \left[y_i \log p_i(\boldsymbol{\beta}) + (1 - y_i) \log \left\{ 1 - p_i(\boldsymbol{\beta}) \right\} \right]\end{aligned}$$

Analytically, there is no general closed-form solution to the MLE $\hat{\beta}_{MLE}$ and iterative procedures are often adopted to find it numerically.

Review:Logistic Regression

A commonly used iterative procedure is Newton's method. Specifically for logistic regression, Newton's method iteratively applies the following formula until $\hat{\beta}^{(t+1)}$ converges

$$\hat{\beta}^{(t+1)} = \hat{\beta}^{(t)} + \left\{ \sum_{i=1}^{n} w_i \left(\hat{\beta}^{(t)} \right) \mathbf{x}_i \mathbf{x}_i^T \right\}^{-1} \frac{\partial \ell \left(\hat{\boldsymbol{\beta}}^{(t)} \right)}{\partial \boldsymbol{\beta}},$$

where $w_i(\boldsymbol{\beta}) = p_i(\boldsymbol{\beta}) \{1 - p_i(\boldsymbol{\beta})\}$.

General Subsampling Algorithm

- Sampling: Assign subsampling probabilities $\pi_i, i = 1, 2, \ldots, n$ for all data points. Draw a random subsample of size $r(\ll n)$ according to the probabilities $\{\pi_i\}_{i=1}^n$ from the full data. Denote the covariates, responses, and subsampling probabilities in the subsample as $\mathbf{x}_i^*, y_i^*$ and π_i^* for $i = 1, 2, \ldots, r$.
- Estimation: Maximize

$$\ell^*(\boldsymbol{\beta}) = \frac{1}{r} \sum_{i=1}^{r} \frac{1}{\pi_i^*} \left[y_i^* \log p_i^*(\boldsymbol{\beta}) + (1 - y_i^*) \log \left\{ 1 - p_i^*(\boldsymbol{\beta}) \right\} \right]$$

where $p_i^*(\beta) = \exp\left(\boldsymbol{\beta}^T \mathbf{x}_i^*\right) / \left\{1 + \exp\left(\boldsymbol{\beta}^T \mathbf{x}_i^*\right)\right\}$.

Continue: General Subsampling Algorithm

Due to the convexity of $\ell^*(\beta)$, themaximization can be implemented by Newton's method, that is, iteratively applying the following formula until convergence

$$\tilde{\boldsymbol{\beta}}^{(t+1)} = \tilde{\boldsymbol{\beta}}^{(t)} + \left\{ \sum_{i=1}^{r} \frac{w_i^* \left(\tilde{\boldsymbol{\beta}}^{(t)} \right) \mathbf{x}_i^* \left(\mathbf{x}_i^* \right)^T}{\pi_i^*} \right\}^{-1} \sum_{i=1}^{r} \frac{\left\{ y_i^* - p_i^* \left(\tilde{\boldsymbol{\beta}}^{(t)} \right) \right\} \mathbf{x}_i^*}{\pi_i^*}$$

where $w_i^*(\boldsymbol{\beta}) = p_i^*(\boldsymbol{\beta}) \{1 - p_i^*(\boldsymbol{\beta})\}$.

Assumptions

Denote the full data matrix as

$$\mathcal{F}_n = (\mathbf{X}, \mathbf{y}), \text{ where } \mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n)^T$$

Assumption 1: As $n \to \infty$, $\mathbf{M}_X = n^{-1} \sum_{i=1}^{n} w_i \left(\hat{\beta}_{\text{MLE}}\right) \mathbf{x}_i \mathbf{x}_i^T$ goes to a positive-definite matrix in probability and $n^{-1} \sum_{i=1}^{n} \|\mathbf{x}_i\|^3 = O_P(1)$.

Assumption 2: $n^{-2} \sum_{i=1}^{n} \pi_i^{-1} \|\mathbf{x}_i\|^k = O_P(1)$ for $k = 2, 4$.

Theorem 1

If assumptions 1 and 2 hold, then as $n \to \infty$ and $r \to \infty$, $\tilde{\beta}$ is consistent to $\hat{\beta}_{MLE}$ in conditional probability, given $\mathcal{F}_n$ in probability. Moreover, the rate of convergence is $r^{-1/2}$. That is, with probability approaching one, for any $\epsilon > 0$, there exists a finite Δ_ϵ and r_ϵ such that

$$P\left(\left\|\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{\mathrm{MLE}}\right\| \geq r^{-1/2}\Delta_\epsilon | \mathcal{F}_n\right) < \epsilon$$

for all $r \geq r_\epsilon$.

Assumption 3.

There exists some $\delta > 0$ such that

$$n^{-(2+\delta)} \sum_{i=1}^{n} \pi_i^{-1-\delta} \left\| \mathbf{x}_i \right\|^{2+\delta} = O_P(1).$$

Assumption 3 is used to verify the Lindeberg-Feller condition. The aforementioned three assumptions are essentially moment conditions and are very general.

Theorem 2

If Assumptions 1,2 and 3 hold, then as $n \to \infty$ and $r \to \infty$, conditional on $\mathcal{F}_n$ in probability,

$$\mathbf{V}^{-1/2}\left(\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{\mathrm{MLE}}\right) \xrightarrow{d} N(0, \mathbf{I}),$$

where

$$\mathbf{V} = \mathbf{M}_X^{-1}\mathbf{V}_c\mathbf{M}_X^{-1} = O_p\left(r^{-1}\right),$$

and

$$\mathbf{V}_c = \frac{1}{rn^2}\sum_{i=1}^{n}\frac{\left\{y_i - p_i\left(\hat{\boldsymbol{\beta}}_{\mathrm{MLE}}\right)\right\}^2 \mathbf{x}_i\mathbf{x}_i^T}{\pi_i}.$$

Optimal Subsampling Strategies

- To implement Algorithm 1, one has to specify the subsampling probability (SSP) $\boldsymbol{\pi} = \{\pi_i\}_{i=1}^n$ for the full data.
- An easy choice is to use the uniform SSP $\pi^{\mathrm{UNI}} = \{\pi_i = n^{-1}\}_{i=1}^n$.
- However, an algorithm with the uniform SSP may not be 'optimal' and a nonuniform SSP may have a better performance.
- In this section, we propose more efficient subsampling procedures by choosing nonuniform π_i's to "'minimize' the asymptotic variance-covariance matrix $\mathbf{V}$.

Minimum Asymptotic MSE of $\tilde{\beta}$.

The distribution of $\tilde{\beta} - \hat{\beta}_{MLE}$ given $\mathcal{F}$ can be approximately by that of $\mathbf{u}$, a normal variable with distribution $N(0, \mathbf{V})$.
The asymptotic MSE of $\tilde{\beta}$ is equal to the trace of $\mathbf{V}$, namely

$$\text{AMSE}(\tilde{\boldsymbol{\beta}}) = \mathrm{E}\left(\|\mathbf{u}\|^2 | \mathcal{F}_n\right) = \text{tr}(\mathbf{V})$$

Theorem 3

In algorithm 1, if the SSP is chosen such that

$$\pi_i^{\mathrm{mMSE}} = \frac{\left|y_i - p_i\left(\hat{\boldsymbol{\beta}}_{\mathrm{MLE}}\right)\right| \left\|\mathbf{M}_X^{-1}\mathbf{x}_i\right\|}{\sum_{j=1}^{n}\left|y_j - p_j\left(\hat{\boldsymbol{\beta}}_{\mathrm{MLE}}\right)\right| \left\|\mathbf{M}_X^{-1}\mathbf{x}_j\right\|}, \quad i = 1, 2, \ldots, n$$

then asymptotic MSE of $\tilde{\beta}$, $tr(\mathbf{V})$, attains its minimum.

Remarks about Theorem 3

- the optimal SSP depends on data through both the covariates and the responses directly.
- For the covariates, the optimal SSP is larger for a larger $\left\|\mathbf{M}_X^{-1}\mathbf{x}_i\right\|$, which is the square root of the ith diagonal element of the matrix $\mathbf{X}\mathbf{M}_X^{-2}\mathbf{X}^T$.
- The effect of the responses on the optimal SSP depends on discrimination difficulties through the term $\left|y_i - p_i\left(\hat{\boldsymbol{\beta}}_{\mathrm{MLE}}\right)\right|$.

The effect of the responses

Let $S_0 = \{i : y_i = 0\}$ and $S_1 = \{i : y_i = 1\}$.

- For the S_0 set, a larger $p_i(\hat{\boldsymbol{\beta}}_{\mathrm{MLE}})$ results in a larger π_i^{mMSE}.
- While for the S_1 set, the effect is negative.
- The optimal subsampling approach is more likely to select data points with smaller $p_i(\hat{\beta}_{MLE})$ when $y_i = 1$ and data points with larger $p_i(\hat{\beta}_{MLE})$ when $y_i = 0$
- Intuitively, it attempts to give preferences to data points that are more likely to be misclassified.

The effect of the responses

$$\begin{aligned}
\mathrm{tr}(\mathbf{V}) &= \frac{1}{rn^2}\,\mathrm{tr}\left(\sum_{i=1}^{n}\frac{\left\{y_i - p_i\left(\hat{\beta}_{\mathrm{MLE}}\right)\right\}^2 \mathbf{M}_X^{-1}\mathbf{x}_i\mathbf{x}_i^T\mathbf{M}_X^{-1}}{\pi_i}\right)\\
&= \frac{1}{rn^2}\sum_{i=1}^{n}\frac{\left\{y_i - p_i\left(\hat{\beta}_{\mathrm{MLE}}\right)\right\}^2 \mathrm{tr}\left(\mathbf{M}_X^{-1}\mathbf{x}_i\mathbf{x}_i^T\mathbf{M}_X^{-1}\right)}{\pi_i}\\
&= \frac{1}{rn^2}\sum_{i\in S_0}\frac{\left\{p_i\left(\hat{\beta}_{\mathrm{MLE}}\right)\right\}^2 \left\|\mathbf{M}_X^{-1}\mathbf{x}_i\right\|^2}{\pi_i}\\
&\quad + \frac{1}{rn^2}\sum_{i\in S_1}\frac{\left\{1 - p_i\left(\hat{\beta}_{\mathrm{MLE}}\right)\right\}^2 \left\|\mathbf{M}_X^{-1}\mathbf{x}_i\right\|^2}{\pi_i}
\end{aligned}$$

Minimum Asymptotic MSE of $M_X\tilde{\beta}$

For two given SSPs $\boldsymbol{\pi}^{(1)}$ and $\boldsymbol{\pi}^{(2)}$, $\mathbf{V}\left(\boldsymbol{\pi}^{(1)}\right) \leq \mathbf{V}\left(\boldsymbol{\pi}^{(2)}\right)$ if and only if $\mathbf{V}_c\left(\boldsymbol{\pi}^{(1)}\right) \leq \mathbf{V}_c\left(\boldsymbol{\pi}^{(2)}\right)$

This gives us guidance to simplify the optimality criterion.

Instead of focusing on the more complicated matrix $\mathbf{V}$, we define an alternative optimality criterion by focusing on $\mathbf{V}_c$.

Specifically, instead of minimizing $tr(\mathbf{V})$, we choose to minimize $tr(\mathbf{V}_c)$.

Theorem 4

In Algorithm 1, if the SSP is chosen such that

$$\pi_i^{\mathrm{mVc}} = \frac{\left|y_i - p_i\left(\hat{\boldsymbol{\beta}}_{\mathrm{MLE}}\right)\right| \|\mathbf{x}_i\|}{\sum_{j=1}^{n}\left|y_j - p_j\left(\hat{\boldsymbol{\beta}}_{\mathrm{MLE}}\right)\right| \|\mathbf{x}_j\|}, \quad i = 1, 2, \ldots, n$$

then $tr(\mathbf{V}_c)$ attains its minimum.

Remark of Theorem 4

- It turns out that the alternative optimality criterion indeed greatly reduces the computing time.
- $\mathrm{tr}\left(\mathbf{V}_c\right) = \mathrm{E}\left(\left\|\mathbf{M}_X\mathbf{u}\right\|^2 |\mathcal{F}_n\right)$ is the AMSE of $\mathbf{M}_X\tilde{\beta}$

Two step algorithm

- The optimal SSP defined before depend on $\hat{\beta}_{MLE}$,which is the full data MLE to be approximated, so an exact OSMAC is not applicable directly.
- We propose a two-step algorithm to approximate the OSMAC.
- In the first step, a subsample of r_0 is taken to get a pilot estimate of $\hat{\beta}_{MLE}$ which is then used to approximate the optimal SSPs for drawing the more informative second step subsample.

Two step algorithm

Algorithm 2 Two-step Algorithm

- **Step 1:** Run Algorithm˜1 with subsample size r_0 to obtain an estimate $\tilde{\boldsymbol{\beta}}_0$, using either the uniform SSP $\boldsymbol{\pi}^{\mathrm{UNI}} = \{n^{-1}\}_{i=1}^{n}$ or SSP $\{\pi_i^{\mathrm{prop}}\}_{i=1}^{n}$, where $\pi_i^{\mathrm{prop}} = (2n_0)^{-1}$ if $i \in S_0$ and $\pi_i^{\mathrm{prop}} = (2n_1)^{-1}$ if $i \in S_1$. Here, n_0 and n_1 are the numbers of elements in sets S_0 and S_1, respectively. Replace $\hat{\boldsymbol{\beta}}_{\mathrm{MLE}}$ with $\tilde{\boldsymbol{\beta}}_0$ in (10) or (13) to get an approximate optimal SSP corresponding to a chosen optimality criterion.
- **Step 2:** Subsample with replacement for a subsample of size r with the approximate optimal SSP calculated in Step 1. Combine the samples from the two steps and obtain the estimate $\breve{\boldsymbol{\beta}}$ based on the total subsample of size $r_0 + r$ according to the Estimation step in Algorithm 1.

Assumption 4

The covariate distribution satisfies that $\mathrm{E}\left(\mathbf{x}\mathbf{x}^T\right)$ positive definite and $\mathrm{E}\left(e^{\mathrm{a}^T\mathrm{x}}\right) < \infty$ for any $\mathbf{a} \in \mathbb{R}^d$.

Assumption 4 imposes two conditions on covariate distribution. The first condition ensures that the asymptotic covariance matrix is full rank. The second condition requires that covariate distributions have light tails.

Theorem 5

Let $r_0 r^{-1/2} \to 0.\ 0.$ Under Assumption 4, if the estimate $\tilde{\beta}_0$ based on the first step sample exists, then, as $r \to \infty$ and $n \to \infty$, with probability approaching one, for any $\epsilon > 0$, there exists a finite δ_ϵ and r_ϵ such that

$$P\left(\left\|\check{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{\mathrm{MLE}}\right\| \geq r^{-1/2}\Delta_\epsilon | \mathcal{F}_n\right) < \epsilon$$

for all $r \geq r_\epsilon$.

Theorem 6

Assume that $r_0 r^{-1/2} \to 0$. 0. Under Assumption 4, as $r_0 \to \infty, r \to \infty$ and $n \to \infty$, conditional on $\mathcal{F}_n$ and $\tilde{\beta}_0$,

$$\mathbf{V}^{-1/2}\left(\check{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{\mathrm{MLE}}\right) \longrightarrow N(0, \mathbf{I})$$

in distribution, in which $\mathbf{V} = \mathbf{M}_X^{-1}\mathbf{V}_c\mathbf{M}_X^{-1}$ with $\mathbf{V}_c$ having the expression of

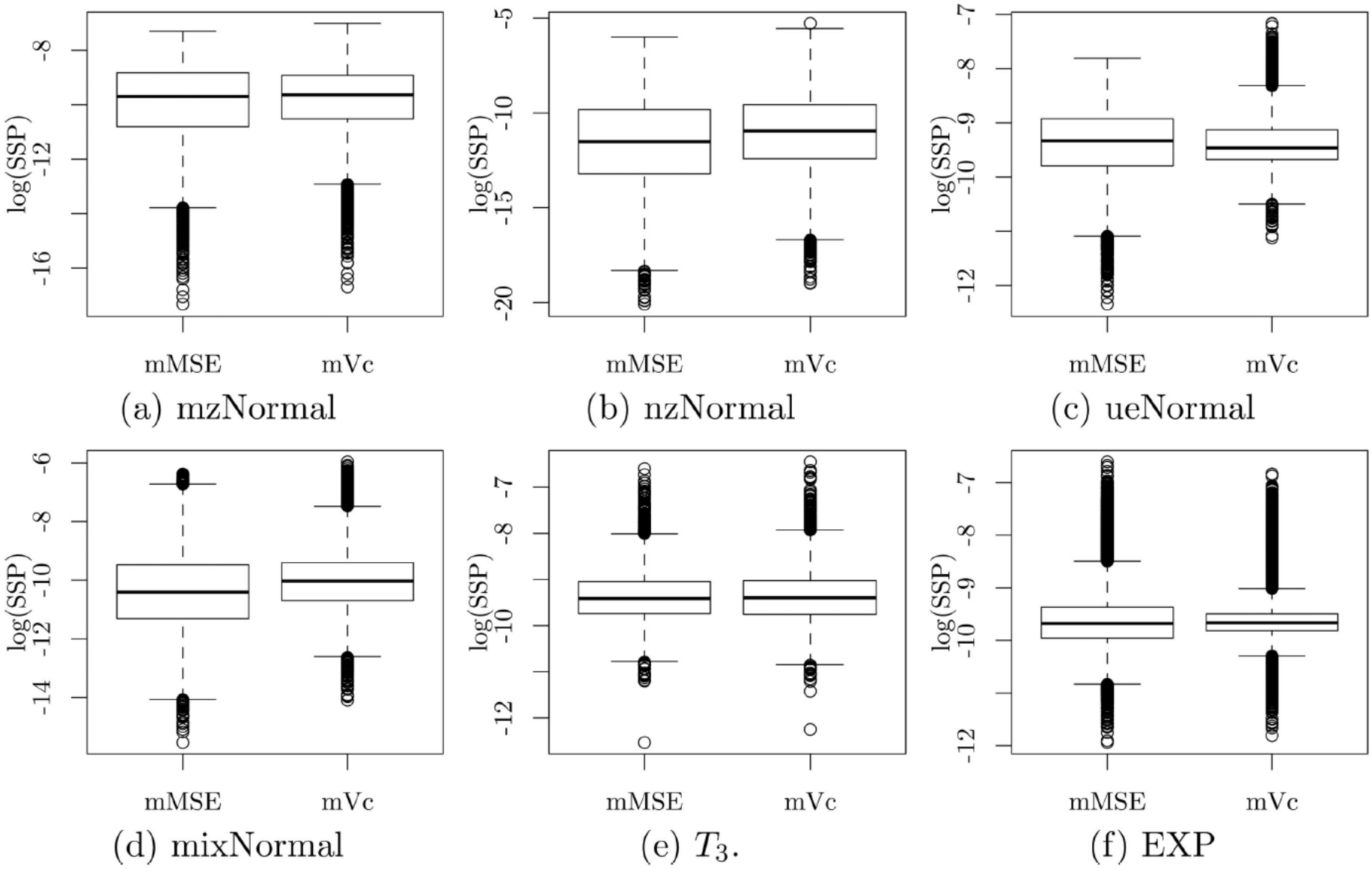
$$\mathbf{V}_c = \frac{1}{rn^2}\left\{\sum_{i=1}^{n}\left|y_i - p_i\left(\hat{\boldsymbol{\beta}}_{\mathrm{MLE}}\right)\right| \|\mathbf{x}_i\|\right\}\left\{\sum_{i=1}^{n}\frac{\left|y_i - p_i\left(\hat{\boldsymbol{\beta}}_{\mathrm{MLE}}\right)\right| \mathbf{x}_i\mathbf{x}_i^T}{\|\mathbf{x}_i\|}\right\}$$

Simulation

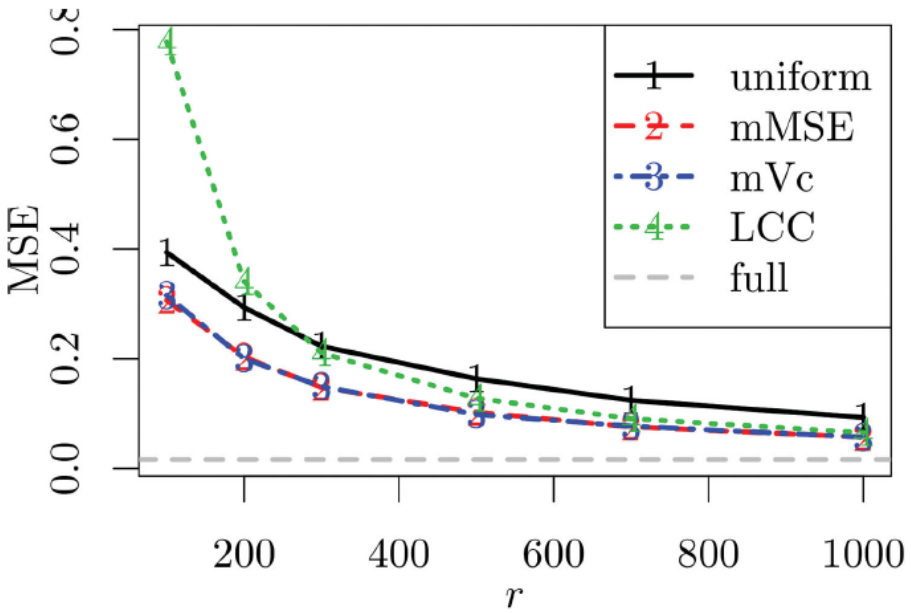
$n = 10000$, true β_0 being a 7×1 vector of 0.5. We consider the following six simulated datasets using different distributions of $\mathbf{x}$

- mznormal. $\mathbf{x}$ follows a multivariate normal distribution $N(\mathbf{0}, \mathbf{\Sigma})$, where $\mathbf{\Sigma}_{ij} = 0.5^{I(i \neq j)}$.
- nzNormal. $\mathbf{x}$ follows a multivariate normal distribution $N(1.5, \Sigma)$. About 95% of the responses are 1's, so this dataset is an example of imbalanced data.
- ueNormal. $\mathbf{x}$ follows a multivariate normal distribution with zero mean but its components have unequal variances.
- mixNormal. $\mathbf{x} \sim 0.5N(\mathbf{1}, \mathbf{\Sigma}) + 0.5N(-\mathbf{1}, \mathbf{\Sigma})$.
- T_3 with covariance Σ
- EXP independent.

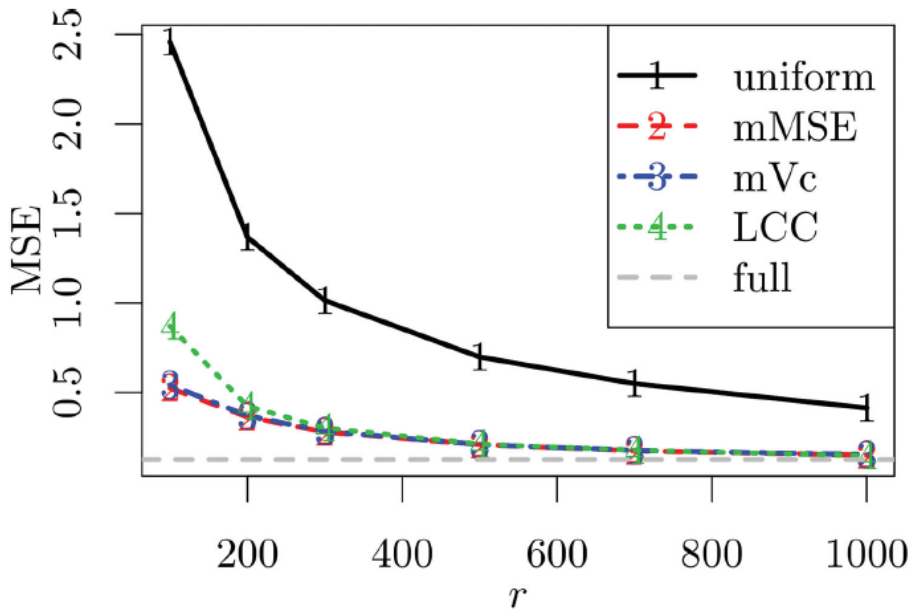
Boxplots of SSPs for different datasets



MSEs for different second step subsample size r with the first step subsample size being fixed at $r_0 = 200$.

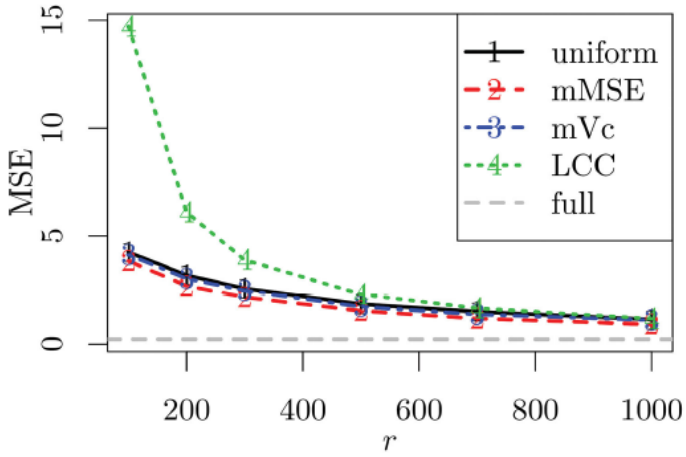


(a) mzNormal

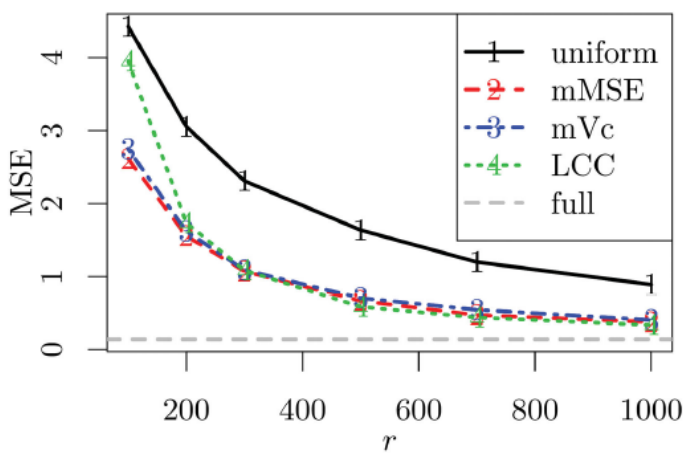


(b) nzNormal

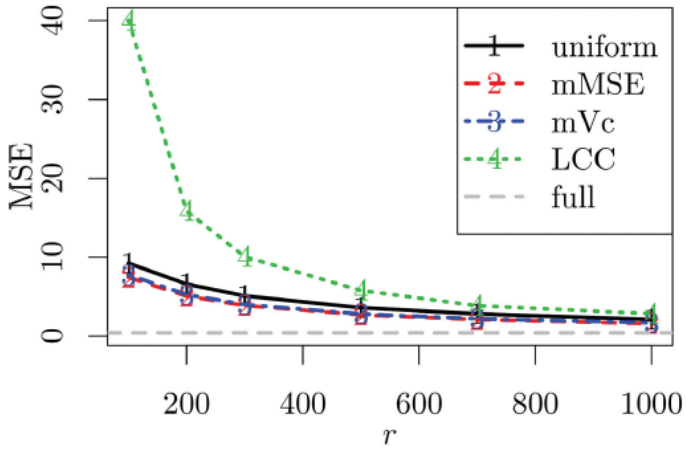
Continue



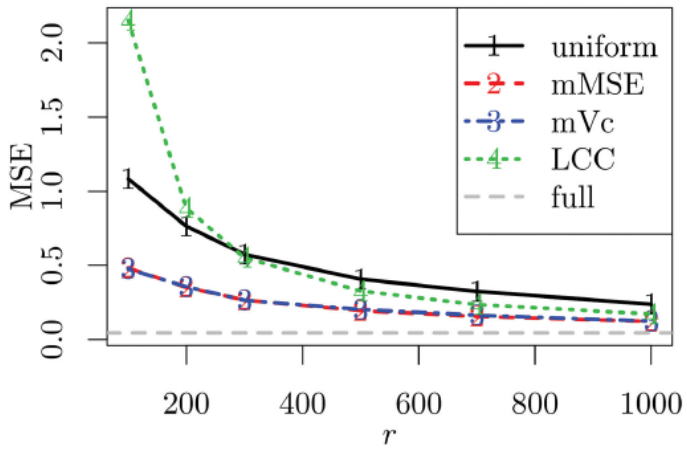
(c) ueNormal



(d) mixNormal

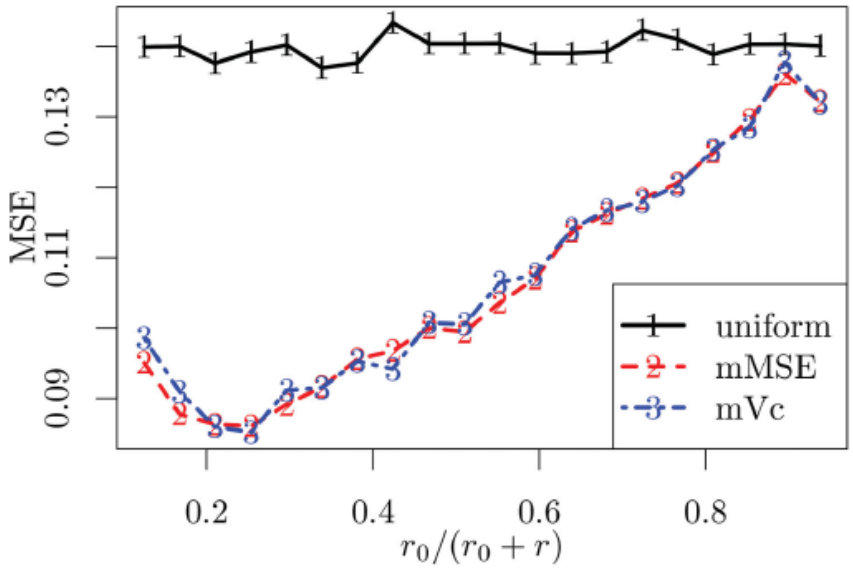


(e) T_3

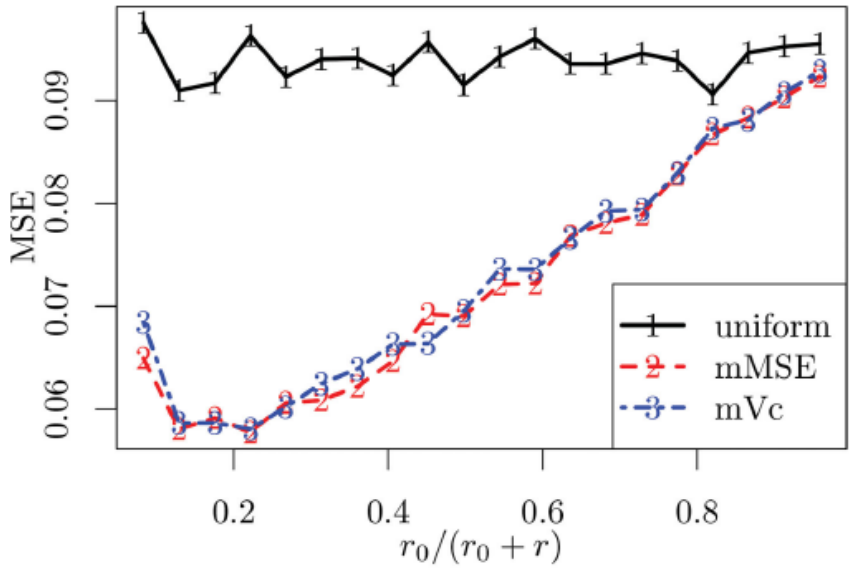


(f) EXP

Figure 3 MSEs versus proportions of the first step subsample with fixed total subsample sizes for the mzNormal dataset.

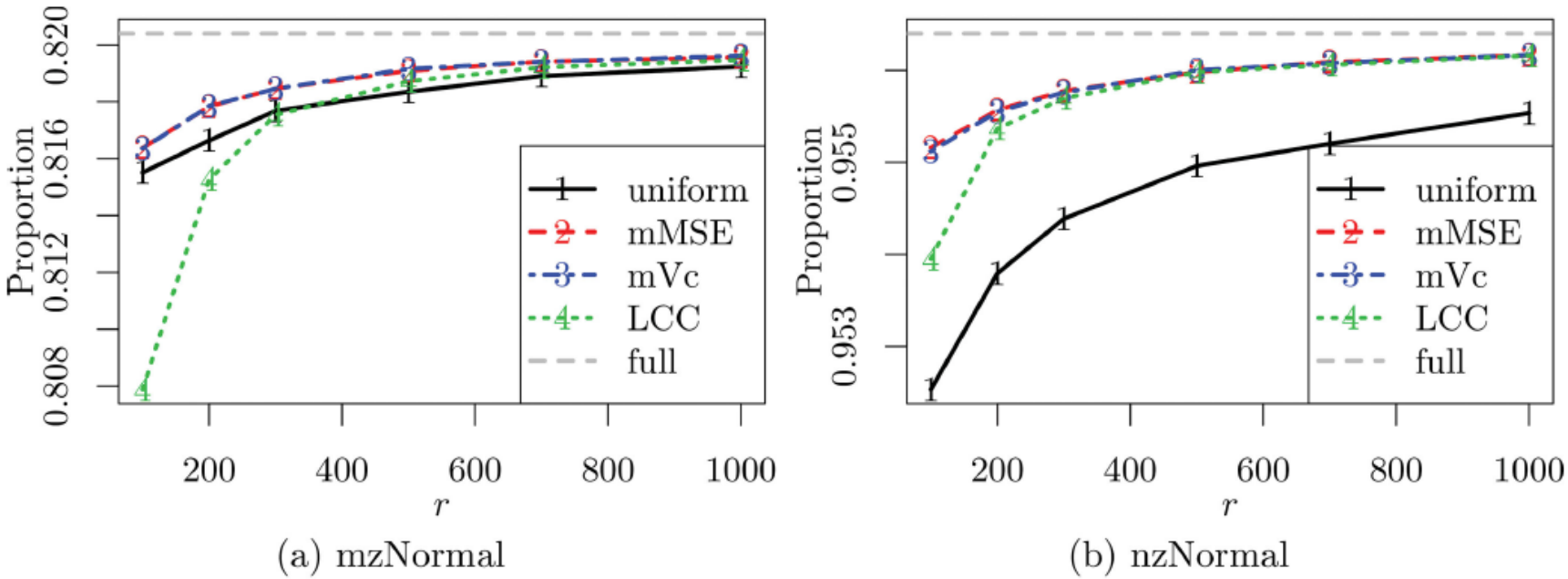


(a) $r_0 + r = 800$



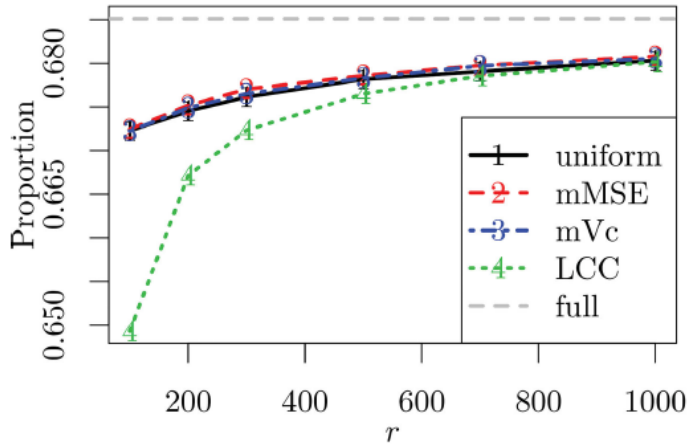
(b) $r_0 + r = 1200$

Proportions of correct classifications for different second step subsample size r with the first step subsample size being fixed at $r_0 = 200$.

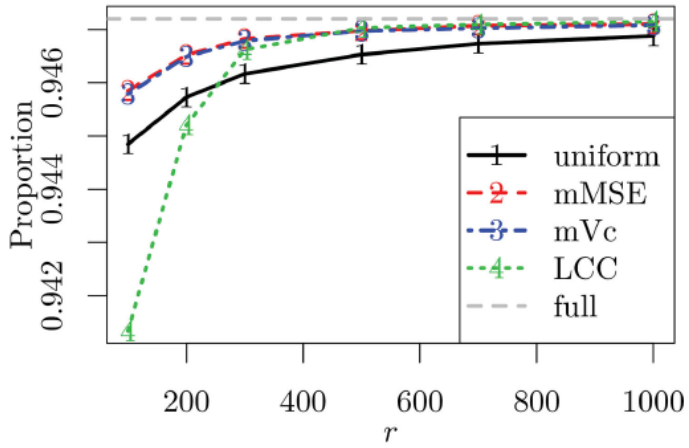


(a) mzNormal (b) nzNormal

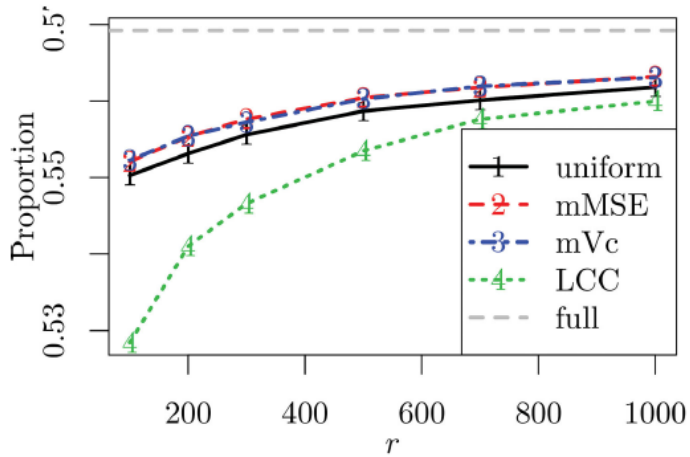
Continue



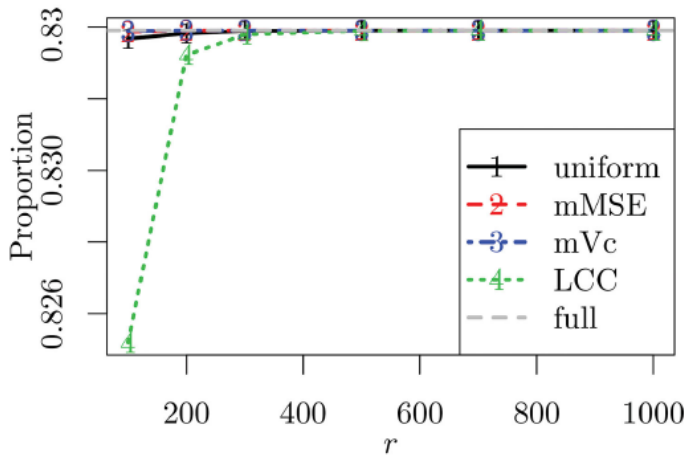
(c) ueNormal



(d) mixNormal

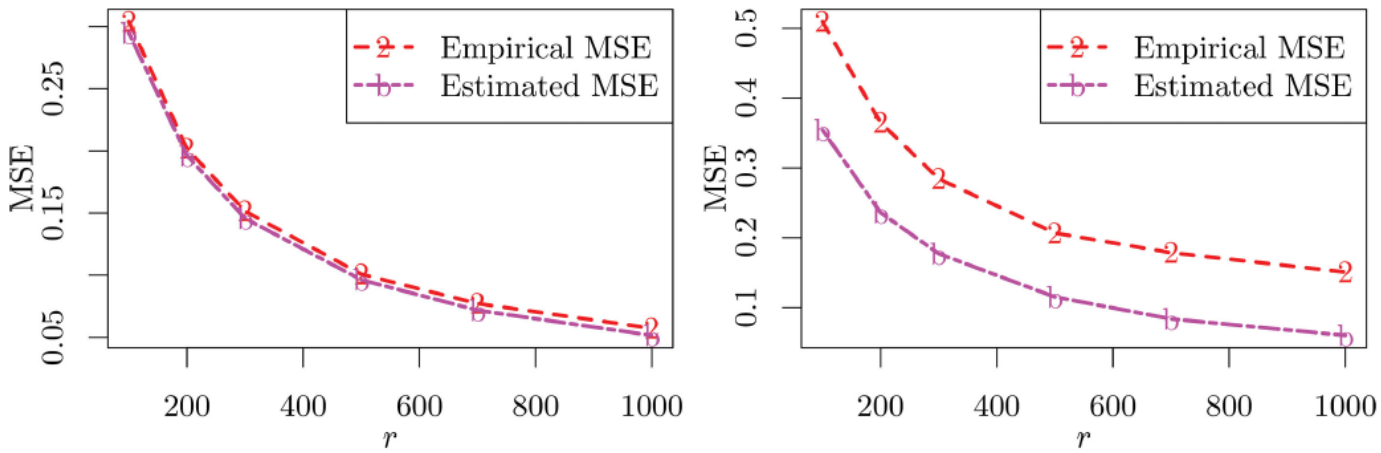


(e) T_3



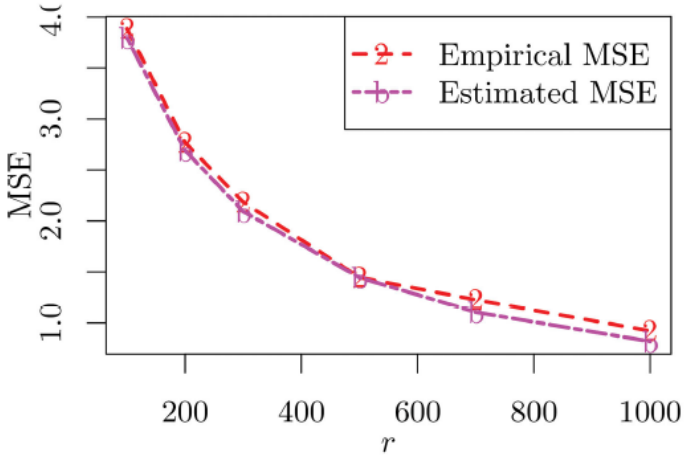
(f) EXP

Estimated and empirical MSEs for the OSMAC with π^{mMSE}, The first step subsample size is fixed at $r_0 = 200$.

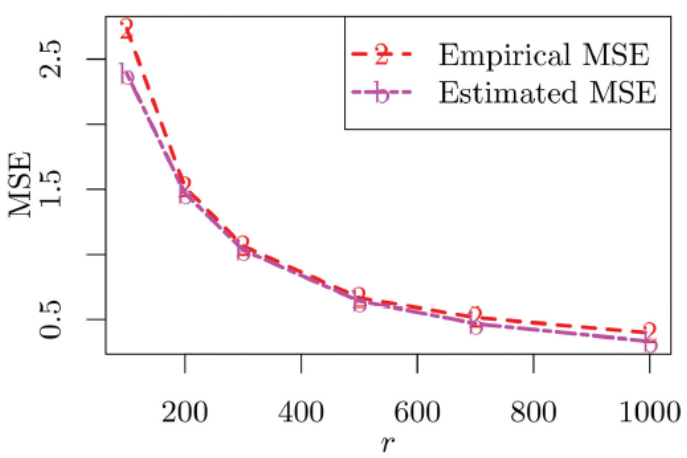


(a) mzNormal (b) nzNormal

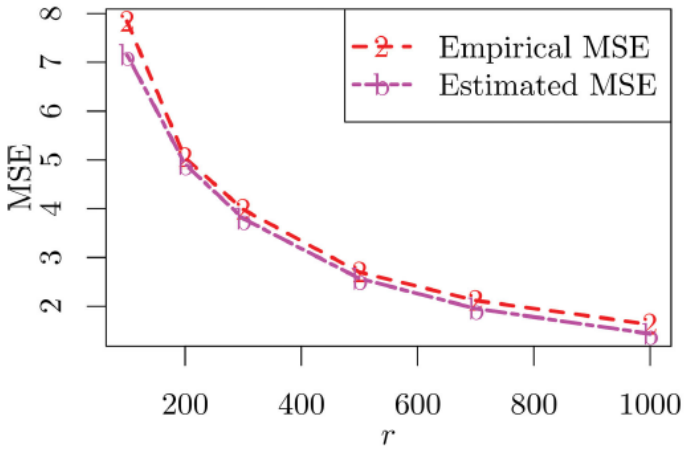
Continue



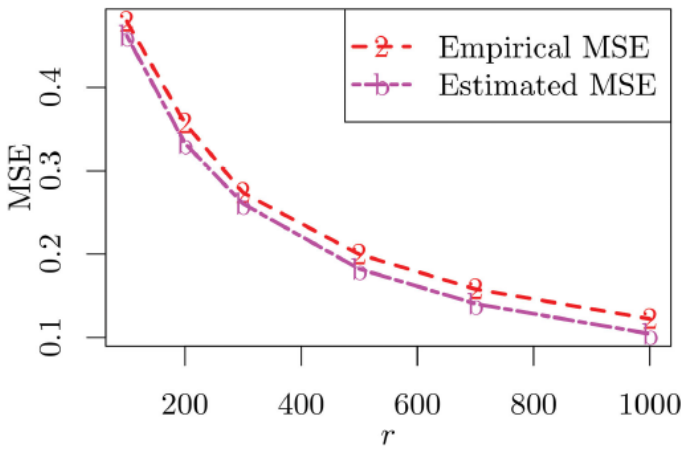
(c) ueNormal



(d) mixNormal

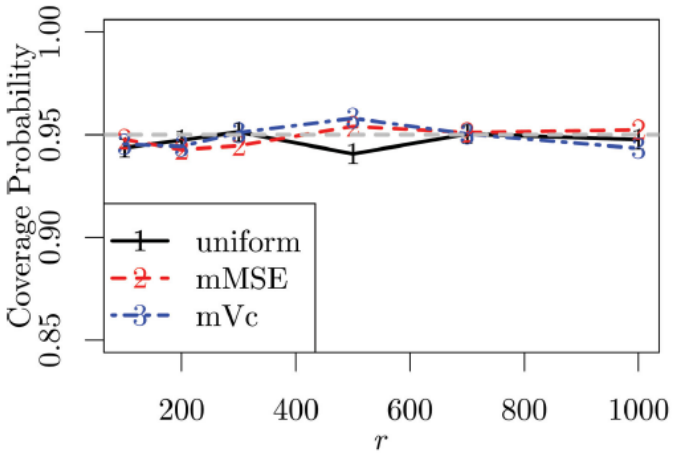


(e) T_3

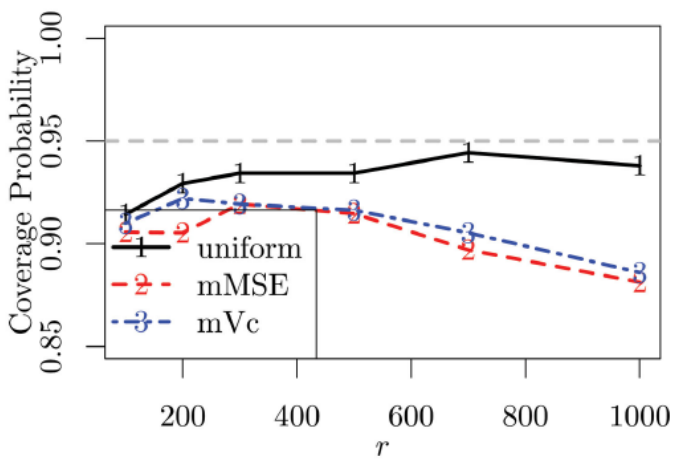


(f) EXP

Empirical coverage probabilities for different second step subsample size r with the first step subsample size being fixed

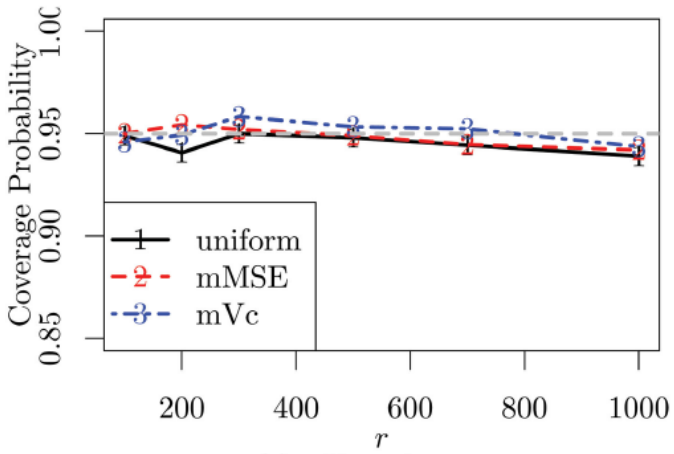


(a) mzNormal

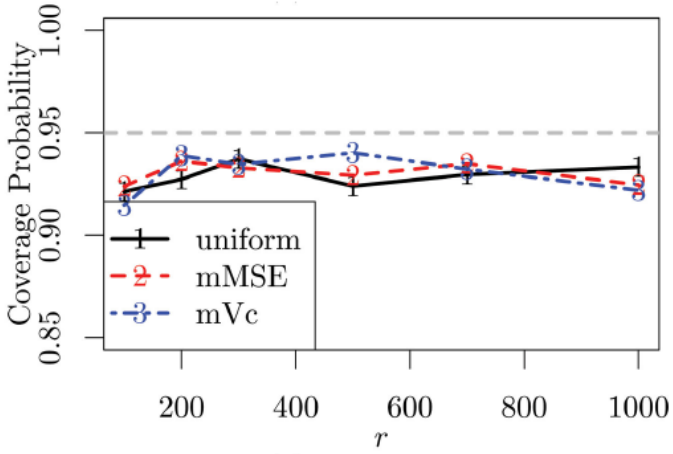


(b) nzNormal

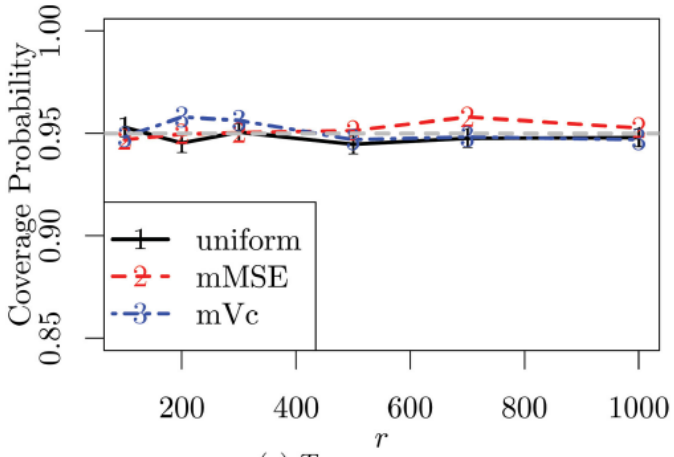
Continue



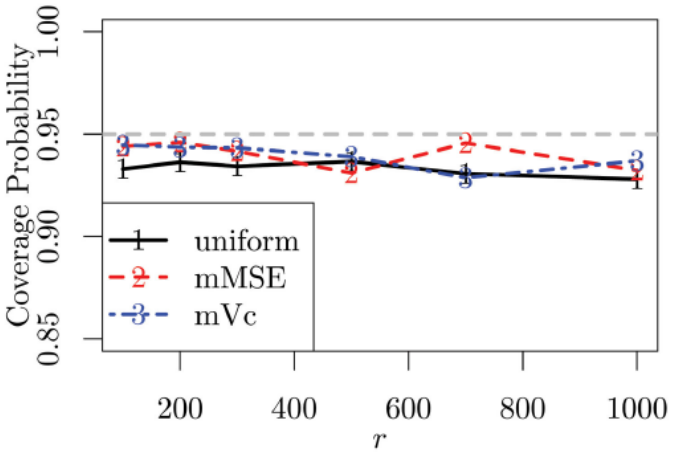
(c) ueNormal



(d) mixNormal



(e) T_3



(f) EXP

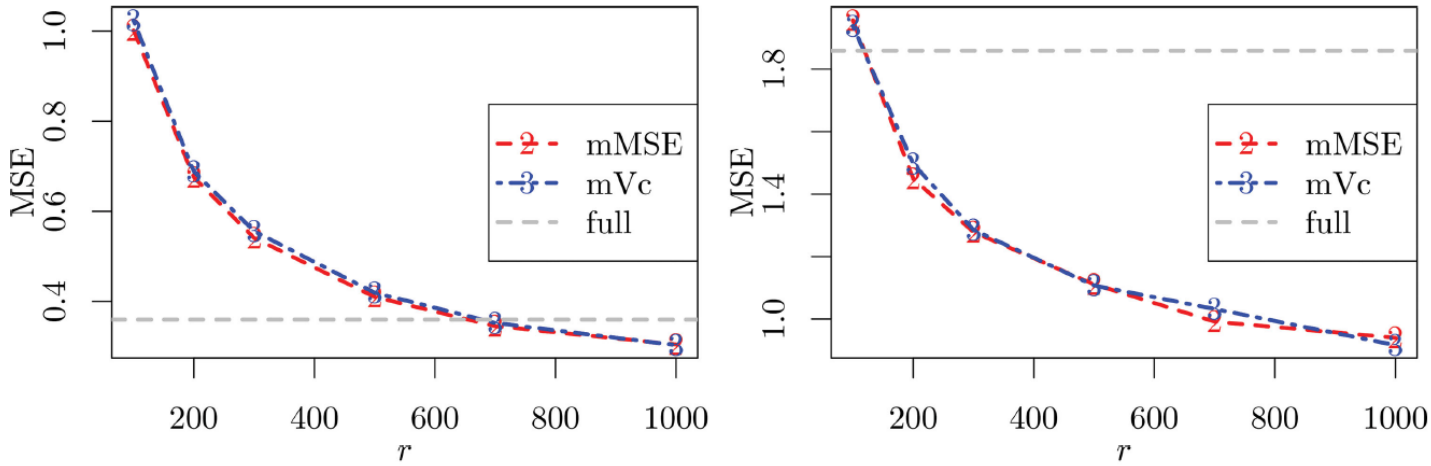
Time

Table 1. CPU seconds for the mzNormal dataset with $r_0 = 200$ and different r. The CPU seconds for using the full data is given in the last row.

	r					
Method	100	200	300	500	700	1000
mMSE	3.340	3.510	3.720	4.100	4.420	4.900
mVc	3.000	3.130	3.330	3.680	4.080	4.580
Uniform	0.690	0.810	0.940	1.190	1.470	1.860
Full data CPU seconds: 13.480						

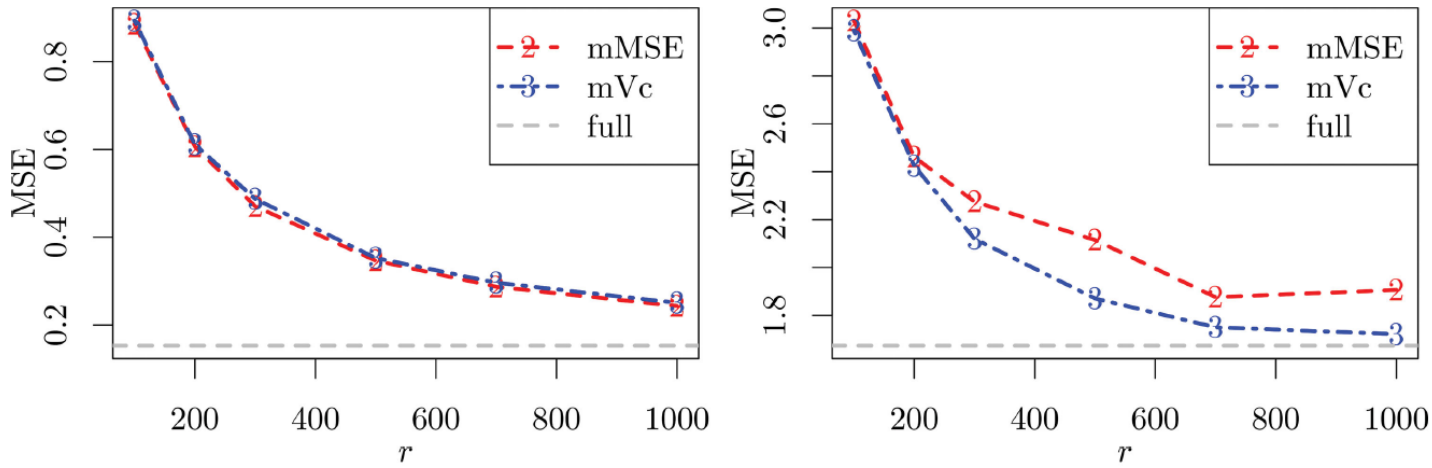
Numerical Evaluations for Rare Events data

To investigate the performance of the proposed method for the case of rare events, we generate rare events data using the same configurations that are used to generate the nzNormal data, except we change the mean of $\mathbf{x}$ to -2.14 or -2.9. With these values, 1.01% and 0.14% responses are 1 in the full data of size $n = 10000$.



(a) 1.01% of y_i's are 1 (b) 0.14% of y_i's are 1

Figure 7. MSEs for rare event data with different second step subsample size r and a fixed first step subsample size $r_0 = 200$, where the covariates follow mult normal distributions.



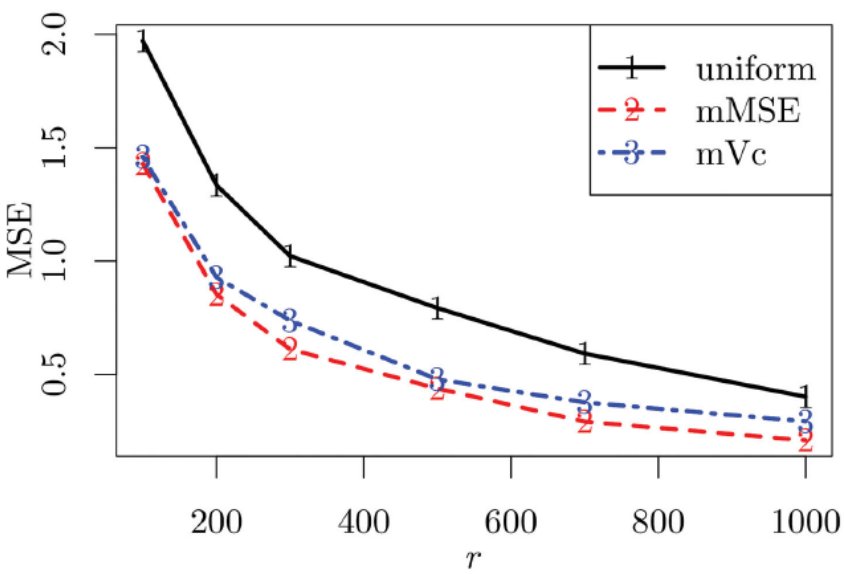
(a) 1.04% of y_i's are 1 (b) 0.11% of y_i's are 1

Figure 8. Unconditional MSEs for rare event data with different second step subsample size r and a fixed first step subsample size $r_0 = 200$, where the covariate multivariate normal distributions.

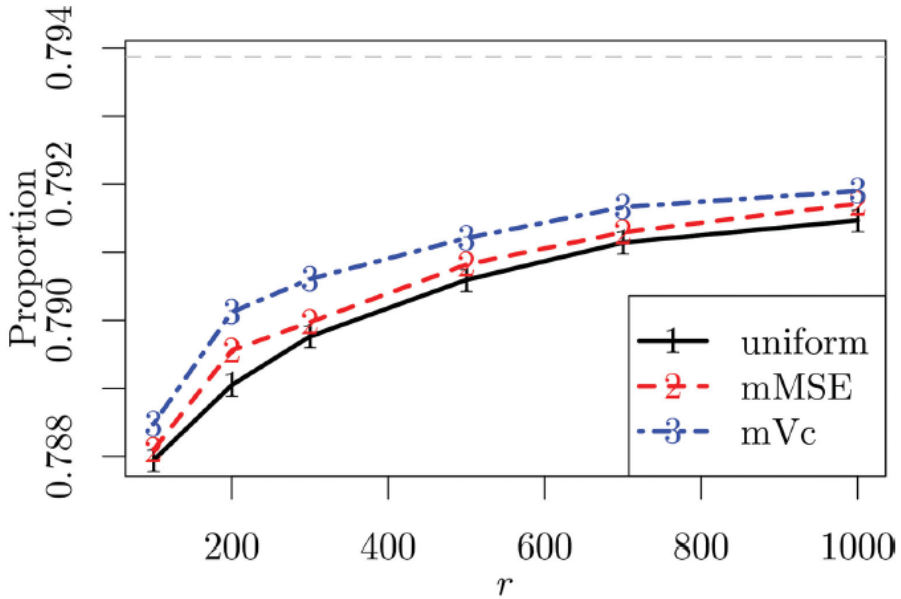
Census Income Dataset

- There are totally 48842 observations in the dataset, and the response variable is whether a person's income exceeds $ 59K a year.
- There are 11,687 individuals (23.93%) in the data whose income exceed $50K a year.
- Inferential task is to estimate the effect on income from the following covariates: x_1, age; x_2, final weight (Fnlwgt); x_3, highest level of education in numerical form; x_4, capital loss (LosCap); x_5, hours worked per week.

MSEs and proportions of correct classifications for the adult income dataset with $r_0 = 200$ and different second step subsample size r



(a) MSEs vs r

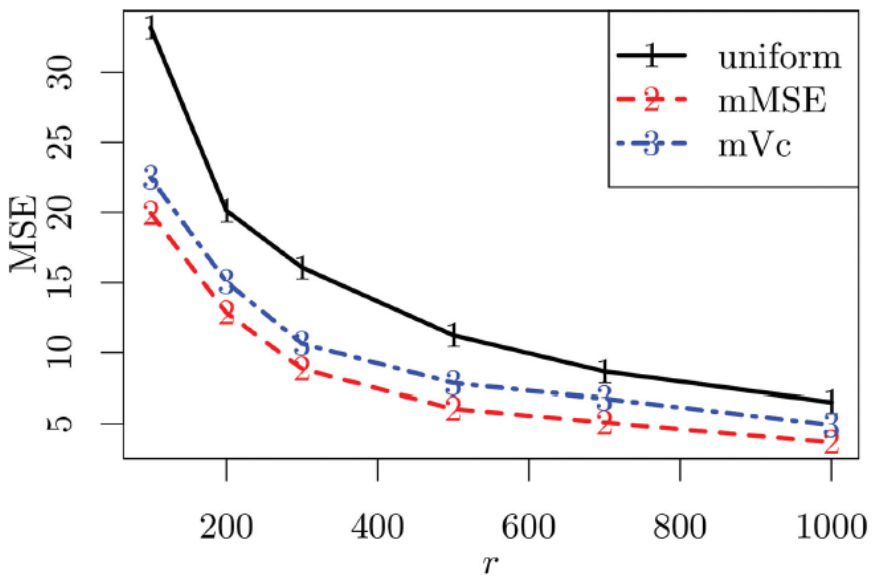


(b) Proportions of correct classifications vs r

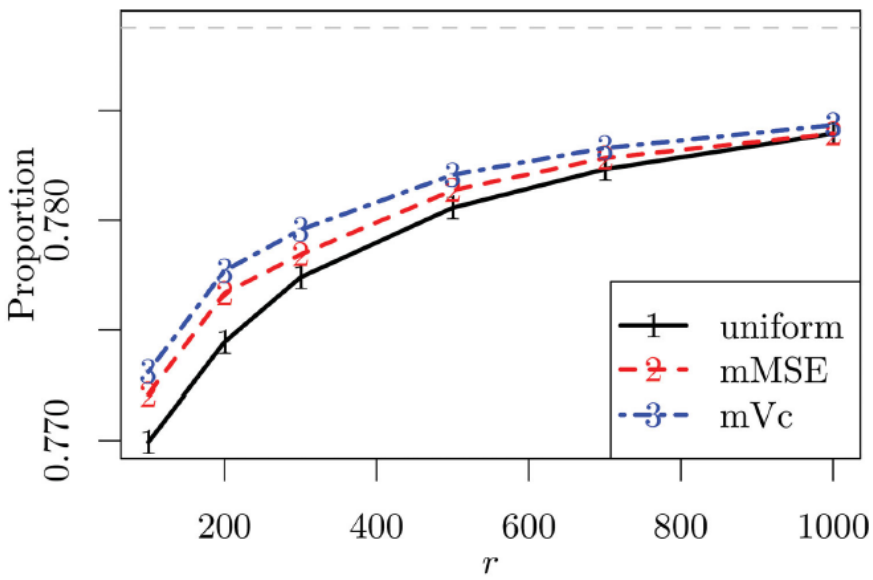
Supersymmetric Benchmark Dataset

- The full sample size is 5,000,000 and the data file is about 2.4 gigabytes.
- About 54.24% of the responses in the full data are fromthe background process.
- We use the first n = 4,500,000 observation as the training set and use the last 500,000 observations as the validation set.

MSEs and proportions of correct classifications for the SUSY dataset with $r_0 = 200$ and different second step subsample size r.



(a) MSEs vs r



(b) Proportions of correct classifications vs r

Average AUC (as percentage) for the SUSY dataset based on 1000 subsamples.

Table 5. Average AUC (as percentage) for the SUSY dataset based on 1000 subsamples.

Method	AUC % (SE)
Uniform	85.06 (0.29)
mMSE	85.08 (0.30)
mVc	85.17 (0.25)
Full	85.75

NOTE: A number in the parentheses is the associated standard error (as percentage) of the 1000 AUCs.