The graph-guided group lasso for genome-wide association studies

Zi Wang
Mathematics Department, Imperial College
180 Queen's Gate, London SW7 2AZ
zi.wang11@imperial.ac.uk

Giovanni Montana
Mathematics Department, Imperial College
180 Queen's Gate, London SW7 2AZ
g.montana@imperial.ac.uk

Abstract: In this work we propose a penalised regression model in which the covariates are known to be clustered into groups, and the clusters are arranged as nodes in a graph. We are motivated by an application to genome-wide association studies in which the objective is to identify important predictors, single nucleotide polymorphisms (SNPs), that account for the variability of a quantitative trait. In this applications, SNPs naturally cluster into SNP sets representing genes, and genes are treated as nodes of a biological network encoding the functional relatedness of genes. Our proposed graph-guided group lasso (GGGL) takes into account such prior knowledge available on the covariates at two different levels, and allows to select important SNPs sets while also favouring the selection of functionally related genes. We describe a computationally efficient algorithm for parameter estimation, provide experimental results and present a GWA study on lipids levels in two Asian populations.

Keywords: sparse group lasso, Laplacian penalty, genome-wide association studies

1 Introduction

Genome-wide association studies (GWAs) are concerned with the search of common genetic variants across the human genome that are associated to a disease status or quantitative trait. The genetic markers are often taken to be single-nucleotide polymorphisms (SNPs), and are treated as covariates in a linear regression model in which the response is a continuous measurement. We let X be the $n \times p$ design matrix containing n independent samples for which p SNPs have been observed, and y be the n-dimensional vector containing the univariate quantitative traits. We further assume that X and y are column-wise normalized to have sum zero and unit length.

Since the objective is to carry our variable selection, the empirical loss is minimised subject to some constraint conditions placed on the coefficients, in order to regularised the solution and carry out variable selection [1, 2].

One way of improving variable selection accuracy in GWAs is to make use of available prior knowledge about the genetic markers and the functional relationships between genes. Such knowledge typically includes the grouping of SNPs into genes, and it has been observed that selecting groups of SNPs in a single block, rather than individual SNPs in isolation, may increase the power to detect true causative and rare variants (e.g. [3]). However, additional information can also be obtained from publicly available data bases in the form of biological networks encoding pairwise interactions between genes or proteins associated to those genes. Under the assumption that such networks describe true biological processes, there are reasons to be believe that using this additional information to guide the SNP selection process may produce results that are biologically more plausible and easy to interpret as well as increase. Regularised regression models that take into consideration the integrated effects of all SNPs that belong to functionally related genes are also believed to achieve superior performance in terms of detecting the true causative markets [4, 5]. In previous GWA studies, this has been accomplished by using variations of the group lasso [6] and the sparse group lasso [7]. When groups overlap, for instance when a SNP is mapped to more than one single gene, variables selected by the overlapping group lasso [8] are the union of some groups.

To the best of our knowledge, graph structures places on genetic markers or genes are not yet used to drive the variable selection process in GWA studies with quantitative traits. In a typical gene network, two nodes are connected by an edge if the associated genes belong to the same genetic pathway or are deemed to share related functions. In the case of a weighted graph, the weights may be a probability measure of the uncertainty of the link between the genes. In this work we consider the case where prior knowledge is available at two different levels: SNPs are grouped into genes, and a weighted gene network encodes the functional relatedness of the all pairs of genes. We propose a penalised regression model, the graph-guided group lasso,

which selects important SNP groups, while also fusing information between adjacent SNPs groups in the given biological network.

2 Graph-guided group lasso

Suppose that the p available SNPs are grouped into mutually exclusive genes $\{R_1, R_2, ..., R_r\}$. The size of a group R_l is denoted by $|R_l|$. We let X_{R_l} denote the $n \times |R_l|$ matrix where the columns correspond to SNPs in R_l, and $\mathcal{G} = \mathcal{G}(V, E)$ the gene network with vertex set V corresponding to the r genes in $\mathcal{R}$. The weight of the edge $k - l$ is denoted by w_{kl}. For simplicity, we assume that all the weights are non-negative.

The regression coefficients are obtaining by minimising $\|y - X\beta\|_2^2$ plus a penalty term given by

$$2\lambda_1 \sum_{g=1}^{r} \sqrt{|R_g|}\|\beta_{R_g}\|_2 + 2\lambda_2\|\beta\|_1 + \mu \sum_{i \in R_k, j \in R_l, R_k \sim R_l} w_{kl}(\beta_i - \beta_j)^2 \quad (1)$$

where λ_1, λ_2, and μ are non-negative regularization parameters, and $R_k \sim R_l$ if and only if they are connected in the network $\mathcal{G}$. This model has two main features. Firstly, by making use of the Laplacian penalty on the complete bipartite graph (R_i, R_j) for all $R_j \sim R_i$, information is fused from all other genes interacting with R_i in $\mathcal{G}$ so that these functionally related genes are encouraged to be selected in and out of the model altogether. ([4]) Secondly, there is a grouping effect, in the sense that all SNPs within a gene R_i are either selected together or not selected. This feature follows from the properties of the sparse group lasso penalty [7], in which sparsity of genes and SNPs are regularized by λ_1 and λ_2 respectively.

Note the prior knowledge represented by grouping and the graph are at heterogeneous levels, hence how the pairwise relations at genes' level influence variable selection for individual SNPs may have different answers. The proposed penalty has also the effect of smoothing the regression coefficients corresponding to all SNPs that belong to interacting genes. When $\mu \to \infty$, all these coefficients are expected to be equal.

In some cases, a modification of the model above may be preferred. If two genes are directly connected in $\mathcal{G}$, it may be preferred to encourage them to be selected or discarded altogether without smoothing the individual SNP coefficients within a gene. For this reason, we also propose a second version of the GGGL model by replacing the last term in (1) by:

$$\mu \sum_{R_k \sim R_l} w_{kl}(\bar{\beta}_{R_k} - \bar{\beta}_{R_l})^2 \quad (2)$$

where $\bar{\beta}_{R_k}$ denotes the average coefficient for predictors in R_k. We show that, using (2), the interacting genes are indeed encouraged to be selected in or out of the model altogether, nonetheless no smoothing effect is imposed on the coefficients corresponding to the SNPs within a gene. In summary, the penalty (1) is more desirable when the interest is only in selecting genes, regardless of the specific SNP effects that drive the gene selection process, whereas (2) is more appropriate for the detection of localised SNP effects.

In summary, we propose a sparse regression model, graph-driven group lasso, for GWA studies that allows to incorporate prior knowledge at two different levels. We describe a computationally efficient estimation algorithm for both version of the model, which is based on coordinate descent methods. We also carry out extensive power studies using realistically simulated data, and compare the proposed model to the original group lasso [6] and a regression model with a network constrained penalty [4]. Finally, we present a real application to detect genetic effects associated to lipids levels in two Asian cohorts.

References

[1] Tibshirani. Regression shrinkage and selection via the lasso. *J.R.Statist.* Soc.B, 58:267-288. 1996.

[2] Wu *et al.* Genome-wide association analysis by lasso penalized logistic regression. *Bioinformatics* 25(6):714-721. 2009.

[3] Zhou *et al.* Association screening of common and rare genetic variants by penalized regression. *Bioinformatics* 26(19): 2375-2382. 2010.

[4] Li and Li. Network-constrained regularization and variable selection for analysis of genomic data. *Bioinformatics.* Vol. 24 no. 9, pages 1175-1182 2008.

[5] Silver and Montana. Fast identification of biological pathways associated with a quantitative trait using group lasso with overlaps. vol. 11, issue 1, article *Statistical Applications in Genetics and Molecular Biology.* vol. 11, issue 1, article 7. 2012.

[6] Yuan and Lin. Model selection and estimation in regression with grouped variables. *J.R.Statist.* Soc.B, 68(1):49-67, 2006.

[7] Friedman *et al.* A note on the group lasso and a sparse group lasso. *arXiv:*1001.0736. 2010.

[8] Jacob *et al.* Group Lasso with overlap and graph Lasso. *International Conference on Machine Learning (ICML 26)* 2009.