



Dimension reduction techniques for classification

Milan, May 2003

Anestis Antoniadis

Laboratoire IMAG-LMC
University Joseph Fourier
Grenoble, France

Outline

- Why using dimension reduction methods with microarray data?

Outline

- **Why using dimension reduction methods with microarray data?**
- thousands of variables (genes, p) and a very small number of (biological) replicates, so regression analysis is difficult in practice.

Outline

- **Why using dimension reduction methods with microarray data?**
- thousands of variables (genes, p) and a very small number of (biological) replicates, so regression analysis is difficult in practice.
- **Dimension reduction for regression**

Outline

- **Why using dimension reduction methods with microarray data?**
- thousands of variables (genes, p) and a very small number of (biological) replicates, so regression analysis is difficult in practice.
- **Dimension reduction for regression**
- PCA (SVD) based methods

Outline

- **Why using dimension reduction methods with microarray data?**
- thousands of variables (genes, p) and a very small number of (biological) replicates, so regression analysis is difficult in practice.
- **Dimension reduction for regression**
- PCA (SVD) based methods
- PLS methods (from chemometrics)

Outline

- **Why using dimension reduction methods with microarray data?**
- thousands of variables (genes, p) and a very small number of (biological) replicates, so regression analysis is difficult in practice.
- **Dimension reduction for regression**
- PCA (SVD) based methods
- PLS methods (from chemometrics)
- Sliced Inverse Regression methods (SIR, SAVE, MAVE)

Outline

- **Why using dimension reduction methods with microarray data?**
- thousands of variables (genes, p) and a very small number of (biological) replicates, so regression analysis is difficult in practice.
- **Dimension reduction for regression**
- PCA (SVD) based methods
- PLS methods (from chemometrics)
- Sliced Inverse Regression methods (SIR, SAVE, MAVE)
- **Applications in classification**

Notation

Microarray data. The set of observed data arrives in two parts:

- a $n \times p$ matrix $\mathbf{X} = (x_{ij})$, where, typically, each row corresponds to a gene ($1 \times p$) expression profile for an individual or subject (i) corresponding to the i th row of \mathbf{X} , so the microarray corresponds to \mathbf{X}^T . We will suppose that the data have already been preprocessed and normalized.
- an n -dimensional vector \mathbf{Y} of group labels, taking values in $\{0, \dots, G - 1\}$.

We will assume that the rows of \mathbf{X} are standardized to have mean zero and variance one for each column (gene).

Two type of problems

Class discovery and class prediction.

We will focus here on the class prediction problem: observations are known to belong to a prespecified class and the task is to build predictors for assigning new observations to these classes.

Standard methods include linear discriminant analysis, diagonal linear discriminant classifiers, classification trees, nearest neighbor (NN), SVM and aggregating classifiers.

Comprehensive account in Dudoit, Fridlyant and Speed (2002).

A regression model based on SVD

Formulate the effects of gene expression on class type using the *multinomial logistic regression model*:

$$\log \frac{\mathbb{P}(Y_i = r)}{\mathbb{P}(Y_i = 0)} = \mathbf{X}_i \cdot \boldsymbol{\beta}_{r0}, \quad r = 1, \dots, G - 1,$$

where $\boldsymbol{\beta}_{r0}$ is a p -dimensional vector of unknown regression coefficients.

Since $p \gg n$ it is not possible to estimate the parameters of the above model using standard statistical methods. A principal component study becomes then a suitable first step to reduce the dimension of $\boldsymbol{\beta}_{r0}$.

Principal Component Regression

We first perform a singular value decomposition of the $p \times n$ matrix \mathbf{X}^T :

$$\mathbf{X}^T = \mathbf{U}\mathbf{D}\mathbf{V},$$

where

- \mathbf{U} is a $p \times n$ matrix whose columns are orthonormal.
- \mathbf{D} is the diagonal matrix containing the ordered singular values d_i of \mathbf{X} . We will assume that $d_i > 0$ for $i = 1, \dots, n$.
- \mathbf{V} is the $n \times n$ singular value decomposition factor matrix and has both orthonormal rows and columns.

PCR


Aim: project the high-dimensional multivariate data into a lower dimensional subspace.

By SVD we now have

$$\log \frac{\mathbb{P}(Y_i = r)}{\mathbb{P}(Y_i = 0)} = \mathbf{W}_{i \cdot} \boldsymbol{\gamma}_{r0},$$

where $\mathbf{W}_{i \cdot}$ is the i th row of $\mathbf{W} = \mathbf{D}\mathbf{V}$ and $\boldsymbol{\gamma}_{r0} = \mathbf{U}^T \boldsymbol{\beta}_{r0}$ is now an n -dimensional vector of regression coefficients.

We have therefore reduced the dimension of space for the predictor variables from p to n , thus making the problem computationally tractable, fitting the model by maximum likelihood methods.



The number of potential predictor genes is only a fraction of the set of genes that have real biological activity.

Suggestion: reduce the initial number p of genes by ANOVA like based prefiltering techniques for improving predictive performance (see e.g. Golub et al. (1999), Dutoit et al. (2002), Nguyen and Rocke (2002)).

Idea: fit an ANOVA model for gene expression versus class for each gene. For each ANOVA model, compute an overall F -statistic and take the m genes with the largest F -statistic as the potential predictors in the model.

Choosing the components

A major issue is determining how many components to retain. One way of performing this is *leave-one-out cross validation*.

One sample is removed from the data set at a time. For a fixed number of components, say k , the regression model is fit to the remaining data. Based on the estimated model, the fit is used to predict the withheld sample. An error measure is then computed and the procedure is repeated to get an estimate of the prediction classification error. This is done for each value of k and the value of k that yields the smallest classification error is then chosen.

Caution

Performance of the classification rules for a selected subset of genes is measured by their errors on the test set and also by their leave-one-out cross validated errors.

If these errors are calculated within the gene preliminary selection process, there is a selection bias in them when they are used as an estimate of the prediction error.

Partial Least Squares

SVD produces orthogonal class descriptors that reduce the high dimensional data (supergenes).

This is achieved without regards to the response variation and may be inefficient. This way of reducing the regressor dimensionality is totally independent of the output variable.

One must not treat the predictors separately from the response. This is the spirit of the methods developed by Nguyen and Rocke (2002) and Gosh (2002), where the PLS components are chosen so that the sample covariance between the response and a linear combination of the p predictors (genes) is maximum.

Partial Least squares

- Popular regression method in chemometrics. It attempts to simultaneously find linear combinations of the predictors whose correlation is maximized with the response and which are uncorrelated over the training sample.

Partial Least squares

- Popular regression method in chemometrics. It attempts to simultaneously find linear combinations of the predictors whose correlation is maximized with the response and which are uncorrelated over the training sample.
- Several algorithms for numerical fitting partial least squares models.

Partial Least squares

- Popular regression method in chemometrics. It attempts to simultaneously find linear combinations of the predictors whose correlation is maximized with the response and which are uncorrelated over the training sample.
- Several algorithms for numerical fitting partial least squares models.
- However, PLS is really designed to handle continuous responses and especially for models that do not really suffer from conditional heteroscedasticity as it is the case for binary or multinomial data.

Sample PLS

Given \mathbf{X} and $\mathbf{Y} \in \mathbb{R}^n$, an important feature of PLS is that the following two decompositions are carried out together:

$$\mathbf{E}_0 = \mathbf{X} = \sum_{j=1}^K \mathbf{t}_j \mathbf{p}_j^T + \mathbf{E}_K$$

and

$$\mathbf{f}_0 = \mathbf{y}^* = \sum_{j=1}^K q_j \mathbf{t}_j + \mathbf{f}_K,$$

where the \mathbf{t}_j are n -vector latent variables (**scores**), \mathbf{p}_j are the p -vector **loadings**, and \mathbf{E}_K is a residual matrix. The q_j are scalar coefficients and \mathbf{f}_K is an n vector of residuals.

Sample PLS

Denoting by \mathbf{T} the $n \times K$ matrix of scores and by \mathbf{P} the $K \times p$ matrix of loadings whose rows are the \mathbf{p}_j^T , one has:

$$\mathbf{X} = \mathbf{TP} + \mathbf{E}_K.$$

One may see that the above decomposition is not unique since for any invertible $K \times K$ matrix \mathbf{C} we have $(\mathbf{TC})(\mathbf{C}^{-1}\mathbf{P})$. The uniqueness of the \mathbf{t}_j 's and \mathbf{p}_j 's comes from imposing conditions of orthogonality, i.e. $\mathbf{PP}^T = \mathbf{D}_P$ and $\mathbf{T}^T\mathbf{T} = \mathbf{D}_T$.

Pseudo code

1. Set $\mathbf{E}_0 = \mathbf{X}$ and $\mathbf{f}_0 = \mathbf{y}$. Compute $\mathbf{p}_1 = \operatorname{argmax}\{\mathbf{p}, \|\mathbf{p}\| = 1, |\langle \mathbf{E}_0 \mathbf{p}, \mathbf{y} \rangle|\}$ (answer $\mathbf{p}_1 = \mathbf{E}_0^T \mathbf{y} / \|\mathbf{E}_0^T \mathbf{y}\|$).
2. $\mathbf{t}_1 = \mathbf{E}_0 \mathbf{p}_1 / \|\mathbf{E}_0 \mathbf{p}_1\|$
3. $\mathbf{X} = \mathbf{t}_1 (\mathbf{X}^T \mathbf{t}_1)^T + \mathbf{E}_1$
4. $\mathbf{y} = \mathbf{t}_1 (\mathbf{y}^T \mathbf{t}_1)^T + \mathbf{f}_1$
5. repeat the above until K . One may show that $K \leq \operatorname{rank}(\mathbf{X})$.

Denote

$$\mathbf{w}_k = \mathbf{X}^T \mathbf{f}_{k-1} / \|\mathbf{X}^T \mathbf{f}_{k-1}\|$$

and let \mathbf{W}_K the $p \times K$ matrix whose columns are the \mathbf{w}_k 's. The fitted PLS regression is then given by

$$\hat{\mathbf{Y}} = \mathbf{XW}_K(\mathbf{W}_K^T \mathbf{X}^T \mathbf{XW}_K)^{-1} \mathbf{W}_K^T \mathbf{X}^T \mathbf{Y}$$

and

$$\hat{\boldsymbol{\beta}}_{PLS} = \mathbf{W}_K(\mathbf{W}_K^T \mathbf{X}^T \mathbf{XW}_K)^{-1} \mathbf{W}_K^T \mathbf{X}^T \mathbf{Y}$$

In contrast to PCR, PLS procedures are nonlinear. Again many ways in choosing K .

Another look

PLS Algorithm (Initialize $\mathbf{E}_0 = \mathbf{X}$ and $\mathbf{f}_0 = \mathbf{y}$)

1. Iterate until $\Delta\hat{\mathbf{y}}$ is small

(a) For $k = 1$ to K

i. $\hat{\omega}_k < -\mathbf{E}_{k-1}^T \mathbf{f}_{k-1} / \|\mathbf{E}_{k-1}^T \mathbf{f}_{k-1}\|_n$

ii. $\mathbf{t}^k < -\mathbf{E}_{k-1} \hat{\omega}_k$


iii. $\hat{q}_k < -$ coefficient lsfit (\mathbf{f}_{k-1} on \mathbf{t}^k with no intercept)

iv. $\mathbf{f}_k < -\mathbf{f}_{k-1} - \mathbf{t}^k \hat{q}_k$

v. $\mathbf{p}_k < -$ coefficient lsfit (\mathbf{E}_{k-1} on \mathbf{t}^k with no intercept)

vi. $\mathbf{E}_k < -$ residual lsfit (\mathbf{E}_{k-1} on \mathbf{t}^k with no intercept)

(b) end For



(c) $\hat{y} < -\text{mean}(\mathbf{f}_0) + \sum_{k=1}^K \hat{q}_k \mathbf{t}_k$

3. Choose $s \leq K$

4. $\text{lm}(y \sim \mathbf{t}_1 \cdots \mathbf{t}_s)$

However, PLS is really designed to handle continuous responses and especially for models that do not really suffer from conditional heteroscedasticity as it is the case for binary or multinomial data.

GLM

The purpose here is to shortly introduce some relevant facts on generalized linear models. For a more thorough description of such models please refer to McCullagh and Nelder (1989) or Fahrmeir and Tutz (1994).

Consider a pair (X, Y) of random variables, where Y is real-valued and \mathbf{X} is possibly real vector-valued; here Y is referred to as a response or dependent variable and \mathbf{X} as the vector of covariates or predictor variables.

Generalized models (GM for short) are particular regression models which describe the dependence of the response variable of interest Y on one or more predictor variables \mathbf{X} .

A basic generalized model analysis starts with a random sample of size n from the distribution of (\mathbf{X}, Y) where the conditional distribution of Y given that $\mathbf{X} = \mathbf{x}$ is assumed to be from a one-parameter exponential family distribution with a density of the form

$$\exp \left(\frac{y\theta(\mathbf{x}) - b(\theta(\mathbf{x}))}{\phi} + c(y, \phi) \right)$$

The natural parameter function $\theta(x)$ specifies how the response depends on the covariates.

The conditional mean and variance of the i th response Y_i are given by

$$\mathbb{E}(Y_i / X = x_i) = \dot{b}(\theta(x_i)) = \mu(x_i)$$

and

$$\text{Var}(Y_i / X = x_i) = \phi \ddot{b}(\theta(x_i))$$

Here a dot denotes differentiation.

In the usual GM framework, the mean is related to the GM regression surface via the link function transformation $g(\mu(x_i)) = \eta(x_i)$ where $\eta(x)$ is referred as the predictor function.

A wide variety of distributions can be modelled using this approach including normal regression with additive normal errors (identity link), logistic regression (logit link) where the response is a binomial variable and Poisson regression models (log link) where the observations are from a Poisson distribution.

MLE algorithm

Consider the case

$$\eta = \mathbf{X}\beta$$

with a monotone and continuously differentiable link function g such that $\eta_i = g(\mu_i)$ (the canonical link if $g(\mu_i) = \theta_i$).

One may show that the ML estimators of the parameters may be obtained by an iterative reweighted least squares procedure, as follows:

Let $\hat{\eta}_0$ the current estimate of the linear predictor, with a corresponding value of the mean $\hat{\mu}_0$ by means of the link g . One forms the adjusted response

$$\mathbf{z}_0 = \hat{\eta}_0 + (\mathbf{y} - \hat{\mu}_0) \left(\frac{d\eta}{d\mu} \right)_0'$$

the derivative of the link being computed at $\hat{\mu}_0$. The weight matrix \mathbf{W} is evaluated at $\hat{\mu}_0$. One then regresses by LS the vector \mathbf{z}_0 on \mathbf{X} with weights \mathbf{W} to get $\hat{\beta}_1$ and the procedure is repeated until convergence.

PLS for GLMs

A close look at the PLS algorithm shows that the adjusted dependent vector residuals, \mathbf{f}_{k-1} (in step $k - 1$), are regressed on the explanatory variable residuals, \mathbf{E}_{k-1} . Next, the adjusted dependent vector residuals (in step $k - 1$) are regressed on the current latent variable. The result of this fitted value is then subtracted from the residuals to form the next sequence of adjusted dependent vector residuals.

Borrowing the equivalence between MLE and IRLS, Marx (1996) treats the iterated adjusted dependent vector as the current dependent variable, in a weighted metric which allows him to express an IRPLS algorithm as the analog of the familiar PLS algorithm.

Pseudocode

Initialize $\mathbf{E}_0 = \mathbf{X}$, $\mathbf{f}_0 = \psi(\mathbf{y})$, $\hat{V} = g'(\psi(\mathbf{y}))\}^2 / (\text{Var}(\mathbf{Y}))$

1. Iterate until $\Delta\hat{\eta}$ is small

(a) For $k = 1$ to K

- i. $\hat{w}_k = \mathbf{E}_{k-1}^T \hat{V} \mathbf{f}_{k-1} / \|\mathbf{E}_{k-1}^T \hat{V} \mathbf{f}_{k-1}\|_n$
- ii. $\mathbf{t}^k = \mathbf{E}_{k-1} \hat{w}_k$
- iii. $\hat{q}_k =$ coefficient lsfit (\mathbf{f}_{k-1} on \mathbf{t}^k with no intercept and weight \hat{V})
- iv. $\mathbf{f}_k = \mathbf{f}_{k-1} - \mathbf{t}^k \hat{q}_k$
- v. $\mathbf{p}_k =$ coefficient lsfit (\mathbf{E}_{k-1} on \mathbf{t}^k with no intercept and weight \hat{V})
- vi. $\mathbf{E}_k =$ residual lsfit (\mathbf{E}_{k-1} on \mathbf{t}^k with no intercept and weight \hat{V})

(b) end For

$$(c) \hat{\boldsymbol{\eta}} < - \text{wt.mean}(\mathbf{f}_0, \text{wt} = \hat{V}) + \sum_{k=1}^K \hat{q}_k \mathbf{t}_k$$

$$(d) \hat{V} = \{g'(\hat{\boldsymbol{\eta}})\}^2 / \text{var}(\mathbf{Y})$$

$$(e) \mathbf{f}_0 = \hat{\boldsymbol{\eta}} + \text{diag}\{1/g'(\hat{\eta}_i)\}(\mathbf{y} - g(\hat{\boldsymbol{\eta}}))$$

(f) Compute the adjusted residuals E_0

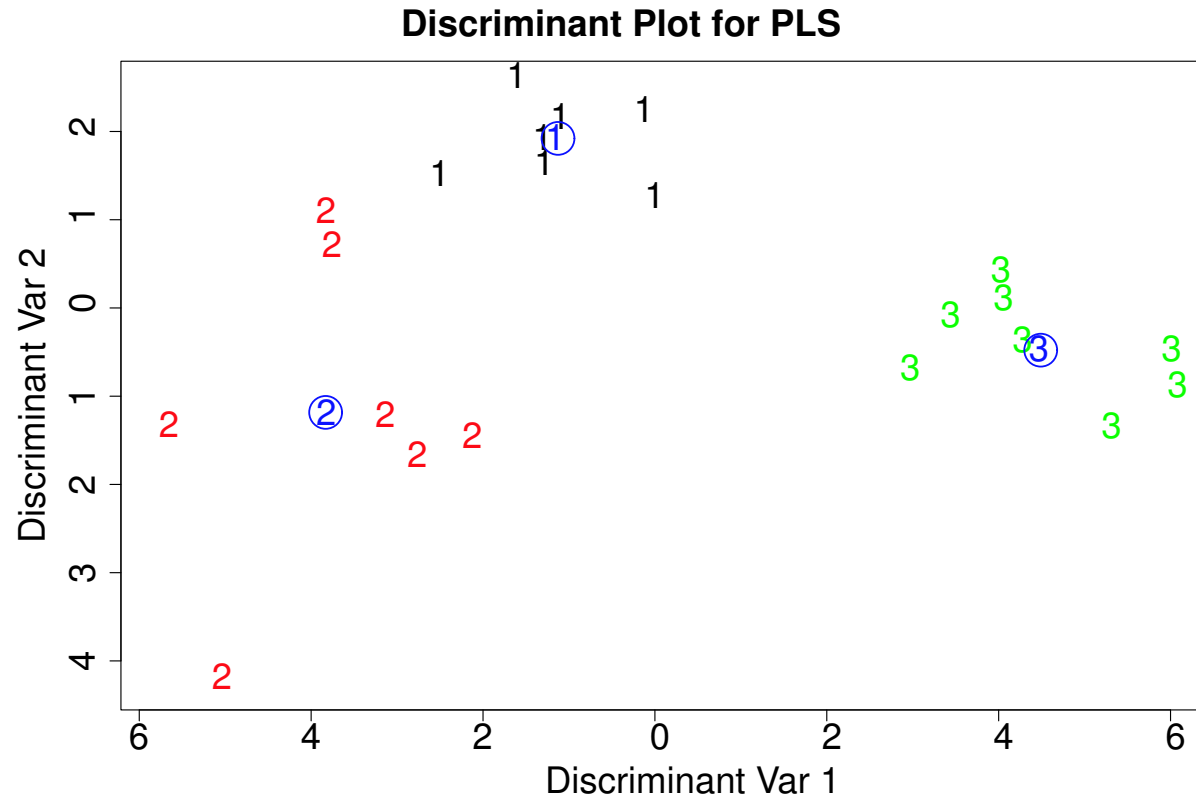
3. Choose $s \leq K$

4. $\text{glm}(y \sim \mathbf{t}_1 \cdots \mathbf{t}_s)$

An example

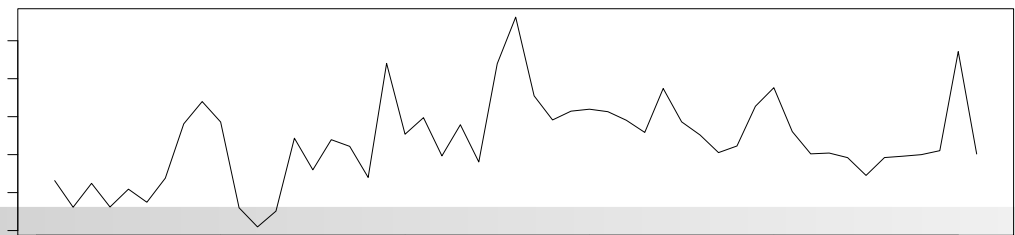
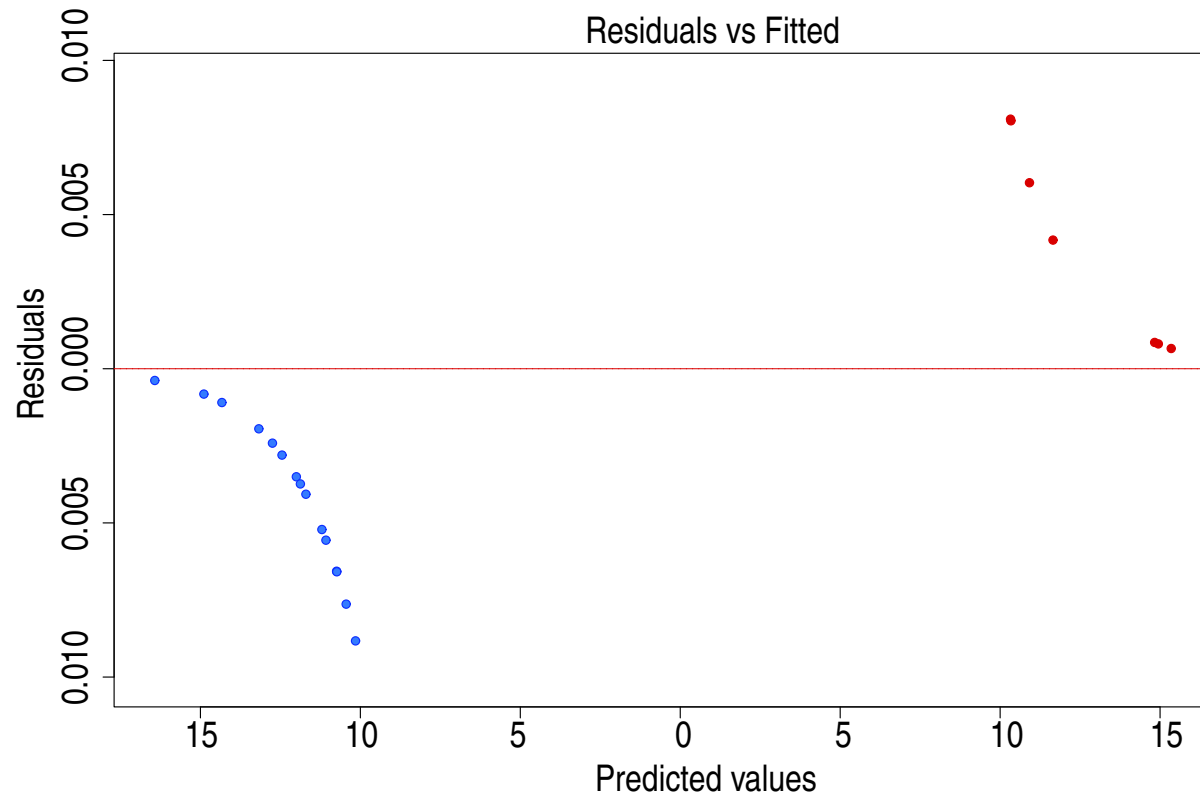
The data consist of 22 cDNA microarrays, each representing 5361 genes based on biopsy specimens of primary breast tumours of 7 patients with germ-line mutations of BRCA1, 8 patients with germline mutations of BRCA2, and 7 with sporadic cases. These data were first presented and analysed by Hedenfalk et al. (2001). Information on the data can be found in <http://www.nejm.org> and <http://www.nhgri.nih.gov/DIR/Microarray>. The analysis focuses on identifying groups of genes that can be used to predict class membership to the two BRCA mutation

PLS regression



Applying (classical) PLS to the BRCA data.

GPLS regression



Sliced inverse regression (SIR)

A convenient data reduction formulation, that accounts for the correlation among genes, is to assume there exists a $p \times k$, $k \leq p$, matrix η so that

$$F(\mathbf{Y}|\mathbf{X}) = F(\mathbf{Y}|\eta^T \mathbf{X})$$

where $F(\cdot|\cdot)$ is the conditional distribution function of the response \mathbf{Y} given the second argument.

The above statement that the $p \times 1$ predictor vector \mathbf{X} can be replaced by the $k \times 1$ predictor vector $\eta^T \mathbf{X}$ without loss of information.

Most importantly, if $k < p$, then sufficient reduction in the dimension of the regression is achieved. The linear subspace $S(\boldsymbol{\eta})$ spanned by the columns of $\boldsymbol{\eta}$ is a dimension-reduction subspace (Li, 1991) and its dimension denotes the number of linear combinations of the components of \mathbf{X} needed to model Y .

Let $S_{Y|X}$ denote the unique smallest dimension reduction subspace, referred to as the central subspace (Cook, 1996). The dimension $d = \dim(S_{Y|X})$ is called the structural dimension of the regression of Y on X , and can take on any value in the $\{0, 1, \dots, p\}$ set.

Estimation of the central subspace

The estimation of the central subspace is based on finding a kernel matrix M so that $S(M) \subset S_{Y|X}$.

- SIR and variations (Li, 1991) $M = Cov(E(X|Y))$,

Estimation of the central subspace

The estimation of the central subspace is based on finding a kernel matrix M so that $S(M) \subset S_{Y|X}$.

- SIR and variations (Li, 1991) $M = Cov(E(X|Y))$,
- polynomial inverse regression (Bura and Cook, 2001) $M = \mathbb{E}(X|Y)$.

Estimation of the central subspace

The estimation of the central subspace is based on finding a kernel matrix M so that $S(M) \subset S_{Y|X}$.

- SIR and variations (Li, 1991) $M = \text{Cov}(E(X|Y))$,
- polynomial inverse regression (Bura and Cook, 2001) $M = \mathbb{E}(X|Y)$.
- pHd (Li, 1991) $M = \mathbb{E}((Y - \mathbb{E}(Y))XX^T)$,

Estimation of the central subspace

The estimation of the central subspace is based on finding a kernel matrix M so that $S(M) \subset S_{Y|X}$.

- SIR and variations (Li, 1991) $M = \text{Cov}(E(X|Y))$,
- polynomial inverse regression (Bura and Cook, 2001) $M = \mathbb{E}(X|Y)$.
- pHd (Li, 1991) $M = \mathbb{E}((Y - \mathbb{E}(Y))XX^T)$,
- SAVE (Cook and Weisberg, 1991)
 $M = \mathbb{E}(\text{Cov}(X) - \text{Cov}(X|Y))^2$, and

Estimation of the central subspace

The estimation of the central subspace is based on finding a kernel matrix M so that $S(M) \subset S_{Y|X}$.

- SIR and variations (Li, 1991) $M = \text{Cov}(E(X|Y))$,
- polynomial inverse regression (Bura and Cook, 2001) $M = \mathbb{E}(X|Y)$.
- pHd (Li, 1991) $M = \mathbb{E}((Y - \mathbb{E}(Y))XX^T)$,
- SAVE (Cook and Weisberg, 1991)
 $M = \mathbb{E}(\text{Cov}(X) - \text{Cov}(X|Y))^2$, and
- SIRII (Li, 1991) with
 $M = \mathbb{E}(\text{Cov}(X|Y) - \mathbb{E}(\text{Cov}(X|Y)))^2$

The two conditions for all kernel matrices M to span subspaces of the central dimension reduction subspace are that $\mathbb{E}(X|\gamma^T X)$ be linear, and that $\text{Var}(X|\gamma^T X)$ be constant. The conditions are empirically checked by considering the scatterplot matrix of the predictors.

- Linearity of $\mathbb{E}(X|\gamma^T X)$ can be ascertained if the scatterplots look roughly linear or random,
- homogeneity of the variance holds if there are no pronounced fluctuations in data density in the scatterplots.

Without loss of generality we use standardised predictors $\mathbf{Z} = \Sigma_X^{-1/2}(\mathbf{X} - \mathbb{E}(\mathbf{X}))$. Let $\boldsymbol{\mu}_j = \mathbb{E}(\mathbf{Z}|Y = j)$ and $\Sigma_j = \text{Var}(\mathbf{Z}|Y = j)$, $j = 0, 1$, denote the conditional means and variances, respectively, for the binary response Y , and let

$$\boldsymbol{\nu} = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_0, \quad \Delta = \Sigma_1 - \Sigma_0$$

The main results by Cook and Lee (1999) state that $S_{SAVE} = S(\boldsymbol{\nu}, \Delta) \subset S_{Y|Z}$ and also showed that $S_{SIR} = S(\boldsymbol{\nu}) \subset S_{Y|Z}$.

Implementation

In implementing the method, ν and Δ are replaced by the corresponding sample moments,

$$\hat{\nu} = \hat{\Sigma}_{\mathbf{x}}^{-1/2}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_0)$$

and

$$\hat{\Delta} = \hat{\Sigma}_{\mathbf{x}}^{-1/2}(\hat{\Sigma}_{\mathbf{x}|1} - \hat{\Sigma}_{\mathbf{x}|0})\hat{\Sigma}_{\mathbf{x}}^{-1/2}$$

to yield $\hat{S}_{SAVE} = S(\hat{\nu}, \hat{\Delta})$, a $k \times (k + 1)$ matrix, and $\hat{S}_{SIR} = S(\hat{\nu})$, a $k \times 1$ vector. The latter has obviously dimension of at most 1.

Estimating d

The test statistic for dimension is given by

$\Lambda_d = n \sum_{\ell=d+1}^k \hat{\lambda}_\ell^2$, where $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \dots \geq \hat{\lambda}_k$ are the singular values of the estimated kernel matrix

$\hat{M}_{SAVE} = (\hat{\mathbf{v}}, \hat{\Delta})$, or $\hat{M}_{SIR} = \hat{\mathbf{v}}$, depending on the method used.

In both cases, the estimation is carried out by performing a series of tests for testing $H_0 : d = m$ against $H_a : d > m$, starting at $m = 0$, which corresponds to independence of Y and \mathbf{Z} .

The test statistic for SAVE has an asymptotic weighted chisquared distribution. The SIR test statistic for dimension has an asymptotic chi-squared distribution (Li, 1991).

Remarks

- SIR and SAVE can be applied to problems with multinomial or multi-valued responses.

Remarks

- SIR and SAVE can be applied to problems with multinomial or multi-valued responses.
- When X are normally distributed, SIR is equivalent to Linear Discriminant Analysis in the sense that they both estimate the same discriminant linear combinations of the predictors.

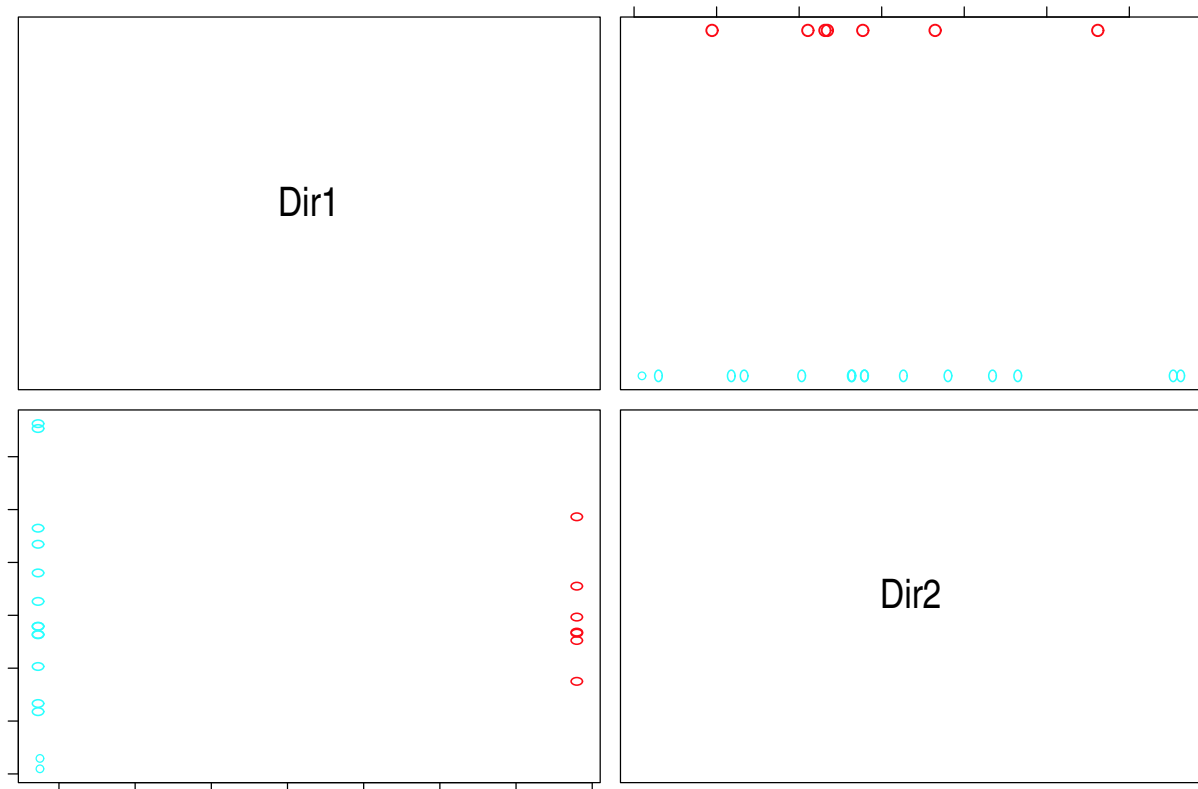
Remarks

- SIR and SAVE can be applied to problems with multinomial or multi-valued responses.
- When X are normally distributed, SIR is equivalent to Linear Discriminant Analysis in the sense that they both estimate the same discriminant linear combinations of the predictors.
- In binary regression both LDA and SIR estimate at most one direction in the central dimension reduction subspace.

Remarks

- SIR and SAVE can be applied to problems with multinomial or multi-valued responses.
- When X are normally distributed, SIR is equivalent to Linear Discriminant Analysis in the sense that they both estimate the same discriminant linear combinations of the predictors.
- In binary regression both LDA and SIR estimate at most one direction in the central dimension reduction subspace.
- When X is a normal vector, SAVE is equivalent to Quadratic Discriminant Analysis.

SIR example



Applying SIR to the BRCA data.

MAVE

A regression-type model for dimension reduction can be written as

$$Y = g(\boldsymbol{\eta}^T \mathbf{X}) + \epsilon$$

where g is an unknown smooth link function, $\boldsymbol{\eta} = (\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_D)$ is a $p \times k$ orthogonal matrix ($\boldsymbol{\eta}^T \boldsymbol{\eta} = I_k$) with $k < p$ and $\mathbb{E}(\epsilon | \mathbf{X}) = 0$ almost surely. The last condition allows ϵ to be dependent on \mathbf{X} and covers, in particular, the binary regression case.

The MAVE algorithm of Xia *et al.* (2002) is devoted to the estimation of the matrix $\boldsymbol{\eta}$ in $\mathbb{E}(Y|\mathbf{X}) = g(\boldsymbol{\eta}^T \mathbf{X})$ with g an unknown smooth function.

The estimated $\boldsymbol{\eta}$ is a solution to

$$\min_B \mathbb{E}\{Y - \mathbb{E}(Y|B^T \mathbf{X})\}^2 = \mathbb{E}(\sigma_B^2(B^T \mathbf{X})),$$

subject to $B^T B = I$. To minimize the above expression one has first to estimate the conditional variance


$\sigma_B^2(B^T \mathbf{X}) = \mathbb{E}[\{Y - \mathbb{E}(Y|B^T \mathbf{X})\}^2 | B^T \mathbf{X}]$. Let

$g_B(\mathbf{v}) = \mathbb{E}(Y | B^T \mathbf{X} = \mathbf{v})$.

Given a sample $\{\mathbf{X}_i, Y_i\}$ a local linear fit is applied to estimate $g_B(\cdot)$ and the EDR directions are estimated by solving the minimization problem

$$\min_{B, a_j, \mathbf{b}_j} \left(\sum_{j=1}^n \sum_{i=1}^n (Y_i - [a_j + \mathbf{b}_j^T B^T (X_i - X_j)])^2 w_{ij} \right),$$

where $w_{ij} = K_h\{\mathbf{B}^T(X_i - X_j)\} / \sum_{\ell=1}^n K_h\{\mathbf{B}^T(X_\ell - X_j)\}$ are multidimensional kernel weights.



We start with the identity matrix as an initial estimator of B to be used in the kernel weights. Then iteratively, we use the multidimensional kernel weights to obtain an estimator $\hat{\mathbf{B}}$ by minimization and refine the kernel weights with the updated value of B and iterate until convergence.

The choices of the bandwidth h and the EDR dimension d are implemented through a cross-validation technique.


Results

We demonstrate the usefulness of the proposed methodology described above on a well known data set: the leukemia data first analyzed in Golub et al. (1999). It consists of absolute measurements from Affymetrix high-density oligonucleotide arrays and contains $n = 72$ tissue samples on $p = 7,129$ genes (47 cases of acute lymphoblastic leukemia (ALL) and 25 cases of acute myeloid leukemia (AML)) .

We have used the MATLAB software environment for preprocessing the data and to implement the classification methodology. The suite of MATLAB functions implementing the procedure is freely available at the URL

<http://www-lmc.imag.fr/SMS/Software/microarrays/>.

We followed exactly the protocol in Dudoit et al. (2002) to pre-process the data by thresholding, filtering, a base 10 logarithmic transformation and standardization, so that the final data is summarized by a 3571×72 matrix $X = (x_{ij})$, where x_{ij} denotes the base 10 logarithm of the expression level for gene i in mRNA sample j .



The data are already divided into a learning set of 38 mRNA samples and a test set of 34 mRNA samples. The observations in the two sets came from different labs and were collected at different times.

When training the rule and for the pre-selection of genes, we first reduced the set of available genes to the top $p^* = 50, 100$ and 200 genes as ranked in terms of a BSS/WSS criterion and used by Dudoit et al. (2002).

To compare our results we have also applied the two discriminant analysis procedures DLDA et DQDA described in Dudoit et al. (2002)

Table 1: Classification rates by the four methods for the leukemia data set with 38 training samples (27 ALL, 11 AML) and 34 test samples (20 ALL, 14 AML). Given are the number of correct classification out of 38 and 34 for the training and test samples respectively.

p^*	Training Data (Leave-out-one CV)			
	MAVE-LD	DLDA	DQDA	MAVE-NPLD
50	37	38	37	37
100	38	38	37	38
200	38	38	36	38


Test Data (Out-of-sample)					
p^*	MAVE-LD		DLDA	DQDA	MAVE-NPLD
50	33		33	33	33
100	33		33	33	33
200	32		33	32	32

References

Bura, E. and Cook, R. D. (2001a). Estimating the structural dimension of regressions via parametric inverse regression. *J. Roy. Statist. Soc. Ser. B*, 63, 393–410.

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


Cook, R. D. and Lee, H. (1999). Dimension reduction in binary response regression. *J. Amer. Statist. Assoc.*, 94, 1187–1200.



Dudoit, S., Fridlyand, J. and Speed, T. P. (2002). Comparison of discrimination methods for the classification of tumours using gene expression data. *J. Amer. Statist. Assoc.*, 97, 77–87.

Golub, T. R., Slonim, D. K., Tamayo, P. et al. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, 286 531–537.

Li, K. C. (1991). Sliced inverse regression for dimension reduction. *J. Am. Statist. Assoc.*, 86, 316–342.



B. D. Marx (1996), Iteratively Reweighted Partial Least Squares Estimation for Generalized Linear Regression, *Technometrics*, Vol. 38, No. 4, 374–392.

Nguyen DV and Rocke DM (2002) Tumor classification by partial least squares using microarray gene expression data. *Bioinformatics* **18**: 39–50.