Please see the online version of this article for supplementary materials.

Variable Selection With the Strong Heredity Constraint and Its Oracle Property

Nam Hee CHOI, William LI, and Ji ZHU

In this paper, we extend the LASSO method (Tibshirani 1996) for simultaneously fitting a regression model and identifying important interaction terms. Unlike most of the existing variable selection methods, our method automatically enforces the heredity constraint, that is, an interaction term can be included in the model only if the corresponding main terms are also included in the model. Furthermore, we extend our method to generalized linear models, and show that it performs as well as if the true model were given in advance, that is, the *oracle* property as in Fan and Li (2001) and Fan and Peng (2004). The proof of the oracle property is given in online supplemental materials. Numerical results on both simulation data and real data indicate that our method tends to remove irrelevant variables more effectively and provide better prediction performance than previous work (Yuan, Joseph, and Lin 2007 and Zhao, Rocha, and Yu 2009 as well as the classical LASSO method).

KEY WORDS: Heredity structure; LASSO; Regularization.

1. INTRODUCTION

Consider the usual regression situation: we have training data $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_i, y_i), \ldots, (\mathbf{x}_n, y_n)$, where $\mathbf{x}_i = (x_{i1}, \ldots, x_{ij}, \ldots, x_{ip})$ are the predictors and y_i is the response. To model the response y in terms of the predictors $x_1, \ldots, x_p$, one may consider the linear model

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \epsilon,$$

where ϵ is the error term. In many important practical problems, however, the main terms $x_1, \ldots, x_p$ alone may not be enough to capture the relationship between the response and the predictors, and higher-order interactions are often of interest to scientific researchers. For example, many complex diseases, such as cancer, involve multiple genetic and environmental risk factors, and scientists are particularly interested in assessing gene–gene and gene–environment interactions.

In this paper, we consider a regression model with main terms and all possible two-way interaction terms, that is,

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \alpha_{12}(x_1 x_2) + \alpha_{13}(x_1 x_3) + \cdots + \alpha_{p-1,p}(x_{p-1} x_p) + \epsilon. \quad (1)$$

The goal here is to find out which terms, especially which interaction terms, have an important effect on the response. For example, $x_1, \ldots, x_p$ may represent different genetic factors, y may represent a certain phenotype, and we are interested in deciphering how these genetic factors "work together" to determine the phenotype. Later, we extend the setting to generalized linear models and develop an asymptotic theory there.

There are two important challenges in this problem: prediction accuracy and interpretation. We would like our model to accurately predict on future data. Prediction accuracy can often be improved by shrinking the regression coefficients. Shrinkage sacrifices unbiasedness to reduce the variance of the predicted value and hence may improve the overall prediction accuracy. Interpretability is often realized via variable selection. With a large number of variables (including both the main terms and the interaction terms), possibly larger than the number of observations, we often would like to determine a smaller subset that exhibits the strongest effects.

Variable selection has been studied extensively in the literature; for example, see Breiman (1995), Tibshirani (1996), Fan and Li (2001), and Shen and Ye (2002). In particular, LASSO (Tibshirani 1996) has gained much attention in recent years. The LASSO criterion penalizes the L_1-norm of the regression coefficients to achieve a sparse model:

$$\min_{\beta_j, \alpha_{jj'}} \sum_{i=1}^{n} \left(\left(y_i - \beta_0 - \sum_j \beta_j x_{ij} - \sum_{j<j'} \alpha_{jj'}(x_{ij} x_{ij'}) \right)^2 + \lambda \left(\sum_j |\beta_j| + \sum_{j<j'} |\alpha_{jj'}| \right) \right). \quad (2)$$

The L_1-norm penalty can shrink some of the fitted coefficients to be exactly zero when making the tuning parameter sufficiently large. However, LASSO and other methods mentioned above are for the case when the candidate variables can be treated individually or "flatly." When interaction terms exist, there is a natural hierarchy among the variables, that is, an interaction term can be included in the model only if both of the corresponding main terms are also included in the model. This is referred to as the marginality in generalized linear models (McCullagh and Nelder 1989; Nelder 1994) or the strong heredity in the analysis of designed experiments (Hamada and Wu 1992). Justifications of effect heredity can be found in Chipman (1996) and Joseph (2006). A generic variable selection method, however, may select an interaction term but not the corresponding main terms, and such models are difficult to interpret in practice.

In this paper, we extend the LASSO method so that it simultaneously fits the regression model and identifies interaction terms obeying the strong heredity constraint. Furthermore,

Nam Hee Choi is Lecturer, Department of Statistics, University of Michigan, Ann Arbor, MI 48109. William Li is Professor, Carlson School of Management, University of Minnesota, Minneapolis, MN 55455. Ji Zhu is Associate Professor, Department of Statistics, University of Michigan, Ann Arbor, MI 48109 (E-mail: *jizhu@umich.edu*). We thank Rayjean Hung, Stefano Porru, Paolo Boffetta, and John Witte for sharing the bladder cancer dataset. Choi and Zhu are partially supported by grants DMS-0705532 and DMS-0748389 from the National Science Foundation.

Journal of the American Statistical Association
March 2010, Vol. 105, No. 489, Theory and Methods
DOI: 10.1198/jasa.2010.tm08281

we show that when the regularization parameters are appropriately chosen, our new method has the oracle property (Fan and Li 2001; Fan and Peng 2004), that is, it performs as well as if the correct underlying model were given in advance. Such theoretical property has not been previously studied for variable selection with heredity constraints.

The rest of the paper is organized as follows. In Section 2, we introduce our new model and an algorithm to fit the model. Asymptotic properties are studied in Section 3, and numerical results are in Section 4. We conclude the paper with Section 5.

2. STRONG HEREDITY INTERACTION MODEL

In this section, we extend the LASSO method for selecting interaction terms while at the same time keeping the strong heredity constraint. We call our model the strong heredity interaction model (SHIM). After introducing the model in Section 2.1, we develop an algorithm to compute the SHIM estimate in Section 2.2. We then extend SHIM to generalized linear models in Section 2.3.

2.1 Model

We reparameterize the coefficients for the interaction terms $\alpha_{jj'}$, $j < j'$, $j, j' = 1, \ldots, p$, as $\alpha_{jj'} = \gamma_{jj'}\beta_j\beta_{j'}$, and consider the following model:

$$g(\mathbf{x}) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \gamma_{12}\beta_1\beta_2(x_1x_2) + \gamma_{13}\beta_1\beta_3(x_1x_3) + \cdots + \gamma_{p-1,p}\beta_{p-1}\beta_p(x_{p-1}x_p). \quad (3)$$

Notice the difference in the coefficients of the interaction terms between (1) and (3). In (3), the coefficient for the interaction term $(x_jx_{j'})$ is expressed as the product of $\gamma_{jj'}$, β_j and $\beta_{j'}$, instead of a single parameter $\alpha_{jj'}$. By writing the coefficient as a product, the model itself enforces the heredity constraint. That is, whenever the coefficient for either x_j or $x_{j'}$, that is, β_j or $\beta_{j'}$, is equal to zero, the coefficient for the interaction term $(x_jx_{j'})$ is automatically set to zero; vice versa, if the coefficient for $(x_jx_{j'})$ is not equal to zero, it implies that both β_j and $\beta_{j'}$ are not equal to zero.

For the purpose of variable selection, we consider the following penalized least squares criterion:

$$\min_{\beta_j,\gamma_{jj'}} \sum_{i=1}^{n}\big((y_i - g(\mathbf{x}_i))^2 + \lambda_\beta(|\beta_1| + \cdots + |\beta_p|) + \lambda_\gamma(|\gamma_{12}| + \cdots + |\gamma_{p-1,p}|)\big), \quad (4)$$

where $g(\mathbf{x})$ is from (3), and the penalty is the L_1-norm of the parameters, as in LASSO (2). There are two tuning parameters, λ_β and λ_γ. The first tuning parameter λ_β controls the estimates at the main effect level: if β_j is shrunken to zero, variable x_j and all its "descendants," that is, the corresponding interaction terms that involve x_j are removed from the model. The second tuning parameter λ_γ controls the estimates at the interaction effect level: if β_j and $\beta_{j'}$ are not equal to zero but the corresponding interaction effect is not strong, $\gamma_{jj'}$ still has the possibility of being zero, so it has the flexibility of selecting only the main terms.

To further improve the criterion (4), we apply the adaptive idea which has been used extensively in the literature, including Breiman (1995), Zou (2006), Wang, Li, and Jiang (2007), and Zhang and Lu (2007), that is, to penalize different parameters differently. We consider

$$\min_{\beta_j,\gamma_{jj'}} \sum_{i=1}^{n}\big((y_i - g(\mathbf{x}_i))^2 + \lambda_\beta(w_1^\beta|\beta_1| + \cdots + w_p^\beta|\beta_p|) + \lambda_\gamma(w_{12}^\gamma|\gamma_{12}| + \cdots + w_{p-1,p}^\gamma|\gamma_{p-1,p}|)\big), \quad (5)$$

where w_j^β and $w_{jj'}^\gamma$ are prespecified weights. The intuition is that if the effect of a variable is strong, we would like the corresponding weight to be small, hence the corresponding parameter is lightly penalized. If the effect of a variable is not strong, we would like the corresponding weight to be large, hence the corresponding parameter is heavily penalized. How to prespecify the weights w_j^β and $w_{jj'}^\gamma$ from the data is discussed below.

Computing Adaptive Weights. Regarding the adaptive weights w_j^β and $w_{jj'}^\gamma$ for the regression parameters in (5), we consider three possibilities:

1. Set all the weights equal to 1. We denote this as "plain."
2. Following Breiman (1995) and Zou (2006), we can compute the weights using the ordinary least squares (OLS) estimates from the training observations:
$$w_j^\beta = \left|\frac{1}{\hat\beta_j^{OLS}}\right|, \qquad w_{jj'}^\gamma = \left|\frac{\hat\beta_j^{OLS}\cdot\hat\beta_{j'}^{OLS}}{\hat\alpha_{jj'}^{OLS}}\right|,$$
where $\hat\beta_j^{OLS}$ and $\hat\alpha_{jj'}^{OLS}$ are the corresponding OLS estimates. We denote this as "Adaptive(OLS)."
3. When $n < p$, the OLS estimates are not available, we can compute the weights using the ridge regression estimates, that is, replacing all the above OLS estimates with the ridge regression estimates, and we denote this as "Adaptive(Ridge)." We recognize that when using the ridge regression estimates as weights, there is an issue of selecting the tuning parameter for the ridge regression. We note that in our simulation studies (Section 4.1), we used a separate validation set to select the tuning parameter, while in real data analysis (Section 4.3), we used cross-validation. We also experimented with GCV and BIC. We found that the result of the Adaptive(Ridge) SHIM is not sensitive to the choice of the ridge penalty.

2.2 Algorithm

To estimate the parameters β_j and $\gamma_{jj'}$, we can use an iterative approach, that is, we first fix β_j and estimate $\gamma_{jj'}$, then we fix $\gamma_{jj'}$ and estimate β_j, and we iterate between these steps until the solution converges. Since at each step, the value of the objective function (5) decreases, the solution is guaranteed to converge.

When β_j, $j = 1, \ldots, p$, are fixed, (5) becomes a LASSO problem, hence we can use either the LARS/LASSO algorithm (Efron et al. 2004) or a quadratic programming package to efficiently solve for $\gamma_{jj'}$, $j < j'$. When $\gamma_{jj'}$, $j < j'$, are fixed, we can sequentially solve for β_j: for each $j = 1, \ldots, p$, we fix $\gamma_{jj'}, j < j'$, and $\boldsymbol{\beta}_{[-j]} = (\beta_1, \ldots, \beta_{j-1}, \beta_{j+1}, \ldots, \beta_p)$, then (5) becomes a simple LASSO problem with only one parameter β_j, and we can solve it with a closed form formula. We note that the sequential strategy of fixing $(p-1)$ β_j's and solving for the other β_j is similar to the shooting algorithm in Fu (1998) and Friedman et al. (2007).

In summary, the algorithm proceeds as follows:

1. *Standardization*. Center $\mathbf{y}$. Center and normalize each term $\mathbf{x}_j, \mathbf{x}_j\mathbf{x}_{j'}, j < j', j, j' = 1, \ldots, p$.
2. *Initialization*. Initialize $\hat{\beta}_j^{(0)}$ and $\hat{\gamma}_{jj'}^{(0)}$, $j < j'$, $j, j' = 1, \ldots, p$, with some plausible values. For example, we can use the least square estimates or the simple regression estimates by regressing the response $\mathbf{y}$ on each of the terms. Let $m = 1$.
3. *Update* $\hat{\gamma}_{jj'}$. Let

$$\begin{aligned}\tilde{y}_i &= y_i - \hat{\beta}_1^{(m-1)} x_{i1} - \cdots - \hat{\beta}_p^{(m-1)} x_{ip}, \\ &\qquad i = 1, \ldots, n, \\ \tilde{x}_{i,jj'} &= \hat{\beta}_j^{(m-1)} \hat{\beta}_{j'}^{(m-1)} (x_{ij} x_{ij'}), \\ &\qquad i = 1, \ldots, n; j < j', j, j' = 1, \ldots, p,\end{aligned}$$

then

$$\hat{\gamma}_{jj'}^{(m)} = \arg\min_{\gamma_{jj'}} \sum_{i=1}^{n} \Bigg(\Bigg(\tilde{y}_i - \sum_{j<j'} \gamma_{jj'} \tilde{x}_{i,jj'} \Bigg)^2 + \lambda_\gamma \sum_{j<j'} w_{jj'}^{\gamma} |\gamma_{jj'}| \Bigg).$$

4. *Update* $\hat{\beta}_j$.
   - Let $\hat{\beta}_j^{(m)} = \hat{\beta}_j^{(m-1)}, j = 1, \ldots, p$.
   - For each j in $1, \ldots, p$, let

$$\begin{aligned}\tilde{y}_i &= y_i - \sum_{j' \neq j} \hat{\beta}_{j'}^{(m)} x_{ij'} - \sum_{j'<j'', j', j'' \neq j} \hat{\beta}_{j'}^{(m)} \hat{\beta}_{j''}^{(m)} (x_{ij'} x_{ij''}), \\ &\qquad i = 1, \ldots, n, \\ \tilde{x}_i &= x_{ij} + \sum_{j'<j} \hat{\gamma}_{j'j}^{(m)} \hat{\beta}_{j'}^{(m)} (x_{ij'} x_{ij}) + \sum_{j'>j} \hat{\gamma}_{jj'}^{(m)} \hat{\beta}_{j'}^{(m)} (x_{ij} x_{ij'}), \\ &\qquad i = 1, \ldots, n,\end{aligned}$$

   then

$$\hat{\beta}_j^{(m)} = \arg\min_{\beta_j} \sum_{i=1}^{n} \big((\tilde{y}_i - \beta_j \tilde{x}_i)^2 + \lambda_\beta w_j^{\beta} |\beta_j| \big).$$

5. Compute the relative difference between $Q_n(\hat{\boldsymbol{\theta}}^{(m-1)})$ and $Q_n(\hat{\boldsymbol{\theta}}^{(m)})$:

$$\Delta^{(m)} = \frac{|Q_n(\hat{\boldsymbol{\theta}}^{(m-1)}) - Q_n(\hat{\boldsymbol{\theta}}^{(m)})|}{|Q_n(\hat{\boldsymbol{\theta}}^{(m-1)})|},$$

   where

$$\begin{aligned}Q_n(\boldsymbol{\theta}) = \sum_{i=1}^{n} (y_i - g(\mathbf{x}_i))^2 &+ \lambda_\beta (w_1^{\beta} |\beta_1| + \cdots + w_p^{\beta} |\beta_p|) \\ &+ \lambda_\gamma (w_{12}^{\gamma} |\gamma_{12}| + \cdots + w_{p-1,p}^{\gamma} |\gamma_{p-1,p}|)\end{aligned}$$

   for $\boldsymbol{\theta} = (\beta_1, \ldots, \beta_p, \gamma_{12}, \ldots, \gamma_{p-1,p})$.
6. Stop the algorithm if $\Delta^{(m)}$ is small enough. Otherwise, let $m = m + 1$ and go back to step 2.

Finally, we recognize that the SHIM criterion, similar as SCAD (Fan and Li 2001) and Bridge (Fu 1998), is nonconvex. Hence convergence of the algorithm to the global minimum is not guaranteed. To assess this limitation, we conducted an empirical investigation on the real dataset (Section 4.3). We ran our fitting procedure 100 times, using randomized starting values (based on OLS estimates from 100 bootstrap samples), and examined the resulting coefficient estimates. The results are surprisingly similar: the mean absolute difference between the original estimate and a new estimate is within the range of 10^{-8}. This experiment implies that our fitting algorithm may not be getting stuck in any local minimum and is reaching a global optimum.

2.3 Extension to Generalized Linear Models

The SHIM method can be naturally extended to likelihood based generalized linear models. Assume that the data $\mathbf{V}_i = \{(\mathbf{x}_i, y_i)\}$, $i = 1, \ldots, n$, are collected independently. Conditioning on $\mathbf{x}_i$, suppose Y_i has a density $f(g(\mathbf{x}_i), y_i)$, where g is a known link function with main terms and all possible interaction terms:

$$\begin{aligned}g(\mathbf{x}) &= \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \alpha_{12}(x_1 x_2) \\ &\quad + \alpha_{13}(x_1 x_3) + \cdots + \alpha_{p-1,p}(x_{p-1} x_p) \\ &= \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \gamma_{12}\beta_1\beta_2(x_1 x_2) + \cdots \\ &\quad + \gamma_{p-1,p}\beta_{p-1}\beta_p(x_{p-1} x_p). \qquad (6)\end{aligned}$$

As before, for the purpose of variable selection, we consider the following penalized negative log-likelihood criterion:

$$\begin{aligned}\min_{\beta_j, \gamma_{jj'}} -\sum_{i=1}^{n} \big(&\ell(g(\mathbf{x}_i), y_i) + \lambda_\beta (w_1^{\beta} |\beta_1| + \cdots + w_p^{\beta} |\beta_p|) \\ &+ \lambda_\gamma (w_{12}^{\gamma} |\gamma_{12}| + \cdots + w_{p-1,p}^{\gamma} |\gamma_{p-1,p}|) \big), \qquad (7)\end{aligned}$$

where $\ell(\cdot, \cdot) = \log f(\cdot, \cdot)$ is the conditional log-likelihood of Y. Similar to what we suggested in Section 2.1, one can specify the weights w_j^{β} and $w_{jj'}^{\gamma}$ using unpenalized maximum likelihood estimates or L_2-penalized maximum likelihood estimates. Later in Section 3, we show that under certain regularity conditions, using the unpenalized maximum likelihood estimates for specifying the weights guarantees that SHIM possesses the asymptotic oracle property.

3. ASYMPTOTIC ORACLE PROPERTY

In this section, we study the asymptotic behavior of SHIM based on the generalized linear model setting introduced in Section 2.3. In Section 3.1, we consider the asymptotic properties of SHIM estimates when the sample size n approaches to infinity. Furthermore, in Section 3.2, we consider the asymptotic properties of SHIM estimates when the number of covariates p_n also increases as the sample size n increases.

3.1 Asymptotic Oracle Property When $n \to \infty$

We show that when the number of predictors is fixed and the sample size approaches to infinity, SHIM possesses the oracle property under certain regularity conditions, that is, it performs as well as if the true model were known in advance (Fan and Li 2001).

Problem Setup. Let β_j^* and $\alpha_{jj'}^*$ denote the underlying true parameters. We further assume that the true model obeys the strong heredity constraint: $\alpha_{jj'}^* = 0$ if $\beta_j^* = 0$ or $\beta_{j'}^* = 0$. Let $\boldsymbol{\theta}^* = (\boldsymbol{\beta}^{*\top}, \boldsymbol{\gamma}^{*\top})^\top$ where

$$\gamma_{jj'}^* = \begin{cases} \dfrac{\alpha_{jj'}^*}{\beta_j^* \beta_{j'}^*} & \text{if } \beta_j^* \neq 0 \text{ and } \beta_{j'}^* \neq 0, \\ 0 & \text{otherwise.} \end{cases}$$

We consider the SHIM estimates $\hat{\boldsymbol{\theta}}_n$:

$$\begin{aligned} \hat{\boldsymbol{\theta}}_n &= \arg\min_{\boldsymbol{\theta}} Q_n(\boldsymbol{\theta}) \\ &= \arg\min_{\boldsymbol{\theta}} - \sum_{i=1}^{n} \Bigg(\ell(g(\mathbf{x}_i), y_i) \\ &\quad + n \sum_{j=1}^{p} \lambda_j^{\beta} |\beta_j| + n \sum_{k<k'} \lambda_{kk'}^{\gamma} |\gamma_{kk'}| \Bigg), \end{aligned} \tag{8}$$

where g is defined in (6). Note that $Q_n(\boldsymbol{\theta})$ in (8) is equivalent to the criterion in (7) by letting $\lambda_j^{\beta} = \frac{1}{n} \lambda_\beta w_j^{\beta}$ and $\lambda_{kk'}^{\gamma} = \frac{1}{n} \lambda_\gamma w_{kk'}^{\gamma}$. Furthermore, we define

$$\begin{aligned} &\mathcal{A}_1 = \{j : \beta_j^* \neq 0\}, \\ &\mathcal{A}_2 = \{(k, k') : \gamma_{kk'}^* \neq 0\}, \qquad \mathcal{A} = \mathcal{A}_1 \cup \mathcal{A}_2, \end{aligned}$$

that is, $\mathcal{A}_1$ contains the indices for main terms whose true coefficients are nonzero, and $\mathcal{A}_2$ contains the indices for interaction terms whose true coefficients are nonzero. Let

$$\begin{aligned} a_n &= \max\{\lambda_j^{\beta}, \lambda_{kk'}^{\gamma} : j \in \mathcal{A}_1,\ (k, k') \in \mathcal{A}_2\}, \\ b_n &= \min\{\lambda_j^{\beta}, \lambda_{kk'}^{\gamma} : j \in \mathcal{A}_1^c,\ (k, k') \in \mathcal{A}_2^c,\ k, k' \in \mathcal{A}_1\}. \end{aligned}$$

Notice that to compute b_n, we do not consider every case of $\gamma_{kk'}^* = 0$, that is, $(k, k') \in \mathcal{A}_2^c$. Instead, we only consider the cases where $\gamma_{kk'}^*$ is zero and the two corresponding β_k^* and $\beta_{k'}^*$ are nonzero, that is, $(k, k') \in \mathcal{A}_2^c$ and $k, k' \in \mathcal{A}_1$.

Oracle Property of SHIM. The asymptotic properties of SHIM when the sample size increases are described in the following lemma and theorems. The regularity conditions (C1)–(C3) and the proofs are given in the Supplemental Material.

Lemma 1. Assume that $a_n = o(1)$ as $n \to \infty$. Then under the regularity conditions (C1)–(C3), there exists a local minimizer $\hat{\boldsymbol{\theta}}_n$ of $Q_n(\boldsymbol{\theta})$ such that $\|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*\| = O_p(n^{-1/2} + a_n)$.

Lemma 1 implies that if the tuning parameters λ_j^{β} and $\lambda_{kk'}^{\gamma}$ associated with the nonzero coefficients converge to 0 at a rate faster than $n^{-1/2}$, then there exists a local minimizer of $Q_n(\boldsymbol{\theta})$, which is $\sqrt{n}$-consistent.

Theorem 1 (Sparsity). Assume that $\sqrt{n} b_n \to \infty$ and the local minimizer $\hat{\boldsymbol{\theta}}_n$ given in Lemma 1 satisfies $\|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*\| = O_p(n^{-1/2})$. Then under the regularity conditions (C1)–(C3),

$$P\big(\hat{\boldsymbol{\beta}}_{\mathcal{A}_1^c} = 0\big) \to 1 \qquad \text{and} \qquad P\big(\hat{\boldsymbol{\gamma}}_{\mathcal{A}_2^c} = 0\big) \to 1.$$

Theorem 1 shows that SHIM can consistently remove the noise terms with probability tending to 1. Specifically, when the tuning parameters for the nonzero coefficients converge to 0 faster than $n^{-1/2}$ and those for zero coefficients are big enough so that $\sqrt{n} a_n \to 0$ and $\sqrt{n} b_n \to \infty$, then Lemma 1 and Theorem 1 imply that the $\sqrt{n}$-consistent estimator $\hat{\boldsymbol{\theta}}_n$ satisfies $P(\hat{\boldsymbol{\theta}}_{\mathcal{A}^c} = 0) \to 1$.

Theorem 2 (Asymptotic normality). Assume that $\sqrt{n} a_n \to 0$ and $\sqrt{n} b_n \to \infty$. Then under the regularity conditions (C1)–(C3), the component $\hat{\boldsymbol{\theta}}_{\mathcal{A}}$ of the local minimizer $\hat{\boldsymbol{\theta}}_n$ given in Lemma 1 satisfies

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_{\mathcal{A}} - \boldsymbol{\theta}_{\mathcal{A}}^*) \to_d \mathrm{N}(0, \mathbf{I}^{-1}(\boldsymbol{\theta}_{\mathcal{A}}^*)),$$

where $\mathbf{I}(\boldsymbol{\theta}_{\mathcal{A}}^*)$ is the Fisher information matrix of $\boldsymbol{\theta}_{\mathcal{A}}$ at $\boldsymbol{\theta}_{\mathcal{A}} = \boldsymbol{\theta}_{\mathcal{A}}^*$ assuming that $\boldsymbol{\theta}_{\mathcal{A}^c}^* = 0$ is known in advance.

In Theorem 2, we find that the SHIM estimates for nonzero coefficients in the true model have the same asymptotic distribution as they would have if the zero coefficients were known in advance. Therefore, based on Theorems 1 and 2, we can conclude that asymptotically SHIM performs as well as if the true underlying model were given in advance, that is, it has the oracle property (Fan and Li 2001), when the tuning parameters satisfy the conditions $\sqrt{n} a_n \to 0$ and $\sqrt{n} b_n \to \infty$.

Now the remaining question is how we specify the adaptive weights so that the conditions $\sqrt{n} a_n \to 0$ and $\sqrt{n} b_n \to \infty$ are satisfied. It turns out that the Adaptive(MLE) weights introduced in Section 2.1 satisfy those conditions. Following the idea in Wang, Li, and Tsai (2007a), let

$$\lambda_j^{\beta} = \frac{\log(n)}{n} \lambda_\beta w_j^{\beta} = \frac{\log(n)}{n} \lambda_\beta \left| \frac{1}{\hat{\beta}_j^{MLE}} \right|,$$

$$\lambda_{kk'}^{\gamma} = \frac{\log(n)}{n} \lambda_\gamma w_{kk'}^{\gamma} = \frac{\log(n)}{n} \lambda_\gamma \left| \frac{\hat{\beta}_k^{MLE} \cdot \hat{\beta}_{k'}^{MLE}}{\hat{\alpha}_{kk'}^{MLE}} \right|.$$

Using the fact that $\hat{\boldsymbol{\beta}}^{MLE}$ and $\hat{\boldsymbol{\alpha}}^{MLE}$ are $\sqrt{n}$-consistent estimates of $\boldsymbol{\beta}^*$ and $\boldsymbol{\alpha}^*$, it can be easily shown that the tuning parameters λ_j^{β} and $\lambda_{kk'}^{\gamma}$ defined above satisfy the conditions for the oracle property. Therefore, we can conclude that by tuning the two regularization parameters λ_β and λ_γ and using the prespecified weights Adaptive(MLE), SHIM asymptotically possesses the oracle property.

3.2 Asymptotic Oracle Property When $p_n \to \infty$ as $n \to \infty$

In this section, we consider the asymptotic behavior of SHIM when the number of predictors p_n is allowed to approach infinity as well as the sample size n. Similar to that of Fan and Peng (2004), we show that under certain regularity conditions, SHIM still possesses the oracle property.

We first redefine some notations because now the number of predictors p_n changes with the sample size n. We denote the total number of parameters $q_n = (p_n + 1)p_n/2$. We add a subscript n to $\mathbf{V}$, $f(\cdot, \cdot)$ and $\boldsymbol{\theta}$ to denote that these quantities now change with n. Similarly for $\mathcal{A}_1$, $\mathcal{A}_2$, and $\mathcal{A}$ which are defined in Section 3.1, and we let $s_n = |\mathcal{A}_n|$.

Oracle Property of SHIM. The asymptotic properties of SHIM when the number of predictors increases as well as the sample size are described in the following lemma and theorems. The regularity conditions (C4)–(C6) and the proofs are given in the Supplemental Material.

Lemma 2. Assume that the density $f_n(\mathbf{V}_n, \boldsymbol{\theta}_n^*)$ satisfies the regularity conditions (C4)–(C6). If $\sqrt{n}a_n \to 0$ and $q_n^5/n \to 0$ as $n \to \infty$, then there exists a local minimizer $\hat{\boldsymbol{\theta}}_n$ of $Q_n(\boldsymbol{\theta}_n)$ such that

$$\|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_n^*\| = O_p\big(\sqrt{q_n}\big(n^{-1/2} + a_n\big)\big).$$

Theorem 3. Suppose that the density $f_n(\mathbf{V}_n, \boldsymbol{\theta}_n^*)$ satisfies the regularity conditions (C4)–(C6). If $\sqrt{nq_n}a_n \to 0$, $\sqrt{n/q_n}b_n \to \infty$, and $q_n^5/n \to 0$ as $n \to \infty$, then with probability tending to 1, the $\sqrt{n/q_n}$-consistent local minimizer $\hat{\boldsymbol{\theta}}_n$ in Lemma 2 satisfies the following:

(a) *Sparsity*. $\hat{\boldsymbol{\theta}}_{n\mathcal{A}_n^c} = \mathbf{0}$.
(b) *Asymptotic normality*.

$$\sqrt{n}\mathbf{A}_n\mathbf{I}_n^{1/2}\big(\boldsymbol{\theta}_{n\mathcal{A}_n}^*\big)\big(\hat{\boldsymbol{\theta}}_{n\mathcal{A}_n} - \boldsymbol{\theta}_{n\mathcal{A}_n}^*\big) \to_d \mathrm{N}(\mathbf{0}, \mathbf{G}),$$

where $\mathbf{A}_n$ is an arbitrary $m \times s_n$ matrix with a finite m such that $\mathbf{A}_n\mathbf{A}_n^\top \to \mathbf{G}$ and $\mathbf{G}$ is a $m \times m$ nonnegative symmetric matrix and $\mathbf{I}_n(\boldsymbol{\theta}_{n\mathcal{A}_n}^*)$ is the Fisher information matrix of $\boldsymbol{\theta}_{n\mathcal{A}_n}$ at $\boldsymbol{\theta}_{n\mathcal{A}_n} = \boldsymbol{\theta}_{n\mathcal{A}_n}^*$.

Note that because the dimension of $\hat{\boldsymbol{\theta}}_{n\mathcal{A}_n}$ approaches to infinity as the sample size n grows, for asymptotic normality of SHIM estimates, we consider an arbitrary linear combination $\mathbf{A}_n\hat{\boldsymbol{\theta}}_{n\mathcal{A}_n}$, where $\mathbf{A}_n$ is an arbitrary $m \times s_n$ matrix with a finite m.

The AE pointed out the reference to Zou and Zhang (2009), in which the asymptotic oracle property of Elastic-Net was studied. We note that the condition for the asymptotic oracle property in Zou and Zhang (2009) is weaker, in the sense that they only require $p_n/n^v \to 0$, where $0 < v < 1$, while we require $q_n^5/n \to 0$. The main reason that Zou and Zhang (2009) could achieve a better rate is that in Elastic-Net, when the regularization parameter for the L_1-norm penalty is set to zero, Elastic-Net reduces to the ridge regression, for which one can obtain an analytic solution, hence achieve a tighter bound on $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|$. Unfortunately, the SHIM criterion is more complicated than Elastic-Net, and we could not obtain an analytic solution when either of the regularization parameters is set to zero.

Similar as in Section 3.1, now the remaining question is whether the Adaptive(MLE) weights introduced in Section 2.1 satisfy the conditions for the oracle property. Let

$$\lambda_{nj}^{\beta} = \frac{\log(n)q_n}{n}\lambda_\beta w_j^\beta = \frac{\log(n)q_n}{n}\lambda_\beta\left|\frac{1}{\hat{\beta}_j^{MLE}}\right|,$$

$$\lambda_{n,kk'}^{\gamma} = \frac{\log(n)q_n}{n}\lambda_\gamma w_{kk'}^\gamma = \frac{\log(n)q_n}{n}\lambda_\gamma\left|\frac{\hat{\beta}_k^{MLE}\cdot\hat{\beta}_{k'}^{MLE}}{\hat{\alpha}_{kk'}^{MLE}}\right|.$$

Using the fact that $\hat{\boldsymbol{\beta}}^{MLE}$ and $\hat{\boldsymbol{\alpha}}^{MLE}$ are $\sqrt{n/q_n}$-consistent estimates of $\boldsymbol{\beta}^*$ and $\boldsymbol{\alpha}^*$ and assuming $q_n^4/n \to 0$, it can be easily shown that the tuning parameters λ_{nj}^{β} and $\lambda_{n,kk'}^{\gamma}$ defined above satisfy the conditions for the oracle property: $\sqrt{nq_n}a_n \to 0$ and $\sqrt{n/q_n}b_n \to \infty$. Therefore, we can conclude that by tuning the two regularization parameters λ_β and λ_γ and using the prespecified weights Adaptive(MLE), SHIM asymptotically possesses the oracle property.

4. NUMERICAL RESULTS

4.1 Simulation Study

In this section, we use simulation data to demonstrate the efficacy of SHIM, and compare the results with those of LASSO, a method that does not guarantee the heredity constraint. Furthermore, we compare the performance of SHIM with two other methods, Yuan, Joseph, and Lin (2007) and Zhao, Rocha, and Yu (2009), which also address the variable selection problem with heredity constraint.

We mimicked and extended the simulations in Zhao, Rocha, and Yu (2009). There are $p = 10$ predictors with only the first 4 affecting the response. The total number of candidate terms (including all possible two-way interaction terms) is $p + p(p-1)/2 = 55$. Each of the 10 predictors is normally distributed with mean zero and variance one. The 10 predictors are generated either independently or with correlation $\mathrm{Corr}(X_j, X_{j'}) = 0.5^{|j-j'|}$. With each of independent predictors and correlated predictors, we considered five different cases with coefficients shown in Table 1. The signal to noise ratio (SNR) was set to 4.0 in every case.

Case 1 is a model with no interaction effect; Case 2 is a model with interaction effects of moderate size; Case 3 represents a model with interaction effects of large size; and Case 4 is a model where the size of interaction effects is larger than that of the main effects. Case 5 is a model that does not even obey the heredity constraint.

We generated $n = 200$ training observations from each of the above models. To select the tuning parameters for SHIM and other methods, we considered three criteria: GCV, BIC, and the validation error on a separate validation set with $m = 200$ observations. The three criteria are defined as follows:

$$\mathrm{GCV} = \frac{\hat{\sigma}^2}{(1 - df/n)^2},$$

$$\mathrm{BIC} = \log\hat{\sigma}^2 + (df)(\log n)/n,$$

$$\text{Validation Error} = \frac{1}{m}\sum_{i=1}^{m}(y_i^{\mathrm{val}} - \hat{f}(\mathbf{x}_i^{\mathrm{val}}))^2,$$

where $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(\mathbf{x}_i))^2$ and df is the number of nonzero estimates. We simulated 100 replicates. For each replicate and each method, we selected three models that minimize GCV, BIC, and the validation error respectively.

In the following subsections, we compare our method SHIM with other methods in terms of the prediction accuracy and the variable selection performance.

Table 1. Simulation study: coefficients of the true models

	x_1	x_2	x_3	x_4	x_1x_2	x_1x_3	x_1x_4	x_2x_3	x_2x_4	x_3x_4
Case 1	7	2	1	1	0	0	0	0	0	0
Case 2	7	2	1	1	1.0	0	0	0.5	0.4	0.1
Case 3	7	2	1	1	7	7	7	2	2	1
Case 4	7	2	1	1	14	14	14	4	4	2
Case 5	0	0	0	0	7	7	7	2	2	1

Prediction Performance

We first compare the prediction accuracy of SHIM with those of other methods: Oracle, OLS, LASSO, CAP, and CARDS. "Oracle" refers to the OLS applied only to the relevant terms, which serves as an optimal bench mark. CAP and CARDS refer to Zhao, Rocha, and Yu (2009) and Yuan, Joseph, and Lin (2007) that also address the heredity constraint. Specifically, Yuan, Joseph, and Lin (2007) extends the LARS algorithm (Efron et al. 2004), and Zhao, Rocha, and Yu (2009) suggests a Composite Absolute Penalty (CAP) to enforce the heredity constraint.

We computed the mean squared error (MSE) on a test set with 10,000 observations to measure the prediction accuracy. We found that for the purpose of prediction, validation error performs the best, GCV is the next, and BIC performs the worst among the three criteria for all methods. Validation error, however, is not always available in real data. Therefore, we chose to report the prediction performance based on both validation error and GCV. Figures 1 and 2 show boxplots of the 100 MSEs from 100 replicates for both independent and correlated cases. We can see that both LASSO and SHIM perform much better than OLS; this illustrates that some regularization or shrinkage is crucial for prediction accuracy. Furthermore, SHIM seems to perform consistently better than LASSO. This is observed not only in Cases 1–4, where the true models obey the heredity constraint, but also in Case 5, where the true model does not obey the heredity constraint. In Case 5, although SHIM would never select the right model and would estimate some irrelevant main terms with nonzero coefficients (in order to include the relevant interaction terms), the magnitude of these estimates are small (due to the flexibility of having two tuning parameters), hence the prediction accuracy is not jeopardized too much. This result is not surprising if one notices that prediction and variable selection are two different aspects of model fitting.

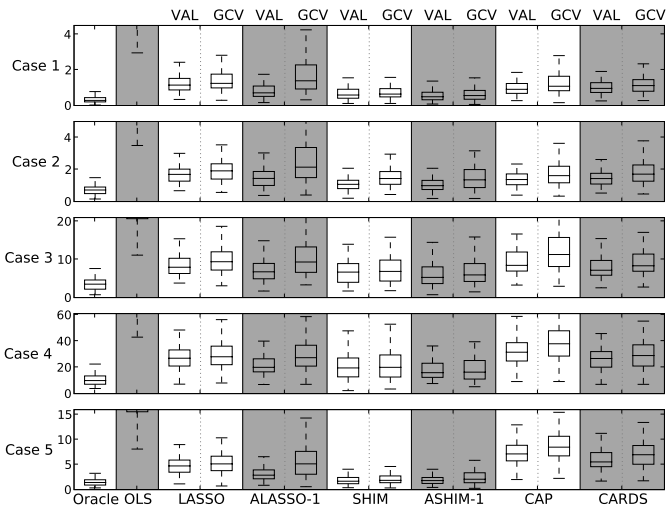

Figure 1. Simulation results: the boxplots of MSE values in independent cases. "VAL" refers to when we select the tuning parameters based on validation errors and "GCV" refers to when we use the GCV criterion. "ALASSO-1" and "ASHIM-1" respectively refer to the adaptive LASSO and the adaptive SHIM with the weights based on the OLS estimates; "CAP" refers to Zhao, Rocha, and Yu (2009) and "CARDS" refers to Yuan, Joseph, and Lin (2007). As a benchmark, "Oracle" refers to the OLS applied only to the relevant terms.

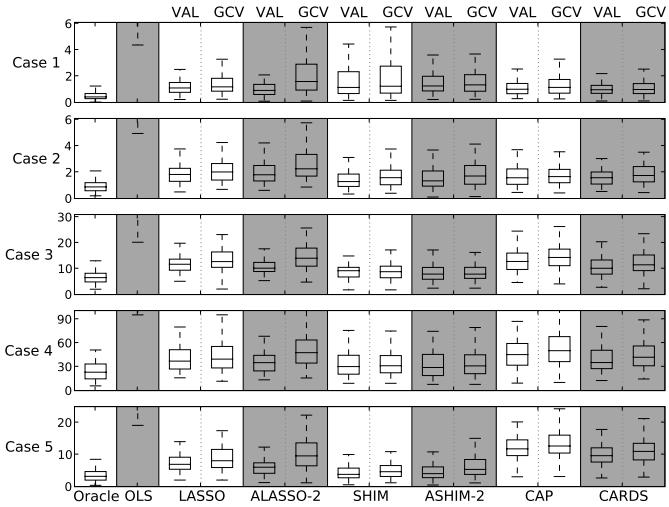

Figure 2. Simulation results: the boxplots of MSE values in correlated cases. "VAL" refers to when we select the tuning parameters based on validation errors and "GCV" refers to when we use the GCV criterion. "ALASSO-2" and "ASHIM-2" respectively refer to the adaptive LASSO and the adaptive SHIM with the weights based on the ridge regression estimates; "CAP" refers to Zhao, Rocha, and Yu (2009) and "CARDS" refers to Yuan, Joseph, and Lin (2007). As a benchmark, "Oracle" refers to the OLS applied only to the relevant terms.

Comparing SHIM (nonadaptive version) with CAP (Zhao, Rocha, and Yu 2009) and CARDS (Yuan, Joseph, and Lin 2007), we can see that the prediction accuracy of SHIM is consistently better than CARDS and CAP in both independent and correlated cases, especially when the effect of interaction terms is large compared to the effect of main terms.

Variable Selection Performance

We also compare the variable selection performance of SHIM with those of the other methods.

Following Wang, Li, and Tsai (2007b), we define "underfitted," "correctly fitted," and "overfitted" models. Suppose there are q candidate terms and only $q_0 \le q$ number of relevant terms in the true model. We let $I_F = \{1, 2, \ldots, q\}$ denote the index set of the full model, $I_T = \{j_1, j_2, \ldots, j_{q_0}\}$ denote the index set of the true model, and I denote the index set of a selected model. Then we define a model as underfitted if $I_T \nsubseteq I$, overfitted if $I_T \subsetneq I$, and correctly fitted if $I = I_T$.

For the purpose of variable selection, we found that BIC outperforms GCV and validation error, which agrees with the discussion in Wang, Li, and Tsai (2007b), Wang and Leng (2007) and Zhang and Lu (2007). Variable selection results based on BIC are shown in Table 2. We can see that when there is no interaction effect (Case 1), the three heredity methods, that is, SHIM, CAP, and CARDS, perform similarly and all better than LASSO. When the interaction effects are relatively strong (Cases 3 and 4), SHIM tends to select the correct model more often than other methods. When the interaction effects are weak (Case 2), hence easily missed, none of the methods is able to select the exact correct model. In Case 5, since the true model does not obey the heredity constraint, SHIM, CAP, and CARDS can never identify the correct model.

Table 2. Simulation results: variable selection based on BIC. "Underfitted," "Correctly fitted," and "Overfitted" respectively represent the numbers of replicates that are underfitted, correctly fitted, and overfitted among the 100 replicates. "ALASSO" and "ASHIM" refer to the adaptive LASSO and the adaptive SHIM with the OLS weights for independent cases and the ridge regression weights for correlated cases; "CAP" refers to Zhao, Rocha, and Yu (2009) and "CARDS" refers to Yuan, Joseph, and Lin (2007)

		LASSO	ALASSO	SHIM	ASHIM	CAP	CARDS
				Independent cases			
Case 1	Underfitted	21	30	3	14	23	14
	Correctly fitted	23	27	17	49	42	44
	Overfitted	56	43	80	37	35	42
Case 2	Underfitted	100	100	98	98	97	99
	Correctly fitted	0	0	0	0	1	0
	Overfitted	0	0	2	2	2	1
Case 3	Underfitted	93	97	17	20	41	59
	Correctly fitted	0	0	78	78	17	13
	Overfitted	7	3	5	2	42	28
Case 4	Underfitted	99	100	10	19	29	44
	Correctly fitted	0	0	87	76	12	17
	Overfitted	1	0	3	5	59	39
Case 5	Underfitted	50	64	1	6	29	42
	Correctly fitted	2	10	0	0	0	0
	Overfitted	48	26	99	94	71	58
				Correlated cases			
Case 1	Underfitted	18	48	22	65	19	17
	Correctly fitted	38	26	53	31	54	61
	Overfitted	44	26	25	4	27	22
Case 2	Underfitted	95	100	88	93	85	93
	Correctly fitted	1	0	0	1	1	3
	Overfitted	4	0	12	6	14	4
Case 3	Underfitted	91	99	9	27	22	44
	Correctly fitted	1	0	88	68	29	36
	Overfitted	8	1	3	5	49	20
Case 4	Underfitted	98	100	3	22	16	33
	Correctly fitted	0	0	97	72	31	38
	Overfitted	2	0	0	6	53	29
Case 5	Underfitted	59	83	1	14	16	33
	Correctly fitted	15	6	0	0	0	0
	Overfitted	26	11	99	86	84	67

To further assess the variable selection results, we also plotted the "sensitivity" and "specificity" of selected models via BIC in Figures 3 and 4, where sensitivity and specificity are defined as follows:

$$\text{sensitivity} = \text{the proportion of the number of selected relevant terms to the total number of relevant terms}$$

and

$$\text{specificity} = \text{the proportion of the number of unselected irrelevant terms to the total number of irrelevant terms.}$$

Therefore, a dot in a figure located at the upper left corner would mean that the corresponding method (via BIC) effectively selects relevant terms and removes irrelevant terms. We can see that overall models selected by SHIM are closer to the upper left corner than other methods, especially when the effects of interaction terms are relatively large. In Case 5 (interaction-only model), because SHIM, CAP, and CARDS select relevant interaction terms as well as irrelevant main terms due to the heredity constraint, their specificities are slightly lower than that of LASSO, but the sensitivities remain high, especially for SHIM. Overall, SHIM (via BIC) seems to perform better than LASSO, CAP, and CARDS on variable selection, and the performance of SHIM also seems to be more stable than those of other methods (Figures 3 and 4).

4.2 Analyzing Designed Experiments Using SHIM

In designed experiments, economic considerations may compel the investigator to use few experiments (runs). Many efficient experimental designs have been proposed in the literature. Among them fractional factorial designs are thoroughly studied and widely used. While the design of experiments literature is replete with research on the construction of the efficient designs, the methodologies of analysis have not received the same amount of attention. Traditional analysis methods (e.g., stepwise, all subset) continue to be a dominating choice for researchers in the DOE area. Wu and Hamada (2000) stated three principles in the analysis of the designed experiment: ef-

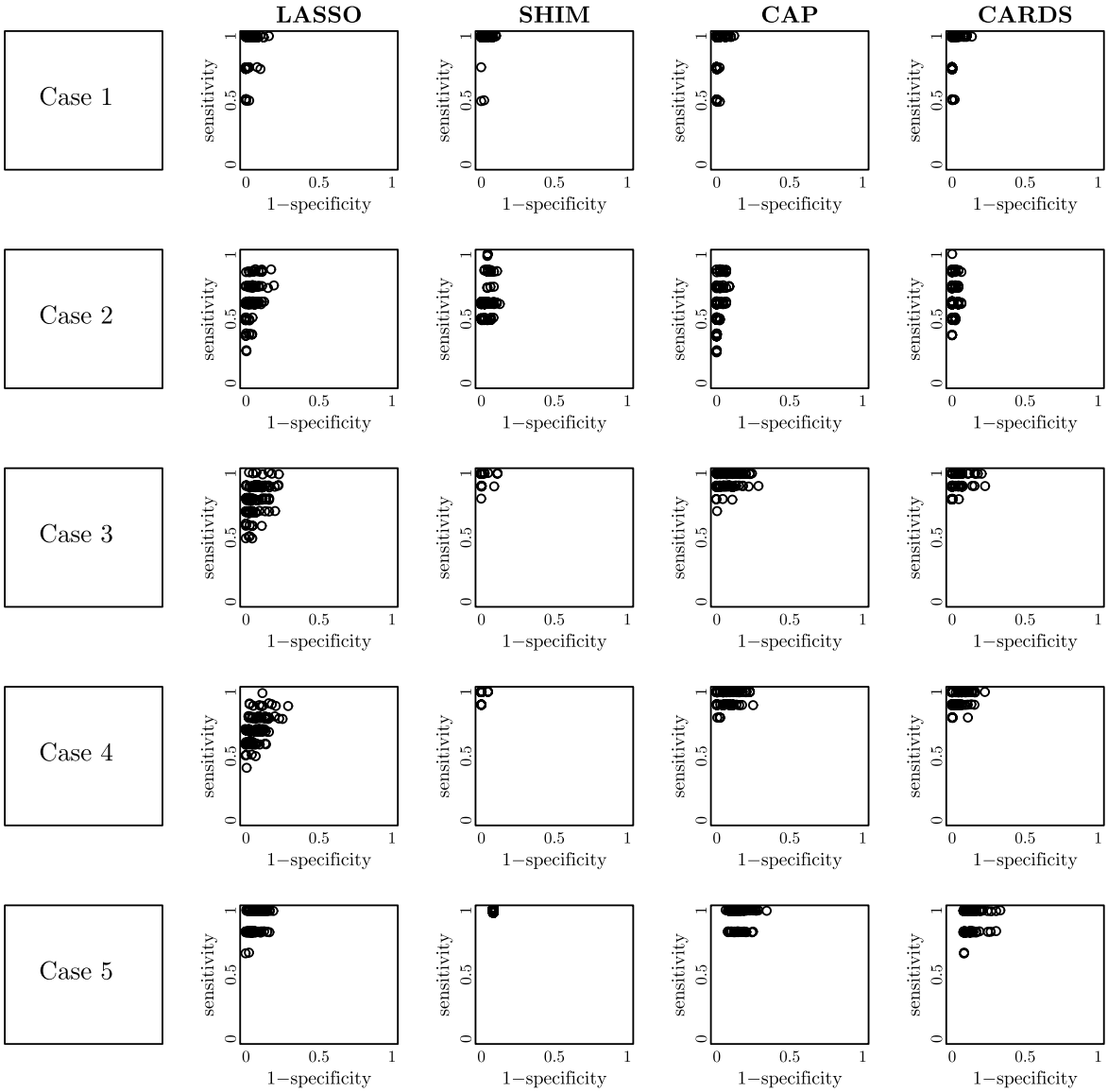

Figure 3. Simulation results: sensitivity and 1-specificity of the selected models based on BIC in independent cases. Each dot corresponds to a replicate among 100 replicates. "CAP" refers to Zhao, Rocha, and Yu (2009) and "CARDS" refers to Yuan, Joseph, and Lin (2007).

fect sparsity (i.e., only a few of all candidate factors are active), effect hierarchy (e.g., main effects are more likely to be significant than two-factor interactions), and effect heredity (e.g., two-factor interaction x_1x_2 should be in the model only if the main effects x_1 and x_2 are also in the model).

The proposed method appears to be particularly suitable for analyzing the designed experiments, as SHIM encourages effect sparsity and requires effect heredity in the model. In this section we explore the use of SHIM in analyzing designed experiments. We consider a simulation study, in which a minimum-aberration 2^{6-2}_{IV} design was used to generate simulated data. Six two-level factors are studied in a 16-run design, which is defined by $x_5 = x_1x_2x_3$ and $x_6 = x_1x_2x_4$. Similar to those in Table 1, four cases of model are considered and shown in Table 3.

To assess SHIM, we generated 1000 simulations and recorded (1-specificity, sensitivity) as defined in Section 4.1. In each simulation, the data are generated by using the true models of Table 3, plus a random error of N(0, 1). We then compare SHIM with LASSO, CARDS, and CAP. The results based on BIC-selected models are shown in Figure 5. One can see that SHIM performs consistently better than other methods in terms of removing irrelevant effects, especially when the heredity property is relatively strong (Cases 3 and 4).

4.3 Real Data Analysis

In this section, we apply our method SHIM to a real dataset. This dataset was from Hung et al. (2004) for a case-control study of bladder cancer. It consists of the genotypes on 14 loci and the status of smoking behavior for 201 bladder cancer patients and 214 controls. Four of the genotypes are two-level factors, nine are three-level factors, and one is a five-level factor. We represent all genotypes with dummy variables, hence a total of $4 + 2 \times 9 + 4 = 26$ dummy variables. The status of smoking behavior is represented with two predictors: one is a three-level factor (nonsmoker, light smoker, and heavy smoker), and the other is a continuous variable, measuring the number of

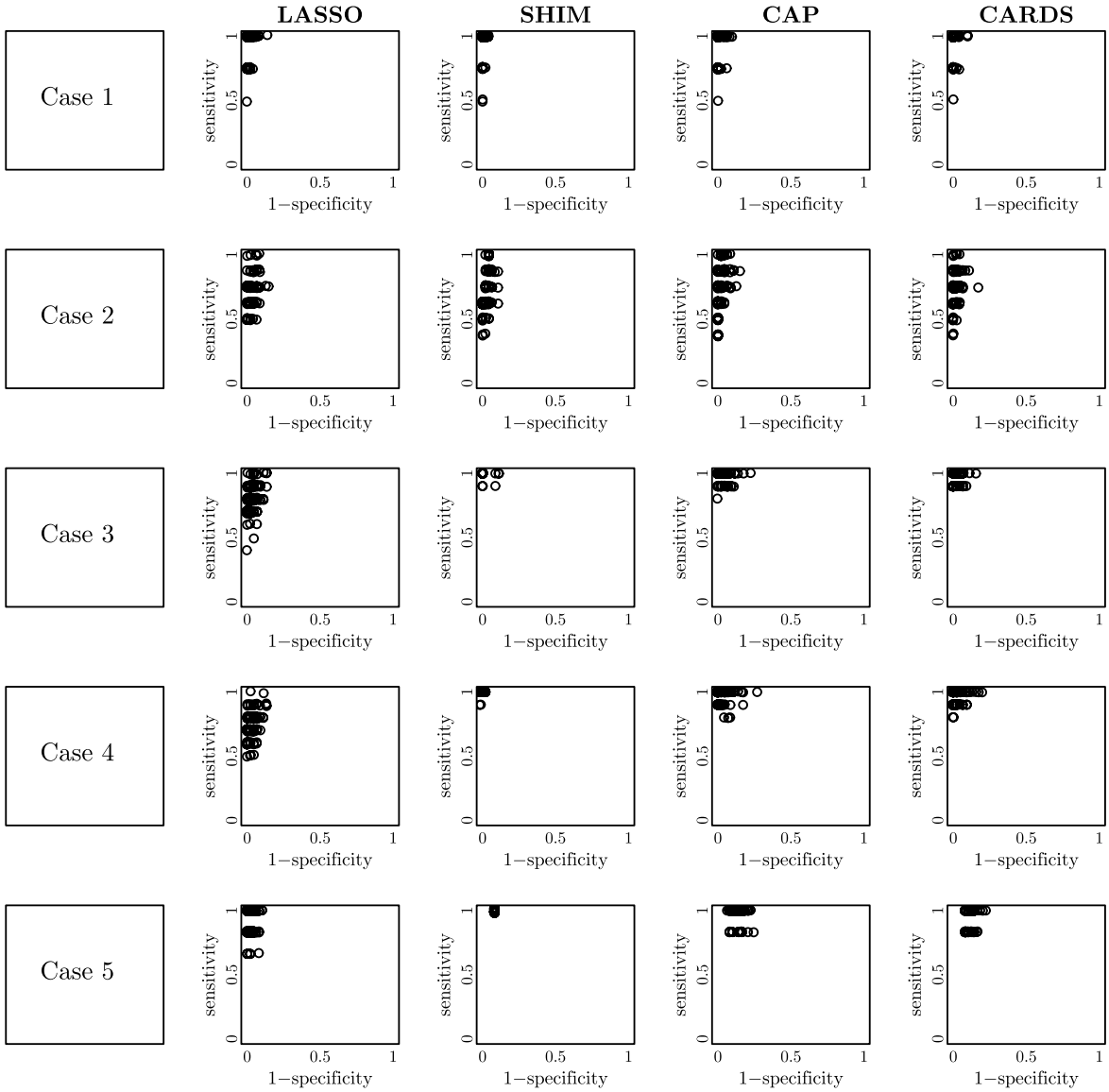

Figure 4. Simulation results: sensitivity and 1-specificity of the selected models based on BIC in correlated cases. Each dot corresponds to a replicate among 100 replicates. "CAP" refers to Zhao, Rocha, and Yu (2009) and "CARDS" refers to Yuan, Joseph, and Lin (2007).

packs consumed per year. Since the response variable is binary (case/control), we used the negative binomial log-likelihood as the loss function rather than the squared error.

We randomly split the data into training ($n = 315$) and testing ($N = 100$). Tuning parameters were chosen via five-fold cross-validation based on the training data. Fitted models were evaluated on the testing data, with the classification rule given by $\operatorname{sgn}(\hat{g}(\mathbf{x}))$.

We considered three cases. In the first case, we used only the genetic information, that is, the 14 loci genetic factors. There are a total of 336 candidate terms, including the main terms and all possible two-way interaction terms (between two different loci). In the second case, we considered the 14 genetic factors and the categorical smoke status. There are a total of 390 candidate terms, including all possible two-way interaction terms among the genetic factors and the interaction terms between genetic factors and the categorical smoke status. In the third case, we replaced the categorical smoke status with the continuous smoke status, where we considered the interactions between genetic factors and the continuous smoke status. For comparison, we fitted both LASSO and SHIM in each case. We used Adaptive(Ridge) as the prespecified weights because the number of terms is larger than the number of observations in the first two cases. Misclassification errors, sensitivities, and specificities (all on the test data) of these models are summarized in Table 4. As we can see, the models that use the genetic factors and the continuous smoke status perform slightly better than other models in terms of the error rate. This may be heuris-

Table 3. DOE example setting: coefficients of the true models

	x_1	x_2	x_3	x_1x_2	x_1x_3	x_2x_3
Case 1	7	2	1	0	0	0
Case 2	7	2	1	1	0	0
Case 3	7	2	1	7	7	7
Case 4	7	2	1	14	14	14

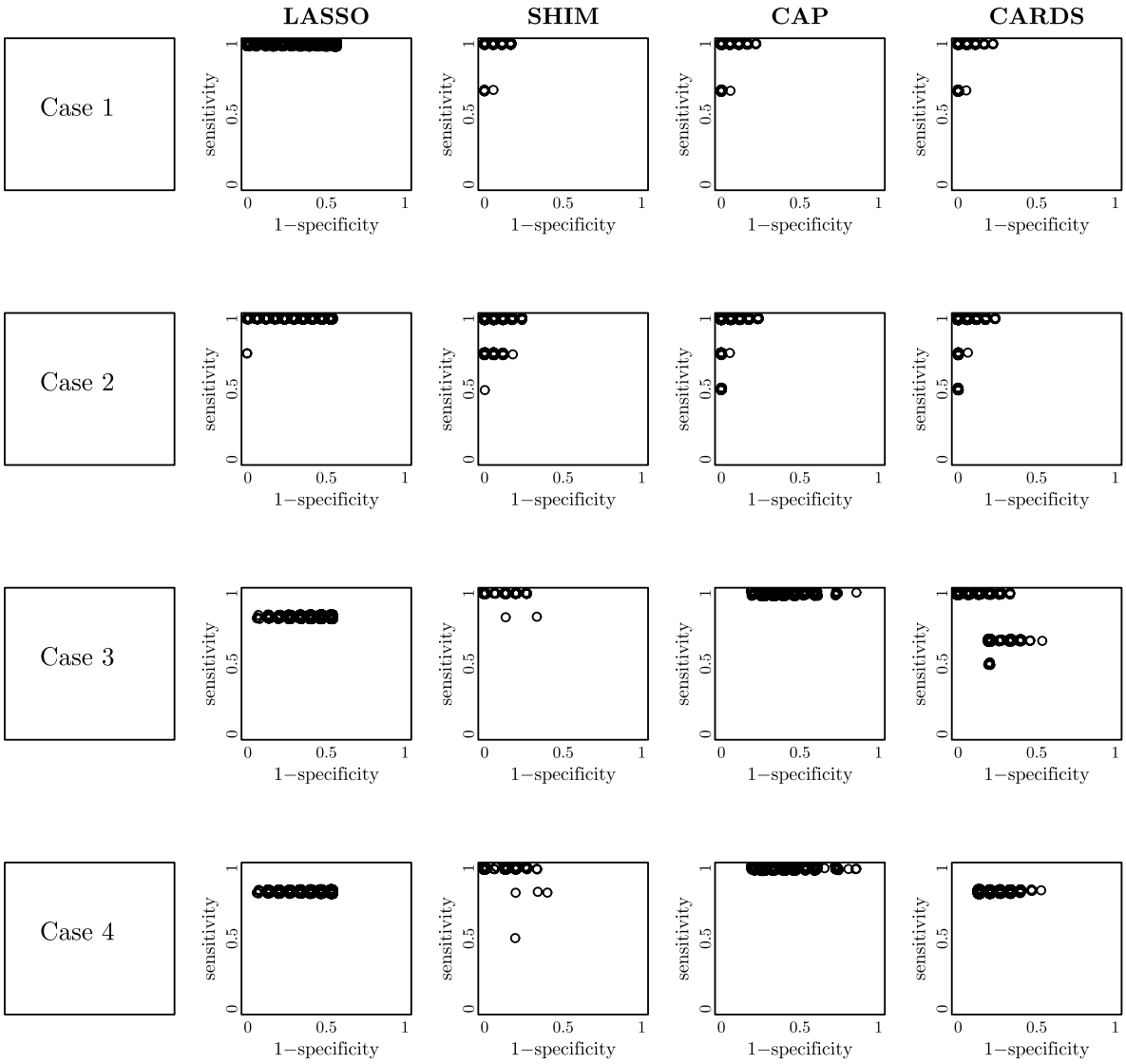

Figure 5. DOE example results: sensitivity and 1-specificity of the selected models based on BIC. Each dot corresponds to a replicate among 1000 replicates. "CAP" refers to Zhao, Rocha, and Yu (2009) and "CARDS" refers to Yuan, Joseph, and Lin (2007).

tically understood as that the continuous smoke-status contains more information than the categorical smoke status.

We then focused on the third case. Terms selected by Adaptive-LASSO and Adaptive-SHIM are shown in the upper part of Table 5. Notice that both methods selected the smoke status (*PackYear*), *GSTM*1, and *MPO*. LASSO also selected an interaction term, $NQO1 \times PackYear$, but it does not obey the heredity constraint; on the other hand, SHIM selected the main term *NQO*1, but not the interaction term.

To further assess the terms that were selected, we applied a bootstrap analysis. The lower part of Table 5 summarizes the terms that were selected with selection frequency higher than 30% based on $B = 100$ bootstrap samples. As we can see, the five terms selected by SHIM using the training data are the only five terms that had the selection frequency higher than 30% in bootstrap samples. So SHIM is fairly stable in terms of selecting terms. We can also see that the smoke status was always selected, followed immediately by *MPO*. The interaction term $NQO1 \times PackYear$ was selected half of the time by LASSO, but never by SHIM; instead, SHIM selected the main term *NQO*1 half of the time.

These results seem to be consistent with the findings in Hung et al. (2004). The five terms selected by SHIM are among the ones that were shown to have a significant effect on increasing the risk of bladder cancer in Hung et al. (2004).

5. CONCLUSION

In this paper, we have extended the LASSO method for simultaneously fitting a regression model and identifying interaction terms. The proposed method automatically enforces the heredity constraint. In addition, it enjoys the "oracle" property under mild regularity conditions. We demonstrate that our new method tends to remove irrelevant variables more effectively and provide better prediction performance than the classical LASSO method, as well as two other more recent work.

The heredity that we have considered in this paper is the so-called strong heredity, that is, an interaction term can be included in the model only if both of the corresponding main

Table 4. Real data analysis results: misclassification error, sensitivity, and specificity on the test data

		Misclassification error	Sensitivity	Specificity
SHIM using the genetic factors				
LASSO	Plain	0.44	0.48	0.63
	Adaptive	0.41	0.52	0.65
SHIM	Plain	0.36	0.54	0.73
	Adaptive	0.38	0.46	0.77
SHIM using the genetic factors and the categorical smoke status variable				
LASSO	Plain	0.35	0.58	0.71
	Adaptive	0.37	0.56	0.69
SHIM	Plain	0.35	0.65	0.65
	Adaptive	0.34	0.65	0.67
SHIM using the genetic factors and the continuous smoke status variable				
LASSO	Plain	0.34	0.60	0.71
	Adaptive	0.32	0.67	0.69
SHIM	Plain	0.33	0.67	0.67
	Adaptive	0.32	0.65	0.71

terms are also included in the model. There is another type of heredity, weak heredity (Hamada and Wu 1992), in which only one of the main terms is required to be present when an interaction term is included in the model. Extending our SHIM framework to enforce the weak heredity is straightforward: instead of reparameterizing the coefficient for $x_j x_{j'}$ as the product $\gamma_{jj'}\beta_j\beta_{j'}$, we may write it as $\gamma_{jj'}(|\beta_j| + |\beta_{j'}|)$. So if the coefficient for $x_j x_{j'}$ is not equal to zero, it implies that at least one of β_j and $\beta_{j'}$ is not equal to zero.

Table 5. Real data analysis results: the upper part lists the terms that were selected using the training data, and the lower part lists the terms that were selected (with selection frequency higher than 30%) based on 100 bootstrap samples. The numbers in the parentheses are the corresponding selection frequencies out of $B = 100$ bootstrap samples. LASSO and SHIM were used with the genetic factors and the continuous smoke-status variable

Adaptive LASSO	Adaptive SHIM
Selected terms using the training data	
PackYear	*PackYear*
*GSTM*1	*GSTM*1
MPO	*MPO*
(*NQO*1) × (*PackYear*)	*NQO*1
–	*MnSOD*
Selected terms using 100 bootstrap samples	
PackYear (100%)	*PackYear* (100%)
MPO (78%)	*MPO* (82%)
(*NQO*1) × (*PackYear*) (49%)	*GSTM*1 (57%)
*GSTM*1 (43%)	*NQO*1 (46%)
*NQO*1 (37%)	*MnSOD* (40%)
MnSOD (36%)	–
(*COMT*) × (*PackYear*) (35%)	–
(*MPO*) × (*PackYear*) (32%)	–
(*XRCC*1) × (*PackYear*) (30%)	–

SUPPLEMENTAL MATERIALS

Conditions and proof: Part 1 describes the regularity conditions that are needed for the asymptotic oracle properties of SHIM (Lemmas 1 and 2, Theorem 1–3) shown in Section 3. In Part 2, the proof of those properties is provided. (shim-supp.pdf)

[Received May 2008. Revised September 2009.]

REFERENCES

Breiman, L. (1995), "Better Subset Regression Using the Non-Negative Garrote," *Technometrics*, 37, 373–384. [354,355]

Chipman, H. (1996), "Bayesian Variable Selection With Related Predictors," *Canadian Journal of Statistics*, 24, 17–36. [354]

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004), "Least Angle Regression" (with discussion), *The Annals of Statistics*, 32 (2), 407–499. [355,359]

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. [354-357]

Fan, J., and Peng, H. (2004), "Nonconcave Penalized Likelihood With a Diverging Number of Parameters," *The Annals of Statistics*, 32 (3), 928–961. [355,357]

Friedman, J., Hastie, T., Hofling, H., and Tibshirani, R. (2007), "Pathwise Coordinate Optimization," *The Annals of Applied Statistics*, 1 (2), 302–332. [355]

Fu, W. (1998), "Penalized Regressions: The Bridge versus the Lasso," *Journal of Computational and Graphical Statistics*, 7 (3), 397–416. [355,356]

Hamada, M., and Wu, C. (1992), "Analysis of Designed Experiments With Complex Aliasing," *Journal of Quality Technology*, 24 (3), 130–137. [354,364]

Hung, R., Brennan, P., Malaveille, C., Porru, S., Donato, F., Boffetta, P., and Witte, J. (2004), "Using Hierarchical Modeling in Genetic Association Studies With Multiple Markers: Application to a Case-Control Study of Bladder Cancer," *Cancer Epidemiology, Biomarkers & Prevention*, 13, 1013–1021. [361,363]

Joseph, V. (2006), "A Bayesian Approach to the Design and Analysis of Fractionated Experiments," *Technometrics*, 48 (2), 219–229. [354]

McCullagh, P., and Nelder, J. (1989), *Generalized Linear Models*, London: Chapman & Hall/CRC. [354]

Nelder, J. (1994), "The Statistics of Linear Models: Back to Basics," *Statistics and Computing*, 4, 221–234. [354]

Shen, X., and Ye, J. (2002), "Adaptive Model Selection," *Journal of the American Statistical Association*, 97, 210–221. [354]

Tibshirani, R. (1996), "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society, Ser. B*, 58, 267–288. [354]

van der Vaart, A. (1998), *Asymptotic Statistics*, Cambridge, U.K.: Cambridge University Press.

Wang, H., and Leng, C. (2007), "Unified LASSO Estimation by Least Squares Approximation," *Journal of the American Statistical Association*, 102 (479), 1039–1048. [359]

Wang, H., Li, G., and Jiang, G. (2007), "Robust Regression Shrinkage and Consistent Variable Selection via the LAD-LASSO," *Journal of Business & Economic Statistics*, 25, 347–355. [355]

Wang, H., Li, G., and Tsai, C. (2007a), "Regression Coefficient and Autoregressive Order Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society, Ser. B*, 69 (1), 63–78. [357]

——— (2007b), "Tuning Parameter Selectors for the Smoothly Clipped Absolute Deviation Method," *Biometrika*, 94 (3), 553–568. [359]

Wu, C., and Hamada, M. (2000), *Experiments: Planning, Analysis and Parameter Design Optimization*, New York: Wiley. [360]

Yuan, M., Joseph, V., and Lin, Y. (2007), "An Efficient Variable Selection Approach for Analyzing Designed Experiments," *Technometrics*, 49 (4), 430–439. [358-363]

Zhang, H., and Lu, W. (2007), "Adaptive-LASSO for Cox's Proportional Hazard Model," *Biometrika*, 94, 691–703. [355,359]

Zhao, P., Rocha, G., and Yu, B. (2009), "The Composite Absolute Penalties Family for Grouped and Hierarchical Variable Selection," *The Annals of Statistics*, 37, 3468–3497. [358-363]

Zou, H. (2006), "The Adaptive Lasso and Its Oracle Properties," *Journal of the American Statistical Association*, 101 (476), 1418–1429. [355]

Zou, H., and Zhang, H. (2009), "On the Adaptive Elastic-Net With a Diverging Number of Parameters," *The Annals of Statistics*, 37, 1733–1751. [358]