Biometrika (2009), **96**, 2, *pp*. 339–355
doi: 10.1093/biomet/asp020

Printed in Great Britain

A group bridge approach for variable selection

BY JIAN HUANG

Department of Statistics and Actuarial Science, University of Iowa, 221 Schaeffer Hall, Iowa City, Iowa 52242, U.S.A.
jian-huang@uiowa.edu

SHUANGE MA

Division of Biostatistics, Department of Epidemiology and Public Health, Yale University, New Haven, Connecticut 06520, U.S.A.
shuangge.ma@yale.edu

HUILIANG XIE

Department of Management Science, University of Miami, Coral Gables, Florida 33124, U.S.A.
hxie@exchange.sba.miami.edu

AND CUN-HUI ZHANG

Department of Statistics, Rutgers University, Piscataway, New Jersey 08854, U.S.A.
cunhui@stat.rutgers.edu

SUMMARY

In multiple regression problems when covariates can be naturally grouped, it is important to carry out feature selection at the group and within-group individual variable levels simultaneously. The existing methods, including the lasso and group lasso, are designed for either variable selection or group selection, but not for both. We propose a group bridge approach that is capable of simultaneous selection at both the group and within-group individual variable levels. The proposed approach is a penalized regularization method that uses a specially designed group bridge penalty. It has the oracle group selection property, in that it can correctly select important groups with probability converging to one. In contrast, the group lasso and group least angle regression methods in general do not possess such an oracle property in group selection. Simulation studies indicate that the group bridge has superior performance in group and individual variable selection relative to several existing methods.

Some key words: Bridge estimator; Iterative lasso; Penalized regression; Two-level selection; Variable-selection consistency.

1. INTRODUCTION

Consider the linear regression model

$$y_i = x_{i1}\beta_1 + \cdots + x_{id}\beta_d + \varepsilon_i \quad (i = 1, \ldots, n), \tag{1}$$

where y_i is the response variable, $x_{i1}, \ldots, x_{id}$ are covariate variables, β_js are regression coefficients and ε_is are error terms. It is assumed that the covariates can be naturally grouped. We are

interested in simultaneously selecting important groups as well as important individual variables within the selected groups.

Traditional approaches to variable selection include C_p (Mallows, 1973), AIC (Akaike, 1973), and BIC (Schwarz, 1978). Recent methods based on regularization include the bridge (Frank & Friedman, 1993), lasso (Tibshirani, 1996), smoothly clipped absolute deviation (Fan & Li, 2001; Fan & Peng, 2004), and elastic net (Zou & Hastie, 2005), among others. These methods are designed for selecting individual variables.

The need to select groups of variables arises in multifactor analysis of variance and nonparametric additive regression. In analysis of variance, a factor with multiple levels can be represented by a group of dummy variables. In nonparametric additive regression, each component can be expressed as a linear combination of a set of basis functions. In both the cases, the selection of important factors or nonparametric components amounts to the selection of groups of variables. Several recent papers have considered selecting important groups of variables using penalized methods. Yuan & Lin (2006) proposed the group lasso, group least angle regression and group nonnegative garrote methods. The group lasso is a natural extension of the lasso, in which an L_2 norm of the coefficients associated with a group of variables is used as a component of the penalty function. The group lasso method was extended to general loss functions by Kim et al. (2006). They used the same penalty as the group lasso penalty and called the extension blockwise sparse regression. P. Zhao, G. Rocha and B. Yu, in an unpublished technical report for the University of California at Berkley, proposed a composite absolute penalty for group selection, which can be regarded as a generalization of the group lasso. These studies only considered group selection, but did not address the question of individual selection within groups. Ma & Huang (2007) proposed a clustering threshold gradient descent regularization method that selects variables at both the group and individual variable levels. However, the method does not optimize a well-defined objective function, and it is therefore difficult to study its theoretical properties.

The existing penalized methods are not capable of simultaneous group and individual variable selection, even though this is desirable in many problems, such as the impact study, studied in § 4·2, that was designed to determine the effects of different risk factors on body mass index of high school students in two Seattle public schools. It is of interest to know which groups have a significant impact on the body mass index as well as the variables in these groups that are important. For example, if food consumption has a significant effect, it is of great interest to know which food types have significant impact.

2. THE GROUP BRIDGE ESTIMATOR

2·1. *Definition and the estimator*

Let $x_k = (x_{1k}, \ldots, x_{nk})'$ $(k = 1, \ldots, d)$ be the design vectors and let $y = (y_1, \ldots, y_n)'$ be the response vector in (1), so that the linear model is written as

$$y = x_1\beta_1 + \cdots + x_d\beta_d + \varepsilon,$$

with an error vector $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)'$. Let $A_1, \ldots, A_J$ be subsets of $\{1, \ldots, d\}$ representing known groupings of the design vectors and denote the regression coefficients in the jth group by $\beta_{A_j} = (\beta_k, k \in A_j)'$. For any $m \times 1$ vector a, denote its L_1 norm by $\|a\|_1 = |a_1| + \cdots + |a_m|$. We consider the objective function

$$L_n(\beta) = \left\| y - \sum_{k=1}^{d} x_k\beta_k \right\|_2^2 + \lambda_n \sum_{j=1}^{J} c_j \|\beta_{A_j}\|_1^{\gamma}, \tag{2}$$

where $\lambda_n > 0$ is the penalty level and c_j are constants for the adjustment of the different dimensions of β_{A_j}. A simple choice is $c_j \propto |A_j|^{1-\gamma}$, where $|A_j|$ is the cardinality of A_j. In (2), the bridge penalty is applied to the L_1 norms of the grouped coefficients. Therefore, we call the $\hat{\beta}_n$ that minimizes (2) a group bridge estimator. Here the groups A_j can overlap and their union is allowed to be a proper subset of the whole so that variables not in $\cup_{j=1}^{J} A_j$ are not penalized. Since the grouping structure is incorporated into model fitting through the penalty function, overlapping is only allowed in the penalty part of the objective function (2). For example, if a variable belongs to two groups, the coefficient associated with this variable appears twice in the penalty. It is clear that the objective function (2) with such a penalty is well defined. When $|A_j| = 1$ $(j = 1, \ldots, J)$, (2) simplifies to the standard bridge criterion. When $\gamma = 1$, (2) is the lasso criterion, which can only do individual variable selection. However, when $0 < \gamma < 1$, the group bridge criterion (2) can be used for variable selection at the group and individual variable levels simultaneously. This will be further explained in Proposition 1 and the remarks following it.

2·2. *Computation*

Direct minimization of $L_n(\beta)$ is difficult, since the group bridge penalty is not a convex function for $0 < \gamma < 1$. We formulate an equivalent minimization problem that is easier to solve computationally. For $0 < \gamma < 1$, define

$$S_{1n}(\beta, \theta) = \left\| y - \sum_{k=1}^{d} x_k \beta_k \right\|_2^2 + \sum_{j=1}^{J} \theta_j^{1-1/\gamma} c_j^{1/\gamma} \|\beta_{A_j}\|_1 + \tau_n \sum_{j=1}^{J} \theta_j, \tag{3}$$

where τ_n is a penalty parameter.

PROPOSITION 1. *Suppose that* $0 < \gamma < 1$. *If*

$$\lambda_n = \tau_n^{1-\gamma} \gamma^{-\gamma} (1-\gamma)^{\gamma-1}, \tag{4}$$

then $\hat{\beta}_n$ *minimizes* $L_n(\beta)$ *if and only if* $(\hat{\beta}_n, \hat{\theta})$ *minimizes* $S_{1n}(\beta, \theta)$ *subject to* $\theta \geqslant 0$, *for some* $\hat{\theta} \geqslant 0$, *where* $\theta \geqslant 0$ *means* $\theta_j \geqslant 0$ $(j = 1, \ldots, J)$.

This proposition is similar to the characterization of the component selection and smoothing method of Lin & Zhang (2006). Examining the form of S_{1n} in (3), we see that the minimization of S_{1n} with respect to (β, θ) yields sparse solutions at the group and individual variable levels. To be specific, the penalty is an adaptively weighted L_1 penalty and, therefore, the solution is sparse in β. On the other hand, for $0 < \gamma < 1$, small θ_j will force $\beta_{A_j} = 0$, which leads to group selection.

Based on Proposition 1, we propose the following iterative algorithm. Obtain an initial estimate $\beta^{(0)}$ and, for $s = 1, 2, \ldots$, carry out Steps 1 and 2 until convergence.

Step 1. Compute

$$\theta_j^{(s)} = c_j \left(\frac{1-\gamma}{\tau_n \gamma} \right)^{\gamma} \left\| \beta_{A_j}^{(s-1)} \right\|_1^{\gamma} \quad (j = 1, \ldots, J). \tag{5}$$

Step 2. Compute

$$\beta^{(s)} = \arg\min_{\beta} \left\| y - \sum_{k=1}^{d} x_k \beta_k \right\|_2^2 + \sum_{j=1}^{J} \left(\theta_j^{(s)} \right)^{1-1/\gamma} c_j^{1/\gamma} \|\beta_{A_j}\|_1. \tag{6}$$

This algorithm always converges, since at each step it decreases the nonnegative objective function (3). The main computational task is Step 2, which is a lasso problem and can be solved efficiently using the least angle regression algorithm (Efron et al., 2004). In general, this algorithm converges to a local minimizer depending on the initial value $\beta^{(0)}$, since the group bridge penalty is not convex. In this paper, we focus on full-rank designs, where the unbiased least squares estimator is a natural initial estimator.

2·3. *Tuning parameter selection*

For a given bridge index γ, there is a one-to-one correspondence between λ_n and τ_n as given in (4). For simplicity, we describe tuning parameter selection in terms of λ_n. For a fixed λ_n, let $\hat{\beta}_n = \hat{\beta}_n(\lambda_n)$ be the group bridge estimator of β. Let $\hat{\theta}_{nj}$ $(j = 1, \ldots, J)$ be the jth component of $\hat{\theta}_n = \hat{\theta}\{\hat{\beta}_n(\lambda_n)\}$, as defined in (5), and let $X = (x_1, \ldots, x_d)$ be an $n \times d$ covariate matrix. The Karush–Kuhn–Tucker condition for (6) implies that

$$2(y - X\hat{\beta}_n)' x_k = \sum_{j: A_j \ni k} \hat{\theta}_{nj}^{1-1/\gamma} c_j^{1/\gamma} \operatorname{sgn}(\hat{\beta}_{nk}) \quad (\hat{\beta}_{nk} \neq 0). \tag{7}$$

Since $\operatorname{sgn}(\beta_{nk}) = \beta_{nk}/|\beta_{nk}|$, this allows us to write the fitted response vector as

$$\hat{y} = X\hat{\beta}_n = X_{\lambda_n}(X'_{\lambda_n} X_{\lambda_n} + 0{\cdot}5 W_{\lambda_n})^{-1} X'_{\lambda_n} y,$$

where X_{λ_n} is the submatrix of X whose columns correspond to the covariates with nonzero estimated coefficients for the given λ_n, and W_{λ_n} is the diagonal matrix with diagonal elements

$$\sum_{A_j \ni k} \hat{\theta}_{nj}^{1-1/\gamma} c_j^{1/\gamma} / |\hat{\beta}_{nk}| \quad (\hat{\beta}_{nk} \neq 0).$$

Therefore, the number of effective parameters with a given λ_n can be approximated by

$$d(\lambda_n) = \operatorname{tr}\{X_{\lambda_n}(X'_{\lambda_n} X_{\lambda_n} + 0{\cdot}5 W_{\lambda_n})^{-1} X'_{\lambda_n}\}.$$

This procedure is not only similar to that described by Fu (1998) but also resembles the tuning parameter selection method of Tibshirani (1996) and Zhang & Lu (2007).

An AIC-type criterion for choosing λ_n is

$$\text{AIC}(\lambda_n) = \log\{\|y - X\hat{\beta}_n(\lambda_n)\|_2^2 / n\} + 2d(\lambda_n)/n.$$

A generalized crossvalidation score (Wahba, 1990) is defined as

$$\text{GCV}(\lambda_n) = \frac{\|y - X\hat{\beta}_n(\lambda_n)\|_2^2}{n\{1 - d(\lambda_n)/n\}^2}.$$

It can be seen that these two criteria are close to each other when $d(\lambda_n)$ is relatively small compared to n.

Although GCV and AIC are reasonable criteria for tuning, they tend to select more variables than the true model contains, and we therefore also consider a BIC-type criterion

$$\text{BIC}(\lambda_n) = \log\{\|y - X\hat{\beta}_n(\lambda_n)\|_2^2 / n\} + \log(n) d(\lambda_n)/n.$$

The tuning parameter λ_n is selected as the minimizer of $\text{AIC}(\lambda_n)$, $\text{GCV}(\lambda_n)$ or $\text{BIC}(\lambda_n)$. In general, AIC-type criteria are better suited if the purpose of the analysis is to minimize the difference between the true distribution of y and the estimate from a candidate model, and the BIC-type criterion should be used if the purpose is to uncover the model structure, but none of these criteria can achieve both goals (Yang, 2005).

2·4. *Variance estimation*

The covariance matrix of $\hat{\beta}_n(\lambda_n)$ is estimated in a similar way to Tibshirani's (1996) estimator of the covariance matrix of the lasso estimator. Let $B_1 = B_1(\lambda_n) = \{k : \hat{\beta}_{nk} \neq 0\}$ be the set of selected variables and let $\hat{\beta}_{nB_1}(\lambda_n) = \{\hat{\beta}_{nk}(\lambda_n) : k \in B_1\}$ be the nonzero components of $\hat{\beta}_n(\lambda_n)$ given λ_n. By (7),

$$\hat{\beta}_{nB_1}(\lambda_n) = \left(X'_{\lambda_n} X_{\lambda_n} + 0{\cdot}5 W_{\lambda_n}\right)^{-1} X'_{\lambda_n} y,$$

so that the covariance matrix of $\hat{\beta}_{nB_1}(\lambda_n)$ can be approximated by

$$\left(X'_{\lambda_n} X_{\lambda_n} + 0{\cdot}5 W_{\lambda_n}\right)^{-1} X'_{\lambda_n} X_{\lambda_n} \left(X'_{\lambda_n} X_{\lambda_n} + 0{\cdot}5 W_{\lambda_n}\right)^{-1} \hat{\sigma}^2, \tag{8}$$

where $\hat{\sigma}^2 = \| y - X\hat{\beta}_n(\lambda_n)\|_2^2 / \{n - d(\lambda_n)\}$.

2·5. *Comparison with existing group selection methods*

Yuan & Lin (2006) proposed the group lasso, group least angle regression and group garrote methods for group selection. These methods do not select individual variables within groups. To illustrate this point, we look at the group lasso in more detail. The group lasso estimator is defined as

$$\tilde{\beta}_n = \arg\min_{\beta} \left\| y - \sum_{k=1}^{d} x_k \beta_k \right\|_2^2 + \lambda_n \sum_{j=1}^{J} \left\| \beta_{A_j} \right\|_{K_j,2}, \tag{9}$$

where K_j is a positive definite matrix and $\|\beta_{A_j}\|_{K_j,2} = (\beta'_{A_j} K_j \beta_{A_j})^{1/2}$. A typical choice of K_j suggested by Yuan & Lin (2006) is $K_j = |A_j| I_j$, where I_j is the $|A_j| \times |A_j|$ identity matrix.

Let τ_n be a penalty parameter and define

$$S_{2n}(\beta, \theta) = \left\| y - \sum_{k=1}^{d} x_k \beta_k \right\|_2^2 + \sum_{j=1}^{J} \theta_j^{-1} \left\| \beta_{A_j} \right\|_{K_j,2}^2 + \tau_n \sum_{j=1}^{J} \theta_j. \tag{10}$$

PROPOSITION 2. *Let $\tau_n = 2^{-2}\lambda_n^2$. Then $\tilde{\beta}_n$ satisfies* (9) *if and only if $(\tilde{\beta}_n, \tilde{\theta})$ minimizes $S_{2n}(\beta, \theta)$ subject to $\theta \geqslant 0$, for some $\tilde{\theta} \geqslant 0$.*

According to this proposition, the group lasso behaves like an adaptively weighted ridge regression, in which the sum of the squared coefficients in group j is penalized by θ_j, and the sum of the θ_js is in turn penalized by τ_n. Therefore, in minimizing (10), we obtain either $\beta_{A_j} = 0$, in which case the corresponding group is dropped from the model, or $\beta_{A_j} \neq 0$, in which case all the elements of β_{A_j} are nonzero and all the variables in group j are retained in the model. The group lasso therefore selects groups of variables, but it does not select individual variables within groups.

3. ASYMPTOTIC PROPERTIES

In this section, we study the asymptotic properties of the group bridge estimators. We show that, for $0 < \gamma < 1$, the group bridge estimators correctly select groups with nonzero coefficients with probability converging to one under reasonable conditions. We also derive the asymptotic distribution of the estimators of the nonzero coefficients.

Without loss of generality, suppose that

$$\beta_{A_j} \neq 0 \quad (j = 1, \ldots, J_1), \qquad \beta_{A_j} = 0 \quad (j = J_1 + 1, \ldots, J).$$

Let $B_2 = \cup_{j=J_1+1}^{J} A_j$ be the union of the groups with zero coefficients, let $B_1 = B_2^c$ and let $\beta_{B_j} = (\beta_k, k \in B_j)'$ $(j = 1, 2)$. Assume without loss of generality that the index k is arranged so that $\beta = (\beta'_{B_1}, \beta'_{B_2})'$. Let β_0 be the true value of β. Since $\beta_{0B_2} = 0$, the true model is fully explained by the first J_1 groups. In this notation, $\hat{\beta}_{nB_1}$ and $\hat{\beta}_{nB_2}$ are respectively, the estimates of β_{0B_1} and β_{0B_2} from the group bridge estimator $\hat{\beta}_n$. Set $X = (x_1, \ldots, x_d)$ and $X_1 = (x_k, k \in B_1)$. Define

$$\Sigma_n = n^{-1}X'X, \quad \Sigma_{1n} = n^{-1}X_1'X_1.$$

Let ρ_n and ρ_n^* be the smallest and largest eigenvalues of Σ_n, respectively. We make the following assumptions.

Assumption 1. The errors $\varepsilon_1, \ldots, \varepsilon_n$ are independent with zero-mean and finite variance σ^2.

Assumption 2. The maximum multiplicity $C_n^* = \max_k \sum_{j=1}^{J} I\{k \in A_j\}$ is bounded and

$$\frac{\lambda_n^2}{n\rho_n} \sum_{j=1}^{J_1} c_j^2 \|\beta_{0A_j}\|_1^{2\gamma-2} |A_j| \leqslant \sigma^2 d M_n, \quad M_n = O(1).$$

Assumption 3. The constants c_j are scaled so that $\min_{j \leqslant J} c_j \geqslant 1$ and

$$\frac{\lambda_n \rho_n^{1-\gamma/2}}{d^{1-\gamma/2}\rho_n^* n^{\gamma/2}} \to \infty.$$

Assumption 1 is standard in linear regression. Assumptions 2 and 3 both require full-rank design with $\rho_n > 0$, which is equivalent to $\operatorname{rank}(X) = d \leqslant n$. Still, we allow the number of covariates $d = d_n$ to grow at a certain rate such that $n > d_n \to \infty$. For fixed unknown $\{B_1, \beta_{0B_1}, J_1\}$, Assumptions 2 and 3 are consequences of

$$\frac{1}{\rho_n} + \rho_n^* + \sum_{j=1}^{J_1} c_j^2 = O(1), \quad \frac{\lambda_n}{n^{1/2}} \to \lambda_0 < \infty, \quad \frac{\lambda_n d^{\gamma/2}}{d n^{\gamma/2}} \to \infty, \tag{11}$$

provided that $c_j \geqslant 1$ and $C_n^* = O(1)$. This allows $d_n = o(1)n^{(1-\gamma)/(2-\gamma)}$. It is clear from (11) that Assumptions 2 and 3 put restrictions on the magnitude of the penalty parameter. In particular, they exclude the case $\gamma \geqslant 1$.

THEOREM 1. *Suppose that* $0 < \gamma < 1$ *and that Assumptions* 1–3 *hold.*

(i) *It holds that* $\hat{\beta}_{nB_2} = 0$ *with probability converging to* 1.

(ii) *Suppose that* $\{B_1, \beta_{0B_1}, J_1\}$ *are fixed unknowns, that* (11) *holds and that* $\Sigma_{1n} \to \Sigma_1$ *and* $n^{-1/2}X_1'\varepsilon \to Z \sim N(0, \sigma^2\Sigma_1)$ *in distribution. Then, in distribution,*

$$\sqrt{n}(\hat{\beta}_{nB_1} - \beta_{0B_1}) \to \arg\min\{V_1(u) : u \in R^{|B_1|}\},$$

where

$$V_1(u) = -2u'Z + u'\Sigma_1 u + \gamma\lambda_0 \sum_{j=1}^{J_1} c_j \|\beta_{0A_j}\|_1^{\gamma-1} \sum_{k \in A_j \cap B_1} \{u_k \operatorname{sgn}(\beta_{0k}) I(\beta_{0k} \neq 0) + |u_k| I(\beta_{0k} = 0)\}.$$

In particular, when $\lambda_0 = 0$, *in distribution,* $\sqrt{n}(\hat{\beta}_{nB_1} - \beta_{0B_1}) \to \Sigma_1^{-1}Z \sim N(0, \sigma^2\Sigma_1^{-1})$.

Part (i) of Theorem 1 is of particular interest. It states that the group bridge estimators of the coefficients of the zero groups are exactly equal to zero with probability converging to one. This, together with part (ii), implies that the group bridge estimator is able to distinguish correctly nonzero groups from zero groups eventually. Therefore, the group bridge estimator has the oracle property in group selection. Part (ii) shows that the estimator of nonzero coefficients is $n^{1/2}$-consistent and in general converges to the argmin of the Gaussian process V_1. When $\lambda_0 > 0$, the limiting distribution puts positive probability at 0.

In the theorem, we require that $0 < \gamma < 1$. The case $\gamma = 1$ is a discontinuity point in the sense that it is a boundary point at which the penalty is convex. If $\gamma = 1$, the penalty reduces to the standard lasso penalty, which does not take the group structure into account.

The proof of Theorem 1 is given in the Appendix. Since Theorem 1 is valid in the case of $A_j = \{j\}$ $(j = 1, \ldots, d)$, it generalizes the result of Huang et al. (2008), who showed selection consistency and the asymptotic distribution for the bridge estimator of Frank & Friedman (1993). In this case, there is no need to select within groups and $\lambda_n/\sqrt{n} \to 0$ seems appropriate.

For independent and identically distributed errors, the assumption $n^{-1/2}X_1'\varepsilon \to Z \sim N(0, \sigma^2\Sigma_1)$, in distribution, follows from the Lindeberg–Feller central limit theorem under $\Sigma_{1n} \to \Sigma_1$ (van der Vaart, 1998, pp. 20–1). To compare the different asymptotic properties of the group bridge and group lasso estimators, we present the following theorem for the group lasso estimator of Yuan & Lin (2006).

THEOREM 2. *Suppose that* $\{\beta, d, A_j, c_j, K_j, j = 1, \ldots, J\}$ *are all fixed as* $n \to \infty$ *and that the* ε_i*s are independent and identically distributed with* $E\varepsilon_i = 0$ *and* $\mathrm{var}(\varepsilon_i) = \sigma^2 \in (0, \infty)$. *Suppose further that the* $d \times d$ *matrices* Σ_n *converge to a positive-definite matrix* Σ *and that* $\lambda_n n^{-1/2} \to \lambda_0 < \infty$. *Then*

$$\sqrt{n}(\tilde{\beta}_n - \beta_0) \to \arg\min\{V(u) : u \in R^d\},$$

in distribution where, for some $Z \sim N(0, \sigma^2\Sigma)$,

$$V(u) = -2u'Z + u'\Sigma u + \lambda_0 \sum_{j=1}^{J} c_j \left\{ \frac{u'_{A_j} K_j \beta_{0A_j}}{\|\beta_{0A_j}\|_{K_j,2}} I(\beta_{A_j} \neq 0) + \|u_{A_j}\|_{K_j,2} I(\beta_{0A_j} = 0) \right\}.$$

By Theorem 2, when $\lambda_0 = 0$, the group lasso estimator has the same asymptotic distribution as the least squares estimator. Therefore, it is required that $\lambda_0 > 0$ for the group lasso to be able to carry out group selection. When $\lambda_0 > 0$, the asymptotic distribution of $\tilde{\beta}_n$ puts positive probability at 0 when $\beta_{A_j} = 0$. However, in general, this positive probability is less than one. Thus, the group lasso is, in general, not consistent in selecting the nonzero groups.

The proof of Theorem 2 is similar to that of part (ii) of Theorem 1 and is therefore omitted. When $|A_j| = 1$ $(j = 1, \ldots, J)$, this theorem simplifies to the result of Knight & Fu (2000) about the lasso.

4. NUMERICAL STUDIES

4·1. *Simulation study*

We use simulation to evaluate the finite-sample performance of the group bridge estimator. We also compare the group bridge with the group lasso, group least angle regression, group garrote and smoothly clipped absolute deviation methods in the simulation. The group lasso, group least angle regression and group garrote estimators are computed using the algorithms in

Yuan & Lin (2006). The smoothly clipped absolute deviation estimator is computed using the majorize-minimize algorithm described by Hunter & Li (2005).

We consider two scenarios. For the generating models in Example 1, the number of groups is moderately large, the group sizes are equal and relatively large, and within each group the coefficients are either all nonzero or all zero. In Example 2, the group sizes vary and there are coefficients equal to zero in a nonzero group. We use $\gamma = 0{\cdot}5$ in the group bridge estimator. The sample size $n = 200$ in each example.

Example 1. In this experiment, there are five groups and each group consists of eight covariates. The covariate vector is $X' = (X_1', \ldots, X_5')'$ and, for any j in $\{1, \ldots, 5\}$, the subvector of covariates that belong to the same group is $X_j' = (x_{8(j-1)+1}, \ldots, x_{8(j-1)+8})$. To generate the covariates $x_1, \ldots, x_{40}$, we first simulate 40 random variables $R_1, \ldots, R_{40}$ independently from the standard normal distribution. Then Z_j $(j = 1, \ldots, 5)$ are simulated with a normal distribution and an AR(1) structure such that $\mathrm{cov}(Z_{j_1}, Z_{j_2}) = 0{\cdot}4^{|j_1 - j_2|}$, for $j_1, j_2 = 1, \ldots, 5$. The covariates $x_1, \ldots, x_{40}$ are generated as

$$x_j = (Z_{g_j} + R_j)/\sqrt{2} \quad (j = 1, \ldots, 40),$$

where g_j is the smallest integer greater than $(j-1)/8$ and the x_js with the same value of g_j belong to the same group. The random error is $\varepsilon \sim N(0, 2^2)$. The response variable Y is generated from $Y = \sum_{j=1}^{40} x_j \beta_j + \varepsilon$, where

$$(\beta_1, \ldots, \beta_8) = (0{\cdot}5, 1, 1{\cdot}5, 2, 2{\cdot}5, 3, 3{\cdot}5, 4),\ (\beta_9, \ldots, \beta_{16}) = (2, 2, \ldots, 2),$$
$$(\beta_{17}, \ldots, \beta_{24}) = (\beta_{25}, \ldots, \beta_{32}) = (\beta_{33}, \ldots, \beta_{40}) = (0, 0, \ldots, 0).$$

Thus, the coefficients in each group are either all nonzero or all zero.

Example 2. In this experiment, the group size differs across groups. There are six groups made up of three groups each of size 10 and three groups each of size 4. The covariate vector is $X' = (X_1', \ldots, X_6')$, where the six subvectors of covariates are $X_j' = (x_{10(j-1)+1}, \ldots, x_{10(j-1)+10})$, for $j = 1, 2, 3$, and $X_j' = (x_{4(j-4)+31}, \ldots, x_{4(j-4)+34})$, for $j = 4, 5, 6$. To generate the covariates $x_1, \ldots, x_{42}$, we first simulate Z_i $(i = 1, \ldots, 6)$ and $R_1, \ldots, R_{42}$ independently from the standard normal distribution. For $j = 1, \ldots, 30$, let g_j be the largest integer less than $j/10 + 1$ and, for $j = 31, \ldots, 42$, let g_j be the largest integer less than $(j - 30)/4 + 1$. The covariates $(x_1, \ldots, x_{42})$ are obtained as

$$x_j = (Z_{g_j} + R_j)/\sqrt{2} \quad (j = 1, \ldots, 42).$$

The response vector is generated from $Y = \sum_{j=1}^{42} x_j \beta_j + \varepsilon$, where

$$(\beta_1, \ldots, \beta_{10}) = (0{\cdot}5, -2, 0{\cdot}5, 2, -1, 1, 2, -1{\cdot}5, 2, -2),$$
$$(\beta_{11}, \ldots, \beta_{20}) = (-1{\cdot}5, 2, 1, -2, 1{\cdot}5, 0, \ldots, 0),$$
$$(\beta_{21}, \ldots, \beta_{30}) = (0, \ldots, 0),\ (\beta_{31}, \ldots, \beta_{34}) = (2, -2, 1, 1{\cdot}5),$$
$$(\beta_{35}, \ldots, \beta_{38}) = (-1{\cdot}5, 1{\cdot}5, 0, 0),\ (\beta_{39}, \ldots, \beta_{42}) = (0, \ldots, 0),$$

and $\varepsilon \sim N(0, 2^2)$. Thus the coefficients in a group can be either all zero, all nonzero or partly zero.

Table 1 summarizes the simulation results based on 400 replications for these two examples. For the group bridge estimators, we considered AIC, BIC and generalized crossvalidation, GCV, for determining the penalty parameter. The variable selection and coefficient estimation results

Table 1. *Simulation study. Number of groups selected, number of variables selected, model error and percentage of occasions on which correct groups are selected, averaged over* 400 *replications with estimated standard errors in parentheses, by various estimators, for* (*a*) *Example* 1 *and* (*b*) *Example* 2

Method (tuning)	No. of groups	No. of variables	Model error	% Corr. sel.
		(a) Example 1		
Group lasso (C_p)	3·91 (0·94)	31·24 (7·52)	0·58 (0·19)	8·50 (1·39)
Group lasso (BIC)	2·32 (0·50)	18·56 (4·03)	0·65 (0·23)	69·75 (2·23)
Group Lars (C_p)	3·54 (0·72)	28·28 (5·77)	0·59 (0·25)	9·75 (1·48)
Group Lars (BIC)	2·30 (0·47)	18·38 (3·79)	0·64 (0·22)	71·00 (1·48)
Group garrote (C_p)	3·56 (0·63)	28·46 (5·05)	0·51 (0·22)	6·50 (1·23)
Group garrote (BIC)	2·12 (0·33)	16·98 (2·63)	0·41 (0·16)	87·75 (1·64)
SCAD (GCV)	4·31 (0·83)	20·87 (3·96)	0·55 (0·25)	1·00 (0·50)
SCAD (BIC)	3·32 (0·72)	17·34 (1·65)	0·51 (0·21)	8·00 (1·36)
Group bridge (AIC)	4·00 (0·82)	24·34 (4·25)	0·55 (0·20)	5·25 (1·12)
Group bridge (BIC)	2·07 (0·31)	16·15 (0·79)	0·47 (0·17)	94·75 (1·12)
True model	2	16	0	100
		(b) Example 2		
Group lasso (C_p)	5·76 (0·48)	40·68 (2·94)	0·88 (0·25)	2·00 (0·70)
Group lasso (BIC)	4·75 (0·65)	33·14 (5·02)	1·09 (0·34)	37·25 (2·42)
Group Lars (C_p)	5·36 (0·59)	37·89 (4·42)	1·00 (0·32)	5·75 (1·16)
Group Lars (BIC)	4·49 (0·51)	31·03 (3·77)	1·04 (0·34)	51·75 (2·45)
Group garrote (C_p)	4·93 (0·52)	34·61 (4·47)	0·83 (0·30)	17·25 (1·89)
Group garrote (BIC)	4·26 (0·44)	29·23 (2·29)	0·73 (0·23)	73·75 (2·20)
SCAD (GCV)	5·65 (0·54)	26·75 (3·38)	0·75 (0·27)	3·00 (0·85)
SCAD (BIC)	5·28 (0·64)	23·18 (1·81)	0·76 (0·28)	10·00 (1·50)
Group bridge (AIC)	5·16 (0·72)	31·06 (3·58)	0·75 (0·24)	19·00 (1·96)
Group bridge (BIC)	4·14 (0·38)	24·78 (1·67)	0·74 (0·26)	87·25 (1·67)
True model	4	23	0	100

AIC, Akaike information criterion; BIC, Bayesian information criterion; GCV, generalized crossvalidation; SCAD, smoothly clipped absolute deviation method; Lars, least angle regression; No. of groups, number of groups selected; No. of variables, the average number of variables selected; % Corr. sel., the percentage of occasions on which the model produced contains exactly the same groups as the underlying model.

based on GCV are similar to those using AIC and are thus omitted. For the group lasso and group least angle regression, we considered C_p, AIC and BIC. The results based on AIC are similar to those based on C_p. As the C_p method was suggested by Yuan & Li (2006), we included the C_p and BIC results in Table 1. As for the smoothly clipped absolute deviation method, GCV and BIC are used in tuning. The former was recommended by Fan & Li (2001).

In Table 1, the model error is computed as $(\hat{\beta} - \beta)' E(XX')(\hat{\beta} - \beta)$, where β is the generating value. Enclosed in parentheses are the corresponding standard deviations. The last line in each panel in the table gives the true values used in the generating model. For example, in Example 1, there are two nonzero groups and 16 nonzero coefficients in the generating model.

A comparison of different tuning parameter selection methods indicates that, for the methods considered, tuning based on BIC in general does better than based on AIC, C_p or GCV in terms of selection at the group and individual variable levels. We therefore focus on the comparisons of the methods with BIC tuning below.

In terms of the number of groups selected, the number of variables selected and the percentage of correct models selected, the group bridge considerably outperforms the group lasso, group least angle regression and group garrote, which tend to select more groups and variables than

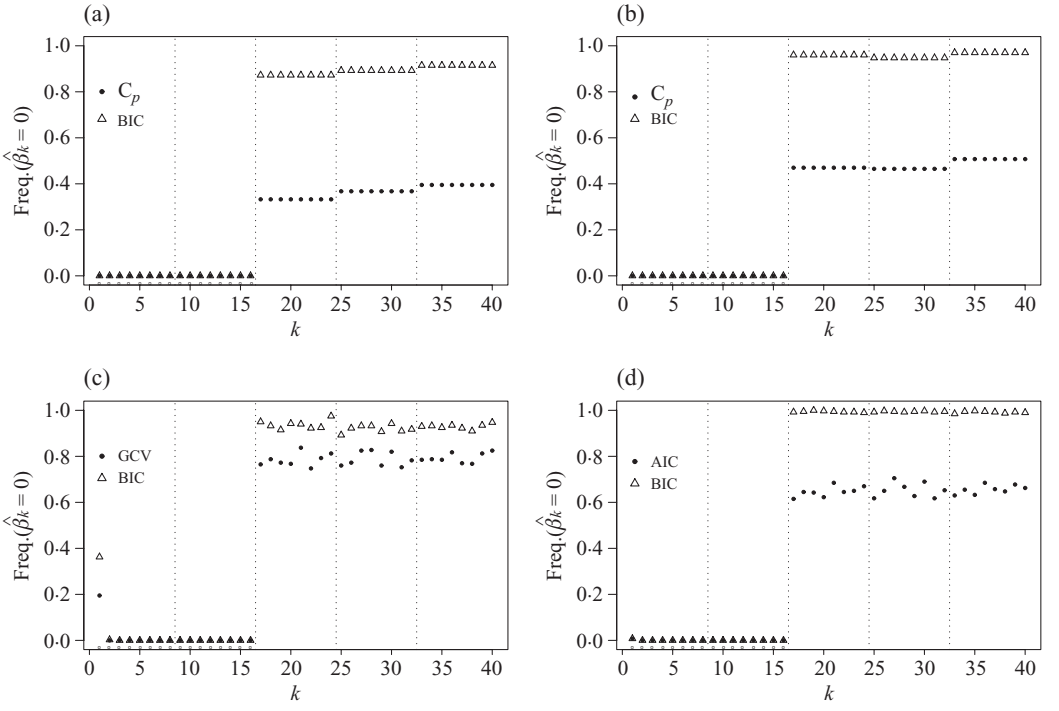

Fig. 1. Example 1. Relative frequency of each component of the group lasso, group garrote, SCAD and group bridge estimates being equal to 0, for (a) group lasso, (b) group garrote, (c) SCAD and (d) group bridge. In each panel, the small circles along the horizontal axis indicate nonzero coefficients. Panels show results from BIC tuning parameter selection (triangles), C_p for the group lasso and group garrote (solid dots), smoothly clipped absolute deviation method (GCV and solid dots) and AIC for the group bridge (AIC and solid dots).

there actually are in the generating models. This agrees with the simulation results reported by Yuan & Lin (2006). In comparison, the numbers of groups and variables in the models selected by the group bridge are close to the generating values. The group bridge incurs smaller model error than the group lasso and group least angle regression, but incurs slightly higher model error than the group garrote. The group bridge outperforms the smoothly clipped absolute deviation method in terms of both selection and model errors. This is not surprising since the smoothly clipped absolute deviation method is designed for individual variable selection and does not take into account the group structure information.

To examine the selection results for each covariate, we plot the percentage of the 400 replications for which the coefficient is estimated exactly at zero, i.e. the associated covariate is not selected. The results are shown in Figs. 1 and 2. We did not include the plots for the group least angle regression, since its performance falls between that of the group lasso and the group garrote. In Fig. 1(a), there are four circles at 1, 2, 3 and 4, indicating that the first four coefficients in Example 1 are nonzero. Figs. 1 and 2 show that the group bridge estimators tend to have higher percentages of correctly identifying zero coefficients than the group lasso and group garrote. In both examples, the group bridge outperforms the smoothly clipped absolute deviation method in correctly identifying zero groups. The smoothly clipped absolute deviation method does better than the group bridge in detecting the zero coefficients in a nonzero group, but it sometimes falsely identifies a nonzero coefficient as zero, as can be seen in the first group in each example. The model errors from the group bridge method are slightly smaller than those from the smoothly clipped absolute deviation method, which may be attributed to the higher false positive rate of

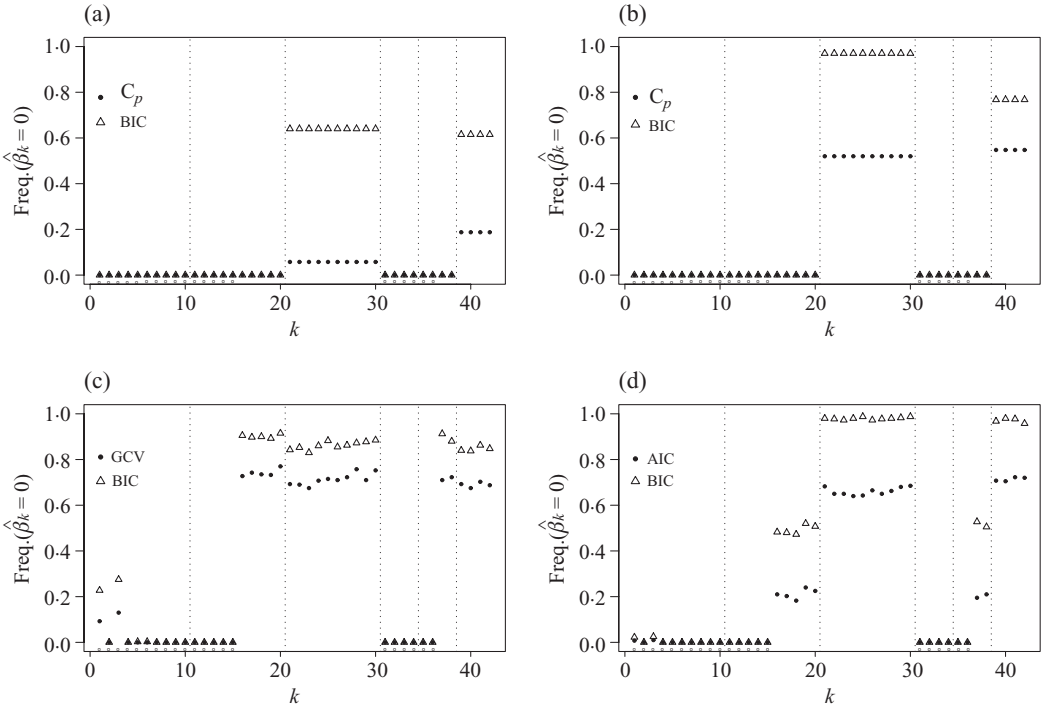

Fig. 2. Example 2. Relative frequency of each component of the group lasso, group garrote, SCAD and group bridge estimates being equal to 0, for (a) group lasso, (b) group garrote, (c) SCAD and (d) group bridge. See Fig. 1 for further details.

the smoothly clipped absolute deviation method with regard to X_1 in Example 1 and X_1 and X_3 in Example 2.

In summary, the group bridge method does better than the existing methods in terms of group and individual variable selection and is competitive when evaluated in terms of model error.

We also looked at the performance of the proposed standard error estimation method. In general, when the BIC is used in tuning parameter selection, the proposed method tends to slightly underestimate true sampling variabilities, but otherwise appears to provide reasonable standard error estimates. The slight underestimation is perhaps due to the effect of choosing tuning parameters, which is not accounted for in (8).

4·2. *Impact study*

The Impact study was part of a three-year project designed to measure the impact of nutritional policies and environmental change on obesity in the high school students enrolled in Seattle Public Schools. The study description can be found at http://depts.washington.edu/uwcphn/activities/projects/noncommercialism.html.

The study was led by the University of Washington Center for Public Health Nutrition and was conducted in two urban high schools. One primary goal of this study is to determine the effects of different risk factors on body mass index. Table 2 provides the definitions of the variables included in the study. Indicators are created for the ethnicity variable. Natural clusters exist for the risk factors. The 25 covariates can be naturally classified into eight different categories, measuring different aspects such as food sources and demographics. The response variable is the logarithm of the body mass index. We focus on the 799 subjects with complete records.

Table 2. *Impact study. Dictionary of covariates*

Group	Variable	Type	Definition
Age	V1	C	Age
	V2	C	Age^2
Gender	V3	B	Female gender
Ethnicity	V4	B	Ethnic (American Indian/Alaska native)
	V5	B	Ethnic (Hispanic/Latino)
	V6	B	Ethnic (Asian)
	V7	B	Ethnic (Native Hawaiian/Pacific Islander)
	V8	B	Ethnic (White)
	V9	B	Ethnic (Do not know)
	V10	B	No answer
	V11	B	Bi/multi-racial
	V12	B	Speaking other language
Food source	V13	B	Breakfast/lunch from cafeteria more than 3 times per week
	V14	B	Food from à la carte more than 3 times per week
	V15	B	Fast food
	V16	B	Food from home more than 3 times per week
Consumption	V17	C	Fizzy drinks
(Unhealthy)	V18	B	Sweets
	V19	B	Crisps
	V20	B	Cake
	V21	B	Ice cream
Consumption	V22	C	Milk
(Healthy)	V23	C	Fruit and vegetable
School	V24	B	School A or B
Physical activity	V25	C	Mild physical activity
	V26	C	Hard physical activity

Type, type of variable; C, continuous; B, binary.

We analyzed the data from the Impact study, using the group bridge method as well as the ordinary least squares, group lasso, group least angle regression, group nonnegative garrote and smoothly clipped absolute deviation methods. The results are given in Table 3. To save space, the following columns are omitted: ordinary least squares, the group lasso and group least angle regression with BIC, which select none of the groups, and the smoothly clipped absolute deviation method with BIC, which selects V7, Hawaiian, and V15, Fastfood. For the group lasso and group garrote, when C_p is used in tuning parameter selection, all the variable groups are included in the final model. The group bridge does not select the group of consumption of healthy food when AIC or GCV is used. The group bridge with BIC leads to the same group selection as the group garrote with BIC and omits three more groups: age, gender and unhealthy food consumption. However, the group bridge gives rise to a sparser model than the group garrote. We conclude from the group bridge estimate that demographics, food source, unhealthy food consumption, school group and physical activity may be associated with body mass index in the Impact cohort. In the ethnicity group, the group bridge using BIC only selects Hawaiian and Asian.

For evaluation, we first randomly select a training set of size 600. The test set is composed of the remaining 199 records. We compute estimates using the training set only, and then compute the prediction mean squared errors for the testing set. The splitting, estimation and prediction are repeated 200 times. The average number of groups selected, average number of variables selected and median of the relative mean squared prediction errors, relative to ordinary least squares, for each method are listed in Table 4. When AIC and C_p are used, the group bridge selects

Table 3. *Impact study. Estimates* ($\times 10^{-2}$) *from different approaches. The standard errors of the group bridge estimates are enclosed in parentheses*

Variable	Group lasso (C_p)	Group Lars (C_p)	Group garrote (C_p)	Group garrote (BIC)	SCAD (GCV)	Group bridge (AIC)	Group bridge (BIC)
Age	9·1	2·3	13·2	0	16·1	0 (–)	0 (–)
Age^2	−0·3	−0·1	−0·5	0	−0·6	−0·0 (0·0)	0 (–)
Gender	1·3	0	2·57	0	0·4	1·0 (0·8)	0 (–)
Native	−2·0	−1·1	−3·2	−0·7	0	0 (–)	0 (–)
Hispanic	4·0	2·2	5·0	1·1	1·8	2·3 (1·8)	0 (–)
Asian	−1·9	−1·3	−3·2	−0·7	−3·4	−3·5 (1·7)	−2.7 (2.2)
Hawaiian	8·4	5·1	11·4	−0·7	9·8	9·1 (2·5)	5.8 (2.9)
White	−1·2	−0·9	−2·2	2·6	0	0 (–)	0 (–)
Unknown	1·8	0·7	2·8	−0·5	0	0 (–)	0 (–)
NoAns	2·1	1·3	3·0	0·7	0	0 (–)	0 (–)
Bi-rac.	−1·6	−0·9	−2·4	−0·6	−0·7	−1·4 (1·2)	0 (–)
Otherlang.	−3·0	−1·8	−4·3	−1·0	−1·5	−1·9 (1·1)	0 (–)
Breaklunch	1·3	1·0	1·1	0·7	0·9	1·5 (1·0)	1.2 (0.8)
Lacarte	0·8	0·5	1·1	0·7	0	0·2 (0·4)	0 (–)
Fastfood	−3·7	−3·2	−4·6	−2·9	−4·6	−4·4 (1·3)	−4.8 (1.2)
Foodhome	−1·4	−0·9	−1·7	−1·1	−0·6	−1·0 (0·9)	−0.9 (0.7)
Fizzydrinks	−0·1	−0·1	−0·2	0	0	0 (–)	0 (–)
Sweets	0·1	0·1	0	0	0	0 (–)	0 (–)
Crisps	−1·5	−0·7	−2·3	0	−1·4	−1·5 (0·9)	0 (–)
Cake	−2·4	−1·2	−3·8	0	−2·7	−2·7 (1·2)	0 (–)
Ice cream	0·5	0·3	0·6	0	0	0 (–)	0 (–)
Milk	0·2	0	0·4	0	0	0 (–)	0 (–)
Fruit	0·1	0	0·1	0	0	0 (–)	0 (–)
School	−2·3	−2·0	−2·7	−2·0	−2·2	−2·7 (1·2)	−2.9 (1.2)
Mildact	−0·1	−0·1	−0·1	0	0	0 (–)	0 (–)
Hardact	0·5	0·3	0·7	0·2	0·4	0·5 (0·2)	0.4 (0.2)

SCAD, smoothly clipped absolute deviation; Lars, least angle regression·

Table 4. *Impact study. Training-testing results*

	OLS	Group lasso C_p	Group lasso BIC	Group Lars C_p	Group Lars BIC	Group garrote C_p	Group garrote BIC	SCAD GCV	SCAD BIC	Group bridge AIC	Group bridge BIC
Avg. no. of groups	8·00	6·38	0·03	5·86	0·01	7·47	2·51	6·76	0·01	6·09	2·68
Avg. no. of variables	25·98	22·89	0·06	21·61	0·03	24·94	8·29	14·03	0·03	14·62	4·50
Median of relative PE	1·00	0·99	1·01	0·99	1·01	1·00	1·01	0·99	1·01	0·99	1·00

SCAD, smoothly clipped absolute deviation method; Lars, least angle regression; PE, prediction error; OLS, ordinary least squares.

fewer groups and has smaller prediction errors. When BIC is used, all the approaches, except the group bridge and the group nonnegative garrote, rarely select any group and yield the null model, but their prediction errors are comparable with those produced by ordinary least squares. This suggests that the variation of this cohort's body mass index is not very well captured by the variables measured in the study and that other variables such as genetic factors may be of greater importance in explaining the variation of body mass index. In fact, with the full model using the least squares, the R^2 value is only 8%. Similar results have been observed in a previous study of factors that may affect body mass index (Storey et al., 2003). We must exercise caution in interpreting the analysis results of this dataset. This is an observational study, and most of

the participating students are of African American origin, so that the results here should not be extrapolated to the general population of high school students.

5. DISCUSSION

The proposed group bridge approach can be applied to other regression problems when both group and individual variable selections are desired; for example, in the context of the general M-estimation, we can use

$$\sum_{i=1}^{n} m\left(y_i, \beta_0 + \sum_{k=1}^{d} x_{ik}\beta_k\right) + \lambda_n \sum_{j=1}^{J} \|\beta_{A_j}\|_1^{\gamma},$$

where m is a given loss function. This formulation includes generalized linear models, censored regression models, including Cox regression and robust regression. For example, for generalized linear models such as logistic regression, we take m to be the negative loglikelihood function. For Cox regression, we take the empirical loss function to be the negative partial likelihood. For loss functions other than least squares, further work is needed to study the computational algorithms and theoretical properties of the group bridge estimators.

A more general view can be adopted regarding the formulation of penalties. The group bridge penalty is a combination of two penalties, the bridge penalty for group selection and the lasso for within-group selection. In general, it is possible to consider combinations of different penalties. For example, we can use the smoothly clipped absolute deviation penalty for within-group selection and the bridge penalty for group selection. Different penalty functions may be preferable under different data and model settings.

Finally, we have only considered the asymptotic properties of the group bridge estimators when the number of covariates is smaller than the sample size. The need for two-level selection also arises when the number of covariates is larger than the sample size. For example, in regression analysis of a clinical outcome, such as disease status or survival, with high-dimensional genomic data, it is natural to consider genes in the same pathway as a group; typically, there are only limited numbers of pathways and genes that will be important to a clinical outcome.

ACKNOWLEDGEMENT

The authors thank the referees, the associate editor and Professor D. M. Titterington for their helpful comments. Huang is also a member of the Department of Biostatistics of the University of Iowa. The research of Huang and Ma is supported by grants from the U.S. National Institutes of Health and National Science Foundation. The research of Zhang is partially supported by grants from the U.S. National Science Foundation and National Security Agency.

APPENDIX

Proofs

Proof of Proposition 1. We have $\min_{\beta,\theta} S_{1n}(\beta, \theta) = \min_{\beta} \hat{S}_{1n}(\beta)$, where $\hat{S}_{1n}(\beta) = \min_{\theta}\{S_{1n}(\beta, \theta) : \theta \geqslant 0\}$. For any β,

$$\hat{\theta}(\beta) \equiv \arg\min\{S_{1n}(\beta, \theta) : \theta \geqslant 0\} = \arg\min\left\{\sum_{j=1}^{J} \theta_j^{1-1/\gamma} c_j^{1/\gamma} \|\beta_{A_j}\|_1 + \tau_n \sum_{j=1}^{J} \theta_j, \theta \geqslant 0\right\}.$$

Therefore, $\hat{\theta}(\beta) = (\hat{\theta}_1(\beta), \ldots, \hat{\theta}_d(\beta))'$ must satisfy

$$(1/\gamma - 1)\theta_j^{-1/\gamma}(\beta)c_j^{1/\gamma}\|\beta_{A_j}\|_1 = \tau_n \quad (j = 1, \ldots, J).$$

Write $\hat{S}_{1n}(\beta) = S_{1n}\{\beta, \hat{\theta}(\beta)\}$ and substitute the expressions

$$\theta_j(\beta) = \left(\frac{1-\gamma}{\gamma}\right)^{\gamma} c_j|\beta_{A_j}\|_1^{\gamma}\tau_n^{-\gamma}, \quad \theta_j^{1-1/\gamma}(\beta) = \left(\frac{\gamma}{1-\gamma}\right)^{1-\gamma} \frac{c_j^{1-1/\gamma}\tau_n^{1-\gamma}}{\|\beta_{A_j}\|_1^{1-\gamma}}$$

into $S_{1n}\{\beta, \hat{\theta}(\beta)\}$. Then we obtain, after some algebra, $\hat{S}_{1n}(\beta) = \| y - X\beta\|_2^2 + \lambda_n \sum_{j=1}^{J} c_j \| \beta_{A_j}\|_1^{\gamma}$. Here we used $\lambda_n = \tau_n^{1-\gamma}\{(1/\gamma - 1)^{\gamma} + (1/\gamma - 1)^{\gamma-1}\}$, so that $\hat{S}_{1n}(\beta) = L_n(\beta)$. □

Theorem 1 is proved by establishing the following results in three steps: Lemma A1 establishes estimation consistency and rate of convergence; Lemma A2 establishes variable-selection consistency, part (i) of Theorem 1, and obtains the asymptotic distribution, part (ii) of Theorem 1.

LEMMA A1. *Suppose that Assumptions* 1 *and* 2 *hold with* $0 < \gamma \leqslant 1$. *Then*

$$E\left(\|\hat{\beta}_n - \beta_0\|_2^2\right) \leqslant \frac{\sigma^2 d}{n\rho_n}(8 + 16C_n^* M_n),$$

where ρ_n *is the smallest eigenvalue of* $\Sigma = X'X/n$.

The proof of this lemma can be found in an accompanying technical report by the authors, available at http://www.stat.uiowa.edu/techrep/tr376.pdf.

LEMMA A2. *Suppose that Assumptions* 1–3 *hold with* $0 < \gamma < 1$. *Then*

$$\mathrm{pr}(\hat{\beta}_{nA_j} = 0 \text{ for all } j > J_1) \to 1.$$

Proof. Let $B_2 = \cup_{j=J_1+1}^{J} A_j$ and define $\tilde{\beta}_n = (\tilde{\beta}_{n1}, \ldots, \tilde{\beta}_{nd})'$ by

$$\tilde{\beta}_{nk} = \begin{cases} \hat{\beta}_{nk} & (k \notin B_2), \\ 0 & (k \in B_2). \end{cases}$$

Since $\hat{\theta}_{nj}^{1-1/\gamma} c_j^{1/\gamma} \| \hat{\beta}_{A_j}\|_1 = \gamma\lambda_n c_j \| \hat{\beta}_{A_j}\|_1^{\gamma}$ by (5), (7) implies that

$$2(Y - \mathcal{X}\hat{\beta}_n)'X_k = \gamma\lambda_n \sum_{A_j \ni k} c_j \|\hat{\beta}_{nA_j}\|_1^{\gamma-1}\mathrm{sgn}(\hat{\beta}_{nk}) \quad (\hat{\beta}_{nk} \neq 0).$$

Since $(\hat{\beta}_{nk} - \tilde{\beta}_{nk})\mathrm{sgn}(\hat{\beta}_{nk}) = |\hat{\beta}_{nk}|I\{k \in B_2\}$, we have

$$\begin{aligned} 2(Y - \mathcal{X}\hat{\beta}_n)'\mathcal{X}(\hat{\beta} - \tilde{\beta}_n) &= \sum_{k \in B_2} |\hat{\beta}_{nk}|\gamma\lambda_n \sum_{A_j \ni k} c_j \|\hat{\beta}_{nA_j}\|_1^{\gamma-1} \\ &= \gamma\lambda_n \sum_{j=1}^{J} c_j \|\hat{\beta}_{A_j}\|_1^{\gamma-1}(\|\hat{\beta}_{nA_j}\|_1 - \|\tilde{\beta}_{nA_j}\|_1). \end{aligned}$$

Since $\gamma b^{\gamma-1}(b - a) \leqslant b^{\gamma} - a^{\gamma}$ for $0 \leqslant a \leqslant b$, for $j \leqslant J_1$ we have

$$\gamma\|\hat{\beta}_{nA_j}\|_1^{\gamma-1}(\|\hat{\beta}_{nA_j}\|_1 - \|\tilde{\beta}_{nA_j}\|_1) \leqslant \|\hat{\beta}_{nA_j}\|_1^{\gamma} - \|\tilde{\beta}_{nA_j}\|_1^{\gamma}.$$

Since $\| \tilde{\beta}_{nA_j}\|_1 = 0$ for $j > J_1$, this implies that

$$2|(Y - \mathcal{X}\hat{\beta}_n)'\mathcal{X}(\hat{\beta} - \tilde{\beta}_n)| \leqslant \lambda_n \sum_{j=1}^{J_1} c_j\left(\|\hat{\beta}_{nA_j}\|_1^{\gamma} - \|\tilde{\beta}_{nA_j}\|_1^{\gamma}\right) + \gamma\lambda_n \sum_{j=J_1+1}^{J} c_j \|\hat{\beta}_{nA_j}\|_1^{\gamma}. \tag{A1}$$

By the definition of $\hat{\beta}_n$, we have

$$\|Y - \mathcal{X}\hat{\beta}_n\|_2^2 + \lambda_n \sum_{j=1}^{J} c_j \|\hat{\beta}_{nA_j}\|_1^\gamma \leqslant \|Y - \mathcal{X}\tilde{\beta}_n\|_2^2 + \lambda_n \sum_{j=1}^{J} c_j \|\tilde{\beta}_{nA_j}\|_1^\gamma.$$

Since $\|\tilde{\beta}_{nA_j}\|_1 = 0$ for $j > J_1$, by (A1),

$$\begin{aligned}
&2|(Y - \mathcal{X}\hat{\beta}_n)'\mathcal{X}(\hat{\beta} - \tilde{\beta}_n)| + (1-\gamma)\lambda_n \sum_{j=J_1+1}^{J} c_j \|\hat{\beta}_{nA_j}\|_1^\gamma \\
&\quad\leqslant \lambda_n \sum_{j=1}^{J} c_j \|\hat{\beta}_{nA_j}\|_1^\gamma - \lambda_n \sum_{j=1}^{J} c_j \|\tilde{\beta}_{nA_j}\|_1^\gamma \\
&\quad\leqslant \|Y - \mathcal{X}\tilde{\beta}_n\|_2^2 - \|Y - \mathcal{X}\hat{\beta}_n\|_2^2 \\
&\quad= \|\mathcal{X}(\hat{\beta}_n - \tilde{\beta}_n)\|_2^2 + 2(Y - \mathcal{X}\hat{\beta}_n)'\mathcal{X}(\hat{\beta}_n - \tilde{\beta}_n).
\end{aligned}$$

Thus, since $n\rho_n^*$ is the largest eigenvalue of $\mathcal{X}'\mathcal{X}$ and $\hat{\beta}_{nk} - \tilde{\beta}_{nk} = \hat{\beta}_{nk} I(k \in B_2)$,

$$(1-\gamma)\lambda_n \sum_{j=J_1+1}^{J} c_j \|\hat{\beta}_{nA_j}\|_1^\gamma \leqslant \|\mathcal{X}(\hat{\beta}_n - \tilde{\beta}_n)\|_2^2 = n\rho_n^* \|\hat{\beta}_{nB_2}\|_2^2 \leqslant n\rho_n^* \|\hat{\beta}_n - \beta_0\|_2^2,$$

which implies, by Lemma A1 and since $C_n^* M_n = O(1)$ in Assumption 2, that

$$(1-\gamma)\lambda_n \sum_{j=J_1+1}^{J} c_j \|\hat{\beta}_{nA_j}\|_1^\gamma \leqslant n\rho_n^* \|\hat{\beta}_{nB_2}\|_2^2 \leqslant O_P(\sigma^2 d\rho_n^*/\rho_n). \tag{A2}$$

We still need to find a lower bound of $\sum_{j=J_1+1}^{J} c_j \|\hat{\beta}_{nA_j}\|_1^\gamma$. Since $c_j \geqslant 1$ by Assumption 3,

$$\sum_{j=J_1+1}^{J} c_j \|\hat{\beta}_{nA_j}\|_1^\gamma \geqslant \left(\sum_{j=J_1+1}^{J} \|\hat{\beta}_{nA_j}\|_1 \right)^\gamma \geqslant \|\hat{\beta}_{nB_2}\|_1^\gamma \geqslant \|\hat{\beta}_{nB_2}\|_2^\gamma. \tag{A3}$$

If $\|\hat{\beta}_{nB_2}\|_2 > 0$, the combination of (A2) and (A3) yields

$$(1-\gamma)\lambda_n \leqslant n\rho_n^* \|\hat{\beta}_{nB_2}\|_2^{2-\gamma} \leqslant O_P(1) n\rho_n^* \{\sigma^2 d/(n\rho_n)\}^{1-\gamma/2}.$$

Since $\lambda_n(\rho_n/d)^{1-\gamma/2}/(\rho_n^* n^{\gamma/2}) \to \infty$ by Assumption 3, this implies that

$$\mathrm{pr}\big(\|\hat{\beta}_{nB_2}\|_2 > 0\big) \leqslant \mathrm{pr}\left\{ \frac{\lambda_n(\rho_n/d)^{1-\gamma/2}}{\rho_n^* n^{\gamma/2}} \leqslant O_P(1) \right\} \to 0. \qquad \square$$

Proof of Theorem 1. Part (i) follows from Lemma A2. Part (ii) can be proved based on Lemma A1 and the argmin continuous mapping theorem of Kim & Pollard (1990). The details can be found in the authors' technical report. $\square$

REFERENCES

AKAIKE, H. (1973). Information theory and an extension of the maximum likelihood principle. In *Second International Symposium on Information Theory*, Ed. B. N. Petrov and F. Csaki, pp. 267–81. Budapest: Akademiai Kiado.

EFRON, B., HASTIE, T., JOHNSTONE, I. & TIBSHIRANI, R. (2004). Least angle regression (with Discussion). *Ann. Statist.* **32**, 407–99.

FAN, J. & LI, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Am. Statist. Assoc.* **96**, 1348–60.

FAN, J. & PENG, H. (2004). Nonconcave penalized likelihood with a diverging number of parameters. *Ann. Statist.* **32**, 928–61.

FRANK, I. E. & FRIEDMAN, J. H. (1993). A statistical view of some chemometrics regression tools (with Discussion). *Technometrics* **35**, 109–48.
FU, W. J. (1998). Penalized regressions: the bridge versus the Lasso. *J. Comp. Graph. Statist.* **7**, 397–16.
HUANG, J., HOROWITZ, J. L. & MA, S. G. (2008). Asymptotic properties of bridge estimators in sparse high-dimensional regression models. *Ann. Statist.* **36**, 587–13.
HUNTER, D. R. & LI, R. (2005). Variable selection using MM algorithms. *Ann. Statist.* **33**, 1617–42.
KIM, Y., KIM, J. & KIM, Y. (2006). The blockwise sparse regression. *Statist. Sinica* **16**, 375–90.
KIM, J. & POLLARD, D. (1990). Cube root asymptotics. *Ann. Statist.* **18**, 191–219.
KNIGHT, K. & FU, W. J. (2000). Asymptotics for Lasso-type estimators. *Ann. Statist.* **28**, 1356–78.
LIN, Y. & ZHANG, H. (2006). Component selection and smoothing in multivariate nonparametric regression. *Ann. Statist.* **34**, 2272–97.
MA, S. & HUANG, J. (2007). Clustering threshold gradient descent regularization: with application to survival analysis using microarray data. *Bioinformatics* **23**, 466–72.
MALLOWS, C. (1973). Some comments on C_p. *Technometrics* **15**, 661–75.
SCHWARZ, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6**, 461–4.
STOREY, M. L., FORSHEE, R. A., WEAVER, A. R. & SANSALONE, W. R. (2003). Demographic and lifestyle factors associated with body mass index among children and adolescents. *Int. J. Food Sci. Nutr.* **54**, 491–503.
TIBSHIRANI, R. (1996). Regression shrinkage and selection via the Lasso. *J. R. Statist. Soc.* B **58**, 267–88.
VAN DER VAART, A. W. (1998). *Asymptotic Statistics*. Cambridge: Cambridge University Press.
WAHBA, G. (1990). *Spline Models for Observational Data.* CBMS-NSF Regional Conference Series in Applied Mathematics. Philadelphia, PA: Society for Industrial and Applied Mathematics.
YANG, Y. (2005). Can the strengths of AIC and BIC be shared? *Biometrika* **92**, 937–50.
YUAN, M. & LIN, Y. (2006). Model selection and estimation in regression with grouped variables. *J. R. Statist. Soc.* B **68**, 49–67.
ZHANG, H. H. & LU, W. B. (2007). Adaptive Lasso for Cox's proportional hazards model. *Biometrika* **94**, 691–703.
ZOU, H. & HASTIE, T. (2005). Regularization and variable selection via the elastic net. *J. R. Statist. Soc.* B **67**, 301–20.

[*Received May* 2007. *Revised October* 2008]