Comparison of mixed model based approaches for correcting for population substructure with application to extreme phenotype sampling

M.Onifade, K.Burkett

University of Ottawa

Abstract

Mixed models have been useful in correcting for confounding due to population stratification and hidden relatedness in genome wide association studies. This class of models includes linear mixed models (LMM) and generalised linear mixed models (GLMM). Existing mixed model approaches to correct for population substructure have been investigated with both continuous and case/control response variables. However, they have not been investigated in the context of 'extreme phenotype sampling' (EPS), where genetic covariates are only collected on samples having extreme response variable values. In this work, we compare the performance of existing mixed model approaches (LTMLM, GMMAT, CARAT) with EPS data analysed as a binary trait. We use simulation to estimate the type 1 error of all approaches when there is confounding. Since linear mixed models are commonly used even with binary traits, we also analysed the data using a LMM (GEMMA). (Describe results here)

Keywords: Population stratification, Extreme phenotype samples, Mixed models, type 1 error.

1. Introduction

In genetic studies involving human populations, researchers are interested in determining how genetic variation contributes to diseases. Genome Wide Association Studies (GWAS), which involve genotyping a large number of individuals at hundreds of thousands of genetic markers have been useful for discovering the relationships between common variants and complex diseases. Recently, rare variants have been identified as important genetic factors contributing to the risk of disease and human traits. . Exome sequencing has been used to discover rare variation in the human genome; although costs have reduced, it remains a relatively expensive technique . Therefore study designs that are powerful at lower sample sizes are advantageous.

reference??

find a reference

An example of a cost saving design is extreme phenotype sampling: a design where genotyping or sequencing is only done on individuals in the tails of the phenotype distribution. The use of this study design can be traced to the work of [11] where it was used in mapping quantitative trait locis (QTLs) during linkage analysis. Extreme phenotype sampling has since found other uses beyond linkage analysis as other authors have adapted the study beyond linkage analysis. Still in linkage analysis, [3] used EPS and advised that the cutoffs shouldn't be more than the upper and lower 25th percentile. In association studies, [15] used extreme selection techinique to test for the association between a genetic variant and intelligence quotient and [1] assessed the association between general cognitive ability as a behavioral trait and variation in candidate genes. [22] explored the power of the study when using extreme samples compared to the whole population. In rare variant study, [6, 9, 14], EPS has been shown to have sufficient power to detect rare variants.

As with all population based genetic association designs, extreme phenotype sampling is prone to confounding by population structure or stratification. Difference in allele frequencies among members of a strata or subgroup in the population may lead to confounding if there are differences in the phenotype distribution between the subgroups. Confounding is known to leading to spurious associations and an inflation of the type 1 error, which has led to a development of methods that can correct for the effects of population stratification. The earliest methods includes the Genomic control method of Devlin et al. [4] and the STRUCTURE approach of Pritchard et al. [19]. Principal components (PC) based corrections, implemented in the program EIGENSTRAT [17] have also been successfully applied in a number of studies [17, 13]. Very recently, mixed models have become popular due to their robustness in tackling other sources of confounding in the study, in particular cryptic relatedness[18]. Over the years, an impressive number of exact and approximate LMM methods have been developed for use in genetic association studies [10, 12, 24]. Each of these methods incorporate different approaches for making LMM-based analyses feasible at the genome wide level.

However in genetic studies involving humans, the phenotype of interest is often a binary trait, which can be obtained from case-control and cohort study designs, for example.

examples of studies involving case control or cohort

Just like continuous traits, binary traits have also been analysed using linear mixed models [5, 20, 21]. These methods have used an additive polygenic model which allows for transformation of the parameters of the linear model and the logistic model. This practice has resulted in a loss of power as the mean and variance behaviour of the binary trait is ignored [8].

Earlier studies that aimed to correct for confounding rates in binary trait have relied on the method of [16] that derived a direct relationship between linear models and logistic regression. In particular, the authors justified the application of a linear mixed model to binary data by introducing a way of transforming the effect size estimates from the linear to the log-odds scale which is the natural scale by which case-control data is measured. Although widely applied to binary traits, the LMM assumes a continous phenotype where it is reasonable to assume that the trait has a constant residual variance. However, for binary traits in the presence of covariates, this assumption is not valid; therefore fitting a binary response with mixed models may fail to correct the type 1 error rate [2].

Mixed model approaches that are applicable under binary traits have also been developed. Example of an applicable method are based on the use of the liability threshold model that associates with each individual a normally distributed latent variable known as the liability. These methods have been implemented in the softwares LTMLM [7] and LEAP [23] and offers an attractive method for association testing and case-control ascertainment in case/control studies. These methods estimates the latent liabilities and tests for association using these estimates. While LTMLM tests for association using the posterior mean liabilities, LEAP uses a maximum posterior estimation. Another suitable method proposed for analysing binary traits is based on the generalised linear mixed model (GLMM). Specifically, the GMMAT uses the logistic mixed model and first fits a null model for all SNPs in the study and uses this model to compute score test statistics for testing the association between the binary traits and the genetic variant. Another binary trait association method known as CARAT uses a retrospective case-control analysis method to account for the analysis of binary traits and covariates. We desire to state here that all these methods have been examined in cases of population structure.

In this work, we aim to accomplish two goals. First, we present an overview and comparison of methods available for analysing binary traits with or without covariates adjustments using liability models and mixed models. Secondly, we investigate their performance when the binary data comes from an EPS design. We also include an LMM approach, which treats the phenotypes as if it were continous, in our comparison. Here, each of the extremes will be treated as a different category. This is motivated by the fact that mixed model based approaches for correcting confounding has not been tested in the context of the EPS. We focus on whether these methods adequately correct the type 1 error rates due to confounding under an extreme phenotype study design. Finally, we also compare the approaches on a real dataset.

add more detail when the real data phenotype is known.

[1] Ball, D., Hill, L., Eley, T. C., Chorney, M. J., Chorney, K., Thompson, L. A., Detterman, D. K., Benbow, C., Lubinski, D., Owen, M., et al. (1998). Dopamine markers and general cognitive ability. *Neuroreport*, 9(2):347–349.

[2] Chen, H., Wang, C., Conomos, M. P., Stilp, A. M., Li, Z., Sofer, T., Szpiro, A. A., Chen, W., Brehm, J. M., Celedón, J. C., et al. (2016). Control for population structure and relatedness for binary traits in genetic

association studies via logistic mixed models. *The American Journal of Human Genetics*, 98(4):653–666.

[3] Darvasi, A. and Soller, M. (1992). Selective genotyping for determination of linkage between a marker locus and a quantitative trait locus. *Theoretical and applied Genetics*, 85(2-3):353–359.

[4] Devlin, B., Roeder, K., and Wasserman, L. (2001). Genomic control, a new approach to genetic-based association studies. *Theoretical population biology*, 60(3):155–166.

[5] Fakiola, M., Strange, A., Cordell, H. J., Miller, E. N., Pirinen, M., Su, Z., Mishra, A., Mehrotra, S., Monteiro, G. R., Band, G., et al. (2013). Common variants in the hla-drb1–hla-dqa1 hla class ii region are associated with susceptibility to visceral leishmaniasis. *Nature genetics*, 45(2):208.

[6] Guey, L. T., Kravic, J., Melander, O., Burtt, N. P., Laramie, J. M., Lyssenko, V., Jonsson, A., Lindholm, E., Tuomi, T., Isomaa, B., et al. (2011). Power in the phenotypic extremes: a simulation study of power in discovery and replication of rare variants. *Genetic epidemiology*, 35(4):236–246.

[7] Hayeck, T. J., Zaitlen, N. A., Loh, P.-R., Vilhjalmsson, B., Pollack, S., Gusev, A., Yang, J., Chen, G.-B., Goddard, M. E., Visscher, P. M., Patterson, N., and Price, A. L. (2015). Mixed Model with Correction for Case-Control Ascertainment Increases Association Power. *American Journal of Human Genetics*, 96(5):720–730.

[8] Jiang, D., Mbatchou, J., and McPeek, M. S. (2015). Retrospective association analysis of binary traits: overcoming some limitations of the additive polygenic model. *Human heredity*, 80(4):187–195.

[9] Kang, G., Lin, D., Hakonarson, H., and Chen, J. (2012). Two-stage extreme phenotype sequencing design for discovering and testing common and rare genetic variants: efficiency and power. *Human heredity*, 73(3):139–147.

[10] Kang, H. M., Zaitlen, N. A., Wade, C. M., Kirby, A., Heckerman, D., Daly, M. J., and Eskin, E. (2008). Efficient control of population structure in model organism association mapping. *Genetics*, 178(3):1709–1723.

[11] Lander, E. S. and Botstein, D. (1989). Mapping mendelian factors underlying quantitative traits using rflp linkage maps. *Genetics*, 121(1):185–199.

[12] Lippert, C., Listgarten, J., Liu, Y., Kadie, C. M., Davidson, R. I., and Heckerman, D. (2011). Fast linear mixed models for genome-wide association studies. *Nature methods*, 8(10):833.

[13] Patterson, N., Price, A. L., and Reich, D. (2006). Population structure and eigenanalysis. *PLoS genetics*, 2(12):e190.

[14] Peloso, G. M., Rader, D. J., Gabriel, S., Kathiresan, S., Daly, M. J., and Neale, B. M. (2016). Phenotypic extremes in rare variant study designs. *European Journal of Human Genetics*, 24(6):924–930.

[15] Petrill, S. A., Plomin, R., McClearn, G. E., Smith, D. L., Vignetti, S., Chorney, M. J., Chorney, K., Thompson, L. A., Detterman, D. K.,

Benbow, C., et al. (1997). No association between general cognitive ability and the a1 allele of the d2 dopamine receptor gene. *Behavior Genetics*, 27(1):29–31.

[16] Pirinen, M., Donnelly, P., Spencer, C. C., et al. (2013). Efficient computation with a linear mixed model on large-scale data sets with applications to genetic studies. *The Annals of Applied Statistics*, 7(1):369–390.

[17] Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature genetics*, 38(8):904.

[18] Price, A. L., Zaitlen, N. A., Reich, D., and Patterson, N. (2010). New approaches to population stratification in genome-wide association studies. *Nature Reviews Genetics*, 11(7):459.

[19] Pritchard, J. K., Stephens, M., and Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics*, 155(2):945–959.

[20] Sawcer, S., Hellenthal, G., Pirinen, M., Spencer, C. C., Patsopoulos, N. A., Moutsianas, L., Dilthey, A., Su, Z., Freeman, C., Hunt, S. E., et al. (2011). Genetic risk and a primary role for cell-mediated immune mechanisms in multiple sclerosis. *Nature*, 476(7359):214.

[21] Tsoi, L. C., Spain, S. L., Knight, J., Ellinghaus, E., Stuart, P. E., Capon, F., Ding, J., Li, Y., Tejasvi, T., Gudjonsson, J. E., et al. (2012). Identi-

fication of 15 new psoriasis susceptibility loci highlights the role of innate immunity. *Nature genetics*, 44(12):1341.

[22] Van Gestel, S., Houwing-Duistermaat, J. J., Adolfsson, R., van Duijn, C. M., and Van Broeckhoven, C. (2000). Power of selective genotyping in genetic association analyses of quantitative traits. *Behavior genetics*, 30(2):141–146.

[23] Weissbrod, O., Lippert, C., Geiger, D., and Heckerman, D. (2015). Accurate liability estimation improves power in ascertained case-control studies. *Nature methods*, 12(4):332.

[24] Zhou, X. and Stephens, M. (2012). Genome-wide efficient mixed-model analysis for association studies. *Nature genetics*, 44(7):821.