Discriminative Feature Grouping

Lei Han[1] **and Yu Zhang**[1,2*]
[1]Department of Computer Science, Hong Kong Baptist University, Hong Kong
[2]The Institute of Research and Continuing Education, Hong Kong Baptist University (Shenzhen)

Abstract

Feature grouping has been demonstrated to be promising in learning with high-dimensional data. It helps reduce the variances in the estimation and improves the stability of feature selection. One major limitation of existing feature grouping approaches is that some similar but different feature groups are often mis-fused, leading to impaired performance. In this paper, we propose a Discriminative Feature Grouping (DFG) method to discover the feature groups with enhanced discrimination. Different from existing methods, DFG adopts a novel regularizer for the feature coefficients to trade-off between fusing and discriminating feature groups. The proposed regularizer consists of a ℓ_1 norm to enforce feature sparsity and a pairwise ℓ_∞ norm to encourage the absolute differences among any three feature coefficients to be similar. To achieve better asymptotic property, we generalize the proposed regularizer to an adaptive one where the feature coefficients are weighted based on the solution of some estimator with root-n consistency. For optimization, we employ the alternating direction method of multipliers to solve the proposed methods efficiently. Experimental results on synthetic and real-world datasets demonstrate that the proposed methods have good performance compared with the state-of-the-art feature grouping methods.

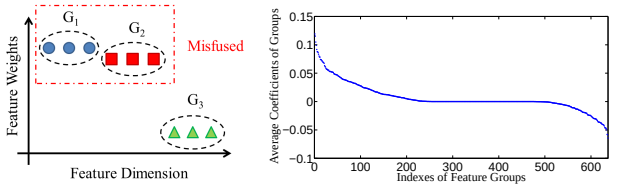
Introduction

Learning with high-dimensional data is a challenge especially when the size of the data is not very large. Sparse modeling, which selects only a relevant subset of the features, has thus received increasing attention. Lasso (Tibshirani 1996) is one of the most popular sparse modeling methods and has been well studied in the literature. However, in the presence of highly correlated features, Lasso tends to select only one or some of those features, leading to unstable estimations and impaired performance. To address this issue, the group lasso (Yuan and Lin 2006) has been proposed to select groups of features by using the ℓ_1/ℓ_2 regularizer. As extensions of the group lasso, several methods are proposed to learn from overlapping groups (Zhao, Rocha, and Yu 2009; Jacob, Obozinski, and Vert 2009; Yuan, Liu, and Ye 2011). Other extensions of the group lasso, e.g., (Kim and Xing 2010; Jenatton et al. 2010), aim to learn from the given tree structured information among features. However, those methods require the feature groups to be given as a priori information. That is, they can utilize the given feature groups to obtain solutions with group sparsity, but lack the ability of learning the feature groups.

Feature grouping techniques, which find the groups of highly correlated features automatically from data, thus have been proposed to address this issue. These techniques help gain additional insights to understand and interpret data, e.g., finding co-regulated genes in microarray analysis (Dettling and Bühlmann 2004). Feature grouping techniques assume that the features with identical coefficients form a feature group. The elastic net (Zou and Hastie 2005) is a representative feature grouping approach, which combines the ℓ_1 and ℓ_2 norms to encourage highly correlated features to have identical coefficients. The fused Lasso family, including the fused Lasso (Tibshirani et al. 2005), graph based fused Lasso (Kim and Xing 2009), and generalized fused Lasso (GFLasso) (Friedman et al. 2007), uses some fused regularizers to directly force the feature coefficients of each pair of features to be close based on the ℓ_1 norm. Recently, the OSCAR method (Bondell and Reich 2008), which combines a ℓ_1 norm and a pairwise ℓ_∞ norm on each pair of features, has shown good performance in learning feature groups. Moreover, some extensions of OSCAR have also been proposed (Shen and Huang 2010; Yang et al. 2012; Jang et al. 2013) to further reduce the estimation bias.

However, when there exist some similar but still different feature groups, we find that empirically all the existing feature grouping methods tend to fuse those groups together as one group, thus leading to impaired learning performance. Figure 1(a) shows an example, where G_1 and G_2 are similar but different feature groups, and they are easy to be mis-fused by existing feature grouping methods. In many real-world applications with high-dimensional data, e.g., microarray analysis, the phenomena that feature groups with similar but different feature coefficients appear frequently. For example, by using the method in (Jacob, Obozinski, and Vert 2009), the averaged coefficients of each feature group among the given 637 groups, which correspond to the biological gene pathways, in the breast cancer data is shown in Figure 1(b) and we can observe that there are a lot of feature

*Both authors contribute equally.



(a) (b)

Figure 1: (a) The misfusion problem; (b) A study of the averaged feature coefficients of the groups in microarray data.

groups with similar but different (averaged) feature coefficients. This problem is also found in some other real-world applications.

In order to solve the aforementioned problem in existing feature grouping methods, we propose a Discriminative Feature Grouping (DFG) method to not only discover feature groups but also discriminate similar feature groups. The DFG method proposes a novel regularizer on the feature coefficients to trade-off between fusing and discriminating feature groups. The proposed regularizer consists of a ℓ_1 norm to enforce feature sparsity and a pairwise ℓ_∞ norm to encourage $|\beta_i - \beta_j|$ and $|\beta_i - \beta_k|$ to be identical for any three feature coefficients β_i, β_j and β_k. As analyzed, the pairwise ℓ_∞ regularizer is capable of both grouping features and discriminating similar feature groups. Moreover, to achieve better asymptotic property, we propose an adaptive version of DFG, the ADFG method, in which the feature coefficients are weighted based on the solution of some estimator with root-n consistency. For optimization, we employ the alternating direction method of multipliers (ADMM) (Boyd et al. 2011) to solve the proposed methods efficiently. For analysis, we study the asymptotic properties of the DFG and ADFG models, where the feature groups obtained by the ADFG method can recover the ground truth with high probability. Experimental results conducted on synthetic and real-world datasets demonstrate that the proposed methods are competitive compared with existing feature grouping techniques.

Notations: Let $\mathbf{X} \in \mathbb{R}^{n\times d}$ be the predictor matrix or the data matrix and $\mathbf{y} \in \mathbb{R}^n$ be the responses or labels, where n is the number of samples and d is the number of features. For any vector $\mathbf{x}$, $\|\mathbf{x}\|_p$ denotes its ℓ_p-norm. $|\mathcal{A}|$ denotes the cardinality of a set $\mathcal{A}$.

Background

In this section, we briefly overview some existing feature grouping techniques.

As a representative of the fused Lasso family, the GFLasso method solves the following optimization problem:

$$\min_{\boldsymbol{\beta}} L(\boldsymbol{\beta}) + \lambda_1\|\boldsymbol{\beta}\|_1 + \lambda_2 \sum_{i<j}^{d} |\beta_i - \beta_j|, \tag{1}$$

where $L(\cdot)$ denotes the loss function and λ_1 and λ_2 are regularization parameters. We use the square loss $L(\boldsymbol{\beta}) = \frac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2$ in this paper. In problem (1), the ℓ_1 norm encourages the sparsity in $\boldsymbol{\beta}$ and the fusion term (i.e., the last term) enforces any two coefficients β_i and β_j to be identical, which is a way to discover feature groups.

The OSCAR method proposes a pairwise ℓ_∞ regularizer and solves the following optimization problem:

$$\min_{\boldsymbol{\beta}} L(\boldsymbol{\beta}) + \lambda_1\|\boldsymbol{\beta}\|_1 + \lambda_2 \sum_{i<j}^{d} \max\{|\beta_i|, |\beta_j|\}. \tag{2}$$

The pairwise ℓ_∞ regularizer encourages the absolute values of every two coefficients $|\beta_i|$ and $|\beta_j|$ to be identical. Based on OSCAR, some non-convex extensions, e.g., the ncFGS and ncTFGS methods (Yang et al. 2012), are proposed to further reduce the estimation bias. The objective functions of the ncFGS and ncTFGS methods are formulated as

$$\min_{\boldsymbol{\beta}} L(\boldsymbol{\beta}) + \lambda_1\|\boldsymbol{\beta}\|_1 + \lambda_2 \sum_{i<j}^{d} \Big||\beta_i| - |\beta_j|\Big|, \tag{3}$$

$$\min_{\boldsymbol{\beta}} L(\boldsymbol{\beta}) + \lambda_1 \sum_{i}^{d} J_\tau(|\beta_i|) + \lambda_2 \sum_{i<j}^{d} J_\tau(||\beta_i| - |\beta_j||), \tag{4}$$

where $J_\tau(x) = \min(\frac{x}{\tau}, 1)$ and τ is a threshold. When $\tau \to \infty$, problem (4) reduces to problem (3). The third terms in both problems encourage any pair of coefficients to be similar. Note that when $\lambda_1 \geq \frac{d-1}{2}\lambda_2$, problem (3) reduces to problem (2) since $\max(|x|, |y|) = \frac{1}{2}(|x| + |y| + ||x| - |y||)$.

Discriminative Feature Grouping

In this section, we introduce the DFG method and its adaptive extension, the ADFG method. Moreover, we also discuss how to incorporate some additional information into our proposed methods.

DFG Method

Problems defined in Eqs. (1-4) impose a fusion-like regularizer for any two feature coefficients β_i and β_j, where features with identical coefficients are assumed to be from the same feature group. One major limitation of those existing feature grouping methods is that some similar but different feature groups are easy to be mis-fused. To address the issue, the DFG method is proposed with the objective function formulated as

$$\min_{\boldsymbol{\beta}} L(\boldsymbol{\beta}) + \lambda_1\|\boldsymbol{\beta}\|_1 + \lambda_2 \sum_{i=1}^{d} \sum_{\substack{j<k \\ j,k\neq i}}^{d} \max\{|\beta_i - \beta_j|, |\beta_i - \beta_k|\}, \tag{5}$$

where λ_1 and λ_2 are positive regularization parameters. We denote the third term in problem (5) as $\Omega_{GD}(\boldsymbol{\beta})$. The ℓ_1 regularizer (i.e., the second term in problem (5)) encourages feature sparsity and the pairwise ℓ_∞ regularizer in $\Omega_{GD}(\cdot)$ encourages $|\beta_i - \beta_j|$ and $|\beta_i - \beta_k|$ to be identical for any triple of feature indices (i, j, k). Note that $\max\{|\beta_i - \beta_j|, |\beta_i - \beta_k|\}$ can be reformulated as

$$\max\{|\beta_i - \beta_j|, |\beta_i - \beta_k|\} = \frac{1}{2}|\beta_j - \beta_k| + |\beta_i - \frac{\beta_j + \beta_k}{2}|. \tag{6}$$

Then we can see two effects of $\Omega_{GD}(\boldsymbol{\beta})$: (1) the first term in the right-hand side of Eq. (6) is the fusion regularizer to enforce β_j and β_k to be grouped similar to the fused Lasso

family (regardless of β_i), which reflects the grouping property; (2) the second term encourages β_i to approach the average of β_j and β_k, making β_i, β_j and β_k stay discriminative unless all the three coefficients become identical, which is the discriminating effect. Therefore, the regularizer $\Omega_{GD}(\boldsymbol{\beta})$ not only groups the features in a similar way to the fused Lasso but also discriminates the similar groups.

ADFG Method

The first and second terms at the right-hand side of Eq. (6) are denoted by $\Omega_G(\cdot)$ and $\Omega_D(\cdot)$ respectively. $\Omega_D(\cdot)$ encourages one feature coefficient to approach the average of another two feature coefficients, which seems too restrictive to model the relations between different feature groups. To capture more flexible relations between groups, we propose an adaptive version of the DFG method, the ADFG method, with a new regularizer corresponding to $\Omega_{GD}(\boldsymbol{\beta})$ defined as

$$\Omega_{GD}^{Ad}(\boldsymbol{\beta}) = \Omega_G^{Ad}(\boldsymbol{\beta}) + \Omega_D^{Ad}(\boldsymbol{\beta}), \tag{7}$$

where the adaptive grouping regularizer $\Omega_G^{Ad}(\cdot)$ and the adaptive discriminating one $\Omega_D^{Ad}(\cdot)$ are defined as

$$\Omega_G^{Ad}(\boldsymbol{\beta}) = \lambda_2 \sum_{j<k}^{d} w_{jk}|\beta_j - \beta_k|,$$

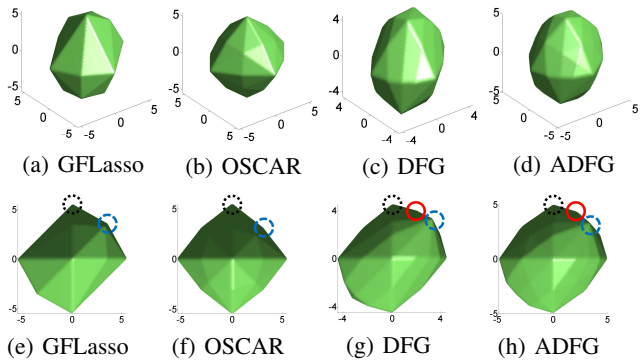
$$\Omega_D^{Ad}(\boldsymbol{\beta}) = \lambda_3 \sum_{i=1}^{d} \sum_{\substack{j<k \\ j,k\neq i}}^{d} w_{ijk}|\beta_i - \alpha_{ijk}\beta_j - (1-\alpha_{ijk})\beta_k|,$$

where w_{ij} is a weight based on an initial estimator $\widehat{\boldsymbol{\beta}}$, i.e., $w_{ij} = |\widehat{\beta}_i - \widehat{\beta}_j|^{-\gamma}$, with γ as a positive constant, $w_{ijk} = w_{ij} + w_{ik}$, and $\alpha_{ijk} = \frac{w_{ij}}{w_{ijk}}$. To achieve good theoretic property as we will see later, $\widehat{\boldsymbol{\beta}}$ is supposed to be the solution of some estimator with root-n consistency, e.g., the ordinary least square estimator which is adopted in our implementation. The larger γ, the more trust the initial estimator $\widehat{\boldsymbol{\beta}}$ gains. The ADFG method can be viewed as a generalization of the DFG method since when $\gamma = 0$ and $\lambda_2 = (d-2)\lambda_3$, the ADFG method reduces to the DFG method. Moreover, to make the ADFG method flexible, we use different regularization parameters λ_2 and λ_3 to weight the grouping and discriminating parts separately. To keep accordance with Ω_{GD}^{Ad}, we also adopt the adaptive version for the ℓ_1 norm, which is denoted by $\Omega_{\ell_1}^{Ad}(\boldsymbol{\beta}) = \lambda_1 \sum_{i=1}^{d} w_i|\beta_i|$ with $w_i = |\widehat{\beta}_i|^{-\gamma}$, in Eq. (5) as in (Zou 2006).

In order to better understand the regularizer $\Omega_{GD}^{Ad}(\boldsymbol{\beta})$, we see that for the discriminating part $\Omega_D^{Ad}(\cdot)$, if $|\beta_i - \alpha_{ijk}\beta_j - (1-\alpha_{ijk})\beta_k| = 0$, and β_j and β_k are close but still different, β_i is different from β_j and β_k since $\alpha_{ijk} \in (0,1)$. For the grouping part $\Omega_G^{Ad}(\cdot)$, it is similar to the adaptive generalized fused lasso regularizer introduced in (Viallon et al. 2013).

Remark 1 *$\Omega_D^{Ad}(\cdot)$ can also be viewed as a regularizer to capture the feature relations by encouraging a linear relationship among any β_i, β_j, and β_k based on the trust of an initial estimator. Recall that $\Omega_G^{Ad}(\cdot)$ generates feature groups. Therefore, $\Omega_{GD}^{Ad}(\cdot)$ can leverage between feature grouping and maintaining the feature relations.*

Remark 2 *Figure 2 provides illustrations for different regularizers in a ball $\mathcal{R}(\boldsymbol{\beta}) \leq 10$ for different methods, where $\mathcal{R}(\cdot)$ is the corresponding regularizer. Since the regularizers in the ncFGS and ncTFGS are similar to that of the OSCAR, they are omitted. In Figures 2(e)-2(h), similar to (Bondell and Reich 2008), the optimal solutions are more likely to hit the sharp points, where sparse solutions appear at the black tiny-dashed circles, feature fusions occur at the blue dashed circles, and features keep discriminative at the red solid circles. We then observe that only the DFG and ADFG methods have both the grouping and discriminating effects.*



(a) GFLasso (b) OSCAR (c) DFG (d) ADFG

(e) GFLasso (f) OSCAR (g) DFG (h) ADFG

Figure 2: Pictorial representations of the regularizers in the ball $\mathcal{R}(\boldsymbol{\beta}) \leq 10$ with $\boldsymbol{\beta} = [\beta_1, \beta_2, \beta_3]^T \in \mathbb{R}^3$: (a) GFLasso ($\lambda_1 = 1$, $\lambda_2 = 0.4$); (b) OSCAR ($\lambda_1 = 1$, $\lambda_2 = 0.4$); (c) DFG ($\lambda_1 = 1$, $\lambda_2 = 0.4$); (d) ADFG ($\lambda_1 = 1$, $\lambda_2 = 0.1$, $\lambda_3 = 0.2$); (e)-(h) the corresponding projections onto the β_1-β_2 plane.

Remark 3 *In addition, the solution paths for different regularizers in the orthogonal case can reveal the properties of the proposed regularizers from another perspective.*

Incorporating Graph Information

Similar to (Yang et al. 2012), some a priori information can be easily incorporated into the proposed DFG and ADFG methods. For example, when the feature correlations are encoded in a given graph, the Ω_{GD} regularizer in DFG can be adapted to

$$\Omega_{GD}(\boldsymbol{\beta}) = \lambda_2 \sum_{\substack{(i,j)\in E \\ (i,k)\in E}} \max\{|\beta_i - \beta_j|, |\beta_i - \beta_k|\}, \tag{8}$$

where a graph $G = (V, E)$ encodes the correlations between pairs of features into the set of edges E. Similar formulations can be derived for Ω_{GD}^{Ad} and are omitted here.

Optimization Procedure

It is easy to show that the objective functions of both the DFG and ADFG methods are convex. We propose to solve the ADFG method using the ADMM, and the same optimization procedure is applicable to the DFG method since the DFG method is a special case of the ADFG method. Note that $\Omega_{\ell_1}^{Ad}(\boldsymbol{\beta})$, $\Omega_G^{Ad}(\boldsymbol{\beta})$ and $\Omega_D^{Ad}(\boldsymbol{\beta})$ can be reformulated as $\Omega_{\ell_1}^{Ad}(\boldsymbol{\beta}) = \|\mathcal{T}_1\boldsymbol{\beta}\|_1$, $\Omega_G^{Ad}(\boldsymbol{\beta}) = \|\mathcal{T}_2\boldsymbol{\beta}\|_1$ and $\Omega_D^{Ad}(\boldsymbol{\beta}) = \|\mathcal{T}_3\boldsymbol{\beta}\|_1$, where $\mathcal{T}_1 \in \mathbb{R}^{d\times d}$, $\mathcal{T}_2 \in \mathbb{R}^{\frac{d(d-1)}{2}\times d}$ and

$\mathcal{T}_3 \in \mathbb{R}^{\frac{d(d-1)(d-2)}{2} \times d}$ are sparse matrices. $\mathcal{T}_1$ is a diagonal matrix with the weights w_i's along the diagonal. In $\mathcal{T}_2$, each row is a $1 \times d$ vector with only two non-zero entries w_{jk} and $-w_{jk}$ at the jth and kth positions respectively. In $\mathcal{T}_3$, each row is a $1 \times d$ vector with only three non-zero entries w_{ijk}, $-\alpha_{ijk} w_{ijk}$ and $(\alpha_{ijk} - 1)w_{ijk}$ at the ith, jth and kth positions respectively. Therefore, the storage and computation w.r.t. $\mathcal{T}_1$, $\mathcal{T}_2$ and $\mathcal{T}_3$ are very efficient since they are sparse matrices. The objective function of the ADFG method can be reformulated as

$$\min_{\boldsymbol{\beta}} \frac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda_1\|\mathcal{T}_1\boldsymbol{\beta}\|_1 + \lambda_2\|\mathcal{T}_2\boldsymbol{\beta}\|_1 + \lambda_3\|\mathcal{T}_3\boldsymbol{\beta}\|_1. \quad (9)$$

Since the regularizers in problem (9) are functions of linear transformations of $\boldsymbol{\beta}$, we introduce some new variables and reformulate problem (9) as

$$\begin{aligned} &\min_{\boldsymbol{\beta},\mathbf{p},\mathbf{q},\mathbf{r}} \frac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda_1\|\mathbf{p}\|_1 + \lambda_2\|\mathbf{q}\|_1 + \lambda_3\|\mathbf{r}\|_1 \\ &\text{s.t.} \quad \mathcal{T}_1\boldsymbol{\beta} - \mathbf{p} = \mathbf{0},\ \mathcal{T}_2\boldsymbol{\beta} - \mathbf{q} = \mathbf{0},\ \mathcal{T}_3\boldsymbol{\beta} - \mathbf{r} = \mathbf{0}. \end{aligned}$$

The augmented Lagrangian is then defined as

$$\begin{aligned} L_\rho(\boldsymbol{\beta}, \mathbf{p}, \mathbf{q}, \mathbf{r}, \boldsymbol{\mu}, \boldsymbol{\upsilon}, \boldsymbol{\nu}) &= \frac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda_1\|\mathbf{p}\|_1 + \lambda_2\|\mathbf{q}\|_1 \\ &+ \lambda_3\|\mathbf{r}\|_1 + \boldsymbol{\mu}^T(\mathcal{T}_1\boldsymbol{\beta} - \mathbf{p}) + \boldsymbol{\upsilon}^T(\mathcal{T}_2\boldsymbol{\beta} - \mathbf{q}) + \boldsymbol{\nu}^T(\mathcal{T}_3\boldsymbol{\beta} - \mathbf{r}) \\ &+ \frac{\rho}{2}\|\mathcal{T}_1\boldsymbol{\beta} - \mathbf{p}\|_2^2 + \frac{\rho}{2}\|\mathcal{T}_2\boldsymbol{\beta} - \mathbf{q}\|_2^2 + \frac{\rho}{2}\|\mathcal{T}_3\boldsymbol{\beta} - \mathbf{r}\|_2^2, \end{aligned}$$

where $\boldsymbol{\mu}$, $\boldsymbol{\upsilon}$, $\boldsymbol{\nu}$ are augmented Lagrangian multipliers. Then we can update all variables, including $\boldsymbol{\beta}$, $\mathbf{p}$, $\mathbf{q}$, $\mathbf{r}$, $\boldsymbol{\mu}$, $\boldsymbol{\upsilon}$, and $\boldsymbol{\nu}$, in one iteration as follows.

Update $\boldsymbol{\beta}$: In the $(k+1)$-th iteration, $\boldsymbol{\beta}^{k+1}$ is computed by minimizing L_ρ with other variables fixed:

$$\begin{aligned} &\arg\min_{\boldsymbol{\beta}} \frac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + (\mathcal{T}_1\boldsymbol{\mu}^k + \mathcal{T}_2^T\boldsymbol{\upsilon}^k + \mathcal{T}_3^T\boldsymbol{\nu}^k)^T\boldsymbol{\beta} \\ &+ \frac{\rho}{2}\|\mathcal{T}_1\boldsymbol{\beta} - \mathbf{p}^k\|_2^2 + \frac{\rho}{2}\|\mathcal{T}_2\boldsymbol{\beta} - \mathbf{q}^k\|_2^2 + \frac{\rho}{2}\|\mathcal{T}_3\boldsymbol{\beta} - \mathbf{r}^k\|_2^2. \end{aligned} \quad (10)$$

Problem (10) is a quadratic problem and has a closed-form solution as $\boldsymbol{\beta}^{k+1} = \mathbf{F}^{-1}\mathbf{b}^k$, where

$$\begin{aligned} &\mathbf{F} = \mathbf{X}^T\mathbf{X} + \rho(\mathbf{I} + \mathcal{T}_1^T\mathcal{T}_1 + \mathcal{T}_2^T\mathcal{T}_2 + \mathcal{T}_3^T\mathcal{T}_3), \\ &\mathbf{b}^k = \mathbf{X}^T\mathbf{y} - \mathcal{T}_1\boldsymbol{\mu}^k - \mathcal{T}_2^T\boldsymbol{\upsilon} - \mathcal{T}_3^T\boldsymbol{\nu} + \rho\mathcal{T}_1\mathbf{p}^k + \rho\mathcal{T}_2^T\mathbf{q}^k + \rho\mathcal{T}_3^T\mathbf{r}^k. \end{aligned}$$

Update $\mathbf{p}$, $\mathbf{q}$ and $\mathbf{r}$: $\mathbf{p}^{k+1}$ can be obtained by solving

$$\arg\min_{\mathbf{p}} \frac{\rho}{2}\|\mathbf{p} - \mathcal{T}_1\boldsymbol{\beta}^{k+1} - \frac{1}{\rho}\boldsymbol{\mu}^k\|_2^2 + \frac{\lambda_1}{\rho}\|\mathbf{p}\|_1,$$

which has a closed-form solution as $\mathbf{p}^{k+1} = S_{\lambda_1/\rho}(\mathcal{T}_1\boldsymbol{\beta}^{k+1} + \frac{1}{\rho}\boldsymbol{\mu}^k)$, where the soft-thresholding operator $S_\lambda(\cdot)$ is defined as $S_\lambda(x) = \text{sign}(x)\max\{|x| - \lambda, 0\}$. Similarly, we have $\mathbf{q}^{k+1} = S_{\lambda_2/\rho}(\mathcal{T}_2\boldsymbol{\beta}^{k+1} + \frac{1}{\rho}\boldsymbol{\upsilon}^k)$ and $\mathbf{r}^{k+1} = S_{\lambda_3/\rho}(\mathcal{T}_3\boldsymbol{\beta}^{k+1} + \frac{1}{\rho}\boldsymbol{\nu}^k)$.

Update $\boldsymbol{\mu}$, $\boldsymbol{\upsilon}$ and $\boldsymbol{\nu}$: $\boldsymbol{\mu}$, $\boldsymbol{\upsilon}$ and $\boldsymbol{\nu}$ can be updated as $\boldsymbol{\mu}^{k+1} = \boldsymbol{\mu}^k + \rho(\mathcal{T}_1\boldsymbol{\beta}^{k+1} - \mathbf{p}^{k+1})$, $\boldsymbol{\upsilon}^{k+1} = \boldsymbol{\upsilon}^k + \rho(\mathcal{T}_2\boldsymbol{\beta}^{k+1} - \mathbf{q}^{k+1})$, and $\boldsymbol{\nu}^{k+1} = \boldsymbol{\nu}^k + \rho(\mathcal{T}_3\boldsymbol{\beta}^{k+1} - \mathbf{r}^{k+1})$.

By noting that $\mathbf{F}^{-1}$ can be pre-computed, the whole learning procedure can be implemented very efficiently.

Theoretical Analysis

In this section, we study the asymptotic behavior of the proposed models as the number of samples $n \to \infty$. Assume $\boldsymbol{\beta}^*$ is the true coefficient vector. Let $\mathcal{A} = \{i : \beta_i^* \neq 0\}$ (the true pattern of non-zero coefficients) and $d_0 = |\mathcal{A}|$, $\mathcal{B} = \{(i,j) : \beta_i^* \neq 0 \text{ and } \beta_i^* = \beta_j^*\}$ (the true pattern of feature groups) and $\mathcal{D} = \{(i,j,k) : \beta_i^* \neq 0, \beta_j^* \neq 0, \beta_k^* \neq 0 \text{ and } \beta_i^* \neq \beta_j^*, \beta_i^* \neq \beta_k^*, \beta_j^* \neq \beta_k^*\}$ (the true pattern of different features). Let s_0 be the number of distinct non-zero coefficients in $\boldsymbol{\beta}^*$. Define $\boldsymbol{\beta}^*_{\mathcal{BD}} = (\beta^*_{i_1}, \cdots, \beta^*_{i_{s_0}})^T$, which is composed of the s_0 distinct non-zero values of $\boldsymbol{\beta}^*$, and let $\boldsymbol{\beta}^{Ad}_{\mathcal{BD}} = (\beta^{Ad}_{i_1}, \cdots, \beta^{Ad}_{i_{s_0}})^T$ be the corresponding estimation. Let $\mathcal{A}_1, \cdots, \mathcal{A}_{s_0}$ be the sets of indices where in each set the corresponding coefficients are equivalent. The learning model is a linear model, i.e., $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ where $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)^T$ is the noise. Moreover, we make two assumptions that are commonly used in the spare learning literature (Zou 2006; Viallon et al. 2013):

- ***A.1*** The noises $\epsilon_1, \ldots, \epsilon_n$ are $i.i.d$ random variables with mean 0 and variance σ^2;
- ***A.2*** $\frac{1}{n}\mathbf{X}^T\mathbf{X} \to \mathbf{C}$ where $\mathbf{C}$ is positive definite.

Let $\mathbf{C}_{\mathcal{A}}$ be the corresponding $d_0 \times d_0$ principal submatrix of $\mathbf{C}$ with the indices of rows and columns defined in $\mathcal{A}$. $\mathbf{X}_{\mathcal{BD}}$ is a matrix of size $n \times s_0$ with the ith column defined as $\mathbf{x}_{\mathcal{BD}_i} = \sum_{j \in \mathcal{A}_i} \mathbf{x}_j$. Then $\mathbf{C}_{\mathcal{BD}}$ is defined as $\mathbf{C}_{\mathcal{BD}} = \frac{1}{n}\mathbf{X}^T_{\mathcal{BD}}\mathbf{X}_{\mathcal{BD}}$. The regularization parameters are assumed to be functions of the sample size n and so they are denoted by $\lambda_m^{(n)}$ $(m = 1, 2, 3)$. For the asymptotic behavior of the DFG method, we have the following result.

Theorem 1 *Let $\widehat{\boldsymbol{\beta}}$ be the estimator of DFG. If $\lambda_m^{(n)}/\sqrt{n} \to \lambda_m^{(0)} \geq 0$ $(m = 1, 2)$, where $\lambda_m^{(0)}$ is some non-negative constant, then under assumptions **A.1** and **A.2** we have*

$$\sqrt{n}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) \to_d \arg\min_{\mathbf{u}} \mathcal{V}(\mathbf{u}),$$

where $\mathcal{V}(\mathbf{u})$ is defined as

$$\begin{aligned} \mathcal{V}(\mathbf{u}) &= \mathbf{u}^T\mathbf{C}\mathbf{u} - 2\mathbf{u}^T\mathbf{W} + \lambda_1^{(0)}\sum_{i=1} f(u_i, \beta_i^*) \\ &+ \frac{\lambda_2^{(0)}}{2}(d-2)\sum_{j<k} f(u'_{jk}, \beta'_{jk}) + \frac{\lambda_2^{(0)}}{2}\sum_{i=1}\sum_{\substack{j<k \\ j,k\neq i}} f(u''_{ijk}, \beta''_{ijk}). \end{aligned}$$

$\mathbb{I}(\cdot)$ is the indicator function, $f(x,y) = \text{sign}(y)x\mathbb{I}(y \neq 0) + |x|\mathbb{I}(y = 0)$, $u'_{jk} = u_j - u_k$, $\beta'_{jk} = \beta_j^* - \beta_k^*$, $u''_{ijk} = (2u_i - u_j - u_k)/2$, $\beta''_{ijk} = (2\beta_i^* - \beta_j^* - \beta_k^*)/2$, and $\mathbf{W}$ is assumed to follow a normal distribution $\mathcal{N}(\mathbf{0}, \sigma^2\mathbf{C})$.

Theorem 1 gives the root-n consistency of DFG. However, the following theorem implies that when $\lambda_m^{(n)} = O(\sqrt{n})$ $(m = 1, 2)$, the support of $\boldsymbol{\beta}^*$, i.e. the non-zero elements in $\boldsymbol{\beta}^*$, cannot be recovered by the DFG method with high probability.

Theorem 2 *Let $\widehat{\boldsymbol{\beta}}$ be the estimator of DFG and $\widehat{\mathcal{A}}_n = \{i : \widehat{\beta}_i \neq 0\}$. If $\lambda_m^{(n)}/\sqrt{n} \to \lambda_m^{(0)} \geq 0$ $(m = 1, 2)$, then under assumptions **A.1** and **A.2**, we have*

$$\limsup_n \mathbb{P}(\widehat{\mathcal{A}}_n = \mathcal{A}) \leq c < 1,$$

where c is a constant depending on the true model.

For the ADFG method, we can prove that with appropriate choices for $\lambda_m^{(n)}$ $(m = 1, 2, 3)$, the estimation $\widehat{\boldsymbol{\beta}}^{Ad}$ obtained from the ADFG method enjoys nice asymptotic oracle properties, which are depicted in the following theorem, in contrast with the DFG method.

Theorem 3 *Let $\widehat{\boldsymbol{\beta}}^{Ad}$ be the estimator of ADFG. Let $\widehat{\mathcal{A}}_n^{Ad}$, $\widehat{\mathcal{B}}_n^{Ad}$, and $\widehat{\mathcal{D}}_n^{Ad}$ be the corresponding sets obtained from $\widehat{\boldsymbol{\beta}}^{Ad}$. If $\lambda_m^{(n)}/\sqrt{n} \to 0$ and $\lambda_m^{(n)} n^{(\gamma-1)/2} \to \infty$ $(m = 1, 2, 3)$, then under assumptions **A.1** and **A.2** we have*

- *Consistency in feature selection and discriminative feature grouping: $\mathbb{P}(\widehat{\mathcal{A}}_n^{Ad} = \mathcal{A}) \to 1$, $\mathbb{P}(\widehat{\mathcal{B}}_n^{Ad} = \mathcal{B}) \to 1$ and $\mathbb{P}(\widehat{\mathcal{D}}_n^{Ad} = \mathcal{D}) \to 1$ as $n \to \infty$.*
- *Asymptotic normality: $\sqrt{n}(\widehat{\boldsymbol{\beta}}_{\mathcal{BD}}^{Ad} - \boldsymbol{\beta}_{\mathcal{BD}}^*) \to_d \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{C}_{\mathcal{BD}}^{-1})$.*

Theorem 3 shows that the ADFG method has good property as stated in the asymptotic normality part, and the estimated sets $\widehat{\mathcal{A}}_n^{Ad}$, $\widehat{\mathcal{B}}_n^{Ad}$ and $\widehat{\mathcal{D}}_n^{Ad}$ can recover the corresponding true sets defined in $\boldsymbol{\beta}^*$ with high probability approaching 1 when n goes to infinity.

Experiments

In this section, we conduct empirical evaluation for the proposed methods by comparing with the Lasso, GFLasso, OSCAR, and the non-convex extensions of OSCAR, i.e. the ncFGS and ncTFGS methods in problems (3) and (4).

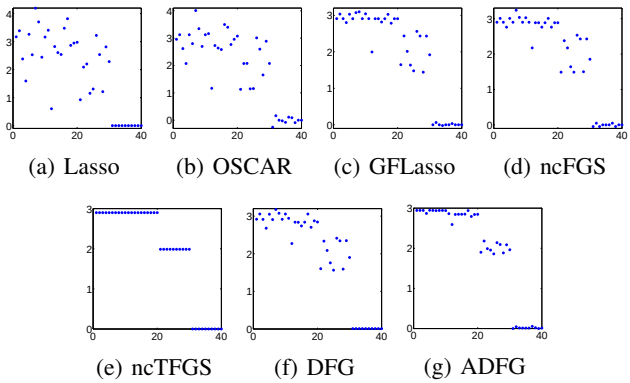
Synthetic Data

We study two synthetic datasets to compare the performance of different methods. The two datasets are generated according to a linear regression model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ with the noises generated as $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$, where $\mathbf{X} \in \mathbb{R}^{n \times d}$, $\boldsymbol{\beta} \in \mathbb{R}^d$, and $\mathbf{y} \in \mathbb{R}^n$. In the first dataset, n, d, and σ are set to be 100, 40 and 2 respectively. The ground truth for the coefficients is $\boldsymbol{\beta}^* = (\underbrace{3, \cdots, 3}_{10(G_1)}, \underbrace{2.8, \cdots, 2.8}_{10(G_2)}, \underbrace{2, \cdots, 2}_{10(G_3)}, \underbrace{0, \cdots, 0}_{10})^T$. In this case, we can see that two feature groups G_1 and G_2 are similar, making them easy to be mis-identified compared with G_3. Each data point corresponding to a row in $\mathbf{X}$ is generated from a normal distribution $\mathcal{N}(\mathbf{0}, \mathbf{S})$ where the ith diagonal element s_{ii} of $\mathbf{S}$ is set to 1 for all i. The (i, j)th element of $\mathbf{S}$, s_{ij}, is set to 0.9 if $i \neq j$ and $\beta_i^* = \beta_j^*$, and otherwise $s_{ij} = 0.25^{|\beta_i^* - \beta_j^*|}$. The settings of the second dataset are almost identical to those of the first dataset except that $\boldsymbol{\beta}^* = (\underbrace{2.8, .., 2.8}_{10(G_1)}, \underbrace{2.6, .., 2.6}_{10(G_2)}, \underbrace{2.4, .., 2.4}_{10(G_3)}, \underbrace{0, .., 0}_{10})^T$, where groups G_1, G_2, and G_3 are easy to be mis-identified.

We use the mean square error (MSE) to measure the performance of the estimation $\boldsymbol{\beta}$ with the MSE defined as $MSE = \frac{1}{n}(\boldsymbol{\beta} - \boldsymbol{\beta}^*)^T \mathbf{X}^T \mathbf{X} (\boldsymbol{\beta} - \boldsymbol{\beta}^*)$. To measure the accuracy of feature grouping and group discriminating, we introduce a metric $S = \sum_{i=1}^K S_i / K$ with K as the number of feature groups and S_i defined as

$$S_i = \frac{\sum_{j \neq k, j,k \in I_i} \mathbb{I}(\beta_i = \beta_j) + \sum_{j \neq k, j \in I_i, k \notin I_i} \mathbb{I}(\beta_j \neq \beta_k)}{|I_i|(d-1)},$$

where I_i $(i = 1, \ldots, K)$ denotes the set of indices of the ith feature group with non-zero coefficients in the ground truth. The numerator in S_i consists of two parts, where the first and second terms represent the recovery of equal and unequal coefficients for I_i separately. The denominator is the total number of possible combinations. Thus, S can measure the performance of feature grouping and discriminating, and a larger value for S indicates better performance. For each dataset, we generate n samples for training, as well as additional n samples for testing. Hyperparameters, including the regularization parameters in all the models, τ in ncTFGS, and γ in ADFG, are tuned using an independent validation set with n samples. We use a grid search method with the resolutions for the λ_i's $(i = 1, 2, 3)$ in all methods as $[10^{-4}, 10^{-3}, \cdots, 10^2]$ and those for γ as $[0, 0.1, \cdots, 1]$. Moreover, the resolution for τ in the ncTFGS method is $[0.05, 0.1, \cdots, 5]$, which is in line with the setting of the original work (Yang et al. 2012).



(a) Lasso (b) OSCAR (c) GFLasso (d) ncFGS

(e) ncTFGS (f) DFG (g) ADFG

Figure 3: Feature coefficients obtained on the first synthetic data.

Figure 3 shows the feature coefficients obtained by different methods on the first dataset. We see that the ncFGS, ncTFGS, DFG, and ADFG methods achieve better parameter estimation than the Lasso and OSCAR methods. The ncTFGS method shows clear recovery of group G_3 but it mis-combines G_1 and G_2 together as one group. In contrast, although the coefficients in one group obtained from the DFG and ADFG methods are not exactly identical which also occurs in the OSCAR, GFLasso, and ncFGS methods, they are able to distinguish G_1 and G_2. Table 1 shows the average performance of different methods in terms of MSE and S over 10 simulations. In the first dataset, due to the existent of a distant group G_3 (from G_1 and G_2), the ncTFGS achieves the best performance in terms of S, while in the second problem where all the three groups are similar, the DFG and ADFG methods achieve a higher S. In both datasets, the ADFG has the best performance in terms of MSE.

Breast Cancer

We conduct experiments on the previously studied breast cancer data, which contains 8141 genes in 295 tumors (78 metastatic and 217 non-metastatic). The tasks here

Table 1: Average results over 10 repetitions in terms of mean and standard deviation on the synthetic datasets.

Dataset	(1)		(2)	
	MSE	S	MSE	S
Lasso	1.929(0.591)	-	1.853(0.831)	-
OSCAR	1.511(0.545)	0.766(0.015)	1.204(0.534)	0.753(0.021)
GFLasso	0.477(0.281)	0.843(0.065)	0.462(0.284)	0.763(0.039)
ncFGS	0.476(0.286)	0.842(0.063)	0.462(0.279)	0.768(0.031)
ncTFGS	0.323(0.191)	**0.857(0.064)**	0.574(0.280)	0.759(0.110)
DFG	0.399(0.194)	0.781(0.016)	0.267(0.191)	0.770(0.028)
ADFG	**0.289(0.152)**	0.815(0.058)	**0.216(0.149)**	**0.776(0.042)**

Table 2: Results averaged over 10 repetitions for different methods on Breast Cancer dataset without a priori information.

	Acc. (%)	Sen. (%)	Pec. (%)
Lasso	73.5(5.5)	83.5(5.6)	63.0(8.1)
OSCAR	76.2(1.9)	87.3(5.3)	64.4(6.6)
GFLasso	76.8(2.0)	87.6(4.9)	65.7(5.1)
ncFGS	77.4(1.4)	88.4(5.0)	66.0(5.8)
ncTFGS	78.1(1.9)	87.3(5.3)	68.3(6.2)
DFG	78.5(3.0)	87.3(5.4)	69.5(8.1)
ADFG	**81.1(3.8)**	**89.9(7.1)**	**71.9(8.5)**

are binary classification problems to distinguish between metastatic and non-metastatic tumors. We use the square loss for all methods. In this data, the group information are known a priori, and we have observed from Figure 1(b) that a large number of similar but different groups exist. In addition to the feature groups, some a priori information about the feature correlations between some pairs of features in terms of a graph is also known. In the following, we conduct two experiments. In the first experiment, we do not utilize any prior information, while the second one compares the variants of OSCAR, ncFGS, ncTFGS, DFG and ADFG by incorporating the graph information. The measurements include accuracy (Acc.), sensitivity (Sen.) and specificity (Pec.) as used in (Yang et al. 2012).

Learning without A Priori Information Similar to (Jacob, Obozinski, and Vert 2009; Zhong and Kwok 2012), we select the 300 most correlated genes to the outputs as the feature representation, and alleviate the class imbalance problem by duplicating the positive samples twice. 50%, 30%, and 20% of data are randomly chosen for training, validation and testing, respectively. Table 2 shows the average results over 10 repetitions. In Table 2, the DFG and ADFG methods show very competitive performance compared with other methods in all the three metrics, and the ADFG method achieves the best performance.

Incorporating Graph Information We investigate the variants of the DFG and ADFG methods introduced previously by utilizing the available priori information on feature correlations in terms of a graph. The data preparation is identical to that in the previous experiment. The results are shown in Table 3. Similar to the experiment without a priori information, the DFG and ADFG methods perform better than the other methods, and the ADFG method enjoys the best performance. In addition, the performance of all the methods improves compared with that in the previous experiment, which implies that the prior information is helpful.

Table 3: Results averaged over 10 repetitions for different methods on Breast Cancer dataset with the given graph information.

	Acc. (%)	Sen. (%)	Pec. (%)
OSCAR	78.7(4.1)	89.0(3.8)	66.5(6.9)
GFLasso	79.1(4.3)	88.1(4.8)	69.4(5.0)
ncFGS	80.6(4.2)	**91.3(3.6)**	68.0(6.0)
ncTFGS	81.1(4.0)	90.2(4.1)	70.3(7.6)
DFG	**82.3(4.5)**	**91.5(6.8)**	**71.4(7.1)**
ADFG	**82.4(4.0)**	**91.3(6.1)**	**71.8(6.4)**

Table 4: Test accuracy (%) averaged over 10 repetitions for different methods on 20-Newsgroups dataset.

Class pairs	(1)	(2)	(3)	(4)	(5)
Lasso	75.1(2.9)	81.9(2.0)	74.8(0.9)	78.5(4.0)	81.1(1.6)
OSCAR	75.5(1.8)	82.7(0.9)	73.8(1.7)	78.9(2.1)	**83.7(1.9)**
GFLasso	**76.2(1.5)**	83.7(1.5)	74.5(1.9)	77.6(1.8)	**83.7(2.0)**
ncFGS	75.3(1.6)	82.8(1.2)	72.7(1.5)	77.4(2.2)	82.9(1.3)
ncTFGS	75.3(1.5)	82.8(1.2)	72.8(1.4)	77.7(2.1)	**83.6(1.7)**
DFG	**76.3(2.1)**	85.0(2.0)	**77.0(1.1)**	79.2(3.3)	**83.9(2.4)**
ADFG	**76.4(2.1)**	**86.0(1.7)**	**77.1(2.4)**	**80.4(3.0)**	**83.8(2.5)**

20-Newsgroups

Following (Yang et al. 2012), we use the data from some pairs of classes in the 20-newsgroups dataset to form binary classification problems. To make the tasks more challenging, we select 5 pairs of very similar classes: (1) baseball vs. hockey; (2) autos vs. motorcycles; (3) mac vs. ibm.pc; (4) christian vs. religion.misc; (5) guns vs. mideast. Therefore, in all the settings, the feature groups are more likely to be similar, posing challenge to identify them. Similar to (Yang et al. 2012), we first use the ridge regression to select the 300 most important features and all the features are centered and scaled to unit variance. Then 20%, 40% and 40% of samples are randomly selected for training, validation, and testing, respectively. Table 4 reports the average classification accuracy over 10 repetitions for all the methods. According to the results, the performance of the OSCAR method is better than that of the ncFGS and ncTFGS methods and the DFG and ADFG methods outperform other methods, which again verifies the effectiveness of our methods.

Conclusion and Future Work

In this paper, we proposed a novel regularizer together with its adaptive extension to achieve discriminative feature grouping. We developed an efficient algorithm and discussed the asymptotic properties for the proposed models. In feature grouping, the assumption that the values of coefficients in a feature group should be exactly identical seems a bit restricted. In our future work, we will relax this assumption and learn more flexible feature groups.

Acknowledgment

This work is supported by NSFC 61305071 and HKBU FRG2/13-14/039.

References

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

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

Dettling, M., and Bühlmann, P. 2004. Finding predictive gene groups from microarray data. *Journal of Multivariate Analysis* 90(1):106–131.

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

Jacob, L.; Obozinski, G.; and Vert, J.-P. 2009. Group lasso with overlap and graph lasso. In *International Conference on Machine Learning*.

Jang, W.; Lim, J.; Lazar, N. A.; Loh, J. M.; and Yu, D. 2013. Regression shrinkage and grouping of highly correlated predictors with HORSES. *arXiv preprint arXiv:1302.0256*.

Jenatton, R.; Mairal, J.; Bach, F. R.; and Obozinski, G. R. 2010. Proximal methods for sparse hierarchical dictionary learning. In *International Conference on Machine Learning*.

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

Kim, S., and Xing, E. P. 2010. Tree-guided group lasso for multi-task regression with structured sparsity. In *International Conference on Machine Learning*.

Shen, X., and Huang, H.-C. 2010. Grouping pursuit through a regularization solution surface. *Journal of the American Statistical Association* 105(490).

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

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

Viallon, V.; Lambert-Lacroix, S.; Höfling, H.; and Picard, F. 2013. Adaptive generalized fused-lasso: Asymptotic properties and applications. *Technical Report*.

Yang, S.; Yuan, L.; Lai, Y.-C.; Shen, X.; Wonka, P.; and Ye, J. 2012. Feature grouping and selection over an undirected graph. In *ACM SIGKDD Conference on Kownledge Discovery and Data Mining*.

Yuan, M., and Lin, Y. 2006. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 68(1):49–67.

Yuan, L.; Liu, J.; and Ye, J. 2011. Efficient methods for overlapping group lasso. In *Advances in Neural Information Processing Systems*.

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

Zhong, L. W., and Kwok, J. T. 2012. Efficient sparse modeling with automatic feature grouping. In *International Conference on Machine Learning*.

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

Zou, H. 2006. The adaptive lasso and its oracle properties. *Journal of the American statistical association* 101(476):1418–1429.