

Bridge regression: adaptivity and group selection

CHEOLWOO PARK AND YOUNG JOO YOON

Department of Statistics, University of Georgia, Athens, GA 30602, USA

January 10, 2011

Abstract

In high dimensional regression problems regularization methods have been a popular choice to address variable selection and multicollinearity. In this paper we study bridge regression that adaptively selects the penalty order from data and produces flexible solutions in various settings. We implement bridge regression based on the local linear and quadratic approximations to circumvent the nonconvex optimization problem. Our numerical study shows that the proposed bridge estimators are a robust choice in various circumstances compared to other penalized regression methods such as the ridge, lasso, and elastic net. In addition, we propose group bridge estimators that select grouped variables and study their asymptotic properties when the number of covariates increases along with the sample size. These estimators are also applied to varying-coefficient models. Numerical examples show superior performances of the proposed group bridge estimators in comparisons with other existing methods.

Key words: Analysis of Variance, Bridge regression, Multicollinearity, Oracle property, Penalized regression, Variable selection, Varying-coefficient models.

1 Introduction

The high dimensional nature of many current data sets has given regularization methods enormous attention in the statistical community. Examples include microarray gene expression data analysis in biology, cloud detection through analysis of satellite images, classification of spam emails, and many others. High dimensionality could lead us to models that are very complex which poses challenges in prediction and interpretation. In such cases, structural information within the data can be incorporated into the model estimation procedure to significantly reduce the actual complexity

involved in the estimation procedure. Regularization methods provide a powerful yet versatile technique for doing so. Great progress has been made in the last decade, but further improvements are still possible.

We consider the linear regression model with p predictors and n observations:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{1.1}$$

where $\mathbf{Y} = (Y_1, \dots, Y_n)^\top$, $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)^\top$, $\boldsymbol{\epsilon} = (\epsilon_1, \dots, \epsilon_n)^\top$, and $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_p)$. Here the ϵ_i 's are independently and identically distributed as normal with mean 0 and variance σ^2 . Assume that the Y_i 's are centered and the covariates \mathbf{x}_i 's are standardized. In estimating the regression coefficients $\boldsymbol{\beta}$, the ordinary least squares (OLS) estimator, the most common method, is unbiased. However, it may still have a large mean squared error when the multicollinearity in the design matrix \mathbf{X} causes unstable solutions.

Penalized regression methods, such as the ridge (Hoerl and Kennard, 1970), lasso (Tibshirani, 1996), elastic net (Zou and Hastie, 2005), and bridge (Frank and Friedman, 1993), have been proposed to solve the problem. The ridge regression utilizes the L_2 penalty and is best used when there are high correlations between predictors. However, it could be influenced by irrelevant variables since it uses all the predictors in hand. The lasso utilizes the L_1 penalty and does both continuous shrinkage and automatic variable selection simultaneously. However, in the presence of multicollinearity, it has empirically been observed that the prediction performance of the lasso is dominated by ridge regression (Tibshirani, 1996). The elastic net attempts to keep the advantages of the ridge and lasso, and overcome their shortcomings by combining the L_1 and L_2 penalties. In addition, it has a grouping effect, i.e. if there is a set of variables among which the pairwise correlations are high, the elastic net groups the correlated variables together.

Bridge regression (Frank and Friedman, 1993; Fu, 1998; Knight and Fu, 2000; Liu et al., 2007; Huang et al., 2008) utilizes the L_q ($q > 0$) penalty and thus it includes the ridge ($q = 2$) and lasso ($q = 1$) as special cases. It is known that if $0 < q \leq 1$, bridge estimators produce sparse models. Due to the general L_q penalty form, bridge regression naturally fits any situation where it needs variable selection or there exists multicollinearity.

Recent developments of penalized methods are remarkable. Fan and Li (2001) suggested the scad (smoothly clipped absolute deviation) penalty and showed that the estimator has the oracle property; it works as well as if the correct submodel were known. Zou (2006) and Zhang and Lu (2007) developed the adaptive lasso and showed that their methods also possess the oracle property.

Park and Casella (2008) proposed the Bayesian lasso which provides interval estimates that can guide variable selection. Zou and Yuan (2008) imposed the F_∞ norm on support vector machines in the classification context. Zou and Zhang (2009) proposed the adaptive elastic net that combines the strengths of the quadratic regularization and the adaptively weighted lasso shrinkage. These methods are not included in our comparisons, but we suggest it as future work.

In this paper, we study the properties of bridge estimators thoroughly. In Section 2, we introduce two bridge regression algorithms based on the local linear and quadratic approximations, and discuss their advantages and drawbacks. We also show the superiority of the proposed estimators in various simulated settings and a real example. In Section 3, we develop bridge regression with adaptive L_q penalty for group variable selections. With the appropriate selection of tuning parameters, we establish the oracle property for the group bridge estimators whose proofs are delayed to Section 5. The comparisons of the proposed and existing methods are conducted via numerical examples. In Section 4, we apply the group bridge regression to varying-coefficient models using basis function approximations. This procedure is capable of simultaneously selecting important variables with time-varying effects and estimating unknown smooth functions using basis function approximations.

2 Bridge estimators with adaptive L_q penalty

In Section 2.1, we briefly review bridge regression. Section 2.2 introduces two bridge regression algorithms with adaptive L_q penalty using the local linear and quadratic approximations. In Section 2.3, we present our numerical results using simulated and real examples and thoroughly compare the OLS, ridge, lasso, adaptive lasso, scad, elastic net and bridge estimators against one another.

2.1 Background

Bridge regression is a broad class of the penalized regression method proposed by Frank and Friedman (1993). The bridge estimate can be obtained by minimizing

$$\hat{\boldsymbol{\beta}} = \arg \min_{\boldsymbol{\beta}} \left\{ \sum_{i=1}^n (Y_i - \mathbf{x}_i^T \boldsymbol{\beta})^2 + \lambda \sum_{j=1}^p |\beta_j|^q \right\}. \quad (2.1)$$

It does variable selection when $0 < q \leq 1$, and shrinks the coefficients when $q > 1$. Frank and Friedman (1993) did not solve for the estimator of bridge regression for any given $q > 0$, but they

pointed out that it is desirable to optimize the parameter q .

Fu (1998) studied the structure of bridge estimators and proposed a general algorithm to solve for $q \geq 1$. The shrinkage parameter q and the tuning parameter λ are selected via generalized cross-validation. Knight and Fu (2000) showed asymptotic properties of bridge estimators with $q > 0$ when p is fixed. Huang et al. (2008) studied the asymptotic properties of bridge estimators in sparse, high dimensional, linear regression models when the number of covariates p may increase along with the sample size n . Liu et al. (2007) introduced an L_q support vector machine algorithm which selects q from the data.

The effect of the L_q penalty with different q 's is well explained in Liu et al. (2007), and we briefly mention some part of it here along with the effect of the elastic net. Figure 1 (a) shows the L_q with $q = 0.1, 1, 2$ and the elastic net penalty is also overlaid. For the L_q , its penalty function is strictly convex if $q > 1$ and strictly nonconvex if $q < 1$. When $q = 1$, it is still convex but not differentiable at the origin. It is clearly shown that the elastic net penalty is between $q = 1$ (lasso) and $q = 2$ (ridge), and it is strictly convex.

In Figure 1 (b) we consider a simple linear regression model with one parameter θ and one observation $z = \theta + \epsilon$, where ϵ is a random error with mean 0 and variance σ^2 . Without any penalty, the OLS estimator $\hat{\theta}$ is z . When a penalty is used, we solve $\arg\min_{\theta} F(\theta)$ where $F(\theta) = (\theta - z)^2 + \lambda|\theta|^q$ for the L_q penalty and $F(\theta) = (\theta - z)^2 + \lambda_1|\theta| + \lambda_2\theta^2$ for the elastic net. In Figure 1 (b), we plot the minimizer of $F(\theta)$ for the L_q with $q = 0.1, 1, 2$ and the elastic net. For the L_q , when $q > 1$, large (small) $|\theta|$'s are more (less, respectively) shrunk toward 0 as q gets larger. When $q = 1$ (lasso), small $|\theta|$'s become exactly zero and large coefficients are shrunk in the same amount of magnitude. When $q < 1$, small $|\theta|$'s become exactly zero but large $|\theta|$'s are close to z (Huang et al., 2008). For the elastic net, small $|\theta|$'s become exactly zero but large coefficients are shrunk in the different amount of magnitude unlike the lasso.

2.2 Computation of bridge estimators

In practice, a learning procedure of the L_q penalty with a fixed q has its advantages over others only under certain situations because different types of penalties may suit best for different data structures. Since the best choice of q varies from problem to problem, we propose to treat q as a tuning parameter and select it adaptively. When there exist many noise or redundant variables, we expect the estimate with $q \leq 1$ for the automatic selection of important variables. When all the

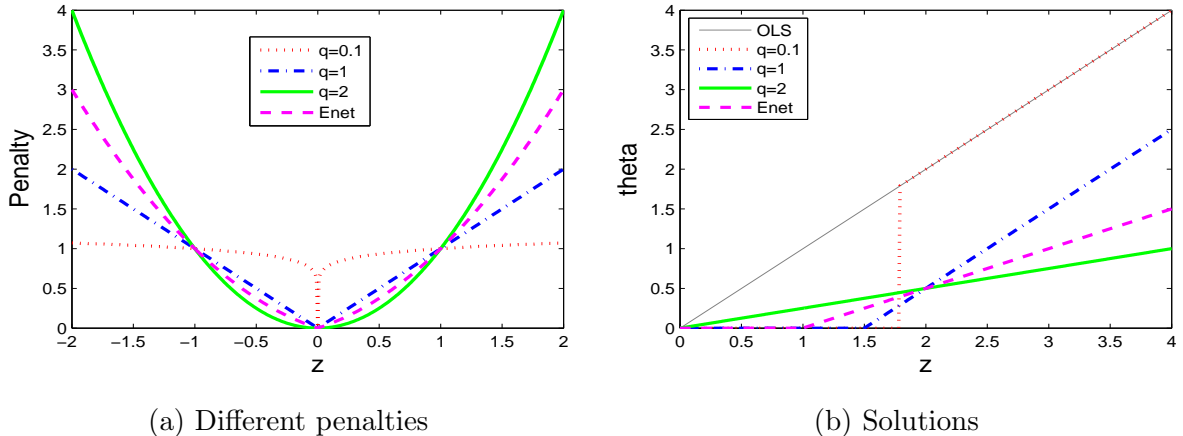


Figure 1: (a) Plots of the L_q penalties with $q = 0.1, 1, 2$ and elastic net penalty. (b) The corresponding solutions $\operatorname{argmin}_{\theta} F(\theta)$ with $\lambda = 3$, $\lambda_1 = 2$, and $\lambda_2 = 1$.

covariates are important or there are high correlations between variables, it may be more preferable to use $q > 1$ to avoid unnecessary variable deletion.

Despite the flexibility of bridge estimators, the nonconvexity of the penalty function may reduce the practical use of the estimators. In order to avoid the nonconvex optimization problem we introduce two algorithms for solving bridge regression. The first method applies the local quadratic approximation (LQA) suggested by Fan and Li (2001) and the second applies the local linear approximation (LLA) suggested by Zou and Li (2008).

For the LQA, under some mild conditions, the penalty term can be locally approximated at $\beta_0 = (\beta_{01}, \dots, \beta_{0p})^T$ by a quadratic function:

$$|\beta_j|^q \approx |\beta_{0j}|^q + \frac{q}{2} \frac{|\beta_{0j}|^{q-1}}{|\beta_{0j}|} (\beta_j^2 - \beta_{0j}^2).$$

Then, the minimization problem of (2.1) can be expressed as a quadratic minimization problem:

$$\hat{\beta} = \operatorname{arg} \min_{\beta} \left\{ \sum_{i=1}^n (Y_i - \mathbf{x}_i^T \beta)^2 + \frac{\lambda q}{2} \sum_{j=1}^p |\beta_{0j}|^{q-2} \beta_j^2 \right\}, \quad (2.2)$$

and the Newton-Raphson algorithm can be used. More specifically, for a given q and λ , we first initialize $\beta_0 = (\beta_{01}, \dots, \beta_{0p})^T$ and then calculate $\hat{\beta}^{(b)} = \left\{ \mathbf{X}^T \mathbf{X} + \Sigma_{\lambda}(\hat{\beta}^{(b-1)}) \right\}^{-1} \mathbf{X}^T \mathbf{Y}$ where $\Sigma_{\lambda}(\hat{\beta}^{(b-1)}) = \operatorname{diag}(\lambda q |\hat{\beta}_1^{(b-1)}|^{q-2}/2, \dots, \lambda q |\hat{\beta}_p^{(b-1)}|^{q-2}/2)$ for $b = 1, 2, \dots$. The algorithm stops when there is little change in $\hat{\beta}^{(b)}$, for example $\|\hat{\beta}^{(b)} - \hat{\beta}^{(b-1)}\| < \eta$ where η is a pre-selected small positive value. In our numerical examples, $\eta = 10^{-3}$ is used. During the iteration if $|\hat{\beta}_j^{(b-1)}| \leq \eta$, we delete

the j th variable to make the algorithm stable and also exclude it from the final model. For the initial value of β , the ridge coefficients provide a good starting value. For $0 < q \leq 1$, Hunter and Li (2005) showed that the LQA is a special case of a minorization-maximization (MM) algorithm and guarantees the ascent property of maximization problems.

Remarks.

- (i) For $q \geq 1$, the problem (2.1) is convex and solvable without using the approximation. In our algorithm, however, we instead solve the problem (2.2) because it provides a unified algorithm for all $q > 0$ and our limited simulation study shows little differences between the two.
- (ii) The LQA algorithm shares a drawback of backward stepwise variable selection, that is, if a variable is deleted at any step in the process of iteration, it would permanently be excluded from the final model. Hunter and Li (2005) proposed a perturbed version of the LQA, which is not forced to delete a covariate permanently, but the selection of the size of perturbation is another nontrivial task as pointed out by Zou and Li (2008).
- (iii) Since the LQA algorithm is a backward deletion procedure, it is desirable for the initial estimate to keep all the original variables, not to mention provide a precise guess. We use the ridge estimate as the initial value because it does not select variables and the simulation study in Section 2.3 shows that it is a robust choice in various settings. The ridge regression requires the selection of a tuning parameter and we choose it via grid search.

For the LLA, we use one-step estimates proposed by Zou and Li (2008), which automatically adopts a sparse representation. The one-step bridge estimator for $0 < q < 1$ is obtained as follows. Define $\mathbf{x}_{ij}^* = \sqrt{2}|\beta_{0j}|^{1-q}\mathbf{x}_{ij}/q$ and $Y_i^* = \sqrt{2}\mathbf{x}_i^T\beta_0$. Using $(\mathbf{x}_{ij}^*, Y_i^*)$, we apply the LARS algorithm (Efron et al., 2004) to solve

$$\hat{\beta}^* = \arg \min_{\beta} \left\{ \sum_{i=1}^n (Y_i^* - \mathbf{x}_i^{*T}\beta)^2 + \lambda \sum_{j=1}^p |\beta_j| \right\}.$$

Then, $\beta_{1,j} = \hat{\beta}_j^*|\beta_{0,j}|^{1-q}/q$.

Zou and Li (2008) showed that the LLA is the best convex MM algorithm, which proves the convergence of the LLA algorithm. The LLA naturally produces a sparse estimates without iterations, which also reduces the computational burden. In addition, it enjoys the efficient algorithm for solving the lasso. However, the automatic variable selection property limits the usefulness of

the LLA algorithm when all the covariates are important or there are high correlations between some sets of variables. We discuss this issue in Section 2.3.1.

For both algorithms, there are two tuning parameters λ and q . The parameter λ controls the tradeoff between minimizing the loss and the penalty, and q determines the order of penalty function. The proper choice of q is important and depends on the nature of data. We find the optimal combination of λ and q by grid search for the accurate assessment of the performance.

2.3 Numerical examples

2.3.1 Simulated examples

We perform a simulation study under similar settings done in Zou and Hastie (2005). The simulation has four settings and we compare the OLS, ridge, lasso, alasso (adaptive lasso), scad, enet (elastic net), and bridge estimators with the LQA and LLA. For each setting, our simulated data consist of three independent data sets: a training set, a validation set and a test set. For Settings 1, 2, and 3, the design matrix \mathbf{X} is simulated from a multivariate normal distribution with mean zero and variance one, and a certain correlation structure which is given at each setting. The signal-to-noise ratios are 2.36, 1.61, 1.20, and 3.27 for Settings 1, 2, 3, and 4, respectively. The specific settings are as follows:

1. Setting 1: 50 data sets, each consisting of 20/20/200 (training/validation/test) observations and 13 predictors with $\beta = (3, 1.5, 0, 0, 2, 0, 0, \dots, 0)$, $\sigma = 3$, and $\text{corr}(\mathbf{x}_i, \mathbf{x}_j) = 0.5^{|i-j|}$.
2. Setting 2: 50 data sets, each consisting of 20/20/200 observations and 8 predictors with $\beta_j = 0.85$ for all j , $\sigma = 3$, and $\text{corr}(\mathbf{x}_i, \mathbf{x}_j) = 0.5^{|i-j|}$.
3. Setting 3: 50 data sets, each consisting of 100/100/400 observations and 40 predictors with

$$\beta = (\underbrace{0, \dots, 0}_{10}, \underbrace{2, \dots, 2}_{10}, \underbrace{0, \dots, 0}_{10}, \underbrace{2, \dots, 2}_{10})$$

and $\sigma = 15$; $\text{corr}(\mathbf{x}_i, \mathbf{x}_j) = 0.5$ for all i and j .

4. Setting 4: 50 data sets, each consisting of 50/50/400 observations and 40 predictors with $\beta = (3, \dots, 3, 0, \dots, 0)$, that is the first 35 coefficients are 3 and the rest are zero and $\sigma = 15$.

Table 1: Mean Squared Errors

Setting	MSE	OLS	ridge	lasso	alasso	scad	enet	LQA (q)	LLA (q)
1	Mean	17.52	5.77	4.23	4.11	5.04	3.54	3.55 (0.66)	3.32 (0.68)
	s.e.	1.91	0.30	0.32	0.34	0.54	0.27	0.34 (0.09)	0.29 (0.03)
	Median	14.06	5.61	3.81	3.66	4.22	3.43	2.87 (0.40)	2.75 (0.70)
2	Mean	7.28	2.38	3.96	4.68	5.59	3.15	2.50 (2.30)	3.16 (0.77)
	s.e.	0.79	0.25	0.31	0.39	0.43	0.22	0.30 (0.13)	0.32 (0.03)
	Median	14.48	2.03	3.25	3.91	4.87	2.78	1.75 (3.00)	2.39 (0.90)
3	Mean	145.99	24.76	47.59	55.12	91.04	35.40	23.91 (2.30)	25.08 (0.89)
	s.e.	5.41	1.09	2.13	2.41	3.50	1.17	1.01 (0.11)	1.19 (0.01)
	Median	147.88	23.85	44.03	51.89	91.45	34.00	23.11 (2.50)	23.24 (0.90)
4	Mean	1148.43	57.37	89.83	78.64	61.26	59.40	55.12 (1.48)	83.95 (0.90)
	s.e.	109.69	3.93	5.99	5.63	4.62	4.52	4.69 (0.18)	7.18 (0.01)
	Median	902.21	53.55	85.67	72.77	56.21	59.33	49.41 (1.30)	71.02 (0.90)

The predictors X were generated as follows:

$$x_i = Z_k + \epsilon_i^x, \quad Z_k \sim N(0, 1), \quad i = 5k - 4, \dots, 5k, \quad k = 1, \dots, 7,$$

$$x_i \sim N(0, 1), \quad iid, \quad i = 36, \dots, 40.$$

where ϵ_i^x are errors in relation to the x and they are independent, identically distributed $N(0, 0.01)$, $i = 1, \dots, 35$. In this model, we have seven equally important groups, and within each group there are five members. There are also 5 pure noise features.

In the simulation we consider $\lambda = 2^{k-6}$ for $k = 1, 2, \dots, 20$ for all penalized methods. We also consider $q = 0.1, 0.4, 0.7, 1, 1.3, 1.7, 2, 2.5, 3$ for the LQA, and $q = 0.1, 0.3, 0.5, 0.7, 0.9$ for the LLA. We choose the best combination of (λ, q) which produces the lowest Mean Squared Error (MSE) over the validation sets.

Table 1 reports the MSE of each estimator for the four simulation settings. The mean, standard error (s.e.), and median of the MSE's are reported. For the bridge estimators, the mean, standard error, and median of the selected q 's are also included in the parentheses. Since the OLS consistently performs the worst, we exclude it from our discussion.

In Setting 1, the elastic net and bridge (both LQA and LLA) produce the lowest MSE's while the lasso, adaptive lasso, scad and ridge estimators do not perform poorly and stay close to the

other three methods. The elastic net performs well in this setting because it is designed for handling both variable selection and multicollinearity. A similar result can be found in Zou (2006). For the bridge estimators note that the mean values for the optimal q 's are 0.66 and 0.68 for the LQA and LLA, respectively. When many noise variables are present, bridge estimators reduce the MSE's by shrinking small coefficients more. Huang et al. (2008) showed that bridge estimators correctly select covariates with nonzero coefficients and the estimators of nonzero coefficients have the same asymptotic distribution that they would have if the zero coefficients were known in advance. Also, it can be seen that the LLA estimator performs better than the LQA in a sparse situation with weak correlations among the predictors.

In Setting 2, only weak correlations are present and no variable selection is intended. In this case, the ridge, elastic net, and bridge with $q > 1$ are expected to perform well (Zou and Hastie, 2005; Zou, 2006). The MSE's of these estimators are lowest as expected. The ridge performs the best and the LQA bridge closely follows the next. Since the variable selection is not needed, the LLA bridge cannot enjoy its automatic sparse representation in this setting and produces the higher MSE. The mean value for the optimal q selected by the LQA is 2.30.

For the third simulation setting a desirable estimator should handle both variable selection and strong correlations among the predictors. The result shows that the ridge and bridge estimators similarly perform the best while the lasso, adaptive lasso, scad and elastic net fail to handle the case in a satisfactory way due to the strong correlation structure. The elastic net does not perform well in Settings 2 and 3, but Zou and Hastie (2005) showed that the naive version of the elastic net yielded the lower MSE's. However, the naive version did not produce a satisfactory result in Setting 4 in their paper. Interestingly the two bridge estimators perform similarly while the mean values of the optimal q 's are 2.30 for the LQA and 0.89 for the LLA. This implies that when high correlations exist among the predictors and some of them are irrelevant to the response, the LQA (and ridge) takes care of multicollinearity while the LLA selects variables to reduce the MSE.

The fourth simulation setting is specifically suitable for the elastic net because it is designed to deal with the grouping effect. Other variable selection methods are not expected to perform well in this setting because the grouping effect forces the correlation among variables. In addition to the grouping effect, 5 out of the 40 coefficients are zeros. The elastic net performs well as expected, but the LQA yields the lowest mean and median MSE's and the ridge follows next. Due to the correlations between the variables the mean value of the optimal q is 1.47. Zou and Hastie (2005)

Table 2: Pollution data: prediction errors

Method	OLS	ridge	lasso	alasso	scad	enet	LQA	LLA
Prediction error	3398.64	3392.62	2201.77	2734.15	2659.04	2201.39	2195.06	2985.28
q							0.70	0.90

stated that “a regression method exhibits the grouping effect if the regression coefficients of a group of highly correlated variables tend to be equal.” We show that bridge estimators with $q > 1$ also possess the grouping effect in Theorem 3 in Section 5, which does not apply to the LLA estimator. Since the LLA automatically adopts a sparse representation, it selects one or two variables within a group, which results in producing the higher MSE in spite of the presence of noise variables in Setting 4. It suggests that the LLA might not always be the best choice in sparse situations because some variables might be grouped together.

In summary, the bridge estimators are the best or second best in all of the settings, which demonstrates that they are robust in various circumstances. It may not be the best performer for a single setting, but it outperforms the others in one way or another. We observe that the LLA can be a better choice over the LQA in some sparse situations, but its usefulness is limited especially when variable selection is not needed or variables are grouped together. Therefore, when the structure of a data set is not known beforehand, the LQA bridge estimators are a preferable candidate since they can adapt to the data and handle most of the cases in a satisfactory manner.

2.3.2 Pollution data

The data were analyzed by McDonald and Schwing (1973); Luo et al. (2006). This dataset is composed of 60 observations and 15 predictors. The response variable is the total age-adjusted mortality rate obtained for the years 1959–1961 for 201 Standard Metropolitan Statistical Areas. McDonald and Schwing (1973) used ridge regression and Luo et al. (2006) used an adaptive variable selection procedure based on a SIMEX method for choosing tuning parameters. In order to validate prediction errors we randomly select 40 observations for model fitting and use the rest as the test set. For the selection of tuning parameters ten-fold cross-validation is used.

Table 2 compares the prediction errors and it shows that the lasso, elastic net, and bridge (LQA) similarly perform the best. The selected $q = 0.7$ for the LQA and $q = 0.9$ for the LLA,

Table 3: Pollution data: selected variables

Method	Variables selected
McDonald and Schwing	(1,2,6,8,9,14)
Luo et al	(1,2,6,9,14)
lasso	(1,2,3,6,7,8,9,14)
alasso	(1,2,3,4,5,6,8,9,10,11,13,14)
scad	(1,2,3,4,8,9,10,14)
enet	(1,2,6,7,8,9,14)
LQA	(1,2,3,6,8,9,14)
LLA	(1,2,3,6,7,8,9,14,15)

which indicates that the data contain some noisy variables needed to be removed. Table 3 displays the variable selection result using the entire data set. The penalized variable selection procedures tend to choose more variables than the previous results and the LQA bridge and elastic net have the smallest model size among them.

3 Group bridge estimators with adaptive L_q penalty

In this section we consider regression problems where some explanatory factors may be represented by a group of derived input variables. In this case the selection of important variables corresponds to the selection of groups of variables. For example, in the multifactor ANOVA problem, each factor may have several levels and can be expressed through a group of dummy variables. The goal of ANOVA is often to select important main effects and interactions for accurate prediction, which amounts to the selection of groups of derived input variables. Yuan and Lin (2006) proposed a group lasso procedure which is capable of selecting meaningful blocks of covariates or categorical variables. Its extension to blockwise sparse regression is done by Kim and Kim (2006), and Nardi and Rinaldo (2008) established estimation and model selection consistency, prediction and estimation bounds and persistence for the group lasso estimator. The group lasso penalty has been applied to logistic regression by Meier and Bühlmann (2008), and to nonparametric problems by Bach (2008).

We propose group bridge estimators that include the group lasso as a special case. Recently

Huang et al. (2009) proposed a group bridge approach with $0 < q < 1$ that carries out both group variable selection and within-group individual variable levels simultaneously. Compared to Huang et al.'s method, the proposed group estimator adaptively selects the order q in the penalty function from the data. Also, the range of q is not restricted to between 0 and 1 so that it provides a more flexible solution in practice since it is unknown in advance whether variable selection is needed or not. In addition, the proposed estimator is applied to varying-coefficient models in Section 4.

Section 3.1 introduces the LQA algorithm. In Section 3.2, we compare the proposed method with the ridge, elastic net, group lasso, and group scad using simulated and real examples. In Section 3.3, we investigate theoretical properties of group bridge estimators.

3.1 Computation of group bridge estimators

Let m be the number of groups which is assumed to be known in advance and d_j is the size of the j th group so that $\sum_{j=1}^m d_j = p$. Then, the estimate can be obtained by minimizing

$$\ell_n(\boldsymbol{\beta}) = \sum_{i=1}^n (Y_i - \mathbf{x}_i^T \boldsymbol{\beta})^2 + \lambda \sum_{j=1}^m \tau_j \|\boldsymbol{\beta}_j\|^q \quad (3.1)$$

where $\lambda, \tau_j \geq 0$, $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^T, \dots, \boldsymbol{\beta}_m^T)^T$, and $\|\boldsymbol{\beta}_j\|^q = \left(\sum_{k=1}^{d_j} \beta_k^{(j)2} \right)^{q/2}$. Here $\beta_k^{(j)}$'s are the regression coefficients in the j th group. Again we let q be estimated from data to gain more flexibility. As we observe in Section 2, the proposed approach automatically adapts to either a sparse situation or multicollinearity. Note that if $q = 1$ in (3.1), the estimator is reduced to the group lasso. The use of τ_j allows the estimator to give different weights to the different coefficients. Yuan and Lin (2006) suggested to use $\tau_j = \sqrt{d_j}$ and Nardi and Rinaldo (2008) to use $\tau_j = \|\hat{\boldsymbol{\beta}}_j^{OLS}\|^{-1}$ for the group lasso. We consider both options in our numerical analysis and theoretical proofs. When $X^T X$ is close to singular, we suggest to use $\tau_j = \|\hat{\boldsymbol{\beta}}_j^{ridge}\|^{-1}$.

If $0 < q < 1$, we employ the LQA introduced in Section 2.2. Given an initial value of $\boldsymbol{\beta}_j^0$, $p_\lambda(\|\boldsymbol{\beta}_j\|) = \lambda \tau_j \|\boldsymbol{\beta}_j\|^q$ can be approximated by a quadratic form

$$p_\lambda(\|\boldsymbol{\beta}_j\|) \approx p_\lambda(\|\boldsymbol{\beta}_j^0\|) + \frac{1}{2} p'_\lambda(\|\boldsymbol{\beta}_j^0\|) \frac{(\boldsymbol{\beta}_j^T \boldsymbol{\beta}_j - \boldsymbol{\beta}_j^{0T} \boldsymbol{\beta}_j^0)}{\|\boldsymbol{\beta}_j^0\|}. \quad (3.2)$$

Therefore, the minimization problem of (3.1) is reduced to a quadratic minimization problem and the Newton-Raphson algorithm can be used. We summarize the proposed algorithm as follows.

For a given q, λ

- (i) initialize $\beta^0 = (\beta_1^0, \dots, \beta_m^0)^\top$.
- (ii) until $(\hat{\beta}^{(b)})$ converges
- $$\hat{\beta}^{(b)} = \left\{ \mathbf{X}^\top \mathbf{X} + \Sigma_\lambda(\hat{\beta}^{(b-1)}) \right\}^{-1} \mathbf{X}^\top \mathbf{Y},$$
- where $\Sigma_\lambda(\hat{\beta}^{(b-1)}) = \text{diag}\{(p'_\lambda(\|\hat{\beta}_1^{(b-1)}\|)/\|\hat{\beta}_1^{(b-1)}\|)I_{d_1}, \dots, (p'_\lambda(\|\hat{\beta}_m^{(b-1)}\|)/\|\hat{\beta}_m^{(b-1)}\|)I_{d_m}\}$
- with I_{d_j} a d_j dimensional identity matrix.

For the initial value of β , the ridge coefficients provide a good starting value as discussed in Section 2.2. In our real data analysis, λ and q are selected from ten-fold cross-validation. For the simulated examples, we select them by generating separate validation sets. In ANOVA examples, the number of groups m is known in advance, but it is not practically true in general problems. We suggest the adaptive selection of m as future work.

3.2 Numerical examples

In this section, we compare the proposed method with the ridge, elastic net, group lasso, and group scad using simulated and real examples. The group scad is implemented by replacing the penalty in (3.1) by the scad penalty proposed by Fan and Li (2001).

3.2.1 Simulation

As in Yuan and Lin (2006), we consider the four simulation settings:

1. Setting 1: 15 latent variables Z_1, \dots, Z_{15} are simulated from a centered multivariate normal distribution with $\text{corr}(Z_i, Z_j) = 0.5^{|i-j|}$. Each Z_i is trichotomized as 0, 1, or 2 given it is smaller than $\Phi^{-1}(\frac{1}{3})$, larger than $\Phi^{-1}(\frac{2}{3})$ or in between. Then Y is generated from

$$Y = 1.8I(Z_1 = 1) - 1.2I(Z_1 = 0) + I(Z_3 = 1) + 0.5I(Z_3 = 0) + I(Z_5 = 1) + I(Z_5 = 0) + \epsilon,$$

where $I(\cdot)$ is the indicator function and the noise ϵ is normally distributed with variance σ^2 chosen so that the signal-to-noise ratio is 1.8. For each run, 50 observations are collected.

2. Setting 2: We consider both main effects and second-order interactions. As in Setting 1, four categorical factors Z_1, Z_2, Z_3 and Z_4 were first generated. The true regression equation is

$$\begin{aligned} Y = & 3I(Z_1 = 1) + 2I(Z_1 = 0) + 3I(Z_2 = 1) + 2I(Z_2 = 0) + I(Z_1 = 1, Z_2 = 1) \\ & + 1.5I(Z_1 = 1, Z_2 = 0) + 2I(Z_1 = 0, Z_2 = 1) + 2.5I(Z_1 = 0, Z_2 = 0) + \epsilon, \end{aligned}$$

with signal-to-noise ratio of 3. For each simulated data set, 100 observations are collected.

3. Setting 3: 17 random variables Z_1, \dots, Z_{16} and W are independently generated from a standard normal distribution. The covariates are defined as $X_i = (Z_i + W)/\sqrt{2}$ and Y follows

$$Y = X_3^3 + X_3^2 + X_3 + \frac{1}{3}X_6^3 - X_6^2 + \frac{2}{3}X_6 + \epsilon,$$

where $\epsilon \sim N(0, 2^2)$. For each run, 100 observations are collected.

4. Setting 4: Covariates X_1, \dots, X_{10} are collected in the same fashion as Setting 3. Then the last 10 covariates X_{11}, \dots, X_{20} are trichotomized as in the first two models. This gives us a total of 10 continuous covariates and 10 categorical covariates. The regression equation is

$$Y = X_3^3 + X_3^2 + X_3 + \frac{1}{3}X_6^3 - X_6^2 + \frac{2}{3}X_6 + 2I(X_{11} = 0) + I(X_{11} = 1) + \epsilon,$$

where $\epsilon \sim N(0, 2^2)$. For each run, we collect 100 observations.

5. Setting 5: The data are collected in the same fashion as Setting 3 except for the diverging parameter $p = p_n = \lceil 4n^{2/3} \rceil - 5$ for $n = 200$ and 400 .

At each simulation setting we generate independent training, validation, and testing data sets with the same size and repeat it 100 times. Table 4 reports the mean, median, and standard error of the MSE's of the five estimators. We only report the result here with $\tau_j = \sqrt{d_j}$ for the group lasso, group scad, and group bridge to save space.

It can be seen that the group bridge estimator consistently performs the best in the five settings. For Settings 1 and 2, important main effects and/or interaction terms should be selected by a procedure and the group bridge with $0 < q < 1$ and group lasso show their superiority over the other methods. In Setting 3, due to the correlated continuous factors, the group bridge with $q > 1$ and ridge perform better than the others. Setting 4 has an additive model involving both continuous and categorical variables, and again the group bridge performs the best. In this case, the estimated q is slightly greater than 1 due to the simultaneous demands of the group selection and the presence of the correlated continuous factors. In Setting 5, we let $p = p_n$ diverge with n in Setting 3. We report the results with $n = 200$ and 400 , and they are similar to that of Setting 3. The gbridge and enet perform the best for $n = 200$ and the gbridge and glasso for $n = 400$. The estimated q is again slightly greater than 1 for both cases.

Table 4: Grouped variables simulation

Setting		ridge	enet	glasso	gscad	gbridge (q)
1 (Main effects)	Mean	3.19	2.97	2.84	3.19	2.79 (0.66)
	s.e.	0.12	0.12	0.11	0.15	0.12 (0.05)
	Median	3.02	2.88	2.79	3.03	2.77 (0.70)
2 (Interactions)	Mean	2.07	1.86	1.73	1.83	1.67 (0.28)
	s.e.	0.04	0.03	0.03	0.06	0.03 (0.03)
	Median	2.06	1.88	1.68	1.69	1.60 (0.10)
3 (Continuous)	Mean	4.13	4.23	4.19	4.19	4.13 (1.52)
	s.e.	0.05	0.07	0.06	0.06	0.06 (0.06)
	Median	4.08	4.15	4.11	4.18	4.10 (1.70)
4 (Mixed)	Mean	4.59	4.48	4.44	4.79	4.34 (1.14)
	s.e.	0.07	0.07	0.07	0.11	0.06 (0.04)
	Median	4.41	4.35	4.34	4.61	4.21 (1.00)
5 (Diverging) $n = 200, p_n = 131$	Mean	4.42	4.23	4.27	4.86	4.22 (1.36)
	s.e.	0.05	0.05	0.05	0.48	0.04 (0.05)
	Median	4.44	4.25	4.24	4.27	4.18 (1.30)
5 (Diverging) $n = 400, p_n = 212$	Mean	4.37	4.21	4.13	4.29	4.12 (1.03)
	s.e.	0.04	0.03	0.03	0.11	0.03 (0.05)
	Median	4.31	4.17	4.13	4.12	4.09 (1.00)

Table 5: Birth weight data

Method	ridge	enet	glasso	gscad	gbridge
Prediction error	557513.2	544363.0	538299.4	596122.8	534069.3 ($q=1.7$)

3.2.2 Birth weight data

We study the birth weight data set analyzed by Hosmer and Lemeshow (1989) and Yuan and Lin (2006). It includes the birth weights of 189 babies and eight predictors concerning the mother. Among the eight predictors, two are continuous (mother’s age in years and mother’s weight in pounds at the last menstrual period) and six are categorical (mother’s race: white, black or other), smoking status during pregnancy (yes or no), number of previous premature labors (0, 1, or 2 or more), history of hypertension (yes or no), presence of uterine irritability (yes or no), number of physician visits during the first trimester (0, 1, 2, or 3 or more). The data were collected at Baystate Medical Center, Springfield, Massachusetts, during 1986. Since a preliminary analysis suggested that nonlinear effects of both mother’s age and weight may exist, we model both effects by using third-order polynomials. We randomly select 151 observations for model fitting and use the rest as the test set to validate the prediction errors.

Table 5 summarizes the comparison of the prediction errors of the five estimators. It shows that the group bridge performs the best and the group lasso closely follows next. The estimated q for the group bridge is 1.7. We repeat the random selection of training and test sets many times and obtain the similar result as in Table 5.

3.3 Asymptotic properties

In this section we investigate theoretical properties of group bridge estimators by allowing the number of covariates $p = p_n$ to go to infinity along with the sample size n . Our study follows the work in Huang et al. (2008). Let the true parameter value be $\beta_0 = (\beta_{10}^T, \beta_{20}^T)^T$, where β_{10} is a $k_n \times 1$ vector and β_{20} is a $r_n (= p_n - k_n) \times 1$ vector. Suppose that $\beta_{10} \neq \mathbf{0}$ and $\beta_{20} = \mathbf{0}$. Assume that $\sum_{j=1}^{L_n} d_j = k_n$. We write $\mathbf{x}_i = (\mathbf{w}_i^T, \mathbf{z}_i^T)^T$, where \mathbf{w}_i consists of the first k_n covariates and \mathbf{z}_i consists of the remaining r_n covariates. Let \mathbf{X}_{1n} and \mathbf{X}_{2n} be the matrices whose transposes are $\mathbf{X}_1^T = (\mathbf{w}_1, \dots, \mathbf{w}_n)$ and $\mathbf{X}_2^T = (\mathbf{z}_1, \dots, \mathbf{z}_n)$, respectively. Let $\Sigma_n = n^{-1} \mathbf{X}^T \mathbf{X}$ and $\Sigma_{1n} =$

$n^{-1}\mathbf{X}_1^T\mathbf{X}_1$. Let ρ_{1n} and ρ_{2n} be the smallest and largest eigenvalues of Σ_n and let δ_{1n} and δ_{2n} be the smallest and largest eigenvalues of Σ_{1n} , respectively.

We state the conditions for consistency and oracle properties of group bridge estimators.

(A1) (a) $\rho_{1n} > 0$ for all n ; (b) $(p_n + \lambda_n \sum_{j=1}^{L_n} \tau_j d_j^{q/2}) / (n\rho_{1n}) \rightarrow 0$.

(A2) (a) $\lambda_n (\sum_{j=1}^{L_n} \tau_j^2 d_j^{q-1} / n)^{1/2} \rightarrow 0$; (b) $\lambda_n ((p_n + \lambda_n \sum_{j=1}^{L_n} \tau_j d_j^{q/2}) / \rho_{1n})^{-\frac{(2-q)}{2}} \sum_{j=1}^{r_n} \tau_j n^{-q/2} \rightarrow \infty$.

(A3) There exist constants $0 < b_0 < b_1 < \infty$ such that

$$b_0 \leq \min\{|\beta_{1j}|, 1 \leq j \leq k_n\} \leq \max\{|\beta_{1j}|, 1 \leq j \leq k_n\} \leq b_1.$$

(A4) (a) There exist constants $0 < \delta_1 < \delta_2 < \infty$ such that $\delta_1 \leq \delta_{1n} \leq \delta_{2n} \leq \delta_2$ for all n ; (b) $n^{-1/2} \max_{1 \leq i \leq n} \mathbf{w}_i^T \mathbf{w}_i \rightarrow 0$.

In what follows we present two theorems under these four conditions .

Theorem 1. (Consistency). Let $\hat{\beta}_n$ denote the minimizer of (3.1) and suppose that $q > 0$. Then,

$$\|\hat{\beta}_n - \beta\| = O_p \left(\left(\left(p_n + \lambda_n \sum_{j=1}^{L_n} \tau_j d_j^{q/2} \right) / (n\rho_{1n}) \right)^{1/2} \right).$$

In this theorem, if $\tau_j = \sqrt{d_j}$ as suggested by Yuan and Lin (2006),

$$\|\hat{\beta}_n - \beta\| = O_p \left(\left(\left(p_n + \lambda_n \sum_{j=1}^{L_n} d_j^{(q+1)/2} \right) / (n\rho_{1n}) \right)^{1/2} \right).$$

If we use $\tau_j = 1/\|\hat{\beta}_j^{OLS}\|$ as suggested by Nardi and Rinaldo (2008),

$$\|\hat{\beta}_n - \beta\| = O_p \left(\left(\left(p_n + \lambda_n \sum_{j=1}^{L_n} d_j^{(q-1)/2} \right) / (n\rho_{1n}) \right)^{1/2} \right).$$

Therefore, $\tau_j = 1/\|\hat{\beta}_j^{OLS}\|$ is more attractive in theory since it yields a faster convergence rate.

However, it can be problematic for sparse models and $\tau_j = \sqrt{d_j}$ is preferable in this case.

The next theorem shows that group bridge estimators have the oracle property, that is, they can perform as well as if the correct submodel were known.

Theorem 2. (Oracle Property). Suppose that $0 < q < 1$ and let $\hat{\beta}_n = (\hat{\beta}_{1n}^T, \hat{\beta}_{2n}^T)^T$, where $\hat{\beta}_{1n}$ and $\hat{\beta}_{2n}$ are estimators of β_{10} and β_{20} , respectively. Then,

(i) (Sparsity) $\hat{\beta}_{2n} = \mathbf{0}$ with probability converging to 1.

(ii) (Asymptotic normality) Let $s_n^2 = \sigma^2 \boldsymbol{\alpha}_n^T \Sigma_{1n}^{-1} \boldsymbol{\alpha}_n$ for any $k_n \times 1$ vector $\boldsymbol{\alpha}_n$ satisfying $\|\boldsymbol{\alpha}_n\| \leq 1$.

Then,

$$n^{-1/2} s_n^{-1} \boldsymbol{\alpha}_n^T (\hat{\beta}_{1n} - \beta_{10}) \rightarrow_D N(0, 1).$$

Proofs of both theorems are given in Section 5.

4 Varying-coefficient models

Nonparametric varying coefficient models (Hastie and Tibshirani, 1993) are a class of generalized regression models in which the coefficients are allowed to vary as smooth functions of other variables. In this way, modeling bias can significantly be reduced and models provide more appealing interpretability. Such models are particularly useful in longitudinal studies where they allow one to explore the extent to which covariates affect responses changing over time. See Hoover et al. (1998), Brumback and Rice (1998) and Fan and Zhang (1999) for details on novel applications of the varying-coefficient models to longitudinal data.

Suppose that we observe $(\mathbf{x}_i(t_{ij}), Y_i(t_{ij}))$ for the i th subject at discrete time point t_{ij} , $i = 1, \dots, n$ and $j = 1, \dots, J_i$. Then, the linear varying coefficient model can be written as

$$Y_i(t_{ij}) = \mathbf{x}_i(t_{ij})^T \boldsymbol{\beta}(t_{ij}) + \epsilon_i(t_{ij}), \quad (4.1)$$

where $\boldsymbol{\beta}(t) = (\beta_1(t), \dots, \beta_m(t))^T$ is a m -dimensional vector of smooth functions of t , and $\epsilon_i(t)$, $i = 1, \dots, n$ are independently identically distributed random processes, independent of $\mathbf{x}_i(t)$. The goal is to estimate $\boldsymbol{\beta}(t)$ nonparametrically and select relevant predictors $x_k(t)$ with nonzero functional coefficient $\beta_k(t)$ if necessary. Wang et al. (2007) proposed a group scad procedure for model selection for varying-coefficient models with time-independent covariates and demonstrated its application in analysis of microarray time course gene expression data. Wang et al. (2008) further developed the group scad for general nonparametric varying-coefficient models and provided theoretical justification. In this section we propose group bridge estimators for varying-coefficient models and compare it with the group scad.

4.1 Group bridge with adaptive L_q penalty for varying-coefficient models

Following Wang et al. (2008), we utilize a method based on basis expansion of $\beta(t)$ and penalized estimation using the group bridge penalty. Suppose that $\beta_k(t) = \sum_{l=1}^{\infty} \gamma_{kl} B_{kl}(t)$ where $\{B_{kl}(t)\}_{l=1}^{\infty}$ are orthonormal basis functions of a function space of interest. Then, $\beta_k(t)$ can be approximated by a truncated series and the model (4.1) becomes

$$Y_i(t_{ij}) = \sum_{k=1}^m \sum_{l=1}^{d_k} \gamma_{kl} x_i^{(k)}(t_{ij}) B_{kl}(t_{ij}) + \epsilon_i(t_{ij}),$$

where d_k is the number of basis functions in approximating the function $\beta_k(t)$ and $x_i^{(k)}(t_{ij})$ represents the k th covariate. Note that each function in (4.1) is characterized by a set of parameters $\boldsymbol{\gamma}_k = (\gamma_{k1}, \dots, \gamma_{kd_k})^T$. When selecting coefficients, we should select nonzero $\boldsymbol{\gamma}_k$ as a group instead of individual nonzero γ_{kl} . Thus, we use the group bridge regression considered in Section 3.1. Then, $\hat{\boldsymbol{\gamma}} = (\hat{\boldsymbol{\gamma}}_1^T, \dots, \hat{\boldsymbol{\gamma}}_m^T)^T$ can be obtained by minimizing

$$\frac{1}{N} \sum_{i=1}^n \sum_{j=1}^{J_i} \left(Y_i(t_{ij}) - \sum_{k=1}^m \sum_{l=1}^{d_k} \gamma_{kl} x_i^{(k)}(t_{ij}) B_{kl}(t_{ij}) \right)^2 + \lambda \sum_{k=1}^m \tau_k \|\boldsymbol{\gamma}_k\|^q, \quad (4.2)$$

where $N = \sum_{i=1}^n J_i$. Then, $\hat{\beta}_k(t) = \sum_{l=1}^{d_k} \hat{\gamma}_{kl} B_{kl}(t)$.

To introduce a simplified algorithm, we follow the notations of Wang et al. (2008). We define

$$\mathbf{B}(t) = \begin{pmatrix} B_{11}(t) & \cdots & B_{1d_1}(t) & 0 & \cdots & 0 & 0 & \cdots & 0 \\ & & \vdots & & & \vdots & & & \\ 0 & \cdots & 0 & 0 & \cdots & 0 & B_{m1}(t) & \cdots & B_{md_m}(t) \end{pmatrix},$$

$\mathbf{U}_i(t_{ij}) = (\mathbf{x}_i(t_{ij})^T \mathbf{B}(t_{ij}))^T$, $\mathbf{U}_i = (\mathbf{U}_i(t_{i1}), \dots, \mathbf{U}_i(t_{iJ_i}))^T$, and $\mathbf{U} = (\mathbf{U}_1, \dots, \mathbf{U}_n)$. We also define $\mathbf{Y} = (Y_1(t_{11}), \dots, Y_n(t_{nJ_n}))^T$. Then, the objective function in (4.2) can be expressed as

$$\frac{1}{N} \sum_{i=1}^n (\mathbf{Y}_i - \mathbf{U}_i \boldsymbol{\gamma})^T (\mathbf{Y}_i - \mathbf{U}_i \boldsymbol{\gamma}) + \lambda \sum_{k=1}^m \tau_k \|\boldsymbol{\gamma}_k\|^q.$$

Using the same local quadratic approximation (LQA) in (3.2) we summarize the proposed algorithm as follows. For a given q, λ

(i) initialize $\boldsymbol{\gamma}^0 = (\boldsymbol{\gamma}_1^0, \dots, \boldsymbol{\gamma}_m^0)^T$.

(ii) until $(\boldsymbol{\gamma}^{(b)})$ converges)

$$\boldsymbol{\gamma}^{(b)} = \left\{ \mathbf{U}^T \mathbf{U} + N/2 \boldsymbol{\Sigma}_\lambda(\hat{\boldsymbol{\gamma}}^{(b-1)}) \right\}^{-1} \mathbf{U}^T \mathbf{Y},$$

where $\Sigma_\lambda(\hat{\gamma}^{(b-1)}) = \text{diag}\{(p'_\lambda(\|\hat{\gamma}_1^{(b-1)}\|)/\|\hat{\gamma}_1^{(b-1)}\|)I_{d_1}, \dots, (p'_\lambda(\|\hat{\gamma}_m^{(b-1)}\|)/\|\hat{\gamma}_m^{(b-1)}\|)I_{d_m}\}$
with I_{d_j} a d_j dimensional identity matrix.

For the initial value of γ , the ridge coefficients provide a good starting value as discussed in Section 2.2. In our analysis we choose $\tau_k = \sqrt{d_k}$.

In the algorithm, d_k , $k = 1, \dots, m$, λ and q are selected from the data. For d_k , we only consider the situation when $d_k = d$ for all $\beta_k(t)$ and try different values of d . In other words, we use the same number of basis functions for all $\beta_k(t)$. Other tuning parameters λ and q can be selected by the approximate cross-validation (ACV) as in Wang et al. (2008):

$$ACV(\lambda, q) = \sum_{i=1}^n \sum_{j=1}^{J_i} (Y_i(t_{ij}) - U^{(-i)}(t_{ij})\hat{\gamma}^{(-i)})^2$$

where $(-i)$ denotes the deletion of the i th subject.

4.2 Simulation

In this section we study the performance of the proposed group bridge estimators via simulated varying-coefficient models considered in Wang et al. (2008). In each run, a simple random sample of 200 subjects is generated according to the model:

$$Y_i(t_{ij}) = \beta_0(t_{ij}) + \sum_{k=1}^{23} \beta_k(t_{ij})x_k(t_{ij}) + \epsilon(t_{ij}), \quad i = 1, \dots, 200, \quad j = 1, \dots, J_i.$$

The first three variables $x_k(t)$, $k = 1, 2, 3$ are the relevant variables to the response variable $Y(t)$. At a given time t , $x_1(t)$ is distributed uniformly from $[t/10, 2+t/10]$. Conditioning on $x_1(t)$, $x_2(t)$ is the gaussian process with mean zero and variance $(1+x_1(t))/(2+x_1(t))$. With a success probability of 0.6, $x_3(t)$ is a Bernoulli random variable, and is independent of $x_1(t)$ and $x_2(t)$. In addition, we generate 20 redundant variables $x_k(t)$, $k = 4, \dots, 23$, where each $x_k(t)$ is a random gaussian process with covariance structure $cov(x_k(t), x_s(t)) = 4 \exp(-|k-s|)$. The error $\epsilon(t)$ is given by $Z(t) + \delta(t)$, where $Z(t)$ has the same distribution as $x_k(t)$, $k = 4, \dots, 23$, and $\delta(t)$ is simulated from $N(0, 4)$ at a given time t . The coefficients corresponding to the constant term and the relevant variables, $\beta_k(t)$, $k = 0, 1, 2, 3$, are given by

$$\begin{aligned} \beta_0(t) &= 15 + 20 \sin\left(\frac{\pi t}{60}\right), & \beta_1(t) &= 2 - 3 \cos\left(\frac{\pi(t-25)}{15}\right), \\ \beta_2(t) &= 6 - 0.2t, & \beta_3(t) &= -4 + \frac{(20-t)^3}{2000}. \end{aligned}$$

Table 6: Varying-coefficient models simulation

	$\beta_0(t)$		$\beta_1(t)$	
Methods	means (s.e.)	median	means (s.e.)	median
gbridge	0.07597 (0.00357)	0.07190	0.00957 (0.00025)	0.00911
gscad	0.08012 (0.00368)	0.07629	0.00998 (0.00026)	0.00949
	$\beta_2(t)$		$\beta_3(t)$	
Methods	means (s.e.)	median	means (s.e.)	median
gbridge	0.00070 (3.55E-05)	0.00067	0.00022 (1.52E-05)	0.00020
gscad	0.00075 (3.88E-05)	0.00073	0.00023 (1.58E-05)	0.00020

The remaining coefficients corresponding to the irrelevant variables are set by $\beta_k(t) = 0$, $k = 4, \dots, 23$. We also generate the observation time points t_{ij} as follows. Each subject has a set of scheduled time points $\{1, \dots, 30\}$, and each scheduled time has a probability of 60% of being skipped. Then the actual observation time t_{ij} is obtained by adding a random perturbation uniformly distributed on $(-0.5, 0.5)$ to the nonskipped scheduled time.

In order to compare the performance, we compute the $MSE(\hat{\beta}_j) = E(\beta_j(t) - \hat{\beta}_j(t))^2$, $j = 0, 1, 2, 3$. The MSE's are estimated via 100 Monte Carlo simulations. We try $d = 1, 2, 3, 4$ and discover that they produce similar results. Thus, we report the result with $d = 2$ to save space.

The mean, standard error, and median of MSE's are reported in Table 6. It indicates that the two methods are competitive but the group bridge performs slightly better than the group scad for the estimates of β_j ($j = 0, 1, 2, 3$). In terms of the selection of variables, both methods select the variables 1–3 in each run. For the variables 4–23, the group scad selects about 9.69 variables in each run but the group bridge selects none, which shows the superior performance of the proposed method. The group bridge selects $q = 0.1$ in each run. Figure 2 displays the group bridge estimates of the time-varying coefficients. True, the 2.5%, 50%, and 97.5% pointwise quantiles of the estimated time-varying coefficients are closely overlaid together, and the plots show that the group bridge gives an excellent fit.

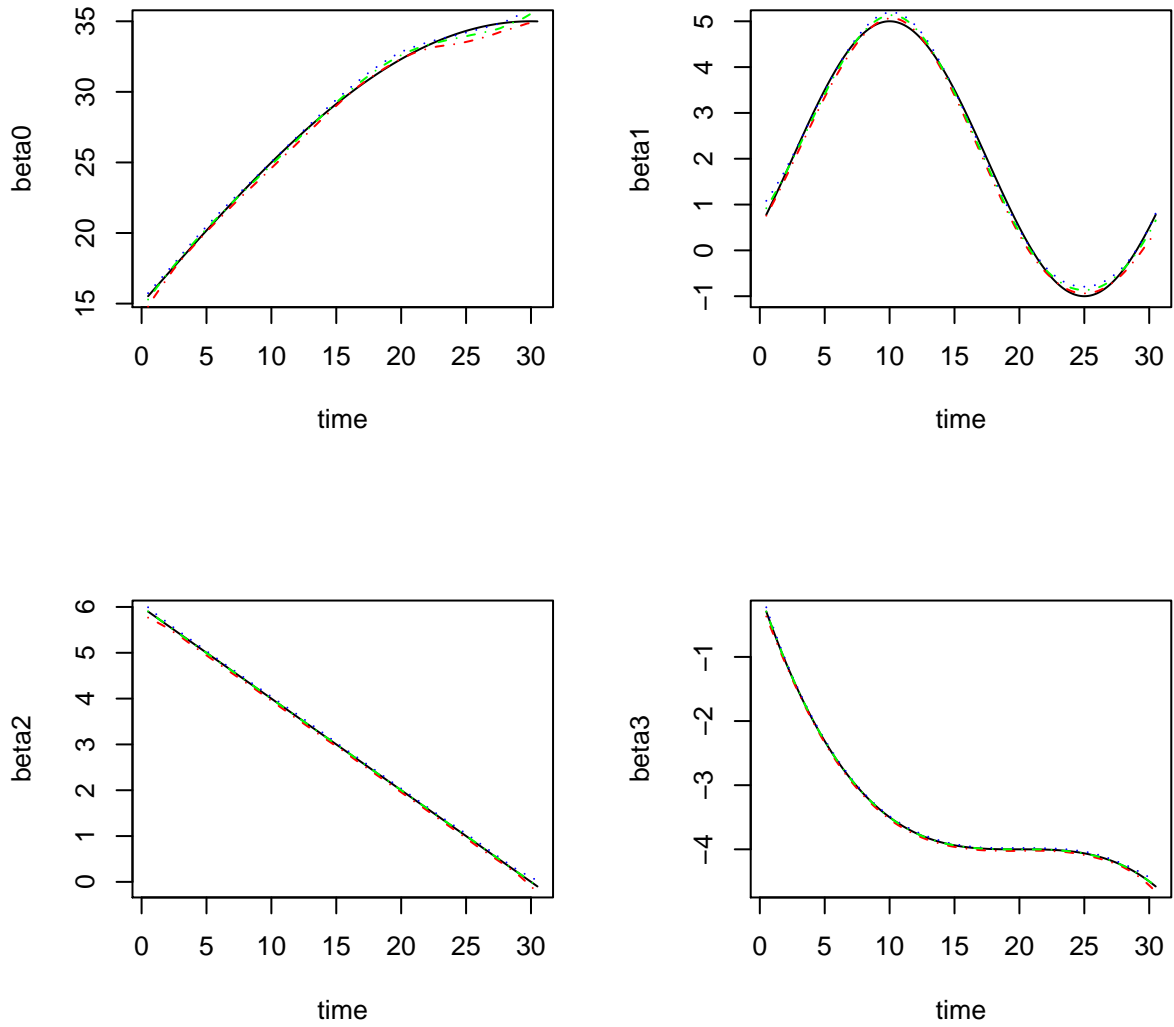


Figure 2: True (solid lines) and the median of the estimated (dashed lines) time-varying coefficients $\beta_0(t)$, $\beta_1(t)$, $\beta_2(t)$, and $\beta_3(t)$ over 100 replications. The 2.5% (dotted-dashed lines) and 97.5% (dotted lines) pointwise quantile curves are overlaid as well.

5 Appendix

In this section we provide proofs for Theorems 1-3.

5.1 Proof of Theorem 1

By allowing the number of groups $m = m_n$ to increase with n , we can show that

$$\sum_{i=1}^n (Y_i - \mathbf{x}_i^\top \hat{\boldsymbol{\beta}})^2 \leq \sum_{i=1}^n (Y_i - \mathbf{x}_i^\top \boldsymbol{\beta}_0)^2 + \lambda_n \sum_{j=1}^{m_n} \tau_j \|\boldsymbol{\beta}_{0j}\|^q.$$

Let $\eta_n = \lambda_n \sum_{j=1}^{m_n} \tau_j \|\boldsymbol{\beta}_{0j}\|^q$, then by Huang et al. (2008)

$$nE[(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0)^\top \Sigma_n (\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0)] \leq 6\sigma^2 p_n + 3\eta_n.$$

Since the number of nonzero coefficients is $k_n = \sum_{j=1}^{L_n} d_j$,

$$\begin{aligned} \eta_n &= \lambda_n \sum_{j=1}^{m_n} \tau_j \|\boldsymbol{\beta}_{0j}\|^q = \lambda_n \sum_{j=1}^{L_n} \tau_j \|\boldsymbol{\beta}_{0j}\|^q \\ &= O\left(\lambda_n \sum_{j=1}^{L_n} \tau_j d_j^{q/2}\right). \end{aligned}$$

Note that ρ_{1n} is the smallest eigenvalue of Σ_{1n} and this leads us to

$$\|\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}\| = O_p\left(\left(\left(p_n + \lambda_n \sum_{j=1}^{L_n} \tau_j d_j^{q/2}\right) / (n\rho_{1n})\right)^{1/2}\right).$$

5.2 Proof of Theorem 2

Proof of (i). Let $h_n = ((p_n + \lambda_n \sum_{j=1}^{L_n} \tau_j d_j^{q/2}) / (n\rho_{1n}))^{1/2}$. By Theorem 1, for a sufficiently large C , $\hat{\boldsymbol{\beta}}_n$ lies in the ball $\{\boldsymbol{\beta} : \|\boldsymbol{\beta} - \boldsymbol{\beta}_0\| \leq h_n C\}$ with probability converging to 1. Let $\boldsymbol{\beta}_{1n} = \boldsymbol{\beta}_{01} + h_n \mathbf{u}_1$ and $\boldsymbol{\beta}_{2n} = \boldsymbol{\beta}_{02} + h_n \mathbf{u}_2 = h_n \mathbf{u}_2$ with $\|\mathbf{u}\|^2 = \|\mathbf{u}_1\|^2 + \|\mathbf{u}_2\|^2 \leq C^2$. Let

$$V_n(\mathbf{u}_1, \mathbf{u}_2) = \ell_n(\boldsymbol{\beta}_{1n}, \boldsymbol{\beta}_{2n}) - \ell_n(\boldsymbol{\beta}_{10}, \mathbf{0}) = \ell_n(\boldsymbol{\beta}_{10} + h_n \mathbf{u}_1, h_n \mathbf{u}_2) - \ell_n(\boldsymbol{\beta}_{10}, \mathbf{0}).$$

Then, $\hat{\boldsymbol{\beta}}_{1n}$ and $\hat{\boldsymbol{\beta}}_{2n}$ can be obtained by minimizing $V_n(\mathbf{u}_1, \mathbf{u}_2)$ over $\|\mathbf{u}\| \leq C$, except on an event with probability converging to zero. To prove (i), it suffices to show that, for any \mathbf{u}_1 and \mathbf{u}_2 with

$\|\mathbf{u}\| \leq C$, if $\|\mathbf{u}_2\| > 0$, $V_n(\mathbf{u}_1, \mathbf{u}_2) - V_n(\mathbf{u}_1, \mathbf{0}) > 0$ with probability converging to 1. By Huang et al. (2008),

$$\begin{aligned} V_n(\mathbf{u}_1, \mathbf{u}_2) - V_n(\mathbf{u}_1, \mathbf{0}) &= h_n^2 \sum_{i=1}^n (\mathbf{z}_i^T \mathbf{u}_2)^2 + 2h_n^2 \sum_{i=1}^n (\mathbf{w}_i^T \mathbf{u}_1)(\mathbf{z}_i^T \mathbf{u}_2) \\ &\quad - 2h_n \sum_{i=1}^n \epsilon_i (\mathbf{z}_i^T \mathbf{u}_2) + \lambda_n h_n^q \sum_{j=1}^{r_n} \tau_j \|\mathbf{u}_{2j}\|^q \\ &\equiv I + II + III + IV. \end{aligned}$$

By (A4)(a), we have

$$I + II \geq -nh_n^2 \delta_{2n} \|\mathbf{u}_1\|^2 \geq -nh_n^2 \delta_2 C^2, \quad (5.1)$$

$$|III| \leq h_n n^{1/2} p_n^{1/2} O_p(1), \quad (5.2)$$

$$IV \leq \lambda_n h_n^q \sum_{j=1}^{r_n} \tau_j C^q. \quad (5.3)$$

Under the condition (A2)(b), $n^{-1} \lambda_n h_n^{-(2-q)} \sum_{j=1}^{r_n} \tau_j \rightarrow \infty$. Combining (5.1), (5.2), and (5.3), we have $V_n(\mathbf{u}) > 0$ with probability converging to 1.

Proof of (ii). By Theorem 1, $\hat{\boldsymbol{\beta}}_{1n}$ is consistent. By condition (A3), each component of $\hat{\boldsymbol{\beta}}_{1n}$ stays away from zero for sufficiently large n . Thus, it satisfies the equation $(\partial/\partial \boldsymbol{\beta}_1) \ell_n(\hat{\boldsymbol{\beta}}_{1n}, \hat{\boldsymbol{\beta}}_{2n}) = 0$. That is,

$$-2 \sum_{i=1}^n (Y_i - \mathbf{w}_i^T \hat{\boldsymbol{\beta}}_{1n} - \mathbf{z}_i^T \hat{\boldsymbol{\beta}}_{2n}) \mathbf{w}_i + \lambda_n q \psi_n = 0,$$

where $\psi_n = (\tau_1 \|\hat{\boldsymbol{\beta}}_1\|^{q-2} \hat{\boldsymbol{\beta}}_1^T, \dots, \tau_{L_n} \|\hat{\boldsymbol{\beta}}_{L_n}\|^{q-2} \hat{\boldsymbol{\beta}}_{L_n}^T)^T$. Since $\boldsymbol{\beta}_{20} = \mathbf{0}$ and $\epsilon_i = Y_i - \mathbf{w}_i^T \boldsymbol{\beta}_{10}$, we have

$$n^{1/2} \boldsymbol{\alpha}_n^T (\hat{\boldsymbol{\beta}}_{1n} - \boldsymbol{\beta}_{10}) = n^{-1/2} \sum_{i=1}^n \epsilon_i \boldsymbol{\alpha}_n^T \Sigma_{1n}^{-1} \mathbf{w}_i - \frac{1}{2} q n^{-1/2} \lambda_n \boldsymbol{\alpha}_n^T \Sigma_{1n}^{-1} \psi_n - n^{-1/2} \sum_{i=1}^n \mathbf{z}_i^T \hat{\boldsymbol{\beta}}_{2n} \mathbf{w}_i.$$

By (i), the last term on the right hand side equals zero with probability converging to 1. When $\|\boldsymbol{\alpha}_n\| \leq 1$, under condition (A3),

$$\begin{aligned} |n^{-1/2} \boldsymbol{\alpha}_n^T \Sigma_{1n}^{-1} \psi_n| &\leq n^{-1/2} \delta_1^{-1} \|\boldsymbol{\alpha}_n\| \cdot \|\psi_n\| \\ &\leq n^{-1/2} \delta_1^{-1} \left(\sum_{j=1}^{L_n} \tau_j^2 \|\hat{\boldsymbol{\beta}}_j\|^{2q-2} \right)^{1/2} \\ &\leq (\text{constant}) \cdot n^{-1/2} \left(\sum_{j=1}^{L_n} \tau_j^2 d_j^{q-1} \right)^{1/2}, \end{aligned}$$

which goes to 0 under (A2)(a). Therefore,

$$n^{1/2} s_n^{-1} \boldsymbol{\alpha}_n^T (\hat{\boldsymbol{\beta}}_{1n} - \boldsymbol{\beta}_{10}) = n^{-1/2} s_n^{-1} \sum_{i=1}^n \epsilon_i \boldsymbol{\alpha}_n^T \boldsymbol{\Sigma}_{1n}^{-1} \mathbf{w}_i + o_p(1).$$

The conditions of Linderg-Feller central limit theorem can be verified as in Huang et al. (2008).

5.3 Theorem 3 and its proof

Theorem 3. *Given data (\mathbf{Y}, \mathbf{X}) and parameters λ and $q > 1$, the response \mathbf{Y} is centered and the predictors \mathbf{X} are standardized. Let $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \dots, \hat{\beta}_p)^T$ be the bridge estimate. Define*

$$D_{(\lambda, q)}(i, j) = \frac{1}{\|\mathbf{Y}\|} \left| \hat{\beta}_i^{q-1} - \hat{\beta}_j^{q-1} \right|,$$

then

$$D_{(\lambda, q)}(i, j) \leq \frac{2}{\lambda q} \sqrt{2(1 - \rho)},$$

where $\|\mathbf{Y}\|^2 = \sum_i Y_i^2$ and $\rho = \mathbf{x}_i^T \mathbf{x}_j$.

The quantity $D_{(\lambda, q)}(i, j)$ describes the difference between the coefficient paths of predictors i and j . If \mathbf{x}_i and \mathbf{x}_j are highly correlated, i.e. $\rho \approx 1$, Theorem 3 implies that the difference $D_{(\lambda, q)}(i, j)$ is almost 0.

Proof.

Let

$$\begin{aligned} L(\lambda, q, \boldsymbol{\beta}) &= \sum_{i=1}^n (Y_i - \mathbf{x}_i^T \boldsymbol{\beta})^2 + \lambda \sum_{j=1}^p |\beta_j|^q \\ &= \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda \|\boldsymbol{\beta}\|^q. \end{aligned} \tag{5.4}$$

Following the proof of Theorem 1 in Zou and Hastie (2005), the solution to (5.4), $\hat{\boldsymbol{\beta}}(\lambda, q)$, satisfies

$$\left. \frac{\partial L(\lambda, q, \boldsymbol{\beta})}{\partial \beta_k} \right|_{\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}(\lambda, q)} = 0, \quad \text{if } \hat{\beta}_k(\lambda, q) \neq 0.$$

Hence we have

$$-2\mathbf{x}_i^T \{\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}(\lambda, q)\} + \lambda q \hat{\beta}_i^{q-1}(\lambda, q) = 0, \tag{5.5}$$

$$-2\mathbf{x}_j^T \{\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}(\lambda, q)\} + \lambda q \hat{\beta}_j^{q-1}(\lambda, q) = 0. \tag{5.6}$$

Subtracting equation (5.5) from equation (5.6) gives

$$\hat{\beta}_i^{q-1}(\lambda, q) - \hat{\beta}_j^{q-1}(\lambda, q) = \frac{2}{\lambda q} (\mathbf{x}_i^T - \mathbf{x}_j^T) \hat{\mathbf{r}}(\lambda, q) \tag{5.7}$$

where $\hat{\mathbf{r}}(\lambda, q) = \mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}(\lambda, q)$. Since \mathbf{X} are standardized, $\|\mathbf{x}_i - \mathbf{x}_j\|^2 = 2(1 - \rho)$ where $\rho = \mathbf{x}_i^T \mathbf{x}_j$. By equation (5.4) we have

$$L(\lambda, q, \hat{\boldsymbol{\beta}}(\lambda, q)) \leq L(\lambda, q, \boldsymbol{\beta} = \mathbf{0}),$$

that is,

$$\|\hat{\mathbf{r}}(\lambda, q)\|^2 + \lambda \|\hat{\boldsymbol{\beta}}(\lambda, q)\|^q \leq \|\mathbf{Y}\|^2,$$

which implies $\|\hat{\mathbf{r}}(\lambda, q)\|^2 \leq \|\mathbf{Y}\|^2$. Then by equation (5.7),

$$D_{(\lambda, q)}(i, j) \leq \frac{2}{\lambda q \|\mathbf{Y}\|} \|\mathbf{x}_i^T - \mathbf{x}_j^T\| \|\hat{\mathbf{r}}(\lambda, q)\| \leq \frac{2}{\lambda q} \sqrt{2(1 - \rho)}.$$

Acknowledgements

We would like to thank Lifeng Wang and Yin Xiong for helping us with the simulation studies.

References

- Bach, F. (2008). Consistency of the group Lasso and multiple kernel learning. *Journal of Machine Learning*, 9:1179–1225.
- Brumback, B. and Rice, J. (1998). Smoothing spline models for the analysis of nested and crossed samples of curves. *Journal of the American Statistical Association*, 93:961–976.
- Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). Least angle regression. *The Annals of Statistics*, 32:407–499.
- Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96:1348–1360.
- Fan, J. and Zhang, W. (1999). Statistical estimation in varying-coefficient models. *The Annals of Statistics*, 27:1491–1518.
- Frank, I. E. and Friedman, J. H. (1993). A statistical view of some chemometrics regression tools. *Technometrics*, 35:109–148.
- Fu, W. J. (1998). Penalized regression: the Bridge versus the Lasso. *Journal of Computational and Graphical Statistics*, 7:397–416.

- Hastie, T. and Tibshirani, R. (1993). Varying-coefficient models. *Journal of the Royal Statistical Society Series B*, 55:757–796.
- Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: biased estimation for nonorthogonal problems. *Technometrics*, 12:55–67.
- Hoover, D. R., Rice, J. A., Wu, C. O., and Yang, L. P. (1998). Nonparametric smoothing estimates of time-varying coefficient models with longitudinal data. *Biometrika*, 85:809–822.
- Hosmer, D. W. and Lemeshow, S. (1989). *Applied Logistic Regression*. New York: Wiley.
- Huang, J., Horowitz, J. L., and Ma, S. (2008). Asymptotic properties of bridge estimators in sparse high-dimensional regression models. *The Annals of Statistics*, 36:587–613.
- Huang, J., Ma, S., Xie, H., and Zhang, C.-H. (2009). A group bridge approach for variable selection. *Biometrika*, 96:339–355.
- Hunter, D. and Li, R. (2005). Variable selection using MM algorithms. *The Annals of Statistics*, 33:1617–1642.
- Kim, Y., K. J. and Kim, Y. (2006). Blockwise sparse regression. *Statistica Sinica*, 16:375–390.
- Knight, K. and Fu, W. J. (2000). Asymptotics for Lasso-type estimators. *The Annals of Statistics*, 28:1356–1378.
- Liu, Y., Zhang, H., Park, C., and Ahn, J. (2007). Support vector machines with adaptive L_q penalty. *Computational Statistics and Data Analysis*, 51:6380–6394.
- Luo, X., Stefanski, L. A., and Boos, D. D. (2006). Tuning variable selection procedures by adding noise. *Technometrics*, 48:165–175.
- McDonald, G. and Schwing, R. (1973). Instabilities of regression estimates relating air pollution to mortality. *Technometrics*, 15:463–482.
- Meier, L., v. d. G. S. and Bühlmann, P. (2008). The group lasso for logistic regression. *Journal of the Royal Statistical Society Series B*, 70:53–71.
- Nardi, Y. and Rinaldo, A. (2008). On the asymptotic properties of the group lasso estimator for linear models. *Electronic Journal Statistics*, 2:605–633.

- Park, T. and Casella, G. (2008). The Bayesian Lasso. *Journal of the American Statistical Association*, 103:681–686.
- Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B*, 58:267–288.
- Wang, L., , Li, H., and Huang, J. (2008). Variable selection for nonparametric varying-coefficient models for analysis of repeated measurements. *Journal of the American Statistical Association*, 103:1556–1569.
- Wang, L., Chen, G., and Li, H. (2007). Group SCAD regression analysis for microarray time course gene expression data. *Bioinformatics*, 23:1486–1494.
- Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society Series B*, 68:49–67.
- Zhang, H. H. and Lu, W. (2007). Adaptive-LASSO for Cox’s proportional hazard model. *Biometrika*, 94:691–703.
- Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101:1418–1429.
- Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B*, 67:301–320.
- Zou, H. and Li, R. (2008). One-step sparse estimates in nonconcave penalized likelihood models. *The Annals of Statistics*, 36:1509–1533.
- Zou, H. and Yuan, M. (2008). The F_∞ -norm support vector machine. *Statistica Sinica*, 18:379–398.
- Zou, H. and Zhang, H. H. (2009). On the adaptive elastic-net with a diverging number of parameters. *The Annals of Statistics*, 37:1733–1751.