Jackknife variance estimation corrections

Xuelong Wang

2019-11-01

Contents

1 Motivation

Jackknife is one of sub-sampling technique to estimate the bias and variance of a statistics $S(X_1, \ldots, X_n)$. Define $S_{(i)} = S(X_1, X_{(i-1)}, X_{(i+1)}X_n)$, then the variance of S estimated by jackknife is

$$Var(S) = \frac{n-1}{n} \sum_i (S_{(i)} - S_.)^2$$

Where $S_. = \frac{1}{n} \sum_i S_{(i)}$. Note that $\tilde{Var}(S(X_1, \ldots, S_{n-1})) = \sum_i (S_{(i)} - S_.)^2$ can be conisder as a estimation for $Var(S(X_1, \ldots, S_{n-1}))$ However, Eforn in 1981 shown that the jackknife variance esimtation is always overestimate the true variance.

$$E(\tilde{Var}(S(X_1, \ldots, S_{n-1}))) \geq Var(S(X_1, \ldots, S_{n-1}))$$

2 One solution

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

3 Simulation study

3.1 without correction

3.1.1 setup

- Independent
- Normal
- p = 21
- with interaction terms

3.1.2 simulation_result

	method	n	est_mean	est_var	var_jack
1:	EigenPrism	100	9.37	17.22	27.81
2:	EigenPrism	150	10.20	7.68	16.40
3:	EigenPrism	231	10.21	5.12	10.12
4:	EigenPrism	500	NaN	NA	NaN
5:	GCTA	100	8.78	17.70	38.81
6:	GCTA	150	9.58	10.20	19.54
7:	GCTA	231	9.69	5.07	9.30
8:	GCTA	500	10.07	2.25	2.85

3.1.3 setup

- Independent
- Normal
- p = 22
- with interaction terms

3.1.4 simulation_result

	method	n	est_mean	est_var	var_jack
1:	EigenPrism	100	9.76	20.26	28.53
2:	EigenPrism	253	10.77	5.61	9.08
3:	EigenPrism	500	NaN	NA	NaN
4:	EigenPrism	600	NaN	NA	NaN
5:	EigenPrism	700	NaN	NA	NaN
6:	GCTA	100	8.76	24.20	40.05
7:	GCTA	253	10.44	5.60	8.44
8:	GCTA	500	10.11	1.53	3.03
9:	GCTA	600	10.31	1.39	2.23
10:	GCTA	700	10.16	1.10	1.61

Note that based on the ideal situation the bias will reduced when sample size is large and distribution is normal.

3.2 correction

3.2.1 setup

- Independent
- Normal
- p = 21
- with interaction terms

3.2.2 simulation result

```
   var_main_effect var_inter_effect cov_main_inter_effect var_total_effect
1:               8                2                0.0123               10
   structure decor x_dist
1:         I FALSE normal
```

```
     n   MSE   est est_jack   var v_jack v_jack_c  v_Eg v_jack_diff
1: 100 17.48  9.37     9.92 17.22   27.8    12.31 22.09       20.06
2: 231  5.11 10.21    10.47  5.12   10.1     2.12  7.81        6.12
3: 100 19.08  8.78    10.03 17.70   38.8    -3.48    NA       17.67
4: 231  5.13  9.69    10.28  5.07    9.3     2.92    NA        6.11
       method
1: EigenPrism
2: EigenPrism
3:       GCTA
4:       GCTA
```

3.2.3 setup

- Independent
- Chi
- p = 21
- with interaction terms

3.2.4 simulation result

```
   var_main_effect var_inter_effect cov_main_inter_effect var_total_effect
1:               8                2                0.0128               10
   structure decor x_dist
1:         I FALSE    chi
```

```
     n   MSE  est est_jack   var v_jack v_jack_c  v_Eg v_jack_diff
1: 100 15.94 10.7     10.3 15.64  23.23    13.15 21.34       18.19
2: 231  6.28 10.5     10.4  6.07  10.01     3.29  8.34        6.65
3: 100 16.27 10.5     10.3 16.19  23.79     9.97    NA       16.88
4: 231  5.55 10.4     10.4  5.48   7.83     4.05    NA        5.94
       method
1: EigenPrism
2: EigenPrism
3:       GCTA
4:       GCTA
```

3.2.5 setup

- PCB
- p = 21
- with interaction terms
- with decorrelation

3.2.6 simulation result

```
   var_main_effect var_inter_effect cov_main_inter_effect var_total_effect
1:               8                2                 0.622             11.2
   structure decor x_dist
1:        un FALSE   1999
```

```
     n MSE est est_jack var v_jack v_jack_c v_Eg v_jack_diff     method
1: 100 206  21       NA 112    208     61.1   NA         134 EigenPrism
2: 100 153  18     23.4 108    236   -757.9   NA        -261       GCTA
```

4 Questions and other methods

4.1 Questions

- Running time is large $n * (n-1)/2$
- Assumptions: quadratic form of S, $Var^n = \frac{n-1}{n} Var^{n-1}$?
- The coefficient about the correction

4.2 Other methods

5 Some modification based on previous simulation

5.1 Main effect only or larger n for combined effect

5.1.1 Normal main n>p

```
   var_main_effect var_inter_effect cov_main_inter_effect var_total_effect
1:               8                0                     0                0
   structure decor x_dist
1:         I FALSE normal
```

```
     n  MSE  est est_jack  var v_jack v_jack_c v_Eg v_jack_diff     method
1: 100  NaN  NaN       NA   NA     NA       NA   NA          NA EigenPrism
2: 231  NaN  NaN       NA   NA     NA       NA   NA          NA EigenPrism
3: 100 3.79 7.66     7.69 3.71   5.03  -18.546   NA       -6.76       GCTA
4: 231 1.11 7.89     7.91 1.10   1.55   -0.906   NA        0.32       GCTA
```

```
   var_main_effect var_inter_effect cov_main_inter_effect var_total_effect
1:               8                0                     0                0
   structure decor x_dist
1:         I FALSE normal
```

	n	MSE	est	est_jack	var	v_jack	v_jack_c	v_Eg	v_jack_diff	method
1:	100	NaN	NaN	NA	NA	NA	NA	NA	NA	EigenPrism
2:	231	NaN	NaN	NA	NA	NA	NA	NA	NA	EigenPrism
3:	50	NaN	NaN	NA	NA	NA	NA	NA	NA	EigenPrism
4:	100	3.78	7.66	7.66	3.70	4.88	3.18	NA	4.03	GCTA
5:	231	1.11	7.88	7.88	1.10	1.54	1.33	NA	1.43	GCTA
6:	50	9.22	7.55	7.42	9.11	16.10	-4.51	NA	5.80	GCTA

5.1.2 chi n >p

	var_main_effect	var_inter_effect	cov_main_inter_effect	var_total_effect
1:	8	0	0	0

	structure	decor	x_dist
1:	I	FALSE	chi

	n	MSE	est	est_jack	var	v_jack	v_jack_c	v_Eg	v_jack_diff	method
1:	100	NaN	NaN	NA	NA	NA	NA	NA	NA	EigenPrism
2:	231	NaN	NaN	NA	NA	NA	NA	NA	NA	EigenPrism
3:	100	4.16	8.30	7.53	4.11	6.42	-14.494	NA	-4.038	GCTA
4:	231	1.27	7.89	7.78	1.27	2.10	-0.489	NA	0.806	GCTA

5.1.3 Normal main n<p, p = 100

	var_main_effect	var_inter_effect	cov_main_inter_effect	var_total_effect
1:	8	0	0	0

	structure	decor	x_dist
1:	I	FALSE	normal

	n	MSE	est	est_jack	var	v_jack	v_jack_c	v_Eg	v_jack_diff	method
1:	50	21.65	8.37	7.63	21.73	48.10	12.675	33.5	30.39	EigenPrism
2:	50	25.58	8.02	8.52	25.84	74.13	-190.666	NA	-58.27	GCTA
3:	100	7.12	7.96	7.06	7.19	15.11	1.977	11.7	8.54	EigenPrism
4:	100	6.25	7.83	7.45	6.29	13.74	-0.966	NA	6.38	GCTA
5:	200	NaN	NaN	NA	NA	NA	NA	NaN	NA	EigenPrism
6:	200	2.45	7.88	7.59	2.46	4.64	-214.514	NA	-104.94	GCTA

	var_main_effect	var_inter_effect	cov_main_inter_effect	var_total_effect
1:	8	0	0	0

	structure	decor	x_dist
1:	I	FALSE	normal

	n	MSE	est	est_jack	var	v_jack	v_jack_c	v_Eg	v_jack_diff
1:	50	21.65	8.37	7.63	21.73	48.10	12.675	33.5	30.39
2:	50	25.58	8.02	8.49	25.84	73.09	-168.830	NA	-47.87
3:	100	7.12	7.96	7.06	7.19	15.11	1.977	11.7	8.54
4:	100	6.25	7.83	7.46	6.29	13.68	-0.742	NA	6.47
5:	200	NaN	NaN	NA	NA	NA	NA	NaN	NA
6:	200	2.48	7.89	7.95	2.49	3.92	1.373	NA	2.65

```
       method
1: EigenPrism
2:       GCTA
3: EigenPrism
4:       GCTA
5: EigenPrism
6:       GCTA
```

5.1.4 Chi combined effect n <p p = 25

```
   var_main_effect var_inter_effect cov_main_inter_effect var_total_effect
1:               8                2                0.0332             10.1
   structure decor x_dist
1:         I FALSE    chi
```

```
     n   MSE   est est_jack   var v_jack v_jack_c  v_Eg v_jack_diff
1: 100 17.29 10.24    10.58 17.43  24.07    14.52 22.02       19.30
2: 150 11.00 10.69    10.72 10.72  15.32     8.16 13.62       11.74
3: 325  4.15 10.69    11.00  3.80   7.34     2.46  5.94        4.90
4: 100 20.82  9.45     9.64 20.64  27.24    10.72    NA       18.98
5: 150 10.09 10.15    10.21 10.19  16.37     8.42    NA       12.39
6: 325  3.45 10.32    10.77  3.43   5.90     3.17    NA        4.54
       method
1: EigenPrism
2: EigenPrism
3: EigenPrism
4:       GCTA
5:       GCTA
6:       GCTA
```

5.1.5 PCB combined effec n<p p = 21

```
   var_main_effect var_inter_effect cov_main_inter_effect var_total_effect
1:               8                2                 0.622             11.2
   structure decor x_dist
1:        un  TRUE   1999
```

```
     n  MSE  est est_jack  var v_jack v_jack_c v_Eg v_jack_diff     method
1: 100 14.6 11.4     9.92 14.7   34.1     16.9 18.6        25.5 EigenPrism
2: 100 14.0 11.2    11.58 14.1   38.1   -341.9   NA      -151.9       GCTA
```

5.1.6 PCB combined effec n<p p = 21 with rank transformation

```
   var_main_effect var_inter_effect cov_main_inter_effect var_total_effect
1:               8                2                 0.207             10.4
   structure decor x_dist
1:        un  TRUE   1999
```

```
     n  MSE   est est_jack  var v_jack  v_jack_c v_Eg v_jack_diff
1: 100 12.9 11.52     10.4 11.8   16.9      7.63 18.5        12.3
2: 100 18.1  8.62     26.2 15.0  211.0 -26482.62   NA    -13135.8
       method
1: EigenPrism
```

```
2:       GCTA
```

5.1.7 PCB combined effec n<p p = 21 with rank transformation for all

```
   var_main_effect var_inter_effect cov_main_inter_effect var_total_effect
1:               8                2                  2.76             15.5
   structure decor x_dist
1:        un  TRUE   1999
```

```
     n   MSE   est est_jack  var v_jack v_jack_c v_Eg v_jack_diff
1: 100  30.4 17.72     13.9 25.8   35.4       17 41.5        26.2
2: 100 174.2  4.93    -40.1 62.8 1651.2  -224762   NA   -111555.4
       method
1: EigenPrism
2:       GCTA
```