Inference methods of high dimensional variance estimator report

Xuelong Wang

2019-12-20

Contents

1 Motivation

2 Subsampling method: Jackknife

2.1 Jackknife Vairance

$S(X_1, \ldots, X_n)$ is a statistic of interest, define

$$S_{(i)} = S(X_1, X_{i-1}, X_{i+1} \ldots, X_n)$$

as the delete-1 result of S. If we delete each observation, then we will get n $S_{(i)}$. We could use those n subsample to estimate the variance of S on original n dataset as following,

$$\widehat{VAR}\ S(X_1, \ldots, X_n) = \frac{n-1}{n} \sum_i^n (S_{(i)} - S_{(.)})^2$$

, where $S_{(.)} = \frac{\sum_i^n S_{(i)}}{n}$. The variance estimation actually can be considered into a two-step process

1. Estimate the variance of S with n-1 sample:

$$\widehat{VAR}\ S(X_1, X_{i-1}, X_{i+1} \ldots, X_n) := \widetilde{VAR}\ S(X_1, X_{i-1}, X_{i+1} \ldots, X_n) = \sum_i^n (S_{(i)} - S_{(.)})^2,$$

which could be considered as an modification of the variance estimation corresponding to the dependency of the n delete-1 subsamples. That is originally we need a coefficient $\frac{1}{n-1}$ for sample variance if the samples are indepedent. But the delete-1 subsamples are high dependent to each other, so intuitively

the sample variance will underestimate the variance. In order to alleviate the underestimation, it seems that we multiply $n-1$.

$$n-1 \cdot \frac{1}{n-1} \cdot \sum_{i}^{n} (S_{(i)} - S_{(.)})^2 = \sum_{i}^{n} (S_{(i)} - S_{(.)})^2.$$

However, by doing this, the result become overestimated and that will be discussed in the following sections.

2. Modification the variance of $n-1$ samples to n samples by:

$$\widehat{VAR}\ S(X_1, \ldots, X_n) = \frac{n-1}{n} \widetilde{VAR}\ S(X_1, X_{i-1}, X_{i+1} \ldots, X_n).$$

2.2 Bias of Variance estimation

In the Efron 1981's paper, it shows that

$$E\left[\widetilde{VAR}\ S(X_1, X_{i-1}, X_{i+1} \ldots, X_n)\right] \geq VAR\ S(X_1, X_{i-1}, X_{i+1} \ldots, X_n).$$

He also he mentioned that the bias of the variance will be reduced by increasing of n.

$$E\hat{Var} = Var^{(n)} + \{\frac{n-1}{n} Var^{(n-1)} - Var^{(n)})\} + O(1/n^2)$$

For example, if $S_n = F_n^{-1}(1/2)$, which is the sample median estimation, then ...

2.3 Bias correction

2.3.1 Using delete-1-2 method

If we assume the S is a smooth functions of emperical CDF, especially a quadratic functions, then it can be shown the leading terms of $E(\tilde{Var}(S(X_1, \ldots, S_{n-1}))) \geq Var(S(X_1, \ldots, S_{n-1}))$ is a quadratic term in expecation. Therefore we could try to estimate the quadratic term and correct the bias for the jackknife variance estimation.

Define $Q_{ii'} \equiv nS - (n-1)(S_i + S_{i'}) + (n-2)S_{(ii')}$, then the correction will be

$$\hat{Var}^{corr}(S(X_1, \ldots, X_n)) = \hat{Var}(S(X_1, \ldots, X_n)) - \frac{1}{n(n-1)} \sum_{i<i'} (Q_{ii'} - \bar{Q})^2$$

where $\bar{Q} = \sum_{i<i'}(Q_{ii'})/(n(n-1)/2)$

2.3.2 Delete-d method

The delete-d jackknife varinace estimator is

$$\mathcal{V}_{J(d)} = \frac{n-d}{d} \cdot \frac{1}{N} \sum_{S} (\hat{\theta}_S - \hat{\theta}_{S.})$$

, where $N = \binom{n}{d}$ and S is subset of $x_1, \ldots, x_n$ with size $n-d$. Note that delete-1 jackknife will be a special case of delete-d case variance estimation:

$$\mathcal{V}_{J(1)} = \frac{n-1}{1} \cdot \frac{1}{N} \sum_{S} (\hat{\theta}_S - \hat{\theta}_{S.})$$

where $N = \binom{n}{1} = n$. But how could we explain the 2-steps estimation in Eforn's 1989 paper?

Note that S could a very large value, so in the following simulation, only $S = 1000$ is used. In Jun Shao's another paper, he proposed an approximation of the deletel-d variance estimation. That is just select m from $S = \binom{n}{d}$ sub-samples and in that paper it recommended $m = n^{1.5}$.

2.4 functional of distribution functions

2.5 variance estimation for high dimensional data

2.6 Jackknife variance estimation on high dimension signal estimation

2.6.1 Jackknife variance estimation's bias and sample size n

2.6.1.1 setup

- Independent
- Normal
- $p = \{100, 1000\}$
- $n = \{50, 75, 100, 150, 200, 500, 750, 1000, 1500, 2000\}$
- $d = 0.5 \times n$
- $n_{repeat} = n^{1.5}$ for delete d jackknife and $n_{repeat} = n$ for delete 1 jackknife
- main effect: $Var(X^T\beta) = 8$

2.6.1.2 result

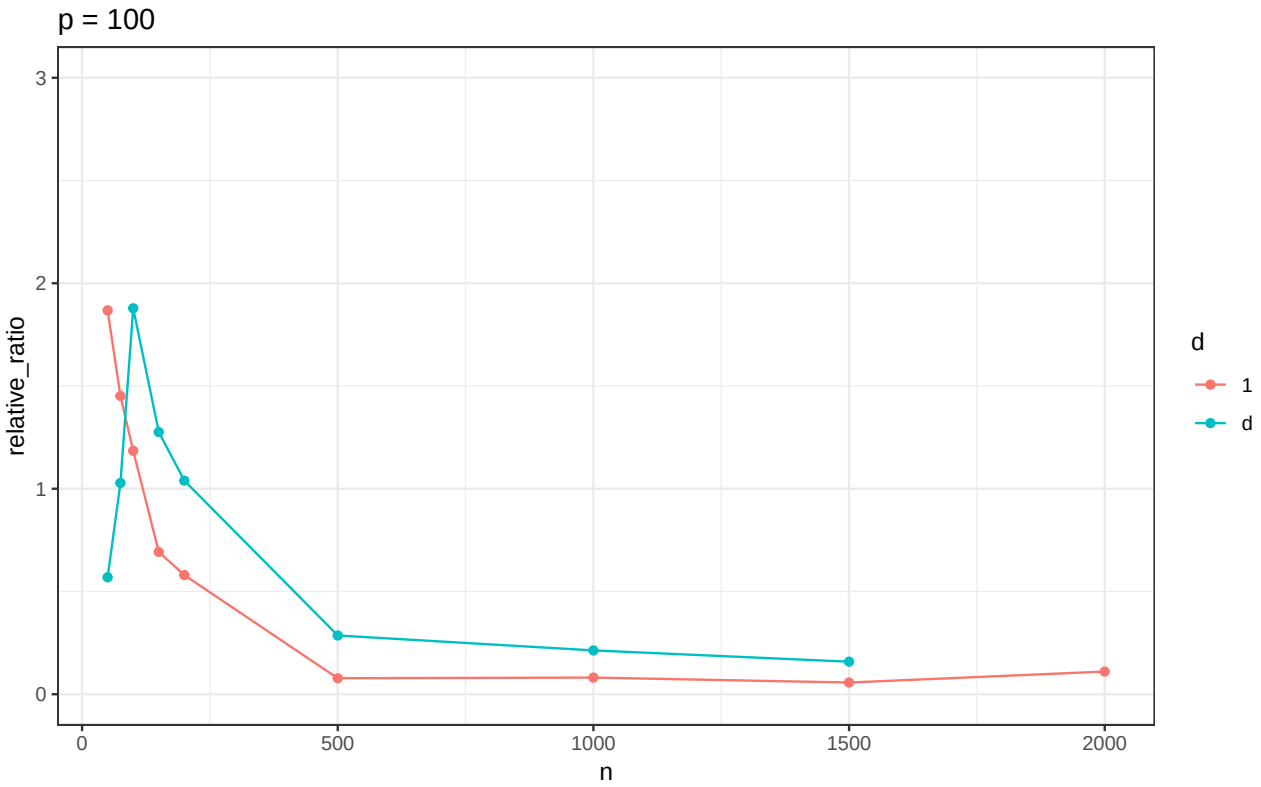
Based on the previous simulation results, we find there is a bias among all the jackknife variance estimation. Based on the Efron's result, the overestimation is because the statistics S is not a smooth function of the distribution function, so that the correct coefficient actually inflate the variance estimation.

The following reuslt is trying to see the relation between the bias and the sample size n

Note: 1. For delete-1 jackknife, the variance estimation becomes better when the sample size is increasing 1. However, for delete-d, it does not show the similar pattern, the relative ratio becomew worse when n is large, which is what we expected. One factor could the number of covariates, that is when p is large then it will be hard to make the jackknife work well??

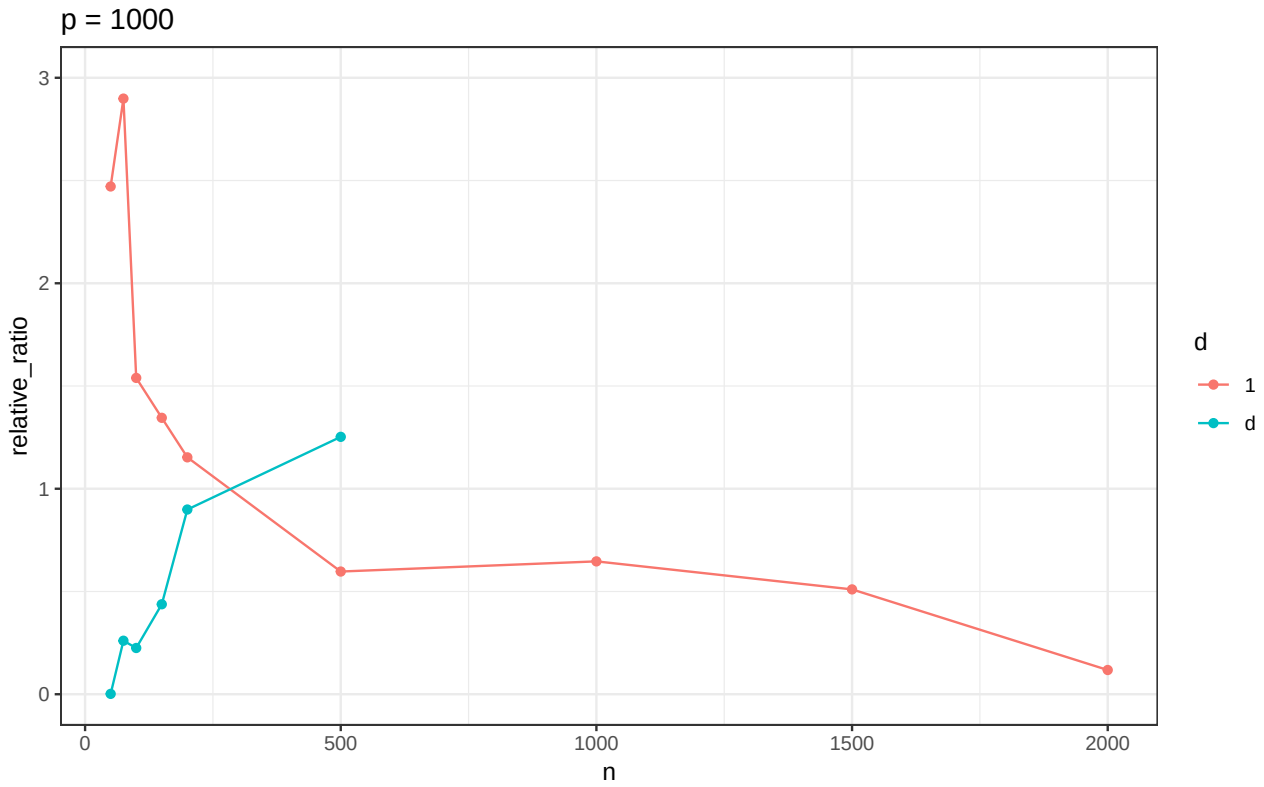
2.6.1.3 GCTA with p = 100

n	MSE	est_var	est_mean	NA_main	GCTA_rr_main_jack	GCTA_rr_v_jack	GCTA_rr_v_jack_var	relative_ratio	relative_ratio_var	N	d
50	25.58	25.84	8.0	0	8.5	74.10	8378.63	1.87	324.31	100	1
50	25.58	25.84	8.0	0	35.2	40.54	363.40	0.57	14.07	100	25
75	13.18	13.10	7.5	0	7.5	32.11	685.27	1.45	52.32	100	1
75	13.18	13.10	7.5	0	-120.5	26.57	100.83	1.03	7.70	100	38
100	6.25	6.29	7.8	0	7.5	13.74	102.20	1.18	16.26	100	1
100	6.25	6.29	7.8	0	-257.6	18.09	38.63	1.88	6.14	100	50
150	4.07	4.09	8.1	0	8.1	6.92	8.46	0.69	2.07	100	1
150	4.07	4.09	8.1	0	-17.0	9.31	7.11	1.28	1.74	100	75
200	2.48	2.49	7.9	0	7.9	3.93	1.32	0.58	0.53	100	1
200	2.48	2.49	7.9	0	76.2	5.08	1.31	1.04	0.53	100	100
500	0.83	0.83	8.1	0	8.1	0.89	0.02	0.08	0.03	100	1
500	0.83	0.83	8.1	0	4.9	1.06	0.03	0.29	0.04	100	250
1000	0.33	0.32	8.1	0	8.1	0.35	0.00	0.08	0.00	100	1
1000	0.33	0.32	8.1	0	95.6	0.39	0.00	0.21	0.01	100	500
1500	0.21	0.20	8.1	0	8.1	0.22	0.00	0.06	0.00	99	1
1500	0.21	0.20	8.1	0	85.9	0.23	0.00	0.16	0.00	99	750
2000	0.14	0.14	8.0	0	8.1	0.15	0.00	0.11	0.00	100	1



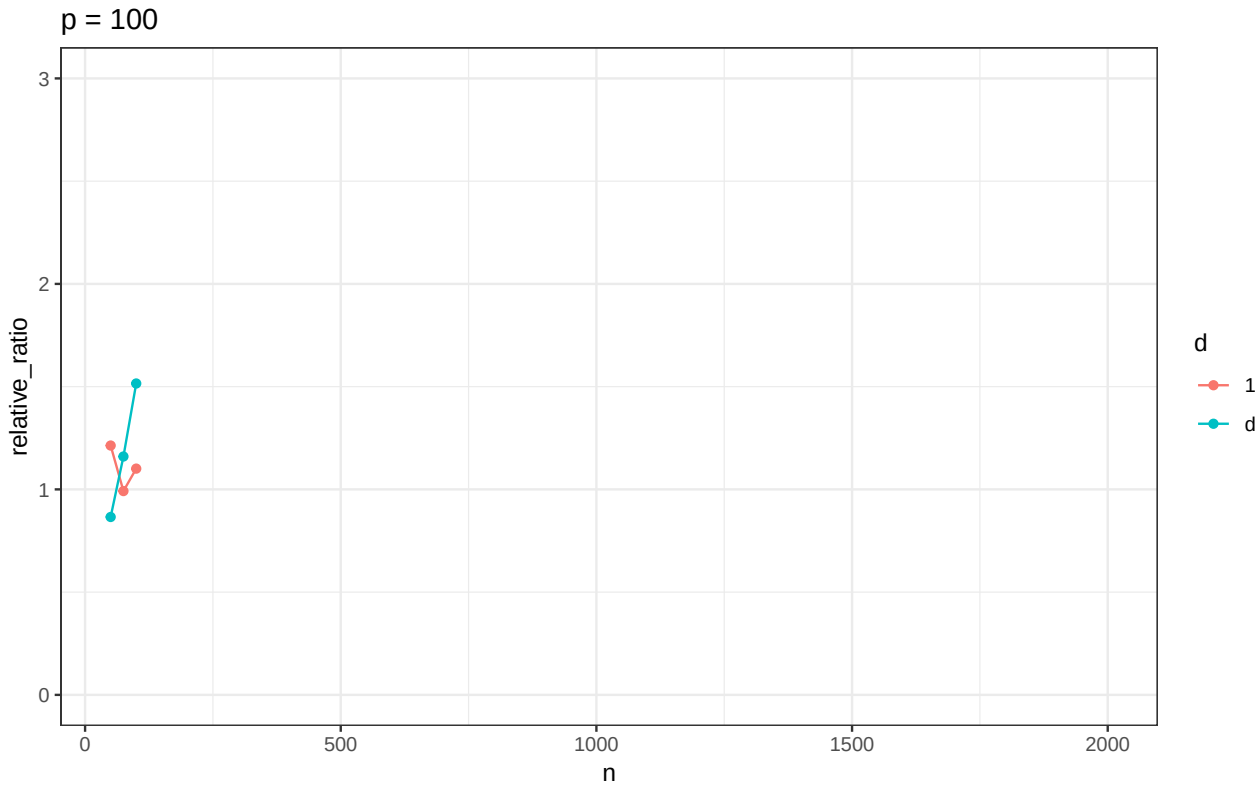
2.6.1.4 GCTA with p = 1000

n	MSE	est_var	est_mean	NA_main	GCTA_rr_main_jack	GCTA_rr_v_jack	GCTA_rr_v_jack_var	relative_ratio	relative_ratio_var	N	d
50	55.85	55.27	9.1	0	8.2	191.85	61399.17	2.47	1110.88	100	1
50	55.85	55.27	9.1	0	136.6	55.37	518.78	0.00	9.39	100	25
75	34.53	34.17	7.2	0	6.8	133.23	13844.25	2.90	405.13	100	1
75	34.53	34.17	7.2	0	-289.0	43.07	340.53	0.26	9.96	100	38
100	32.50	32.08	7.1	0	6.6	81.45	4350.37	1.54	135.63	100	1
100	32.50	32.08	7.1	0	-268.2	39.29	243.29	0.23	7.58	100	50
150	21.94	21.72	7.3	0	7.5	50.94	759.61	1.35	34.97	100	1
150	21.94	21.72	7.3	0	-267.1	31.23	67.58	0.44	3.11	100	75
200	13.23	13.25	7.7	0	7.5	28.52	121.00	1.15	9.14	100	1
200	13.23	13.25	7.7	0	-340.4	25.15	39.00	0.90	2.94	100	100
500	2.88	2.91	8.0	0	7.8	4.65	1.08	0.60	0.37	100	1
500	2.88	2.91	8.0	0	-1102.5	6.56	1.05	1.25	0.36	100	250
1000	0.77	0.78	7.9	0	8.0	1.28	0.04	0.65	0.05	99	1
1500	0.41	0.41	8.0	0	8.0	0.62	0.00	0.51	0.01	100	1
2000	0.33	0.34	8.0	0	8.1	0.38	0.00	0.12	0.00	62	1


p = 1000
relative_ratio
3
2
1
0
0
500
1000
1500
2000
n
d
1
d

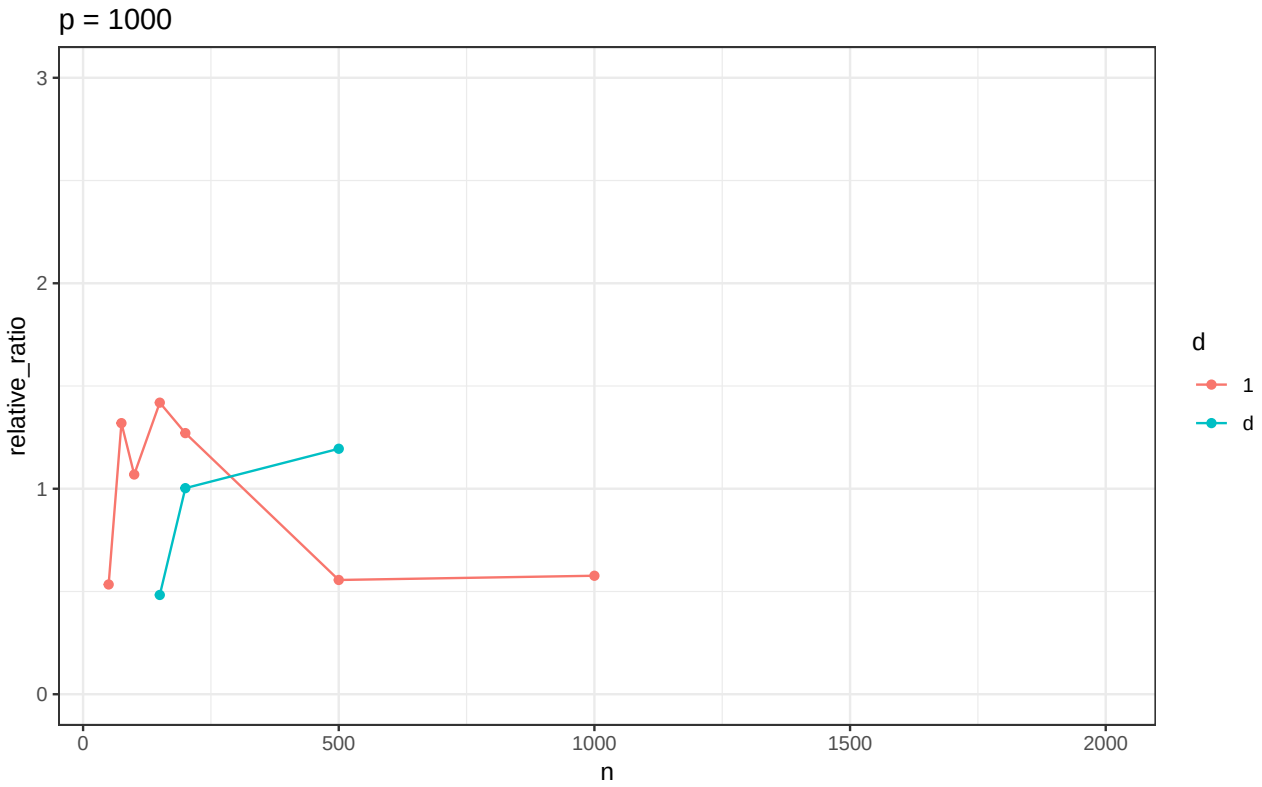
2.6.1.5 Eg with p = 100

n	MSE	est_var	est_mean	NA_main	EigenPrism_main_jack	EigenPrism_v_jack	EigenPrism_v_jack_var	relative_ratio	relative_ratio_var	N	d
50	21.6	21.7	8.4	0	7.6	48.10	454.15	1.21	20.9	100	1
50	21.6	21.7	8.4	0	-366.9	40.54	363.40	0.87	16.7	100	25
75	12.2	12.3	7.9	0	7.3	24.50	95.00	0.99	7.7	100	1
75	12.2	12.3	7.9	0	-463.3	26.57	100.83	1.16	8.2	100	38
100	7.1	7.2	8.0	0	7.1	15.11	40.84	1.10	5.7	100	1
100	7.1	7.2	8.0	0	-710.9	18.09	38.63	1.52	5.4	100	50
150	NaN	NA	NaN	100	NaN	NaN	NA	NaN	NA	100	1
150	NaN	NA	NaN	100	NaN	9.31	7.11	NA	NA	100	75
200	NaN	NA	NaN	100	NaN	NaN	NA	NaN	NA	100	1
200	NaN	NA	NaN	100	NaN	5.08	1.31	NA	NA	100	100
500	NaN	NA	NaN	100	NaN	NaN	NA	NaN	NA	100	1
500	NaN	NA	NaN	100	NaN	1.06	0.03	NA	NA	100	250
1000	NaN	NA	NaN	100	NaN	NaN	NA	NaN	NA	100	1
1000	NaN	NA	NaN	100	NaN	0.39	0.00	NA	NA	100	500
1500	NaN	NA	NaN	99	NaN	NaN	NA	NaN	NA	99	1
1500	NaN	NA	NaN	99	NaN	0.23	0.00	NA	NA	99	750
2000	NaN	NA	NaN	100	NaN	NaN	NA	NaN	NA	100	1



2.6.1.6 Eg with p = 1000

n	MSE	est_var	est_mean	NA_main	EigenPrism_main_jack	EigenPrism_v_jack	EigenPrism_v_jack_var	relative_ratio	relative_ratio_var	N	d
50	139.7	125.05	12.0	0	9.3	191.85	61399.17	0.53	490.99	100	1
50	139.7	125.05	12.0	0	-784.3	55.37	518.78	-0.56	4.15	100	25
75	57.4	57.44	8.7	0	6.9	133.23	13844.25	1.32	241.00	100	1
75	57.4	57.44	8.7	0	-1435.4	43.07	340.53	-0.25	5.93	100	38
100	39.0	39.36	7.7	0	6.1	81.45	4350.37	1.07	110.52	100	1
100	39.0	39.36	7.7	0	-1558.8	39.29	243.29	0.00	6.18	100	50
150	20.9	21.06	7.9	0	7.3	50.94	759.61	1.42	36.07	100	1
150	20.9	21.06	7.9	0	-1415.7	31.23	67.58	0.48	3.21	100	75
200	12.4	12.56	7.9	0	7.2	28.52	121.00	1.27	9.64	100	1
200	12.4	12.56	7.9	0	-2090.4	25.15	39.00	1.00	3.11	100	100
500	3.0	2.99	8.0	0	7.7	4.65	1.08	0.56	0.36	100	1
500	3.0	2.99	8.0	0	-3215.4	6.56	1.05	1.19	0.35	100	250
1000	0.8	0.81	8.0	0	8.0	1.28	0.04	0.58	0.05	99	1
1500	NaN	NA	NaN	100	NaN	0.62	0.00	NaN	NA	100	1
2000	NaN	NA	NaN	62	NaN	0.38	0.00	NaN	NA	62	1



2.6.2 After bias correction

3 Subsampling method: bootstrap

3.1 non-parameteric bootstrap

3.2 parametric bootstrap