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
[2] Computer Science Department, Purdue University, West Lafayette IN, USA
[3] Department of Statistics, Purdue University, West Lafayette IN, USA
[4] Department of Biological Sciences, Purdue University, West Lafayette, IN, USA
a.bose@ibm.com; {bose6, mcburch, agniva, pdrineas, ppaschou}@purdue.edu

Abstract. Genome-wide association studies (GWAS) have been extensively used to estimate the signed effects of trait-associated alleles. Recent independent studies failed to replicate the strong evidence of selection for height across Europe implying the shortcomings of standard population stratification correction approaches. Here, we present CluStrat, a stratification correction algorithm for complex population structure that leverages the linkage disequilibrium (LD)-induced distances between individuals. CluStrat performs agglomerative hierarchical clustering using the Mahalanobis distance and then applies sketching-based randomized ridge regression on the genotype data to obtain the association statistics. With the growing size of data, computing and storing the genome wide covariance matrix is a non-trivial task. We get around this overhead by computing the GRM directly using a connection between statistical leverage scores and the Mahalanobis distance. We test CluStrat on a large simulation study of discrete and admixed, arbitrarily-structured sub-populations identifying two to three-fold more true causal variants when compared to Principal Component (PC) based stratification correction methods while trading off for a slightly higher spurious associations. Applying CluStrat on WTCCC2 Parkinson's disease (PD) data, we identified loci mapped to a host of genes associated with PD such as BACH2, MAP2, NR4A2, SLC11A1, UNC5C to name a few.
Availability and Implementation: CluStrat source code and user manual is available at: `https://github.com/aritra90/CluStrat`

Keywords: Population Structure · Association Studies · Clustering · Ridge Regression.

1 Introduction

The basic principle underlying Genome Wide Association Studies (GWAS) is a test for association between genotyped variants for each individual and the trait

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

of interest. GWAS have been extensively used to estimate the signed effects of trait-associated alleles, mapping genes to disorders and over the past decade about 10,000 strong associations between genetic variants and one (or more) complex traits have been reported [48,51,45,17]. One unambiguous conclusion from GWAS is that for almost any complex trait that has been studied so far, genetic variation is linked with many loci contributing to the polygenic nature of the traits. Hence, on average, the proportion of variance explained at the single marker is very small [45].

One of the key challenges in GWAS are confounding factors, such as population stratification, which can lead to spurious genotype-trait associations [37,39,33]. If a dataset consists of individuals from different ethnic groups, then the genotype data will be characterized by genome-wide *linkage disequilibrium (LD)* between variants. LD models the fact that alleles at different loci are correlated in individuals from the same ethnic group. Population structure causes genuine genetic signals in causal variants for a particular trait of interest to be mirrored in numerous non-causal loci because of LD [23], resulting in spurious associations. A related phenomenon, the so-called cryptic relatedness, is caused by individuals who are closely related and often grouped together by standard population structure correction strategies, and also poses a serious confounding problem [18]. Two popular approaches for stratification correction while building the *Genetic Relationship Matrix (GRM)* [2,41] involve (i) including the principal components of the genotypes as adjustment variables [37,38], and (ii) fitting a *Linear Mixed Model (LMM)* with an estimated kinship or GRM from the individual's genotypes [51]. Recently, three independent studies [40,5,43] failed to replicate the previously reported signals of directional selection on height in European populations, as seen in the GIANT consortium (253,288 individuals [49]) in the independent and more recently UK Biobank cohort (500,000 individuals [8]). They further showed that the GIANT GWAS is confounded due to stratification along the north to south axis, where strong signals of selection were previously reported. These recent studies highlight the need for more sophisticated tools for correcting for population stratification.

Our work proposes a simple clustering-based approach to correct for stratification better than existing methods. This method takes into account the linkage disequilibrium while computing the distance between the individuals in a sample. Our approach, called CluStrat, performs *Agglomerative Hierarchical Clustering (AHC)* using a regularized Mahalanobis distance-based GRM, which captures the population-level covariance (LD) matrix for the available genotype data. We test CluStrat on large simulation studies of discrete and admixed, complex-structured populations of over 1,000 individuals genotyped on over one million genetics markers (Single Nucleotide Polymorphisms or SNPs for short) and we observe that our approach has the lowest number of spurious associations in our simulations. Our approach also identifies two to three-fold more rare variants at causal loci when compared to standard stratification correction strategies. Of independent interest is a simple, but not necessarily well-known, connection between the regularized Mahalanobis distance-based GRM that is used in our

approach and the leverage and cross-leverage scores of the genotype matrix (see Methods and Appendix A for details).

2 Materials and Methods

Notation In the remainder of the paper we let matrix $\mathbf{X} \in \mathbb{R}^{m \times n}$ denote the genotype matrix (e.g., the minor allele frequency (MAF) matrix on m samples genotyped on n SNPs). The matrix is appropriately normalized as is common in population genetics analyses to have zero mean and variance one (columnwise). The vector $y \in \mathbb{R}^m$ represents the trait of interest and its i-th entry is set to one for cases and to zero for controls (for binary traits). We let $\mathbf{X}_{i*}$ denote the i-th row of the matrix $\mathbf{X}$ as a row vector and $\mathbf{X}_{*i}$ denote the i-th column of the matrix $\mathbf{X}$ as a column vector. We represent the top k left singular vectors of the matrix $\mathbf{X}$ by the matrix $\mathbf{U}_k \in \mathbb{R}^{m \times k}$ and we will use the notation $(\mathbf{U}_k)_{i*}$ to denote the i-th row of $\mathbf{U}_k$ as a row vector.

2.1 CluStrat

CluStrat provides an LD based clustering framework to capture the population structure and tests for association within each cluster, as described in Algorithm 1.

Algorithm 1 Structure informed clustering to correct for population stratification

Input: Genotype matrix $\mathbf{X} \in \mathbb{R}^{m \times n}$, trait vector $y \in \mathbb{R}^m$, p-value threshold p, number of clusters k

Output: Set of significantly associated SNPs M

1: $\mathbf{D} = MahDist(\mathbf{X})$
2: $\mathbf{C}$: Cluster membership vector (output of agglomerative hierarchical clustering on $\mathbf{D}$, k clusters)
3: **for** $i = 1 \dots k$
4: $Y_i = y_{C_i}$ and $\mathbf{X}^{(C_i)} = \mathbf{X}_{C_i *}$
5: Find $\hat{\beta}_i^{ridge} = \left(\mathbf{X}^{(C_i)^\top}\mathbf{X}^{(C_i)} + \lambda I\right)^{-1}\mathbf{X}^{(C_i)^\top} Y_i$.
6: Obtain set of significant p-value indices P_i from $\hat{\beta}_i^{ridge}$.
7: **end for**
8: $P = \bigcup_{i \in C} P_i$ and get $\mathbf{X}^{(P_1)} = \mathbf{X}_{*P}$
9: Find $\hat{\beta}^{ridge} = \left(\mathbf{X}^{(P_1)^\top}\mathbf{X}^{(P_1)} + \lambda I\right)^{-1}\mathbf{X}^{(P_1)^\top} y$.
10: Obtain set of p-values P_2 for $\hat{\beta}^{ridge}$.
11: Return M, set of markers corresponding to significant p-values from P_2.

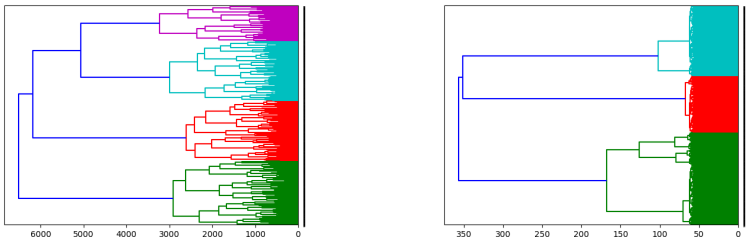
It computes the distance matrix $\mathbf{D}$ from the normalized genotype matrix $\mathbf{X}$ and performs AHC for a number of clusters k, selected by a cross validation.

For each cluster, it runs an association test using ridge regression and obtains p-values for each marker. Thereafter, it computes P_1 the union of intersections of significant associations across all clusters and select the corresponding markers from $\mathbf{X}$ to form $\mathbf{X}^{(P_1)}$. We can interpret this step as a scheme for variable selection. We run another association test with ridge regression on $\mathbf{X}^{(P_1)}$ to obtain M, the final set of significant associations for all meta-analysis p-values below p.

We now briefly discuss the use of the Mahalanobis distance at the first step of the proposed algorithm. In an arbitrarily structured breeding population, correlation between loci due to LD often results in block-diagonal structures in the genetic relationship matrix. Thus, it is important to account for this LD structure in the computation of the distance matrix [34]. One way to account for the LD structure is to use the squared Mahalanobis distance [32,36] (denoted as $\mathbf{D}$ in eqn. 1). Given a matrix $\mathbf{G} \in \mathbb{R}^{n \times n}$ which contains the covariance structure of LD (covariance due to LD between genetic markers), the LD-corrected GRM implementing the Mahalanobis distance is defined as

$$\mathbf{D} = \mathbf{X}\mathbf{G}^{-1}\mathbf{X}^{\top} \tag{1}$$

The Mahalanobis distance is useful in high-dimensional settings where the Euclidean distances fail to capture the true distances between observations (see Appendix A for relationships between Mahalanobis and Euclidean distances). It achieves this by taking the correlation structure between the features into account. We perform the association test in CluStrat by running ridge regression



(a) Mahalanobis Distance (b) Euclidean Distance

Fig. 1: Dendrograms obtained after running AHC with Ward's linkage on Pritchard-Stephens-Donnelly (PSD) model ($\alpha = \{0.1, 0.1, 0.1\}$) shows Mahalanobis distance with fine grained interactions between the individuals inside a cluster recovering population substructure and cryptic relatedness which Euclidean distance based GRM fails to recover.

on each cluster. The regularizer, λ, is chosen by 5-fold cross validation. It is

worth noting that we use ridge regression for each cluster as the number of samples is significantly smaller than the number of SNPs, thus making the overall system under-determined. We find the ridge-estimates as follows:

$$\hat{\beta}^{ridge} = (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}_n)^{-1} \mathbf{X}^\top y = \mathbf{X}^\top (\mathbf{X}\mathbf{X}^\top + \lambda \mathbf{I}_m)^{-1} y \tag{2}$$

We emphasize that the above operation is run for each cluster. We simply dropped the superscripts from $\mathbf{X}$ in the above equation for simplicity. Then, we find the standard error of the estimates in order to calculate the p-values associated with each marker to compute the significance of its association with the trait. The standard error for each marker i in ridge regression is given by

$$SE(\hat{\beta}_i^{ridge}) = \frac{\sigma}{\nu} \left\| (\mathbf{X}\mathbf{X}^\top + \lambda \mathbf{I}_m)^{-1} \mathbf{X}_{*i} \right\|_2. \tag{3}$$

Recall that $\mathbf{X}_{*i}$ is the i-th column of $\mathbf{X}$ and ν is known as the residual degrees of freedom. We set ν as shown in previous work [26] to the following,

$$\nu = m - c\mathbf{X}(\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^\top \tag{4}$$

for a small scalar constant $c > 0$.

For biobank-scale datasets requiring terabytes of memory, computing the standard error can be a challenge. However, we can use random projection based approaches to sketch the input matrix $\mathbf{X}$ in order to approximate the standard error for each marker. This is indeed a novel contribution of our approach. We delegate details to Appendix A. We do note that our work is heavily based on previous work on Randomized Linear Algebra (RLA) [19,22,20,50]). To the best of our knowledge, this is the first approximation of the standard error in penalized regression using a sketching based framework and is of independent interest; see also [11] for related work.

2.2 Computing Mahalanobis Distance

Mahalanobis distance is known to be connected to statistical leverage [47]. We discuss the connection between a regularized version of the Mahalanobis distance and a regularized notion of statisical leverage scores below. We first note that the Mahalanobis distance is invariant to linear transformations, which means that the standard normalizations of the genotype matrix $\mathbf{X}$ do not affect the Mahalanobis distance between two vectors. Recall the definition of the Mahalanobis distance between samples i and j:

$$\mathbf{D}(\mathbf{X}_{i*}, \mathbf{X}_{j*}) = (\mathbf{X}_{i*} - \mathbf{X}_{j*})\mathbf{G}^{-1}(\mathbf{X}_{i*} - \mathbf{X}_{j*})^\top. \tag{5}$$

Now, recall that the rank-k leverage scores of the genotype matrix $\mathbf{X} \in \mathbb{R}^{m \times n}$ with $n \gg m$ are defined by the row norms of the matrix of its top k left singular vectors $\mathbf{U}_k \in \mathbb{R}^{m \times k}$. Let $(\mathbf{U}_k)_{i*}$ denote the i-th row of the matrix $\mathbf{U}_k$. Then the rank-k statistical leverage scores of the rows of $\mathbf{A}$, for $i \in 1, \cdots, n$ are given by

$$\mathbf{H}_i = \|(\mathbf{U}_k)_{i*}\|_2^2.$$

Similarly, the rank-k (i,j)-th cross-leverage score, $\mathbf{H}_{ij}$, is equal to the dot product of the i-th and j-th rows of $\mathbf{U}_k$, namely

$$\mathbf{H}_{ij} = \langle (\mathbf{U}_k)_{i*}, (\mathbf{U}_k)_{j*} \rangle. \tag{6}$$

Here, $\mathbf{H} \in \mathbb{R}^{m \times m}$ is the matrix of all leverage and cross-leverage scores. We note that $\mathbf{H}_i = \mathbf{H}_{ii} = \|(\mathbf{U}_k)_{i*}\|_2^2 = \left(\mathbf{U}_k \mathbf{U}_k^\top\right)_{ii}$ is a special case of the dot product in eqn. 6 for the diagonal leverage scores. We show that the Mahalanobis distance can be written in terms of the rank-k leverage and cross-leverage scores (see Appendix A for details on the relationship between Mahalanobis distance and leverage scores). Indeed, the final formulas are:

$$\mathbf{D}_i = \mathbf{D}(\mathbf{X}_{i*}, 0) = (m-1)\left(\mathbf{H}_i - \frac{1}{m}\right), \text{ and} \tag{7}$$

$$\mathbf{D}_{ij} = \mathbf{D}(\mathbf{X}_{i*}, \mathbf{X}_{j*}) = (m-1)(\mathbf{H}_i + \mathbf{H}_j - 2\mathbf{H}_{ij}). \tag{8}$$

Thus, we show that Mahalanobis distance between two vectors can be computed efficiently without storing or inverting $\mathbf{G}$, by the corresponding rank-k leverage and cross-leverage scores. By computing the rank-k Mahalanobis distance with respect to the top k-left singular vectors of the genotype matrix $\mathbf{X}$, we make this computation feasible for UK Biobank-scale datasets using methods such as `TeraPCA` [7] to approximate the matrix $\mathbf{U}_k$ accurately and efficiently.

Algorithm 2 MahDist : Compute Mahalanobis distance based GRM

Input: $\mathbf{X} \in \mathbb{R}^{m \times n}$ where $n > m$, k number of PCs to retain
Output: Mahalanobis GRM $\mathbf{D}$

1: Compute $\mathbf{U}_k$, the matrix of the top k left singular vectors of the genotype matrix $\mathbf{X}$
2: $\mathbf{H} = \mathbf{U}_k \mathbf{U}_k^\top$
3: $\mathbf{D}(\mathbf{X}_{i*}, \mathbf{X}_{j*}) = (m-1)\left(\mathbf{H}_{ii} + \mathbf{H}_{jj} + 2\mathbf{H}_{ij}\right)$
4: Return $\mathbf{D}$

2.3 Agglomerative Hierarchical Clustering

We performed AHC using the LD induced Mahalanobis distance with a varying number of clusters. We set the expected number of clusters to $d+q$ where d is the number of populations in the data and q is a user-defined range. We performed a five-fold crossvalidation to choose the optimal number of clusters and retain the cluster which maximizes the intersection of associations across all the clusters. The observed number of clusters is obtained by the inconsistency method of pruning according to the depth of the dendrogram. We note that for the simple case where q is set to zero, the clustering essentially attempts to recover the

populations. In practice, we observed that the number of qualitative clusters obtained by running PCA on the genotype data serves as a good heuristic for the number of user defined clusters using the AHC procedure.

2.4 Datasets

We generated an extensive set of simulations with challenging scenarios to demonstrate the robustness to different real-world scenarios and power to detect few spurious associations.

We simulated and analyzed 100 GWAS datasets from a quantitative trait model (and it's equivalent binary trait model using the Odds Ratio (OR) as the classifier for disease status from the continuous variable y) based on previous work [41].

$$y_j = \alpha + \sum_{i=1}^{m} \beta_i \mathbf{X}_{ij} + \lambda_j + \epsilon_j \tag{9}$$

where β_i is the genetic effect of SNP i on the trait, λ_j is the random non-genetic effect and ϵ_j is the random noise variation for individual j. $\mathbf{X}_{ij}$ is the i^{th} marker for the j^{th} individual and $y \in \mathbb{R}^m$ is the trait response variable (binary or continuous). For the genotype data, we simulated allele frequencies using (i) Balding-Nichols (BN) model [4] based on allele-frequency and F_{ST} estimates calculated on the HapMap data set, (ii) three different levels of admixture by varying the parameter α from {0.01,0.1,0.5} in Pritchard-Stephens-Donnelly model (PSD) [39] and (iii) structure estimated from 1000 Genomes Project (TGP) [3] (see Appendix A for details). To capture real world population structure, we

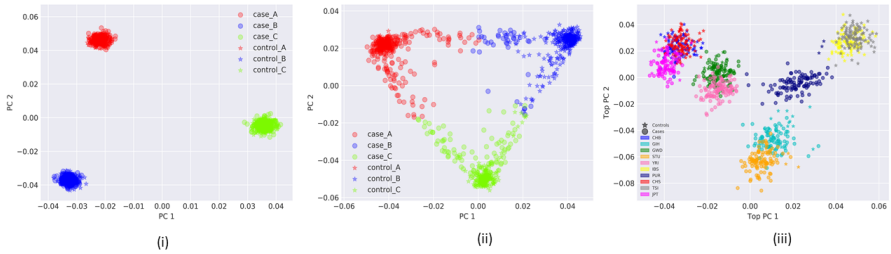
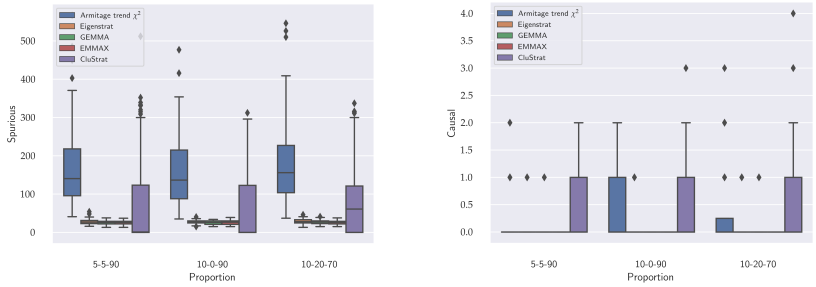


Fig. 2: Projection of the samples from three populations simulated from (i) BN (ii) PSD ($\alpha = \{0.1, 0.1, 0.1\}$) and (iii) TGP model on the top two axes of variation.

applied CluStrat on the Parkinson's Disease (PD) data available from the The Wellcome Trust Case Control Consortium (WTCCC2) study containing 4706 individuals (2837 controls and 1869 cases) across 517,672 SNPs. After performing quality control (details in Appendix) and pruning for LD between variants, we obtained 99,631 markers.

3 Results

We applied CluStrat to 30 simulation scenarios, modelling variable proportions of true genetic effect and admixture and compared its performance to standard population structure correction approaches such as EIGENSTRAT [38], GEMMA [52], and EMMAX [28]. We compared all the methods on all of the above scenarios with the p-value threshold set to $p = \frac{25}{n_i} = 0.0025$; here n_i is the number of SNPs, which was set to 10,000. The expected number of spurious association as mentioned in [41] is $n_0 \times p$ where n_0 is equal to n minus number of causal SNPs (10 in our case). In all of the above scenarios, CluStrat outperformed the standard approaches in detecting the true causal variants while reporting slightly more spurious associations in the simulation scenarios.



(a) Spurious associations (b) Causal associations

Fig. 3: Box plots for spurious and causal associations on the TGP model shows Armitage trend χ^2 has the maximum number of spurious associations (Appendix figure 4 and 6). CluStrat outperforms both the methods in this scenario by detecting two fold more causal loci while allowing slightly more spurious associations.

The BN and PSD model simulates scenarios with unrelated isolated populations (Figure 2 (i) and (ii)). The samples when projected on the top two PCs clearly resemble three isolated clusters with no connections between them in BN and with admixed populations between the clusters in PSD, respectively. This serves as a "base case" for arbitrarily structured population with and without admixture. Armitage trend χ^2 test with no population structure correction renders almost half of the SNPs in the simulation study as true associations, resulting in a vast number of spurious associations, clearly highlighting the need for population structure correction. PCA or LMM based approaches on the other hand return roughly the expected number of spurious associations as also shown in prior work [38]. Yet, PCA and LMM approaches are very stringent and detect zero causal SNPs in almost our experiments (Appendix figure 4, 6 and 7). CluStrat, however, strikes a balance between the two: it generates more spurious

associations than the expected value, though far less than the Armitage trend χ^2 test, and recovers almost similar number of causal SNPs. This shows that in an ideal case of population stratification correction, CluStrat can identify more causal SNPs mainly due to the use of the Mahalanobis distance and the simple clustering algorithm that we propose.

The TGP model is a more realistic model, drawing genotypes from allele frequency distributions from the 1000 Genomes Phase 3 dataset [3]. Projection of genotypes drawn from the 1000 Genomes (TGP) dataset on the top two axes of variations shows the distribution of samples across the world (see Figure 2 (iii). CluStrat (Figure 3) captures two-to-three fold increase in detection of causal variants while allowing for slightly more number of spurious associations. This shows that structure informed clustering of the genotype data followed by regularized association tests outperforms genotype and phenotype adjustments with the top k PCs, which is what EIGENSTRAT and LMMs often do.

Applying CluStrat on WTCCC2's with a p-value threshold set to 10^{-7} we found a host of associated regions mapped to genes previously known to be associated with PD in literature. Our strongest associated loci rs10177996 (p-value $= 2.2 \times 10^{-16}$) maps to WNT10 in the Wignless-type MMTV integration site (Wnt) signalling cascade which has emerged as a very important pathway in major neurodegenerative pathologies including PD [29]. Another significant loci appeared to be in Chromosome 6 at rs176713 associated to the transcriptional inhibitor BACH2 which is known to interfere with Nrf2 function which when activated is a promising protective mechanism for progressive neurodegeneration in PD [27]. Other significant associated markers which are mapped to genes shown to be associated in previous work are SLC11A1 [6], UNC5C [30], MAP2 [16], EFNA5 [46], NR4A2, GRM7 [24], CNTN2, etc. (see Table 1 in Appendix B for details). We conclude that CluStrat not only works better in simulation scenarios but can also replicate previously recorded associations in real data sets such as PD.

4 Discussion

CluStrat provides a structure informed clustering approach to correct for population stratification in GWAS. In our experiments, we verified the power of our approach in a variety of simulated data and observed that CluStrat outperforms the widely used EIGENSTRAT, GEMMA, and EMMAX methods in all settings, by detecting two to four times more causal SNPs. Adversely, our approach detects more spurious associations than standard approaches; however, it still performs much better than the uncorrected Armitage trend χ^2 tests. Principal component based methods have been under scrutiny recently as independent studies [5,40] on the UK Biobank [8] failed to replicate the genetic associations of heritable height in Europeans, where a positive selection signal was observed in a north to south gradient [15,42,35] in the GIANT [49] cohort. These studies attributed the failure to replicate the results to cryptic relatedness among individuals, which PCA-based approaches for population stratification correction do

not always capture. CluStrat provides a fine structure-based clustering approach to tackle cryptic relatedness and ancestral differences among the individuals between and within populations.

We chose to use the low-rank Mahalanobis distance metric in CluStrat because it captures the LD-induced structure information in the GRM. We established a link between the low-rank Mahalanobis distance and the low-rank leverage/cross-leverage scores, which allows us to get around the storage and computational bottlenecks of Mahalanobis distance. Prior work [34] computed the Mahalanobis distance by randomly subsampling a small number of SNPs to estimate the covariance matrix and circumvent to computational time and space requirements. Mahalanobis distance is also shown to remove bias in heritability estimates in presence of LD, therefore finding true causal variants [31]. An interesting topic for future work is to make CluStrat even faster by approximating the leverage and cross-leverage scores as shown in [19]. We also want to explore meta-analysis strategies to combine p-values from each cluster and obtain a cumulative significance across clusters as done in GWA studies. We showed that the Mahalanobis distance performs better (Figure 1) in capturing cryptic relatedness than the Euclidean distance based GRM. In upcoming work, we intend to evaluate CluStrat on the UK Biobank data to explore whether it succeeds or fails to replicate the north to south gradient of positive selection of height in Europeans [5,40]. CluStrat being a clustering based strategy is computationally slower than the PCA based approaches, we intend to explore Mahalanobis distance based GRM in other statistical models for association tests. Another future direction for CluStrat is to extend it to compute Polygenic Risk Scores (PRS) on a discovery or validation dataset and compare it with widely used packages such as PRSice2 [10] and LDPred [44], which compute PRS from GWAS summary statistics as well as raw genotypes.

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. We evaluated CluStrat on a host of simulation scenarios for arbitrarily structured populations with and without admixture. The choice of the number of clusters does not change the results drastically and one can use the number of broad clusters that are visually apparent when plotting the data on the top two or three principal components as a initial choice of clusters. We implemented a five-fold cross validation approach to obtain the optimal choice for the number of clusters and the regularization parameters. We concluded that CluStrat outperforms PCA or LMM based population stratification correction techniques in a variety of simulated datasets.

5 Acknowledgements

AB carried out this work as a part of his PhD dissertation in the Computer Science Department, Purdue University, West Lafayette, IN, USA.

References

1. Achlioptas, D.: Database-friendly random projections: Johnson-lindenstrauss with binary coins. Journal of computer and System Sciences **66**(4), 671–687 (2003)
2. Astle, W., Balding, D.J., et al.: Population structure and cryptic relatedness in genetic association studies. Statistical Science **24**(4), 451–471 (2009)
3. Auton, A., Abecasis, G.R., Altshuler, D.M., et al.: A global reference for human genetic variation. Nature **526**(7571), 68–74 (2015). https://doi.org/10.1038/nature15393, `http://www.nature.com/doifinder/10.1038/nature15393`
4. Balding, D.J., Nichols, R.A.: A method for quantifying differentiation between populations at multi-allelic loci and its implications for investigating identity and paternity. Genetica **96**(1-2), 3–12 (1995)
5. Berg, J.J., Harpak, A., Sinnott-Armstrong, N., et al.: Reduced signal for polygenic adaptation of height in uk biobank. eLife **8**, e39725 (2019)
6. Blackwell, J.M., Goswami, T., Evans, C.A., et al.: Slc11a1 (formerly nramp1) and disease resistance: Microreview. Cellular microbiology **3**(12), 773–784 (2001)
7. 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 (2019)
8. Bycroft, C., Freeman, C., Petkova, D., et al.: The uk biobank resource with deep phenotyping and genomic data. Nature **562**(7726), 203 (2018)
9. Charikar, M., Chen, K., Farach-Colton, M.: Finding frequent items in data streams. In: International Colloquium on Automata, Languages, and Programming. pp. 693–703. Springer (2002)
10. Choi, S.W., O'Reilly, P.F.: Prsice-2: Polygenic risk score software for biobank-scale data. GigaScience **8**(7), giz082 (2019)
11. Chowdhury, A., Yang, J., Drineas, P.: An iterative, sketching-based framework for ridge regression. In: International Conference on Machine Learning. pp. 988–997 (2018)
12. Clarkson, K.L., Woodruff, D.P.: Low-rank approximation and regression in input sparsity time. Journal of the ACM (JACM) **63**(6), 54 (2017)
13. Cohen, M.B., Nelson, J., Woodruff, D.P.: Optimal approximate matrix product in terms of stable rank. arXiv preprint arXiv:1507.02268 (2015)
14. Consortium, I.H.., et al.: Integrating common and rare genetic variation in diverse human populations. Nature **467**(7311), 52 (2010)
15. Coop, G., Pickrell, J.K., Novembre, J., et al.: The role of geography in human adaptation. PLOS Genetics **5**(6), 1–16 (06 2009). https://doi.org/10.1371/journal.pgen.1000500, `https://doi.org/10.1371/journal.pgen.1000500`
16. D'Andrea, M.R., Ilyin, S., Plata-Salaman, C.R.: Abnormal patterns of microtubule-associated protein-2 (map-2) immunolabeling in neuronal nuclei and lewy bodies in parkinson's disease substantia nigra brain tissues. Neuroscience letters **306**(3), 137–140 (2001)
17. Demontis, D., Walters, R.K., Martin, J., et al.: Discovery of the first genome-wide significant risk loci for attention deficit/hyperactivity disorder. Nature genetics **51**(1), 63 (2019)
18. Devlin, B., Roeder, K.: Genomic control for association studies. Biometrics **55**(4), 997–1004 (1999)

19. Drineas, P., Magdon-Ismail, M., Mahoney, M.W., Woodruff, D.P.: Fast approximation of matrix coherence and statistical leverage. Journal of Machine Learning Research **13**(Dec), 3475–3506 (2012)
20. Drineas, P., Mahoney, M.W., Muthukrishnan, S.: Sampling algorithms for l 2 regression and applications. In: Proceedings of the seventeenth annual ACM-SIAM symposium on Discrete algorithm. pp. 1127–1136. Society for Industrial and Applied Mathematics (2006)
21. Drineas, P., Mahoney, M.W., Muthukrishnan, S.: Relative-error cur matrix decompositions. SIAM Journal on Matrix Analysis and Applications **30**(2), 844–881 (2008)
22. Drineas, P., Mahoney, M.W., Muthukrishnan, S., Sarlós, T.: Faster least squares approximation. Numerische mathematik **117**(2), 219–249 (2011)
23. Ewens, W.J., Spielman, R.S.: The transmission/disequilibrium test: history, subdivision, and admixture. American journal of human genetics **57**(2), 455 (1995)
24. Fisher, N.M., Seto, M., Lindsley, C.W., Niswender, C.M.: Metabotropic glutamate receptor 7: A new therapeutic target in neurodevelopmental disorders. Frontiers in molecular neuroscience **11**, 387 (2018)
25. Hao, W., Song, M., Storey, J.D.: Probabilistic models of genetic variation in structured populations applied to global human studies. Bioinformatics **32**(5), 713–721 (2015)
26. Hastie, T.J.: Generalized additive models. In: Statistical models in S, pp. 249–307. Routledge (2017)
27. Johnson, D.A., Johnson, J.A.: Nrf2—a therapeutic target for the treatment of neurodegenerative diseases. Free Radical Biology and Medicine **88**, 253–267 (2015)
28. Kang, H.M., Sul, J.H., Service, S.K., et al.: Variance component model to account for sample structure in genome-wide association studies. Nature genetics **42**(4), 348 (2010)
29. L'Episcopo, F., Tirolo, C., Caniglia, S., et al.: Targeting wnt signaling at the neuroimmune interface for dopaminergic neuroprotection/repair in parkinson's disease. Journal of molecular cell biology **6**(1), 13–26 (2014)
30. Li, Q., Wang, B.L., Sun, F.R., et al.: The role of unc5c in alzheimer's disease. Annals of translational medicine **6**(10) (2018)
31. Ma, R., Dicker, L.H.: The mahalanobis kernel for heritability estimation in genome-wide association studies: fixed-effects and random-effects methods. arXiv preprint arXiv:1901.02936 (2019)
32. Mahalanobis, P.C.: On the generalized distance in statistics. National Institute of Science of India (1936)
33. Marchini, J., Cardon, L.R., Phillips, M.S., Donnelly, P.: The effects of human population structure on large genetic association studies. Nature Genetics **36**(5), 512–517 (2004). https://doi.org/10.1038/ng1337, `https://doi.org/10.1038/ng1337`
34. 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)
35. Mathieson, I., Lazaridis, I., Rohland, N., et al.: Genome-wide patterns of selection in 230 ancient eurasians. Nature **528**(7583), 499 (2015)
36. Mitchell, A.F., Krzanowski, W.J.: The mahalanobis distance and elliptic distributions. Biometrika **72**(2), 464–467 (1985)
37. Patterson, N., Price, A.L., Reich, D.: Population structure and eigenanalysis. PLoS genetics **2**(12), e190 (2006)

38. Price, A.L., Patterson, N.J., Plenge, R.M., et al.: Principal components analysis corrects for stratification in genome-wide association studies. Nature genetics **38**(8), 904 (2006)
39. Pritchard, J.K., Stephens, M., Donnelly, P.: Inference of population structure using multilocus genotype data. Genetics **155**(2), 945–959 (Jun 2000), `http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1461096/`, 10835412[pmid]
40. 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)
41. Song, M., Hao, W., Storey, J.D.: Testing for genetic associations in arbitrarily structured populations. Nature genetics **47**(5), 550 (2015)
42. Turchin, M.C., Chiang, C.W., Palmer, C.D., et al.: Evidence of widespread selection on standing variation in europe at height-associated snps. Nature genetics **44**(9), 1015 (2012)
43. Uricchio, L.H., Kitano, H.C., Gusev, A., Zaitlen, N.A.: An evolutionary compass for detecting signals of polygenic selection and mutational bias. Evolution letters **3**(1), 69–79 (2019)
44. Vilhjálmsson, B.J., Yang, J., Finucane, H.K., et al.: Modeling linkage disequilibrium increases accuracy of polygenic risk scores. The American Journal of Human Genetics **97**(4), 576–592 (2015)
45. Visscher, P.M., Wray, N.R., Zhang, Q., et al.: 10 years of gwas discovery: Biology, function, and translation. The American Journal of Human Genetics **101**(1), 5 – 22 (2017). https://doi.org/https://doi.org/10.1016/j.ajhg.2017.06.005, `http://www.sciencedirect.com/science/article/pii/S0002929717302409`
46. Wang, T., Chen, J., Tang, C.X., Zhou, X.Y., Gao, D.S.: Inverse expression levels of ephrina3 and ephrina5 contribute to dopaminergic differentiation of human sh-sy5y cells. Journal of Molecular Neuroscience **59**(4), 483–492 (2016)
47. Weiner, I.B.: Handbook of psychology, history of psychology, vol. 1. John Wiley & Sons (2003)
48. Welter, D., MacArthur, J., Morales, J., et al.: The nhgri gwas catalog, a curated resource of snp-trait associations. Nucleic acids research **42**(D1), D1001–D1006 (2013)
49. Wood, A.R., Esko, T., Yang, J., et al.: Defining the role of common variation in the genomic and biological architecture of adult human height. Nature genetics **46**(11), 1173 (2014)
50. Woodruff, D.P., et al.: Sketching as a tool for numerical linear algebra. Foundations and Trends® in Theoretical Computer Science **10**(1–2), 1–157 (2014)
51. Yang, J., Benyamin, B., McEvoy, B.P., et al.: Common snps explain a large proportion of the heritability for human height. Nature genetics **42**(7), 565 (2010)
52. Zhou, X., Stephens, M.: Genome-wide efficient mixed-model analysis for association studies. Nature genetics **44**(7), 821 (2012)

Appendix A

Data simulator

The complete simulation study on quantitative traits with population structure latent variable is constructed in 5 different ways for 3 different proportions of variance among genetic effects, non-genetic effects and random noise, all of which contributing to the trait. We simulated 100 independent datasets containing $m = 1,000$ individuals and $n = 100,000$ markers from a quantitative trait model 9. Let Z be a latent variable which captures environmental factors contributed by population structure. Equation 9 allows interdependence of structure, lifestyle and environment. We assume $\mathbf{E}[\epsilon_j|z_j] \sim \mathcal{N}(0, \sigma^2(z_j))$ allowing for heteroskedasticity of the random noise variation [41]. Therefore, $x^j = (x_{1j}, x_{2j}, \cdots, x_{mj})^\top$, λ_j and σ^2 can be thought of as functions of z_j where $Z = (z_1, z_2, \cdots, z_m)$. λ_j is unspecified but along with z_j, they are assumed to be dependent, random variables. Thus, the population genetic model is dependent on the structure variable z_j for each individual. We define the corresponding binary trait model as

$$\log\left(\frac{\Pr(y_j = 1)}{\Pr(y_j = 0)}\right) = \alpha + \sum_{i=1}^{m} \beta_i x_{ij} + \lambda_j \tag{10}$$

using the Odds Ratio (OR) as the classifier for disease status from the continuous variable y.

The complete simulation study on quantitative traits with population structure latent variable is constructed in 5 different ways for 3 different proportions of variance among genetic effects, non-genetic environmental effects and random noise, all of which contributing to the trait. Therefore $\mathbf{Var}\left[\sum_{i=1}^{n} \beta_i x_{ij}\right]$, $\mathbf{Var}\left[\sum_{j=1}^{n} \lambda_j\right]$ and $\mathbf{Var}[\epsilon_j]$ are assigned in proportions of (5%,5%,90%), (10%,0%,90%) and (10%,20%,70%), respectively. Thus, we varied the amount of genetic contribution to the trait for each simulation scenarios and capture variable amounts of population structure confounding. We simulated ten truly associated SNPs whose effect sizes were distributed according to a Normal distribution and we set $\beta_i = 0$ for all other non-causal SNPs.

The genotype matrix $\mathbf{X} \in \mathbb{R}^{m \times n}$ consisting of the simulated allele frequencies was simulated using the algorithm from a previous study [25,41]. Specifically, we set $\mathbf{F} = \mathbf{TS}$ where $\mathbf{T} \in \mathbb{R}^{m \times d}$ and $\mathbf{S} \in \mathbb{R}^{d \times n}$ where $d \leq n$ is the number of population groups. $\mathbf{S}$ is the matrix containing the population groups encompassing the structure for the individuals shared across all SNPs. On the other hand, $\mathbf{T}$ characterizes how the structure is manifested in the allele frequencies of each SNP [25]. Finally, projecting $\mathbf{S}$ onto the column space of $\mathbf{T}$ we obtain the allele frequency matrix $\mathbf{F}$. We sample $\mathbf{X}$ as a special case of $\mathbf{F}$ for Balding-Nichols (BN), Pritchard-Stephens-Donelly (PSD) and TGP (1000 Genomes Project), respectively. We formed $\mathbf{T}$ and $\mathbf{S}$ for the above 5 simulations with 3 scenarios each and continuous traits, resulting in, 15 different evaluation scenarios each for

continuous and binary traits. The algorithm for constructing $\mathbf{T}$ and $\mathbf{S}$ is detailed in reference [25,41].

For BN, the allele frequency matrix is simulated from the HapMap phase 3 dataset [14] using three unrelated populations. The final genotype matrix, $\mathbf{X}$, is drawn independently at random from the Binomial distribution with parameters n set to 2, denoting the allele status (0,1 or 2) corresponding to homozygous major/minor or heterozygous with probability p set to the simulated allele frequency for each individual-SNP pair. For PSD, the allele frequency matrix was drawn from the BN frequency distribution. However, it differs from BN in simulating $\mathbf{S}$ by i.i.d draws from Dirichlet distribution with varying α which denotes the parameter influencing the relatedness between the individuals. We show results for $\alpha = \{0.01, 0.1, 0.5\}$ here and conducted simulations on a wide range of α values from 0.01 to 0.5.

Real dataset

To capture real world population structure, we applied CluStrat on the Parkinson's Disease (PD) data available from the The Wellcome Trust Case Control Consortium (WTCCC2) study containing 4706 individuals (2837 controls and 1869 cases) across 517,672 SNPs. After performing quality control by filtering for genotyping rate lower than 99%, MAF less than 0.01 and Hardy-Weinberg equilibrium less than 0.001 and pruning for LD between variants higher than 0.2 squared correlation we obtained 99,631 markers.

Distance metrics for Hierarchical clustering

CluStrat computes the distance matrix $\mathbf{D}$ from $\mathbf{X}$ to perform the AHC. The choice of distance metric is user defined. However, we choose the distance metric based on LD induced distances to capture the cryptic relatedness between individuals in a population which is not otherwise captured by other stratification methods. We use the normalized genotype matrix $\mathbf{X}$ following the standard normalization procedure by minor allele frequency of each marker. Let us consider the unscaled GRM which captures the Euclidean distances as $\mathbf{D} = \mathbf{X}\mathbf{X}^\top$ and let $\mathbf{I} \in \mathbb{R}^{n \times n}$ (in the order of the number of markers n). Thus $\mathbf{D}$ can be rewritten as

$$\mathbf{D} = \mathbf{X}\mathbf{I}\mathbf{X}^\top \tag{11}$$

Thus we can see the unscaled GRM as the same weighting on the diagonal for all markers. In an arbitrarily structured breeding population, there exists correlation between loci due to linkage resulting in varying values along the diagonal or a block-diagonal structure in the GRM. Thus, it is important to account for this LD covariance structure in the computation of the GRM [34]. One way to account for the LD structure in GRM is to use the squared Mahalanobis distance [32,36] (denoted as $\mathbf{D}$ for simplification). Given a matrix $\mathbf{G} \in \mathbb{R}^{n \times n}$ which contains the covariance structure of LD (covariance due to markers), then the

LD-corrected GRM with Mahalanobis distance is defined as

$$\mathbf{D} = \mathbf{X}\mathbf{G}^{-1}\mathbf{X}^\top \tag{12}$$

The RHS of equation 1 represents the squared multivariate Mahalanobis distance between individuals. Mahalanobis distance is useful in a high-dimensional setting where the Euclidean distances fail to capture the true distances between observations. It achieves this by taking correlation between the features captured in the SNP covariance matrix into account. The Cholesky factorization of the covariance matrix $\mathbf{G} = \mathbf{L}\mathbf{L}^\top$ where $\mathbf{L}$ is the lower diagonal matrix known as the Cholesky factor of $\mathbf{G}$ [34]. We can represent equation 1 as

$$\begin{aligned}
\mathbf{X}\mathbf{G}^{-1}\mathbf{X}^\top &= \mathbf{X}\left(\mathbf{L}\mathbf{L}^\top\right)^{-1}\mathbf{X}^\top \\
&= \mathbf{X}(\mathbf{L}^\top)^{-1}(\mathbf{L})^{-1}\mathbf{X}^\top \\
&= \left(\mathbf{X}(\mathbf{L}^{-1})^\top\right)\left(\mathbf{L}^{-1}\mathbf{X}\right)^\top \\
&= \left(\mathbf{L}^{-1}\mathbf{X}^\top\right)^\top\left(\mathbf{L}^{-1}\mathbf{X}^\top\right) \\
&= \mathbf{Q}^\top\mathbf{Q}
\end{aligned}$$

$\mathbf{Q} = \mathbf{L}^{-1}\mathbf{X}^\top$ represents the transformed variables and $\mathbf{Q}^\top\mathbf{Q}$ is the squared Euclidean distance between the transformed variables. Thus, Mahalanobis distance accounts for covariance between variables by transforming the data into an uncorrelated form and computing the euclidean distances between them.

Mahalanobis Distance and Leverage Scores Mahalanobis distance is known to be connected to statistical leverage [47], which is extended in the RandNLA framework as leverage scores. We show this relationship by first noting that Mahalanobis distance is invariant to linear transformations, which means the Mahalanobis distance between two vectors,

$$\mathbf{D}(\mathbf{X}_{i*}, \mathbf{X}_{j*}) = (\mathbf{X}_{i*} - \mathbf{X}_{j*})\mathbf{G}^{-1}(\mathbf{X}_{i*} - \mathbf{X}_{j*})^\top \tag{13}$$

can have zero means for each vector. In our genotype matrix, $\mathbf{X} \in \mathbb{R}^{m\times n}$, we have n markers and m observations. The design matrix $\mathbf{X}$ on which we intend to fit the model, however, must contain an intercept and thus we refer to $\mathbf{X}$ here as the design matrix containing the intercept column followed by one column for each SNP for all the individuals in rows. Furthermore, as we compute the Mahalanobis distance with respect to the *low-rank* genotype matrix $\mathbf{X}_k$, we only consider the *low-rank leverage scores* (rather than the leverage scores of the original matrix $\mathbf{X}$) which are essentially the diagonal elements of the following projection-matrix:

$$\mathbf{H} = \mathbf{X}_k\left(\mathbf{X}_k^\top\mathbf{X}_k\right)^{-1}\mathbf{X}_k^\top \tag{14}$$

and similarly, the off-diagonal elements of $\mathbf{H}$ are called *cross-leverage scores* of $\mathbf{X}_k$. Now, we will give a clean connection between Mahalanobis distance and these leverage and cross-leverage scores.

First, consider the diagonal elements of $\mathbf{H}$ *i.e.* when $i = j$, we have

$$\mathbf{H}_{ii} = (1; \mathbf{X}_{k_{i*}}) \left(\mathbf{X}_k^\top \mathbf{X}_k\right)^{-1} (1; \mathbf{X}_{k_{i*}})^\top . \tag{15}$$

Exploiting the structure of $\left(\mathbf{X}_k^\top \mathbf{X}_k\right)^{-1}$, we can reformulate it in terms of a block matrix as follows

$$\mathbf{X}_k^\top \mathbf{X}_k = m \begin{pmatrix} 1 & \mathbf{0}^\top \\ \mathbf{0} & \mathbf{C} \end{pmatrix}$$

where $\mathbf{C}_{ij} = \frac{1}{m} \sum_{\ell=1}^{m} \mathbf{X}_{k_{\ell i}} \mathbf{X}_{k_{\ell j}} = \frac{m-1}{m} Cov(\mathbf{X}_{k_{*i}}, \mathbf{X}_{k_{*j}}) = \frac{m-1}{m} \mathbf{\Sigma}_{ij}$. $\mathbf{\Sigma}$ here is the corresponding sample covariance matrix. Thus,

$$\left(\mathbf{X}_k^\top \mathbf{X}_k\right)^{-1} = \frac{1}{n} \begin{pmatrix} 1 & \mathbf{0}^\top \\ \mathbf{0} & \mathbf{C}^{-1} \end{pmatrix} = \begin{pmatrix} \frac{1}{n} & \mathbf{0}^\top \\ \mathbf{0} & \frac{1}{n-1} \mathbf{\Sigma^{-1}} \end{pmatrix}$$

From Equation 15 we obtain

$$\mathbf{H}_i = (1; \mathbf{X}_{k_{i*}}) \begin{pmatrix} \frac{1}{m} & \mathbf{0}^\top \\ \mathbf{0} & \frac{1}{m-1} \mathbf{\Sigma^{-1}} \end{pmatrix} (1; \mathbf{X}_{k_{i*}})^\top \tag{16}$$

$$= \frac{1}{m} + \frac{1}{m-1} \mathbf{X}_{k_{i*}} \mathbf{\Sigma^{-1}} \mathbf{X}_{k_{i*}}^\top \tag{17}$$

$$= \frac{1}{m} + \frac{1}{m-1} \mathbf{D} \left(\mathbf{X}_{k_{i*}}, 0\right) \tag{18}$$

Solving for

$$\mathbf{D}_i = \mathbf{D}(\mathbf{X}_{k_{i*}}, 0)$$

yields,

$$\mathbf{D}_i = (m-1) \left(\mathbf{H}_i - \frac{1}{m}\right)$$

Similarly, we can prove the cross-leverage scores

$$\mathbf{H}_{ij} = \frac{1}{m} + \frac{1}{m-1} \mathbf{X}_{k_{i*}} \mathbf{\Sigma^{-1}} \mathbf{X}_{k_{j*}} \tag{19}$$

To prove the relationship of $\mathbf{H}_{ij}$ with $\mathbf{D}_{ij}$ we see,

$$\begin{aligned} \mathbf{D}(\mathbf{X}_{k_{i*}}, \mathbf{X}_{k_{j*}}) &= (\mathbf{X}_{k_{i*}} - \mathbf{X}_{k_{j*}}^\top) \mathbf{\Sigma}^{-1} (\mathbf{X}_{k_{i*}} - \mathbf{X}_{k_{j*}}) \\ &= \mathbf{D}(\mathbf{X}_{k_{i*}}, 0) + \mathbf{D}(\mathbf{X}_{k_{j*}}, 0) - 2\mathbf{X}_{k_{i*}} \mathbf{\Sigma}^{-1} \mathbf{X}_{k_{j*}} \\ &= (m-1)(\mathbf{H}_i - \frac{1}{m}) + (m-1)(\mathbf{H}_j - \frac{1}{m}) - 2(m-1)(\mathbf{H}_{ij} - \frac{1}{m}) \\ &= (m-1)(\mathbf{H}_i + \mathbf{H}_j - 2\mathbf{H}_{ij}) \end{aligned}$$

If we take $\mathbf{X}_{k_{i*}} = \mathbf{X}_{k_{j*}}$ then we find $\mathbf{D}(\mathbf{X}_{k_{i*}}, \mathbf{X}_{k_{j*}}) = 0$. Thus, we show that Mahalanobis distance between two vectors can be computed by the corresponding vector's leverage scores.

One of the key computational bottlenecks of Mahalanobis distance is computing the inverse of the SNP covariance matrix $\mathbf{G}$ as required in Equation 1. In real datasets, with the improvements in genotyping and sequencing technologies, the number of SNPs can be in the millions, thereby making $\mathbf{G}$ in the order of million times million and infeasible to store in secondary memory. Here, we propose the first approximation of Mahalanobis distance by computing leverage and cross-leverage scores in a faster and efficient way. As we have shown in Equation 19 and 16 following up from previous work [47], Mahalanobis distance can be written in terms of leverage scores. Advances in RandNLA community have brought about faster computations for leverage scores as well as cross-leverage scores; hence, we can compute approximations to these scores using random sampling algorithms with theoretical guarantees [19]. For our purposes of demonstrating the proof-of-concept, we work with simulated data as described above for 1,000 individuals and 500,000 SNPs which could be feasibly processed in a personal workstation to compute the deterministic leverage and cross-leverage scores.

Computing leverage and cross-leverage scores. In fact, we do not need to compute the *rank-k* leverage and cross-leverage scores exactly. Using the idea of [19,12], they can be well-approximated in a much faster way with high-probability. In particular, computing m row-leverage scores takes time

$$\mathcal{O}\left(\text{nnz}(\mathbf{X}_k)\log n + k^3 \log^2 k + k^2 \log n\right) ,$$

where nnz means the non-zero entries of the matrix, and computation of the high-valued cross-leverage scores can be done in time

$$\mathcal{O}(\text{nnz}(\mathbf{X}_k)\log^3 n) .$$

Fast Computation of Standard errors

For biobank-scale data-sets requiring terabytes of memory, computing the standard error can be a challenge. However, we can use random projection based sketching matrices to find an approximate standard error for each marker by projecting the genotype matrix $\mathbf{X}$ on a sketching matrix $\mathbf{S} \in \mathbb{R}^{n\times r}$ to form a sketch $\mathbf{XS}$. We can rewrite the standard error in Equation 3 to find it's approximate as,

$$\tilde{SE}(\hat{\beta}_i) = \sigma^2 \| \left(\mathbf{XSS}^\top \mathbf{X}^\top + \lambda \mathbf{I}_m\right)^{-1} \mathbf{X}^{(i)} \|_2^2 \tag{20}$$

The sketched matrix $\mathbf{XS}$ generically has the same rank but much fewer columns than $\mathbf{X}$, satisfying $1 \leq r \leq min\{m, n\}$. Sketching, in general, is used to speed up solving systems of linear equations [20,22,12]. The sketching dimension, r, is directly proportional to the accuracy obtained by the approximate standard errors. Some prior knowledge of the design matrix, $\mathbf{X}$, helps determine the target rank, r,

that will result in satisfactory error guarantees. The sketching matrix, $\mathbf{S}$, can be chosen simply as i.i.d normal random variables with mean equal to zero and variance equal to $\frac{1}{r}$. There exists other ways to choose $\mathbf{S}$ based on random projections as shown in previous work involving Fast Johnson-Lindenstrauss Transform [1], Subsampled Randomized Hadamard Transform [20,21] and Count-Sketch matrices [9] from streaming setting involving faster computation with sparse matrices.

Time to compute eqn. (20)**.** Following the discussion as in [11], let the time to compute the sketch $\mathbf{XS} \in \mathbb{R}^{m\times s}$ be $T(\mathbf{X},\mathbf{S})$ which depends on the particular construction of $\mathbf{S}$. In order to invert the matrix $\mathbf{Q} = \mathbf{XSS}^\top\mathbf{X}^\top$, it suffices to compute the SVD of the matrix $\mathbf{XS}$. Notice that given the singular values of $\mathbf{XS}$, we can compute the singular values of $\mathbf{Q}$ and also notice that the left and right singular vectors of $\mathbf{Q}$ are the same as the left singular vectors of $\mathbf{XS}$. Interestingly, we do not need to compute $\mathbf{Q}^{-1}$. Instead, we can store it implicitly by storing the left (and right) singular vectors of $\mathbf{Q}$ along with its singular values, $\mathbf{\Sigma_Q}$. Then, we can compute all necessary matrix-vector products using this implicit representation of $\mathbf{Q}^{-1}$. Thus, inverting $\mathbf{Q}$ takes $\mathcal{O}\left(sm^2\right)$ time and this will eventually dominate the computation of all other matrix-vector products and the Euclidean-norm. Therefore, total running time to compute eqn. (20) is given by

$$\mathcal{T} = \mathcal{O}(sm^2) + T(\mathbf{X},\mathbf{S})$$

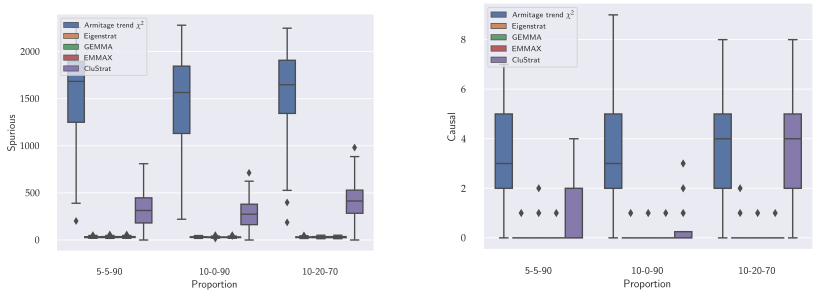
Clearly, specific constructions of the sketching matrix $\mathbf{S}$ will determine both s and $T(\mathbf{X},\mathbf{S})$, and therefore $\mathcal{T}$. For example, if $\mathbf{S}$ is a subsampled randomized Hadamard transform (SRHT) matrix, then we have, $T(\mathbf{X},\mathbf{S}) = \mathcal{O}(mn\log n)$ and $s = \Omega((m\log(m/\delta))/\varepsilon^2)$; therefore $\mathcal{T} = \mathcal{O}((m^3\log(m/\delta))/\varepsilon^2) + \mathcal{O}(mn\log n)$. Similarly, if $\mathbf{S}$ has sub-Gaussian entries, then $s = \mathcal{O}(m/\varepsilon^2)$ and $T(\mathbf{X},\mathbf{S}) = \mathcal{O}(m^2n)$; therefore $\mathcal{T} = \mathcal{O}(m^2n)$. Furthermore, if $\mathbf{S}$ is a count-sketch matrix of [12], then, in this case, $s = \Omega(\frac{m^2}{\varepsilon^2\delta})$ and $T(\mathbf{X},\mathbf{S}) = \text{nnz}(\mathbf{X})$. So, total running time $\mathcal{T} = \mathcal{O}(\text{nnz}(\mathbf{X}) + \frac{m^4}{\varepsilon^2\delta})$. Here, ε is the accuracy parameter and δ is the corresponding failure probability.

Note that sketching-dimension, s, for sub-Gaussian is optimal, but $T(\mathbf{X},\mathbf{S})$ takes much time. On the other hand, for count-sketch, $T(\mathbf{X},\mathbf{S})$ is much faster (only $\text{nnz}(\mathbf{X})$), but sketching-dimension is huge $\mathcal{O}(m^2/\varepsilon^2)$. In a recent work [13], the authors showed that we can actually use all the sketches discussed here in conjunction with each other to get the best performance both in terms of sketching-dimension as well as computation time. More precisely, if one set $\mathbf{S} = \mathbf{S}_1\mathbf{S}_2\mathbf{S}_3$ with $\mathbf{S}_1$ being the count-sketch, $\mathbf{S}_2$ being the SRHT and $\mathbf{S}_3$ being the sub-Gaussian , then $\mathbf{XS}$ will have $\mathcal{O}(m/\varepsilon^2)$ columns with running time $T(\mathbf{X},\mathbf{S}) = \mathcal{O}(\text{nnz}(A)) + \tilde{\mathcal{O}}\left(\varepsilon^{-\mathcal{O}(1)}\left(m^3 + m^2d\right)\right)$.

Appendix B

Comparing Stratification Methods

BN model The BN model simulates scenarios with unrelated isolated populations (Figure 2 (i)) and serves as the basic case for arbitrarily structured population with no admixture.



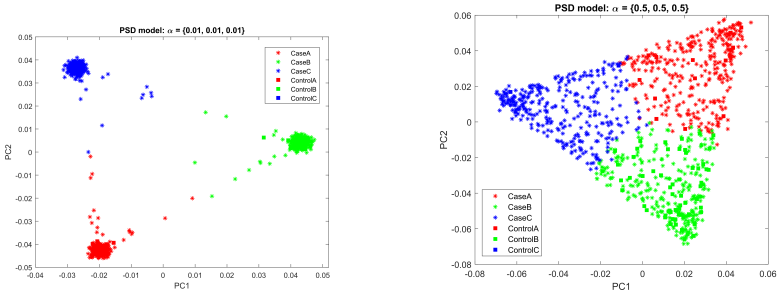
(a) Spurious associations (b) Causal associations

Fig. 4: Box plots for spurious and causal associations on the BN model shows that Armitage trend χ^2 has the maximum number of spurious associations containing about 4-5 causal SNPs whereas EIGENSTRAT has minimum number of spurious associations while detecting almost zero causal SNPs. CluStrat has more spurious associations than EIGENSTRAT and considerably less than Armitage trend χ^2 recovering slightly more number of causal SNPs than the latter.

The samples when projected on the top two PCs clearly resembles three isolated clusters with no connections between them. This is an ideal case when the populations are not mixing due to environmental factors acting as barriers of gene flow between populations. GWAS has shown to be robust in these settings [45]; however, the cryptic relatedness for each cluster remains a plaguing issue [5]. We ran CluStrat on this scenario with p-value threshold set to $p = \frac{25}{m_i} = 0.0025$ (m_i is the number of SNPs in each iteration, set to 10,000 for 100 iterations). The expected number of spurious association as mentioned in [41] is $m_0 \times p$ where $m_0 = m-$ number of causal SNPs. In our case, as we set the number of causal SNPs to 10 as per [41], $m_0 = 9990$ and therefore, the number of spurious associations to be approximately 25 with degree of freedom set to 1 for genotypes.

Armitage trend χ^2 with no population structure correction renders almost half of the SNPs in the simulation study as true associations resulting in a considerable amount of spurious associations highlighting the need for population structure correction. EIGENSTRAT on the other hand results in the expected number of spurious associations as also shown in previous work [38]. However,

it behaves stringently and detects zero causal SNPs almost all of the time (Figure 4). CluStrat, however, strikes a balance between the two and generates far more spurious associations than the expected value but about 5 folds less than Armitage trend χ^2 recovering a slightly higher number of causal SNPs. This shows that in the ideal case of population structure correction, CluStrat can identify more causal SNPs due to the structure informed clustering setup which widely used stratification correction methods lack.

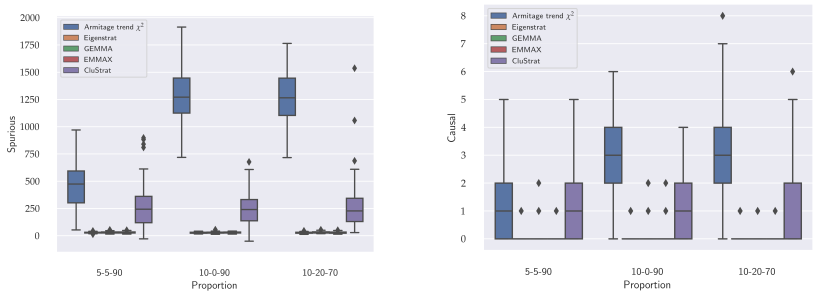


(a) $\alpha = \{0.01, 0.01, 0.01\}$ (b) $\alpha = \{0.5, 0.5, 0.5\}$

Fig. 5: Projection of the samples from PSD model with varying sets of values of α. We observe that increasing α increases the density between individuals leading to admixture and creates a uniform gradient as all values of α_i are equal.

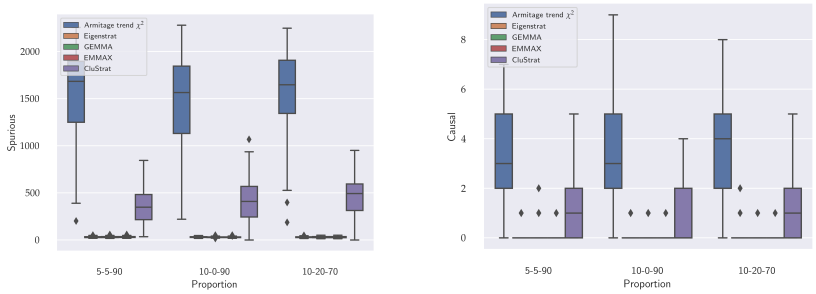
PSD model The PSD model emulates real world datasets more closely than BN model. It allows for admixing individuals and gradients across the populations. It is sampled from the Dirichlet distribution parameterized by a concentration parameter $\alpha \in \mathbb{R}^d$ where $d = 3$ (the number of populations for all simulations conducted). A higher value of α_i corresponds to greater weight of i^{th} population.

We ran CluStrat on the PSD model with varying number of α from 0.01 to 0.5 and kept equal α_i for a symmetric distribution. We report the boxplots of spurious and causal associations (Figure 6 and 7) for $\alpha = 0.1, 0.5$ and and observe that for the first case of variance, $(5\%, 5\%, 90\%)$, Armitage trend χ^2 and CluStrat performs almost similarly in terms of spurious associations. This is due to the fact that only 5% of the trait is explained by true genetic associations in presence of LD and the rest is noise and environmental factors. However, CluStrat outnumbers EIGENSTRAT, GEMMA and EMMAX in terms of causal associations and detects four to six fold more true causal SNPs. For the other two variance proportions, CluStrat performed better than the other methods in detecting the causal associations and strikes a balance in terms of spurious associations.



(a) Spurious associations (b) Causal associations

Fig. 6: Box plots for spurious and causal associations on the PSD model ($\alpha = \{0.1, 0.1, 0.1\}$) shows Armitage trend χ^2 has maximum number of spurious associations containing less causal SNPs than the BN model (Figure 4) owing to the admixed nature of the individuals in PSD. EIGENSTRAT, GEMMA and EMMAX has least number of spurious associations while detecting almost zero causal SNPs. CluStrat has more spurious associations than the standard approaches and less than Armitage trend χ^2 while recovering two to three fold more causal SNPs.



(a) Spurious associations (b) Causal associations

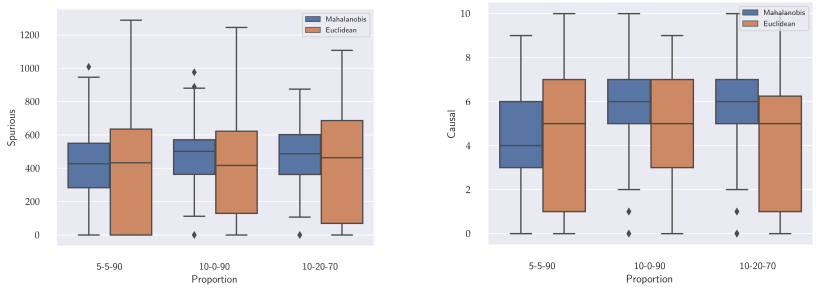
Fig. 7: Box plots for spurious and causal associations on the PSD model ($\alpha = \{0.5, 0.5, 0.5\}$) shows Armitage trend χ^2 has maximum number of spurious associations containing less causal SNPs than the BN model (Figure 4) owing to the overtly admixed nature of the individuals in PSD. EIGENSTRAT, GEMMA and EMMAX has least number of spurious associations while detecting almost zero causal SNPs. CluStrat has more spurious associations than the standard approaches and slightly more than $\alpha = 0.1$ owing to more admixed nature of the data. It has considerably less spurious associations than Armitage trend χ^2 while recovering two to three fold more causal SNPs.

Comparing Distance metrics

CluStrat with Euclidean distance metric based GRM (sample covariance matrix) also contains structure information as part of the relationships between the individuals within and between population groups. The GRM with Euclidean distance is straightforward to compute as shown below

$$\mathbf{D} = \mathbf{X}^\top \mathbf{X}$$

where $\mathbf{X} \in \mathbb{R}^{n \times m}$, with number of markers, n and number of samples, m ($n >> m$). We show that although Euclidean distances between individuals is straightforward to compute, it fails to distinguish fine-grained distances between individuals in the same cluster owing to cryptic relatedness. This is highlighted after performing AHC using Ward's linkage method which minimizes the increase in sum of squares between two cluster centroids in order to decide when to merge them. (Figure 1).



(a) Spurious associations (b) Causal associations

Fig. 8: Box plots for spurious and causal associations obtained by running AHC with Mahalanobis and Euclidean distances on the PSD model ($\alpha = \{0.1, 0.1, 0.1\}$). We observe similar performance on both the distance metrics in terms of identifying true causal variants. Mahalanobis distance discovers less spurious associations than Euclidean distance.

When Mahalanobis distance based GRM is used instead of Euclidean distance in AHC on PSD model with 1,000 individuals and 10,000 SNPs across 3 admixed arbitrarily structured ethnic groups, it reveals four broad clusters with various fine-grained sub-clusters revealing how Mahalanobis distance help recover cryptic relatedness and substructure within a population.

Due to admixture in the PSD model ($\alpha = \{0.1, 0.1, 0.1\}$) as shown in Figure 5 the dendrogram finds three broad clusters owing to the three populations in the simulation. It subsequently finds different sub-clusters at different depth on the horizontal axis. Thus, identifying interaction between individuals inside a cluster. This is a significant advantage of using Mahalanobis distance over it's Euclidean

counterpart as the latter only reveals three broad clusters with indistinguishable interactions in each cluster (Figure 1).

When we ran AHC with both the distances, we observe similar performance on the PSD model with Mahalanobis distance based GRM performing slightly better with respect to it's Euclidean counterpart (Figure 8). We note that as we increase the scale of admixed genotype data with more complex structure, Mahalanobis distance is better suited as it is known to project correlated high dimensional data to an uncorrelated lower dimensional space where it recovers the hidden Euclidean distances [32].

Table 1: Table showing strongest associations after running CluStrat on WTCCC2 PD data

Chrom#	SNPID	GeneID	p-value
2	rs10177996	WNT10A	2.22×10^{-16}
2	rs1059823	SLC11A1	2.4×10^{-16}
4	rs11936554	UNC5C	4×10^{-16}
2	rs13013415	WDR33	4.5×10^{-16}
2	rs1509467	MAP2	5×10^{-15}
3	rs1516570	GRM7	5.7×10^{-14}
6	rs176713	BACH2	6×10^{-12}
6	rs176713	BACH2	6×10^{-12}
4	rs2322559	SLIT2	6.1×10^{-12}
4	rs2328457	AIG1	6.13×10^{-12}
3	rs3816969	NMNAT3	6.15×10^{-12}
3	rs4677964	PARP15	7.15×10^{-10}