Literature Review

Sahir Rai Bhatnagar

July 15, 2018

It is easy to write more than you truly need, so try to keep it as limited as possible. (1) The problems of high dimension regressions such as overfitting and redundancy of variables. (2) then penalization generally as a solution to (1). You might I suppose mention very briefly the alternative solutions like apriori dimension reduction or forward selection, but I would spend as little time as possible on other methods (acknowledging their existence merely and saying your thesis focuses on penalization). (3) then L1 methods including lasso & group lasso (do you discuss elastic net? if so then you might need to include this too). Here I think you should be quite detailed in terms of the theory, how the penalization parameters are chosen, convergence, etc. (4) Then something about other structured L1 penalizations that have been proposed. There are quite a few examples of penalties built for specific applications, and maybe you could find a few such examples and cite them. Not comprehensively. Then you can point to the sail chapter as a new structured penalty. (5) A brief intro to linear mixed models followed by why naive penalization violates the normality of residuals to motivate your ggmix chapter. I'd stop there

1 The problems of high dimension regressions such as overfitting and redundancy of variables.

Computational approaches to variable selection have become increasingly important with the advent of high-throughput technologies in genomics and brain imaging studies, where the data has become massive, yet where it is believed that the number of truly important variables is small relative to the total number of variables.

Predicting a phenotype and understanding which variables improve that prediction are two very challenging and overlapping problems in analysis of high-dimensional data such as those arising from genomic and brain imaging studies. It is often believed that the number of truly important predictors is small relative to the total number of variables, making computational approaches to variable selection and dimension reduction extremely important.

Genome-wide association studies (GWAS) have become the standard method for analyzing genetic datasets owing to their success in identifying thousands of genetic variants associated with complex diseases (https://www.genome.gov/gwastudies/). Despite these impressive

findings, the discovered markers have only been able to explain a small proportion of the phenotypic variance known as the missing heritability problem (**?**). One plausible explanation is that there are many causal variants that each explain a small amount of variation with small effect sizes (**?**). Methods such GWAS, which test each variant or single nucleotide polymorphism (SNP) independently, are likely to miss these true associations due to the stringent significance thresholds required to reduce the number of false positives (**?**).

1.1 Current methods overview and their limitations

1.1.1 Single-marker or single-variable tests

Many of the methods used in the studies mentioned in Sections **??**, **??** and **??** are limited to marginal regression models, i.e., looking at one locus at a time and then subsequently applying a multiple testing adjustment. Although this approach is simple and easy to implement, the single-marker test approach has several limitations. First, the system level changes which are believed to be initiated by the environment, will induce very strong correlations between the genes (e.g. Figure **??**). However, most statistical methods for performing multiple testing adjustments assume weak dependence among the variables being tested (**?**). Dependence among multiple tests can lead to incorrect Type 1 error rates (**?**) and highly variable significance measures (**?**). It is even more difficult to declare significant true-positive interaction terms, i.e., interaction with the environmental factor, because the environmental variable is common to all models being tested. Second, the single marker gene environment interaction model does not allow for modeling the joint effect of many genes which is biologically more plausible given that whole networks are more likely to be associated with disease than just a single gene. Third, even in the presence of weakly dependent variables, adjusting for multiple tests in whole genome studies can result in low power.

1.1.2 Multivariate regression approaches including penalization methods

For n observations and p covariates, consider the multiple linear regression model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where $\mathbf{Y}$ is a $n \times 1$ phenotype vector, $\mathbf{X}$ is an $n \times p$ design matrix, $\boldsymbol{\beta}$ is the $p \times 1$ coefficient vector and $\boldsymbol{\varepsilon}$ is the $n \times 1$ error vector. The least squares estimate is given by $\widehat{\boldsymbol{\beta}} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{Y}$. In genomics data, the problem is that $\mathbf{X}^T\mathbf{X}$ is singular because the number of covariates greatly exceeds the number of subjects. For example DNA microarrays measure the expression of approximately 20,000 genes. However, due to funding constraints, the sample size is often less than a few hundred. A common solution to this problem is through penalized regression, i.e., apply a constraint on the values of $\boldsymbol{\beta}$. The problem can be formulated as finding the vector $\boldsymbol{\beta}$ that minimizes the penalized sum of squares:

$$\underbrace{\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j\right)^2}_{\text{goodness of fit}} + \underbrace{\sum_{j=1}^{p} p(\beta_j; \lambda, \gamma)}_{\text{penalty}} \tag{1}$$

The first term of (1) is the squared loss of the data and can be generalized to any loss function while the second term is a penalty which depends on non-negative tuning parameters γ and λ that control the amount of shrinkage to be applied to $\boldsymbol{\beta}$ and the degree of concavity of the penalty function, respectively. Several penalty terms have been developed in the literature. Ridge regression places a bound on the square of the coefficients (ℓ_2 penalty) (**?**) which has the effect of shrinking the magnitude of the coefficients. This however does not produce parsimonious models as none of the coefficients can be shrunk to exactly 0. The Lasso (**?**) overcomes this problem by placing a bound on the sum of the absolute values of the coefficients (ℓ_1 penalty) which sets some of them to 0, thereby simultaneously performing model selection. The Lasso, along with other forms of penalization (e.g. SCAD **?**, Fused Lasso (**?**), Adaptive Lasso (**?**), Relaxed Lasso (**?**), MCP (**?**)) have proven successful in many practical problems. Despite these encouraging results, such methods have low sensitivity in the presence of high empirical correlations between covariates because only one variable tends to be selected from the group of correlated or nearly linearly dependent variables (**?**). As a consequence, there is rarely consistency on which variable is chosen from one dataset to another (e.g. in cross-validation folds). This behavior is not well suited to genomic data in which large sets of predictors are highly correlated (e.g. a regulatory module) and are also associated with the response. The elastic net was proposed to benefit from the strengths of ridge regression's treatment of correlated variables and lasso's sparsity (**?**). By placing both an ℓ_1 and ℓ_2 penalty on $\boldsymbol{\beta}$, the elastic net achieves model parsimony while yielding similar regression coefficients for correlated variables. These methods however do not take advantage of the grouping structure of the data. For example, cortical thickness measurements from magnetic resonance imaging (MRI) scans are often grouped into cortical regions of the Automated Anatomical Labelling (AAL) atlas (**?**). Genes involved in the same cellular process (e.g. KEGG pathway (**?**)) can also be placed into biologically meaningful groups. When regularizing with the ℓ_1 penalty, each variable is selected individually regardless of its position in the design matrix. Existing structures between the variables (e.g. spatial, networks, pathways) are ignored even though in many real-life applications the estimation can benefit from this prior knowledge in terms of both prediction accuracy and interpretability (**?**). The group lasso (**?**) (and generalizations thereof) overcomes this problem by producing a structured sparsity (**?**), i.e., given a predetermined grouping of non-overlapping variables, all members of the group are either zero or non-zero. The main drawback when applying these methods to genomic data is that these groups may not be known *a priori*. Known pathways may not be relevant to the response of interest and the study of inferring gene networks is still in its infancy.

1.1.3 Clustering together with regression

Due to the unknown grouping problem, several authors have suggested a two-step procedure where they first cluster or group variables in the design matrix and then subsequently proceed to model fitting where the feature space is some summary measure of each group. This idea dates back to 1957 when Kendall (**?**) first proposed using principal components in regression. Hierarchical clustering based on the correlation of the design matrix has also been used to create groups of genes in microarray studies and for each level of hierarchy, the cluster

average was used as the new set of potential predictors in forward-backward selection (**?**) or the lasso (**?**). Bühlmann *et al.* (**?**) proposed a bottom-up agglomerative clustering algorithm based on canonical correlations and used the group lasso on the derived clusters. There are several advantages to these methods over the ones previously mentioned in Sections 1.1.1 and 1.1.2. First, the results are more interpretable than the traditional lasso (and related methods) because the non-zero components of the prediction model represent sets of genes as opposed to individual ones. Second, by using genes which cluster well, we bias the inputs towards correlated sets of genes which are more likely to have similar function. Third, taking a summary measure of the resulting clusters can reduce the variance in prediction (overfitting) due to the compressed dimension of the feature space. Lastly, from a practical point of view this approach is flexible and easy to implement because efficient algorithms exist for both clustering (**?**) and model fitting (**??**). A limitation of these approaches is that the clustering is done in an unsupervised manner, i.e., the clusters do not use the response information. This has the effect of assigning similar coefficient values to correlated features. Witten *et al.* (**?**) recently proposed a method which encourages features that share an association with the response to take on similar coefficient values. This is useful in situations where only a fraction of the features in a cluster are associated with the response.

In the context of studying gene environment interactions, all previously mentioned methods can yield results where the main effects are 0 but the cross terms are not. While there may exist situations where this can occur, such behavior is generally not biologically plausible in genomics (**?**). Two other arguments against such behavior have to do with statistical power and practical importance; 1) large main effects are more likely to lead to detectable interactions than small ones (**?**) and 2) a data collector cares about the number of variables they need to *measure* to make predictions at a future time (**?**). Moreover, the proposed clustering methods are based on features from all observations. Any correlation patterns specific to a subgroup of patients (e.g. Figures **??** and **??**) may become diluted when looking at the overall correlation matrix. Therefore the focus of my doctoral research will be on developing a method that can identify previously unknown sets of highly correlated genes that interact with the environment to explain phenotypic variation, with a focus on prediction accuracy and model interpretability.

2 penalization generally as a solution to (1)

You might I suppose mention very briefly the alternative solutions like apriori dimension reduction or forward selection, but I would spend as little time as possible on other methods (acknowledging their existence merely and saying your thesis focuses on penalization).

3 then L1 methods including lasso & group lasso (do you discuss elastic net? if so then you might need to include this too)

Here I think you should be quite detailed in terms of the theory, how the penalization parameters are chosen, convergence, etc.

4 Then something about other structured L1 penalizations that have been proposed

There are quite a few examples of penalties built for specific applications, and maybe you could find a few such examples and cite them. Not comprehensively. Then you can point to the sail chapter as a new structured penalty.

5 (5) A brief intro to linear mixed models followed by why naive penalization violates the normality of residuals to motivate your ggmix chapter.

6 you will need a section on things like gradient descent and other algorithms for finding solutions efficiently