 DOI: 10.1111/j.1541-0420.2007.00836.x

Sliced Inverse Regression with Regularizations

Lexin Li[1,*] and Xiangrong Yin[2,]**

[1]Department of Statistics, North Carolina State University, Raleigh, North Carolina 27695, U.S.A.
[2]Department of Statistics, University of Georgia, Athens, Georgia 30602, U.S.A.
[*]*email:* li@stat.ncsu.edu
[**]*email:* xryin@stat.uga.edu

Summary. In high-dimensional data analysis, sliced inverse regression (SIR) has proven to be an effective dimension reduction tool and has enjoyed wide applications. The usual SIR, however, cannot work with problems where the number of predictors, p, exceeds the sample size, n, and can suffer when there is high collinearity among the predictors. In addition, the reduced dimensional space consists of linear combinations of all the original predictors and no variable selection is achieved. In this article, we propose a regularized SIR approach based on the least-squares formulation of SIR. The L_2 regularization is introduced, and an alternating least-squares algorithm is developed, to enable SIR to work with $n < p$ and highly correlated predictors. The L_1 regularization is further introduced to achieve simultaneous reduction estimation and predictor selection. Both simulations and the analysis of a microarray expression data set demonstrate the usefulness of the proposed method.

Key words: Regularized least squares; Sliced inverse regression; Sufficient dimension reduction.

1. Introduction

There has recently been a surge of interest in analyzing high-throughput genomic data such as whole genome-wide SNP data and microarray-based gene expression data. The data set often consists of a phenotypic response, denoted by $Y \in \mathbb{R}$, which can be a binary disease indicator (e.g., tumor versus normal tissue), or a continuous measurement of a patient's response to a drug, or time to cancer recurrence (or death) that is subject to censoring. Meanwhile, high-dimensional genomic profiles of individual subjects, denoted by $X \in \mathbb{R}^p$, with p indicating the dimension, are recorded; for instance, expression levels of thousands of genes are measured simultaneously with microarray technology. There have been extensive studies to model the phenotypic response Y given the genomic predictors X; see, for instance, Golub et al. (1999), Dudoit, Fridlyand, and Speed (2002), and Bura and Pfeiffer (2003).

When p is large, classical modeling approaches often suffer the curse of dimensionality. It is thus natural to consider dimension reduction prior to model formulation. Sufficient dimension reduction (SDR) theory (Cook, 1998) has been developed to reduce the predictor dimension, meanwhile preserving full regression information and imposing few probabilistic assumptions. More specifically, SDR seeks to replace the p-dimensional predictor X with a d-dimensional vector $\eta^{\mathsf{T}}X$, where η is a $p \times d$ matrix with $d \leq p$, such that the conditional distribution function $F(Y \mid X) = F(Y \mid \eta^{\mathsf{T}}X)$. The subspace spanned by the columns of η, $\mathrm{Span}(\eta)$, is called a dimension reduction subspace, and the intersection of such spaces is itself a dimension reduction subspace under minor conditions (Cook, 1996). Such an intersection, by definition, is a unique and parsimonious population parameter that captures all regression information of Y given X, and thus is the main object of interest in our dimension reduction inquiry. We call it the *central subspace*, denote it by $\mathcal{S}_{Y|X}$, and call its dimension, $d = \dim(\mathcal{S}_{Y|X})$, the structural dimension of regression (Cook, 1998).

There have been a number of methods proposed to estimate the central subspace. Among them, sliced inverse regression (SIR) is one of the first and perhaps the most commonly used SDR method. Under appropriate conditions, Li (1991) showed that an estimate of the basis of $\mathcal{S}_{Y|X}$ can be obtained by the first d eigenvectors, $\eta_1, \ldots, \eta_d$, of the decomposition,

$$\mathrm{Cov}(\mathrm{E}(X \mid Y))\eta_j = \theta_j \Sigma_x \eta_j, \tag{1}$$

where $\theta_1 \geq \cdots \geq \theta_d > 0$ are the corresponding positive eigenvalues, and $\Sigma_x = \mathrm{Cov}(X)$. There are both asymptotic and permutation tests available to determine d, the dimension of the central subspace (Li, 1991; Cook and Yin, 2001).

SIR estimation in (1) requires the inversion of the predictor covariance Σ_x. In many applications such as microarray studies, the number of predictors (i.e., genes) often exceeds the number of sample observations. In those cases, the usual sample estimate of Σ_x is singular and is noninvertible. In addition, the predictors may be highly correlated, which is often expected in gene expression data. While the collinearity does not introduce any theoretical difficulty to SIR, the sample estimate may become highly variable. To partly circumvent these problems, singular value decomposition (SVD) has been employed prior to SIR (Chiaromonte and Martinelli, 2002; Li and Li, 2004). However, the SVD-based methods mainly focused on building a predictive model, and it is difficult to perform individual predictor selection. Alternatively, Zhong

et al. (2005) proposed a modified version of SIR by replacing the covariance matrix Σ_x in (1) with $(\Sigma_x + \tau I_p)$, where τ is a nonnegative constant and I_p is a p-dimensional identity matrix. They recommended an approximate formula to select the tuning parameter τ. Their idea is intuitive but the solution, in particular the formula for selecting τ, is ad hoc.

In this article, motivated by the least-squares formulation of SIR (Cook, 2004), we propose a regularized SIR method combining both L_1 and L_2 regularizations. The L_2 regularization enables SIR to work with $n < p$ and highly correlated predictors. The L_1 regularization achieves simultaneous reduction estimation and predictor selection. The rest of the article is organized as follows. Development of the regularized SIR is presented in Section 2. Effectiveness of the proposed method is demonstrated by simulation studies in Section 3, and a real microarray data analysis in Section 4. We conclude the article with a discussion in Section 5.

2. Regularized Sliced Inverse Regression

2.1 *Least-Squares Formulation of SIR*

To describe SIR, consider the standardized predictor $Z = \Sigma_x^{-1/2}(X - \mathrm{E}(X))$. There is no loss of generality of working in the Z-scale, because $\mathcal{S}_{Y|X} = \Sigma_x^{-1/2}\mathcal{S}_{Y|Z}$, where $\mathcal{S}_{Y|Z}$ denotes the central subspace of regression of Y on Z. Suppose we have n independent and identically distributed realizations of $(X,\ Y)$. The sample version of Z is $\hat{Z} = \hat{\Sigma}_x^{-1/2}(X - \bar{X})$, where $\bar{X}$ is the grand average of X and $\hat{\Sigma}_x$ is the usual sample covariance matrix. Furthermore, suppose the range of the response Y is partitioned into h nonoverlapping slices, with n_y observations in the yth slice, $y = 1, \ldots, h$, and let $\bar{Z}_y$ denote the average of $\hat{Z}$ in the yth slice, and $\hat{f}_y = n_y/n$. Cook (2004) showed that the SIR estimate specified by (1) can be obtained by minimizing

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

over $B \in \mathbb{R}^{p\times d}$ and $C = (C_1, \ldots, C_h) \in \mathbb{R}^{d\times h}$, where the norm is defined with respect to a standard inner product. The solution $\hat{B}$ forms an estimation of the basis of $\mathcal{S}_{Y|Z}$.

SIR does not impose any traditional model assumption on the conditional distribution of $Y|\,X$, but instead requires a condition on the marginal distribution of X. It is called the linearity condition, which states that, for any $b \in \mathbb{R}^p, \mathrm{E}(b^{\mathsf{T}}X \,|\, \eta^{\mathsf{T}}X) = c_0 + c_1\eta_1^{\mathsf{T}}X + \cdots + c_d\eta_d^{\mathsf{T}}X$, for some constants $c_0, \ldots, c_d$, where $\eta = (\eta_1, \ldots, \eta_d)$ forms a basis of $\mathcal{S}_{Y|X}$. Hall and Li (1993) argued that this is not a restrictive assumption, because it holds to a reasonable approximation as p increases. In addition, when X is elliptically symmetrically distributed, and particularly, when X follows a multivariate normal distribution, the linearity condition holds (Eaton, 1986). The condition can also be induced by predictor transformation, reweighting (Cook and Nachtsheim, 1994), and clustering (Li, Cook, and Nachtsheim, 2004).

2.2 *SIR with L_2 Regularization*

Since the standardized predictor Z involves the inverse of Σ_x, the SIR formulation (2) is not directly applicable when the sample covariance matrix $\hat{\Sigma}_x$ is singular. Its solution may also become unstable when the predictors are highly correlated. Ridge regression with L_2 regularization has been proposed to address similar issues in the ordinary least-squares setup. We adopt the same strategy to SIR.

To avoid the issue of singularity of $\hat{\Sigma}_x$ in the Z scale, we first derive another least-squares formulation of SIR, which is equivalent to (2) but is in the original predictor X scale. Note that in the X scale, $G(B,\ C)$ in (2) becomes

$$G(A, C) = \sum_{y=1}^{h} \hat{f}_y\{(\bar{X}_y - \bar{X}) - \hat{\Sigma}_x AC_y\}^{\mathsf{T}} \times \hat{\Sigma}_x^{-1}\{(\bar{X}_y - \bar{X}) - \hat{\Sigma}_x AC_y\}, \tag{3}$$

where $\bar{X}_y$ denotes the average of X in the yth slice, and $A = \hat{\Sigma}_x^{-1/2}B$. Letting $\hat{A}$ be the value of A that minimizes $G(A,\ C)$, then $\mathrm{Span}(\hat{A})$ estimates the central subspace $\mathcal{S}_{Y|X}$. We derive the following equivalent form of $G(A,\ C)$.

Lemma 1. *Let*

$$\tilde{G}(A, C) = \sum_{y=1}^{h} \hat{f}_y \|(\bar{X}_y - \bar{X}) - \hat{\Sigma}_x AC_y\|^2. \tag{4}$$

For a given C, the minimizer $\hat{A}$ of $\tilde{G}(A, C)$ in (4) *also minimizes $G(A,\ C)$ in* (3).

The proof is given in the Appendix. We note that, as shown in the proof, the equivalence between (3) and (4) requires the existence of $\hat{\Sigma}_x^{-1}$. However, (4) avoids the $\hat{\Sigma}_x^{-1}$ term in (3), thus it can be easily extended to incorporate the regularization paradigm similarly as in the ordinary least-squares regression.

Based on Lemma 1, we next propose the following *ridge SIR estimator.*

Definition 1. *For a nonnegative constant τ, let*

$$G_\tau(A, C) = \sum_{y=1}^{h} \hat{f}_y \|(\bar{X}_y - \bar{X}) - \hat{\Sigma}_x AC_y\|^2 + \tau \mathrm{vec}(A)^{\mathsf{T}}\mathrm{vec}(A), \tag{5}$$

where $\mathrm{vec}(\cdot)$ *is a matrix operator that stacks all columns of the matrix to a single vector. Let $(\hat{A}, \hat{C}) = \arg\min_{A,C} G_\tau(A, C)$. Then $\mathrm{Span}(\hat{A})$ is called a ridge SIR estimator of the central subspace $\mathcal{S}_{Y|X}$.*

When $\hat{\Sigma}_x^{-1}$ exists and $\tau = 0$, $G_\tau(A,\ C)$ reduces to $\tilde{G}(A, C)$, which is in turn equivalent to $G(A,\ C)$, thus the ridge SIR estimator reduces to a usual SIR estimator. When $\hat{\Sigma}_x$ is not invertible, a positive τ is incorporated to deal with the singularity of $\hat{\Sigma}_x$ in the estimation procedure.

To minimize (5) for a fixed τ, we propose an alternating least-squares algorithm as stated below. Straightforward calculation shows that, given A, solution of C can be obtained by h usual least squares,

$$\hat{C} = (\hat{C}_1, \ldots, \hat{C}_h), \quad \text{with}$$
$$\hat{C}_y = \left(A^{\mathsf{T}}\hat{\Sigma}_x^2 A\right)^{-1} A^{\mathsf{T}}\hat{\Sigma}_x(\bar{X}_y - \bar{X}), \quad y = 1, \ldots, h.$$

Next, rewrite $G_\tau(A, C)$ in the form of least-squares regression,

$$\begin{aligned} G_\tau(A,C) &= \sum_{y=1}^{h} \hat{f}_y \left\| (\bar{X}_y - \bar{X}) - \left(C^\mathsf{T} \otimes \hat{\Sigma}_x\right) \mathrm{vec}(A) \right\|^2 \\ &\quad + \tau \mathrm{vec}(A)^\mathsf{T} \mathrm{vec}(A) \\ &= \left\| \tilde{W}^{1/2} \tilde{Y} - \tilde{W}^{1/2} \left(C^\mathsf{T} \otimes \hat{\Sigma}_x\right) \mathrm{vec}(A) \right\|^2 \\ &\quad + \tau \mathrm{vec}(A)^\mathsf{T} \mathrm{vec}(A), \end{aligned} \tag{6}$$

where $\otimes$ is the Kronecker product, $\tilde{Y} = \mathrm{vec}(\bar{X}_1 - \bar{X}, \ldots, \bar{X}_h - \bar{X})$, $\tilde{W}^{1/2} = D_f^{1/2} \otimes I_p$, and $D_f = \mathrm{diag}(\hat{f}_1, \ldots, \hat{f}_h)$. Consequently, given C, the solution of A in (6) is,

$$\mathrm{vec}(\hat{A}) = \left(CD_f C^\mathsf{T} \otimes \hat{\Sigma}_x^2 + \tau I_{pd}\right)^{-1} \left(CD_f \otimes \hat{\Sigma}_x\right) \tilde{Y}. \tag{7}$$

We cycle between minimizing A and C until convergence. The algorithm produces a monotonically decreasing series of evaluations of $G_\tau(A,\ C)$, and because $G_\tau(A,\ C) \geq 0$, it is guaranteed to converge. In our limited simulations, we have tried various starting values and the algorithm converges to the same solution, which suggests that the initial values chosen for A do not affect the ultimate result.

To select the ridge parameter τ in (5), we derive a generalized crossvalidation criterion (GCV), following Golub, Heath, and Wahba (1979):

$$\mathrm{GCV} = \frac{\left\| (I_{ph} - S_\tau) \tilde{W}^{1/2} \tilde{Y} \right\|^2}{ph\{1 - \mathrm{trace}(S_\tau)/ph\}^2}, \tag{8}$$

where

$$S_\tau = \left(D_f^{1/2} \hat{C}^\mathsf{T} \otimes \hat{\Sigma}_x\right) \left(\hat{C} D_f \hat{C}^\mathsf{T} \otimes \hat{\Sigma}_x^2 + \tau I_{pd}\right)^{-1} \left(\hat{C} D_f^{1/2} \otimes \hat{\Sigma}_x\right).$$

A detailed derivation of (8) is given in the Appendix. Note that (8) follows a typical GCV definition, where the numerator is the first term in (6) by plugging in $\mathrm{vec}(\hat{A})$ in (7), whereas the term S_τ is symmetric. The value of τ is selected to minimize (8). The structural dimension, $d = \dim(\mathcal{S}_{Y|X})$, is regarded as known in the proposed algorithm. Estimation of d will be discussed in Section 2.4.

2.3 *SIR with Both L_1 and L_2 Regularizations*

An equally important goal of the analysis, in addition to reduction estimation, is to select active predictors. Estimates of ridge SIR are linear combinations of *all* the predictors, and no variable selection is achieved. Following the least absolute shrinkage and selection operator (Lasso) idea (Tibshirani, 1996), we further introduce L_1 regularization to the ridge SIR estimator to induce sparsity in the estimated linear combinations. A similar strategy has been proposed by Ni, Cook, and Tsai (2005), who coupled L_1 regularization with the usual SIR.

Let $(\hat{A}, \hat{C})$ denote the ridge SIR estimator, that is, $(\hat{A}, \hat{C}) = \arg\min_{A,C} G_\tau(A, C)$. We next propose the *sparse ridge SIR estimator*.

Definition 2. *A sparse ridge SIR estimator of the central subspace $\mathcal{S}_{Y|X}$ is defined as* $\mathrm{Span}(\mathrm{diag}(\hat{\alpha})\hat{A})$, *where the shrinkage index vector $\hat{\alpha} = (\hat{\alpha}_1, \ldots, \hat{\alpha}_p)^\mathsf{T} \in \mathbb{R}^p$ is obtained by minimizing*

$$G_\lambda(\alpha) = \sum_{y=1}^{h} \hat{f}_y \left\| (\bar{X}_y - \bar{X}) - \hat{\Sigma}_x \mathrm{diag}(\alpha) \hat{A} \hat{C}_y \right\|^2 \tag{9}$$

over α, subject to $\sum_{j=1}^p |\alpha_j| \leq \lambda$, for some nonnegative constant λ.

The constrained optimization of (9) can be done by employing a standard Lasso algorithm. To see this, we first note that $\mathrm{diag}(\alpha)\hat{A}\hat{C}_y = \mathrm{diag}(\hat{A}\hat{C}_y)\alpha$, and thus,

$$G_\lambda(\alpha) = \sum_{y=1}^{h} \hat{f}_y \| (\bar{X}_y - \bar{X}) - \hat{\Sigma}_x \mathrm{diag}(\hat{A}\hat{C}_y)\alpha \|^2.$$

Next we write:

$$\begin{aligned} \tilde{Y} &= \mathrm{vec}(\bar{X}_1 - \bar{X}, \ldots, \bar{X}_h - \bar{X}) \in \mathbb{R}^{ph}, \\ \tilde{X} &= (\mathrm{diag}(\hat{A}\hat{C}_1)\hat{\Sigma}_x, \ldots, \mathrm{diag}(\hat{A}\hat{C}_h)\hat{\Sigma}_x)^\mathsf{T} \in \mathbb{R}^{ph \times p}. \end{aligned}$$

Then the shrinkage vector α is exactly the Lasso estimator for the regression of $\tilde{Y}$ with ph "observations" on the p-dimensional "data matrix" $\tilde{X}$. Several Lasso algorithms, such as Tibshirani (1996), Fu (1998), and Osborne, Presnell, and Turlach (2000), can be employed to estimate α.

When the Lasso parameter $\lambda \geq p, \hat{\alpha}_j = 1$ for $j = 1, \ldots, p$, and we get a ridge SIR estimator. As λ gradually decreases, some indices α_j are shrunk to zero, indicating the corresponding predictors are not needed for the regression given other predictors. To select λ, we adopt the family of information criteria suggested in Ni et al. (2005), including Akaike information criterion (AIC; Akaike, 1973), Bayesian information criterion (BIC; Schwarz, 1978),and residual information criterion (RIC; Shi and Tsai, 2002):

$$\begin{aligned} \mathrm{AIC} &= ph \log\left(\frac{G_\lambda(\hat{\alpha})}{ph}\right) + 2p_\lambda, \\ \mathrm{BIC} &= ph \log\left(\frac{G_\lambda(\hat{\alpha})}{ph}\right) + \log(ph) p_\lambda, \\ \mathrm{RIC} &= (ph - p_\lambda) \log\left(\frac{G_\lambda(\hat{\alpha})}{ph - p_\lambda}\right) \\ &\qquad + p_\lambda(\log(ph) - 1) + \frac{4}{ph - p_\lambda - 2}, \end{aligned}$$

where p_λ denotes the effective number of parameters in the Lasso estimator. Following Zou, Hastie, and Tibshirani's (2004) discussion on the degrees of freedom of Lasso, we approximate p_λ by the number of nonzero components in the estimated $\hat{\alpha}$.

2.4 *Estimation of Structural Dimension*

In the estimation procedure of the proposed regularized SIR, we regard $d = \dim(\mathcal{S}_{Y|X})$ as known. In practice, however, d needs to be estimated given the data. There exists a number of asymptotic tests to determine d (Li, 1991; Schott, 1994; Bura and Cook, 2001), but none is directly applicable for $n < p$. The permutation test of Cook and Yin (2001) may be employed, but it requires an additional independence assumption, and is computationally intensive. Alternatively, we adopt a criterion proposed by Zhu, Miao, and Peng (2006),

which estimates d via the number of nonzero eigenvalues of the matrix $\mathrm{Cov}(\mathrm{E}(X|\, Y))$, or equivalently, the number of eigenvalues of the matrix $\Omega = \mathrm{Cov}(\mathrm{E}(X\,|\,Y)) + I_p$ that are greater than one.

Letting $\hat{\delta}_1, \ldots, \hat{\delta}_p$ denote the eigenvalues of the sample estimate $\hat{\Omega}$ of Ω, κ denote the number of $\hat{\delta}_i$'s that are greater than one, and C_n denote a penalty constant, Zhu et al. (2006, equations (10) to (12)) suggested the following estimator of d,

$$\hat{d} = \arg \max_{m \in \{0,1,\ldots,p-1\}} \left\{ \begin{array}{l} \dfrac{n}{2} \displaystyle\sum_{i=1+\min(\kappa,m)}^{p} (\log(\hat{\delta}_i) + 1 - \hat{\delta}_i) \\ \quad - \dfrac{C_n m(2p-m+1)}{2} \end{array} \right\}. \quad (10)$$

They have recommended several forms for the penalty constant C_n. In our case, we simply take $C_n = \log(n)h/n$, and have found it works well in our simulations.

3. Simulations

3.1 *Sample Size Less Than Number of Predictors*

Simulation studies were carried out to demonstrate the effectiveness of the proposed method. We first consider the case when $n < p$. The data are generated from the following model, with $n = 100$ and $p = 200$.

$$Y_1 = x_1 + x_1 \times x_2 + \sigma_0 \varepsilon.$$

Components of $X = (x_1, x_2, \ldots, x_p)^{\mathsf{T}}$ and the error term ε follow independent standard normal distributions. The parameter σ_0 controls the relative strength of signal to noise, and it is chosen such that $\mathrm{Var}(\mathrm{E}(Y_1\,|\,X))/\sigma_0^2 = 20$. The central subspace is spanned by $(\beta_1,\, \beta_2)$, where $\beta_1 = (1, 0, 0, \ldots, 0)^{\mathsf{T}}, \beta_2 = (0, 1, 0, \ldots, 0)^{\mathsf{T}}$, and the true structural dimension $d = 2$. The active predictors are x_1 and x_2. The response model includes both a main effect (x_1) and an interaction term ($x_1 \times x_2$).

For this model we focus on the performance of the estimation method in selecting active predictors. We employ two measures commonly used in biomedical literature, that is, the true positive rate (TPR), which is defined as the ratio of the number of correctly identified active predictors to the number of truly active predictors, and the false positive rate (FPR), which is defined as the ratio of the number of falsely identified active predictors to the total number of inactive predictors. The measures TPR and FPR are also known as sensitivity and 1-specificity, and ideally, we wish to have TPR to be close to 1 and FPR to be close to 0 at the same time.

Table 1 reports the average of TPR and FPR, evaluated for the estimated $\hat{\beta}_1$ and $\hat{\beta}_2$, respectively, based on 100 data replications. As a comparison, we also reported the performance of Lasso (Tibshirani, 1996) and Elastic net (Zou and Hastie, 2005). Both approaches are based on the ordinary least-squares (OLS) regression, but Lasso introduced L_1 regularization, whereas Elastic net introduced both L_1 and L_2 regularizations.

All procedures successfully identified the main effect x_1 in the model; TPR of $\hat{\beta}_1$ is 1 for all cases. However, both Lasso and Elastic net failed to select x_2 in the interaction term; TPR of $\hat{\beta}_2$ is only 0.140 for Lasso, and 0.260 for Elastic net. This is because both methods are based on the OLS, and in this example, the population OLS estimator equals $\Sigma_x^{-1}\mathrm{Cov}(X, Y_1) = \Sigma_x^{-1}\mathrm{E}(XY_1) = \Sigma_x^{-1}(1, 0, 0, \ldots, 0)^{\mathsf{T}}$. Consequently, neither OLS nor OLS-based Lasso or Elastic net could identify x_2. By contrast, sparse ridge SIR selected x_2 with a high successful rate, with TPR of $\hat{\beta}_2$ ranging from 75% to 89%, although it paid the price of selecting more false positives. In particular, AIC-based sparse ridge SIR has the highest FPR, while RIC-based estimator has the lowest FPR that is comparable to Lasso. In applications like gene microarray analysis, it is often deemed more important to identify all the true positives, whereas the false positives may be further screened by refined biological experiments. We thus believe the sparse ridge SIR offers a useful solution.

Table 1
Average of the true positive rate and false positive rate based on 100 *data replications, for the sparse ridge sliced inverse regression estimator (SR-SIR), the Lasso, and the Elastic net, when* $n < p$

Method	TPR $\hat{\beta}_1$	FPR $\hat{\beta}_1$	TPR $\hat{\beta}_2$	FPR $\hat{\beta}_2$
SR-SIR (AIC)	1.000	0.460	0.890	0.460
SR-SIR (BIC)	1.000	0.181	0.850	0.182
SR-SIR (RIC)	1.000	0.053	0.750	0.054
Lasso	1.000	0.051	0.140	0.055
Elastic net	1.000	0.173	0.260	0.176

We also investigated estimation of the structural dimension d for this example. Table 2, the second row, reports the percentages of estimated d values out of 100 data replications. It is seen that the estimation method given in (10) works reasonably well.

3.2 *Correlated Predictors*

We next examined the performance of the sparse ridge SIR estimator for correlated predictors, by considering the following response model,

$$\beta_1 = (1, 1, 1, 0, \ldots, 0)^{\mathsf{T}}/\sqrt{3}, \quad \text{and } U = \beta_1^{\mathsf{T}} X,$$
$$Y_2 = 2U + U^2 + \sigma_0 \varepsilon.$$

The predictor vector $X = (x_1, \ldots, x_p)$ follows a multivariate normal distribution with mean 0, and the correlation between x_i and x_j is $\rho^{|i-j|}$, with ρ taking the values $\{0, 0.1, 0.2, \ldots, 0.9\}$. The error ε is standard normal and is independent of X. The parameter σ_0 is chosen in the same way as in the previous example. We chose $n = 200$ and $p = 20$ in this example to focus on the effect of the predictor correlation on the estimation methods.

Table 2
The percentages of estimated structural dimension d based on 100 *data replications*

Response model		$\hat{d} = 0$	$\hat{d} = 1$	$\hat{d} = 2$	$\hat{d} = 3$	$\hat{d} = 4$
$Y_1 (d = 2)$		0.00	0.06	**0.54**	0.33	0.07
$Y_2 (d = 1)$	$\rho = 0.3$	0.00	**0.99**	0.01	0.00	0.00
	$\rho = 0.6$	0.00	**0.95**	0.05	0.00	0.00
	$\rho = 0.9$	0.00	**0.91**	0.09	0.00	0.00

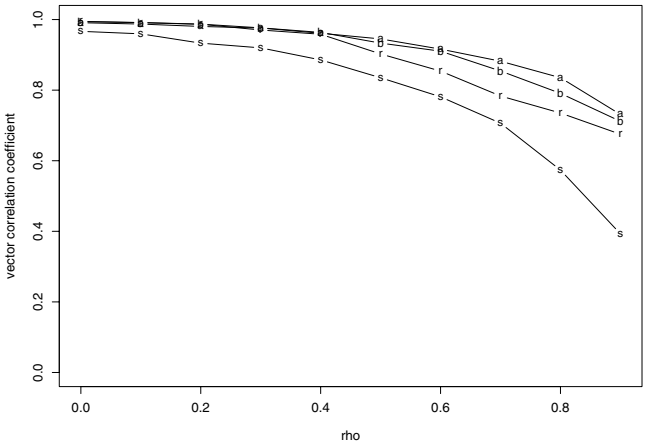

Figure 1. Comparison of the sparse ridge SIR estimator and the usual SIR estimator, measured by the vector correlation coefficient between the true and the estimated central subspace, as the predictor correlation ρ varies. s denotes SIR, a denotes sparse ridge SIR based on AIC, b denotes sparse ridge SIR based on BIC, and r denotes sparse ridge SIR based on RIC.

For comparison purposes, we included the usual SIR estimator. Accuracy of the estimated basis for the central subspace, $\mathcal{S}_{Y|X} = \text{Span}(\beta_1)$, under different correlation values of ρ, is of the primary interest here. The vector correlation coefficient ψ_q given in Ye and Weiss [2003] is employed as an evaluation criterion, which is defined as $\psi_q = (\prod_{i=1}^{d} \phi_i^2)^{1/2}$, where ϕ^2s are the eigenvalues of the matrix $B_2^{\mathsf{T}} B_1 B_1^{\mathsf{T}} B_2$, with B_1, B_2 denoting the orthonormal bases of the true and the estimated central subspace, respectively. (Another criterion, the trace correlation coefficient, in Ye and Weiss [2003] equals ψ_q when $d = 1$, and is thus omitted here.) The criterion ψ_q describes the "closeness" of the two subspaces, and ranges between 0 and 1, with a larger value indicating a better estimate. When $\psi_q = 1$, $\text{Span}(B_1) = \text{Span}(B_2)$.

Figure 1 shows the average of ψ_q for the response model Y_2, as the correlation coefficient ρ varies. Both SIR and the regularized SIR performed similarly when the predictor correlation is small, whereas the regularized SIR estimator outperformed SIR when the correlation ρ is large. The advantage of the regularized estimator over the usual SIR estimator is clearly seen in the presence of the highly correlated predictors.

We also examined the dimension selection method (10) in estimating d of the central subspace. In this example, $d = 1$. Table 2, rows three to five, report the percentages of estimated d values for three different ρ values. Estimations are seen to be quite accurate.

4. Diffuse Large B-Cell Lymphoma Data

Diffuse large-B-cell lymphoma (DLBCL) is the most common type of lymphoma in adults, and has an annual incidence of more than 25,000 cases in the United States (Jaffe, 1998). The survival rate of the standard chemotherapy for DLBCL is only about 35% to 40%. Thus it is important to predict the outcome of the chemotherapy and to understand the factors that influence the survival outcome. Rosenwald et al. (2002) reported the survival time of 240 DLBCL patients after the chemotherapy, along with measurements of 7399 genes obtained from cDNA microarrays for each individual patient. The survival times of these data ranged from about 0 to 21.8 years, and there were 138 deceased patients during the follow-ups. The goal of the analysis is to predict the survival time based on gene expression information, and to identify potential genes that may be related to the patients' survival.

Following Rosenwald et al. (2002), we divided the patients into a training group of 160 samples and a testing group of 80 samples. To expedite the computation of the regularized SIR method, we employed a gene preselection procedure that has been commonly used in microarray studies, to preselect 329 genes using a univariate Cox test. To further accommodate the censored response, we adopted the double slicing method for SIR (Li, Wang, and Chen, 1999; Li and Li, 2004).

Sparse ridge SIR was applied to the training data, obtaining the estimated structural dimension $\hat{d} = 1$. The resulting AIC-based sparse ridge SIR estimator selected 34 genes; both BIC- and RIC-based estimators selected the same group of 12 genes. Out of those 12 genes selected by BIC and RIC, 11 were selected by AIC as well. Additionally, Rosenwald et al. (2002) reported four gene signature groups that are believed to contain potentially important genes related to the risk of death caused by DLBCL. Among the 34 genes selected by AIC-based sparse ridge SIR estimator, 11 genes belong to the four gene signature groups specified by Rosenwald et al. (2002); among the 12 genes selected by BIC- and RIC-based estimators, 6 genes belong to the signature groups. We believe the identified genes would make a good list of candidate genes for further biological investigation.

We also evaluated the predictive performance of the proposed method. A Cox proportional hazards model was fit with the derived sparse ridge SIR estimator as the predictor. Three risk groups of patients, the low-risk patients, the intermediate-risk patients, and the high-risk patients, were defined according to the 33% and 66% quantiles of the estimated risk scores. Figure 2 shows the Kaplan–Meier estimates of survival curves for the three groups. It is noted that the sparse ridge SIR estimator based on all three information criteria achieved good separation of the three risk groups, which indicates a good model fit to the training data. The log-rank test of difference among three survival curves yielded the p-value of 0 for all cases, which confirms our visual examination. The fitted Cox model was then applied to the testing data, and the same cutoff values used in the training data set were used to assign the test samples to the three risk groups. Figure 3 shows the corresponding Kaplan–Meier estimates of survival curves. The sparse ridge SIR estimator based on AIC again achieved a good stratification of all three risk groups, with a p-value of 0.003 for the log-rank test, whereas the estimator based on BIC and RIC separated the low-risk group with the intermediate and high-risk groups, resulting in a p-value of 0.017. Overall, the sparse ridge SIR estimator in conjunction with a Cox proportional hazards model demonstrate competent model fitting and predictive performance.

5. Discussion

In this article we have proposed an extension of SIR based on its least-squares formulation. By introducing the L_2

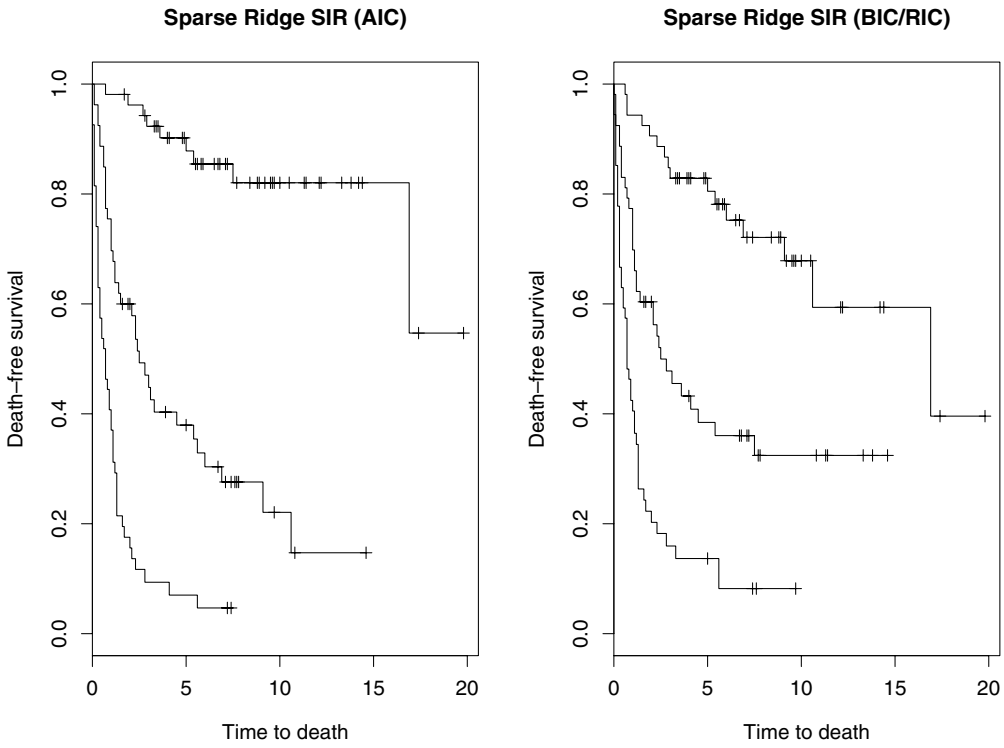

Figure 2. Kaplan–Meier estimate of survival curves for the three risk groups of patients in the *training data*. The left-hand side panel is the AIC-based sparse ridge SIR estimator, and the right-hand side panel is the BIC- and RIC-based sparse ridge SIR estimator.

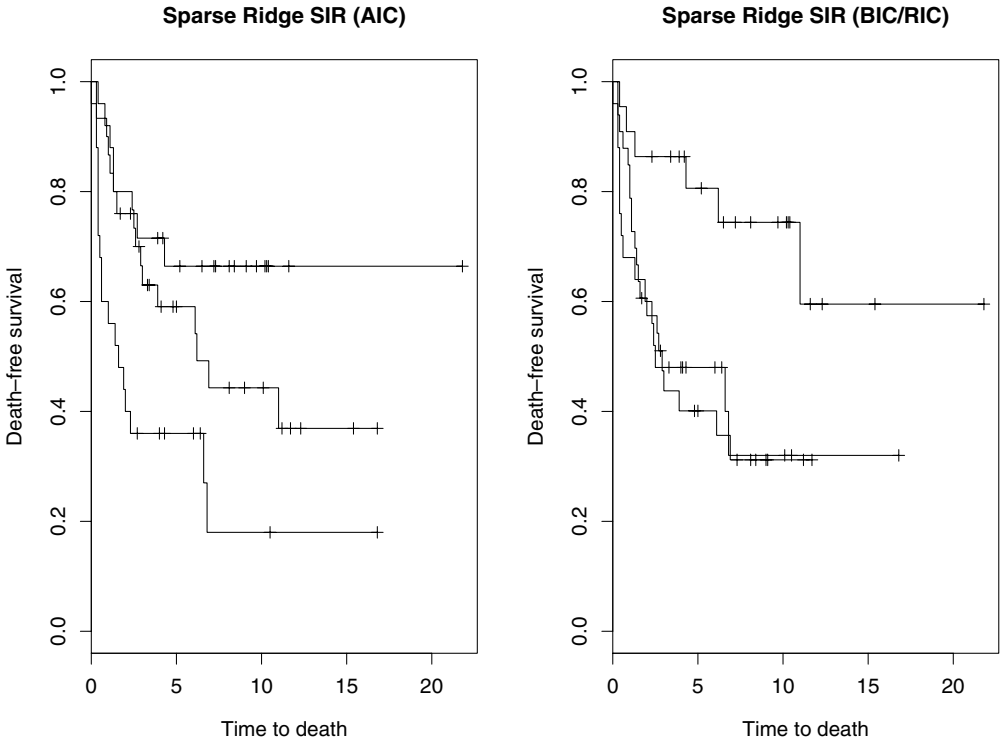

Figure 3. Kaplan–Meier estimate of survival curves for the three risk groups of patients in the *testing data*. The left-hand side panel is the AIC-based sparse ridge SIR estimator, and the right-hand side panel is the BIC- and RIC-based sparse ridge SIR estimator.

regularization, the SIR method is enabled to work for both $n < p$, and for highly correlated predictors. The L_1 regularization further achieves simultaneous reduction estimation and variable selection. The proposed method has been shown to work effectively through both simulations and the real data application. Our experience also suggests that the proposed method works reasonably fast. For instance, for the simulation example in Section 3.1 where $n = 100$ and $p = 200$, the computing time was less than 5 minutes on a standard personal computer, whereas the computing time for the DLBCL data in Section 4 was about one and a half hours.

We believe this method of coupling L_1 and L_2 regularizations with SIR would provide a useful dimension reduction tool. It would also strengthen the applicability of SDR in general, by considering possible extensions to other SDR estimation methods, such as sliced average variance estimation (Cook and Weisberg, 1991) and principal Hessian directions (Li, 1992).

Acknowledgements

The authors are grateful to the editor, the associate editor, and the referee for their constructive comments that have greatly improved the article.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *Second International Symposium on Information Theory*, B. N. Petrov and F. Csaki (eds), 267–281. Budapest: Akademiai Kiado.

Bura, E. and Cook, R. D. (2001). Extending sliced inverse regression: The weighted chi-squared test. *Journal of the American Statistical Association* **96,** 996–1003.

Bura, E. and Pfeiffer, R. M. (2003). Graphical methods for class prediction using dimension reduction techniques on DNA microarray data. *Bioinformatics* **19,** 1252–1258.

Chiaromonte, F. and Martinelli, J. (2002). Dimension reduction strategies for analyzing global gene expression data with a response. *Mathematical Biosciences* **176,** 123–144.

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 Nachtsheim, C. J. (1994). Re-weighting to achieve elliptically contoured covariates in regression. *Journal of the American Statistical Association* **89,** 592–600.

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

Cook, R. D. and Yin, X. (2001). Dimension reduction and visualization in discriminant analysis. *Australian and New Zealand Journal of Statistics* **43,** 147–177.

Dudoit, S., Fridlyand, J., and Speed, T. P. (2002). Comparison of discrimination methods for the classification of tumors using gene expression data. *Journal of American Statistical Association* **97,** 77–87.

Eaton, M. (1986). A characterization of spherical distributions. *Journal of Multivariate Analysis* **20,** 272–276.

Fu, W. J. (1998). Penalized regression: The bridge versus the lasso. *Journal of Computational and Graphical Statistics* **7,** 397–416.

Golub, G. H., Heath, M., and Wahba, G. (1979). Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics* **21,** 215–223.

Golub, T. R. Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., Bloomfield, C. D., and Lander, E. S. (1999). Molecular classification of cancer discovery and class prediction by gene expression monitoring. *Science* **286,** 531–537.

Hall, P. and Li, K. C. (1993). On almost linearity of low dimensional projections from high dimensional data. *Annals of Statistics* **21,** 867–889.

Jaffe, E. S. (1998). Histopathology of the non-Hodgkin's lymphomas and Hodgkin's disease. In *The Lymphomas*, G. P. Canellos, T. A. Lister, and J. L. Sklar (eds), 77–106. Philadelphia: W.B. Saunders.

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

Li, K. C. (1992). On principal Hessian directions for data visualization and dimension reduction: Another application of Stein's Lemma. *Annals of Statistics* **87,** 1025–1039.

Li, K. C., Wang, J. L., and Chen, C. H. (1999). Dimension reduction for censored regression data. *Annals of Statistics* **27,** 1–23.

Li, L. and Li, H. (2004). Dimension reduction methods for microarrays with application to censored survival data. *Bioinformatics* **20,** 3406–3412.

Li, L., Cook, R. D., and Nachtsheim, C. J. (2004). Cluster-based estimation for sufficient dimension reduction. *Computational Statistics and Data Analysis* **47,** 175–193.

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

Osborne, M. R., Presnell, B., and Turlach, B. A. (2000). On the lasso and its dual. *Journal of Computational and Graphical Statistics* **9,** 319–337.

Rosenwald, A., Wright, G., Chan, W. C., Connors, J. M., Campo, E., Fisher, R. I., Gascoyne, R. D., Muller-Hermelink, H. K., Smeland, E. B., and Staudt, L. M. (2002). The use of molecular profiling to predict survival after chemotherapy for diffuse large-B-cell lymphoma. *The New England Journal of Medicine* **346,** 1937–1947.

Schott, J. R. (1994). Determining the dimensionality in sliced inverse regression. *Journal of the American Statistical Association* **89,** 141–148.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics* **6,** 461–464.

Shi, P. and Tsai, C.-L. (2002). Regression model selection—A residual likelihood approach. *Journal of the Royal Statistical Society, Series B* **64,** 237–252.

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society, Series B* **58,** 267–288.

Ye, Z. and Weiss, R. E. (2003). Using the bootstrap to select one of a new class of dimension reduction methods. *Journal of the American Statistical Association* **98,** 968–979.

Zhong, W., Zeng, P., Ma, P., Liu, J. S., and Zhu, Y. (2005). RSIR: Regularized sliced inverse regression for motif discovery. *Bioinformatics* **21,** 4169–4175.

Zhu, L., Miao, B., and Peng, H. (2006). On sliced inverse regression with large dimensional covariates. *Journal of American Statistical Association* **101,** 630–643.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B* **67,** 301–320.

Zou, H., Hastie, T., and Tibshirani, R. (2004). *On the "Degrees of Freedom" of the Lasso*. Technical Report, Department of Statistics, Stanford University.

Received April 2006. *Revised January* 2007. *Accepted March* 2007.

Appendix

Proof of Lemma 1. It is first noted that $\tilde{G}(A, C)$ in (4) can be written as

$$\tilde{G}(A,C) = \left\{\tilde{Y} - (C^{\mathsf{T}} \otimes \hat{\Sigma}_x)\mathrm{vec}(A)\right\}^{\mathsf{T}} \tilde{W} \left\{\tilde{Y} - (C^{\mathsf{T}} \otimes \hat{\Sigma}_x)\mathrm{vec}(A)\right\},$$

where $\tilde{Y} = \mathrm{vec}(\bar{X}_1 - \bar{X}, \ldots, \bar{X}_h - \bar{X}), \tilde{W} = D_f \otimes I_p$, and $D_f = \mathrm{diag}(\hat{f}_1, \ldots, \hat{f}_h)$. Given C, the solution of A for (4) can be obtained as

$$\mathrm{vec}(\hat{A}) = \left(\left(CD_fC^{\mathsf{T}}\right)^{-1} CD_f \otimes \hat{\Sigma}_x^{-1}\right)\tilde{Y}.$$

Similarly, $G(A,\ C)$ in (3) can be written as

$$G(A,C) = \left\{\tilde{Y} - (C^{\mathsf{T}} \otimes \hat{\Sigma}_x)\mathrm{vec}(A)\right\}^{\mathsf{T}} \tilde{\tilde{W}} \left\{\tilde{Y} - (C^{\mathsf{T}} \otimes \hat{\Sigma}_x)\mathrm{vec}(A)\right\},$$

where $\tilde{\tilde{W}} = D_f \otimes \hat{\Sigma}_x^{-1}$, and $\tilde{Y}$ and D_f are as defined above. Then, given C, the solution of A for (3) can be obtained as

$$\begin{aligned}\mathrm{vec}(\hat{A}) &= \left(CD_fC^{\mathsf{T}} \otimes \hat{\Sigma}_x\right)^{-1}(C \otimes \hat{\Sigma}_x)\tilde{\tilde{W}}\tilde{Y} \\ &= \left(\left(CD_fC^{\mathsf{T}}\right)^{-1} CD_f \otimes \hat{\Sigma}_x^{-1}\right)\tilde{Y}.\end{aligned}$$

The conclusion follows.

Derivation of GCV criterion (8). Note that formula (6) can be viewed as a typical ridge regression problem, by treating $\tilde{W}^{1/2}\tilde{Y}$ as the response vector (the y in Golub et al., 1979), and $\tilde{W}^{1/2}(C^{\mathsf{T}} \otimes \hat{\Sigma}_x)$ as the predictor matrix (the X in Golub et al., 1979). With τ replacing the term $n\lambda$ in Golub et al. (1979), simple algebra shows that $A(\lambda)$ in (1.5) of Golub et al. (1979) becomes S_τ in our GCV definition. With ph denoting the sample size, and plugging in the corresponding terms in (1.4) of Golub et al. (1979), we obtain the GCV formula (8).