Heritability Estimation and Risk Prediction in Schizophrenia

Choi Shing Wan

A thesis submitted in partial fulfillment of the requirements for the Degree of Doctor of Philosophy

Department of Psychiatry
University of Hong Kong
Hong Kong

September 3, 2015

Declaration

Acknowledgements

Abbreviations

GWAS Genome Wide Association Study. 7, 8

LD Linkage Disequilibrium. 8, 10

PGS Polygenic Risk Score. 17

SCZ Schizophrenia. 13

SNP Single Nucleotide Polymorphism. 7–9

SVD Singular Value Decomposition. 11

tSVD Truncated Singular Value Decomposition. 11

Contents

Introduction

Some considerations

1. PRSice requires the phenotype to aid its selection (More information= stronger)
2. It seems like LDSC doesn't necessary perform badly in oligogenic situation. Rather, it is that when the trait is oligogenic, it is more likely for LDSC to behaviour in a strange way.
3. For each condition: extreme phenotype, quantitative trait, case control, we can have a separated review. Discuss on the benefits and challenges of each condition and the method we deal with them. So we can have two chapters (case control, quantitative trait) where extreme phenotype can be a big subsection within quantitative trait.
4. For each chapter, there will be this introduction (review on the method), our methodology (Calculation, implementation and also simulation), result (the simulation result). Then we can have the application (PGC, network)

Chapter 1

Literature Review

1.1 Twin Studies

Should briefly talk about how Twin modeling was used for finding the GE contribution. Should also mention the ACE model. At the end, we can talk about the heritability estimates of SCZ and AD

1.2 Searching for Genetic Variants

1.2.1 Role of Common Variants

Genome Wide Association Study

Should talk about what is GWAS and how it is used. Should also talk about the current GWAS studies in SCZ and AD

1.2.2 Role of Rare Variants

Exome Sequencing

Similar to the GWAS. Talk about the Pros and Cons. Need to briefly mention the Denovo paper and Shaun's paper.

Whole Genome Sequencing

Very very brief description of WGS and the current status.

1.3 Narrow Sense Heritability

1.4 Risk Prediction

1.5 Summary

Chapter 2

Heritability Estimation

This chapter should be used in similar way as the general method section in Clara's thesis. Considering that the subsequent chapters all rely on this implementation.

2.1 Introduction

2.2 Methodology

The work in this chapter were done in collaboration with my colleagues who have kindly provide their support and knowledges to make this piece of work possible.

2.2.1 Heritability Estimation

The narrow-sense heritability is defined as

$$h^2 = \frac{\mathrm{Var}(X)}{\mathrm{Var}(Y)}$$

where $\mathrm{Var}(X)$ is the variance of the genotype and $\mathrm{Var}r(Y)$ is the variance of the phenotype. In a Genome Wide Association Study (GWAS), regression were performed between the Single Nucleotide Polymorphisms (SNPs) and the phenotypes, giving

$$Y = \beta X + \epsilon \tag{2.1}$$

where Y and X are the standardized phenotype and genotype respectively. ϵ is then the error term, accounting for the non-genetic elements contributing to the phenotype (e.g. Environment factors). Based on eq. (2.1), one can then have

$$\begin{aligned}\mathrm{Var}(Y) &= \mathrm{Var}(\beta X) + \mathrm{Var}(\epsilon) \\ \mathrm{Var}(Y) &= \beta^{\mathrm{Var}}(X) \\ \beta^2\frac{\mathrm{Var}(X)}{\mathrm{Var}(Y)} &= 1 \qquad (2.2)\end{aligned}$$

β^2 is then considered as the portion of phenotype variance explained by the variance of genotype, which can also be considered as the narrow-sense heritability of the phenotype.

A challenge in calculating the heritability from GWAS data is that usually only the test-statistic or p-value were provided and one will not be able to directly calculate the heritability based on eq. (2.2). In order to estimation the heritability of a trait from the GWAS test-statistic, we first observed that when both X and Y are standardized, β^2 will be equal to the coefficient of determination (r^2). Then, based on properties of the Pearson product-moment correlation coefficient:

$$r = \frac{t}{\sqrt{n-2+t^2}} \tag{2.3}$$

where t follows the student-t distribution and n is the number of samples. One can then obtain the r^2 by taking the square of eq. (2.3)

$$r^2 = \frac{t^2}{n-2+t^2} \tag{2.4}$$

It is observed that t^2 will follow the F-distribution and when n is big, t^2 will converge into χ^2 distribution.

When the effect size is small and n is big, r^2 will be approximately χ^2 distributed with mean ~ 1. We can then approximate eq. (2.4) as

$$r^2 = \frac{\chi^2}{n} \tag{2.5}$$

and define the *observed* effect size of each SNP to be

$$f = \frac{\chi^2 - 1}{n} \tag{2.6}$$

When there are Linkage Disequilibrium (LD) between each individual SNPs, the situation will become more complicated as each SNPs' observed effect will contains effect coming from other SNPs in LD with it.

$$f_{observed} = f_{true} + f_{LD} \tag{2.7}$$

To account for the LD structure, we first assume our phenotype $\boldsymbol{Y}$ and genotype $\boldsymbol{X} = (X_1, X_2, \ldots, X_m)^t$ are standardized and that

$$\begin{aligned} \boldsymbol{Y} &\sim f(0,1) \\ \boldsymbol{X} &\sim f(0,\boldsymbol{R}) \end{aligned}$$

Where $\boldsymbol{R}$ is the LD matrix between SNPs.

We can then express eq. (2.1) in matrix form:

$$\boldsymbol{Y} = \boldsymbol{\beta}^t\boldsymbol{X} + \epsilon \tag{2.8}$$

Definition of heritability will then become

$$\begin{aligned} Heritability &= \frac{\mathrm{Var}(\boldsymbol{\beta}^t\boldsymbol{X})}{\mathrm{Var}(\boldsymbol{Y})} \\ &= \mathrm{Var}(\boldsymbol{\beta}^t\boldsymbol{X}) \end{aligned} \tag{2.9}$$

If we then assume now that $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_m)^t$ has distribution

$$\begin{aligned}\boldsymbol{\beta} &\sim f(0, H)\\ \boldsymbol{H} &= diag(\boldsymbol{h})\\ \boldsymbol{h} &= (h_1^2, h_2^2, \ldots, h_m^2)^t\end{aligned}$$

where $\boldsymbol{H}$ is the variance of the true effect. It is shown that heritability can be expressed as

$$\begin{aligned}\mathrm{Var}(\boldsymbol{\beta}^t\boldsymbol{X}) &= \mathrm{E}_X\mathrm{Var}_{\beta|X}(\boldsymbol{X}^t\boldsymbol{\beta}) + \mathrm{Var}_X\mathrm{E}_{(\beta|X)}(\boldsymbol{\beta}^2\boldsymbol{X})\\ &= \mathrm{E}_X(\boldsymbol{X}^t\boldsymbol{\beta}\boldsymbol{\beta}^T\boldsymbol{X})\\ &= \mathrm{E}_X(\boldsymbol{X}^t\boldsymbol{H}\boldsymbol{X})\\ &= \mathrm{E}(\boldsymbol{X})^t\boldsymbol{H}\mathrm{E}(\boldsymbol{X}) + \mathrm{Tr}(\mathrm{Var}(\boldsymbol{X}\boldsymbol{H}))\\ &= \mathrm{Tr}(\mathrm{Var}(\boldsymbol{X}\boldsymbol{H}))\\ &= \sum_i h_i^2 \qquad (2.10)\end{aligned}$$

Now if we consider the covariance between SNP i (X_i) and Y, we have

$$\begin{aligned}\mathrm{Cov}(\boldsymbol{X}_i, \boldsymbol{Y}) &= \mathrm{Cov}(\boldsymbol{X}_i, \boldsymbol{\beta}^t\boldsymbol{X} + \epsilon)\\ &= \mathrm{Cov}(\boldsymbol{X}_i, \boldsymbol{\beta}^t\boldsymbol{X})\\ &= \sum_j \mathrm{Cov}(\boldsymbol{X}_i, \boldsymbol{X}_j)\boldsymbol{\beta}_j\\ &= \boldsymbol{R}_i\boldsymbol{\beta}_j \qquad (2.11)\end{aligned}$$

As both X and Y are standardized, the covariance will equal to the correlation and we can define the correlation between SNP i and Y as

$$\rho_i = \boldsymbol{R}_i\boldsymbol{\beta}_j \qquad (2.12)$$

In reality, the *observed* correlation usually contains error. Therefore we define the *observed* correlation to be

$$\hat{\rho_i} = \rho_i + \frac{\epsilon_i}{\sqrt{n}} \qquad (2.13)$$

for some error ϵ_i. The distribution of the correlation coefficient about the true correlation ρ is approximately

$$\hat{\rho_i} \sim f(\rho_i, \frac{(1-\rho^2)^2}{n})$$

By making the assumption that ρ_i is close to 0 for all i, we have

$$\begin{aligned}\mathrm{E}(\epsilon_i|\rho_i) &\sim 0\\ \mathrm{Var}(\epsilon_i|\rho_i) &\sim 1\end{aligned}$$

We then define our z-statistic and χ^2-statistic as

$$\begin{aligned} z_i &= \hat{\rho}_i\sqrt{n} \\ \chi^2 &= z_i^2 \\ &= \hat{\rho_i}^2 n \end{aligned}$$

From eq. (2.13) and eq. (2.12), χ^2 can then be expressed as

$$\begin{aligned} \chi^2 &= \hat{\rho}^2 n \\ &= n(\boldsymbol{R}_i\boldsymbol{\beta}_j + \frac{\epsilon_i}{\sqrt{n}})^2 \end{aligned}$$

The expectation of χ^2 is then

$$\begin{aligned} \mathrm{E}(\chi^2) &= n(\boldsymbol{R_i}\boldsymbol{\beta}\boldsymbol{\beta}^t\boldsymbol{R}_i + 2\boldsymbol{R}_i\boldsymbol{\beta}\frac{\epsilon_i}{\sqrt{n}} + \frac{\epsilon_i^2}{n}) \\ &= n\boldsymbol{R}_i\boldsymbol{H}\boldsymbol{R}_i + 1 \end{aligned}$$

To derive least square estimates of h_i^2, we need to find $\hat{h_i^2}$ which minimizes

$$\begin{aligned} \sum_i(\chi_i^2 - \mathrm{E}(\chi_i^2))^2 &= \sum_i(\chi_i^2 - (n\boldsymbol{R}_i\boldsymbol{H}\boldsymbol{R}_i + 1))^2 \\ &= \sum_i(\chi_i^2 - 1 - n\boldsymbol{R}_i\boldsymbol{H}\boldsymbol{R}_i)^2 \end{aligned}$$

If we define

$$f_i = \frac{\chi_i^2 - 1}{n} \tag{2.14}$$

we got

$$\begin{aligned} \sum_i(\chi_i^2 - \mathrm{E}(\chi_i^2))^2 &= \sum_i(f_i - \boldsymbol{R}_i\boldsymbol{H}\boldsymbol{R}_i)^2 \\ &= \boldsymbol{f}\boldsymbol{f}^t - 2\boldsymbol{f}^t\boldsymbol{R_{sq}}\hat{\boldsymbol{h}} + \hat{\boldsymbol{h}}^t\boldsymbol{R}^t_{\boldsymbol{sq}}\boldsymbol{R_{sq}}\hat{\boldsymbol{h}} \end{aligned} \tag{2.15}$$

where $\boldsymbol{R_{sq}} = \boldsymbol{R} \circ \boldsymbol{R}$. By differentiating eq. (2.15) w.r.t $\hat{h}$ and set to 0, we get

$$\begin{aligned} 2\boldsymbol{R}^t_{\boldsymbol{sq}}\boldsymbol{R_{sq}}\hat{\boldsymbol{h^2}} - 2\boldsymbol{R_{sq}}\boldsymbol{f} &= 0 \\ \boldsymbol{R_{sq}}\hat{\boldsymbol{h^2}} &= \boldsymbol{f} \end{aligned} \tag{2.16}$$

And the heritability is then defined as

$$\widehat{Heritability} = \boldsymbol{1}^t\boldsymbol{R}^{-1}_{\boldsymbol{sq}}\boldsymbol{f} \tag{2.17}$$

2.2.2 Inverse of the Linkage Disequilibrium matrix

In order to obtain the heritability estimation, we will require to solve eq. (2.17). If $\boldsymbol{R_{sq}}$ is of full rank and positive semi-definite, it will be straight-forward to solve the matrix equation. However, more often than not, the LD matrix are rank-deficient and suffer from multicollinearity, making it ill-conditioned, therefore highly sensitive to changes or errors in the input. To be exact, we can view eq. (2.17) as calculating the sum of $\hat{\boldsymbol{h^2}}$ from eq. (2.16). This will involve solving for

$$\hat{\boldsymbol{h^2}} = \boldsymbol{R_{sq}^{-1}f} \tag{2.18}$$

where an inverse of $\boldsymbol{R_{sq}}$ is observed.

In normal circumstances (e.g. when $\boldsymbol{R_{sq}}$ is full rank and positive semi-definite), one can easily solve eq. (2.18) using the QR decomposition or LU decomposition. However, when $\boldsymbol{R_{sq}}$ is ill-conditioned, the traditional decomposition method will fail. Even if the decomposition is successfully performed, the result tends to be a meaningless approximation to the true $\hat{\boldsymbol{h^2}}$.

Therefore, to obtain a meaningful solution, regularization techniques such as the Tikhonov Regularization (also known as Ridge Regression) and Truncated Singular Value Decomposition (tSVD) has to be performed[1]. Arguably, there are a large variety of regularization techniques, yet the discussion of which is beyond the scope of this study. In this study, we will focus on the use of tSVD in the regularization of the LD matrix. This is because the Singular Value Decomposition (SVD) routine has been implemented in the EIGEN C++ library [2], allowing us to implement the tSVD method without much concern with regard to the algorithm.

In this study, tSVD were used as the SVD was implemented in the EIGEN C++ library[2], allow for a quick and simple implementation of the method.

2.2.3 Quantitative Trait

2.2.4 Case Control Studies

2.2.5 Extreme Phenotype Selections

2.3 Simulation

2.3.1 Quantitative Trait

2.3.2 Case Control Studies

2.3.3 Exreme Phenotype Selections

2.4 Result

2.5 Discussion

Chapter 3

Heritability of Schizophrenia

3.1 Introduction

3.2 Heritability Estimation

This will be a very simple section, focused on how to perform the heritability estimation on Schizophrenia (SCZ). Should also tokenize the heritability into subcategories (e.g. immune, neuron, etc)

3.2.1 Methodology

3.2.2 Result

3.3 Brain development and Schizophrenia

Here we will perform the WGCNA and brain development network. Seeing how the whether if any brain development network were enriched with SNPs that explain the variance of phenotype

3.3.1 Methodology

3.3.2 Result

3.4 Discussion

Chapter 4

Heritability of Response to antipsychotic treatment

4.1 Introduction

Here we try to use Beatrice's data and estimate the heritability explained in drug response. Should also repeat the region-wise heritability

4.2 Methodology

4.3 Result

4.4 Discussion

Chapter 5

Risk Prediction

5.1 Methodology

We can define the traditional Polygenic Risk Score (PGS) as

$$\hat{Y} = diag(\beta)X \tag{5.1}$$

where X is the standardized genotype, β is the test-statistic calculated from other studies.

5.1.1 Simulation

5.2 Result

5.3 Discussion

Chapter 6

Conclusion

Bibliography

[1] Arnold Neumaier. "Solving Ill-Conditioned and Singular Linear Systems: A Tutorial on Regularization". In: *SIAM Review* 40.3 (1998), pp. 636–666. ISSN: 0036-1445. DOI: 10.1137/S0036144597321909.

[2] Gaël Guennebaud, Benoît Jacob, et al. *Eigen v3*. http://eigen.tuxfamily.org. 2010.

Appendix