Projection Pursuit for Exploratory Supervised Classification

Eun-Kyung Lee*
Dianne Cook*
Sigbert Klinke**
Thomas Lumley***

BERLIN

ECONOMIC RISK

SFB 649

* Department of Statistics, Iowa State University, USA
** Institute for Statistics and Econometrics,
Humboldt-Universität zu Berlin, Germany
*** Department of Biostatistics, University of Washington, USA

Projection Pursuit for Exploratory Supervised Classification [0]

EUN-KYUNG LEE[1], DIANNE COOK[2], SIGBERT KLINKE[3], and THOMAS LUMLEY[4]

1 Department of Statistics, Ewha Womans University

2 Department of Statistics, Iowa State University

3 Institute for Statistics and Econometrics, Humboldt-Universität zu Berlin

4 Department of Biostatistics, University of Washington

ABSTRACT

In high-dimensional data, one often seeks a few interesting low-dimensional projections that reveal important features of the data. Projection pursuit is a procedure for searching high-dimensional data for interesting low-dimensional projections via the optimization of a criterion function called the projection pursuit index. Very few projection pursuit indices incorporate class or group information in the calculation. Hence, they cannot be adequately applied in supervised classification problems to provide low-dimensional projections revealing class differences in the data . We introduce new indices derived from linear discriminant analysis that can be used for exploratory supervised classification.

Key Words: Data mining; Exploratory multivariate data analysis; Gene expression data; Discriminant analysis;

1. Introduction

This paper is about methods for finding interesting projections of multivariate data when the observations belong to one of several known groups. The type of data is denoted as a p-dimensional vector $\mathbf{X}_{ij}$ representing the jth observation of the ith class, $i = 1, \ldots, g$, g is the number of classes, $j = 1, \ldots, n_i$, and n_i is the number of observations in class i. Let $\bar{\mathbf{X}}_{i.} = \sum_{j=1}^{n_i} \mathbf{X}_{ij}/n_i$ be the ith group mean and $\bar{\mathbf{X}}_{..} = \sum_{i=1}^{g} \sum_{j=1}^{n_i} \mathbf{X}_{ij}/n$ be the total mean, where $n = \sum_{i=1}^{g} n_i$. Interesting projections correspond to views where there are the biggest difference between the observations from different classes, that is, the classes are clustered in the view. In this paper, the approach to finding interesting projections uses the measures of between group variation, relative to within-group variation. These new methods are important for exploratory data analysis and data mining purposes when the task is to (1) examine the nature of clustering in the space of the data due to class information, and (2) to build a classifier for predicting the class of new data.

Projection pursuit is a method to search for interesting linear projections by optimizing some pre-determined criterion function, called a projection pursuit index. This idea originated with Kruskal (1969), and Friedman and Tukey (1974) first used the term "projection pursuit" describing a technique for exploratory analysis of multivariate data. It is useful for an initial data analysis, especially when data is in a high dimensional space. A problem many multivariate analysis techniques face is "the curse of dimensionality", that is, most of high dimensional space is empty. Projection pursuit methods help us explore multivariate data in interesting low dimensional spaces. The definition of an "interesting" projection depends on the projection pursuit index and on the application or purpose.

Many projection pursuit indices have been developed to define interesting projections. Because most low-dimensional projections are approximately normal (Huber, 1985), most of the projection pursuit indices are focused on non-normality. For example, the entropy index and the moment index (Jones and Sibson, 1987), the Legendre index (Friedman, 1987), the Hermite index (Hall, 1989), and the Natural Hermite index (Cook, et al, 1993), all search for projections where the data exhibit a high degree of non-normality.

Visual inspection of high dimensional data using projections is helpful to understand data, especially when it is combined with dynamic graphics. GGobi is an interactive and dynamic software system for data visualization and projection pursuit is implemented in it dynamically (Swayne, et al, 2003). The Holes index and the central mass index in GGobi are helpful in finding projections with few observations in the center and projections containing an abundance of points in the center, respectively (Cook, et al, 1993).

As the data mining area has grown, projection pursuit methods are increasingly used in classification and clustering to escape the curse of dimensionality. Posse (1992) suggested a method for projection pursuit discriminant analysis for two groups. He used kernel density estimation of the projected data instead of the original data and used the total probability of misclassification of the projected data as a projection pursuit index. Polzehl (1995) considered the cost of misclassification and used the expected overall loss as a projection pursuit index. Flick, et al (1990) uses a basis function expansion to estimate density and minimizes a measure of scatter. These projection pursuit methods for classification focus on 1-dimensional projections and it is hard to extend them to k-dimensional projections. Examining higher than 1-dimensional projections is important for visual inspection of high-dimensional data.

The methods presented in this paper start from a well-known classification method called linear discriminant analysis(LDA). This approach is extended to provide new projection pursuit indices for exploratory supervised classification. These indices use Fisher's linear discriminant ideas and expand Huber's ideas on projection pursuit for classification. These new indices are helpful for building understanding about how class structure relates to measured variables and they can be used to provide graphics to assess and verify supervised classification results. These indices are implemented as an R package, and these indices are available in GGobi for dynamic graphics (Swayne, et al, 2003)

This paper is organized as follows. Section 2 introduces the new projection pursuit indices and describes their properties. The optimization method to find the interesting projections is discussed in Section 3. Section 4 describes how to apply these indices using two gene expression data sets.

2. Index Definition

2.1 LDA projection pursuit index

The first index is derived from classical linear discriminant analysis (LDA). The approach, first developed by Fisher (1938), finds linear combinations of the data which have large between-group sums of squares relative to within-group sums of squares. (For detailed explanations, see Johnson and Wichern, 2002 and Duda et al. 2001) Let

$$\mathbf{B} = \sum_{i=1}^{g} n_i (\bar{\mathbf{X}}_{i.} - \bar{\mathbf{X}}_{..})(\bar{\mathbf{X}}_{i.} - \bar{\mathbf{X}}_{..})^T : \text{between-group sums of squares,} \tag{1}$$

$$\mathbf{W} = \sum_{i=1}^{g}\sum_{j=1}^{n_i} (\mathbf{X}_{ij} - \bar{\mathbf{X}}_{i.})(\mathbf{X}_{ij} - \bar{\mathbf{X}}_{i.})^T : \text{within-group sums of squares.} \tag{2}$$

Dimension reduction is achieved by finding the linear projection, $\mathbf{a}$, that maximizes $(\mathbf{a}^T\mathbf{B}\mathbf{a})/(\mathbf{a}^T\mathbf{W}\mathbf{a})$, which leads to the natural definition of a projection pursuit index. $(\mathbf{a}^T\mathbf{B}\mathbf{a})/(\mathbf{a}^T\mathbf{W}\mathbf{a})$ ranges between 0 and 1, where low values correspond to projections that display little class difference and high values correspond to projections that have large differences between the classes. To extend to an arbitrary-dimensional projection, we consider a test statistic used in multivariate analysis of variance (MANOVA) called Wilks $\Lambda^* = |\mathbf{W}|/|\mathbf{W}+\mathbf{B}|$. This quantity also ranges between 0 and 1, although the interpretation of numerical values are reversed from the 1-dimensional measure defined above. Small values of Λ^* correspond to large differences between the classes.

Let $\mathbf{A} = [\mathbf{a}_1 \ \mathbf{a}_2 \ \cdots \ \mathbf{a}_k]$ define an orthogonal projection onto a k-dimensional space. In projection pursuit the convention is that interesting projections are the ones that maximize the projection pursuit index, so we use the negative value of Wilks Lambda and add 1 to keep this index between 0 and 1. This gives the LDA projection pursuit index (LDA index) as:

$$I_{LDA}(\mathbf{A}) = \begin{cases} 1 - \dfrac{|\mathbf{A}^T\mathbf{W}\mathbf{A}|}{|\mathbf{A}^T(\mathbf{W}+\mathbf{B})\mathbf{A}|} & \text{for } |\mathbf{A}^T(\mathbf{W}+\mathbf{B})\mathbf{A}| \neq 0 \\[2em] 0 & \text{for } |\mathbf{A}^T(\mathbf{W}+\mathbf{B})\mathbf{A}| = 0 \end{cases} \tag{3}$$

Low index values correspond to little difference between classes and high values correspond to large differences between classes. The next proposition quantifies the minimum and maximum values. For simplicity, we denote $\mathbf{W} + \mathbf{B}$ as $\mathbf{\Phi}$.

Proposition 1. Let rank$(\mathbf{\Phi}) = p, \ k \leq \min(p, g)$. Then,

$$1 - \prod_{i=1}^{k} \lambda_i \leq I_{LDA}(\mathbf{A}) \leq 1 - \prod_{i=p-k+1}^{p} \lambda_i \tag{4}$$

where $\lambda_1, \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$: eigenvalues of $\mathbf{\Phi}^{-1/2}\mathbf{W}\mathbf{\Phi}^{-1/2}$,

$\mathbf{e}_1, \mathbf{e}_2, \cdots, \mathbf{e}_p$: corresponding eigenvectors of $\mathbf{\Phi}^{-1/2}\mathbf{W}\mathbf{\Phi}^{-1/2}$,

$\mathbf{f}_1, \mathbf{f}_2, \cdots, \mathbf{f}_p$: eigenvectors of $\mathbf{\Phi}^{-1/2}\mathbf{B}\mathbf{\Phi}^{-1/2}$.

In (4), the right equality holds when $\mathbf{A} = \mathbf{\Phi}^{-1/2}[\mathbf{e}_p \ \mathbf{e}_{p-1} \ \cdots \ \mathbf{e}_{p-k+1}] = \mathbf{\Phi}^{-1/2}[\mathbf{f}_1 \ \mathbf{f}_2 \ \cdots \ \mathbf{f}_k]$ and the left equality holds when $\mathbf{A} = \mathbf{\Phi}^{-1/2}[\mathbf{e}_k \ \mathbf{e}_{k-1} \ \cdots \ \mathbf{e}_1] = \mathbf{\Phi}^{-1/2}[\mathbf{f}_{p-k+1} \ \mathbf{f}_{p-k+2} \ \cdots \mathbf{f}_p]$.

A problem arises for LDA when rank$(\mathbf{W}) = r < p$. We need to remove collinearity by removing variables, before applying LDA. Otherwise, we need to modify the $\mathbf{W}^{-1}$ calculation, for example, to use the pseudo-inverse (pseudo LDA : Fukunaga, 1990) , or to use a ridge estimate instead of $\mathbf{W}$ such as regularized discriminant analysis (Friedman, 1989). For projection pursuit, because we make calculations in k-dimensional space instead of p-dimensional space, we can find interesting projections without an initial dimension reduction or modified $\mathbf{W}$ calculation. The next proposition shows how the LDA index works when rank$(\mathbf{W}) < p$.

Proposition 2. Let rank$(\mathbf{\Phi}) = r < p, \ k \leq \min(r, g)$. Then,

$$1 - \prod_{i=1}^{k} \delta_i \leq I_{LDA}(\mathbf{A}) \leq 1 - \prod_{i=r-k+1}^{r} \delta_i \tag{5}$$

where $\mathbf{\Phi} = \begin{bmatrix} \mathbf{P} & \mathbf{Q} \end{bmatrix} \begin{bmatrix} \mathbf{\Lambda} & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \mathbf{P}^T \\ \mathbf{Q}^T \end{bmatrix} = \mathbf{P}\mathbf{\Lambda}\mathbf{P}^{\mathbf{T}}$: spectral decomposition of $\mathbf{\Phi}$,

$\mathbf{P} : k \times r$ matrix, $\mathbf{P}^T\mathbf{P} = \mathbf{I}_r$,

$\mathbf{Q} : k \times (k-r)$ matrix, $\mathbf{Q}^T\mathbf{Q} = \mathbf{I}_{k-r}$,

$\mathbf{\Lambda} = \text{diag}[\delta_1, \delta_2, \cdots, \delta_r] : r \times r$ diagonal matrix,

$\delta_1, \delta_2, \cdots, \delta_r$: eigenvalues of $\mathbf{\Lambda}^{-1/2}\mathbf{P}^T\mathbf{W}\mathbf{P}\mathbf{\Lambda}^{-1/2}$,

$\mathbf{e}_1, \mathbf{e}_2, \cdots, \mathbf{e}_r$: corresponding eigenvectors of $\mathbf{\Lambda}^{-1/2}\mathbf{P}^T\mathbf{W}\mathbf{P}\mathbf{\Lambda}^{-1/2}$.

In (5), the right equality holds when $\mathbf{A} = \mathbf{P}\mathbf{\Lambda}^{-1/2}[\mathbf{e}_r \ \mathbf{e}_{r-1} \ \cdots \ \mathbf{e}_{r-k+1}]$, and the left equality holds when $\mathbf{A} = \mathbf{P}\mathbf{\Lambda}^{-1/2}[\mathbf{e}_k \ \mathbf{e}_{k-1} \ \cdots \ \mathbf{e}_1]$.
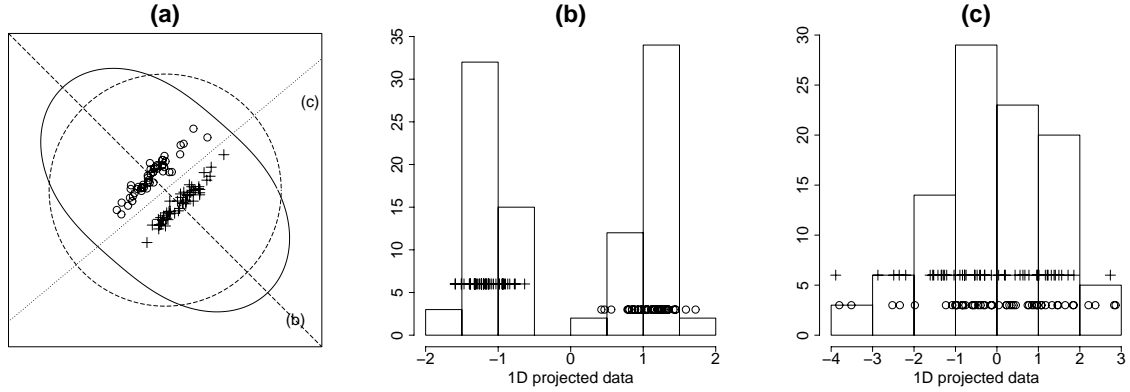

Figure 1. (a) Huber's plot(1990) using I_{LDA} on data simulated from two bivariate normal population. The symbols $\bigcirc$ and $+$ represent two different classes. The solid ellipse represents I_{LDA} value for all 1-dimensional projections, and the dashed circle is a guide set at the median I_{LDA} value. The straight dotted line labelled (b) is the optimal projection direction using I_{LDA} and the histogram of the projected data is shown in the correspondingly labelled plot (b). In plot (a) the dotted line labelled (c) is the first principal component direction and the the histogram of the projected data is shown in the correspondingly labelled plot (c).

The proofs of these two propositions are provided in Lee (2003). To illustrate the behavior of the LDA

index (Figure 1), we use a type of plot that was introduced by Huber (1990). In one-dimensional projections from a 2-dimensional space, for $\theta = 0°, \cdots, 179°$, the projection pursuit index is calculated using projection $\mathbf{a}_\theta = (cos\theta, sin\theta)$ and displayed radially as a function of θ. In each figure, the data points are plotted in the center. The solid ellipse represents the index value, I_{LDA}, plotted at distances relative to the center. The dotted circle is a guide line plotted at the median index value.

Figure 1 shows how the LDA index works. Data are simulated from two normal distributions with the same variance, $\Sigma = \begin{pmatrix} 1 & 0.95 \\ 0.95 & 1 \end{pmatrix}$, and different means, $\mu_1 = \begin{pmatrix} -1 \\ 0.6 \end{pmatrix}$ and $\mu_2 = \begin{pmatrix} 1 \\ -0.6 \end{pmatrix}$. Each group has 50 samples. Figure 1(a) shows that the LDA index function (solid line) is smooth and has a maximum value when the projected data reveals two separated classes. Figure 1(b) and Figure 1(c) are the histograms of the optimal projected data using the LDA index and the data projected onto the first principal component. The LDA index finds separated class structure. Principal component analysis is commonly used to find revealing low-dimensional projections, but it really does not work well in classification problems. Here, principal component analysis solves a different problem: It finds a projection that shows large variation (Johnson and Wichern, 2002).

The LDA index works well generally, but it has some problems in special circumstances. One special situation is 2-dimensional data generated from a uniform mixture of three Gaussian distributions, with identity variance-covariance matrices and centers at the vertices of an equilateral triangle. Figure 2(a) shows the theoretical case where three classes have the exact same variance-covariance matrix and three class means are the vertices of an equilateral triangle. In this case, all directions have the same LDA index values. The best projection is the full 2-dimensional data space. Figure 2(b) shows data simulated from this distribution. Because of the sampling, variances are slightly different in each class and the three means do not lie exactly on an equilateral triangle. Therefore the optimal direction(the dotted straight line in Figure 2(b)) depends on the sampling variation. If a new sample is generated, a completely different optional projection will occur. This is not what we want in exploratory methods. We would like to be able to find all the interesting data

structures, which in this case would be the three 1-dimensional projections revealing each group separated from the other two groups. We extend this problem of the LDA index to define a new index that is able to detect interesting structures in this situation.

2.2 LDA extended projection pursuit index using L_r-norm

We start from the 1-dimensional index. Let $y_{ij} = \mathbf{a}^T\mathbf{X}_{ij}$ be a projected observation onto a 1-dimensional space. In the LDA index, we use $\mathbf{a}^T\mathbf{B}\mathbf{a}$ and $\mathbf{a}^T\mathbf{W}\mathbf{a}$ as the measures of between-group and within-group variations, respectively. These two measures can be explained as the square of L_2 vector norm, as follows.
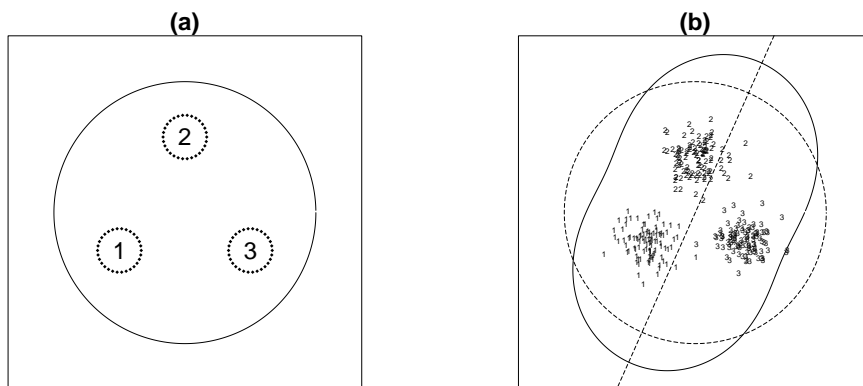

Figure 2. This plot illustrates a problem situation for I_{LDA}. (a) The theoretical case where the three classes have the exact same variance and the three class means come are located on the vertices of an equilateral triangle. All directions have exactly same I_{LDA} values (solid circle). The best projection is really the full 2-dimensional data space! (b) What happens in practice? This plot contains data generated from the theoretical distribution. An optimal projection is found purely due to sampling variation. If a new sample were generated a completely different optimal projection will be found.

$$\mathbf{a}^T \mathbf{B} \mathbf{a} \quad = \quad \sum_{i=1}^{g} \sum_{j=1}^{n_i} (\bar{y}_{i.} - \bar{y}_{..})^2 = \left\{ ||\mathbf{M}\bar{\mathbf{y}}_g - \mathbf{1}_n \bar{y}_{..}||_2 \right\}^2 \tag{6}$$

$$\mathbf{a}^T \mathbf{W} \mathbf{a} \quad = \quad \sum_{i=1}^{g} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i.})^2 = \left\{ ||\mathbf{y} - \mathbf{M}\bar{\mathbf{y}}_g||_2 \right\}^2 \tag{7}$$

$$\mathbf{a}^T \boldsymbol{\Phi} \mathbf{a} \quad = \quad \sum_{i=1}^{g} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{..})^2 = \left\{ ||\mathbf{y} - \mathbf{1}_n \bar{y}_{..}||_2 \right\}^2 = \left\{ ||\mathbf{M}\bar{\mathbf{y}}_g - \mathbf{1}_n \bar{y}_{..}||_2 \right\}^2 + \left\{ ||\mathbf{y} - \mathbf{M}\bar{\mathbf{y}}_g||_2 \right\}^2 \tag{8}$$

where $\mathbf{M} = diag(\mathbf{1}_{n_1}, \cdots, \mathbf{1}_{n_g})$, $\bar{\mathbf{y}}_g = [\bar{y}_{1.}, \bar{y}_{2.}, \cdots, \bar{y}_{g.}]^T$, $\mathbf{y} = \left[\mathbf{y}_1{}^T, \mathbf{y}_2{}^T, \cdots, \mathbf{y}_g{}^T \right]^T$, $\mathbf{y}_i = [y_{i1}, y_{i2}, \cdots, y_{in_g}]^T$,

and $\mathbf{1}_n = [1, 1, \cdots, 1]^T : n \times 1$ vector. We extend to the L_r norm. Let

$$\mathbf{B}_r \quad = \quad \left\{ ||\mathbf{M}\bar{\mathbf{y}}_g - \mathbf{1}_n \bar{y}_{..}||_r \right\}^r = \sum_{i=1}^{g} \sum_{j=1}^{n_i} |\bar{y}_{i.} - \bar{y}_{..}|^r \tag{9}$$

$$\mathbf{W}_r \quad = \quad \left\{ ||\mathbf{y} - \mathbf{M}\bar{\mathbf{y}}_g||_r \right\}^r = \sum_{i=1}^{g} \sum_{j=1}^{n_i} |y_{ij} - \bar{y}_{i.}|^r. \tag{10}$$

Then

$$\left\{ ||\mathbf{y} - \mathbf{1}_n \bar{y}_{..}||_r \right\}^r = \sum_{i=1}^{g} \sum_{j=1}^{n_i} |y_{ij} - \bar{y}_{..}|^r \quad \leq \quad \sum_{i=1}^{g} \sum_{j=1}^{n_i} |\bar{y}_{i.} - \bar{y}_{..}|^r + \sum_{i=1}^{g} \sum_{j=1}^{n_i} |y_{ij} - \bar{y}_{i.}|^r = \mathbf{B}_r + \mathbf{W}_r. \tag{11}$$

Even though the additivity does not hold for the L_r vector norm, $\mathbf{B}_r$ and $\mathbf{W}_r$ can be substitutes for the

measures of between-group and within-group variabilities. We use these measures to define our new index.

The 1-dimensional L_r projection pursuit index (L_r index) is defined by

$$I_{L_r}(\mathbf{a}) = \left(\frac{\mathbf{B}_r}{\mathbf{W}_r} \right)^{1/r} \quad = \quad \frac{||\mathbf{M}\bar{\mathbf{y}}_g - \mathbf{1}_n \bar{y}_{..}||_r}{||\mathbf{y} - \mathbf{M}\bar{\mathbf{y}}_g||_r} \tag{12}$$

$$= \quad \left(\frac{\sum_{i=1}^{g} \sum_{j=1}^{n_i} (\bar{y}_{i.} - \bar{y}_{..})^r}{\sum_{i=1}^{g} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i.})^r} \right)^{1/r}. \tag{13}$$

Taking the ratio to the $1/r$ power, prevents this index value from getting too big. The 1-dimensional LDA

index is a special case of this index when $r = 2$.

For a k-dimensional projection $\mathbf{A}$, let $\mathbf{Y}_{ij} = \mathbf{A}^T \mathbf{X}_{ij} = [y_{ij1}, y_{ij2}, \cdots, y_{ijk}]^T$ be a projected observation
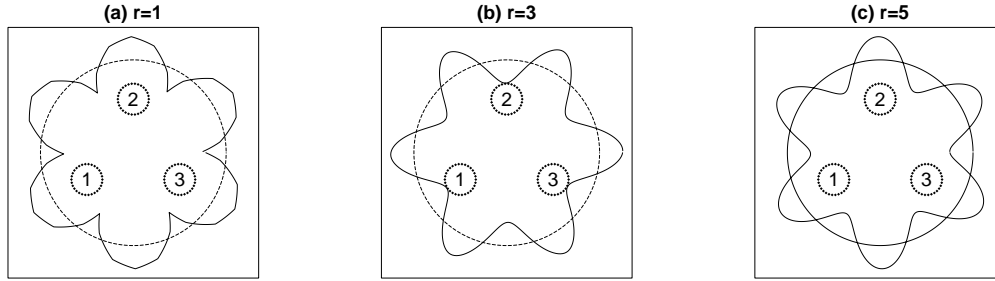
Figure 3. Huber's plot showing behavior of I_{L_r} index on the special data that caused problems for I_{LDA}. (a) I_{L_1}: The optimal projections separate each class from the other two. (b) I_{L_3}: The optimal projections separate all three classes. (c) I_{L_5}: The optimal projections separate each class from the other two. When r=2 and r=4, the index is the same as I_{LDA}, shown in Figure 2(a).

onto the k dimensional space spanned by $\mathbf{A}$. Then

$$[\mathbf{A}^T\mathbf{B}\mathbf{A}]_{lm} = \sum_{i=1}^{g}\sum_{j=1}^{n_i}\left(\bar{y}_{i.l} - \bar{y}_{..l}\right)\left(\bar{y}_{i.m} - \bar{y}_{..m}\right), \qquad (14)$$

$$[\mathbf{A}^T\mathbf{W}\mathbf{A}]_{lm} = \sum_{i=1}^{g}\sum_{j=1}^{n_i}\left(y_{ijl} - \bar{y}_{i.l}\right)\left(y_{ijm} - \bar{y}_{i.m}\right) \qquad (15)$$

where $l, m = 1, 2, \cdots, k$. The diagonals of these matrices represent the variances of the between (or within) group for each variable and the off-diagonals represent covariances between variables. We take only the diagonal parts of these between-group and within-group variance and extend these sums of squares to L_r norms. Then,

$$I_{L_r}(\mathbf{A}) = \left(\frac{\sum_{l=1}^{k}\sum_{i=1}^{g}\sum_{j=1}^{n_i}\left(\bar{y}_{i.l} - \bar{y}_{..l}\right)^r}{\sum_{l=1}^{k}\sum_{i=1}^{g}\sum_{j=1}^{n_i}\left(y_{ijl} - \bar{y}_{i.l}\right)^r}\right)^{1/r}. \qquad (16)$$

For detailed explanations, see Lee (2003).

Figure 3 shows how the new index I_{L_r} $(r = 1, 2, 3)$ performs for the special situation that caused problem for I_{LDA}. When $r = 1$, all three optimal projections separate one class from the other two classes. When $r = 3$, the optimal projections separate the three classes. With the L_5 index, we found the same optimal

projections as the L_1 index but the index function is smoother than the L_1 index. When r is 2 and 4, these indices have the same value for all directions, just like the LDA index.

The LDA index and the L_r index ($r \geq 2$) are usually sensitive to outliers, mainly due to use the sums of squares or higher power, which are sensitive measure to outliers. On the other hand, the L_1 index uses the sums of absolute values. Therefore it is more robust to outliers than other indices. Figure 4 shows how these indices work in the presence of an outlier. In each plot, there are two classes (1 and 2). The class 1 has 21 observations with one outlier and the class 2 has 20 observations. The histogram of the optimal 1-dimensional projected data using the L_1 index (Figure 4 (a-1)) shows that the outlier is separated from two groups and the best projection is not affected by this outlier. When $r \geq 2$, the best projections are leveraged towards the direction of the outlier. With the exception of the outlier, the L_1 index provides a more separated view of the two classes than the best projection of the $L_r(r \geq 2)$ index.
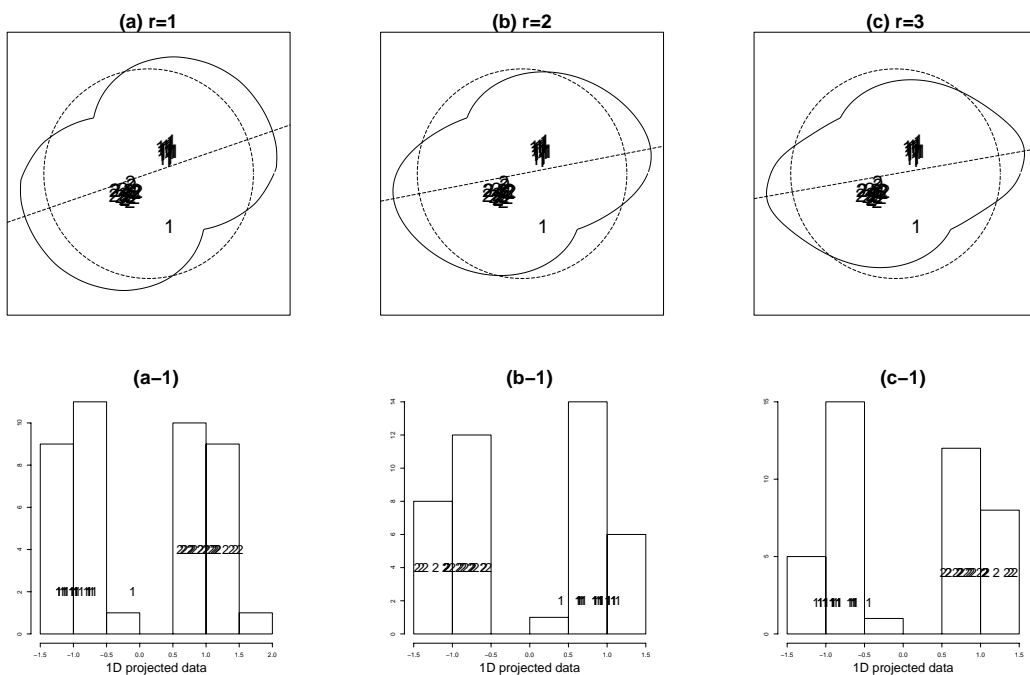

Figure 4. The behavior of the I_{L_r} in the presence of a outlier, using simulated data with 2 classes, where class 1

has an outlier. (a) Huber's plot of I_{L_1}, (a-1) Histogram of the projected data onto the L_1 optimal projection (b) Huber's plot of I_{L_2}, (b-1) Histogram of the projected data onto the L_2 optimal projection (c) Huber's plot of I_{L_3}, (c-1) Histogram of the projected data onto the L_3 optimal projection.

3. Optimization

A good optimization procedure is an important part of projection pursuit. The purpose of projection pursuit optimization is to find all of the interesting projections, not only to find one global maximum, because sometimes the local maximum can reveal unexpectedly interesting data structure. For this reason, the projection pursuit optimization algorithm needs to be flexible enough to find global and local maxima.

Posse (1990) compared the several optimization procedures, and suggest a random search for finding the global maximum of a projection pursuit index. Cook, et al (1995) use a grand tour alternated with a simulated annealing optimization of a projection pursuit index, to creating a continuous stream of projections that are displayed for exploratory visualization of multivariate data. Klein and Dubes (1989) showed that simulated annealing can produce results as good as those obtained by conventional optimization methods and this method performs well for large data sets.

Simulated annealing was first proposed by Kirkpatrick, et al (1983) as a method to minimize objective functions that have many variables. The fundamental idea of simulated annealing is that a re-scaling parameter, called the "temperature", allows control of the speed of convergence to a minimum value. For an objective function $h(\theta)$, called the "energy", we start from the initial value θ_0. A value, θ^* is generated in the neighborhood of θ_0. Then, θ^* is accepted as a new value with probability ρ, defined by the temperature and the energy difference between θ_0 and θ^*. This probability ρ guards against getting trapped into a local minimum allowing the algorithm to visit a local minimum and then jump out and explore for other minima. For detailed explanations, see Bertsimas and Tsitsiklis (1993).

For our projection pursuit optimization we maximize an objective function. We use two different tem-

peratures, one (D_i) is for neighborhood definition, and the other (T_i) is for the probability ρ . D_i is re-scaled by the predetermined cooling parameter c and T_i is defined by $T_0/\log(i+1)$. Before we start, we need to choose the cooling parameter, c, and the initial temperature, T_0. The cooling parameter, c, determines how many iterations are needed to converge and whether the maximum is likely to be a local maximum or a global maximum. The initial temperature, T_0, also controls the speed of convergence. Even if the cooling parameter c is small, there is a chance that the algorithm will stop before it reaches the peak. If c is large, more iterations are needed to get a final value, but this final solution is more likely to be at the peak value, and that it is a global maximum. Therefore this algorithm is quite flexible for finding interesting projections. (For detailed discussion, see Lee, 2003.)

Simulated Annealing Optimization Algorithm for Projection Pursuit

1. Set an initial projection, $\mathbf{A}_0$, and calculate the initial projection pursuit index value $I_0 = I(\mathbf{A}_0)$.

For the ith iteration,

2. Generate a projection $\mathbf{A}_i$ from $N_{D_i}(\mathbf{A}_0)$,

   where $D_i = c^i$, c is the predetermined cooling parameter in the range (0,1),

   $N_{D_i}(\mathbf{A}_0) = \{\mathbf{A} : \mathbf{A}$ is an orthonormal projection with direction $\mathbf{A}_0 + D_i\mathbf{B}$, $\forall$ random projections $\mathbf{B}\}$

3. Calculate $I_i = I(\mathbf{A}_i)$, $\Delta I_i = I_i - I_0$, $T_i = \frac{T_0}{\log(i+1)}$,

4. Set $\mathbf{A}_0 = \mathbf{A}_i$ and $I_0 = I_i$ with probability $\rho = \min\left(\exp\left(\frac{\Delta I_i}{T_i}\right), 1\right)$ and increase i to i+1

   Repeat 2-4 until ΔI_i is small.

4. Application

DNA microarray technologies provide a powerful tool for analyzing thousands of genes simultaneously. Comparison of gene expression levels between samples can be used to obtain information about important

genes and their functions. Because microarrays contain a large number of genes on each chip but typically few chips are used, analyzing DNA microarray data usually means dealing with large p, small n challenges. A recent publication that compares classification methods for gene expression data (Dudoit, et al., 2002) has focused on the classification error. We will use the same data sets to demonstrate the use of new projection pursuit indices.

4.1 Data sets

Leukemia This data set originated from a study of gene expression in two types of acute leukemias, acute lymphoblastic leukemia(ALL) and acute myeloid leukemia(AML). The data set consists of $n_1 = 25$ cases of AML and $n_2 = 47$ cases of ALL(38 cases of B-cell ALL and 9 cases of T-cell ALL), giving $n = 72$. After preprocessing, we have $p = 3571$ human genes. This data set is available at http://www-genome.wi.mit.edu/mpr and was described by Golub, et al. (1999).

NCI60 This data set consists of 8 different tissue types where cancer was found : $n_1 = 9$ cases from breast, $n_2 = 5$ cases from central nervous system(CNS), $n_3 = 7$ cases from colon, $n_4 = 8$ cases from leukemia, $n_5 = 8$ cases from melanoma, $n_6 = 9$ cases from non-small-cell lung carcinoma(NSCLC), $n_7 = 6$ cases from ovarian, and $n_8 = 9$ cases from renal, and p=6830 human genes. Missing values are imputed by a simple k nearest-neighbor algorithm ($k = 5$). We use these data to show how to use exploratory projection pursuit classification when the number of classes is large. This data set is available at http://genome-www.stanford.edu/sutech/download/nci60/index.html and was described by Ross, et al. (2000).

Standardization and Gene Selection The gene expression data were standardized so that each observation has mean 0 and variance 1. For gene selection, we use the ratio of between-group to within-group sums of squares.

$$BW(j) = \frac{\sum_{i=1}^{n} \sum_{k=1}^{g} I(y_i = k)(\bar{x}_{k,j} - \bar{x}_{.,j})^2}{\sum_{i=1}^{n} \sum_{k=1}^{g} I(y_i = k)(x_{i,j} - \bar{x}_{k,j})^2} \tag{17}$$

where $\bar{x}_{.,j} = (1/n) \sum_{i=1}^{n} x_{i,j}$ and $\bar{x}_{k,j} = (\sum_{i=1}^{n} I(y_i = k)x_{i,j})/(\sum_{i=1}^{n} I(y_i = k))$. At the beginning, we follow
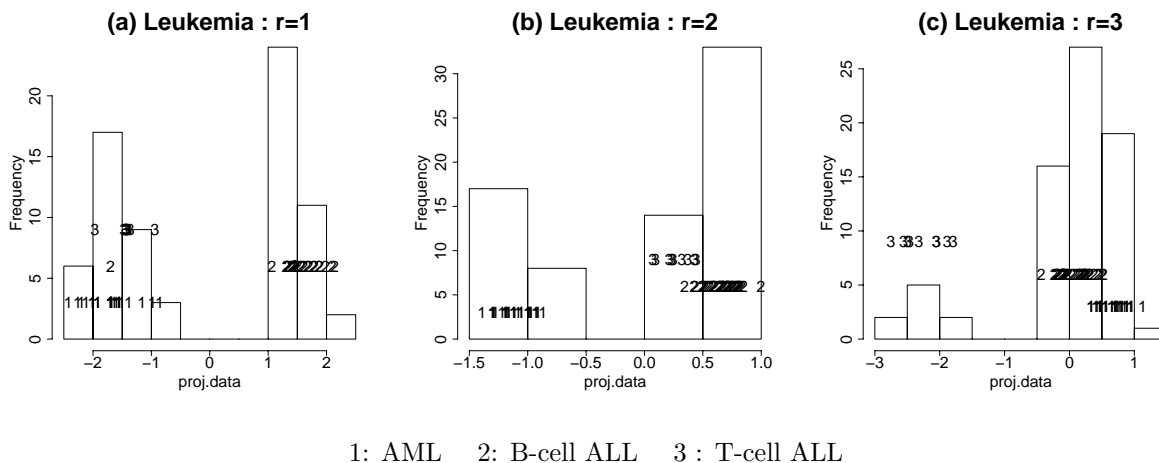
(a) Leukemia : r=1   **(b) Leukemia : r=2**   **(c) Leukemia : r=3**

1: AML   2: B-cell ALL   3 : T-cell ALL

Figure 5. Leukemia data : 1-dimensional projection (p=40) (a) the histogram of the optimal projected data using I_{L_1}

(b) the histogram of the optimal projected data using I_{L_2} (c) the histogram of the optimal projected data using I_{L_3}

the original study (Dudoit, et al, 2002) and start with $p = 40$ for the leukemia data and $p = 30$ for the NCI60 data and discuss different numbers of genes later.

4.2 Results

1-dimensional projection

Figure 5 displays the histograms of the projected data onto the optimal 1-dimensional projections. For this application, we choose a very large cooling parameter (0.999) which gives us the global maximum. In the Leukemia data, when $r=1$ (Figure 5-a), the B-cell ALL class is separated from the other classes except for one case. When $r = 2$ (Figure 5-b), the three classes are almost separable when the L_2 index is used, which is the same result as for the LDA index. As r is increased, the index tends to separate the T-cell ALL from the others (Figure 5-c).

The NCI60 data is a quite challenging example. For such a small number of observations, there are too
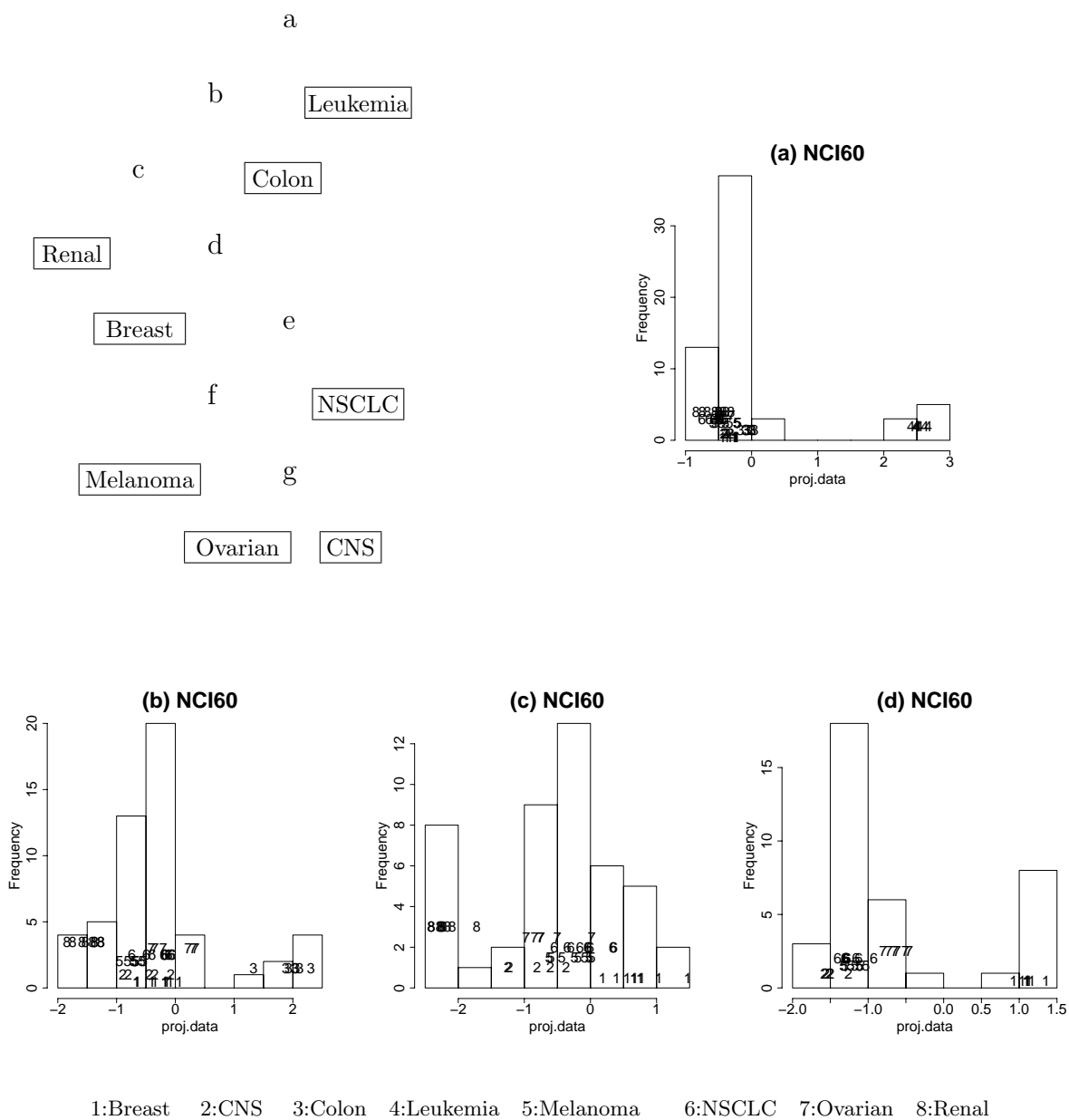
a

b   Leukemia

c   Colon

Renal   d

Breast   e

f   NSCLC

Melanoma   g

Ovarian   CNS

(a) NCI60

(b) NCI60   **(c) NCI60**   **(d) NCI60**

1:Breast   2:CNS   3:Colon   4:Leukemia   5:Melanoma   6:NSCLC   7:Ovarian   8:Renal

Figure 6. NCI60 data : the 1-dimensional projection (p=30). (a) The histogram of the optimal projection using the LDA index. Leukemia group is separated from the others - peel off Leukemia group. (b)Colon group is separated (c) Renal group is separated (d) Breast group is separated.

16

many classes. For these data, we try an isolation method that applies projection pursuit iteratively and takes off one class at a time (Friedman and Tukey, 1974). The 8 classes are too many to separate with a single 1-dimensional projection. After finding one split, we apply projection pursuit to each partition. Usually one class is peeled off from the others in each step. The tree diagram in Figure 6 illustrates the steps. In the first step (Figure 6-a), we separate the Leukemia class from the others. At the second step, Colon class is separated (Figure 6-b). Then, the Renal, the Breast, the NSCLC, and the Melanoma classes are separated sequentially. Finally, the Ovarian and the CNS classes are separated.
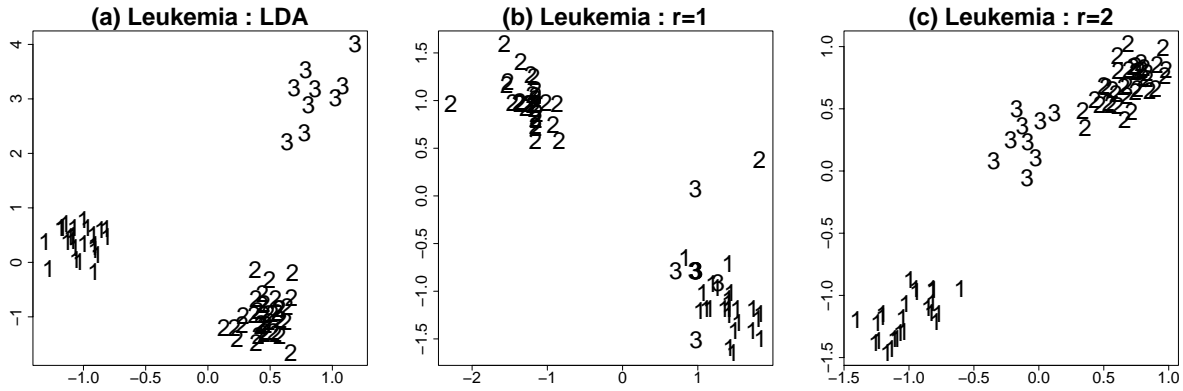
2-dimensional projection

Figures 7 and 8 show the plot of the data projected onto the optimal 2-dimensional projections for the Leukemia data. All three classes separate easily using the LDA index. Using the L_1 index, the B-cell ALL class is separated with one exception - the same outlier of the result of the 1-dimensional projection in Figure 5(c). In the 2-dimensional case, the LDA index is only the same as the L_2 index if $\mathbf{B}$ and $\mathbf{W}$ are diagonal matrices. The best result is obtained using the I_{LDA} index, where all three classes are clearly separated. In the NCI60 data, the Leukemia class is clearly separated from the others for all indices (Figure 8).

Classification

Table 1. Test set error. Median and Upper quartile of the misclassified samples from 200 replications. ($n_{test} = 24$)

	Median	Upper quartile
Fisher's Linear Discriminant Analysis (LDA)	3	4
Diagonal linear discriminant analysis (DLDA)	0	1
Diagonal quadratic discriminant analysis (DQDA)	1	2
LDA projection pursuit method	2.5	4
L_1 projection pursuit method	1	2

Even though our projection pursuit indices are developed for the exploratory data analysis, especially for

1: AML   2: B-cell ALL   3 : T-cell ALL

Figure 7. Leukemia data : 2-dimensional projection (p=40). (a) I_{LDA} : The three classes are separated. (b) I_{L_1} : The B-cell ALL class is separated from the other two except for one case. (c) I_{L_2} : The three classes are separated, although the gap between classes 2 and 3 is small.
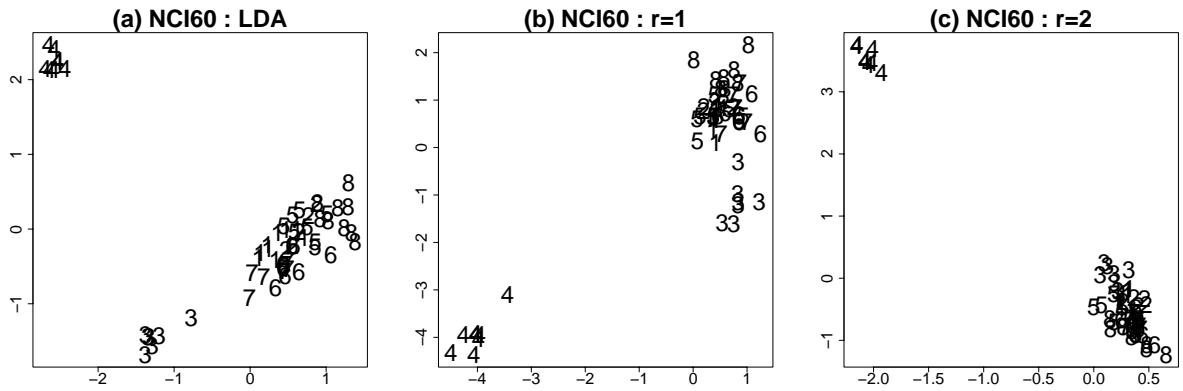
1:Breast   2:CNS   3:Colon   4:Leukemia   5:Melanoma   6:NSCLC   7:Ovarian   8:Renal

Figure 8. NCI60 : the 2-dimensional projection (p=30). (a) I_{LDA} : The Leukemia and Colon classes are separated from the others. (b) I_{L_1} : The Leukemia class and Colon classes are separated from the others, but Colon class is not clearly separated. (c) Only the Leukemia class is separated from the others.

the visual inspection, they can be used for classification. For comparison to the other methods, we show the results of Dudoit, et al(2001) on the Leukemia data with two classes : AML and ALL. For a 2/3 training set ($n_{train} = 48$), we calculate BW(Equation 17) values for each gene and select the 40 genes with the larger BW values. Using this 40 gene training set, find the optimal projection. Let $\mathbf{a}^*$ be the optimal projection, $\bar{\mathbf{X}}_{AML}$ be the mean of the observations in AML groups, $\bar{\mathbf{X}}_{ALL}$ be the mean of the observations in ALL groups, and $\mathbf{X}$ be an observation. Then, we build a classifier : If $\mathbf{a}^{*T}(\mathbf{X} - \bar{\mathbf{X}}_{AML}) < \mathbf{a}^{*T}(\mathbf{X} - \bar{\mathbf{X}}_{ALL})$, then $\mathbf{X}$ belongs to the AML group. Else, $\mathbf{X}$ belongs to the ALL group. (For detailed explanation, see Johnson and Wichern, 2002). Using this classifier, we compute the test error. This is repeated 200 times. The median and upper quartile of the test errors are summarized in Table 1. The results of Fisher's LDA, DLDA, and DQDA are from Dudoit, et al (2001). As we expect, I_{LDA} has similar results to Fisher's LDA. The L_1 compares well with the other methods.

5. Discussion

We have proposed new projection pursuit indices for exploratory supervised classification and examined their properties. In most applications, the LDA index works well to find a projection that has well-separated class structure. The L_r index can lead us to projections that have special features. With the L_1 index, we can get a projection that is robust to outliers. This index is useful for discovering outliers. As r is increased, the L_r index tends to be more sensitive to outliers. For exploratory supervised classification, we need to use several projection pursuit indices (at least LDA and L_1 indices) and examine different results. These indices can be used to obtain a better understanding of the class structure in the data space and their projection coefficients help find the important variables that best separate classes (Lee, 2003). The insights learned from plotting the optimal projections are useful when building a classifier and for assessing classifiers.

Projection pursuit methods can be applied to multivariate tree methods. Several authors have considered the problem of constructing tree-structured classifiers that have linear discriminants at each node. Friedman

19

(1977) reported that applying Fisher's linear discriminants, instead of univariate features, at some internal nodes was useful in building better trees. This is a similar approach to the isolation method that we applied to NCI 60 data (Figure 6).

A major issue revealed by the gene expression application is that when there are too few cases for variables the reliability of the classifications is questionable. There is a high probability of a separating hyperplane purely by chance when the number of genes is larger than half the sample size (the perceptron capacity bound). When the number of genes is larger than the sample size, most of high dimensional space is empty and we can find a separating hyperplane that divides groups purely by chance (see Ripley, 1996). For more detailed discussion, see Lee (2003).

For a large number of variables, our simulated annealing optimization algorithm for projection pursuit method is quite slow to find the global optimal projection. A faster annealing algorithm described by Ingber (1989) may be better.

Finally we have used the R language for this research and provide the classPP package (available at CRAN). These indices are also available for the guided tour in the software *GGobi* (http://www.ggobi.org).

REFERENCE

[1] Bertsimas, D., Tsitsiklis, J. (1993). "Simulated Annealing." *Statistical Science*, **8**(1), 10-15.

[2] Cook, D., Buja, A., and Cabrera, J. (1993). "Projection Pursuit Indexes Based on Orthogonal Function Expansions." *Journal of Computational and Graphical Statistics*, **2**(3), 225-250.

[3] Cook, D., Buja, A., Cabrera, J., and Hurley, C. (1995). "Grand Tour and Projection Pursuit." *Journal of Computational and Graphical Statistics*, **4**(3), 155-172.

[4] Duda, R. O., Hart, P. E., and Stork, D. G. (2001). *Pattern Classification, 2nd edition*, Wiley-Interscience Publication .

[5] Dudoit, S., Fridlyand, J., and Speed, T. P. (2002) "Comparison of Discrimination Methods for the Clas-

sification of Tumors Using Gene Expression Data." *Journal of the American Statistical Association*, **97**(1), 77-87.

[6] Fisher, R. A. (1938). "The Statistical Utilization of Multiple Measurements." *Annals of Eugenics*, **8**, 376-386.

[7] Flick, T. E., Jones, L. K., Priest, R. G., and Herman, C. (1990). "Pattern Classification using Projection Pursuit." *Pattern Recognition*, **23**(12), 1367-1376.

[8] Friedman, J. H. (1977). "A Recursive Partitioning Decision Rule for Nonparametric Classification." *IEEE Transactions on Computers*, **26**, 404-408.

[9] Friedman, J. H. (1987). "Exploratory Projection Pursuit." *Journal of the American Statistical Association*, **82**(1), 249-266.

[10] Friedman, J. H. (1989). "Regularized Discriminant Analysis." *Journal of the American Statistical Association*, **84**(1), 165-175.

[11] Friedman, J. H., and Tukey, J. W. (1974). "A Projection Pursuit Algorithm for Exploratory Data Analysis." *IEEE Transactions on Computers*, **C-23**, 881-890.

[12] Fukunaga, K. (1990). *Introduction to Statistical Pattern Recognition*, Academic Press, San-Diego.

[13] Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M., Downing, J. R., Caligiuri, M. A., Bloomfield, C. D., and Lander, E. S. (1999). "Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring." *Science*, **268**, 531-537.

[14] Hall, P. (1989). "On Polynomial-Based Projection Indices for Exploratory Projection Pursuit." *The Annals of Statistics*, **17**(2), 589-605.

[15] Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The Elements of Statistical Learning*, Springer.

[16] Huber, P. J. (1985). "Projection Pursuit." (with discussion) *The Annals of Statistics*, **13**(2), 435-525.

[17] Huber, P. J. (1990). "Data Analysis and Projection Pursuit." *Technical Report PJH-90-1, MIT*

[18] Ingber, L. (1989). "Very fast simulated re-annealing." *Mathl. Comput. Modelling*, **12**(8), 967-973.

[19] Johnson, R. A., and Wichern, D. W. (2002). *Applied Multivariate Statistical Analysis, 5th edition*, Prentice-Hall, New Jersey.

[20] Jones, M. C., and Sibson, R. (1987). "What is Projection Pursuit?" (with discussion) *Journal of the Royal Statistical Society, A*, **150**(1), 1-36.

[21] Kirkpatrick, S., Gelatt, C. D., and Vecchi, M. P. (1983). "Optimization by Simulated Annealing." *Science*, **220**, 671-680.

[22] Klein, R. W., Dubes, R. C. (1989). "Experiments in projection and clustering by simulated annealing." *Pattern Recognition*, **22**(2), 213-220.

[23] Kruskal, J. B.(1969) "Toward a Practical Method which Helps Uncover the Structure of a Set of Multivariate Observations by Finding the Linear Transformation which Optimizes a New 'Index of condensation'." *Statistical Computing*, New York : Academic press, 427-440

[24] Lee, E. (2003). "Projection Pursuit for Exploratory Supervised Classification." *Ph.D thesis, Department of Statistics, Iowa State University.*

[25] Piers, A. M. (2003). "Robust Linear Discriminant Analysis and the Projection Pursuit Approach, practical aspects." *Developments in Robust Statistics : International Conference on Robust Statistics 2001* , Ed. Dutter, Filzmoser, Gather and Rousseeuw. Springer-Verlag, 317-329.

[26] Polzehl, J. (1995). "Projection pursuit discriminant analysis." *Computational statistics and data analysis*, **20**(2), 141-157.

[27] Posse, C. (1990). "An Effective Two-dimensional Projection Pursuit Algorithm." *Communications in Statistics - Simulation and Computation*, **19**(4), 1143-1164.

[28] Posse, C. (1992). "Projection Pursuit Discriminant Analysis for Two Groups." *Communications in Statistics - Theory and Method*, **21**(1), 1- 19.

[29] Ripley, B. D. (1996). *Pattern Recognition and Neural Networks*, Cambridge University Press.

[30] Ross, D. T., Scherf, U., Eisen, M. B., Perou, C. M., Spellman, P., Iyer, V., Jeffrey, S. S., de Rijn, M. V.,

Waltham, M., Pergamenschikov, A., Lee, J. C. F., Lashkari, D., Shalon, D., Myers, T. G., Weinstein, J. N., Botstein, D., and Brown, P. O. (2000). "Systematic variation in gene expression patterns in human cancer cell lines." *Nature Genetics*, **24**(3), 227-234.

[31] Swayne, D.F., Lang, D.T., Buja, A., and Cook, D. (2003). "GGobi : evolving from XGobi into an extensible framework for interactive data visualization." *Computational Statistics and Data Analysis*, **43**(4), 423-444.

SFB 649 Discussion Paper Series

For a complete list of Discussion Papers published by the SFB 649,
please visit http://sfb649.wiwi.hu-berlin.de.

001 "Nonparametric Risk Management with Generalized Hyperbolic Distributions" by Ying Chen, Wolfgang Härdle and Seok-Oh Jeong, January 2005.

002 "Selecting Comparables for the Valuation of the European Firms" by Ingolf Dittmann and Christian Weiner, February 2005.

003 "Competitive Risk Sharing Contracts with One-sided Commitment" by Dirk Krueger and Harald Uhlig, February 2005.

004 "Value-at-Risk Calculations with Time Varying Copulae" by Enzo Giacomini and Wolfgang Härdle, February 2005.

005 "An Optimal Stopping Problem in a Diffusion-type Model with Delay" by Pavel V. Gapeev and Markus Reiß, February 2005.

006 "Conditional and Dynamic Convex Risk Measures" by Kai Detlefsen and Giacomo Scandolo, February 2005.

007 "Implied Trinomial Trees" by Pavel Čížek and Karel Komorád, February 2005.

008 "Stable Distributions" by Szymon Borak, Wolfgang Härdle and Rafal Weron, February 2005.

009 "Predicting Bankruptcy with Support Vector Machines" by Wolfgang Härdle, Rouslan A. Moro and Dorothea Schäfer, February 2005.

010 "Working with the XQC" by Wolfgang Härdle and Heiko Lehmann, February 2005.

011 "FFT Based Option Pricing" by Szymon Borak, Kai Detlefsen and Wolfgang Härdle, February 2005.

012 "Common Functional Implied Volatility Analysis" by Michal Benko and Wolfgang Härdle, February 2005.

013 "Nonparametric Productivity Analysis" by Wolfgang Härdle and Seok-Oh Jeong, March 2005.

014 "Are Eastern European Countries Catching Up? Time Series Evidence for Czech Republic, Hungary, and Poland" by Ralf Brüggemann and Carsten Trenkler, March 2005.

015 "Robust Estimation of Dimension Reduction Space" by Pavel Čížek and Wolfgang Härdle, March 2005.

016 "Common Functional Component Modelling" by Alois Kneip and Michal Benko, March 2005.

017 "A Two State Model for Noise-induced Resonance in Bistable Systems with Delay" by Markus Fischer and Peter Imkeller, March 2005.

018 "Yxilon – a Modular Open-source Statistical Programming Language" by Sigbert Klinke, Uwe Ziegenhagen and Yuval Guri, March 2005.

019 "Arbitrage-free Smoothing of the Implied Volatility Surface" by Matthias R. Fengler, March 2005.

020 "A Dynamic Semiparametric Factor Model for Implied Volatility String Dynamics" by Matthias R. Fengler, Wolfgang Härdle and Enno Mammen, March 2005.

021 "Dynamics of State Price Densities" by Wolfgang Härdle and Zdeněk Hlávka, March 2005.

022 "DSFM fitting of Implied Volatility Surfaces" by Szymon Borak, Matthias R. Fengler and Wolfgang Härdle, March 2005.

023 "Towards a Monthly Business Cycle Chronology for the Euro Area" by Emanuel Mönch and Harald Uhlig, April 2005.

024 "Modeling the FIBOR/EURIBOR Swap Term Structure: An Empirical Approach" by Oliver Blaskowitz, Helmut Herwartz and Gonzalo de Cadenas Santiago, April 2005.

025 "Duality Theory for Optimal Investments under Model Uncertainty" by Alexander Schied and Ching-Tang Wu, April 2005.

026 "Projection Pursuit For Exploratory Supervised Classification" by Eun-Kyung Lee, Dianne Cook, Sigbert Klinke and Thomas Lumley, May 2005.