Simultaneous SNP selection and adjustment for population structure in high dimensional prediction models

Sahir R Bhatnagar[1,2,*], Yi Yang[3], Tianyuan Lu[4,5¤], Erwin Schurr[6], JC Loredo-Osti[7], Marie Forest[8], Karim Oualkacha[9], Celia MT Greenwood[1,4,5,10,11]

1 Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Montréal, Québec, Canada
2 Department of Diagnostic Radiology, McGill University, Montréal, Québec, Canada
3 Department of Mathematics and Statistics, McGill University, Montréal, Québec, Canada
4 Quantitative Life Sciences, McGill University, Montréal, Québec, Canada
5 Lady Davis Institute, Jewish General Hospital, Montréal, Québec, Canada
6 Department of Medicine, McGill University, Montréal, Québec, Canada
7 Department of Mathematics and Statistics, Memorial University, St. John's, Newfoundland, Canada
8 École de Technologie Supérieure, Montréal, Québec, Canada
9 Département de Mathématiques, UQÀM, Montréal, Québec, Canada
10 Department of Oncology, McGill University, Montréal, Québec, Canada
11 Department of Human Genetics, McGill University, Montréal, Québec, Canada

* sahir.bhatnagar@mcgill.ca

Abstract

Complex traits are known to be influenced by a combination of environmental factors and rare and common genetic variants. However, detection of such multivariate associations can be compromised by low statistical power and confounding by population structure. Linear mixed effects models (LMM) can account for correlations due to relatedness but have not been applicable in high-dimensional (HD) settings where the number of fixed effect predictors greatly exceeds the number of samples. False positives or false negatives can result from two-stage approaches, where the residuals estimated from a null model adjusted for the subjects' relationship structure are subsequently used as the response in a standard penalized regression model. To overcome these challenges, we develop a general penalized LMM with a single random effect called `ggmix` for simultaneous SNP selection and adjustment for population structure in high dimensional prediction models. We develop a blockwise coordinate descent algorithm with automatic tuning parameter selection which is highly scalable, computationally efficient and has theoretical guarantees of convergence. Through simulations and three real data examples, we show that `ggmix` leads to more parsimonious models compared to the two-stage approach or principal component adjustment with better prediction accuracy. Our method performs well even in the presence of highly correlated markers, and when the causal SNPs are included in the kinship matrix. `ggmix` can be used to construct polygenic risk scores and select instrumental variables in Mendelian randomization studies. Our algorithms are available in an R package available on CRAN (`https://cran.r-project.org/package=ggmix`).

Author summary

This work addresses a recurring challenge in the analysis and interpretation of genetic association studies: which genetic variants can best predict and are independently associated with a given phenotype in the presence of population structure ? Not controlling confounding due to geographic population structure, family and/or cryptic relatedness can lead to spurious associations. Much of the existing research has therefore focused on modeling the association between a phenotype and a single genetic variant in a linear mixed model with a random effect. However, this univariate approach may miss true associations due to the stringent significance thresholds required to reduce the number of false positives and also ignores the correlations between markers. We propose an alternative method for fitting high-dimensional multivariable models, which selects SNPs that are independently associated with the phenotype while also accounting for population structure. We provide an efficient implementation of our algorithm and show through simulation studies and real data examples that our method outperforms existing methods in terms of prediction accuracy and controlling the false discovery rate.

Introduction

Genome-wide association studies (GWAS) have become the standard method for analyzing genetic datasets owing to their success in identifying thousands of genetic variants associated with complex diseases (`https://www.genome.gov/gwastudies/`). Despite these impressive findings, the discovered markers have only been able to explain a small proportion of the phenotypic variance; this is known as the missing heritability problem [**?**]. One plausible reason is that there are many causal variants that each explain a small amount of variation with small effect sizes [**?**]. Methods such GWAS, which test each variant or single nucleotide polymorphism (SNP) independently, may miss these true associations due to the stringent significance thresholds required to reduce the number of false positives [**?**]. Another major issue to overcome is that of confounding due to geographic population structure, family and/or cryptic relatedness which can lead to spurious associations [**?**]. For example, there may be subpopulations within a study that differ with respect to their genotype frequencies at a particular locus due to geographical location or their ancestry. This heterogeneity in genotype frequency can cause correlations with other loci and consequently mimic the signal of association even though there is no biological association [**?**, **?**]. Studies that separate their sample by ethnicity to address this confounding suffer from a loss in statistical power due to the drop in sample size.

To address the first problem, multivariable regression methods have been proposed which simultaneously fit many SNPs in a single model [**?**, **?**]. Indeed, the power to detect an association for a given SNP may be increased when other causal SNPs have been accounted for. Conversely, a stronger signal from a causal SNP may weaken false signals when modeled jointly [**?**].

Solutions for confounding by population structure have also received significant attention in the literature [**?**, **?**, **?**, **?**]. There are two main approaches to account for the relatedness between subjects: 1) the principal component (PC) adjustment method and 2) the linear mixed model (LMM). The PC adjustment method includes the top PCs of genome-wide SNP genotypes as additional covariates in the model [**?**]. The LMM uses an estimated covariance matrix from the individuals' genotypes and includes this information in the form of a random effect [**?**].

While these problems have been addressed in isolation, there has been relatively little progress towards addressing them jointly at a large scale. Region-based tests of association have been developed where a linear combination of p variants is regressed on

the response variable in a mixed model framework [**?**]. In case-control data, a stepwise logistic-regression procedure was used to evaluate the relative importance of variants within a small genetic region [**?**]. These methods however are not applicable in the high-dimensional setting, i.e., when the number of variables p is much larger than the sample size n, as is often the case in genetic studies where millions of variants are measured on thousands of individuals.

There has been recent interest in using penalized linear mixed models, which place a constraint on the magnitude of the effect sizes while controlling for confounding factors such as population structure. For example, the LMM-lasso [**?**] places a Laplace prior on all main effects while the adaptive mixed lasso [**?**] uses the L_1 penalty [**?**] with adaptively chosen weights [**?**] to allow for differential shrinkage amongst the variables in the model. Another method applied a combination of both the lasso and group lasso penalties in order to select variants within a gene most associated with the response [**?**]. However, methods such as the LMM-lasso are normally performed in two steps. First, the variance components are estimated once from a LMM with a single random effect. These LMMs normally use the estimated covariance matrix from the individuals' genotypes to account for the relatedness but assumes no SNP main effects (i.e. a null model). The residuals from this null model with a single random effect can be treated as independent observations because the relatedness has been effectively removed from the original response. In the second step, these residuals are used as the response in any high-dimensional model that assumes uncorrelated errors. This approach has both computational and practical advantages since existing penalized regression software such as `glmnet` [**?**] and `gglasso` [**?**], which assume independent observations, can be applied directly to the residuals. However, recent work has shown that there can be a loss in power if a causal variant is included in the calculation of the covariance matrix as its effect will have been removed in the first step [**?**, **?**].

In this paper we develop a general penalized LMM framework called `ggmix` that simultaneously selects variables and estimates their effects, accounting for between-individual correlations. We develop a blockwise coordinate descent algorithm with automatic tuning parameter selection which is highly scalable, computationally efficient and has theoretical guarantees of convergence. Our method can handle several sparsity inducing penalties such as the lasso [**?**] and elastic net [**?**]. Through simulations and three real data examples, we show that `ggmix` leads to more parsimonious models compared to the two-stage approach or principal component adjustment with better prediction accuracy. Our method performs well even in the presence of highly correlated markers, and when the causal SNPs are included in the kinship matrix.

All of our algorithms are implemented in the `ggmix` R package hosted on CRAN with extensive documentation (`https://sahirbhatnagar.com/ggmix`). We provide a brief demonstration of the `ggmix` package in Appendix **??**.

The rest of the paper is organized as follows. In Section 3, we compare the performance of our proposed approach and demonstrate the scenarios where it can be advantageous to use over existing methods through simulation studies and three real data analyses. This is followed by a discussion of our results, some limitations and future directions in Section 4. Section 5 describes the `ggmix` model, the optimization procedure and the algorithm used to fit it.

Lorem ipsum dolor sit [1] amet, consectetur adipiscing elit. Curabitur eget porta erat. Morbi consectetur est vel gravida pretium. Suspendisse ut dui eu ante cursus gravida non sed sem. Nullam Eq (1) sapien tellus, commodo id velit id, eleifend volutpat quam. Phasellus mauris velit, dapibus finibus elementum vel, pulvinar non tellus. Nunc pellentesque pretium diam, quis maximus dolor faucibus id. [2] Nunc convallis sodales ante, ut ullamcorper est egestas vitae. Nam sit amet enim ultrices, ultrices elit pulvinar, volutpat risus.

$$\mathrm{P_Y} = \underbrace{H(Y_n) - H(Y_n|\mathbf{V}_n^Y)}_{S_Y} + \underbrace{H(Y_n|\mathbf{V}_n^Y) - H(Y_n|\mathbf{V}_n^{X,Y})}_{T_{X\to Y}}, \quad (1)$$

Materials and methods

Etiam eget sapien nibh

Nulla mi mi, Fig 1 venenatis sed ipsum varius, volutpat euismod diam. Proin rutrum vel massa non gravida. Quisque tempor sem et dignissim rutrum. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Morbi at justo vitae nulla elementum commodo eu id massa. In vitae diam ac augue semper tincidunt eu ut eros. Fusce fringilla erat porttitor lectus cursus, S1 Video vel sagittis arcu lobortis. Aliquam in enim semper, aliquam massa id, cursus neque. Praesent faucibus semper libero.

Fig 1. Bold the figure title. Figure caption text here, please use this space for the figure panel descriptions instead of using subfigure commands. A: Lorem ipsum dolor sit amet. B: Consectetur adipiscing elit.

Results

Nulla mi mi, venenatis sed ipsum varius, Table 1 volutpat euismod diam. Proin rutrum vel massa non gravida. Quisque tempor sem et dignissim rutrum. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Morbi at justo vitae nulla elementum commodo eu id massa. In vitae diam ac augue semper tincidunt eu ut eros. Fusce fringilla erat porttitor lectus cursus, vel sagittis arcu lobortis. Aliquam in enim semper, aliquam massa id, cursus neque. Praesent faucibus semper libero.

Table 1. Table caption Nulla mi mi, venenatis sed ipsum varius, volutpat euismod diam.

Heading1				**Heading2**			
*cell*1*row*1	cell2 row 1	cell3 row 1	cell4 row 1	cell5 row 1	cell6 row 1	cell7 row 1	cell8 row 1
*cell*1*row*2	cell2 row 2	cell3 row 2	cell4 row 2	cell5 row 2	cell6 row 2	cell7 row 2	cell8 row 2
*cell*1*row*3	cell2 row 3	cell3 row 3	cell4 row 3	cell5 row 3	cell6 row 3	cell7 row 3	cell8 row 3

Table notes Phasellus venenatis, tortor nec vestibulum mattis, massa tortor interdum felis, nec pellentesque metus tortor nec nisl. Ut ornare mauris tellus, vel dapibus arcu suscipit sed.

LOREM and IPSUM nunc blandit a tortor

3rd level heading

Maecenas convallis mauris sit amet sem ultrices gravida. Etiam eget sapien nibh. Sed ac ipsum eget enim egestas ullamcorper nec euismod ligula. Curabitur fringilla pulvinar lectus consectetur pellentesque. Quisque augue sem, tincidunt sit amet feugiat eget, ullamcorper sed velit. Sed non aliquet felis. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Mauris commodo justo ac dui pretium imperdiet. Sed suscipit iaculis mi at feugiat.

1. react
2. diffuse free particles
3. increment time by dt and go to 1

Sed ac quam id nisi malesuada congue

Nulla mi mi, venenatis sed ipsum varius, volutpat euismod diam. Proin rutrum vel massa non gravida. Quisque tempor sem et dignissim rutrum. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Morbi at justo vitae nulla elementum commodo eu id massa. In vitae diam ac augue semper tincidunt eu ut eros. Fusce fringilla erat porttitor lectus cursus, vel sagittis arcu lobortis. Aliquam in enim semper, aliquam massa id, cursus neque. Praesent faucibus semper libero.

- First bulleted item.
- Second bulleted item.
- Third bulleted item.

Discussion

Nulla mi mi, venenatis sed ipsum varius, Table 1 volutpat euismod diam. Proin rutrum vel massa non gravida. Quisque tempor sem et dignissim rutrum. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Morbi at justo vitae nulla elementum commodo eu id massa. In vitae diam ac augue semper tincidunt eu ut eros. Fusce fringilla erat porttitor lectus cursus, vel sagittis arcu lobortis. Aliquam in enim semper, aliquam massa id, cursus neque. Praesent faucibus semper libero [3].

Conclusion

CO_2 Maecenas convallis mauris sit amet sem ultrices gravida. Etiam eget sapien nibh. Sed ac ipsum eget enim egestas ullamcorper nec euismod ligula. Curabitur fringilla pulvinar lectus consectetur pellentesque. Quisque augue sem, tincidunt sit amet feugiat eget, ullamcorper sed velit.

Sed non aliquet felis. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Mauris commodo justo ac dui pretium imperdiet. Sed suscipit iaculis mi at feugiat. Ut neque ipsum, luctus id lacus ut, laoreet scelerisque urna. Phasellus venenatis, tortor nec vestibulum mattis, massa tortor interdum felis, nec pellentesque metus tortor nec nisl. Ut ornare mauris tellus, vel dapibus arcu suscipit sed. Nam condimentum sem eget mollis euismod. Nullam dui urna, gravida venenatis dui et, tincidunt sodales ex. Nunc est dui, sodales sed mauris nec, auctor sagittis leo. Aliquam tincidunt, ex in facilisis elementum, libero lectus luctus est, non vulputate nisl augue at dolor. For more information, see S1 Appendix.

Supporting information

S1 Fig. Bold the title sentence. Add descriptive text after the title of the item (optional).

S2 Fig. Lorem ipsum. Maecenas convallis mauris sit amet sem ultrices gravida. Etiam eget sapien nibh. Sed ac ipsum eget enim egestas ullamcorper nec euismod ligula. Curabitur fringilla pulvinar lectus consectetur pellentesque.

S1 File. Lorem ipsum. Maecenas convallis mauris sit amet sem ultrices gravida. Etiam eget sapien nibh. Sed ac ipsum eget enim egestas ullamcorper nec euismod ligula. Curabitur fringilla pulvinar lectus consectetur pellentesque.

S1 Video. Lorem ipsum. Maecenas convallis mauris sit amet sem ultrices gravida. Etiam eget sapien nibh. Sed ac ipsum eget enim egestas ullamcorper nec euismod ligula. Curabitur fringilla pulvinar lectus consectetur pellentesque.

S1 Appendix. Lorem ipsum. Maecenas convallis mauris sit amet sem ultrices gravida. Etiam eget sapien nibh. Sed ac ipsum eget enim egestas ullamcorper nec euismod ligula. Curabitur fringilla pulvinar lectus consectetur pellentesque.

S1 Table. Lorem ipsum. Maecenas convallis mauris sit amet sem ultrices gravida. Etiam eget sapien nibh. Sed ac ipsum eget enim egestas ullamcorper nec euismod ligula. Curabitur fringilla pulvinar lectus consectetur pellentesque.

Acknowledgments

Cras egestas velit mauris, eu mollis turpis pellentesque sit amet. Interdum et malesuada fames ac ante ipsum primis in faucibus. Nam id pretium nisi. Sed ac quam id nisi malesuada congue. Sed interdum aliquet augue, at pellentesque quam rhoncus vitae.

References

References

1. Conant GC, Wolfe KH. Turning a hobby into a job: how duplicated genes find new functions. Nat Rev Genet. 2008 Dec;9(12):938–950.
2. Ohno S. Evolution by gene duplication. London: George Alien & Unwin Ltd. Berlin, Heidelberg and New York: Springer-Verlag.; 1970.
3. Magwire MM, Bayer F, Webster CL, Cao C, Jiggins FM. Successive increases in the resistance of Drosophila to viral infection through a transposon insertion followed by a Duplication. PLoS Genet. 2011 Oct;7(10):e1002337.