Sparse covariance estimation

Xuelong Wang

2019-10-03

Contents

1 Motivation

After decorrelation with the information of the historical data, the correlation between the covariate is reduced. However, there are still some correlation coeifficients are large. That may suggest that after using the historical data, the decorrelated data still is not uncorrelated. But the correlation structure becomes a sparse and symmetric. Therefore, we could apply another decorrelation to further reudce the non-zero correlation, so that we may have a better performance on the following variance estimation procedure.

2 Simulation

2.1 Simulation procedure

2.1.1 Standardization will not change total variance

1. Standardize the X $\tilde{Z}_m = (X - \mu)A_1$
2. Generate the interaction based on the $\tilde{Z}_{int} = \tilde{Z}_m * \tilde{Z}_m$ without the square terms and set $\tilde{Z}_t = (\tilde{Z}_m, \tilde{Z}_{int})$
3. Generate the Y based on the $\tilde{Z}_t$
4. Estimate the $Var(\tilde{Z}_t\beta_t)$ by $Z = \tilde{Z}_t A_2$, where A_2 is for decorrelation

Note that the $Var(\tilde{Z}_t\beta_t) = Var(Z\gamma)$, $A_2 = \hat{\Sigma}_h^{-1/2}$ or $A_2 = \hat{\Sigma}_h^{-1/2}\hat{\Sigma}_s^{-1/2}$

2.2 Decorrelation steps

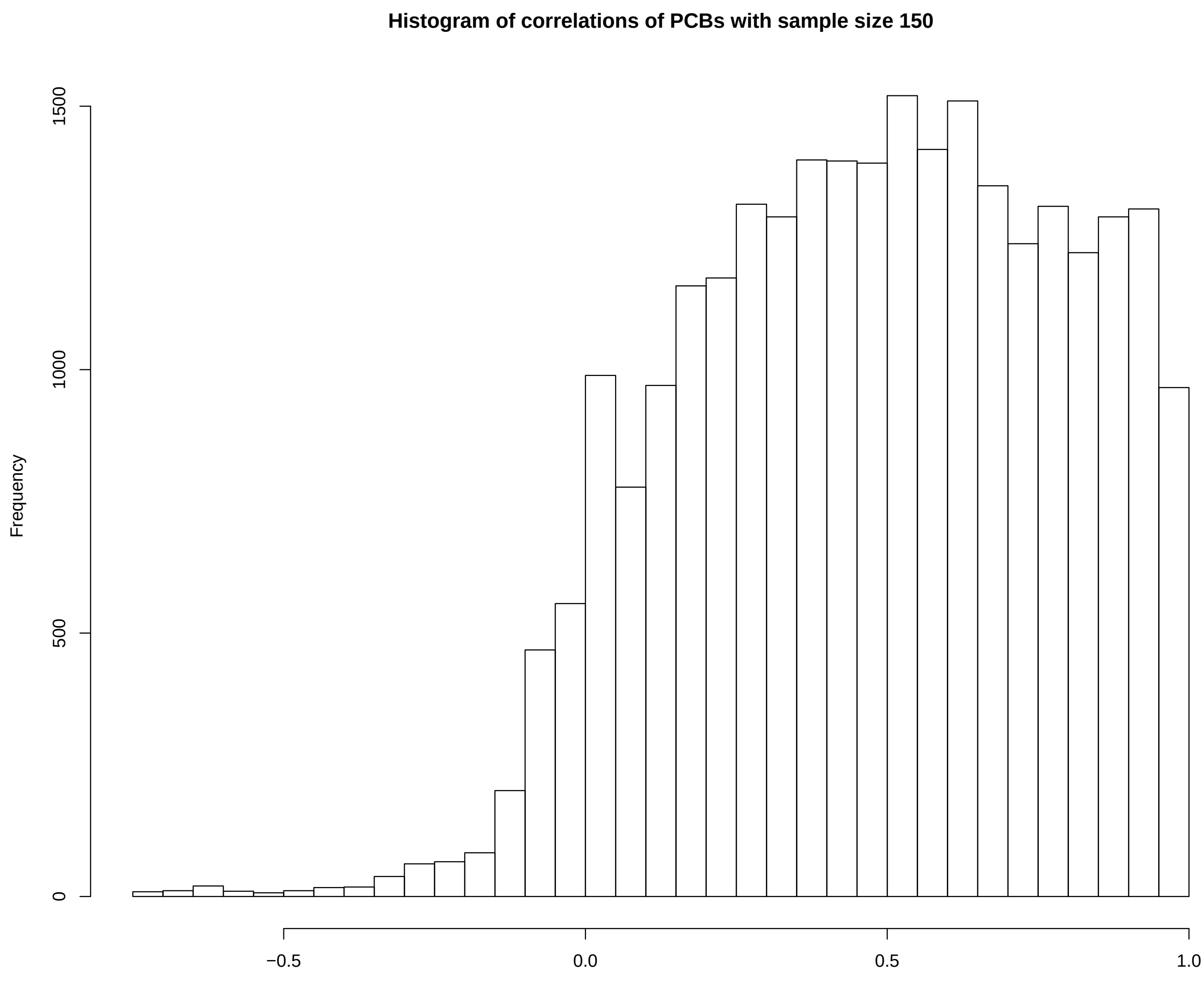
2.3 two steps decorrelation

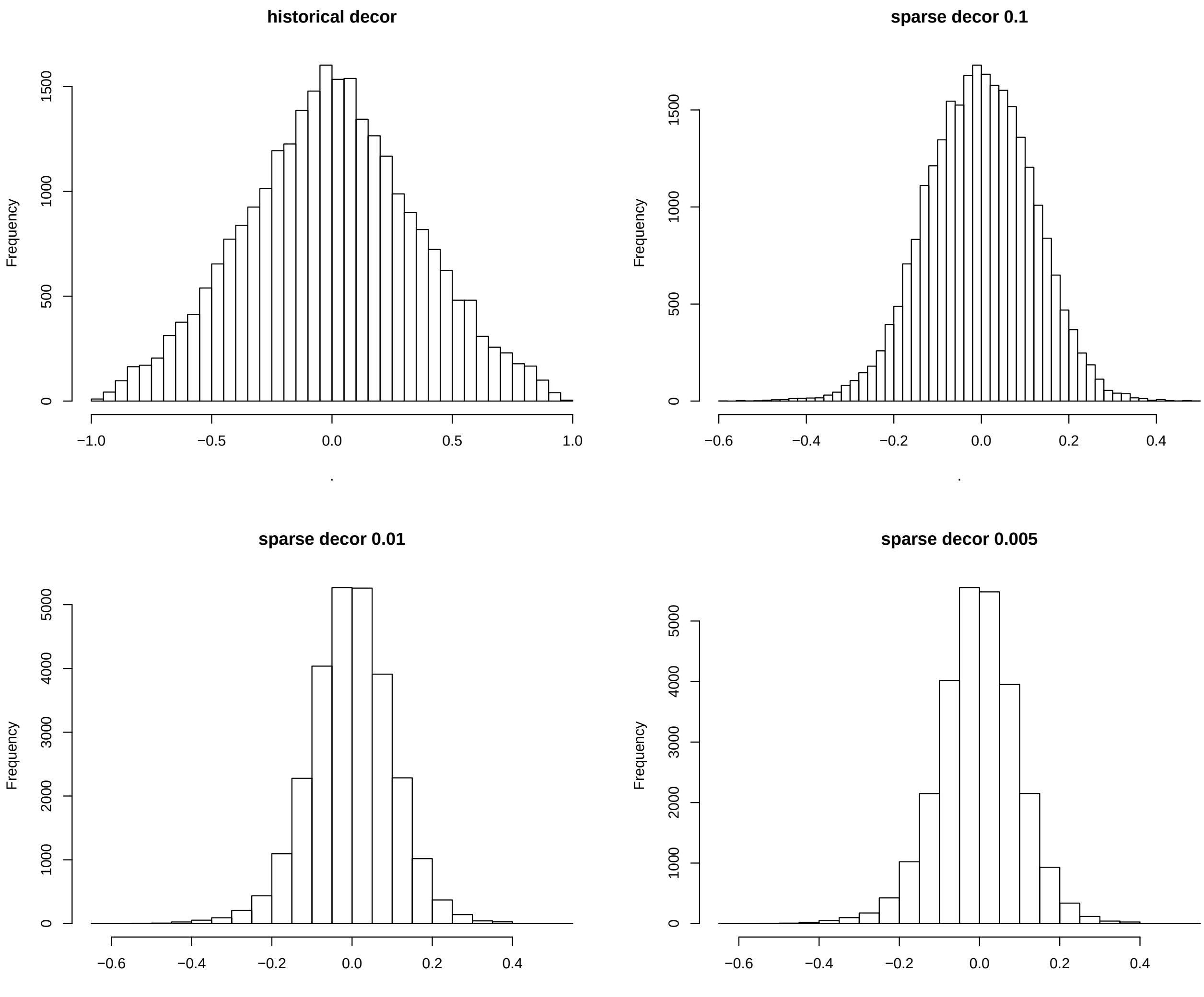
1. Decorrelation by covariance matrix estimated by historical data

2. If after the step 1, the correlation is still large, then we may need a second decorrelation by sparse precision method

2.4 Dpglasso: The Graphical Lasso: New Insights and Alternatives

1. An Alternatives for Glasso
2. Glasso works on W the covariance matrix, but its alg cannot make sure the precision matrix Θ is positive definite.
3. Dpglasso can provide both W and Θ be positive definite




historical decor
sparse decor 0.1
sparse decor 0.01
sparse decor 0.005
Frequency
0
500
1000
1500
2000
3000
4000
5000
−1.0
−0.5
0.0
0.5
1.0
−0.6
−0.4
−0.2
0.2
0.4

2.5 PCB

2.5.1 None

	var_main_effect	var_inter_effect	cov_main_inter_effect	var_total_effect
1:	8	2	0.62	11

	structure	decor	x_dist
1:	un	FALSE	1999

	n	MSE	est_var	est_mean	NA_total	method
1:	100	206	112	21	3	EigenPrism
2:	100	153	108	18	0	GCTA
3:	150	304	112	25	0	EigenPrism
4:	150	278	149	23	0	GCTA
5:	231	276	83	25	0	EigenPrism
6:	231	217	86	23	0	GCTA
7:	500	NaN	NA	NaN	100	EigenPrism
8:	500	472	166	29	0	GCTA
9:	1000	NaN	NA	NaN	100	EigenPrism
10:	1000	495	113	31	0	GCTA

2.5.2 Hist

	var_main_effect	var_inter_effect	cov_main_inter_effect	var_total_effect
1:	8	2	0.62	11

	structure	decor	x_dist
1:	un	TRUE	1999

	n	MSE	est_var	est_mean	NA_total	method
1:	100	31	32	11	3	EigenPrism
2:	100	22	22	11	0	GCTA
3:	150	26	26	11	0	EigenPrism
4:	150	21	21	11	0	GCTA
5:	231	19	20	11	0	EigenPrism
6:	231	14	14	11	0	GCTA
7:	500	NaN	NA	NaN	100	EigenPrism
8:	500	31	31	12	0	GCTA
9:	1000	NaN	NA	NaN	100	EigenPrism
10:	1000	44	39	13	1	GCTA

2.5.3 Hist + sparse

	var_main_effect	var_inter_effect	cov_main_inter_effect	var_total_effect
1:	8	2	0.62	11

	structure	decor	x_dist
1:	un	TRUE	1999

	n	MSE	est_var	est_mean	NA_total	method
1:	100	14.6	14.68	11.4	0	EigenPrism
2:	100	14.0	14.11	11.2	0	GCTA
3:	150	8.6	7.91	10.4	0	EigenPrism
4:	150	6.9	6.78	10.8	0	GCTA
5:	231	11.6	7.42	9.2	0	EigenPrism
6:	231	5.2	4.37	10.3	0	GCTA

```
 7:  500  NaN      NA      NaN      100 EigenPrism
 8:  500  3.1    1.53     10.0        0       GCTA
 9: 1000  NaN      NA      NaN      100 EigenPrism
10: 1000  2.4    0.99     10.1        1       GCTA
```

2.6 Chi

2.6.1 None

```
   var_main_effect var_inter_effect cov_main_inter_effect var_total_effect
1:               8                2                   1.8               14
   structure decor x_dist
1:        un FALSE    chi
```

```
      n MSE est_var est_mean NA_total     method
1:  100 108      67       20        3 EigenPrism
2:  100  77      68       17        0       GCTA
3:  200 213      90       25        7 EigenPrism
4:  200 189     109       23        0       GCTA
5:  500 NaN      NA      NaN      100 EigenPrism
6:  500 131      31       24        0       GCTA
7: 1000 NaN      NA      NaN      100 EigenPrism
8: 1000 134      20       24        0       GCTA
```

2.6.2 hist

```
   var_main_effect var_inter_effect cov_main_inter_effect var_total_effect
1:               8                2                   1.8               14
   structure decor x_dist
1:        un  TRUE    chi
```

```
      n  MSE est_var est_mean NA_total     method
1:  100 23.4    21.5       12        0 EigenPrism
2:  100 18.5    16.5       12        0       GCTA
3:  200 15.5    12.9       12        0 EigenPrism
4:  200  9.3     8.2       13        0       GCTA
5:  500  NaN      NA      NaN      100 EigenPrism
6:  500  4.1     2.8       13        0       GCTA
7: 1000  NaN      NA      NaN      100 EigenPrism
8: 1000  3.0     1.3       12        1       GCTA
```

2.6.3 hist + sparse

```
   var_main_effect var_inter_effect cov_main_inter_effect var_total_effect
1:               8                2                   1.8               14
   structure decor x_dist
1:        un  TRUE    chi
```

```
      n  MSE est_var est_mean NA_total     method
1:  100 27.7    27.6       14        0 EigenPrism
2:  100 26.9    27.1       14        0       GCTA
3:  200 12.4    10.7       12        0 EigenPrism
4:  200  8.9     8.8       13        0       GCTA
```

```
5:  500  NaN      NA      NaN      100 EigenPrism
6:  500  4.5     2.5       12        0       GCTA
7: 1000  NaN      NA      NaN      100 EigenPrism
8: 1000  4.1     1.8       12        0       GCTA
```

3 Mimic the histrocial covariance situation

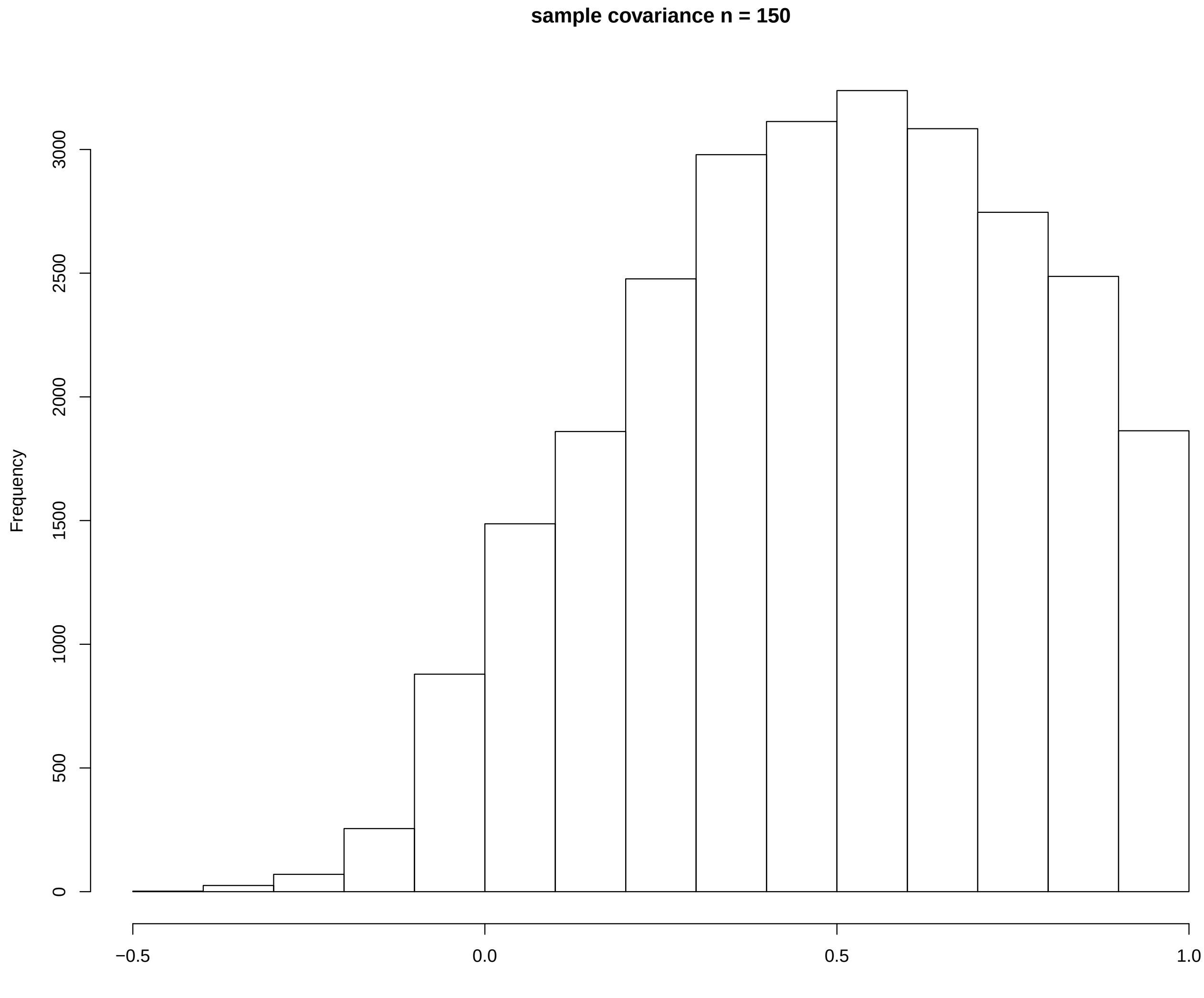
Based on the previous simulation, we found that if the historical data is from extact the same distribution, then the sparse decorrelation may not be neccessary. That is after doing the second step of decorrelation the variance estimation does not get much improvement. So to mimic the situation situation we change the simulation so that there is historical covariance is not perfect.

3.1 Simulation setup

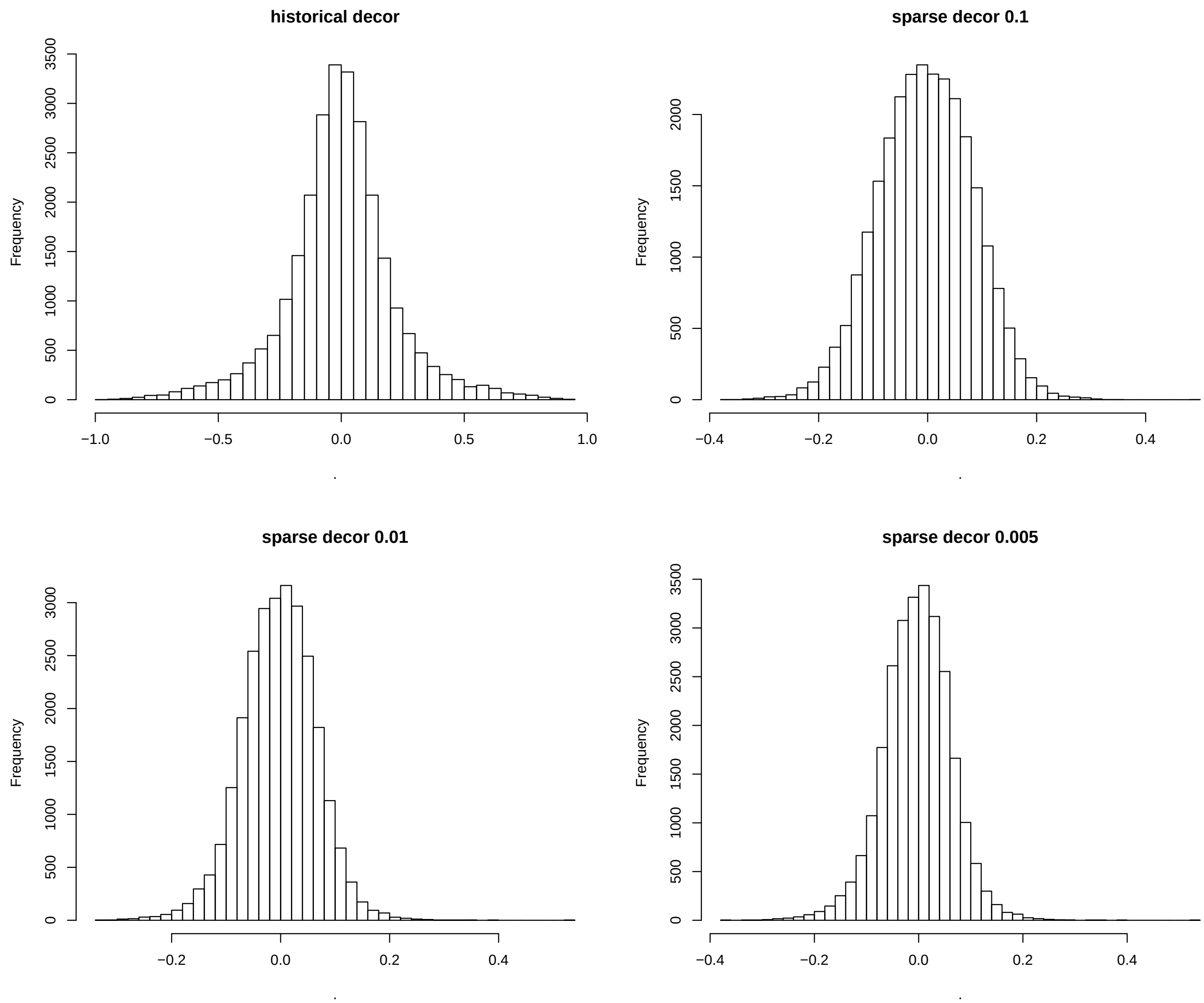
- $p = 21$
- $X \sim \chi^2_1$
- $Cor(X)$ is the sample covariance of subset of standardized PCB data with $n = 150$
- $X_h = XB$, where $B = \begin{bmatrix} 1 & 1 & 1 & 1 & 1\dots \\ 0 & 1 & 1 & 1 & 1\dots \\ 0 & & 1 & 1 & 1\dots \\ 0 & & & 1 & 1\dots \\ 0 & & & & 1\dots \end{bmatrix}$. So that the X_h is a column transfromation of X, therefore the its covariance is different from the true value.

3.2 Decorrelation result

3.2.1 $X_h = X$

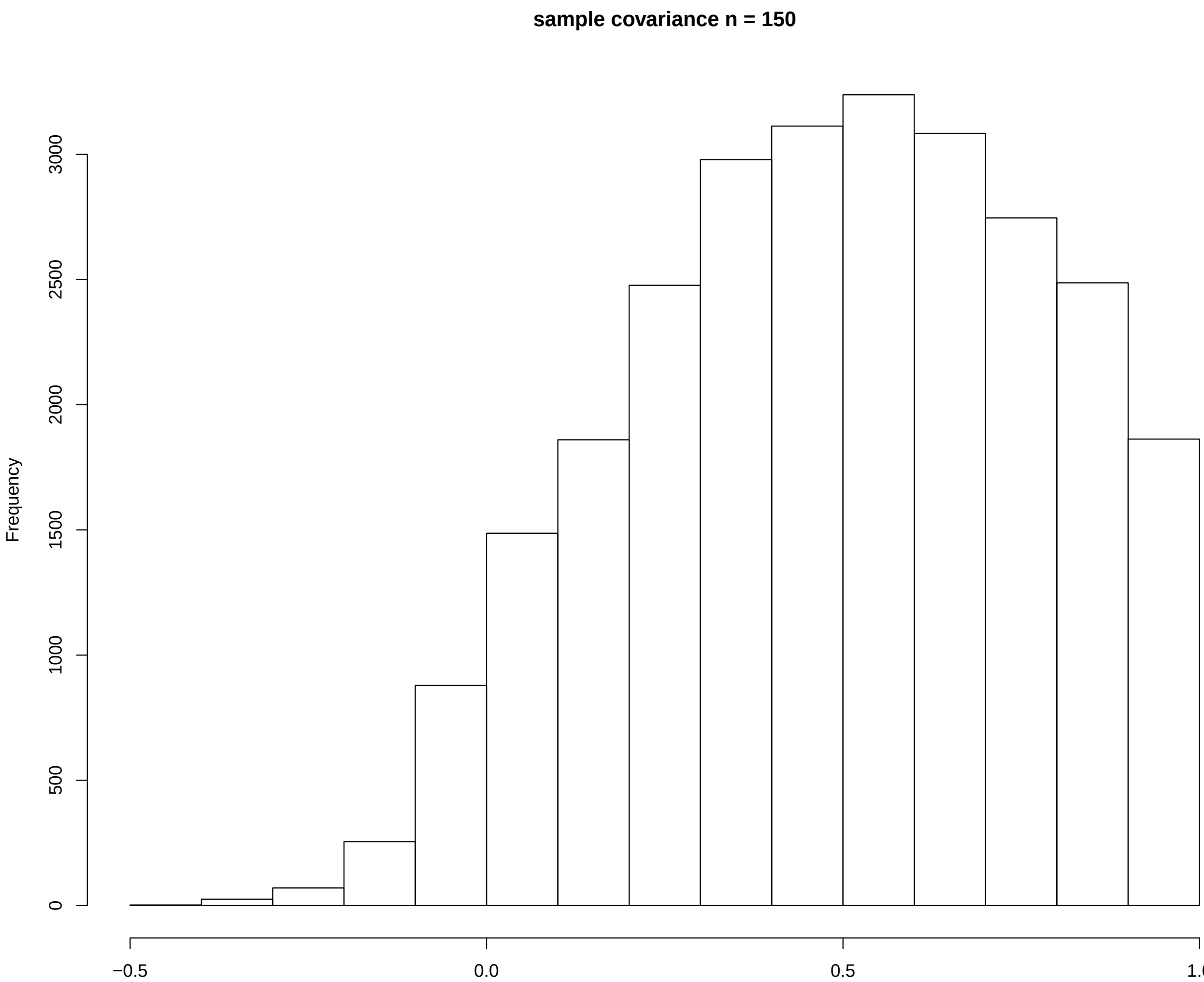



```
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-0.42    0.29    0.50    0.49    0.72    1.00

 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-0.96   -0.11    0.00    0.00    0.11    0.94

 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-0.37   -0.06    0.00    0.00    0.06    0.48

 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-0.33   -0.05    0.00    0.00    0.04    0.54
```

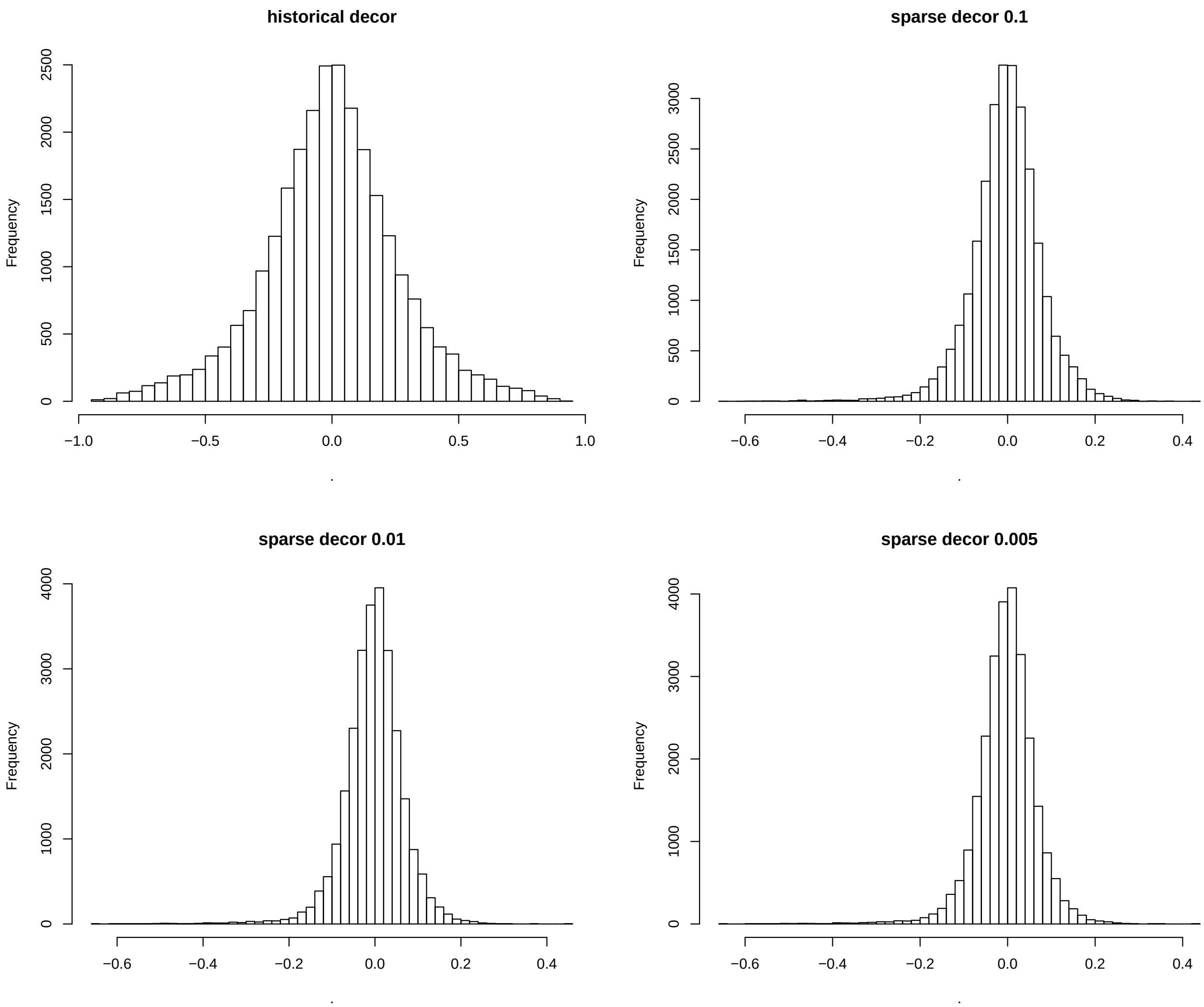


Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
-0.37	-0.04	0.00	0.00	0.04	0.53

3.2.2 $X_h = XB, t = 8$




```
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-0.42    0.29    0.50    0.49    0.72    1.00

 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-0.93   -0.15    0.00    0.00    0.15    0.91

 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-0.65   -0.04    0.00    0.00    0.04    0.42

 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-0.64   -0.04    0.00    0.00    0.04    0.44
```




```
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 -0.64   -0.04    0.00    0.00    0.03    0.44
```

3.3 Simulation result on the variance estimation process

3.3.1 None

```
   var_main_effect var_inter_effect cov_main_inter_effect var_total_effect
1:               8                2                   1.8               14
   structure decor x_dist
1:        un FALSE    chi
```

```
      n MSE est_var est_mean NA_total     method
1:  100 108      67       20        3 EigenPrism
2:  100  77      68       17        0       GCTA
3:  200 213      90       25        7 EigenPrism
4:  200 189     109       23        0       GCTA
5:  500 NaN      NA      NaN      100 EigenPrism
6:  500 131      31       24        0       GCTA
7: 1000 NaN      NA      NaN      100 EigenPrism
```

```
8: 1000 134      20       24        0       GCTA
```

3.3.2 Hist

```
   var_main_effect var_inter_effect cov_main_inter_effect var_total_effect
1:               8                2                   1.8               14
   structure decor x_dist
1:        un  TRUE    chi

      n MSE est_var est_mean NA_total     method
1:  100  32    27.3       11        0 EigenPrism
2:  100  25    18.1       11        0       GCTA
3:  200  29    22.8       11        0 EigenPrism
4:  200  19    13.4       11        0       GCTA
5:  500 NaN      NA      NaN      100 EigenPrism
6:  500  14     3.5       10        0       GCTA
7: 1000 NaN      NA      NaN      100 EigenPrism
8: 1000  15     2.9       10        0       GCTA
```

3.3.3 Hist + Sparse(0.1)

```
   var_main_effect var_inter_effect cov_main_inter_effect var_total_effect
1:               8                2                   1.8               14
   structure decor x_dist
1:        un  TRUE    chi

      n  MSE est_var est_mean NA_total     method
1:  100 19.8    20.0       14        0 EigenPrism
2:  100 22.0    22.0       13        0       GCTA
3:  200 13.6     8.5       11        0 EigenPrism
4:  200  7.7     6.9       13        0       GCTA
5:  500  NaN      NA      NaN      100 EigenPrism
6:  500  6.0     1.5       12        0       GCTA
7: 1000  NaN      NA      NaN      100 EigenPrism
8: 1000 10.1     0.9       11        1       GCTA
```

3.3.4 Hist + Sparse(0.01)

```
   var_main_effect var_inter_effect cov_main_inter_effect var_total_effect
1:               8                2                   1.8               14
   structure decor x_dist
1:        un  TRUE    chi

      n  MSE est_var est_mean NA_total     method
1:  100 39.5   39.37       14        0 EigenPrism
2:  100 22.5   22.69       13        0       GCTA
3:  200 13.7    8.43       11        0 EigenPrism
4:  200  7.5    6.19       13        0       GCTA
5:  500  NaN      NA      NaN      100 EigenPrism
6:  500  7.5    1.43       11        0       GCTA
7: 1000  NaN      NA      NaN      100 EigenPrism
8: 1000 10.6    0.93       11        1       GCTA
```

3.3.5 Hist + Sparse(0.001)

	var_main_effect	var_inter_effect	cov_main_inter_effect	var_total_effect
1:	8	2	1.8	14

	structure	decor	x_dist
1:	un	TRUE	chi

	n	MSE	est_var	est_mean	NA_total	method
1:	100	74.2	67.14	16	0	EigenPrism
2:	100	32.0	32.36	14	0	GCTA
3:	200	12.5	8.37	12	0	EigenPrism
4:	200	7.4	6.25	13	0	GCTA
5:	500	NaN	NA	NaN	100	EigenPrism
6:	500	7.7	1.43	11	0	GCTA
7:	1000	NaN	NA	NaN	100	EigenPrism
8:	1000	10.4	0.96	11	0	GCTA

3.3.6 Perfect Hist

	var_main_effect	var_inter_effect	cov_main_inter_effect	var_total_effect
1:	8	2	1.8	14

	structure	decor	x_dist
1:	un	TRUE	chi

	n	MSE	est_var	est_mean	NA_total	method
1:	100	23.4	21.5	12	0	EigenPrism
2:	100	18.5	16.5	12	0	GCTA
3:	200	15.5	12.9	12	0	EigenPrism
4:	200	9.3	8.2	13	0	GCTA
5:	500	NaN	NA	NaN	100	EigenPrism
6:	500	4.1	2.8	13	0	GCTA
7:	1000	NaN	NA	NaN	100	EigenPrism
8:	1000	3.0	1.3	12	1	GCTA

3.3.7 Perfect Hist + Sparse(0.1)

	var_main_effect	var_inter_effect	cov_main_inter_effect	var_total_effect
1:	8	2	1.8	14

	structure	decor	x_dist
1:	un	TRUE	chi

	n	MSE	est_var	est_mean	NA_total	method
1:	100	27.7	27.6	14	0	EigenPrism
2:	100	26.9	27.1	14	0	GCTA
3:	200	12.4	10.7	12	0	EigenPrism
4:	200	8.9	8.8	13	0	GCTA
5:	500	NaN	NA	NaN	100	EigenPrism
6:	500	4.5	2.5	12	0	GCTA
7:	1000	NaN	NA	NaN	100	EigenPrism
8:	1000	4.1	1.8	12	0	GCTA

3.3.8 Perfect Hist + Sparse(0.01)

```
   var_main_effect var_inter_effect cov_main_inter_effect var_total_effect
1:               8                2                   1.8               14
   structure decor x_dist
1:        un  TRUE    chi
```

```
      n  MSE est_var est_mean NA_total     method
1:  100 59.9    54.4       11        3 EigenPrism
2:  100 60.5    59.7       13        0       GCTA
3:  200 13.6    13.0       15        0 EigenPrism
4:  200 12.4    12.4       14        0       GCTA
5:  500  NaN      NA      NaN      100 EigenPrism
6:  500  3.6     3.1       13        0       GCTA
7: 1000  NaN      NA      NaN      100 EigenPrism
8: 1000  2.9     2.1       13        0       GCTA
```

3.3.9 Perfect Hist + Sparse(0.001)

```
   var_main_effect var_inter_effect cov_main_inter_effect var_total_effect
1:               8                2                   1.8               14
   structure decor x_dist
1:        un  TRUE    chi
```

```
      n   MSE est_var est_mean NA_total     method
1:  100 148.2   146.6       15        0 EigenPrism
2:  100  76.4    76.7       13        0       GCTA
3:  200  32.2    27.3       16        3 EigenPrism
4:  200  12.5    12.4       14        0       GCTA
5:  500   NaN      NA      NaN      100 EigenPrism
6:  500   3.8     3.1       13        0       GCTA
7: 1000   NaN      NA      NaN      100 EigenPrism
8: 1000   2.7     2.4       13        0       GCTA
```