Biometrics 000, 000–000 DOI: 000
000 0000

Regularization and variable selection for data with interdependent structures

Lingmin Zeng* and Jun Xie

Department of Statistics, Purdue University,

250 N. University Street, West Lafayette, IN 47907.

**email:* lzeng@stat.purdue.edu

Summary: Variable selection methods are powerful tools in analysis of high dimensional massive data. Specifically, the methods have been applied in gene expression microarray data analysis. It is well known that for genes sharing a common biological pathway or a similar function, the correlations among them can be very high. However, most of the available variable selection methods cannot deal with complicated interdependence among data. We propose two new algorithms, namely gLars and gRidge, to select groups of highly correlated variables together in regression models. The new approaches intent to conduct grouping and selecting at the same time. Simulations and a real example show that our proposed methods often outperform the existing variable selection methods, including LARS and elastic net, in terms of both prediction error and preserving sparsity of representation.

Key words: Elastic net algorithm; Grouping effect; LARS algorithm; Ridge regression; Variable selection

1. Introduction and motivation

Regularization and variable selection are traditional statistical problems that attract much attention recently. Large demands are from the analysis of high dimensional massive data, for example, gene expression microarray data and single nucleotide polymorphism (SNP) data. A special feature of the genomic data is that genes sharing a common pathway or having a similar biological function tend to have high pairwise correlations (Segal et al., 2003). It would be desirable for a variable selection method to form those genes into a group and select the whole group for an appropriate data analysis.

Consider the usual linear regression model with p predictors $\mathbf{x}_1, \ldots, \mathbf{x}_p$ and a response $\mathbf{y}$. Best subset selection gives a global optimum unbiased model estimate but is computation intensive and lack of stability (Breiman, 1996). The least absolute shrinkage and selection operator (Lasso) proposed by Tibshirani (1996) is an ad-hoc method used widely. Lasso imposes a

L_1 norm bound on the regression coefficients while minimizes the residual sum of squares, so that it gives continuous and sparse estimates. However, for a highly correlated gene group, Lasso tends to select only one gene from the group instead of the whole group. Zou and Hastie (2005) propose the elastic net method by combining L_2 and L_1 penalties on the regression coefficients. Elastic net aims to achieve the grouping effect that highly correlated variables will be in or out of the model together. It works well when the absolute values of pairwise correlations are extremely high (close to 1) among the group. Elastic net outperforms Lasso in terms of prediction error on correlated data in many circumstances. However, elastic net does not reveal the underlying group structure in its solution and may perform poorly for variable groups with moderate pairwise correlations.

Alternatively, for data with interdependent structures, a naive approach is a two-step procedure: first cluster highly correlated variables into groups then select among the groups. Park et al. (2006) apply hierarchical clustering on gene expression data in the first step. For each cluster (group), they then average the genes and take the average as a supergene to fit a regression model. Nevertheless, results of hierarchical clustering highly depend on group size and correlation threshold. Averaging genes into a supergene would also increase bias of the estimates, which affects the variable selection substantially. Another method is supervised group Lasso proposed by Ma et al. (2007), which is also a two-step procedure. After dividing genes into clusters using the K-means method, they first identify important genes within each group by Lasso then select important clusters using the group Lasso (Yuan and Lin, 2006).

Efron et al. (2004) propose a less greedy forward variable selection algorithm - LARS. With a slight modification, LARS gives a complete Lasso solution path. Its fast computational speed makes Lasso widely applied in many areas. Since LARS and Lasso are designed to select individual variables, Yuan and Lin (2006) develop group LARS and group Lasso, which extend LARS and Lasso to select groups of variables. They have shown that those extensions have superior performance to the traditional stepwise method. However, group LARS (group Lasso) requires the underlying group structure information in advance. Hence the method is in fact the second step of the aforementioned two-step procedure.

Inspired by Yuan and Lin (2006), we propose a new group selection algorithm - gLars. gLars does not need to define any group structure beforehand. It does grouping and selection at the same time. Similar to LARS and group Lars, gLars also has a computational advantage by providing a piecewise linear solution path. To select the final model on the solution path of gLars, we use 10th-fold cross validation. Since LARS and gLars move along ordinary least square (OLS) direction at each step, it may suffer the shortage of OLS. When the correlations between the predictors are extremely high and noise level is high, the variance of the coefficient estimates of gLars may be large. We propose gRidge algorithm by a small change of gLars algorithm. gRidge shares all the good properties of gLars but reduces variance of the gLars estimates. Simulations and a real example show that gRidge and gLars works well comparing with LARS and elastic net.

In the following sections, we present our algorithms in more details. We introduce gLars and gRidge algorithms in Section 2 and 3 respectively. Section 4 is on the selection of tuning parameters. We illustrate our methods with several simulations in Section 5 and a real example in Section 6. A summary and discussions are given in Section 7.

2. The algorithm of gLars

Consider the usual linear regression model $\mathbf{y} = \mathbf{X}\beta + \varepsilon$, where $\mathbf{y} = (\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n)^T$ is the response variable, $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_p)$ is the predictor matrix, and ε is a vector of independent and identically distributed random errors with mean 0 and variance σ^2. There are n observations and p predictors. Each $\mathbf{x}_j = (x_{1j}, x_{2j}, \ldots, x_{nj})^T, j = 1, \ldots p$ is the column vector of the predictor matrix. We center the response variable and standardize all the column vectors of the predictor matrix. Hence, there is no intercept in our model.

$$\sum_{i=1}^{n} y_i = 0, \qquad \sum_{i=1}^{n} x_{ij} = 0, \qquad \sum_{i=1}^{n} x_{ij}^2 = 1 \qquad \text{for } j = 1, 2, \ldots, p$$

Group LARS algorithm selects a group of predictors based on pre-specified grouping information. In many practical situations where grouping information is unknown, it is difficult to implement Group LARS. Our proposed gLars avoids this difficulty by forming groups simultaneously along the variable selection process according to certain criterions.

2.1 *Grouping definition*

Zou and Hastie (2005) propose elastic net, which encourages highly correlated predictors in or out of the model together. When applying to the gene selection problem, they claim that once one gene among a group is selected, the whole group would automatically be included into the model (group selection). Furthermore, the grouping effect exhibits if the regression coefficients of a group of highly correlated variables tend to be equal (up to a change of sign). These discussions give an idea of grouping effect. However, it is unclear how groups are defined. In our procedure, we give an explicit definition of group. Predictors form a group if they satisfy both of the two criterions:

- they are highly correlated with the response variable (or current residual);
- they are highly correlated with each other.

The correlation threshold (in absolute value) for Criterion 1 is suggested to be the greater than 75th percentile of all correlations for a large data set or greater than 50th percentile of all correlations for a small data set. The correlation threshold (absolute value) for Criterion 2 is from a set of grids, for example $0.9, 0.8, 0.7, 0.6$. We select an optimal one by cross validation. If we set up the correlation threshold of a group in Criterion 2 to be 1, then gLars degenerates to the LARS algorithm.

2.2 *gLars algorithm*

The LARS algorithm proposed by Efron et al. (2004) is a less greedy forward model selection procedure. At the beginning of LARS, a predictor enters the model if its absolute correlation with the response is the largest one among all the predictors. The coefficient of this predictor grows in its ordinary least square direction until another predictor has the same correlation with the current residual (i.e. equal-angle). Next, both coefficients of the two picked predictors begin to move along their ordinary least square directions until a third predictor has the same correlation with the current residual as the first two. The whole process continues until all the predictors enter the model. In each step, one variable adds into the model and the regularization solution paths are extended in a piecewise linear way. After all the variables enter the model, the whole LARS solution paths are completed.

In the gLars algorithm, we start as LARS to pick up a predictor which has the largest correlation with the response. We call this predictor a "leader element". We then build a group based on this leader element and the current residual according to the criterions in Section 2.1. Note that both criterions have to be satisfied when selecting a variable into a group. Once a group has been constructed, it will be represented by a unique direction in R^n as the linear combination of the ordinary least square directions of all variables in the group. Next, we choose another leader element, analogous to the equal-angle requirement of the LARS algorithm. A new group is formed again following the grouping definition. We refine the solution paths in a piecewise linear format. The whole process continues until all the predictors enter the model. The detailed algorithm can be found in Zeng (2008).

3. The algorithm of gRidge

Ordinary least square often does poorly when the correlations among the predictors are very high and the noise level is high. Since both LARS and gLars move towards ordinary least square direction in each step, they face the same shortage. Ridge estimates, on the other hand, perform better in this situation (Hoerl and Kennard (1970)). We propose a gRidge algorithm, which moves towards ridge estimates direction in each step.

The relationship between ridge estimates $\hat{\beta}(\lambda)$ and ordinary least square estimates $\hat{\beta}$ can be shown as

$$\hat{\beta}(\lambda) = (X'X + \lambda I)^{-1}X'Y = (I + \lambda(X'X)^{-1})^{-1}\hat{\beta} = C\hat{\beta},$$

where $C = (I + \lambda(X'X)^{-1})^{-1}$ and λ is the ridge parameter. The gRidge algorithm is thus a simple modification of the gLars algorithm. When a group is constructed, gRidge represents the group by a unique direction from the linear combination of the ridge directions of all variables in the group. The variable coefficients are moving towards the ridge directions.

As we run simulations, we notice that gRidge outperforms other methods in terms of relative prediction errors (RPEs, defined in Section 5). However, this method is limited by its comparably larger false positives due to an over-grouping effect.

We propose to add a hard threshold δ to gRidge estimates so that small (but nonzero) coefficients will be removed. Based on simulations, we define a threshold $\delta = \sqrt{\sigma \log(p)/n}$. Thus smaller error terms, smaller number of predictors, or larger sample size give smaller threshold. We name the modified gRidge algorithm gRidge_new, after this hard threshold filtering. Simulation studies show that gRidge_new not only preserves low RPE but also largely reduces the false positive rate.

4. Choice of tuning parameters

Both gLars and gRidge produce the entire piecewise linear solution paths as group LARS does. Groups of variables are selected when we stop the paths after a certain number of steps. The number of step k is the tuning parameter. Equivalently, we may use a tuning parameter as the fraction of the L_1 norm of the coefficients

$$s = \Sigma_{j\_selected} ||\beta_j||_{L_1} / \Sigma_j ||\beta_j||_{L_1}.$$

For gLars (similar to LARS), s (or k) is the only tuning parameter. It is determined by standard five-fold cross-validation (CV). For gRidge, there are two tuning parameters: the ridge parameter λ in addition to s (or k). Similar to elastic net, we cross-validate on two-dimension. First, we choose a grid for λ, say (0.01,0.1,1,10,100,1000). Then for each λ, gRidge produces the entire solution path. The parameter s (or k) is selected by five-fold CV. At the end, we choose the λ value which gives the smallest CV error.

5. Simulation Studies

Simulation studies are used to compare the proposed gLars and gRidge with ordinary least squares, ridge regression, LARS and elastic net. The simulated data are generated from the true model

$$\mathbf{y} = \mathbf{X}\beta + \sigma\varepsilon, \qquad \varepsilon \sim N(0, 1).$$

Five examples are presented for 5 scenarios below. For each example, we simulate 100 data sets. Each data set consists of a training set and a test set. The tuning parameters are selected on training set by five-fold cross validation. Then the models are fitted after selecting variables by a method. The variable selection methods are compared in terms of relative prediction error (RPE) (Zou (2006)) on the test set. The relative prediction error is defined by

$$RPE = \frac{1}{\sigma^2}(\hat{\beta} - \beta)^T V (\hat{\beta} - \beta),$$

where V is the population covariance matrix of $\mathbf{X}$.

The 5 scenarios are given by:

1 Example 1 (adopted from Zou and Hastie (2005)), there are 100 and 200 observations in the training and test set

respectively. The true parameter $\beta = (3, 1.5, 0, 0, 2, 0, 0, 0)$ and $\sigma = 3$. The pairwise correlation between $\mathbf{x}_i$ and $\mathbf{x}_j$ was set to be $corr(\mathbf{x}_i, \mathbf{x}_j) = 0.5^{|i-j|}$. This example creates a sparse model with a few large effects and covariates with first order autoregressive correlation structure.

2 Example 2 (adopted from Zou and Hastie (2005)) is the same as example 1, except that $\beta_j = 0.85$ for all j, which creates a non-sparse underlying model with many small effects.

3 Example 3, we simulate 100 and 400 observations in the training and test set respectively. We set the true parameters as

$$\beta = (\underbrace{3, \ldots, 3}_{15}, \underbrace{1.5, \ldots, 1.5}_{5}, \underbrace{0, \ldots, 0}_{20})$$

and $\sigma = 6$. The predictors were generated as:

$$\mathbf{x}_i = Z + \varepsilon_i^x,\ Z \sim N(0,1),\ i = 1, \ldots, 15,$$

$$\mathbf{x}_i \sim N(0,1), i.i.d.,\ i = 16, \ldots, 40,$$

where ε_i^x are independent identically distributed $N(0, 0.01)$, $i = 1, \ldots, 15$. This example creates one group from the first 15 highly correlated covariates. The next 5 covariates are independent but provide signals on the response variable.

4 Example 4 (adopted from Zou and Hastie (2005)), we simulate 100 and 400 observations in the training and test set respectively. We set the true parameters as

$$\beta = (\underbrace{3, \ldots, 3}_{15}, \underbrace{0, \ldots, 0}_{25})$$

and $\sigma = 15$. The predictors were generated as:

$$\begin{aligned}
\mathbf{x}_i &= Z_1 + \varepsilon_i^x, & Z_1 \sim N(0,1), \quad & i = 1, \ldots, 5,\\
\mathbf{x}_i &= Z_2 + \varepsilon_i^x, & Z_2 \sim N(0,1), \quad & i = 6, \ldots, 10,\\
\mathbf{x}_i &= Z_3 + \varepsilon_i^x, & Z_3 \sim N(0,1), \quad & i = 11, \ldots, 15,\\
\mathbf{x}_i &\sim N(0,1), i.i.d., & & i = 16, \ldots, 40,
\end{aligned}$$

where ε_i^x are independent identically distributed $N(0, 0.01)$, $i = 1, \ldots, 15$. There are three equally important groups with 5 members in each. There are also 25 pure noise variables.

5 Example 5, we simulate 100 and 200 observations in the training and test set respectively. We set the true parameters as

$$\beta = (\underbrace{3, 3, 3, 0, 0}_{5}, \underbrace{3, 3, 3, 0, 0}_{5}, \underbrace{3, 3, 3, 0, 0}_{5}, \underbrace{0, \ldots, 0}_{25})$$

The predictors and the error term are the same as Example 4. There are also three equally important groups with 5 members in each of them. However, in each group, there are 2 noise variables, which have no effect on the response variable but are highly related with the other three important variables. There are totally 31 pure noise variables.

Table 1 and 2 summarize the prediction results. The popular Lars performs poorly in almost all examples. As the number of covariates increases, RPE of Lars increases dramatically in example 3-5, meanwhile, its estimators are highly unsteady with large standard deviations. The simulation results indicate that Lars performs badly under collinearity. Elastic net outperforms Lars in all examples in terms of prediction error, and identifies true signals with very few false positives. But Elastic net method cannot find all true signals in Example 3 (the five true signals with correspondence covariates are independently identically distributed). Our first proposed method gLars is slightly worse than Elastic net in terms of prediction error while producing correct sparse solutions as Elastic net. Different from elastic net, gLars can identify all true signals in Example 4. The gRidge improves gLars in all examples in terms of prediction error. Its RPEs are either the smallest or the second smallest across all methods. Especially as covariates increase, RPEs of gRidge are always the best. The reductions of RPEs in example 3-5 are 80.9%, 92.7% and 91.6% respectively compared with elastic net. It indicates that gRidge has unbiasedness (Fan and Li, 2001) in the long run. We also notice that while preserving the large coefficients close to the true coefficients, gRidge tends to select more variables than elastic net, owing to its over grouping effect. After we add a hard threshold to gRidge, the new gRidge_new estimators achieves the ideal performance.

6. Prostate cancer example

The prostate cancer data come from a study by Stamey et al. (1989). There are 97 observations collected from men who were about to receive a radical prostatectomy. We want to discover the relationship between the level of prostate specific antigen and several clinical endpoints. Tibshirani (1996) fits a linear model by Lasso to reveal this relationship. The response variable is log(prostate specific antigen) (lpsa). Those clinical endpoints are x_1 log(cancer volume) (lcavol), x_2 log(prostate weight) (lweight), x_3 age, x_4 log(benign prostatic hyperplasia amount) (lbph), x_5 seminal vesicle invasion (svi), x_6 log(capsular penetration) (lcp), x_7 Gleason score (gleason) and x_8 percentage Gleason scores 4 or 5 (pgg45).

We randomly split data into two parts: a training set with 67 observations and a test set with 30 observations. Model fitting and tuning parameter selection by 10th-fold cross validation were carried out on training data. The prediction error were then calculated on test data to compare model performance.

All covariates were centered and standardized to have mean 0 and standard deviation 1. There is certain correlation presented between variables. For example, the pairwise coefficients between x_7 gleason and x_8 pgg45 is 0.752 and 0.675 between x_1 lcavol and x_6 lcp and so on. Those indicates a moderate collinearity among predictors.

Table 4 shows the coefficients estimates for ordinary least square(OLS), Lars, Elastic Net, gLars and gRidge after tenth-fold cross validation. Fig. 1 gives the solution paths of Lars, Elastic Net, gLars and gRidge. And the order of covariates which enter to the model is given in Table 5. All those methods suggest that covariates log(cancer volume) (lcavol), log(prostate weight)

Table 1

Median relative prediction errors (RPE) for 6 simulated examples based on 100 replications. The numbers in parentheses are the corresponding standard errors of RPEs estimated by using the bootstrap with $B = 1000$ resampling on the 100 RPEs.

Methods	Example 1	Example 2	Example 3	Example 4	Example 5
OLS	0.5843	0.5696	0.6368	0.6390	0.6458
	(0.0547)	(0.0502)	(0.0267)	(0.0240)	(0.0265)
Ridge	0.2832	0.1884	0.2519	0.0993	0.1971
	(0.0194)	(0.0150)	(0.0100)	(0.0050)	(0.0087)
LARS	0.4640	0.4529	9.3793	3.3287	7.9525
	(0.0497)	(0.0460)	(0.4450)	(0.1321)	(0.3168)
Elastic net	0.1714	0.1481	0.5884	0.2096	0.3308
	(0.0130)	(0.0155)	(0.0546)	(0.0377)	(0.0610)
gLars	0.2616	0.3158	0.1917	0.2363	0.4800
	(0.0336)	(0.0359)	(0.0535)	(0.0518)	(0.1057)
gRidge	0.1806	0.2301	0.1125	0.0152	0.0277
	(0.0298)	(0.0254)	(0.0068)	(0.0037)	(0.0047)
gRidge_new	0.1816	0.2407	0.1113	0.0141	0.0267
	(0.0297)	(0.0255)	(0.0083)	(0.0039)	(0.0049)

(lweight), seminal vesicle invasion (svi) and Gleason score(gleason) are important in explaining the level of prostate specific antigen. Both Elastic net and LARS also choose age and in their final model. However, gLars and gRidge choose log(capsular penetration) (lcp) and Gleason scores 4 or 5 (pgg45)instead. This situation is due to the group effect. We notice that the correlation of gleason and pgg45 is 0.752 *i.e.* they form a small group. As all 4 methods picks gleason in their final model, it is better to include pgg45 in the model too. The similar case happens for covariates lcavol and lcp too. We think those two form another group. The test prediction errors for the 4 methods on the test set are reported in Table 3. It is clearly that gLars, gRidge and gRidge_new perform better than Lars and Elastic net methods in terms of prediction error.

7. Summary and discussion

Group LARS(Yuan and Lin, 2006) is a natural extension of LARS. It is successful in selecting groups of variables once the grouping structure is known. gLar aims to select groups of predictors even though the underlying grouping structure is unknown in advance. The performance of gLar is competitive with elastic net. However, because of its dependence on the

Table 2

Median number of nonzero coefficients / median number of zero coefficients misspecified as nonzero coefficients of 6 simulations based on 100 replications

Methods	Example 1	Example 2	Example 3	Example 4	Example 5
OLS	3/5	8/0	20/20	15/25	9/31
Ridge	3/5	8/0	20/20	15/25	9/31
LARS	3/0	7/0	9/0	4/0	5/1
Elastic net	3/1	8/0	15/0	15/0	9/6
gLars	3/2	8/0	20/0	15/0	9/6
gRidge	3/3	8/0	20/12.5	15/8	9/11.5
gRidge_new	3/1	7/0	20/0	15/0	9/6

Table 3

The test prediction error of models selected by OLS, Lars, Elastic Net, gLars and gRidge

Methods	OLS	LARS	Elastic net	gLars	gRidge	gRidge_new
Prediction error	0.688	0.586	0.610	0.562	0.550	0.547

Table 4

Coefficients estimator for OLS, Lars, Elastic Net, gLars and gRidge methods

Coefficents	OLS	Lars	Elastic Net	gLars	gRidge	gRidge_new
Intercept	1.92	1.92	1.92	1.92	1.92	1.92
lcavol	0.562	0.481	0.487	0.484	0.464	0.464
lweight	0.426	0.358	0.358	0.299	0.293	0.293
age	-0.214	-0.076	-0.108	0	0	0
lbph	0.126	0.069	0.087	0	0.009	0
svi	0.444	0.338	0.354	0.339	0.311	0.311
lcp	-0.165	0	0	-0.113	-0.055	-0.055
gleason	0.226	0.158	0.174	0.141	0.123	0.123
pgg45	0.04	0	0	0.011	0.015	0

Table 5

The order of covariates enter the model

Lars	lcavol→ svi→ lweight→ gleason→ lbph→ age→ lcp→ pgg45
Elastic net	lcavol→ svi→ gleason → lcp → lweight → pgg45→ lbph→ age
gLar	lcavol, lcp→ svi→ lweight → gleason, pgg45 → lbph → age
gRidge	lcavol, lcp→ svi→ lweight→ gleason, pgg45 → lbph → age

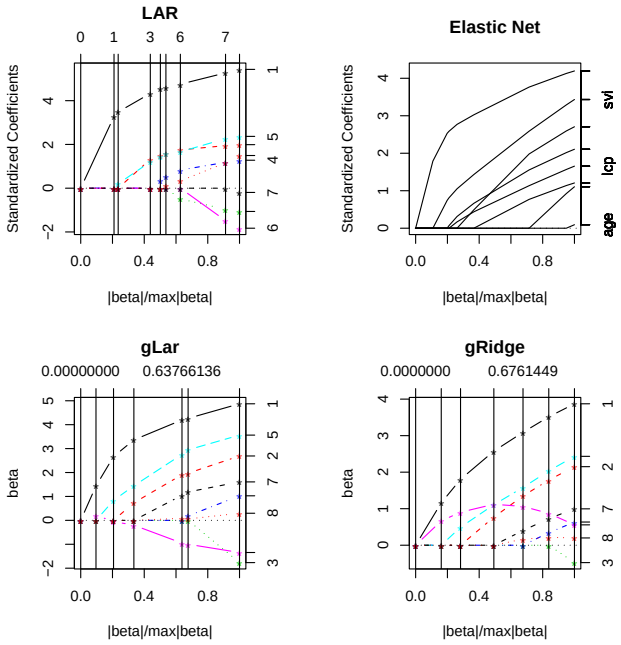

Figure 1. Solution paths of LARS, Elastic net, gLars and gRidge.

full least squares estimates, the gLars estimates may vary when the noise level is high. We develop an extended method gRidge which overcomes this shortage. gRidge performs well when the number of predictors is large, and the predictor matrix is close to singular. In some cases, gRidge tends to select more variables due to its over-grouping effect. We further add a hard threshold to the gRidge solutions to make the estimates more sparse. gLars and gRidge are computationally as fast as LARS. But solutions of both methods would depend on the two grouping criterions. gLars and gRidge algorithm can be easily extended to generalized linear models. Our algorithms offer new tools of variable selection for data with complicated interdependent structures.

References

Breiman, L.(1996).Heuristics of instability and stabilization in model selection. *Annals of Statistics*. **24,** 2350-2383.

Efron B., Johnstone I., Hastie T. and Tibshirani R. (2004). Least angle regression. *Annals of Statistics*. **32,** 407-499.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Am. Statist. Ass.*. **96,** 1348-1360.

Hoerl, A. and Kennard, R. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, **12,** 55-67.

Hoerl, A. and Kennard, R. (1970). Ridge regression: Applications to nonorthogonal problems. *Technometrics*, **12,** 69-82.

Ma, S., Song, X., and Huang, J. (2007). Supervised group Lasso with applications to microarray data analysis. *BMC Bioinformatics*, **8,** 60.

Park, M., Hastie, T., and Tibshirani, R. (2006). Averaged gene expression for regression. *Biostatistics*, **0,** 1-16.

Park, M., Hastie, T. (2006). Regularization Path Algorithm for Generalized Linear Models. *Technical Report* .

Segal, M., Dahlquist, K. and Conklin, B. (2003). Regression approach for microarray data analysis. *J. Comp. Biol.*, **10,** 961-980.

Stamey, T., Kabalin, J., McNeal J., Johnstone, I., Freiha, F., Redwine, E. and Young, N. (1989). Prostate specific antigen in the diagnosis and treatment of adenocarcinoma of the prostate, ii: Radical prostatectomy treated patients. *J. Urol.*, **16,** 1076-1083.

Tibshirani R. (1996). Regression shrinkage and selection via the lasso. *J. Royal. Statist. Soc B.* **58,** 267-288.

Yuan, M. and Lin, Y. (2006).Model selection and estimation in regression with grouped variables. *J. R. Statist. Soc. B*, **68,** 49-67.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *J. R. Statist. Soc. B.* **67,** 301-320.

Zou, H. (2006). The adpative lasso and its oracle properties. *J. R. Statist. Soc. B.* **101,** 1418-1429.

Zeng, L.(2008). Model selection on correlated data. *Unpublished Ph.D. thesis*, University of Purdue, Dept. of Statistics.