CluStrat: A Structure Informed Clustering Strategy for Population Stratification

Aritra Bose[1,2], Myson C. Burch[2], Agniva Chowdhury[3], Peristera Paschou[4(✉)], and Petros Drineas[2]

[1] Computational Genomics, IBM T.J. Watson Research Center, Yorktown Heights, NY, USA
a.bose@ibm.com

[2] Computer Science Department, Purdue University, West Lafayette, IN, USA
{bose6,mcburch,pdrineas}@purdue.edu

[3] Department of Statistics, Purdue University, West Lafayette, IN, USA
agniva@purdue.edu

[4] Department of Biological Sciences, Purdue University, West Lafayette, IN, USA
ppaschou@purdue.edu

Genome-wide association studies (GWAS) have been extensively used to estimate the signed effects of trait-associated alleles. One of the key challenges in GWAS are confounding factors, such as population stratification, which can lead to spurious genotype-trait associations. Recent independent studies [1,8,10] failed to replicate the strong evidence of previously reported signals of directional selection on height in Europeans in the UK Biobank cohort, and attributed the loss of signal to cryptic relatedness in populations. Population structure causes genuine genetic signals in causal variants to be mirrored in numerous non-causal loci due to *linkage disequilibrium (LD)* [3], resulting in spurious associations. Thus, it is important to account for LD in the computation of the distance matrix [6]. One way to account for the LD structure is to use the squared Mahalanobis distance [5]. Here, we present CluStrat, a stratification correction algorithm for complex population structure that leverages the LD-induced distances between individuals. It performs agglomerative hierarchical clustering using the Mahalanobis distance based *Genetic Relationship Matrix (GRM)* which captures the population-level covariance of the genotypes. Thereafter, we apply sketching-based randomized ridge regression on the clusters and perform a meta-analysis to obtain the association statistics.

With the growing size of data, computing and storing the genome wide GRM is a non-trivial task. We get around this overhead by computing the Mahalanobis distance between two vectors efficiently without storing or inverting the covariance matrix, but instead computing the corresponding rank-k leverage and cross-leverage scores. We compute the rank-k Mahalanobis distance with respect to the top k-left singular vectors of the genotype matrix, thus making the computation feasible for UK Biobank-scale datasets using methods such as `TeraPCA` [2] to approximate the left singular vectors accurately and efficiently.

Supported by NSF IIS 1715202 and NFS DMS 1760353 awarded to PD and PP.

R. Schwartz (Ed.): RECOMB 2020, LNBI 12074, pp. 234–236, 2020.
https://doi.org/10.1007/978-3-030-45257-5_19

We test CluStrat on a large simulation study of arbitrarily-structured, admixed sub-populations by generating 100 GWAS datasets (with 1,000 individuals genotyped on one million genetic markers) from a quantitative trait model (and it's equivalent binary trait) based on previous work [9]. We simulated 30 different scenarios, varying proportions of true genetic effect and admixture and compared it's performance to standard population structure correction approaches such as EIGENSTRAT [7], GEMMA [11], and EMMAX [4]. We identified two to three-fold more true causal variants when compared to the above methods for almost all scenarios, while trading off for a slightly higher spurious associations, but, far less than the uncorrected Armitage trend χ^2 test. Applying CluStrat on WTCCC2 Parkinson's disease (PD) data with a p-value threshold set to 10^{-7}, we identified loci mapped to a host of genes known to be associated with PD such as BACH2, MAP2, NR4A2, SLC11A1, UNC5C to name a few. In summary, CluStrat highlights the advantages of biologically relevant distance metrics, such as the Mahalanobis distance, which seems to capture the cryptic interactions within populations in the presence of LD better than the Euclidean distance. Of independent interest is a simple, but not necessarily well-known, connection between the regularized Mahalanobis distance-based GRM and the leverage and cross-leverage scores of the genotype matrix. CluStrat source code and user manual is available at: https://github.com/aritra90/CluStrat and the full version is available at https://doi.org/10.1101/2020.01.15.908228.

References

1. Berg, J.J., Harpak, A., Sinnott-Armstrong, N., et al.: Reduced signal for polygenic adaptation of height in UK Biobank. eLife **8**, e39725 (2019)
2. Bose, A., Kalantzis, V., Kontopoulou, E.M., et al.: TeraPCA: a fast and scalable software package to study genetic variation in tera-scale genotypes. Bioinformatics **35**, 3679–3683 (2019)
3. Ewens, W.J., Spielman, R.S.: The transmission/disequilibrium test: history, subdivision, and admixture. Am. J. Hum. Genet. **57**(2), 455 (1995)
4. Kang, H.M., Sul, J.H., Service, S.K., et al.: Variance component model to account for sample structure in genome-wide association studies. Nat. Genet. **42**(4), 348 (2010)
5. Mahalanobis, P.C.: On the generalized distance in statistics. National Institute of Science of India (1936)
6. Mathew, B., Léon, J., Sillanpää, M.J.: A novel linkage-disequilibrium corrected genomic relationship matrix for SNP-heritability estimation and genomic prediction. Heredity **120**(4), 356 (2018)
7. Price, A.L., Patterson, N.J., Plenge, R.M., et al.: Principal components analysis corrects for stratification in genome-wide association studies. Nat. Genet. **38**(8), 904 (2006)
8. Sohail, M., Maier, R.M., Ganna, A., et al.: Polygenic adaptation on height is overestimated due to uncorrected stratification in genome-wide association studies. eLife **8**, e39702 (2019)
9. Song, M., Hao, W., Storey, J.D.: Testing for genetic associations in arbitrarily structured populations. Nat. Genet. **47**(5), 550 (2015)

10. Uricchio, L.H., Kitano, H.C., Gusev, A., et al.: An evolutionary compass for detecting signals of polygenic selection and mutational bias. Evol. Lett. **3**(1), 69–79 (2019)
11. Zhou, X., Stephens, M.: Genome-wide efficient mixed-model analysis for association studies. Nat. Genet. **44**(7), 821 (2012)