Published in final edited form as:
J Am Stat Assoc. 2010 June 1; 105(490): 727–739. doi:10.1198/jasa.2010.tm09380.

Grouping pursuit through a regularization solution surface [*]

Xiaotong Shen and **Hsin-Cheng Huang**

Summary

Extracting grouping structure or identifying homogenous subgroups of predictors in regression is crucial for high-dimensional data analysis. A low-dimensional structure in particular–grouping, when captured in a regression model, enables to enhance predictive performance and to facilitate a model's interpretability Grouping pursuit extracts homogenous subgroups of predictors most responsible for outcomes of a response. This is the case in gene network analysis, where grouping reveals gene functionalities with regard to progression of a disease. To address challenges in grouping pursuit, we introduce a novel homotopy method for computing an entire solution surface through regularization involving a piecewise linear penalty. This nonconvex and overcomplete penalty permits adaptive grouping and nearly unbiased estimation, which is treated with a novel concept of grouped subdifferentials and difference convex programming for efficient computation. Finally, the proposed method not only achieves high performance as suggested by numerical analysis, but also has the desired optimality with regard to grouping pursuit and prediction as showed by our theoretical results.

Keywords

Gene networks; large *p* but small *n*; nonconvex minimization; prediction; supervised clustering

1 Introduction

Essential to high-dimensional data analysis is seeking a certain lower-dimensional structure in knowledge discovery, as in web mining. Extracting one-kind of lower-dimensional structure–grouping, remains largely unexplored in regression. In gene network analysis, a large amount of current genetic knowledge has been organized in terms of networks, for instance, the Kyoto Encyclopedia of Genes and Genomes (KEGG), a collection of manually drawn pathway maps representing the knowledge about molecular interactions and reactions. In situations as such, extracting homogenous subnetworks from a network of dependent predictors, most responsible for predicating outcomes of a response, has been one key challenge of biomedical research. There homogenous subnetworks of genes are usually estimated for understanding a disease's progression. The central issue this article addresses is automatic identification of homogenous subgroups in regression, what we call grouping pursuit.

Now consider a linear model in which response Y_i depends on a vector of p predictors:

[*] Xiaotong Shen is Professor, School of Statistics, University of Minnesota, 224 Church Street S.E., Minneapolis, MN 55455 (E-mail: xshen@stat.umn.edu). His research is supported in part by National Science Foundation Grant DMS-0906616 and National Institute of Health Grant 1R01GM081535. Hsin-Cheng Huang is Research Fellow, Institute of Statistical Science, Academia Sinica, Taipei 115, Taiwan (E-mail: hchuang@stat.sinica.edu.tw). He is supported in part by Grant NSC 97-2118-M-001-001-MY3. The authors thank the editor, the associate editor and three referees for their helpful comments and suggestions.

$$Y_i \equiv \mu(\boldsymbol{x}_i)+\boldsymbol{\varepsilon}_i, \quad \mu(\boldsymbol{x}) \equiv \boldsymbol{x}^T \beta=\sum_{j=1}^{p} x_j\beta_j, \quad E(\varepsilon_i)=0, \quad Var(\varepsilon_i)=\sigma^2; \quad i=1,\ldots,n, \tag{1}$$

where $\boldsymbol{\beta} \equiv (\beta_1, \ldots, \beta_p)^T$ is a vector of regression coefficients, $\boldsymbol{x}_i$ is independent of ε_i, and $\mu(\boldsymbol{x})$ is in a generic form, including linear, and nonlinear predictors expressed in terms of linear combinations of known bases. Our objective is to identify all possible homogenous subgroups of predictors, for optimal prediction of the outcome of $\boldsymbol{Y}$. Here homogeneity means that regression coefficients are of similar (same) values, that is, $\beta_{j_1} \approx \cdots \approx \beta_{j_K}$ within each group $\{j_1, \ldots, j_K\} \subset \{1, \ldots, p\}$. In (1), grouping pursuit estimates all distinct values of $\boldsymbol{\beta}$ as well as all corresponding subgroups of homogenous predictors.

Grouping pursuit seeks variance reduction of estimation while retaining roughly the same amount of bias, which is advantageous in high-dimensional analysis. First, it collapses predictors whose sample covariances between the residual and predictors are of similar values, for best predicting outcomes of $\boldsymbol{Y}$; c.f., Theorem 4. Moreover, it goes beyond the notion of feature selection. This is because it seeks not only a set of redundant predictors, or a single group of zero-coefficient predictors, but also additional homogenous subgroups for further variance reduction. As a result, it yields higher predictive performance. These aspects are confirmed by numerical and theoretical results in Sections 4 and 5. Second, the price to be paid for adaptive grouping pursuit is estimation of tuning parameters, which is small as compared to its potential gain of a simpler model with higher predictive accuracy.

Grouping pursuit considered in here is one kind of supervised clustering. Papers that investigate groping pursuit are those of Tibshirani et al. (2005), where the Fused Lasso is proposed using an L_1-penalty with respect to a certain serial order; Bondell and Reich (2008), where the OSCAR penalty involves pairwise L_∞-penalties for grouping variables in terms of absolute values, in addition to variable selection. Grouping pursuit dramatically differs from feature selection for grouped predictors. This is in the sense that the former only groups predictors *without* removing redundancy, whereas the latter removes redundancy by encouraging grouped predictors stay together in selection; see Yuan and Lin (2006), and Zhao, Rocha and Yu (2009).

Our primary objective is achieving high accuracy in both grouping and prediction through a computationally efficient method, which seems to be difficult, if not impossible, with existing methods, especially those through enumeration. To achieve our objective, we employ the regularized least squares method with a piecewise linear nonconvex penalty. The penalty to be introduced in (2) involves one thresholding parameter determining which pairs to be shrunk towards a common group, which works jointly with one regularization parameter for shrinkage towards unknown location. These two tuning parameters combine thresholding with shrinkage for achieving adaptive grouping, which is otherwise not possible with shrinkage alone. The penalty is overcomplete in that the number of individual penalty terms in the penalty may be redundant with regard to certain grouping structures, and is continuous but with three nondifferentiable points, leading to significant computational advantage, in addition to the desired optimality for grouping pursuit (Theorems 3 and Corollary 1).

Computationally, the proposed penalty imposes great challenges in two aspects: (a) potential discontinuities and (b) overcompleteness of the penalty, where an effective treatment does not seem to exist in the literature; see Friedman et al. (2007) about computational challenges for a pathwise coordinate method in this type of situation. To meet the challenges, we design a novel homotopy algorithm to compute the regularization solution surface. The algorithm uses a novel concept of grouped subdifferentials to deal with overcompleteness for tracking the process of grouping, and difference convex (DC) programming to treat discontinuities due to nonconvex minimization. This, together with a model selection routine for estimators that can be discontinuous, permits adaptive grouping pursuit.

Theoretically, we derive a finite-sample probability error bound of our DC estimator, what we call DCE, computed from the homotopy algorithm for grouping pursuit. On this basis, we prove that DCE is consistent with regard to grouping pursuit as well as reconstructing the unbiased least squares estimate under the true grouping, roughly for nearly exponentially many predictors in n as long as $\frac{\log p}{n} \to 0$, c.f., Theorem 3 for details.

For subnetwork analysis, we apply our proposed method to study predictability of a protein-protein interaction (PPI) network of genes on the time to breast cancer metastasis through gene expression profiles. In Section 5.2, 27 homogenous subnetworks are identified through a Laplacian network weight vector, which surround three tumor suppressor genes TP53, BRACA1 and BRACA2 for metastasis. There 17 disease genes that were identified in the study of Wang et al. (2005) belong to 5 groups containing 1, 1, 1, 1, and 13 disease genes, indicating gene functionalities with regard to breast cancer survivability.

This article is organized in seven sections. Section 2 introduces the proposed method and the homotopy algorithm. Section 3 is devoted to selection of tuning parameter. Section 4 presents a theory concerning optimal properties of DCE in grouping pursuit and prediction, followed by some numerical examples and an application to breast cancer data in Section 5. Section 6 discusses the proposed method. Finally, the appendix contains technical proofs.

2 Grouping pursuit

In (1), let the true coefficient vector $\beta^0=(\beta_1^0, \ldots, \beta_p^0)^T$ be $(\alpha_1^0 \mathbf{1}^T_{|\mathcal{G}_1^0|}, \ldots, \alpha_{K^0}^0 \mathbf{1}^T_{|\mathcal{G}_{K^0}^0|})^T$, where K^0 is the number of distinct groups, $\alpha_1^0 < \cdots < \alpha_{K^0}^0$, and $\mathbf{1}_{|\mathcal{G}_1^0|}$ denotes a vector of 1's with length $|\mathcal{G}_1^0|$. Grouping pursuit, as defined early, estimates true grouping $\mathcal{G}^0=(\mathcal{G}_1^0, \ldots, \mathcal{G}_{K^0}^0)$ as well as $\alpha^0=(\alpha_1^0, \ldots, \alpha_{K^0}^0)^T$. Without loss of generality, assume that the response and predictors are centered, that is, $\boldsymbol{Y}^T\mathbf{1} = 0$ and $(x_{1j}, \ldots, x_{nj})\mathbf{1} = 0$; $j = 1, \ldots, p$.

Ideally, one may enumerate over all possible least squares regressions for identifying the best grouping. However, the total number of all possible groupings, which is the pth Bell number (Rota, 1964), is much larger than that of all possible subsets in feature selection, hence that it is computationally infeasible even for moderate p. For instance, the 10th order Bell number is 115975. To circumvent this difficulty, we develop an automatic nonconvex regularization method to obtain (1) accurate grouping, (2) the least squares estimate based on the true grouping, (3) an efficient homotopy algorithm, and (4) high predictive performance.

Our approach utilizes a penalty involving pairwise comparisons: $\{\beta_j - \beta_{j'} : 1 \le j < j' \le p\}$. When $\beta_j - \beta_{j'} = 0$, X_j and $X_{j'}$ are grouped. By transitivity, that is, $\beta_{j1} - \beta_{j2} = 0$ and $\beta_{j2} - \beta_{j3} = 0$ imply that $\beta_{j1} = \beta_{j2} = \beta_{j3}$, we identify all homogenous groups through $p(p - 1)/2$ comparisons. Naturally, these comparisons can be conducted through penalized least squares with penalty $\Sigma_{j<j'} |\beta_j - \beta_{j'}|$. However, this convex penalty is not desirable for predictive performance, because it is not adaptive for discriminating large from small pairwise differences. As a result, overpenalizing large differences due to shrinking small differences towards zero impedes predictive performance. We thus introduce its nonconvex counterpart

$J(\boldsymbol{\beta}) = \Sigma_{j<j'} G(\beta_j - \beta_{j'})$ for adaptive grouping pursuit, where $G(z) = \lambda_2$ if $|z| > \lambda_2$ and $G(z) = |z|$ otherwise, with $\lambda_2 > 0$ being the thresholding parameter. For $G(z)$, one locally convex and two locally concave points at $z = 0, \pm\lambda_2$ enable us to achieve computational advantage, as well as to realize sharp statistical properties. First, the piecewise linearity and the two locally concave points of $G(z)$ yield an efficient method (Algorithm 1), and fast finite-step convergence of the surface algorithm (Algorithm 2). Second, they yield a sharp finite-sample error bound in Theorem 3. These aspects are unique for $G(z)$, which may not be shared by other penalties such as SCAD (Fan and Li, 2001); see the discussion after Theorem 2. A function like $G(z)$ was considered in other contexts such as wavelet denoising (Fan, 1997) and a combined L_0 and L_1 penalty via integer programming (Liu and Wu, 2007).

We now propose our penalized least squares criterion for automatic grouping pursuit:

$$S(\beta) = \frac{1}{2n}\sum_{i=1}^{n}(Y_i - \boldsymbol{x}_i^T\beta)^2 + \lambda_1 J(\beta), \quad J(\beta) = \sum_{j<j'} G(\beta_j - \beta_{j'}), \tag{2}$$

where $\lambda_1 > 0$ is the regularization parameter controlling the degree of grouping. For (2), any local/global minimizer can not attain at any of non-smooth locally concave points of $J(\boldsymbol{\beta})$.

Lemma 1 Let $h(\cdot)$ be any differentiable function in $\mathbb{R}^p$ and $\beta * = (\beta_1^*, \ldots, \beta_p^*)^T$ be a local minimizer of $f(\boldsymbol{\beta}) = h(\boldsymbol{\beta}) + \lambda_1 J(\boldsymbol{\beta})$ with $J(\cdot)$ given in (2). Then $|\beta_j^* - \beta_{j'}^*| \neq \lambda_2$ for $j \neq j'$.

2.1 Grouped subdifferentials

We now introduce a novel concept of grouped subdifferentials for a convex function, which constitutes a basis of our homotopy algorithm for tracking the process of grouping.

A subgradient of a convex function $f(\boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}$ at $\boldsymbol{\beta}$ is any vector $\boldsymbol{b} \in \mathbb{R}^p$ satisfying $f(\boldsymbol{\beta}^*) \geq f(\boldsymbol{\beta}) + b^T(\boldsymbol{\beta}^* - \boldsymbol{\beta})$ for any $\boldsymbol{\beta}^*$ with sufficiently small $\boldsymbol{\beta}^* - \boldsymbol{\beta}$, and reduces to the derivative at a smooth point. The subdifferential of $S(\boldsymbol{\beta})$ at any $\boldsymbol{\beta}$ is the set of all such $\boldsymbol{b}$'s, which is either a singleton or a non-singleton compact set. Let $\hat{\boldsymbol{\beta}}$ be a local minimizer of (2) and $(\mathcal{G}_1, \ldots, \mathcal{G}_K)$ be the corresponding grouping, where K is the number of distinct groups. The subgradient of $|\beta_j - \beta_{j'}|$ with respect to β_j at $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}$ is given by $b_{jj'} = \text{Sign}(\hat{\beta}_j - \hat{\beta}_{j'})$ if $|\hat{\beta}_j - \hat{\beta}_{j'}| > 0$, and $|b_{jj'}| \leq 1$ otherwise. Here $b_{jj'}$ is a singleton everywhere except at $\hat{\beta}_j - \hat{\beta}_{j'} = 0$.

To proceed, write $\hat{\boldsymbol{\beta}}$ as $(\widehat{\alpha}_1 \mathbf{1}_{|\mathcal{G}_1|}^T, \ldots, \widehat{\alpha}_K \mathbf{1}_{|\mathcal{G}_K|}^T)^T$, where the index in each group is arranged increasingly, $\hat{\alpha}_1 < \cdots < \hat{\alpha}_K$. Define $g(j) \equiv k$ if $\hat{\beta}_j = \hat{\alpha}_k$; $j = 1, \ldots, p$, mapping indices from $\hat{\boldsymbol{\beta}}$ to $\hat{\boldsymbol{\alpha}}$. Then grouping $(\mathcal{G}_1, \ldots, \mathcal{G}_K)$ partitions index set $\{1, \ldots, p\}$, with $\mathcal{G}_k \equiv \{j : g(j) = k\}$; $k = 1, \ldots, K$.

Ordinarily, group splitting can be tracked through certain transition conditions for $\{b_{jj'} : j \neq j'\}$, c.f., Rosset and Zhu (2007). However, $\{b_{jj'} : j \neq j'\}$ are not estimable from data when an overcomplete penalty is used. To overcome this difficulty, we define the grouped subgradient of index $j \in \mathcal{G}_k$ at $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}$ as $B_j \equiv \Sigma_{j' \in \mathcal{G}_k \setminus \{j\}} b_{jj'}$ if $|\mathcal{G}_k| > 1$, and $B_j \equiv 0$ if $|\mathcal{G}_k| = 1$; $j = 1, \ldots, p$. Note that $\Sigma_{j \in \mathcal{G}_k} B_j = 0$; $k = 1, \ldots, K$, because $b_{jj'} = -b_{j'j}$; $j \neq j'$. Moreover, we define the grouped subgradient of a subset $A \subset \mathcal{G}_k$ at $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}$ as $B_A \equiv \Sigma_{j \in A} B_j = \Sigma_{(j,j') \in A \times (\mathcal{G}_k \setminus A)} b_{jj'}$. Then

$$|B_A| \le |A|(|\mathcal{G}_k| - |A|). \tag{3}$$

Subsequently, we work with $\{B_A : A \subset \{1, \dots, p\}\}$ that can be uniquely determined; see Theorem 1.

2.2 Difference convex programming

This section treats non-differentiable nonconvex minimization (2) through DC programming, which is a principle for nonconvex minimization, relying on decomposing an objective function into a difference of two convex functions. The reader may consult An and Tao (1997) for DC programming. Through this DC method, we will design a novel homotopy algorithm for a DC solution of (2) in Section 2.3, which is a solution through DC programming.

First, we decompose $S(\boldsymbol{\beta})$ in (2) into a difference of two convex functions $S_1(\beta)=\frac{1}{2n}\sum_{i=1}^{n}(Y_i - \boldsymbol{x}_i^T\beta)^2+\lambda_1\sum_{j<j'}|\beta_j - \beta_{j'}|$ and $S_2(\boldsymbol{\beta}) = \lambda_1 \Sigma_{j<j'} G_2(\beta_j - \beta_{j'})$, through a DC decomposition of $G(\cdot) = G_1(\cdot) - G_2(\cdot)$ with $G_1(z) = |z|$ and $G_2(z) = (|z| - \lambda_2)_+$, where z_+ is the positive part of z. This DC decomposition is interpretable in that $S_2(\cdot)$ corrects the estimation bias due to use of convex penalty $\lambda_1 \Sigma_{j<j'} |\beta_j - \beta_{j'}|$ for a nonconvex problem (2).

Second, we construct a sequence of upper approximations by successively replacing $S_2(\boldsymbol{\beta})$ at iteration $m = 0, 1, \dots$, by its affine minorization based on iteration $m - 1$, leading to an upper convex approximating function at iteration m:

$$S_1(\beta) - S_2\left(\widehat{\widehat{\beta}}^{(m-1)}(\lambda_1, \lambda_2)\right) - \left(\beta - \widehat{\widehat{\beta}}^{(m-1)}(\lambda_1, \lambda_2)\right)^T \nabla S_2\left(\widehat{\widehat{\beta}}^{(m-1)}(\lambda_1, \lambda_2)\right), \tag{4}$$

where ∇ is the subgradient operator, $\hat{\beta}^{(m-1)}$ (λ_1, λ_2) is the minimizer of (4) at iteration $m - 1$, and $\hat{\beta}^{(-1)}$ $(\lambda_1, \lambda_2) \equiv \mathbf{0}$. The last term in (4) becomes $\lambda_1\sum_{j=1}^{p}\left(\beta_j - \widehat{\widehat{\beta}}_j^{(m-1)}(\lambda_1, \lambda_2)\right) \times \sum_{j':j' \neq j} \nabla G_2\left(\widehat{\widehat{\beta}}_j^{(m-1)}(\lambda_1, \lambda_2) - \widehat{\widehat{\beta}}_{j'}^{(m-1)}(\lambda_1, \lambda_2)\right)$ with $\nabla G_2(z) = \mathrm{Sign}(z)I(|z| > \lambda_2)$ being a subgradient of G_2 at z.

Third, we utilize the grouped subdifferentials to track the entire solution surface iteratively. One technical difficulty is that $\hat{\beta}^{(m)}$ (λ_1, λ_2) would have jumps in λ_1 if $\hat{\beta}^{(0)}$ (λ_1, λ_2) were piecewise linear in λ_1 given λ_2, in view of (6) and (7) in Theorem 1. This is undesirable for tracking by continuity through homotopy. For grouping pursuit, we therefore replace $\hat{\beta}^{(0)}$ (λ_1, λ_2) in (4) by $\hat{\beta}^{(0)}$ (λ_0, λ_2), where rough tuning for λ_0 suffices because a DC algorithm is not sensitive to an initial value (An and Tao, 1997). This choice leads to a piecewise linear and continuous minimizer $\hat{\boldsymbol{\beta}}^{(1)}(\boldsymbol{\lambda})$ of (4) in λ_1 given (λ_0, λ_2), where $\boldsymbol{\lambda} = (\lambda_0, \lambda_1, \lambda_2)^T$. Successively replacing $\hat{\beta}^{(m-1)}$ (λ_1, λ_2) in (4) by $\hat{\boldsymbol{\beta}}^{(m-1)}$ $(\lambda_0, \lambda_0, \lambda_2)$ for $m \in \mathbb{N}$, we obtain a modified version of (4):

$$S^{(m)}(\beta)=S_1(\beta) - S_2\left(\widehat{\beta}^{(m-1)}(\lambda_0, \lambda_0, \lambda_2)\right) - \left(\beta - \widehat{\beta}^{(m-1)}(\lambda_0, \lambda_0, \lambda_2)\right)^T \nabla S_2\left(\widehat{\beta}^{(m-1)}(\lambda_0, \lambda_0, \lambda_2)\right), \tag{5}$$

which yields its minimizer $\hat{\boldsymbol{\beta}}^{(m)}(\boldsymbol{\lambda})$ and the estimated grouping $\mathcal{G}^{(m)}(\boldsymbol{\lambda})$. As suggested by Theorems 1 and 2, $\hat{\boldsymbol{\beta}}^{(m)}(\boldsymbol{\lambda})$ converges in finite steps. Most importantly, the iterative scheme yields an estimator having the desired properties of a global minimizer, c.f., Theorem 3.

Given grouping $\mathcal{G} = (\mathcal{G}_1, \ldots, \mathcal{G}_K)$, let $\boldsymbol{Z}_{\mathcal{G}} = (\boldsymbol{z}_{\mathcal{G}_1}, \ldots, \boldsymbol{z}_{\mathcal{G}_K})$ be an $n \times K$ matrix with $\boldsymbol{z}_{\mathcal{G}_k} = \boldsymbol{X}_{\mathcal{G}_k}\boldsymbol{1}$, and $\boldsymbol{X}_{\mathcal{G}_k}$ be the design matrix spanned by the predictors of $\mathcal{G}_k$; $k = 1, \ldots, K$.

Theorem 1 Assume that $\boldsymbol{Z}^T_{\mathcal{G}^{(m)}(\boldsymbol{\lambda})}\boldsymbol{Z}_{\mathcal{G}^{(m)}(\boldsymbol{\lambda})}$ is invertible. Then $\hat{\boldsymbol{\beta}}^{(m)}(\boldsymbol{\lambda})$ defined by (5) is piecewise linear in $(\boldsymbol{Y}, \boldsymbol{\lambda})$ and continuous in λ_1. In addition,

$$\widehat{\alpha}^{(m)}(\boldsymbol{\lambda}) \equiv \left(\widehat{\alpha}_1^{(m)}(\boldsymbol{\lambda}), \ldots, \widehat{\alpha}^{(m)}_{K^{(m)}(\boldsymbol{\lambda})}(\boldsymbol{\lambda})\right)^T = \left(\boldsymbol{Z}^T_{\mathcal{G}^{(m)}(\boldsymbol{\lambda})}\boldsymbol{Z}_{\mathcal{G}^{(m)}(\boldsymbol{\lambda})}\right)^{-1}\left(\boldsymbol{Z}^T_{\mathcal{G}^{(m)}(\boldsymbol{\lambda})}\boldsymbol{Y} - n\lambda_1\delta^{(m)}(\boldsymbol{\lambda})\right), \tag{6}$$

where $\delta^{(m)}(\boldsymbol{\lambda}) \equiv \left(\delta_1^{(m)}(\boldsymbol{\lambda}), \ldots, \delta^{(m)}_{K^{(m)}(\boldsymbol{\lambda})}(\boldsymbol{\lambda})\right)^T, \delta_k^{(m)}(\boldsymbol{\lambda}) \equiv \sum_{j \in \mathcal{G}_k^{(m)}(\boldsymbol{\lambda})} \Delta_j^{(m)}(\boldsymbol{\lambda})$, and

$$\Delta_j^{(m)}(\boldsymbol{\lambda}) \equiv \sum_{j':j' \neq j} \left\{\text{sign}\left(\widehat{\beta}_j^{(m)}(\boldsymbol{\lambda}) - \widehat{\beta}_{j'}^{(m)}(\boldsymbol{\lambda})\right) - \nabla G_2\left(\widehat{\beta}_j^{(m-1)}(\lambda_0, \lambda_0, \lambda_2) - \widehat{\beta}_{j'}^{(m-1)}(\lambda_0, \lambda_0, \lambda_2)\right)\right\}. \tag{7}$$

Moreover, for $j \in \mathcal{G}_k^{(m)}(\boldsymbol{\lambda})$ with $|\mathcal{G}_k^{(m)}(\boldsymbol{\lambda})| \geq 2$, and $k = 1, \ldots, K^{(m)}(\boldsymbol{\lambda})$,

$$B_j^{(m)}(\boldsymbol{\lambda}) = \frac{1}{n\lambda_1}\boldsymbol{x}_j^T\left(\boldsymbol{I} - \boldsymbol{Z}_{\mathcal{G}^{(m)}(\boldsymbol{\lambda})}\left(\boldsymbol{Z}^T_{\mathcal{G}^{(m)}(\boldsymbol{\lambda})}\boldsymbol{Z}_{\mathcal{G}^{(m)}(\boldsymbol{\lambda})}\right)^{-1}\boldsymbol{Z}^T_{\mathcal{G}^{(m)}(\boldsymbol{\lambda})}\right)\boldsymbol{Y} + \boldsymbol{x}_j^T\boldsymbol{Z}_{\mathcal{G}^{(m)}(\boldsymbol{\lambda})}\left(\boldsymbol{Z}^T_{\mathcal{G}^{(m)}(\boldsymbol{\lambda})}\boldsymbol{Z}_{\mathcal{G}^{(m)}(\boldsymbol{\lambda})}\right)^{-1}\delta^{(m)}(\boldsymbol{\lambda}) - \Delta_j^{(m)}(\boldsymbol{\lambda}). \tag{8}$$

Theorem 1 reveals two important aspects of $\hat{\boldsymbol{\beta}}^{(m)}(\boldsymbol{\lambda})$ from (5). First, $\hat{\boldsymbol{\alpha}}^{(m)}(\boldsymbol{\lambda})$ and $B_j^{(m)}(\boldsymbol{\lambda})$ are continuous and piecewise linear in λ_1 and λ_1^{-1}, piecewise constant in λ_2 with possible jumps, and piecewise linear in $\boldsymbol{Y}$ with possible jumps. In other words, $(\widehat{\alpha}^{(m)}(\boldsymbol{\lambda}), B_j^{(m)}(\boldsymbol{\lambda}))$ is continuous in λ_1 given $(\boldsymbol{Y}, \lambda_0, \lambda_2)$, but may contain jumps with respect to $(\boldsymbol{Y}, \lambda_2)$. Second, $J(\cdot)$ shrinks $\hat{\boldsymbol{\beta}}^{(m)}(\boldsymbol{\lambda})$ towards the least squares estimate $\left(\widehat{\alpha}_1^0, \ldots, \widehat{\alpha}^0_{K^{(m)}(\boldsymbol{\lambda})}\right)^T$, with the amount of shrinkage controlled by $\lambda_1 > 0$. This occurs only when a pairwise difference stays below the thresholding value of λ_2. As $\lambda_2 \to 0$, the bias due to penalization becomes ignorable, yielding a nearly unbiased estimate for grouping and parameter estimation.

2.3 Algorithms for difference convex solution surface

One efficient computational tool is a homotopy method (Allgower and Georg, 2003; Wu et al., 2009), which utilizes continuity of a solution in $\boldsymbol{\lambda}$ to compute the entire solution surface simultaneously. To our knowledge, homotopy methods for nonconvex problems have not yet received attention in the literature. This section develops a homotopy method for a regularization solution surface for nonconvex minimization (2) through DC programming. One major computational challenge is that the solution may be piecewise linear with jumps in $(\boldsymbol{Y}, \boldsymbol{\lambda})$, which is difficult to treat with homotopy. To overcome this difficulty, we design a DC algorithm to obtain an easily computed solution $\hat{\boldsymbol{\beta}}(\boldsymbol{\lambda})$, which could be local or global. Note that a DC method guarantees a global solution when it is combined with the branch-

and-bound method, c.f., Liu, Shen and Wong (2005). However, seeking a global minimizer of (2) is unnecessary, because the DC solution has the desired statistical properties for grouping (Theorem 3), and can be computed more efficiently (Theorem 2).

The main gradients of our DC homotopy algorithm are (1) iterating the entire DC solution surfaces, (2) utilizing the piecewise linear continuity of $(\widehat{\alpha}^{(m)}(\boldsymbol{\lambda}), B_j^{(m)}(\boldsymbol{\lambda}))$ in $(\lambda_1, \lambda_1^{-1})$ for given $(\lambda_0, \lambda_2, m)$, and (3) tracking transition points (joints for a piecewise linear function) through the grouped subdifferentials. This algorithm permits efficient computation of a DC solution for nonconvex minimization (2) with an overcomplete penalty, which is otherwise difficult to treat. To compute $\hat{\boldsymbol{\beta}}^{(m)}(\boldsymbol{\lambda})$, we proceed as follows. First, we fix at one evaluation point of (λ_0, λ_2), then move along path from $\lambda_1 = \infty$ towards $\lambda_1 = 0$ given (λ_0, λ_2). By Theorem 1, $\hat{\boldsymbol{\beta}}^{(m)}(\boldsymbol{\lambda})$ is piecewise linear and continuous in λ_1 given (λ_0, λ_2). Along this path, we compute transition points at which the derivative of $\hat{\boldsymbol{\beta}}^{(m)}(\boldsymbol{\lambda})$ with respect to λ_1 changes. Second, we move to other evaluation points of (λ_0, λ_2) and repeat the above process.

For given (λ_0, λ_2), transition in λ_1 occurs when either of the following conditions is met:

A. Merging: Groups $\mathcal{G}_l^{(m)}(\boldsymbol{\lambda})$ and $\mathcal{G}_k^{(m)}(\boldsymbol{\lambda})$ are combined at $\boldsymbol{\lambda}$, when $\widehat{\alpha}_l^{(m)}(\boldsymbol{\lambda})=\widehat{\alpha}_k^{(m)}(\boldsymbol{\lambda})$;

B. Splitting: Group $\mathcal{G}_k^{(m)}(\boldsymbol{\lambda})$ is split into two disjoint sets A_1 and A_2 with $A_1 \cup A_2=\mathcal{G}_k^{(m)}(\boldsymbol{\lambda})$ at $\boldsymbol{\lambda}$, when $|B_{A_1}^{(m)}(\boldsymbol{\lambda})|=|B_{A_2}^{(m)}(\boldsymbol{\lambda})|=|A_1||A_2|$, according to (3).

In (A) and (B), at a transition point, two or more groups may be merged, and a single group may be split into two or multiple subgroups.

We now describe the basic idea for computing $\{\hat{\boldsymbol{\beta}}^{(m)}(\boldsymbol{\lambda}) : \lambda_1 > 0\}$ given $(\lambda_0, \lambda_2, m)$. From (6), we track $(\mathcal{G}^{(m)}(\boldsymbol{\lambda}), \Delta_j^{(m)}(\boldsymbol{\lambda}))$ along the path from $\lambda_1 = \infty$ towards $\lambda_1 = 0$ for given (λ_0, λ_2). Let $\lambda_1^*>0$ be the current transition point. Our algorithm successively identifies the next transition point λ_1^{**} along the path. For notational ease, we use $(\mathcal{G} = (\mathcal{G}_1, \ldots, \mathcal{G}_K), \Delta_j)$ to denote the current $(\mathcal{G}^{(m)}(\boldsymbol{\lambda}), \Delta_j^{(m)}(\boldsymbol{\lambda}))$ after the current transition; $j = 1, \ldots, p$. Note that $(\mathcal{G}^{(m)}(\boldsymbol{\lambda}), \Delta_j^{(m)}(\boldsymbol{\lambda}))$; $j = 1, \ldots, p$, remain unchanged before the next transition is reached.

For merging in (A), we compute potential merge points:

$$n(\boldsymbol{e}_k - \boldsymbol{e}_l)^T \left(\boldsymbol{Z}_{\mathcal{G}}^T \boldsymbol{Z}_{\mathcal{G}}\right)^{-1} \delta, \quad \delta=(\delta_1, \ldots, \delta_K)^T, \quad \delta_k=\sum_{j:\, j\in\mathcal{G}_k} \Delta_j; \tag{9}$$

$1 \le k < l \le K$, where $\boldsymbol{e}_k$ is the kth column of $\boldsymbol{I}_p$. Then

$$\lambda_{1,\mathcal{A}} = \max \left\{m_{kl} \in (0, \lambda_1^*]:1 \le k<l \le K\right\} \tag{10}$$

is a potential transition point at which $\mathcal{G}_{k'}$ and $\mathcal{G}_{l'}$ are combined into one group. Define $\lambda_{1,\mathcal{A}}=0$ if $\left\{m_{kl} \in (0, \lambda_1^*):1 \le k<l \le K\right\}=\varnothing$, with $\varnothing$ denoting the empty set.

For splitting in (B), we utilize (3) and the subdifferentials in (8), to compute, for each $k = 1, \ldots, K$, the largest $\lambda_1 \in (\lambda_{1,\mathcal{A}}, \lambda_1^*]$ and $A \subset \mathcal{G}_k$ with $|A| < |\mathcal{G}_k|/2$ such that

$$L_k^+(\lambda_1, A)L_k^-(\lambda_1, A)=0, \tag{11}$$

where $L_k^{\pm}(\lambda_1, A) \equiv \sum_{j\in A} \left\{\frac{1}{n\lambda_1}\xi_j+\eta_j\right\} \mp |A|\,(|\mathcal{G}_k| - |A|)$, $\eta_j \equiv \boldsymbol{x}_j^T \boldsymbol{Z}_{\mathcal{G}} \left(\boldsymbol{Z}_{\mathcal{G}}^T \boldsymbol{Z}_{\mathcal{G}}\right)^{-1}\delta - \Delta_j$, and $\xi_j \equiv \boldsymbol{x}_j^T \left(I - \boldsymbol{Z}_{\mathcal{G}}\left(\boldsymbol{Z}_{\mathcal{G}}^T \boldsymbol{Z}_{\mathcal{G}}\right)^{-1} \boldsymbol{Z}_{\mathcal{G}}^T\right)\boldsymbol{Y}$. It follows from (8) that $B_j^{(m)}(\boldsymbol{\lambda})=\frac{1}{n\lambda_1}\xi_j+\eta_j$ before the next transition occurs, and hence that $L_k^{\pm}(\lambda_1, \mathcal{G}_k)=\sum_{j\in\mathcal{G}_k} B_j^{(m)}(\boldsymbol{\lambda})=0$ for any $\lambda_1 \in \mathbb{R}$. Unfortunately, solving (11) through enumeration is infeasible over all subsets $A \subset \mathcal{G}_k$. In Algorithm 1 below, we develop an efficient strategy utilizing piecewise linearity of $L_k^{\pm}(\lambda_1, A)$ in λ_1^{-1} for computing the potential transition point $\lambda_{1,B} \equiv \max\left\{s_k \in \left(\lambda_{1,A}, \lambda_1^*\right] : k=1, \ldots, K\right\}$, as well as its corresponding grouping, where s_k is the solution of (11); $k = 1, \ldots, K$. This strategy requires roughly $O(p^2 \log p)$ operations, c.f. Proposition 1.

To describe our strategy for solving (11), let $A_{k\ell}^+(\lambda_1)$ and $A_{k\ell}^-(\lambda_1)$ be the two subsets of $\mathcal{G}_k$ of size ℓ corresponding to the ℓ largest and smallest values of $D_{k\ell}^+(\lambda_1) \equiv \left\{\frac{1}{n\lambda_1}\xi_j+\eta_j : \xi_j>0,\ j \in \mathcal{G}_k\right\}$ and $D_{k\ell}^-(\lambda_1) \equiv \left\{\frac{1}{n\lambda_1}\xi_j+\eta_j : \xi_j<0,\ j \in \mathcal{G}_k\right\}$, for $\ell = 1, \ldots, |\mathcal{G}_k|$ and $k = 1, \ldots, K$, where $A_{k\ell}^{\pm}(\lambda_1) \equiv A_{k,\ell-1}^{\pm}(\lambda_1)$ if $\left|D_{k\ell}^{\pm}(\lambda_1)\right| < \ell$. Since $L_k^{\pm}(\lambda_1^*, \mathcal{G}_k)=0$, we can refine our search to $\left\{A_{k\ell}^{\pm}(\lambda_1) : \ell=1, \ldots, [\mathcal{G}_k/2], \lambda_1 \in (\lambda_{1,A}, \lambda_1^*]\right\}$ for solving (11). Note that $L_k^+\left(\lambda_1, A_{k\ell}^+(\lambda_1)\right) \le 0$ and $L_k^-\left(\lambda_1, A_{k\ell}^-(\lambda_1)\right) \ge 0$ by the definition of $B_j^{(m)}(\boldsymbol{\lambda})$'s for $\lambda_1 \in \left[\lambda_1^{**}, \lambda_1^*\right]$. For $k = 1, \ldots, K$ and $\ell = 1, \ldots, [\mathcal{G}_k/2]$, we seek the first zero-crossing λ_1 values for $L_k^{\pm}(\lambda_1, A_{k\ell}^{\pm}(\lambda_1))$, denoted as $s_{k\ell}^{\pm}$, as λ_1 decreases from λ_1^*. For each $k = 1, \ldots, K$, we start with $\ell = 1$ and compute

$$s_{k1}^{\pm} = \max\left\{\frac{\xi_j}{\pm(|\mathcal{G}_k| - 1) - \eta_j} \in \left(\lambda_{1,A}, \lambda_1^*\right] : j=1, \ldots, p\right\}. \tag{12}$$

For $\ell = 2, \ldots, [\mathcal{G}_k/2]$, we observe that elements in $A_{k\ell}^{\pm}(\lambda_1)$ need to be updated as λ_1 decreases due to rank changes in $\left\{\frac{1}{n\lambda_1}\xi_j+\eta_j : j \in \mathcal{G}_k\right\}$. This occurs at switching points:

$$h_{jj'} \equiv \frac{\xi_j - \xi_{j'}}{\eta_{j'} - \eta_j}; \quad \text{at which } \frac{1}{n\lambda_1}\xi_j+\eta_j=\frac{1}{n\lambda_1}\xi_{j'}+\eta_{j'}; \quad 1 \le j<j' \le p. \tag{13}$$

On this basis, $s_{k\ell}^{\pm}$ can be computed. First, calculate the largest $\lambda_1 \in \{\lambda_{1,A}\} \cup \{h_{jj'} : 1 \le j < j' \le p\}$ at which $L_k^+(\lambda_1, A_{k\ell}^+(\lambda_1)) \ge 0$ or $L_k^-(\lambda_1, A_{k\ell}^-(\lambda_1)) \le 0$. Second, compute the exact crossing point for $L_k^{\pm}(\lambda_1, A_{k\ell}^{\pm}(\lambda_1))=0$ through linear interpolation due to the fact that $L_k^{\pm}(\lambda_1, A_{k\ell}^{\pm}(\lambda_1))$ is piecewise linear and continuous in λ_1^{-1} with joints at $h_{jj'}$, $1 \le j < j' \le p$.

Algorithm 1 computes the next transition point, $\mathcal{G}$ and Δ_j's.

Algorithm 1: Computation of next transition point—Given the current grouping $\mathcal{G}$ and Δ_j's with invertible $\boldsymbol{Z}_{\mathcal{G}}^T \boldsymbol{Z}_{\mathcal{G}}$,

Step 1 (Potential transition for merging) Compute $\lambda_{1,A}$ as defined in (10) as well as the corresponding grouping.

Step 2 Compute $s_{k1}^{\pm}$; $k = 1, \dots, K$, as defined in (12), and $H \equiv \{\lambda_{1,A}\} \cup \{h_{jj'} : 1 \le j < j' \le p\}$ based on (13).

Step 3 (Splitting points) Starting with $\ell = 1$, and for $k = 1, \dots, K$, we

- compute the largest $\lambda_1 \in H$ such that $L_k^+(\lambda_1, A_{k\ell}^+(\lambda_1)) \ge 0$, and the largest $\lambda_1 \in H$ such that $L_k^-(\lambda_1, A_{k\ell}^-(\lambda_1)) \le 0$, where bisection or Fibonacci search (e.g., Gill, Murray and Wright, 1981) may be applied;
- interpolate $L_k^+(\lambda_1, A_{k\ell}^+(\lambda_1))$ (resp. $L_k^-(\lambda_1, A_{k\ell}^-(\lambda_1))$) linearly to obtain $s_{k\ell}^+$ (resp. $s_{k\ell}^-$) satisfying $L_k^+\left(s_{k\ell}^+, A_{k\ell}^+(s_{k\ell}^+)\right)=0$ (resp. $L_k^-\left(s_{k\ell}^-, A_{k\ell}^-(s_{k\ell}^-)\right)=0$); set $s_{k\ell}^+=0$ (resp. $s_{k\ell}^-=0$) if $\max_{\lambda_1 \in H} L_k^+(\lambda_1, A_{k\ell}^+(\lambda_1))<0$ (resp. $\max_{\lambda_1 \in H} L_k^-(\lambda_1, A_{k\ell}^-(\lambda_1))>0$);
- compute $s_{k\ell}=\max\left\{s_{k\ell}^+, s_{k\ell}^-\right\}$ and the corresponding index set;
- if $s_{k,\ell-1} > s_{k,\ell}$ and $\ell \le [\mathcal{G}_k/2] - 1$, then go to Step 3 with ℓ replaced by $\ell + 1$. Otherwise, set $s_{k,\ell+1} = \cdots = s_{k,[\mathcal{G}_k/2]} = 0$.

Step 4 (Potential transition for splitting) Compute $\lambda_{1,B}=\max\left\{s_{k\ell} \in \left[0, \lambda_1^*\right] : 1<k<K, l=1, \dots, [\mathcal{G}_k/2]\right\}$, as well as the corresponding grouping.

Step 5 (Transition) Compute $\lambda_1^{**}=\max\left\{\lambda_{1,A}, \lambda_{1,B}\right\}$. If $\lambda_1^{**}=\lambda_{1,A}>0$, two groups are merged at λ_1^{**}. If $\lambda_1^{**}=\lambda_{1,B}>0$, a group is split into two at λ_1^{**}. Update $\mathcal{G}$ and Δ_j's. If $\lambda_1^{**}=0$, no further transition can be obtained.

Algorithm 2: Main algorithm

Step 1 (Parameter initialization): Specify the upper bound parameter K^*, and evaluation points of (λ_0, λ_2), where $K^* \le \min\{n, p\}$.

Step 2 (Initialization for DC iterations): Given (λ_0, λ_2), compute $\hat{\boldsymbol{\beta}}^{(0)}(\lambda_0, +\infty, \lambda_2)$, the corresponding $\mathcal{G}$ and Δ's by solving (5) with $m = 0$. Compute $\hat{\boldsymbol{\beta}}^{(0)}(\boldsymbol{\lambda})$ along the path from $\lambda_1 = \infty$ to $\lambda_1 = 0$ until $|\mathcal{G}^{(0)}(\boldsymbol{\lambda})| = K^*$ using Algorithm 1 while holding (λ_0, λ_2) fixed.

Step 3 (DC iterations): Starting from $m = 1$, compute $\hat{\boldsymbol{\beta}}^{(m)}(\lambda_0, +\infty, \lambda_2)$, the corresponding $\mathcal{G}$ and Δ's by solving (5); then successively compute $\hat{\boldsymbol{\beta}}^{(m)}(\boldsymbol{\lambda})$ along the path from $\lambda_1 = \infty$ to $\lambda_1 = 0$ until $|\mathcal{G}| = K^*$ using Algorithm 1.

Step 4 (Stopping rule): If $S(\hat{\boldsymbol{\beta}}^{(m-1)}(\boldsymbol{\lambda})) - S(\hat{\boldsymbol{\beta}}^{(m)}(\boldsymbol{\lambda})) = 0$, then go to Step 3 with m replaced by $m + 1$. Otherwise, move to next evaluation point of (λ_0, λ_2) and go to Step 2 until all the evaluation points have been computed.

Denote by m^* the termination step of Algorithm 2, which may depend on (λ_0, λ_2). Our estimate DCE of $\boldsymbol{\beta}$ is $\hat{\boldsymbol{\beta}}(\boldsymbol{\lambda}) \equiv \hat{\boldsymbol{\beta}}^{(m^*)}(\boldsymbol{\lambda})$. The corresponding grouping of $\hat{\boldsymbol{\beta}}(\boldsymbol{\lambda})$ is $\mathcal{G}(\boldsymbol{\lambda})$. In practice, Algorithm 2 is applicable to unstandardized or standardized predictors.

Proposition 1 (Computational properties). For Algorithm 1, $L_k^{\pm}(\lambda_1, A_{k\ell}^{\pm}(\lambda_1))$ is continuous, piecewise linear, and strictly monotone in λ_1; $k = 1, \ldots, K$, $\ell = 1, \ldots, [\mathcal{G}_k/2]$. Moreover, the computational complexities of Algorithms 1 and 2 are no greater than p^2 (log $p + n$) and $O(m^* n^* p^2(\log p + n))$, where n^* is the number of transition points.

In general, it is difficult to bound n^* precisely. However, an application of a heuristic argument similar to that of Rosset and Zhu (2007, Section 3.2, p. 1019) suggests that n^* is $O(\min\{n, p\})$ on average for group combining and splitting.

Theorem 2 (Computation). Assume that $\boldsymbol{Z}^T_{\mathcal{G}^{(m)}(\boldsymbol{\lambda})}\boldsymbol{Z}_{\mathcal{G}^{(m)}(\boldsymbol{\lambda})}$ is invertible for $m \le m^*$. Then the solution of Algorithm 2 is unique, and sequence $S(\hat{\boldsymbol{\beta}}^{(m)}(\lambda_0, \lambda_0, \lambda_2))$ decreases strictly in m unless $\hat{\boldsymbol{\beta}}^{(m)}(\lambda_0, \lambda_0, \lambda_2) = \hat{\boldsymbol{\beta}}^{(m-1)}(\lambda_0, \lambda_0, \lambda_2)$. In addition, Algorithm 2 terminates in finite steps, i.e., $m^* < \infty$ with

$$\widehat{\beta}^{(m)}(\boldsymbol{\lambda}) = \widehat{\beta}^{(m^*-1)}(\boldsymbol{\lambda}) \tag{14}$$

for all $\boldsymbol{\lambda}$ over the evaluation region of $\boldsymbol{\lambda}$ and all $m \ge m^*$.

Two distinctive properties of $\hat{\boldsymbol{\beta}}^{(m^*)}(\boldsymbol{\lambda})$ are revealed by (14), leading to fast convergence and a sharp result for consistency of $\hat{\boldsymbol{\beta}}(\boldsymbol{\lambda})$. First, the stopping rule of Algorithm 2, which is reinforced at one fixed λ_0, controls the entire surface for all (λ_1, λ_2) simultaneously, owing to the replacement of $\hat{\beta}^{(m-1)}(\lambda_0, \lambda_2)$ by $\hat{\boldsymbol{\beta}}^{(m-1)}(\lambda_0, \lambda_0, \lambda_2)$ for iteration m in (4). Second, the iteration process terminates finitely with $\hat{\boldsymbol{\beta}}^{(m^*)}(\boldsymbol{\lambda})$ satisfying (14), because of the step function of $\nabla S_2(\hat{\boldsymbol{\beta}}^{(m-1)}(\boldsymbol{\lambda}))$ resulted from the locally concave points $z = \pm\lambda_2$ of $G(z)$. Most critically, (14) is not expected for any penalty that is not piecewise linear with non-differentiable but continuous points.

3 Estimation of tuning parameters and σ^2

Selection of tuning parameters is important for DCE, which involves $\boldsymbol{\lambda} = (\lambda_0, \lambda_1, \lambda_2)$. In (1), predictive performance of estimator $\hat{\boldsymbol{\beta}}(\boldsymbol{\lambda})$ is measured by MSE($\hat{\boldsymbol{\beta}}(\boldsymbol{\lambda})$), defined as $\frac{1}{n}EL\left(\widehat{\beta}(\boldsymbol{\lambda}), \beta^0\right)$, where $L\left(\widehat{\beta}(\boldsymbol{\lambda}), \beta^0\right) = \sum_{i=1}^{n}\left(\widehat{\mu}(\boldsymbol{\lambda}, \boldsymbol{x}_i) - \mu^0(\boldsymbol{x}_i)\right)^2$, $\hat{\mu}(\boldsymbol{\lambda}, \boldsymbol{x}) = \hat{\boldsymbol{\beta}}^T(\boldsymbol{\lambda})\boldsymbol{x}$, and $\mu^0(\boldsymbol{x}) = (\boldsymbol{\beta}^0)^T\boldsymbol{x}$.

One critical aspect of tuning is that DCE is a piecewise continuous estimator with jumps in $\boldsymbol{Y}$. Therefore, any model selection routine may be used for tuning of DCE, which allows for estimators with discontinuities. For instance, cross-validation and the generalized degrees of freedom (GDF, Shen and Huang, 2006) are applicable, but Stein's unbiased risk estimator (Stein, 1981) is not suited because of the requirement of continuity. Then the tuning parameters are estimated by minimizing the model selection criterion.

In practice, σ^2 needs to be estimated when it is unknown. In the literature, there have been many proposals for the case of $p < n$, for instance, σ^2 can be estimated by the residual sum squares over $(n - p)$. In general, estimation of σ^2 in the case of $p > n$ has not yet received much attention. In our case, we propose a simple estimator $\widehat{\sigma}^2 = \frac{1}{n - K^*/2}\sum_{i=1}^{n}(Y_i - \widehat{\mu}_i(\boldsymbol{\lambda}^*))^2$, where $\boldsymbol{\lambda}^* = (\lambda_0, \tilde{\lambda}_1, \infty)$, $\tilde{\lambda}_1$ is the smallest λ_1 reaching the upper bound $K(\lambda_0, \lambda_1, \infty) = K^*/2$, and K^* is defined in Step 2 of Algorithm 2. Note that when $\lambda_2 = \infty$, $\hat{\mu}_i(\boldsymbol{\lambda}^*)$ is independent of λ_0. The quality of estimation depends on the bias of $\hat{\mu}_i(\boldsymbol{\lambda}^*)$ as well as the variance. By choosing a tight value of $K^* \ge K_0$, one may achieve good performance.

4 Theory

This section derives a finite-sample probability error bound, based on which we prove that $\hat{\boldsymbol{\beta}}(\boldsymbol{\lambda})$ is consistent with regard to grouping pursuit and predictive optimality simultaneously for the same set of values of $\boldsymbol{\lambda}$. As a result, the true grouping $\mathcal{G}^0$ is reconstructed, as well as the unbiased least squares estimate $\widehat{\beta}^{(ols)} \equiv \left(\widehat{\beta}_1^{(ols)}, \ldots, \widehat{\beta}_p^{(ols)}\right)^T = \left(\widehat{\alpha}_1^{(ols)} \mathbf{1}_{|\mathcal{G}_1^0|}, \ldots, \widehat{\alpha}_{K^0}^{(ols)} \mathbf{1}_{|\mathcal{G}_{K^0}^0|}\right)^T$ given $\mathcal{G}^0$. Here $\widehat{\alpha}^{(ols)} \equiv \left(\widehat{\alpha}_1^{(ols)}, \ldots, \widehat{\alpha}_{K^0}^{(ols)}\right)^T = \left(\boldsymbol{Z}_{\mathcal{G}^0}^T \boldsymbol{Z}_{\mathcal{G}^0}\right)^{-1} \boldsymbol{Z}_{\mathcal{G}^0}^T \boldsymbol{Y}$ with $\boldsymbol{Z}_{\mathcal{G}^0}^T \boldsymbol{Z}_{\mathcal{G}^0}$ being invertible.

Denote by $c_{\min}(\mathcal{G}) > 0$ the smallest eigenvalue of $\boldsymbol{Z}_{\mathcal{G}}^T \boldsymbol{Z}_{\mathcal{G}}/n$, where $\boldsymbol{Z}_{\mathcal{G}}$ is the design matrix based on grouping $\mathcal{G}$ Denote by $\gamma_{\min} \equiv \min\{|\alpha_k^0 - \alpha_l^0|>0:1 \le k<l \le K^0\}$, the resolution level that may depend on (p, n), or the level of difficulty of grouping pursuit, with a small value of $\gamma_{\min}$ being difficult. The following result is established for $\hat{\boldsymbol{\beta}}(\boldsymbol{\lambda})$ from Algorithm 2.

Theorem 3 (Error bounds for grouping pursuit and consistency). Under the model assumptions of (1) with $\varepsilon_i \sim N(0, \sigma^2)$, assume that $\lambda_0 = \lambda_1$, $(2K^* + 1)\lambda_1/\lambda_2 < \min_{|\mathcal{G}|\le(K^*)^2} c_{\min}(\mathcal{G})$, where $K_0<K^* \le \min\{\sqrt{n}, p\}$. Then for any n and p, we have

$$\begin{aligned} &P(\mathcal{G}(\boldsymbol{\lambda}) \ne \mathcal{G}^0) \le P(\widehat{\beta}(\boldsymbol{\lambda}) \ne \widehat{\beta}^{(ols)}) \\ &\le \frac{K^0(K^0-1)}{2}\Phi\left(\frac{-n^{1/2}(\gamma_{\min}-3\lambda_2/2)}{2\sigma c_{\min}^{-1/2}(\mathcal{G}^0)}\right)+p\Phi\left(\frac{-n\lambda_1}{\sigma\max_{1\le j\le p}\|\boldsymbol{x}_j\|}\right), \end{aligned} \tag{15}$$

where $\Phi(z)=\int_{-\infty}^{z}\exp(-u^2/2)du$ is the cumulative distribution function of $N(0, 1)$, and $\|\boldsymbol{x}_j\|$ is the L_2-norm of $\boldsymbol{x}_j \in \mathcal{R}^n$. Moreover, as $p, n \to +\infty$, if

i.
$$\frac{n(\gamma_{\min} - 3\lambda_2/2)^2}{8c_{\min}(\mathcal{G}^0)\sigma^2} - 2\log K^0 \to \infty, \quad 0<\lambda_2<\frac{2}{3}\gamma_{\min},$$

ii.
$$\frac{n\lambda_1^2}{2\sigma^2\max_{1\le j\le p}\|\boldsymbol{x}_j\|^2/n} - \log p \to \infty,$$

then $P(\mathcal{G}(\boldsymbol{\lambda}) \ne (\mathcal{G}^0) \le P(\hat{\boldsymbol{\beta}}(\boldsymbol{\lambda}) \ne \hat{\boldsymbol{\beta}}^{(ols)}) \to 0$. In other words, $\mathcal{G}(\boldsymbol{\lambda}) = \mathcal{G}^0$ and $\hat{\boldsymbol{\beta}}(\boldsymbol{\lambda}) = \hat{\boldsymbol{\beta}}^{(ols)}$ with probability tending to 1.

Corollary 1 (Predictive performance) Under the assumption of Theorem 3, $\frac{L(\widehat{\beta}(\boldsymbol{\lambda}),\beta^0)}{L(\widehat{\beta}^{(ols)},\beta^0)} \to 1$, and $L(\hat{\boldsymbol{\beta}}(\boldsymbol{\lambda}), \boldsymbol{\beta}^0) = O_p(K_0/n)$, as $p, n \to \infty$.

Theorem 3 and Corollary 1 say that DCE consistently identifies the true grouping $\mathcal{G}^0$ and reconstructs the unbiased least squares estimator $\hat{\boldsymbol{\beta}}^{(ols)}$ based on $\mathcal{G}^0$ when $p, n \to \infty$. They also confirm the assertion made in the Introduction for consistency with regard to nearly exponentially many predictors in n. Specifically, consistency occurs when (a) $p=O(\exp(n\lambda_1^2))$ (Condition (ii)) or $\frac{\log p}{n} \to 0$, (b) $\lambda_1(2K^* + 1) < \lambda_2\min_{|\mathcal{G}|\le(K^*)^2} c_{\min}(\mathcal{G})$, (c) $0<\lambda_2<\frac{2}{3}\gamma_{\min}$ and $nc_{\min}^{-1}(\mathcal{G}^0)(\gamma_{\min} - 3\lambda_2/2)^2 - 16\sigma^2\log K^0 \to \infty$ (Condition (i)), provided that $\max_{j:1\le j\le p} \|\boldsymbol{x}_j\|^2/n$ is bounded. Note that $c_{\min}(\mathcal{G}^0)$ may tend to zero as $p, n \to \infty$ even if the number of true groups K^0 is independent of (p, n). To understand the conditions (a)-(c), we examine the simplest case in which $\min_{|\mathcal{G}|\leq(K^*)^2} c_{\min}(\mathcal{G})$, K^0 and K^* are independent of (p, n). Then (a)-(c) reduce to that $p = O(\exp(n\lambda_1^2))$, $n^{1/2}\lambda_1 \to \infty$, and $\lambda_2 > c_1\lambda_1$ but $\lambda_2 \leq \frac{3}{2}\gamma_{\min} - d_n$ for some constant $c_1 > 0$ and some sequence $d_n > 0$ with $n^{1/2}d_n \to \infty$, where the resolution level $\gamma_{\min}$ is required to be not too low in that $n^{1/2}\gamma_{\min} \to \infty$. Interestingly, there is a trade-off between p and the resolution level $\gamma_{\min}$. Note that $p = O(\exp(n^{2\delta}))$ when $\boldsymbol{\lambda}$ is tuned: $\lambda_1 = c_2\lambda_2 = c_3\gamma_{\min}$, given that $\gamma_{\min} = O(n^{-1/2+\delta})$, for some positive constants c_2, $c_3 > 0$ and $0 < \delta \leq 1/2$. Depending on the value of δ or $\gamma_{\min}$, p can be nearly exponentially many for high-resolution regression functions with $\delta = 1/2$, whereas p can be $\Delta O(\exp(n^{2\delta}))$ for low-resolution functions when δ is close to 0. The resolution level for DCE can be as low as nearly $O(n^{-1/2})$, which, to our knowledge, compares favorably with existing penalties for feature selection.

We now describe characteristics of grouping, in particular—how predictors are grouped. Denote by $\rho_j(\boldsymbol{\lambda}) \equiv \boldsymbol{x}_j^T\left(\boldsymbol{Y} - \boldsymbol{X}^T\widehat{\beta}(\boldsymbol{\lambda})\right)$, the sample covariance between $\boldsymbol{x}_j$ and the residual.

Theorem 4 (Grouping). Let $\Delta_j(\boldsymbol{\lambda}) = \Delta_j^{(m^*)}(\boldsymbol{\lambda})$ be defined in Theorem 1, where $\Delta_j(\boldsymbol{\lambda}) = \Delta_{j'}(\boldsymbol{\lambda})$ if $j, j' \in \mathcal{G}_k(\boldsymbol{\lambda})$. Let $E_k(\boldsymbol{\lambda}) \equiv [\Delta_j(\boldsymbol{\lambda}) - (|\mathcal{G}_k(\boldsymbol{\lambda})| - 1), \Delta_j(\boldsymbol{\lambda}) + (|\mathcal{G}_k(\boldsymbol{\lambda})| - 1)]$ be an interval or a point for any $j \in \mathcal{G}_k(\boldsymbol{\lambda})$. Then $E_1(\boldsymbol{\lambda}), \ldots, E_{K(\boldsymbol{\lambda})}(\boldsymbol{\lambda})$ are disjoint. Finally, j belongs to $\mathcal{G}_k(\boldsymbol{\lambda})$ if and only if $\frac{1}{n\lambda_1}\rho_j(\boldsymbol{\lambda}) \in E_k(\boldsymbol{\lambda})$; $k = 1, \ldots, K(\boldsymbol{\lambda})$.

Theorem 4 says that predictors are grouped according to if their sample covariance values fall into the same intervals, where these disjoint intervals $E_1(\boldsymbol{\lambda}), \ldots, E_{K(\boldsymbol{\lambda})}(\boldsymbol{\lambda})$ characterize grouping. As $\boldsymbol{\lambda}$ varies, group splitting or combining may take place when these intervals split or combine.

5 Numerical examples

This section examines effectiveness of the proposed method on three simulated examples and one real application to gene network analysis.

For a fair comparison, we compare DCE with the estimator obtained from its convex counterpart–an ultra-fused version of the fused Lasso based on convex penalty $\Sigma_{j<j'} |\beta_j - \beta_{j'}|$. This allows us to understand the role λ_2 plays in grouping. In addition, we examine the Lasso to investigate the connection between grouping pursuit and feature selection, to confirm our intuition in the foregoing discussion. For reference, least squares estimators based on the full model, and the true grouping, are reported as well, in addition to the average number of iterations in Algorithm 2. Finally, we compare DCE with OSCAR using two examples from Bondell and Reich (2008) in Example 3.

5.1 Benchmarks

We perform simulations in several scenarios, including correlated predictors, different noise levels and situations of "small p but large n" and "small n but large p". Note that a decrease in the value of σ^2 implies an increase in the sample size in this case. We therefore fix $n = 50$ and vary the value of σ^2 in Examples 1 and 2 and use different sample sizes in Example 3.

For estimating $\boldsymbol{\lambda}$ for DCE, we generate an independent tuning set $(\boldsymbol{x}_i, \tilde{y}_i)_{i=1}^n$ of size n in each example. Specially, the estimated $\boldsymbol{\lambda}$, denoted by $\hat{\boldsymbol{\lambda}}$, is obtained by minimizing tuning error $n^{-1}\sum_{i=1}^n \left(\tilde{y}_i - \widehat{\mu}(\boldsymbol{\lambda}, \boldsymbol{x}_i)\right)^2$ over the tuning set with respect to $\boldsymbol{\lambda}$ over the path of $\lambda_1 > 0$,

$\lambda_0 \in \{i\lambda_1^*/10{:}i{=}1,\ldots,10\}$ and $\lambda_2 \in \{0.5, 1, 1.5, 2, 2.5, 3, 5, 10, \infty\}$ in Examples 1-3, where λ_1^* is the largest transition point corresponding to $\lambda_2 = \infty$. The predictive performance is evaluated by MSE($\hat{\boldsymbol{\beta}}(\hat{\boldsymbol{\lambda}})$) as defined in Section 3. The accuracy of grouping is measured by the percentage of matchings between estimated and true pairs of indices (j, j'), for $j \neq j'$. Similarly, the tuning parameter of Lasso and the convex counterpart of DCE is estimated over $\lambda_1 > 0$ based on the same tuning set for a fair comparison. All numerical analyses are conducted in R 2.9.1.

Example 1 (Sparse Grouping)—This example was used previously for feature selection in Zou and Hastie (2005). A random sample of $\{(\boldsymbol{x}_i, Y_i) : i = 1, \ldots, n\}$ is obtained with n = 50, where Y_i follows (1) with $\varepsilon_i \sim N(0, \sigma^2)$ and σ^2 = 2, 1, .5, and $\boldsymbol{x}_i$ is sampled from $N(\mathbf{0}, \boldsymbol{\Sigma}_{40\times 40})$ with p = 20, having the diagonal and off-diagonal elements 1 and 0.5. Here

$$\beta=(\underbrace{0,\ldots,0}_{5},\underbrace{2,\ldots,2}_{5},\underbrace{0,\ldots,0}_{5},\underbrace{2,\ldots,2}_{5})^T.$$

As suggested in Table 1, DCE outperforms its convex counterpart and the least squares estimate based on estimated grouping across three levels of σ^2. This is due primarily to the nonconvex penalty, which corrects the estimation bias due to use of its convex counterpart. This is evident from the fact that the convex counterpart of DCE (m = 0) performs worst than the least square estimates based on the estimated grouping. Furthermore, grouping indeed offers additional improvement in predictive performance in view of the result of the Lasso. Most importantly, DCE reconstructs the least square estimates based on true grouping well, confirming the asymptotic results in Theorem 3 and Corollary 1. The reconstruction is nearly perfectly when σ^2 = .5 but is less so when the value of σ^2 increases towards 2, indicating that the noise level does impact the accuracy of reconstruction. Overall, grouping identification and reconstruction appear to be accurate, agreeing with the theoretical results.

Figure 1 displays the paths in λ_1 given various values of λ_2 with λ_0 = .2. Clearly, $\hat{\boldsymbol{\beta}}(\boldsymbol{\lambda})$ is continuous in λ_1 given (λ_0, λ_2) and has jumps in λ_2 given (λ_0, λ_1), as discussed early. Figure 2 shows four two-dimensional DCE solution surfaces for $(\hat{\beta}_1(\boldsymbol{\lambda}), \hat{\beta}_2(\boldsymbol{\lambda}), \hat{\beta}_{19}(\boldsymbol{\lambda}), \hat{\beta}_{20}(\boldsymbol{\lambda}))$ with respect to $\boldsymbol{\lambda}$. In Figure 2, the four estimates are close to their corresponding least squares estimates when either λ_1 or λ_2 becomes small. On the other hand, they tend to be close to each other when both λ_1 and λ_2 become large. Note that some jumps in λ_2 are visible for $\hat{\beta}_1(\boldsymbol{\lambda})$ around $\boldsymbol{\lambda}$ = (0.2, 2.5, 2.5).

Example 2 (Large *p* but small *n*)—A random sample of $\{(\boldsymbol{x}_i, Y_i) : i = 1, \ldots, n\}$ with n = 50, is obtained, where Y_i follows (2) with $\varepsilon_i \sim N(0, \sigma^2)$, σ = .41, .58, p = 50, 100, and $\boldsymbol{x}_i$ is sampled from $N(\mathbf{0}, \boldsymbol{\Sigma})$ with the (j, k)th element of $\boldsymbol{\Sigma}$ being $0.5^{|j-k|}$. Here

$$\beta=(\underbrace{3,\ldots,3}_{5},\underbrace{-1.5,\ldots,-1.5}_{5},\underbrace{1,\ldots,1}_{5},\underbrace{2,\ldots,2}_{5},\underbrace{0,\ldots,0}_{p-20})^T.$$

In this "large p but small n" example, DCE outperforms its convex counterpart with regard to predictive performance in all the cases, but the amounts of improvement vary. Here, the Lasso performs slightly better in some cases, which is expected because of the large group size of zero-coefficient predictors. Interestingly, the average number of iterations for Algorithm 2 is about 3 as compared to 4 in Example 1. Finally, the matching proportion for grouping is reasonably high.

Example 3 (Small *p* but large *n*)—Consider Examples 4 and 5 of Bondell and Reich (2008), which are low-dimensional with relatively large dimensions. Their Example 4 is the same as Example 1 except that n = 100, p = 40, $\sigma^2 = 15^2$ and

$\beta=(\underbrace{0,\ldots,0}_{10},\underbrace{2,\ldots,2}_{10},\underbrace{0,\ldots,0}_{10},\underbrace{2,\ldots,2}_{10})^T$. Their Example 5 has a similar setting, but with n = 50, p = 40, $\sigma^2 = 15^2$, and $\beta=(\underbrace{3,\ldots,3}_{15},\underbrace{0,\ldots,0}_{25})^T$, where $\boldsymbol{x}_i \sim N(\mathbf{0}, \boldsymbol{V})$, and $\boldsymbol{V}$ is a block-diagonal matrix with diagonal blocks $\mathbf{1}_5\mathbf{1}_5^T$, $\mathbf{1}_5\mathbf{1}_5^T$, $\mathbf{1}_5\mathbf{1}_5^T$ and $\boldsymbol{I}_{p-15}$. In addition, the first 15 components of $\boldsymbol{x}_i$ are added to independent noise distributed as $N(0, .16)$ to generate three equally important groups having pairwise correlations being around 0.85.

Overall DCE performs comparably with OSCAR in these noisy situations with $\sigma^2 = 15^2$, which performs better but worst, respectively in Examples 4 and 5 of Bondell and Reich (2008). This is mainly due to the fact that DCE does not select variables beyond grouping. This aspect was also evident from Table 2, where DCE performs worse than Lasso in some cases. Most noticeably, DCE performs slightly worse than its convex counterpart when σ is very large. This is expected because an adaptive method tends to perform worse than its non-adaptive counterpart in a noisy situation.

5.2 Breast cancer metastasis and gene network

Mapping the pathways giving rise to metastasis is important in breast cancer research. Recent studies suggest that gene expression profiles are useful in identifying gene subnetworks correlated with metastasis. Here we apply our proposed method to understand the functionality of subnetworks of genes for predicting the time to metastasis, which may provide novel hypotheses and confirm the existing theory for pathways involved in tumor progression.

The breast cancer metastasis data (Wang et al., 2005; Chuang et al., 2007) contain gene expression levels of 8141 genes for 286 patients, 107 of whom were detected to develop metastasis within a five year follow-up after surgery. To utilize the present gene network knowledge, we explore the PPI network previously constructed in Chuang et al. (2007).

For breast metastasis, three tumor suppressor genes–TP53, BRACA1 and BRACA2, are known to be crucial in preventing uncontrolled cell proliferation and repairing the chromosomal damage. Certain mutations of these genes increase risk of breast center, c.f., Soussi (2003). In our analysis, we construct a subnetwork of Chuang et al. (2007), consisting of genes TP53, BRCA1 and BRCA2, as well as genes that were regulated by them. This leads to 294 expressed genes for 107 patients who developed metastasis.

For subnetwork analysis, consider a vector of p predictors, each corresponding to one node in an undirected graph together with edges connecting nodes. Also available is a vector of network weights $\boldsymbol{w} \equiv (w_1, \ldots, w_p)^T$, indicating relative importance of the predictors. The weight vector reflects the biological importance of a “hub” gene. Given predictors $(\tilde{\boldsymbol{x}}_i=(\tilde{x}_{i1},\ldots,\tilde{x}_{ip})^T)_{i=1}^n$,

$$Y_i \equiv \mu(\tilde{\boldsymbol{x}}_i)+\varepsilon_i, \quad \mu(\tilde{\boldsymbol{x}}) \equiv \tilde{x}^T\tilde{\beta}, \quad E(\varepsilon_i)=0, \quad \mathrm{Var}(\varepsilon_i)=\sigma^2; \quad i=1,\ldots,n, \tag{16}$$

where $\tilde{\boldsymbol{\beta}} \equiv (\tilde{\beta}_1, \ldots, \tilde{\beta}_p)^T$ is a vector of regression coefficients, and $\tilde{x}_i$ is independent of ε_i. In (16), we aim to identify all possible homogenous subnetworks of predictors with respect to $\boldsymbol{w}$. That is, $w_{j_1}^{-1}\tilde{\beta}_{j_1} \approx \cdots \approx w_{j_K}^{-1}\tilde{\beta}_{j_K}$ within each group $\{j_1, \ldots, j_K\} \subset \{1, \ldots, p\}$. Model (16) reduces to (1) by letting $\boldsymbol{x}_i = (w_1\tilde{x}_{i1}, \ldots, w_p\tilde{x}_{ip})^T$ and $\beta=(w_1^{-1}\tilde{\beta}_1,\ldots,w_p^{-1}\tilde{\beta}_p)^T$. Note that existence of a path between nodes j and j' in the undirected graph indicates if predictors x_j and $x_{j'}$ can be grouped. However, our network is a complete graph in this application.

For data analysis, the 107 patients are divided randomly into two groups with 70 and 37 patients, respectively for model-building and validation. For model building, an estimated MSE based on the generalized degrees of freedom is minimized with regard to a set of grid points over the path of $\lambda_1 > 0$, $\lambda_0 \in \{i\lambda_1^*/10{:}i=1,\ldots,10\}$ and $\lambda_2 \in \{0.05, 0.1, 0.15, 0.2, 0.25, 0.3, 0.35, \infty\}$, where λ_1^* is the largest transition point corresponding to $\lambda_2 = \infty$, to obtain the optimal tuning parameters based on the data perturbation method, c.f., Shen and Huang (2007). Moreover, we take Y_i as the log time to metastasis (in months) and x_i to be the expression levels with dimension $p = 294$, together with a Laplacian weight vector $\{w_j=d_j^{1/2},{:}j=1,\ldots,294\}$, where d_j is the number of directed nodes connecting to node j, c.f., Li and Li (2008).

Table 3 summarizes the estimated grouped regression coefficients based on 27 estimated groups with the corresponding subnetworks displayed by different colors in Figure 3. Interestingly TP53 and BRACA1 are similar but they differ from BRACA2, as evident by the corresponding color intensities in Figure 3, indicating the roles that they play in the process of the metastasis. To make a sense of the estimated grouping, we examine the disease genes causing the metastasis, which were identified in Wang et al. (2005). Among the 17 disease genes expressed in our network, 1, 1, 1, 1 and 13 genes belong to the second, 13th, 14th, 17th and 18th groups, respectively. In fact, three disease genes form single groups, and one disease gene belongs to a small group, and 13 of them are in a large group, indicating that genes work in groups according to their functionalities with regard to survivability of breast cancer. In our analysis, the mean squared prediction error is .81 based on the testing data set of 37 observations, which yields the MSE of .3. This is reasonably good relative to the estimated $\sigma^2 = .51$.

6 Discussion

This article proposes a novel grouping pursuit method for high-dimensional least squares regression. The proposed method is placed in the framework of likelihood estimation and model selection. It offers a general treatment to a continuous but non-differentiable nonconvex likelihood problem, where the global penalized maximum likelihood estimate is difficult to obtain. Remarkably, our DC treatment yields desired statistical properties expected from the global penalized maximum likelihood estimate. This is mainly because the DC score equation defined through subdifferentials equals to that of the original nonconvex problem, when the termination criterion is met, c.f., Theorem 2. In this process, the continuous but non-differentiable penalty is essential, At present the proposed method is not designed for feature selection. To generalize, one may replace $J(\boldsymbol{\beta})$ by $\sum_{j=1}^{p} G(\beta_j)+\sum_{j<j'} G(\beta_j-\beta_{j'})$ in (2). Moreover, estimation of the tuning parameters needs to be further investigated with regard to the accuracy of selection. Further research is therefore necessary.

7 Technical proofs

Proof of Lemma 1

We prove by contradiction. Without loss of generality, assume that $\beta^*=\left(\beta_1^*,\ldots\beta_m^*\right)^T$ is a local minimum of $f(\boldsymbol{\beta}) = h(\boldsymbol{\beta}) + \lambda_1 J(\boldsymbol{\beta})$ attaining at the locally non-smooth concave point λ_2

for pair (1, 2) in that $\beta_1^* - \beta_2^* = \lambda_2$. Let $f_1(\beta_1) = f(\beta_1, \beta_2^*, \ldots, \beta_p^*)$, $h_1(\beta_1) = h(\beta_1, \beta_2^*, \ldots, \beta_p^*)$ and $J_1(\beta_1) = J(\beta_1, \beta_2^*, \ldots, \beta_p^*)$. Denote by the right derivative of $J_1(\beta_1)$ at β_1^* to be b. By assumption, its left derivative at β_1^* must be $b + \lambda_1$. Consider the derivative of $h_1(\beta_1)$ at β_1^*. Note that $f_1(\beta_1)$ achieves a local minimum at β_1^*, implying that its right and left derivatives at β_1^* are larger than 0 and less than 0. Hence the right and left derivatives of $h_1(\beta_1)$ at β_1^* is larger than $-b$ and smaller than $-b - \lambda_1$. This contradicts to the fact that $h_1(\beta_1)$ is differentiable in β_1 for any $\lambda_2 > 0$. This completes the proof.

Proof of Theorem 1

To prove continuity in λ_1, note that the derivative of (5) with respect to $\boldsymbol{\beta}$ is continuous in λ_1 because $\sum_{j':j' \neq j} \nabla G_2 \left(\widehat{\beta}_j^{(m-1)}(\lambda_0, \lambda_0, \lambda_2) - \widehat{\beta}_{j'}^{(m-1)}(\lambda_0, \lambda_0, \lambda_2) \right)$ is independent of λ_1. By convexity of (5) in $\boldsymbol{\beta}$ and uniqueness of $\hat{\boldsymbol{\beta}}^{(m)}(\boldsymbol{\lambda})$, $\hat{\boldsymbol{\beta}}^{(m)}(\boldsymbol{\lambda})$ is continuous in λ_1 for each m.

Next we derive an expression for $\hat{\boldsymbol{\alpha}}^{(m)}(\boldsymbol{\lambda})$. If $K^{(m)}(\boldsymbol{\lambda}) = 1$ and $|\mathcal{G}_k^{(m)}(\boldsymbol{\lambda})| \geq 2$, then the result follows from (5). Now consider the case of $K^{(m)}(\boldsymbol{\lambda}) \geq 2$. For any $j = 1, \ldots, p$, write $\sum_{j' \neq j} b_{jj'}^{(m)}(\boldsymbol{\lambda}) = \sum_{j':j' \sim j} b_{jj'}^{(m)}(\boldsymbol{\lambda}) + \sum_{j':j' \nsim j} b_{jj'}^{(m)}(\boldsymbol{\lambda})$ as $B_j^{(m)}(\boldsymbol{\lambda}) + \sum_{k:k \neq g(\boldsymbol{\lambda},j)} |\mathcal{G}_{g(\boldsymbol{\lambda},j)}^{(m)}(\boldsymbol{\lambda})|$ Sign $\left(\widehat{\alpha}_k^{(m)}(\boldsymbol{\lambda}) - \widehat{\alpha}_{g(\boldsymbol{\lambda},j)}^{(m)}(\boldsymbol{\lambda}) \right)$, where $j' \sim j$ if j' and j are in the same group and $j' \nsim j$ otherwise. Differentiating (5) with respect to $\boldsymbol{\beta}$, we obtain

$$-\boldsymbol{x}_j^T (\boldsymbol{Y} - \boldsymbol{Z}_{\mathcal{G}^{(m)}(\boldsymbol{\lambda})} \widehat{\alpha}^{(m)}(\boldsymbol{\lambda})) + n\lambda_1 \left(\Delta_j^{(m)}(\boldsymbol{\lambda}) + B_j^{(m)}(\boldsymbol{\lambda}) \right) = 0, \quad j = 1, \ldots, p, \tag{17}$$

which is an optimality condition (Rockafellar and Wets, 2003). For $k = 1, \ldots, K^{(m)}(\boldsymbol{\lambda})$, invoking the sum-to-zero constraint $\sum_{j \in \mathcal{G}_k^{(m)}(\boldsymbol{\lambda})} B_j^{(m)}(\boldsymbol{\lambda}) = 0$, we have $-\boldsymbol{z}_{k,\mathcal{G}^{(m)}(\boldsymbol{\lambda})} (\boldsymbol{Y} - \boldsymbol{Z}_{\mathcal{G}^{(m)}(\boldsymbol{\lambda})} \widehat{\alpha}^{(m)}(\boldsymbol{\lambda})) + n\lambda_1 \delta_k^{(m)}(\boldsymbol{\lambda}) = 0$, implying (6). Thus (8) follows from (17).

Proof of Proposition 1

By definition, $L_k^{\pm}\left(\lambda_1, A_{k\ell}^{\pm}(\lambda_1)\right)$ is piecewise linear and continuous in λ_1, and is strictly monotone because $\sum_{j \in A_{k\ell}^+(\lambda_1)} \xi_j > 0$ and $\sum_{j \in A_{k\ell}^-(\lambda_1)} \xi_j < 0$.

Lets $g_k = |\mathcal{G}_k| - 1$. For Algorithm 1, the complexities for Steps 1 and 2 are $O(np^2)$. In Step 3, sorting for computing search points is no greater than $O(g_k^2 \log g_k)$ for group $\mathcal{G}_k$. Hence the complexity for Step 3 is $O(p^2 \log p)$ using the fact that $\sum_{k=1}^{K} g_k < p$, because the complexity of search in Step 3 is no great than $\log p$ for the bisection (Fibonacci) search. Then the complexity for Algorithm 2 is $O(m^* n^* p^2 (\log p + n))$. This completes the proof.

Proof of Theorem 2

Uniqueness of the solution follows from strict convexity of $S^{(m)}(\boldsymbol{\beta})$ in $\boldsymbol{\beta}$ for each m under the assumption that $\boldsymbol{Z}_{\mathcal{G}^{(m)}(\boldsymbol{\lambda})}^T \boldsymbol{Z}_{\mathcal{G}^{(m)}(\boldsymbol{\lambda})}$ is invertible.

Our plan is to prove the result for $\lambda_1 = \lambda_0$ with $\boldsymbol{\lambda} = (\lambda_0, \lambda_0, \lambda_2)^T$. Then controlling at this point implies the desirable result for all $\boldsymbol{\lambda}$. In what follows, we set $\lambda_1 = \lambda_0$ unless indicated otherwise. For convergence of Algorithm 2, it follows from (2) and (5) that for $m \in \mathbb{N}$, $0 \le S(\hat{\boldsymbol{\beta}}^{(m)}(\boldsymbol{\lambda})) = S^{(m+1)}(\hat{\boldsymbol{\beta}}^{(m)}(\boldsymbol{\lambda})) \le S^{(m)}(\hat{\boldsymbol{\beta}}^{(m)}(\boldsymbol{\lambda})) \le S^{(m)}(\hat{\boldsymbol{\beta}}^{(m-1)}(\boldsymbol{\lambda})) = S(\hat{\boldsymbol{\beta}}^{(m-1)}(\boldsymbol{\lambda}))$. This implies that $\lim_{m\to\infty} S(\hat{\boldsymbol{\beta}}^{(m)}(\boldsymbol{\lambda}))$ exists, thus leading to convergence. To study the number of steps to termination, note that $S^{(m)}(\hat{\boldsymbol{\beta}}^{(m-1)}(\boldsymbol{\lambda})) - S^{(m)}(\hat{\boldsymbol{\beta}}^{(m)}(\boldsymbol{\lambda}))$ can be written as

$$\begin{aligned} & {}^{T}\boldsymbol{X}\left(\widehat{\beta}^{(m)}(\boldsymbol{\lambda}) - \widehat{\beta}^{(m-1)}(\boldsymbol{\lambda})\right) \\ & +\lambda_1 \left\{\sum_{j=1}^{p}\left(\widehat{\beta}_j^{(m)}(\boldsymbol{\lambda}) - \widehat{\beta}_j^{(m-1)}(\boldsymbol{\lambda})\right)\sum_{j' \neq j}\nabla G_2\left(\widehat{\beta}_j^{(m-1)}(\boldsymbol{\lambda}) - \widehat{\beta}_{j'}^{(m-1)}(\boldsymbol{\lambda})\right)\right\} \\ & +\lambda_1 \left\{\sum_{j<j'}\left(\left|\widehat{\beta}_j^{(m-1)}(\boldsymbol{\lambda}) - \widehat{\beta}_{j'}^{(m-1)}(\boldsymbol{\lambda})\right| - \left|\widehat{\beta}_j^{(m)}(\widehat{\boldsymbol{\lambda}}) - \widehat{\beta}_{j'}^{(m-1)}(\boldsymbol{\lambda})\right|\right)\right\}, \end{aligned}$$

which can be simplified, using the following equality from (17), $-\boldsymbol{x}_j^T(\boldsymbol{Y} - \boldsymbol{X}\widehat{\beta}^{(m)}(\boldsymbol{\lambda})) = n\lambda_1\sum_{j':j'\neq j}\left\{\nabla G_2\left(\widehat{\beta}_j^{(m-1)}(\boldsymbol{\lambda}) - \widehat{\beta}_{j'}^{(m-1)}(\boldsymbol{\lambda})\right) - b_{jj'}^{(m)}(\boldsymbol{\lambda})\right\}$, as $\frac{1}{2n}\|\boldsymbol{X}\left(\widehat{\beta}^{(m)}(\boldsymbol{\lambda}) - \widehat{\beta}^{(m-1)}(\boldsymbol{\lambda})\right)\|^2 + \lambda_1\sum_{j<j'}\left\{\left|z_{jj'}^{(m-1)}(\boldsymbol{\lambda})\right| - \left|z_{jj'}^{(m)}(\boldsymbol{\lambda})\right| - b_{jj'}^{(m)}(\boldsymbol{\lambda})\left(z_{jj'}^{(m-1)}(\boldsymbol{\lambda}) - z_{jj'}^{(m)}(\boldsymbol{\lambda})\right)\right\}$, where $z_{jj'}^{(m)}(\boldsymbol{\lambda}) = \widehat{\beta}_j^{(m)}(\boldsymbol{\lambda}) - \widehat{\beta}_{j'}^{(m)}(\boldsymbol{\lambda})$. By convexity of $|z|$, $\left|z_{jj'}^{(m-1)}(\boldsymbol{\lambda})\right| - \left|z_{jj'}^{(m)}(\boldsymbol{\lambda})\right| - b_{jj'}^{(m)}(\widehat{\boldsymbol{\lambda}})\left(z_{jj'}^{(m-1)}(\boldsymbol{\lambda}) - z_{jj'}^{(m)}(\boldsymbol{\lambda})\right) \ge 0$, implying $S(\hat{\boldsymbol{\beta}}^{(m-1)}(\boldsymbol{\lambda})) - S(\hat{\boldsymbol{\beta}}^{(m)}(\boldsymbol{\lambda})) \ge S^{(m)}(\hat{\boldsymbol{\beta}}^{(m-1)}(\boldsymbol{\lambda})) - S^{(m)}(\hat{\boldsymbol{\beta}}^{(m)}(\boldsymbol{\lambda}))$ is bounded below by $\frac{1}{2n}\|\boldsymbol{X}\left(\widehat{\beta}^{(m)}(\boldsymbol{\lambda}) - \widehat{\beta}^{(m-1)}(\boldsymbol{\lambda})\right)\|^2$. That is, $\frac{1}{2n}\left(\widehat{\alpha}^{(m)}(\boldsymbol{\lambda}) - \alpha^{(m-1)}(\widehat{\boldsymbol{\lambda}})\right)^T \boldsymbol{Z}_{\mathcal{G}^{(m)}(\boldsymbol{\lambda})}^T \boldsymbol{Z}_{\mathcal{G}^{(m)}(\boldsymbol{\lambda})}\left(\widehat{\alpha}^{(m)}(\lambda) - \widehat{\alpha}^{(m-1)}(\boldsymbol{\lambda})\right)$ is greater than zero unless $\hat{\boldsymbol{\alpha}}^{(m)}(\boldsymbol{\lambda}) = \hat{\boldsymbol{\alpha}}^{(m-1)}(\boldsymbol{\lambda})$.

Finally, finite step convergence follows from strict decreasingness of $S^{(m)}(\hat{\boldsymbol{\beta}}(\boldsymbol{\lambda}))$ in m and finite possible values of $\nabla S_2(\hat{\boldsymbol{\beta}}^{(m-1)}(\boldsymbol{\lambda}))$ in (5). For (14), note that when termination, $\nabla S_2(\hat{\boldsymbol{\beta}}^{(m-1)}(\boldsymbol{\lambda}))$ remains unchanged for $m \ge m^*$, so does the cost function (5) for $m \ge m^*$. This implies termination for all $\boldsymbol{\lambda} = (\lambda_0, \lambda_1, \lambda_2)^T$ in (14). This completes the proof.

Proof of Theorem 3

Define event

$$F \equiv \left\{\min_{k<l}\left|\widehat{\alpha}_k^{(ols)} - \widehat{\alpha}_l^{(ols)}\right| > 3\lambda_2/2\right\} \cap_{k:|\mathcal{G}_k^0|>1}\left\{\max_{j:j\in\mathcal{G}_k^0}\left|\boldsymbol{x}_j^T\left(\boldsymbol{Y} - \boldsymbol{X}^T\widehat{\beta}^{(ols)}\right)\right| \le n\lambda_1(|\mathcal{G}_k^0| - 1)\right\}.$$

By (17) with $m = m^*$ and (14), for $k = 1, \ldots, K$, $\hat{\boldsymbol{\beta}}(\boldsymbol{\lambda}) = \hat{\boldsymbol{\beta}}^{(m)}(\boldsymbol{\lambda})$ satisfies

$$\begin{cases} -\left(\sum_{j\in\mathcal{G}_k}\boldsymbol{x}_j\right)^T(\boldsymbol{Y} - \boldsymbol{X}\beta) - n\lambda_1\left(\sum_{j\in\mathcal{G}_k}\Delta_j(\beta)\right) = 0; & \\ |\boldsymbol{x}_j^T(\boldsymbol{Y} - \boldsymbol{X}\beta) + n\lambda_1\Delta_j(\beta)| \le n\lambda_1(|\mathcal{G}_k| - 1); & j \in \mathcal{G}_k,\ |\mathcal{G}_k| > 1, \end{cases} \quad (18)$$

for some partition ($\mathcal{G}_1, \ldots, \mathcal{G}_K$) of $\{1, \ldots, p\}$ with $K \le \min\{n, p\}$, where $\Delta_j(\boldsymbol{\beta}) \equiv \Sigma_{j':j'\neq j}\{\mathrm{Sign}(\beta_j - \beta_{j'}) - \nabla G_2(\beta_j - \beta_{j'})\}$; $j = 1, \ldots, p$.

Note that the first event in F, together with the grouped subdifferentials, yields that $\sum_{j\in\mathcal{G}_k^0}\Delta_j(\widehat{\beta}^{(ols)})=0$; $k = 1, \ldots, K^0$. This, together with the least squares property that $\left(\sum_{j\in\mathcal{G}_k^0}\boldsymbol{x}_j\right)^T(\boldsymbol{Y}-\boldsymbol{X}\widehat{\beta}^{(ols)})=0$, implies that the first equation of (18) is fulfilled with $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}^{(ols)}$. Moreover, the events in F imply the second equation of (18) with $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}^{(ols)}$. Consequently $\hat{\boldsymbol{\beta}}^{(ols)}$ is a solution of (18) on F.

It remains to show that (18) yields the unique minimizer on F. Define

$$\tilde{G}(z)=\begin{cases} G(z); & \text{if } |z| \le \lambda_2(1-\nu) \text{ or } |z| \ge \lambda_2(1+\nu),\\ -\frac{1}{4\lambda_2\nu}(z-\lambda_2)^2+\frac{1}{2}(z-\lambda_2)+\lambda_2(1-\frac{\nu}{4}); & \text{if } |z-\lambda_2|<\lambda_2\nu,\\ -\frac{1}{4\lambda_2\nu}(z+\lambda_2)^2-\frac{1}{2}(z+\lambda_2)+\lambda_2(1-\frac{\nu}{4}); & \text{if } |z+\lambda_2|<\nu\lambda_2,\end{cases}$$

for $\nu = 1/2$. Given any grouping $\mathcal{G}$ with $|\mathcal{G}| \le K^*$, $\tilde{S}(\boldsymbol{\beta})$ is a function of $\boldsymbol{\alpha}_{\mathcal{G}}$ with $\boldsymbol{\beta} = (\alpha_1\mathbf{1}_{|\mathcal{G}_1|}, \ldots, \alpha_K\mathbf{1}_{|\mathcal{G}_K|})^T$. Then $\tilde{S}(\beta)=\frac{1}{2n}\sum_{i=1}^{n}(Y_i-\boldsymbol{x}_i^T\beta)^2+\lambda_1\sum_{j<j'}\tilde{G}(\beta_j-\beta_{j'})$ is strictly convex in $\boldsymbol{\alpha}_{\mathcal{G}} \in \mathbb{R}^{|\mathcal{G}|}$ when $\frac{1}{n}\boldsymbol{Z}_{\mathcal{G}}^T\boldsymbol{Z}_{\mathcal{G}}>\frac{\lambda_1}{\lambda_2}\left\{(|\mathcal{G}|+1)\boldsymbol{I}_{|\mathcal{G}|}-\mathbf{1}_{|\mathcal{G}|}\mathbf{1}_{|\mathcal{G}|}^T\right\}$, which occurs when $c_{\min}(\mathcal{G})>\frac{\lambda_1}{\lambda_2}(|\mathcal{G}|+1)$. To prove that $\hat{\boldsymbol{\beta}}^{(ols)}$ is the unique minimizer of $\tilde{S}(\boldsymbol{\beta})$, suppose $\tilde{\boldsymbol{\beta}}$ is another minimizer of $\tilde{S}(\boldsymbol{\beta})$ with $\tilde{\mathcal{G}}$ the corresponding grouping and $|\tilde{\mathcal{G}}| < K^*$. Because $\min_{|\mathcal{G}|\le(K^*)^2} c_{\min}(\mathcal{G})>\frac{\lambda_1}{\lambda_2}(K^*+1)$ and $(K^*)^2 \le n$, it follows that $\tilde{S}(\boldsymbol{\beta})$ is strictly convex in $\boldsymbol{\alpha}_{\mathcal{G}^0\vee\tilde{\mathcal{G}}} \in \mathbb{R}^{|\mathcal{G}^0\vee\&\tilde{\mathcal{G}}|}$, implying $\tilde{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}^{(ols)}$, where $\mathcal{G}^0 \vee \tilde{\mathcal{G}}$ is the coarsest common refinement of $\mathcal{G}^0$ and $\tilde{\mathcal{G}}$ with $|\mathcal{G}^0 \vee \tilde{\mathcal{G}}| \le \min\{n, p\}$.

To prove that $\hat{\boldsymbol{\beta}}^{(ols)}$ is the unique minimizer of $S(\boldsymbol{\beta})$ on F, we let $\mathcal{G}^* = \mathcal{G}^0 \vee \mathcal{G}$ with $|\mathcal{G}| \le K^*$. Let $\widehat{\alpha}_{\mathcal{G}*}^{(ols)}$ be the estimate corresponding to $\hat{\boldsymbol{\beta}}^{(ols)}$. By the mean value theorem, $\|\frac{\partial}{\partial\alpha_{\mathcal{G}*}}\tilde{S}(\beta)-\frac{\partial}{\partial\alpha_{\mathcal{G}*}}\tilde{S}(\beta)|_{\alpha_{\mathcal{G}*}=\widehat{\alpha}_{\mathcal{G}*}^{(ols)}}\|$ is lower bounded by

$$\left\{\min_{|\mathcal{G}|\le(K*)^2}c_{\min}(\mathcal{G})-\frac{\lambda_1}{\lambda_2}(K^*+1)\right\}\|\alpha_{\mathcal{G}*}-\widehat{\alpha}_{\mathcal{G}*}^{(ols)}\|>0. \tag{19}$$

Note that $\tilde{S}(\boldsymbol{\beta}) = S(\boldsymbol{\beta})$ over $E = \{\boldsymbol{\beta} : ||\alpha_k - \alpha_l| - \lambda_2| > \lambda_2/2 : 1 \le k < l \le |\mathcal{G}|\}$. Moreover, by construction, $\sup_{\alpha_{\mathcal{G}*}}\|\frac{\partial}{\partial\alpha_{\mathcal{G}*}}S(\beta)-\frac{\partial}{\partial\alpha_{\mathcal{G}*}}\tilde{S}(\beta)\| \le \frac{\lambda_1}{2}K^*$ on F, which implies, together with (19), for any $\boldsymbol{\beta} \neq \hat{\boldsymbol{\beta}}^{(ols)} \in E^c$, $\|\frac{\partial}{\partial\alpha_{\mathcal{G}*}}S(\beta)\|$ is

$$\|\left(\frac{\partial}{\partial\alpha_{\mathcal{G}*}}\tilde{S}(\beta)-\frac{\partial}{\partial\alpha_{\mathcal{G}*}}\tilde{S}(\beta)|_{\alpha_{\mathcal{G}*}=\widehat{\alpha}_{\mathcal{G}*}^{(ols)}}\right)+\left(\frac{\partial}{\partial\alpha_{\mathcal{G}*}}S(\beta)-\frac{\partial}{\partial\alpha_{\mathcal{G}*}}\tilde{S}(\beta)\right)\| \ge \left(\min_{|\mathcal{G}|\le(K^*)^2}c_{\min}(\mathcal{G})-\frac{\lambda_1}{\lambda_2}(K^*+1)\right)\|\alpha_{\mathcal{G}*}-\widehat{\alpha}_{\mathcal{G}*}^{(ols)}\|-\frac{\lambda_1}{2}K^*,$$

which is lower bounded by $\left(\min_{|\mathcal{G}|\le(K^*)^2} c_{\min}(\mathcal{G}) - \frac{\lambda_1}{\lambda_2}(2K^*+1)\right)\frac{\lambda_2}{2}>0$, because $\min_{|\mathcal{G}|\le(K^*)^2} c_{\min}(\mathcal{G})>\frac{\lambda_1}{\lambda_2}(2K^*+1)$. This, together with Lemma 1, implies that $S(\boldsymbol{\beta})$ has no local minimal in E^c on F, and hence it has the unique local minimal on F. On the other hand, $\hat{\boldsymbol{\beta}}(\boldsymbol{\lambda})$ is a local minimizer of $S(\boldsymbol{\beta})$ on F. Consequently, $\hat{\boldsymbol{\beta}}^{(ols)} = \hat{\boldsymbol{\beta}}(\boldsymbol{\lambda})$ on F.

Note that $\widehat{\alpha}_k^{(ols)} - \widehat{\alpha}_l^{(ols)} \sim N(\alpha_k^0 - \alpha_l^0,\ Var(\widehat{\alpha}_k^{(ols)} - \widehat{\alpha}_l^{(ols)}))$ with $Var(\widehat{\alpha}_k^{(ols)} - \widehat{\alpha}_l^{(ols)}) \le 4c_{\min}^{-1}(\mathcal{G}^0)\sigma^2/n$, and $\boldsymbol{x}_j^T\left(\boldsymbol{Y} - \boldsymbol{X}^T\widehat{\beta}^{(ols)}\right) \sim N\left(0, \sigma^2\|\left(\boldsymbol{I} - \boldsymbol{Z}_{\mathcal{G}^0}\left(\boldsymbol{Z}_{\mathcal{G}^0}^T\boldsymbol{Z}_{\mathcal{G}^0}\right)^{-1}\boldsymbol{Z}_{\mathcal{G}^0}^T\right)\boldsymbol{x}_j\|^2\right)$ with $\|\left(\boldsymbol{I} - \boldsymbol{Z}_{\mathcal{G}^0}\left(\boldsymbol{Z}_{\mathcal{G}^0}^T\boldsymbol{Z}_{\mathcal{G}^0}\right)^{-1}\boldsymbol{Z}_{\mathcal{G}^0}^T\right)\boldsymbol{x}_j\|^2 \le \|x_j\|^2$. It follows that $P(\mathcal{G}(\boldsymbol{\lambda}) \ne \mathcal{G}^0) \le P(\hat{\boldsymbol{\beta}}(\boldsymbol{\lambda}) \ne \hat{\boldsymbol{\beta}}^{(ols)}) \le P(F^c)$, which is upper bounded by

$$-\boldsymbol{X}^T\widehat{\beta}^{(ols)})|>n\lambda_1(|\mathcal{G}_k^0| - 1)\} \le \frac{K^0(K^0-1)}{2}\Phi\left(\frac{n^{1/2}(3\lambda_2/2-\gamma_{\min}}{2\sigma c_{\min}^{-1/2}(\mathcal{G}^0)}\right)$$
$$+p\Phi\left(\frac{-n\lambda_1}{\sigma\max_{1\le j\le p}\|x_j\|}\right),$$

where $\boldsymbol{q}_{kl}=\boldsymbol{Z}_{\mathcal{G}^0}\left(\boldsymbol{Z}_{\mathcal{G}^0}^T\boldsymbol{Z}_{\mathcal{G}^0}\right)^{-1}(\boldsymbol{e}_k - \boldsymbol{e}_l)$ and $\boldsymbol{e}_k$ is the kth column of $\boldsymbol{I}_p$. Using an inequality that $\Phi(-|z|) \le \sqrt{2/\pi}|z|^{-1}\exp(-z^2/2)$, we obtain the desired bound.

Proof of Corollary 1

It is a direct consequence of Theorem 3 and the least squares property. The proof is thus omitted.

Proof of Theorem 4

By Theorem 2, $\hat{\beta}(\boldsymbol{\lambda}) = \hat{\beta}^{(m^*)}(\boldsymbol{\lambda})$. From (3), for any $j \in \mathcal{G}_k(\boldsymbol{\lambda})$; $k = 1, \ldots, K(\boldsymbol{\lambda})$, we have $|B_j(\boldsymbol{\lambda})| \le |\mathcal{G}_k(\boldsymbol{\lambda})| - 1$. Note further that for $j \in \mathcal{G}_k(\boldsymbol{\lambda})$, $\Delta_j(\boldsymbol{\lambda})$ can be rewritten as $\Delta_j(\boldsymbol{\lambda}) = \Sigma_{k':k'\ne k}\{|\mathcal{G}_{k'}(\boldsymbol{\lambda})|(\text{Sign}\,(\hat{\alpha}_k(\boldsymbol{\lambda}) - \hat{\alpha}_{k'}(\boldsymbol{\lambda})) - \nabla G_2(\hat{\alpha}_k(\hat{\boldsymbol{\lambda}}^{(0)}) - \hat{\alpha}_{k'}(\hat{\boldsymbol{\lambda}}^{(0)})))\}$. By (17), $|r_j(\hat{\boldsymbol{\beta}}(\boldsymbol{\lambda})) - n\lambda_1\Delta_j(\boldsymbol{\lambda})| \le n\lambda_1|B_j(\boldsymbol{\lambda})|$, implying $\frac{1}{n\lambda_1}\rho_j(\boldsymbol{\lambda}) \in E_k(\boldsymbol{\lambda})$.

For disjointness of $E_k(\boldsymbol{\lambda})$'s, assume, without loss of generality, that $\hat{\alpha}_1(\boldsymbol{\lambda}) < \cdots < \hat{\alpha}_{K(\boldsymbol{\lambda})}(\boldsymbol{\lambda})$. For any $j \in \mathcal{G}_k(\boldsymbol{\lambda})$, $j' \in \mathcal{G}_{k'}(\boldsymbol{\lambda})$, and $k < k'$, $\Delta_j(\boldsymbol{\lambda}) - \Delta_{j'}(\boldsymbol{\lambda}) = (|\mathcal{G}_{k'}(\boldsymbol{\lambda})| + |\mathcal{G}_{k'}(\boldsymbol{\lambda})|)(k' - k) > (|\mathcal{G}_{k'}(\boldsymbol{\lambda})| - 1) + (|\mathcal{G}_k(\boldsymbol{\lambda})| - 1)$, implying disjointness.

References

1. An HLT, Tao PD. Solving a class of linearly constrained indefinite quadratic problems by D.C. algorithms. J Global Optim 1997;11:253–85.
2. Allgower, EL.; George, K. Introduction to Numerical Continuation Methods. SIAM; 2003.
3. Bondell HD, Reich BJ. Simultaneous regression shrinkage, variable selection, and supervised clustering of predictors with OSCAR. Biometrics 2008;64:115–23. [PubMed: 17608783]
4. Chuang HY, Lee EJ, et al. Network-based classification of breast cancer metastasis. Molecular Systems Biology 2007;3:140. [PubMed: 17940530]

5. Efron B. The estimation of prediction error: covariance penalties and cross-validation. J Amer Statist Assoc 2004;99:619–32.

6. Fan J, Li R. Variable selection via nonconcave penalized likelihood and its Oracle properties. J Amer Statist Assoc 2001;96:1348–60.

7. Fan J. Comments on “Wavelets in statistics: a review” by A. Antoniadis. J Italian Statist Assoc 1997;6:131–138.

8. Friedman J, Haste T, Hofling H, Tibshirani R. Pathwise coordinate optimization. Ann Applied Statist 2007;1:302–332.

9. Li C, Li H. Network-constraint regularization and variable selection for analysis of genomic data. Bioinformatics 2008;24:1175–82. [PubMed: 18310618]

10. Liu, S.; Shen, X.; Wong, W. Computational developments of ψ-learning. Proc 5th SIAM Intern Conf on Data Mining; Newport, CA. April, 2005; 2005. p. 1-12.

11. Liu Y, Wu Y. Variable selection via a combination of the L_0 and L_1 penalties. J Comput Graph Statist 2007;16:782–798.

12. Gill, PE.; Murray, W.; Wright, MH. Practical Optimization. Academic Press; London: 1981.

13. Rockafellar, RT.; Wets, RJ. Variational Analysis. Springer-Verlag; 2003.

14. Rosset S, Zhu J. Piecewise linear regularized solution paths. Ann Statist 2007;35:1012–30.

15. Rota GC. The number of partitions of a set. American Mathematical Monthly 1964;71:498–504.

16. Shen X, Huang HC. Optimal model assessment, selection and combination. J Amer Statist Assoc 2006;101:554–68.

17. Stein C. Estimation of the mean of a multivariate normal distribution. Ann Statist 1981;9:1135–51.

18. Soussi T. Focus on the P53 gene and cancer: advances in TP53 mutation research. Human mutation 2003;21:173–5. [PubMed: 12619102]

19. Tibshirani R, Saunders M, Rosset S, Zhu J, Knight K. Sparsity and smoothness via the fused lasso. J Royal Statist Soc, Ser B 2005;67:91–108.

20. Wang Y, Klijin JG, et al. Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. Lancet 2005;365:671–79. [PubMed: 15721472]

21. Yuan M, Lin Y. Model selection and estimation in regression with grouped variables. J Royal Statist Soc Ser B 2006;68:49–67.

22. Wu S, Shen X, Geyer C. Adaptive regularization through entire solution surface. Biometrika 2009;96:513–527.

23. Zhao P, Rocha G, Yu B. The composite absolute penalties family for grouped and hierarchical variable selection. Ann Statist 2009;37:3468–3497.

24. Zou H, Hastie T. Regularization and variable selection via the elastic net. J Royal Statist Assoc, Ser B 2005;67:301–20.

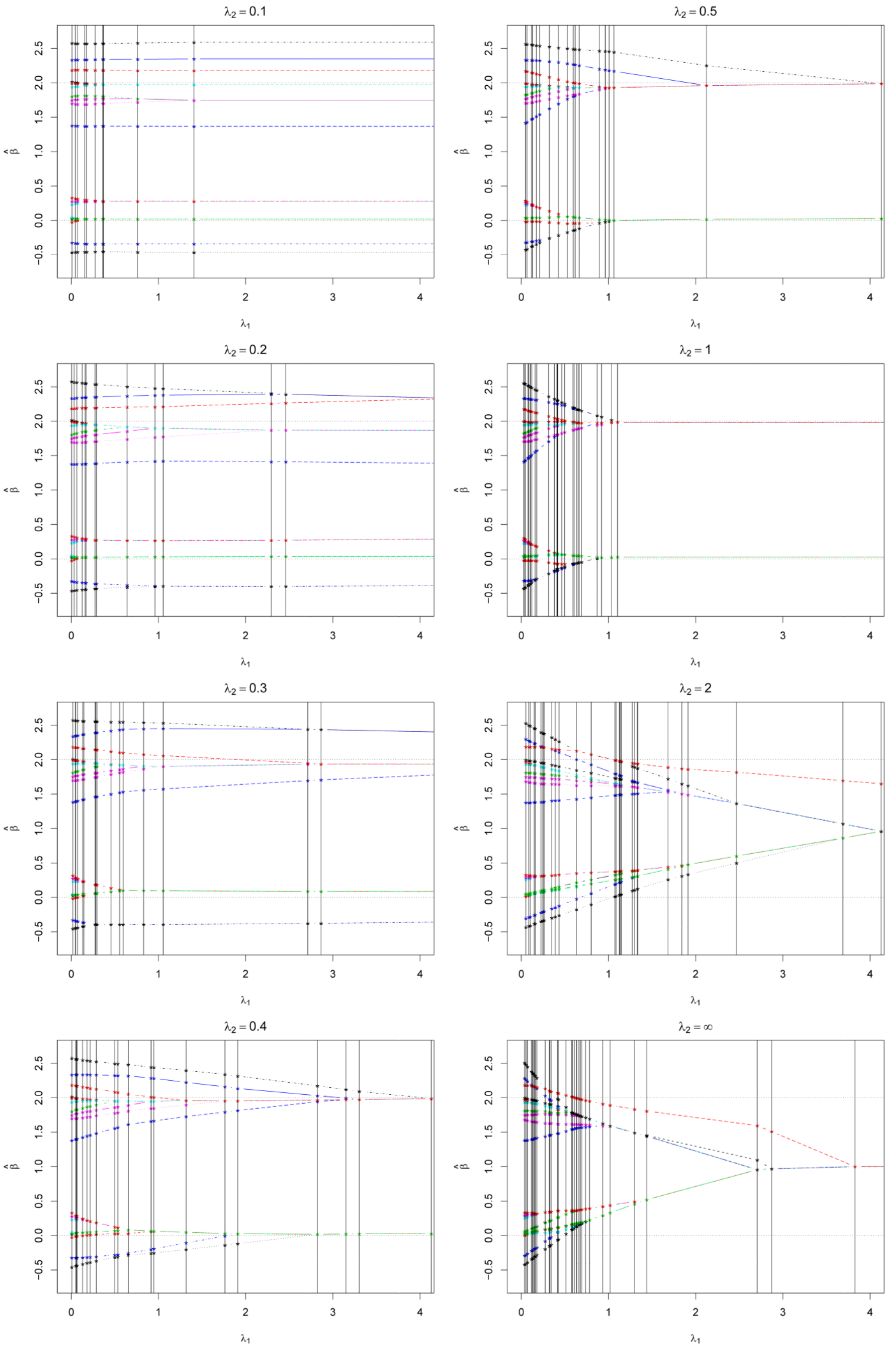

Figure 1.

Plots of $\hat{\boldsymbol{\beta}}(\boldsymbol{\lambda})$ as a function of λ_1 for various λ_2 values with $\lambda_0 = .2$ and $\sigma = 1.2$ in Example 1. Different components of $\hat{\boldsymbol{\beta}}(\boldsymbol{\lambda})$ are represented by different types of lines and colors.

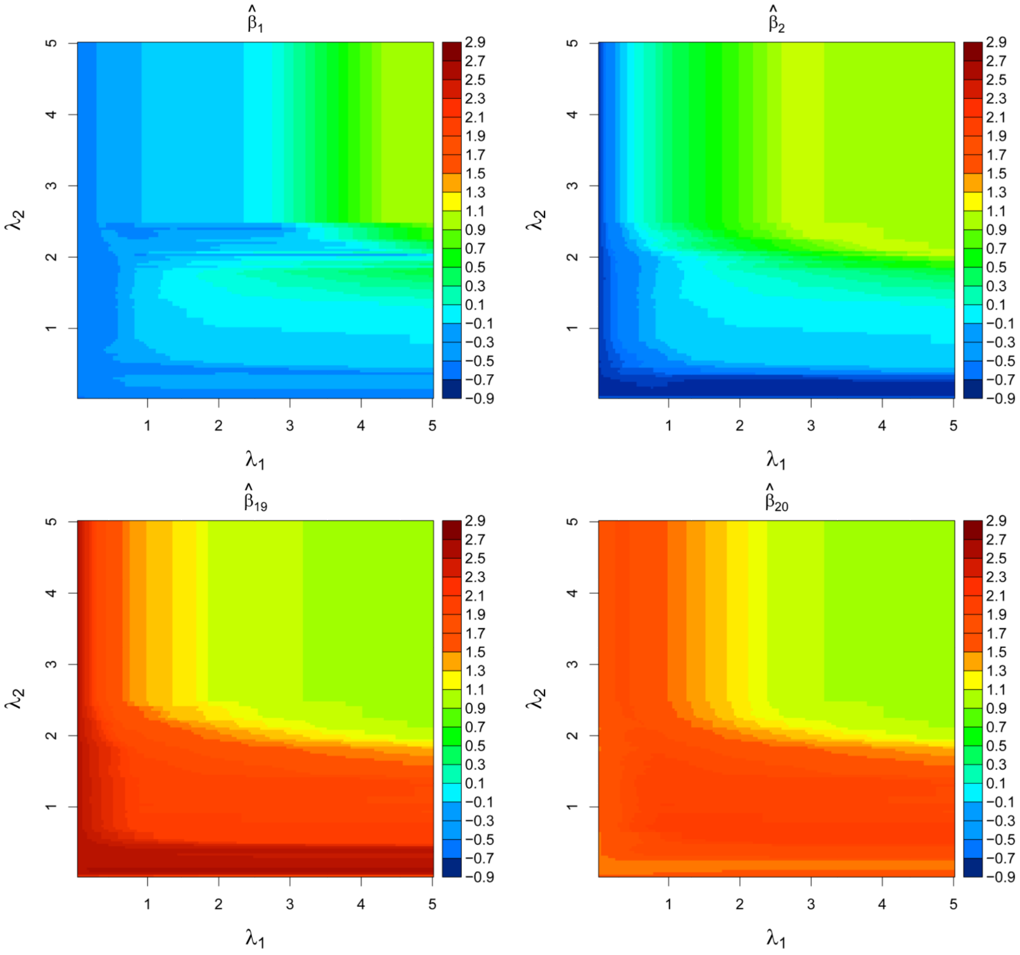

Figure 2.

Image plots of regularization solution surfaces of four components $\hat\beta_1(\boldsymbol{\lambda})$, $\hat\beta_2(\boldsymbol{\lambda})$, $\hat\beta_{19}(\boldsymbol{\lambda})$, and $\hat\beta_{20}(\boldsymbol{\lambda})$ as a function of (λ_1, λ_2) for $\lambda_0 = 0.2$ and $\sigma = 1.2$ in Example 1.

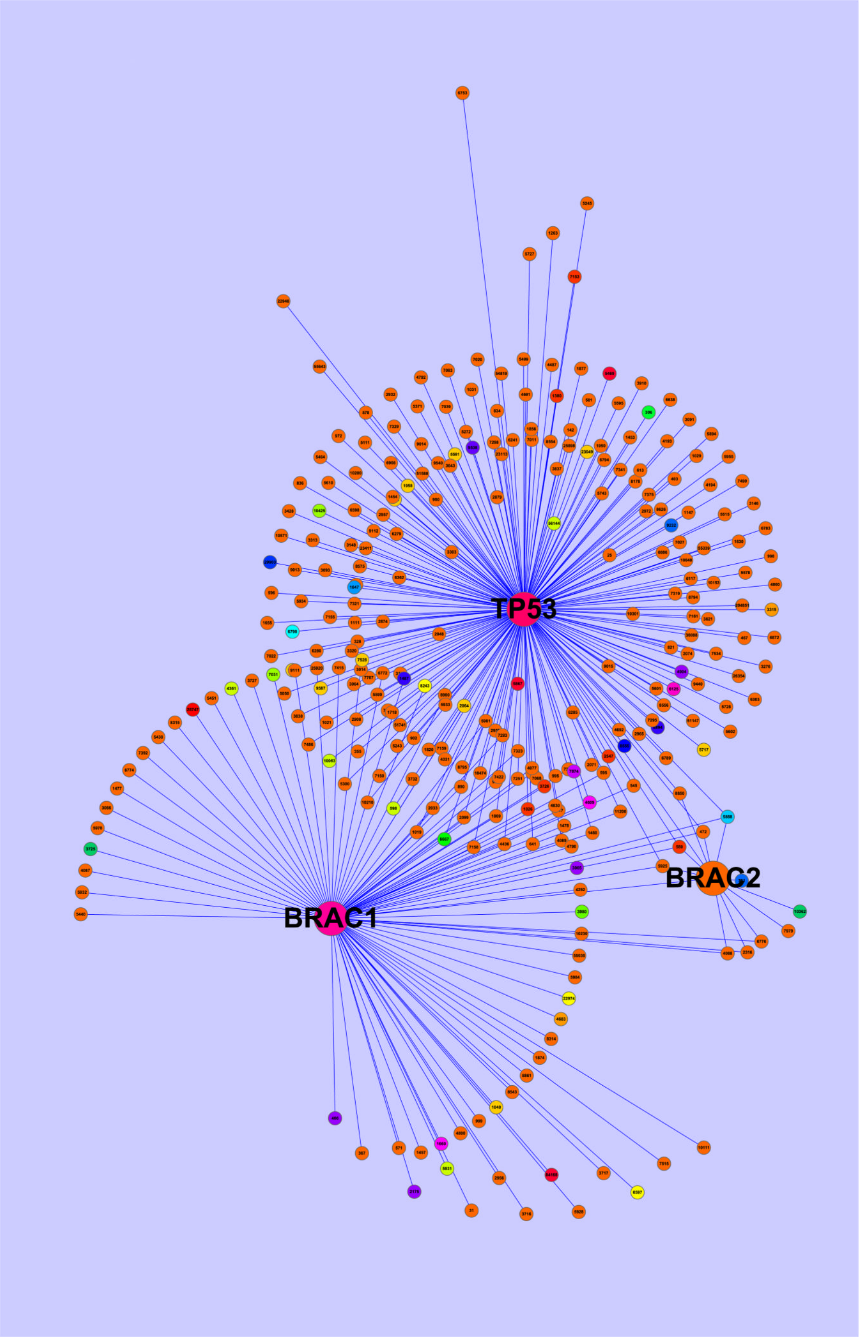

Figure 3.

Plot of the PPI subnetwork for the metastasis data, as described by an undirected graph with 294 nodes and 326 edges. Three regulating genes TP53, BRACA1 and BRACA2 are represented by large nodes. There are 27 estimated groups colored over the color spectrum with dark color corresponding to small estimated group regression coefficients. The plot is produced in Cytoscape.

Table 1

MSEs as well as estimated standard errors (in parentheses) of grouping pursuit for various methods based on 100 simulation replications in Example 1. Here Full, True, Lasso, Convex, DCE, denote the least squares estimates based on the full model, the true model, the Lasso estimate, our convex counterpart based on iteration *m* = 0, and our estimate.

n	***σ***	**Full**	**True model**		**Lasso**	**Our**		**Ave # Iter**	**Ave match proportion**
			Grouping	**Variable**		**Convex**	**DCE**		
50	2.0	1.607 (0.0446)	0.235 (0.0150)	0.845 (0.0310)	1.318 (0.0449)	1.418 (0.0483)	0.837 (0.0721)	4.31 (0.17)	0.633
	1.0	0.402 (0.0111)	0.059 (0.0037)	0.211 (0.0077)	0.330 (0.0112)	0.362 (0.0128)	0.070 (0.0050)	4.08 (0.14)	0.716
	0.5	0.100 (0.0028)	0.015 (0.0009)	0.053 (0.0019)	0.083 (0.0029)	0.091 (0.0032)	0.019 (0.0016)	3.71 (0.10)	0.733
100	2.0	0.833 (0.0244)	0.129 (0.0083)	0.456 (0.0176)	0.658 (0.0229)	0.699 (0.0234)	0.183 (0.0157)	4.32 (0.16)	0.788
	1.0	0.208 (0.0061)	0.032 (0.0021)	0.114 (0.0044)	0.164 (0.0058)	0.175 (0.0059)	0.040 (0.0035)	3.94 (0.12)	0.867
	0.5	0.052 (0.0015)	0.008 (0.0005)	0.028 (0.0011)	0.041 (0.0014)	0.044 (0.0015)	0.010 (0.0008)	3.16 (0.06)	0.887
200	2.0	0.411 (0.0123)	0.060 (0.0041)	0.223 (0.0092)	0.335 (0.0117)	0.364 (0.0129)	0.080 (0.0057)	4.03 (0.13)	0.926
	1.0	0.103 (0.0031)	0.015 (0.0010)	0.056 (0.0023)	0.084 (0.0029)	0.091 (0.0032)	0.020 (0.0014)	3.33 (0.06)	0.950
	0.5	0.026 (0.0008)	0.004 (0.0003)	0.014 (0.0006)	0.022 (0.0008)	0.023 (0.0008)	0.005 (0.0003)	2.98 (0.01)	0.960

Table 2

MSEs as well as estimated standard errors (in parentheses) for various methods based on 100 simulation replications in Example 2. Here Full, True, Lasso, Convex, DCE denote the least squares estimates based on the full model, the true model, the Lasso estimate, our convex counterpart based on iteration $m = 0$, and our estimate.

p	n	σ	**Full**	**True model**		**Lasso**	**Our**		**Ave # Iter**	**Ave match proportion**
				Grouping	**Variable**		**Convex**	**DCE**		
50	50	0.58	0.327 (0.007)	0.041 (0.002)	0.142 (0.004)	0.227 (0.006)	0.276 (0.006)	0.123 (0.010)	3.01 (0.07)	0.799
50	100	0.41	0.089 (0.002)	0.010 (0.001)	0.037 (0.001)	0.056 (0.002)	0.075 (0.002)	0.011 (0.001)	3.12 (0.03)	0.862
100	50	0.58	0.325 (0.007)	0.037 (0.002)	0.136 (0.004)	0.311 (0.007)	0.327 (0.007)	0.325 (0.007)	2.17 (0.04)	0.778
100	100	0.41	0.165 (0.003)	0.010 (0.001)	0.034 (0.001)	0.073 (0.002)	0.116 (0.003)	0.110 (0.003)	2.06 (0.02)	0.722

Table 3

Median MSEs based on 100 simulation replications in Example 3. Here Full, True, Lasso, OSCAR, E-NET, Convex, DCE denote the least squares estimates based on the full model, the true model, the Lasso estimate, the elastic net estimate, the OSCAR estimate, our convex counterpart based on iteration $m = 0$, and our estimate. Note that the results for OSCAR and E-NET in Examples 4 and 5 of Bondell and Reich (2008) with $\sigma = 15$ are taken from Table 1 there.

Bondell & Reich	σ	Full	True model		Lasso	E-NET	OSCAR	Our		Ave # Iter	Ave match proportion
			Group	Variable				Conv	DCE		
Ex 4	15	95.1	6.2	47.4	45.4	34.4	25.9	21.4	22.0	2	0.516
Ex 5	15	174.8	11.6	67.4	64.7	40.7	51.8	67.6	70.0	4	0.535
	10	77.7	5.1	30.0	33.1			35.3	38.0	4	0.575
	5	19.4	1.3	7.5	10.7			10.4	6.0	4.5	0.626
	1	0.86	0.05	0.30	0.47			0.43	0.06	3	0.703

Table 4

Estimated group coefficients and group sizes for breast cancer data in Section 5.3.

k	1	2	3	4	5	6	7	8	9
$\hat{\alpha}_k$	−0.634	−0.507	−0.449	−0.416	−0.381	−0.370	−0.244	−0.225	−0.205
$\lvert\mathcal{G}_k\rvert$	1	1	1	2	1	1	2	1	4
k	10	11	12	13	14	15	16	17	18
$\hat{\alpha}_k$	−0.116	−0.059	−0.041	−0.018	−0.017	−0.017	0.006	0.039	0.048
$\lvert\mathcal{G}_k\rvert$	1	2	1	1	1	3	1	6	237
k	19	20	21	22	23	24	25	26	27
$\hat{\alpha}_k$	0.060	0.105	0.110	0.127	0.140	0.247	0.301	0.361	0.392
$\lvert\mathcal{G}_k\rvert$	2	10	3	5	2	1	1	1	2