

Boiling Points Predictions Study via Dimension Reduction Methods: SIR, PCR and PLSR

Hong Yin^{1,2}, Yizeng Liang³ and Qinnan Hu³

¹*Hong Kong Baptist University*, ²*Wuhan University* and

³*Central South University*

Abstract: Variable selection is an important tool in QSAR. In this article, we employ three known techniques: sliced inverse regression (SIR), principal components regression (PCR) and partial least squares regression (PLSR) for models to predict the boiling points of 530 saturated hydrocarbons. With 122 topological indices as input variables our results show that these three methods have good performance and perform better than some existing methods in the literature.

Key words: Cross-validation, dimension reduction, partial least squares regression, principal component regression, sliced inverse regression.

1. Introduction

The goal of quantitative structure and activity relationship (QSAR) research is to establish a relationship between certain molecular activity and molecular descriptors by means of statistic tool (Devillers and Balaban 1999):

$$A(\text{activity}) = f(\text{molecular structure}) = f(\text{molecular descriptors}).$$

For example, one can predict the molecular physicochemical, biological and toxicological properties from the statistic model based on chemical characteristics. One of the most frequently used chemical characteristics is the

so-called topological index (TI). A TI is a numerical value which is intended to characterize the chemical structure of the underlying chemical compound. This numerical value quantifies a link between molecular activity and molecular structure of this compound. The concept of TI is so attractive that we are flooded with various TI's in the literature. Since the Wiener index, the first TI, was proposed in 1947, more than four hundred TI's have been defined by researchers. The TI's are proposed with a mind to capture the most important characteristic of molecular activity and structure of the compound. But molecular activity and structure are complex matters so no such TI that can completely determinate molecular activity for a certain class of chemical compounds. A natural approach, in fact a popular direction used by many authors, is to consider all possible TI's and choose a set of TI's such that one can predict a certain molecular activity based on this set.

Along this line of approach there exist many difficulties, such as (a) there are strong collinear among TI's; (b) many useful techniques of model selection, like the backward elimination and best subset, cannot be implemented as the number of possible models is 2^p which become large exponentially fast. It is clear that, with this approach, it is difficult to reach an ideal model. As a result, many more new TI's have been proposed every year. It leads more serious situation in difficulties (a) and (b). Thus, selection of the efficient variables becomes very important in QSAR studies. How can we select efficient variables and use them to construct an ideal model? In this paper, we take the view that the performance of a learning method is measured by its prediction capability on independent test data. From this view, the better prediction on the new test data, the better the model is. We will use cross-validation for performance assessment in this paper. But first let us discuss about the interplay between bias, variance and model complexity.

1.1 Model selection and bias-variance tradeoff

Assume we have p input variables for each of the n responses. For example, in this paper, each variable may be a certain TI and each response is a boiling point of a compound under consideration. The output of n samples

can be summarized as a n -vector $\mathbf{y} = (y_1, y_2, \dots, y_n)'$. The i -th observed value is written in lowercase as $\mathbf{x}_i = (x_{i1}, x_{i2}, \dots, x_{ip})$, corresponds to the p TI's of this compound. In this paper, matrices are denoted by bold uppercase letters, for example, a set of n input p -vectors \mathbf{x}_i , $i = 1, 2, \dots, n$ would be written by the $n \times p$ matrix $\mathbf{X} = (x_{ij})$ and the j -th column of \mathbf{X} is denoted by $\mathbf{x}_{(j)}$. Give an input matrix \mathbf{X} , the prediction of output \mathbf{y} is typically denoted by $\hat{\mathbf{y}}$. We will use the design matrix \mathbf{X} and the response data \mathbf{y} to find new variables for better prediction of the boiling points of compounds in the testing samples. These new variables are also called new features and the methods we employ are basically dimension reduction techniques. All the new features derived by the dimension reduction methods will be denoted as \mathbf{z}_m , $m = 1, 2, \dots, M$. The meaning of these notation will remain unchanged through this paper unless specified otherwise.

Assume that the true model is $\mathbf{y} = f(\mathbf{X}) + \epsilon$ where $E(\epsilon) = 0$, $Var(\epsilon) = \sigma_\epsilon^2$. It is unknown and we want to find an approximate model $\hat{f}(\mathbf{X})$ to replace the true one. The square residual at $\mathbf{X} = \mathbf{x}_0$ under the $\hat{f}(\mathbf{X})$ is

$$\begin{aligned} Err(\mathbf{x}_0) &= E[(y_0 - \hat{f}(\mathbf{x}_0))^2 | \mathbf{X} = \mathbf{x}_0] \\ &= \sigma_\epsilon^2 + [f(\mathbf{x}_0) - E\hat{f}(\mathbf{x}_0)]^2 + E[\hat{f}(\mathbf{x}_0) - E\hat{f}(\mathbf{x}_0)]^2. \\ &= \sigma_\epsilon^2 + Bias^2(\hat{f}(\mathbf{x}_0)) + Var(\hat{f}(\mathbf{x}_0)) \\ &= Irreducible\ Error + Bias^2 + Variance. \end{aligned}$$

For a linear approximate model $\hat{f}(\mathbf{x}) = \mathbf{x}\hat{\beta}$, where $\hat{\beta}$ is the least square estimator of β , we have

$$\begin{aligned} Err(\mathbf{x}_0) &= E[(y_0 - \hat{f}(\mathbf{x}_0))^2 | \mathbf{X} = \mathbf{x}_0] \\ &= \sigma_\epsilon^2 + [f(\mathbf{x}_0) - E\hat{f}(\mathbf{x}_0)]^2 + Var[\hat{f}(\mathbf{x}_0)] \\ &= \sigma_\epsilon^2 + [f(\mathbf{x}_0) - E\hat{f}(\mathbf{x}_0)]^2 + \|h(\mathbf{x}_0)\|^2 \sigma_\epsilon^2. \end{aligned}$$

Here $h(\mathbf{x}_0) = \mathbf{x}_0(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$. While this variance changes with \mathbf{x}_0 , its average (over n_1 testing sample values) is $\frac{p}{n_1}\sigma_\epsilon^2$, hence,

$$\frac{1}{n_1} \sum_{i=1}^{n_1} Err(\mathbf{x}_i) = \sigma_\epsilon^2 + \frac{1}{n_1} \sum_{i=1}^{n_1} [f(\mathbf{x}_i) - E\hat{f}(\mathbf{x}_i)]^2 + \frac{p}{n_1} \sigma_\epsilon^2 \quad (1)$$

How to minimize $\frac{1}{n_1} \sum_{i=1}^{n_1} Err(\mathbf{x}_i)$ is the goal of constructing a satisfied model. Let's study the three parts on the right hand side of equation (1)

above:

- (1) The first term on the right hand side of (1) cannot be avoided no matter how well $f(\mathbf{x}_i)$ to be estimated;
 (2) The second term is:

$$Bias^2 = \frac{1}{n_1} \sum_{i=1}^{n_1} [f(\mathbf{x}_i) - E\hat{f}(\mathbf{x}_i)]^2.$$

Let β_* denote the parameters of the best fitting linear approximate to f :

$$\beta_* = \operatorname{argmin}_{\beta_*} E(f(\mathbf{x}) - \mathbf{x}\beta)^2,$$

then,

$$\begin{aligned} Bias^2 &= \frac{1}{n_1} \sum_{i=1}^{n_1} [f(\mathbf{x}_i) - E\hat{f}(\mathbf{x}_i)]^2 \\ &= \frac{1}{n_1} \sum_{i=1}^{n_1} [f(\mathbf{x}_i) - \mathbf{x}_i\beta_*]^2 + \frac{1}{n_1} \sum_{i=1}^{n_1} [\mathbf{x}_i\beta_* - E(\mathbf{x}_i\hat{\beta})]^2 \\ &= [Model\ Bias]^2 + [Estimation\ Bias]^2. \end{aligned}$$

If $\hat{\beta}$ is ordinary least squares estimation, the “*Estimation Bias*” is zero, while “*Model Bias*” can only be reduced by enlarging the dimension of the input variables space, p .

- (3) The third term $\frac{p}{n_1}\sigma_\epsilon^2$ obviously increases as p increases.

So there is a bias-variance tradeoff behavior. The Figure 1 shows the typical behavior of the test and training error, as model complexity is varied.

1.2 Cross-validation

In this subsection we describe the cross validation method that we shall use to find a suitable set of p variables in our eventual model which minimizes prediction error of testing samples.

Ideally if we have enough data, we would set aside a validation set and use it to assess the performance of our prediction model. Since data are often scarce, this is usually not possible. To finesse the problem, K -fold cross-validation uses part of the available data to fit the model, and a different part to test it. We split the whole data into K roughly equal-sized parts; for

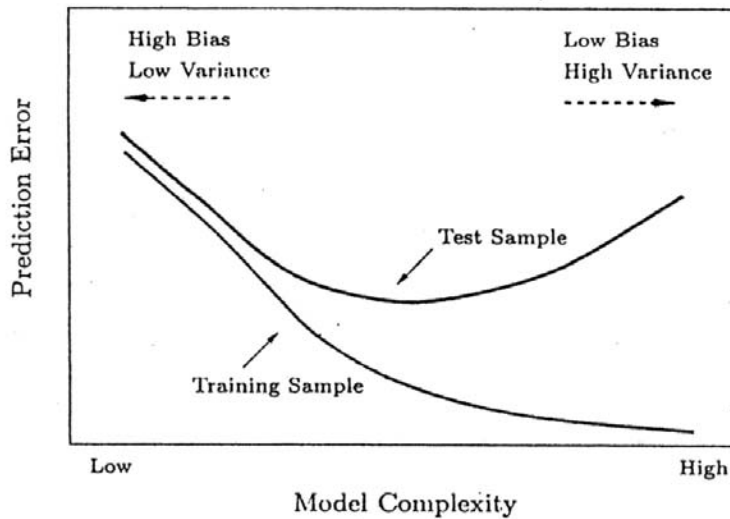


Figure 1: Test and training error as a function of model complexity.

example, denote the full data set by T , divide T into K mutually disjoint subsets with approximately the same size. For each $k = 1, 2, \dots, K$, we use $T - T^k$ and T^k as the training set and testing set respectively. We find the fitted function $\hat{f}^{-k}(\beta)$ using the training set $T - T^k$ and then calculate the prediction error of the fitted model when predicting k -th part of the data. We repeat this for $k = 1, 2, \dots, K$ and combine the K estimates of prediction error. So the cross-validation estimate of prediction error is:

$$CV(\beta) = \frac{1}{N} \sum_{i=1}^N (y_i - \hat{f}^{-k(i)}(\mathbf{x}_i; \beta))^2.$$

Typical choices of K are 5 or 10 when the number of observations is very large. The case $K = N$ is known as *leave-one-out* cross-validation. Our data has 530 observations and we employ *leave-one-out* cross-validation to assess the ability of every method. At the same time, the function $CV(\beta)$ provides an estimate of the test error curve, and we find the tuning parameter β that minimizes $CV(\beta)$. The final model is thus obtained for fitting all the data.

In this paper three function approximation methods, i.e. sliced inverse regression (SIR), principal components regression (PCR) and partial least squares regression (PLSR) are compared. These methods are evaluated under the framework of chemometrics, especially in QSAR research. We further analyze similarities, difference and efficiency of these methods from their basic reasoning and concept. Finally, we compare our results with that of previous work and conclude that SIR, PCR and PLSR are promising methods in reducing the dimension of variable space.

This paper is organized as follows. In Section 2, we will briefly introduce the methods of SIR, PCR and PLSR. Then, in Section 3 these three methods are applied to a real set of chemical data. The cross-validation is employed for comparisons among the models recommended by the three methods.

2. Dimension Reduction

In this section, three methods, sliced inverse regression (SIR), principal components regression (PCR) and partial least regression (PLSR) are briefly introduced.

2.1 Sliced inverse regression (SIR)

The data set for SIR is of the same form as for a typical linear regression. There are n observations on p variables. A typical observation is of the form $(y_i, x_{i1}, x_{i2}, \dots, x_{ip})$, $i = 1, 2, \dots, n$. We can, of course, rewrite it with the familiar matrix notation: the response vector \mathbf{y} is a $n \times 1$ together with a design matrix \mathbf{X} which is $n \times p$ in size. The distribution of \mathbf{X} is assumed elliptically symmetric (e.g., the normal distribution). Li (1991) suggests a model with the structure

$$y = g(\mathbf{x}\beta_1, \mathbf{x}\beta_2, \dots, \mathbf{x}\beta_m, \epsilon),$$

which is used to approximate the true one, where g is an unknown function. In SIR, $\beta_1, \beta_2, \dots, \beta_m$ are m p -dimensional column vectors, called projection directions. A salient feature of SIR is the concept of the effective dimension reduction (e. d. r.) space \mathbf{B} which is generated by the m vectors $\beta_1, \beta_2, \dots, \beta_m$. Any non-zero vector in the e.d.r. space is called

an e.d.r. direction, and in a typical SIR application, it is hoped that m is much smaller than p . If this is the situation, then we only need to find estimates of these m p -dimensional vectors $\beta_1, \beta_2, \dots, \beta_m$ determined up to an equivalence relation defined by $(\mathbf{x}\beta_1, \mathbf{x}\beta_2, \dots, \mathbf{x}\beta_m)$.

In exploring non-linear relationship between y and \mathbf{x} , SIR considers the inverse regression that treats y as if it were the independent variable and treat \mathbf{x} as if it were the dependent variable. In fact, it is the curve $\eta(y) = E(\mathbf{x}|y)$, treated as a function of \mathbf{x} but conditioning on y , that is to be explored. We denote conditional covariance matrix of \mathbf{x} by $\Sigma_\eta = \text{Cov}(E(\mathbf{x}|y))$. Note that the point $E(\mathbf{x}|y = y)$, being p -dimensional by definition, stays in p -dimensional space, it can be shown that it is contained in the linear subspace spanned by $\Sigma_{\mathbf{x}}\beta_i$, $i = 1, 2, \dots, m$, where $\Sigma_{\mathbf{x}}$ is the covariance matrix of \mathbf{x} . The details of the theory can refer to Li (1991).

The estimations of $\beta_1, \beta_2, \dots, \beta_m$ can be obtained from the eigenvalue decomposition of Σ_η with respect to $\Sigma_{\mathbf{x}}$: $\Sigma_\eta\beta_i = \lambda_i\Sigma_{\mathbf{x}}\beta_i$, ($\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p$) where λ_i is the i -th eigenvalue and β_i is the corresponding eigenvector. For detailed statistical theories justifying the inverse regression view, readers are referred to Li (1991), and Duan and Li (1991), Zhu and Fang (1996) and Chen and Li (1998).

Algorithm 1. Sliced inverse regression

1. Standardize each observation $\mathbf{x}_i = (x_{i1}, x_{i2}, \dots, x_{ip})$ by $\tilde{\mathbf{x}}_i = (\mathbf{x}_i - \bar{\mathbf{x}})\hat{\Sigma}_{\mathbf{x}}^{-\frac{1}{2}}$ for $i = 1, 2, \dots, n$, where $\hat{\Sigma}_{\mathbf{x}}$ and $\bar{\mathbf{x}}$ are the sample variance matrix and sample mean, $\hat{\Sigma}_{\mathbf{x}}^{1/2}$ is the positive defined squared root of $\hat{\Sigma}_{\mathbf{x}}$ and $\hat{\Sigma}_{\mathbf{x}}^{-1/2}$ is the inverse of $\hat{\Sigma}_{\mathbf{x}}^{1/2}$.

2. Sort the response $\mathbf{y} = (y_1, y_2, \dots, y_n)'$ from the smallest to the largest.

3. Partition the standardized data set $\tilde{\mathbf{x}}_i$'s into H slices according to the sorted sequence of \mathbf{y} : n_h is the sample size in the slice I_h , $h = 1, 2, \dots, H$. We choose I 's so that n_h are approximately equal.

4. Compute the sample mean of each slice:

$$\hat{\eta}_h = \frac{1}{n_h} \sum_{i \in I_h} \tilde{\mathbf{x}}_i.$$

5. Find the eigenvalues and eigenvectors for the weighted covariance

matrix:

$$\hat{\mathbf{V}} = \frac{1}{n} \sum_{h=1}^H n_h \hat{\eta}'_h \hat{\eta}_h$$

denote its sorted eigenvalues by $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \dots \geq \hat{\lambda}_p$ and their corresponding eigenvectors by $\hat{\mathbf{v}}_1, \hat{\mathbf{v}}_2, \dots, \hat{\mathbf{v}}_p$.

6. Compute the estimation of e.d.r. direction by $\hat{\beta}_i = \hat{\Sigma}_{\mathbf{X}}^{-\frac{1}{2}} \hat{\mathbf{v}}_i$, $i = 1, 2, \dots, m$.

We call the eigenvector $\hat{\beta}_1$ the first SIR direction and $\mathbf{X}\hat{\beta}_1$ the first SIR variate, $\mathbf{X}\hat{\beta}_2$ the second SIR variate, and so on. After finding a good e. d. r. space, we can project data into this smaller space and then estimate the response surface applying smoothing techniques to this projected variables. For the sake of simpleness, the models which we will estimate are assumed linear regression models based on the new input variables.

2.2 Principal components regression (PCR)

When a regression model has a large number of input variables, the collinearity between variables is always the problem. One approach to overcome this difficulty is to transfer the original variables to be new orthogonal variables and hoping just a few of them will accommodate most of the variations of the data set. This approach has been long employed in the statistical literature. For example, the principal components analysis accords with this idea. It “models out” the main part of the variation from the original data set \mathbf{X} by:

$$\mathbf{z}_m = \mathbf{X}\mathbf{v}_m, \quad m = 1, 2, \dots, M \leq p,$$

where the vector \mathbf{v}_m is the eigenvector associated to the m -th largest eigenvalue of the covariance matrix $\text{Cov}(\mathbf{X})$. If this set of eigenvalues can be found, then a model based on \mathbf{y} , regressed over $\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_M$, can serve as a simpler model for prediction. Since the principal components \mathbf{z}_m 's are orthogonal, this regression is just a sum of univariate regressions,

$$\hat{\mathbf{y}}^{pre} = \bar{\mathbf{y}} + \sum_{m=1}^M \hat{\theta}_m \mathbf{z}_m,$$

where $\hat{\theta}_m = \frac{\langle \mathbf{z}_m, \mathbf{y} \rangle}{\langle \mathbf{z}_m, \mathbf{z}_m \rangle}$. And because the \mathbf{z}_m 's are linear combinations of the original $\mathbf{x}_{(j)}$'s, we can express the least squares solution in term of coefficients of the $\mathbf{x}_{(j)}$'s

$$\hat{\beta}^{pre}(M) = \sum_{m=1}^M \hat{\theta}_m \mathbf{v}_m$$

Note that if $M = p$, we would just get back the usual least squares estimates. If M is less than p , the number of dimension of original variable space can be reduced.

2.3 Partial least squares regression (PLSR)

Partial least squares regression (PLSR) is another approach aimed for dimension reduction which was introduced by Wold (1975). The idea of PLSR is similar to that of PCA, but with the modification that both \mathbf{y} and \mathbf{X} are considered in the process. In PCA, only \mathbf{X} is used. It begins by computing the univariate regression coefficient $\hat{\gamma}_{1j}$ of \mathbf{y} on each $\mathbf{x}_{(j)}$, that is, $\hat{\gamma}_{1j} = \langle \mathbf{x}_{(j)}, \mathbf{y} \rangle / \langle \mathbf{x}_{(j)}, \mathbf{x}_{(j)} \rangle$. When $\mathbf{x}_{(j)}$ is standardized to have mean 0 and variance 1, we will have a simpler formula: $\hat{\gamma}_{1j} = \langle \mathbf{x}_{(j)}, \mathbf{y} \rangle$. The quantity $\mathbf{z}_1 = \sum \hat{\gamma}_{1j} \mathbf{x}_{(j)}$ is called a derived input which is the first partial least squares variate. From this expression, it can be seen that each inputs is weighted by the strength of their univariate effect on \mathbf{y} .

If we use the first new feature to fit the response \mathbf{y} , the least square estimate $\hat{\theta}_1$ on \mathbf{z}_1 can be found. Then we orthogonalize $\mathbf{x}_{(1)}, \mathbf{x}_{(2)}, \dots, \mathbf{x}_{(p)}$ with respect to \mathbf{z}_1 to get another set of input variables. We use this set of input variables as before, and based on which we produce the second PLSR variate. Repeat the process until $M \leq p$ new PLSR variables have been computed.

So we obtain a sequence of derived inputs $\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_M$. If the model involves using all $\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_p$, it reduces to the usual least squares estimates. The algorithm 2 gives the process and Chart 1 is the flow chart of PLSR:

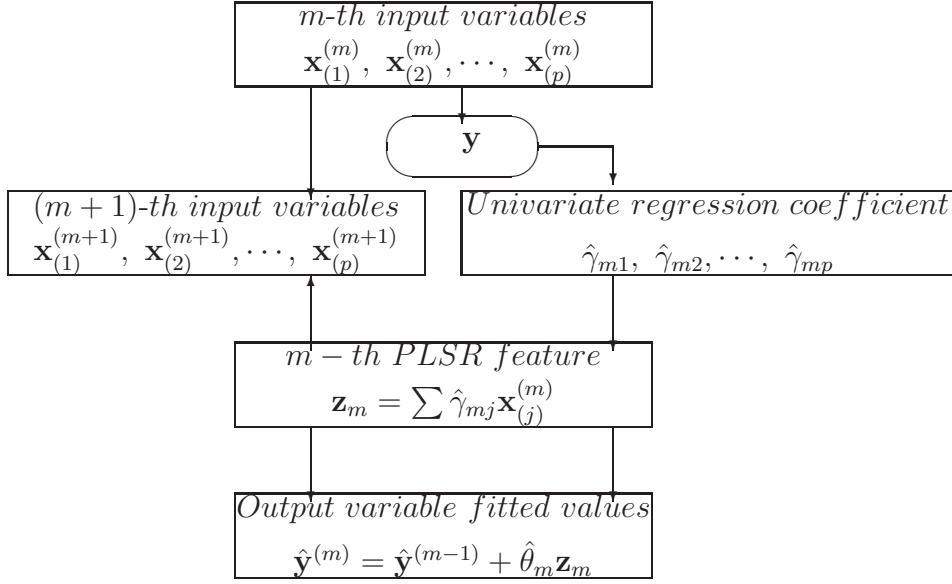


Chart 1 : Flow chart of PLSR

Algorithm 2. Partial least squares regression

1. Standardize each $\mathbf{x}_{(j)}$ and set $\hat{\mathbf{y}}^{(0)} = \bar{\mathbf{y}}, \mathbf{x}_{(j)}^{(0)} = \mathbf{x}_{(j)}, j = 1, 2, \dots, p$;
2. For $m = 1$ to p

$$\mathbf{z}_m = \sum_{j=1}^p \hat{\gamma}_{mj} \mathbf{x}_{(j)}^{(m-1)}, \text{ where } \hat{\gamma}_{mj} = \langle \mathbf{x}_{(j)}^{(m-1)}, \mathbf{y} \rangle,$$

$$\hat{\theta}_m = \langle \mathbf{z}_m, \mathbf{y} \rangle / \langle \mathbf{z}_m, \mathbf{z}_m \rangle,$$

$$\hat{\mathbf{y}}^{(m)} = \hat{\mathbf{y}}^{(m-1)} + \hat{\theta}_m \mathbf{z}_m,$$

orthogonalize each $\mathbf{x}_{(j)}^{(m-1)}$ with respect to \mathbf{z}_m :

$$\mathbf{x}_{(j)}^{(m)} = \mathbf{x}_{(j)}^{(m-1)} - \left[\langle \mathbf{z}_m, \mathbf{x}_{(j)}^{(m-1)} \rangle / \langle \mathbf{z}_m, \mathbf{z}_m \rangle \right] \mathbf{z}_m, \quad j = 1, 2, \dots, p.$$

3. Select the optimal m according to test prediction error.

3. A Case Study on Boiling Points via Topological Indices

This section concerns with boiling points of 530 saturated hydrocarbons with 2-10 carbons. The 122 TIs are chosen in Table 1 and \mathbf{y} is one of their physical activity: boiling points at normal pressure. The three dimension reduction methods mentioned in the previous section are employed to this data.

Table 1: 122 indices names used in this data and corresponding references

Names	References
9 Atis	Moreau and Broto, 1980a and 1980b
1 Bdi	Balaban 1983
7 Chi	Randic 1975; Kier and Hall 1976
1 Dddi	Lukovits 1996
1 Dddqi	Balaban 1983
1 Di	Lukovits 1996; Razinger 1997
1 Dvali	Balaban <i>et al.</i> 1992
1 Hari	Plavsic <i>et al.</i> 1993
1 Hsyi	Hosoya 1971
1 Idi	Randic 1997
1 J	Balaban 1982
1 Hwi	Klein <i>et al.</i> 1995
4 Infi	Bonchev and Trinajstic 1977
4 Kappai	Kier 1985, 1986
30 Medi	Liu <i>et al.</i> 1998; Liu <i>et al.</i> 2001.
2 Mti	Schultz 1989
1 Ordi	Balaban 1982
1 Pgi	Lukovits 1998 (all-path version of graph)
1 Qwi	Mohar <i>et al.</i> 1993
14 Seti	Filip <i>et al.</i> 1987
14 Speceiti	Lovasz and Pelikanm 1973
10 Szi	Khadikar <i>et al.</i> 1995
1 Tci	Carbon number (no reference)
3 Tri	Filip <i>et al.</i> 1987
4 Uvxyi	Balaban and Balaban 1991 and 1992
4 Uvxyoi	Ivanciuc and Balaban 1999
1 Wi	Wiener 1947
2 Zi	Gutman <i>et al.</i> 1975

The data and related study can be found in Rucker and Rucker (1999). They chose TIs by the use of variable selection procedures in SAS under

multi-linear regression model.

3.1 Data clearing

Data clearing is important in data mining. There are a lot of preprocess before making an analysis of data, such as filling in missing values, smoothing out noise, detecting outliers, deleting unnecessary variables, and so on. According to the characteristics of our chemical data, we carry through the following actions.

If we code the TIs from 1 to 122, there are 23 couples of indices whose correlation coefficient are 1. These 23 couples indices are: (1, 53), (2, 54), (21, 35), (22, 36), (23, 37), (24, 38), (25, 39), (26, 40), (27, 41), (28, 42), (29, 43), (30, 44), (31, 45), (32, 46), (33, 47), (34, 48), (79, 99), (82, 102), (83, 103), (85, 105), (86, 106), (87, 107), (88, 108). It indicates that the two indices of each pair contain the same information. So we should get rid of 23 indices from each pair of variables. The 99 remaining indices form a reduced matrix, without loss of generality, we still denote it as \mathbf{X} . This reduced \mathbf{X} is, however, still not full rank. We throw off 7 additional variables to make the remaining \mathbf{X} full rank. Our study is based on the remaining 92 variables. There are 30 variables removed in this preprocess. They are (1 19 28 32 35 36 37 38 39 40 41 42 43 44 45 46 47 48 53 54 80 81 84 99 102 103 105 106 107 108).

3.2 Results of dimension reduction methods

For convenience, we use M to denote the number of variables which are used in the reducing dimension methods. We first use SIR to get 92 projection directions and these are used for input variables. Figure 2 gives plot of test and training error against the number of variables. It shows that the training error decreases when M increases while the same behavior does not always occur for test error. The minimal test root mean squared error ($RSME = \sqrt{ave(\mathbf{y} - \hat{\mathbf{y}})^2}$) of SIR is 4.3625 at $M = 57$ and when M ranges in 30 to 70 the behavior of test error is more stable. Figure 3 gives the performance of PCR. When M increases to 30, the test error begins to level off until M is about 60 and the minimal test error of PCR is 4.3267

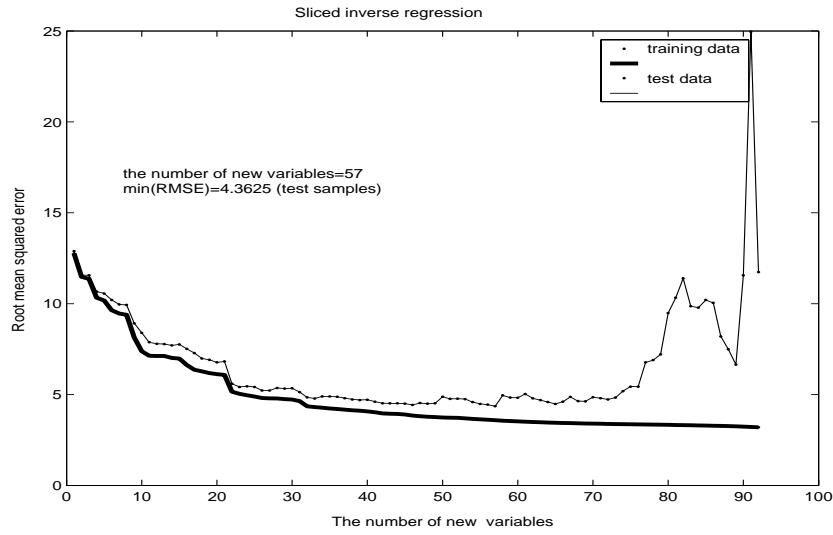


Figure 2: Test and training error of SIR through cross-validation

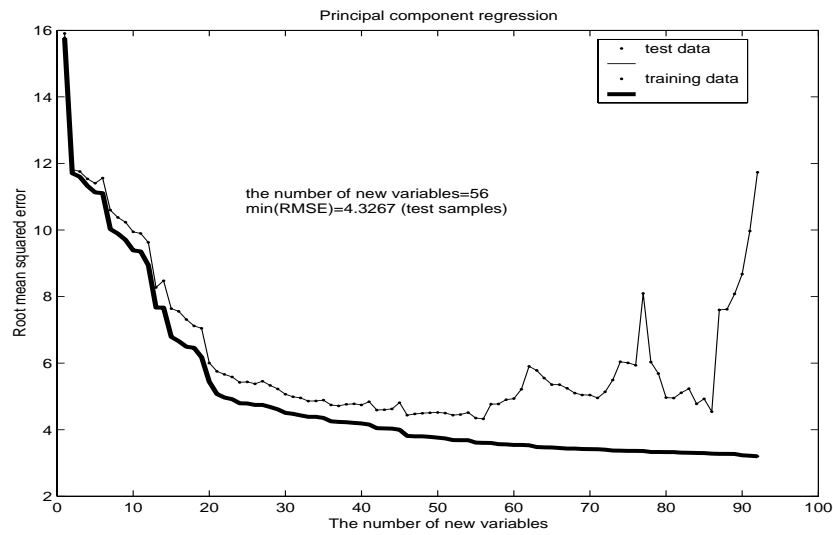


Figure 3: Test and training error of PCR through cross-validation

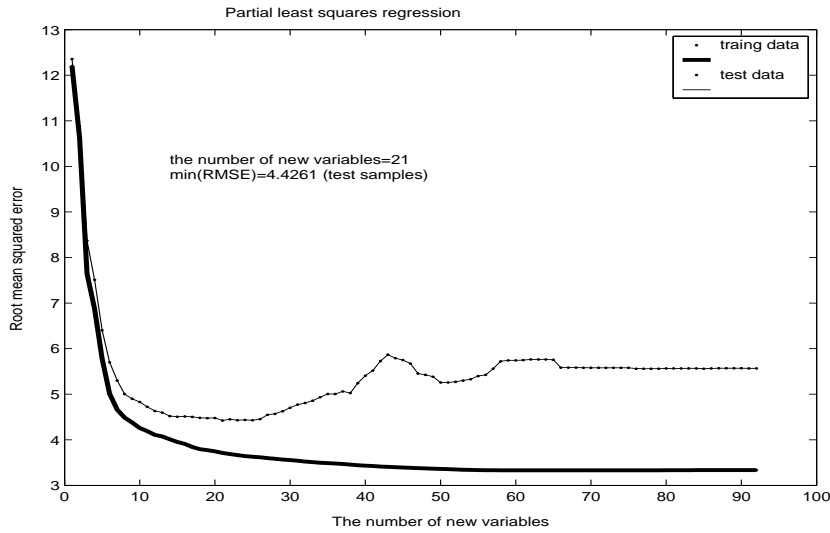


Figure 4: Test and training error of PLSR through cross-validation

at $M = 56$. Now let us look into the Figure 4, which presents the result of PLSR. The minimal test error of PLSR is 4.4261 at $M = 21$. When M exceeds 65, the test error levels off. These three figures have the common ground that the track of RMSE of test samples follows the track of RMSE of training samples originally and at the some point, it begins to departure the other. This phenomena is due to bias-variance tradeoff we discussed in Subsection 1.1.

For convenience, we also list RMSE of the same data in Table 2 and Table 3 presents comparisons of the three dimension reduction methods.

Table 2: The previous work

Topological Indices	RMSE(training samples)
χ	7.97
$W \chi$	7.53
$W \chi ap^4$	6.63
$W 1Z mwc_4 n^{0.5}$	6.19
$W 1Z 1twc ap^4 n$	5.75
$W 1Z 1twc ap^4 dia n^{0.5}$	5.64

Table 3: The dimension reduction results

methods	number of variables	RMSE(training)	RMSE(test)
SIR	57	3.5907	4.3625
PCR	56	3.6037	4.3267
PLSR	21	3.7102	4.4261

The results of Table 2 come from Rücher and Rücher (1999). These are the best results at present using linear regression models. Table 2 includes only RMSE of training samples and Table 3 gives RMSE of test samples through leave-one-out cross-validation. Usually test errors are larger than training errors. So there is a comparability between our results and the previous work. From the two tables, it shows dimension reduction methods we used are better.

Table 3 gives relative results of SIR, PCA and PLSR. The three results come from the same model structures: $\mathbf{y} = f(\mathbf{X}\alpha_1, \mathbf{X}\alpha_2, \dots, \mathbf{X}\alpha_m, \epsilon)$, here f is assumed a linear function. The difference among them is the technique which seeks direction $\{\alpha_i\}_{i=1}^m$. As we all know, in principal components regression, the m -th principal component direction α_m solves:

$$\max(\text{Var}(\mathbf{X}\alpha)), (\|\alpha\| = 1, \text{Corr}(\mathbf{X}\alpha_i, \mathbf{X}\alpha) = 0, i = 1, 2, \dots, m - 1).$$

The condition $\text{Corr}(\mathbf{X}\alpha_i, \mathbf{X}\alpha) = 0$ ensures that $\mathbf{z}_m = \mathbf{X}\alpha$ is uncorrelated with all the previous linear combinations $\mathbf{z}_i = \mathbf{X}\alpha_i, i = 1, 2, \dots, m - 1$. What optimization problem is the partial least squares solving? It can be shown that the partial least squares method seeks directions that have high variance and have high correlation with the response, in contrast to principal components regression (Stone and Brooks 1990 and Frank and Friedman 1993). That is to say, the m -th PLS direction α_m solves:

$$\max(\text{Corr}^2(\mathbf{y}, \mathbf{X}\alpha)\text{Var}(\mathbf{X}\alpha)), (\|\alpha\| = 1,$$

where $\text{Corr}(\mathbf{X}\alpha_i, \mathbf{X}\alpha) = 0, i = 1, 2, \dots, m - 1$).

Unlike PCR, the $\{\alpha_m\}_{i=1}^m$ PLS finds are not orthogonal owing to the different criterion. For the sliced inverse regression, reversing the role of \mathbf{y} and \mathbf{X} , the similar process is to find a variable (derivable from $\mathbf{x}_{(j)}$ linearly)

which is most predictable from \mathbf{y} . It is easily proved that the best prediction is $E(\mathbf{X}\alpha|\mathbf{y})$, a nonlinear function of \mathbf{y} in general. Thus the most predictable variable is the one which maximizes:

$$\frac{Var(E(\mathbf{X}\alpha|\mathbf{y}))}{Var(\mathbf{X}\alpha)}, (\|\alpha\| = 1, Corr(\mathbf{X}\alpha_i, \mathbf{X}\alpha) = 0, i = 1, 2, \dots, m - 1).$$

From Table 3, we find that the results of SIR's and PCR's are very similar. According to theoretical analysis, we can conclude that the results of SIR's may be better than that of PCR's because SIR considers the information of the response when it constructs new variables. But this superiority of SIR over PCR seems not to represent obviously to its performance. There are two reasons. First, this chemical data has an important characteristic different from other data that most of its topological indices are high collinear with its boiling points. That is to say, the design matrix has already included most information of the response. Under this circumstance, SIR and PCR just have the same ability in dimension reduction. Second, all results come from linear models. If we employ generalized additive model or other nonlinear model, the results of SIR should be prior to that of PCR.

Now let's compare the results of PCR's and PLSR's. PLSR use fewer components to achieve about the same result as PCR, generally about half as many components. This property has been empirically observed for some time and is often considered as an argument in favor of the superiority of PLSR over PCR. So we can fit the data to the same degree of closeness with fewer components, thus producing more parsimonious modes. The superiority of PLSR over PCR can be explained from their different criteria. PCR only uses the input variables information to determine its components, whereas PLSR uses the response values as well. Obviously, the response variable information contributes to the construction of the new input. So PLSR can fit the training data and predict the test data to a higher degree of accuracy than PCR with the same number of components. In conclusion, PLSR is a better satisfied dimension reduction method applied to high-dimensional chemical data, we will make a further study using more general model based on the directions derived by PLSR technique.

Acknowledgement

The work was partially supported by the Hong Kong UGC grant RGC/HKBU 2044/02P and Statistics Research and Consultancy Centre, Hong Kong Baptist University. The authors thank Prof. Kai-Tai Fang and Prof. Min-Te Chao for their valuable comments.

References

- Balaban, A. T. (1982). Highly discriminating distance-based topological index. *Chem. Phys. Lett.*, **89**, 399-404.
- Balaban, A. T. (1983). Topological indices based on topological distances in molecular graphs. *Pure Appl. Chem.*, **55**, 199-206.
- Balaban, A. T. and Balaban, T. S. (1991). New vertex invariants and topological indices of chemical graphs based on information on distances. *J. Math. Chem.*, **8**, 383-397.
- Balaban, A. T. and Balaban, T. S. (1992). Correlations using topological indices based on real graph invariants. *J. Chem. Phys.*, **89**, 1735-1745.
- Balaban, T. S., Filip, P. A. and Ivanciuc, D. (1992). Computer generation of acyclic graphs based on local vertex invariants and topological indices. Derived canonical labelling and coding of trees and alkane. *J. Math. Chem.*, **11**, 79-105.
- Bonchev, D. and Trinajstić, N. (1977). Information theory, distance matrix, and molecular branching. *J. Chem. Phys.*, **67**, 4517-4533.
- Chen, C. H. and Li, K. C. (1998). *Model Selection and Inference: A Practical Information-Theoretic Approach*. New York: Springer.
- Devillers, J. and Balaban, A. T. (1999). *Topological indices and related descriptors in QSAR and QSPR*. Gordon and Breach Science Publishers.
- Duan, N. H. and Li, K. C. (1991). Slicing regression: a link-free regression method. *The Annals of Statistics.*, **19**, 505-530.
- Filip, P. A., Balaban, T. S. and Balaban, A. T. (1987). A new approach for devising local graph invariants: derived topological indices with low degeneracy and good correlational ability. *J. Math. Chem.*, **1**, 61-83.

- Frank, I. and Friedman, J. (1993). A statistical view of some chemometrics regression tools (with discussion). *Technometrics.*, **35**(2), 109-148.
- Gutman, I. ruscic, B., Trinajstic, N. and Wilcox, C. F. (1975). Graph theory and molecular orbital. XII. Acyclic polyenes. *J. Chem. Phys.*, **62**, 3399-3409.
- Hosoya, H. (1971). A newly proposed quantity characterizing topological nature of structural isomers of saturated hydrocarbons. *Bull. Chem. Soc. Japan.*, **44**, 2332-2339.
- Ivanciuc, O. and Balaban, A. T. (1999). The graph description of chemical structures. Design on topological indices. Part 20. Molecular structure descriptors computed with information on distances operators. *Rev. Roum. Chim.*, **44**.
- Kier, L. B. (1985). A shape index from chemical graphs, *Quant. Struct.-Act. Relat.*, **4**, 109-116.
- Kier, L. B. (1986). Shape indexes of orders one and three from molecular graphs. *Quant. Struct.-Act. Relat.*, **5**, 1-7.
- Kier, L. B. and Hall, L. H. (1976). *Molecular Connectivity in Chemistry and Drug Research*. Academic Press, New York.
- Khadikar, P. V., Deshpande, N. V., Kale, P. P., Dobrynin, A. A. and Gutman, I. The szeged index and an analogy with the Wiener index. *J. Chem. Inf. Comput. Sci.*, (1995). **35**, 547-550.
- Klein, D. J., Lukovits, I. and Gutman, I. (1995). On the definition of the Hyper-Wiener index for cycle-containing structures. *J. Chem. Inf. Comp. Sci.*, **35**, 50-52.
- Li, K. C. (1991). Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association.*, **86**, 316-327.
- Liu, S. S., Cao, C. Z. and Li, Z. L. (1998). Approach to estimation and prediction for normal boiling point (NBP) of alkanes based on a novel molecular distance-edge (MDE) vector. *J. Chem. Inf. Chem. Soc.*, **38**, 387-394.
- Liu, S. S., Yin, C. S., Li, Z. L. and Cai, S. X. (2001). QSAR study of steroid benchmark and dipeptides based on MEDV-13. *J. Chem. Inf. Comp. Sci.*, **41**, 321-329.

- Lovasz, L. and Pelikanm, J. (1973). On the eigenvalues of trees. *Period. Math. Hung.*, **3**, 175-182.
- Lukovits, I. (1996). The Detour index. *Croat. Chem. Acta.*, **69**, 873-883
- Lukovits, I. (1998). An all-path version of the Wiener index. *J. Chem. Inf. Comput. Sci.*, **38**, 125-129.
- Moreau, G. and Broto, P. (1980a). Autocorrelation of a topological structure: a new molecular descriptor. *Nouv. J. Chim.*, **4**, 359-360.
- Moreau, G. and Broto, P. (1980b). Autocorrelation of molecular structures, application to QSAR studies. *Nouv. J. Chim.*, **4**, 757-764.
- Mohar, B., Babic, D. and Trinajstic, N. (1993). A novel definition of the Wiener index for trees, *J. Chem. Inf. Comput. Sci.*, **33**, 153-154
- Plavsic, D., Nikoplic, S., Trinajstic, N. and Mihalic, Z. (1993). On the Harary index for the characterization of chemical graphs. *J. Math. Chem.*, **12**, 235-250.
- Randic, M. (1975). On characterization of molecular branching. *J. Amer. Chem. Soc.*, **97**, 6609-6013.
- Randic, M. (1997). On canonical numbering of atoms in a molecule and graph isomorphism. *J. Chem. Inf. Comput. Sci.*, **17**, 171-180.
- Razinger, M. (1997). On calculation of the Detour index. *J. Chem. Inf. Comp. Sci.*, **37**, 283-286.
- Rücker, G. and Rücker, C. (1999). On topological indices, boiling points, and cycloalkanes. *J. Chem. Inf. Compt. Sci.*, **39**, 788-802.
- Schultz, H. P. (1989). *J. Chem. Inf. Comp. Sci.*, **29**, 227-228.
- Stone, M. and Brooks, R. J. (1990). Continuum regression: cross-validated sequentially constructed prediction embracing ordinary least squares, partial least squares and principal components regression (Corr: V54 p906-907). *Journal of the Royal Statistical Society, Series B, Methodological.*, **52**, 237-269.
- Wiener, H. (1947). Structural determination of paraffin boiling points. *J. Am. Chem. Soc.*, **69**, 17-20.

Wold, H. (1975). Soft modelling by latent variables: the nonlinear iterative partial least squares (NIPALS) approach. *Perspectives in Probability and Statistics.*, In Honor of M. S. Bartlett, pp. 117-144.

Zhu, L. X. and Fang, K. T. (1996), Asymptotics for kernel estimate of sliced inverse regression, *Annals Statistics.*, **24**, 1053-1068.

Received May 20, 2003; accepted August 2, 2003

Hong Yin
Department of Mathematics
Hong Kong Baptist University
Hong Kong, P. R. China
01400924@hkbu.edu.hk

Yi-Zeng Liang
College of Chemistry and Chemical Engineering
Central South University
Changsha, P. R. China
yzliang@cs.hn.cn

Qin-Nan Hu
College of Chemistry and Chemical Engineering
Central South University
Changsha, P. R. China