Multi-task group lasso for Genome Wide Association Studies

Asma Nouira[1,2,3] and Chloé-Agathe Azencott[1,2,3]

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

[2] Institut Curie, PSL Research University, F-75005 Paris, France

[3] INSERM, U900, F-75005 Paris, France

Statistical Methods for Post Genomic Data 2020

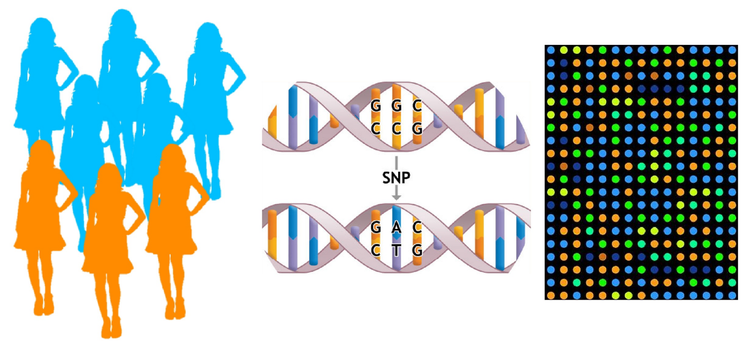
Population Stratification in case-control studies

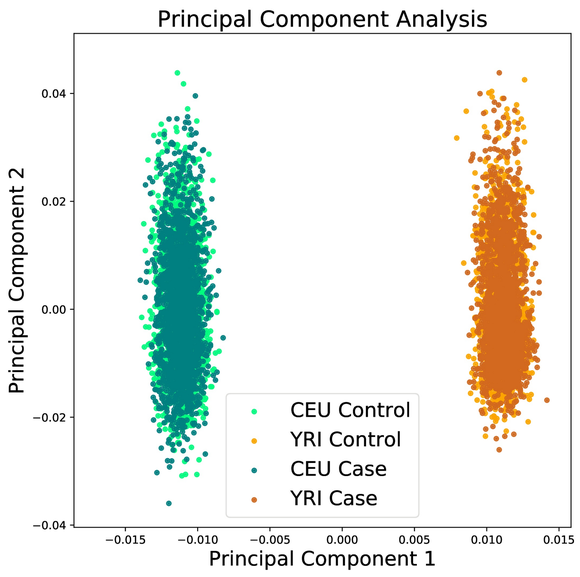
- Correct for associations due to ancestries rather than associations between genotype and phenotype.
- The state-of-the-art adjustment methods lead to over correction of some population-specific causal SNPs.

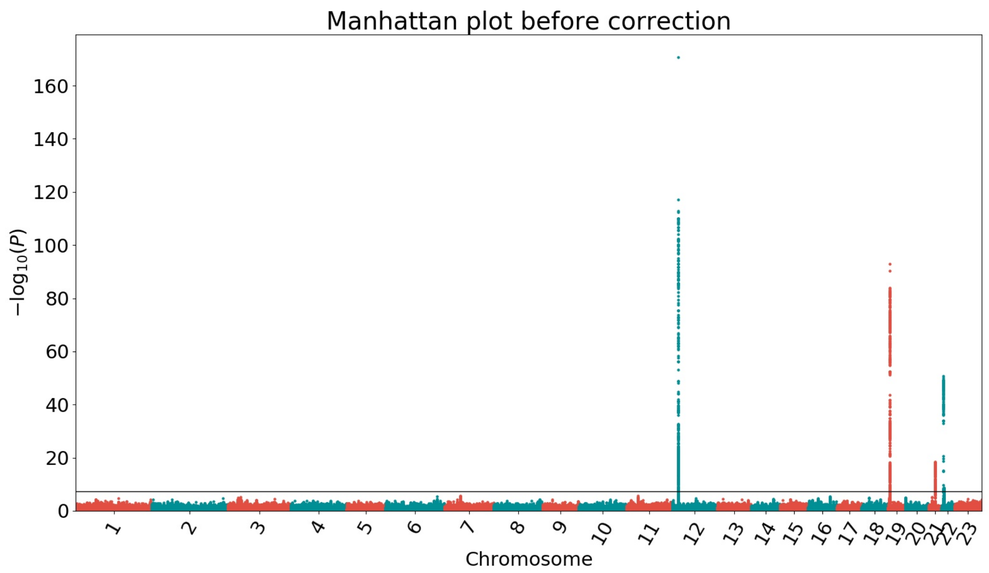
Simulated data with GWAsimulator[1]

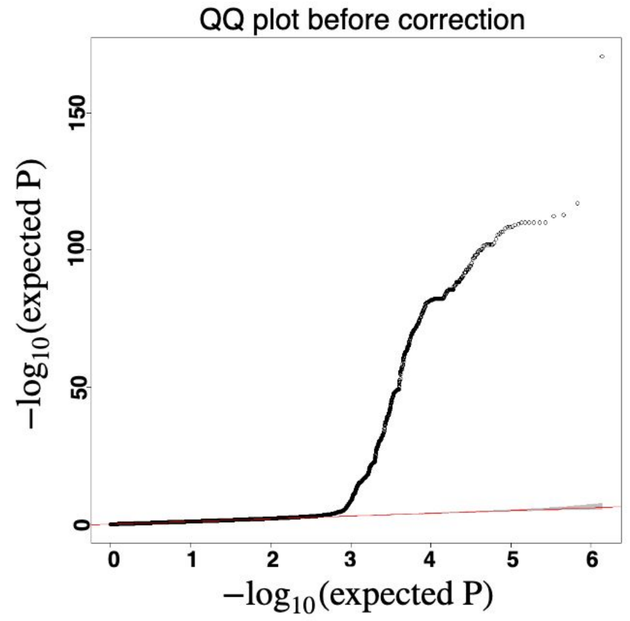
- **Population:** 4000 Yoruba Nigerian (YRI)
  4000 Northern and Western European (CEU)
- **Categorical phenotype**: 4000 cases and 4000 controls
- **Disease loci information:**

Population	Chromosome	Position	Start	End
CEU	12	9067	8000	10000
	19	2885	1000	4000
	21	9659	6000	10000
	22	3357	3000	8000
YRI	12	9067	4000	11000
	19	2885	1000	4000
	21	9659	5000	11000
	22	3357	1000	5000









PCA-based adjustments methods

- **EIGENSTRAT[2]:** Multivariate linear model including PCs

$$x_{ij}^{adj} = x_{ij} - \gamma_j a_i \;;\; \gamma_j = \frac{\sum_i a_i x_{ij}}{\sum_i a_i^2} \text{ and } \sum_i a_i^2 = 1$$

- **Logistic regression with PCs covariates:**

$$\log\left(\frac{q}{1-q}\right) = \beta x + b_1\Phi_1 + b_2\Phi_2 + \ldots + b_d\Phi_d$$

LD blocks partitioning using BALD[3]

- **Ward's linkage criterion:** Compute distance between markers

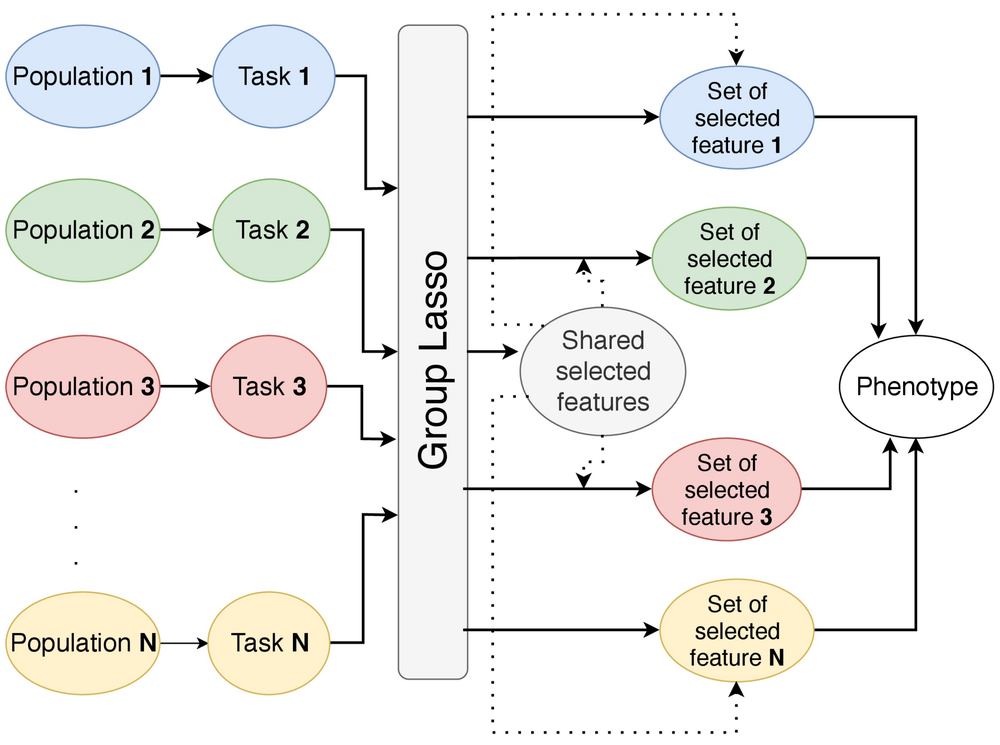
$$d_{wl}(A,B) = \frac{p_A \times p_B}{p_A + p_B}\|g_A - g_B\|_2^2$$

- **Gap statistic approach:** Estimate the blocks number

$$\mathrm{Gap}(G) = \frac{1}{B}\sum_{b=1}^{B}\log(w_G^b) - \log(w_G)$$

Multi-task Group lasso architecture

Samples populations are used as input tasks.



- **Loss function**

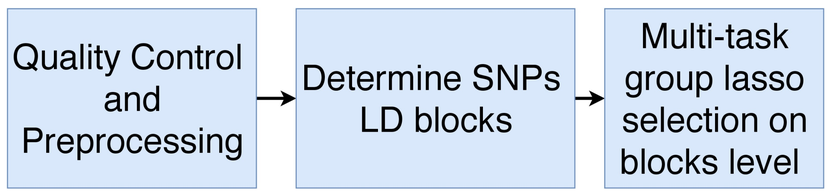
$$\min_{\beta\in\mathbb{R}^{T\times p}} \sum_{t=1}^{T}\frac{1}{n_t}\sum_{m=1}^{n_t}\left\| Y^{(tm)} - \left(\beta_{t0} + \sum_{j=1}^{p}\beta_j^{(t)}X_j^{(tm)}\right)\right\|_2^2 + \lambda\sum_{g=1}^{G}\sum_{t=1}^{T}\sqrt{pg}\left\|\beta_g^{(t)}\right\|_2$$

Correspondance to:
asma.nouira@mines-paristech.fr

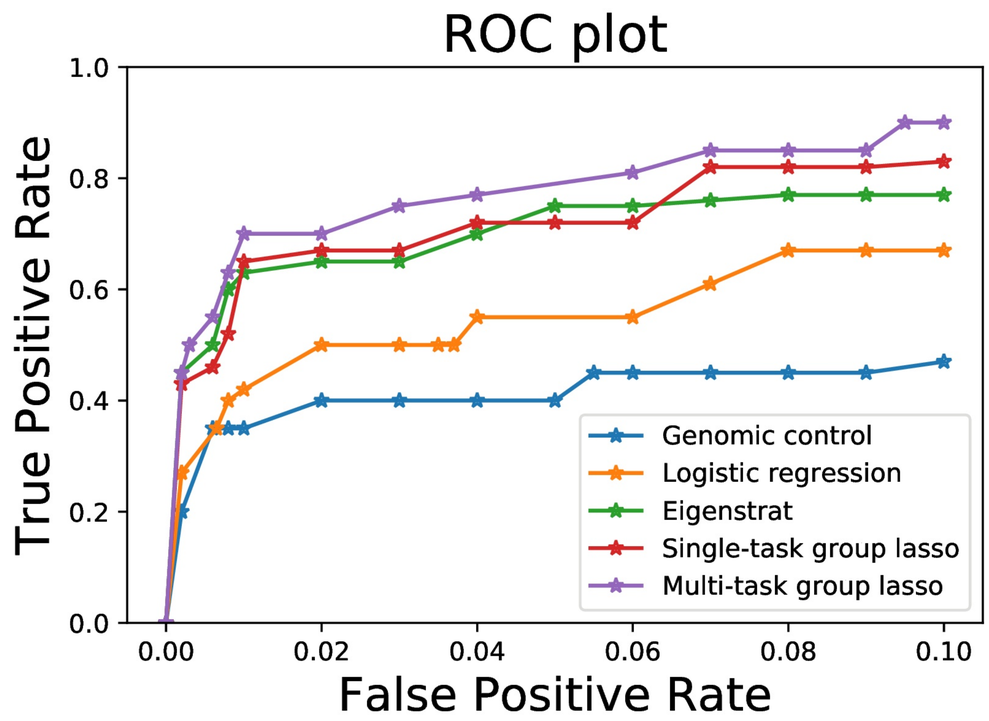
Methods

The proposed Multi-task method aims to adjust population stratification and consists of:

- Clustering of SNPs into blocks following Linkage disequilibrium (LD) patterns.
- Feature Selection at the block level.
- Multi-task group lasso where tasks are populations and groups are LD blocks.



Results



Conclusion

- Multi-task feature selection model deals with population stratification issue.

Future work

- Measure the stability of the selection of the multi-task group lasso.
- Enforce the stability of the selection in the multi-task approach.
- Apply the proposed method to other simulated data cases.

Acknowledgment

This work was supported by the French National Research Agency (ANR-18-CE45-0021-01).

References

[1] Li C et al., GWAsimulator: a rapid whole-genome simulation program, Bioinformatics (2008).
[2] Price et al., Principal components analysis corrects for stratification in genome-wide association studies, Nature Genetics (2006).
[3] Dehman et al., Performance of a blockwise approach in variable selection using linkage disequilibrium information, BMC Bioinformatics (2015).
[4] Puniyani et al., Multi-population GWA mapping via multi-task regularized regression, Bioinformatics (2010).