Shrinkage Inverse Regression Estimation for Model Free Variable Selection

Howard D. Bondell and **Lexin Li**[1]

North Carolina State University, Raleigh, USA

Summary. The family of inverse regression estimators recently proposed by Cook and Ni (2005) have proven effective in dimension reduction by transforming the high-dimensional predictor vector to its low-dimensional projections. In this article, we propose a general shrinkage estimation strategy for the entire inverse regression estimation family that is capable of simultaneous dimension reduction and variable selection. We demonstrate that the new estimators achieve consistency in variable selection without requiring any traditional model, meanwhile retaining the root-n estimation consistency of the dimension reduction basis. We also show the effectiveness of the new estimators through both simulation and real data analysis.

Keywords: Inverse regression estimation; Nonnegative garrote; Sliced inverse regression; Sufficient dimension reduction; Variable selection

[1] *Address for correspondence:* Lexin Li, Department of Statistics, North Carolina State University, Box 8203. Raleigh, NC 27695, USA. E-mail: li@stat.ncsu.edu

1. Introduction

Sufficient dimension reduction (SDR) has generated considerable interest in recent years as a way to study regression of a univariate response Y on a p-dimensional predictor X. It aims to replace the usually high-dimensional predictor vector with its low-dimensional projection onto an appropriate subspace, meanwhile preserving full regression information and imposing no parametric model. Inquiry of SDR hinges on a population meta-parameter called the *central subspace*, denoted as $\mathcal{S}_{Y|X}$, which is the minimum subspace among all subspaces $\mathcal{S} \subseteq \mathbb{R}^p$ satisfying $Y \perp\!\!\!\perp X | P_{\mathcal{S}} X$. Here $\perp\!\!\!\perp$ stands for independence and $P_{\mathcal{S}}$ denotes the orthogonal projection onto $\mathcal{S}$. Letting η denote a $p \times d$ matrix whose columns form a basis of $\mathcal{S}_{Y|X}$ and $d = \dim(\mathcal{S}_{Y|X})$, the definition of $\mathcal{S}_{Y|X}$ indicates that Y is independent of X given $\eta^{\mathsf{T}} X$, and thus $\eta^{\mathsf{T}} X$ represents the smallest number of linear combinations of X that extract all the information about $Y|X$. $\mathcal{S}_{Y|X}$ uniquely exists under mild conditions (Cook, 1996), and we assume its existence throughout this article.

The general framework of sufficient dimension reduction is also particularly useful for variable selection. Given the growing number of variables collected in experimental and observational studies, variable selection is becoming increasingly important, so that the analyst can distinguish between the relevant variables from those which are irrelevant for the particular objectives of the analysis. There has been an enormous literature on variable selection (see Miller, 2002, for a review). Most existing approaches assume that the true underlying model is known up to a finite dimensional parameter, or the imposed working model is usefully close to the true model. Selection of relevant predictors then becomes part of a process of building and selecting a good model given the observed data. However, the true model is generally unknown, and model formulation can be complex and difficult, especially when there are a large number of potential predictors. Assessing the goodness of the fitted model

can be even more elusive when model building and model selection are interweaved. Alternatively, based on the framework of SDR, Li, Cook, and Nachtsheim (2005) proposed the concept of model free variable selection, and demonstrated that it is possible to construct practically useful variable selection approaches that do *not* require any traditional model. As a result such approaches may help relieve the analyst of model building efforts before the selection considerations, and subsequent model formulation may become easier given the selected subset of variables.

There have primarily been two categories of model free variable selection approaches developed within the SDR framework. The first is test-based, including, for example, the approximate sliced inverse regression-based t test (Chen and Li, 1998), the marginal coordinate test (Cook, 2004), and the gridded chi-squared test (Li, Cook, and Nachtsheim, 2005). The tests are typically incorporated into a variable subset search procedure, e.g., a stepwise backward or forward search. However, such subset selection methods are not only computationally intensive, but may also be unsatisfactory in terms of prediction accuracy and stability (Breiman, 1995). An alternative class of model free selection methods integrate SDR with the regularization paradigm, and examples include shrinkage sliced inverse regression (Ni, Cook, and Tsai, 2005), and sparse sliced inverse regression (Li and Nachtsheim, 2006). However, theoretical properties of these shrinkage selection methods have not yet been studied.

In this article, we will first review a family of *inverse regression estimators* (IRE) of the central subspace that have been recently proposed. This family includes both long-standing SDR estimators and the more recent state-of-art estimators of $\mathcal{S}_{Y|X}$. We then integrate a general regularization paradigm with the entire IRE family, and refer to the new methods collectively as the *shrinkage inverse regression estimators*. The focus of this article is to demonstrate that the proposed class of shrinkage estimators possess the desirable property of consistency in variable selection, while simultane-

ously retaining root-n estimation consistency. Since the new methods are based on the SDR estimators that do not require any traditional model, model free variable selection is achieved along with dimension reduction. The rest of the article is organized as follows. The IRE family of SDR estimators are briefly reviewed in Section 2, followed by a rigorous definition of the notion of variable selection in the absence of a traditional model. The shrinkage inverse regression estimators are proposed in Section 3, and the asymptotic properties are studied. In Section 4, we examine the finite sample performance of some specific instance of the shrinkage inverse regression estimators through both simulation and real data analysis. We conclude the paper in Section 5 with a discussion. Technical proofs are relegated to the Appendix.

2. Sufficient Dimension Reduction and Variable Selection

2.1. Inverse regression estimator

There have been a number of estimation methods proposed to estimate the central subspace $\mathcal{S}_{Y|X}$, including sliced inverse regression (SIR, Li, 1991) and sliced average variance estimation (SAVE, Cook and Weisberg, 1991). More recently, Cook and Ni (2005) proposed a family of minimum discrepancy based inverse regression estimators for estimating $\mathcal{S}_{Y|X}$. Both SIR and SAVE can be cast into this family. Moreover, a number of new estimators within this family have been developed, for instance, the optimal inverse regression estimator (Cook and Ni, 2005), and the covariance inverse regression estimator (CIRE, Cook and Ni, 2006).

The IRE family starts with the construction of a $p \times h$ matrix $\theta = (\theta_1, \ldots, \theta_h)$ having the property that $\mathrm{Span}(\theta) = \mathcal{S}_{Y|X}$. The conditions to ensure that the columns of θ span the central subspace are typically mild, and imposed on the marginal distribution of X, rather than on $Y|X$. For this reason, the inverse regression estimators

4

are viewed as model free dimension reduction methods. Next IRE estimates the basis $\eta \in \mathbb{R}^{p \times d}$ of $\mathcal{S}_{Y|X}$ by decomposing $\theta = \eta\gamma$, where $\gamma \in \mathbb{R}^{d \times h}$. The dimension d of the central subspace can be readily estimated given the data, with method-specific asymptotic tests available. Thus d is treated as known in the subsequent development.

Given n i.i.d. observations $\{(x_i, y_i), i = 1, \ldots, n\}$ of (X, Y), IRE estimates (η, γ) by minimizing over $B \in \mathbb{R}^{p \times d}$ and $C \in \mathbb{R}^{d \times h}$ a quadratic discrepancy function

$$G(B, C) = \left\{ \text{vec}(\hat{\theta}) - \text{vec}(BC) \right\}^{\mathsf{T}} V_n \left\{ \text{vec}(\hat{\theta}) - \text{vec}(BC) \right\}, \tag{1}$$

where $\hat{\theta}$ is a usual $\sqrt{n}$-consistent estimator of θ, V_n is a consistent estimator of some user-selected positive definite matrix $V \in \mathbb{R}^{ph \times ph}$, and $\text{vec}(\cdot)$ denotes the matrix operator that stacks all columns of a matrix to a vector. Equation (1) represents a class of estimators, with its individual member determined by the choice of the pair $(\hat{\theta}, V_n)$. An alternating least squares algorithm was devised to minimize (1) (Cook and Ni, 2005). Letting $(\hat{\eta}, \hat{\gamma}) = \arg\min_{B,C} G(B, C)$, $\text{Span}(\hat{\eta})$ is a consistent *inverse regression estimator* of $\mathcal{S}_{Y|X}$, and $\hat{\vartheta} = \hat{\eta}\hat{\gamma}$ is a $\sqrt{n}$-consistent estimator of θ for any choice of positive definite V_n (Cook and Ni, 2005).

2.2. Variable selection in SDR

We next rigorously define the notion of variable selection in the absence of a traditional model. The goal of variable selection is to seek the smallest subset of the predictors $X_{\mathcal{A}}$, with partition $X = (X_{\mathcal{A}}^{\mathsf{T}}, X_{\mathcal{A}^c}^{\mathsf{T}})^{\mathsf{T}}$, such that

$$Y \perp\!\!\!\perp X_{\mathcal{A}^c} | X_{\mathcal{A}}. \tag{2}$$

Here $\mathcal{A}$ denotes a subset of indices of $\{1, \ldots, p\}$ corresponding to the relevant predictor set $X_{\mathcal{A}}$, and $\mathcal{A}^c$ is the compliment of $\mathcal{A}$.

It is important to note that the conditional independence statement (2) can be directly connected with the basis η of the central subspace $\mathcal{S}_{Y|X}$. Following the

partition of $X = (X_{\mathcal{A}}^{\mathsf{T}}, X_{\mathcal{A}^c}^{\mathsf{T}})^{\mathsf{T}}$, we can partition η accordingly as,

$$\eta = \begin{pmatrix} \eta_{\mathcal{A}} \\ \eta_{\mathcal{A}^c} \end{pmatrix}, \eta_{\mathcal{A}} \in \mathbb{R}^{(p-p_0) \times d}, \eta_{\mathcal{A}^c} \in \mathbb{R}^{p_0 \times d}.$$

Let $\mathcal{H} = \mathrm{Span}((0_{p_0 \times (p-p_0)}, I_{\mathcal{A}^c})^{\mathsf{T}})$, where $p_0 = |\mathcal{A}^c|$, and $I_{\mathcal{A}^c}$ is a $p_0 \times p_0$ identity matrix. Let $P_{\mathcal{H}}$ denote the orthogonal projection onto $\mathcal{H}$ and $Q_{\mathcal{H}} = I_p - P_{\mathcal{H}}$. We then have $P_{\mathcal{H}} X = (0, X_{\mathcal{A}^c}^{\mathsf{T}})^{\mathsf{T}}$, $Q_{\mathcal{H}} X = (X_{\mathcal{A}}^{\mathsf{T}}, 0)^{\mathsf{T}}$. Similarly, $P_{\mathcal{H}} \eta$ corresponds to $\eta_{\mathcal{A}^c}$, and $Q_{\mathcal{H}} \eta$ corresponds to $\eta_{\mathcal{A}}$. The following proposition, given in Cook (2004), then connects (2) and η.

Proposition 1 *(Proposition 1, Cook, 2004)* $Y \perp\!\!\!\perp P_{\mathcal{H}} X | Q_{\mathcal{H}} X$ *if and only if* $P_{\mathcal{H}} \eta = \mathcal{O}_p$, *where* $\mathcal{O}_p$ *denotes the origin in* $\mathbb{R}^p$.

This proposition indicates that, the rows of a basis of the central subspace corresponding to $X_{\mathcal{A}^c}$ are all zero vectors; and all the predictors whose corresponding rows of the central subspace basis equal zero are irrelevant.

Proposition 1 also guarantees that the subset $X_{\mathcal{A}}$ uniquely exists provided that the central subspace $\mathcal{S}_{Y|X}$ exists. We thus have a well-defined population parameter $\mathcal{A}$. We also note that, in the IRE family, $\mathrm{Span}(\theta) = \mathcal{S}_{Y|X}$, from which we can partition θ conformly:

$$\theta = \begin{pmatrix} \theta_{\mathcal{A}} \\ \theta_{\mathcal{A}^c} \end{pmatrix}, \theta_{\mathcal{A}} \in \mathbb{R}^{(p-p_0) \times h}, \theta_{\mathcal{A}^c} \in \mathbb{R}^{p_0 \times h},$$

Then Proposition 1 implies that $\theta_{\mathcal{A}^c} = 0_{p_0 \times h}$. Consequently, we can describe the relevant predictor index set $\mathcal{A}$ as $\mathcal{A} = \{j : \theta_{jk} \neq 0 \text{ for some } k, 1 \leq j \leq p, 1 \leq k \leq h\}$. Such a definition is useful for later development.

3. Shrinkage Inverse Regression Estimation

3.1. Shrinkage inverse regression estimator

The IRE basis estimator $\hat{\eta}$ of $\mathcal{S}_{Y|X}$ are linear combinations of *all* the predictors under inquiry. When a subset of predictors are irrelevant, as specified in (2), it would be desirable to have the corresponding row estimates of $\hat{\eta}$ to equal zero, and consequently to achieve variable selection. For this purpose, we introduce a p-dimensional shrinkage vector $\alpha = (\alpha_1, \ldots, \alpha_p)^{\mathsf{T}}$, with $\alpha_j \in \mathbb{R}$, $j = 1, \ldots, p$. Given $(\hat{\eta}, \hat{\gamma}) = \arg\min_{B,C} G(B, C)$ of (1), we propose to minimize the following quadratic discrepancy function over α,

$$G_s(\alpha) = n\left\{\text{vec}(\hat{\theta}) - \text{vec}(\text{diag}(\alpha)\hat{\eta}\hat{\gamma})\right\}^{\mathsf{T}} V_n \left\{\text{vec}(\hat{\theta}) - \text{vec}(\text{diag}(\alpha)\hat{\eta}\hat{\gamma})\right\}, \qquad (3)$$

$$\text{subject to } \sum_{j=1}^{p} |\alpha_j| \leq \tau, \tau \geq 0.$$

Let $\hat{\alpha} = \arg\min_{\alpha} G_s(\alpha)$. We call $\text{Span}\{\text{diag}(\hat{\alpha})\hat{\eta}\}$ the *shrinkage inverse regression estimator* of the central subspace $\mathcal{S}_{Y|X}$.

We note that the proposed method is similar in spirit to the nonnegative garrote formulation in multiple linear regression (Breiman, 1995). Although the elements of α are allowed to be negative in (3), it is straightforward to show that this is asymptotically equivalent to enforcing a nonnegativity constraint on the shrinkage factors. This result is analogous to the connection between the nonnegative garotte and the adaptive LASSO discussed in the context of linear models (Zou, 2006, Corollary 2), in that with probability tending to 1, the resulting $\hat{\alpha}$ will all be nonnegative. Given such a connection between the shrinkage IRE and nonnegative garrote, we observe that, when the constraint parameter $\tau \geq p$, $\hat{\alpha}_j = 1$ for all j's, and we get back a usual IRE solution. As τ gradually decreases, some indices $\hat{\alpha}_j$ are shrunk to zero, which in turn shrinks the entire rows of the estimated basis of the central subspace to zero vectors, and consequently variable selection is achieved. A key difference in

7

estimating the basis of the central subspace, as compared to the typical linear model is that a variable corresponds to an entire row of the matrix θ, so that although θ consists of ph components to estimate, the shrinkage vector α is only of length p.

We also note that the proposed shrinkage inverse regression estimator reduces to shrinkage SIR estimator proposed by Ni, Cook and Tsai (2005) for a particular choice of the pair $(\hat{\theta}, V_n)$. Ni, et al. demonstrated the good empirical performance of shrinkage SIR but did not explore its theoretical properties. We next briefly discuss computations of the proposed shrinkage estimator and then study its asymptotic behavior for the entire IRE family, including shrinkage SIR as a special case.

3.2. Computations

Solving (3) is straightforward by noting that $G_s(\alpha)$ in (3) can be re-written as

$$
G_s(\alpha) = n \left\{ \mathrm{vec}(\hat{\theta}) - \begin{pmatrix} \mathrm{diag}(\hat{\eta}\hat{\gamma}_1) \\ \vdots \\ \mathrm{diag}(\hat{\eta}\hat{\gamma}_h) \end{pmatrix} \alpha \right\}^{\mathsf{T}} V_n \left\{ \mathrm{vec}(\hat{\theta}) - \begin{pmatrix} \mathrm{diag}(\hat{\eta}\hat{\gamma}_1) \\ \vdots \\ \mathrm{diag}(\hat{\eta}\hat{\gamma}_h) \end{pmatrix} \alpha \right\}.
$$

Thus it becomes a quadratic programming problem with linear constraint, with the "response" $U \in \mathbb{R}^{ph}$, and the "predictors" $W \in \mathbb{R}^{ph \times p}$:

$$
U = \sqrt{n} V_n^{1/2} \mathrm{vec}(\hat{\theta}), \qquad W = \sqrt{n} V_n^{1/2} \begin{pmatrix} \mathrm{diag}(\hat{\eta}\hat{\gamma}_1) \\ \vdots \\ \mathrm{diag}(\hat{\eta}\hat{\gamma}_h) \end{pmatrix}, \tag{4}
$$

where $V_n^{1/2}$ represents the symmetric square root of V_n.

Practically it is important to choose the constraint parameter τ in (3). We adopt a generalized Mallows' C_p criterion for the purpose of parameter tuning. More specifically, we choose τ to minimize the following criterion:

$$
\frac{\|U - \hat{U}(\tau)\|^2}{\hat{\sigma}^2} + \phi\, p_e, \tag{5}
$$

for some constant ϕ, where $\hat{U}(\tau) = W\hat{\alpha}(\tau)$, with $\hat{\alpha}(\tau)$ denoting the solution of α given τ; $\hat{\sigma}^2$ is the usual variance estimator obtained from the OLS fit of U on W; p_e is the effective number of parameters as given by Yuan and Lin (2006, Eq 6.5) for the grouped nonnegative garrote, $p_e = 2\sum_{j=1}^{p}\text{Ind}(\hat{\alpha}_j(\tau) > 0) + \sum_{j=1}^{p}\hat{\alpha}_j(\tau)(h-2)$, with $\text{Ind}(\cdot)$ denoting the indicator function. Choosing $\phi = 2$ is the typical Mallows' C_p criterion. However, our extensive simulations have suggested that choosing $\phi = \log(n_e)$, where n_e is the effective sample size, which equals ph in our setup, yields better empirical performance. This choice of ϕ is motivated via its relationship to the BIC criterion, and is the choice used in our examples.

3.3. Asymptotic properties

To study the asymptotic behavior, we consider the equivalent Lagrangian formulation of the constrained optimization problem. Specifically, the optimization problem in (3) can be reformulated as

$$\hat{\alpha} = \arg\min_{\alpha} \|U - W\alpha\|^2 + \lambda_n \sum_{j=1}^{p} |\alpha_j|,$$

for some nonnegative penalty constant λ_n. Let $\tilde{\theta}$ denote the resulting shrinkage estimator of θ, i.e., $\tilde{\theta} = \text{diag}(\hat{\alpha})\hat{\eta}\hat{\gamma}$, along with corresponding partitions $\tilde{\theta}_{\mathcal{A}} \in \mathbb{R}^{(p-p_0)\times h}$ and $\tilde{\theta}_{\mathcal{A}^c} \in \mathbb{R}^{p_0 \times h}$. Moreover, following the notation of Section 2.2, let $\mathcal{A} = \{j : \theta_{jk} \neq 0 \text{ for some } k, 1 \leq j \leq p, 1 \leq k \leq h\}$, and let $\mathcal{A}_n = \{j : \tilde{\theta}_{jk} \neq 0 \text{ for some } k, 1 \leq j \leq p, 1 \leq k \leq h\}$. We now show that, for an appropriate choice of λ_n, the shrinkage estimator yields consistency in variable selection along with asymptotic normality of the estimator for $\theta_{\mathcal{A}}$.

Theorem 1 *Assume that the initial estimator in the IRE family satisfies that $\sqrt{n}\left\{\text{vec}(\hat{\theta}) - \text{vec}(\theta)\right\} \to N(0, \Gamma)$, for some $\Gamma > 0$, and that $V_n^{1/2} = V^{1/2} + o(1/\sqrt{n})$. Suppose that $\lambda_n \to \infty$ and $\lambda_n/\sqrt{n} \to 0$, then the shrinkage estimator $\tilde{\theta}$ satisfies*

9

(a) Consistency in variable selection: $Pr(\mathcal{A}_n = \mathcal{A}) \to 1$

(b) Asymptotic normality: $\sqrt{n}\left\{\text{vec}(\tilde{\theta}_\mathcal{A}) - \text{vec}(\theta_\mathcal{A})\right\} \to N(0, \Lambda)$, *for some* $\Lambda > 0$

Theorem 1 (a) indicates that the shrinkage inverse regression estimator can select relevant predictors consistently. That is, for all $j \notin \mathcal{A}$ we have $Pr(\hat{\alpha}_j \neq 0) \to 0$, and for all $j \in \mathcal{A}$ we have $Pr(\hat{\alpha}_j \neq 0) \to 1$. Theorem 1 (b) further shows that the estimator for $\theta_\mathcal{A}$ that corresponds to the relevant predictors is $\sqrt{n}$-consistent. The proof of Theorem 1 is given in the Appendix.

The form of the asymptotic variance depends on both the choice of the initial estimator $\hat{\theta}$, and the matrix V that determines the particular member of the shrinkage IRE family. For efficiency, one would choose $V = \Gamma^{-1}$ in (1) as in Cook and Ni (2006). However, unlike the case of linear regression where a shrinkage estimator can obtain the identical asymptotic variance as if the true set $\mathcal{A}$ were known beforehand, this is no longer true here. It is due to the form of the matrix V, in that the block of Γ^{-1} corresponding to $\theta_\mathcal{A}$ is not the same as inverting the matrix Γ after only retaining the block corresponding to $\theta_\mathcal{A}$, in general.

3.4. Shrinkage CIRE and shrinkage SIR

While the proposed shrinkage estimation strategy works for the entire IRE family, in this section we consider some specific members of this family to help fix the ideas. Since the covariance inverse regression estimator (CIRE, Cook and Ni, 2006) is the state-of-art estimator for the central subspace, we first examine the shrinkage version of CIRE. We then briefly discuss the shrinkage sliced inverse regression estimator (Ni, Cook, and Tsai, 2005).

CIRE constructs θ with columns $\theta_s = \Sigma^{-1}\text{Cov}(Y J_s, X)$, $s = 1, \ldots, h$. Following the usual SDR protocol, we assume that the response Y is categorical with h levels, or

10

it has been discretized by constructing h slices, so that Y takes values in $\{1, 2, \ldots, h\}$. $J_s(Y) = 1$ if Y is in slice s and 0 otherwise, $f_s = Pr(J_s = 1)$, and $\Sigma = \text{Var}(X)$ which is assumed to be positive definite. Given the data, a consistent estimator, $\hat{\theta}$, of θ, is obtained by the slope of the OLS fit of $y_i J_s(y_i)$ on x_i, i.e.,

$$\hat{\theta}_s = \hat{\Sigma}^{-1} \frac{1}{n} \sum_{i=1}^{n} y_i J_s(y_i)(x_i - \bar{x}),$$

where $\bar{x} = \sum_{i=1}^{n} x_i / n$ and $\hat{\Sigma}$ is the usual sample covariance estimator of Σ. Note that $\sqrt{n}\{\text{vec}(\hat{\theta}) - \text{vec}(\theta)\} \to N(0, \Gamma)$, and the covariance matrix Γ of the limiting distribution is of the form

$$\Gamma = \text{E}\left[\delta\delta^{\mathsf{T}} \otimes \Sigma^{-1}\{X - \text{E}(X)\}\{X - \text{E}(X)\}^{\mathsf{T}}\Sigma^{-1}\right],$$

where δ is the population residual vector from the ordinary least squares fit of $Y(J_1, \ldots, J_h)^{\mathsf{T}}$ on X, and $\otimes$ indicates the Kronecker product. Letting $\hat{\delta}_i$ denote the i-th residual vector from the OLS fit of $y_i J_s(y_i)$ on x_i, a consistent estimator of Γ can then be constructed as

$$\Gamma_n = \frac{1}{n} \sum_{i=1}^{n} \left\{ \hat{\delta}_i \hat{\delta}_i^{\mathsf{T}} \otimes \hat{\Sigma}^{-1}\{x_i - \bar{x}\}\{x_i - \bar{x}\}^{\mathsf{T}}\hat{\Sigma}^{-1} \right\}.$$

Cook and Ni (2006) recommended to choose $V_n = \Gamma_n^{-1}$ to obtain the asymptotically efficient CIRE. For this choice of $(\hat{\theta}, V_n)$, it is straightforward to verify that the conditions in Theorem 1 are satisfied, and thus the resulting shrinkage CIRE enjoys the desirable properties of consistency in variable selection as well as $\sqrt{n}$-consistency.

Cook and Ni (2005) also showed that another widely used central subspace estimator SIR can be formulated as a member of the IRE family, by choosing $\theta_s = f_s \Sigma^{-1}\{\text{E}(X|J_s = 1) - \text{E}(X)\}$ and $V = D_f^{-1} \otimes \Sigma$, where $D_f = \text{diag}(f_s)$, and $f_s = Pr(J_s = 1)$, $s = 1, \ldots, h$. The corresponding consistent estimators are constructed as

$$\hat{\theta}_s = \hat{\Sigma}^{-1} \frac{1}{n} \left\{ \sum_{i=1}^{n} J_s(y_i)(x_i - \bar{x}) \right\},$$

11

and $V_n = D_{\hat{f}}^{-1} \otimes \hat{\Sigma}$, with $\hat{f} = \sum_{i=1}^{n} J_s(y_i)/n$. Plugging the above $\hat{\theta}$ and V_n in the general form of shrinkage inverse regression estimator in (3) leads to the shrinkage sliced inverse regression estimator of Ni, Cook, and Tsai (2005). Again Theorem 1 ensures that the shrinkage SIR estimator enjoys the consistency properties in both variable selection and basis estimation.

4. Simulation and Real Data Analysis

Given that CIRE is the state-of-art method for estimating $\mathcal{S}_{Y|X}$, in this section we examine the finite sample performance of the shrinkage CIRE. Our experience gained through simulation indicates that the method works quite well in terms of both variable selection and basis estimation. We also compare the shrinkage method with the test-based model free variable selection. A real data analysis is given to further illustrate the proposed method.

4.1. Finite sample performance

We first consider an example adopted from Chen and Li (1998).

$$Y = \text{sign}(\eta_1^\mathsf{T} X) \log(|\eta_2^\mathsf{T} X + 5|) + 0.2\varepsilon, \tag{6}$$

where $X = (X_1, \ldots, X_{20})^\mathsf{T}$ has $p = 20$ dimensions, and each component follows a standard normal distribution. The error ε is standard normally distributed and is independent of X. The central subspace $\mathcal{S}_{Y|X}$ is spanned by (η_1, η_2), where $\eta_1 = (1, \ldots, 1, 0, \ldots, 0)^\mathsf{T}$, $\eta_2 = (0, \ldots, 0, 1, \ldots, 1)^\mathsf{T}$, with q ones in each direction. We vary q to take values in $\{1, 5, 10\}$, representing the scenario with a very sparse basis ($q = 1$) to the case where all the predictors are relevant ($q = 10$). We denote these three setups as Cases 1, 2, and 3, respectively. In addition, we examine the case where $\eta_1 = (1, 0.75, 0.75, 0.5, 0.25, 0, \ldots, 0)^\mathsf{T}$ and $\eta_2 = (0, \ldots, 0, 1, 0.75, 0.75, 0.5, 0.25)^\mathsf{T}$, indicating

different magnitude of the active predictors, and denote this by Case 4. The sample size was set as $n = 200$ and $n = 400$.

*** TABLE 1 GOES HERE ***

Table 1 summarizes the average results based on 100 data replications. The first evaluation criterion is the number of active predictors. Both the true and the estimated numbers are reported in the table, where the estimated numbers are seen to be very close to the true ones. Table 1 next reports the accuracy of the shrinkage estimator in terms of the true positive rate, which is defined as the ratio of the number of correctly identified active predictors to the total number of truly relevant predictors, and the false positive rate, which is the ratio of the number of falsely identified active predictors to the total number of irrelevant predictors. Those two measures are commonly used in the biomedical literature, and ideally one wishes to have the true positive rate to be close to one and the false positive rate to be close to zero simultaneously. The resulting true positive rates of the proposed method were equal or close to one in all cases, while the false positive rates were reasonably low, which indicates that the shrinkage CIRE worked quite well in variable selection. The last two columns of the table reports the vector correlation coefficient (Hotelling, 1936), which measures the "closeness" of the estimated subspace and the true central subspace. It ranges between 0 and 1, with a larger value indicating a better estimate. For comparison, the results of both shrinkage CIRE and the usual CIRE estimators are recorded. As anticipated, the shrinkage estimator achieved higher accuracy than the solution without shrinkage when the true basis is sparse ($q = 1$ and 5), and was only slightly inferior when all the predictors are relevant ($q = 10$). Moreover, the shrinkage estimator performed well in the case where the relative magnitude of the active

13

predictors differs. As the sample size increased, the performance of the shrinkage estimator improved substantially, which agrees well with our theoretical results.

We have also examined a heteroscedastic model, $Y = \eta_1^\intercal X + \exp(\eta_2^\intercal X) \times \varepsilon$, where the central subspace is spanned by $\eta_1 = (1, 1, 0, \ldots, 0)^\intercal$ and $\eta_2 = (0, \ldots, 0, 1, 1)^\intercal$. In this model, the predictor effect presents not only in the conditional mean $\mathrm{E}(Y|X)$, but also in the conditional variance $\mathrm{Var}(Y|X)$. The same qualitative phenomenon as those reported in Table 1 were observed, and thus the results are omitted here.

4.2. Comparison with test-based model free variable selection

Predictor conditional independence hypothesis tests have been developed for SIR, CIRE, and other members of the IRE family (Cook, 2004, Cook and Ni, 2005, 2006). Such tests, when coupled with some subset search procedure, can be used for variable selection. In this section we compare the shrinkage method with those test-based approaches, and again we focus on the methods based on CIRE. Cook and Ni (2006) developed two versions of chi-squared tests of predictor conditional independence hypothesis, one assumes that $d = \dim(\mathcal{S}_{Y|X})$ is unknown and is referred as the *marginal predictor test*, and the other assumes d known and is referred as the *conditional predictor test*. In our simulation, both tests were implemented in the standard backward elimination search procedure, and the nominal level was fixed at 0.05, which is commonly used in practice. Note that the shrinkage method is also based on fixed dimension, d, and as such, the most direct comparison is with the conditional predictor test.

*** TABLE 2 GOES HERE ***

Simulation model (6) was re-examined, and a correlation structure was introduced to the predictors, where $\mathrm{corr}(X_i, X_j) = \rho^{|i-j|}, 1 \leq i < j \leq p$. With no correlation,

14

i.e. $\rho = 0$, the shrinkage method and both the conditional and marginal test-based approaches performed similarly in terms of variable selection. As ρ increased, the shrinkage method was observed to perform better when compared with the test-based selection. Table 2 reports the average results measured in terms of the true and false positive rates based on 100 data replications when $\rho = 0.5$. As seen from the table, compared with the conditional test, the shrinkage method was competitive in terms of both the true positive and false positive rates, generally yielding more true positives as well as less false positives. Moreover the shrinkage method is much less computationally intensive than the stepwise selection that must be incorporated into each of the test-based methods. It is also noted that both shrinkage method and the conditional predictor test yielded a higher true positive rate when compared with the marginal predictor test in all cases, with the marginal test often also being more conservative in terms of the false positive rate as well. This is because both the conditional predictor test and the shrinkage method are equipped with additional knowledge of the dimension of the central subspace.

4.3. A real data example

As an illustration we applied the shrinkage CIRE to the automobile data, which has recently been analyzed by Zhu and Zeng (2006). The data is available at the UCI machine learning repository. Following Zhu and Zeng (2006), we focused our analysis on the 13 car attributes with continuous measurements as predictors, including wheelbase, length, width, height, curb weight, engine size, bore, stroke, compression ratio, horsepower, peak rpm, city mpg, and highway mpg. The objective is to depict the relationship between those features and the car price. The data consist of 195 instances with complete records, and Zhu and Zeng (2006) concluded it reasonable to infer that the dimension of the central subspace equals two. We thus began our

shrinkage analysis with $\dim(\mathcal{S}_{Y|X}) = 2$.

*** FIGURE 1 GOES HERE ***

Figure 1 shows the full solution paths of the shrinkage CIRE as the tuning parameter τ is varied between 0 and $p = 13$. On the vertical axis is the estimated shrinkage parameter α_j, $j = 1, ..., p$, obtained from (3). It indicates the amount of shrinkage for each row of the basis of the central subspace. Figure 1 also shows the order when each predictor entered the basis estimate as τ increased. The first batch of predictors that entered include curb weight, city mpg, engine size, and horsepower, which appear to be the most important features to determine a car's price. The vertical line in Figure 1 indicates the choice of τ by minimizing the tuning criterion (5). Another three features, highway mpg, length, and height were selected. When $\tau = p$, there was no shrinkage for any of the predictors, as all elements of the shrinkage vector converged to one.

5. Discussion

In this article we have proposed a general shrinkage estimation strategy for the entire IRE family that is capable of simultaneous estimation of the dimension reduction basis and variable selection. The new estimators are shown to yield consistency in variable selection while retaining root-n estimation consistency. Both simulations and real data analysis demonstrate effectiveness of the new methods.

Development of theoretical properties of the proposed shrinkage IRE estimators has taken advantage of of the recent progress in linear model variable selection; see for instance, Zou (2006) and Yuan and Lin (2007). However, Zou (2006) and Yuan and

16

Lin (2007) both focused attention on the linear regression model, which assumes linear mean $E(Y|X)$ and homoscedastic variance $Var(Y|X)$, whereas shrinkage IRE permits a much more flexible class of regression models, e.g., nonlinear mean, heteroscedastic variance, as well as cases where the distribution of $Y|X$ relies on multiple linear combinations of the predictors. Mainly attributed to the flexible structure of the central subspace, the new methods can handle variable selection in a variety of regression forms, for instance, the single-index model $Y = f(\eta_1^\intercal X) + \varepsilon$, the heteroscedastic model $Y = f_1(\eta_1^\intercal X) + f_2(\eta_2^\intercal X) \times \varepsilon$, the additive model $Y = \sum_{j=1}^d f_j(\eta_j^\intercal X) + \varepsilon$, and the generalized linear model $\log\{Pr(Y = 1|X)/Pr(Y = 0|X)\} = \eta_1^\intercal X$. In the above cases, f's are the smooth link functions, ε represents a random error independent of the predictors, and η's form the basis of the central subspace $\mathcal{S}_{Y|X}$. Moreover, thanks to the model free nature of the IRE estimators, the shrinkage estimators achieve variable selection without requiring knowledge of the underlying model. These characteristics distinguish the proposed methods from the majority of model based variable selection approaches.

The results in this article also extend the existing model free variable selection development in at least two ways. First, it has been demonstrated that the shrinkage approach performs competitively when compared with the stepwise selection that is equipped with some predictor conditional independence test. In practice we generally prefer the shrinkage approach over the stepwise selection, since the shrinkage method is shown to be both consistent and quickly computed, and also avoids the potential issue of multiple testing. Secondly, asymptotic properties have been derived for the entire family of inverse regression estimators. As a direct consequence, the large sample behavior of the existing shrinkage SIR can be obtained straightforwardly.

The proposed shrinkage inverse regression estimators are closely related to the nonnegative garrote formulation in linear models. The estimation can be viewed

17

as a two-stage procedure, which first obtains an unconstrained estimator and then multiplies by a shrinkage vector to produce the final shrinkage estimator. As with the usual nonnegative garrote (see also a discussion in Yuan and Lin, 2007), the finite sample empirical performance of the shrinkage estimator rely on the initial estimator, which is the inverse regression estimator without shrinkage in our case. The development of a shrinkage estimation and variable selection method that depends less on the initial estimator can be practically useful, and work along this line is in progress.

Appendix: Justifications

Outline of Proof of Theorem 1

We first note that neither the proof of Yuan and Lin (2007) for the nonnegative garotte, nor that of Zou (2006) for the adaptive LASSO for linear models can be used directly in that the shrinkage factor is of dimension p while the initial estimate is of dimension ph. For estimation of the central subspace, the shrinkage factor is applied to complete rows of the initial basis estimate. Specifically, each row of the matrix θ corresponds to a predictor, which is given a single shrinkage factor, whereas in the typical linear model setup there are p parameters (one for each predictor) and p shrinkage factors.

Our proof proceeds as follows. We first show that the shrinkage IRE estimator can be re-expressed as an adaptive LASSO-type formulation on a constrained p-dimensional parameter space, followed by an estimated linear transformation to the ph-dimensional space. This reparameterization is a key step in allowing for the proof of the asymptotic normality of the shrinkage IRE estimator and is given in Lemma 1. Next Lemma 2 shows the asymptotic normality of the 'residual' in this reparameter-

ized lower dimensional formulation. Lemma 3 then gives some additional properties that allow us to adopt the method of proof of Zou (2006) for the p-dimensional problem. We then further show that the theorem holds for the full ph-dimensional parameter estimate.

Without loss of generality, we assume that the first column of $\theta_{\mathcal{A}}$ is fully non-zero, i.e. $\prod_{i=1}^{p-p_0} \theta_{i1} \neq 0$. This may be assumed, as any right non-singular transformation of the basis θ combined with the corresponding equivariance property of any reasonable choice of the matrix V_n leaves the discrepancy function in (1) unchanged.

Three lemmas

We will need the following three lemmas to prove Theorem 1.

Lemma 1 *Define the $ph \times p$ matrix*

$$
\hat{P} \equiv \begin{pmatrix} \mathrm{diag}(\hat{\vartheta}_1) \\ \vdots \\ \mathrm{diag}(\hat{\vartheta}_h) \end{pmatrix} \left\{ \mathrm{diag}(\hat{\vartheta}_1) \right\}^{-1},
$$

and let

$$
P \equiv \begin{pmatrix} \mathrm{diag}(\theta_1) \\ \vdots \\ \mathrm{diag}(\theta_h) \end{pmatrix} \left\{ \mathrm{diag}(\theta_1) \right\}^{-},
$$

where the generalized inverse of the diagonal matrix is given by the reciprocal of the non-zero elements and zero otherwise. Then the solution, $\tilde{\theta}$, to the optimization problem given by (3) is equivalent to the solution of the following problem.

$$
\tilde{\theta}_1 = \arg\min_{\theta_1} \|U - M\theta_1\|^2, \quad subject\ to\ \sum_{j=1}^{p} \frac{|\theta_{j1}|}{|\hat{\vartheta}_{j1}|} \leq t,
$$

with $M = \sqrt{n} V_n^{1/2} \hat{P}$, and θ_1 denotes the first column of θ. Then set $\hat{\alpha}_j = \tilde{\theta}_{j1}/\hat{\vartheta}_{j1}$ and $\tilde{\theta} = diag(\hat{\alpha})\hat{\vartheta} = \hat{P}\tilde{\theta}_1$.

Proof of Lemma 1: By definition $\hat{\alpha}_j = \tilde{\theta}_{j1}/\hat{\vartheta}_{j1} = ... = \tilde{\theta}_{jh}/\hat{\vartheta}_{jh}$. Thus the optimization problem in (3) in terms of the solution $\hat{\alpha}$ can be re-expressed as an optimization problem with solution $\tilde{\theta}$ as

$$\tilde{\theta} = \arg\min_{\theta} \|U - \sqrt{n}V_n^{1/2}\mathrm{vec}(\theta)\|^2, \quad \text{subject to } \sum_{j=1}^{p} \frac{|\theta_{j1}|}{|\hat{\vartheta}_{j1}|} \le t,$$

$$\text{and } \frac{\theta_{jk}}{\hat{\vartheta}_{jk}} = \frac{\theta_{j1}}{\hat{\vartheta}_{j1}}, \text{ for all } j = 1, ..., p \text{ and } k = 1, ..., h,$$

where U is given in (4).

There are only p free parameters in the optimization as the additional constraints restrict the space. Note that, by construction, every element of $\hat{\vartheta}$ is non-zero with probability one, since this is true of the initial estimate $\hat{\theta}$. In addition, the rows of P corresponding to $\theta_{\mathcal{A}^c}$ are identically zero. This along with the fact that θ_{j1} is non-zero for all $j \in \mathcal{A}$ allows for the estimator $\tilde{\theta}$ to be completely determined by its first column as given by the lemma.

Lemma 2 *Assume that $\sqrt{n}\{\mathrm{vec}(\hat{\theta}) - \mathrm{vec}(\theta)\} \to N(0, \Gamma)$, for some $\Gamma > 0$, and that $V_n^{1/2} = V^{1/2} + o(1/\sqrt{n})$. Then*

$$\epsilon = U - M\theta_1 \to N(0, S),$$

with

$$S = V^{1/2}\Gamma V^{1/2} + V^{1/2}Q\Delta \left(\Delta^{\mathsf{T}}\Gamma^{-1}\Delta\right)^{-} \Delta^{\mathsf{T}}Q^{\mathsf{T}}V^{1/2} + 2V^{1/2}DQ^{\mathsf{T}}V^{1/2},$$

where $\Delta = \left(\frac{\partial \mathrm{vec}(\theta)}{\partial \mathrm{vec}(\eta)}, \frac{\partial \mathrm{vec}(\theta)}{\partial \mathrm{vec}(\gamma)}\right)$, $Q_{ph \times ph} = [P\ 0] - \mathrm{diag}\left(Ind\{\mathrm{vec}(\theta) \ne 0\}\right)$, and $D = \mathrm{Cov}(\sqrt{n}\mathrm{vec}(\hat{\theta}), \sqrt{n}\mathrm{vec}(\hat{\vartheta}))$, with $Ind\{\cdot\}$ being the indicator function and $\mathrm{vec}(\theta) \ne 0$ is interpreted componentwise.

Proof of Lemma 2: First note that $\epsilon = V_n^{1/2} \sqrt{n} \left\{ \text{vec}(\hat{\theta}) - \hat{P}\theta_1 \right\}$. For each $j \notin \mathcal{A}$, one has that $\theta_{j1} = ... = \theta_{jh} = 0$ and hence on $\mathcal{A}^c$, $\text{vec}(\hat{\theta}) - \hat{P}\theta_1 = \text{vec}(\hat{\theta}) - \text{vec}(\theta)$, while on $\mathcal{A}$ one has

$$\text{vec}(\hat{\theta}) - \hat{P}\theta_1 = \{\text{vec}(\hat{\theta}) - \text{vec}(\theta)\} - (\hat{P} - P)\theta_1.$$

Now $\hat{P} - P = A_1 + A_2$, with

$$A_1 = \begin{pmatrix} \text{diag}(\hat{\vartheta}_1 - \theta_1) \\ \vdots \\ \text{diag}(\hat{\vartheta}_h - \theta_h) \end{pmatrix} \{\text{diag}(\theta_1)\}^{-1}$$

and

$$A_2 = \begin{pmatrix} \text{diag}(\hat{\vartheta}_1) \\ \vdots \\ \text{diag}(\hat{\vartheta}_h) \end{pmatrix} \left[\left\{ \text{diag}(\hat{\vartheta}_1) \right\}^{-1} - \{\text{diag}(\theta_1)\}^{-1} \right].$$

On $\mathcal{A}$, we may expand the second term as

$$A_2 = - \begin{pmatrix} \text{diag}(\theta_1) \\ \vdots \\ \text{diag}(\theta_h) \end{pmatrix} \{\text{diag}(\theta_1)\}^{-1} \text{diag}(\hat{\vartheta}_1 - \theta_1) \{\text{diag}(\theta_1)\}^{-1} + o(\sqrt{n}).$$

Thus we obtain on $\mathcal{A}$,

$$\text{vec}(\hat{\theta}) - \hat{P}\theta_1 = \left\{ \text{vec}(\hat{\theta}) - \text{vec}(\theta) \right\} + P \left\{ \hat{\vartheta}_1 - \theta_1 \right\} - \left\{ \text{vec}(\hat{\vartheta}) - \text{vec}(\theta) \right\} + o(\sqrt{n}).$$

Hence for all j, we can write

$$\text{vec}(\hat{\theta}) - \hat{P}\theta_1 = \left\{ \text{vec}(\hat{\theta}) - \text{vec}(\theta) \right\} + Q \left\{ \text{vec}(\hat{\vartheta}) - \text{vec}(\theta) \right\} + o(\sqrt{n}),$$

with Q as given in the lemma. The result then follows directly from the fact that $\sqrt{n} \left\{ \text{vec}(\hat{\vartheta}) - \text{vec}(\theta) \right\} \to N(0, \Delta (\Delta^{\mathsf{T}}\Gamma^{-1}\Delta)^- \Delta^{\mathsf{T}})$, as in Cook and Ni (2005), and that $V_n^{1/2} \to V^{1/2}$. This completes the proof.

Lemma 3 *Under the conditions of Lemmas 1 and 2,*

1. *$\frac{1}{n}M^\mathsf{T}M = \hat{P}^\mathsf{T}V_n\hat{P} = O_p(1)$ and $\left(\frac{1}{n}M^\mathsf{T}M\right)_{11} \to (P^\mathsf{T}VP)_{11}$, where the subscript denotes the upper left block that corresponds to θ_A.*

2. *$\frac{\epsilon^\mathsf{T}M}{\sqrt{n}} \to Z$, where $Z \sim N(0, P^\mathsf{T}V^{1/2}SV^{1/2}P)$ and ϵ, S are given in Lemma 2.*

Proof of Lemma 3: Both parts of the lemma follow directly from Lemma 2 and the $\sqrt{n}$-consistency of both V_n and $\hat{\vartheta}$.

Proof of Theorem 1

Consider the 'pseudo' linear model $U = M\theta_1 + \epsilon$. For estimation of θ_1, Lemma 1 shows that $\tilde{\theta}_1$ is an adaptive LASSO estimator with weights given by $\hat{\vartheta}_1$. Lemma 2 shows that the 'residual' term in this 'pseudo' linear model is asymptotically $N(0, S)$. The conditions in Lemma 3 allow the approach of Zou (2006) to be used to prove consistency in both variable selection and estimation for this adaptive LASSO estimator. Using $Var(\epsilon) = S$, along with Lemma 3, we obtain the asymptotic variance of $\sqrt{n}(\tilde{\theta}_1 - \theta_1)$ on $\mathcal{A}$ as $(P^\mathsf{T}VP)_{11}^{-1}\left(P^\mathsf{T}V^{1/2}SV^{1/2}P\right)_{11}(P^\mathsf{T}VP)_{11}^{-1}$. This shows that the first column of the estimated basis has the desired properties.

To complete the proof it remains to show that the remaining columns also obtain consistency in both variable selection and estimation. For consistency in variable selection, it follows directly from the fact that $\alpha_j = 0 \Leftrightarrow j \in \mathcal{A}^c$ due to the consistency of $\tilde{\theta}_1$. We now prove the asymptotic normality of $\text{vec}(\tilde{\theta})$.

Now, $\sqrt{n}\left\{\text{vec}(\tilde{\theta}) - \text{vec}(\theta)\right\} = \sqrt{n}(\hat{P}\tilde{\theta}_1 - P\theta_1) =$

$$\sqrt{n}\left(\hat{P} - P\right)\tilde{\theta}_1 + \sqrt{n}P(\tilde{\theta}_1 - \theta_1).$$

From above, the second term is asymptotically normal. Now the first term is $B_1 + B_2$, with $B_1 = \sqrt{n}\left(\hat{P} - P\right)\theta_1$, and $B_2 = \sqrt{n}\left(\hat{P} - P\right)\left(\tilde{\theta}_1 - \theta_1\right)$. In the course of the

proof of Lemma 2 it was shown that B_1 is asymptotically normal and that $\left(\hat{P} - P\right) \to$ 0 on $\theta_{\mathcal{A}}$, hence $B_2 \to 0$, since $\sqrt{n}\left(\tilde{\theta}_1 - \theta_1\right) = O_p(1)$. Thus we have the asymptotic normality of $\sqrt{n}\left\{\text{vec}(\tilde{\theta}) - \text{vec}(\theta)\right\}$. This completes the proof.

References

Breiman, L. (1995). Better subset selection using the nonnegative garrote. *Technometrics*, **37**, 373-384.

Chen, C.H. and Li, K.C. (1998). Can SIR be as popular as multiple linear regression? *Statistica Sinica*, **8**, 289-316.

Cook, R.D. (1996). Graphics for regressions with a binary response. *Journal of the American Statistical Association*, **91**, 983-992.

Cook, R.D. (1998). *Regression Graphics: Ideas for Studying Regressions Through Graphics*, New York: Wiley.

Cook, R.D. (2004). Testing predictor contributions in sufficient dimension reduction. *Annals of Statistics*, **32**, 1061-1092.

Cook, R.D. and Ni, L. (2005). Sufficient dimension reduction via inverse regression: a minimum discrepancy approach. *Journal of the American Statistical Association*, **100**, 410-428.

Cook, R.D. and Ni, L. (2006). Using intraslice covariances for improved estimation of the central subspace in regression. *Biometrika*, **93**, 65-74.

Cook, R.D. and Weisberg, S. (1991). Discussion of Li (1991). *Journal of the American Statistical Association*, **86**, 328-332.

Hotelling, H. (1936). Relations between two sets of variates. *Biometrika*, **28**, 321-377.

Li, K.C. (1991). Sliced inverse regression for dimension reduction (with discussion). *Journal of the American Statistical Association*, **86**, 316-327.

Li, L., Cook, R.D., and Nachtsheim, C.J. (2005). Model-free variable selection. *Journal of the Royal Statistical Society*, Series B, **67**, 285-299.

Li, L. and Nachtsheim, C.J. (2006). Sparse sliced inverse regression. *Technometrics*, **48**, 503-510.

Miller, A.J. (2002). *Subset Selection in Regression*. Chapman and Hall, 2nd Edition.

Ni, L., Cook, R.D., and Tsai, C.-L. (2005). A note on shrinkage sliced inverse regression. *Biometrika*, **92**, 242-247.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society*, Series B, **68**, 49-67.

Yuan, M. and Lin, Y. (2007). On the nonnegative garrote estimator. *Journal of the Royal Statistical Society*, Series B, **69**, 143-161.

Zhu, Y., and Zeng, P. (2006). Fourier methods for estimating the central subspace and the central mean subspace in regression. *Journal of the American Statistical Association*, **101**, 1638-1651.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, **101**, 1418-1429.

Table 1: Finite sample performance of the shrinkage CIRE estimator. Reported are the average out of 100 data replications of the true and the estimated number of relevant predictors, the true and false positive rates, and the vector correlation of the shrinkage CIRE (S-CIRE) and the usual CIRE.

		# actives		positive rate		vector correlation	
	sample size	true	estimate	true	false	S-CIRE	CIRE
Case 1	$n = 200$	2	3.31	1.000	0.073	0.989	0.879
	$n = 400$	2	2.49	1.000	0.027	0.999	0.951
Case 2	$n = 200$	10	11.19	0.997	0.122	0.934	0.909
	$n = 400$	10	10.40	1.000	0.040	0.979	0.961
Case 3	$n = 200$	20	18.91	0.946	–	0.794	0.884
	$n = 400$	20	19.96	0.998	–	0.932	0.953
Case 4	$n = 200$	10	10.26	0.913	0.113	0.933	0.911
	$n = 400$	10	10.01	0.969	0.032	0.978	0.966

Table 2: Comparison of the shrinkage CIRE estimator (S-CIRE) and the stepwise selection approach based on the conditional predictor test (c-test) and the marginal predictor test(m-test). Reported are the average out of 100 data replications of the true and false positive rates.

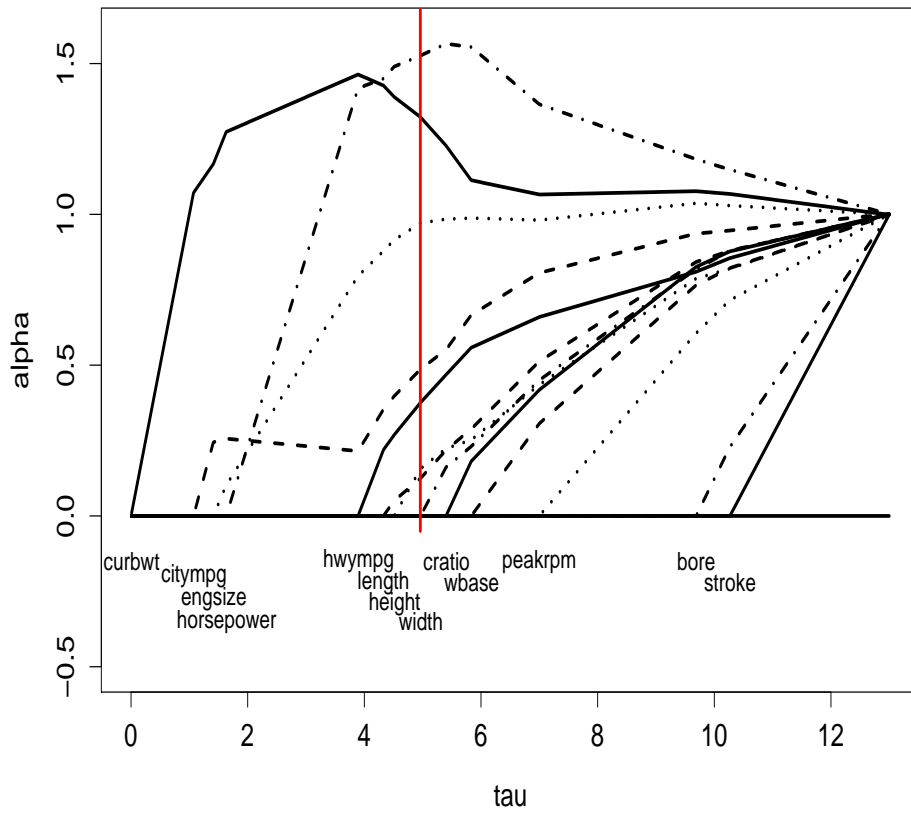|		true positive rate			false positive rate		
	sample size	S-CIRE	c-test	m-test	S-CIRE	c-test	m-test
Case 1	$n = 200$	1.000	1.000	1.000	0.076	0.089	0.093
	$n = 400$	1.000	1.000	1.000	0.026	0.074	0.071
Case 2	$n = 200$	0.923	0.882	0.821	0.168	0.164	0.103
	$n = 400$	0.984	0.987	0.962	0.062	0.095	0.064
Case 3	$n = 200$	0.670	0.611	0.546	—	—	—
	$n = 400$	0.824	0.800	0.708	—	—	—
Case 4	$n = 200$	0.850	0.849	0.812	0.133	0.146	0.105
	$n = 400$	0.912	0.917	0.882	0.055	0.095	0.072

Figure 1: Solution paths for the automobile data. The vertical line denotes the value of τ that minimizes the tuning criterion (5).