Multitask group Lasso for Genome Wide association Studies in admixed populations

Asma Nouira* and Chloé-Agathe Azencott
MINES ParisTech, PSL Research University, CBIO-Centre for Computational Biology, F-75006 Paris, France
Institut Curie, PSL Research University, F-75005 Paris, France
INSERM, U900, F-75005 Paris, France
*E-mail: asma.nouira@mines-paristech.fr**

Genome-Wide Association Studies, or GWAS, aim at finding Single Nucleotide Polymorphisms (SNPs) that are associated with a phenotype of interest. GWAS are known to suffer from the large dimensionality of the data with respect to the number of available samples. Other limiting factors include the dependency between SNPs, due to linkage disequilibrium (LD), and the need to account for population structure, that is to say, confounding due to genetic ancestry.

We propose an efficient approach for the multivariate analysis of admixed GWAS data based on a multitask group Lasso formulation. Each task corresponds to a subpopulation of the data, and each group to an LD-block. This formulation alleviates the curse of dimensionality, and makes it possible to identify disease LD-blocks shared across populations/tasks, as well as some that are specific to one population/task. In addition, we use stability selection to increase the robustness of our approach. Finally, gap safe screening rules speed up computations enough that our method can run at a genome-wide scale.

To our knowledge, this is the first framework for GWAS on admixed populations combining feature selection at the LD-groups level, a multitask approach to address population structure, stability selection, and safe screening rules. We show that our approach outperforms state-of-the-art methods on both a simulated and a real-world cancer datasets.

Keywords: Genome Wide Association Studies, Feature selection, Multitask group Lasso, Stability selection, Safe screening rules

1. Introduction

Over the last 15 years, Genome-Wide Association Studies (GWAS) have become one of the most prevalent methods to identify regions of the genome associated with complex phenotypic traits, and in particular complex diseases in humans.[1] One of the major concerns in GWAS is population stratification, which arises when allele frequency differences between cases and controls are due to differences in genetic ancestry rather than to association with the phenotype. Many correction methods have been proposed to adjust the inflation of associations in admixed populations, including methods based on principal components analysis or on linear mixed models.[2] However, it is possible that these techniques lead to overcorrection, in

particular by masking population-specific disease loci.

An additional issue in GWAS is Linkage Disequilibrium (LD), which manifests as correlation between adjacent Single Nucleotide Polymorphisms (SNPs), creating statistical dependence between those markers and reducing statistical power.[3] Combining strongly correlated SNPs into blocks and modeling the association signal over an entire region is one way to address this limitation.

Classical approaches for GWAS are based on single-marker analyses, testing for association between each SNP and the phenotype independently. This may prevent the detection of effects that are due to SNPs acting additively, leading many authors to favor fitting a linear model to all SNPs jointly.[4] Penalized regression approaches, such as the Lasso, which uses an ℓ_1-norm regularization to shrink some coefficients of the model to zero, effectively removing them from the model, are particularly suited to this task.

Additional regularizers can be used to enforce additional prior hypotheses on the coefficients of such a linear model. Among them, the group Lasso[3,5] ensures sparsity at the level of pre-defined groups of features, and the multitask Lasso[6,7] fits models on related tasks jointly, encouraging similar sparsity patterns across all tasks.

In this work, we propose to combine both approaches into a multitask group Lasso framework, in which groups correspond to pre-defined LD patterns, and each task corresponds to a subpopulation, therefore simultaneously addressing the limitations of single-marker analyses and the issues of both LD and population structure.

In addition, we draw on the stability selection framework[8] to improve the stability of the results, that is to say, their robustness to small perturbations in the input data. Indeed, because the number of SNPs is typically much larger than that of samples, penalized regression approaches tend to select different sets of SNPs when presented with different subsets of the data, which severely limits their interpretability.

Finally, we use the recently proposed gap safe screening rules[9] to improve computational complexity, and scale our approach to on about one million SNPs.

In what follow, we present our proposed approach in details, place it in the context of existing work, and evaluate it on both a simulated data set and a real-world cancer GWAS data set.

2. Methods

In this section, we introduce our proposed approach, MuGLasso, which follows four steps: (1) assigning each sample to a genetic population, hence forming subsets of the data that will be considered as different but related tasks (Section 2.1); (2) leveraging the correlations between SNPs to create LD-groups, hence performing feature selection at the level of groups rather than individual SNPs, increasing interpretability and alleviating the curse of dimensionality (Section 2.2); (3) fitting a regularized model per task, using an $\ell_{2,1}$ penalty that enforces sparsity at the level of LD-groups and ties the solutions across tasks, and gap safe screening rules to speed up the optimization procedure (Section 2.3); and (4) implementing stability selection to improve the robustness of the solution (Section 2.4).

2.1. *Population stratification*

Population structure, whereby the data is made of subsets of individuals that differ systematically both in genetic ancestry and in the phenotype under investigation, is a major confounding factor in GWAS. Indeed, it leads to detecting allele frequency differences in cases and controls that correspond to differences in ancestry, instead of a more direct association between genotype and phenotype. Several approaches have been developed to adjust for population structure.

Among them, a large number of of methods rely on Principal Component Analysis (PCA),[10–12] and consist in including top Principal Components (PCs) of the genotypes as covariates in regression models. In addition, linear mixed models[13] can be used to model the phenotype as a combination of fixed and random effects, with the covariance of the latter being computed from a genetic similarity matrix. Although they often outperform PCA-based methods, the mixed model approaches tend to be more computationally demanding. Both approaches are similar in that regressing out principal components can be seen as approximation of a linear mixed model.[2]

However, these techniques may lead to ignoring population-specific SNPs, which is why we propose a multitask approach that can identify disease loci that are either population-specific or shared between populations. We therefore form tasks by separating the data into subpopulations, identified as clusters on the projection of the genotypes on their top PCs.

2.2. *Linkage disequilibrium groups*

Linkage disequilibrium (LD) is the non-random association of alleles of at least two loci.[14] LD can be leveraged to form groups of correlated SNPs. Grouping SNPs helps to alleviate the curse of dimensionality in GWAS by reducing the number of testing possibilities. This can be achieved by combining p-values within a group of correlated SNPs, or through the use of penalized regression approaches that perform feature selection at the level of groups, rather than at the level of individual SNPs.[3] The latter has the advantage over individual statistical testing of modeling the additive effects of multiple genetic markers simultaneously.

2.2.1. *Adjacency-constrained hierarchical clustering*

In many species, including humans,[15] LD is known to be correlated to the physical distance between SNPs. Hence, genomes can be clustered in LD blocks of strongly correlated adjacent SNPs, called in this paper LD-groups. Such LD-groups can be obtained using adjacency-constrained hierarchical agglomerative clustering,[3] in which only physically adjacent clusters can be merged.

2.2.2. *LD-groups across populations*

Because LD patterns may be influenced by genetic ancestry,[16] we perform LD-groups partitioning for each population separately. We then combine those LD-groups into common shared LD-groups. More specifically, the set of coordinates of the boundaries of the shared LD-groups is obtained as the union of the sets of coordinates of the boundaries of the LD-groups for each

population. This procedure is described on Supplementary Figure B1.

2.3. *Multitask group Lasso*

2.3.1. *General framework and problem formulation*

We use a penalized regression approach to fit a multivariate linear model between the phenotype and the SNPs, with a regularization term that ensures that (1) the solution is sparse at the level of LD-groups and (2) the regression coefficients are smoothed within groups and across tasks. Such an approach provides shared LD-groups associated with the phenotype across all tasks, and allows for some LD-groups to be specific to each task.

Problem formulation Given a set of p SNPs measured for n samples, we split the n samples in T subpopulations/tasks, each of size n_t for $t = 1, \dots, T$, and the p SNPs in G LD-groups, each of size p_g for $g = 1, \dots, G$. For each population t, we denote by $\boldsymbol{x}_m^{(t)}$ the p-dimensional vectors of SNPs of the m-th sample in the population ($m = 1, \dots, n_t$), and by $y_m^{(t)}$ its phenotype. We then formulate the following optimization problem:

$$\min_{B \in \mathbb{R}^{p \times T}} \frac{1}{n} \sum_{t=1}^{T} \sum_{m=1}^{n_t} \mathcal{L}\left(y_m^{(t)}, \sum_{j=1}^{p} \beta_j^{(t)} x_{mj}^{(t)}\right) + \lambda \sum_{g=1}^{G} \sqrt{p_g} \left\| B^{(g)} \right\|_F, \tag{1}$$

where $\boldsymbol{\beta}^{(t)} \in \mathbb{R}^p$ is the vector of regression coefficients specific to task t : $\boldsymbol{\beta}^{(t)} = (B_{1t}, \dots, B_{pt})$, $\mathcal{L}$ is the quadratic loss if the phenotype is quantitative ($y \in \mathbb{R}$) and the logistic loss if it is qualitative ($y \in \{0, 1\}$), $\|.\|_F$ denotes the Frobenius norm, and $B^{(g)}$ is a $p_g \times T$ matrix containing the regression coefficients, across all tasks, for the SNPs of group g.

2.3.2. *Related work*

$\boldsymbol{\ell_{2,1}}$-norm regularization Our approach is closely related to the group Lasso[5] and multitask Lasso,[6] which both make use of an $\ell_{2,1}$-norm regularization. More precisely, the group Lasso corresponds to a special case of Equation (1), with a single task ($T = 1$), resulting in sparsity at the group levels. Using a group Lasso where the groups are defined based on LD blocks has been successfully applied to GWAS on up to 20 000 SNPs.[3] The multitask Lasso corresponds to a special case of Equation (1), with each group containing exactly one SNP. This formulation ties sparsity patterns across tasks and has been applied before to multi-population GWAS, although only a few thousand SNPs.[7]

The multitask group Lasso we propose can also be reformulated as an $\ell_{2,1}$-norm regularization problem, through the creation of a new dataset $(\widetilde{X}, \widetilde{\boldsymbol{y}})$ where $\widetilde{X} \in \mathbb{R}^{n \times pT}$ is a block-diagonal matrix such that each of the T diagonal blocks corresponds to the SNP matrix $X^{(t)} \in \mathbb{R}^{n_t \times p}$ for task t, and $\widetilde{\boldsymbol{y}}$ is a n-dimensional vector obtained by stacking the phenotype vectors for each task. Equation (1) can then be rewritten as:

$$\min_{\boldsymbol{b} \in \mathbb{R}^{pT}} \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}\left(\widetilde{y}_i, \sum_{k=1}^{pT} b_k \widetilde{x}_{ik}\right) + \lambda \sum_{g=1}^{G} \sqrt{p_g} \left\| \boldsymbol{b}^{(g)} \right\|_2, \tag{2}$$

with $\boldsymbol{b}^{(g)} \in \mathbb{R}^{p_g T}$ the regression coefficients corresponding to all SNPs of group (g) for all tasks.

In essence, this is a group Lasso with G groups each containing T copies (one per task) of the p_g features of SNP group g. Thus $B_{jt} = \boldsymbol{b}_{p(t-1)+j}$.

Other multitask group Lassos Other authors have proposed variations on the idea of a multitask group Lasso before. Several publications[17,18] add a second regularization term to our formulation, increasing within-group or across-task sparsity. Unfortunately, this dramatically increases computational time, and indeed none of these publications analyze genome-wide data sets. In addition, because interpretation will be done at the group level rather than at the SNP level, within-group sparsity is not necessarily desirable in this context.

Several authors have built on these propositions and add a third regularization term, either enforcing group-independent task sparsity[19] or overall sparsity (with an ℓ_1-norm over all coefficients).[20] Again, the addition of these regularizers severely hinders the applicability of these methods at a genome-wide scale.

Hence none of these methods is readily applicable to our setting. In addition, their stability has never been evaluated, even though it is an important criterion for the reliability and interpretability of the results (see Section 2.4).

2.3.3. *Gap safe screening rules*

To speed up the computation of the solution of Equation (2), we call upon gap safe screening rules,[9] which are used to efficiently identify features for which the regression coefficients will be zero and hence ignore them when solving the problem. Such screening rules have been proposed for a large number of popular regularized regressions,[9] including $\ell_{2,1}$-norm regularizations. In particular, Equation (2) can be solved using the `Gap_Safe_Rule` package[a]. We briefly summarize the idea underlying gap safe screening rules in Appendix B.3.

2.4. *Stability selection*

Unfortunately, in GWAS, penalized regression approaches often lack stability, that is to say, robustness to slight variations in the input dataset.[21] However, stability increases both the reliability of the results and the interpretability. To address this limitation, *stability selection*[8,21] consistst in performing feature selection repeatedly on subsamples of the data and only retains the features most often selected. More specifically, given a subsample $I \subset \{1, \ldots, n\}$ of size $\lfloor n/2 \rfloor$ of the data, we call $\widehat{S}^\lambda(I)$ the set of features selected by the selection procedure of interest (for example, a Lasso), with hyperparameter λ, on this subsample of the data. For any feature $j \in \{1, \ldots, p\}$, we call $\widehat{\Pi}_j^\lambda$ the probability that feature j is selected on a random subsample of size $\lfloor n/2 \rfloor$ of the data. This probability is determined, given m such random subsamples $I_1, I_2, \ldots, I_m$, as the proportion of those subsamples for which the feature selection procedure selects feature j: $\widehat{\Pi}_j^\lambda = \frac{1}{m} \sum_{k=1}^m \mathbf{1}_{j \in S^\lambda(I_k)}$. Finally, given a a threshold $\frac{1}{2} < \pi_{\text{cutoff}} \leq 1$ (in this work, we used $\pi_{\text{cutoff}} = 0.75$), the stable set of selected features is determined as $\widehat{S}^{\text{stable}} = \{j : \max_{\lambda \in \Lambda} \widehat{\Pi}_j^\lambda \geq \pi_{\text{cutoff}}\}$.

[a] `https://github.com/EugeneNdiaye/Gap_Safe_Rules`

3. Experiments

3.1. *Data*

Simulated data Using GWAsimulator,[22] we simulated GWAS data with realistic LD patterns from two populations (CEU and YRI) of the HapMap 3 data. We induced population structure by varying the case:control ratio within each subpopulation (CEU 1100:900 and YRI 900:1100), as well as by simulating population-specific disease loci. We simulated a total of 149 970 disease SNPs, 2 999 (resp 4 999) of which are specific to the CEU (resp. YRI) population. More details are available in Appendix A.1. We obtained a final dataset of 4,000 samples of admixed populations and 1,400,000 SNPs.

DRIVE Breast Cancer OncoArray The DRIVE OncoArray dataset (dbGap study accession phs001265/GRU) contains 28 281 individuals that were genotyped for 582 620 SNPs. 13 846 samples are cases and 14 435 are controls. More details are available in Appendix A.2.

3.2. *Preprocessing*

Quality control and imputation We removed SNPs with a minor allele frequency lower than 5%, a p-value for Hardy-Weinberg Equilibrium in controls lower than 10%, or a missing genotyping rate larger than 10%. We removed duplicate SNPs and excluded samples with more than 10% of SNPs missing. We imputed missing genotypes in DRIVE using IMPUTE2.[23]

LD pruning We performed LD pruning prior to our analysis, both to reduce the curse of dimensionality and to better capture population structure using PCA.[24] More specifically, we used PLINK[25] with a LD cutoff of $r^2 > 0.85$ and a sliding window size of 50Mb. 1 000 000 SNPs remain in the simulated data and 313 237 in DRIVE.

PCA and population structure We used PLINK[25] to compute principal components of the genotypes. We thus identify two populations in the simulated data, matching the CEU and YRI populations (see Supplementary Figure C1a). In DRIVE, we identify two populations (see Supplementary Figure C1b), which we refer to as POP1 (corresponding to samples from the USA, Australia and Denmark) and POP2 (corresponding to samples from Cameroon, Nigeria and Uganda, as well as a few samples from the USA).

LD-groups choice We obtain LD-groups for each of the PCA-based populations using adjclust[26] and obtain shared LD-groups as described in Section 2.2.2. Table 1 shows the number of LD-groups obtained for each subpopulation and the final number of shared groups.

3.3. *Comparison partners*

As a baseline, we use PLINK[25] to perform tests of association between each SNP and the phenotype, using the top PCs as covariates. We refer to this method as **Adjusted GWAS**, by contrast with **Stratified GWAS**, in which we perform such tests of association separately within each population. In addition, we computed a PCA-adjusted phenotype as the residuals of a regression between the top PCs and the phenotype.

Table 1. For each subpopulation of both datasets (simulated and real), LD-groups number is given and the shared LD-groups number after combination

Data	Subpopulations	LD-groups number	Shared LD-groups number
Simulated data	CEU YRI	25,281 15,636	35,792
DRIVE real data	POP1 POP2	8,152 5,032	17,782

We also compare MuGLasso to other Lasso approaches, so as to better understand the respective effects of grouping correlated SNPs, on the one hand, and using a multitask model to address population structure, on the other hand. More specifically, we used a Lasso (single task, no groups) on each population separately (**Stratified Lasso**) or on the adjusted phenotype (**Adjusted Lasso**), as well as a group Lasso (single task) on each population separately (**Stratified group Lasso**) or on the adjusted phenotype (**Adjusted group Lasso**). For computational efficiency, we use the bigLasso package[27] for the Lasso, and the `Gap_Safe_Rule` package[9] for the group Lasso. For all variants of the Lasso, including MuGLasso, we set the regularization hyperparameter λ by cross-validation.

To compare these methods, we report runtime, ability to recover true causal SNPs (in the case of simulated data), and stability of the selection. To measure selection stability, we repeat the feature selection procedure on 10 subsamples of the data, and report the average Pearson's correlation between all pairs of indicator vectors representing the selected features for each subsample (see Appendix B.4 for details).

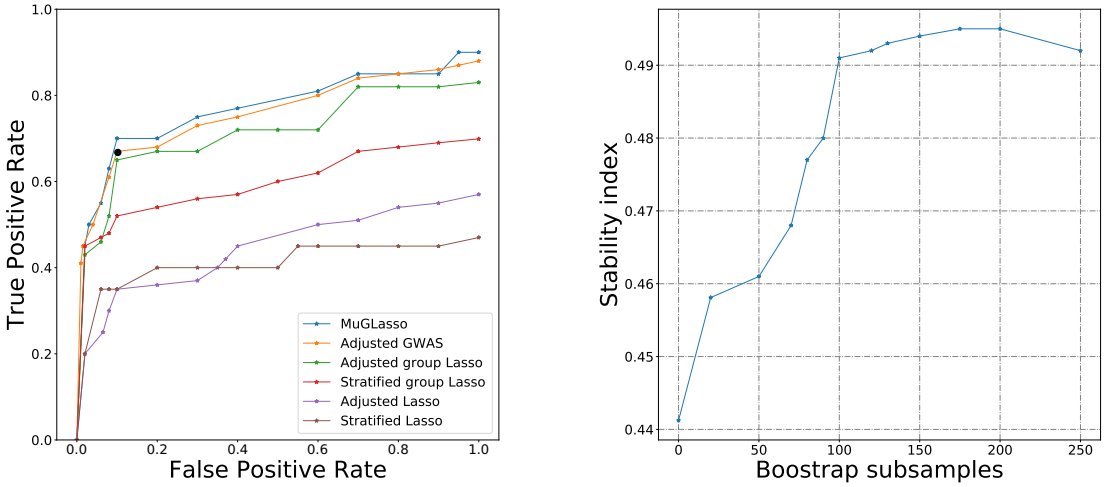
4. Results

4.1. *MuGLasso draws on both LD-groups and the multitask approach to recover disease SNPs*

On the simulated data, we observe (Figure 1a) that MuGLasso is better than any other method at recovering the true disease SNPs. Performing feature selection at the level of LD-groups, rather than individual SNPs, improves performance. Indeed, the group Lassos and MuGLasso outperform the SNP-level Lassos. In addition, treating all samples simultaneously (as in MuGLasso or the adjusted approaches) also improves performance. This confirms our hypothesis that grouping features and using all samples simultaneously both alleviate the curse of dimensionality. However, this comes at an added computational time (see Supplementary Figure C2 on simulated data and Figure 2a on DRIVE), as even with gap safe screening rules, using an $\ell_{2,1}$-regularization over several tens of thousands of groups is much more computationally intensive than a Lasso. However, the implementation is efficient enough to allow computations on 10^6 SNPs, even with the added cost of repeated subsampling to increase stability.

4.2. *MuGLasso provides the most stable selection*

Figures 1b (simulated data) and 2b (DRIVE) show the stability index of MuGLasso as a function of the number of subsamples. Increasing the number of subsamples increases the stability of the selection; 100 bootstrap samples appears to be an acceptable trade-off between



(a) Recovery of disease loci (b) Stability of MuGLasso

Fig. 1. On simulated data, ability of different methods to retrieve causal disease SNPs as a ROC plot (1a), and stability index of MuGLasso as a function of the number of bootstrap samples (1b). On the ROC plot, the black dot indicates the performance of the stratified GWAS at the Bonferonni-corrected significance threshold.

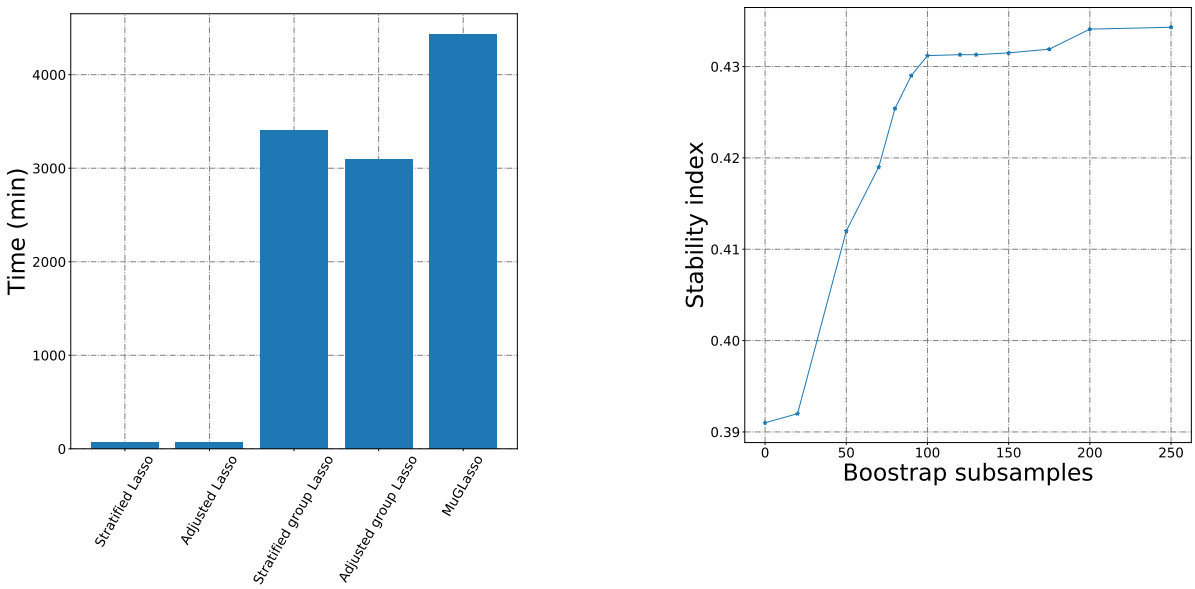
runtime and improved stability, and this is therefore the number of subsamples we use for all other experiments.

Tables 2 and 3 show the stability index of the different methods, on simulated data and DRIVE, respectively. We ran the adjusted GWAS once on the entire data set, as would usually be done, and therefore cannot report its stability. Our results again illustrate that stability selection does increase the stability of Lasso methods. We confirm this by running MuGLasso without stability selection as well as Adjusted group Lasso with stability selection on top. In both cases, the stability index increases when stability selection is used. In addition, we report the total number of selected SNPs and LD-groups. For methods that select individual SNPs, we obtain the number of selected LD-groups by considering that each selected SNP selects its entire LD-group. Our results illustrate that the improved stability of MuGLasso does not come at the expense of selecting more features. On the contrary, stability selection provides fewer SNPs/LD-groups with better stability.

4.3. *MuGLasso selects both task-specific and global LD-groups*

For both datasets, the LD-groups selected by MuGLasso are a mixture between population-specific LD-groups (identified as those with near-zero regression coefficients for one task) and LD-groups that are shared between both populations. Table 4 shows the number of LD-groups/SNPs in each of these categories for MuGLasso. By contrast, the adjusted approaches do not provide population-specific LD-groups or SNPs.

Finally, we report on Figure 3 the precision and recall of MuGLasso and the stratified



(a) Runtimes of the Lasso approaches

(b) Stability of MuGLasso

Fig. 2. On DRIVE, runtimes of the different Lasso approaches (2a) and stability index of MuGLasso as a function of the number of bootstrap samples (2b).

Table 2. Stability index and number of selected features for different methods, on simulated data

Methods	Number of selected LD-groups	Number of selected SNPs	Stability index	Selection level
MuGLasso	5,623	155,312	0.4912	LD-groups
MuGLasso without stab sel	6,124	161,221	0.4412	LD-groups
Adjusted group Lasso + stab sel	6,054	162,104	0.4134	LD-groups
Adjusted group Lasso	6,347	167,204	0.3714	LD-groups
Stratified group Lasso	4,836	154,732	0.3398	LD-groups
Adjusted Lasso	5,379	158,856	0.2368	Single-SNP
Stratified Lasso	5,704	168,158	0.1742	Single-SNP
Adjusted GWAS	5,063	141,340	-	Single-SNP

approaches on the population-specific SNPs. MuGLasso outperforms all other approaches in both precision and recall.

5. Discussion and Conclusions

We presented MuGLasso, an efficient approach for detecting disease loci in GWAS data from admixed populations. Our approach is based on a multitask framework, where input tasks correspond to subpopulations, and feature selection is performed at the level of LD-groups. Assigning samples from PCA-identified populations to different tasks addresses the issue of population stratification, and retains the flexibility of identifying population-specific disease loci. Treating all samples together, by contrast with stratified approaches, alleviates the curse

Table 3. Stability index and number of selected features for different methods, on DRIVE

Methods	Number of selected LD-groups	Number of selected SNPs	Stability index	Selection level
MuGLasso	62	1,357	0.4312	LD-groups
MuGLasso without stab sel	72	1,524	0.3911	LD-groups
Adjusted group Lasso + stab sel	59	1,293	0.3234	LD-groups
Adjusted group Lasso	68	1,466	0.2613	LD-groups
Stratified group Lasso	58	1,119	0.2498	LD-groups
Adjusted Lasso	41	874	0.2068	Single-SNP
Stratified Lasso	38	789	0.1581	Single-SNP
Adjusted GWAS	16	306	-	Single-SNP

Table 4. For MuGLasso, number of selected LD-groups/SNPs, across and per population

Data	Population	Number of selected LD-groups (and SNPs)
Simulated data	CEU	95 (2,418 SNPs)
	YRI	103 (3,081 SNPs)
	shared (CEU and YRI)	5,227 (149,813 SNPs)
DRIVE	POP1	6 (148 SNPs)
	POP2	2 (43 SNPs)
	shared (POP1 and POP2)	54 (1166 SNPs)

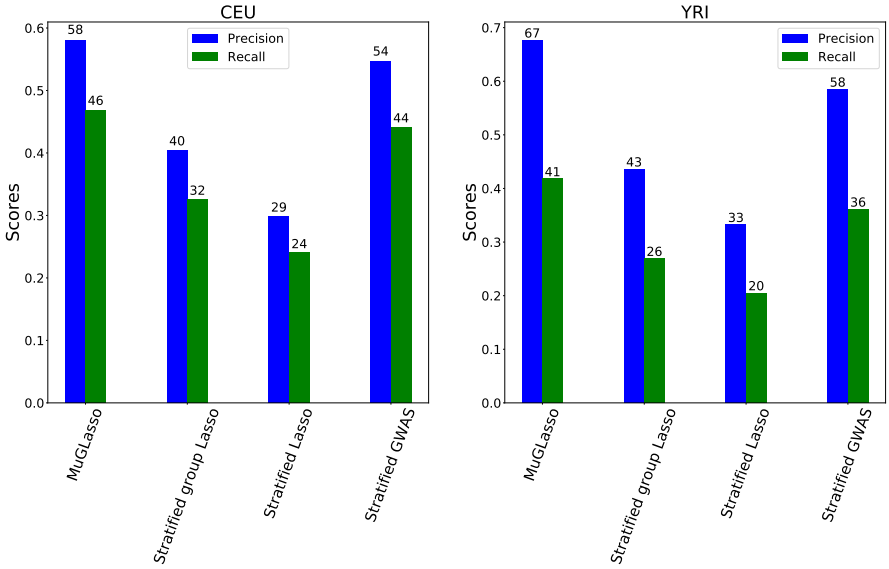


Fig. 3. For simulated data, precision and recall of MuGLasso and the stratified approaches on the populations-specific SNPs

of dimensionality. Ensuring sparsity at the level of LD-groups addresses the high correlation between nearby SNPs and also alleviates the curse of dimensionality. Although more time-consuming than a classical GWAS, our implementation is computationally efficient enough to scale to the analysis of entire GWAS data sets of about one million SNPs.

On simulated data, MuGLasso outperforms state-of-the-art approaches in its ability to re-

cover disease loci. Importantly, this also holds for population-specific SNPs; hence performance is not driven solely by the ability of recover disease loci that are common to all populations. In addition, MuGLasso is the most stable of all evaluated method, which increases the interpretability of the results.

Finally, although we presented MuGLasso in the context of admixed populations, our tool could be used in other multitask settings. In particular, tasks can stem from related phenotypes[17] or from different studies pertaining to the same trait, in a meta-analysis approach.[18] Groups could also be defined according to different prior biological knowledge, for example based on functional units such as genes, in the spirit of gene-set analyses of GWAS data. In addition, although we only presented results on case-control studies with two populations, the method directly applies to quantitative phenotypes and any number of tasks.

An important outcome of our study is that, although we have not included in MuGLasso a regularization term that would enforce sparsity at the level of tasks as in,[20] we still obtain task-specific groups. Including such an additional term in Equation (1) would perhaps improve the already state-of-the-art task-specific precision and recall of MuGLasso, but this would unfortunately come at the expense of a notable increase in computational time, if only because of the cross-validation needed to set the value of a second hyperparameter.

An in-depth biological analysis of the loci identified by MuGLasso on DRIVE would illustrate the biological relevance of our method, but is out of the scope of this methodological paper.

In the future, we are looking forward to making use of the post-inference selection framework for group-sparse linear models[28] to provide p-values for the selected loci. As of now, it is unclear how to apply these ideas to case-control studies in a computationally efficient manner.

Acknowledgments

This work was supported by the French Agence Nationale de la Recherche (references ANR-18-CE45-0021-01 and ANR19-P3IA-0001). OncoArray genotyping and phenotype data harmonization for the Discovery, Biology, and Risk of Inherited Variants in Breast Cancer (DRIVE) breast-cancer case control samples was supported by X01 HG007491 and U19 CA148065 and by Cancer Research UK (C1287/A16563).

Supplementary Materials and code

Code is available at `https://github.com/asmanouira/MuGLasso_GWAS`.

References

1. P. M. Visscher *et al.*, 10 years of gwas discovery: Biology, function, and translation, *Am J Human Genet* **101** (2017).
2. Y. Zhang and W. Pan, Principal component regression and linear mixed model in association analysis of structured samples: competitors or complements?, *Genet Epidemiol* **39**, 149 (2015).
3. A. Dehman, C. Ambroise and P. Neuvial, Performance of a blockwise approach in variable selection using linkage disequilibrium information, *BMC Bioinformatics* (2015).
4. S. Okser *et al.*, Regularized machine learning in the genetic prediction of complex traits, *PLoS Genetics* **10**, p. e1004754 (2014).

5. M. Yuan and Y. Lin, Model selection and estimation in regression with grouped variables, *J R Stat Soc B* (2006).
6. G. Obozinski, B. Taskar and M. Jordan, Multi-task feature selection, *Technical report, UC Berkeley* (2006).
7. K. Puniyani, S. Kim and E. P. Xing, Multi-population GWA mapping via multi-task regularized regression, *Bioinformatics* **26**, i208 (2010).
8. N. Meinshausen and P. Bühlmann, Stability selection, *J R Stat Soc B* (2009).
9. E. Ndiaye *et al.*, Gap safe screening rules for sparsity enforcing penalties, *Journal of Machine Learning Research 18* (2017).
10. E. Zeggini *et al.*, Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes, *Nat Genet* (2008).
11. A. C. Need *et al.*, A genome-wide investigation of SNPs and CNVs in schizophrenia, *PLOS Genetics* (2009).
12. A. L. Price *et al.*, Principal components analysis corrects for stratification in genome-wide association studies, *Nat Genet* (2006).
13. J. Yu *et al.*, A unified mixed-model method for association mapping that accounts for multiple levels of relatedness, *Nat Genet* (2006).
14. M. Slatkin, Linkage disequilibrium – understanding the evolutionary past and mapping the medical future, *Nat Rev Genet* (2008).
15. D. E. Reich *et al.*, Linkage disequilibrium in the human genome, *Nature* **411**, 199 (2001).
16. M. Boehnke, A look at linkage disequilibrium, *Nat Genet* **25**, 246 (2000).
17. H. Wang *et al.*, Identifying quantitative trait loci via group-sparse multitask regression and feature selection: an imaging genetics study of the adni cohort, *Bioinformatics* **28**, 229 (2012).
18. D. Lin *et al.*, Integrative analysis of multiple diverse omics datasets by sparse group multitask regression, *Front Cell Dev Biol* **2**, p. 62 (2014).
19. X. Liu *et al.*, Group guided sparse group lasso multi-task learning for cognitive performance prediction of Alzheimer's disease, in *Int Conf on Brain Inform*, (Springer, 2017).
20. L. Li *et al.*, Multi-task learning sparse group lasso: a method for quantifying antigenicity of influenza A (H1N1) virus using mutations and variations in glycosylation of hemagglutinin, *BMC Bioinformatics* **21**, 1 (2020).
21. D. H. Alexander and K. Lange, Stability selection for genome-wide association, *Genetic Epidemiology* **35** (2011).
22. C. Li and M. Li, GWAsimulator: a rapid whole-genome simulation program, *Bioinformatics* (2008).
23. B. N. Howie *et al.*, A flexible and accurate genotype imputation method for the next generation of genome-wide association studies, *PLoS Genetics* (2009).
24. A. Abdellaoui *et al.*, Population structure, migration and diversifying selection in the Netherlands, *Eur J Hum Genet 21* (2013).
25. S. Purcell *et al.*, PLINK: A tool set for whole-genome association and population-based linkage analyses, *Am J Human Genet* (2007).
26. C. Ambroise *et al.*, Adjacency-constrained hierarchical clustering of a band similarity matrix with application to genomics, *Algorithms Mol Biol* (2019).
27. Z. Yaohui and B. Patrick, The biglasso package: A memory- and computation-efficient solver for lasso model fitting with big data in R, *The R Journal* (2017).
28. F. Yang, R. Foygel Barber, P. Jain and J. Lafferty, Selective inference for group-sparse linear models, *Adv Neural Inf Process Syst* **29**, 2469 (2016).

Supplementary Materials: Multitask group Lasso for Genome Wide association Studies in admixed populations

Asma Nouira* and Chloé-Agathe Azencott

MINES ParisTech, PSL Research University, CBIO-Centre for Computational Biology, F-75006 Paris, France

Institut Curie, PSL Research University, F-75005 Paris, France

INSERM, U900, F-75005 Paris, France

*E-mail: asma.nouira@mines-paristech.fr**

Appendix A. Data availability

Appendix A.1. *Simulated data*

Code to reproduce our simulations is available on `https://github.com/asmanouira/MuGLasso_GWAS`

Table A1 shows the location of the predefined disease loci, for each population. Table A2 shows the number of predefined disease loci, both common to both population and specific to each population.

Table A1. For simulated data, location of predefined disease loci represented by start/end positions information in each subpopulation through chromosomes: 2, 12, 19, 21 and 22.

Chromosome	Subpopulations	
	CEU	YRI
2	1,000 - 50,000	1,000 - 50,000
12	10 - 37,000	10 - 40,000
19	1,000 - 50,000	1,000 - 50,000
21	10 - 10,000	10 - 7,000
22	-	10 - 2,000

Table A2. For simulated data, number of predefined causal SNPs

Populations	Number of SNPs
Specific-CEU	2,999
Specific-YRI	4,989
Shared (CEU+YRI)	141,982
Total	149,970

Appendix A.2. *DRIVE*

Data access The dataset "General Research Use" in DRIVE Breast Cancer OncoArray Genotypes is available from the dbGaP controlled-access portal, under Study Accession phs001265.v1.p1 (`https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study\_id=phs001265.v1.p1`). Researchers can gain access the data by applying to the data access committee, see `https://dbgap.ncbi.nlm.nih.gov`.

Ethics approval The dataset was obtained from NIH after ethical review of project #17707, titled "Network-guided multi-locus biomarker discovery", and used under approval of this request (#67806-4).

Acknowledgments OncoArray genotyping and phenotype data harmonization for the Discovery, Biology, and Risk of Inherited Variants in Breast Cancer (DRIVE) breast-cancer case control samples was supported by X01 HG007491 andU19 CA148065 and by Cancer Research UK (C1287/A16563). Genotyping was conducted by the Center for Inherited Disease Research (CIDR), Centre for Cancer Genetic Epidemiology, University of Cambridge, and theNational Cancer Institute. The following studies contributed germline DNA from breast cancer cases and controls:the Two Sister Study (2SISTER), Breast Oncology Galicia Network (BREOGAN), Copenhagen GeneralPopulation Study (CGPS), Cancer Prevention Study 2 (CPSII), The European Prospective Investigation intoCancer and Nutrition (EPIC), Melbourne Collaborative Cohort Study (MCCS), Multiethnic Cohort (MEC), NashvilleBreast Health Study (NBHS), Nurses Health Study (NHS), Nurses Health Study 2 (NHS2), Polish Breast CancerStudy (PBCS), Prostate Lung Colorectal and Ovarian Cancer Screening Trial (PLCO), Studies of Epidemiologyand Risk Factors in Cancer Heredity (SEARCH), The Sister Study (SISTER), Swedish Mammographic Cohort(SMC), Women of African Ancestry Breast Cancer Study (WAABCS), Women's Health Initiative (WHI).

Appendix B. Supplementary Methods

Appendix B.1. *LD groups across populations*

Figure B1 illustrates the process by which we obtain LD-groups across populations, from LD-groups obtained on each population separately using adjacency-constrained hierarchical clustering (see Section 2.2.1)

Appendix B.2. *Multitask group lasso*

Figure B2 illustrates the architecture of the multitask group Lasso described in Section 2.3.

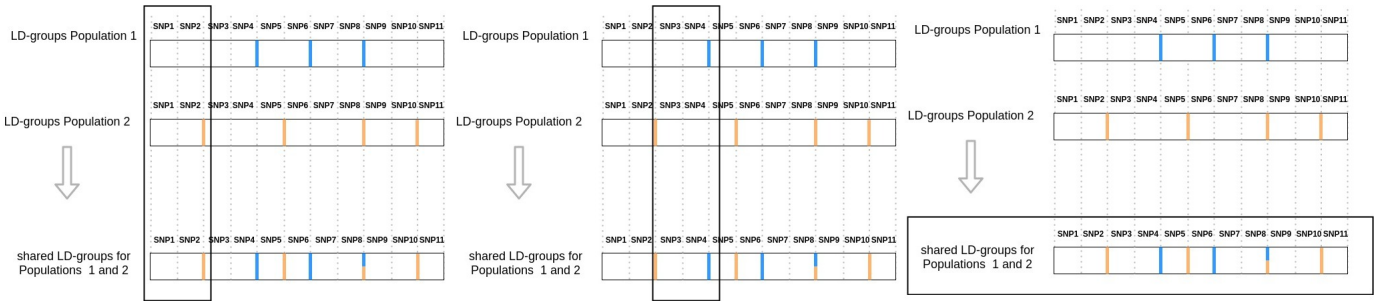


Fig. B1. Choice of shared LD-groups choice after adjacency-constrained hierarchical clustering for each population

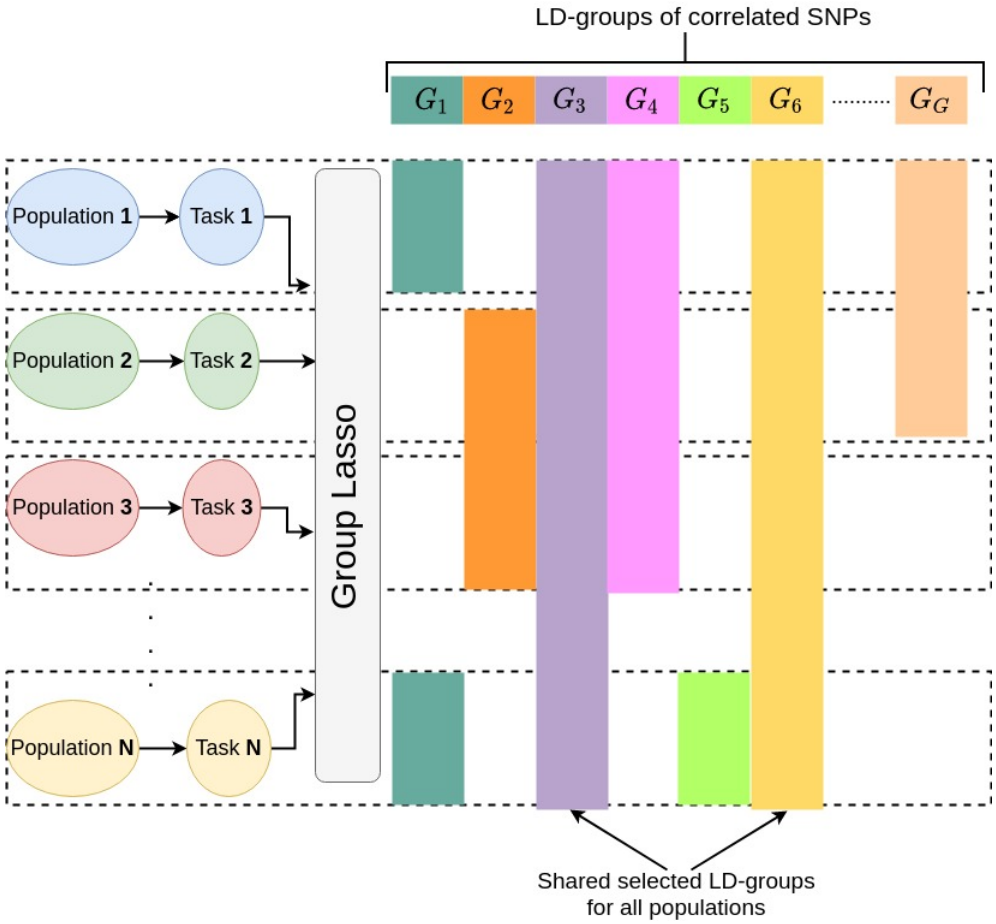


Fig. B2. Multitask group Lasso architecture

Appendix B.3. *Gap safe screening rules*

Let $X \in \mathbb{R}^{n \times d}$ be a design matrix and $\boldsymbol{y} \in \mathbb{R}^n$ the corresponding vector of outcomes, which can be binary or real-valued. We consider the following optimization problem:

$$\widehat{\boldsymbol{\beta}}^{(\lambda)} \in \underset{\boldsymbol{\beta} \in \mathbb{R}^d}{\arg\min}\, P_\lambda(\boldsymbol{\beta}) := \sum_{i=1}^{n} f_i\left(X_{i.}^\top \boldsymbol{\beta}\right) + \lambda\, \Omega(\boldsymbol{\beta}), \tag{B.1}$$

where all $f_i : \mathbb{R} \to \mathbb{R}$ are convex and differentiable functions with $1/\gamma-$Lipschitz gradient, and $\Omega : \mathbb{R}^d \to \mathbb{R}_+$ is a norm that is group-decomposable, i.e., the set of d features is partitioned in G groups of sizes $d_1, d_2, \ldots, d_G$, and

$$\Omega(\boldsymbol{\beta}) = \sum_{g=1}^{G} \Omega_g \left(\boldsymbol{\beta}^{(g)} \right),$$

where each Ω_g is a norm on $\mathbb{R}^{d_g}$ and, as previously, $\boldsymbol{\beta}^{(g)}$ corresponds to the coefficients of $\boldsymbol{\beta}$ restricted to the features in group g. As before, the λ parameter is a non-negative constant controlling the trade-off between the data fitting term and the regularization term.

Equation (2) is a special case of Equation (B.1) because the squared loss and the logistic loss are convex and differentiable.

Safe screening rules make it possible to solve such problems more efficiently by discarding features whose coefficients are guaranteed to be zero at the optimum, prior to using a solver. They usual rely on the dual formulation of Equation (B.1):

$$\widehat{\boldsymbol{\theta}}^{(\lambda)} = \underset{\boldsymbol{\theta} \in \Delta_X}{\arg\max}\, D_\lambda(\boldsymbol{\theta}) := -\sum_{i=1}^{n} f_i^*(-\lambda\theta_i), \tag{B.2}$$

where $f_i^* : \mathbb{R} \to \mathbb{R}$ is the Fenchel-Legendre transform of f_i, defined by $f_i^*(u) = \sup_{z \in \mathbb{R}} \langle z, u \rangle - f_i(z)$ and $\Delta_X \subset \mathbb{R}^n$ is defined by $\Delta_X = \left\{ \boldsymbol{\theta} \in \mathbb{R}^n : \forall g = 1, \ldots, G, \Omega_g^D(X^{(g)\top}\boldsymbol{\theta}) \leq 1 \right\}$, where $\Omega_g^D : \mathbb{R}^{p_g} \to \mathbb{R}$ is the conjugate norm of Ω_g, defined by $\Omega_g^D(\boldsymbol{u}) = \max_{\boldsymbol{z} in \mathbb{R}^{p_g} : \Omega_g(\boldsymbol{z}) \leq 1} \langle \boldsymbol{z}, \boldsymbol{u} \rangle$, and $X^{(g)} \in \mathbb{R}^{n \times p_g}$ is the design matrix X restricted to the features/columns in group g.

In our setting,

- $\Omega_g^D(\boldsymbol{u}) = \left\| \boldsymbol{\beta}^{(g)} \right\|_2$ and $\Omega^D(\boldsymbol{u}) = \max_{g=1,\ldots,G} \frac{1}{w_g} \left\| \boldsymbol{u}^{(g)} \right\|_2$.
- If one uses the squared loss, that is to say, $f_i(z) = \frac{1}{2}(y_i - z)^2$, then $f_i^*(z) = \frac{1}{2}z^2 + y_i z$ and the Lipschitz constant is $\gamma = 1$.
- If one uses the logistic loss, that is to say, $\boldsymbol{y} \in \{0, 1\}^n$ and $f_i(z) = -y_i z + \log(1 + \exp(z))$, then

  $$f_i^*(z) = \begin{cases} (z + y_i)\log(z + y_i) + (1 - (z + y_i))\log(1 - (z + y_i)) & \text{if } 0 \leq (z + y_i) \leq 1 \\ +\infty & \text{otherwise,} \end{cases}$$

  and the Lipschitz constant is $\gamma = 4$.

The general idea of safe-screening rules, introduced by [EGVR10], is to find a region $\mathcal{R} \subset \mathbb{R}^n$ such that if $\widehat{\boldsymbol{\theta}}^{(\lambda)} \in \mathcal{R}$, for any $g \in \{1, \ldots, G\}$,

$$\Omega_g^D \left(X^{(g)\top} \widehat{\boldsymbol{\theta}}^{(\lambda)} \right) < 1 \Rightarrow \widehat{\boldsymbol{\beta}}^{(\lambda)} = 0.$$

Gap safe screening rules [N+17] exploit the duality gap $(P_\lambda(\boldsymbol{\beta}) - D_\lambda(\boldsymbol{\theta}))$ to obtain the radius of the safe region $\mathcal{R}$. More specifically, Ndiaye et al. show that $\forall \boldsymbol{\beta} \in \mathbb{R}^p, \forall \boldsymbol{\theta} \in \Delta_X$,

$$\left\| \widehat{\boldsymbol{\theta}}^{(\lambda)} - \boldsymbol{\theta} \right\|_2 \leq \sqrt{\frac{2(P_\lambda(\boldsymbol{\beta}) - D_\lambda(\boldsymbol{\theta}))}{\gamma\lambda^2}},$$

which leads them to define, for any $\boldsymbol{\beta} \in \mathbb{R}^p$ and $\boldsymbol{\theta} \in \Delta_X$, the ball centered in $\boldsymbol{\theta}$ and of radius $\sqrt{\frac{2P_\lambda(\boldsymbol{\beta}) - D_\lambda(\boldsymbol{\theta})}{\gamma\lambda^2}}$ as a safe region, that is to say a region that is guaranteed to contain $\widehat{\boldsymbol{\theta}}^{(\lambda)}$.

Appendix B.4. *Measuring selection stability*

To measure the stability of a feature selection property, we use the sample's Pearson coefficient [NB16]. This stability index is closely related to that proposed by Kuncheva [Kun08] and is appropriate for the comparison of feature sets of different sizes. This index relies on repeating the feature selection procedure M time (in this work, $M = 10$) and evaluating the overlap if the M resulting feature sets.

Each of the M sets of selected features can be represented by an indicator vector $\boldsymbol{s} \in \{0,1\}^p$, where $s_j = 1$ if feature j is selected and 0 otherwise. The stability index between two feature sets $\mathcal{S}$ and $\mathcal{S}'$, represented by their indicator vectors $\boldsymbol{s}$ and $\boldsymbol{s}'$, is computed as the Pearsons's correlation between these two vectors:

$$\phi(\mathcal{S}, \mathcal{S}') = \frac{\sum_{j=1}^{p}(s_j - \bar{\boldsymbol{s}})(s'_j - \bar{\boldsymbol{s}'})}{\sqrt{\sum_{j=1}^{p}(s_j - \bar{\boldsymbol{s}})^2}\sqrt{\sum_{j=1}^{p}(s'_j - \bar{\boldsymbol{s}'})^2}}, \tag{B.3}$$

where $\bar{\boldsymbol{s}} = \frac{1}{p}\sum_{j=1}^{p} s_j$ and $\bar{\boldsymbol{s}'} = \frac{1}{p}\sum_{j=1}^{p} s'_j$.

Note that, because $\sum_{j=1}^{p} s_j = |\mathcal{S}|$, $\sum_{j=1}^{p} s_j s'_j = |\mathcal{S} \cap \mathcal{S}'|$, and $s_j^2 = s_j$, we can rewrite Equation (B.3) as

$$\phi(\mathcal{S}, \mathcal{S}') = \frac{|\mathcal{S} \cap \mathcal{S}'| - \frac{1}{p}|\mathcal{S}|\,|\mathcal{S}'|}{\sqrt{|\mathcal{S}|\left(1 - \frac{|\mathcal{S}|}{p}\right)}\sqrt{|\mathcal{S}'|\left(1 - \frac{|\mathcal{S}'|}{p}\right)}},$$

hence interpreting this index as the size of the intersection of the two sets, corrected by chance, that is to say, ensuring that the expected value of the index is 0 when the two selections are random.

The stability index between M sets of selected features is computed as the average pairwise stability index between all possible pairs of sets of selected features:

$$\phi(\mathcal{S}_1, \mathcal{S}_2, \ldots, \mathcal{S}_M) = \frac{M(M-1)}{2}\sum_{k=1}^{M}\sum_{l=k+1}^{M}\phi(\mathcal{S}_k, \mathcal{S}_l). \tag{B.4}$$

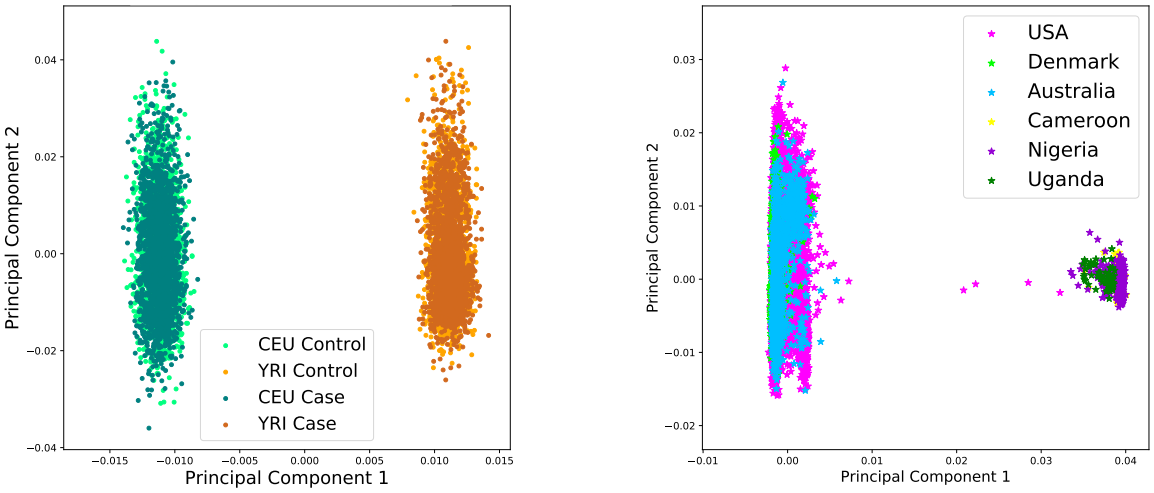
Appendix C. Supplementary Results

Appendix C.1. *PCA of the genotypes*

Figure C1 shows the genotypes of the simulated data (Figure C1a) and the DRIVE data (Figure C1b) projected on the two first principal components of the data.

Appendix C.2. *Runtimes*

Figure C2 shows the runtimes of the different Lasso methods on simulated data.



(a) Population structure in simulated data (b) Population structure in the DRIVE data

Fig. C1. PCA for simulated and real datasets

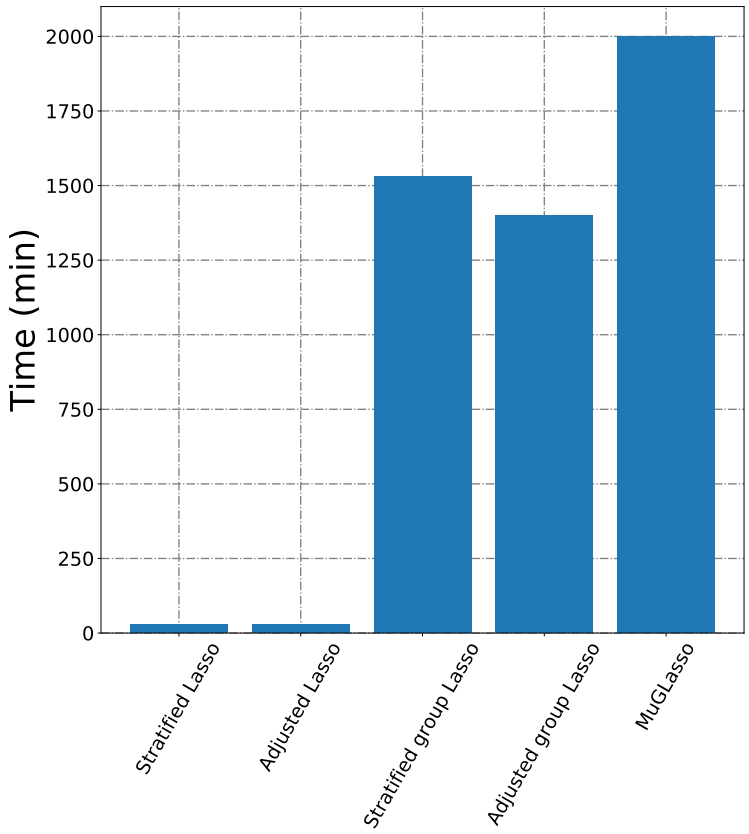


Fig. C2. Runtimes of the different Lasso approaches.

Supplementary References

EGVR10. Laurent El Ghaoui, Vivian Viallon, and Tarek Rabbani. Safe feature elimination for the lasso and sparse supervised learning problems. *arXiv preprint arXiv:1009.4219*, 2010.

Kun08. Ludmila I. Kuncheva. A stability index for feature selection. *IASTED ICAIA*, 2008.

N$^+$17. Eugene Ndiaye et al. Gap safe screening rules for sparsity enforcing penalties. *Journal of Machine Learning Research 18*, 2017.

NB16. Sarah Nogueira and Gavin Brown. Measuring the stability of feature selection. *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 2016.