Biometrika (2007), **94**, 3, *pp*. 603–613
© 2007 Biometrika Trust
Printed in Great Britain

doi:10.1093/biomet/asm044
Advance Access publication 5 August 2007

Sparse sufficient dimension reduction

BY LEXIN LI

Department of Statistics, North Carolina State University, Raleigh, North Carolina 27695, U.S.A.

li@stat.ncsu.edu

SUMMARY

Existing sufficient dimension reduction methods suffer from the fact that each dimension reduction component is a linear combination of all the original predictors, so that it is difficult to interpret the resulting estimates. We propose a unified estimation strategy, which combines a regression-type formulation of sufficient dimension reduction methods and shrinkage estimation, to produce sparse and accurate solutions. The method can be applied to most existing sufficient dimension reduction methods such as sliced inverse regression, sliced average variance estimation and principal Hessian directions. We demonstrate the effectiveness of the proposed method by both simulations and real data analysis.

Some key words: Lasso; Shrinkage sparse estimator; Sufficient dimension reduction.

1. INTRODUCTION

In the analysis of high-dimensional data, the theory of sufficient dimension reduction (Li, 1991; Cook, 1998a) has been developed to reduce the dimension of the problem prior to model formulation, while preserving full regression information and imposing few probabilistic assumptions. For regression problems involving a univariate response Y and a p-dimensional predictor vector $X \in \mathbb{R}^p$, sufficient dimension reduction seeks a subspace $\mathcal{S}$ of minimal dimension such that $Y \perp\!\!\!\perp X | P_{\mathcal{S}} X$, where $\perp\!\!\!\perp$ stands for independence, and $P_{\mathcal{S}}$ denotes the orthogonal projection on to $\mathcal{S}$. Such a space exists and is unique under mild conditions (Cook, 1996). We call it the central subspace for the regression of Y on X, denote it by $\mathcal{S}_{Y|X}$, and call its dimension, $d = \dim(\mathcal{S}_{Y|X})$, which is often far less than p, the structural dimension of regression (Cook, 1998a). If $\eta \in \mathbb{R}^{p \times d}$ denotes a basis of $\mathcal{S}_{Y|X}$, $\eta^{\mathsf{T}} X$ carries all information that X has about Y.

Methods for estimating $\mathcal{S}_{Y|X}$ include sliced inverse regression (Li, 1991), sliced average variance estimation (Cook & Weisberg, 1991) and canonical correlation estimation (Fung et al., 2002). Li (1992) & Cook (1998b) proposed principal Hessian directions, and Cook & Li (2002) proposed iterative Hessian transformation, both of which estimate a variant of the central subspace that contains all the information about Y that is available through $E(Y|X)$. Also, Chiaromonte et al. (2002) proposed partial sliced inverse regression for estimating the partial central subspace, and Yin & Cook (2002) proposed a covariance method for estimating the central kth moment subspace.

However, all the methods mentioned above produce linear combinations of all the original predictors, and this often makes it difficult to interpret the extracted components. As a simple illustration, consider a response model $Y = \exp(-0{\cdot}5\beta_1^{\mathsf{T}} X) + 0{\cdot}5\varepsilon$, where X has $p = 6$ dimensions, and all predictors and the error ε are independent standard normal variables.

The central subspace is spanned by $\beta_1 = (1, -1, 0, 0, 0, 0)^{\mathsf{T}}/\sqrt{2}$. Sliced inverse regression, for instance, yielded an estimate $\hat{\beta}_1 = (0{\cdot}651, -0{\cdot}745, -0{\cdot}063, 0{\cdot}134, 0{\cdot}014, 0{\cdot}003)^{\mathsf{T}}$ based on simulated data with 100 samples. Although the last four coefficients are relatively small, all predictors are included in the estimate, which obscures the fact that the true β_1 only involves the first two predictors. To improve interpretability, Chen & Li (1998) proposed an approximate formula for standard deviations of sliced inverse regression estimates. Cook (2004) developed a rigorous conditional independence test procedure to assess the contribution of individual predictors in the extracted sliced inverse regression components. Ni et al. (2005) and Li & Nachtsheim (2006) both combined the least absolute shrinkage and selection operator, or lasso (Tibshirani, 1996), with sliced inverse regression to produce sparse estimates. These methods all rely on special characteristics of sliced inverse regression, and extensions to other dimension reduction methods are difficult.

Motivated by the recent development of sparse principal components analysis by Zou et al. (2006), we propose in this article a unified approach to produce sparse estimates of the basis of the central subspace and its variants.

2. SPARSE SUFFICIENT DIMENSION REDUCTION

2·1. *Sufficient dimension reduction estimator*

Sufficient dimension reduction methods can often be formulated as a generalized eigenvalue problem of the form

$$M\upsilon_i = \rho_i G\upsilon_i, \quad \text{for } i = 1, \ldots, p, \tag{1}$$

where M is a method-specific symmetric kernel matrix and is nonnegative definite; G is a symmetric and positive definite matrix, often taking the form of the covariance matrix Σ_x of X; vectors $\upsilon_1, \ldots, \upsilon_p$ are eigenvectors satisfying $\upsilon_i^{\mathsf{T}} G \upsilon_j = 1$ if $i = j$, and 0 if $i \neq j$; and $\rho_1 \geqslant \cdots \geqslant \rho_p \geqslant 0$ are corresponding eigenvalues. Table 1 gives a summary of M and G matrices for a number of commonly used dimension reduction estimators. When $M = \Sigma_x$ and $G = I_p$, (1) becomes principal components analysis.

Under method-specific conditions, which are usually imposed on the marginal distribution of X only, it can be shown that the eigenvectors $\{\upsilon_1, \ldots, \upsilon_d\}$ in (1) that correspond to the nonzero eigenvalues $\{\rho_1 \geqslant \cdots \geqslant \rho_d > 0\}$ form a basis for the central subspace under enquiry. There are both asymptotic tests and a permutation test (Cook & Yin, 2001) for determining the number of nonzero eigenvalues in (1), which in

Table 1. *The generalized eigenvalue formulation of sufficient dimension reduction methods.*

Sliced inverse regression:	
$M = \mathrm{cov}[E\{X - E(X)\|Y\}]$	$G = \Sigma_x$
Sliced average variance estimation:	
$M = \Sigma_x^{1/2} E[\{I - \mathrm{cov}(Z\|Y)\}^2]\Sigma_x^{1/2}$, where $Z = \Sigma_x^{-1/2}\{X - E(X)\}$	$G = \Sigma_x$
Principal Hessian directions (y-based):	
$M = \Sigma_x^{1/2}\Sigma_{yzz}\Sigma_{yzz}\Sigma_x^{1/2}$, where $\Sigma_{yzz} = E[\{Y - E(Y)\}ZZ^{\mathsf{T}}]$	$G = \Sigma_x$
Principal Hessian directions (r-based):	
$M = \Sigma_x^{1/2}\Sigma_{rzz}\Sigma_{rzz}\Sigma_x^{1/2}$, where $\Sigma_{rzz} = E[\{Y - E(Y) - E(YZ^{\mathsf{T}})Z\}ZZ^{\mathsf{T}}]$	$G = \Sigma_x$
Iterative Hessian transformation:	
$M = \Sigma_x^{1/2}\tilde{\Sigma}_{yzz}\tilde{\Sigma}_{yzz}^{\mathsf{T}}\Sigma_x^{1/2}$, where $\tilde{\Sigma}_{yzz} = (\beta_{yz}, \ldots, \Sigma_{yzz}^{p-1}\beta_{yz})$, $\beta_{yz} = E(YZ)$	$G = \Sigma_x$

turn determines the structural dimension $d = \dim(\mathcal{S}_{Y|X})$. Thus d is treated as known in the following derivation of sparse dimension reduction.

2·2. *The sparse sufficient dimension reduction estimator*

We first transform the eigenvalue problem (1) to a regression-type optimization problem.

PROPOSITION 1. *Let $m_i, i = 1, \ldots, p$, denote the columns of $M^{1/2}$, the square-root of the symmetric sufficient dimension reduction kernel matrix M, and let β be a $p \times d$ matrix. Let*

$$\hat{\beta} = \arg\min_{\beta} \sum_{i=1}^{p} \|G^{-1}m_i - \beta\beta^{\mathsf{T}}m_i\|_G^2, \tag{2}$$

subject to $\beta^{\mathsf{T}}G\beta = I_d$, where the norm is with respect to the G inner product. Then $\hat{\beta}_j = \upsilon_j$, $j = 1, \ldots, d$, where $\hat{\beta}_j$ is the jth column of $\hat{\beta}$.

The proof is straightforward since the objective function in (2) can be written as

$$\begin{aligned}\sum_{i=1}^{p} \|G^{-1}m_i - \beta\beta^{\mathsf{T}}m_i\|_G^2 &= \sum_{i=1}^{p} \|G^{-1/2}m_i - (G^{1/2}\beta)(G^{1/2}\beta)^{\mathsf{T}}G^{-1/2}m_i\|^2 \\ &= \sum_{i=1}^{p} \|\tilde{m}_i - \tilde{\beta}\tilde{\beta}^{\mathsf{T}}\tilde{m}_i\|^2,\end{aligned}$$

where $\tilde{m}_i = G^{-1/2}m_i$ and $\tilde{\beta} = G^{1/2}\beta$. Since $\tilde{\beta}\tilde{\beta}^{\mathsf{T}}$ is the projection operator on to the space spanned by the columns of $\tilde{\beta}$, $\sum_{i=1}^{p} \|\tilde{m}_i - \tilde{\beta}\tilde{\beta}^{\mathsf{T}}\tilde{m}_i\|^2 = \mathrm{tr}(\tilde{M}) - \mathrm{tr}(\tilde{\beta}^{\mathsf{T}}\tilde{M}\tilde{\beta})$, where $\tilde{M} = \sum_{i=1}^{p} \tilde{m}_i\tilde{m}_i^{\mathsf{T}} = G^{-1/2}MG^{-1/2}$. Thus the minimizer is composed of the eigenvectors of the matrix $\tilde{M}$. After some algebra, the conclusion of Proposition 1 follows.

Based on (2), one may impose the lasso constraint directly on β to produce a sparse estimate; that is, for shrinkage parameters τ_j, consider the optimization problem

$$\min_{\beta} \sum_{i=1}^{p} \|G^{-1}m_i - \beta\beta^{\mathsf{T}}m_i\|_G^2, \text{ subject to } \beta^{\mathsf{T}}G\beta = I_d, \text{ and } |\beta_j|_1 \leqslant \tau_j, \tag{3}$$

where β_j denotes the jth column of β, $j = 1, \ldots, d$, and $|\beta_j|_1 = \sum_{k=1}^{p} |\beta_{jk}|$ is the sum of absolute values of all the components of β_j. This is the idea behind a method of Jolliffe et al. (2003) for producing sparse principal components. However, optimization of (3) is complicated, with many local optima; see Jolliffe et al. (2003) & Zou et al. (2006). Instead of optimizing (3) directly, we consider an alternative formulation of (2), following the idea of Zou et al. (2006), that can result in a more efficient algorithm when we introduce the lasso constraint.

PROPOSITION 2. *Let $m_i, i = 1, \ldots, p$, denote the columns of $M^{1/2}$ as defined before, and let α and β be $p \times d$ matrices. For any $\lambda_2 > 0$, let*

$$(\hat{\alpha}, \hat{\beta}^*) = \arg\min_{\alpha,\beta} \left\{ \sum_{i=1}^{p} \|G^{-1}m_i - \alpha\beta^{\mathsf{T}}m_i\|_G^2 + \lambda_2 \,\mathrm{tr}(\beta^{\mathsf{T}}G\beta) \right\}, \tag{4}$$

subject to $\alpha^{\mathsf{T}}G\alpha = I_d$, where the norm is with respect to the G inner product. Then $\hat{\beta}_j^* \propto \upsilon_j$, $j = 1, \ldots, d$, where $\hat{\beta}_j^*$ is the jth column of $\hat{\beta}^*$.

The proof is given in the Appendix. A strictly positive ridge penalty, $\lambda_2 \operatorname{tr}(\beta^{\mathsf{T}} G \beta)$, is introduced to ensure the reconstruction of eigenvectors υ_j through the formulation (4). Based on this, we consider the optimization problem

$$\min_{\alpha,\beta} \left\{ \sum_{i=1}^{p} \| G^{-1} m_i - \alpha \beta^{\mathsf{T}} m_i \|_G^2 + \lambda_2 \operatorname{tr}(\beta^{\mathsf{T}} G \beta) + \sum_{j=1}^{d} \lambda_{1j} |\beta_j|_1 \right\}, \tag{5}$$

subject to $\alpha^{\mathsf{T}} G \alpha = I_d$, where $\lambda_{1j} \geqslant 0$, $j = 1, \ldots, d$, are lasso shrinkage parameters. We call the solution $\hat{\beta}$ of (5) the sparse sufficient dimension reduction estimator. As a result of the lasso constraint in (5), the resulting estimator is expected to have some coefficients shrunk to zero, which may lead to easier interpretation.

2·3. *Numerical algorithm*

The objective function in (5) can be written as

$$\mathcal{L} = \operatorname{tr}(G^{-1/2} M G^{-1/2}) + \sum_{j=1}^{d} \left\{ \beta_j^{\mathsf{T}} (M + \lambda_2 G) \beta_j - 2 \alpha_j^{\mathsf{T}} M \beta_j + \lambda_{1j} |\beta_j|_1 \right\}, \tag{6}$$

where α_j is the jth column of the matrix α, $j = 1, \ldots, d$. Equation (6) can be easily obtained by following equation (A1) in the Appendix. Furthermore we have the following proposition.

PROPOSITION 3. *Given α, let*

$$\hat{\beta}_{\alpha j} = \arg\min_{\beta_j} \left\{ \beta_j^{\mathsf{T}} (M + \lambda_2 G) \beta_j - 2 \alpha_j^{\mathsf{T}} M \beta_j + \lambda_{1j} |\beta_j|_1 \right\}, \tag{7}$$

and let

$$\hat{\theta}_{\alpha j} = \arg\min_{\theta_j} \{ \| u^* - m^* \theta_j \|^2 + \lambda_{1j} |\theta_j|_1 \}, \tag{8}$$

where

$$m^* = \begin{pmatrix} M^{1/2} \\ \sqrt{\lambda_2} G^{1/2} \end{pmatrix}_{2p \times p}, \qquad u^* = \begin{pmatrix} M^{1/2} \alpha_j \\ 0 \end{pmatrix}_{2p \times 1}.$$

Then $\hat{\beta}_{\alpha j} = \hat{\theta}_{\alpha j}$.

Proposition 3 follows because $m^{*\mathsf{T}} m^* = M + \lambda_2 G$ and $u^{*\mathsf{T}} m^* = \alpha_j^{\mathsf{T}} M$. Since the solution to (8) is simply a lasso estimate with u^* as the response and m^* as the predictors, Proposition 3 implies that, given α, minimization of (6) over the β_j's is equivalent to d independent lasso optimization problems, and the solutions can be obtained by any lasso algorithm. In our sparse estimation procedure, we employ the recently proposed least angle regression method (Efron et al., 2004), which can solve the whole lasso solution path efficiently with the same order of computational complexity as a single ordinary least squares.

Next, given β, the objective function $\mathcal{L}$ can be written as

$$\mathcal{L} = \operatorname{tr}(G^{-1/2} M G^{-1/2}) - 2 \operatorname{tr}(\alpha^{\mathsf{T}} M \beta) + \operatorname{tr}\{\beta^{\mathsf{T}} (M + \lambda_2 G) \beta\} + \sum_{j=1}^{d} \lambda_{1j} |\beta_j|_1.$$

Its minimizer, subject to the constraint that $\alpha^{\mathsf{T}} G \alpha = I_d$, is given by the next proposition.

PROPOSITION 4. *Let α and β be $p \times d$ matrices, let β have rank d, let*

$$\hat{\alpha}_\beta = \arg\max_\alpha \operatorname{tr}(\alpha^\mathsf{T} M \beta), \tag{9}$$

subject to $\alpha^\mathsf{T} G \alpha = I_d$, and let U, D and V denote the matrices from the singular value decomposition of the matrix $G^{-1/2} M \beta$, i.e., $G^{-1/2} M \beta = U D V^\mathsf{T}$. Then $\hat{\alpha}_\beta = G^{-1/2} U V^\mathsf{T}$.

The proof is given in the Appendix. Based on Propositions 3 and 4, we propose the following alternating minimization algorithm for solving (5).

Step 1. Choose the usual sufficient dimension reduction estimator with no lasso constraint as an initial value for α.

Step 2. Given fixed α, solve d independent lasso problems (7) to obtain the estimate of $\beta = (\beta_1, \ldots, \beta_d)$.

Step 3. For fixed β, carry out singular value decomposition of $G^{-1/2} M \beta = U D V^\mathsf{T}$, and update $\alpha = G^{-1/2} U V^\mathsf{T}$.

Step 4. Repeat Steps 2 and 3 until β converges.

Step 5. Normalize β as $\beta_j = \beta_j / \|\beta_j\|$, $j = 1, \ldots, d$.

For the given shrinkage parameters $\lambda_1 = (\lambda_{11}, \ldots, \lambda_{1d})$ and λ_2, the algorithm produces a monotonically decreasing series of evaluations of the nonnegative objective function $\mathcal{L}$ in (5), and thus the algorithm is guaranteed to converge. Our experience through extensive simulations suggests that the algorithm often converges fast, and it seems typically to converge to the global minimum.

To choose the tuning parameters λ_1 and λ_2, we propose a criterion function which has a form similar to the Akaike information criterion (Akaike, 1973),

$$\sum_{i=1}^{p} \|G^{-1} m_i - \hat{\beta}_\lambda \hat{\beta}_\lambda^\mathsf{T} m_i\|_G^2 + 2 p_\lambda / n, \tag{10}$$

where $\hat{\beta}_\lambda$ denotes the solution for β given λ_1 and λ_2, p_λ denotes the effective number of parameters, and n is the sample size. Following the discussion of Zou et al. (2007) about the degrees of freedom of the lasso estimator, we estimate p_λ by the number of nonzero components of $\hat{\beta}_\lambda$. The first term in (10) is minimized by the usual dimension reduction estimator with no lasso constraint, and thus favours a less sparse estimator, whereas the second term in (10) penalizes the number of nonzero components. Empirical evidence suggests that (10) works fairly well. In practice, we often set $\lambda_{11} = \cdots = \lambda_{1d}$, and among a grid values of (λ_1, λ_2) we choose the pair such that (10) is minimized.

Consider briefly the illustrative example in § 1 again, where 100 samples of the model $Y = \exp(-0{\cdot}5\beta_1^\mathsf{T} X) + 0{\cdot}5\varepsilon$ were obtained, with $\beta_1 = (1, -1, 0, 0, 0, 0)^\mathsf{T}/\sqrt{2}$. Whereas sliced inverse regression yielded the estimate $\hat{\beta}_1 = (0{\cdot}651, -0{\cdot}745, -0{\cdot}063, 0{\cdot}134, 0{\cdot}014, 0{\cdot}003)^\mathsf{T}$, the sparse version proposed in this section produced the estimate $\hat{\beta}_1 = (0{\cdot}716, -0{\cdot}698, 0, 0, 0, 0)^\mathsf{T}$, in which the last four coordinates were correctly shrunk to zero.

3. SIMULATIONS AND EXAMPLES

3·1. *Sparse sliced inverse regression estimator*

We first consider an example demonstrating sparse sliced inverse regression. Let

$$Y_1 = \operatorname{sign}(\beta_1^\mathsf{T} X) \log(|\beta_2^\mathsf{T} X + 5|) + 0{\cdot}2\varepsilon,$$

where X has $p = 20$ dimensions, and all predictors and the error ε are independent standard normal variables. The central subspace is spanned by $\beta = (\beta_1, \beta_2)$, and the two vectors take three different forms: (i) $\beta_1 = (1, 1, 1, 1, 0, \ldots, 0)^{\mathsf{T}}$ and $\beta_2 = (0, \ldots, 0, 1, 1, 1, 1)^{\mathsf{T}}$; (ii) $\beta_1 = (1, 1, 0{\cdot}1, 0{\cdot}1, 0, \ldots, 0)^{\mathsf{T}}$ and $\beta_2 = (0, \ldots, 0, 0{\cdot}1, 0{\cdot}1, 1, 1)^{\mathsf{T}}$; (iii) $\beta_1 = (1, \ldots, 1, 0, \ldots, 0)^{\mathsf{T}}$ and $\beta_2 = (0, \ldots, 0, 1, \ldots, 1)^{\mathsf{T}}$, where there are 10 coordinates equal to one in each direction. The sample size was taken as $n = 200$. To evaluate estimation accuracy, we recorded the number of zero components in each estimated direction $\hat{\beta}_1$ and $\hat{\beta}_2$. We also computed the absolute correlation between the estimated predictor $\hat{\beta}_j^{\mathsf{T}} X$ and the true one $\beta_j^{\mathsf{T}} X$, plus the mean squared error, $\hat{E}(\hat{\beta}_j^{\mathsf{T}} X - \beta_j^{\mathsf{T}} X)^2$, for $j = 1, 2$. Furthermore, to evaluate $\hat{\beta} = (\hat{\beta}_1, \hat{\beta}_2)$ jointly, we employed the vector correlation coefficient, $(\prod_{i=1}^{d} \phi_i^2)^{1/2}$, where the ϕ_i^2's are the eigenvalues of the matrix $\hat{\beta}^{\mathsf{T}}{}_{\mathrm{on}} \beta_{\mathrm{on}} \beta^{\mathsf{T}}{}_{\mathrm{on}} \hat{\beta}_{\mathrm{on}}$, with β_{on} and $\hat{\beta}_{\mathrm{on}}$ denoting the orthonormalized versions of β and $\hat{\beta}$ respectively (Ye & Weiss, 2003). This measure ranges between 0 and 1, with a larger value indicating a better estimate.

Table 2 shows the average results based on 100 data replications. In case (i), in which the active predictors are sparse in each direction, the sparse estimator performed well, achieving both sparsity and high estimation accuracy. For instance, the true number of zero components in $\hat{\beta}_1$ was 16, and the average estimated number was 15·16. In addition, the sparse estimator achieved higher absolute correlation and lower mean squared error compared with the usual estimator with no lasso constraint. For case (ii), the sparse estimator seems to shrink more components to zero than it should; for $\hat{\beta}_1$, the estimated number of zeros was 17·67, and this is because the magnitude of the third and the fourth components in β_1 was ten times less than those of the first two components in this direction. However, the sparse estimator still yielded higher correlation and lower mean squared error than the unconstrained estimator. In case (iii), both estimators performed similarly, with the unconstrained one showing a slight edge; this is because a fair number of components are truly active. These observations reflect the recognized nature of the lasso, that it generally favours and is most useful for the case where there are many inactive predictors.

3·2. *Sparse principal Hessian directions estimator*

We next consider an example from Li (1992), in which

$$Y_2 = \cos(2\beta_1^{\mathsf{T}} X) - \cos(\beta_2^{\mathsf{T}} X) + 0{\cdot}5\varepsilon,$$

Table 2. *Simulations comparing sliced inverse regression* (SIR) *and sparse sliced inverse regression* (S-SIR). *The average of the number of zero components* (NUM)*, the absolute correlation* (COR)*, the mean squared error* (MSE)*, and the vector correlation coefficient* (VCC)*, based on* 100 *data replications, are reported*

		$\hat{\beta}_1$			$\hat{\beta}_2$			$(\hat{\beta}_1, \hat{\beta}_2)$
		NUM	COR	MSE	NUM	COR	MSE	VCC
Case (i)	SIR	0·000	0·926	1·604	0·000	0·911	1·245	0·934
	S-SIR	15·16	0·975	1·352	15·38	0·974	1·026	0·946
Case (ii)	SIR	0·000	0·884	0·551	0·000	0·856	0·544	0·932
	S-SIR	17·67	0·984	0·245	17·68	0·986	0·205	0·968
Case (iii)	SIR	0·000	0·916	4·793	0·000	0·942	4·168	0·917
	S-SIR	9·220	0·877	5·006	9·630	0·908	4·329	0·816

where X has $p = 10$ dimensions, and all predictors and the error ε are independent standard normal random variables. The central subspace is spanned by $\beta_1 = (1, 0, \ldots, 0)^{\mathsf{T}}$ and $\beta_2 = (0, 1, 0, \ldots, 0)^{\mathsf{T}}$. Following Cook (1998b), we employed the residual-based principal Hessian directions estimator. Table 3 presents the average results out of 100 data replications. Two sample sizes, $n = 100$ and $n = 200$, were examined. It is clearly seen that the sparse principal Hessian directions estimator has led to substantial improvement in estimation accuracy in this very sparse case. For instance, when $n = 100$, the average absolute correlation of the unconstrained estimator for $\hat{\beta}_2$ was 0·726, and it became 0·902 for the sparse estimator. The estimated number of zeros by the sparse estimator was 8·140, close to the true value 9. When the sample size n increased, the performance of each estimator improved, with the sparse estimator consistently outperforming the other.

3·3. *Sparse sliced average variance estimator*

Cook & Critchley (2000) and Cook & Yin (2001) both applied the sliced average variance estimator to a dataset on counterfeit Swiss bank notes. The data consist of 200 observations of a binary response variable, which indicates a note's authenticity, with $Y = 0$ for genuine notes and $Y = 1$ for counterfeit notes, and six predictors measuring the size of a note: length at centre, left-edge length, right-edge length, length of bottom edge, length of top edge and diagonal length. The permutation test suggested that the structural dimension of this problem is two. Figure 1(a) shows a summary plot of the first two sliced average variance estimates, with circles indicating genuine notes and dots indicating counterfeit notes. The resulting estimates were $(-0{\cdot}033 \times \text{Length} - 0{\cdot}200 \times \text{Left} + 0{\cdot}250 \times \text{Right} + 0{\cdot}594 \times \text{Bottom} + 0{\cdot}571 \times \text{Top} - 0{\cdot}466 \times \text{Diagonal})$, and $(-0{\cdot}284 \times \text{Length} - 0{\cdot}055 \times \text{Left} - 0{\cdot}158 \times \text{Right} + 0{\cdot}505 \times \text{Bottom} + 0{\cdot}333 \times \text{Top} + 0{\cdot}725 \times \text{Diagonal})$. Figure 1(a) shows that the genuine and counterfeit notes are well separated by those two estimates. An outlying authentic note and the bimodal distribution within the counterfeit notes are also clearly shown.

The sparse sliced average variance estimation method yielded the estimates $(0 \times \text{Length} + 0 \times \text{Left} + 0 \times \text{Right} + 0{\cdot}785 \times \text{Bottom} + 0{\cdot}619 \times \text{Top} + 0 \times \text{Diagonal})$ and $(0 \times \text{Length} + 0 \times \text{Left} + 0 \times \text{Right} + 0{\cdot}400 \times \text{Bottom} + 0 \times \text{Top} + 0{\cdot}917 \times \text{Diagonal})$. Interpretation of the sparse estimates becomes simpler: the first estimate measures the length of top and bottom edges, and the second one measures the bottom length and the diagonal length. Figure 1(b) plots the two sparse estimates, indicating the same features as in Fig. 1(a). The sparse method also suggests that it may be useful to test formally the conditional

Table 3. *Simulations comparing principal Hessian directions* (PHD) *and sparse principal Hessian directions* (S-PHD)*. The average of the number of zero components* (NUM)*, the absolute correlation* (COR)*, the mean squared error* (MSE)*, and the vector correlation coefficient* (VCC)*, based on* 100 *data replications, are reported.*

		$\hat{\beta}_1$			$\hat{\beta}_2$			$(\hat{\beta}_1, \hat{\beta}_2)$
		NUM	COR	MSE	NUM	COR	MSE	VCC
$n = 100$	PHD	0·000	0·793	0·437	0·000	0·726	0·406	0·651
	S-PHD	8·280	0·911	0·194	8·140	0·902	0·158	0·837
$n = 200$	PHD	0·000	0·894	0·184	0·000	0·933	0·150	0·848
	S-PHD	8·360	0·970	0·055	8·230	0·991	0·020	0·963

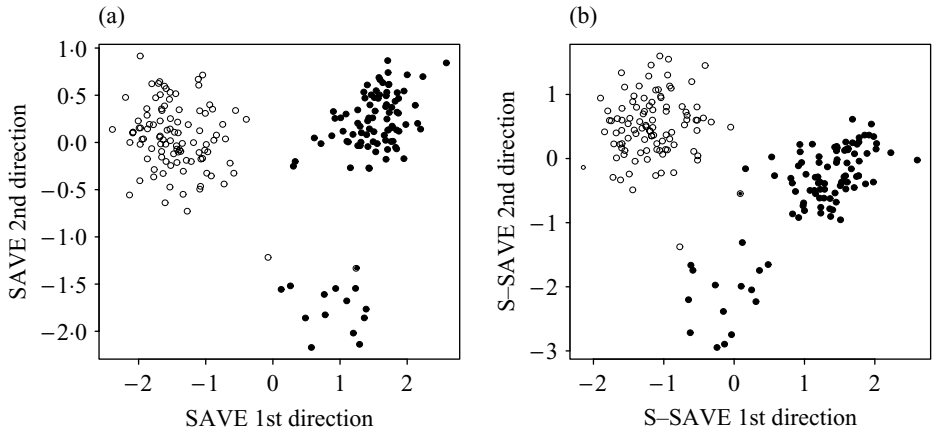

Fig. 1. Summary plot for Swiss bank notes data. Panel (a) shows the two sliced average variance estimates, and panel (b) shows the two sparse sliced average variance estimates. Circles indicate authentic notes and dots indicate counterfeit notes.

independence between the response and (Length, Left, Right), given the other three predictors.

3·4. *Wisconsin breast cancer data*

The Wisconsin breast cancer dataset was developed to study a diagnosis method based on fine-needle aspiration. The data contain 569 cases with two diagnoses, 212 malignant and 357 benign. There are 30 predictors, which correspond to the mean, the standard deviation and a tail average of the empirical distributions of ten characteristics of the cell nucleus: radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry and fractal dimension. It has been found that the first sliced inverse regression direction provides a good separation of malignant and benign cases. This estimated direction was $(-0{\cdot}508, 0{\cdot}013, 0{\cdot}382, 0{\cdot}074, 0{\cdot}001, -0{\cdot}147, 0{\cdot}074, 0{\cdot}055, 0{\cdot}002, 0{\cdot}000, 0{\cdot}080, -0{\cdot}002, -0{\cdot}030, -0{\cdot}028, 0{\cdot}031, 0{\cdot}001\ -0{\cdot}071, 0{\cdot}043, 0{\cdot}009, -0{\cdot}013, 0{\cdot}624, 0{\cdot}029, -0{\cdot}054, -0{\cdot}381, 0{\cdot}008, 0{\cdot}007, 0{\cdot}053, 0{\cdot}020, 0{\cdot}023, 0{\cdot}051)^{\mathsf{T}}$. We applied sparse sliced inverse regression, obtaining the estimate $(0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0{\cdot}685, 0{\cdot}294, 0, 0, 0, 0, 0, 0{\cdot}667, 0, 0)^{\mathsf{T}}$; that is, the separating direction depends on the tail averages of radius, texture and concave points. Figure 2 plots the data by the usual sliced inverse regression estimate and the sparse estimate, with circles indicating benign cases and crosses indicating malignant cases. Both estimates provided good separation of the two diagnosis groups. The two directions are also highly correlated, with the correlation equal to 0·959.

4. DISCUSSION

Our unified strategy for producing sparse dimension reduction estimators can be coupled with any of the methods mentioned in §1, so long as they can be presented in the generalized eigenvalue formulation (1). In addition, the strategy can be applied straightforwardly to dimension reduction methods for multivariate responses (Setodji & Cook, 2004). Our method can strengthen application of sufficient dimension reduction in the following ways. First, it is useful in contexts such as bioinformatics, in which sparsity in regression is often expected. Interpretation of the resulting sparse estimates is typically easier. Secondly,

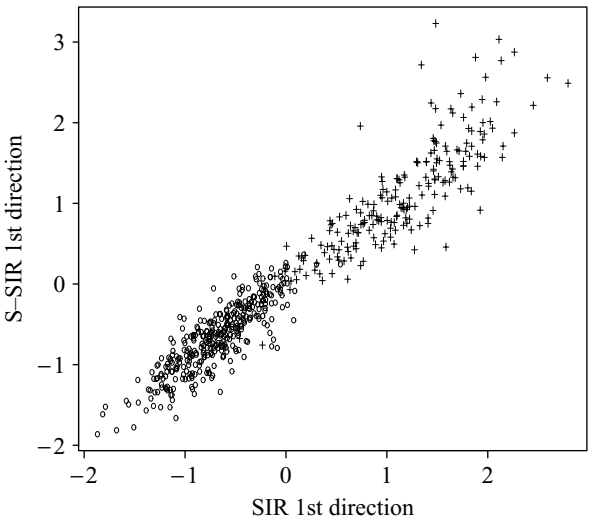

Fig. 2. Summary plot for Wisconsin breast cancer data. The horizontal axis is the sliced inverse regression estimator, and the vertical axis is the sparse sliced inverse regression estimator. Circles indicate benign cases and crosses indicate malignant cases.

although it cannot and should not serve as a substitute for the formal conditional independence test for predictors, the sparse method can possibly avoid a stepwise search procedure and point to relevant independence tests for the dimension reduction method, such as sliced inverse regression, where the predictor-independence test exists. The method can also be useful for those dimension reduction estimators where the corresponding tests are not yet available.

Recently, Cook & Ni (2005) proposed a new class of inverse regression estimators based on minimum discrepancy functions. Since the new estimator does not admit the eigen-decomposition form (1), the proposed sparse estimation strategy is not directly applicable. Extension of sparse estimation to this new family is currently under investigation, as is the predictor selection consistency of the sparse method.

Acknowledgement

The author is grateful to the editor and the referees for their constructive comments, which have greatly improved the paper.

Appendix

Proofs

Proof of Proposition 2. Let $\tilde{m}_i = G^{-1/2}m_i$, $\tilde{\beta} = G^{1/2}\beta$, $\tilde{\alpha} = G^{1/2}\alpha$ and $\tilde{M} = \sum_{i=1}^{p} \tilde{m}_i\tilde{m}_i^{\mathsf{T}}$. Then we have

$$\sum_{i=1}^{p} \|G^{-1}m_i - \alpha\beta^{\mathsf{T}}m_i\|_G^2 + \lambda_2 \operatorname{tr}(\beta^{\mathsf{T}}G\beta) = \sum_{i=1}^{p} \|\tilde{m}_i - \tilde{\alpha}\tilde{\beta}^{\mathsf{T}}\tilde{m}_i\|_{I_p}^2 + \lambda_2 \operatorname{tr}(\tilde{\beta}^{\mathsf{T}}\tilde{\beta})$$

$$= \operatorname{tr}(\tilde{M}) + \sum_{j=1}^{d} \left\{\tilde{\beta}_j^{\mathsf{T}}(\tilde{M} + \lambda_2 I_p)\tilde{\beta}_j - 2\tilde{\alpha}_j^{\mathsf{T}}\tilde{M}\tilde{\beta}_j\right\}. \quad \text{(A1)}$$

Thus, for a given $\tilde{\alpha}$, (A1) is minimized at $\tilde{\beta}_j^* = (\tilde{M} + \lambda_2 I_p)^{-1}\tilde{M}\tilde{\alpha}_j$, $j = 1, \ldots, d$, and consequently $\tilde{\beta}^* = (\tilde{M} + \lambda_2 I_p)^{-1}\tilde{M}\tilde{\alpha}$.

Minimization of (4) then becomes

$$\max_{\tilde{\alpha}} \operatorname{tr}\left\{\tilde{\alpha}^{\mathsf{T}}\tilde{M}(\tilde{M} + \lambda_2 I_p)^{-1}\tilde{M}\tilde{\alpha}\right\}, \quad \text{subject to } \tilde{\alpha}^{\mathsf{T}}\tilde{\alpha} = I_d.$$

The solution $\tilde{\alpha}^*$ consists of the first d eigenvectors of the matrix $\tilde{M}(\tilde{M} + \lambda_2 I_p)^{-1}\tilde{M}$, which are the same as the first d eigenvectors $\{\tilde{\upsilon}_1, \ldots, \tilde{\upsilon}_d\}$ of the matrix $\tilde{M}$. Therefore,

$$\tilde{\beta}_j^* = (\tilde{M} + \lambda_2 I_p)^{-1}\tilde{M}\tilde{\alpha}_j^* = (\tilde{M} + \lambda_2 I_p)^{-1}\tilde{M}\tilde{\upsilon}_j \propto \tilde{\upsilon}_j, \quad j = 1, \ldots, d.$$

Thus, $\hat{\beta}_j^* = G^{-1/2}\tilde{\beta}_j^* \propto G^{-1/2}\tilde{\upsilon}_j = \upsilon_j$, $j = 1, \ldots, d$. □

Proof of Proposition 4. Based on the notation in the proof of Proposition 2, the optimization problem (9) becomes

$$\arg\max_{\tilde{\alpha}} \operatorname{tr}(\tilde{\alpha}^{\mathsf{T}}\tilde{M}\tilde{\beta}), \quad \text{subject to } \tilde{\alpha}^{\mathsf{T}}\tilde{\alpha} = I_d. \tag{A2}$$

As a direct consequence of Theorem 4 of Zou et al. (2006), the solution to (A2) is $\tilde{\alpha}^* = UV^{\mathsf{T}}$, where U and V are obtained from singular value decomposition of $\tilde{M}\tilde{\beta} = G^{-1/2}M\beta$, and thus $\hat{\alpha}_\beta = G^{-1/2}\tilde{\alpha}^*$. □

REFERENCES

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

CHEN, C. H. & LI, K. C. (1998). Can SIR be as popular as multiple linear regression? *Statist. Sinica* **8**, 289–316.

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. *J. Am. Statist. Assoc.* **93**, 84–94.

COOK, R. D. (2004). Testing predictor contributions in sufficient dimension reduction. *Ann. Statist.* **32**, 1061–92.

COOK, R. D. & CRITCHLEY, F. (2000). Identifying outliers and regression mixtures graphically. *J. Am. Statist. Assoc.* **95**, 781–94.

COOK, R. D. & LI, B. (2002). Dimension reduction for the conditional mean in regression. *Ann. Statist.* **30**, 455–74.

COOK, R. D. & NI, L. (2005). Sufficient dimension reduction via inverse regression: a minimum discrepancy approach. *J. Am. Statist. Assoc.* **100**, 410–28.

COOK, R. D. & WEISBERG, S. (1991). Discussion of Li (1991). *J. Am. Statist. Assoc.* **86**, 328–32.

COOK, R. D. & YIN, X. (2001). Dimension reduction and visualization in discriminant analysis. *Aust. New Zeal. J. Statist.* **43**, 147–77.

EFRON, B., HASTIE, T., JOHNSTONE, I. & TIBSHIRANI, R. (2004). Least angle regression (with Discussion). *Ann. Statist.* **32**, 407–99.

FUNG, W. K., HE, X., LIU, L. & SHI, P. (2002). Dimension reduction based on canonical correlation. *Statist. Sinica* **12**, 1093–113.

JOLLIFFE, I. T., TRENDAFILOV, N. T. & UDDIN, M. (2003). A modified principal component technique based on the lasso. *J. Comp. Graph. Statist.* **12**, 531–47.

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. *Ann. Statist.* **87**, 1025–39.

LI, L. & NACHTSHEIM, C. J. (2006). Sparse sliced inverse regression. *Technometrics* **48**, 503–10.

NI, L., COOK, R. D. & TSAI, C. L. (2005). A note on shrinkage sliced inverse regression. *Biometrika* **92**, 242–7.

SETODJI, C. M. & COOK, R. D. (2004). K-means inverse regression. *Technometrics* **46**, 421–9.

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

YE, Z. & WEISS, R. E. (2003). Using the bootstrap to select one of a new class of dimension reduction methods. *J. Am. Statist. Assoc.* **98**, 968–79.
YIN, X. & COOK, R. D. (2002). Dimension reduction for the conditional *k*th moment in regression. *J. R. Statist. Soc.* B **64**, 159–75.
ZOU, H., HASTIE, T. & TIBSHIRANI, R. (2006). Sparse principal component analysis. *J. Comp. Graph. Statist.* **15**, 265–86.
ZOU, H., HASTIE, T. & TIBSHIRANI, R. (2007). On the degrees of freedom of the Lasso. *Ann. Statist.* In press.

[*Received December* 2005. *Revised December* 2006]