**Estimation for correlation from GWAS summary statistics: when treated as random effects.**

Recently, some studies try to estimate correlation between a pair of test statistics for GWAS summary statistics. Often, the correlation is estimated from z-score test statistics or genetic effects [1, 2]. However, it is unclear the interpretation for the correlation. As a proof-of-principle, we derive the structure of the correlation estimated from summary statistics, and show that it is a mixed distribution for at least two components: the part due to genetic architecture, and the one due to overlapping samples.

Furthermore, caveats should be taken into account when using the estimated correlation matrix for conducting meta-analysis: it may inflate the type-I error rate.

|  |  |
| --- | --- |
| **Notations** | **Definitions** |
|  | Phenotypic correlation for a pair of traits for two cohorts. |
|  | , is the correlation due to overlapping samples. is the number of overlapping samples, and are samples sizes for two cohorts. |
|  | Genetic correlation. for the same trait. |
|  | The correlation estimated from summary statistics, and |
|  | The correlation component due to overlapping samples for the estimated correlation from summary statistics. |
|  | The correlation component due to genetic component for the estimated correlation from summary statistics. The upper bound for . |
|  | , the proportion of the correlation due to overlapping samples. |
|  | The mean of the variance for the markers. |
|  | The mean of the variance for the QTLs |

Pink ones are real parameters that are not observable; cyan ones are estimates for corresponding parameters.

**Results for the correlation of summary statistics**

|  |  |  |  |
| --- | --- | --- | --- |
| Source | Different traits | Correlation |  |
| (genetic effects) | Yes |  |  |
|  | No |  |  |
| (test statistics) | Yes |  |  |
|  | No |  |  |

Notes: 1) when convert the into a linear regression, it becomes the recent proposed LD regression for summary stats.

2) the technical used here for deriving correlation is based on covariance structure for mixture distribution, as exercised in [my comment for “Nature Selection and Friendship”](http://journal.frontiersin.org/Journal/10.3389/fgene.2014.00400/full).

**Results for Meta-analysis**

**Method I: Multi-variants chisq**

**The type I error rate**

The type I error rate will be well controlled if

For the same trait ,

When , no overlapping samples, , the test statistic is inflated.

When , no heritability (), , , not biased.

When and , , which is naïve meta-analysis.

When ignoring , regardless of its value, , but its sampling variance will be inflated

The NCP is

When

**Method II: Squared inverse-variance-weighted meta-analysis**

The type I error rate will be controlled if

When , , and

It can be further simplified if

When , , because when heritability is zero.

When but , , , the type I error rate will be deflated.

When but , , the type I error rate will be deflated.

When ignoring , regardless it is zero no not, it is naïve meta-analysis,, no inflation if there is no overlapping samples, otherwise inflation due to overlapping samples.

The NCP is

If , and ,

When

**Conclusion:**

1) Regardless of whether using multi-variant chisq test or inverse variance weighted meta-analysis, it is not suggested to use the correlation matrix that is estimated by treating summary statistics as random variables. In meta-analysis, it estimates each effect as fixed effects, but using the matrix treating them as random variables does not fit the context.

2) when the sample sizes for two cohorts are same, the NCP for multi-variate chisq is the same to inverse-variance test.

As a proof-of-principle, we assume the genetic effect is estimated from the single-marker regression,

, . Under a typical polygenic model, , in which is the number of QTLs, and as , , and

as , , under a typical polygenic model, , in which is the number of QTLs. Assume summary statistics are from a pair of cohorts, which have and samples. In addition, there are overlapping samples between the cohorts.

**Method 1: Correlation based on genetic effects**

Depending on whether the locus is causal or not, scores can be split into two distributions

Now, for a pair of cohorts

As we only discuss for the same trait, and .

For a mixture distribution, the variance is

and the covariance is

in which and .

For the case above, as ,

We have

in which is the overlapping samples between two cohorts, and is the mean of the variance for the QTLs, for all markers (including QTLs), , and are genetic correlation and phenotypic correlation for the traits, and .

The correlation between two z-score sequences is

**(Equation 1)**

The correlation can be partitioned into two components

in which

**(Equation 2)**

represents the correlation related to heritability,

**(Equation 3)**

represents the correlation related to overlapping samples.

The proportion due to overlapping sample is

**(Equation 4)**

when ,

and

**Method 2: Correlation from scores**

In addition, the correction can also be estimated from z-scores for the pair of cohorts. Similarly, depending on whether the locus is causal or not, scores can be split into two distributions. Now, for a pair of cohorts

For a mixture distribution, its variance is

and the covariance is

in which and .

For the case above, as ,

We have

in which is the overlapping samples between two cohorts.

The correlation between two z-score sequences is

**(Equation 5)**

As the overlapping samples are confounded with genetic architecture, it can be decomposed into two components

**(Equation 6)**

**Remark 1:** The correlation will be unit if the sample sizes of both cohorts go infinite as long as .

**Remark 2:** When , . It shows some connection to the prediction theory[3, 4], the prediction accuracy is .

The correlation part due to overlapping sample is

**(Equation 7)**

**Remark 3:** the estimated correlation due to overlapping samples is not greater than . When heritability is zero, and the traits are same, .

**Remark 4:**  increases with the number of markers, and has its upper bound .

The correlation can be split into two components: due to heritability and due to overlapping samples. The proportion due to overlapping samples are

**(Equation 8)**

The proportion of the correlation raised by overlapping samples.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |  |  |  |
| 1000 | 1000 | 100 | 0.25 | 30000 | 1000 | 0.1 | 0.0083 | 0.0992 | 0.1075 |
| 1000 | 2000 | 100 | 0.25 | 30000 | 1000 | 0.0707 | 0.0116 | 0.0698 | 0.0814 |
| 1000 | 5000 | 100 | 0.25 | 30000 | 1000 | 0.0447 | 0.0182 | 0.0436 | 0.0618 |
| 1000 | 10000 | 100 | 0.25 | 30000 | 1000 | 0.0316 | 0.0252 | 0.0303 | 0.0555 |
|  |  |  |  |  |  |  |  |  |  |
| 1000 | 1000 | 100 | 0 | 30000 | 1000 | 0.1 | 0 | 0.1 | 0.1 |
| 1000 | 2000 | 100 | 0 | 30000 | 1000 | 0.0707 | 0 | 0.0707 | 0.0707 |
| 1000 | 5000 | 100 | 0 | 30000 | 1000 | 0.0447 | 0 | 0.0447 | 0.0447 |
| 1000 | 10000 | 100 | 0 | 30000 | 1000 | 0.0316 | 0 | 0.0316 | 0.0316 |

It seems clear that the contribution of the overlapping samples to the proportion of is determined by many factors, such as sample sizes, overlapping samples, and heritability, as well as genetic architecture. The correlation between test statistics between a pair of cohorts is not a good choice if the purpose is to detect overlapping samples.

When there is no overlapping samples,

, and subtracting 1 from each element at the denominator , which is the unbiased estimated of the genetic correlation.

Method 3: correlation from chi-sq

, and . , and . , and .

and

**Using the estimated correlation for GWAS meta-analysis**

**Method I: multi-variants chi-square test**

A chisq test for combining traits has be used to deliver an overall test for association [1].

in which is a vector and is a correlation matrix.

As a proof-of-principle analysis, we assume for a pair of cohort for the same trait, and the test is for a target marker

and furthermore, , in which (NCP) is the correlation between the marker and the trait .

**(Equ 9)**

Three parts, baseline for chisq test, NCP, and overlapping samples. , chi-square test with 2-degree of freedom.

**Type I error rate**

For a null locus, , the test statistics is

We can define , in which is the degree of freedom

For the same trait ,

When , no overlapping samples, , the test statistic is inflated.

When , no heritability (), , , not biased.

When and , , which is naïve meta-analysis.

When ignoring , regardless of its value, , but its sampling variance will be inflated

For non-centrality parameter

When

**Method II: Squared Invert-variance estimate (See Zhu, AJHG, 2015, 96:21-36)**

For convenience, assume meta-analysis for a pair of cohorts, which have the same sample size.

The correlation matrix is , and its inverse is .

The weights is calculated as .

The estimate of the effect is

The sampling variance of the estimate is

And the t-test is

and , which follows .

in which the two terms represent NCP and the baseline part of the test.

**The type I error rate should be evaluated on the baseline part**

Under the null hypothesis

Similarly, we can define .

If and ,

Under a much simpler case that ,

When , , because when heritability is zero.

When but , , , the type I error rate will be deflated.

When but , , the type I error rate will be deflated.

When ignoring , regardless it is zero no not, it is naïve meta-analysis,, no inflation if there is no overlapping samples, otherwise inflation due to uncorrected inflation due to overlapping samples.

It corrects for such as overlapping samples, as the impact of overlapping samples are same to all loci,

When , it corrects for heterogeneity, but I so far can’t think of a scenario under which heterogeneity equally influences cohorts as overlapping samples do.

**NCP**

Depending on the genetic architecture, the estimated genetic effect varies.

If , and ,

When

References

1. Bolormaa S, Pryce JE, Reverter A, Zhang Y, Barendse W, Kemper K, Tier B, Savin K, Hayes BJ, Goddard ME: **A multi-trait, meta-analysis for detecting pleiotropic polymorphisms for stature, fatness and reproduction in beef cattle.** *PLoS Genet* 2014, **10**:e1004198.

2. Zhu X, Feng T, Tayo BO, Liang J, Young JH, Franceschini N, Smith JA, Yanek LR, Sun Y V., Edwards TL, Chen W, Nalls M, Fox E, Sale M, Bottinger E, Rotimi C, Liu Y, McKnight B, Liu K, Arnett DK, Chakravati A, Cooper RS, Redline S: **Meta-analysis of Correlated Traits via Summary Statistics from GWASs with an Application in Hypertension**. *Am J Hum Genet* 2015, **96**:21–36.

3. Daetwyler HD, Villanueva B, Woolliams J a: **Accuracy of predicting the genetic risk of disease using a genome-wide approach.** *PLoS One* 2008, **3**:e3395.

4. Dudbridge F: **Power and predictive accuracy of polygenic risk scores.** *PLoS Genet* 2013, **9**:e1003348.