Summer 2022

Statistical Genetic Discoveries Using Restricted Maximum Likelihood Method

Erika Wu
Princess Anne High School, Virginia Beach

STATISTICAL GENETIC DISCOVERIES USING RESTRICTED MAXIMUM LIKELIHOOD METHOD

Erika Wu
Princess Anne High School
Virginia Beach, USA
E-mail: erikawu20@gmail.com

Lesly Cerezo
Instituto Politécnico Nacional
Mexico City, Mexico
E-mail: lortizc1401@alumno.ipn.mx

Abstract

In statistical genetics, genetic association and genomic prediction become more successful with a highly heritable trait. Identifying highly heritable components of a complex disease can thus advance scientific understanding of the disease and potentially lead to effective prevention and treatments. Using Matlab and existing large-scale genome datasets, we evaluate a restricted maximum likelihood approach to identify highly heritable components of a complex disease as a function of multiple clinical variables.

Keywords: chip heritability, genome-wide association study, restricted maximum likelihood, mixed-effect linear model, quadratic programming

1. Introduction

Statistical genetics is a scientific field focused on developing mathematical models and statistical inference methodologies to draw inferences from genetic data [1, 2]. Statistical methods are often used for genetic association and genomic prediction. Genetic association is commonly defined as the relationship between a phenotype and a genetic polymorphism (such as single nucleotide polymorphism) or between two genetic polymorphisms [3]. Genomic prediction is to predict a future genetic value or phenotypic trait of an individual. The ability to translate genotype information into the prediction of disease phenotypes is important for precision medicine [4].

The success of both genetic association and genomic prediction are positively associated with the heritability of the trait used in the analysis. Identifying highly heritable components of a complex disease can thus advance scientific understanding of the disease and potentially lead to effective prevention and treatments.

Existing heritable component analysis methods use data from related individuals to compute linearly-combined traits to maximize heritability. However, very limited data sets from related individuals are available. Furthermore, data sets from twins and families often involve built-in biases. With privacy and other data collection constraints, most available genotypic data are from

apparently unrelated individuals. Recent advances in acquiring genome-wide markers have enhanced heritability estimation using genotypic data from apparently unrelated individuals.

A genome-wide association study (GWAS) is a research approach used to identify genomic variants that are statistically associated with a risk for a disease or a particular trait. The method involves surveying the genomes of many people, looking for genomic variants that occur more frequently in those with a specific disease or trait compared to those without the disease or trait. Once such genomic variants are identified, they are typically used to search for nearby variants that contribute directly to the disease or trait [5].

Genome-wide genetic data has also provided new opportunities to estimate genetic relatedness using mixed-model analysis. The principle behind mixed-model analyses is the same as twin and family-based heritability analyses — heritability is estimated via correlation between genetic sharing and phenotypic sharing. The key difference is that rather than using the theoretical estimates of genetic sharing based on Mendel's laws, in mixed-model analysis, empirical estimates of genetic sharing are used and are directly observed from genotype data. Estimates of genetic sharing, often represented as a genetic relationship matrix, can then be used in a variety of statistical analyses. These analyses are typically based on data from large scale genome-wide single nucleotide polymorphism (SNP) genotyping arrays [6].

Genetic relationship matrices use genotyped SNPs from GWAS datasets to estimate genomic sharing among individuals in a study sample. Using these estimates of genomic sharing, investigators can statistically model the proportion of trait variance explained by this sharing. If the information captured by the GWAS dataset represented 100% of all genetic variation, this analysis would have yielded a perfectly accurate estimate of trait heritability. Because genotyping technologies do not capture all genetic variation, the estimates of genomic sharing are limited to genetic variants directly genotyped (or variants in strong linkage disequilibrium). Therefore, when properly adjusted for confounding factors, the variance explained by genetic sharing of GWAS-genotyped SNPs — often referred to as chip heritability or pseudo-heritability — can be considered a surrogate for narrow-sense heritability (or heritability due to additive genetic effects).

Novel statistical models are thus needed to identify disease components (subtypes) with high chip heritability from very large scale data. The research problem in this project is to evaluate a novel statistical model based on Restricted Maximum Likelihood (REML). Compared to the common maximum likelihood estimate (MLE) methods in statistics, REML estimation is often used in the more complicated context of mixed models with fixed effects from covariates.

2. Optimization Problem

Given a set of n subjects, we denote their trait values of a quantitative trait y by a vector y of length n. We use a matrix $Zn\times m$ to represent their standardized genotypic data at m genetic markers, and $Cn\times p$ to represent their data on p covariates. The chip heritability estimation method [7] assumes the following mixed-effect linear model that characterizes how a phenotype is related to genotypes and covariates:

$$y=C\beta+Zu+\varepsilon \quad (1)$$

where ε is a vector of length n, which specifies residual effects. In Eq.(1), all covariates create fixed effects (fixed β) on the phenotype whereas genetic effects are random (random u). Assume that u and ε are independent and follow Gaussian distributions: $u \sim N(0, \sigma_e^2)$ and $\varepsilon \sim N(0, I\sigma_u^2)$. Then, the covariance of y between individuals, denoted by $\Omega_{n\times n}$, can be calculated as:

$$\Omega=ZZ^T\sigma_u^2+I\sigma_e^2 \quad (2)$$

A definitive quantitative trait y is not known beforehand but needs to be derived from a set of known clinical variables. Let $X_{n\times d}$ be the data matrix of d clinical variables x for the same n subjects as in Z. A trait y is defined by a linear function of $y = w^\top x$ where w is the vector of combination coefficients. Correspondingly, the trait values $y = Xw$.

The problem now is to search for the best w to form a trait y that maximizes the likelihood, i.e., a large σ_u^2 but small σ_e^2.

3. Methodology

Using Matlab and existing large-scale genome datasets, we evaluate an REML-based quadratic programming approach to identification of highly heritable components of a complex disease as a function of multiple clinical variables [7]. The heritability of the components is estimated directly from unrelated individuals using their genome-wide single nucleotide polymorphisms (SNPs). This REML-based approach is designed to accommodate fixed effects due to covariates, such as age and race, so that the derived traits have high chip heritability after correcting for fixed effects.

In this REML-based approach, a sequential quadratic programming algorithm is used to efficiently solve the optimization problem. This project validates the algorithm both in simulations and on a real-world dataset that was aggregated from genetic studies of cocaine, opioid, and alcohol dependence [7].

4. Implementation of Quadratic Programming Algorithm

The algorithm proposed by Sun et al [7] is a sequential quadratic programming approach to solving the optimization problem.

We decompose w into two parts: $w = u - v$, where both u and v are vectors of the same size as that of w, and all the components in u and v are required to be non-negative (i.e., $u \geq 0$, and $v \geq 0$). Because $Xw=Xu-Xv$, we denote $\gamma= [u^T, v^T]^T$, $H = [X, -X]$, and then we have $Xw = H\gamma$. By the change of variables, and using the L^1 vector norm of w as a regularizing factor to prevent overfitting [8], the optimization problem can be equivalently rewritten as:

$$\begin{aligned} \min_{\gamma} \quad & f: \ \tfrac{1}{n}\boldsymbol{\gamma}^T\left(\mathbf{H}^T\mathbf{P}\mathbf{H}\right)\boldsymbol{\gamma} + \tfrac{\lambda}{d}\textstyle\sum_{i=1}^{2d}\gamma_i \\ \text{subject to} \quad & g_1: \ \boldsymbol{\gamma}^T\left(\mathbf{H}^T\mathbf{Q}\mathbf{H}\right)\boldsymbol{\gamma} - 1 = 0 \\ & g_{2:e}: \ \boldsymbol{\gamma} \geq 0, \end{aligned}$$

where *f* denotes the objective function, H = [X , − X], X $n \times d$ is the data matrix of *d* clinical variables x for the same *n* subjects, *g*'s denote the constraints, and $e = 2d + 1$, indicating the number of constraints in that group, and P and Q defined as follows:

$$P = G^{-1} - G^{-1}C(C^TG^{-1}C)^{-1}C^TG^{-1} \qquad (3)$$

where $G = ZZ^T/m$, which is also the genetic relationship matrix (GRM) among subjects determined by the causal variants.

$$Q = 1/n * J^TJ \qquad (4)$$

where *J* is calculated by $J = I - C(C^T\Omega^{-1}C)^{-1}C^T\Omega^{-1}$ and Ω is the covariance matrix in Eq.(2).

Below is the quadratic programming algorithm [7]:

Input: Z, C, X, λ
Output: γ 1.

The steps followed:

1. Calculate P according to Eq.(3), and Q according to Eq.(4).
2. Initialize with u=1, v=0.
3. Initialize the Lagrange multipliers α =1.
4. Evaluate f,▽f, ▽gi and ▽2L with the current γ and α.
5. Obtain $\hat{p}$ and $\hat{q}$ as detailed below.
6. Perform a line search to find the searching step size s.
7. Update γ and α as in Eq.(5). Repeat 4-7 until γ reaches a fixed point.

$\hat{p}$ and $\hat{q}$ are obtained as follows:

$$\begin{aligned} & min\, p \ \ f(\gamma_t) + \nabla f(\gamma_t)^T p + \tfrac{1}{2}p^T \nabla^T L(\gamma_t, \alpha_t)p \\ & s.t. \ \ \nabla g_1(\gamma_t)^T p + g_1(\gamma_t) = 0 \\ & \quad \nabla g_1(\gamma_t)^T p + g_1(\gamma_t) \geqslant 0, \ i \in [2:e] \end{aligned}$$

$$\gamma_{t+1} = \gamma_t + s\hat{p}_t \ , \ \ \alpha_{t+1} = \alpha_t + s(\hat{q}_t - \alpha_t) \qquad (5)$$

The above update to *γ* follows the direction along which the minimization objective of the optimization problem can be decreased the most.

5. Results and Conclusion

Using Matlab to implement the algorithm, simulation studies demonstrate that the evaluated approach can identify the hypothesized component from multiple synthesized features. A case study on cocaine dependence using an existing dataset [7] confirms a quantitative trait that achieved high chip heritability, where the quantitative trait corresponds to the likelihood of an individual's membership in a cocaine dependence subtype.

Limitations of Study

One underlying assumption, as in other chip heritability studies, is that the genetic variants captured by the genome-wide SNPs tag rare and structural variation with enough accuracy to properly estimate their effects along with those of common SNPs.

Acknowledgement

We deeply thank our mentor, Dr. Jiangwen Sun, and his graduate assistant, Mr. Kai Buckhalter, for their guidance in this project.

References

[1] Lange K, Papp JC, Sinsheimer JS, Sobel EM. Next Generation Statistical Genetics: Modeling, Penalization, and Optimization in High-Dimensional Data. Annu Rev Stat Appl. 2014 Jan 1;1(1):279-300. doi: 10.1146/annurev-statistics-022513-115638. PMID: 24955378; PMCID: PMC4062304.

[2] Schork NJ, Greenwood TA, Braff DL. Statistical genetics concepts and approaches in schizophrenia and related neuropsychiatric research. Schizophr Bull. 2007 Jan;33(1):95-104. doi: 10.1093/schbul/sbl045. Epub 2006 Oct 11. PMID: 17035359; PMCID: PMC2632283.

[3] Kumar P, Song Z-H, Chapter 61 - Polymorphisms of the CB2 Cannabinoid Receptor, Editor(s): V.R. Preedy, Handbook of Cannabis and Related Pathologies, Academic Press, 2017, Pages 584-591, ISBN 9780128007563, doi: 10.1016/B978-0-12-800756-3.00071-5.

[4] de los Campos G, Gianola D, Allison DB. Predicting genetic predisposition in humans: the promise of whole-genome markers. Nat Rev Genet. 2010 Dec;11(12):880-6. doi: 10.1038/nrg2898. Epub 2010 Nov 3. PMID: 21045869.

[5] "Genome-Wide Association Studies (GWAS)" National Human Genome Research Institute - Genetics Glossary. https://www.genome.gov/genetics-glossary/Genome-Wide-Association-Studies (Retrieved August 2022).

[6] Hall JB, Bush WS. Analysis of Heritability Using Genome-Wide Data. Curr Protoc Hum Genet. 2016 Oct 11;91:1.30.1-1.30.10. doi: 10.1002/cphg.25. PMID: 27727439; PMCID: PMC5127448.

[7] Sun J, Kranzler HR, Bi H. Refining multivariate disease phenotypes for high chip heritability. BMC Med Gen. 2015 Sep;8(S3). doi: 10.1186/1755-8794-8-S3-S3.

[8] Vapnik VN. An overview of statistical learning theory. IEEE Transactions on Neural Networks. 1999, 10 (5): 988-999. doi: 10.1109/72.788640.