PCB sub sampling simulation

Xuelong Wang

2019-09-03

Contents

1 Motivation

Based on the previous simulation results, we found that after using the historical data to decorrelate the data, the covariates still have some low correlations among them. But it seems that GCTA and EigenPrism could work well even the covariate have some low correlations among them. To further mimic the real data situation, we using the combined PCBs 1999 - 2004 data to simulate the high dimensional data and try to estimate the combined total vairance by using GCTA and EigenPrism.

2 Simulation

2.1 Simulation result

2.1.1 Chi

- $cov(X) = cor(PCB|year = 1999)$
- $p = 21$
- target is the combined main and interaction effect
- X ~ χ^2_1
- $n = 100, 150, 231$

```
     n MSE est_var est_mean NA_total     method total decor x_dist
 1: 100  82      66       18        0 EigenPrism    14 FALSE    chi
 2: 150 110      72       21        0 EigenPrism    14 FALSE    chi
 3: 231 114      64       22        0 EigenPrism    14 FALSE    chi
 4: 100  79      66       18        0       GCTA    14 FALSE    chi
 5: 150 108      77       20        0       GCTA    14 FALSE    chi
 6: 231 124      80       21        0       GCTA    14 FALSE    chi
 7: 100  26      26       13        0 EigenPrism    14  TRUE    chi
 8: 150  24      25       14        0 EigenPrism    14  TRUE    chi
 9: 231  15      15       14        0 EigenPrism    14  TRUE    chi
10: 100  28      27       13        0       GCTA    14  TRUE    chi
11: 150  20      20       14        0       GCTA    14  TRUE    chi
12: 231  14      14       14        0       GCTA    14  TRUE    chi
```

2.1.2 PCBs

- X: will be 21 PCBs or after adding interaction terms 231
- n: 100,150,231
- traget: $\beta^T \hat{\Sigma}_h \beta$. Since we don't know the extact covariance matrix of the PCBs so we are using the all the historical data to estate the covariance matrix $\hat{\Sigma}_h$

	n	MSE	est_var	est_mean	NA_total	method	total	decor	x_dist
1:	100	193	134	19.0	4	EigenPrism	11	FALSE	1999
2:	150	483	332	23.7	0	EigenPrism	11	FALSE	1999
3:	231	320	177	23.3	0	EigenPrism	11	FALSE	1999
4:	100	153	132	15.9	0	GCTA	11	FALSE	1999
5:	150	693	587	21.8	0	GCTA	11	FALSE	1999
6:	231	292	194	21.2	0	GCTA	11	FALSE	1999
7:	100	47	47	10.9	0	EigenPrism	11	TRUE	1999
8:	150	37	37	10.7	0	EigenPrism	11	TRUE	1999
9:	231	18	15	9.5	0	EigenPrism	11	TRUE	1999
10:	100	50	49	10.1	0	GCTA	11	TRUE	1999
11:	150	39	39	10.2	0	GCTA	11	TRUE	1999
12:	231	18	16	9.5	0	GCTA	11	TRUE	1999

2.1.2.1 Simulation result with larger n

If we subset data as the corresponding years, then it seems that we can get covariance matrix which is very close to I. However, if the size of the sub-sample is small, e.i.100, then the decorrelated covariance matrix may be not close to I

- $n \in \{100, 500, 1000, 2000\}$

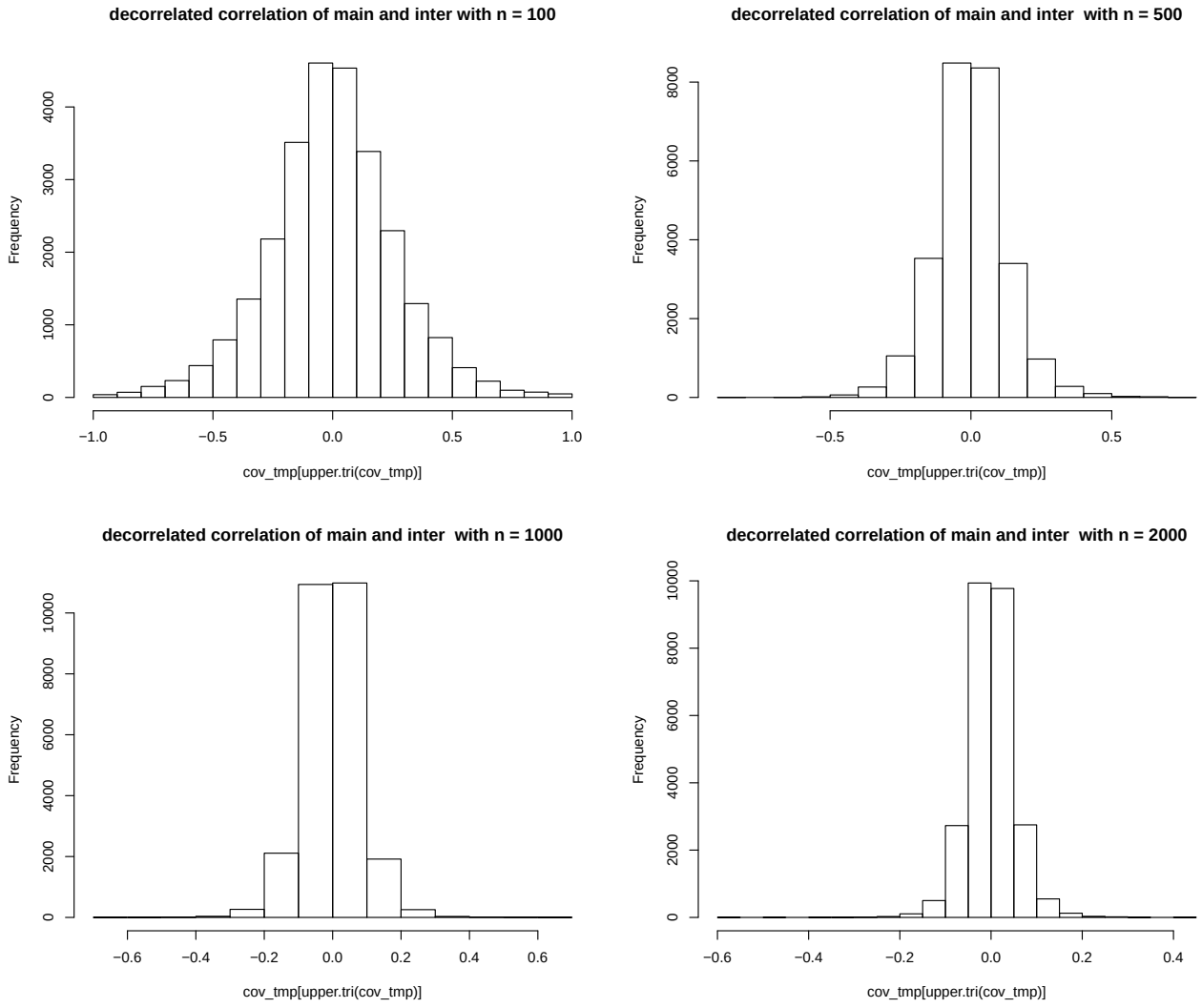

Figure 1: Combined main and interaction 1999-2000

```
      n  MSE est_var est_mean NA_total method total decor x_dist
1:  100 49.9    49.1     10.1        0   GCTA    11  TRUE   1999
2:  500 11.2     9.2      9.8        0   GCTA    11  TRUE   1999
3: 1000  6.9     5.0      9.9        1   GCTA    11  TRUE   1999
4: 2000  3.8     2.4     10.0        2   GCTA    11  TRUE   1999
```

3 problems

3.1 Interaction and decorrelation

The procedure to generate and estimate the combined main and interaction effect is followings:

1. $X \to X_t = (X, X_{inter}) \to Y = X^t\beta_t + \epsilon$
2. $X \to X_t = (X, X_{inter}) \to Z_t = X_t^T \Sigma_h^{-1/2} \to \hat{Var}(X_t^T \beta_t)$

The problem I encounter is that if I did not standerdize the PCBs then it seems that the estimated combined effect is close to the target even if we don't apply the decorrelation procedure.

	n	MSE	est_var	est_mean	NA_total	method	total	decor	x_dist
1:	100	129	112	10.5	1	EigenPrism	15	FALSE	1999
2:	150	221	219	12.8	0	EigenPrism	15	FALSE	1999
3:	231	126	114	11.2	0	EigenPrism	15	FALSE	1999
4:	100	88	55	9.1	0	GCTA	15	FALSE	1999
5:	150	207	190	10.4	0	GCTA	15	FALSE	1999
6:	231	93	63	9.3	0	GCTA	15	FALSE	1999
7:	100	97	82	10.8	0	EigenPrism	15	TRUE	1999
8:	150	114	99	10.8	0	EigenPrism	15	TRUE	1999
9:	231	61	28	9.1	0	EigenPrism	15	TRUE	1999
10:	100	132	113	10.3	0	GCTA	15	TRUE	1999
11:	150	130	113	10.6	0	GCTA	15	TRUE	1999
12:	231	64	31	9.1	0	GCTA	15	TRUE	1999

After some simulation, I found it is possible because of the small values of the PCBs. Although the correlation of PBCs are high, the covariance of PCBs are small because of the small values of PCB. decorerlation i