2011 International Conference on Bioscience, Biochemistry and Bioinformatics
IPCBEE vol.5 (2011) © (2011) IACSIT Press, Singapore

Pathway Selection for GWAS Using the Group Lasso with Overlaps

Matt Silver, Giovanni Montana
Statistics Section, Dept. of Mathematics,
Imperial College, London, UK g.montana@imperial.ac.uk
Alzheimer's Disease Neuroimaging Initiative*

***Abstract*—We present a statistical model for the ranking of significant pathways from genome-wide association studies (GWAS) in which SNPs are grouped into functionally related gene sets or path- ways. We use a sparse regression model with the distinguishing feature that it is able to account for the presence of overlapping gene sets, arising from the typically large number of genes that are assigned to multiple pathways. Pathways selection is carried out using a resampling approach, stability selection. The final algorithm is highly computation- ally efficient when compared with other methods that use permutations to rank significance. With simulated quantitative phenotypes and real genotype and pathway data, we demonstrate that our method performs well when compared with other widely- used pathway selection strategies.**

Keywords-GWAS, pathways, penalized regression, group lasso*

I. INTRODUCTION

The mixed success of attempts to identify genetic variants that account for a large part of the heritability of common disease has focused attention on the need to develop new methodological approaches for the analysis of GWAS data [1]. One promising approach, first developed for the analysis of gene expression data, uses prior information on gene function to group genes and associated SNPs into gene sets or pathways [2]. The motivation here is that by jointly considering the effects of multiple SNPs or genes within a biological pathway, significant associations might be identified that would otherwise be missed when considering markers individually. As well as offering the potential for increased statistical power, pathways-based GWAS (PGWAS) can ease the biological interpretation of results, and may also facilitate the comparison of results between different datasets [3, 4, 5]. A typical PGWAS begins with a univariate test of association in which individual SNPs are scored according to their degree of association with disease status or a quantitative trait. Various techniques are then used to combine these univariate statistics into pathway scores. In the GenGen pathways analysis program for example [6], all genes are first ranked according to the value of the highest scoring SNP within 500kb. Pathway significance is then assessed by determining the degree to which high-ranking genes are over-represented in a given pathway. As a final step, where more than one pathway is considered a correction for multiple testing is made.

In addition to methods based on univariate and multi-locus test statistics, a number of multivariate penalized regression techniques have recently been proposed for the analysis of GWAS data. In contrast to other methods which focus on hypothesis testing, penalized regression attempts to identify subsets of SNPs that best describe the variation in response by enforcing sparse solutions to the regression equation. Penalized regression methods offer a number of potential advantages over conventional, univariate tests of association. These include the ability to jointly consider all predictors in the model at the same time, as opposed to working only with marginal associations; the ability to deal with correlations between predictors in a principled way; and to incorporate model covariates. Recently, penalized logistic regression has been used to select SNPs and analyze two-way and higher-order SNP-SNP interactions [7], and to identify both common and rare variants by grouping SNPs into genes [8]. Another recent study also groups SNPs into genes within a pathway [9], but uses a lasso penalty at the SNP level, together with group ridge regression at the gene level, to obtain a statistic for pathway association. A pathway p-value is then obtained through permutation.

Our focus here is on the identification of biological pathways that are associated with a quantitative trait, although our method can be extended to case-control studies with minimal effort. The method we propose includes a number of distinguishing features not present in previous sparse regression-based PGWAS methods. These include the use of group lasso penalized regression, with SNPs aggregated into pathways, enabling us to consider the joint effects of multiple pathways at the same time. We find that the group lasso allows us to deal with correlations between SNPs within genes in an elegant way, with the proviso that suitable pathway weighting schemes are used. An important feature of our proposed strategy is its ability to account for overlaps between groups, arising from the large number genes that belong to multiple pathways. We find that overlaps between pathways can significantly bias pathway rankings by inflating false positive rates. Group lasso was originally proposed for disjoint groups, and does not handle

*Data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (www.loni.ucla.edu/ADNI). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or pro- vided data but did not participate in analysis or writing of this report. A complete list of ADNI investigators is available at http://www.loni.ucla.edu/ADNI/Collaboration/ADNI Manuscript Citations.pdf.

model selection with overlapping groups well. We overcome this problem by utilizing an extension of the group lasso that was first proposed in the context of pathway selection in gene expression analysis [10]. To our knowledge this is the first application of this method to GWAS. Finally, we use a resampling approach called stability selection [11] to rank pathways. Stability selection was originally proposed as an alternative to methods such as cross validation and AIC for the selection of the optimal number of variables to include in the model. We instead use stability selection to rank pathways by selection frequency, and find it offers considerable benefits in terms of computational efficiency.

The paper is structured as follows. We describe our method and the data sets used in this study in the next section. The performance of the proposed penalized regression model is assessed by means of Monte Carlo simulations and compared to the GenGen method [6] in the following section. We end with a brief discussion and some remarks on future extensions.

II. Materials and Methods

A. Genotypes, pathways and phenotypes

We use real genotype data from n = 500 individuals obtained from the Alzheimer's Disease Neuroimaging Initiative[1]. We remove variants with a call rate < 95%, minor allele frequency (MAF) < 0.1 and significant deviation from Hardy-Weinberg equilibrium ($p < 5.7 \times 10{-}7$). For the purpose of this assessment study we utilize only the first 5000 SNPs on chromosome 1. In order to map genes to pathways, we use the Molecular Signals Database[2] which at the time of our analysis contained 639 pathways mapped to 5390 genes. We map the 5000 SNPs to all genes within 10kb of annotated genes (3006 SNPs mapped to 405 genes), and additionally exclude pathways with less than 10 mapped SNPs (143 pathways) and less than 3 mapped genes (41 pathways). Finally we exclude duplicate pathways containing identical SNPs (24 pathways). After all pre-processing we are left with p = 632 SNPs mapped to 60 pathways, with a significant number of overlaps, corresponding to SNPs belonging to multiple pathways. We denote by x_{ij} SNP j observed on sample i, with $i = 1, \ldots, n$ and $j = 1, \ldots, p$. We also denote by p_g the number of SNPs in pathway g, with $g = 1, \ldots, 60$. Using these real genotypes, we are then able to simulate a vector of quantitative phenotypes y_i with $i = 1, \ldots, n$. We evaluate the performance of the methods over 400 Monte Carlo simulations, each with a randomly selected pathway containing 5 randomly selected 'causal' SNPs under an additive model.

We consider SNP effect sizes of 0.01, 0.03 and 0.05, with SNP effect size defined as the mean proportionate change in phenotype. SNP effect size is held constant irrespective of the SNP's MAF. The effect of different sample sizes is assessed by adding normally-distributed noise to the phenotypes, and controlling signal to noise ratios (SNR).

[1]www.loni.ucla.edu/ADNI
[2]http://www.broadinstitute.org/gsea/msigdb/index.jsp

B. Group lasso regression with overlaps

The p SNPs observed on the n subjects can be arranged in a (n x p) design matrix X. We consider the usual linear regression setting $y = X\beta + \varepsilon$ where $y \in \Re$ is the quantitative trait or phenotype treated as response, $\beta \in \Re^p$ is a vector of regression parameters to be estimated, and $\varepsilon \in \Re$ is i.i.d. noise. In the group lasso, variables are assumed to partition into G groups, each with group parameter vector β_g [12]. The optimal solution $\hat{\beta}$ satisfies

$$\arg\min_{\beta}\left(\left\|y - X\beta\right\|_2^2\right) + \lambda\sum_{g=1}^{G} w_g \left\|\beta_g\right\|_2$$

where the first term corresponds to the OLS solution, with $\|\cdot\|_2^2$ being the square of the l_2 norm, and the second term is the penalty function which enforces sparsity in the model. The regularization constant λ controls the number of groups that enter the model. A group weighting factor, $w_g = \sqrt{p_g}$ is usually applied to adjust group penalties according to group size. The optimization equation is solved using coordinate gradient descent [17]. The group lasso encourages sparsity at the group level by applying an l_1 or lasso penalty, while shrinking parameters values within selected groups with the application of an l_2 ridge-type penalty. For groups not retained in the model, all coefficients are set to zero. A potential advantage of the group lasso is in the situation where groups of variables are correlated, as is the case for SNPs within genes. In this case the within-group ridge penalty ensures that all correlated variables are selected, whereas the lasso penalty selects only one [8].

One potential limitation of the group lasso in the context of PGWAS is in the situation where groups overlap, i.e. where one or more predictors belong to multiple groups. Where this is the case, the group lasso may be unable to distinguish overlapping groups. For example, if a variable has a non-zero parameter value in one selected group, all other groups to which it belongs must also be selected. Conversely, where a variable has a zero parameter value in one (non-selected) group, all groups to which it belongs cannot be selected either. We find with data constructed using real biological pathways, that this means the standard group lasso is often unable to converge. A solution to this problem has been proposed by [10]. In this method, the design matrix of predictors is transformed by duplicating multiple columns to ensure that each group in the expanded variable space is disjoint by construction. This technique is illustrated in Fig. 1 with real data mapping SNPs to pathways used in our analysis.

The choice of $w_g = \sqrt{p_g}$ is motivated by the desire to ensure that group l_2 norms are unbiased by group size. In the context of PGWAS, within group correlation between SNPs due to LD (linkage disequilibrium) is also expected to

influence the size of group norms. A number of solutions have been proposed to deal with the issue of correlation between variables in ordinary lasso regression [14, 15]. Here we seek a means of accounting for correlation in the context of group lasso, by adjusting the group weighting. An obvious correlation measure is LD, but since this can only measure pairwise correlations between SNPs, we consider instead an adjustment to the standard group size weighting based on pathway mutual information (MI). MI is an entropy-based measure which captures the amount of information present in a set of variables. In our analysis we use the normalized mutual information [16] of a pathway, nMI_g to adjust the group penalty, such that

$$w_g = \sqrt{\frac{p_g}{\mathrm{nMI}_g}}$$

C. *Stability selection*

A key challenge with penalized regression is the choice of regularization constant, λ, which controls how many variables (here pathways) are to be retained in the model. Common methods for determining an optimal choice for λ include cross validation and the Akaike Information Criterion. We use an alternative method known as stability selection [11]. In the context of model selection, this emphasises model stability, over model size, by measuring the frequency at which variables are selected across multiple subsamples of the data. Provided that sufficient variables are selected by the model, stability selection performs well irrespective of the actual value of the regularisation constant used [11]. Our algorithm proceeds by first tuning λ in an initial learning phase, so that an average of 5 pathways are selected across a small number of subsamples. This λ value is then used in the full pathway selection phase, using stability selection to rank pathways across 100 subsamples.

III. Results

We compare our method with the recently published GenGen PGWAS method [6]. We use the default settings for GenGen with 1000 permutations, and rank pathways by normalised enrichment score. Results showing the average performance across 400 Monte Carlo simulations with a SNP effect size of 0.05, and with SNRs of 2 and 8 are shown in Table I. Group weighting for the group lasso is scaled by nMI. Performance is measured by the sensitivity of each method, that is the proportion of simulations in which the correct causal pathway is identified, for a given number of false positives (false positive rate). The group lasso with overlaps shows consistently superior performance at all false positive rates, and remarkably so for lower false positive rates. This advantage is maintained at a number of other effect sizes (data not shown). We also find that group weighting scaled by nMI shows superior performance, compared with the standard weighting by group size (data not shown).

IV. Discussion

We applied group lasso, with SNPs aggregated into pathways, to the task of pathway selection. We found that this method performs well in comparison with GenGen. By jointly modelling effects of multiple SNPs within genes across multiple pathways, the group lasso with overlaps is able to rank pathways well, even where there is considerable overlap between them. In contrast, methods such as GenGen which combine scores from single-SNP association tests may be prone to inflated false positive rates by giving undue weight to individual SNP scores.

The method we present here has some attractive features which point to a number of possible extensions. For example, using stability selection it may be possibly to simultaneously select SNPs within selected pathways, including rare variants that are retained in the model due to the within-pathway ridge penalty. In addition a theoretical bound on the number of selected false positives under stability selection has also been developed [11]. While this rests on an assumption that might not strictly apply in the context of pathway and SNP selection, it would be interesting to explore this further.

Acknowledgment

Matt Silver is supported by a grant from the Wellcome Trust. We wish to thank the Alzheimer's Disease Neuroimaging Initiative (ADNI), who provided all genotype data.

References

[1] R. Manolio et al. "Finding the missing heritability of complex diseases," Nature, vol. 461, no. 7265, pp. 747–753, 2009.

[2] V. K. Mootha et al. "PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes." Nature genetics, vol. 34, no. 3, pp. 267–73, July 2003.

[3] [3] R. M. Cantor, K. Lange, and J. S. Sinsheimer, "Prioritizing GWAS Results: A Review of Statistical Methods and Recommendations for Their Application," American Journal of Human Genetics, pp. 6–22, 2010.

[4] A. Subramanian et al. "Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles." Proceedings of the National Academy of Sciences of the United States of America, vol. 102, no. 43, pp. 15 545–50, Oct. 2005.

[5] S. Ma and M. R. Kosorok, "Detection of gene pathways with predictive power for breast cancer prognosis." BMC Bioinformatics, vol. 11, no. 1, p. 1, 2010.

[6] K. Wang, M. Li, and M. Bucan, "Pathway-Based Approaches for Analysis of Genomewide Association Studies." American Journal of Human Genetics, vol. 81, no. 6, pp. 1278–1283, Oct. 2007.

[7] T. T. Wu, Y. F. Chen, T. Hastie, E. Sobel, and K. Lange, "Genome-wide association analysis by lasso penalized logistic regression." Bioinformatics, vol. 25, no. 6, pp. 714–21, 2009.

[8] H. Zhou, M. E. Sehl, J. S. Sinsheimer, and K. Lange, "Association Screening of Common and Rare Genetic Variants by Penalized Regression." Bioinformatics, vol. 26, no. 19, pp. 2375–2382, 2010.

[9] L. S. Chen et al., "Insights into Colon Cancer Etiology via a Regularized Approach to Gene Set Analysis of GWAS Data," Am J of Human Genetics, vol. 86, no. 6, pp. 860–871, 2010.

[10] L. Jacob, G. Obozinski, and J. Vert, "Group Lasso with Overlap and Graph Lasso," Proceedings of the 26th International Conference on Machine Learning, 2009.

[11] N. Meinshausen and P. Bu hlmann, "Stability selection," Journal of the Royal Statistical Society: Series B, vol. 72, no. 4, pp. 417–473, July 2010.

[12] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," J Royal Statistical Society: Series B, vol. 68, no. 1, pp. 49–67, Feb. 2006.

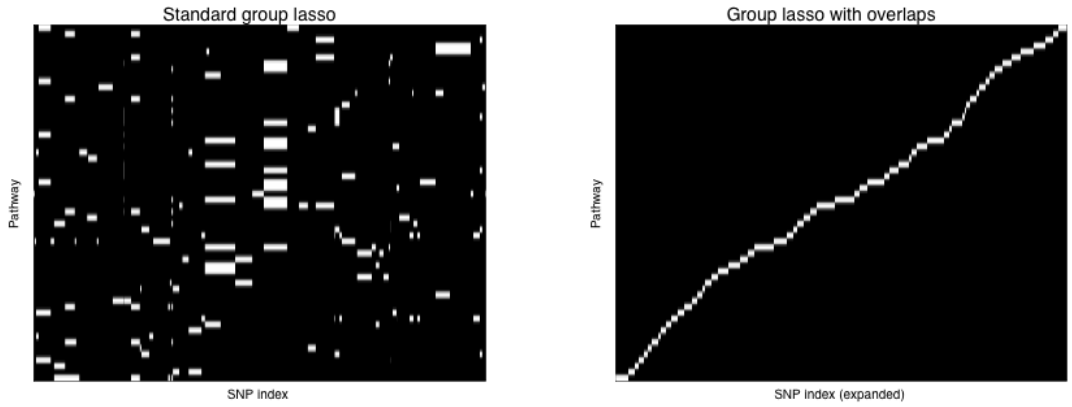
[13] J. Friedman, T. Hastie, H. Hfling, and R. Tibshirani, "Pathwise coordinate optimization." Annals of Applied Statistics, vol. 1, no. 2 pp. 302–332, 2007

[14] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," J Royal Statistical Society: Series B, vol. 67, no. 2, pp. 301–320, Apr. 2005.

[15] Z. Daye and X. Jeng, "Shrinkage and model selection with correlated variables via weighted fusion," Computational Statistics & Data Analysis, vol. 53, no. 4, pp. 1284–1298, Feb. 2009.

[16] Z. Liu and S. Lin, "Multilocus LD measure and tagging SNP selection with generalized mutual information." Genetic epidemiology, vol. 29, no. 4, pp. 353–64, Dec. 2005

[17] P. Tseng and S. Yun, "A coordinate gradient descent method for nonsmooth separable minimization." Mathematical Programming, vol. 117, no. 1-2, pp. 387-423, Aug 2007.



(a) (b)

Figure 1. Assignment of SNPs to pathways with a real dataset. (a) Standard group lasso. Adjacent SNPs cluster into genes within pathways (x-axis), with many genes overlapping across multiple pathways (y-axis). (b) Group lasso with overlaps. The 632 SNPs represented along the x-axis in (a) are expanded to 1877 SNPs each having a unique index, so that no overlaps occur.

TABLE I. PATHWAY SELECTION SENSITIVITY FOR TWO SIGNAL TO NOISE RATIOS

	SNR = 8			**SNR = 2**		
nfp[1] (fpr[2])	***Group lasso***	***GenGen[3]***	***Sensitivity ratio[4]***	***Group lasso***	***GenGen***	***Sensitivity ratio***
0 (0.000)	0.51	0.23	2.20	0.52	0.22	2.35
1 (0.017)	0.79	0.37	2.12	0.66	0.39	1.69
2 (0.033)	0.86	0.56	1.55	0.72	0.53	1.37
3 (0.050)	0.91	0.68	1.33	0.79	0.63	1.24
4 (0.067)	0.94	0.75	1.25	0.82	0.70	1.17
5 (0.083)	0.95	0.80	1.19	0.84	0.77	1.09
6 (0.100)	0.96	0.87	1.11	0.86	0.81	1.07

[1]Number of false positives 2False positive rate 3Ranked by normalized enrichment score 4Ratio of group lasso and GenGen sensitivities