CHAPTER

Sparse models for imaging genetics

5

J. Wang[1], T. Yang[2], P. Thompson[3], J. Ye[1]

University of Michigan, Ann Arbor, MI, United States[1] Arizona State University, Tempe, AZ, United States[2] University of Southern California, Los Angeles, CA, United States[3]

CHAPTER OUTLINE

5.1 INTRODUCTION

Imaging genetics studies neuroimaging-related genetic variation. In the past decade, neuroimaging techniques—for example, computed tomography (CT), magnetic resonance imaging (MRI), functional MRI (fMRI), and positron emission tomography (PET)—provide both anatomical and functional visualizations of the nervous system, which greatly advance modern medicine, neuroscience, and psychology. As an emerging promising technique, imaging genetics research has attracted extensive attention. With the integration of molecular genetics and disorder-related neuroimaging phenotypes, imaging genetics provides a unique opportunity to reveal the impact

Machine Learning and Medical Imaging. http://dx.doi.org/10.1016/B978-0-12-804076-8.00005-0

of genetic variation in neuroimaging, that is, how individual differences in single nucleotide polymorphisms (SNPs) affect brain development, structure, and function (Hariri et al., 2006; Thompson et al., 2013). Molecular geneticists believe that some common genetic variants in SNPs may lead to common disorders (Cirulli and Goldstein, 2010). Moreover, as another benefit of exploiting neuroimaging in genetics, imaging phenotypes are closer to the biology of genetic function (Meyer-Lindenberg, 2012) than disease or cognitive phenotypes.

Previous studies show the great promise of imaging genetics. For example, the $\epsilon 4$ allele of apolipoprotein E (ApoE4) is one of the well-known genetic risk factors for Alzheimer's disease (AD). From a neuroimaging perspective, the degeneration of brain tissue of ApoE4 carriers is faster as they age; young adult ApoE4 carriers often exhibit thinner cortical gray matter than noncarriers (Shaw et al., 2007). It has been verified in a series of genome-wide association (GWA) studies of AD that ApoE4 is strongly associated with the volumes of key brain regions, such as the hippocampus and entorhinal cortex (Potkin et al., 2009; Stein et al., 2012; Yang et al., 2015). Recent worldwide consortium efforts, such as ENIGMA (Enhancing Neuroimaging Genetics through Meta-Analysis (Stein et al., 2012)) and CHARGE (Cohorts for Heart and Aging Research in Genomic Epidemiology (Bis et al., 2012; Psaty et al., 2009)), enable us to detect robust common neuroimaging-genetic associations (Medland et al., 2014).

Imaging genetic studies are challenging in practice due to the relatively small number of subjects and extremely high dimensionality of imaging as well as genetic data. Neuroimaging data, for example, contains hundreds of thousands of voxels. Advances in modern sequencing techniques lead to huge scale (whole) genome sequencing data with tens of millions of SNPs. However, most traditional statistical methods are intended for low-dimensional data sets (James et al., 2013), in which the number of subjects is much larger than the number of features. This significantly limits the practical usage of traditional methods to the high-dimensional imaging data sets, as they are prone to overfitting.

The high-dimensional data sets involved in many imaging genetics studies confront researchers and scientists with an urgent need for novel methods that can effectively uncover the predictive patterns from these types of data. A useful observation from many real-world applications is that data with complex structures often has sparse underlying representations. More specifically, although the data may have millions of features, it may be well interpreted by a few of the most relevant explanatory features. For example, the neural representation of natural scenes in the visual cortex is sparse, as only a small number of neurons are active at a given instant (Vinje and Gallant, 2000); images have very sparse representations with respect to an overcomplete dictionary because they lie on or close to low-dimensional subspaces or submanifolds (Wright et al., 2010); although humans have millions of SNPs, only a small number of them are relevant to certain diseases such as leukemia and Alzheimer's disease (Golub et al., 1999; Guyon et al., 2002; Mu and Gage, 2011). Moreover, sparsity has been shown to be an effective approach to alleviate overfitting, from which most traditional statistical methods suffer. Therefore finding sparse

representations is particularly important in discovering the underlying mechanisms of many complex systems.

As an emerging and powerful technique, sparse models have attracted increasing research interest in image genetics in the past decade. As well as their robustness to overfitting, sparse models are also promising in enhancing the interpretability of the model by automatically identifying a small subset of features that can best explain the outcome. Indeed, we can categorize existing methodological approaches for imaging genetics into three classes (Thompson et al., 2013).

The first one is the so-called *univariate-imaging univariate-genetic association* analysis that performs a univariate statistical test on each SNP-voxel pair individually. This type of approach has been widely used in previous GWA studies. However, these approaches fail to reveal scenarios such as SNP-SNP interactions and the joint effects of multiple SNPs, which occur commonly in gene expression (Dinu et al., 2012; Cornelis et al., 2009; Singh et al., 2011; Yang et al., 2012). In addition, it is worth mentioning that this kind of analysis is computationally inefficient.

The second class is the *univariate-imaging multivariate-genetic association* method. Based on a candidate imaging phenotype, a common multivariate approach utilizes sparse models, for example, Lasso (least absolute shrinkage and selection operator (Tibshirani, 1996; Yang et al., 2015)), to perform simultaneous model fitting and variable (causal SNPs) selection. Moreover, by incorporating biological prior knowledge such as linkage disequilibrium (LD) information, we can employ group Lasso to locate groups of candidate SNPs (Wang et al., 2012; Yuan and Lin, 2006). In the sequel, tree-structured group Lasso can also be applied if the hierarchical structure of SNPs is further available (Liu and Ye, 2010).

The third class of methodology in imaging genetics is *joint multivariate association* analysis, for example, canonical correlation analysis (CCA) and partial least squares (PLS) regression. However, a clear drawback of this kind of approach is that the detected genetic variants and imaging features may not be immediately related to a disorder (Batmanghelich et al., 2013).

In this chapter, we focus on univariate-imaging multivariate-genetic association studies in imaging genetics. We first introduce two simple sparse models, that is, Lasso and sparse logistic regression, in Section 5.2. Then, in Section 5.3, we introduce a series of popular structured sparse methods, which incorporate some prior knowledge. We will also review some popular optimization algorithms in Section 5.4. In Section 5.5, we pay particular attention to a suite of novel techniques, that is, screening rules, for sparse models (Hastie et al., 2015; Wang et al., 2015b), which can improve the computational efficiency by several orders of magnitude.

5.2 BASIC SPARSE MODELS

To illustrate the basic idea of sparse models, in this section we introduce two simple but widely used sparse models: Lasso (Tibshirani, 1996) that is for regression and

sparse logistic regression (Sun et al., 2009; Wu et al., 2009; Zhu and Hastie, 2004) that is for classification.

Suppose that the training samples contain N observations with p features. We denote the outcome by a vector $\mathbf{y} \in \mathbb{R}^p$ and the feature matrix by $\mathbf{X} \in \mathbb{R}^{N\times p}$. By convention, each row $\mathbf{x}^i \in \mathbb{R}^p$, $i = 1, \ldots, N$, of $\mathbf{X}$ represents a data sample and each column $\mathbf{x}_j \in \mathbb{R}^N$, $j = 1, \ldots, p$, of $\mathbf{X}$ represents a feature. In this chapter, we mainly focus on linear models $h : \mathbb{R}^p \to \mathbb{R}$ with

$$h(\mathbf{x}) = \beta^{\mathrm{T}}\mathbf{x}, \tag{5.1}$$

where $\beta \in \mathbb{R}^p$ is the coefficient vector that needs to be estimated.

Many traditional regression and classification methods like least squares and logistic regression are developed for low-dimensional data sets (James et al., 2013), in which the number of observations N is much larger than the number of features. However, as new technologies have advanced in the past two decades, we are frequently confronted with extremely high-dimensional data sets (like fMRI, PET, and GWAS), in which the number of features p is much greater than the number of samples N. Directly applying traditional regression or classification methods to the high-dimensional data sets may be inappropriate. Take the least squares regression as an example. When $p \gg N$, we can find a regression hyperplane that fits the data exactly (the training error is zero). In many applications, a perfect fit on the training data usually implies overfitting, which may lead to poor performance on the testing data.

Regularization has been shown to be a promising approach to alleviate overfitting. Many sparse models estimate the coefficient vector β by incorporating various sparse-inducing regularizers:

$$\min_{\beta} f(\beta) = \ell(\beta) + \lambda\Omega(\beta), \tag{5.2}$$

where $\ell(\beta)$ is a loss function measuring the fitness of the model on the training data, $\Omega(\beta)$ is the regularizer penalizing the complexity of the model, and $\lambda > 0$ is a regularization parameter controlling the trade-off between the loss $\ell(\cdot)$ and the penalty $\Omega(\cdot)$. We note that the sparse-inducing penalty $\Omega(\cdot)$ is a typically nonsmooth function of the coefficient vector.

Lasso is a widely used regression technique to find sparse representations of a given signal with respect to a set of basis vectors. Standard Lasso employs least squares loss and $\Omega(\beta) = \|\beta\|_1$ as its regularizer, that is,

$$\min_{\beta} \frac{1}{2}\|\mathbf{y} - \mathbf{X}\beta\|^2 + \lambda\|\beta\|_1. \tag{5.3}$$

Due to the ℓ_1-norm penalty, many components of the solution vector of Lasso are zero when the value of λ is large. The features corresponding to these nonzero components are considered to be important to explain the outcome. Therefore, in a wide range of real applications, Lasso serves as an effective feature selection method and has

achieved great success (Bruckstein et al., 2009; Chen et al., 2001; Candès, 2006; Wright et al., 2010; Zhao and Yu, 2006).

Similar to Lasso, sparse logistic regression also employ the ℓ_1-norm regularization, while utilizing the logistic loss. Specifically, sparse logistic regression takes the form:

$$\min_{\beta} \sum_{i=1}^{N} \log\left(\frac{1}{1+\mathrm{e}^{-y_i(\beta^{\mathrm{T}}\mathbf{x}^i)}}\right) + \lambda\|\beta\|_1. \tag{5.4}$$

Sparse logistic regression has received much attention in the last few years and the interest in it is growing (Sun et al., 2009; Wu et al., 2009; Zhu and Hastie, 2004) due to the increasing prevalence of high-dimensional data. The popularity of sparse logistic regression is also due to the fact that it can simultaneously achieve the goals of classification and feature selection.

5.3 STRUCTURED SPARSE MODELS

A major drawback of Lasso and sparse logistic regression is that they do not take the feature structure into account. In other words, the sparse representation obtained by Lasso or sparse logistic regression remains the same if we shuffle the features. However, in many real applications, this is undesirable, as the features frequently exhibit certain intrinsic structures, for example, trees, graphs, spatial or temporal smoothness, and disjoint/overlapping groups.

In this section, we introduce several structured sparse models, which incorporate different prior knowledge of feature structures by carefully designed sparse-inducing regularizers.

5.3.1 GROUP LASSO AND SPARSE GROUP LASSO

In many applications, the features form groups or clusters. For example, features with discrete values are usually transformed into groups of dummy variables; in the study of Alzheimer's disease, we divide the voxels of the PET images into a set of nonoverlapping groups according to the brain regions. To select groups of features, Yuan and Lin (2006) proposed the nonoverlapping group Lasso, in which the groups do not share features. Assume that the features are partitioned into k disjoint groups $\{G_1, \ldots, G_k\}$, where G_i contains the indices of features belonging to the ith group. The regularizer of group Lasso takes the form:

$$\Omega_{\mathrm{gL}}(\beta) = \sum_{i=1}^{k} w_i \|\beta_{G_i}\|_q, \tag{5.5}$$

where w_i is the weight for the ith group, $\mathbf{X}_{G_i}$ is the submatrix whose columns consist of the features belonging to the ith group, and $\|\cdot\|_q$ with $q > 1$ is the

ℓ_q-norm (the value of q is usually set to be 2 or ∞) (Wang et al., 2013). Group Lasso has been widely used in applications with group structure available, for example, regression (Kowalski, 2009; Negahban and Wainwright, 2008; Yuan and Lin, 2006), classification (Meier et al., 2008), joint covariate selection for group selection (Obozinski et al., 2007), and multitask learning (Argyriou et al., 2008; Liu et al., 2009; Quattoni et al., 2009).

Group Lasso performs group selection. However, for some applications, it is desirable to identify features within each group that exhibit the strongest effects. To achieve this goal, sparse-group Lasso (SGL) (Friedman et al., 2010; Simon et al., 2013) combines the Lasso (Tibshirani, 1996) and group Lasso (Yuan and Lin, 2006) penalties to identify important groups and features simultaneously. Specifically, the sparse group Lasso penalty can be written as follows:

$$\Omega_{\text{SGL}}(\beta) = \alpha\|\beta\|_1 + (1-\alpha)\sum_{i=1}^{k} w_i\|\beta_{G_i}\|_q, \tag{5.6}$$

where $\alpha \in [0, 1]$ balances the sparsity in the feature level and the sparsity in the group level. In recent years, SGL has found great success in many real-world applications, including but not limited to machine learning (Vidyasagar; Yogatama and Smith, 2014), signal processing (Sprechmann et al., 2011), bioinformatics (Peng et al., 2010), etc.

5.3.2 OVERLAPPING GROUP LASSO AND TREE LASSO

Group Lasso assumes that the feature groups are disjoint. However, in certain applications, some features may be shared across different groups. For example, in the study of biologically meaningful gene/proteins, we say that the proteins/genes in the same groups are related in the sense that: (1) the proteins/genes appear in the same pathway; (2) the proteins/genes belong to the same Gene Ontology (GO) term (Ashburner et al., 2000; Harris et al., 2004); (3) the proteins/genes are related from gene set enrichment analysis (GSEA) (Subramanian et al., 2005). As the same gene may be involved in different pathways, it may be shared across different groups. The overlapping group Lasso penalty (Zhao et al., 2009) takes the form:

$$\Omega_{\text{ogL}}(\beta) = \alpha\|\beta\|_1 + (1-\alpha)\sum_{i=1}^{k} w_i\|\beta_{G_i}\|_q, \tag{5.7}$$

where $\alpha \in [0, 1]$, w_i is the nonnegative weight for the ith group, and G_i consists of the feature indices from the ith group. The difference that distinguishes (5.7) and (5.6) is that G_i may overlap with G_j for $i \neq j$.

A particularly interesting special case of overlapping group Lasso is the so-called tree structured group Lasso (tgLasso) (Kim and Xing, 2010; Zhao et al., 2009). In some applications, the data may exhibit hierarchical tree-structured sparse patterns among features. For example, based on the spatial locality (Liu and Ye, 2010), we can

represent an image by a tree whose leaf node corresponds to a single feature (pixel) and whose internal node corresponds to a group of features (pixels). Another interesting application of tgLasso is to identify risk SNPs regarding AD from the GWAS data (Li et al., 2016). It is known that we can measure the association of alleles at different loci by linkage disequilibrium (LD). Thus by taking the LD information and the chromosomal loci of SNPs into account, we can build the tree structure for SNPs. If the tree structure is available, the tree-structured group Lasso penalty takes the form:

$$\Omega_{\text{tgL}}(\beta) = \sum_{i,j} w_j^i \|\beta_{G_j^i}\|_q, \tag{5.8}$$

where G_j^i is the group of features corresponding to the jth node of depth i and w_j^i is the positive weight for G_j^i. We note that every node in the tree is a superset of its descendant nodes. Therefore if the features in a node are absent from the sparse representation, so are the features in all its descendant nodes.

5.3.3 FUSED LASSO AND GRAPH LASSO

In many applications, the data come with spatial or temporal smoothness. For example, in the study of arrayCGH (Tibshirani et al., 2005; Tibshirani and Wang, 2008), the features—the DNA copy numbers along the genome—have the natural spatial order. The fused Lasso penalty encodes the structure of smoothness by penalizing the differences between the adjacent coefficients, that is,

$$\Omega_{\text{fL}}(\beta) = \alpha\|\beta\|_1 + (1-\alpha)\sum_{i=1}^{p-1} |\beta_i - \beta_{i-1}|, \tag{5.9}$$

where $\alpha \in [0, 1]$. We can see that the fused Lasso penalty would lead to solutions in which adjacent components are close or identical to each other.

In certain applications, the data may exhibit a more complex smoothness structure. Specifically, the features may form an undirected graph structure, in which connected features may share some common properties. For example, much biological evidence suggested that genes tend to work in groups if they have similar biological functions (Li and Li, 2008). This prior knowledge can be encoded by a graph, in which each node represents a gene and the edges denote the regulatory relationships between genes. Many recent works have shown that the structure information encoded as a graph can significantly improve the predictive power of the model. Let (V, E) be a given graph, where V denotes the set of nodes and E denotes the edges. By noting that an open chain is a special example of a graph, we can generalize the fused Lasso penalty to a graph Lasso penalty—known as the ℓ_1 graph Lasso—as follows:

$$\Omega_{\text{grL}}^{\ell_1}(\beta) = \alpha\|\beta\|_1 + (1-\alpha)\sum_{(i,j)\in E}|\beta_i - \beta_j|, \tag{5.10}$$

where the second term penalizes the difference between the coefficients of connected features. Thus the coefficients of connected features tend to be close or identical to each other. The graph Lasso penalty in (5.10) entails a significant computational challenge as both terms are nonsmooth and the graph structure may be complicated. An efficient alternative, called Laplacian Lasso (or ℓ_2 graph Lasso), employs the following penalty:

$$\Omega_{\text{grL}}^{\ell_2}(\beta) = \alpha\|\beta\|_1 + (1-\alpha)\beta^{\text{T}}L\beta, \tag{5.11}$$

where L is the graph Laplacian matrix (Belkin and Niyogi, 2003; Chung, 1997). The graph Laplacian matrix is positive semidefinite and well captures the local geometric structure of the data. We note that the Laplacian Lasso penalty reduces to the elastic net penalty (Zou and Hastie, 2005) if the Laplacian matrix is an identity matrix. As the second term of the Laplacian penalty is quadratic, we can incorporate it with the least square loss, and thus many solvers for Lasso are applicable to the Laplacian Lasso.

In view of (5.10) and (5.11), both ℓ_1 graph Lasso and Laplacian Lasso encourage positive correlation of the coefficients of connected features, that is, they tend to have the same sign. However, in certain applications, connected features may be negatively correlated. To deal with this challenge, GFLasso incorporates the sample correlation into the penalty:

$$\Omega_{\text{GFL}}(\beta) = \alpha\|\beta\|_1 + (1-\alpha)\sum_{(i,j)\in E}|\beta_i - \text{sign}(r_{i,j})\beta_j|, \tag{5.12}$$

where $r_{i,j}$ is the sample correlation between the ith and jth features. We can see that GLLasso encourages positive correlation between connected features if $r_{i,j} > 0$ and negative correlation if $r_{i,j} < 0$. However, if the sample correlation is inaccurate, GFLasso may introduce additional bias.

An alternative, which is called graph OSCAR (GOSCAR), avoids the usage of the sample correlation. The GOSCAR penalty can be written as

$$\Omega_{\text{GFL}}(\beta) = \alpha\|\beta\|_1 + (1-\alpha)\sum_{(i,j)\in E}\max\{|\beta_i|, |\beta_j|\}. \tag{5.13}$$

The ℓ_∞ penalty encourages the magnitude of the coefficients of connected features to be close or identical to each other. However, the ℓ_∞ penalty may overpenalize the coefficients, leading to additional bias. This motivates a nonconvex version of graph Lasso penalty:

$$\Omega_{\text{ncFGS}}(\beta) = \alpha\|\beta\|_1 + (1-\alpha)\sum_{(i,j)\in E}||\beta_i| - |\beta_j||, \tag{5.14}$$

which can reduce the bias in many applications compared to the aforementioned convex penalties. Similar to GOSCAR, ncFGS penalty does not make use of the sample correlation either.

5.4 OPTIMIZATION METHODS

Many sparse models in the form of (5.2) are nonsmooth and nondifferentiable, which imposes a serious challenge to the corresponding optimization algorithms. In the past few years, as sparse models have become increasingly popular, extensive research efforts have been devoted to developing efficient solvers for the sparse models. In this section, we briefly review two particularly popular first-order methods: proximal gradient descent and accelerated gradient methods, which are especially useful for large-scale problems.

5.4.1 PROXIMAL GRADIENT DESCENT

In this section, we briefly review the well-known proximal gradient descent algorithm for (5.2). For many sparse models, the loss function $\ell(\cdot)$ is convex and differentiable, and the regularizer $\Omega(\cdot)$ is convex but nondifferentiable. The major challenge in developing optimization algorithms for (5.2) is due to the nondifferentiable regularizer $\Omega(\cdot)$.

The key idea (Beck and Teboulle, 2009; Hastie et al., 2015) of proximal gradient descent is that, in each iteration, we minimize a local approximation of $f(\cdot)$ consisting of the nondifferentiable component $\Omega(\cdot)$ and a linear approximation of the differentiable component $\ell(\cdot)$. Specifically, in the kth iteration, we update β^k by the following generalized gradient update:

$$\beta^{k+1} = \underset{\beta}{\operatorname{argmin}} \left\{ \ell(\beta^k) + \langle \nabla \ell(\beta^k), \beta - \beta^k \rangle + \frac{1}{2t^k} \|\beta - \beta^k\|^2 + \Omega(\beta) \right\}. \tag{5.15}$$

For a convex function h, we can define the proximal map:

$$\mathbf{prox}_h(u) = \underset{\mathbf{v}}{\operatorname{argmin}} \left\{ \frac{1}{2} \|\mathbf{v} - \mathbf{u}\|^2 + h(\mathbf{v}) \right\}. \tag{5.16}$$

Then, it follows that

$$\beta^{k+1} = \mathbf{prox}_{t^k \Omega} \left(\beta^k - t^k \nabla \ell(\beta^k) \right). \tag{5.17}$$

Sufficient conditions (Nesterov, 2007) for the convergence of the update in (5.17) are as follows:

1. The gradient of the differentiable component $\ell(\cdot)$ is Lipschitz continuous, that is, for any $\beta, \beta' \in \mathbb{R}^p$, the following inequality holds:

$$\|\nabla \ell(\beta) - \nabla \ell(\beta')\|_2 \leq L \|\beta - \beta'\|_2. \tag{5.18}$$

2. The step size t^k is a constant that satisfies $t^k \in (0, 1/L]$.

Then, it can be shown that

$$f(\beta^k) - f(\beta^*) \leq \frac{L\|\beta^0 - \beta^*\|^2}{2k}, \tag{5.19}$$

where β^* is an optimal solution. Thus (5.19) implies that the proximal gradient descent in (5.17) leads to a convergence rate of $O(1/k)$.

5.4.2 ACCELERATED GRADIENT METHOD

When the proximal mapping in (5.17) can be computed efficiently, the proximal gradient descent approach is a very popular tool in solving the corresponding sparse models, especially for large-scale problems. However, the convergence can be slow for certain objective functions, as the update by proximal gradient descent may lead to an undesirable type of zig-zagging behavior from step to step (Hastie et al., 2015). To improve the convergence property, Nesterov (Nesterov, 1983, 2007) proposed a class of accelerated gradient methods with a convergence rate $O(1/k^2)$. We summarize the accelerated gradient method in Algorithm 5.1.

ALGORITHM 5.1 ACCELERATED GRADIENT METHOD

Input: A constant $t \in (0, 1/L]$, where L is a Lipschitz constant of $\nabla \ell$.

1: Set $\beta^0 = \theta^1 \in \mathbb{R}^p$, $s^1 = 1$, and $k = 1$.
2: **while** termination condition is not satisfied **do**
3: $\quad \beta^k = \mathbf{prox}_{t\Omega}(\theta^k - t\nabla \ell(\theta^k))$,
4: $\quad s^{k+1} = \frac{1+\sqrt{1+4(s^k)^2}}{2}$,
5: $\quad \theta^{k+1} = \beta^k + \left(\frac{s^k - 1}{s^{k+1}}\right)(\beta^k - \beta^{k-1})$,
6: $\quad k = k + 1$.
7: **end while**

Let β^k be generated by Algorithm 5.1. Then, it is shown that

$$f(\beta^k) - f(\beta^*) \leq \frac{2L\|\beta^0 - \beta^*\|^2}{(k+1)^2}, \tag{5.20}$$

where β^* is an optimum.

We note that, besides the convergence rates, a key difference—that distinguishes the accelerated gradient method from proximal gradient descent—is that the function values computed by the former may be increasing, that is, $f(\beta^{k+1})$ may be larger than $f(\beta^k)$, while they keep decreasing for the latter.

5.5 SCREENING

In the past few years, many algorithms have been proposed to efficiently solve the sparse models. However, when the feature dimension is extremely large, the applications of sparse models to large-scale problems remain challenging due to their nondifferentiable and complicated regularizers.

In the past few years, the idea of screening (El Ghaoui et al., 2012; Tibshirani et al., 2012; Wang et al., 2015b) has been found to be a very promising approach to improve the efficiency of sparse models. Essentially, screening aims to quickly identify the zero coefficients in the sparse solutions by simple testing rules. Then, we can remove the corresponding features from the optimization without sacrificing accuracy. Thus the size of the data matrix can be significantly reduced, leading to substantial savings in computational cost and memory usage. In many applications, the speedup gained by screening methods can be several orders of magnitude.

In this section, we focus on the screening method for Lasso. We also briefly review some screening methods for other, more complicated sparse models.

5.5.1 SCREENING FOR LASSO

We can roughly divide existing screening methods of Lasso into two categories: the heuristic screening methods and the safe screening methods.

As implied by the name, the heuristic screening methods may mistakenly discard features that have nonzero coefficients in the sparse representations. This type of method includes SIS (Fan et al., 2008) and strong rule (Tibshirani et al., 2012). SIS removes features based on the correlation between features and the outcome, but not from the perspective of optimization. Strong rule assumes that the inner products between features and the residue are nonexpansive (Bauschke and Combettes, 2011) with respect to the parameter values. However, this assumption may not hold in real applications. Thus strong rule needs to postprocess the results by KKT conditions to check if it makes mistakes.

In contrast to the heuristic screening methods, the safe screening methods can guarantee that the coefficients of the discarded features are zero in the solution vector. Existing safe screening methods include SAFE (El Ghaoui et al., 2012), DOME (Xiang and Ramadge, 2012), and EDPP (Wang et al., 2015b), which are inspired by the KKT conditions.

We note that, although heuristic in theory, strong rule seldom makes mistakes in practice and it significantly outperforms many safe screening methods like SAFE and DOME. Therefore, in this section, we focus on the EDPP screening rule, whose performance is comparable to or even better than strong rule. For details of EDPP (see Wang et al., 2015b).

5.5.1.1 Background

Recall that the Lasso problem is given by (5.3). It is known that the dual problem of Lasso is equivalent to

$$\inf_{\theta} \quad \left\{ \frac{1}{2} \left\| \theta - \frac{\mathbf{y}}{\lambda} \right\|_2^2 : \ |\mathbf{x}_i^T \theta| \leq 1, \ i = 1, 2, \ldots, p \right\}. \tag{5.21}$$

Let $\beta^*(\lambda)$ and $\theta^*(\lambda)$ be the optimal solutions of problems (5.3) and (5.21) respectively, and $\mathcal{F}$ be the feasible set of problems (5.21). For notational convenience, we define the projection operator by

$$\mathbf{P}_C(\mathbf{w}) = \underset{\mathbf{u} \in C}{\operatorname{argmin}} \|\mathbf{u} - \mathbf{w}\|_2, \tag{5.22}$$

where C is a closed and convex set. We can see that $\theta^*(\lambda)$ is the projection of $\mathbf{y}/\lambda$ onto $\mathcal{F}$, that is,

$$\theta^*(\lambda) = \mathbf{P}_{\mathcal{F}} \left(\frac{\mathbf{y}}{\lambda} \right). \tag{5.23}$$

Moreover, the primal optimum and dual optimum are related by the KKT conditions:

$$\mathbf{y} = \mathbf{X}\beta^*(\lambda) + \lambda\theta^*(\lambda), \tag{5.24}$$

$$(\theta^*(\lambda))^{\mathrm{T}} \mathbf{x}_i \in \begin{cases} \operatorname{sign}([\beta^*(\lambda)]_i), & \text{if } [\beta^*(\lambda)]_i \neq 0, \\ [-1, 1], & \text{if } [\beta^*(\lambda)]_i = 0, \end{cases} \tag{5.25}$$

where $[\cdot]_k$ denotes the kth component.

Inspired by the KKT condition in Eq. (5.25), we can see that

$$|(\theta^*(\lambda))^{\mathrm{T}} \mathbf{x}_i| < 1 \Rightarrow [\beta^*(\lambda)]_i = 0, \text{ie}, \mathbf{x}_i \text{ is an inactive feature.} \tag{R1}$$

Thus we can potentially utilize (R1) to identify the inactive features for the Lasso problem. However, by a closer look at (R1), we can see that (R1) is not applicable to identify the inactive features, as it involves $\theta^*(\lambda)$. Inspired by the SAFE rules (El Ghaoui et al., 2012), we can relax (R1) as follows:

$$\sup_{\theta \in \Theta} |\mathbf{x}_i^{\mathrm{T}} \theta| < 1 \Rightarrow [\beta^*(\lambda)]_i = 0, \text{ie}, \mathbf{x}_i \text{ is an inactive feature,} \tag{R1'}$$

where Θ is a set that contains $\theta^*(\lambda)$.

Thus, without the knowledge of $\theta^*(\lambda)$, (R1') implies that an estimation of the dual optimum is sufficient to develop an applicable screening rule for the Lasso problem. Nevertheless, in view of (R1) and (R1'), we can see that a small region Θ implies an accurate estimation of $\theta^*(\lambda)$, and thus a more aggressive screening rule for identifying the inactive features.

A useful consequence of (R1) is that we can find the smallest value of λ such that $\beta^*(\lambda) = 0$. Indeed, we have (Wang et al., 2015b)

$$\lambda \geq \lambda_{\max} = \|\mathbf{X}^{\mathrm{T}} \mathbf{y}\|_\infty \Leftrightarrow \beta^*(\lambda) = 0. \tag{5.26}$$

5.5.1.2 Enhanced DPP (EDPP) screening rules

Following (R1'), the framework of the EDPP screening rule (Wang et al., 2015b) for Lasso can be divided into the following three steps:

1. We first estimate a region Θ which contains the dual optimum $\theta^*(\lambda)$.
2. We solve the maximization problem in (R1'), that is, $\sup_{\theta\in\Theta} |\mathbf{x}_i^{\mathrm{T}}\theta|$.
3. By plugging in the upper bound we find, in the last step, that it is straightforward to develop the screening rule based on (R1').

The key step of EDPP is the estimation of the dual optimum, which determines the performance of the screening rule. Based on the geometric properties of the dual problem, EDPP provides a very accurate estimation of the dual optimum.

The first geometric property that EDPP utilizes is the so-called firmly nonexpansiveness of the projection operators.

Theorem 5.1 (Bauschke and Combettes, 2011)**.** *Let C be a nonempty closed convex subset of a Hilbert space $\mathcal{H}$. Then the projection operator defined in Eq. (5.22) is continuous and firmly nonexpansive. In other words, for any $\mathbf{w}_1, \mathbf{w}_2 \in \mathcal{H}$, we have*

$$\|\mathbf{P}_C(\mathbf{w}_1) - \mathbf{P}_C(\mathbf{w}_2)\|_2^2 + \|(Id - \mathbf{P}_C)(\mathbf{w}_1) - (Id - \mathbf{P}_C)(\mathbf{w}_2)\|_2^2 \leq \|\mathbf{w}_1 - \mathbf{w}_2\|_2^2,$$

where Id is the identity operator.

Another useful geometric property of the projection operators is related to the projection of rays.

Lemma 5.1 (Bauschke and Combettes, 2011)**.** *Let C be a nonempty closed convex subset of a Hilbert space $\mathcal{H}$. For a point $\mathbf{w} \in \mathcal{H}$, let $\mathbf{w}(t) = \mathbf{P}_C(\mathbf{w}) + t(\mathbf{w} - \mathbf{P}_C(\mathbf{w}))$. Then, the projection of the point $\mathbf{w}(t)$ is $\mathbf{P}_C(\mathbf{w})$ for all $t \geq 0$, that is,*

$$\mathbf{P}_C(\mathbf{w}(t)) = \mathbf{P}_C(\mathbf{w}), \quad \forall t \geq 0.$$

Based on Theorem 5.1 and Lemma 5.1, EDPP estimates the dual optimum as follows.

Theorem 5.2 (Wang et al., 2015b)**.** *For the Lasso problem, suppose that the dual optimal solution $\theta^*(\cdot)$ at $\lambda_0 \in (0, \lambda_{\max}]$ is known. For any $\lambda \in (0, \lambda_0]$, let us define*

$$\mathbf{x}_* = \mathrm{argmax}_{\mathbf{x}_i} |\mathbf{x}_i^T \mathbf{y}|, \tag{5.27}$$

$$\mathbf{v}_1(\lambda_0) = \begin{cases} \dfrac{\mathbf{y}}{\lambda_0} - \theta^*(\lambda_0), & \text{if } \lambda_0 \in (0, \lambda_{\max}), \\ \mathrm{sign}(\mathbf{x}_*^T \mathbf{y})\mathbf{x}_*, & \text{if } \lambda_0 = \lambda_{\max}, \end{cases} \tag{5.28}$$

$$\mathbf{v}_2(\lambda, \lambda_0) = \frac{\mathbf{y}}{\lambda} - \theta^*(\lambda_0), \tag{5.29}$$

$$\mathbf{v}_2^{\perp}(\lambda, \lambda_0) = \mathbf{v}_2(\lambda, \lambda_0) - \frac{\langle \mathbf{v}_1(\lambda_0), \mathbf{v}_2(\lambda, \lambda_0)\rangle}{\|\mathbf{v}_1(\lambda_0)\|_2^2} \mathbf{v}_1(\lambda_0). \tag{5.30}$$

Then, the dual optimal solution $\theta^*(\lambda)$ can be estimated as follows:

$$\left\|\theta^*(\lambda) - \left(\theta^*(\lambda_0) + \frac{1}{2}\mathbf{v}_2^{\perp}(\lambda, \lambda_0)\right)\right\|_2 \leq \frac{1}{2}\|\mathbf{v}_2^{\perp}(\lambda, \lambda_0)\|_2.$$

For notational convenience, let

$$\begin{aligned}
\mathbf{o}(\lambda, \lambda_0) &= \theta^*(\lambda_0) + \frac{1}{2}\mathbf{v}_2^{\perp}(\lambda, \lambda_0), \\
r(\lambda, \lambda_0) &= \frac{1}{2}\|\mathbf{v}_2^{\perp}(\lambda, \lambda_0)\|_2, \\
\Theta(\lambda, \lambda_0) &= \{\theta : \|\theta - \mathbf{o}(\lambda, \lambda_0)\| \leq r(\lambda, \lambda_0)\}.
\end{aligned}$$

Theorem 5.2 implies that

$$\theta^*(\lambda) \in \Theta(\lambda, \lambda_0). \tag{5.31}$$

By Cauchy-Schwartz inequality, it is easy to solve the optimization problem $\max_{\theta \in \Theta(\lambda, \lambda_0)} |\mathbf{x}_i^{\mathrm{T}}\theta|$. Thus, by (R1') and Eq. (5.24), we immediately have the following EDPP screening rules.

Theorem 5.3. *For the Lasso problem, assume that the dual optimum $\theta^*(\cdot)$ at $\lambda_0 \in (0, \lambda_{\max}]$ is known, and $\lambda \in (0, \lambda_0]$. Then, we have $[\beta^*(\lambda)]_i = 0$ if the following holds:*

$$\left|\mathbf{x}_i^T\left(\theta^*(\lambda_0) + \frac{1}{2}\mathbf{v}_2^{\perp}(\lambda, \lambda_0)\right)\right| < 1 - \frac{1}{2}\|\mathbf{v}_2^{\perp}(\lambda, \lambda_0)\|_2\|\mathbf{x}_i\|_2.$$

In many real applications, the optimal values of the parameters are usually unknown. To determine an appropriate parameter value, commonly used approaches such as cross-validation and stability selection solve the optimization problem along a grid of parameter values, which can be very time consuming. Motivated by the ideas of (Tibshirani et al., 2012; El Ghaoui et al., 2012), we can develop a sequential version of EDPP rules. Specifically, if we need to solve the Lasso problem along a sequence of parameter values $\lambda_1 > \lambda_2 > \cdots > \lambda_m$, we can first apply EDPP to discard inactive features for the Lasso problem with parameter value being λ_1. After solving the reduced optimization problem at λ_1, we obtain the exact solution $\beta^*(\lambda_1)$. Then, by Eq. (5.24), we can find $\theta^*(\lambda_1)$. According to (R1'), once we know the optimal dual solution $\theta^*(\lambda_1)$, we can construct a new screening rule to identify inactive features for Lasso at λ_2 based on $\theta^*(\lambda_1)$. By repeating the above process, we obtain the sequential version of the EDPP rule.

We formulate the sequential version of EDPP as follows.

Corollary 5.1 (EDPP)**.** *For the Lasso problem, suppose that we are given a sequence of parameter values $\lambda_{\max} = \lambda_0 > \lambda_1 > \cdots > \lambda_{\mathcal{K}}$. Then for any integer $0 \leq k < \mathcal{K}$, we have $[\beta^*(\lambda_{k+1})]_i = 0$ if $\beta^*(\lambda_k)$ is known and the following holds:*

$$\left|\mathbf{x}_i^T\left(\frac{\mathbf{y} - \mathbf{X}\beta^*(\lambda_k)}{\lambda_k} + \frac{1}{2}\mathbf{v}_2^{\perp}(\lambda_{k+1}, \lambda_k)\right)\right| < 1 - \frac{1}{2}\|\mathbf{v}_2^{\perp}(\lambda_{k+1}, \lambda_k)\|_2\|\mathbf{x}_i\|_2. \tag{5.32}$$

5.5.1.3 Applications of EDPP to imaging genetics

In this section, we apply Lasso to identify potential risk SNPs—that are related to AD imaging phenotypes—from the ADNI WGS data by stability selection. As stability selection usually involves solving the Lasso problem many times, this process can be very time consuming. Thus we utilize EDPP to speedup the computations. We can see that the speed up gained by EDPP can be orders of magnitude. For more discussions, see Yang et al. (2015).

The ADNI WGS data contains 717 subjects. We choose the baseline hippocampal volume to be the response. To illustrate the performance of EDPP in terms of speedup, we vary the number of features p from 0.1 million to one million with a step size of 0.1 million. For each value of p, we solve the Lasso problems along a sequence of 100 parameter values equally spaced in the logarithmic scale of $\lambda/\lambda_{\max}$ from 1.0 to 0.05.

Fig. 5.1 reports the speedup gained by EDPP for data sets with different dimensions. We can observe that the speedup is up to 406 times. Moreover, Fig. 5.1 shows that the speedup gained by EDPP increases with the feature dimension growth. This implies that EDPP is a promising approach to improve the efficiency of Lasso, especially for large-scale data sets.

We next apply Lasso to explore the imaging genetics association between imaging phenotypes and SNPs from the ADNI WGS SNPs data with 329 subjects and 5,906,152 features. We utilize EDPP to facilitate the computation of Lasso problems. Specifically, for each of the enthorhinal cortex (EC) and hippocampus (HIPP) brain regions, we choose the volume at baseline and volume changes over a 24-month interval as the response vectors. We employ stability selection (Meinshausen and

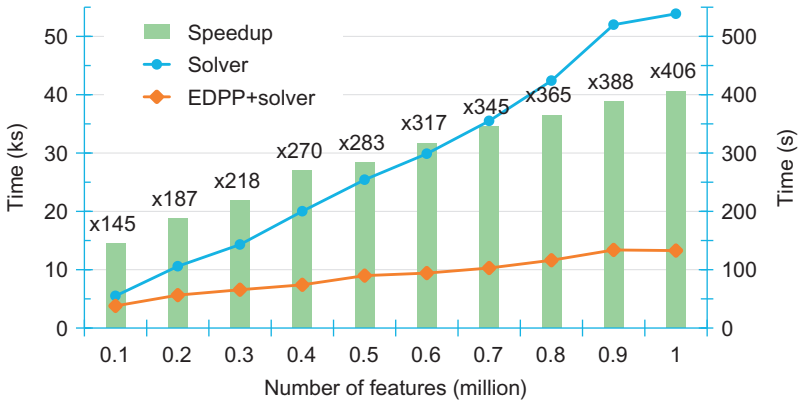

FIG. 5.1

Speedup gained by EDPP (Yang et al., 2015). Run time of solver and solver combined with EDPP are in units of kiloseconds and seconds, respectively.

Bühlmann, 2010) to identify the risk SNPs. For each response vector, we perform 100 simulations. In each simulation, we first randomly select half of the samples, and then we apply the solver combined with EDPP to solve the Lasso problem along a sequence of 100 parameter values equally spaced on the logarithmic scale of $\lambda/\lambda_{\max}$ from 1.0 to 0.05.

Tables 5.1 and 5.2 report the top 10 SNPs that are most frequently selected for each outcome. Table 5.1 shows that the genes corresponding to the selected top 10 SNPs contain those genes that are implicated in the risk of AD or other neuropsychiatric disorders. For example, APOE—that ranks among the top predictors of baseline enthorhinal and hippocampal volumes—is a top AD risk gene. This is consistent with the observations from much of the existing literature that APOE has an effect on temporal lobe structures not just in old age but also in children and adolescents (Shaw et al., 2007). We can also observe from Table 5.2 that the fourth ranked gene related to hippocampal volume change is CACNA1C. The gene CACNA1C is involved in calcium channel function that is associated with DTI measures in the hippocampus. It is known (Strohmaier et al., 2013) that CACNA1C is among the top genes associated with anxiety, depression, and obsessive compulsive disorder. Moreover, as shown in Table 5.2, Lasso identifies a gene named BACE2 that is a close homolog of BACE1. The existing literature shows that BACE1 encodes a key enzyme involved in the cellular pathways of AD.

5.5.2 SCREENING METHODS FOR OTHER SPARSE MODELS

As an emerging and promising technique in dealing with large-scale problems, *screening* has achieved great success in improving the efficiency of many popular sparse models, for example, Lasso (El Ghaoui et al., 2012; Tibshirani et al., 2012; Wang et al., 2015b; Xiang et al., 2011), nonnegative Lasso (Wang and Ye, 2014), group Lasso (Tibshirani et al., 2012; Wang et al., 2015b), mixed-norm regression (Wang et al., 2013), ℓ_1-regularized logistic regression (Wang et al., 2014b), sparse-group Lasso (Wang and Ye, 2014), tree-structured group Lasso (Wang and Ye, 2015), fused Lasso (Wang et al., 2015a), support vector machine (SVM) (Ogawa et al., 2013; Wang et al., 2014a), and least absolute deviation (LAD) (Wang et al., 2014a). The speedup gained by screening rules can be orders of magnitude.

We note that the framework of EDPP is very flexible. We have extended EDPP to many popular sparse models including all of the aforementioned ones. The package, named DPC (Dual Projection to Convex sets), is available at http://dpc-screening.github.io/. Interestingly, the recently proposed screening method for tree-structured group Lasso, MLFre (Wang and Ye, 2015), covers EDPP for Lasso as a special case, as Lasso, group Lasso, and sparse group Lasso are all special cases of tree-structured group Lasso.

Moreover, we can easily implement the screening rule EDPP and its variants in parallel, as the identification of inactive features/groups is independent from each other. This makes EDPP and its variants particularly suitable for distributed computation for data privacy and more efficiency (Li et al., 2016).

Table 5.1 Top 10 SNPs Associated With Baseline Volumes

	EC Baseline				HIPP Baseline			
	Chr	**Pos**	**RS_ID**	**Gene**	**Chr**	**Pos**	**RS_ID**	**Gene**
Rank 1	chr6	72869836	rs201890142	RIMS1	chr10	71969989	rs12412466	PPA1
Rank 2	chr19	15136345	Unknown	Unknown	chr19	45411941	rs429358	APOE
Rank 3	chr1	142555416	rs6672189	Unknown	chr11	11317240	rs10831576	GALNT18
Rank 4	chr19	45411941	rs429358	APOE	chr4	49147785	rs151073945	Unknown
Rank 5	chr2	96630311	rs369756382	ANKRD36C	chr8	145158607	rs34173062	MAF1
Rank 6	chr2	95552943	rs199536016	LOC442028	chr6	168107162	rs71573413	Unknown
Rank 7	chr2	95552986	rs200710055	LOC442028	chrX	143403234	rs4825209	Unknown
Rank 8	chr1	142545571	LOC442028	Unknown	chr2	231846840	rs4973360	Unknown
Rank 9	chr13	94122100	rs76403280	GPC6	chr14	20710095	rs35055545	OR11H4
Rank 10	chr5	97913219	rs202036446	Unknown	chr6	69775654	rs2343398	BAI3

Table 5.2 Top 10 SNPs Associated With Volume Changes

	EC Changes				HIPP Changes			
	Chr	**Pos**	**RS_ID**	**Gene**	**Chr**	**Pos**	**RS_ID**	**Gene**
Rank 1	chr7	7333294	rs1317198	Unknown	chr15	23060281	rs11636690	NIPA1
Rank 2	chr10	9216956	rs1149952	Unknown	chr21	42613255	rs74977559	BACE2
Rank 3	chr6	115278412	rs146156795	Unknown	chr9	91366940	rs79543088	Unknown
Rank 4	chr10	117854524	rs2530339	GFRA1	chr12	2342248	rs7303977	CACNA1C
Rank 5	chr8	6557130	rs2912047	LOC100507530	chr6	169666147	rs6605518	Unknown
Rank 6	chr12	23117263	rs12581794	Unknown	chr18	48740822	rs34794713	Unknown
Rank 7	chr16	77841030	rs16946521	VAT1L	chr13	102307271	rs9518474	ITGBL1
Rank 8	chr3	175784869	rs9845573	Unknown	chrX	2372296	rs7889210	DHRSX
Rank 9	chr4	7494498	rs4308363	SORCS2	chr4	174673036	rs12646029	LOC101928478
Rank 10	chr13	102616850	rs17502999	FGF14	chr4	170638416	rs149287207	CLCN3

5.6 CONCLUSIONS

In this chapter, we review many popular sparse models for imaging genetics. Due to their capability of incorporating various prior knowledge, sparse models are very effective in identifying the predictors that exhibit the strongest effects on the imaging phenotypes. However, as the regularizers of the sparse models are usually nonsmooth and complex, applications of sparse models to large-scale problems entail great challenges to existing optimization algorithms. To deal with this challenge, we introduce a suite of novel techniques, called sparse screening, to effectively scale existing algorithms to large-scale problems. This usually leads to substantial savings in memory usage and the resulting speedup can be several orders of magnitude. Thus we expect that sparse screening will be a powerful tool in facilitating the research of imaging genetics.

REFERENCES

Argyriou, A., Evgeniou, T., Pontil, M., 2008. Convex multi-task feature learning. Mach. Learn. 73 (3), 243–272.

Ashburner, M., Ball, C., Blake, J., Botstein, D., Butler, H., 2000. Gene ontology: tool for the unification of biology. the gene ontology consortium. Nat. Genet. 25, 25–29.

Batmanghelich, N.K., Dalca, A.V., Sabuncu, M.R., Golland, P., 2013. Joint modeling of imaging and genetics. In: Information Processing in Medical Imaging. Springer, Heidelberg, pp. 766–777.

Bauschke, H.H., Combettes, P.L., 2011. Convex Analysis and Monotone Operator Theory in Hilbert Spaces. Springer, New York.

Beck, A., Teboulle, M., 2009. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. SIAM J. Imaging Sci. 2 (1), 183–202.

Belkin, M., Niyogi, P., 2003. Laplacian eigenmaps for dimensionality reduction and data representation. Neural Comput. 15, 1373–1396.

Bis, J., et al., 2012. Common variants at 12q14 and 12q24 are associated with hippocampal volume. Nat. Genet. 44 (5), 545–551.

Bruckstein, A., Donoho, D., Elad, M., 2009. From sparse solutions of systems of equations to sparse modeling of signals and images. SIAM Rev. 51, 34–81.

Candès, E.J., 2006. Compressive sampling. In: Proceedings of the International Congress of Mathematicians, Madrid, Spain, vol. 3, pp. 1433–1452.

Chen, S.S., Donoho, D.L., Saunders, M.A., 2001. Atomic decomposition by basis pursuit. SIAM Rev. 43, 129–159.

Chung, F., 1997. Spectral Graph Theory. American Mathematical Society Providence, RI.

Cirulli, E.T., Goldstein, D.B., 2010. Uncovering the roles of rare variants in common disease through whole-genome sequencing. Nat. Rev. Genet. 11 (6), 415–425.

Cornelis, M., Qi, L., Zhang, C., Kraft, P., Manson, J., Cai, T., Hunter, D., Hu, F., 2009. Joint effects of common genetic variants on the risk for type 2 diabetes in US men and women of European ancestry. Ann. Intern. Med. 150, 541–550.

Dinu, I., Mahasirimongkol, S., Liu, Q., Yanai, H., Eldin, N., Kreiter, E., Wu, X., Jabbari, S., Tokunaga, K., Yasui, Y., 2012. SNP-SNP interactions discovered by logic regression explain Crohns disease genetics. PLoS ONE 7, e43035.

El Ghaoui, L., Viallon, V., Rabbani, T., 2012. Safe feature elimination in sparse supervised learning. Pac. J. Optim. 8, 667–698.

Fan, R., Chang, K., Lv, J., 2008. Sure independence screening for ultrahigh dimensional feature spaces. J. R. Stat. Soc. B 70, 849–911.

Friedman, J., Hastie, T., Tibshirani, R., 2010. A note on the group lasso and a sparse group lasso. arXiv preprint, arXiv 1001.0736.

Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J.P., Coller, H., Loh, M.L., Downing, J.R., Caligiuri, M.A., Bloomfield, C.D., 1999. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 286 (5439), 531–537.

Guyon, I., Weston, J., Barnhill, S., Vapnik, V., 2002. Gene selection for cancer classification using support vector machines. Mach. Learn. 46 (1–3), 389–422.

Hariri, A.R., Drabant, E.M., Weinberger, D.R., 2006. Imaging genetics: perspectives from studies of genetically driven variation in serotonin function and corticolimbic affective processing. Biol. Psychiat. 59 (10), 888–897.

Harris, M., et al., 2004. The gene ontology database and informatics resource. Nucleic Acids Res. 32, 258–261.

Hastie, T., Tibshirani, R., Wainwright, M., 2015. Statistical Learning With Sparsity: The Lasso and Generalizations. CRC Press, Boca Raton, FL.

James, G., Witten, D., Hastie, T., Tibshirani, R., 2013. An Introduction to Statistical Learning, vol. 112. Springer, New York.

Kim, S., Xing, E.P., 2010. Tree-guided group lasso for multi-task regression with structured sparsity. In: Proceedings of the 27th International Conference on Machine Learning (ICML-10), pp. 543–550.

Kowalski, M., 2009. Spare regression using mixed norms. Appl. Comput. Harmon. Anal. 27, 303–324.

Li, C., Li, H., 2008. Network-constrained regularization and variable selection for analysis of genomic data. Bioinformatics 24, 1175–1182.

Li, Q., Yang, T., Zhan, L., Hibar, D., Jananshad, N., Ye, J., Thompson, P., Wang, J., 2015. Large-scale collaborative genetic studies of risk SNPs for Alzheimer's disease across multiple institutions (under submission).

Li, Y., Wang, J., Yang, T., Chen, J., Liu, L., Thompson, P., Ye, J., 2016. Detection of Alzheimer's disease risk factors by tree-structured group lasso screening. In: 2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI), in press.

Liu, J., Ye, J., 2010. Moreau-Yosida regularization for grouped tree structure learning. In: Lafferty, J.D., Williams, C.K.I., Shawe-Taylor, J., Zemel, R.S., Culotta, A. (Eds.), Advances in Neural Information Processing Systems 23. Curran Associates, Inc., Red Hook, NY, pp. 1459–1467.

Liu, H., Palatucci, M., Zhang, J., 2009. Blockwise coordinate descent procedures for the multi-task lasso, with applications to neural semantic basis discovery. In: Proceedings of the 26th Annual International Conference on Machine Learning, ICML '09, Montreal, Quebec, Canada. ACM, New York, NY, pp. 649–656. ISBN 978-1-60558-516-1

Medland, S.E., Jahanshad, N., Neale, B.M., Thompson, P.M., 2014. Whole-genome analyses of whole-brain data: working within an expanded search space. Nat. Neurosci. 17 (6), 791–800.

Meier, L., Geer, S., Bühlmann, P., 2008. The group lasso for logistic regression. J. R. Stat. Soc. B 70, 53–71.

Meinshausen, N., Bühlmann, P., 2010. Stability selection. J. R. Stat. Soc. B 72, 417–473.

Meyer-Lindenberg, A., 2012. The future of fMRI and genetics research. NeuroImage 62 (2), 1286–1292.

Mu, Y., Gage, F., 2011. Adult hippocampal neurogenesis and its role in Alzheimers disease. Mole. Neurodegen. 6, 85.

Negahban, S., Wainwright, M., 2008. Joint support recovery under high-dimensional scaling: benefits and perils of $\ell_{1,\infty}$-regularization. In: Advances in Neural Information Processing Systems, pp. 1161–1168.

Nesterov, Y., 1983. A method for solving a convex programming problem with convergence rate $1/k^2$. Sov. Math. Dokl. 27 (2), 372–376.

Nesterov, Y., 2007. Gradient methods for minimizing composite objective function. Center for Operations Research and Econometrics (CORE), Université Catholique de Louvain. CORE Discussion Papers No. 2007076, http://EconPapers.repec.org/RePEc:cor:louvco:2007076.

Obozinski, G., Taskar, B., Jordan, M.I., 2007. Joint covariate selection for grouped classification. Statistics Department, UC Berkeley.

Ogawa, K., Suzuki, Y., Takeuchi, I., 2013. Safe screening of non-support vectors in pathwise SVM computation. In: Proceedings of the 30th International Conference on Machine Learning, pp. 1382–1390.

Peng, J., Zhu, J., Bergamaschi, A., Han, W., Noh, D., Pollack, J., Wang, P., 2010. Regularized multivariate regression for identifying master predictors with application to integrative genomics study of breast cancer. Ann. Appl. Stat. 4, 53–77.

Potkin, S.G., Guffanti, G., Lakatos, A., Turner, J.A., Kruggel, F., Fallon, J.H., Saykin, A.J., Orro, A., Lupoli, S., Salvi, E., et al., 2009. Hippocampal atrophy as a quantitative trait in a genome-wide association study identifying novel susceptibility genes for Alzheimer's disease. PLoS ONE 4 (8), e6501.

Psaty, B.M., O'Donnell, C.J., Gudnason, V., Lunetta, K.L., Folsom, A.R., Rotter, J.I., Uitterlinden, A.G., Harris, T.B., Witteman, J.C., Boerwinkle, E., et al., 2009. Cohorts for heart and aging research in genomic epidemiology (charge) consortium design of prospective meta-analyses of genome-wide association studies from 5 cohorts. Circ. Cardiovasc. Genet. 2 (1), 73–80.

Quattoni, A., Carreras, X., Collins, M., Darrell, T., 2009. An efficient projection for $\ell_{1,\infty}$, infinity regularization. In: Proceedings of the 26th Annual International Conference on Machine Learning, ICML '09, Montreal, Quebec, Canada. ACM, New York, NY, pp. 857–864. ISBN 978-1-60558-516-1.

Shaw, P., Lerch, J.P., Pruessner, J.C., Taylor, K.N., Rose, A.B., Greenstein, D., Clasen, L., Evans, A., Rapoport, J.L., Giedd, J.N., 2007. Cortical morphology in children and adolescents with different apolipoprotein e gene polymorphisms: an observational study. Lancet Neurol. 6 (6), 494–500.

Simon, N., Friedman., J., Hastie., T., Tibshirani, R., 2013. A sparse-group lasso. J. Comput. Graph. Stat. 22, 231–245.

Singh, M., Singh, P., Juneja, P., Singh, S., Kaur, T., 2011. SNP–SNP interactions within APOE gene influence plasma lipids in postmenopausal osteoporosis. Rheumat. Int. 31, 421–423.

Sprechmann, P., Ramírez, I., Sapiro., G., Eldar, Y., 2011. C-HiLasso: a collaborative hierarchical sparse modeling framework. IEEE Trans. Signal Process. 59, 4183–4198.

Stein, J.L., Medland, S.E., Vasquez, A.A., Hibar, D.P., Senstad, R.E., Winkler, A.M., Toro, R., Appel, K., Bartecek, R., Bergmann, Ø., et al., 2012. Identification of common variants associated with human hippocampal and intracranial volumes. Nat. Genet. 44 (5), 552–561.

Strohmaier, J., Amelang, M., Hothorn, L.A., Witt, S.H., Nieratschker, V., Gerhard, D., Meier, S., Wust, S., Frank, J., Loerbroks, A., Rietschel, M., Sturmer, T., Schulze, T.G., 2012. The psychiatric vulnerability gene CACNA1C and its sex-specific relationship with personality traits, resilience factors and depressive symptoms in the general population. Mol. Psychiatry 18 (5), 607–613.

Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., Mesirov, J.P., 2005. Gene set enrichment analysis: A knowledge-based approach for interpreting genomewide expression profiles. Proc. Natl. Acad. Sci. USA 102 (43), 15545–15550.

Sun, L., Liu, J., Chen, J., Ye, J., 2009. Efficient recovery of jointly sparse vectors. In: Bengio, Y., Schuurmans, D., Lafferty, J.D., Williams, C.K.I., Culotta, A. (Eds.), Advances in Neural Information Processing Systems 22. Curran Associates, Inc., Red Hook, NY, pp. 1812–1820.

Thompson, P.M., Ge, T., Glahn, D.C., Jahanshad, N., Nichols, T.E., 2013. Genetics of the connectome. NeuroImage 80, 475–488.

Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. J. R. Stat. Soc. B 58, 267–288.

Tibshirani, R., Wang, P., 2008. Spatial smoothing and hot spot detection for CGH data using the fused lasso. Biostatistics 9 (1), 18–29.

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., Knight, K., 2005. Sparsity and smoothness via the fused lasso. J. R. Stat. Soc. B 67 (1), 91–108.

Tibshirani, R., Bien, J., Friedman, J., Hastie, T., Simon, N., Taylor, J., Tibshirani, R., 2012. Strong rules for discarding predictors in lasso-type problems. J. R. Stat. Soc. B 74, 245–266.

Vidyasagar, M., 2014. Machine learning methods in the computational biology of cancer. Proc. R. Soc. Lond. A 471 (2173), 20140805. http://dx.doi.org/10.1098/rspa.2014.0805.

Vinje, W., Gallant, J., 2000. Sparse coding and decorrelation in primary visual cortex during natural vision. Science 287, 1273–1276.

Wang, J., Ye, J., 2014. Two-layer feature reduction for sparse-group lasso via decomposition of convex sets. In: Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N.D., Weinberger, K.Q. (Eds.), Advances in Neural Information Processing Systems 27. Curran Associates, Inc., Red Hook, NY, pp. 2132–2140.

Wang, J., Ye, J., 2015. Multi-layer feature reduction for tree structured group lasso via hierarchical projection. In: Cortes, C., Lawrence, N.D., Lee, D.D., Sugiyama, M., Garnett, R. (Eds.), Advances in Neural Information Processing Systems 28. Curran Associates, Inc., Red Hook, NY, pp. 1279–1287.

Wang, H., Nie, F., Huang, H., Risacher, S.L., Saykin, A.J., Shen, L., et al., 2012. Identifying disease sensitive and quantitative trait-relevant biomarkers from multidimensional heterogeneous imaging genetics data via sparse multimodal multitask learning. Bioinformatics 28 (12), i127–i136.

Wang, J., Jun, J., Ye, J., 2013. Efficient mixed-norm regularization: algorithms and safe screening methods. CoRR, abs/1307.4156. http://arxiv.org/abs/1307.4156.

Wang, J., Wonka, P., Ye, J., 2014a, Scaling SVM and least absolute deviations via exact data reduction. In: Proceedings of the 31st International Conference on Machine Learning, ICML 2014, Beijing, China, June 21–26, 2014, pp. 523–531.

Wang, J., Zhou, J., Liu, J., Wonka, P., Ye, J., 2014b. A safe screening rule for sparse logistic regression. In: Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N.D., Weinberger, K.Q. (Eds.), Advances in Neural Information Processing Systems 27. Curran Associates, Inc., Red Hook, NY, pp. 1053–1061.

Wang, J., Fan, W., Ye, J., 2015a. Fused lasso screening rules via the monotonicity of subdifferentials. IEEE Trans. Pattern Anal. Mach. Intell. 37 (9), 1806–1820.

Wang, J., Wonka, P., Ye, J., 2015b. Lasso screening rules via dual polytope projection. J. Mach. Learn. Res. 16, 1063–1101. http://jmlr.org/papers/v16/wang15a.html.

Wright, J., Ma, Y., Mairal, J., Sapiro, G., Huang, T.S., Yan, S., 2010. Sparse representation for computer vision and pattern recognition. Proc. IEEE 98, 1031–1044.

Wu, T.T., Chen, Y.F., Hastie, T., Sobel, E., Lange, K., 2009. Genomewide association analysis by lasso penalized logistic regression. Bioinformatics 25, 714–721.

Xiang, Z.J., Ramadge, P.J., 2012. Fast lasso screening tests based on correlations. In: 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 2137–2140. http://dx.doi.org/10.1109/ICASSP.2012.6288334.

Xiang, Z.J., Xu, H., Ramadge, P.J., 2011. Learning sparse representations of high dimensional data on large scale dictionaries. In: Shawe-Taylor, J., Zemel, R.S., Bartlett, P.L., Pereira, F., Weinberger, K.Q. (Eds.), Advances in Neural Information Processing Systems 24. Curran Associates, Inc., Red Hook, NY, pp. 900–908.

Yang, J., Ferreira, T., Morris, A., Medland, S., Madden, P., Heath, A., Martin, N., Montgomery, G., Weedon, M., Loos, R., Frayling, T., McCarthy, M., Hirschhorn, J., Goddard, M., Visscher, P., 2012. Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. Nat. Genet. 44, 369–375.

Yang, T., Wang, J., Sun, Q., Hibar, D., Jahanshad, N., Liu, L., Wang, Y., Zhan, L., Thompson, P., Ye, J., 2015, April. Detecting genetic risk factors for Alzheimer's disease in whole genome sequence data via lasso screening. In: IEEE 12th International Symposium on Biomedical Imaging (ISBI), pp. 985–989.

Yogatama, D., Smith, N.A., 2014. Linguistic structured sparsity in text categorization. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Vol. 1: Long Papers), Baltimore, MD. Association for Computational Linguistics, Berlin, Germany, pp. 786–796.

Yuan, M., Lin, Y., 2006. Model selection and estimation in regression with grouped variables. J. R. Stat. Soc. B 68, 49–67.

Zhao, P., Yu, B., 2006. On model selection consistency of lasso. J. Mach. Learn. Res. 7, 2541–2563.

Zhao, P., Rocha, G., Yu, B., 2009. The composite absolute penalties family for grouped and hierarchical variable selection. Ann. Stat. 37 (6A), 3468–3497.

Zhu, J., Hastie, T., 2004. Classification of gene microarrays by penalized logistic regression. Biostatistics 5, 427–443.

Zou, H., Hastie, T., 2005. Regularization and variable selection via the elastic net. J. R. Stat. Soc. B 67, 301–320.