Inference methods of high dimensional variance estimator report

Xuelong Wang

2020-02-13

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

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

The details of proof could be found by (Efron and Stein 1981), the idea of the proof is the ANOVA decomposition:

$$\begin{aligned} S\left(X_1, X_2, \cdots, X_n\right) = \mu + \sum_{i} A_i\left(X_i\right) + \sum_{i<i'} B_{i'}\left(X_i, X_{i'}\right) \\ + \sum_{i<i'<i''} C_{ii'i''}\left(X_i, X_{i'}, X_{i''}\right) + \cdots + H\left(X_1, X_2, \cdots, X_n\right) \end{aligned}$$

, where $\mu = E(S)$, $A_i\left(x_i\right) = E\left\{S \mid X_i = x_i\right\} - \mu$ and $B_{it}\left(x_i, x_i\right) = E\left(S \mid X_i = x_i, X_i = x_{it}\right\} - E\left\{S \mid X_i = x_i\right) - E\left\{S \mid X_i = x_i\right\} + \mu$. A is the analogy of main effect and B is for the two-term interaction effects. Note that after the ANOVA decompostion, all the terms has mean **0** and correlation **0**. Therefore we have

$$\begin{aligned} S\left(X_1, X_2, \cdots, X_n\right) =& \mu + \frac{1}{n} \sum_{i} \alpha_i + \frac{1}{n^2} \sum_{i<i'} \beta_{ii} \\ & + \frac{1}{n^3} \sum_{i<i<i''} \gamma_{ii'i''} + \cdots + \frac{1}{n^n} \eta_{1,2,3,\ldots,n} \end{aligned},$$

where $\begin{array}{l} \alpha_i \equiv \alpha\left(X_i\right) \equiv nA\left(X_i\right), \quad \beta_{ii} = \beta\left(X_i, X_i\right) \equiv n^2 B\left(X_i, X_{i'}\right) \\ \gamma_{ui^*} = \gamma\left(X_i, X_{i'}, X_{i^*}\right) = n^3 C\left(X_i, X_i, X_{i^*}\right), \cdots \end{array}$. Then since all of them are uncorrelated, we could take the variance on both side and have

$$\operatorname{Var} S\left(X_1, X_2, \cdots, X_n\right) = \frac{\sigma_a^2}{n} + \left(\begin{array}{c} n-1 \\ 1 \end{array}\right) \frac{\sigma_\beta^2}{2n^3} + \left(\begin{array}{c} n-1 \\ 2 \end{array}\right) \frac{\sigma_\gamma^2}{3n^5} + \cdots + \frac{\sigma_n^2}{n^{2n}}.$$

It can also shown that

$$\begin{aligned} E\left(\widetilde{\mathrm{VAR}} S\left(X_1, X_2, \cdots, X_{n-1}\right)\right) =& \frac{\sigma_\alpha^2}{n-1} \\ & + \left(\begin{array}{c} n-2 \\ 1 \end{array}\right) \frac{\sigma_\beta^2}{(n-1)^2} + \left(\begin{array}{c} n-2 \\ 2 \end{array}\right) \frac{\sigma_r^2}{(n-1)^3} + \cdots \end{aligned}$$

, so we have

$$\begin{aligned} & E\left(\widetilde{\mathrm{VAR}} S\left(X_1, X_2, \cdots, X_{n-1}\right)\right\} - \operatorname{Var} S\left(X_1, X_2, \ldots, X_{n-1}\right) \\ & = \frac{1}{2}\left(\begin{array}{c} n-2 \\ 1 \end{array}\right) \frac{\sigma_N^2}{(n-1)^3} + \frac{2}{3}\left(\begin{array}{c} n-2 \\ 2 \end{array}\right) \frac{\sigma_r^2}{(n-1)^s} + \cdots \end{aligned}$$

. Note that bias of the variance comes from the variance of high order interactions. If S is a **linear** functional the emprical cumulative density function, the the bias is 0. However, if it is not, then there will be a non-zero bias. Although Efron suggested a bias correction method, but it is not very practical which I will mention in the next section.

For certain types of S, the bias of the variance will be reduced by increasing of n.

$$E\hat{Var} = Var^{(n)} + \{\frac{n-1}{n}Var^{(n-1)} - Var^{(n)})\} + O(1/n^3),$$

2.2.1 Functionals of emprical distribution function

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

1. One potential issue of this method is that it cannot guarantee the corrected variance is positive. In other words, some times the bias correction is overestimating the bias so that ending a negative variance. This issue is not unexpected, because the correction is based only on the quadratic form.
2. Another issue is the computational time. To calculate the variance correction, one needs to do $\binom{n}{2}$ times iteration, which will be time comsuming for large n.

2.3.2 Delete-d method

The delete-d jackknife method is porposed In (Shao, Wu, and others 1989), 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 $m = n^{1.5}$.

2.3.2.1 An example of delete-d and delete-1: median

$S_n = F_n^{-1}(1/2)$ The simulation setup is following

n	MSE	est_var	est_mean	NA_main	median_main_jack	median_v_jack	relative_ratio	relative_ratio_var	d
50	0.03	0.03	-0.01	0	0.42	0.03	0.11	0.01	0.5
75	0.02	0.02	0.00	0	1.15	0.02	0.02	0.01	0.5
100	0.02	0.02	0.00	0	0.46	0.02	0.04	0.00	0.5
150	0.01	0.01	0.00	0	-0.48	0.01	0.04	0.00	0.5
200	0.01	0.01	0.00	0	2.34	0.01	0.03	0.00	0.5
50	0.03	0.03	-0.01	0	-0.01	0.07	1.17	1.05	1.0
75	0.02	0.02	0.00	0	0.05	0.03	0.41	0.09	1.0
100	0.02	0.02	0.00	0	0.00	0.03	0.79	0.20	1.0
150	0.01	0.01	0.00	0	0.00	0.02	0.76	0.13	1.0
200	0.01	0.01	0.00	0	0.00	0.02	1.01	0.18	1.0

2.4 functional of distribution functions

2.5 Jackknife variance estimation on high dimension signal estimation

Different methods have their own

2.5.1 Jackknife variance estimation's bias and sample size n

2.5.1.1 setup

- Independent
- Normal
- $p = \{100, 1000\}$
- $n = \{50, 75, 100, 150, 200, 500, 750, 1000, 1500\}$
- $d = \{0.5, 0.75\} \times n$ or $d = 25$
- $n_{repeat} = n^{1.5}$ for delete d jackknife and $n_{repeat} = n$ for delete 1 jackknife
- main effect: $Var(X^T\beta) = 8$

2.5.1.2 result

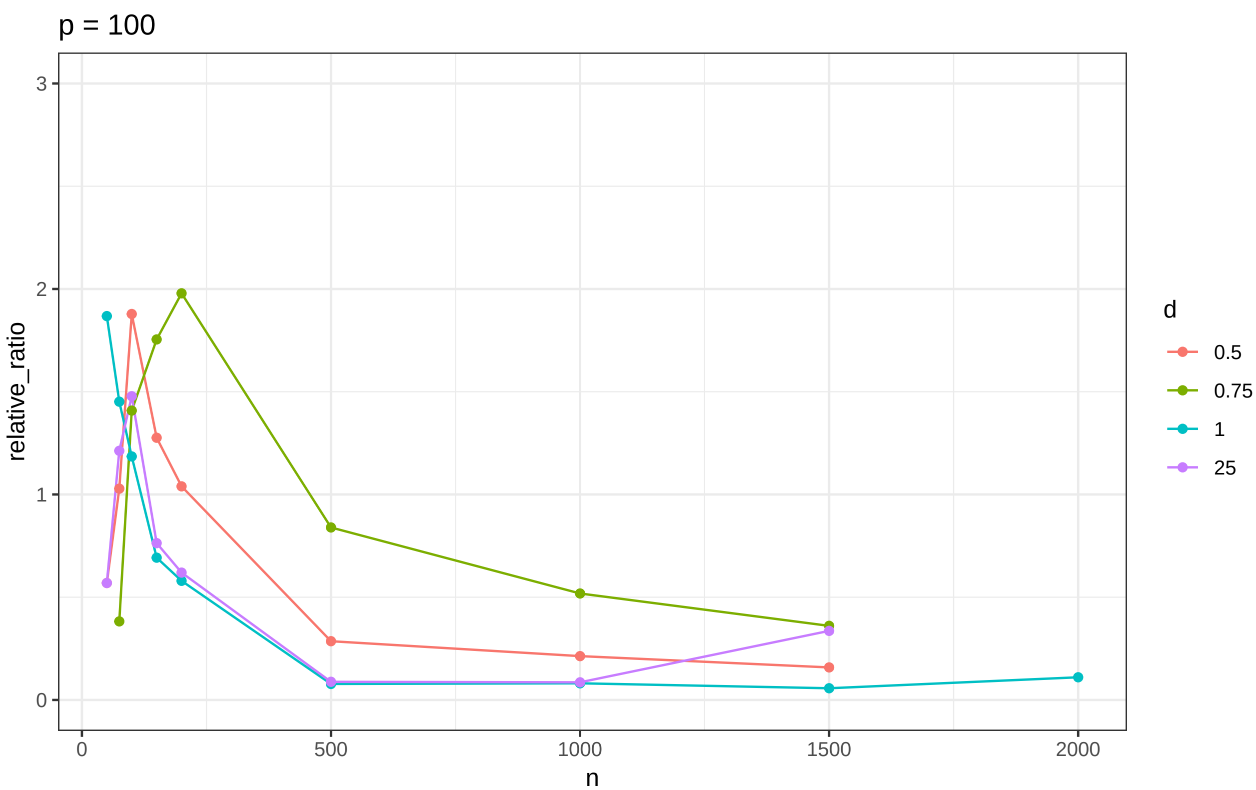
Based on the previous simulation results, we find there is a bias among all the jackknife variance estimation. Based on the Efron's result, the overestimation is because the statistics S is not a smooth function of the distribution function, so that the correct coefficient actually inflate the variance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

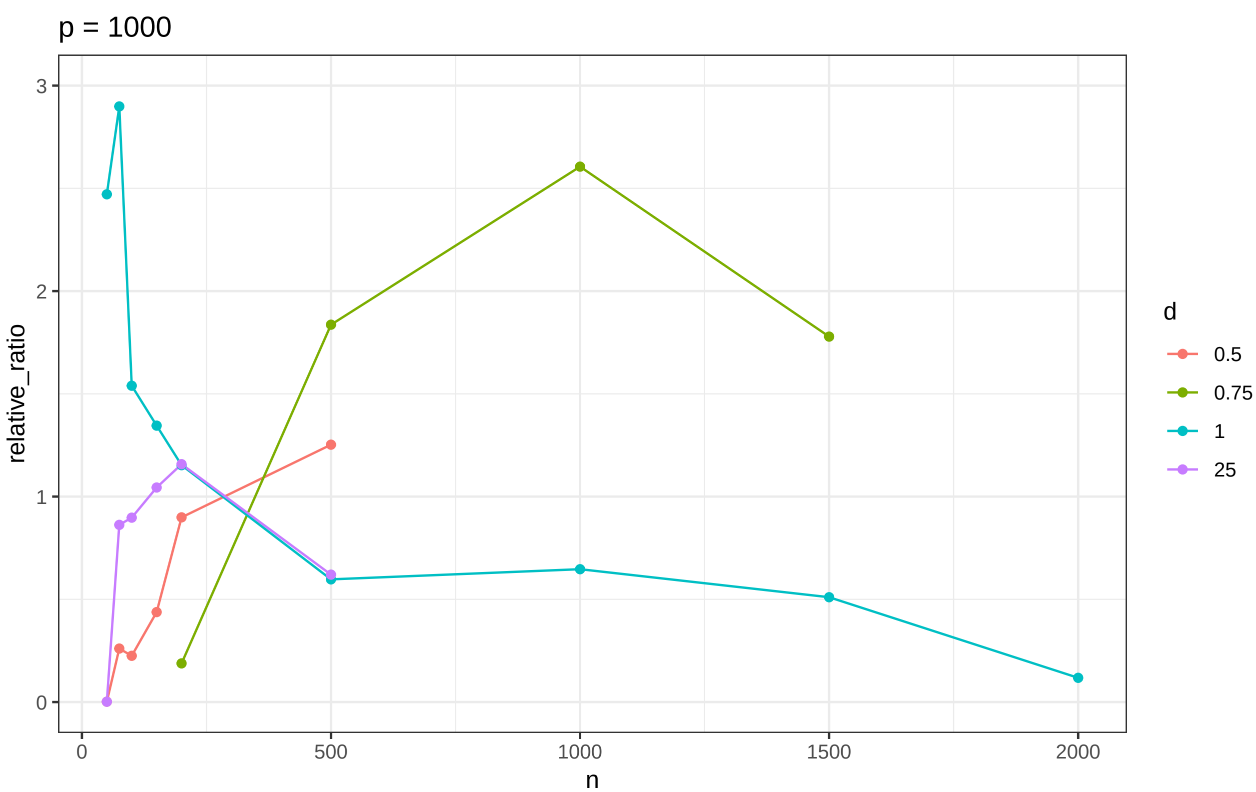
2.5.1.3 GCTA with p = 100

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	35.2	40.54	363.40	0.57	14.07	100	0.50
75	13.18	13.10	7.5	0	-120.5	26.57	100.83	1.03	7.70	100	0.50
100	6.25	6.29	7.8	0	-257.6	18.09	38.63	1.88	6.14	100	0.50
150	4.07	4.09	8.1	0	-17.0	9.31	7.11	1.28	1.74	100	0.50
200	2.48	2.49	7.9	0	76.2	5.08	1.31	1.04	0.53	100	0.50
500	0.83	0.83	8.1	0	4.9	1.06	0.03	0.29	0.04	100	0.50
1000	0.33	0.32	8.1	0	95.6	0.39	0.00	0.21	0.01	100	0.50
1500	0.21	0.20	8.1	0	85.9	0.23	0.00	0.16	0.00	99	0.50
50	25.58	25.84	8.0	0	53.8	21.44	75.63	-0.17	2.93	100	0.75
75	13.18	13.10	7.5	0	-154.6	18.11	39.61	0.38	3.02	100	0.75
100	6.25	6.29	7.8	0	-403.7	15.14	20.26	1.41	3.22	100	0.75
150	4.07	4.09	8.1	0	-342.2	11.27	8.25	1.75	2.02	100	0.75
200	2.48	2.49	7.9	0	-243.4	7.41	2.18	1.98	0.88	100	0.75
500	0.83	0.83	8.1	0	-98.3	1.52	0.05	0.84	0.06	100	0.75
1000	0.33	0.32	8.1	0	246.4	0.49	0.00	0.52	0.01	100	0.75
1500	0.21	0.20	8.1	0	128.3	0.28	0.00	0.36	0.00	100	0.75
50	25.58	25.84	8.0	0	8.5	74.10	8378.63	1.87	324.31	100	1.00
75	13.18	13.10	7.5	0	7.5	32.11	685.27	1.45	52.32	100	1.00
100	6.25	6.29	7.8	0	7.5	13.74	102.20	1.18	16.26	100	1.00
150	4.07	4.09	8.1	0	8.1	6.92	8.46	0.69	2.07	100	1.00
200	2.48	2.49	7.9	0	7.9	3.93	1.32	0.58	0.53	100	1.00
500	0.83	0.83	8.1	0	8.1	0.89	0.02	0.08	0.03	100	1.00
1000	0.33	0.32	8.1	0	8.1	0.35	0.00	0.08	0.00	100	1.00
1500	0.21	0.20	8.1	0	8.1	0.22	0.00	0.06	0.00	99	1.00
2000	0.14	0.14	8.0	0	8.1	0.15	0.00	0.11	0.00	100	1.00
50	25.58	25.84	8.0	0	35.2	40.54	363.40	0.57	14.07	100	25.00
75	13.18	13.10	7.5	0	-60.2	28.98	185.76	1.21	14.18	100	25.00
100	6.25	6.29	7.8	0	-98.7	15.58	52.81	1.48	8.40	100	25.00
150	4.07	4.09	8.1	0	3.1	7.21	7.25	0.76	1.77	100	25.00
200	2.48	2.49	7.9	0	26.6	4.03	1.29	0.62	0.52	100	25.00
500	0.83	0.83	8.1	0	6.6	0.90	0.03	0.09	0.03	100	25.00
1000	0.33	0.32	8.1	0	8.7	0.35	0.00	0.09	0.00	100	25.00
1500	0.15	0.16	8.0	0	11.4	0.21	0.00	0.34	0.00	20	25.00



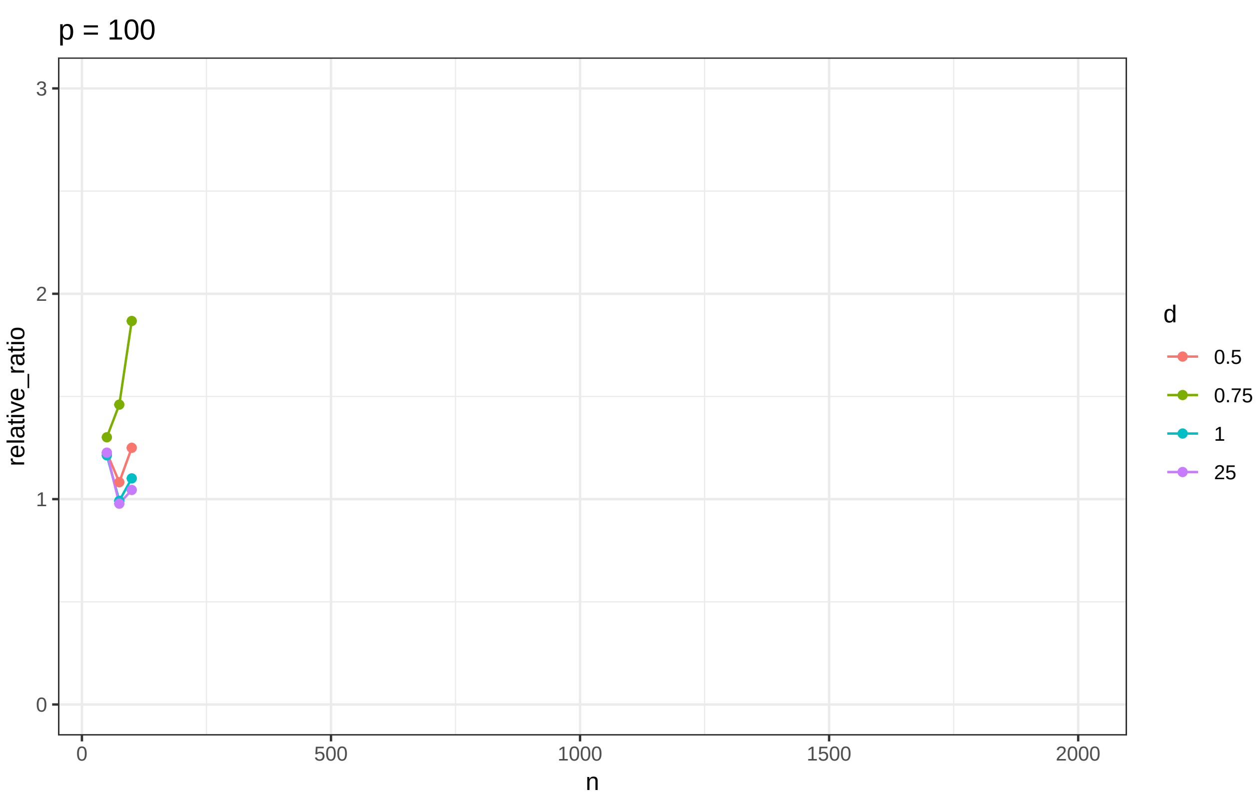
2.5.1.4 GCTA with p = 1000

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	136.6	55.37	518.78	0.00	9.39	100	0.50
75	34.53	34.17	7.2	0	-289.0	43.07	340.53	0.26	9.96	100	0.50
100	32.50	32.08	7.1	0	-268.2	39.29	243.29	0.23	7.58	100	0.50
150	21.94	21.72	7.3	0	-267.1	31.23	67.58	0.44	3.11	100	0.50
200	13.23	13.25	7.7	0	-340.4	25.15	39.00	0.90	2.94	100	0.50
500	2.88	2.91	8.0	0	-1102.5	6.56	1.05	1.25	0.36	100	0.50
50	55.85	55.27	9.1	0	263.1	25.39	109.31	-0.54	1.98	100	0.75
75	34.53	34.17	7.2	0	-345.6	21.20	63.47	-0.38	1.86	100	0.75
100	32.50	32.08	7.1	0	-464.3	19.65	36.61	-0.39	1.14	100	0.75
150	21.94	21.72	7.3	0	-579.1	17.45	14.49	-0.20	0.67	100	0.75
200	13.23	13.25	7.7	0	-637.8	15.74	9.71	0.19	0.73	100	0.75
500	2.88	2.91	8.0	0	-1993.2	8.26	1.24	1.84	0.43	100	0.75
1000	0.77	0.78	8.0	0	1213.7	2.80	0.09	2.61	0.11	100	0.75
1500	0.46	0.46	8.0	0	178.8	1.27	0.01	1.78	0.02	81	0.75
50	55.85	55.27	9.1	0	8.2	191.85	61399.17	2.47	1110.88	100	1.00
75	34.53	34.17	7.2	0	6.8	133.23	13844.25	2.90	405.13	100	1.00
100	32.50	32.08	7.1	0	6.6	81.45	4350.37	1.54	135.63	100	1.00
150	21.94	21.72	7.3	0	7.5	50.94	759.61	1.35	34.97	100	1.00
200	13.23	13.25	7.7	0	7.5	28.52	121.00	1.15	9.14	100	1.00
500	2.88	2.91	8.0	0	7.8	4.65	1.08	0.60	0.37	100	1.00
1000	0.77	0.78	7.9	0	8.0	1.28	0.04	0.65	0.05	99	1.00
1500	0.41	0.41	8.0	0	8.0	0.62	0.00	0.51	0.01	100	1.00
2000	0.33	0.34	8.0	0	8.1	0.38	0.00	0.12	0.00	62	1.00
50	55.85	55.27	9.1	0	136.6	55.37	518.78	0.00	9.39	100	25.00
75	34.53	34.17	7.2	0	-205.2	63.64	1084.96	0.86	31.75	100	25.00
100	32.50	32.08	7.1	0	-104.4	60.86	1277.26	0.90	39.82	100	25.00
150	21.94	21.72	7.3	0	-30.2	44.40	377.52	1.04	17.38	100	25.00
200	13.23	13.25	7.7	0	-47.3	28.58	101.25	1.16	7.64	100	25.00
500	2.88	2.91	8.0	0	-90.7	4.72	1.02	0.62	0.35	100	25.00



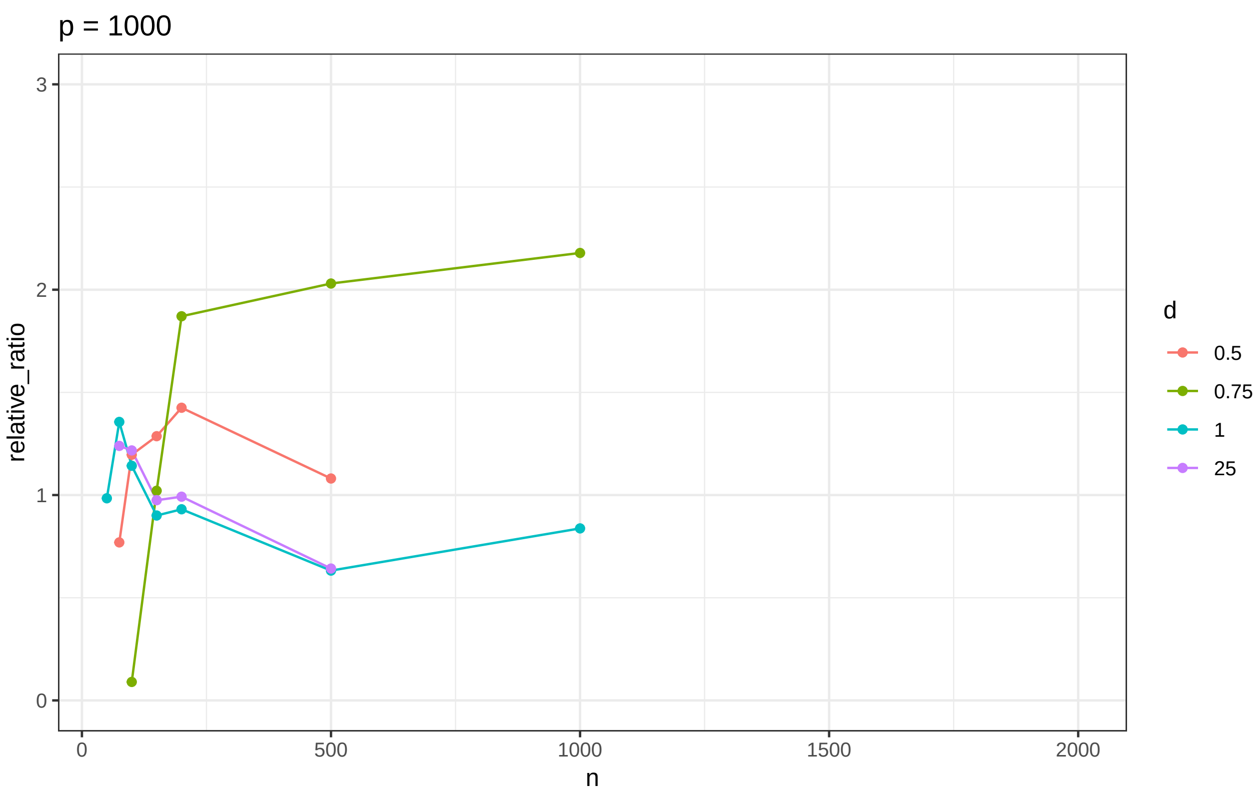
2.5.1.5 Eg with p = 100

n	MSE	est_mean	est_var_m	est_var	NA_main	EigenPrism_main_jack	EigenPrism_v_jack	EigenPrism_v_jack_var	relative_ratio	relative_ratio_var	N	d
50	21.6	8.4	33	21.7	0	-366.9	48.4	463.04	1.23	21.3	100	0.50
75	12.2	7.9	17	12.3	0	-463.3	25.6	77.41	1.08	6.3	100	0.50
100	7.1	8.0	12	7.2	0	-710.9	16.2	33.41	1.25	4.6	100	0.50
150	NaN	NaN	NaN	NA	100	NaN	8.4	4.89	NA	NA	100	0.50
200	NaN	NaN	NaN	NA	100	NaN	5.0	0.95	NA	NA	100	0.50
500	NaN	NaN	NaN	NA	100	NaN	NaN	NA	NA	NA	100	0.50
1000	NaN	NaN	NaN	NA	100	NaN	NaN	NA	NA	NA	100	0.50
1500	NaN	NaN	NaN	NA	99	NaN	NaN	NA	NA	NA	99	0.50
50	21.6	8.4	33	21.7	0	-1085.6	50.0	577.93	1.30	26.6	100	0.75
75	12.2	7.9	17	12.3	0	-1300.7	30.3	119.81	1.46	9.7	100	0.75
100	7.1	8.0	12	7.2	0	-1968.3	20.6	58.55	1.87	8.1	100	0.75
150	NaN	NaN	NaN	NA	100	NaN	11.2	9.62	NA	NA	100	0.75
200	NaN	NaN	NaN	NA	100	NaN	6.6	1.81	NA	NA	100	0.75
500	NaN	NaN	NaN	NA	100	NaN	NaN	NA	NA	NA	100	0.75
1000	NaN	NaN	NaN	NA	100	NaN	NaN	NA	NA	NA	100	0.75
1500	NaN	NaN	NaN	NA	100	NaN	NaN	NA	NA	NA	100	0.75
50	21.6	8.4	33	21.7	0	7.6	48.1	454.15	1.21	20.9	100	1.00
75	12.2	7.9	17	12.3	0	7.3	24.5	95.00	0.99	7.7	100	1.00
100	7.1	8.0	12	7.2	0	7.1	15.1	40.84	1.10	5.7	100	1.00
150	NaN	NaN	NaN	NA	100	NaN	NaN	NA	NA	NA	100	1.00
200	NaN	NaN	NaN	NA	100	NaN	NaN	NA	NA	NA	100	1.00
500	NaN	NaN	NaN	NA	100	NaN	NaN	NA	NA	NA	100	1.00
1000	NaN	NaN	NaN	NA	100	NaN	NaN	NA	NA	NA	100	1.00
1500	NaN	NaN	NaN	NA	99	NaN	NaN	NA	NA	NA	99	1.00
2000	NaN	NaN	NaN	NA	100	NaN	NaN	NA	NA	NA	100	1.00
50	21.6	8.4	33	21.7	0	-366.9	48.4	463.04	1.23	21.3	100	25.00
75	12.2	7.9	17	12.3	0	-213.1	24.3	71.89	0.98	5.8	100	25.00
100	7.1	8.0	12	7.2	0	-255.6	14.7	30.17	1.04	4.2	100	25.00
150	NaN	NaN	NaN	NA	100	NaN	NaN	NA	NA	NA	100	25.00
200	NaN	NaN	NaN	NA	100	NaN	NaN	NA	NA	NA	100	25.00
500	NaN	NaN	NaN	NA	100	NaN	NaN	NA	NA	NA	100	25.00
1000	NaN	NaN	NaN	NA	100	NaN	NaN	NA	NA	NA	100	25.00
1500	NaN	NaN	NaN	NA	20	NaN	NaN	NA	NA	NA	20	25.00



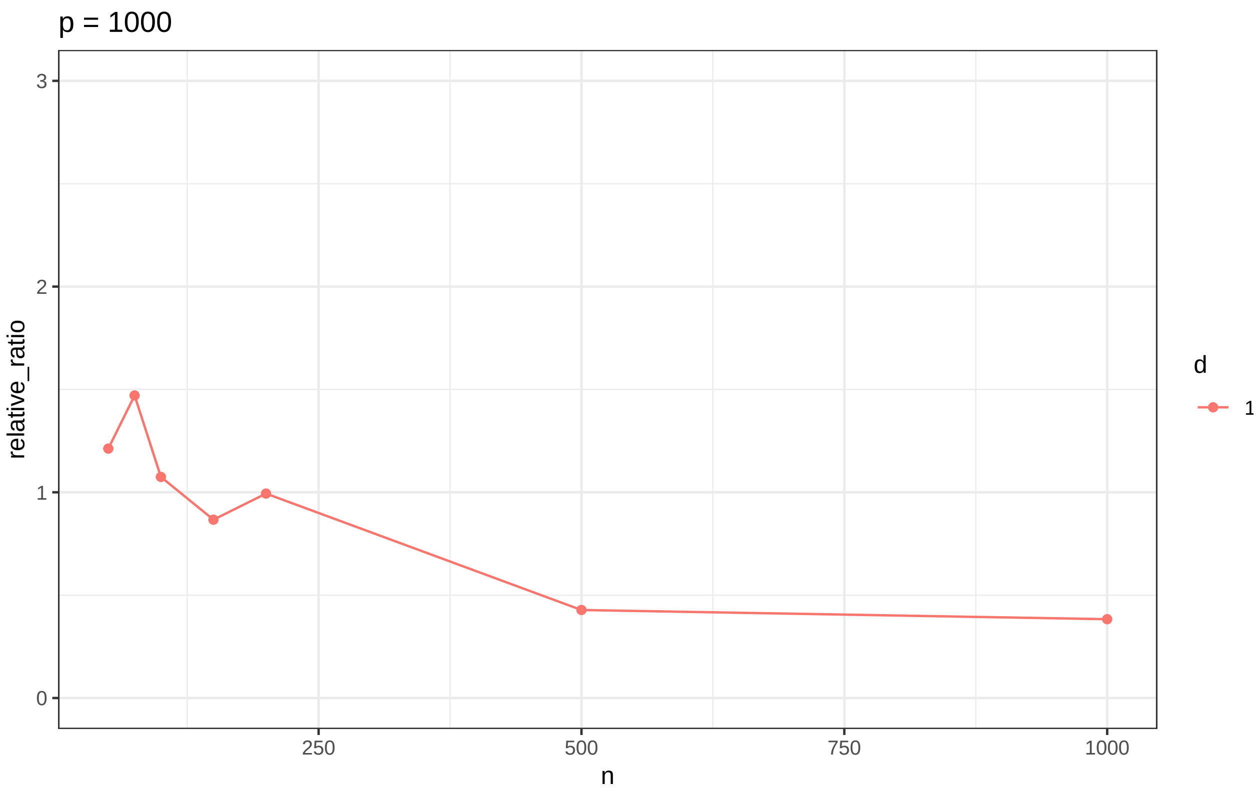
2.5.1.6 Eg with p = 1000

n	MSE	est_mean	est_var_m	est_var	NA_main	EigenPrism_main_jack	EigenPrism_v_jack	EigenPrism_v_jack_var	relative_ratio	relative_ratio_var	N	d
50	139.7	12.0	161.9	125.05	0	-784.3	113.0	3039.25	-0.10	24.30	100	0.50
75	57.4	8.7	77.1	57.44	0	-1435.4	101.6	2347.86	0.77	40.87	100	0.50
100	39.0	7.7	49.3	39.36	0	-1558.8	86.5	853.63	1.20	21.69	100	0.50
150	20.9	7.9	24.3	21.06	0	-1415.7	48.2	142.18	1.29	6.75	100	0.50
200	12.4	7.9	15.0	12.56	0	-2090.4	30.4	49.02	1.43	3.90	100	0.50
500	3.0	8.0	3.2	2.99	0	-3215.4	6.2	0.86	1.08	0.29	100	0.50
50	139.7	12.0	161.9	125.05	0	-1330.9	29.0	203.55	-0.77	1.63	100	0.75
75	57.4	8.7	77.1	57.44	0	-3181.2	37.7	298.49	-0.34	5.20	100	0.75
100	39.0	7.7	49.3	39.36	0	-4683.8	42.9	178.16	0.09	4.53	100	0.75
150	20.9	7.9	24.3	21.06	0	-5285.8	42.6	102.62	1.02	4.87	100	0.75
200	12.4	7.9	15.0	12.56	0	-6109.9	36.0	66.79	1.87	5.32	100	0.75
500	3.0	8.0	3.2	2.99	0	-9625.8	9.1	1.70	2.03	0.57	100	0.75
1000	0.8	8.0	1.1	0.81	0	-5670.4	2.6	0.07	2.18	0.08	100	0.75
1500	NaN	NaN	NaN	NA	81	NaN	1.2	0.01	NA	NA	81	0.75
50	139.7	12.0	161.9	125.05	0	9.3	248.1	18031.74	0.98	144.19	100	1.00
75	57.4	8.7	77.1	57.44	0	6.9	135.3	4335.86	1.36	75.48	100	1.00
100	39.0	7.7	49.3	39.36	0	6.1	84.3	954.62	1.14	24.25	100	1.00
150	20.9	7.9	24.3	21.06	0	7.3	40.0	120.60	0.90	5.73	100	1.00
200	12.4	7.9	15.0	12.56	0	7.2	24.2	39.43	0.93	3.14	100	1.00
500	3.0	8.0	3.2	2.99	0	7.7	4.9	0.83	0.63	0.28	100	1.00
1000	0.8	8.0	1.1	0.81	0	8.0	1.5	0.03	0.84	0.04	99	1.00
1500	NaN	NaN	NaN	NA	100	NaN	NaN	NA	NA	NA	100	1.00
2000	NaN	NaN	NaN	NA	62	NaN	NaN	NA	NA	NA	62	1.00
50	139.7	12.0	161.9	125.05	0	-784.3	113.0	3039.25	-0.10	24.30	100	25.00
75	57.4	8.7	77.1	57.44	0	-613.1	128.6	4033.63	1.24	70.22	100	25.00
100	39.0	7.7	49.3	39.36	0	-487.3	87.3	895.86	1.22	22.76	100	25.00
150	20.9	7.9	24.3	21.06	0	-258.1	41.6	119.32	0.97	5.67	100	25.00
200	12.4	7.9	15.0	12.56	0	-288.7	25.0	42.04	0.99	3.35	100	25.00
500	3.0	8.0	3.2	2.99	0	-175.6	4.9	0.80	0.64	0.27	100	25.00



2.5.1.7 Dicker 2013 with p = 1000

n	MSE	est_mean	est_var_m	est_var	NA_main	Dicker_main_jack	Dicker_v_jack	relative_ratio	relative_ratio_var	d
50	214.7	7.0	250.5	215.8	0	13.3	477.5	1.21	285.16	1
75	81.5	5.7	98.1	77.1	0	8.2	190.5	1.47	116.05	1
100	59.7	5.4	57.7	53.3	0	7.0	110.6	1.07	35.10	1
150	26.7	6.6	27.0	25.0	0	7.2	46.6	0.87	8.50	1
200	14.4	7.2	16.5	14.0	0	7.6	27.9	0.99	4.20	1
500	4.0	8.1	3.6	4.0	0	8.2	5.7	0.43	0.39	1
1000	1.3	7.8	1.3	1.3	0	7.8	1.7	0.38	0.06	1



3 Subsampling method: bootstrap

3.1 non-parameteric bootstrap

3.2 Parametric bootstrap

As the previous section mentioned, the non-parametric bootstrap may not work well for the total signal estimation. A parametric bootstrap method was proposed for CI estimation of the Heritability. Detials at (Schweiger et al. 2016). The main idea of the parametric is following: If we assume the mixed effect model, then we have

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{s} + \mathbf{e}$$

where X is the fixed covaraite and coefficient, and Z and s are for the random effect, $\mathbf{s} \sim N\left(\mathbf{0}_m, \sigma_g^2 \mathbf{I}_m/m\right)$ and $\mathbf{e} \sim N\left(\mathbf{0}_n, \sigma_e^2 \mathbf{I}_n\right)$. Note in here we assume that Z is fixed with columns' mean 0 and varinace 1.

$$\mathbf{y} \sim N\left(\mathbf{X}\boldsymbol{\beta}, \sigma_g^2 \mathbf{K} + \sigma_e^2 \mathbf{I}_n\right),$$

where $\mathbf{K} = \mathbf{Z}\mathbf{Z}^{\mathbf{T}}/m$. Let h^2 as the narrow-sense of heritability

$$h^2 = \frac{\sigma_g^2}{\sigma_g^2 + \sigma_e^2},$$

then we will have

$$\mathbf{y} \sim N\left(\mathbf{X}\boldsymbol{\beta}, \sigma_p^2\left(h^2\mathbf{K} + \left(1 - h^2\right)\mathbf{I}_n\right)\right).$$

In the paper, the author showed that the distribution $\hat{h}^2$ only depends on h^2, i.e. $h^2 = H^2$. Therefore, we could just fix $\sigma_p^2 = 1$ and $\beta = 0_p$. So we have the distribution of y as

$$N\left(\mathbf{0}_n, h^2\mathbf{K} + \left(1 - h^2\right)\mathbf{I}_n\right)$$

. So the direct parametric bootstrap will be

1. Random sampling: draw $N($ e.g. $, 10,000)$ phenotype vectors $\mathbf{y}_1^*, \ldots, \mathbf{y}_N^*$ from the multivariate normal distribution $N\left(\mathbf{0}_n, h^2\mathbf{K} + \left(1 - h^2\right)\mathbf{I}_n\right)$

2. REML estimation: calculate the REML estimates $\widehat{h}^2\left(\mathbf{y}_1^*\right), \ldots, \widehat{h}^2\left(\mathbf{y}_N^*\right)$ for each of these phenotype vectors by using a software package such as GCTA (Genome-wide Complex Trait Analysis)

3. Density estimation: for each one of the bins above, count the proportion of estimates $\widehat{h}^2\left(\mathrm{y}_i^*\right)$ that fall in that bin; similarly, compute the fraction of estimates equal to a boundary estimate $\hat{h}^2\left(\mathrm{y}_i^*\right) = 0$ or 1. Use these fractions as an estimate of the density of $\widehat{h}^2$ for this value of h^2.

Based on our context, we have $Y = X\beta + \epsilon$ and under certain conditions , $Var(X\beta) = \sum \beta_i^2$we have

$$Y \sim N(0, \sum \beta_i^2 + \sigma_\epsilon^2)$$

so the extend the parametric bootstrap in our simulation setup, we may need try

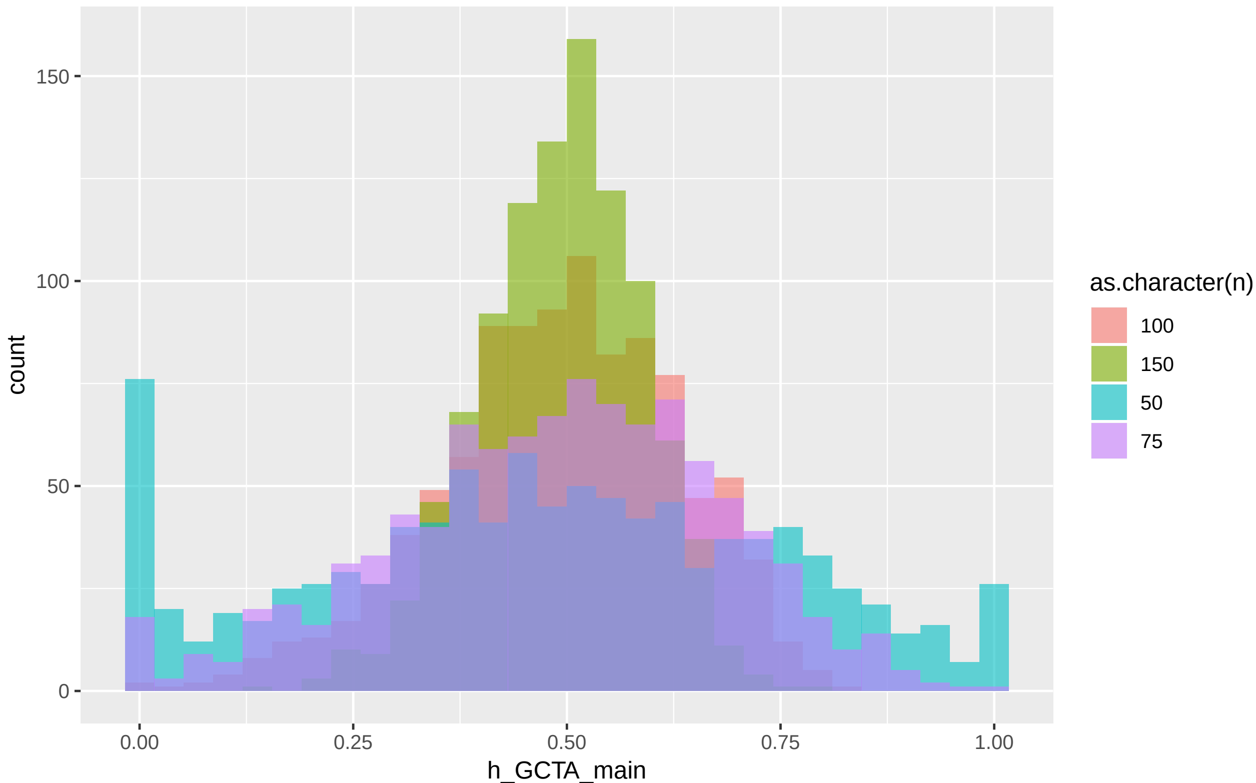
$$Y^* \sim N(0, v\hat{a}r(X\beta) + \hat{\sigma}_\epsilon^2)$$

For GCTA, we could use the Delta thoery to find the estimator's variance of its asymptotic distribution. However, simulation results suggest that in most of the case the $\hat{h^2}$ will have a skew distribution with many values around $0, 1$. Therefore, the Delta method may not give us an accurate variance esitmation of the estimatiors.

3.2.1 simulation result

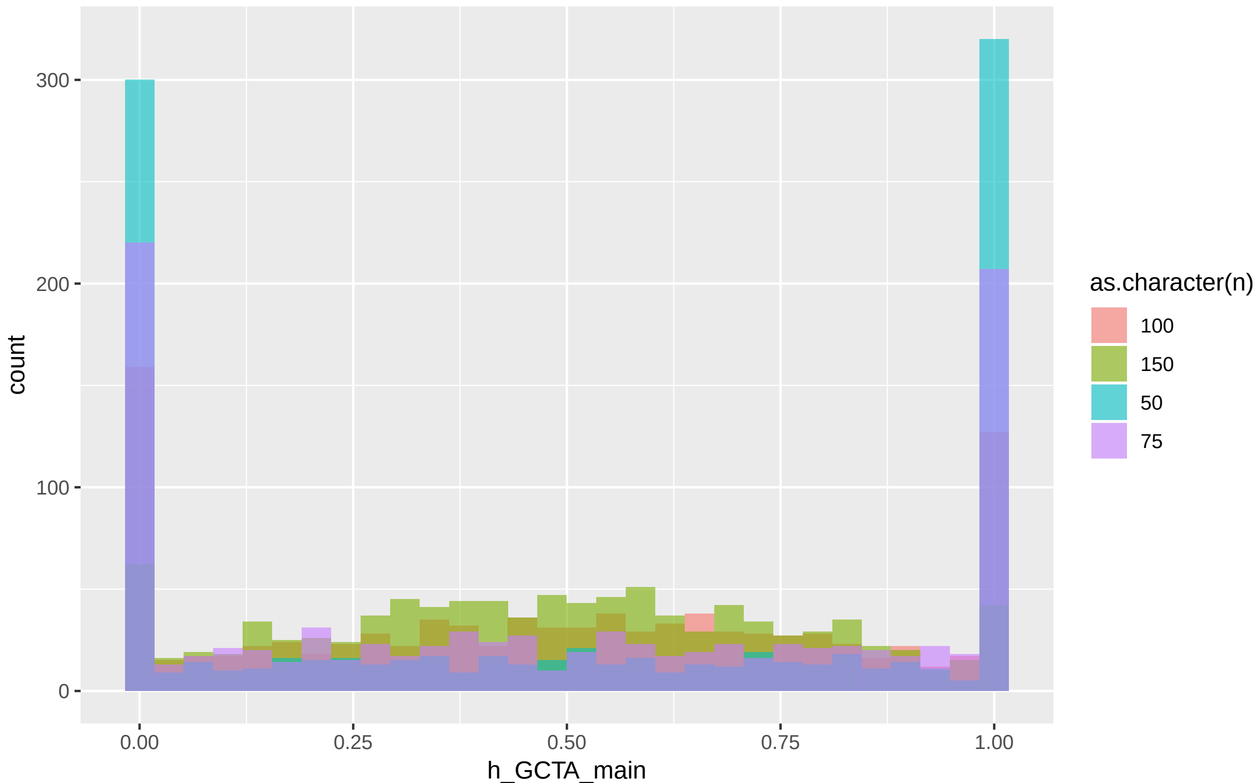
Note that we adopt ρ from (Janson, Barber, and Candes 2017), which in our case $\rho = h^2$. For $h^2 = 0.5$ and $X \sim N(0_p I_p)$, the parameteric bootstrap seems to work well when $p = 100$. However, when $p = 1000$, the relative ratio is large. The possible reason is that the sample size is too small consider the n.

3.2.1.1 p = 100 rho = 0.5



n	MSE	est_h2_var	est_h2	N	NA_main	est_h2_bs	est_h2_var_bs	relative_ratio	relative_ratio_var
50	0.07	0.07	0.47	1000	0	0.48	0.06	-0.18	0.01
75	0.04	0.04	0.49	1000	0	0.48	0.03	-0.14	0.00
100	0.02	0.02	0.49	1000	0	0.48	0.02	-0.02	0.00
150	0.01	0.01	0.49	1000	0	0.49	0.01	0.09	0.00

3.2.1.2 p = 1000 rho = 0.5



n	MSE	est_h2_var	est_h2	N	NA_main	est_h2_bs	est_h2_var_bs	relative_ratio	relative_ratio_var
50	0.18	0.18	0.51	999	0	0.51	0.16	-0.11	0.00
75	0.15	0.15	0.50	999	0	0.50	0.12	-0.18	0.00
100	0.12	0.12	0.49	999	0	0.50	0.10	-0.17	0.01
150	0.08	0.08	0.48	1000	0	0.49	0.06	-0.21	0.00

3.2.2 Simulation results for compare bootstrap with other inference methods

Based on the simulation result from last section, we found that the exisiting inference methods are similar with each other. So in this section, we consider GCTA with parametric method as one method and compare its inference ability with dickers and EigenPrism.

- $n = 250, 500$ and $p = 500$, and $h^2 = \{0.1, 0.3, \ldots, 0.9\}$

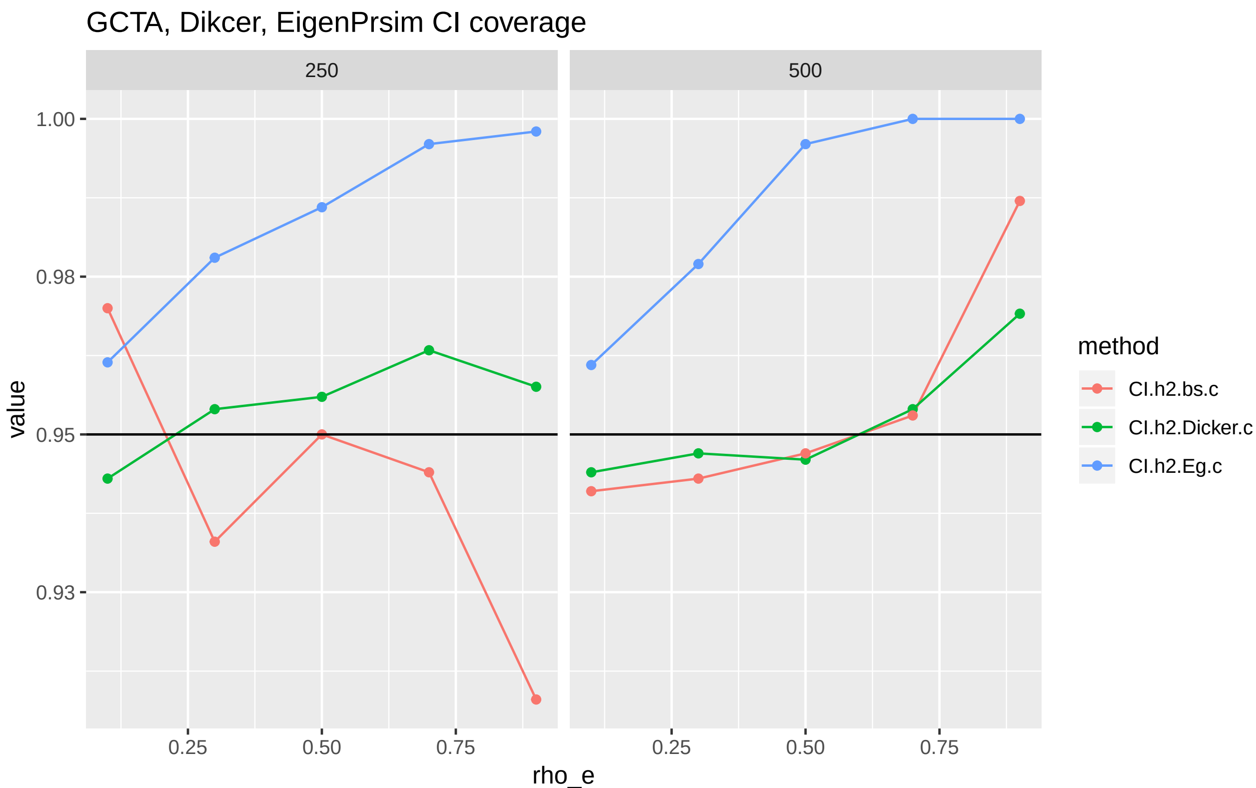
- $X \sim N(0_p, I_p)$.
- β's sparsity rate is 0.5 The inference accuracy is defined as

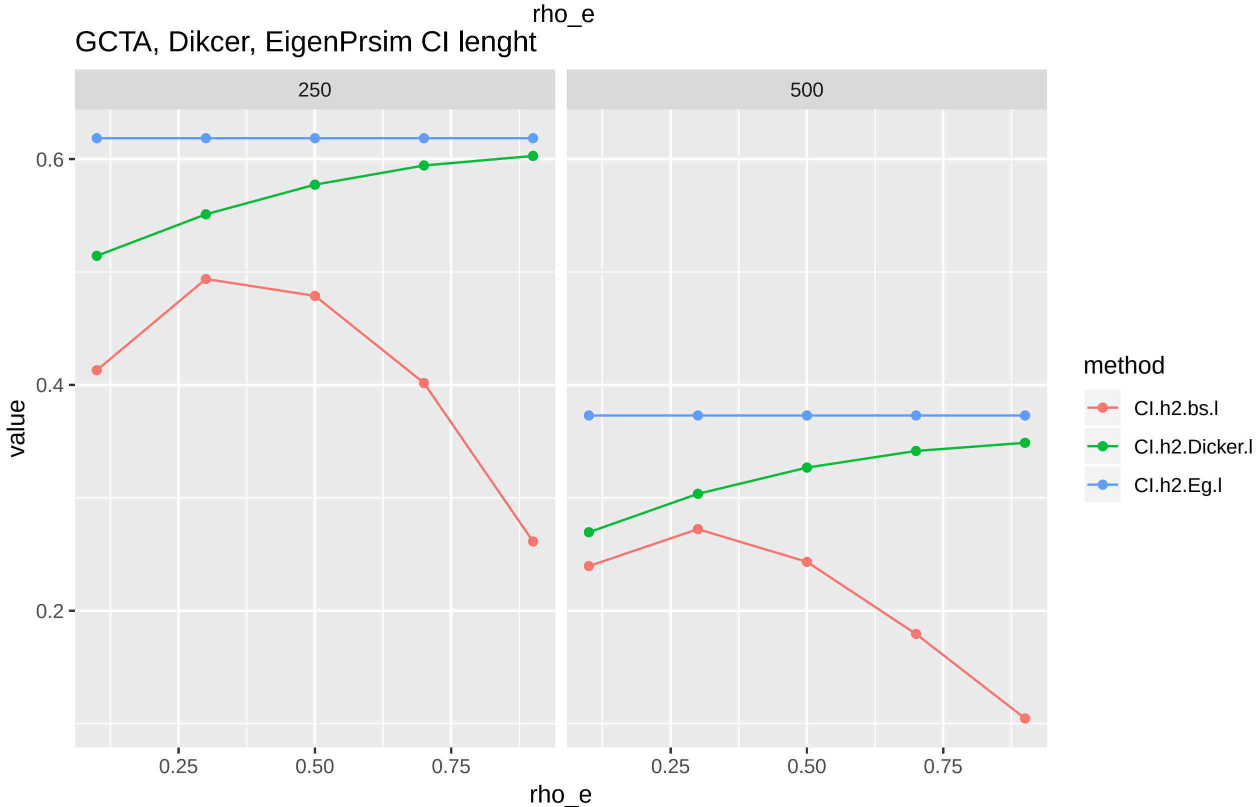
- CI's length

- CI's coverage rate

3.2.2.1 CI coverage rate and lenght

n	rho_e	CI.h2.Eg.c	CI.h2.Eg.l	CI.h2.Dicker.c	CI.h2.Dicker.l	CI.h2.bs.c	CI.h2.bs.l
250	0.1	0.96	0.62	0.94	0.51	0.97	0.41
250	0.3	0.98	0.62	0.95	0.55	0.93	0.49
250	0.5	0.99	0.62	0.96	0.58	0.95	0.48
250	0.7	1.00	0.62	0.96	0.59	0.94	0.40
250	0.9	1.00	0.62	0.96	0.60	0.91	0.26
500	0.1	0.96	0.37	0.94	0.27	0.94	0.24
500	0.3	0.98	0.37	0.95	0.30	0.94	0.27
500	0.5	1.00	0.37	0.95	0.33	0.95	0.24
500	0.7	1.00	0.37	0.95	0.34	0.95	0.18
500	0.9	1.00	0.37	0.97	0.35	0.99	0.10





3.2.3 Simulation results for GCTA with para-bs under total normal

In this section, we test the para-bs under the interaction effect.

- $n = 250, 500$ and $p = 31$, and $h^2 = \{0.1, 0.3, \dots, 0.9\}$

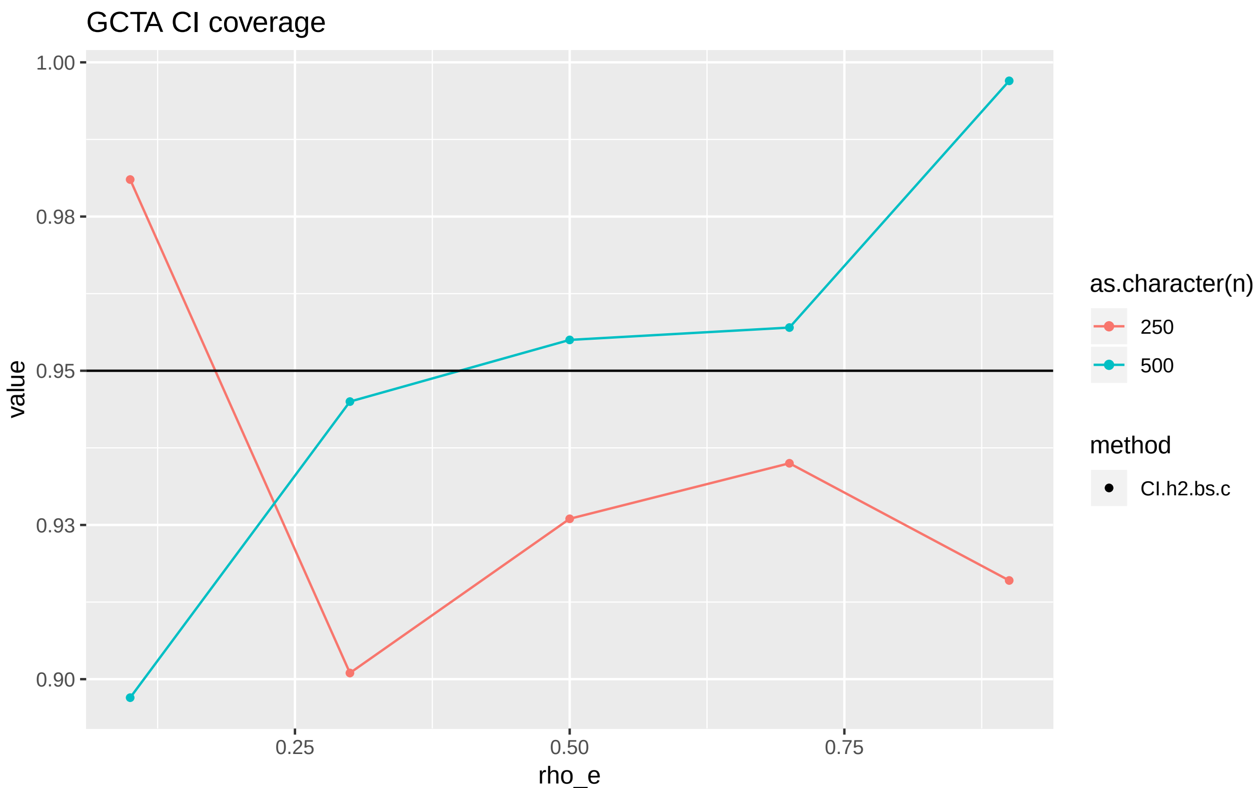
- $X \sim N(0_p, I_p)$.
- β's sparsity rate is 0.5 The inference accuracy is defined as

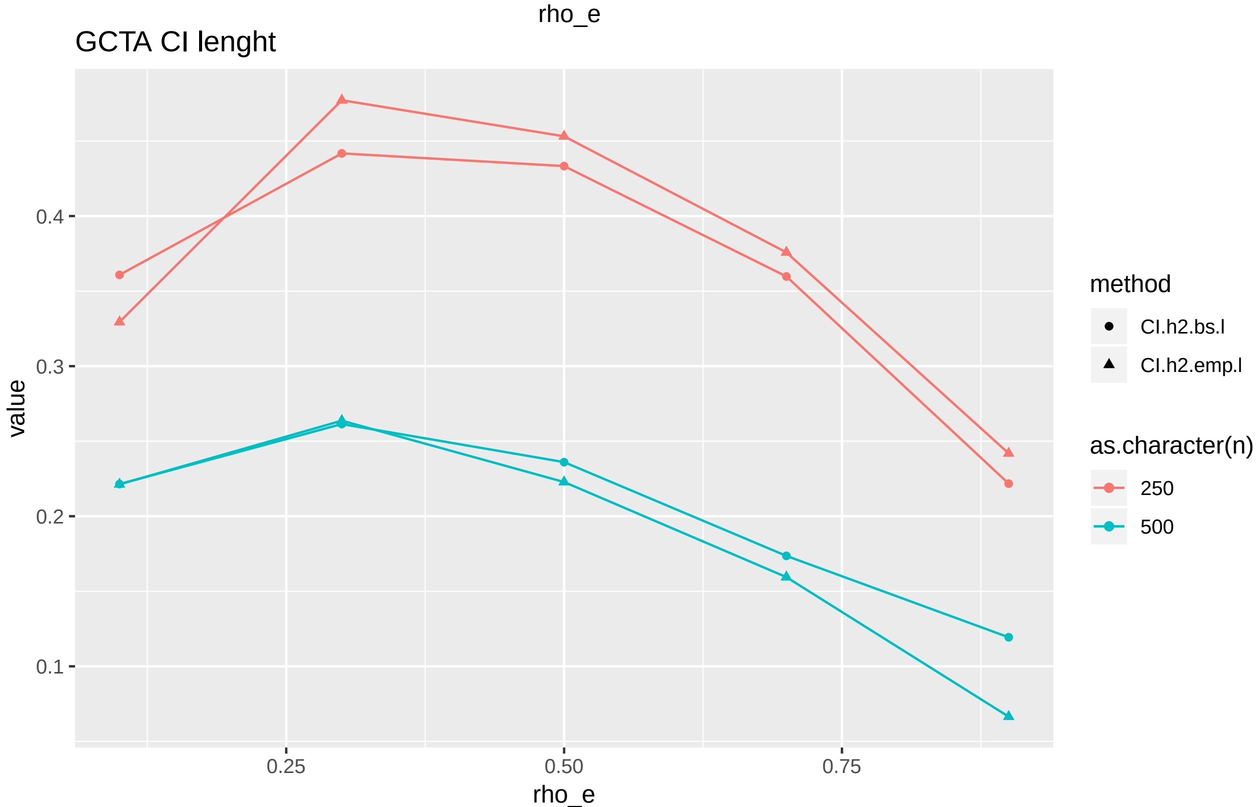
- CI's length

- CI's coverage rate

3.2.3.1 CI coverage rate and lenght

n	rho_e	CI.h2.emp.L	CI.h2.emp.R	CI.h2.emp.l	var.h2.emp	h2	N	NA_total	h2.bs	var.h2.bs	CI.h2.bs.c	CI.h2.bs.L	CI.h2.bs.R	CI.h2.bs.l
250	0.1	0.00	0.33	0.33	0.01	0.11	1000	0	0.12	0.01	0.98	-0.06	0.30	0.36
250	0.3	0.03	0.51	0.48	0.01	0.28	1000	0	0.28	0.01	0.90	0.06	0.50	0.44
250	0.5	0.23	0.69	0.45	0.01	0.48	1000	0	0.47	0.01	0.93	0.25	0.69	0.43
250	0.7	0.47	0.85	0.38	0.01	0.68	1000	0	0.68	0.01	0.94	0.50	0.86	0.36
250	0.9	0.76	1.00	0.24	0.00	0.89	1000	0	0.89	0.00	0.92	0.78	1.00	0.22
500	0.1	0.00	0.22	0.22	0.00	0.09	1000	0	0.10	0.00	0.90	-0.01	0.21	0.22
500	0.3	0.17	0.43	0.26	0.00	0.29	1000	0	0.29	0.00	0.94	0.16	0.42	0.26
500	0.5	0.39	0.61	0.22	0.00	0.50	1000	0	0.50	0.00	0.96	0.38	0.61	0.24
500	0.7	0.62	0.78	0.16	0.00	0.71	1000	0	0.70	0.00	0.96	0.61	0.79	0.17
500	0.9	0.87	0.93	0.07	0.00	0.90	1000	0	0.89	0.00	1.00	0.83	0.95	0.12





3.2.4 Simulation results for GCTA with para-bs under PCN

In this section, we test the para-bs under the interaction effect.

- $n = 100, 150, 231$ and $p = 21$, and $h^2 = \{0.1, 0.3, \ldots, 0.9\}$

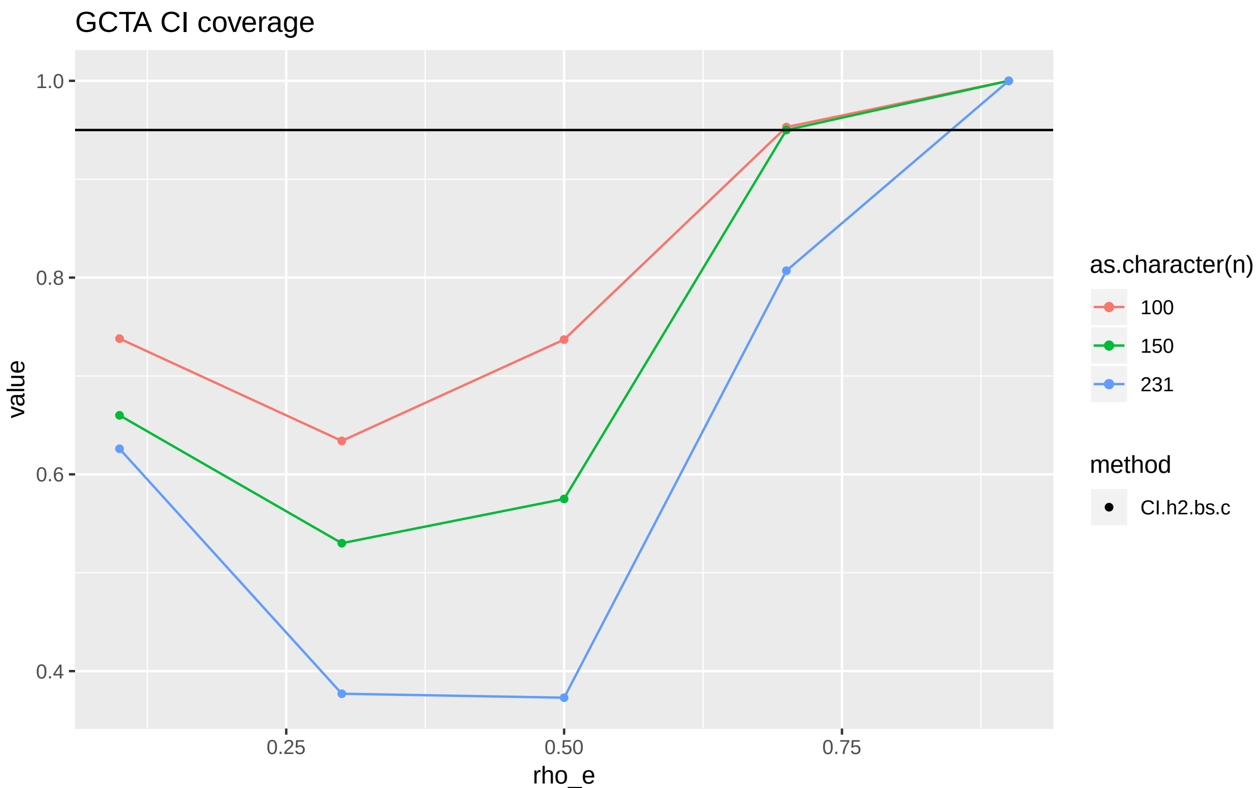
- X PCB from 1999-2004 years.
- β's sparsity rate is 0.5 The inference accuracy is defined as

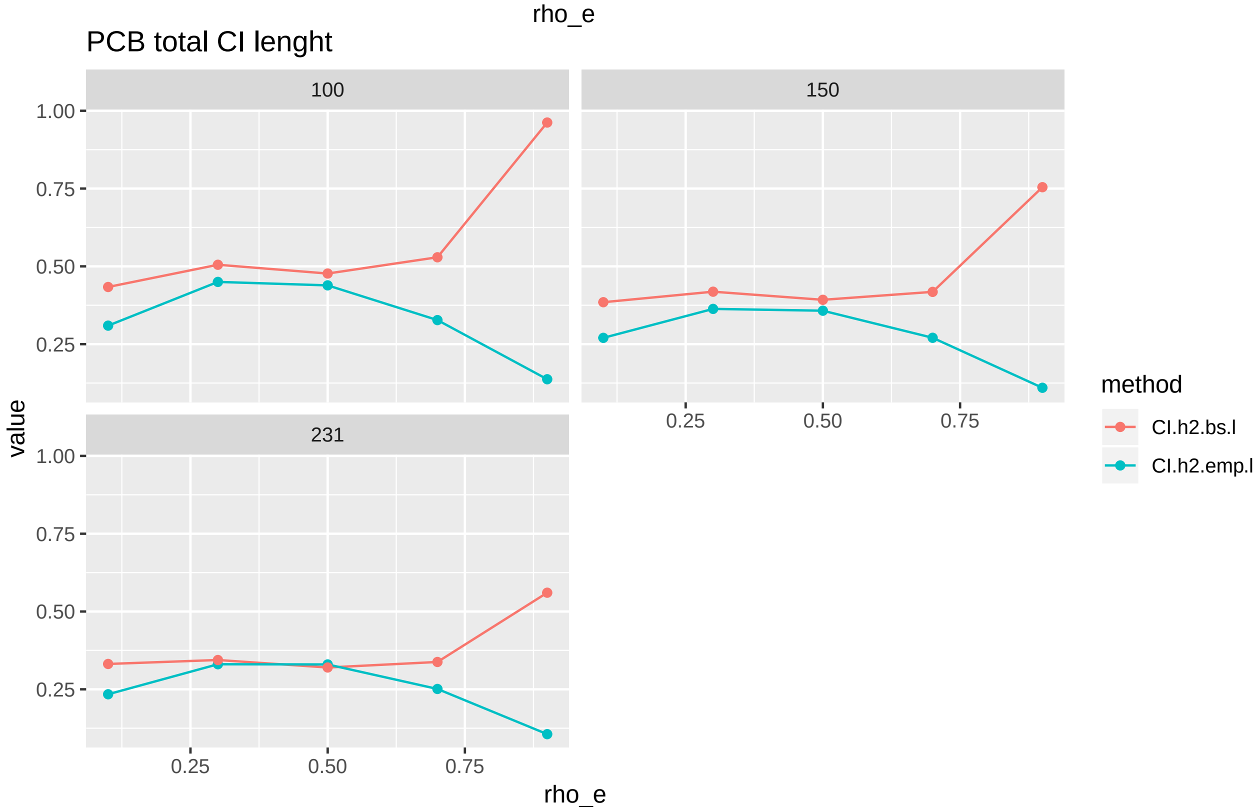
- CI's length

- CI's coverage rate

3.2.4.1 CI coverage rate and lenght

n	rho_e	CI.h2.emp.L	CI.h2.emp.R	CI.h2.emp.l	var.h2.emp	h2	N	NA_total	h2.bs	var.h2.bs	CI.h2.bs.c	CI.h2.bs.L	CI.h2.bs.R	CI.h2.bs.l
100	0.1	0.00	0.31	0.31	0.01	0.10	1000	0	0.23	0.01	0.74	0.02	0.45	0.43
100	0.3	0.06	0.51	0.45	0.01	0.28	1000	0	0.49	0.02	0.63	0.24	0.74	0.51
100	0.5	0.26	0.69	0.44	0.01	0.47	1000	0	0.65	0.02	0.74	0.41	0.89	0.48
100	0.7	0.51	0.83	0.33	0.01	0.68	1000	0	0.76	0.02	0.95	0.50	1.03	0.53
100	0.9	0.81	0.95	0.14	0.00	0.89	1000	0	0.80	0.07	1.00	0.32	1.28	0.96
150	0.1	0.00	0.27	0.27	0.01	0.10	1000	0	0.24	0.01	0.66	0.05	0.43	0.38
150	0.3	0.12	0.49	0.36	0.01	0.29	1000	0	0.51	0.01	0.53	0.30	0.71	0.42
150	0.5	0.31	0.67	0.36	0.01	0.49	1000	0	0.67	0.01	0.57	0.47	0.87	0.39
150	0.7	0.55	0.82	0.27	0.00	0.69	1000	0	0.78	0.01	0.95	0.57	0.99	0.42
150	0.9	0.83	0.94	0.11	0.00	0.89	1000	0	0.84	0.04	1.00	0.46	1.21	0.75
231	0.1	0.00	0.23	0.23	0.00	0.09	1000	0	0.24	0.01	0.63	0.07	0.41	0.33
231	0.3	0.16	0.49	0.33	0.01	0.29	1000	0	0.52	0.01	0.38	0.35	0.69	0.34
231	0.5	0.35	0.68	0.33	0.01	0.49	1000	0	0.69	0.01	0.37	0.53	0.85	0.32
231	0.7	0.58	0.83	0.25	0.00	0.69	1000	0	0.80	0.01	0.81	0.63	0.97	0.34
231	0.9	0.84	0.95	0.11	0.00	0.89	1000	0	0.87	0.03	1.00	0.59	1.15	0.56


GCTA CI coverage
value
rho_e
as.character(n)
100
150
231
method
CI.h2.bs.c
0.4
0.6
0.8
1.0
0.25
0.50
0.75


PCB total CI lenght
100
150
231
value
rho_e
method
CI.h2.bs.l
CI.h2.emp.l
0.25
0.50
0.75
1.00

Delta method

Efron, Bradley, and Charles Stein. 1981. "The Jackknife Estimate of Variance." *The Annals of Statistics*. JSTOR, 586–96.

Janson, Lucas, Rina Foygel Barber, and Emmanuel Candes. 2017. "EigenPrism: Inference for High Dimensional Signal-to-Noise Ratios." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 79 (4). Wiley Online Library: 1037–65.

Schweiger, Regev, Shachar Kaufman, Reijo Laaksonen, Marcus E Kleber, Winfried März, Eleazar Eskin, Saharon Rosset, and Eran Halperin. 2016. "Fast and Accurate Construction of Confidence Intervals for Heritability." *The American Journal of Human Genetics* 98 (6). Elsevier: 1181–92.

Shao, Jun, CF Jeff Wu, and others. 1989. "A General Theory for Jackknife Variance Estimation." *The Annals of Statistics* 17 (3). Institute of Mathematical Statistics: 1176–97.