Simultaneous supervised clustering and feature selection over a graph

BY XIAOTONG SHEN

School of Statistics, University of Minnesota, Minneapolis, Minnesota 55455, U.S.A.
xshen@umn.edu

HSIN-CHENG HUANG

Institute of Statistical Science, Academia Sinica, Taipei 115, Taiwan
hchuang@stat.sinica.edu.tw

AND WEI PAN

Division of Biostatistics, University of Minnesota, Minneapolis, Minnesota 55455, U.S.A.
weip@biostat.umn.edu

SUMMARY

In this article, we propose a regression method for simultaneous supervised clustering and feature selection over a given undirected graph, where homogeneous groups or clusters are estimated as well as informative predictors, with each predictor corresponding to one node in the graph and a connecting path indicating a priori possible grouping among the corresponding predictors. The method seeks a parsimonious model with high predictive power through identifying and collapsing homogeneous groups of regression coefficients. To address computational challenges, we present an efficient algorithm integrating the augmented Lagrange multipliers, coordinate descent and difference convex methods. We prove that the proposed method not only identifies the true homogeneous groups and informative features consistently but also leads to accurate parameter estimation. A gene network dataset is analysed to demonstrate that the method can make a difference by exploring dependency structures among the genes.

Some key words: Expression quantitative trait loci data; High-dimensional data; Nonconvex minimization; Prediction.

1. INTRODUCTION

In high-dimensional data analysis, it is critical to identify and utilize certain lower-dimensional structures to battle the curse of dimensionality. Extracting one kind of lower-dimensional structure defined by clusters has received increasing interest in regression, but less so than exploiting sparseness for feature selection, which has been extensively studied. In gene network analysis, a large amount of genetic knowledge has been organized in networks, for instance, the Kyoto Encyclopedia of Genes and Genomes, a collection of pathway maps representing current knowledge about molecular interactions and reactions. In such situations, identifying homogeneous subnetworks of the genes with similar effect sizes on a biological outcome becomes critical. These homogeneous subnetworks of genes are used to improve our understanding of gene functions and molecular mechanisms underlying a disease. In this article, we study simultaneous identification of homogeneous subgroups of predictors in regression, including

the zero-coefficient group, in the context of simultaneous supervised clustering and feature selection over a given undirected graph.

Supervised clustering of this sort has been under-studied, particularly over an arbitrary undirected graph. The fused lasso (Tibshirani et al., 2005) introduces an L_1 -regularization method for estimating homogeneous subgroups where the predictors can be described a priori in a serial order. Shen & Huang (2010) propose a nonconvex homotopy method for all possible homogeneous subgroups without prior graph information. She (2010) and Jang et al. (2011) propose an L_1 -clustering penalty without prior graph information. Tibshirani & Taylor (2011) develop a convex homotopy method for a generalized lasso problem. Other types of supervised clustering have been considered with an L_∞ -norm-based penalty (Bondell & Reich, 2008), or by a separate step of clustering predictors (Park et al., 2007). Many aspects of supervised clustering remain largely unknown, including its theoretical properties as well as its connection with feature selection, and the potential impact of a given graph on supervised clustering.

Our primary objective in this article is to achieve high accuracy in identification of homogeneous subgroups of coefficients, in addition to that in parameter estimation. To this end, supervised clustering is integrated with feature selection, seeking variance reduction in estimation while controlling bias, leading to higher predictive performance by striking a balanced bias-variance trade-off. On the one hand, feature selection is cast into the framework of supervised clustering by treating a set of noninformative predictors as a zero-coefficient group to be identified. This overcomes the difficulty of selection instability with other feature selection approaches. On the other hand, simultaneous supervised clustering and feature selection are preferable to supervised clustering alone because feature selection achieves variance reduction by removing noninformative predictors. As a result, highly positively correlated predictors with similar regression coefficients are clustered together for both feature selection and parameter estimation, giving better feature selection. The prior knowledge concerning clustering is expressed in a graph, which may not be correct. In an extreme case without any prior knowledge, we simply use a complete graph with edges connecting any pairs of nodes, implying that any two coefficients may be equal. As shown in simulations, in such a situation, our method, by exploiting such potential grouping structures, performs well. In a more extreme case, our method may still have improved performance, even if there is in truth no exact grouping structure. This performance is partly due to the bias-variance trade-off from treating any nonzero coefficients with a similar size as one group, and partly due to the fact that zero coefficients always form a group.

We develop a novel path-following algorithm that integrates difference convex programming with the augmented Lagrange method and the coordinate descent method. This allows us to handle large-scale problems over arbitrary graphs.

There is a paucity of theory on supervised clustering to guide its application in practice. At present, only the fused lasso has been studied for parameter estimation in Rinaldo (2009), where the issue of selection accuracy remains open. This article develops a theory to quantify accuracy of the proposed method for supervised clustering and parameter estimation. We show that the proposed method consistently identifies the true grouping as well as the sparseness structures, leading to recovery of the optimal parameter estimation, where exponentially many candidate predictors may be allowed in the sample size n .

2. PROPOSED METHOD

2.1. Method

Consider linear regression in which an independent random sample $(Y_i, x_i)_{i=1}^n$ follows

$$Y_i = \beta^T x_i + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2) \quad (i = 1, \dots, n), \quad (1)$$

where $\beta = (\beta_1, \dots, \beta_p)^\top$ and $x_i = (x_{i1}, \dots, x_{ip})^\top$ are p -dimensional vectors of regression coefficients and predictors, and x_i is independent of ϵ_i . Here each predictor corresponds to one node in a given undirected graph describing prior knowledge about possible clustering structures, which may or may not be correctly specified. That is, an edge connecting two nodes indicates that the corresponding two predictors may be grouped. In (1), rewrite β as $(0_{|\mathcal{G}_0|}, \alpha_1 1_{|\mathcal{G}_1|}, \dots, \alpha_K 1_{|\mathcal{G}_K|})^\top$, where $K + 1$ is the number of distinct groups, $\alpha_1 < \dots < \alpha_K$ are nonzero coefficients, and $0_{|\mathcal{G}_0|}$ and $1_{|\mathcal{G}_1|}$ denote vectors having 0 and 1 elements of lengths $|\mathcal{G}_0|$ and $|\mathcal{G}_1|$, respectively. This formulation is general; in particular, if $K = p$, or each group $|\mathcal{G}_1| = \dots = |\mathcal{G}_K| = 1$, it corresponds to no grouping. Simultaneous supervised clustering and feature selection estimate $\beta = (0_{|\mathcal{G}_0|}, \alpha_1 1_{|\mathcal{G}_1|}, \dots, \alpha_K 1_{|\mathcal{G}_K|})^\top$, or equivalently, $\mathcal{G} = (\mathcal{G}_0, \mathcal{G}_1, \dots, \mathcal{G}_K)^\top$ and $\alpha = (\alpha_1, \dots, \alpha_K)^\top$.

Let $\mathcal{E} \subset \{1, \dots, p\}^2$ be a set of edges between two distinct nodes $j \neq j'$ for an undirected graph, with $j \sim j'$ indicating an edge directly connecting the two nodes. We propose a constrained least squares criterion for simultaneous supervised clustering and feature selection:

$$\min_{\beta} (2n)^{-1} \sum_{i=1}^n (Y_i - x_i^\top \beta)^2,$$

subject to

$$\sum_{j=1}^p \min\left(\frac{|\beta_j|}{\tau}, 1\right) \leq s_1, \quad \sum_{j < j': (j, j') \in \mathcal{E}} \min\left(\frac{|\beta_j - \beta_{j'}|}{\tau}, 1\right) \leq s_2, \quad (2)$$

where s_1, s_2 are positive tuning parameters controlling feature selection and supervised clustering, and τ is a thresholding parameter determining when a large regression coefficient or a large difference between two coefficients will not be penalized. This criterion can be thought of as a surrogate of the L_0 -method for model selection, where $\min(|\beta_j|/\tau, 1)$ approximates the L_0 -function $I(\beta_j \neq 0)$ when $\tau \rightarrow 0^+$ (Shen et al., 2012). The L_0 -function and this surrogate are not smooth, in contrast to other approximations (Dicker et al., 2012). This surrogate overcomes the computational infeasibility of the L_0 -method through an efficient method, yielding efficient computation and enjoying desirable theoretical properties of the L_0 -method as described in Theorems 2 and 3.

Next we characterize a global minimizer of (2).

LEMMA 1. *A global minimizer of (2) is a local minimizer of its unconstrained version:*

$$S(\beta) = (2n)^{-1} \sum_{i=1}^n (Y_i - x_i^\top \beta)^2 + \lambda_1 \sum_{j=1}^p \min\left(\frac{|\beta_j|}{\tau}, 1\right) + \lambda_2 \sum_{j < j': (j, j') \in \mathcal{E}} \min\left(\frac{|\beta_j - \beta_{j'}|}{\tau}, 1\right), \quad (3)$$

where λ_1, λ_2 correspond to s_1, s_2 in (2). Moreover, the global minimizer satisfies the local optimality condition

$$-n^{-1} (x^{(j)})^\top (Y - X\beta) + \frac{\lambda_1}{\tau} b_j + \frac{\lambda_2}{\tau} \sum_{j': (j, j') \in \mathcal{E}} b_{jj'} = 0 \quad (j = 1, \dots, p), \quad (4)$$

where $x^{(j)} = (x_{1j}, \dots, x_{nj})^\top$, $b_j = \text{sign}(\beta_j) I(|\beta_j| < \tau)$ if $\beta_j \neq 0$, $b_{jj'} = \text{sign}(\beta_j - \beta_{j'}) I(|\beta_j - \beta_{j'}| < \tau)$ if $\beta_j - \beta_{j'} \neq 0$, $b_j \in [-1, 1]$ if $\beta_j = 0$, $b_{jj'} \in [-1, 1]$ if $\beta_j = \beta_{j'}$, $b_j = 0$ if $|\beta_j| = \tau$, and

$b_{jj'} = \emptyset$ if $|\beta_j - \beta_{j'}| = \tau$ are the regular subdifferentials of $\min(|\beta_j|, \tau)$ and $\min(|\beta_j - \beta_{j'}|, \tau)$, and $\emptyset$ indicates the empty set; see Rockafellar & Wets (2003).

Based on (4), we develop a difference convex method leading to a local minimizer of (3) satisfying (4). This strategy produces a local minimizer sharing key properties of its global minimizer; see Theorem 3. For (2), developing effective computational strategies is necessary. It is known that the coordinate descent strategy is computationally efficient for a lasso problem; see Friedman et al. (2007) for a comparison with the homotopy strategy. However, as pointed out by Friedman et al. (2007), the coordinate descent strategy breaks down when the objective or penalty function is not separable in terms of each variable as for the fused lasso and our problem here. Our method is difference convex based, integrated with the augmented Lagrange and coordinate descent methods, for efficient computation.

2.2. Difference convex programming and the augmented Lagrange method

Our difference convex method approximates a nonconvex cost function by linearizing the trailing convex function through a difference convex decomposition of the cost function; see An & Tao (1997). At each iteration, a quadratic problem with equality constraints is solved by the augmented Lagrange method; see Fortin & Glowinski (1983).

To proceed, we decompose $S(\beta)$ in (3) into a difference $S_1(\beta) - S_2(\beta)$ of two convex functions:

$$S_1(\beta) = (2n)^{-1} \sum_{i=1}^n (Y_i - x_i^\top \beta)^2 + \frac{\lambda_1}{\tau} \sum_{j=1}^p |\beta_j| + \frac{\lambda_2}{\tau} \sum_{j < j': (j, j') \in \mathcal{E}} |\beta_j - \beta_{j'}|,$$

$$S_2(\beta) = \frac{\lambda_1}{\tau} \sum_{j=1}^p (|\beta_j| - \tau)_+ + \frac{\lambda_2}{\tau} \sum_{j < j': (j, j') \in \mathcal{E}} (|\beta_j - \beta_{j'}| - \tau)_+,$$

where z_+ is the positive part of z . This decomposition is interpretable in that $S_2(\cdot)$ corrects the estimation bias due to imposing convexity on a nonconvex problem. We then construct a sequence of upper approximations iteratively by replacing $S_2(\beta)$ at iteration m by its piecewise affine minorization $S_2(\hat{\beta}^{(m-1)}) + \tau^{-1} \lambda_2 \sum_{j < j': (j, j') \in \mathcal{E}} I(|\hat{\beta}_j^{(m-1)} - \hat{\beta}_{j'}^{(m-1)}| \geq \tau)(|\beta_j - \beta_{j'}| - |\hat{\beta}_j^{(m-1)} - \hat{\beta}_{j'}^{(m-1)}|) + \tau^{-1} \lambda_1 \sum_{j=1}^p I(|\hat{\beta}_j^{(m-1)}| \geq \tau)(|\beta_j| - |\hat{\beta}_j^{(m-1)}|)$ at iteration $m - 1$, leading to an upper convex approximating function at iteration m :

$$S^{(m)}(\beta) = (2n)^{-1} \sum_{i=1}^n (Y_i - x_i^\top \beta)^2 + \frac{\lambda_1}{\tau} \sum_{j: j \in \mathcal{F}^{(m-1)}} |\beta_j| + \frac{\lambda_2}{\tau} \sum_{j < j': (j, j') \in \mathcal{E}^{(m-1)}} |\beta_j - \beta_{j'}|, \quad (5)$$

where

$$\mathcal{E}^{(m-1)} = \{(j, j') \in \mathcal{E}, |\hat{\beta}_j^{(m-1)} - \hat{\beta}_{j'}^{(m-1)}| < \tau\},$$

$$\mathcal{F}^{(m-1)} = \{j \in (1, \dots, p), |\hat{\beta}_j^{(m-1)}| < \tau\}.$$

Minimization in (5) can be solved through quadratic programming. However, this becomes inefficient for large-scale problems because it is designed for both inequality and equality constraints. Here we develop an efficient algorithm through the augmented Lagrange and coordinate

descent methods. First, we convert a problem with linear constraints to its unconstrained version through slack variables $\beta_{jj'} = \beta_j - \beta_{j'}$ for $j \neq j'$ to work with an augmented equivalent problem of (5) in $\zeta = (\beta_1, \dots, \beta_p, \beta_{12}, \dots, \beta_{1p}, \dots, \beta_{(p-1)p})^T$,

$$\tilde{S}^{(m)}(\zeta) = (2n)^{-1} \sum_{i=1}^n (Y_i - x_i^T \beta)^2 + \frac{\lambda_1}{\tau} \sum_{j: j \in \mathcal{F}^{(m-1)}} |\beta_j| + \frac{\lambda_2}{\tau} \sum_{j < j': (j, j') \in \mathcal{E}^{(m-1)}} |\beta_{jj'}|, \quad (6)$$

subject to the linear constraints $\beta_{jj'} = \beta_j - \beta_{j'}$ for $j \neq j'$. For (6), we employ the augmented Lagrange method to solve its equivalent unconstrained problem iteratively with respect to t for iteration m . At iteration t , we minimize

$$\begin{aligned} \bar{S}^{(m)}(\zeta) = \tilde{S}^{(m)}(\zeta) + \sum_{j < j': (j, j') \in \mathcal{E}} \tau_{jj'}^{(t)} (\beta_j - \beta_{j'} - \beta_{jj'}) + 2^{-1} \\ \times \sum_{j < j': (j, j') \in \mathcal{E}} v_{jj'}^{(t)} (\beta_j - \beta_{j'} - \beta_{jj'})^2, \end{aligned} \quad (7)$$

where $\tau_{jj'}^{(t)}$ and $v_{jj'}^{(t)}$ are Lagrange multipliers for the linear constraints and for computational acceleration. Minimizing (7) over ζ yields $\hat{\zeta}^{(m,t)}$ for a given value of $(\tau_{jj'}^{(t)}, v_{jj'}^{(t)})$, which is updated: $\tau_{jj'}^{(t+1)} = \tau_{jj'}^{(t)} + v_{jj'}^{(t)} (\hat{\beta}_j^{(m,t)} - \hat{\beta}_{j'}^{(m,t)} - \hat{\beta}_{jj'}^{(m,t)})$; $v_{jj'}^{(t+1)} = \rho v_{jj'}^{(t)}$, where $\rho > 1$ will be chosen to speed up convergence. To compute $\hat{\zeta}^{(m,t)}$ from (7), we use a coordinate descent method to minimize (7) in each component of ζ while fixing the other components at their current values, where an analytic updating formula is derived:

$$\begin{aligned} \tilde{\beta}_j^{(m,t)} = \left[n^{-1} \sum_{i=1}^n x_{ij}^2 + \frac{1}{\tau} |\{j' : (j, j') \in \mathcal{E}\}| \right]^{-1} \text{ST} \left\{ n^{-1} \sum_{i=1}^n x_{ij} (y_i - \tilde{y}_i^{(j)}) \right. \\ \left. - \sum_{(j, j') \in \mathcal{E}} \tau_{jj'}^{(t)} + v_{jj'}^{(t)} \sum_{j': (j, j') \in \mathcal{E}} (\tilde{\beta}_{j'}^{(m,t)} + \tilde{\beta}_{jj'}^{(m,t)}), \delta_j \right\} \quad (j = 1, \dots, p), \end{aligned} \quad (8)$$

$$\tilde{\beta}_{jj'}^{(m,t)} = (v_{jj'}^{(t)})^{-1} \text{ST} \{ v_{jj'}^{(t)} (\tilde{\beta}_j^{(m,t)} - \tilde{\beta}_{j'}^{(m,t)}) + \tau_{jj'}^{(t)}, \delta_{jj'} \}, \quad (j, j') \in \mathcal{E}. \quad (9)$$

Here $\tau_{j'j}^{(t)} = -\tau_{jj'}^{(t)}$, $\tilde{\beta}_{j'j}^{(m,t)} = -\tilde{\beta}_{jj'}^{(m,t)}$, $\text{ST}(b, \lambda) = \text{sign}(b)(|b| - \lambda)_+$ is the soft-thresholding operator, $\tilde{y}_i^{(j)} = \sum_{j' \neq j} x_{ij'} \hat{\beta}_{j'}^{(m-1)}$, $\delta_j = \tau^{-1} \lambda_1 I(|\hat{\beta}_j^{(m-1)}| \leq \tau)$, and $\delta_{jj'} = \tau^{-1} \lambda_2 I(|\hat{\beta}_{jj'}^{(m-1)}| \leq \tau)$ for $(j, j') \in \mathcal{E}$. This process of coordinate descent iterates until convergence, giving a global minimizer $\hat{\zeta}^{(m)} = \hat{\zeta}^{(m, t^*)}$, or $\beta^{(m)} = \hat{\beta}^{(m, t^*)}$, with t^* denoting the iteration index at termination, which is assured by Theorem 1 of Tseng (2001). Then the difference convex iteration proceeds in m until convergence. The final solution $\hat{\beta} = \hat{\beta}^{(m^*)}$, where m^* is the iteration index at termination. Algorithm 1 below summarizes the proposed method.

2.3. Algorithm 1: path-following algorithm

Specify an upper bound K^* of K , which controls the maximum number of distinct groups allowed for Algorithm 1, where $K^* \leq \min(n, p)$. Specify tolerance error $\varepsilon > 0$, say $\varepsilon = 10^{-3}$, $\rho = 2$, $\tau_{jj'}^{(0)} = 0$ and $v_{jj'}^{(0)} = 1$. For given τ ,

Step 1. Specify evaluation points of λ_1 and λ_2 with each in a descent order from the largest to the smallest. Supply a good initial estimate $\hat{\beta}^{(0)}$, such as $\hat{\beta}^{(0)}$ at the previous evaluation points or $\hat{\beta}^{(0)} = 0$ when $\lambda_1 = \lambda_2 = \tau = \infty$.

Step 2. Iteration begins with $m = 1$. At iteration m , compute $\hat{\beta}^{(m)}$ by solving (7) through the coordinate descent method as described in (8) and (9), with tolerance error $\varepsilon > 0$. For coordinate updates, proceed with the active set of nonzero coordinates first, followed by zero coordinates, as in Friedman et al. (2007).

Step 3. Terminate when $S(\hat{\beta}^{(m-1)}) - S(\hat{\beta}^{(m)}) \leq 0$ and no component of $\hat{\beta}^{(m)}$ is at $\pm\tau$. Otherwise, perturb the components at $\pm\tau$ slightly and go to Step 2. Then the final estimate is $\hat{\beta}^{(m^*-1)}$, where m^* is the smallest index at termination.

THEOREM 1. *In Algorithm 1, the rate of convergence for the augmented Lagrange component is linear and $S(\hat{\beta}^{(m)})$ decreases strictly in m unless $\hat{\beta}^{(m)} = \hat{\beta}^{(m-1)}$. Algorithm 1 terminates in a finite number of steps, that is, $m^* < \infty$ with*

$$\hat{\beta}^{(m)} = \hat{\beta}^{(m^*-1)} \quad (m \geq m^*). \quad (10)$$

3. NUMERICAL EXAMPLES

3.1. Set-up

We now examine the effectiveness of the proposed method on three simulated examples and a real application to gene network analysis. For comparison, we consider the lasso, generalized lasso, fused lasso and elastic net, whenever appropriate. In addition, we apply the proposed method for supervised clustering alone and for feature selection alone respectively, to demonstrate the need for simultaneous supervised clustering and feature selection. Moreover, we examine the effect of prior knowledge, expressed in terms of $\mathcal{E}$, on the performance of the proposed method, in Examples 2 and 3. Finally, the ideal optimal performance of the least squares estimator given the true grouping and sparseness structures is supplied for reference.

For accuracy of parameter estimation, we use the mean squared error, $\text{MSE}(\hat{\beta}) = n^{-1} \|X(\hat{\beta} - \beta^0)\|^2$. For accuracy of simultaneous supervised clustering and feature selection, we define $\text{Acc}_{\text{all}} = \sum_{k=0}^{K^0} s_k / (K^0 + 1)$, with $s_0 = \{\sum_{j \in \mathcal{G}_0^0} I(\hat{\beta}_j = 0) + \sum_{j \notin \mathcal{G}_0^0} I(\hat{\beta}_j \neq 0)\} / p$ concerning feature selection and $s_k = 1 - \{\sum_{j \neq j': j, j' \in \mathcal{G}_k^0} I(\hat{\beta}_{j'} = \hat{\beta}_j) + \sum_{j' \neq j: j \in \mathcal{G}_k^0, j' \notin \mathcal{G}_k^0} I(\hat{\beta}_{j'} \neq \hat{\beta}_j)\} / \{|\mathcal{G}_k^0|(p-1)\}$ measuring errors in identifying the true group $\mathcal{G}_k^0$, where $\mathcal{G}^0 = (\mathcal{G}_0^0, \mathcal{G}_1^0, \dots, \mathcal{G}_{K^0}^0)^\top$ is the true grouping structure. Note that Acc_{all} lies between 0 and 1.

3.2. Simulations

Simulations are performed in several scenarios, including correlated predictors, situations of small p and large n or small n but large p , and with various graphs. In each simulated example, we minimize the mean squared error with respect to tuning parameters for all competing methods over a set of grid points based on an independent tuning set of size n , and report the minimal mean squared error value as well as Acc_{all} at the corresponding selected tuning parameters. Specifically, $(\lambda_1, \lambda_2) \in A^2 \setminus \{(0, 0)\}$ for the proposed method, the fused lasso and the generalized lasso, where $\setminus$ denotes the set difference. In addition, $\tau \in \{0.01, 0.1, 1, 10, 100\}$ for the proposed method, where $A = \{\tau_m, \tau_m/a, \tau_m/a^2, \dots, \tau_m/a^{11}, 0\}$ with τ_m the smallest τ value setting all

Table 1. *Parameter estimation in Example 1: mean squared errors, their largest standard error over the seven competitors, and the numbers of difference convex iterations of Algorithm 1 based on the selected tuning parameters, over 100 simulation replications with $n = 100$*

p	ρ	k	σ	Select	Group	Fused		E-Net	Lasso	True	Max	Number of Iterations		
				Alone	Alone	Both	Lasso				SE	Select	Group	Both
50	0.5	1	1	0.07	0.15	0.07	0.20	0.18	0.20	0.05	0.01	3.3	8.2	6.0
50	0.5	1	5	3.14	5.47	2.29	3.65	3.14	3.65	1.27	0.19	4.8	6.0	4.7
50	0.5	5	1	0.23	0.10	0.09	0.17	0.32	0.33	0.05	0.01	6.6	7.1	4.9
50	0.5	5	5	8.24	3.22	2.80	3.87	7.67	8.22	1.29	0.22	6.7	3.4	4.1
50	-0.5	1	1	0.07	0.10	0.07	0.16	0.14	0.17	0.05	0.01	3.6	10.0	7.5
50	-0.5	1	5	3.46	4.50	2.95	3.44	3.21	4.01	1.27	0.16	4.1	4.5	6.7
50	-0.5	5	1	0.28	0.09	0.08	0.18	0.42	0.40	0.05	0.01	10.1	9.3	10.5
50	-0.5	5	5	9.45	3.35	2.94	4.37	9.40	9.59	1.34	0.21	6.6	5.5	6.0
200	0.5	4	1	0.23	0.15	0.08	0.26	0.44	0.46	0.06	0.01	7.5	4.1	4.6
200	0.5	4	5	10.02	5.49	3.10	5.99	9.50	9.59	1.37	0.24	8.5	3.6	4.3
200	0.5	20	1	0.91	0.11	0.09	0.28	0.99	0.88	0.05	0.01	35.1	4.1	4.7
200	0.5	20	5	21.62	3.47	3.05	5.30	21.56	20.47	1.32	0.34	46.1	3.4	4.3
200	-0.5	4	1	0.46	0.13	0.07	0.24	0.76	0.72	0.05	0.01	15.8	7.9	6.1
200	-0.5	4	5	13.12	4.75	3.17	6.18	12.64	13.25	1.29	0.24	8.3	4.7	5.4
200	-0.5	20	1	0.95	0.11	0.09	0.27	0.99	0.91	0.05	0.01	33.0	8.4	8.7
200	-0.5	20	5	21.87	4.15	3.46	5.70	21.27	20.56	1.31	0.30	38.5	5.6	6.4
500	0.5	10	1	0.71	0.19	0.10	0.36	0.96	0.89	0.06	0.02	17.9	4.9	4.7
500	0.5	10	5	19.80	5.77	3.31	7.52	18.69	19.25	1.50	0.33	23.9	4.1	4.1
500	0.5	50	1	1.00	0.15	0.10	0.44	1.02	0.97	0.03	0.01	31.1	5.2	5.1
500	0.5	50	5	24.73	4.00	3.62	6.66	24.79	23.97	0.78	0.33	47.5	4.1	4.3
500	-0.5	10	1	0.92	0.16	0.08	0.35	1.00	0.96	0.06	0.02	17.9	4.9	4.7
500	-0.5	10	5	19.90	6.05	3.50	7.80	19.65	17.84	1.51	0.27	10.8	5.6	5.8
500	-0.5	50	1	1.00	0.13	0.10	0.40	1.01	0.98	0.04	0.01	33.5	12.4	9.9
500	-0.5	50	5	24.16	4.76	4.05	6.75	23.95	23.73	0.88	0.31	41.9	5.5	5.2

True, the oracle estimates; Select Alone, Group Alone and Both, our method for feature selection alone, supervised clustering alone and both; Max SE, the largest standard error over the seven competitors.

the lasso parameter estimates to be zeros, and a is chosen such that $\tau_m/a^{11} = 0.001$ for Example 1, and 0.1 for Examples 2 and 3. For the elastic net, we use the default values of the glmnet package in R (R Development Core Team, 2012) for tuning the L_1 -penalty, while six additional values $\{0, 0.01, 0.1, 1, 10, 100\}$ are used for tuning the L_2 -penalty, as in Zou & Hastie (2005). All numerical analyses are conducted in R.

Example 1. A random sample of $\{(x_i, Y_i) : i = 1, \dots, n\}$ is obtained, where Y_i follows (1), x_i is sampled from $N(0, \Sigma)$ with the (j, k) th element of Σ being $\rho^{|j-k|}$, $n = 100$, $\rho = 0.5, -0.5$, $p = 50, 200, 500$, $\sigma = 1, 5$, and

$$\beta = (\underbrace{3, \dots, 3}_k, \underbrace{-1.5, \dots, -1.5}_k, \underbrace{1, \dots, 1}_k, \underbrace{2, \dots, 2}_k, \underbrace{0, \dots, 0}_{p-4k})^T,$$

with $k = p/10$. To mimic a lack of prior knowledge on grouping, a complete graph is used for $\mathcal{E}$, where any two nodes are directly connected by an edge.

Concerning parameter estimation, the improvement is significant over the fused lasso, elastic net and lasso, as indicated in Table 1. There the largest improvement occurs for larger p and positive ρ . This is primarily due to the nonconvex constraint/regularization, and to the fact that grouping takes place towards more positively than negatively correlated predictors. In view of

Table 2. Accuracy of simultaneous supervised clustering and feature selection, as defined by Acc_{all} , and their largest standard error over the seven competitors, in Example 1 for various methods based on 100 simulation replications with $n = 100$

p	ρ	k	σ	Select Alone	Group Alone	Both	Fused Lasso	Elastic Net	Lasso	Max SE
50	0.5	1	1	95.9	79.0	92.5	89.0	92.3	91.4	0.6
50	0.5	1	5	72.8	77.4	83.3	79.3	73.2	75.2	1.5
50	0.5	5	1	92.3	86.6	92.5	84.2	88.5	87.8	0.6
50	0.5	5	5	83.9	85.8	89.5	85.2	86.1	85.9	0.4
50	-0.5	1	1	95.9	78.7	90.4	89.1	93.6	92.3	0.7
50	-0.5	1	5	71.8	77.3	79.7	83.1	74.3	80.8	1.6
50	-0.5	5	1	91.1	86.7	92.2	83.7	85.3	84.8	0.5
50	-0.5	5	5	80.6	85.2	89.4	84.9	81.3	81.8	0.6
200	0.5	4	1	98.3	82.7	96.9	89.5	96.1	93.1	0.5
200	0.5	4	5	79.5	81.2	92.5	89.7	83.5	82.7	0.8
200	0.5	20	1	89.5	87.7	93.7	83.8	81.1	77.0	0.5
200	0.5	20	5	77.5	86.0	89.8	84.3	75.3	74.0	0.4
200	-0.5	4	1	95.8	82.1	96.2	87.8	92.1	89.5	0.5
200	-0.5	4	5	64.3	81.0	91.9	86.6	65.9	61.9	1.8
200	-0.5	20	1	86.9	86.3	92.5	82.8	72.7	69.3	0.5
200	-0.5	20	5	66.4	84.7	89.5	83.4	64.4	66.9	0.6
500	0.5	10	1	97.7	82.9	96.7	90.7	91.4	89.9	0.6
500	0.5	10	5	75.8	81.6	93.1	86.8	77.7	77.6	0.8
500	0.5	50	1	81.5	84.8	91.4	82.6	61.7	47.2	2.0
500	0.5	50	5	59.3	84.3	88.5	82.5	52.0	47.4	1.4
500	-0.5	10	1	92.0	82.1	95.2	86.1	60.0	54.9	0.9
500	-0.5	10	5	48.5	81.2	92.4	84.7	46.3	46.9	0.7
500	-0.5	50	1	75.1	82.5	89.9	81.7	54.5	42.9	1.7
500	-0.5	50	5	46.2	82.3	85.6	81.8	43.3	41.7	0.8

Select Alone, Group Alone and Both, our method for feature selection alone, supervised clustering alone and both; Max SE, the largest standard error over the seven competitors.

the mean squared error value of the oracle estimator, the least squares estimator based on the true grouping, the proposed estimator performs reasonably well but there may be room for improvement for larger p and positive ρ .

Concerning accuracy of supervised clustering and feature selection, the proposed method is again best except when $(p, \rho, k, \sigma) = (50, -0.5, 1, 5)$, as indicated in Table 2. To understand the impact of supervised clustering on feature selection, we observe that the former does help to improve performance of the latter, and vice versa. This is evident from comparing cases of selection alone and grouping alone against those of both, where the latter generally outperforms either of the first two cases. In many situations, the amount of improvement is rather substantial. In summary, simultaneous supervised clustering and feature selection are advantageous over either alone when both sparse and grouping structures are present.

With regard to the computational efficiency of Algorithm 1, the number of difference convex iterations required to converge is on average about six even when p gets large. Algorithm 1 converges faster for simultaneous supervised clustering and feature selection than for feature selection alone, suggesting that supervised learning and feature selection are complementary to each other in the selection process.

Example 2. Consider a regulatory gene network in Li & Li (2008), where the entire network consists of n_{TF} subnetworks, each with one transcription factor and ten regulatory target genes,

Table 3. *Parameter estimation in Example 2: mean squared errors, their largest standard error over the seven competitors, and the numbers of difference convex iterations of Algorithm 1 based on the selected tuning parameters, over 100 simulation replications with $n = 100$*

p	Select Alone	Group Alone	Both	Gen Lasso	Elastic Net	Lasso	Nonzero Clusters	Max SE	Number of Iterations		
									Select	Group	Both
110	15.6	6.1	5.0	5.5	14.8	15.3	3.3	0.37	35.4	2.8	3.6
220	17.1	9.6	4.9	6.6	15.8	16.1	3.2	0.36	30.6	3.0	4.0
550	19.0	19.0	4.8	7.8	16.9	18.8	3.2	0.40	23.1	3.8	3.5
1100	20.6	34.0	4.6	9.0	17.8	21.5	3.1	0.47	25.4	49.4	3.3
2200	21.3	33.7	4.6	10.2	18.7	21.0	3.1	0.71	21.5	9.1	3.4

Select Alone, Group Alone and Both, our method for feature selection alone, supervised clustering alone and both; Nonzero Clusters, the ordinary least squares estimates under the model with only four clusters of variables corresponding to the first four connected components; Max SE, the largest standard error over the seven competitors.

and each transcription factor predictor follows the $N(0, 1)$ distribution; see Li & Li (2008) for a display of the network. To mimic a regulatory relationship, the predictor of each target gene and the transcription factor predictor follow a bivariate normal distribution with correlation $\rho = 0.7$, and target genes are independent, conditional on the transcription factor predictor. A random sample $\{(x_i, Y_i) : i = 1, \dots, n\}$ is obtained, where Y_i follows (1) with $\sigma^2 = \sum_{j=1}^p \beta_j^2/4$, and β equals

$$10^{-1/2} \left\{ 5(10)^{1/2}, \underbrace{5, \dots, 5}_{10}, -5(10)^{1/2}, \underbrace{-5, \dots, -5}_{10}, 3(10)^{1/2}, \underbrace{3, \dots, 3}_{10}, -3(10)^{1/2}, \right. \\ \left. \underbrace{-3, \dots, -3}_{10}, \underbrace{0, \dots, 0}_{p-44} \right\}^T.$$

Here set-up 3 of Li & Li (2008) is examined with $n = 100$ and $p = 110, 220, 550, 1100, 2200$, and $\mathcal{E}$ is a graph consisting of $p/11$ subgraphs; each subgraph contains a transcription factor connecting to each of its ten target genes but with no other edges. The graph $\mathcal{E}$ is incorrectly specified because the target genes sharing the same coefficient value are not directly connected. The penalty matrix for the generalized lasso is specified by $\mathcal{E}$.

The proposed method continues to perform well. For parameter estimation, the amount of improvement in terms of the mean squared error is evident over the elastic net, the lasso and the generalized lasso, as suggested by Table 3. The relative performance of the proposed method over its competitors becomes better and more stable, as p increases, due to the nonconvex constraint and regularization. The overall accuracy Acc_{all} appears to be quite high across all situations, as indicated in Table 4. For computational efficiency, the number of difference convex iterations required for Algorithm 1 is on average about four even when p becomes large, in contrast to 21–35 for feature selection alone; see Table 4. Overall, the conclusions in Example 1 remain valid with a misspecified graph or grouping structure.

Example 3. This example considers the set-up in Example 2 except that the first 44 regression coefficients are independently generated from continuous uniform distributions with the length of support equal to ℓ , where $\beta_1, \dots, \beta_{11} \sim \text{Un}(5 - \ell/2, 5 + \ell/2)$, $\beta_{12}, \dots, \beta_{22} \sim \text{Un}(-5 - \ell/2, -5 + \ell/2)$, $\beta_{23}, \dots, \beta_{33} \sim \text{Un}(3 - \ell/2, 3 + \ell/2)$, and $\beta_{34}, \dots, \beta_{44} \sim \text{Un}(-3 -$

Table 4. *Accuracy of simultaneous supervised clustering and feature selection, as defined by Acc_{all} , and their largest standard error over the seven competitors, in Example 2 for various methods based on 100 simulation replications with $n = 100$*

p	Select Alone	Group Alone	Both	Generalized Lasso	Elastic Net	Lasso	Max SE
110	73.1	87.0	94.4	90.8	74.3	79.7	1.3
220	72.5	83.6	96.7	92.3	73.3	75.3	1.4
550	63.7	81.6	97.7	94.7	66.4	69.5	1.5
1100	58.8	80.8	99.1	96.4	62.4	64.7	1.4
2200	52.0	80.6	99.3	97.2	57.1	61.3	1.4

Table 5. *Parameter estimation in Example 3: mean squared errors and their largest standard error over the seven competitors based on the selected tuning parameters over 100 simulation replications with $n = 100$*

n	p	ℓ	Select Alone	Group Alone	Both	Gen Lasso	E-Net	L_0 E-Net	Lasso	Nonzero Coeff	Nonzero Clusters	Max SE
2000	110	1	4.3	2.6	2.2	2.3	5.4	4.4	5.5	4.2	2.4	0.11
2000	110	2	4.3	4.8	3.7	4.0	5.5	4.5	5.6	4.3	7.1	0.12
2000	110	4	4.7	6.8	4.4	5.3	5.8	4.8	5.9	4.5	25.7	0.45
500	110	1	17.5	5.9	4.3	5.1	22.4	18.7	22.0	16.7	4.9	0.44
500	110	2	18.1	10.4	9.0	9.4	22.7	19.4	22.3	17.0	9.5	0.45
500	110	4	19.3	20.3	16.4	16.8	25.6	20.8	23.3	17.9	28.3	0.49
100	110	1	95.4	24.1	15.4	19.3	93.4	95.7	92.8	85.8	18.7	1.76
100	110	2	95.9	28.6	20.6	23.6	94.5	96.3	93.6	87.0	23.0	1.76
100	110	4	98.7	44.8	38.3	39.6	96.9	98.6	96.1	91.6	40.1	2.06
100	220	1	99.9	42.4	15.7	22.7	98.9	101.1	99.2	84.1	17.7	1.76
100	220	2	100.1	46.4	20.9	26.5	99.7	102.0	100.1	85.2	22.0	1.82
100	220	4	103.2	62.2	39.2	42.0	103.1	104.5	103.5	89.6	39.1	2.06
100	550	1	111.7	96.0	15.8	30.1	106.8	110.6	109.0	82.4	18.5	2.09
100	550	2	113.5	100.2	20.8	34.6	108.0	112.0	110.2	83.7	23.1	2.17
100	550	4	118.2	114.5	40.0	50.4	112.0	115.9	113.8	88.4	41.9	2.46

Select Alone, Group Alone and Both, our method for feature selection alone, supervised clustering alone and both; Nonzero Coefficients and Nonzero Clusters, the ordinary least squares estimates under the model with only the nonzero coefficient variables and with only four clusters of variables corresponding to the first four connected components; Max SE, the largest standard error over the seven competitors.

$\ell/2, -3 + \ell/2$). This generates an extreme situation in which none of the nonzero coefficients are identical; that is, there is no other grouping structure except for the zero-coefficient group.

We consider $n = 100, 500, 2000$, $p = 110, 220, 550$ and $\ell = 1, 2, 4$ and examine the performance of the methods in Example 2. In addition, we study the clustering effect by comparing an L_0 -version of the elastic net with its L_1 penalty replaced by our computational surrogate.

As shown in Table 5, the proposed method continues to outperform its competitors in all the situations, even when all nonzero coefficients differ. Interestingly, the L_0 -version of the elastic net does slightly worse than the elastic net in this situation of absence of the clustering effect, except when n is larger than p . By comparison, the proposed method is robust even in this case, because the grouping structure is estimated. Collapsing similar coefficients may improve prediction accuracy, even though the true coefficients are not identical, particularly so in a high-dimensional situation. Ordinary least squares estimates under the true model based on the first 44 variables perform much worse than those based on four clusters of variables corresponding to the first four connected components, except when n is much larger than p . This can be explained by the bias/variance trade-off. Overall, the proposed method's robust performance is due to its flexibility of allowing different degrees of clustering.

Fig. 1. Plot of a coexpression network from data in Ghazalpour et al. (2006), as an undirected graph with 17 nodes and 22 edges. The plot is generated in Cytoscape (Shannon et al., 2003). Here the proposed method identifies 13 genes, which are marked in grey and are associated with at least one genomic marker, including the two hexagonal ones identified by the lasso.

3.3. Network-based expression quantitative trait loci data analysis

In expression quantitative trait loci analysis, the goal is to identify genomic loci that are linked to gene expression traits. This section analyses a mouse dataset in Lan et al. (2006), consisting of 60 F2 mice from B6 and BTBR founder strains, where the B6 and BTBR strains are diabetes resistant and nonresistant, respectively. About 45 000 gene expression traits are measured, together with genotypes of 194 markers distributed across the mouse genome with an average marker interval of approximately 10 cM. Our goal is to improve power in detecting expression quantitative trait loci for a group of co-regulated genes.

Here we consider a subgroup of the genes in the G protein-coupled receptor protein signaling pathway. We employ a coexpression network as in Pan (2009). In particular, we use another dataset of 135 mice from different strains in Ghazalpour et al. (2006); we identify 17 G protein-coupled receptor genes appearing in both datasets, and construct a network with edges corresponding to pairwise gene expression correlations of 0.4 or higher; see Fig. 1. Since coexpression is related to coregulation, it is reasonable to assume that the genes connected in a coexpression network are likely to share some common expression quantitative trait loci; that is, if two genes are connected in a coexpression network, their expression levels are likely to be associated with the genotypes at some common genomic loci.

Let Y_g denote a vector of length 60 representing the expression levels of gene g , linking to marker data through a linear model:

$$Y_g = X_0 \beta_g + \epsilon_g, \quad \epsilon_g \sim N(0, \sigma^2 I) \quad (g = 1, \dots, G), \quad (11)$$

where X_0 is a 60×194 design matrix, representing the genotypes of the 194 markers across the 60 mouse genomes. Each genotype takes the value -1 , 0 or 1 , indicating one of the three alleles, and β_g is a vector of regression coefficients of length 194. Note (11) specifies G gene-specific models as opposed to a single model (1). Also available is a predictor network induced by the coexpression network of the seventeen genes; that is, if genes g_1 and g_2 are connected in the coexpression network, then β_{g_1} and β_{g_2} can be grouped component wise. It is of particular

Table 6. *Markers with nonzero regression coefficients for each gene as estimated by the proposed method in § 3.3, the lasso and the group lasso*

Gene	Ours			Lasso			Group Lasso		
Calcr1				154					
Ccr5			101						
Rps6ka4	30				26	27	30		
Rgs3				153	154				
1200007D18Rik	30								
Rassf1									
Lphn1									
Kcnq1				153	154				
Gnail					154				
Cysltrl									
Cxcr3				153	154				
Apln	30								
Rgs6	30		113						
P2ry12									
Gabbr1	30								
Rgs3(2)			113		154				
Dok4	30	39	113	153	154		152	153	154

Marker IDs 26, 27, 30, 39, 101, 113, 152, 153 and 154 correspond to markers D2Mit263, D2Mit51, D2Mit148, D3Mit44, D9Mit15, D11Mit227, D15Mit174, D15Mit136 and D15Mit5.

interest to identify the genomic markers that are associated with the expression levels of each gene, as well as homogeneous groups of the genes associated with a common subset of genomic markers. Two typical approaches are either examining each gene individually and separately as in the lasso, or treating each of the connected coexpression subnetworks as a group, as in the group lasso. In contrast to these two approaches, we perform a joint expression quantitative trait loci analysis of all seventeen genes while leveraging their specific dependency structures through the coexpression network.

Placing (11) in the framework of (1), we obtain $Y = X\beta + \epsilon$, where $\epsilon \sim N(0, \sigma^2 I)$, $Y = (Y_1^T, \dots, Y_G^T)^T$ with $G = 17$, $\beta = (\beta_1^T, \dots, \beta_G^T)^T$, and $X = \text{diag}(X_0, \dots, X_0)$ is a 1020×3298 block diagonal matrix. Then we apply the proposed method to perform simultaneous supervised learning and feature selection, with tuning performed through a grid search of $(\lambda_1, \lambda_2, \tau)$ using five-fold crossvalidation. The results are summarized in Table 6 and Fig. 1.

As indicated in Table 6, the proposed method identifies potential expression quantitative trait loci for thirteen genes, including two isolated genes Rps6ka4 and Dok4 selected by the lasso. Among these genes, the expression of genes Rps6ka4, 1200007D18Rik, Apln, Rgs6, Gabbr1 and Dok4, that of genes Rgs3, Kcnq1, Cxcr3 and Dok4, that of genes Calcr1, Rgs3, Kcnq1, Gnail, Cxcr3, Rgs3(2) and Dok4i, and that of genes Rgs6, Rgs3(s) and Dok4 are linked to marker D2Mit148 on chromosome 2, marker D15Mit136 on chromosome 15, marker D15Mit5 on chromosome 15, and marker D11Mit227 on chromosome 11, respectively. The discovered linkage to marker D2Mit148 is consistent with previous studies (Lan et al., 2006; Pan, 2009). Concerning supervised clustering, three genes Rgs3, Kcnq1 and Cxcr3 form one cluster, where the latter two are both connected to gene Rgs3; genes Rps6ka4, 1200007D18Tik, Rgs6, Apln, and Gabbr1 form another cluster, where the five genes are connected over the coexpression network; see Fig. 1. Evidently, grouping occurs mostly for direct neighbours over the network. In contrast, none of the genes are found by the group lasso to be associated with any markers. This may be due to the group selection property of the group lasso: a marker is associated with either all or none of

the genes in a group. Although a few genes share associations with a small number of common markers, as shown in the previous studies (Lan et al., 2006; Pan, 2009), the majority do not which presumably leads to the nonselection of any gene-marker association by the group lasso.

4. THEORY

This section derives a finite-sample probability error bound for reconstructing the oracle estimator by the proposed method. As a result, the proposed method not only reconstructs the true grouping $\mathcal{G}^0$, including zero and nonzero groups, but also offers optimal estimation performance. The oracle estimator, whose existence is assumed throughout the paper, is the least squares estimator based on $\mathcal{G}^0 = (\mathcal{G}_0^0, \dots, \mathcal{G}_{K^0}^0)$, that is, $\hat{\beta}^{\text{ols}} \equiv (\hat{\beta}_1^{\text{ols}}, \dots, \hat{\beta}_p^{\text{ols}})^T = (0_{|\mathcal{G}_0^0|}, \hat{\alpha}_1^{\text{ols}} 1_{|\mathcal{G}_1^0|}, \dots, \hat{\alpha}_{K^0}^{\text{ols}} 1_{|\mathcal{G}_{K^0}^0|})^T$, where $\hat{\alpha}^{\text{ols}} \equiv (\hat{\alpha}_1^{\text{ols}}, \dots, \hat{\alpha}_{K^0}^{\text{ols}})^T = (Z_{\mathcal{G}^0}^T Z_{\mathcal{G}^0})^{-1} Z_{\mathcal{G}^0}^T Y$ with $Z_{\mathcal{G}^0}^T Z_{\mathcal{G}^0}$ invertible, $Z_{\mathcal{G}^0} \equiv (z_{\mathcal{G}_1}, \dots, z_{\mathcal{G}_{K^0}})$ is an $n \times K^0$ matrix with $z_{\mathcal{G}_k} \equiv X_{\mathcal{G}_k} 1$, and $X_{\mathcal{G}_k}$ is the design matrix spanned by the predictors of $\mathcal{G}_k$ ($k = 1, \dots, K^0$).

Before proceeding, define the level of difficulty for simultaneous supervised learning and feature selection with regard to a given graph, describing the least favorable situation, as

$$C_{\min} \equiv \inf_{\mathcal{G}=(\mathcal{G}_0, \mathcal{G}_0^c) \in \mathcal{T}} \frac{\|(I - P_{\mathcal{G}_0^c})X\beta^0\|^2}{n \max(|\mathcal{G}_0 \setminus \mathcal{G}_0^0|, 1)},$$

where $P_{\mathcal{G}_0^c}$ is the $n \times n$ projection matrix onto design matrix $X_{\mathcal{G}_0^c}$ spanned over nonzero coefficients of grouping $\mathcal{G} = (\mathcal{G}_0, \mathcal{G}_0^c) = (\mathcal{G}_0, \dots, \mathcal{G}_K)$, the superscript c denotes the complement, and $\mathcal{T} = \{\mathcal{G} \neq \mathcal{G}^0 : \sum_{j=1}^p I(\beta_j \neq 0) \leq s_1^0, \sum_{j < j': (j, j') \in \mathcal{E}} I(\beta_j \neq \beta_{j'}) \leq s_2^0\}$ is a constrained set of (2) with $\tau = 0^+$, $s_1^0 = \sum_{j=1}^p I(\beta_j^0 \neq 0) = p - |\mathcal{G}_0^0|$, and $s_2^0 = \sum_{j < j': (j, j') \in \mathcal{E}} I(\beta_j^0 \neq \beta_{j'}^0)$. Also defined is the resolution level of the true regression coefficients,

$$\gamma_{\min} \equiv \min_{\{|\beta_j^0 - \beta_{j'}^0| > 0, |\beta_l^0| > 0: (j, j') \in \mathcal{E}, 1 \leq l \leq p\}} (|\beta_j^0 - \beta_{j'}^0|, |\beta_l^0|).$$

For $C_{\min}$ and $\gamma_{\min}$, small values correspond to a difficult situation. In the above definition, $C_{\min} = 0$ if models are not identifiable. Interestingly, $C_{\min}$ is related to $\gamma_{\min}$ through the minimal eigenvalue. It can be shown that $C_{\min} \geq c_{\min} \gamma_{\min}^2$, where $c_{\min} = \min_{\{B: |B| \leq 2(p - |\mathcal{G}_0^0|), \mathcal{G}_0^c \subseteq B\}} \Lambda_{\min}(X_B)$ and $\Lambda_{\min}(X_B)$ denotes the smallest eigenvalue of a nonnegative definite matrix of $n^{-1} X_B^T X_B$. Note that $C_{\min}$ captures two important aspects: dependency among the predictors as well as the resolution level $\gamma_{\min}$. Evidently, highly dependent predictors and weak true regression coefficients make the problem more challenging.

We now present a finite-sample error bound of a global minimizer $\hat{\beta}^{\text{gl}}$ of (2), based on which properties (A)–(C) are established with regard to selection consistency and parameter estimation in Theorem 2. Specifically, $\hat{\beta}^{\text{gl}}$ reconstructs $\hat{\beta}^{\text{ols}}$ consistently as long as $C_{\min}$ exceeds a certain thresholding value in (13), yielding consistent reconstruction of the true grouping in (A), and is asymptotic minimax optimal in parameter estimation in (B), where these results can be made uniform over an L_0 -band defined in (C).

Let $\lambda_{\max}(M)$ denote the maximum eigenvalue of matrix M .

THEOREM 2 (Global minimizer). *Assume that any elements of each true group are connected over $\mathcal{E}$ through a path. For $\hat{\beta}^{\text{gl}}$ with estimated grouping $\hat{\mathcal{G}}^{\text{gl}}$ through tuning: $(s_1, s_2) = (s_1^0, s_2^0)$*

and $\tau \leq 2\sigma [\log p / \{2np^3 \lambda_{\max}(X^T X)\}]^{1/2}$, we have

$$\text{pr}(\hat{\beta}^{\text{gl}} \neq \hat{\beta}^{\text{ols}}) \leq \exp \left\{ -\frac{n}{10\sigma^2} \left(C_{\min} - 10\sigma^2 \frac{2\log p + \bar{K} + 2\log S^*}{n} \right) \right\}, \quad (12)$$

where $\bar{K} = \max_{1 \leq i \leq s_1^0} K_i^* / i$ with $K_i^* \equiv \max_{\mathcal{G} \in \mathcal{T}, |\mathcal{G}_0 \setminus \mathcal{G}_0^0| = i} K(\mathcal{G})$ and with $K(\mathcal{G})$ being the number of groups for grouping $\mathcal{G}$, and $\log S^* = \max_{1 \leq i \leq s_1^0} \log T_i / i$ with $T_i = \max_{\{\mathcal{G}_0: |\mathcal{G}_0 \setminus \mathcal{G}_0^0| = i\}} |T_{\mathcal{G}_0}|$ and with $T_{\mathcal{G}_0} = \{\mathcal{G} = (\mathcal{G}_0^*, \mathcal{G}_1, \dots, \mathcal{G}_K) \in \mathcal{T} : \mathcal{G}_0^* = \mathcal{G}_0^0\}$. In addition, if

$$C_{\min} \geq d_1 \sigma^2 \frac{2\log p + \bar{K} + 2\log S^*}{n}, \text{ for any constant } d_1 > 10, \quad (13)$$

the following oracle properties hold for $\hat{\beta}^{\text{gl}}$ as $n, p \rightarrow \infty$,

- (A) $\hat{\beta}^{\text{gl}}$ reconstructs the oracle estimator $\hat{\beta}^{\text{ols}}$ with probability tending to one, that is, $\text{pr}(\hat{\beta}^{\text{gl}} \neq \hat{\beta}^{\text{ols}}) \rightarrow 0$, implying consistency for supervised clustering and feature selection, i.e., $\text{pr}(\hat{\mathcal{G}}^{\text{gl}} \neq \mathcal{G}^0) \rightarrow 0$.
- (B) The ideal estimation performance of the oracle estimator $\hat{\beta}^{\text{ols}}$ is recovered by $\hat{\beta}^{\text{gl}}$. That is, for any true β^0 , $n^{-1} E \|\hat{\beta}^{\text{gl}} - \beta^0\|_2^2 = (1 + o(1)) n^{-1} E \|\hat{\beta}^{\text{ols}} - \beta^0\|_2^2 = (1 + o(1)) K^0 \sigma^2 / n$, where $\|\cdot\|_2$ is the L_2 -norm in $\mathcal{R}^n$.
- (C) Let $B_0(u, l) \equiv \{\beta \in \mathcal{R}^p : K^0 \leq u, C_{\min}(\beta) \geq l\}$ be an L_0 -band formed by two concentric balls with upper radius $u > 0$ and lower radius $l > 0$. Then $\sup_{\beta^0 \in B_0(u, l)} \text{pr}(\hat{\beta}^{\text{gl}} \neq \hat{\beta}^{\text{ols}}) \rightarrow 0$, for $l = d_1 \sigma^2 (2\log p + \bar{K} + 2\log S^*) / n$, and $u < \min(n, p)$, implying that (A) and (B) hold uniformly over $\beta^0 \in B_0(u, l)$. Moreover, $\hat{\beta}^{\text{gl}}$ is asymptotically minimax over $B_0(u, l)$ in that $\sup_{\beta^0 \in B_0(u, l)} n^{-1} E \|\hat{\beta}^{\text{gl}} - \beta^0\|_2^2 = \inf_{\hat{\beta}} \sup_{\beta^0 \in B_0(u, l)} n^{-1} E \|\hat{\beta} - \beta^0\|_2^2 = \sigma^2 u / n$.

In Theorem 2, S^* describes the complexity of simultaneous supervised clustering and feature selection over a graph, where the prior information expressed in terms of a graph enters (12) through T_i and $\bar{K}$. In general, the graph information reduces the complexity. As shown in Corollary 1, S^* could be of an order that is substantially smaller than the Bell number that is of order $(s_1^0 / \log s_1^0)^{s_1^0}$; see Berend & Tassa (2010).

In order for (13) to hold, $s_1^0 \log s_1^0 / n \rightarrow 0$ as $p, n \rightarrow \infty$. In other words, the number of non-zero regression coefficients s_1^0 cannot be too large relative to the sample size n , depending on the graph information. This is analogous to the fact that the number of informative features cannot exceed the sample size in order for any feature selection method to work.

COROLLARY 1. If $\mathcal{E}$ is a fused graph, that is if $\mathcal{E} = \{(i, i+1) : i = 1, \dots, p-1\}$, then

$$|T_{\mathcal{G}_0}| \leq \sum_{j=1}^{K^0-1} \binom{|\mathcal{G}_0^c|}{j} \leq \sum_{j=q}^{K^0-1} \binom{s_1^0}{j} \leq (s_1^0)^{K^0}, \quad K_i^* \leq K^0 \quad (i = 1, \dots, K^0 - 1).$$

Hence $S^* \leq (s_1^0)^{K^0}$ and $\bar{K} \leq K^0$, and (12) reduces to

$$\text{pr}(\hat{\beta}^{\text{gl}} \neq \hat{\beta}^{\text{ol}}) \leq \exp \left\{ -\frac{n}{10\sigma^2} \left(C_{\min} - 20\sigma^2 \frac{\log p}{n} - 20K^0 \sigma^2 \frac{\log s_1^0}{n} \right) \right\}.$$

Next we present a parallel result for a local minimizer $\hat{\beta}^{lo}$ of (3) under an assumption stronger than (13). Given grouping $\mathcal{G}$, let $Z_{\mathcal{G}}$ be the collapsed design matrix, spanned by predictors of $\mathcal{G}_k$; $k = 1, \dots, K$ and $X_{\mathcal{G}^0}$ be the design matrix of nonzero predictors of the true grouping $\mathcal{G}^0$. Let K^* be the maximum number of distinct groups permitted, or $|\mathcal{G}| \leq K^*$.

THEOREM 3 (Local minimizer). *Assume that any elements of each true group are connected over $\mathcal{E}$ through a path. For any local solution $\hat{\beta}^{lo}$, and in particular the solution from Algorithm 1, if tuning satisfies $4\{\lambda_1^2(p - |\mathcal{G}_0^0|)^2 + \lambda_2^2|\mathcal{N}|^2\}^{1/2} \leq \tau \min_{|\mathcal{G}| \leq K^*} \Lambda_{\min}(Z_{\mathcal{G}})$, then $\text{pr}(\hat{\beta}^{lo} \neq \hat{\beta}^{ols})$ is bounded above by*

$$\begin{aligned} & (p - |\mathcal{G}_0^0|) \Phi \left\{ -\frac{n^{1/2}(\gamma_{\min} - 2\tau)}{\sigma \Lambda_{\min}^{-1/2}(X_{\mathcal{G}^0 \setminus \mathcal{G}_0^0})} \right\} + |\mathcal{G}_0^0| \Phi \left(-\frac{n^{\frac{\lambda_1}{\tau}}}{\sigma \max_{1 \leq j \leq p} \|x^{(j)}\|} \right) \\ & + \frac{K^0(K^0 - 1)}{2} \Phi \left\{ -\frac{n^{1/2}(\gamma_{\min} - 2\tau)}{2\sigma \Lambda_{\min}^{-1/2}(Z_{\mathcal{G}^0 \setminus \mathcal{G}_0^0})} \right\} + |\mathcal{N}| \Phi \left[\frac{-n^{\frac{\lambda_2}{\tau}}}{\sigma \max_{k, A: A \subseteq \mathcal{G}_k^0} \frac{2\{\text{Tr}(X_A^T X_A)\}^{1/2}}{|\mathcal{E} \cap \{A \times (\mathcal{G}_k^0 \setminus A)\}|}} \right], \quad (14) \end{aligned}$$

where $\Phi(\cdot)$ is the cumulative distribution of $N(0, 1)$, $\times$ denotes the Cartesian product, and $|\mathcal{N}|$ is the maximum number of direct neighbours of each node of $\mathcal{E}$. For the risk properties of (B) and (C) in Theorem 2, assume that for some constants $c_i > 0$ ($i = 1, 2$)

$$\frac{9(\lambda_1 + \lambda_2 p)^2 p K^*}{\min_{|\mathcal{G}| \leq K^*} \Lambda_{\min}(Z_{\mathcal{G} \setminus \mathcal{G}_0})} + \frac{9}{n} \|X\beta^0\|^2 \leq c_1 \exp\{c_2 \max(\log K^0, \log |\mathcal{N}|)\}.$$

If $\gamma_{\min} > 3\tau$,

$$\min \left[\Lambda_{\min}(Z_{\mathcal{G}^0 \setminus \mathcal{G}_0^0}) \gamma_{\min}^2, \frac{\min_{|\mathcal{G}| \leq K^*} \Lambda_{\min}^2(Z_{\mathcal{G}})}{128\{(p - |\mathcal{G}_0^0|)^2 + |\mathcal{N}|^2\} \min(\zeta_1, \zeta_2)} \right] \geq l = \frac{12\sigma^2 \log \max(|\mathcal{N}|, p)}{n},$$

with $\zeta_1 = \max_{k, A: A \subseteq \mathcal{G}_k^0} n^{-1} \text{tr}(X_A^T X_A) / |\mathcal{E} \cap \{A \times (\mathcal{G}_k^0 \setminus A)\}|^2$ and $\zeta_2 = n^{-1} \max_{1 \leq j \leq p} \|x^{(j)}\|$, there exist some $(\lambda_1, \lambda_2, \tau)$ such that properties (A)–(C) in Theorem 2 hold for $\hat{\beta}^{lo}$, where (C) and (14) hold uniformly over an L_0 -band,

$$\begin{aligned} B_0^*(u, l) = & \left\{ \beta \in \mathcal{R}^p : K^0 \leq u, \min \left[\Lambda_{\min}(Z_{\mathcal{G}^0 \setminus \mathcal{G}_0^0}) \gamma_{\min}^2(\beta), \right. \right. \\ & \left. \left. \times \frac{\min_{|\mathcal{G}| \leq K^*} \Lambda_{\min}^2(Z_{\mathcal{G}})}{128\{(p - |\mathcal{G}_0^0|)^2 + |\mathcal{N}|^2\} \min(\zeta_1, \zeta_2)} g \right] \geq l \right\}. \end{aligned}$$

Unlike Theorem 2, the graph information enters (14) through $\mathcal{E}$, where $|\mathcal{N}| = 2$ for a fused graph, and is bounded above by $\max_{0 \leq k \leq K^0} (|\mathcal{G}_k^0| / \log |\mathcal{G}_k^0|)^{|\mathcal{G}_k^0|}$.

ACKNOWLEDGMENT

This research was supported in part by grants from the U.S. National Science Foundation and the National Institutes of Health, and from the ROC National Science Council. The authors thank the editors and the reviewers for helpful comments and suggestions.

SUPPLEMENTARY MATERIAL

Supplementary material available at *Biometrika* online includes proofs of all theorems, corollaries, and lemmas.

REFERENCES

- AN, H. L. T. & TAO, P. D. (1997). Solving a class of linearly constrained indefinite quadratic problems by D.C. algorithms. *J. Global Optimiz.* **11**, 253–85.
- BEREND, D. & TASSA, T. (2010). Improved bounds on Bell numbers and on moments of sums of random variables. *Prob. Math. Statist.* **30**, 185–205.
- BONDELL, H. & REICH, B. J. (2008). Simultaneous regression shrinkage, variable selection, and supervised clustering of predictors with OSCAR. *Biometrics* **64**, 115–23.
- DICKER, L., HUANG, B. & LIN, X. (2012). Variable selection and estimation with the seamless- L_0 penalty. *Statist. Sinica*, to appear.
- FORTIN, M. & GLOWINSKI, R. (1983). *Augmented Lagrangian Methods: Applications to the Numerical Solution of Boundary Value Problems* (Studies in Mathematics and Its Application), vol. 15. Amsterdam: North-Holland.
- FRIEDMAN, J., HASTIE, T. J., HOFLING, H. & TIBSHIRANI, R. J. (2007). Pathwise coordinate optimization. *Ann. Appl. Statist.* **1**, 302–32.
- GHAZALPOUR, A., DOSS, S., ZHANG, B., WANG, S., PLAISIER, C., CASTELLANOS, R., BROZELL, A., SCHADT, E. E., DRAKE, T. A., LUSIS, A. J. & HORVATH, S. (2006). Integrating genetic and network analysis to characterize genes related to mouse weight. *PLoS Genet.* **2**, e130.
- JANG, W., LIM, J., LAZAR, N., LOH, J. M., McDOWELL, J. & YU, D. (2011). Regression shrinkage and equality selection for highly correlated predictors with HORSES. *Biometrics*, to appear.
- LAN, H., CHEN, M., FLOWERS, J. B., YANDELL, B. S., STAPLETON, D. S., MATA, C. M., MUI, E. T., FLOWERS, M. T., SCHUELER, K. L., MANLY, K. F., WILLIAMS, R. W., KENDZIORSKI, C. & ATTIE, A. D. (2006). Combined expression trait correlations and expression quantitative trait locus mapping. *PLoS Genet.* **2**, 51–61.
- LI, C. & LI, H. (2008). Network-constrained regularization and variable selection for analysis of genomic data. *Bioinformatics* **24**, 1175–82.
- PAN, W. (2009). Network-based multiple locus linkage analysis of expression traits. *Bioinformatics* **25**, 1390–6.
- PARK, M. Y., HASTIE, T. J. & TIBSHIRANI, R. J. (2007). Averaged gene expressions for regression. *Biostatistics* **8**, 212–27.
- RINALDO, A. (2009). Properties and refinements of the fused lasso. *Ann. Statist.* **37**, 2922–52.
- R DEVELOPMENT CORE TEAM. (2012). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0, <http://www.R-project.org>.
- ROKAFELLAR, R. T. & WETS, R. J. (2003). *Variational Analysis*. New York: Springer-Verlag.
- SHANNON, P., MARKIEL, A., OZIER, O., BALIGA, N. S., WANG, J. T., RAMAGE, D., AMIN, N., SCHWIKOWSKI, B. & IDEKER, T. (2003). Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13**, 2498–504.
- SHE, Y. (2010). Sparse regression with exact clustering. *Elect. J. Statist.* **4**, 1055–96.
- SHEN, X. & HUANG, H. (2010). Grouping pursuit through a regularization solution surface. *J. Am. Statist. Assoc.* **105**, 727–39.
- SHEN, X., PAN, W. & ZHU, Y. (2012). Likelihood-based selection and sharp parameter estimation. *J. Am. Statist. Assoc.* **107**, 223–32.
- TIBSHIRANI, R. J., SAUNDERS, M., ROSSET, S., ZHU, J. & KNIGHT, K. (2005). Sparsity and smoothness via the fused lasso. *J. R. Statist. Soc. B* **67**, 91–108.
- TIBSHIRANI, R. J. & TAYLOR, J. (2011). The solution path of the generalized lasso. *Ann. Statist.* **39**, 1335–71.
- TSENG, P. (2001). Coordinate ascent for maximizing nondifferentiable concave functions. *J. Optimiz. Theory Appl.* **109**, 474–94.
- ZOU, H. & HASTIE, T. J. (2005). Regularization and variable selection via the elastic net. *J. R. Statist. Assoc. B* **67**, 301–20.

[Received June 2011. Revised May 2012]