Biostatistics (2010), **11**, 2, *pp.* 254–264
doi:10.1093/biostatistics/kxp056
Advance Access publication on January 11, 2010

Robust depth-based tools for the analysis of gene expression data

SARA LÓPEZ-PINTADO*

Departamento de Economía, Métodos Cuantitativos e Historia Económica, Universidad Pablo de Olavide, Sevilla 41013, Spain
sloppin@upo.es

JUAN ROMO

Departamento de Estadística, Universidad Carlos III de Madrid, Madrid 28903, Spain

AURORA TORRENTE

Departamento de Matemáticas, Universidad Autónoma de Madrid, Madrid 28050, Spain

Summary

Microarray experiments provide data on the expression levels of thousands of genes and, therefore, statistical methods applicable to the analysis of such high-dimensional data are needed. In this paper, we propose robust nonparametric tools for the description and analysis of microarray data based on the concept of functional depth, which measures the centrality of an observation within a sample. We show that this concept can be easily adapted to high-dimensional observations and, in particular, to gene expression data. This allows the development of the following depth-based inference tools: (1) a scale curve for measuring and visualizing the dispersion of a set of points, (2) a rank test for deciding if 2 groups of multidimensional observations come from the same population, and (3) supervised classification techniques for assigning a new sample to one of G given groups. We apply these methods to microarray data, and to simulated data including contaminated models, and show that they are robust, efficient, and competitive with other procedures proposed in the literature, outperforming them in some situations.

Keywords: Classification; Data depth; High-dimensional data; Microarray; Rank test; Scale curve.

1. Introduction

DNA microarrays and oligonucleotide chips of high density are broadly used in modern biomedical research and can serve as a guide for the diagnosis and treatment of some diseases. One of their most interesting current applications is the characterization and classification of different types of cancers. Traditionally, tumors have been classified according to their morphologic appearance, but since tumors with similar histologic features often follow different clinical courses, a classification based on molecular analysis is more reliable and informative. Microarray analysis of cancer cells results in a better understanding

*To whom correspondence should be addressed.

of molecular variation between tumors, thus they provide a more reliable classification and facilitate the development of more specific and efficient treatments (see Alon *and others*, 1999; Golub *and others*, 1999; Perou *and others*, 1999; Pollack *and others*, 1999; Ross *and others*, 2000; Singh *and others*, 2002; Dopazo, 2006).

Microarray experiments have attracted wide interest. Apart from the biological insight that they provide, they suggest numerous statistical problems in different areas such as image analysis, missing data imputation, clustering, discriminant analysis, design of experiments, and variable selection. Microarray data present the expression levels of many genes (that we will consider as variables) with respect to a number of observations (samples) and therefore, they can be considered as high dimensional. Many classical multivariate statistical procedures do not behave well when the dimension of the data is very high with respect to the sample size. In addition, for this type of data, it is very useful to introduce robust statistical techniques since outliers are difficult to detect and they can affect the analysis in many different ways.

In this paper, we propose robust depth-based statistical tools for the analysis of microarray data. The idea of depth for multivariate data provides a way of measuring how representative or central an observation is within a sample. For a given sample or distribution P, a notion of depth assigns for every observation $\mathbf{x}$ a real number $D(\mathbf{x}, P)$ satisfying that the closer a point is to the mass center the higher its depth is. Based on a notion of depth, a sample of multivariate points can be ordered from center outward and robust statistics such as the median or trimmed mean can be defined. Most of the notions of multivariate depths introduced in the literature (see Tukey, 1975; Liu, 1990; Liu *and others*, 1999) are not tailored to high-dimensional data and are intractable when the dimension is greater than 4. López-Pintado and Romo (2007, 2009) proposed a nonparametric robust methodology for analyzing functional data based on new notions of depth that can be easily adapted to high-dimensional data and are computationally feasible. These new ideas can be used to define the most representative sample within a group of observations of high dimension (e.g., the expression levels of a set of genes in a tumor type affecting a group of individuals). In addition, a scale curve can be defined for this type of data providing a tool for measuring and visualizing the dispersion of a sample of observations. We also propose a rank test for microarray data to decide whether 2 groups of samples come from the same "population" (e.g., type of cancer). Finally, we have adapted the depth-based classification techniques for functional data introduced in López-Pintado and Romo (2006) to high-dimensional data. To assess the performance of these techniques, we have applied them to real and simulated data. In particular, our results show that the classification procedures we propose for gene expression data are often more robust than the ones frequently used in the literature (see Dudoit *and others*, 2002).

2. Representation of the data and a depth measure

Let $\mathbf{y}_1, \ldots, \mathbf{y}_n$ be a sample of points in $\mathbb{R}^d$, where d is much larger than the size of the sample n ($n << d$). Denote by $y(k)$ the kth component of the vector $\mathbf{y}$. A way of representing or visualizing high-dimensional data is using parallel coordinates (see Inselberg, 1985; Wegman, 1990), where the d axes are now parallel and equidistant, and the coordinates of a d-dimensional vector are represented as points on these axes connected by straight lines, as shown in Figure 1, left panel. The ordering of the genes in the x-axis is arbitrary, but all the concepts introduced in this paper are invariant under permutations of the genes (see López-Pintado and Romo, 2009); therefore, the results are not affected by the arbitrary ordering. In what follows, we will denote by $y_i(k)$ the level of expression of the kth gene (variable) in the ith sample (observation).

Although, as stated before, several notions of depth for multivariate data have been proposed (see for a review, Liu *and others*, 1999; Zuo and Serìng, 2000), most of them are computationally intensive and are not appropriate for high-dimensional data. More recently, alternative notions of depth for functional

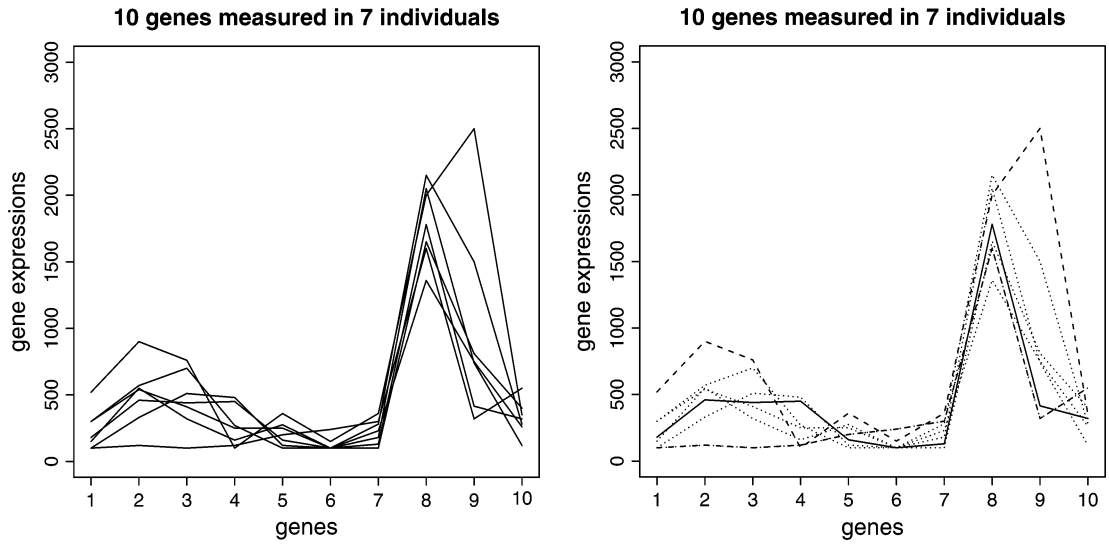

Fig. 1. Left panel: expression level of 10 genes, represented with numbers 1–10 in the x-axis, from 7 individuals. Right panel: the depth index of the solid curve with respect to the dashed ones is 5/10 since the expression levels of 5 of the 10 genes are inside the band determined by the dashed curves.

data have been reported in the literature (see Fraiman and Muniz, 2001; Cuevas *and others*, 2006, 2007; López-Pintado and Jörnsten, 2007; López-Pintado and Romo, 2009), which in general can be adapted to high-dimensional data without a large computational burden. In this paper, we will focus on the finite-dimensional version of the modified band depth (MBD) introduced in López-Pintado and Romo (2009) because it is easy to compute and particularly convenient for irregular curves. Nevertheless, all the statistical tools described here could be applied using any other feasible depth. Let $f_1, \ldots, f_n$ be a set of continuous functions defined on the interval I. The MBD of any f within the sample is as follows:

$$\mathrm{MBD}(f) = \binom{n}{2}^{-1} \frac{1}{\lambda(I)} \sum_{1 \leqslant i_1 < i_2 \leqslant n} \lambda(A(f; f_{i_1}, f_{i_2})), \tag{2.1}$$

where

$$A(f; f_{i_1}, f_{i_2}) = \left\{ t \in I : \min_{r=i_1,i_2} f_r(t) \leqslant f(t) \leqslant \max_{r=i_1,i_2} f_r(t) \right\} \tag{2.2}$$

and λ is the Lebesgue measure in $\mathbb{R}$.

The finite-dimensional version of this depth for a d-dimensional point $\mathbf{y}$ in the sample $\mathbf{y}_1, \ldots, \mathbf{y}_n$, is as follows:

$$\mathrm{MBD}_d(\mathbf{y}) = \binom{n}{2}^{-1} \sum_{1 \leqslant i_1 < i_2 \leqslant n} d^{-1} \times \sum_{k=1}^{d} I_{\{\min\{y_{i_1}(k), y_{i_2}(k)\} \leqslant y(k) \leqslant \max\{y_{i_1}(k), y_{i_2}(k)\}\}}. \tag{2.3}$$

The theoretical properties of this notion of depth are analyzed in López-Pintado and Romo (2009).

For microarray data, the depth of an observation $\mathbf{y}$ with respect to the sample $\mathbf{y}_1, \ldots, \mathbf{y}_n$ can be interpreted as the mean, over all possible pairs of observations, of the proportion of coordinates (or genes) whose expression is between the minimum and the maximum of the expressions of 2 observations from the sample $\mathbf{y}_1, \ldots, \mathbf{y}_n$.

Figure 1, right panel, shows the way of calculating the depth MBD_d of our illustrative example. The depth index of the solid curve with respect to the 2 dashed curves is defined as the proportion of its coordinates that are inside the band determined by the 2 dashed ones (in this case, 5 over 10). To obtain the depth of the solid curve with respect to the sample, we would have to consider all the possible bands determined by pairs of curves from the sample and calculate the mean of these indexes. It can be easily shown that the computational cost of the MBD of n d-dimensional points is $O(n^2 \cdot d)$. For instance, the average CPU time to compute MBD_d in data sets of $n = 25$ points in dimension $d = 250$ was 0.084s and in data sets of $n = 50$ points in dimension $d = 500$ was 1.258s, using a personal computer with an Intel Core 2 processor, 2.40 GHz and 2.00 GB of RAM memory.

3. Statistical tools

The notion of depth provides an order in a high-dimensional data sample, and therefore, robust statistics, such as the median or the trimmed mean, can be defined. Let $\mathbf{y}_{(1)}, \ldots, \mathbf{y}_{(n)}$ be the ordered sample from the deepest observation(s) to the most extreme one(s).

3.1 *The scale curve*

The notion of the scale curve introduced by Liu *and others* (1999) can be extended to functional data by defining the scale curve $A(p)$ of a set of functions $f_1, \ldots, f_n$ in $C(I)$ as the area of the band delimited by the $\lfloor np \rfloor$ most central curves, where $\lfloor np \rfloor$ is the largest integer smaller than np (see López-Pintado and Romo, 2007). The scale curve measures the increase in the area of the band determined by the fraction p most central curves, where p moves from 0 to 1. This notion can be easily adapted to high-dimensional data if we represent each vector as a curve based on the parallel coordinates representation and use the trapezoid formula to compute the area of the band.

The scale curve is an easy tool that helps visualize and compare the dispersion of different samples of high-dimensional data. See section 4 in the supplementary material available at *Biostatistics* online for intuitive examples showing its usefulness.

3.2 *The rank test*

One of the most important characteristics of the notion of depth is that it provides an extension of the idea of rank in the real line to higher dimensions. Let P be a population from which a sample of n points is drawn, and for a given data point $\mathbf{x}_i$, let $R(P_n, \mathbf{x}_i)$ be the proportion of observations from the sample with depth smaller than or equal to the depth of x_i. We can order the observations x_i according to increasing values of R, assigning them an integer rank from 1 to n. The higher the rank of an observation the deeper it is within the sample. Liu and Singh (1993) generalized to multivariate data the univariate Wilcoxon rank test through the order induced by a multivariate depth. López-Pintado and Romo (2009) extended this rank test to functional data, and in this paper, we have adapted it to high-dimensional data, in particular, to microarray data, using the modified band depth MBD_d defined in Section 2.

This test can be summarized as follows. Let $\mathbf{x}_1, \ldots, \mathbf{x}_n$ be a sample of high-dimensional vectors from population P_1 and let $\mathbf{y}_1, \ldots, \mathbf{y}_m$ be a sample of vectors from population P_2. Assume that there is a third reference sample $Z = \{\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_{n_0}\}$ from 1 of the 2 populations, for example, P_1, with $n_0 > n, m$. Let P_{n_0} be the corresponding empirical distribution. Order the values $R(P_{n_0}, \mathbf{x}_i)$ and $R(P_{n_0}, \mathbf{y}_i)$, from smallest to highest, giving them a rank from 1 to $n + m$.

The statistic used to test H_0: $P_1 = P_2$ is as follows:

$$W = \sum_{j=1}^{m} \text{ranks}(R(P_{n_0}, \mathbf{y}_j)), \tag{3.1}$$

whose distribution under H_0 is that of the sum of m elements drawn without replacement from $\{1, 2, \ldots, n+m\}$. We reject the null hypothesis (that both groups come from the same population) when W is smaller than the critical value.

3.3 *Classification methods*

In this section, we describe the methods that will be used to classify microarray data. The first 2 are based on data depth (in particular, on MBD_d) and are derived from the classification rules introduced for functional data in López-Pintado and Romo (2006): the distance to the trimmed mean, DS, and the weighted trimmed average distance, TAD. Both approaches are based on obtaining a "similarity" or "distance" measure between the new observation to be classified and the most representative data points in each group, given by the functional depth that naturally measures how representative an observation is in the class it belongs to. These procedures are tested and, following the exhaustive study of Dudoit *and others* (2002), are compared to the k nearest neighbors method (kNN) and the diagonal linear discriminant analysis (DLDA), classification methods commonly used in the literature that often provide the best results (Romualdi *and others*, 2003; Wessels *and others*, 2005; Tárraga *and others*, 2008).

Distance to the trimmed mean, DS. Let $\mathbf{x}_{(1)}, \ldots, \mathbf{x}_{(n)}$ be the observations in the sample ordered from center outward using any notion of depth (i.e., starting with the deepest one(s) and ending with the shallowest one(s)). The α-trimmed mean is defined as the average of the $n - \lfloor n\alpha \rfloor$ deepest points:

$$\widehat{\mathbf{m}}_n^{\alpha} = \frac{\sum_{i=1}^{n-\lfloor n\alpha \rfloor} \mathbf{x}_{(i)}}{(n - \lfloor n\alpha \rfloor)}. \tag{3.2}$$

The proposed method for the assignment of a new observation to one of G given groups of data, $A_1, \ldots, A_G$, consists of finding the distance from the new observation to the trimmed mean in each group and classifying it in the group that minimizes such distance. This method is denoted as DS. For a particular α, we denote it as DS_α.

Weighted average distance, TAD. This second classification method is based on the weighted average distance of an observation $\mathbf{x}$ to a given group A_g, which we define as the sum of the weighted distances from $\mathbf{x}$ to each element in the group. The weights are computed using the depth of each observation with respect to its own group. Thus, the deepest points have a larger influence on the final distance. Let $A_g = \{\mathbf{x}_1, \ldots, \mathbf{x}_{n_g}\}$ be a group from population P_g. The weighted average distance from $\mathbf{x}$ to A_g is defined as follows:

$$AD(\mathbf{x}, A_g) = \frac{\sum_{i=1}^{n_g} d(\mathbf{x}, \mathbf{x}_i) D(\mathbf{x}_i, P_g)}{\sum_{i=1}^{n_g} D(\mathbf{x}_i, P_g)}, \tag{3.3}$$

where n_g is the size of group A_g and D is any depth notion. The observation $\mathbf{x}$ will be assigned to the group A_k if $AD(\mathbf{x}, A_k) = \min_{g=1,\ldots,G} \{AD(\mathbf{x}, A_g)\}$.

This index of cluster membership depends on the group size. This can be overcome taking into account only of the m deepest points in each group, $x_{(1)}, \ldots, x_{(m)}$, to compute the weighted average distance

(where $m \leqslant n_1, n_2, \ldots, n_g$). This version of the procedure is denoted as TAD, standing for (weighted) trimmed average distance:

$$\mathrm{TAD}(\mathbf{x}, A_g) = \frac{\sum_{i=1}^{m} d(\mathbf{x}, \mathbf{x}_{(i)}) D(\mathbf{x}_{(i)}, P_g)}{\sum_{i=1}^{m} D(\mathbf{x}_{(i)}, P_g)}. \tag{3.4}$$

The other 2 methods that we will apply to the data are kNN and DLDA.

k nearest neighbors, kNN. The method is based on computing the values of a distance function (e.g., the Euclidean distance or one minus the correlation coefficient) for every pair of data points, which correspond to the expression levels of different genes in different (e.g., tumoral) samples. The kNN rule, introduced by Fix and Hodges (1951), classifies each element in the test set by finding the k nearest observations in the training set and assigning the observation to the most frequent class among these k neighbors. The value of k is selected by cross-validation.

Diagonal linear discriminant analysis, DLDA. This is a particular case of a maximum likelihood discriminant rule, which assigns the observation $\mathbf{x}$ to the group where its likelihood is maximized. If the conditional densities of the different groups are unknown but normal, they can be estimated using the training set, where only the mean and the covariance matrix of each group have to be estimated. The DLDA method assumes in addition that all the groups have the same diagonal covariance matrix, thus the discriminant rule is easy to compute.

3.4 *Simulation results*

To validate the procedures presented above, we carried out an extensive simulation study by generating data sets from different schemes. The main findings in this study can be summarized as follows. (1) In data sets generated from distributions with the same or similar covariance structure, the scale curves are indistinguishable, whereas when the groups come from distributions with different covariance structure, the scale curves are distinct. (2) The rank test is able to detect differences between groups coming from distinct populations. (3) We have tested the performance of the depth-based classification methods, also considering contaminated models and show that they are competitive with other procedures frequently used in the literature. See section 3 in the supplementary material available at *Biostatistics* online, for a detailed description of the models, the simulation process, and the results obtained. The methods proposed were implemented in R (code available in section 5 in the supplementary material available at *Biostatistics* online). For the comparison of our classification techniques to kNN and DLDA, we used the code included in the R packages "class" and "sma," respectively.

4. Applications

We have analyzed the following publicly available gene expression data sets.

4.1 *Leukemia*

This microarray data set (Golub *and others*, 1999) comes from a study of the expression levels of 6817 genes, using Affymetrix high-density oligonucleotides arrays, in 2 types of acute leukemia: lymphoblastic (ALL) and myeloid (AML), and consists of 47 ALL samples, comprising 38 from B cells and 9 from T cells, and 25 AML samples. The data were preprocessed and filtered following the 3-step procedure described by Dudoit *and others* (2002): (1) genes with values less than 100 or larger than 16 000 are

thresholded, (2) genes with max/min $\leqslant$ 5 or (max $-$ min) $\leqslant$ 500 (where max and min refer to the maximum and minimum expression levels of a given gene across the samples) are discarded, and (c) the data are transformed taking the base 10 logarithm. In addition, to prevent a single experiment from dominating in the analysis, we standardized each experiment to zero mean and unit variance across the genes, as in Dettling and Bühlmann (2002). With these steps, the dimension of the data is reduced to 3571 genes. We analyzed this data set, first considering only 2 groups, corresponding to the 2 types of leukemia, AML and ALL, and second, distinguishing the 3 groups, AML, B-ALL, and T-ALL.

In each case, we selected the 50 most representative genes, following the B/W criterion described in Dudoit *and others* (2002). Figure 2(a) shows the scale curve for such genes, for individuals with AML and for individuals with AML. The dispersion of the gene expression curves of individuals in the ALL group is greater than that of individuals in the AML group. Figure 3(a) shows the 3 scale curves for the corresponding 50 most representative genes when considering the 3 groups. Here, the B-ALL samples show a greater variability than the AML samples, whereas both have a higher variability than the T-ALL samples. This could be partially caused by the difference in class sizes.

We have also applied the rank test described in Section 3.2 to this data set. Since the number of observations in the ALL group is bigger than the number of observations in the AML group, we have considered the AML group and 22 randomly selected observations from the ALL group as the test sets, and the remaining 25 observations from the ALL group as the reference set. The p value obtained is very close to 0 (in the order of 10^{-13}), so we reject the null hypothesis that both groups of curves come from the same population.

For the 3-group study, we tested the hypothesis of having a common parent distribution. For every possible pair of groups, we ran the rank test choosing test groups of sample size 9 and using the remaining observations from the biggest group as the reference group. The null hypothesis in all the cases is rejected with p values in the order of 10^{-5}. Notice that this is not a multiple testing, and therefore, we only run pairwise comparisons. As an exercise, we also tested the equality of the parent population of 2 samples, randomly selected from the same group and obtained, as expected, nonsignificant results.

To estimate the classification error distribution, we used internal cross-validation so as to avoid an overoptimistic estimation of the error rate (see Simon *and others* (2003)). We randomly chose 2/3 (or 9/10) of the sample as the training set and the remaining 1/3 (or 1/10) of the data as the test set. We considered 200 simulations of the training and test sets, and in each iteration, the dimension was reduced

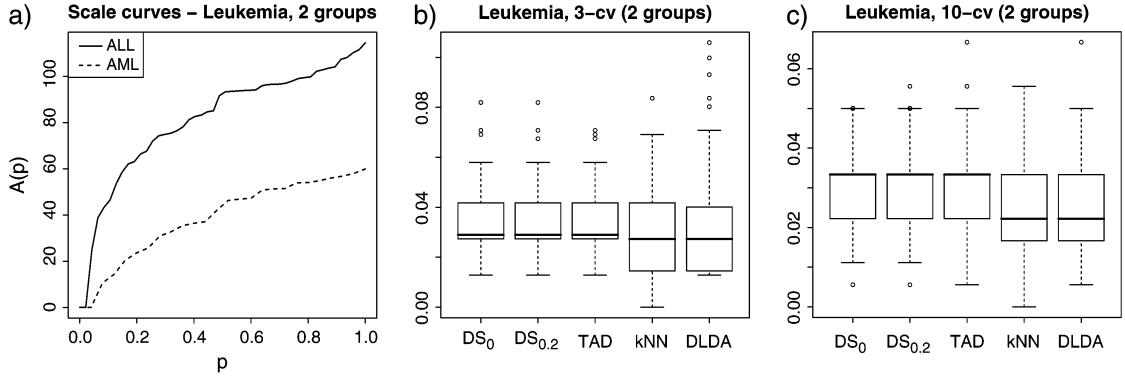

Fig. 2. (a) Scale curves for ALL samples (solid line), and AML samples (dashed line), showing that the ALL group has a larger variability. (b) and (c): Distributions (boxplots) of the error rates for the leukemia data set, considering 2 groups and 2 cross-validation parameters (3-CV and 10-CV, respectively).

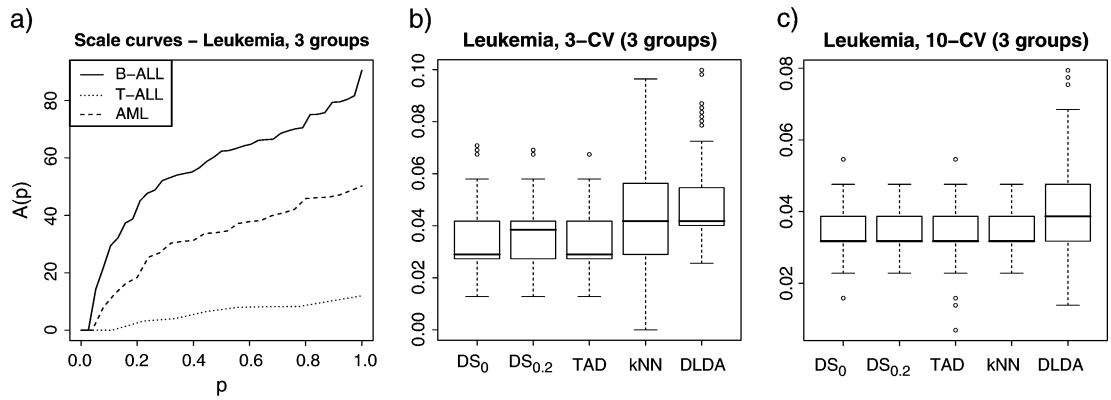

Fig. 3. (a) Scale curves for the B-ALL samples (solid line), T-ALL samples (dotted line), and AML samples (dashed line) showing the different variability of the classes. (b) and (c) Distributions of the error rates for the leukemia data set, considering 3 groups and 2 cross-validation parameters (3-CV and 10-CV, respectively).

to $d = 50$ genes following the criterion B/W applied to the corresponding training set (Dudoit *and others* (2002)). Finally, for each classification method, the error rate in the test set was computed as the proportion of wrongly classified elements.

The different classification methods that we used are the distance to the trimmed mean, considering $\alpha = 0$ (distance to the mean, DS_0) and $\alpha = 0.2$ (distance to the 0.2-trimmed mean, $\mathrm{DS}_{0.2}$), the weighted trimmed average distance, TAD, the nearest neighbors, kNN, and the diagonal linear discriminant analysis, DLDA. In all these methods, the depth and/or distance used are the MBD_d depth and the Euclidean distance.

We have described the error rate distribution using boxplots. The results for the 2-group case are shown in Figures 2(b) and (c), where we denote as 3-CV the cases in which the sample is divided in thirds and as 10-CV when it is divided in tenths. For this data set, all the methods have a good behavior, with low error rates. The best methods are kNN and DLDA, followed by TAD, but with only slight differences in the error rates.

For the 3-group case, the corresponding error rates are shown in Figures 3(b) and (c). Although all the methods again perform similarly, the best one is TAD, followed by the 2 versions of DS, in contrast to the previous results.

4.2 *Prostate*

The raw data (Singh *and others*, 2002) comprise the expression of 52 prostate tumors and 50 nontumor prostate samples, obtained using the Affymetrix technology. Similarly to the leukemia data set, we preprocessed the data by setting thresholds at 10 and 16 000 units, excluding genes whose expression varied less than 5-fold relatively, or less than 500 units absolutely, between the sample, applying a base 10 logarithmic transformation, and finally, standardizing each experiment to zero mean and unit variance across the genes. This leads to a data set containing 102 samples and 6033 genes, divided into 2 classes, tumoral (T) and normal (N), which we analyzed using the proposed techniques.

The scale curves for the 50 most representative genes, selected by the B/W criterion, show that the 2 classes have similar variability (see Figure 4(a)), but if we increase the dimension of the data set by including more and more genes, the normal class variability becomes slightly larger than that of the T class (see the supplementary material available at *Biostatistics* online, section 2.1).

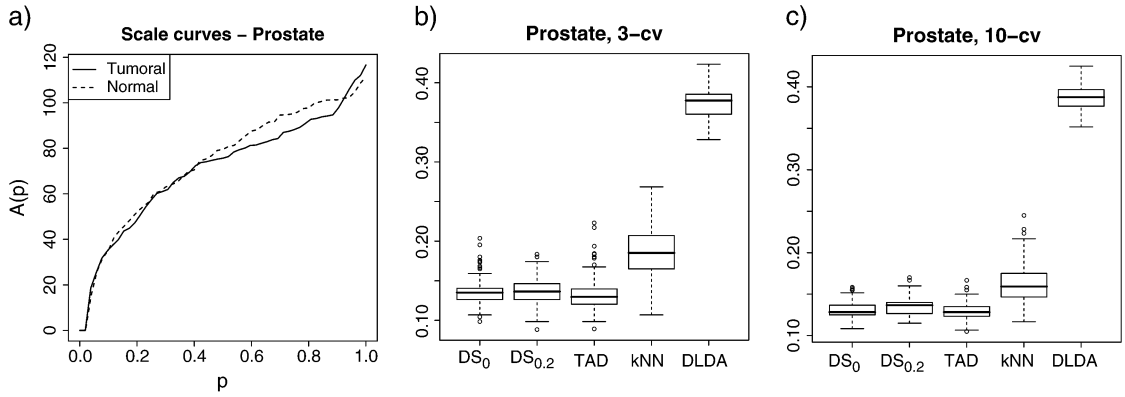

Fig. 4. (a) Scale curves for the tumoral samples (solid line), and normal samples (dashed line), from the prostate data set. Both groups show a similar dispersion for the selected 50 most representative genes. (b) and (c): Distributions of the error rates for the prostate data set and 2 cross-validation parameters (3-CV and 10-CV, respectively).

Considering random samples of size 25 from each group, and a reference set of 27 tumoral samples, the rank test shows that both classes are significantly different, with p values in the order of 10^{-7}.

Finally, we estimated the classification error distribution through internal cross-validation as in the previous example, and the results are shown in the boxplots of Figures 4(b) and (c); in this case, the best method is TAD, followed by the 2 variants of DS; DLDA has a poor performance, with large error rates; kNN improves DLDA results, but it is clearly outperformed by the methods based on data depth.

These results illustrate the good performance of the proposed techniques in different data sets. We have also tested them using the same data sets but increasing the dimension of the problem, that is, selecting a larger number of genes. For a detailed analysis of higher dimensional data, see section 2 in the supplementary material available at *Biostatistics* online.

5. CONCLUSIONS

We have presented several robust statistical tools based on the notion of data depth, which are applicable to high-dimensional data, and in particular to microarray data. First, we propose a scale curve to describe, visualize, and compare the dispersion within samples of high-dimensional points. Second, we introduce a rank test to decide whether there is significant evidence that 2 samples come from 2 different populations (e.g., genes from 2 types of tumor cells). Finally, we propose 2 classification methods that are tested and compared with the most frequently used algorithms in the literature. We analyze simulated and real data to illustrate the usefulness of the scale curve and the rank test. We also show that the proposed classification methods are accurate and comparable to other procedures used in the classification of microarray gene expression data. Nevertheless, for simulated models with outliers, our depth-based methods are robust and competitive in comparison with other techniques, outperforming them in some cases.

SUPPLEMENTARY MATERIAL

Supplementary material is available at http://biostatistics.oxfordjournals.org.

ACKNOWLEDGMENT

Conìct of Interest: None declared.

FUNDING

Spanish Ministry of Education and Science (BEC2002-03769, SEJ2005-06454, SEJ2007-67734, and ECO2008-05080) and Andalucian Government (SEJ2905).

REFERENCES

ALON, U., BARKAI, N., NOTTERMAN, D. A., GISH, K., YBARRA, S., MACK, D. AND LEVINE, A. J. (1999). Broad patterns of gene expression revealed by clustering analysis of tumour and colon tissues. *Proceedings of the National Academy of Sciences of the United States of America* **96**, 6745–6750.

CUEVAS, A., FEBRERO, M. AND FRAIMAN, R. (2006). On the use of bootstrap for estimating functions with functional data. *Computational Statistics and Data Analysis* **51**, 1063–1074.

CUEVAS, A., FEBRERO, M. AND FRAIMAN, R. (2007). Robust estimation and classification for functional data via projection-based depth notions. *Computational Statistics* **22**, 481–496.

DETTLING, M. AND BÜHLMANN, P. (2002). Supervised clustering of genes. *Genome Biology* **3**, 0069.1–0069.15.

DOPAZO, J. (2006). Bioinformatics and cancer: an essential alliance. *Clinical and Translational Oncology* **8**, 409–415.

DUDOIT, S., FRIDLYAND, J. AND SPEED, T. P. (2002). Comparison of discrimination methods for the classification of tumors using gene expression data. *Journal of the American Statistical Association* **97**, 77–87.

FIX, E. AND HODGES, J. L. (1951). Discriminatory analysis, nonparametric discrimination: consistency properties. *Technical Report*. Randolph Field, TX: USAF School of Aviation Medicine.

FRAIMAN, R. AND MUNIZ, G. (2001). Trimmed means for functional data. *Test* **10**, 419–440.

GOLUB, T. R., SLONIM, D. K., TAMAYO, P., HUARD, C., GAASENBEEK, M., MESIROV, J. P., COLLER, H., LOH, M. L., DOWNING, J. R., CALIGIURI, M. A. *and others* (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* **286**, 531–537.

INSELBERG, A. (1985). The plane parallel coordinates. *Visual Computer* **1**, 69–91.

LIU, R. Y. (1990). On a notion of data depth based upon random simplices. *Annals of Statistics* **18**, 405–414.

LIU, R. Y., PARELIUS, J. M. AND SINGH, K. (1999). Multivariate analysis by data depth: descriptive statistics, graphics and inference. *Annals of Statistics* **27**, 783–858.

LIU, R. Y. AND SINGH, K. (1993). A quality index based on data depth and multivariate rank tests. *Journal of the American Statistical Association* **88**, 252–260.

LÓPEZ-PINTADO, S. AND JÖRNSTEN, R. (2007). Functional analysis via extensions of the band depth. In: Liu, R., Strawderman, W. and Zhang, C. H. (editors), *Complex Datasets and Inverse Problems: Tomography, Networks and Beyond.* IMS Lecture Notes—Monograph Series, Volume 54. Beachwood, OH: Institute of Mathematical Statistics, pp. 103–120.

LÓPEZ-PINTADO, S. AND ROMO, J. (2006). Depth-based classification for functional data. In: Liu, R., Serìng, R. and Souvaine, D. L. (editors), *Data Depth: Robust Multivariate Analysis, Computational Geometry and Applications.* DIMACS Series in Discrete Mathematics and Theoretical Computer Science, Volume 72. Providence, RI: American Mathematical Society, pp. 103–119.

LÓPEZ-PINTADO, S. AND ROMO, J. (2007). Depth-based inference for functional data. *Computational Statistics and Data Analysis* **51**, 4957–4968.

LÓPEZ-PINTADO, S. AND ROMO, J. (2009). On the concept of depth for functional data. *Journal of the American Statistical Association* **104**, 718–734.

PEROU, C. M., JEFFREY, S. S., VAN DE RIJN, M., REES, C. A., EISEN, M. B., ROSS, D. T., PERGAMENSCHIKOV, A., WILLIAMS, C. F., ZHU, S. X., LEE, J. C. *and others* (1999). Distinctive gene expression patterns in human

mammary epithelial cells and breast cancers. *Proceedings of the National Academy of Sciences of the United States of America* **96**, 9212–9217.

POLLACK, J., PEROU, C., ALIZADEH, A., EISEN, M., PERGAMENSCHIKOV, A., WILLIAMS, C., JEFFREY, S., BOTSTEIN, D. AND BROWN, P. (1999). Genome-wide analysis of DNA copy-number changes using cDNA microarrays. *Nature Genetics* **23**, 41–46.

ROMUALDI, C., CAMPANARO, S., CAMPAGNA, D., CELEGATO, B., CANNATA, N., TOPPO, S., VALLE, G. AND LANFRANCHI, G. (2003). Pattern recognition in gene expression profiling using DNA array: a comparative study of different statistical methods applied to cancer classification. *Human Molecular Genetics* **12**, 823–836.

ROSS, D. T., SCHERF, U., EISEN, M. B., PEROU, C. M., REES, C., SPELLMAN, P., IYER, V., JEFFREY, S. S., VAN DE RIJN, M., WALTHAM, M. *and others* (2000). Systematic variation in gene expression patterns in human cancer cell lines. *Nature Genetics* **24**, 227–235.

SIMON, R., RADMACHER, M. D., DOBBIN, K. AND MCSHANE, L. M. (2003). Pitfalls in the use of DNA microarray data for diagnostic and prognostic classification. *Journal of the National Cancer Institute* **95**, 14–18.

SINGH, D., FEBBO, P. G., ROSS, K., JACKSON, D. G., MANOLA, J., LADD, C., TAMAYO, P., RENSHAW, A. A., D'AMICO, A. V., RICHIE, J. P. *and others* (2002). Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell* **1**, 203–209.

TÁRRAGA, J., MEDINA, I., CARBONELL, J., HUERTA-CEPAS, J., MÍNGUEZ, P., ALLOZA, E., AL-SHAHROUR, F., Vegas-Azcarate, S., Gotz, S., Escobar, P. *and others* (2008). GEPAS, a web-based tool for microarray data analysis and interpretation. *Nucleic Acids Research* **36**, W308–W314.

TUKEY, J. (1975). Mathematics and the picturing of data. In: *Proceedings of the 1975 International Congress of Mathematics*, Volume 2. Vancouver, pp. 523–531.

WEGMAN, E. (1990). Hyperdimensional data analysis using parallel coordinates. *Journal of The American Statistical Association* **85**, 664–675.

WESSELS, L. F. A., REINDERS, M. J. T., HART, A. A. M., VEENMAN, C. J., DAI, H., HE, Y. D. AND VAN'T VEER, L. J. (2005). A protocol for building and evaluating predictors of disease state based on microarray data. *Bioinformatics* **21**, 3755–3762.

ZUO, Y. AND SERFLING, R. (2000). General notions of statistical depth functions. *Annals of Statistics* **28**, 461–482.

[*Received June 7, 2009; revised December 1, 2009; accepted for publication December 8, 2009*]