Shrinkage and Model Selection with Correlated Variables via Weighted Fusion

by

Z.J. Daye and X.J. Jeng
Purdue University

Technical Report #07-14

Department of Statistics
Purdue University

December 2007

Shrinkage and Model Selection with Correlated Variables via Weighted Fusion

Zhongyin John Daye and Xinge Jessie Jeng *
Purdue University

December 14, 2007

Abstract

Summary. We propose the weighted fusion, a new penalized regression and variable selection method for data with correlated variables. The weighted fusion can incorporate information redundancy among correlated variables for estimation and variable selection. Weighted fusion is also useful when the number of predictors p is much larger than the number of observations n. It allows the selection of more than n variables in a new way. Studies show that weighted fusion often outperforms lasso and elastic net.

Keywords: Elastic net; Lasso; Multicollinearity; $p \gg n$ problem; Regression; Variable Selection

1 Introduction

We consider the problem of variable selection and estimation for the linear regression model,

$$\mathbf{y} = \mathbf{X}\beta^* + \epsilon, \tag{1}$$

where $\mathbf{y}$ is an n-dimensional vector of random responses, $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_p)$ an $n \times p$ design matrix, $\beta^* = (\beta_1^*, \beta_2^*, \ldots, \beta_p^*)$ a vector of regression parameters, and ϵ an n-dimensional vector of independent and identically distributed (i.i.d.) random variables with mean 0 and variance σ^2. In this paper, we assume the response to be

* *Address for correspondence:* Zhongyin John Daye (E-mail: zdaye@stat.purdue.edu) and Xinge Jessie Jeng (Xinge Zheng) (E-mail: zheng4@stat.purdue.edu), Department of Statistics, Purdue University, 250 N. University Street, West Lafayette, IN 47907.

centered and each predictor $\mathbf{x}_j = (x_{1j}, x_{2j}, \ldots, x_{nj})$ normalized so that $\sum_{i=1}^{n} x_{ij} = 0$ and $\|\mathbf{x}_j\|^2 = 1$.

Variable selection and estimation in high-dimensionality have become an integral topic in modern Statistics. This is largely driven by the availability of massive data due to recent technological advancements (Fan and Li, 2006; Yu, 2007). For instance, bioengineering innovations have presented new statistical challenges by introducing functional MRI and gene microarray data. In many of these applications, we wish to achieve prediction accuracy and variable reduction. Due to its simplicity, interpretability, and computational efficiency, linear models have become the prevailing choice for high-dimensional data analysis.

It is well-known that ordinary least squares (OLS) performs poorly both in prediction and variable selection. Hoerl and Kennard (1970a; 1970b), in proposing the ridge, has successfully applied penalized regression in improving prediction accuracy. Recently, Tibshirani (1996), in proposing the lasso, has put forth penalized regression as a viable approach towards variable selection. The related technique of soft thresholding for wavelet approximation was proposed by Donoho and Johnstone (1994) and Donoho et al. (1995).

Regression with correlated variables presents a challenging problem for variable selection. The reason is that under collinearity the data has little information to distinguish the effect of one variable versus another (Harrell, 2001; Mosteller and Tukey, 1977). This often leads to arbitrary selection and under-representation of important variables.

Under high-dimensionality, the situation is particularly dire. In a seminal paper, Zou and Hastie (2005) proposed the elastic net for variable selection with highly correlated variables. The elastic net,

$$\hat{\beta}(\text{EN}) = (1 + \lambda_2) \{\arg\min_{\beta} \|\mathbf{y} - \mathbf{X}\beta\|^2 + \lambda_1 \|\beta\|_1 + \lambda_2 \|\beta\|^2\}, \tag{2}$$

utilizes the lasso $\sum_{j=1}^{p} |\beta_j|$ and ridge $\sum_{j=1}^{p} \beta_j^2$ penalty to enable group selection. When $p > n$, the ridge penalty further allowed elastic net to select more than n variables, whereas lasso regression selects at most n variables before it saturates. The multiplier $(1 + \lambda_2)$ mitigates the double shrinkage imposed by using both the lasso and ridge penalty to penalize β towards zero. The naïve elastic net is obtained by removing the multiplier.

The elastic net has the interpretation of a stabilized version of the lasso (Theorem 2 in Zou and Hastie (2005)), represented as

$$\hat{\beta}(\text{EN}) = \arg\min_{\beta} \beta^T \left(\frac{\mathbf{X}^T\mathbf{X} + \lambda_2 \mathbf{I}}{1 + \lambda_2}\right)\beta - 2\mathbf{y}^T\mathbf{X}\beta + \lambda_1 \|\beta\|_1. \tag{3}$$

We note that when $\lambda_2 \to 0$, the elastic net is equivalent to lasso, and when

$\lambda_2 \to \infty$, the elastic net is equivalent to univariate soft thresholding (UST), $\mathbf{u}_j = \text{sgn}(\mathbf{y}^T\mathbf{x}_j)(|\mathbf{y}^T\mathbf{x}_j| - \lambda_1/2)_+$.

The elastic net penalty has shown improvements over lasso in many situations (Wang et al., 2006). However, we observe that it has several important limitations.

1. The elastic net solution becomes strongly biased as it approaches that of UST. In fact, univariate estimators are usually strongly biased except when predictors are nearly independent.

2. The elastic net exhibits poor performance in selecting related variables as a group when within-group correlations are non-extreme.

3. The elastic net penalty does not explicitly contain correlation.

Apparently, biasedness deteriorates prediction accuracy. It also results in poor selection when used with techniques, such as cross-validation, etc., that depend upon reasonable validation error for robust performance (Fan and Li, 2001). Moreover, in many applications, the within group correlations are not extreme, $|\rho| \approx 0.85$. The elastic net, as implied by its relation to the univariate estimate and Theorem 1 of Zou and Hastie (2005), may have poor grouping effect when correlations are not very close to ± 1. Furthermore, as its penalty does not explicitly contain correlation, elastic net cannot incorporate correlation prior knowledge nor handle complex correlation structures.

On consideration of the limitations of elastic net mentioned above, we believe that a new approach is needed for estimation and variable selection under correlated variables.

In this paper, we present the *weighted fusion*, a new penalized regression method that addresses the limitations of elastic net. Recall that the problem of regression under collinearity is that predictors containing similar information are treated as though they are distinct. The weighted fusion penalty, proposed herein, fuses or clusters coefficients of related variables via correlation-driven weights which induces correlated predictors to be treated similarly in regression. The weighted fusion penalty further solves the $p > n$ dilemma by allowing the selection of more than n variables in a new way. Both simulation and real data examples show that weighted fusion often outperforms lasso and elastic net in prediction and variable selection.

In Section 2, we define the weighted fusion estimator and discuss the grouping effect induced by weighted fusion. In Section 3, we propose the *generalized ridge-lasso estimator* (GRIL) that generalizes both the elastic net and weighted fusion. We present variable selection consistency results and a standard error formula for GRIL. Section 4 discusses computational strategies for weighted fusion and analyzes the effect of the weighted fusion penalty on solution path. Solution paths for the

prostrate cancer data are used to demonstrate our method in Section 5. In Section 6, we compare the performance of weighted fusion with lasso and elastic net.

2 Weighted fusion

2.1 Definition

We define the weighted fusion estimate as,

$$\hat{\beta}(\lambda_1, \lambda_2) = \arg\min_{\beta} \left\{ ||\mathbf{y} - \mathbf{X}\beta||^2 + \lambda_1 \sum_{j=1}^{p} |\beta_j| + \lambda_2 J(\beta) \right\}, \tag{4}$$

where $\lambda_1 \geq 0$, $\lambda_2 \geq 0$ are tuning parameters, and J a penalty function applied to the pairwise signed differences of regression coefficients. We call the difference, $\beta_j - s_{ij}\beta_j$ where $s_{ij} \in \{-1, 0, 1\}$, signed difference.

Let $\rho_{ij} = \mathbf{x}_i^T \mathbf{x}_j$ be the sample correlation between predictor variables, $s_{ij} = \text{sgn}(\rho_{ij})$, and $w_{ij} \geq 0$ are predictors correlation driven weights that will be defined in Section 2.2. The weighted fusion employs the penalty function,

$$J(\beta) = \frac{1}{p} \sum_{i<j} w_{ij} (\beta_i - s_{ij}\beta_j)^2. \tag{5}$$

Notice that when $w_{ij} = 0$ for all i, j we obtain the lasso estimate. When $w_{ij} = 1_{\{j-i=1\}}$ and $\rho_{ij} > 0$ for all i, j, we have the $L2$-norm fused lasso estimate (Tibshirani et al., 2005).

The weighted fusion formulation penalizes for the pairwise signed differences of coefficients according to w_{ij}. Its penalty contour is shown in Figure 1 for the case when $p = 2$. We note, in particular, that the predictors correlation driven weights w_{ij} impose only tendencies and not constraints on the signed-difference fusion. Predictors correlations express themselves in statistical models in complex ways and are difficult to determine beforehand. Though intuition deems that stronger correlations may result in stronger grouping effect, this is not a guarantee, and a coefficients pair with a weaker correlation may express itself with a stronger grouping effect than one with stronger predictors correlation.

Thus, we must balance our belief in the coefficients structure, that is formed on the basis of predictors, with that of the loss $\ell(\mathbf{y}, \mathbf{X}\beta) = \|\mathbf{y} - \mathbf{X}\beta\|^2$. In a sense, weighted fusion (4) discovers predictors correlation based grouping effect under the supervision of the response and predictors joint distribution.

Consider the weighted fusion estimator (4) with $\lambda_1 = 0$. This results in,

$$\hat{\beta}^{RF}(\lambda_2) = (\mathbf{X}^T\mathbf{X} + \frac{\lambda_2}{p}\mathbf{W})^{-1}\mathbf{X}^T\mathbf{y}, \tag{6}$$

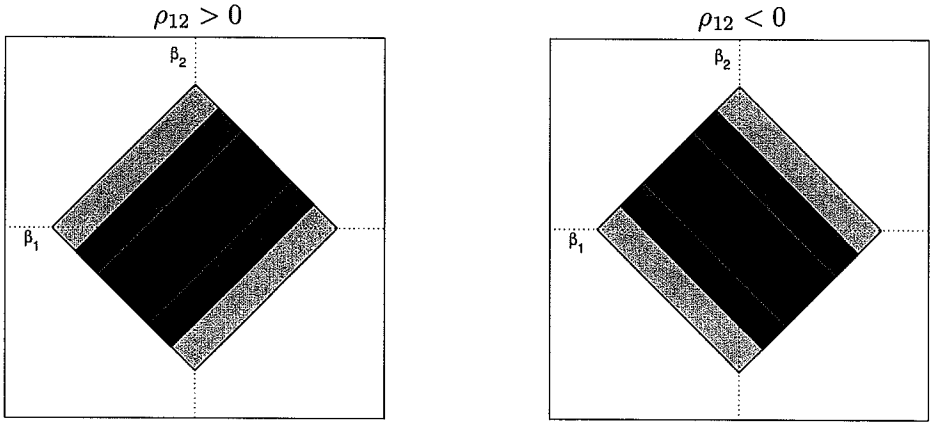

Figure 1: Weighted fusion contour.

where,

$$\mathbf{W} = \begin{pmatrix} \sum_{j\neq 1} w_{1j} & -s_{12}w_{12} & \dots & -s_{1p}w_{1p} \\ & \sum_{j\neq 2} w_{2j} & \cdot & \vdots \\ & & \ddots & -s_{p-1,p}w_{p-1,p} \\ & & & \sum_{j\neq p} w_{pj} \end{pmatrix}. \tag{7}$$

The above formulation is of interest on its own. We call it the *ridge fusion* estimator. If the identity matrix $\mathbf{I}$ is used in place of $\mathbf{W}/p$ in (6), we have the ridge estimator. When $p > n$, the covariance matrix $\mathbf{X}^T\mathbf{X}$ is singular. Ridge regression increases the rank of $\mathbf{X}^T\mathbf{X}$ by adding a constant λ_2 to each eigenvalue. This allows elastic net to handle the $p > n$ situation. Ridge fusion, on the other hand, increases the rank of $\mathbf{X}^T\mathbf{X}$ by adding a constant factor of the matrix $\mathbf{W}$. We note that ridge employs a simple but not well-motivated strategy, whereas ridge fusion has the interpretation of increasing the rank of $\mathbf{X}^T\mathbf{X}$ by drawing correlated coefficients closer together in magnitude. This allowed weighted fusion to resolve the $p > n$ situation in a motivated way. In Section 6, we will show that ridge fusion may sometimes outperform the ridge estimator.

Penalty function (5) has an alternative formulation, $(1/p)\sum_{i<j} w_{ij}|\beta_i - s_{ij}\beta_j|$, using $L1$-norm. This alternative has the benefit of equating the magnitudes of coefficients of highly correlated variables. However, the $L1$-norm penalty introduces $O(p^2)$ additional sources of non-differentiability that prevents the construction of efficient algorithms. Further, experimental studies did not show significant difference between the $L1$-norm and $L2$-norm weighted fusion in prediction and variable selection for correlated variables. Therefore, in this paper, we restrict ourselves to the

$L2$-norm weighted fusion penalty (5), and we may explore the $L1$-norm alternative in a future work.

The fusion penalty function, $\sum_j |\beta_j - \beta_{j-1}|^q$ for $q = 0, 1$, was proposed by Land and Friedman (1996) for signal regression, whereas the fused lasso, that compounds the lasso penalty, was proposed by Tibshirani et al. (2005) for variable selection with ordered features. Their methods employ vanilla fusion penalty on sequential differences. Weighted fusion, on the other hand, uses the pairwise signed-difference fusion penalty function modulated by correlation-driven weights for variable selection with correlated variables.

2.2 Determining the weights

The following general properties are necessary for determining correlation driven weights $w_{ij} \in [0, \infty)$ for the weighted fusion penalty (5):

1. $w_{ij} = w_{ji}$,
2. $w_{ij} = 0$ when $\rho_{ij} = 0$,
3. w_{ij} is non-decreasing in ρ_{ij},
4. $|w_{ik} - w_{jk}| \to 0$ when $|\rho_{ij}| \to 1$ for all k.

Property 1 retains the symmetry of the correlation matrix; property 2 annuls the effects of the weighted fusion penalty when predictors are independent; property 3 imposes the belief that stronger correlation results in stronger tendency for a coefficient pair to be fused; and property 4 equates the effects of two predictors when their correlation approaches ± 1.

We propose a weight for (5) that possesses the general weight properties,

$$w_{ij} = \frac{|\rho_{ij}|^\gamma}{1 - |\rho_{ij}|}, \tag{8}$$

where $\gamma > 0$ is a tuning parameter. Moreover, the weight w_{ij} has the specific property of $w_{ij} \to \infty$ when $|\rho_{ij}| \to 1$.

We note that the weight function (8) imposes a continuous modulation on $|\rho_{ij}|$. Discrete alternatives are possible. For example, a simple alternative may threshold w_{ij} at a fixed value of ρ_{ij}, that is $w_{ij} = c1_{\{|\rho_{ij}|>t\}}$ for c, t fixed. However, discrete modulation is sensitive to errors in estimation, and continuous alternatives can usually achieve better performance.

Figure 2 plots the weight function for varying γ. We note that as γ increases $w_{ij} \approx 0$ for a broader range of $|\rho_{ij}|$. Thus, we see that the data-driven parameter γ indicates how large must the magnitude of a correlation be for its corresponding

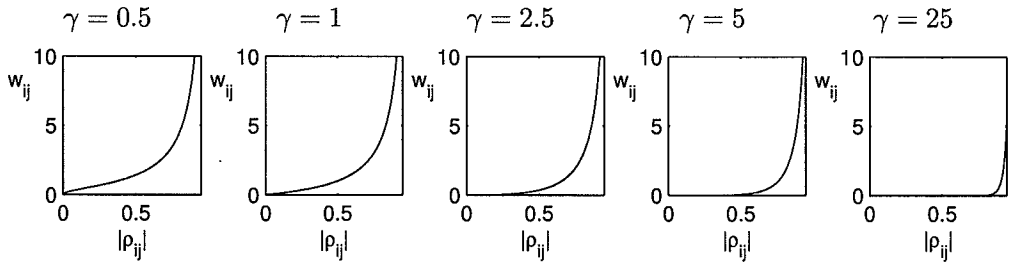

Figure 2: Continuous weight function.

coefficients pair to be fused. As correlation effects are complex, when must correlations be recognized often varies from data to data, and it is imperative to determine γ from the data at hand. We find that only a few representative values for γ need to be cross-validated to obtain good results. We use only 5 values $\{0.5, 1, 2.5, 5, 25\}$ in this paper. By using continuous modulation, the technique is not sensitive to small differences in γ. If previous experience with the data type is available, a reasonable value for γ can be chosen beforehand.

We note that the weighted fusion penalty (5) is very general and can incorporate any nonnegative weights. Prior knowledge, when available, can be used to assign weights accordingly. Robust estimates of correlation, such as Jackknife correlation, etc., can also be used to improve performance.

2.3 Grouping effect

Grouping effect is expressed when the magnitudes of regression coefficients of a group of highly correlated variables tend to be equal. Weighted fusion estimator has the natural tendency of fusing each pair of regression coefficients according to their correlations. We establish the grouping effect of weighted fusion in the following two theorems; the proofs are in Appendix 9.1 and 9.2 for Theorem 1 and 2, respectively.

Theorem 1. *Given data* $(\mathbf{y}, \mathbf{X})$, $\lambda_1 \geq 0$, $\lambda_2 \geq 0$, w_{ij} *satisfying general properties,* $\mathbf{X}$ *normalized, and* $\mathbf{y}$ *centered. Denote the weighted fusion estimator as* $\hat{\beta} = \hat{\beta}(\lambda_1, \lambda_2)$. *If additionally* $\rho_{ij} \to 1$ *implies* $w_{ij} \to \infty$, *then* $|\hat{\beta}_i - s_{ij}\hat{\beta}_j| \to 0$ *as* $\rho_{ij} \to 1$.

Remark 1: The weight function, that we proposed in (8), satisfies the requirements in Theorem 1.

Remark 2: Theorem 1 can be applied to $\hat{\beta}(\lambda_1, \lambda_2)$ with more general penalty function than (5), for example, $J^q(\beta) = \frac{1}{p}\sum_{i<j} w_{ij}(\beta_i - s_{ij}\beta_j)^q$, $q > 0$.

In real situations, correlation may not be extremely close to ± 1. For moderate correlation, we establish an upper bound for the signed difference of $\hat{\beta}(\lambda_1, \lambda_2)$ without the additional assumption: $\rho_{ij} \to 1$ implies $w_{ij} \to \infty$. This result, shown in Theorem 2, is analogous to the grouping effect result of elastic net in Zou and Hastie (2005), Theorem 1.

Theorem 2. *Given data* $(\mathbf{y}, \mathbf{X})$, $\lambda_1 > 0$, $\lambda_2 > 0$, w_{ij} *satisfying general properties,* $\mathbf{X}$ *normalized,* $\mathbf{y}$ *centered. Denote* $\hat{\beta} = \hat{\beta}(\lambda_1, \lambda_2)$. *If* $\hat{\beta}_i \hat{\beta}_j > 0$, *then*

$$|\hat{\beta}_i - \hat{\beta}_j| \leq \frac{||\mathbf{y}||\sqrt{2(1-\rho_{ij})}}{\lambda_2 \bar{w}_{i\cdot}} + \frac{||\mathbf{y}|| + \lambda_1/2}{\lambda_2} |\frac{1}{\bar{w}_{i\cdot}} - \frac{1}{\bar{w}_{j\cdot}}| + \frac{1}{p} \sum_{1 \leq k \leq p} |(\frac{w_{ik}s_{ik}}{\bar{w}_{i\cdot}} - \frac{w_{jk}s_{jk}}{\bar{w}_{j\cdot}})\hat{\beta}_k|, \tag{9}$$

where $\bar{w}_{i\cdot} = (\sum_{1 \leq k \leq p} w_{ik})/p$.

Remark 1: Comparing with Theorem 1, Theorem 2 works for any ρ_{ij} and does not assume $w_{ij} \to \infty$ for $\rho_{ij} \to 1$. But when $\rho_{ij} \to 1$, we still have $|\hat{\beta}_i - \hat{\beta}_j| \to 0$ because of the general properties and the fact $s_{ik} \approx s_{jk}$, $\forall k$. On the other hand, Theorem 2 only works for penalty function (5), whereas Theorem 1 works for $J^q(\beta)$, $q > 0$.

Remark 2: For fixed ρ_{ij}, the dominating term of the weighted fusion upper bound is

$$\frac{||\mathbf{y}||}{\lambda_2}(\frac{\sqrt{2(1-\rho_{ij})}}{\bar{w}_{i\cdot}} + |\frac{1}{\bar{w}_{i\cdot}} - \frac{1}{\bar{w}_{j\cdot}}|) \tag{10}$$

for large n. Compared with naïve elastic net, which has an upper bound of

$$\frac{||\mathbf{y}||\sqrt{2(1-\rho_{ij})}}{\lambda_2} \tag{11}$$

(Theorem 1 in Zou and Hastie (2005)), our method has potentially stronger grouping effect as large ρ_{ij} can induce large $\bar{w}_{i\cdot}$ and $\bar{w}_{j\cdot}$.

3 Generalized ridge-lasso estimator (GRIL)

In this section, we study elastic net and weighted fusion regression via a generalized formulation, which we call the generalized ridge-lasso estimator (GRIL).

$$\hat{\beta}^{\text{GRIL}}(\lambda_1, \lambda_2) = \arg\min_{\beta} ||\mathbf{y} - \mathbf{X}\beta||^2 + \lambda_1 ||\beta||_1 + \lambda_2 \beta^T \mathbf{Q} \beta, \tag{12}$$

where $\mathbf{Q}$ is a positive semi-definite matrix with Cholesky decomposition $\mathbf{Q} = \mathbf{R}^T\mathbf{R}$. Let,

$$\mathbf{y}^* = \begin{pmatrix} \mathbf{y} \\ \mathbf{0} \end{pmatrix}, \qquad \mathbf{X}^* = \begin{pmatrix} \mathbf{X} \\ \sqrt{\lambda_2}\mathbf{R} \end{pmatrix}, \tag{13}$$

then, GRIL may be defined as,

$$\hat{\beta}^{\text{GRIL}}(\lambda_1, \lambda_2) = \arg\min_{\beta} ||\mathbf{y}^* - \mathbf{X}^*\beta||^2 + \lambda_1\|\beta\|_1. \tag{14}$$

We note that GRIL becomes the naïve elastic net when $\mathbf{Q} = \mathbf{I}$ and the weighted fusion when

$$\mathbf{Q} = \frac{1}{p}\mathbf{W}, \tag{15}$$

where $\mathbf{W}$ is defined as in (7).

3.1 GRIL variable selection consistency

We study the variable selection consistency of GRIL by analyzing its sign consistency, which is a stronger property. The results here can be applied to both elastic net and weighted fusion by specifying $\mathbf{Q}$. In order to compare with the results of lasso sign consistency in Zhao and Yu (2006), $\mathbf{X}$ is not normalized in this section. Suppose $\mathbf{X}$ satisfies the regularity conditions: (1) $(\mathbf{X}^T\mathbf{X})/n = \mathbf{C}^n \to \mathbf{C}$, $\mathbf{C}$ is a positive definite matrix. (2) $\{\max_{1\le i\le n}(\mathbf{x}_i^T\mathbf{x}_i)\}/n \to 0$. Denote the true parameter vector $\beta^* = \{\beta^*_{(1)}, \beta^*_{(2)}\}^T$, where $\beta^*_{(1)} = \{\beta^*_j : \beta^*_j \neq 0\}$, and $\beta^*_{(2)} = \{\beta^*_j : \beta^*_j = 0\}$. We call a GRIL estimator $\hat{\beta}^{\text{GRIL}}(\lambda_1, \lambda_2)$ sign consistent if there exists $\lambda_1 = \lambda_1(n) = o(n)$ and $\lambda_2 = \lambda_2(n) = o(n)$ such that

$$\lim_{n\to\infty} P(\text{sgn}(\hat{\beta}^{\text{GRIL}}(\lambda_1, \lambda_2) = \text{sgn}(\beta^*)) = 1.$$

The sufficient and necessary sign consistency conditions are analogous to the Irrepresentable Conditions in Zhao and Yu (2006).

Sufficient condition for GRIL sign consistency:

$$|\mathbf{C}^n_{21}(\mathbf{C}^n_{11})^{-1}\text{sgn}(\beta^*_{(1)}) - \frac{2\lambda_2}{\lambda_1}(\mathbf{C}^n_{21}(\mathbf{C}^n_{11})^{-1}\mathbf{Q}_{11} - \mathbf{Q}_{21})\beta^*_{(1)}| \le 1 - \eta \tag{16}$$

for some $\eta > 0$, and some λ_1, λ_2 satisfying $\lambda_1/n \to 0$, $\lambda_2/n \to 0$, $\lambda_1/(n^{(1+c_1)/2}) \to \infty$, $\lambda_2/(n^{(1+c_2)/2}) \to \infty$.

Necessary condition for GRIL sign consistency:

$$|\mathbf{C}^n_{21}(\mathbf{C}^n_{11})^{-1}\text{sgn}(\beta^*_{(1)}) - \frac{2\lambda_2}{\lambda_1}(\mathbf{C}^n_{21}(\mathbf{C}^n_{11})^{-1}\mathbf{Q}_{11} - \mathbf{Q}_{21})\beta^*_{(1)}| < 1. \tag{17}$$

Theorem 3. *Given data* $(\mathbf{y}, \mathbf{X})$ *under regularity conditions on* $\mathbf{X}$, $\hat{\beta}^{GRIL}(\lambda_1, \lambda_2)$ *is sign consistent if condition (16) holds.*

Theorem 4. *Given data* $(\mathbf{y}, \mathbf{X})$ *under regularity conditions on* $\mathbf{X}$, $\hat{\beta}^{GRIL}(\lambda_1, \lambda_2)$ *is sign consistent only if condition (17) holds.*

Remark 1: In real situations, the values of λ_1 and λ_2 are determined by data. To gain some insights for the weighted fusion estimator, consider two extreme cases. First, if $\lambda_2 = o(\lambda_1)$, then Conditions (16) and (17) are the same as Irrepresentable Conditions in Zhao and Yu (2006) for lasso sign consistency. In this case, penalty $J(\beta)$ is negligible, only lasso penalty matters. On the other hand, if $\lambda_1 = o(\lambda_2)$, then necessary condition (17) is violated; wherefore, $\hat{\beta}^{\text{GRIL}}(\lambda_1, \lambda_2)$ is not sign consistent, and over-selection occurs. This agrees with our intuition, because when lasso penalty is negligible, $\hat{\beta}^{\text{GRIL}}(\lambda_1, \lambda_2)$ degenerates to $\hat{\beta}^{RF}(\lambda_2)$ in (6), and $\hat{\beta}^{RF}(\lambda_2)$ is unlikely to be sparse.

Remark 2: With $\mathbf{Q}$ chosen as the identity matrix $\mathbf{I}$, we have variable selection consistency results for elastic net. Similar results are also derived in Yuan and Lin (2007), Theorem 4.

Remark 3: With $\mathbf{Q}$ chosen as in (15), condition (16) is satisfied when $\mathbf{C}_{21}^n$ is close to $\mathbf{0}$, because that implies $\mathbf{Q}_{21}$ to be close to $\mathbf{0}$.

3.2 Standard error formula

We use the local quadratic approximation (LQA) technique in Fan and Li (2001) to obtain a standard error formula for GRIL. The LQA method has been proven to be consistent (Fan and Peng, 2004).

For a nonzero β_j, the LQA of the lasso penalty is

$$|\beta_j| \approx |\beta_{j0}| + \frac{1}{2|\beta_{j0}|}(\beta_j^2 - \beta_{j0}^2).$$

Suppose the first s elements of β are nonzero, then the GRIL estimator can be derived by iteratively computing the ridge regression

$$(\beta_1, \ldots, \beta_s)^T = (\mathbf{X}_{(1)}^T\mathbf{X}_{(1)} + \lambda_2\mathbf{Q}_{11} + \lambda_1\Sigma(\beta_0))^{-1}\tilde{\mathbf{X}}_{(1)}^T y,$$

where $\Sigma(\beta_0) = \text{diag}(1/|\beta_{10}|, \ldots, 1/|\beta_{s0}|)$, and $\tilde{\mathbf{X}}_{(1)}$ is the cholesky decomposition of $\mathbf{X}_{(1)}^T\mathbf{X}_{(1)} + \lambda_2\mathbf{Q}_{11}$. This leads to the estimated covariance matrix for nonzero coefficients of the GRIL estimator,

$$\begin{aligned}\widehat{cov}(\hat{\beta}_{(1)}) &= \sigma^2(\mathbf{X}_{(1)}^T\mathbf{X}_{(1)} + \lambda_2\mathbf{Q}_{11} + \lambda_1\Sigma(\hat{\beta}_{(1)}))^{-1} \cdot (\mathbf{X}_{(1)}^T\mathbf{X}_{(1)} + \lambda_2\mathbf{Q}_{11}) \\ &\quad \cdot(\mathbf{X}_{(1)}^T\mathbf{X}_{(1)} + \lambda_2\mathbf{Q}_{11} + \lambda_1\Sigma(\hat{\beta}_{(1)}))^{-1}. \qquad (18)\end{aligned}$$

The estimated standard errors are 0 for zero-valued coefficients (Tibshirani, 1996; Fan and Li, 2001).

4 Algorithms

4.1 Weighted fusion

In Section 3 we have introduced the GRIL estimator. The weighted fusion is a special case of GRIL with $\mathbf{Q}$ defined as in (15). One of the attractive features of GRIL is its computational efficiency. By introducing a data augmentation (13), we have reformulated the GRIL into a lasso problem (14) when λ_2 is kept constant. This allows the computation of the GRIL estimator via available efficient algorithms for lasso.

The lasso solution paths were shown to be piecewise linear by Efron et al. (2004) and Rosset and Zhu (2007). This property allows Efron et al. (2004) to propose the LARS algorithm that can compute the entire solution path of the lasso in $O(\min(n,p))$ steps with a complexity of $O(np)$ at each step. This achieves the same order of computation as a single OLS.

The GRIL data augmentation for weighted fusion increases the row dimension of the design matrix from n to $n+p$. For $p \gg n$, we usually adopt an early stopping strategy. In practice, an optimal solution is often achieved early in the LARS algorithm. If we stop the algorithm in m steps, then the weighted fusion and the GRIL estimator, in general, requires $O(mp^2)$ operations.

We further note that many alternative algorithms are available for lasso: blasso (Zhao and Yu, 2007), glasso (Kim and Kim, 2004), etc. These algorithms may also be applied for GRIL.

4.2 Effect of weighted fusion penalty on solution path

In this section, we will study the effect of the weighted fusion penalty on solution path. In particular, we will demonstrate that highly correlated variables can enter the weighted fusion solution path at close proximity.

We examine how the weighted fusion penalty is expressed in the LARS algorithm. Suppose $\mathbf{x}_i$ and $\mathbf{x}_j$ are highly correlated, we want to show that once one of them, say $\mathbf{x}_i$, is included in the active set, $\mathbf{x}_j$ will be the next. We adopt similar notations as in Efron et al. (2004), except that $\mathbf{y}$ and $\mathbf{X}$ are replaced by the reformulated $\mathbf{y}^*$ and $\mathbf{X}^*$ as in (13). Let $\mathcal{A}$ represent the active set that includes all variables already selected, and $\hat{\mu}_{\mathcal{A}}$ the LARS estimate immediately after $\mathbf{x}_i$ is included in the active set. Let

$$\hat{\mathbf{c}} = (\mathbf{X}^*)^T(\mathbf{y}^* - \hat{\mu}_{\mathcal{A}}), \qquad \hat{\mathbf{C}} = \max_{1\le k\le p}\{|\hat{c}_k|\}, \qquad \mathcal{A} = \{k : |\hat{c}_k| = \hat{\mathbf{C}}\},$$

and

$$s_k = \operatorname{sgn}(\hat{c}_k), \forall k \in \mathcal{A}, \qquad \mathbf{X}^*_{\mathcal{A}} = \begin{pmatrix} \mathbf{X}_{\mathcal{A}} \\ \sqrt{\lambda_2}\mathbf{R}_{\mathcal{A}} \end{pmatrix} = (\cdots s_k x^*_k \cdots)_{k\in\mathcal{A}}.$$

Define

$$\mathbf{G}_{\mathcal{A}} = (\mathbf{X}^*_{\mathcal{A}})^T\mathbf{X}^*_{\mathcal{A}}, \qquad \mathbf{A}_{\mathcal{A}} = (\mathbf{1}_{\mathcal{A}}^T\mathbf{G}_{\mathcal{A}}^{-1}\mathbf{1}_{\mathcal{A}})^{-1/2},$$

then $\mathbf{u}_{\mathcal{A}} = \mathbf{X}^*_{\mathcal{A}}\mathbf{A}_{\mathcal{A}}\mathbf{G}_{\mathcal{A}}^{-1}\mathbf{1}_{\mathcal{A}}$ is the unit vector making equal acute angles with columns of $\mathbf{X}^*_{\mathcal{A}}$. i.e.

$$(\mathbf{X}^*_{\mathcal{A}})^T\mathbf{u}_{\mathcal{A}} = \mathbf{A}_{\mathcal{A}}\mathbf{1}_{\mathcal{A}}, \qquad ||\mathbf{u}_{\mathcal{A}}||^2 = 1.$$

Let $\mathbf{a} = (\mathbf{X}^*)^T\mathbf{u}_{\mathcal{A}}$, then according to the LARS algorithm, the index of the next variable entering the active set is

$$\hat{k} = \arg\min_{k\in\mathcal{A}^c}\{\frac{\hat{\mathbf{C}} - \hat{c}_k}{\mathbf{A}_{\mathcal{A}} - a_k}, \frac{\hat{\mathbf{C}} + \hat{c}_k}{\mathbf{A}_{\mathcal{A}} + a_k}\}_+. \tag{19}$$

In order to have $\hat{k} = j$, it is enough to show that

$$\min\{\frac{\hat{\mathbf{C}} - \hat{c}_j}{\mathbf{A}_{\mathcal{A}} - a_j}, \frac{\hat{\mathbf{C}} + \hat{c}_j}{\mathbf{A}_{\mathcal{A}} + a_j}\}_+ < \min_{k\in\mathcal{A}^c, k\neq j}\{\frac{\hat{\mathbf{C}} - \hat{c}_k}{\mathbf{A}_{\mathcal{A}} - a_k}, \frac{\hat{\mathbf{C}} + \hat{c}_k}{\mathbf{A}_{\mathcal{A}} + a_k}\}_+$$

We analyze the numerators and denominators in (19) separately. Since $\hat{\mathbf{c}} = (\mathbf{X}^*)^T(\mathbf{y}^* - \hat{\mu}_{\mathcal{A}}) = \mathbf{X}^T(\mathbf{y} - \hat{\mu}^n_{\mathcal{A}})$, where $\hat{\mu}^n_{\mathcal{A}}$ is the first n entries of $\hat{\mu}_{\mathcal{A}}$, weighted fusion penalty does not change the order of numerators. For the denominators, let us consider $\mathbf{A}_{\mathcal{A}} - a_j$ first. The goal is to show that when $s_i > 0$ and $\rho_{ij} \to 1$, we have $\mathbf{A}_{\mathcal{A}} - a_j \to \infty$, which implies that c_j equals the maximal value of correlations in the updated active set, so $\mathbf{x}_j$ is the next variable entering the active set. Since $\mathbf{A}_{\mathcal{A}} = (s_i\mathbf{x}^*_i)^T\mathbf{u}_{\mathcal{A}}$, direct calculation shows that

$$\begin{aligned}
\mathbf{A}_{\mathcal{A}} - a_j &= (s_i\mathbf{x}^*_i - \mathbf{x}^*_j)^T\mathbf{u}_{\mathcal{A}} \\
&= \mathbf{A}_{\mathcal{A}}(s_i\mathbf{x}_i - \mathbf{x}_j)^T\frac{\mathbf{X}_{\mathcal{A}}}{\mathbf{X}_{\mathcal{A}}^T\mathbf{X}_{\mathcal{A}} + \lambda_2\mathbf{Q}_{\mathcal{A}}}\mathbf{1}_{\mathcal{A}} \\
&+ \mathbf{A}_{\mathcal{A}}\lambda_2(s_i\mathbf{R}_i - \mathbf{R}_j)^T\frac{\mathbf{R}_{\mathcal{A}}}{\mathbf{X}_{\mathcal{A}}^T\mathbf{X}_{\mathcal{A}} + \lambda_2\mathbf{Q}_{\mathcal{A}}}\mathbf{1}_{\mathcal{A}}.
\end{aligned}$$

Consider $(s_i\mathbf{R}_i - \mathbf{R}_j)^T\mathbf{R}_{\mathcal{A}}$ in the second term. Since $\mathbf{R}_i^T\mathbf{R}_j = Q_{ij}$, and $\mathbf{Q}$ is determined by (15), we have

$$(s_i\mathbf{R}_i - \mathbf{R}_j)^T\mathbf{R}_{\mathcal{A}} = (\ldots, s_i(s_iQ_{ii} - Q_{ij})) = (\ldots, \frac{\sum w_{ik}}{p} + \frac{s_is_{ij}w_{ij}}{p}).$$

When $s_i > 0$ and $\rho_{ij} \to 1$, we have $s_is_{ij}w_{ij} \to \infty$, which implies $\mathbf{A}_{\mathcal{A}} - a_j \to \infty$.

Similar argument can be applied to $\mathbf{A}_{\mathcal{A}} + a_j$, which shows that when $s_i > 0$ and $\rho_{ij} \to -1$, we have $\mathbf{A}_{\mathcal{A}} + a_j \to \infty$. This implies that $-c_j$ equals the maximal value of correlations in the updated active set, so $\mathbf{x}_j$ is the next variable entering the active set.

For the case with $s_i < 0$, the argument above with changes of signs leads to the same conclusion that $\mathbf{x}_j$ is the next selected variable.

5 Analysis of prostate cancer data

In this section, we study the solution path of weighted fusion using the prostate cancer data. The prostate cancer data is drawn from Stamey et al. (1989) and consists of 97 observations. The predictors are clinical measurements: log(cancer volume) (lcavol), log(prostate weight) (lweight), age, log(benign prostatic hyperplasia amount) (lbph), seminal vesicle invasion (svi), log(capsular penetration) (lcp), Gleason score (gleason), and percentage Gleason scores 4 or 5 (pgg45). The original data set does not contain strongly correlated groups. We added a noisy duplicate, svi+0.06e, where $\mathbf{e} \sim N(0,1)$, to the data set. This induces a correlation of approximately .99 between svi and its duplicate. The response is log(prostate specific antigen) (lpsa). We used tenfold cross-validation to estimate tuning parameters with a testing set of 67 observations (Hastie et al., 2001).

In Figure 3, we compare the solution paths for lasso, elastic net, UST, and weighted fusion. The highly correlated variables, $\mathbf{x}_5$ (svi) and $\mathbf{x}_9$ (svi+0.06e), are represented in light and dashed lines, respectively. First of all, we note that for lasso $\mathbf{x}_9$ enters the solution path only at the final OLS step where its coefficient estimate is strongly negative. This demonstrates that lasso performs poorly under strong collinearity. Secondly, the elastic net does not select $\mathbf{x}_9$. The paths for coefficients of $\mathbf{x}_5$ and $\mathbf{x}_9$ are far apart and cross-validation can not pick up both variables. Thirdly, UST estimates are shown to be highly biased. Elastic net that bridges lasso and UST may not work well when both lasso and UST perform poorly. Next, we portrayed the solution paths of weighted fusion. We note that the paths for the coefficients of $\mathbf{x}_5$ and $\mathbf{x}_9$ are very close and both are selected by cross-validation. The OLS estimate of β_5 on the original data without $\mathbf{x}_9$ is around 2.5. Interestingly, the maximum values of β_5 and β_9 for weighted fusion halves the original OLS estimate of β_5 at around 1.25. This shows that the effects of β_5 and β_9 in weighted fusion are evenly divided and are not exaggerated when variables are highly correlated. Further, the solution paths for all other variables are very similar to that of lasso. This indicates that weighted fusion preserves the good selection and prediction properties of lasso and simultaneously deals with the problem of collinearity.

The test errors for lasso, elastic net, UST, and weighted fusion are 0.7610, 0.6491, 1.2502, and 0.6543, respectively. We note that lasso has a relatively large test error, whereas UST has the largest. The errors for elastic net and weighted fusion are very close with the error for weighted fusion a little higher due to selection of an additional noisy duplicate.

We have shown weighted fusion to be a more reasonable method under collinearity than lasso and elastic net using solution path analysis. In the next section, we will demonstrate with simulation results.

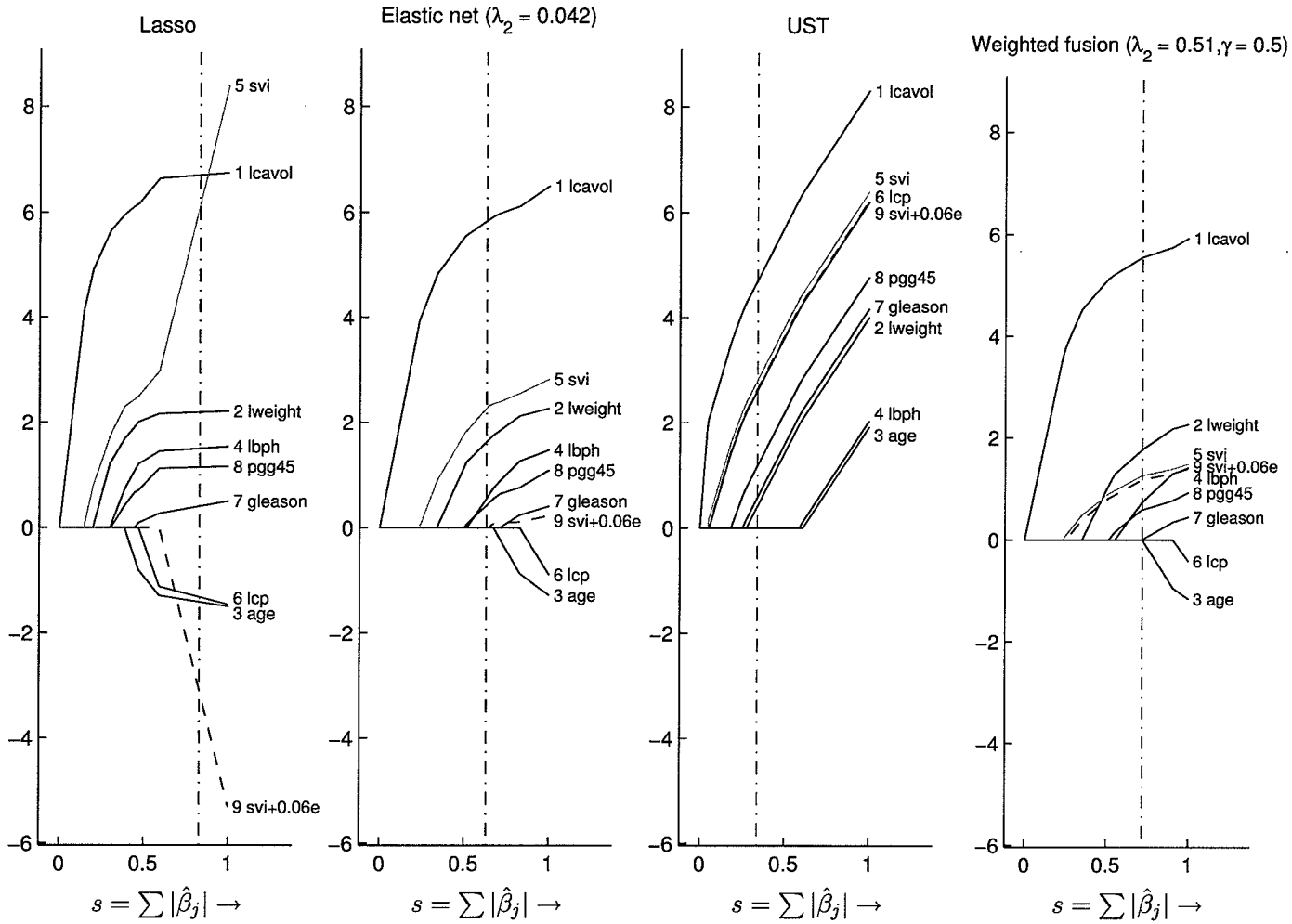

Figure 3: Estimates of regression coefficients for prostate cancer data. From the left to right are, the solution path for lasso, elastic net, UST, and weighted fusion. In each panel, the vertical dash-dotted line indicates the tuning parameter $\hat{s}$ chosen by tenfold cross-validation.

6 Simulation study

In the following examples, we compare the performance of weighted fusion and ridge fusion with that of OLS, ridge, lasso, elastic net, UST, and oracle methods. The oracle is an ideal estimator obtained a priori by OLS regression on the important predictors. All algorithms were written in MATLAB codes. We used fivefold cross-validation to estimate tuning parameters for each procedure.

For each example, we simulated 200 data sets from the true model,

$$\mathbf{y} = \mathbf{X}\beta + \sigma\epsilon,\ \epsilon \sim N(\mathbf{0}, \mathbf{I}),$$

consisting of the training and an independent testing set. Our tuning parameters were estimated using fivefold cross-validation on the training set. We used only the representative values $\{0.5, 1, 2.5, 5, 25\}$ to select the thresholding parameter γ for weighted fusion. Estimated coefficients from the training set were then used to compute the relative prediction error (RPE) on the testing set. The relative prediction error (RPE) is defined by

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

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

The details of the four scenarios are given below.

1. In example 1, each training set consists of 20 observations, and each testing set has 200 observations. We use $\beta = (3, 1.5, 0, 0, 2, 0, 0, 0)$ and $\sigma = 3$. In addition, we set the predictors correlations as $\rho_{ij} = 0.5^{|i-j|}$ for all i and j. This creates a sparse model with a few large effects and predictors with first-order autoregressive correlation structure.

2. Example 2 is the same as example 1, except that $\beta_j = 0.85$ for all j. This creates a non-sparse model with many small effects.

3. In example 3, each training set consists of 100 observations, and each testing set has 400 observations. We use true coefficients

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

and $\sigma = 15$ with $\rho_{ij} = 0.5$ for all i and j such that $i \neq j$. This induces blocked effects and predictors with constant correlation structure.

Table 1: Median RPEs based on 200 replications

Method	*Example 1*	*Example 2*	*Example 3*	*Example 4*
OLS	0.7669 (0.0364)	0.7361 (0.0379)	0.6837 (0.0164)	0.6893 (0.0260)
Ridge	0.5433 (0.0310)	0.3088 (0.0167)	0.1153 (0.0025)	0.2967 (0.0099)
Ridge fusion	0.5397 (0.0246)	0.2632 (0.0119)	0.1069 (0.0021)	0.3618 (0.0098)
Lasso	0.5984 (0.0282)	0.6095 (0.0224)	0.2893 (0.0081)	1.8443 (0.1131)
Elastic net	0.5712 (0.0273)	0.5431 (0.0222)	0.5822 (0.0422)	1.2344 (0.0142)
UST	0.7551 (0.0378)	0.6811 (0.0423)	1.1512 (0.0255)	7.8487 (0.1278)
Weighted fusion	0.5060 (0.0255)	0.3799 (0.0290)	0.2197 (0.0069)	0.3191 (0.0143)
Oracle	0.1982 (0.0127)	0.7361 (0.0357)	0.2465 (0.0074)	0.2529 (0.0083)

NOTE: The numbers in parentheses are the corresponding standard errors of the medians estimated with 500 bootstrapped resamplings on the 200 RPEs.

4. In example 4, each training set consists of 100 observations, and each testing set has 400 observations. We set $\sigma = 6$ and the true coefficients

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

The predictors $\mathbf{X}$ are generated as,

$$\begin{aligned} \mathbf{x}_j &= Z_1 + 0.1\epsilon_x \quad \text{for } j = 1, \ldots, 15 \\ \mathbf{x}_j &\sim N(0,1) \quad \text{for } j = 16, \ldots, 40, \end{aligned}$$

where $Z_1 \sim N(0,1)$ and $\epsilon_x \sim N(0,1)$ independent. This creates a within group correlation of approximately .99 between the first 15 predictors.

Examples 1-3 were used in the original lasso paper by Tibshirani (1996), and example 4 is modified from Zou and Hastie (2005).

In Table 1, we summarize the prediction results of our simulations. First of all, we note that UST shows poor and unstable performances in all examples. The increase of UST in RPE from that of lasso in examples 1, 2, 3, and 4 are 26%, 12%, 298%, and 326%, respectively. Furthermore, elastic net shows relatively small reduction in RPE from that of lasso in examples 1, 2, and 4, with 5%, 11%, 33% reduction, respectively, whereas in example 3 elastic net under-performs lasso with 101% increase in RPE. Interestingly, the prediction accuracy of ridge fusion is similar to ridge, and it even outperforms ridge in examples 2 and 3. Finally, weighted fusion outperforms elastic net and lasso. The reduction in RPE from elastic net in examples 1, 2, 3, and 4 are 11%, 30%, 62%, and 74%, respectively; the reduction in RPE from lasso in examples 1, 2, 3, and 4 are 15%, 38%, 24%, and 83%, respectively.

Table 2: Median number of selected variables based on 200 replications

Method	*Example 1*		*Example 2*		*Example 3*		*Example 4*	
	C	*I*	*C*	*I*	*C*	*I*	*C*	*I*
Lasso	3	1	5	0	12	6	9	0
Elastic net	3	2.5	7	0	16	12	15	0
UST	3	3	7	0	17	14	15	0
Weighted fusion	3	3	7.5	0	14	8	19	3
Oracle	3	0	8	0	20	0	20	0

NOTE: "C" represents the median number of selected nonzero coefficients, whereas "I" represents the median number of zero coefficients selected incorrectly.

The simulation results indicate UST as an extremely poor estimator for prediction. The elastic net bridges the lasso and UST and thus is unlikely to show strong advantage in prediction as indicated by the simulation results. Ridge fusion is stable and sometimes may outperform ridge. The results indicate that weighted fusion dominates both lasso and elastic net for correlated variables.

Table 2 presents the variable selection results for our simulations. The results indicate that weighted fusion produces sparse solutions. Further, by considering grouping effect, weighted fusion tends to select more variables than lasso. In examples 1, 2, and 3, weighted fusion has similar selection behavior as elastic net. In example 4, which has 15 highly correlated and 5 independent significant variables, lasso performs very poorly and selects only 9 important variables. UST selects only 15 variables, and elastic net is the same. On the other hand, weighted fusion selects 19 variables.

7 Discussion

In this paper, we have proposed the weighted fusion, a new penalized regression method that can incorporate information redundancy among correlated variables in regression and variable selection. Both simulation and real examples demonstrate that weighted fusion often outperforms lasso and elastic net in prediction and variable selection.

The weighted fusion penalty (5) may also be applied in conjunction with other optimization criteria in addition to least squares. For example, Huber loss can be used for robust estimation and hinge loss for classification.

Further, we utilized the lasso penalty to produce sparse solutions in (4). However, this is not the only choice. If one's interest is in unbiasedness or consistency,

SCAD (Fan and Li, 2001) or adaptive lasso (Zou, 2006), respectively, can be applied. Another interesting choice is the method of Dantzig selector by Candes and Tao (2007).

We note that the key in producing good performance for estimation and selection for correlated variables is the weighted fusion penalty (5). The weighted fusion penalty introduces predictors correlations-induced structure on regression coefficients. Interestingly, performance results of ridge fusion, which employs just the weighted fusion penalty, show that it can sometimes outperform ridge regression. This demonstrates the stability of the weighted fusion penalty for regression under correlation.

In this paper, we have introduced the GRIL estimator. We believe that GRIL encompasses an important class of computationally efficient estimators for penalized regression.

8 Acknowledgements

The authors are grateful to Jayanta K. Ghosh, Jian Zhang, Ji Zhu, and Yu Zhu for very helpful suggestions and detailed discussions. In addition, we thank Mary Ellen Bock, Jiashun Jin, and Jun Xie for discussions. Xinge Jessie Jeng is supported by a grant from NSF (DMS-0639980). Computing resources and support were provided by the Department of Statistics, Purdue University, and the Rosen Center for Advanced Computing of Information Technology at Purdue.

9 Appendix

9.1 Proof of Theorem 1

By the construction of (4), it is straightforward to see that when $\rho_{ij} \to 1$ implies $w_{ij} \to \infty$, the term $w_{ij}(\hat{\beta}_i - s_{ij}\hat{\beta}_j)^2$ diverges unless $|\hat{\beta}_i - s_{ij}\hat{\beta}_j| \to 0$. Since $\hat{\beta}$ is the minimizer of $||\mathbf{y} - \mathbf{X}\beta||^2 + \lambda_1 \sum_{j=1}^{p} |\beta_j| + (\lambda_2/p) \sum_{i<j} w_{ij}(\beta_i - s_{ij}\beta_j)^2$, we must have $|\hat{\beta}_i - s_{ij}\hat{\beta}_j| \to 0$. □

9.2 Proof of Theorem 2

First, let

$$L(\beta) = ||\mathbf{y} - \mathbf{X}\beta||^2 + \lambda_1 \sum_{j=1}^{p} |\beta_j| + \frac{\lambda_2}{p} \sum_{i<j} w_{ij}(\beta_i - s_{ij}\beta_j)^2.$$

When $\hat{\beta}_i \neq 0$, solving

$$\frac{\partial}{\partial \beta_i} L(\beta) = -2\mathbf{x}_i^T(\mathbf{y} - \mathbf{X}\beta) + \lambda_1 \text{sgn}(\beta_i) + \frac{2\lambda_2}{p}[(\sum_k w_{ik})\beta_i - \sum_k w_{ik}s_{ik}\beta_k] = 0,$$

implies

$$\hat{\beta}_i = \frac{1}{2\lambda_2 \bar{w}_{i\cdot}}[2\mathbf{x}_i^T(\mathbf{y} - \mathbf{X}\hat{\beta}) - \lambda_1 \text{sgn}(\hat{\beta}_i)] + \frac{1}{p\bar{w}_{i\cdot}} \sum_k w_{ik}s_{ik}\hat{\beta}_k,$$

and similarly,

$$\hat{\beta}_j = \frac{1}{2\lambda_2 \bar{w}_{j\cdot}}[2\mathbf{x}_j^T(\mathbf{y} - \mathbf{X}\hat{\beta}) - \lambda_1 \text{sgn}(\hat{\beta}_j)] + \frac{1}{p\bar{w}_{j\cdot}} \sum_k w_{jk}s_{jk}\hat{\beta}_k.$$

Denote

$$I = \Big| \frac{2\mathbf{x}_i^T(\mathbf{y} - \mathbf{X}\hat{\beta}) - \lambda_1 \text{sgn}(\hat{\beta}_i)}{\bar{w}_{i\cdot}} - \frac{2\mathbf{x}_j^T(\mathbf{y} - \mathbf{X}\hat{\beta}) - \lambda_1 \text{sgn}(\hat{\beta}_j)}{\bar{w}_{j\cdot}} \Big|, \tag{20}$$

then

$$|\hat{\beta}_i - \hat{\beta}_j| \leq \frac{1}{2\lambda_2} \cdot I + \frac{1}{p} \sum_{1 \leq k \leq p} |(\frac{w_{ik}s_{ik}}{\bar{w}_{i\cdot}} - \frac{w_{jk}s_{jk}}{\bar{w}_{j\cdot}})\hat{\beta}_k|,$$

so, it is enough to show

$$I \leq \frac{2||\mathbf{y}||\sqrt{2(1-\rho_{ij})}}{\bar{w}_{i\cdot}} + (2||\mathbf{y}|| + \lambda_1) \cdot |\frac{1}{\bar{w}_{i\cdot}} - \frac{1}{\bar{w}_{j\cdot}}|. \tag{21}$$

Since $|a/b - c/d| \leq |(a-c)/b| + |c| \cdot |1/b - 1/d|$, $\hat{\beta}_i\hat{\beta}_j > 0$, and $|(\mathbf{x}_i - \mathbf{x}_j)^T(\mathbf{y} - \mathbf{X}\hat{\beta})| \leq ||\mathbf{x}_i - \mathbf{x}_j|| \cdot ||\mathbf{y} - \mathbf{X}\hat{\beta}|| \leq \sqrt{2(1-\rho_{ij})}||\mathbf{y}||$, then

$$I \leq \frac{2||\mathbf{y}||\sqrt{2(1-\rho_{ij})}}{\bar{w}_{i\cdot}} + |2\mathbf{x}_j^T(\mathbf{y} - \mathbf{X}\hat{\beta}) - \lambda_1 \text{sgn}(\hat{\beta}_j)| \cdot |\frac{1}{\bar{w}_{i\cdot}} - \frac{1}{\bar{w}_{j\cdot}}|.$$

For the second term above, we have $|2\mathbf{x}_j^T(\mathbf{y} - \mathbf{X}\hat{\beta}) - \lambda_1 \text{sgn}(\beta_j)| \leq |2\mathbf{x}_j^T(\mathbf{y} - \mathbf{X}\hat{\beta})| + \lambda_1 \leq 2||\mathbf{x}_j|| \cdot ||\mathbf{y}|| + \lambda_1 \leq 2||\mathbf{y}|| + \lambda_1$, then, (21) follows. □

9.3 Proof of Theorem 3

Let $\hat{u} = \hat{\beta}^{\text{GRIL}}(\lambda_1, \lambda_2) - \beta^* = \{\hat{\mathbf{u}}_{(1)}, \hat{\mathbf{u}}_{(2)}\}$, then sign consistency is implied by

$$|\hat{\mathbf{u}}_{(1)}| < |\beta^*_{(1)}|, \qquad \hat{\mathbf{u}}_{(2)} = 0. \tag{22}$$

Let

$$V(u) = ||\epsilon - \mathbf{X}u||^2 + \lambda_2(u+\beta^*)^T\mathbf{Q}(u+\beta^*) + \lambda_1||(u+\beta^*)||_1,$$

then $\hat{u} = \arg\min_u V(u)$. For notation simplicity, let

$$\mathbf{W} = \mathbf{X}^T\epsilon/\sqrt{n} = \{\mathbf{W}(1), \mathbf{W}(2)\}^T. \tag{23}$$

Since

$$\frac{\partial}{\partial u}||\epsilon - \mathbf{X}u||^2 = \frac{\partial}{\partial u}(u^T\mathbf{X}^T\mathbf{X}u - 2\epsilon^T\mathbf{X}u) = 2(\mathbf{X}^T\mathbf{X})u - 2\sqrt{n}\mathbf{W},$$

$$\frac{\partial}{\partial u}(\lambda_2(u+\beta^*)^T\mathbf{Q}(u+\beta^*)) = \lambda_2\frac{\partial}{\partial u}(u^T\mathbf{Q}u + 2\beta^{*T}\mathbf{Q}u) = 2\lambda_2(\mathbf{Q}u + \mathbf{Q}^T\beta^*),$$

then

$$\frac{\partial}{\partial u}(||\epsilon-\mathbf{X}u||^2+\lambda_2(u+\beta^*)^T\mathbf{Q}(u+\beta^*)) = 2(\sqrt{n}\mathbf{C}^n+\frac{\lambda_2}{\sqrt{n}}\mathbf{Q})\sqrt{n}u-2\sqrt{n}\mathbf{W}+2\lambda_2\mathbf{Q}^T\beta^*. \tag{24}$$

By KKT condition and (24), (22) is equivalent to

$$(\mathbf{C}^n_{11} + \frac{\lambda_2}{n}\mathbf{Q}_{11})\sqrt{n}\hat{\mathbf{u}}_{(1)} - \mathbf{W}(1) + \frac{\lambda_2}{\sqrt{n}}\mathbf{Q}_{11}\beta^*_{(1)} = -\frac{\lambda_1}{2\sqrt{n}}\text{sgn}(\beta^*_{(1)}), \tag{25}$$

$$|\hat{\mathbf{u}}_{(1)}| < |\beta^*_{(1)}|, \tag{26}$$

$$-\frac{\lambda_1}{2\sqrt{n}}\mathbf{1} \le (\mathbf{C}^n_{21} + \frac{\lambda_2}{n}\mathbf{Q}_{21})\sqrt{n}\hat{\mathbf{u}}_{(1)} - \mathbf{W}(2) + \frac{\lambda_2}{\sqrt{n}}\mathbf{Q}_{21}\beta^*_{(1)} \le \frac{\lambda_1}{2\sqrt{n}}\mathbf{1}. \tag{27}$$

Denote $\tilde{\mathbf{C}^n_{11}} = \mathbf{C}^n_{11} + \frac{\lambda_2}{n}\mathbf{Q}_{11}$, $\tilde{\mathbf{C}^n_{21}} = \mathbf{C}^n_{21} + \frac{\lambda_2}{n}\mathbf{Q}_{21}$, and replace $\hat{\mathbf{u}}_{(1)}$ in (26) and (27) by expression of $\hat{\mathbf{u}}_{(1)}$ in (25), then the above is implied by

$$|\tilde{\mathbf{C}^n_{11}}^{-1}\mathbf{W}(1)| < \sqrt{n}\Big(|\beta^*_{(1)}| - \tilde{\mathbf{C}^n_{11}}^{-1}|\frac{\lambda_1}{2n}\text{sgn}(\beta^*_{(1)}) + \frac{\lambda_2}{n}\mathbf{Q}_{11}\beta^*_{(1)}|\Big)$$

$$\begin{aligned}|\tilde{\mathbf{C}^n_{21}}\tilde{\mathbf{C}^n_{11}}^{-1}\mathbf{W}(1) - \mathbf{W}(2)| \le \frac{\lambda_1}{2\sqrt{n}}\Big\{1 \quad &- \quad |\tilde{\mathbf{C}^n_{21}}\tilde{\mathbf{C}^n_{11}}^{-1}\text{sgn}(\beta^*_{(1)}) \\ &- \quad \frac{2\lambda_2}{\lambda_1}(\tilde{\mathbf{C}^n_{21}}\tilde{\mathbf{C}^n_{11}}^{-1}\mathbf{Q}_{11} - \mathbf{Q}_{21})\beta^*_{(1)}|\Big\},\end{aligned}$$

where the left hand sides

$$\tilde{\mathbf{C}^n_{11}}^{-1}\mathbf{W}(1) \to_d N(\mathbf{0}, \mathbf{C}^{-1}_{11})$$

$$\tilde{\mathbf{C}^n_{21}}\tilde{\mathbf{C}^n_{11}}^{-1}\mathbf{W}(1) - \mathbf{W}(2) \to_d N(\mathbf{0}, \mathbf{C}_{22} - \mathbf{C}_{21}\mathbf{C}^{-1}_{11}\mathbf{C}_{12}) \tag{28}$$

by $\tilde{\mathbf{C}^n_{11}} \to \mathbf{C}_{11}$ for $\lambda_2/n \to 0$. Result follows after applying condition (16). □

9.4 Proof of Theorem 4

Sign consistency implies KKT condition, then we have

$$(\mathbf{C}_{11}^n + \frac{\lambda_2}{n}\mathbf{Q}_{11})\sqrt{n}\hat{\mathbf{u}}_{(1)} - \mathbf{W}(1) + \frac{\lambda_2}{\sqrt{n}}\mathbf{Q}_{11}\beta_{(1)}^* = -\frac{\lambda_1}{2\sqrt{n}}\text{sign}(\beta_{(1)}^*) \tag{29}$$

$$-\frac{\lambda_1}{2\sqrt{n}}\mathbf{1} \leq (\mathbf{C}_{21}^n + \frac{\lambda_2}{n}\mathbf{Q}_{21})\sqrt{n}\hat{\mathbf{u}}_{(1)} - \mathbf{W}(2) + \frac{\lambda_2}{\sqrt{n}}\mathbf{Q}_{21}\beta_{(1)}^* \leq \frac{\lambda_1}{2\sqrt{n}}\mathbf{1} \tag{30}$$

held with probability 1. Combine (29) and (30), we have

$$\tilde{\mathbf{C}_{21}^n}\tilde{\mathbf{C}_{11}^n}^{-1}\mathbf{W}(1) - \mathbf{W}(2) \geq \frac{\lambda_1}{2\sqrt{n}}(-\mathbf{1} + \mathbf{v}) \tag{31}$$

with probability 1, where $\mathbf{v} = \mathbf{C}_{21}^n(\mathbf{C}_{11}^n)^{-1}\text{sign}(\beta_{(1)}^*) - (2\lambda_2/\lambda_1)(\mathbf{C}_{21}^n(\mathbf{C}_{11}^n)^{-1}\mathbf{Q}_{11} - \mathbf{Q}_{21})\beta_{(1)}^*$. Suppose (17) fails, then there exists an element in $\mathbf{v}$ that is greater than 1, therefore, the right hand side of (31) has at least one positive element. On the other hand, (28) implies that there is a non-vanishing probability that any element of the left hand side of (31) is negative. Result follows by contradiction. □

References

Candes, E., and Tao, T. (2007). The dantzig selector: Statistical estimation when p is much larger than n. *Ann. Statist. (To Appear).*

Donoho, D., and Johnstone, I. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika, 81*, 425–455.

Donoho, D., Johnstone, I., Kerkyacharian, G., and Picard, D. (1995). Wavelet shrinkage; asymptopia? *J. R. Statist. Soc. B, 57*, 301–337.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). Least angle regression. *Ann. Statist., 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.

Fan, J., and Li, R. (2006). Statistical challenges with high dimensionality: Feature selection in knowledge discovery. *Proceedings of the International Congress of Mathematicians* (pp. 595–622). Zurich.

Fan, J., and Peng, H. (2004). On nonconcave penalized likelihood with diverging number of parameters. *Ann. Statist., 32*, 928–961.

Harrell, F. (2001). *Regression modeling strategies.* Springer.

Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The elements of statistical learning.* Springer.

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

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

Kim, Y., and Kim, J. (2004). Gradient lasso for feature selection. *Proceedings of the 21st International Conference on Machine Learning.*

Land, S., and Friedman, J. (1996). *Variable fusion: a new method of adaptive signal regression* (Technical Report). Department of Statistics, Stanford University.

Mosteller, F., and Tukey, J. (1977). *Data analysis and regression: A second course in statistics.* Addison-Wesley.

Rosset, S., and Zhu, J. (2007). Piecewise linear regularized solution. *Ann. Statist., 35*, 1012–1030.

Stamey, T., Kabalin, J., McNeal, J., Johnstone, I., Freiha, F., Redwine, E., and Yang, 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. R. Statist. Soc. B, 58*, 267–288.

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., and Knight, K. (2005). Sparsity and smoothness via the fused lasso. *J. R. Statist. Soc. B, 67*, 91–108.

Wang, L., Zhu, J., and Zou, H. (2006). The doubly regularized support vector machine. *Statistica Sinica, 16*, 589–615.

Yu, B. (2007). Embracing statistical challenges in the information technology age. *Technometrics, 49*, 237–248.

Yuan, M., and Lin, Y. (2007). On the non-negative garrotte estimator. *J. R. Statist. Soc. B, 69*, 143–161.

Zhao, P., and Yu, B. (2006). On model selection consistency of lasso. *Journal of Machine Learning Research, 7*, 2541–2567.

Zhao, P., and Yu, B. (2007). Stagewise lasso. *Journal of Machine Learning Research (To Appear).*

Zou, H. (2006). The adaptive lasso and its oracle properties. *J. Am. Statist. Ass.*, *101*, 1418–1429.

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