Incorporating linkage disequilibrium blocks in Genome-Wide Association Studies

Alia DEHMAN[1], Christophe AMBROISE[2] and Pierre NEUVIAL[2]

Laboratoire Statistique et Génome, Université d'Évry Val d'Essonne, UMR CNRS 8071 – USC INRA
{alia.dehman, christophe.ambroise, pierre.neuvial}@genopole.cnrs.fr

Abstract *In genome-wide association studies, we are interested in finding genetic markers that are significantly associated with a phenotype of interest. Whole-genome single nucleotide polymorphism (SNP) data are collected for many thousands of SNP markers, leading to high-dimensional regression problems where the number of predictors greatly exceeds the number of observations. Moreover, these predictors are highly dependent, in particular due to linkage disequilibrium (LD).*

We propose a two-step approach that explicitly takes advantage of the grouping structure induced by LD. In the first step, we infer LD blocks by performing a clustering of LD estimates with an adjacency constraint. In the second step, we perform Group Lasso regression on the inferred LD blocks.

We argue that it is relevant to assess performance both at the scale of individual SNPs and at the scale of LD blocks. We investigate the efficiency of this approach compared to state-of-the art penalized regression methods (Lasso and Elastic-Net) at these two scales. Our numerical experiments show that the proposed approach not only activates the groups containing the associated SNPs but is also as precise as the penalized models as for selecting individual associated predictors.

Keywords Genome-wide association studies, penalized regression, Group Lasso, linkage disequilibrium.

1 Introduction

With recent advances in hight-throughput genotyping technology, genome-wide association studies (GWAS) have become an important tool for identifying genetic markers underlying a variation in a given phenotype. In GWAS, it is expected that only a subset of SNPs are significantly associated with the phenotype. Therefore, the SNP selection problem can be formulated as a variable selection problem in sparse and high-dimensional settings. SNPs can also be highly correlated due to the phenomenon of linkage disequilibrium (LD) and a natural group structure can thus be considered among the variables.

The Lasso [6] is an efficient multivariate variable selection method in high-dimensional problems. However, in presence of a group structure on the predictors with high pairwise correlations among them, the Lasso tends to select only one variable from each group and discard the others. Therefore, the "group version" of the Lasso, the Group Lasso [10], was introduced in order to account for this structure in penalized regression model. Like other group-adapted regression methods, such as structured elastic-net [5] , group-MCP [2], group-SCAD [7] or the group SMCP [4], the Group Lasso involves a structured penalty allowing sparse group regression.

As SNPs of a LD block can be correlated regardless of the phenotype, and as causal SNPs can in turn be correlated to non causal ones, we argue that it may make more sense to look for *SNP groups* (that is, LD blocks) that are significantly associated with the phenotype, rather than looking for *individual SNPs*. From a biological point of view, this is particularly relevant in a context where "causal SNP" (or, more generally, causal loci) need not be observed: it is possible that the observed data only contain SNPs that are in LD with causal ones. From a statistical perspective, the distinction between SNP-level and LD block-level associations is related to an identifiability issue: assuming that causal SNPs are observed, is their association to the phenotype strong enough so that they can be distinguished from indirect associations between SNPs in strong LD with causal ones ?

In this paper, we propose a two-step method consisting on inferring LD blocks using a spatially-constrained hierarchical agglomerative algorithm before applying the Group Lasso regression model. This approach is compared do state-of-the-art penalized regression approaches used in high-dimensional problems, for which prior group structure information is ignored (Lasso) or incorporated less directly (Elastic-Net). Competing methods are evaluated in terms of their ability to retrieve *groups of SNPs* associated to the phenotype, and to retrieve *individual SNPs* associated to the phenotype.

The rest of the paper is organized as follows: in Section 2 we describe the proposed two-step approach. The competing mehods and the evaluation are described in Section 3. Results of our simulation studies are presented in Section 4. Finally, Section 5 provides a summary and discusses some related issues.

2 A Two-Step Approach Taking the Group Structure Into Account

We consider the problem of predicting a continuous response $\mathbf{y} \in \mathbb{R}^n$ from covariates $\mathbf{X} \in \mathbb{R}^{n \times p}$. For $i \in \{1, \ldots, n\}$ and $j \in \{1, \ldots, p\}$, $\mathbf{X}_{i\cdot}$ is a p-dimensional vector of covariates for observation i and $\mathbf{X}_{\cdot j}$ is a n-dimensional vector of observations for covariate j. For each $i \in \{1, \ldots, n\}$, we assume that $\mathbf{X}_{i\cdot}$ has a block structure with G blocks of sizes $p_1, \ldots p_G$, with $\sum_{g=1}^{G} p_g = p$. Thus $\mathbf{X}_i = (\mathbf{X}_i^1, \ldots, \mathbf{X}_i^G)$ with each $\mathbf{X}_i^g \in \mathbb{R}^{p_g}, g = 1, \ldots, G$. We note β_g the coefficient vector corresponding to the g^{th} group. We propose a two-step method consisting in inferring the LD blocks using only the genotype data, and then performing a Group Lasso regression on the inferred blocks.

2.1 Inference of Blocks from Genotypes

For this algorithm's first step of inferring groups, only the genotype data are used. The LD measure D' is calculated from the genotype matrix [3] to obtain a $p \times p$ matrix of pairwise D' measures. Then, the Fisher transformation is applied to the LD matrix to obtain quantities that are approximately normally distributed. Finally, we perform a constrained hierarchical clustering of the LD matrix, as now described.

Our clustering procedure is based on the one of the most widely used methods of cluster analysis, the Ward's incremental sum of squares method [8]. The general goal of sum of squares clustering is to minimize the total within-cluster dispersion for G groups around G centroids. If we denote by D_k the sum of squares (or dispersion) for cluster k, then the increase of dispersion after merging clusters k and l is $I_{kl} = D_{kl} - D_k - D_l$. The standard agglomerative hierarchical approach is to start with p clusters of size 1, and to sucessively merge the two clusters k and l which yield the smallest increase in dispersion I_{kl}, until only 1 cluster of size p remains. Our constrained clustering is a simple modification that takes advantage of the fact that LD matrix can be modeled as block-diagonal: at each step of the agglomerative process, we only merge clusters that are *adjacent on the genome*. By construction, this constrained clustering is much faster than the standard hierarchical clustering. The desired number of blocks currently has to be set by the user.

For the numerical experiments reported below, we have used the R package `rioja` [1] which implements this constrained hierarchical clustering. We note that this implementation requires a $p \times p$ matrix of similarities to be computed. This can be quite problematic in the context of GWAS where p can be as large as 10^5 or 10^6. In order to overcome this limitation, it is possible to implement an algorithm where LD measures are not passed as a matrix but calculated on the fly.

2.2 Selection of Blocks Associated with Phenotype

Once LD blocks have been identified, we use the Group Lasso [10] to identify blocks associated with the phenotype. Well adapted to group-structured variables, the Group Lasso estimator is defined as:

$$\hat{\boldsymbol{\beta}}_\lambda = \arg\min_{\boldsymbol{\beta}} (||\mathbf{y} - \mathbf{X}\boldsymbol{\beta}||_2^2 + \lambda \sum_{g=1}^{G} \sqrt{p_g} ||\boldsymbol{\beta_g}||_2),$$

where $||.||_2$ is the Euclidean norm and λ is a penalty parameter. The Group Lasso has the specificity of being a group selection method: by construction, the estimated coefficients within a group will either all be zero or all nonzero.

3 Performance Evaluation

3.1 Simulation Study

Our simulation setting is adapted from the model used in Wu et. al.[9]. We set $n = 200$ and $p = 512$, with 9 groups of sizes $(2, 2, 4, 8, 16, 32, 64, 128, 256)$. The ordering of the groups is drawn at random for each simulation. If $j \neq j'$ are in the same group, $cov(\mathbf{X}_{\cdot j}, \mathbf{X}_{\cdot j'}) = \rho$ else $cov(\mathbf{X}_{\cdot j}, \mathbf{X}_{\cdot j'}) = 0$. For all $j \in \{1, \dots, p\}, \mathbf{X}_{\cdot j}$ is generated from a p-dimensional multivariate normal distribution whose covariance matrix is a block diagonal matrix. Then, we set X_{ij} to $0, 1$ or 2 according to whether $X_{ij} < -c$, $-c < X_{ij} < c$ or $X_{ij} > c$ with c the first quartile of a standard normal distribution. Finally, the first two SNPs of groups of sizes $2, 2, 4, 8$ are chosen to be associated with the phenotype. The strength of the association is calibrated by the coefficient of determination R^2 of the model. The parameters of the simulation are therefore R^2 and ρ.

3.2 ROC-Based Evaluation

Our performance assessment aims at evaluating the ability of our proposed method to distinguish true signals from noise. As the association study is block-oriented, some definitions of "true signal" need to be specified. We define a *causal SNP* as a SNP that is simulated with a non-zero regression parameter. We also define a *block-associated SNP* as a predictor that is not directly associated with the phenotype but simulated in the same block that a causal SNP, and then can be highly correlated with it. As explained in Section 1, we are interested in two types of evaluations: a *block-level evaluation*, to assess how well a given method retrieves block-associated SNPs, and a *SNP-level evaluation*, to assess how well a given method retrieves causal SNPs.

Performance is evaluated using receiver operator characteristics (ROC) curves. For a given threshold, we plot the true positive rate (TPR), which is the fraction of true positives out of the positives versus the false positive rate (FPR), which is the fraction of false positives out of the negatives. To plot the ROC curves, we first evaluate, for each method, the TPR and FPR for a grid of underlying regularization parameter values and for each simulation. Then, we aggregate the curves at fixed parameter i.e, we calculate average TPR and FPR across all simulation runs for each underlying parameter value.

3.3 Competing Approaches Based on Penalized Regression

The proposed approach is compared to two state-of-the art competitors that do not explicitly take the block-structure information into account: Lasso [6] and Elastic-Net [11]. The estimators of Lasso and Elastic-Net, respectively noted β_{lasso} and β_{EN} are defined as:

$$\hat{\boldsymbol{\beta}}_{lasso} = \arg\min_{\boldsymbol{\beta}} ||\mathbf{y} - \mathbf{X}\boldsymbol{\beta}||_2^2 + \lambda||\boldsymbol{\beta}||$$

$$\hat{\boldsymbol{\beta}}_{EN} = \arg\min_{\boldsymbol{\beta}} ||\mathbf{y} - \mathbf{X}\boldsymbol{\beta}||_2^2 + \lambda_1||\boldsymbol{\beta}|| + \lambda_2||\boldsymbol{\beta}||_2^2$$

with λ, λ_1 and λ_2 three regularization parameters. Thanks to the l_1 penalty, the Lasso model encourages sparsity by setting many regression coefficients for irrelevant SNPs to exactly zero. However, the Lasso does not incorporate any information on the group structure induced by LD blocks. Like the Lasso, the Elastic-Net simultaneously performs both automatic variable selection and continuous shrinkage. Unlike the Lasso, the Elastic-Net includes a ridge (ℓ_2) penalty which tends to select groups of correlated variables. Therefore, the Elastic-Net incorporates some prior information regarding the block structure of the data. However, unlike the proposed method, it does not take advantage of the fact that blocks are adjacent along the genome.

4 Results

We have performed a comprehensive simulation study, where the correlation coefficient $\rho \in \{0, 0.1, 0.2, 0.3, 0.5, 0.8\}$ and the determination coefficient $R^2 \in \{0.1, 0.2, 0.3, 0.5, 0.8\}$. We summarize below the obtained results for $R^2 = 0.2$ only: indeed, we believe that this is a low but (unfortunately) realistic value of R^2 in GWAS studies. The results obtained for larger values of R^2 were similar, in the sense that the ranking of the methods for a given ρ was the same, although the performance of each particular method is obviously an increasing function of R^2. Each of the ROC curves has been obtained by averaging $B = 50$ simulation runs.

4.1 Influence of LD

In order to evaluate the influence of LD on the performance of the methods, we assessed the different approaches on data sets with different degrees of correlation ρ among variables. We set the number of inferred clusters for the Group Lasso to 9 groups, that is, the (oracle) number of groups actually used for simulations. The results are summarized in Fig. 1 for $\rho \in \{0, 0.1, 0.2\}$. For larger values of ρ the results are not shown because they were quite similar to the case $\rho = 0.2$, except that the performance of the Lasso deteriorates for $\rho \geq 0.5$. In our experiments, the LD within block is of the same order of magnitude as ρ; therefore, we see the case $\rho = 0.2$ as the most realistic.

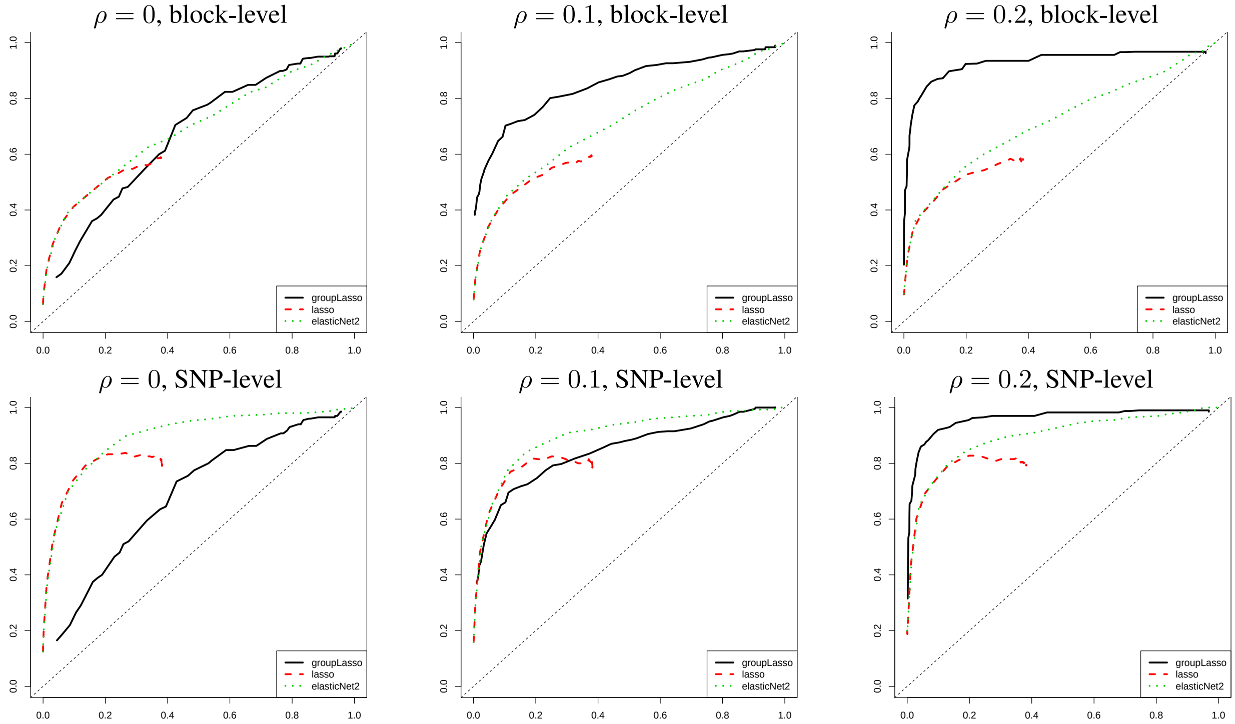

Figure 1. ROC curves (TPR in function of FPR) for the proposed method ("group-lasso", black solid lines), Lasso (dashed red lines) and Elastic Net (green dotted lines) for $\rho \in \{0, 0.1, 0.2\}$. Top row: block-level evaluation; bottom row: SNP-level evaluation. The number of clusters for the proposed method is set to the true number of clusters.

For all positive values of correlation, the proposed method outperforms the other methods in the block-level evaluation. This is not surprising, as this method has been constructed to work in such situations. However, we believe that good performance at the block level is a very encouraging features for GWAS.

For the SNP-level evaluation, the proposed method also outperforms Lasso and Elastic Net as soon as $\rho \geq 0.2$. This is quite remarkable, because by construction the proposed method is bound to select either all SNPs of a block, or none, while there are at most 2 causal SNPs per block. Therefore, even if our proposed method is advantaged by the fact that it is given the true number of blocks, it is also a priori clearly disadvantaged by the SNP-level evaluation. We also observe that for specificities larger than 80% (that is, FPR $<$ 20%) there is virtually no difference between the Lasso and the Elastic-Net, suggesting that the correlation structure is too weak (for small but realistic values of ρ) for the grouping effect of the Elastic-Net to be effective. The grouping effect of the proposed method is more effective because this method specifically looks for blocks of *adjacent SNPs*.

4.2 Varying the Number of Clusters

The setting considered in the preceding section may be seen as an Oracle setting for the proposed method, in the sense that the number of blocks was set to the true number of blocks. In practice however, the true number of blocks is unknown, and it currently has to be set by the user, so we have investigated the robustness of the proposed method to this parameter. Using the same simulation setting as above where the true number of clusters is 9, we have evaluated the proposed method when applied with 5 target clusters, corresponding to an "under-clustering" situation and with 13 target clusters, corresponding to an "over-clustering" situation. The results are shown in Fig. 2 for the under-clustering, and Fig. 3 for over-clustering. By construction, the performance of Lasso and Elastic Net does not depend of the target number of clusters.

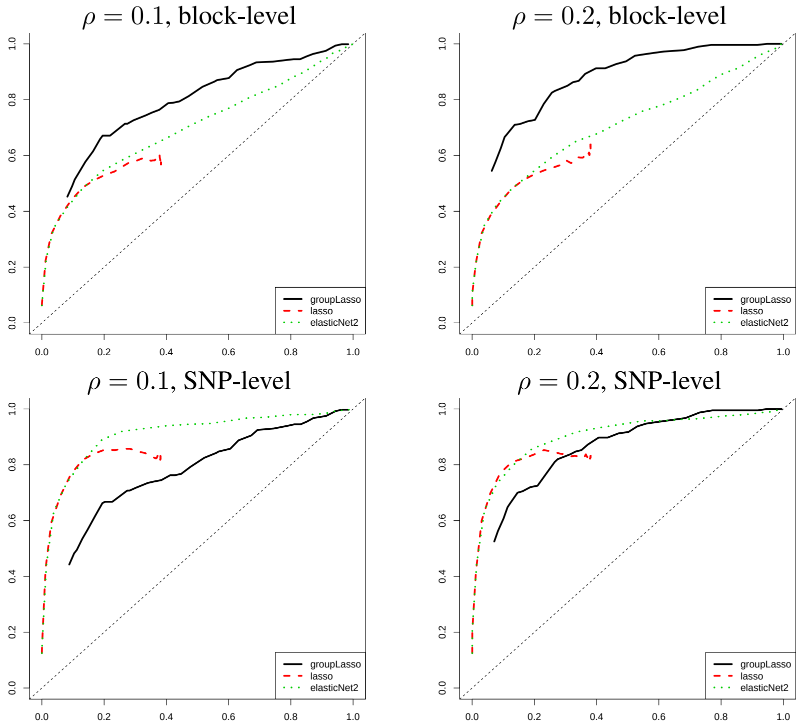

Figure 2. ROC curves (TPR in function of FPR) for the proposed method ("group-lasso", black solid lines), Lasso (dashed red lines) and Elastic Net (green dotted lines) for $\rho \in \{0.1, 0.2\}$. Top row: block-level evaluation; bottom row: SNP-level evaluation. The number of clusters for the proposed method is set to 5.

As in Fig. 1, the proposed two-step approach outperforms its competitors for block-level evaluation, both for under- and over-clustering. For SNP-level evaluation, the proposed method is outperformed by Lasso and Elastic net for under-clustering (Fig. 2). Indeed, with a target number of groups smaller than the actual one, the Group Lasso makes mistakes by canceling or activating too large groups. As for the over-clustering (Fig. 3) it is remarkable that the proposed method outperforms both Lasso and Elastic Net: the Group Lasso does not suffer from over-clustering, as it can activate the "good" blocks among the ones clustered.

5 Conclusion and Perspective

In this paper, we have proposed a two-step approach that takes into account the biological information of the linkage disequilibrium between variables by firstly inferring LD blocks and then performing Group Lasso regression. State-of-the-art one-stage variable selection regression methods Lasso and Elastic-Net are outperformed by our proposed method for the purpose of identifying blocks containing causal SNPs, which we argue is a quite interesting feature in practice given the underdetermination of most GWAS. Interestingly, although the proposed method can only select groups of SNPs and not individual SNPs, it also achieves similar or better performance than its competitors in terms of selection of "causal SNPs". We believe that these results illustrate the relevance of the approach, and thereby the importance of tailored integration of biological knowledge in high-dimensional genomic studies such as GWAS.

A current limitation of the method is that it does not perform automatic model selection at the clustering stage, meaning that the used has to explicitly specify a target number of blocks. Our results in the case where the target number of blocks is misspecified (Section 4.2) are encouraging, as they show that the proposed method is fairly robust to situations where the target number of blocks is over-estimated. In order to improve this aspect of the method, several directions can be investigated:

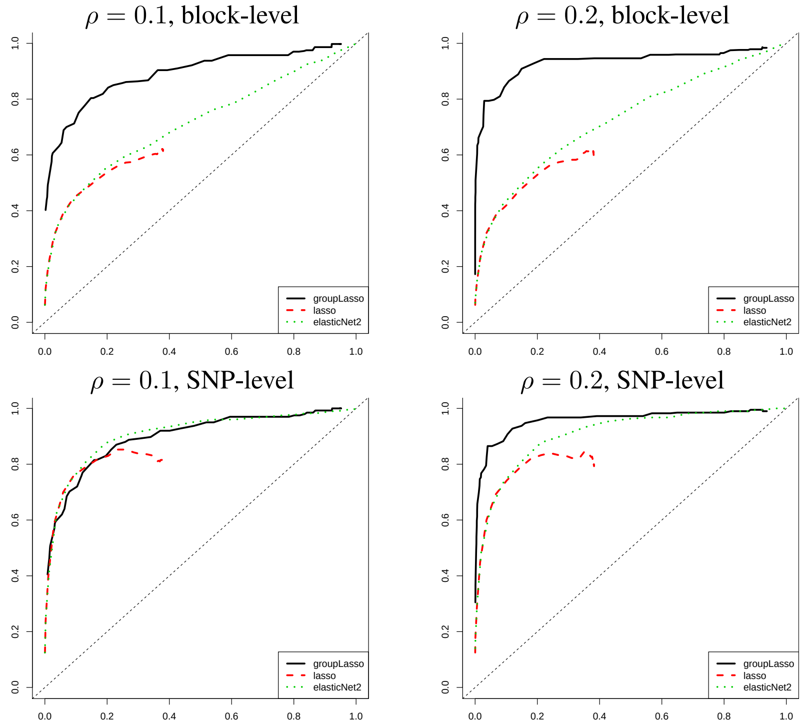

Figure 3. ROC curves (TPR in function of FPR) for the proposed method ("group-lasso", black solid lines), Lasso (dashed red lines) and Elastic Net (green dotted lines) for $\rho \in \{0.1, 0.2\}$. Top row: block-level evaluation; bottom row: SNP-level evaluation. The number of clusters for the proposed method is set to 13.

– deriving model selection criteria that are adapted to the proposed constrained clustering;
– avoiding model selection by replacing the Group Lasso by a *Hierarchical* Group Lasso.
– replacing the current two-stage approach by a one-stage penalized regression method, by constructing a penalty that takes full advantage of the prior information that relevant groups of predictors can be expected to be adjacent along the genome.

Acknowledgements

This work was partially funded by Institut National du Cancer (INCa) and Cancéropôle Ile-de-France.

The authors warmly thank Cyril Dalmasso and Julien Chiquet for very helpful discussions.

References

[1] K. D. Bennett. Determination of the number of zones in a biostratigraphical sequence. *New Phytologist*, 132(1):155–170, 2006.

[2] P. Breheny and J. Huang. Penalized methods for bi-level variable selection. *Statistics and its interface*, 2(3):369, 2009.

[3] D. Clayton and H.-T. Leung. An R package for analysis of whole-genome association studies. *Human heredity*, 64(1):45–51, 2007.

[4] J. Liu, J. Huang, S. Ma, and K. Wang. Incorporating group correlations in genome-wide association studies using smoothed group lasso. *Biostatistics*, 2012.

[5] M. Slawski, W. Zu Castell, and G. Tutz. Feature selection guided by structural information. *The Annals of Applied Statistics*, 4(2):1056–1080, 2010.

[6] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

[7] L. Wang, G. Chen, and H. Li. Group SCAD regression analysis for microarray time course gene expression data. *Bioinformatics*, 23(12):1486–1494, 2007.

[8] J. H. Ward Jr. Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*, 58(301):236–244, 1963.

[9] T. T. Wu, Y. F. Chen, T. Hastie, E. Sobel, and K. Lange. Genome-wide association analysis by lasso penalized logistic regression. *Bioinformatics*, 25(6):714–721, 2009.

[10] M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67, 2005.

[11] H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005.