Biometrika (2005), **92,** 1, *pp.* 242–247

Printed in Great Britain

A note on shrinkage sliced inverse regression

By LIQIANG NI

Department of Statistics and Actuarial Science, University of Central Florida, Orlando, Florida 32816, U.S.A.

lni@mail.ucf.edu

R. DENNIS COOK

School of Statistics, University of Minnesota, Minneapolis, Minnesota 55455, U.S.A.

dennis@stat.umn.edu

AND CHIH-LING TSAI

Graduate School of Management, University of California, Davis, California 95616, U.S.A.

cltsai@ucdavis.edu

Summary

We employ Lasso shrinkage within the context of sufficient dimension reduction to obtain a shrinkage sliced inverse regression estimator, which provides easier interpretations and better prediction accuracy without assuming a parametric model. The shrinkage sliced inverse regression approach can be employed for both single-index and multiple-index models. Simulation studies suggest that the new estimator performs well when its tuning parameter is selected by either the Bayesian information criterion or the residual information criterion.

Some key words: Garotte; Lasso; Shrinkage estimator; Sliced inverse regression; Sufficient dimension reduction.

1. Introduction

In the last decade, sufficient dimension reduction has generated considerable interest in high-dimensional regressions involving a univariate response Y and a p-dimensional predictor $X = (x_1, \ldots, x_p)^{\mathrm{T}}$. The basic idea is to replace X with a lower-dimensional projection $P_{\mathscr{S}}X$ without loss of information about the conditional distribution of $Y|X$, where $P_{\mathscr{S}}$ is the orthogonal projection on to $\mathscr{S}$ in the usual inner product. No pre-specified model for $Y|X$ is required. The parsimonious target of a sufficient dimension reduction enquiry is the central subspace, $\mathscr{S}_{Y|X}$ (Cook, 1996, 1998a), defined as the intersection of all subspaces $\mathscr{S} \subseteq \mathbb{R}^p$ having the property $Y \perp\!\!\!\perp X | P_{\mathscr{S}}X$, where $\perp\!\!\!\perp$ indicates independence. Consequently, $P_\beta X$ extracts all of the information from X about Y, where β is a basis of $\mathscr{S}_{Y|X}$. For further background on the central subspace, see Cook (1998a, Ch. 6) and the references contained therein.

One of the most widely used methods for estimating the central subspace is sliced inverse regression (Li, 1991). In order to improve its estimator and the accuracy of predictions, Naik & Tsai (2001) proposed model-selection methods for single-index models. In contrast to the variable-selection approach, Cook (2004) recently applied a model-free method for assessing the contribution of variables. In addition to variable selection and hypothesis-testing approaches, we employ Tibshirani's (1996) least absolute shrinkage and selection operator, Lasso, to develop a new method, the shrinkage sliced inverse regression estimator.

To better understand the motivation underlying the shrinkage sliced inverse regression, we briefly review shrinkage estimators in linear regression models. Consider a linear model $y=\sum_{k=1}^{p}\eta_k x_k+\varepsilon$, which has been well studied with different assumptions regarding the predictors and the random error ε. Breiman (1995) considered a linear model in which all but a few of the $\{\eta_k\}$ are nearly zero. To deal with such situations, he proposed a nonnegative 'Garotte' method as follows: letting $\{\hat{\eta}_k\}$ be the ordinary least squares estimators, find $\{c_k\}$ to minimise $\sum_i(y_i-\sum_k c_k\hat{\eta}_k x_{ik})^2$ under the constraints $c_k\geqslant 0$, $\sum_k c_k\leqslant t$, where t is a tuning parameter. The resulting estimators, $\tilde{\eta}_k=c_k\hat{\eta}_k$, are the Garotte predictor coefficients. Decreasing t results in more of the estimated coefficients being zero, which may lead to a more interpretable and accurate prediction model. A drawback of the Garotte is that its solution depends on the sign of the ordinary least squares estimate. In the same spirit, Tibshirani (1996) proposed the Lasso estimator, defined as the set of $\{\eta_k\}$ that minimises $\sum_i(y_i-\sum_k\eta_k x_{ik})^2$ subject to $\sum_k|\eta_k|\leqslant t$. These shrinkage methods can work well when model assumptions between y and x are specified.

To eliminate the need for a model for $Y|X$, we employ Lasso shrinkage in the context of sufficient dimension reduction to obtain the shrinkage sliced inverse regression estimators, which are often parsimonious and accurate even when the form of the model is unknown. In addition, this approach enables us to estimate parameters of a multiple-index model, whereas Garotte and Lasso do not.

2. Lasso and shrinkage sliced inverse regression

2·1. *Algorithm and tuning parameter for Lasso*

Several algorithms have been proposed for computing Lasso estimates, such as Tibshirani (1996), Fu (1998) and Osborne et al. (2000). In this paper, we focus on Fu's shooting algorithm, which is based on the fact that the Lasso estimate is equivalent to the solution of the penalised least squares problem

$$\min_{\eta}\left\{\sum_i\left(y_i-\sum_k\eta_k x_{ik}\right)^2+\lambda\sum_k|\eta_k|\right\}, \tag{1}$$

where λ is the tuning parameter, Fu adopted Tibshirani's (1996) generalised crossvalidation method for selecting the tuning parameter λ: for a given λ, solve (1) to obtain the Lasso estimate $\tilde{\eta}$ and then compute the effective number of parameters, $p(\lambda)=\mathrm{tr}\,\{\mathcal{X}(\mathcal{X}^{\mathrm{T}}\mathcal{X}+\lambda W^{-})^{-1}\mathcal{X}^{\mathrm{T}}\}-n_0$, where $\mathcal{X}$ is the $n\times p$ data matrix, W^{-} is the generalised inverse of $W=\mathrm{diag}(2|\tilde{\eta}_k|)$, and n_0 is the number of $\tilde{\eta}_k$ such that $\tilde{\eta}_k=0$. The value of λ is selected by minimising

$$\textsc{gcv}=\frac{\textsc{rss}}{n\{1-p(\lambda)/n\}^2},$$

where $\textsc{rss}=\sum_i(y_i-\sum_k\tilde{\eta}_k x_{ik})^2$.

In addition to generalised crossvalidation, we also consider Akaike's information criterion (Akaike, 1973), the Bayesian information criterion (Schwarz, 1978), and the residual information criterion (Shi & Tsai, 2002):

$$\textsc{aic}=n\log(\textsc{rss}/n)+2p(\lambda),$$

$$\textsc{bic}=n\log(\textsc{rss}/n)+\log(n)p(\lambda),$$

$$\textsc{ric}=\{n-p(\lambda)\}\log(\hat{\sigma}^2)+p(\lambda)\{\log(n)-1\}+4/\{n-p(\lambda)-2\},$$

where $\hat{\sigma}^2=\textsc{rss}/\{n-p(\lambda)\}$. The performance of these four criteria will be discussed in § 3.

2·2. *Sliced inverse regression*

Sliced inverse regression is a widely used method for estimating the central subspace. Usually it requires that the standardised predictor, $Z=\Sigma^{-1/2}\{X-E(X)\}$, satisfy the linearity condition $E(Z|P_\gamma Z)=P_\gamma Z$, where $\Sigma=\mathrm{cov}(X)$ and γ is a basis for $\mathcal{S}_{Y|Z}$. This condition connects the central subspace with the inverse regression of Z on Y, that is $\mathrm{Span}(M_{\mathrm{SIR}}) \subseteq \mathcal{S}_{Y|Z}$, where the kernel matrix $M_{\mathrm{SIR}} = \mathrm{cov}\,\{E(Z|Y)\}$. For the sake of convenience, we further assume that $\mathrm{span}\,(M_{\mathrm{SIR}}) = \mathcal{S}_{Y|Z}$, but the results presented in this paper do not rely on this coverage condition.

Suppose we have a simple random sample of size n of realisations of (X, Y), which has a joint distribution. Let $\bar{X}$ be the grand average of X. The sample version of Z is $\hat{Z} = \hat{\Sigma}^{-1/2}(X - \bar{X})$, where $\hat{\Sigma}$ is the usual sample covariance matrix of X. Suppose there are h slices with n_y observations in the yth slice. Thus, the sample version of M_{SIR} can be represented as $\hat{M}_{\mathrm{SIR}} = \sum_{y=1}^{h} \hat{f}_y \hat{Z}_y \hat{Z}_y^{\mathrm{T}}$, where $\hat{f}_y = n_y/n$ and $\hat{Z}_y$ is the average of $\hat{Z}$ in the yth slice (Li, 1991). Let $\{\hat{v}_1, \hat{v}_2, \ldots, \hat{v}_p\}$ be the eigenvectors of $\hat{M}_{\mathrm{SIR}}$ corresponding to the eigenvalues $\hat{\lambda}_1 > \hat{\lambda}_2 > \ldots > \hat{\lambda}_p \geqslant 0$. If the dimension d of $\mathcal{S}_{Y|Z}$ is known, $\mathrm{Span}(\hat{\beta}) = \mathrm{Span}(\hat{\beta}_1, \ldots, \hat{\beta}_d)$ is a consistent estimator of $\mathcal{S}_{Y|X}$, where $\hat{\beta}_i = \hat{\Sigma}^{-1/2}\hat{v}_i$. For determining d, see Li (1991), Schott (1994) and Bura & Cook (2001).

2·3. *Shrinkage sliced inverse regression*

The sliced inverse regression approach provides an estimator $\mathrm{Span}(\hat{\beta})$ of the central subspace, of which the elements of $\hat{\beta} \in \mathbb{R}^{p\times d}$ are usually nonzero. When a large number of predictors or highly-correlated predictors occur in data analysis, we would expect that only a subset of the predictors are needed in the construction of 'sufficient predictors', $\beta_i^{\mathrm{T}} X$ $(i = 1, \ldots, d)$, suggesting that some rows of β are all zeros. To this end, we employ the Lasso technique to develop a shrinkage sliced inverse regression method which compresses some rows of $\hat{\beta}$ to 0's.

It follows from Cook (2004) that a basis for the span of the eigenvectors corresponding to the d largest eigenvalues of $\hat{M}_{\mathrm{SIR}}$ can be obtained by minimising $F(A, C) = \sum_{y=1}^{h} \| \hat{f}_y^{1/2} \hat{Z}_y - AC_y \|^2$ over $A \in \mathbb{R}^{p\times d}$ and $C_y \in \mathbb{R}^d$, with $C = (C_1, \ldots, C_h)$. Let $\hat{A}$ and $\hat{C}$ be the values of A and C that minimise F. Then $\mathrm{Span}(\hat{A})$ equals the space spanned by the d largest eienvectors of $\hat{M}_{\mathrm{SIR}}$. The value of $\hat{A}$ is not necessarily unique, but $\mathrm{Span}(\hat{A})$ is. Focusing on the coefficients of the X variables, we restate $F(A, C)$ as

$$G(B, C) = \sum_{y=1}^{h} (\hat{f}_y^{1/2} \hat{\Sigma}^{-1/2} \hat{Z}_y - BC_y)^{\mathrm{T}} \hat{\Sigma} (\hat{f}_y^{1/2} \hat{\Sigma}^{-1/2} \hat{Z}_y - BC_y), \tag{2}$$

where $B \in \mathbb{R}^{p\times d}$. The value of B which minimises (2) is exactly $\hat{\beta}$ and the estimator of $\mathcal{S}_{Y|X}$ is $\mathrm{Span}(\hat{\beta}) = \mathrm{Span}(\hat{\Sigma}^{-1/2}\hat{A})$. We now can present the following lemma.

Lemma. *A shrinkage sliced inverse regression estimator of* $\mathcal{S}_{Y|X}$ *is* $\mathrm{Span}\{\mathrm{diag}\,(\tilde{\alpha})\hat{\beta}\}$, *where the shrinkage indices* $\tilde{\alpha} = (\tilde{\alpha}_1, \ldots, \tilde{\alpha}_p)^{\mathrm{T}} \in \mathbb{R}^p$ *are determined by minimising*

$$\sum_{y=1}^{h} \| \hat{f}_y^{1/2} \hat{Z}_y - \hat{\Sigma}^{1/2}\,\mathrm{diag}\,(\hat{B}\hat{C}_y)\alpha \|^2 \tag{3}$$

subject to $\sum_{i=1}^{p} |\alpha_i| \leqslant t$, *and the values of* $\hat{B}$ *and* $\hat{C} = (\hat{C}_1, \ldots, \hat{C}_h)$ *minimise* (2).

Proof. Since sliced inverse regression provides $\hat{B}$ and $\hat{C}$, we are able to adopt the Lasso approach to obtain the shrinkage indices $\tilde{\alpha}$ as the argument α which minimises

$$\sum_{y=1}^{h} \{ \hat{f}_y^{1/2} \hat{\Sigma}^{-1/2} \hat{Z}_y - \mathrm{diag}\,(\alpha)\hat{B}\hat{C}_y \}^{\mathrm{T}} \hat{\Sigma} \{ \hat{f}_y^{1/2} \hat{\Sigma}^{-1/2} \hat{Z}_y - \mathrm{diag}\,(\alpha)\hat{B}\hat{C}_y \} \tag{4}$$

subject to $\sum_{i=1}^{p} |\alpha_i| \leqslant t$. After algebraic simplifications, equation (4) is the same as equation (3). □

The shrinkage estimator $\mathrm{diag}\,(\tilde{\alpha})\hat{B}$ constitutes an estimated basis for the central subspace. Note that the shrinkage indices $\tilde{\alpha}$ can be obtained using Lasso algorithms. To be specific, let

$$\xi_n = \mathrm{vec}\,(\hat{f}_1^{1/2}\hat{Z}_1, \ldots, \hat{f}_h^{1/2}\hat{Z}_h) \in \mathbb{R}^{ph},$$

$$D_n = \{\mathrm{diag}\,(\hat{B}\hat{C}_1)\hat{\Sigma}^{1/2}, \ldots, \mathrm{diag}\,(\hat{B}\hat{C}_h)\hat{\Sigma}^{1/2}\}^{\mathrm{T}} \in \mathbb{R}^{ph\times p},$$

where vec(.) is a matrix operator that stacks the matrix's columns one by one. Then $\tilde{\alpha}$ is exactly the Lasso estimator for the regression of ξ_n on the data matrix D_n. When $t \geqslant p$, $\tilde{\alpha}_i = 1$ $(i = 1, \ldots, d)$; that is shrinkage sliced inverse regression becomes sliced inverse regression. As t gradually decreases, or λ increases in a version of (1), some indices α_i, tend to be 0, which indicates that their corresponding predictors are not needed for the regression given the other predictors. Thus, shrinkage sliced inverse regression enables one to shrink the coefficient of the sufficient predictor without assuming a model.

In summary, shrinkage sliced inverse regression is a two-step procedure: first, apply sliced inverse regression to obtain the structural dimension d, the response ξ_n and the data matrix D_n; secondly, compute $\tilde{\alpha}$ via the Lasso approach by choosing the tuning parameter λ as discussed in § 2·1.

3. Simulation results

We consider scenarios involving linear, single-index and two-index models, and correlated predictors. For each setting, 100 replications were generated. We report the performance of shrinkage estimators for shrinkage sliced inverse regression and Lasso when it is applicable.

First, we consider a linear model $y = \eta^{\mathrm{T}} X + 0{\cdot}5\varepsilon$, where $\eta = (1, 1, 1, 0, \ldots, 0)^{\mathrm{T}}$, $X \in \mathbb{R}^{24}$, and x_i $(i = 1, \ldots, 24)$ and ε are independent and identically distributed standard normal variates. Hence, there are 21 zeros out of a total of 24 coefficients. We computed Lasso and shrinkage sliced inverse regression estimates where the tuning parameter λ is selected via criteria discussed in § 2·1. To make comparisons, we report the average number of zero coefficients, the mean and standard deviation of the correlation r between the estimated $\hat{\eta}^{\mathrm{T}} X$ and the true predictor $\eta^{\mathrm{T}} X$, and the median of the mean squared error $\hat{E}(\hat{\eta}^{\mathrm{T}} X - \eta^{\mathrm{T}} X)^2$, where $\hat{\eta}$ is the estimate of η. For the Lasso approach, Table 1 shows that the Akaike information criterion is most conservative, the residual information criterion is most liberal, and all Lasso estimates perform better than ordinary least squares estimates. However Lasso methods underestimate the number of zero coefficients. In contrast, shrinkage sliced inverse regression results in much closer counts. It also performs better than Lasso in terms of the accuracy of the estimated sufficient predictors measured by the correlations.

Table 1. *Simulation results for Lasso and shrinkage sliced inverse regression based on the linear model $y = \eta^{\mathrm{T}} X + 0{\cdot}5\varepsilon$, where 21 η_i's are zeros*

	Lasso				SSIR			
Criterion	Ave 0's	r mean	r SD	Median MSE	Ave 0's	$\lvert r\rvert$ mean	$\lvert r\rvert$ SD	Median MSE
				$n = 60$				
OLS/SIR	0·00	0·981	0·007	0·114	0·00	0·958	0·015	0·258
GCV	10·86	0·990	0·005	0·064	19·64	0·994	0·006	0·040
AIC	10·11	0·990	0·006	0·065	19·52	0·994	0·006	0·040
BIC	14·02	0·993	0·005	0·058	20·96	0·994	0·006	0·040
RIC	15·21	0·993	0·005	0·057	20·99	0·991	0·009	0·064
				$n = 120$				
OLS/SIR	0·00	0·990	0·003	0·060	0·00	0·981	0·006	0·130
GCV	10·53	0·995	0·003	0·030	19·87	0·998	0·002	0·016
AIC	10·27	0·995	0·003	0·031	19·83	0·998	0·002	0·016
BIC	14·95	0·997	0·003	0·026	20·91	0·998	0·002	0·010
RIC	15·63	0·997	0·002	0·026	21·00	0·998	0·003	0·012

OLS, ordinary least squares; SIR, sliced inverse regression; GCV, generalised cross-validation; AIC, Akaike information criterion; BIC, Bayesian information criterion; RIC, residual information criterion; n, sample size; SSIR, shrinkage sliced inverse regression; Ave 0's, the average number of zero coffficients in $\hat{\eta}$; SD, sample standard deviation of r or $\lvert r\rvert$; Median MSE, the median of mean squared error $\hat{E}(\hat{\eta}^{\mathrm{T}} X - \eta^{\mathrm{T}} X)^2$.

Lasso methods generally target linear models, while shrinkage sliced inverse regression methods can be applied to single-index and multiple-index regressions as long as the linearity condition holds approximately. For example, we applied shrinkage sliced inverse regression to single-index models and heteroscedastic error models such as $y = g(\eta^{\mathrm{T}} X) + 0{\cdot}5\varepsilon$ and $y = g(\eta^{\mathrm{T}} X)\varepsilon$, where $g(\eta^{\mathrm{T}} X) = \exp(x_1 + 0{\cdot}5x_2 + x_3)$. Unreported results are qualitatively similar to those of the linear model. Finally, we consider the two-index model

$$y = \beta_1^{\mathrm{T}} X/\{0{\cdot}5 + (\beta_2^{\mathrm{T}} X + 1{\cdot}5)^2\} + 0{\cdot}2\varepsilon,$$

where $\beta_1 = (1, 0, \ldots, 0)^{\mathrm{T}}$, $\beta_2 = (0, 1, 0, \ldots, 0)^{\mathrm{T}}$, $X \in \mathbb{R}^{24}$, and x_i $(i = 1, \ldots, 24)$ and ε are independent and identically distributed standard normal variates. To assess the impact of correlated predictors on the performance of shrinkage sliced inverse regression, we also adopt Tibshirani's (1996) correlated predictors setting to generate a multivariate normal vector $X \in \mathbb{R}^{24}$ with $\mathrm{cov}(x_i, x_j) = 0{\cdot}5^{|i-j|}$ for this model. The results of both cases are presented in Table 2, where $r_i = \mathrm{corr}(\beta_i^{\mathrm{T}} X, \hat{\beta}_i^{\mathrm{T}} X)$ $(i = 1, 2)$. The performance of shrinkage sliced inverse regression estimators is only slightly affected by the correlations among the predictors. The residual information criterion performs best, followed by the Bayesian information criterion. Unreported results also show that increasing sample sizes improves the performance.

Table 2. *Simulation results for shrinkage sliced inverse regression based on the two-index model* $y = \beta_1^{\mathrm{T}} X/\{0{\cdot}5 + (\beta_2^{\mathrm{T}} X + 1{\cdot}5)^2\} + 0{\cdot}2\varepsilon$, *where 22 rows of* $\beta = (\beta_1, \beta_2)$ *are all zeros*

	Independent predictors					Correlated predictors				
Criterion	Ave 0's	$\|r_1\|$ mean	$\|r_1\|$ SD	$\|r_2\|$ mean	$\|r_2\|$ SD	Ave 0's	$\|r_1\|$ mean	$\|r_1\|$ SD	$\|r_2\|$ mean	$\|r_2\|$ SD
					$n = 60$					
SIR	0·00	0·857	0·050	0·576	0·193	0·00	0·871	0·055	0·648	0·114
GCV	5·98	0·912	0·045	0·632	0·224	5·34	0·915	0·051	0·691	0·124
AIC	5·33	0·907	0·046	0·627	0·221	4·66	0·910	0·052	0·686	0·126
BIC	14·06	0·961	0·035	0·697	0·262	13·39	0·961	0·040	0·752	0·154
RIC	18·67	0·984	0·025	0·738	0·301	18·42	0·985	0·022	0·781	0·189
					$n = 120$					
SIR	0·00	0·934	0·021	0·776	0·113	0·00	0·936	0·018	0·799	0·074
GCV	7·69	0·967	0·017	0·843	0·121	7·03	0·966	0·017	0·851	0·078
AIC	7·14	0·965	0·017	0·838	0·121	6·43	0·964	0·017	0·847	0·079
BIC	14·64	0·989	0·009	0·906	0·121	14·23	0·986	0·012	0·906	0·080
RIC	19·71	0·998	0·003	0·953	0·122	19·25	0·996	0·007	0·948	0·076

SIR, sliced inverse regression; GCV, generalised crossvalidation; AIC, Akaike information criterion; BIC, Bayesian information criterion; RIC, residual information criterion; n, sample size; Ave 0's, the average number of zero rows in $\hat{\beta}$; SD, sample standard deviation of $|r_i|$.

4. DISCUSSION

The shrinkage approach introduced here can be extended to other dimension reduction methods, including sliced average variance (Cook & Weisberg, 1991) and principal Hessian direction (Li, 1992; Cook, 1998b) estimators. The shrinkage sliced inverse regression approach can also be extended to categorical covariates (Chiaromonte et al., 2002) and binary response models (Cook & Lee, 1999). We believe these efforts would strengthen the applicability of shrinkage methods to dimension reduction.

Acknowledgement

We thank the referee and editor for their useful comments. R. D. Cook's research was supported in part by the U.S. National Science Foundation and Chih-Ling Tsai's research was supported in part by the U.S. National Institutes of Health.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *2nd International Symposium on Information Theory*, Ed. B. N. Petrov and F. Csaki, pp. 267–81. Budapest: Akademia Kiado.

Breiman, L. (1995). Better subset regression using the nonnegative garrote. *Technometrics* **37**, 373–84.

Bura, E. & Cook, R. D. (2001). Extending sliced inverse regression: the weighted chi-squared test. *J. Am. Statist. Assoc.* **96**, 996–1003.

Chiaromonte, F., Cook, R. D. & Li, B. (2002). Sufficient dimension reduction in regressions with categorical predictors. *Ann. Statist.* **30**, 475–97.

Cook, R. D. (1996). Graphics for regressions with a binary response. *J. Am. Statist. Assoc.* **91**, 983–92.

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

Cook, R. D. (1998b). Principal hessian directions revisited (with Discussion). *J. Am. Statist. Assoc.* **93**, 84–100.

Cook, R. D. (2004). Testing predictor contributions in sufficient dimension reduction. *Ann. Statist.* **32**, 1062–92.

Cook, R. D. & Lee, H. (1999). Dimension reduction in binary response regression. *J. Am. Statist. Assoc.* **94**, 1187–200.

Cook, R. D. & Weisberg, S. (1991). Discussion of a paper by K.-C. Li. *J. Am. Statist. Assoc.* **86**, 328–32.

Fu, W. J. (1998). Penalized regression: the bridge versus the lasso. *J. Comp. Graph. Statist.* **7**, 397–416.

Li, K.-C. (1991). Sliced inverse regression for dimension reduction (with Discussion). *J. Am. Statist. Assoc.* **86**, 316–42.

Li, K.-C. (1992). On principal hessian directions for data visualization and dimension reduction: another application of Stein's lemma. *J. Am. Statist. Assoc.* **87**, 1025–39.

Naik, P. A. & Tsai, C.-L. (2001). Single-index model selections. *Biometrika* **88**, 821–32.

Osborne, M. R., Presnell, B. & Turlach, B. A. (2000). On the lasso and its dual. *J. Comp. Graph. Statist.* **9**, 319–37.

Schott, J. R. (1994). Determining the dimensionality in sliced inverse regression. *J. Am. Statist. Assoc.* **89**, 141–8.

Schwarz, G. (1978). Estimation the dimension of a model. *Ann. Statist.* **6**, 461–4.

Shi, P. & Tsai, C.-L. (2002). Regression model selection—a residual likelihood approach. *J. R. Statist. Soc.* B **64**, 237–52.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Statist. Soc.* B **58**, 267–88.

[*Received June* 2004. *Revised August* 2004]