The Korean Communications in Statistics **Vol. 14** No. 1, 2007, pp. 215–227

Variable Selection in Sliced Inverse Regression Using Generalized Eigenvalue Problem with Penalties

Chongsun Park[1)]

Abstract

Variable selection algorithm for Sliced Inverse Regression using penalty function is proposed. We noted SIR models can be expressed as generalized eigenvalue decompositions and incorporated penalty functions on them. We found from small simulation that the HARD penalty function seems to be the best in preserving original directions compared with other well-known penalty functions. Also it turned out to be effective in forcing coefficient estimates zero for irrelevant predictors in regression analysis. Results from illustrative examples of simulated and real data sets will be provided.

Keywords: Sliced inverse regression; variable selection; penalty functions; simulated annealing.

1. Motivation

Quite possibly most of statisticians agree that the regression analysis is one of the most popular and powerful tool in predictive modeling area. Nevertheless they also agree that usual linear or generalized linear model have several drawbacks in the model itself due to the lack of flexibility resulted from strict assumptions on model parameters and distribution. Sliced Inverse Regression (SIR; Li, 1991) is known as an efficient tool finding relevant linear combinations of predictors in regression analysis with minimal assumptions on the model. With assumptions that unknown number of linear combinations of predictors are needed in the regression model and a restriction on predictors SIR has been known to be successful in finding those directions effectively. In order to find underlying regression structures it is possible to do some further analysis using graphical or other well-established statistical tools with estimated directions from SIR.

1) Professor, Department of Statistics, Sungkyunkwan University, 3-53 Myungryun-Dong, Jongno-Gu, Seoul 110-745, Korea.
E-mail : cspark@skku.edu

In practice, a large number of predictors are introduced in the initial stage of regression modeling and one of the most difficult aspects of SIR in this case is the interpretation of e.d.r. (effective dimension reduction) directions which normally have all non-zero coefficient estimates as in usual regression problems. Li (2000) recommended a try-and-error procedure for variable selection in the SIR which uses R^2 as its criterion and hardly no further researches have been done regarding this issue so far.

As noted in Li's original paper, SIR can be expressed as a generalized eigenvalue problem which is equivalent to principal component analysis (PCA). Further a number of methods are available to aid interpretation in PCA by ignoring irrelevant predictors or forcing coefficient estimates to zero somehow. Hence by combining these two, new algorithms for variable selection in the SIR would be possible. A common approach in variable selection in PCA is ignoring any coefficients less than some threshold value, so that the function becomes simple and the interpretation becomes easier. Jolliffe (1972, 1973) examines some of possible methods which discard irrelevant variables using multiple correlation, PCA itself, and clustering. Cadima and Jolliffe (1995) noted that this can be misleading. More formal ways of making some of the coefficients zero are to restrict the coefficients to a smaller number of possible values in the derivation of the linear functions like -1, 0, 1 (Hausman, 1982). And a variation on this theme (Vines, 2000) is also possible. Rotation method used in factor analysis is also applicable but has its drawbacks (Jolliffe, 1989, 1995). McCabe (1984) introduced a new strategy to select a subset of the variables themselves and called it 'principal variables'.

Other possible way would be introducing penalty function as in regression analysis. Fan and Li (2001) proposed a variable selection method using a penalized likelihood functions and argued that by using a unified approach via penalized least squares it is possible to retaining good features of both subset selection and ridge regression. Further they showed with proper choice of regularization parameters that the proposed estimators perform as well as the oracle procedure in variable selection. Recently, Jolliffe *et al.* (2003) applied L_1 penalty function method to maximization problem of PCA in order to force any irrelevant coefficients in the principal components to zero. He included L_1 penalty function as an extra constraint to a optimization problem which maximizes variance of linear combination of variables and showed that it is more preferable to rotation methods and several others.

In this paper, we propose a new technique of choosing linear combinations

of predictors in the SIR which successively maximizes variance, as in PCA, but we impose extra constraints which sacrifices some variance in order to improve interpretability. By including penalty functions in probabilistic PCA models for SIR, it is possible to force estimates related with unnecessary predictors to zero safely so making interpretations easier. We noticed from small simulation that hard thresholding (HARD) penalty function seems to be preferable to other well-known penalty functions like L_1 and Smoothly Clipped Absolute Deviation (SCAD; Fan and Li, 2001) functions. In order to find successive linear combination of predictors which should be included in regression models well known simulated annealing algorithm for function optimization could be adopted. This idea may be readily applicable to variable selection in principal component analysis itself even when there are some missing values in predictors at random.

In Section 2, basics of SIR will be introduced and penalized SIR, main idea of including penalty functions to generalized eigenvalue problem from the SIR, is in Section 3. An algorithm using simulated annealing (Aarts and Korst, 1989) to find regression coefficient estimates together with parameter settings will be included in Section 4. Results with illustrative examples from simulated and real data sets are in Section 5. Finally, concluding remarks follow in Section 6.

2. Sliced Inverse Regression (SIR)

Suppose we have a univariate response variable y and p predictors $\mathbf{x} = (x_1, x_2, \ldots, x_p)$, and observed independent n cases for these variables. Then the model assumed in the SIR becomes

$$y = f(\boldsymbol{\beta}^1\mathbf{x}, \boldsymbol{\beta}^2\mathbf{x}, \ldots, \boldsymbol{\beta}^K\mathbf{x}, \epsilon) \quad \text{with} \quad (K \leq p), \tag{2.1}$$

where ϵ is independent of $\mathbf{x}$, and f is an arbitrary unknown function on R^{p+1}. And $\boldsymbol{\beta}^1, \ldots, \boldsymbol{\beta}^K$ are K $p-$dimensional unknown parameter vectors. The model assumed in the SIR is more flexible than usual regression models at least in two aspects. First, it does not need to assume any functional form for the regression curve, and secondly it is possible to include more than two linear combinations of predictors in the model.

Definition 2.1 *Under the model (2.1), the space $\boldsymbol{\beta}$ generated by $\boldsymbol{\beta}^1, \ldots, \boldsymbol{\beta}^K$ is called the e.d.r. space. Any non-zero vector in the e.d.r. space is called an e.d.r. direction.*

Li (1991) considered inverse regression $E(\mathbf{x}|y)$ instead of forward one and showed centered inverse regression curve, $E(\mathbf{x}|y)-E(\mathbf{x})$, which is a $p-$dimensional curve in $\mathbf{R}^p$, lies on a $K-$dimensional subspace of e.d.r. directions. However, predictors should satisfy Condition 2.1 to get e.d.r. directions via centered inverse regression curve. This is called linear condition by Li and several well-known distributions like normal and elliptically contoured distributions are known to satisfy this condition.

Condition 2.1 *For any* $\mathbf{b}$ *in* R^p, *the conditional expectation* $E(\mathbf{bx}|\boldsymbol{\beta}^1\mathbf{x}, \boldsymbol{\beta}^2\mathbf{x}, \ldots, \boldsymbol{\beta}^K\mathbf{x})$ *is linear in* $\boldsymbol{\beta}^1\mathbf{x}, \ldots, \boldsymbol{\beta}^K\mathbf{x}$; *that is, for some constants* $c_0, c_1, \ldots, c_K$, $E(\mathbf{bx}|\boldsymbol{\beta}^1\mathbf{x}, \ldots, \boldsymbol{\beta}^K\mathbf{x}) = c_0 + c_1\boldsymbol{\beta}^1\mathbf{x} + \cdots + c_K\boldsymbol{\beta}^K\mathbf{x}$.

The main result of the SIR can be expressed as in the following theorem.

Theorem 2.1 (Li, 1991). *Under the condition 2.1, and the model (2.1) the centered inverse regression curve* $E(\mathbf{x}|y)-E(\mathbf{x})$ *is contained in the linear subspace spanned by* $\boldsymbol{\beta}^k V_{\mathbf{x}}$ $(k = 1, \ldots, K)$, *where* $V_{\mathbf{x}}$ *denotes the covariance matrix of* $\mathbf{x}$.

We see, therefore, that the eigenvectors, $\boldsymbol{\beta}^k, (k = 1, \ldots, K)$ associated with the largest K eigenvalues of $V_{\mathbf{x}|y} = cov[E(\mathbf{x}|y)]$ with respect to $V_{\mathbf{x}}$ are the standardized e.d.r. directions. Now suppose that we could get an estimate for $V_{\mathbf{x}|y}$ in some way, then we can apply generalized eigenvalue decomposition on this to get e.d.r. directions for the model (2.1).

Here is an algorithm suggested by Li in his original paper.

- Divide range of y into H slices, $I_1, \ldots, I_H$; let the proportion of the y_i that falls in slice h be $\hat{p}_h$.
- Within each slice, compute the sample mean of the $\mathbf{x}$'s and denote it by $\bar{\mathbf{x}}_h$.
- Form the weighted covariance matrix $\hat{V}_{\mathbf{x}|y} = \Sigma_{h=1}^{H}\hat{p}_h(\bar{\mathbf{x}}_h - \bar{\mathbf{x}})(\bar{\mathbf{x}}_h - \bar{\mathbf{x}})^t$.
- Compute the sample covariance for $\mathbf{x}$'s, $V_{\mathbf{x}} = n^{-1}\sum_{i=1}^{n}(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^t$.
- Find the SIR directions by conducting the eigenvalue decomposition of $V_{\mathbf{x}|y}$ with respect to $V_{\mathbf{x}}$.

$$V_{\mathbf{x}|y}\hat{\boldsymbol{\beta}}^i = \hat{\lambda}_i V_{\mathbf{x}}\hat{\boldsymbol{\beta}}^i, \tag{2.2}$$

$$\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \cdots \geq \hat{\lambda}_p \ .$$

Equation (2.2) above is known as a generalized eigenvalue decomposition and is equivalent to maximize

$$-\frac{\boldsymbol{\beta}^k V_{\mathbf{x}|y} \boldsymbol{\beta}^{kt}}{\boldsymbol{\beta}^k V_{\mathbf{x}} \boldsymbol{\beta}^{kt}}$$

subject to

$$\boldsymbol{\beta}^h V_x \boldsymbol{\beta}^{kt} = 0, \ h < k.$$

Or it is the same as maximizing

$$-\boldsymbol{\beta}^k V_{\mathbf{x}|y} \boldsymbol{\beta}^{kt}$$

subject to

$$\boldsymbol{\beta}^k V_{\mathbf{x}} \boldsymbol{\beta}^{kt} = 1 \ \text{ and } \ \boldsymbol{\beta}^h V_{\mathbf{x}} \boldsymbol{\beta}^{kt} = 0, \quad h < k \quad \text{for} \quad k \geq 2.$$

3. Penalized SIR

Now consider applying penalty function idea to one of the above generalized eigenvalue problems of the SIR. Using penalized least squares or likelihood idea in variable selection is known to be better in that they are retaining good features of both subset selection and ridge regression (Fan and Li, 2001). The stepwise deletion and subset selection as alternatives to using penalty idea tend to ignore stochastic errors inherited in the stages of variable selections and further it suffers from several drawbacks including its lack of stability as analyzed, for example, by Breiman (1995).

Suppose we have a penalty function $p_\lambda(\theta)$. Then the typical regression problems with penalized least squares or likelihood becomes to find coefficients which minimizes the following unified "Loss+Penalty" function

$$l(\boldsymbol{\beta}) + n\Sigma_{j=1}^{p} p_\lambda(|\beta_j|). \tag{3.1}$$

The first term in the above objective function (3.1) may be regarded as a loss function of $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)$, and would be $\Sigma(y - \boldsymbol{\beta}\mathbf{x})^2$ for least squares method and it becomes the negative of likelihood for generalized linear models for example.

Consider problem of applying penalized method to one of relevant SIR directions. As defined in previous Section if we let $\mathrm{E}[\mathrm{cov}(\mathbf{x}|y)] = V_{\mathbf{x}|y}$, and $\mathrm{Cov}(\mathbf{x}) = V_{\mathbf{x}}$ then the objective function which need to be minimized becomes

$$-\frac{\boldsymbol{\beta} V_{\mathbf{x}|y} \boldsymbol{\beta}^t}{\boldsymbol{\beta} V_{\mathbf{x}} \boldsymbol{\beta}^t} + \Sigma_{j=1}^{p} p_\lambda(|\beta_j|).$$

Now if we can get consistent estimate for E[cov($\mathbf{x}|y$)] in a certain way like by slicing and averaging as suggested by Li in his original SIR paper, then we have

$$-\frac{\boldsymbol{\beta}\hat{V}_{\mathbf{x}|y}\boldsymbol{\beta}^t}{\boldsymbol{\beta}\hat{V}_{\mathbf{x}}\boldsymbol{\beta}^t} + \Sigma_{j=1}^{p} p_\lambda(|\beta_j|),$$

as our objective function with appropriate penalty functions for $p_\lambda(\cdot)$. Now we can combine "Loss+Penalty" functions for all K relevant components with orthonormal constraints to successive optimization problems as follows.

Penalized SIR (PENSIR) on a covariance matrix E[cov($\mathbf{x}|y$)] finds linear combinations $\boldsymbol{\beta}^1\mathbf{x}, \boldsymbol{\beta}^2\mathbf{x}, \ldots, \boldsymbol{\beta}^K\mathbf{x}$ of the p measured variables $\mathbf{x}$ which successively have minimum "Loss + Penalty" function

$$-\frac{\boldsymbol{\beta}^k V_{\mathbf{x}|y}\boldsymbol{\beta}^{k^t}}{\boldsymbol{\beta}^k V_{\mathbf{x}}\boldsymbol{\beta}^{k^t}} + \Sigma_{j=1}^{p} p_\lambda(|\beta_j^k|) \quad (k = 1, \ldots, K)$$

subject to

$$\beta^h V_x \beta^{k^t} = 0, \ h < k.$$

Or equivalently it is the same as minimizing

$$-\boldsymbol{\beta}^k V_{\mathbf{x}|y}\boldsymbol{\beta}^{k^t} + \Sigma_{j=1}^{p} p_\lambda(|\beta_j^k|) \quad (k = 1, \ldots, K)$$

subject to

$$\boldsymbol{\beta}^k V_{\mathbf{x}}\boldsymbol{\beta}^{k^t} = 1 \text{ and (for } k \geq 2\text{)} \ \boldsymbol{\beta}^h V_{\mathbf{x}}\boldsymbol{\beta}^{k^t} = 0, h < k.$$

Several well-known penalty functions including SCAD penalty function are as follows.

L_p: $p_\lambda(|\beta_j^k|) = \lambda|\beta_j^k|^p$ and it becomes LASSO (Tibshirani, 1996) with $p = 1$ for least squares case.

Hard Thresholding (HARD) Penalty: $p_\lambda(\beta_j^k) = \lambda^2 - (|\beta_j^k| - \lambda)^2 I(|\beta_j^k| < \lambda)$.

Smoothly Clipped Absolute Deviation (SCAD) Penalty:

$$p_\lambda(\beta_j^k) = \begin{cases} \lambda\beta_j^k & \text{if } \beta_j^k < \lambda, \\ -\dfrac{\beta_j^{k^2} - 2a\lambda\beta_j^k + \lambda^2}{2(a-1)} & \text{if } \lambda \leq \beta_j^k < a\lambda, \\ \dfrac{(a+1)\lambda^2}{2} & \text{if } \beta_j^k \geq a\lambda, \end{cases}$$

where β_j^k is j's coefficient of k's component in the regression set up, and a (> 2) and λ are tuning parameters.

Fan and Li (2001) mentioned unibiasedness, sparsity, and continuity as three properties that a good penalty function should have, and suggested Smoothly Clipped Absolute Deviation (SCAD) penalty function as the best one for regression problems. Unfortunately, none of three penalty functions satisfy all those three properties simultaneously. L_p penalty function is biased and this cause some serious problem especially when applied to successive optimization problems in which coefficients compete with each other due to orthonormality conditions on them. We have seen from small simulations that bias problem in L_1 penalty is so serious that including penalty function usually resulted in domination of one or few variables with relatively large coefficients compared to others. By the way, hard thresholding (HARD) penalty function is unbiased and has sparsity but it is not continuous. SCAD behaves like something between L_1 and HARD and needs two dimensional burdensome Generalized Cross-Validation (GCV) or usual Cross-Validation (CV) to find optimal values for two parameters, a and λ.

Overall, it looks reasonable to use HARD penalty for the Penalized SIR (PENSIR) problem since it looks better in forcing coefficients of irrelevant variables to zero and at the same time minimizing bias problems after introducing penalty function in the optimization procedure. We will consider HARD penalty function only in further discussions.

4. An Algorithm

To solve two sets of optimization problems in the previous section we need to use one of numerical algorithms for multivariate function. Two sets of problems are identical in nature so one can choose either set which is easier in implementation. Simulated Annealing (SA) method, introduced by Kirkpatric *et al.* (1983) is known to give near optimal solutions for multivariate optimization problems with many local optimum so it could be effectively applicable to our problems. These methods have been known to be applicable to combinatorial optimization problems like salesman traveling problem and even applicable to continuous multivariate optimizations also. The main idea of SA is that random samples from a distribution generated from a given multivariate objective function converge to near optimal solution as scale parameter, usually known as temperature, of the distribution becomes lower.

Now, let's define the distribution

$$u(\boldsymbol{\beta}) = C \exp\left(\frac{1}{\gamma} D(\boldsymbol{\beta})\right)$$

with $D(\boldsymbol{\beta}) = l(\boldsymbol{\beta}) + n\Sigma_{j=1}^{p} p_\lambda(|\beta_j|)$. The C is normalizing constant and γ is usually called temperature. Then the algorithm for finding 1st e.d.r. direction is as follows.

STEP1 : Initialization

1. Set initial $\boldsymbol{\beta}$ from ordinary SIR
2. $k = 0$ (Step function)
3. Set initial temperature γ_0 (temperature at $k = 0$.)
4. Set initial number of iteration L_0 (number of sample at $k = 0$.)
5. Set $\gamma = \gamma_0$.

STEP2 : Repeat until convergence.

1. For $l = 1$ to L_k
   (a) Set $\boldsymbol{\beta}^{\text{new}}$ from neighborhood of $\boldsymbol{\beta}^{\text{old}}$.
   (b) Set $\boldsymbol{\beta}^{\text{old}} = \boldsymbol{\beta}^{\text{new}}$ if $D(\boldsymbol{\beta}^{\text{new}}) \leq D(\boldsymbol{\beta}^{\text{old}})$.
   (c) Set $\boldsymbol{\beta}^{\text{old}} = \boldsymbol{\beta}^{\text{new}}$ if $D(\boldsymbol{\beta}^{\text{new}}) > D(\boldsymbol{\beta}^{\text{old}})$ and $\exp([D(\boldsymbol{\beta}^{\text{old}}) - D(\boldsymbol{\beta}^{\text{new}})]/\gamma_k) > U[0, 1]$.
2. $k = k + 1$
3. Set $\gamma_k = \gamma_0 \times (0.9)^k$.

OUTPUT : $\boldsymbol{\beta}^{\text{old}}$.

In implementing SA, it is known to be very important to set parameters and initial values for the model properly in order to get resonable solutions. For the objective function, we need to provide appropriate intial values for $\boldsymbol{\beta}$'s first and then it needs to choose neighborhood of them for the next iteration carefully. Also parameters in the SA algorithm should be calibrated properly, too. Here are some details for these issues.

Coefficient estimates from ordinary SIR would be a very good initial values for $\boldsymbol{\beta}$'s in the first and each subsequent e.d.r. directions in the PENSIR. Second and subsequent components should be orthogonal to all previously obtained e.d.r. directions. Hence, it should be considered in finding possible neighbor of $\boldsymbol{\beta}$'s

for each iteration process so that new components should still be very close to previous one to guarantee convergence of the algorithm.

We clearly have advantage in deciding on optimal λ since for any SIR problems we will have the value of loss function roughly close to -1. So it seems to be enough to find resonable λ which should be greater than 0 and at the same time less than eigen value for each e.d.r. direction found from ordinary SIR. It would be better try to find an optimal or near optimal value as the percentage of eigen value from SIR. It could be an option to set λ as a function of $|l|$, absolute value of the loss function, p, the number of predictors, and an appropriate multiplier, so that

$$\lambda = r \times |l| \sqrt{2 \log(p)}$$

with $r \geq 0$. Clearly, when $r = 0$ it would give the same results from ordinary SIR without any constrains. And when $r \times \sqrt{2 \log(p)} \geq 1$ the algorithm forces all coefficient estimates except only one variable to zero. Hence optimal r should be between 0 and $1/\sqrt{2 \log(p)}$. Small experiments only with simulated and real data sets included in this paper suggest r should be between 0 and 0.2 and the solution tends to include one dominating coefficient as r is greater than 0.2 or so. Details of setting parameters in the SA algorithm can be found, as an example, in Aarts and Korst (1989) and will not be discussed further here.

5. Illustrations

In this section we generated simulated data sets with several regression models and applied proposed penalized SIR approaches to demonstrate how it works. In all examples we obtained penalized SIR estimates with hard thresholding penalty functions only. We also include results of our penalized SIR approach applied to a real data set.

5.1. Simulated Examples

Data sets for the first and the second simulated examples are generated from linear regression model with independent and correlated predictors, respectively. And the third data set is generated from a regression model in which the regression function is ratio of two linear combination of predictors with usual additive errors so that the the dimension of the model is two.

Example 5.1 In this example we simulated a dataset consisting of 400 ob-

servations from the model

$$y = 3x_1 + 1.5x_2 - 2x_5 + \epsilon,$$

where $x = (x_1, x_2, \ldots, x_8) \sim \mathrm{N}(0, 1)$ and $\epsilon \sim \mathrm{N}(0, 0.01)$.

Table 5.1: Coefficient estimates for linear regression models for several λ's

	Variable							
λ	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
SIR	0.767	0.379	0.002	−0.019	−0.517	−0.006	0.003	0.007
0	0.767	0.374	0.001	−0.019	−0.521	−0.022	−0.009	0.005
0.1	0.768	0.373	0	0	−0.520	0	0	0
0.2	1	0	0	0	0	0	0	0

We can successfully force all coefficients to zero for irrelevant predictor with $\lambda = 0.1$ as shown in the Table 5.1. Coefficient estimates on the top with $\lambda =$"SIR" correspond to estimates with $\lambda = 0$. Hence, in this case no penalty is imposed on the model. Even with HARD penalty functions the predictor x_1 dominates with λ larger than 0.2.

Example 5.2 The linear regression function in this example is the same as in the first example except that predictors are correlated and two more unnecessary predictors are added. This example is a simulated one with $(x_1, x_3, x_4, x_5, x_7, x_8, x_9, x_{10}) \sim \mathrm{N}(0, 1)$, $x_2 = 2x_5 + N(0, 1)$ and $x_6 = x_1 + N(0, 1)$ with same model and error terms. The number of observation is 400 also.

Results (see, Table 5.2 for details.) are quite similar to the first one regardless of correlated predictors and unnecessary predictors added. When the λ is 0.1 all coefficient estimates for irrelevant predictors become 0's and firt two predictors tend to dominate with $\lambda = 0.3$ or larger.

Example 5.3 In this example we simulated a data set consisting of 400 observations from the model

$$y = \frac{x_1 - 2x_3}{(x_2 + x_4 + 1.5)^2} + \epsilon$$

with $x = (x_1, x_2, \ldots, x_{10}) \sim \mathrm{N}(0, 1)$ and $\epsilon \sim \mathrm{N}(0, 0.01)$.

Table 5.2: Coefficient estimates for linear regression models with correlated predictors

	Variable									
λ	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
SIR	0.767	0.390	0.002	−0.005	−0.510	−0.004	−0.004	0.004	−0.010	0.006
0	0.766	0.390	0.002	−0.006	−0.511	−0.005	−0.004	0.003	−0.009	0.007
0.1	0.767	0.389	0	0	−0.510	0	0	0	0	0
0.2	0.797	0.371	0	0	−0.477	0	0	0	0	0
0.3	0.997	0.081	0	0	0	0	0	0	0	0

Table 5.3: Coefficient estimates of first two components for the rational model

	Variable									
λ	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
Component 1										
SIR	0.462	0.134	−0.860	0.139	0.038	−0.045	−0.042	0.023	−0.041	0.043
0.09	−0.495	0	0.869	0	0	0	0	0	0	0
0.13	−0.493	0	0.870	0	0	0	0	0	0	0
0.18	−0.494	0	0.870	0	0	0	0	0	0	0
0.27	−0.493	0	0.870	0	0	0	0	0	0	0
Component 2										
SIR	−0.089	0.640	0.327	0.654	0.008	−0.135	−0.102	0.110	0.039	0.076
0.09	0	0.678	0	0.689	0	−0.165	−0.090	0.144	0	0.100
0.13	0	0.700	0	0.704	0	0	0	0.143	0	0
0.18	0	0.709	0	0.705	0	0	0	0	0	0
0.27	0	0.707	0	0.705	0	0	0	0	0	0

Coefficient estimates for the first two components are in Table 5.3. Clearly, the proposed algorithm seems to work well on an unusual regression function too. Estimates with $\lambda = 0.18$ are similar to true values and it successfully forces all irrelevant coefficients to zero for both components. Further we can see it still works fine with larger λ of 0.27.

5.2. Real Example

Results for application of our approach to only one real data set known as Ozone data set is included. As similar to simulated illustraions we included estimates for e.d.r. directions.

Example 5.4 This is well-known Ozone data set with Ozone as a response,

Table 5.4: Coefficient estimates of Penalized SIR for OZONE data set

	Variable							
λ	Temp.	InvHT	Pres.	Visi.	Height	Humidity	Temp2	WindSp.
SIR	−0.962	0.003	−0.042	0.022	0.012	−0.195	−0.184	0.003
0	−0.962	0.003	−0.041	0.022	0.013	−0.196	−0.182	0.016
0.1	−0.968	0.003	0	0	0	−0.250	0	0
0.125	−0.969	0.003	0	0	0	−0.249	0	0

and Temparature, InversionHT, Pressure, Visibility, Height, Humidity, Temp2, WindSpeed as predictors. The number of observation is 330. We only look at the first direction only since only the first direction is known to be significant from ordinary SIR.

It looks like three predictors, Temparature, InversionHT, and Humidity are significant as in Table 5.4. Further it is clear from coefficient estimate of InvHT that a important factor which gurantees non-zero coefficient is not the magnitude of estimate itself.

6. Concluding Remarks

In this paper we propose a variable selection method for Sliced Inverse Regression analysis. We incoporated penalty functions for each coefficient estimates and solve penalized optimization problem using simulated annealing algorithm. According to results from simulated and real data sets we found our method turned out to be very effective in forcing coefficient estimates zero for irrelevant predictors for diverse regression models not only with linear but also with non-linear regression functions composed of more than one linear combination of predictors. And HARD penalty function is appropriate with relatively small bias for wide range of λ values for the SIR problem than well-known SCAD and L_1 penalty functions which is known to be effective in usual linear and generalized linear regression problems.

In practical regression problems with large number of predictors the proposed method can be used effectively to obtain smaller set of relevant predictors for further analysis especially when most of predictors are known to be irrelevant. More detailed research for theoretical results regarding properties and asmptotics for coefficient estimates is necessary.

References

Aarts, E. and Korst, J. (1989). *Simulated Annealing and Boltzmann Machines.* John Wiley & Sons.

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

Cadima, J. and Jolliffe, I. T. (1995). Loadings and correlations in the interpretation of principal components. *Journal of Applied Statistics*, **22**, 203–214.

Fan, J. and Li, R. (2001). Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties. *Journal of The American Statistical Association*, **96**, 1348–1360.

Hausman, R. E. Jr. (1982). Constrained multivariate analysis. *Optimisation in Statistics* (Zanckis, S. H. and Rustagi, J. S., eds.), 137-151, North-Holland: Amsterdam.

Jolliffe, I. T. (1972). Discarding variables in a principal component analysis. I: artificial data. *Applied Statistics*, **21**, 160–173.

Jolliffe, I. T. (1973). Discarding variables in a principal component analysis. ii: real data. *Applied Statistics*, **22**, 21–31.

Jolliffe, I. T. (1989). Rotation of Ill-defined principal components. *Applied Statistics*, **38**, 139–147.

Jolliffe, I. T. (1995). Rotation of principal components: choice of normalization constraints. *Journal of Applied Statistics*, **22**, 29–35.

Jolliffe, I. T., Trendafilov, N. T. and Uddin, M. (2003). A modified principal component technique based on the Lasso. *Journal of Computational and Graphical Statistics*, **12**, 531–547.

Kirkpatrick, S., Gelatt, C. D. Jr. and Vecchi, M. P. (1983). Optimization by simulated annealing. *Science*, **220**, 671–680.

Li, K. C. (1991). Sliced inverse regression for dimension reduction. *Journal of The American Statistical Association*, **86**, 316–342.

Li, K. C. (2000). High dimensional data analysis via the SIR/PHD approach. *unpublished manuscript.*

McCabe, G. P. (1984). Principal variables. *Technometrics*, **26**, 137–144.

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

Vines, S. K. (2000). Simple principal components. *Applied Statistics*, **49**, 441–451.

[Received January 2007, Accepted March 2007]