Sparse Methods for Biomedical Data

Jieping Ye
Arizona State University
Tempe, AZ 85287
jieping.ye@asu.edu

Jun Liu
Siemens Corporate Research
Princeton, NJ 08540
jun-liu@siemens.com

ABSTRACT

Following recent technological revolutions, the investigation of massive biomedical data with growing scale, diversity, and complexity has taken a center stage in modern data analysis. Although complex, the underlying representations of many biomedical data are often sparse. For example, for a certain disease such as leukemia, even though humans have tens of thousands of genes, only a few genes are relevant to the disease; a gene network is sparse since a regulatory pathway involves only a small number of genes; many biomedical signals are sparse or compressible in the sense that they have concise representations when expressed in a proper basis. Therefore, finding sparse representations is fundamentally important for scientific discovery. Sparse methods based on the ℓ_1 norm have attracted a great amount of research efforts in the past decade due to its sparsity-inducing property, convenient convexity, and strong theoretical guarantees. They have achieved great success in various applications such as biomarker selection, biological network construction, and magnetic resonance imaging. In this paper, we review state-of-the-art sparse methods and their applications to biomedical data.

Keywords

Sparse learning, structured sparsity, Gaussian graphical model, magnetic resonance imaging

1. INTRODUCTION

Recent technological revolutions have unleashed a torrent of biomedical data with growing scale, diversity, and complexity [24; 27; 77; 86; 101]. The wealth of data confronts scientists with an urgent need for new methods and tools that can intelligently and automatically extract useful information from data and synthesize knowledge [17; 32; 56; 74]. Although complex, the underlying representations of many real-world data are often sparse [32; 38; 41]. For example, for a certain disease such as leukemia, even though humans have tens of thousands of genes, only a small number of them are relevant to the disease; a gene network is sparse since a regulatory pathway involves only a small number of genes; the neural representation of sounds in the auditory cortex of unanesthetized animals is sparse, since the fraction of neurons active at a given instant is small; many biomedical signals have sparse representations when expressed in a proper basis. Therefore, finding sparse representations is fundamentally important for scientific discovery. The last decade has witnessed a growing interest in the search for sparse representations of data.

The quest for sparsity is further motivated for various reasons. First, sparse representations enhance the interpretability of the model. For example, in many biological applications, the selection of genes or proteins which are related to the study, is crucial to facilitate the biological interpretation [18; 38]. In addition, the resulting gene/protein selection might enable a feasible biological validation with a reduced experimental cost. Second, sparseness is one way to measure the complexity of the learning model [84]. Regularization is commonly employed to penalize the complexity of a learning model and alleviate overfitting. Regularization based on the ℓ_0 norm maximizes sparseness, which, however, leads to an NP-hard problem. As a computationally efficient alternative, the ℓ_1 norm regularization, which also leads to a sparse model, is widely used in many areas including signal processing, statistics, and machine learning [13; 23; 52; 93; 98; 124; 127]. Finally, finding sparse representations has recently received increasing attention due to the current burst of research in Compressed Sensing (CS) [4; 6; 16; 25; 26; 102]. CS is a technique for acquiring and reconstructing a signal utilizing the prior knowledge that it is sparse or compressible. It encodes a large sparse signal using a relatively small number of linear measurements, and minimizing the ℓ_1 norm in order to decode the signal. Recent theories [13; 14; 15; 16; 25] assert that one can recover certain signals and images from far fewer samples or measurements than traditional methods.

In this paper, we review sparse methods for (1) incorporating *a priori* knowledge on feature structures for feature selection, (2) constructing undirected Gaussian graphical models, and (3) parallel magnetic resonance imaging.

Structured Feature Selection. Although sparse learning models based on the ℓ_1 norm such as the Lasso [98] have achieved great success in many applications, they do not take the existing feature structure into consideration. Specifically, these models yield the same solution after randomly reshuffling the features. However, in many applications, the features exhibit certain intrinsic structures, e.g., spatial or temporal smoothness, disjoint/overlapping groups, trees, and graphs [42; 45; 51; 65; 116]. The *a priori* structure information may significantly improve the classification/regression performance and help identify the important features. For example, in the study of arrayCGH [99; 100], the features—the DNA copy numbers along the genome—

have the natural spatial order, and the fused Lasso, which incorporates the structure information using an extension of the ℓ_1-norm, outperforms the Lasso in both classification and feature selection. In this paper, we review various structured sparse learning models including group Lasso, sparse group Lasso, overlapping group Lasso, tree Lasso, fused Lasso, and graph Lasso.

Sparse Undirected Gaussian Graphical Models. Undirected graphical models explore the relationships among a set of random variables through their joint distribution. The estimation of undirected graphical models has applications in many domains, such as computer vision, biology, and medicine. An instance is the analysis of gene expression data. As shown in many biological studies, genes tend to work in groups based on their biological functions, and there exist some regulatory relationships between genes [19]. Such biological knowledge can be represented as a graph, where nodes are the genes, and edges describe the regulatory relationships. Graphical models provide a useful tool for modeling these relationships, and can be used to explore gene activities. One of the popular graphical models is the Gaussian graphical model (GGM), which assumes the variables to be Gaussian distributed [5]. In GGM, the problem of learning a graph is equivalent to estimating the inverse of the covariance matrix (precision matrix), since the nonzero off-diagonal elements of the precision matrix represent edges in the graph [5]. In some applications, we need to estimate multiple related precision matrices. For example, in the modeling of brain networks for Alzheimer's disease using neuroimaging data [43], we want to estimate graphical models for three groups: normal controls (NC), patients of mild cognitive impairment (MCI), and Alzheimer's patients (AD). These graphs are expected to share some common connections, but they are not identical. It is thus desirable to jointly estimate the three graphs. In this paper, we review sparse methods for estimating a single undirected graphical model and for estimating multiple related undirected graphical models and discuss their properties.

Parallel Magnetic Resonance Imaging. Parallel imaging has been the single biggest innovation in magnetic resonance imaging in the last decade. It exploits the difference in sensitivities between individual coil elements in a receive array to reduce the number of gradient encodings required for imaging, and the increase in speed comes at a time when other approaches to acquisition time reduction were reaching engineering and human limits [59]. In the SENSE-type reconstruction approach, researchers have taken advantage of the sparsity promoting penalties (e.g., wavelets and total variations) to reduce the acquisition time while maintaining the image quality. Key components of sparse learning include the estimation of the coil sensitivity profiles, the design of the sparsity promoting regularization, the development of the sampling pattern that takes advantage of sparse learning, and the efficient optimization of the non-smooth inverse problem. In this paper, we review different components of sparse learning in magnetic resonance imaging.

The rest of the paper is organized as follows. We review structured sparse learning for feature selection in Section 2. The estimation of sparse undirected Gaussian graphical models is presented in Section 3. We discuss sparse learning in parallel magnetic resonance imaging in Section 4. Finally, we conclude the paper in Section 5.

2. STRUCTURED FEATURE SELECTION

We are given a set of training samples $\{\mathbf{a}_i, b_i\}_{i=1}^n$, where $\mathbf{a}_i \in \mathbb{R}^p$ denotes the p-dimensional features for the i-th sample, and $b_i \in \mathbb{R}$ is its response (numeric for regression, and categorical for classification). In addition, we are given a feature structure, e.g., a group structure, a tree structure, or a graph structure, as part of the input data. We focus on a linear model $h : \mathbb{R}^p \to \mathbb{R}$ with $h(\mathbf{a}) = \mathbf{x}^{\mathrm{T}}\mathbf{a}$, where $\mathbf{x} \in \mathbb{R}^p$ is the vector of model parameters. To fit the model with the training samples, we learn the model parameter vector $\mathbf{x}$ by solving the following optimization problem:

$$\min_{\mathbf{x}} f(\mathbf{x}) \equiv L(\mathbf{x}) + \lambda\Omega(\mathbf{x}), \tag{1}$$

where $L(\mathbf{x})$ is a loss function, $\Omega(\mathbf{x})$ is a regularization term encoding the prior knowledge on the input features, and $\lambda > 0$ is the regularization parameter controlling the trade-off between the loss $L(\cdot)$ and the penalty $\Omega(\cdot)$.
The formulation in (1) can be applied for regression, classification, and longitudinal data analysis:

- Regression: The outcome b is a continuous value, e.g., the hippocampus volume or the minimental state examination (MMSE) score of a subject in the study of Alzheimer's disease. The least squares loss is commonly used for regression.
- Classification: The outcome b is a discrete value, e.g., disease status, including normal controls and disease patients. The logistic loss is commonly used for classification.
- Longitudinal Data Analysis: The outcome b is the observed failure/censoring time. If an event occurs at time t, then the subject has a failure time t. If a patient drops from the study at time t, we consider he/she is censored at time t. The Cox model is a popular approach for longitudinal data analysis, in which the negative log-likelihood function of the proportional hazard is used as the loss function [21].

The regularization term $\Omega(\mathbf{x})$ in (1) is commonly employed to penalize the complexity of a learning model and alleviate overfitting, e.g., the ℓ_2-norm regularization used in ridge regression. However, the commonly used ℓ_2-norm regularization leads to a dense model, i.e., almost all model parameters in $\mathbf{x}$ are non-zero. To enhance the interpretability of the model, a sparse model is desired. One popular sparse model, known as the Lasso, is based on the ℓ_1-norm penalty:

$$\Omega_{\text{lasso}}(\mathbf{x}) = \|\mathbf{x}\|_1. \tag{2}$$

The Lasso has been applied widely in many biomedical applications [91; 94; 107; 111; 123]. In many applications, the features exhibit certain intrinsic structures, e.g., spatial or temporal smoothness, graphs, trees, and disjoint/overlapping groups. The *a priori* structure information may significantly improve the classification/regression performance and help to identify the important features.

2.1 Group Lasso and Sparse Group Lasso

In many applications, the features form a natural group structure. For example, the voxels of the positron emission tomography (PET) images in the Alzheimer's Disease study can be divided into a set of non-overlapping groups according to the brain regions [43]; in the multi-factor ANOVA

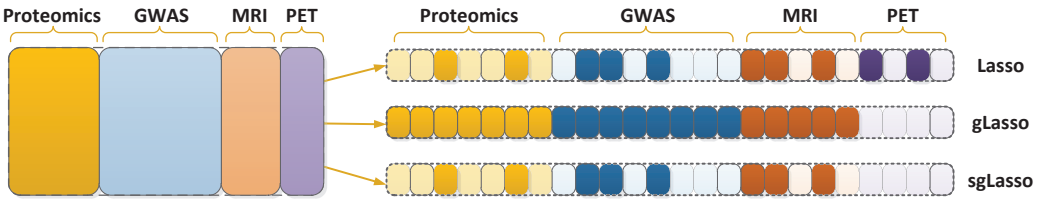

Figure 1: Illustration of Lasso, group Lasso (gLasso), and sparse group Lasso (sgLasso). Four types of data sources, including Proteomics, GWAS (genome-wide association study), MRI (magnetic resonance imaging), and PET from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database are used for illustration. There are four feature groups, one for each data source. The features selected by each algorithm are highlighted.

problem, each factor may have several levels and can be represented using a group of dummy variables [117]. The selection of group structures has recently received increasing attention in the literature [3; 44; 45; 64; 78; 117; 120]. The pioneer work [117] focused on the non-overlapping group Lasso, i.e., the groups are disjoint. Assume the features are partitioned into k disjoint groups $\{G_1, \cdots, G_k\}$. The group Lasso formulation uses the $\ell_{q,1}$-norm penalty on the model parameters:

$$\Omega_{\text{gLasso}}(\mathbf{x}) = \sum_{i=1}^{k} w_i \|\mathbf{x}_{G_i}\|_q, \tag{3}$$

where $\|\cdot\|_q$ is the ℓ_q-norm with $q > 1$ (most existing work focus on $q = 2$ or ∞) [68], and w_i is the weight for the i-th group. The group selection distinguishes the group Lasso from the Lasso which does not take group information into account and does not support group selection. The group Lasso has been applied for regression [55; 80; 117], classification [78], joint covariate selection for grouped classification [85], and multi-task learning [2; 62; 89].

The group Lasso does not perform feature selection within each feature group. For certain applications, it is desirable to perform simultaneous group selection and feature selection. The sparse group Lasso (sgLasso) incorporates the strengthens from both Lasso and group Lasso, and it yields a solution with simultaneous between- and within- group sparsity [30; 87]. The sparse group Lasso penalty is based on a composition of the $\ell_{q,1}$-norm and the ℓ_1-norm:

$$\Omega_{\text{sgLasso}}(\mathbf{x}) = \alpha\|\mathbf{x}\|_1 + (1-\alpha)\sum_{i=1}^{k} w_i \|\mathbf{x}_{G_i}\|_q, \tag{4}$$

where $\alpha \in [0, 1]$, the first term controls the sparsity in the feature level, and the second term controls the sparsity in the group level. The sparse group Lasso has been applied to analyze multiple types of high dimensional genomic data for biomarker discovery [87].

Figure 1 illustrates Lasso, group Lasso, and sparse group Lasso; we use four types of data sources including Proteomics, GWAS (genome-wide association study), MRI (magnetic resonance imaging), and PET from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database[1]. We construct four feature groups, one for each each data source. As shown in the figure, the Lasso does not consider the group (source) information and selects a subset of features from all four groups; the group Lasso selects a subset of the groups (3 in this example) and all features from these 3 groups are selected; the sparse group Lasso simultaneously selects a subset of the groups and a subset of the features within each selected group.

[1]http://www.adni-info.org/

2.2 Overlapping Group Lasso and Tree Lasso

In group Lasso [117], the groups are disjoint. Some recent work [44; 45; 46; 51; 69; 120] studied the more general case where the groups may overlap. One motivating example is the use of biologically meaningful gene/protein sets (groups). The proteins/genes in the same groups are related if they either appear in the same pathway, or are semantically related in terms of Gene Ontology (GO) hierarchy, or are related from gene set enrichment analysis (GSEA) [97]. The canonical pathway in MSigDB, for example, has provided 639 groups of genes [97]. It has been shown that the group (of proteins/genes) markers are more reproducible than individual protein/gene markers and the use of such group information improves classification performance [19]. Groups may overlap - one protein/gene may belong to multiple groups - and the group Lasso formulation is not applicable. For the general overlapping group patterns, we can make use of the following overlapping group Lasso penalty [120]:

$$\Omega_{\text{overlapping}}(\mathbf{x}) = \alpha\|\mathbf{x}\|_1 + (1-\alpha)\sum_{i=1}^{k} w_i \|\mathbf{x}_{G_i}\|_q \tag{5}$$

where $\alpha \in [0, 1]$, $w_i > 0$ $(i = 1, 2, \ldots, k)$, and G_i consists of the indices from the i-th group of features. The k groups of features are pre-specified, and they may overlap. A different overlapping group Lasso formulation was proposed in [44].

In some applications, the features follow a tree structure. For example, an image can be represented using a tree structure where each leaf node corresponds to a feature (pixel) and each internal node corresponds to a group of features (pixels) based on the spatial locality [69]. In such a case, we can make use of the tree structured group Lasso penalty [46; 51; 69; 120]:

$$\Omega_{\text{tree}}(\mathbf{x}) = \sum_{i,j} w_j^i \|\mathbf{x}_{G_j^i}\|_q, \tag{6}$$

where $w_j^i > 0$ is a constant weight, and G_j^i, a node at the depth i, consists of all features in the subtree. Note that any parent node is a superset of its children. Thus, if a specific node is not selected (i.e., its corresponding model coefficient is zero), then all its children will not be selected. It is clear that the tree structured group Lasso is a special case of the overlapping group Lasso with a specific tree structure.

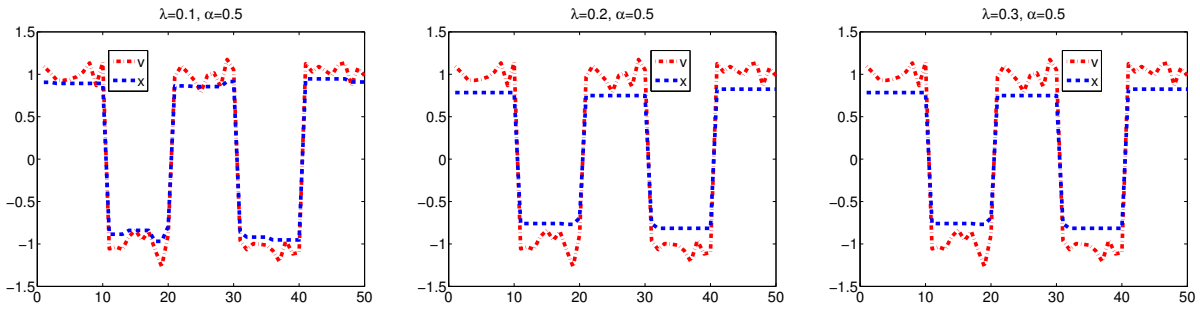

Figure 2: Illustration of the solution $\mathbf{x} = \pi^{\lambda,\alpha}_{\text{fused}}(\mathbf{v})$ of (8), the fused Lasso signal approximator.

2.3 Fused Lasso

In many applications, the features enjoy certain smoothness properties. For example, the adjacent features in the arrayCGH data are close to each other along the genome. Therefore, it is desirable to enforce the model parameters in $\mathbf{x}$ to have the structure of smoothness. Such a structure can be induced by the fused Lasso penalty [28; 99]:

$$\Omega_{\text{fused}}(\mathbf{x}) = \alpha\|\mathbf{x}\|_1 + (1-\alpha)\sum_{i=1}^{p-1}|x_i - x_{i+1}|, \qquad (7)$$

where $\alpha \in [0,1]$. The fused Lasso penalty in (7) shall induce a solution that x_i tends to be close or identical to x_{i+1} for $i = 1, \cdots, p-1$. The smoothness structure can also be revealed from the fused Lasso signal approximator [28]:

$$\pi^{\lambda,\alpha}_{\text{fused}}(\mathbf{v}) = \arg\min_{\mathbf{x}} \left\{ \frac{1}{2}\|\mathbf{x}-\mathbf{v}\|^2 + \lambda\Omega_{\text{fused}}(\mathbf{x}) \right\}. \qquad (8)$$

Figure 2 illustrates the fused Lasso signal approximator (8) under different values of λ with $\alpha = 0.5$. We can observe that the solution is piecewise constant.

2.4 Graph Lasso

In certain applications, the features form an undirected graph structure, in which two features connected by an edge in the graph are more likely to be selected together. As an example, many biological studies have suggested that genes tend to work in groups according to their biological functions, and there are some regulatory relationships between genes [60]. This biological knowledge can be represented as a graph, where the nodes represent the genes, and the edges imply the regulatory relationships between genes. Figure 3 shows a subgraph consisting of 80 nodes (genes) of the network described in [19]. Several recent studies have shown that the estimation accuracy can be improved using dependency information encoded as a graph. Let (N, E) be a given graph, where $N = \{1, 2, \cdots, p\}$ is a set of nodes, and E is a set of edges. Node i corresponds to the i-th feature. If nodes i and j are connected by an edge in E, then the i-th feature and the j-th feature tend to be grouped.

The fussed Lasso penalty in (7) can be extended to a general graph structure; we call it the ℓ_1 graph Lasso:

$$\Omega_{\text{graph},\ell_1}(\mathbf{x}) = \alpha\|\mathbf{x}\|_1 + (1-\alpha)\sum_{(i,j)\in E}|x_i - x_j|, \qquad (9)$$

where the second regularization term penalizes a large deviation between two model parameters whose corresponding nodes are connected in the graph. Intuitively, if two genes/proteins are connected in a network, their model parameters are likely to be close to each other, satisfying the so-called smoothness property on a graph. The ℓ_1 graph Lasso formulation is computationally expensive to solve. The ℓ_2 graph Lasso, or the Laplacian Lasso, is an efficient alternative, which uses the following penalty:

$$\Omega_{\text{graph},\ell_2}(\mathbf{x}) = \alpha\|\mathbf{x}\|_1 + (1-\alpha)\mathbf{x}^{\mathrm{T}}L\mathbf{x}, \qquad (10)$$

where L is the Laplacian matrix [7; 20] constructed from the graph. It is known that the Laplacian matrix is positive semi-definite, and captures the underlying local geometric structure of the data. When L is an identity matrix, (10) reduces to the elastic net penalty [126]. Existing efficient algorithms for solving the Lasso can be applied to solve the ℓ_2 graph Lasso by grouping the loss term $L(\mathbf{x})$ and the Laplacian regularization $\lambda(1-\alpha)\mathbf{x}^{\mathrm{T}}L\mathbf{x}$ together, as the latter is both convex and differentiable.

Both ℓ_1 and ℓ_2 graph Lasso encourage positive correlation between the values of coefficients for the features connected by an edge in the graph. However, in certain applications, two features connected may be negatively correlated. To overcome this limitation, GFlasso employs a different ℓ_1 reg-

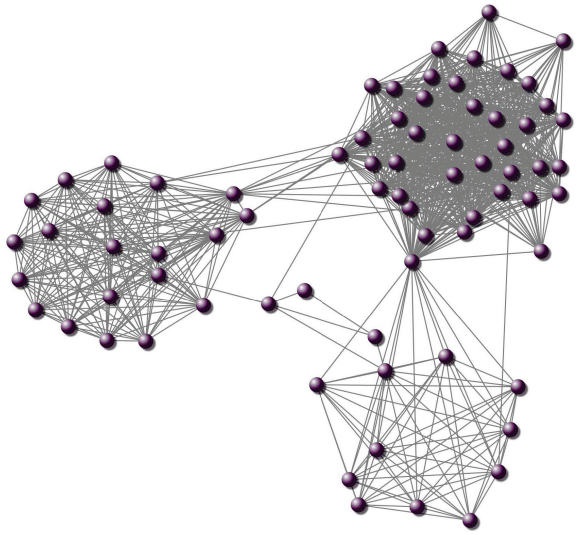

Figure 3: Illustration of a subgraph of the network consisting of 80 nodes.

ularization over a graph:

$$\Omega_{\text{GFlasso}}(\mathbf{x}) = \alpha\|\mathbf{x}\|_1 + (1-\alpha)\sum_{(i,j)\in E}|x_i - \text{sign}(r_{ij})x_j|, \tag{11}$$

where r_{ij} is the sample correlation between two features [50]. The penalty in (11) encourages the coefficients x_i, x_j for features i, j connected by an edge in the graph to be similar when $r_{ij} > 0$, but dissimilar when $r_{ij} < 0$. GFlasso would introduce additional estimation bias due to possible graph misspecification. For example, additional bias may occur when the sign of r_{ij} is inaccurate.

Another alternative is the so-called graph OSCAR (GOSCAR) penalty given by [110]:

$$\Omega_{\text{GOSCAR}}(\mathbf{x}) = \alpha\|\mathbf{x}\|_1 + (1-\alpha)\sum_{(i,j)\in E}\max\{|x_i|, |x_j|\}, \tag{12}$$

where a pairwise ℓ_∞ regularizer is used to encourage the coefficients to be equal [9], but the grouping constraints are imposed on the nodes connected over the given graph. The ℓ_1 regularizer encourages sparseness. The pairwise ℓ_∞ regularizer puts more penalty on the larger coefficients. Note that $\max\{|x_i|, |x_j|\}$ can be decomposed as

$$\max\{|x_i|, |x_j|\} = \frac{1}{2}(|x_i + x_j| + |x_i - x_j|).$$

The GOSCAR formulation is closely related to OSCAR [9]. The penalty of OSCAR is

$$\Omega_{\text{OSCAR}}(\mathbf{x})v = \alpha\|\mathbf{x}\|_1 + (1-\alpha)\sum_{i<j}\max\{|x_i|, |x_j|\}. \tag{13}$$

The ℓ_1 regularizer leads to a sparse solution, and the ℓ_∞ regularizer encourages the coefficients to be equal. OSCAR can be efficiently solved by accelerated gradient methods, whose key projection can be solved by a simple iterative group merging algorithm [121]. However, OSCAR assumes each node is connected to all the other nodes, which is not sufficient for many applications. Note that OSCAR is a special case of GOSCAR when the graph is complete. GOSCAR, incorporating an arbitrary undirected graph, is much more challenging to solve [110].

The penalty in GOSCAR overcomes the limitation of the Laplacian Lasso that the different signs of coefficients can introduce additional penalty. However, under the ℓ_∞ regularizer, even if $|x_i|$ and $|x_j|$ are close to each other, the penalty on this pair may still be large due to the property of the max operator, resulting in the coefficient x_i or x_j being over penalized. The additional penalty would result in biased estimation, especially for large coefficients, as in the Lasso case [98]. In GFlasso, when the pairwise sample correlation wrongly estimates the sign between x_i and x_j, an additional penalty on x_i and x_j would occur, introducing estimation bias. This motivates the following non-convex feature grouping and selection penalty:

$$\Omega_{\text{ncFGS}}(\mathbf{x}) = \alpha\|\mathbf{x}\|_1 + (1-\alpha)\sum_{(i,j)\in E}||x_i| - |x_j|| \tag{14}$$

which shrinks only small differences in absolutes values [110; 125]. As a result, estimation bias is reduced as compared to those convex grouping penalties. Note that the non-convex penalty does not assume the sign of an edge is given; it only relies on the graph structure.

3. SPARSE UNDIRECTED GAUSSIAN GRAPHICAL MODELS

Undirected graphical models are commonly used to describe and explain the relationships among a set of variables based on a collection of observations. In the Gaussian case, the graphical Lasso [29] is a popular approach for learning the structure in an undirected Gaussian graphical model [5]. The basic model for continuous data assumes that the observations have a multivariate Gaussian distribution with mean μ and covariance matrix Σ. If the ijth entry of $\Theta = \Sigma^{-1}$ is zero, then variables i and j are conditionally independent, given the other variables. Here, Θ is called the precision matrix. Thus, the problem of identifying the structure of the undirected Gaussian graphical model is equivalent to finding the nonzero entries of Θ. In [5], the ℓ_1 penalty is imposed on the precision matrix to increase its sparsity. The sparse undirected graphical model has been applied to construct biological networks [5] and brain networks [43].

3.1 Graphical Lasso

Suppose we have n samples independently drawn from a multivariate Gaussian distribution, and these samples are denoted as $\mathbf{y}_1, \cdots, \mathbf{y}_n \sim \mathcal{N}(\mu, \Sigma)$, where $\mathbf{y}_i$ is a p dimensional vector, $\mu \in \mathbb{R}^p$ is the mean, and $\Sigma \in \mathbb{R}^{p\times p}$ is the covariance matrix. Let $\Theta = \Sigma^{-1}$ be the inverse covariance matrix. The empirical mean is denoted as $\hat{\mu} = \frac{1}{n}\sum_{i=1}^n \mathbf{y}_i$, and the empirical covariance is denoted as S:

$$S = \frac{1}{n}\sum_{i=1}^n (\mathbf{y}_i - \hat{\mu})(\mathbf{y}_i - \hat{\mu})^T.$$

It can be shown that under a multivariate Gaussian model, the maximum likelihood estimate of $\Theta = \Sigma^{-1}$ can be obtained by solving the following maximization problem:

$$\max_{\Theta\succ 0} \log\det\Theta - \text{tr}\,(S\Theta)\,, \tag{15}$$

where $\text{tr}\,(S\Theta)$ is the trace of $S\Theta$, given by the summation of the diagonal entries of $S\Theta$. Assume that S is nonsingular. The maximum likelihood estimate of the inverse covariance Θ is $\Theta = S^{-1}$. If the dimensionality is larger than the sample size, i.e., $p > n$, S is singular. In such a case, regularization is commonly applied, and we estimate $\Theta = \Sigma^{-1}$ by maximizing the following objective function:

$$\log\det\Theta - \text{tr}\,(S\Theta) - \lambda J(\Theta), \tag{16}$$

where $J(\Theta)$ is a penalty function. The graphical Lasso employs the ℓ_1 penalty and solves the following optimization problem [5]:

$$\max_{\Theta\succ 0} \log\det\Theta - \text{tr}\,(S\Theta) - \lambda\|\Theta\|_1. \tag{17}$$

It is known that a larger value of λ leads to a sparser Θ that fits the data less well, while a smaller value of λ leads to a less sparse Θ that fits the data well. Thus, the choice of λ is an important issue in practical application of the graphical Lasso [63; 79].

Banerjee et al. [5] employed the interior point method to solve the optimization problem in (17). Friedman et al. [29] developed the graphical Lasso (GLasso) which applied the blockwise coordinate descent method to solve (17). The GLasso fails to converge with warm-starts. To resolve this issue, Mazumder and Hastie [76] proposed a new algorithm called DP-GLasso, each step of which is a box-constrained

QP problem. The main challenge of estimating a sparse precision matrix is its high computational complexity. Witten et al. [106] and Mazumder and Hastie [75] independently derived a screening rule, which dramatically reduced the computational cost especially for large regularization parameter values.

3.2 The Monotone Property

Huang et al. [43] derived the monotone property of the graphical Lasso. We first introduce the following definition.

DEFINITION 1. *In the graphical representation of the inverse covariance, if node i is connected to node j by an arc, then node i is called a "neighbor" of node j. If node i is connected to node k though some chain of arcs, then node i is called a "connectivity component" of node k.*

Intuitively, two nodes are neighbors if they are directly connected, whereas two nodes belong to the same connectivity component if they are indirectly connected, i.e., the connection is mediated through other nodes. In other words, if two nodes do not belong to the same connectivity component (i.e., two nodes completely separated in the graph), then they are completely independent of each other. Huang et al. [43] showed that the connectivity components have the following monotone property:

PROPOSITION 1. *Let $C_k(\lambda_1)$ and $C_k(\lambda_2)$ be the sets of all the connectivity components of node k with $\lambda = \lambda_1$ and $\lambda = \lambda_2$, respectively. If $\lambda_1 < \lambda_2$, then $C_k(\lambda_2) \subseteq C_k(\lambda_1)$.*

Intuitively, if two nodes are connected (either directly or indirectly) at one level of sparseness, they will be connected at all lower levels of sparseness. This monotone property can be used to identify how strongly connected each node k is to its connectivity components [43].

3.3 Simultaneous Estimation of Multiple Graphs

In some applications, we need to estimate multiple related precision matrices. A motivating example is the modeling of brain networks for Alzheimer's disease using neuroimaging data such as PET, in which, we want to estimate graphical models for three groups: normal controls (NC), patients of mild cognitive impairment (MCI), and Alzheimer's patients (AD). These graphs are expected to share some common connections, but they are not identical. Furthermore, the graphs are expected to evolve over time, in the order of disease severity from NC to MCI to AD. Estimating the graphical models separately fails to exploit the common structures among them. It is thus beneficial to jointly estimate the three graphs, especially when the number of subjects in each group is small. There is some recent work on the estimation of multiple precision matrices. Guo et al. [36] proposed to jointly estimate multiple graphical models using a hierarchical penalty. The time-varying graphical models were studied by Zhu et al. [122], and Kolar et al. [53; 54]. Danaher et al. [22] estimated multiple precision matrices simultaneously using a pairwise fused penalty and grouping penalty.

Assume we are given K data sets, $X^{(k)} \in \mathbb{R}^{n_k \times p}$, $k = 1, \cdots, K$ with $K \geq 2$, where n_k is the number of samples of the ith dataset, and p is the number of features. The p features are common for all K data sets, and all samples are independent. Furthermore, the samples within each data set $X^{(k)}$ are identically distributed with a p-variate Gaussian distribution with zero mean and covariance matrix $\Sigma^{(k)}$. We assume that there are many conditionally independent pairs of features, i.e., the precision matrix $\Theta^{(k)} = (\Sigma^{(k)})^{-1}$ is sparse. Denote the sample covariance matrix for each data set $X^{(k)}$ as $S^{(k)}$ and $\Theta = \{\Theta^{(1)}, \ldots, \Theta^{(K)}\}$. We can learn multiple precision matrices together by solving the following optimization problem [22; 109]:

$$\min_{\Theta^{(k)} \succ 0, k=1 \ldots K} \sum_{k=1}^{K} \Big(-\log\det(\Theta^{(k)}) + \mathrm{tr}(S^{(k)}\Theta^{(k)}) \Big) + P(\Theta), \tag{18}$$

where $\Theta^{(k)} = \left(\theta_{ij}^{(k)}\right)$,

$$P(\Theta) = \lambda_1 \sum_{k=1}^{K} \sum_{i \neq j} |\theta_{ij}^{(k)}| + \lambda_2 \sum_{k=1}^{K-1} \sum_{i \neq j} |\theta_{ij}^{(k)} - \theta_{ij}^{(k+1)}|,$$

and λ_1 and λ_2 are nonnegative regularization parameters. The ℓ_1 regularization leads to a sparse solution, and the fused penalty encourages $\Theta^{(k)}$ to be similar to its neighbors. The optimization in (18) is computationally expensive to solve. Danaher et al. [22] developed a screening rule for the two graph case to speed up the computation. The screening rule was recently extended to the more general case with more than two graphs in [109]. Specifically, Yang et al. [109] considered the problem of estimating multiple graphical models by maximizing a penalized log likelihood with ℓ_1 and fused regularization as in [22]. The ℓ_1 regularization yields a sparse solution, and the fused regularization encourages adjacent graphs to be similar. The blockwise coordinate descent method was employed to solve the fused multiple graphical Lasso (FMGL), where each step was solved by the accelerated gradient method [83]. In addition, a screening rule was developed which enabled the efficient estimation of multiple large precision matrices. Specifically, a set of necessary conditions were derived for the solution of FMGL to be block diagonal. These conditions were shown to be sufficient when $K \leq 3$. Yang et al. also performed extensive simulation studies; results indicate that these conditions are likely sufficient for any $K > 3$ as well.

4. PARALLEL MAGNETIC RESONANCE IMAGING

Magnetic resonance imaging (MRI) [39; 105] is a medical imaging technique used in radiology to visualize internal structures of the body in detail. As a non-invasive imaging technique, MRI makes use of the property of nuclear magnetic resonance to image nuclei of atoms inside the body. MRI has been applied to image the brain, muscles, the heart, cancers, etc.

4.1 Undersampled k-space

The acquired raw data by an MR scanner are the Fourier coefficients, or the so-called k-space data (see Figure 4 (a) for illustration). The k-space data are typically acquired by a series of phase encodings (each phase encoding covers a given amount of k-space data that are related to the trajectory, e.g., Cartesian sampling, radial sampling). For example, with Cartesian sampling, we need 256 frequency encodings to cover the full k-space of one 256×256 image. The time between the repetitions of the sequence is called the repetition time (TR) and it measures the time for acquiring one phase encoding. If TR=50 ms, it takes about

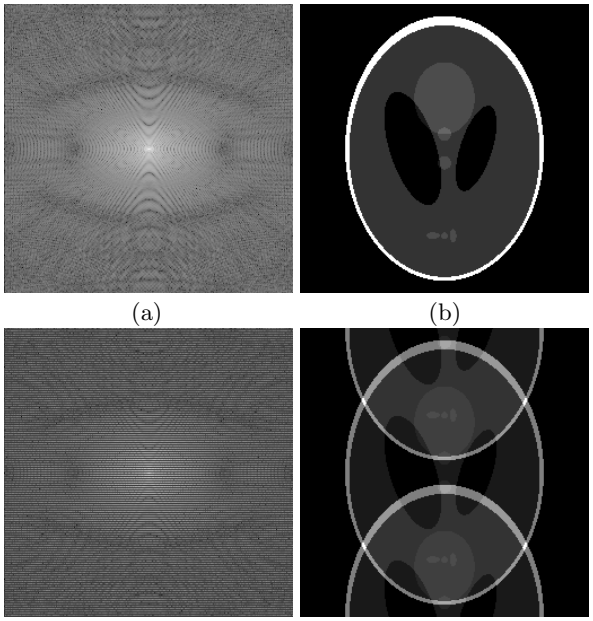

Figure 4: Illustration of MR image and the k-space data: (a) the full k-space data (displayed in logarithmic scale), (b) the image obtained by applying inverse Fourier transform to (a), (c) the undersampled k-space (displayed in logarithmic scale), and (d) the image obtained by applying inverse Fourier transform to (c).

12.8 seconds to acquire the full k-space data of one 256×256 image with the Cartesian trajectory. With the same TR, it takes about 15.4 minutes to acquire the full k-space of a $256 \times 256 \times 72$ volume. With higher spatial resolution, the time for acquiring the full k-space can be even longer. In addition, in dynamic cine imaging, we are interested in the study of the motion of the object (heart, blood, etc) over time. This leads to an increased number of phase encodings and increased acquisition time, and one usually has to compromise between spatial resolution and temporal resolution. To save the acquisition time, one has to undersample the k-space, i.e., reducing the number of acquired phase encodings. For example, if the k-space data are acquired every other line, as shown in Figure 4 (c), half of the acquisition time can be saved. The relationship between the acquired k-space data and the image to be reconstructed can be written as

$$\mathbf{y} = F_u \mathbf{f} + \mathbf{n}, \tag{19}$$

where F_u is a given undersampled Fourier transform operator, $\mathbf{f}$ denotes the MR image, $\mathbf{y}$ is the acquired k-space data, and $\mathbf{n}$ depicts the noise introduced in the acquisition. Unlike the full k-space scenario, one cannot directly apply the inverse Fourier transform to the undersampled data acquired in Figure 4 (c), since otherwise an aliased image shown in Figure 4 (d) will be obtained.

4.2 Parallel MR Imaging

Parallel imaging [34; 47; 88; 95] has been proven effective for reducing the acquisition time. It exploits the difference in sensitivities between individual coil elements in a receive array to reduce the number of gradient encodings required for imaging. Figure 5 illustrates parallel imaging with 8 coils. Specifically, the first two rows show the coil images seen by the individual coil/channel, and the last two rows show the coil profiles of these 8 coils. It can be observed that the 8 coils have different sensitivities. Parallel imaging tries to reconstruct the target image with the undersampled k-space data.

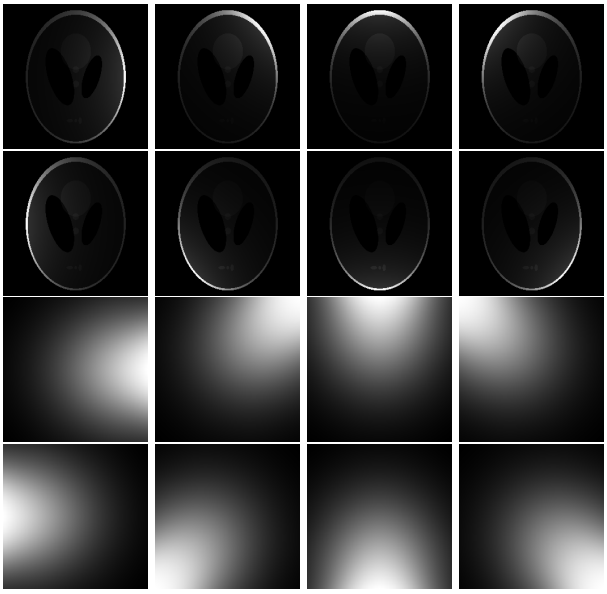

Figure 5: Illustration of the coil images and the coil sensitivity profiles (coil images of 8 channels are shown in the first two rows, and the corresponding coil profiles are shown in the last two rows).

Based on how the coil sensitivities are used, parallel imaging can be roughly divided into the following two main categories: 1) the approaches that implicitly make use of the coil sensitivities, represented by GRAPPA [34], and 2) the approaches that explicitly make use of the coil sensitivities, represented by SENSE [88]. In the GRAPPA type approaches, one usually estimates the missing phase encoding lines with the kernels that are estimated by implicitly using the coil sensitivities. In the SENSE type approach, one models the relationship between the target image and the acquired k-space data as:

$$\mathbf{y}_i = F_u S_i \mathbf{f} + \mathbf{n}_i, \tag{20}$$

where $\mathbf{y}_i$ is the acquired undersampled k-space data by the i-th coil, and S_i is the coil sensitivity maps (see the last two rows of Figure 5). The relationships between GRAPPA and SENSE have been studied in the literature [8; 35; 47], and several recent work [57; 58; 72; 73] have shown that GRAPPA and SENSE can be combined to give improved reconstruction performance.

4.3 Coil Profile Estimation

The most common way to determine the sensitivity maps is to obtain low-resolution pre-scans. However, when the object is not static, the sensitivity functions are different between pre-scan and under-sampled scans, and this could lead to reconstruction errors. To compensate for this, joint estimation approaches [103; 113] have been proposed. How-

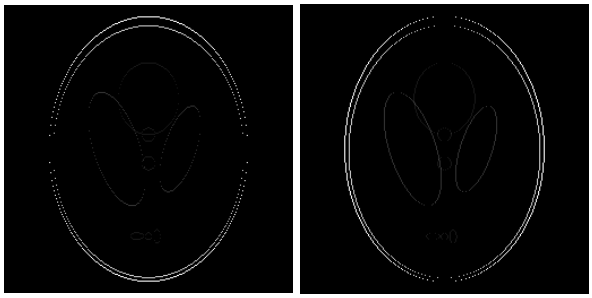

Figure 6: Illustration of the gradient of the phantom (shown in Figure 4) along the vertical direction (left) and horizonal direction (right), respectively.

ever, these approaches usually have high computation cost and are restricted to the SENSE type reconstruction.

The eigen-vector approach proposed in [72] is a very promising approach for sensitivity maps estimation. It tried to build a connection between GRAPPA and SENSE-type approaches, by showing that the Coil Profile used in SENSE can be computed with the GRAPPA-type calibration. Such idea was also used in [57; 58]. It was shown in [72] that the coil sensitivities can be computed as the eigen-vector of a given matrix in the image space corresponding to eigenvalues "1"s.

4.4 Sampling Pattern and Fourier Transform

Cartesian sampling is the most natural scheme which undersamples the k-space by skipping some lines. In cardiac MR imaging, TSENSE [37; 48] is a well-known approach that is based on time interleaving of k-space lines in sequential images, and there are studies that makes use of variable density to optimize the sampling scheme, e.g., [12]. The Fourier transform associated with the Cartesian sampling can be efficiently computed.

Spiral and projection (radial) are the most widely used non-Cartesian sampling patterns, among many others. It was observed in several works (e.g., [40]) that the radial sampling exhibits advantages over Cartesian Sampling. The Fourier transform in the non-Cartesian case is much more challenging than the Cartesian one, and gridding is usually employed for performing Non-Uniform FFT [33].

4.5 Incorporating Prior Knowledge and Optimization

To recover $\mathbf{f}$ from (19), it is important to note that our target $\mathbf{f}$ has certain structures, with which we can better reconstruct $\mathbf{f}$ from the undersampled data $\mathbf{y}$. This is where sparse learning can play a role. Typically, we are interested in computing $\mathbf{f}$ by solving the following problem

$$\min_{\mathbf{f}} \sum_i \text{loss}(\mathbf{y}, F_u S_i \mathbf{f}) + \lambda \phi(\mathbf{f}), \qquad (21)$$

where $\text{loss}(\mathbf{y}, F_u\mathbf{f})$ depicts the data fidelity, and $\phi(\mathbf{f})$ incorporates our prior knowledge about the image to be reconstructed.

For the data fidelity term, a commonly used one is the squared distance between the acquired data and the prediction: $\text{loss}(\mathbf{y}, F_u\mathbf{f}) = \frac{1}{2}\|\mathbf{y} - F_u S_i \mathbf{f}\|_2^2$. Recent studies have shown that the usage of self-consistency [57; 58; 73] can benefit reconstruction.

For $\phi(\mathbf{f})$, one needs to take advantage of the structure in the target image $\mathbf{f}$. Figure 6 shows the gradient of the phantom, and it is easy to observe that such gradient is sparse. Candès et al. [14] proposed to set $\phi(\mathbf{f}) = \|\mathbf{f}\|_{TV}$, showed the effectiveness of the sparsity promoting penalty in the scenario of single coil, and proved the exact recovery under the so-called Robust Uncertainty Principles (RIP). Later on, compressed sensing was used widely in the reconstruction of MR images, e.g., [1; 57; 61; 71; 112]. When applying sparse learning to parallel MR imaging, one key task is to develop a suitable $\phi(\cdot)$ that adapts the structure of the image(s) to be reconstructed. Group sparsity [117] has been used for accelerating dynamic MRI [104], and total variation and wavelet transformation have also been used for parallel MR imaging [14; 66; 67; 90; 103; 112]. An important and hot research topic is to develop better sparsity promoting penalties that adapt to the images to be reconstructed.

The efficient optimization of problem (21) is crucial for parallel imaging. Several popular approaches include conjugate gradient [40], Newton-type methods [103], Nesterov-type approaches [81; 82; 66; 49], and the alternating direction method of multipliers [1; 10; 31; 112].

5. CONCLUSIONS

In this paper, we review sparse methods for biomedical data in three specific applications. Sparse methods have also been applied to many other applications, e.g., incomplete multi-source data fusion [114] and biological image annotation and retrieval [115]. As with many other data mining and machine learning techniques, the selection of the appropriate sparse method and proper tuning of the associated parameters are critical for finding meaningful and useful results. To this end, one needs to understand the data in a domain specific context and understand the strengths and weaknesses of various sparse methods.

Most existing work on sparse learning focus on prediction, parameter estimation, and variable selection. Very few work address the problem of assigning statistical significance or confidence [11; 118]. However, such significance or confidence measures are crucial in biomedical applications where interpretation of parameters and variables is very important [11]. Most sparse methods in the literature are based on a convex regularizer. Sparse methods based on a non-convex regularizer have recently been proposed and efficient methods based on the difference of convex functions (DC) have been developed [92; 119]. However, their theoretical properties have not been well understood yet, although some recent work demonstrate the advantage of non-convex methods over their convex counterparts [92; 108; 119]. Finally, missing data is ubiquitous in biomedical applications. One important issue that has not been well addressed is how to adapt sparse methods to deal with missing data [70; 96].

6. ACKNOWLEDGEMENTS

This work was supported in part by NSF (IIS-0953662, MCB-1026710, CCF-1025177) and NIH (R01LM010730).

7. REFERENCES

[1] M. Afonso, J. Bioucas-Dias, and M. Figueiredo. An augmented lagrangian approach to the constrained optimization formulation of imaging inverse problems.

IEEE Transactions on Image Processing, 20:681–695, 2011.

[2] A. Argyriou, T. Evgeniou, and M. Pontil. Convex multi-task feature learning. *Machine Learning*, 73(3):243–272, 2008.

[3] F. R. Bach. Consistency of the group lasso and multiple kernel learning. *Journal of Machine Learning Research*, 9:1179–1225, 2008.

[4] W. Bajwa, J. Haupt, A. Sayeed, and R. Nowak. Compressive wireless sensing. In *International Conference on Information Processing in Sensor Networks*, 2006.

[5] O. Banerjee, L. El Ghaoui, and A. d'Aspremont. Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data. *The Journal of Machine Learning Research*, 9:485–516, 2008.

[6] R. Baraniuk. Compressive sensing. *IEEE Signal Processing Magazine*, 24(4):118–121, 2007.

[7] M. Belkin and P. Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, 15:1373–1396, 2003.

[8] M. Blaimer, F. Breuer, M. Muller, R. Heidemann, M. A. Griswold, and P. M. Jakob. SMASH, SENSE, PILS, GRAPPA. *Top Magn Reson Imaging*, 15:223–236, 2004.

[9] H. Bondell and B. Reich. Simultaneous regression shrinkage, variable selection, and supervised clustering of predictors with oscar. *Biometrics*, 64(1):115–123, 2008.

[10] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, 3:1–122, 2011.

[11] P. Bühlmann. Statistical significance in high-dimensional linear models. *Arxiv preprint arXiv:1202.1377v1*, 2012.

[12] R. Busse, K. Wang, J. Holmes, J. Brittain, and F. Korosec. Optimization of variable-density cartesian sampling for time-resolved imaging. In *International Society for Magnetic Resonance in Medicine*, 2009.

[13] E. Candès and J. Romberg. Quantitative robust uncertainty principles and optimally sparse decompositions. *Foundations of Computational Mathematics*, 6(2):227–254, 2006.

[14] E. Candès, J. Romberg, and T. Tao. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Transactions on Information Theory*, 52(2):489–509, 2006.

[15] E. Candès and T. Tao. Near optimal signal recovery from random projections: Universal encoding strategies? *IEEE Transactions on Information Theory*, 52(12):5406–5425, 2006.

[16] E. Candès and M. Wakin. An introduction to compressive sampling. *IEEE Signal Processing Magazine*, 25(2):21–30, 2008.

[17] S. Carroll, J. Grenier, and S. Weatherbee. *From DNA to Diversity: Molecular Genetics and the Evolution of Animal Design. 2nd edition.* Malden, MA: Blackwell Pub, 2005.

[18] W. Chu, Z. Ghahramani, F. Falciani, and D. Wild. Biomarker discovery in microarray gene expression data with gaussian processes. *Bioinformatics*, 21(16):3385–3393, 2005.

[19] H. Chuang, E. Lee, Y. Liu, D. Lee, and T. Ideker. Network-based classification of breast cancer metastasis. *Molecular systems biology*, 3:140, 2007.

[20] F. Chung. *Spectral Graph Theory*. American Mathematical Society, 1997.

[21] D. Cox. Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 34(2):187–220, 1972.

[22] P. Danaher, P. Wang, and D. Daniela. The joint graphical lasso for inverse covariance estimation across multiple classes. *Arxiv preprint arXiv:1111.0324*, 2011.

[23] A. d'Aspremont, L. El Ghaoui, M. Jordan, and G. R. G. Lanckriet. A direct formulation for sparse PCA using semidefinite programming. In *NIPS*, 2005.

[24] D. Donoho. High-dimensional data analysis: The curses and blessings of dimensionality. 2000.

[25] D. Donoho. Compressed sensing. *IEEE Transactions on Information Theory*, 52:1289–1306, 2006.

[26] M. Duarte, M. Davenport, M. Wakin, and R. Baraniuk. Sparse signal detection from incoherent projections. In *ICASSP*, 2006.

[27] J. Fan and J. Lv. A selective overview of variable selection in high dimensional feature space. *Statistica Sinica*, 20:101–148, 2010.

[28] J. Friedman, T. Hastie, H. Höfling, and R. Tibshirani. Pathwise coordinate optimization. *Annals of Applied Statistics*, 1(2):302–332, 2007.

[29] J. Friedman, T. Hastie, and R. Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.

[30] J. Friedman, T. Hastie, and R. Tibshirani. A note on the group lasso and a sparse group lasso. Technical report, Department of Statistics, Stanford University, 2010.

[31] T. Goldstein and S. Osher. The split bregman method for l1-regularized problems. *SIAM Journal on Imaging Sciences*, 2:323–343, 2009.

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

[33] L. Greengard and J. Lee. Accelerating the nonuniform fast fourier transform. *SIAM Review*, 46:443–454, 2004.

[34] M. A. Griswold, P. M. Jakob, R. M. Heidemann, M. Nittka, V. Jellus, J. Wang, K. B., and A. Haase. Generalized autocalibrating partially parallel acquisitions (GRAPPA). *Magnetic Resonance in Medicine*, 47:1202–1210, 2002.

[35] M. A. Griswold, S. Kannengiesser, R. M. Heidemann, J. Wang, and P. M. Jakob. Field-of-view limitations in parallel imaging. *Magnetic Resonance in Medicine*, 52:1118–1126, 2004.

[36] J. Guo, E. Levina, G. Michailidis, and J. Zhu. Joint estimation of multiple graphical models. *Biometrika*, 98(1):1–15, 2011.

[37] M. A. Guttman, P. Kellman, A. J. Dick, R. J. Lederman, and E. R. McVeigh. Real-time accelerated interactive MRI with adaptive TSENSE and UNFOLD. *Magnetic Resonance in Medicine*, 50:315–321, 2003.

[38] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik. Gene selection for cancer classification using support vector machines. *Machine Learning*, 46(1-3):389–422, 2002.

[39] E. M. Haacke, R. W. Brown, M. R. Thompson, and R. Venkatesan, editors. *Magnetic Resonance Imaging: Physical Principles and Sequence Design*. Wiley-Liss, 1999.

[40] M. S. Hansen, C. Baltes, J. Tsao, S. Kozerke, K. P. Pruessmann, and H. Eggers. k-t BLAST reconstruction from non-cartesian k-t space sampling. *Magnetic Resonance in Medicine*, 55:85–91, 2006.

[41] T. Hromádka, M. DeWeese, and A. Zador. Sparse representation of sounds in the unanesthetized auditory cortex. *PLoS Biol*, 6(1):e16, 2008.

[42] J. Huang, T. Zhang, and D. Metaxas. Learning with structured sparsity. *Journal of Machine Learning Research*, 12:3371–3412, 2011.

[43] S. Huang, J. Li, L. Sun, J. Liu, T. Wu, K. Chen, A. Fleisher, E. Reiman, and J. Ye. Learning brain connectivity of alzheimer's disease from neuroimaging data. In *NIPS*, pages 808–816, 2009.

[44] L. Jacob, G. Obozinski, and J. Vert. Group lasso with overlap and graph lasso. In *ICML*, 2009.

[45] R. Jenatton, J.-Y. Audibert, and F. Bach. Structured variable selection with sparsity-inducing norms. *Journal of Machine Learning Research*, 12:2777–2824, 2011.

[46] R. Jenatton, J. Mairal, G. Obozinski, and F. Bach. Proximal methods for sparse hierarchical dictionary learning. In *ICML*, 2010.

[47] P. Kellman. Parallel imaging: the basics. In *ISMRM Educational Course: MR Physics for Physicists*, 2004.

[48] P. Kellman, F. H. Epstein, and E. R. McVeigh. Adaptive sensitivity encoding incorporating temporal filtering (tsense). *Magnetic Resonance in Medicine*, 45:846–852, 2001.

[49] K. Khare, C. J. Hardy, K. F. King, P. A. Turski, and L. Marinelli. Accelerated MR imaging using compressive sensing with no free parameters. *Magnetic Resonance in Medicine*, 2012.

[50] S. Kim and E. Xing. Statistical estimation of correlated genome associations to a quantitative trait network. *PLoS genetics*, 5(8):e1000587, 2009.

[51] S. Kim and E. P. Xing. Tree-guided group lasso for multi-task regression with structured sparsity. In *ICML*, 2010.

[52] K. Koh, S. Kim, and S. Boyd. An interior-point method for large-scale l1-regularized logistic regression. *Journal of Machine Learning Research*, 8:1519–1555, 2007.

[53] M. Kolar, L. Song, A. Ahmed, and E. Xing. Estimating time-varying networks. *The Annals of Applied Statistics*, 4(1):94–123, 2010.

[54] M. Kolar and E. Xing. On time varying undirected graphs. In *AISTAT*, 2011.

[55] M. Kowalski. Sparse regression using mixed norms. *Applied and Computational Harmonic Analysis*, 27(3):303–324, 2009.

[56] S. Kumar, K. Jayaraman, S. Panchanathan, R. Gurunathan, A. Marti-Subirana, and S. Newfeld. BEST: A novel computational approach for comparing gene expression patterns from early stages of *Drosophila* melanogaster development. *Genetics*, 162(4):2037–2047, 2002.

[57] P. Lai, M. Lustig, B. A. C., V. S. S., B. P. J., and A. M. Efficient L1SPIRiT reconstruction (ESPIRiT) for highly accelerated 3d volumetric MRI with parallel imaging and compressed sensing. In *ISMRM*, 2010.

[58] P. Lai, M. Lustig, V. S. S., and B. A. C. ESPIRiT (efficient eigenvector-based l1spirit) for compressed sensing parallel imaging - theoretical interpretation and improved robustness for overlapped FOV prescription. In *ISMRM*, 2011.

[59] D. J. Larkman and R. G. Nunes. Parallel magnetic resonance imaging. *Physics in Medicine and Biology*, 52:R15–55, 2007.

[60] C. Li and H. Li. Network-constrained regularization and variable selection for analysis of genomic data. *Bioinformatics*, 24(9):1175–1182, 2008.

[61] D. Liang, B. Liu, J. Wang, and L. Ying. Accelerating SENSE using compressed sensing. *Magnetic Resonance in Medicine*, 62:1574–1584, 2009.

[62] H. Liu, M. Palatucci, and J. Zhang. Blockwise coordinate descent procedures for the multi-task lasso, with applications to neural semantic basis discovery. In *ICML*, 2009.

[63] H. Liu, K. Roeder, and L. Wasserman. Stability approach to regularization selection (StARS) for high dimensional graphical models. In *NIPS*, 2011.

[64] J. Liu, S. Ji, and J. Ye. Multi-task feature learning via efficient $\ell_{2,1}$-norm minimization. In *UAI*, 2009.

[65] J. Liu, S. Ji, and J. Ye. *SLEP: Sparse Learning with Efficient Projections*. Arizona State University, 2009.

[66] J. Liu, J. Rapin, T. Chang, A. Lefebvre, M. Zenge, E. Mueller, and M. S. Nadar. Dynamic cardiac MRI reconstruction with weighted redundant haar wavelets. In *ISMRM*, 2012.

[67] J. Liu, J. Rapin, T. Chang, P. Schmitt, X. Bi, A. Lefebvre, M. Zenge, E. Mueller, and M. S. Nadar. Regularized reconstruction using redundant haar wavelets: A means to achieve high undersampling factors in non-contrast-enhanced 4D MRA. In *ISMRM*, 2012.

[68] J. Liu and J. Ye. Efficient ℓ_1/ℓ_q norm regularization. *Arxiv preprint arXiv:1009.4766v1*, 2010.

[69] J. Liu and J. Ye. Moreau-Yosida regularization for grouped tree structure learning. In *NIPS*, 2010.

[70] P. Loh and M. Wainwright. High-dimension regression with noisy and missing data: Provable guarantees with non-convexity. In *NIPS*, 2011.

[71] M. Lustig, D. L. Donoho, and J. M. Pauly. Sparse MRI: The application of compressed sensing for rapid MR imaging. *Magnetic Resonance in Medicine*, 58:1182–1195, 2007.

[72] M. Lustig, P. Lai, M. Murphy, S. Vasanawala, M. Elad, J. Zhang, and J. Pauly. An eigen-vector approach to autocalibrating parallel MRI, where SENSE meets GRAPPA. In *ISMRM*, 2011.

[73] M. Lustig and J. M. Pauly. SPIRiT: Iterative self-consistent parallel imaging reconstruction from arbitrary k-space. *Magnetic Resonance in Medicine*, 64:457–471, 2010.

[74] M. Marton et al. Drug target validation and identification of secondary drug target effects using DNA microarrays. *Nature Medicine*, 4(11):1293–1301, 1998.

[75] R. Mazumder and T. Hastie. Exact covariance thresholding into connected components for large-scale graphical lasso. *Arxiv preprint arXiv:1108.3829*, 2011.

[76] R. Mazumder and T. Hastie. The graphical lasso: New insights and alternatives. *Arxiv preprint arXiv:1111.5479*, 2011.

[77] S. Megason and S. Fraser. Imaging in systems biology. *Cell*, 130(5):784–795, 2007.

[78] L. Meier, S. Geer, and P. Bühlmann. The group lasso for logistic regression. *Journal of the Royal Statistical Society: Series B*, 70:53–71, 2008.

[79] N. Meinshausen and P. Bühlmann. Stability selection. *Journal of the Royal Statistical Society: Series B*, 72:417–473, 2010.

[80] S. Negahban and M. Wainwright. Joint support recovery under high-dimensional scaling: Benefits and perils of $\ell_{1,\infty}$-regularization. In *NIPS*, pages 1161–1168. 2008.

[81] A. Nemirovski. *Efficient methods in convex programming*. Lecture Notes, 1994.

[82] Y. Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*. Kluwer Academic Publishers, 2004.

[83] Y. Nesterov. *Gradient methods for minimizing composite objective function*. CORE, 2007.

[84] A. Ng. Feature selection, ℓ_1 vs. ℓ_2 regularization, and rotational invariance. In *ICML*, 2004.

[85] G. Obozinski, B. Taskar, and M. I. Jordan. Joint covariate selection for grouped classification. Technical report, Statistics Department, UC Berkeley, 2007.

[86] H. Peng. Bioimage informatics: a new area of engineering biology. *Bioinformatics*, 24(17):1827–1836, 2008.

[87] J. Peng, J. Zhu, B. A., W. Han, D.-Y. Noh, J. R. Pollack, and P. Wang. Regularized multivariate regression for identifying master predictors with application to integrative genomics study of breast cancer. *Annals of Applied Statistics*, 4(1):53–77, 2010.

[88] K. Pruessmann, M. Weiger, M. Scheidegger, and P. Boesiger. SENSE: sensitivity endcoding for fast MRI. *Magnetic Resonance in Medicine*, 42:952–962, 1999.

[89] A. Quattoni, X. Carreras, M. Collins, and T. Darrell. An efficient projection for $\ell_{1,\infty}$ regularization. In *ICML*, 2009.

[90] S. Ramani and J. A. Fessler. Parallel MR image reconstruction using augmented lagrangian methods. *IEEE Transactions on Medical Imaging*, 30:694–706, 2011.

[91] S. Ryali, K. Supekar, D. Abrams, and V. Menon. Sparse logistic regression for whole-brain classification of fMRI data. *Neuroimage*, 51(2):752–764, 2010.

[92] X. Shen, W. Pan, and Y. Zhu. Likelihood-based selection and sharp parameter estimation. *Journal of American Statistical Association*, 107:223–232, 2012.

[93] J. Shi, W. Yin, S. Osher, and P. Sajda. A fast algorithm for large scale ℓ_1-regularized logistic regression. Technical report, CAAM TR08-07, 2008.

[94] W. Shi, K. Lee, and G. Wahba. Detecting disease-causing genes by lasso-patternsearch algorithm. *BMC Proceedings*, 1(Suppl 1):S60, 2007.

[95] D. Sodickson and W. Manning. Simultaneous acquisition of spatial harmonics (SMASH): fast imaging with radiofrequency coil arrays. *Magnetic Resonance in Medicine*, 38:591–603, 1997.

[96] N. Städler and P. Bühlmann. Missing values: sparse inverse covariance estimation and an extension to sparse regression. *Statistics and Computing*, 22:219–235, 2012.

[97] 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*, 102(43):15545–15550, 2005.

[98] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B*, 58(1):267–288, 1996.

[99] R. Tibshirani, M. Saunders, S. Rosset, J. Zhu, and K. Knight. Sparsity and smoothness via the fused lasso. *Journal Of The Royal Statistical Society Series B*, 67(1):91–108, 2005.

[100] R. Tibshirani and P. Wang. Spatial smoothing and hot spot detection for cgh data using the fused lasso. *Biostatistics*, 9(1):18–29, 2008.

[101] P. Tomancak, A. Beaton, R. Weiszmann, E. Kwan, S. Shu, S. E. Lewis, S. Richards, M. Ashburner, V. Hartenstein, S. E. Celniker, and G. M. Rubin. Systematic determination of patterns of gene expression during Drosophila embryogenesis. *Genome Biology*, 3(12):research0088.1–14, 2002.

[102] A. Tropp, A. Gilbert, and M. Strauss. Algorithms for simultaneous sparse approximation: part I: Greedy pursuit. *Signal Processing*, 86(3):572–588, 2006.

[103] M. Uecker, T. Hohage, K. T. Block, and J. Frahm. Image reconstruction by regularized nonlinear inversion - joint estimation of coil sensitivities and image content. *Magnetic Resonance in Medicine*, 60:674–682, 2008.

[104] M. Usman, C. Prieto, T. Schaeffter, and P. G. Batchelor. k-t group sparse: A method for accelerating dynamic MRI. *Magnetic Resonance in Medicine*, 66:1163–1176, 2011.

[105] M. T. Vlaardingerbroek and J. A. Boer, editors. *Magnetic Resonance Imaging*. Spinger, 2004.

[106] D. Witten, J. Friedman, and N. Simon. New insights and faster computations for the graphical lasso. *Journal of Computational and Graphical Statistics*, 20(4):892–900, 2011.

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

[108] S. Xiang, X. Shen, and J. Ye. Efficient sparse group feature selection via nonconvex optimization. *Arxiv preprint arXiv:1205.5075*, 2012.

[109] S. Yang, Z. Pan, X. Shen, P. Wonka, and J. Ye. Fused multiple graphical lasso. *Technical Report, Arizona State University*, 2012.

[110] S. Yang, L. Yuan, Y.-C. Lai, X. Shen, P. Wonka, and J. Ye. Feature grouping and selection over an undirected graph. In *KDD*, 2012.

[111] J. Ye, M. Farnum, E. Yang, R. Verbeeck, V. Lobanov, N. Raghavan, G. Novak, A. DiBernardo, and V. Narayan. Sparse learning and stability selection for predicting MCI to AD conversion using baseline ADNI data. *BMC Neurology*, 2012.

[112] X. Ye, Y. Chen, and F. Huang. Computational acceleration for MR image reconstruction in partially parallel imaging. *IEEE Transactions on Medical Imaging*, 30:1055–1063, 2011.

[113] L. Ying and J. Sheng. Joint image reconstruction and sensitivity estimation in sense (JSENSE). *Magnetic Resonance in Medicine*, 57:1196–1202, 2007.

[114] L. Yuan, Y. Wang, P. Thompson, V. Narayand, and J. Ye. Multi-source feature learning for joint analysis of incomplete multiple heterogeneous neuroimaging data. *NeuroImage*, 61(3):622–632, 2012.

[115] L. Yuan, A. Woodard, S. Ji, Y. Jiang, Z.-H. Zhou, S. Kumar, and J. Ye. Learning sparse representations for fruit-fly gene expression pattern image annotation and retrieval. *BMC Bioinformatics*, 13:107, 2012.

[116] M. Yuan, V. R. Joseph, and H. Zou. Structured variable selection and estimation. *Annals of Applied Statistics*, 3:1738–1757, 2009.

[117] M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal Of The Royal Statistical Society Series B*, 68(1):49–67, 2006.

[118] C.-H. Zhang and S. Zhang. Confidence intervals for low-dimensional parameters with high-dimensional data. *Arxiv preprint arXiv:1110.2563v1*, 2011.

[119] T. Zhang. Analysis of multi-stage convex relaxation for sparse regularization. *Journal of Machine Learning Research*, 11:1081–1107, 2010.

[120] P. Zhao, G. Rocha, and B. Yu. The composite absolute penalties family for grouped and hierarchical variable selection. *Annals of Statistics*, 37(6A):3468–3497, 2009.

[121] L. Zhong and J. Kwok. Efficient sparse modeling with automatic feature grouping. *ICML*, 2011.

[122] S. Zhou, J. Lafferty, and L. Wasserman. Time varying undirected graphs. *COLT*, 2008.

[123] J. Zhu and T. Hastie. Classification of gene microarrays by penalized logistic regression. *Biostatistics*, 5(3):427–443, 2004.

[124] J. Zhu, S. Rosset, T. Hastie, and R. Tibshirani. 1-norm support vector machines. In *Neural Information Processing Systems*, pages 49–56, 2003.

[125] Y. Zhu, X. Shen, and W. Pan. Simultaneous grouping pursuit and feature selection in regression over an undirected graph. *Preprint*, 2012.

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

[127] H. Zou, T. Hastie, and R. Tibshirani. Sparse principle component analysis. *Journal of Computational and Graphical Statistics*, 15(2):262–286, 2006.