Working Paper Series, N. 3, April 2010

Clustering via nonparametric density estimation: an application to microarray data

Riccardo De Bin
Department of Statistical Sciences
University of Padua
Italy

Davide Risso
Department of Statistical Sciences
University of Padua
Italy

Abstract: Cluster analysis is a crucial tool in several biological and medical studies dealing with microarray data. Such studies pose challenging statistical problems due to dimensionality issues, being the number of variables much higher than the number of observations. Here, we present a novel approach to clustering of microarray data via nonparametric density estimation, based on the following steps: (i) selection of relevant variables; (ii) dimensionality reduction; (iii) clustering of observations in the reduced space. Applications on simulated and real data show promising results in comparison with those produced by two standard approaches, *k-means* and *Mclust*. In the simulation studies, our nonparametric approach shows performances comparable to those of models based on normality assumption, even in Gaussian settings. On the other hand, in two benchmarking real datasets, it outperforms the existing parametric approaches.

Keywords: Cluster analysis,dimensionality reduction, kernel method, microarray, nonparametric density estimation

Contents

Department of Statistical Sciences
Via Cesare Battisti, 241
35121 Padova
Italy

tel: +39 049 8274168
fax: +39 049 8274170
http://www.stat.unipd.it

Corresponding author:
Davide Risso
tel: +39 049 827 4111
davide.risso@unipd.it
http://www.stat.unipd.it/CV/davide

Clustering via nonparametric density estimation: an application to microarray data

Riccardo De Bin
Department of Statistical Sciences
University of Padua
Italy

Davide Risso
Department of Statistical Sciences
University of Padua
Italy

Abstract: Cluster analysis is a crucial tool in several biological and medical studies dealing with microarray data. Such studies pose challenging statistical problems due to dimensionality issues, being the number of variables much higher than the number of observations. Here, we present a novel approach to clustering of microarray data via nonparametric density estimation, based on the following steps: (i) selection of relevant variables; (ii) dimensionality reduction; (iii) clustering of observations in the reduced space. Applications on simulated and real data show promising results in comparison with those produced by two standard approaches, *k-means* and *Mclust*. In the simulation studies, our nonparametric approach shows performances comparable to those of models based on normality assumption, even in Gaussian settings. On the other hand, in two benchmarking real datasets, it outperforms the existing parametric approaches.

Keywords: Cluster analysis,dimensionality reduction, kernel method, microarray, nonparametric density estimation

1 Introduction

The analysis of gene expression microarray data using clustering techniques plays an important role, for instance, in the discovery, validation, and understanding of various classes and subclasses of cancer. Three main types of statistical problems arise in classification of cancer samples in a microarray experiment (Dudoit et al., 2002): (a) identification of new classes using gene expression profiles (cluster analysis/unsupervised learning); (b) classification of samples into known classes (discriminant analysis/supervised learning); (c) identification of "marker" genes which characterize the difference among the classes (variable selection). For a recent review of clustering techniques for gene expression data, see Kerr et al. (2008).

In this paper, we present a method to handle simultaneously problems (a) and (c) previously mentioned. In particular, we discuss an application of *pdfClust* (Azzalini

& Torelli, 2007) to microarray data. We compare it with the traditional algorithms, such as *k-means* algorithm and its direct competitor, *Mclust* (Fraley & Raftery, 2002, 2006), a state-of-the-art mixture-model-based clustering tool. We follow the strategy presented in McLachlan et al. (2002), which consists on a data dimensional reduction in order to focus on most significant dimensions. We modify their method by using a nonparametric approach, achieving improvements in clustering of samples both in simulated and in real experiments.

The paper is organized as follows. In Section 2, we present the novel algorithm to clustering of microarray data. In Section 3, we describe two simulation models, while in Section 4 we present two real experiments. Section 5 contains the results based on simulated and real data. Finally, some conclusions are given in Section 6.

2 A novel algorithm to clustering of expression data

In recent years, the improvement of computational resources has enabled to pursue new clustering techniques, or to develop ideas putted aside due to computational problems. In this context, Azzalini & Torelli (2007) propose a method, called *pdfClust*, based on a nonparametric estimate of the underlying density function (the method is briefly reviewed in Appendix A).

Unfortunately, *pdfClust*, in order to compute Delaunay triangulation, exploits the *Quickhull* algorithm (Barber et al., 2006), which, computationally speaking, represents a critical issue in problems characterized by high dimensionality.

In this work, we address the problem of clustering high-dimensional data in a nonparametric fashion, following the stream of the algorithm by McLachlan et al. (2002). McLachlan et al. (2002) propose a mixture model-based approach to cluster microarray expression data. Their scheme accounts for gene selection, through mixtures of t distributions, and dimensionality reduction, through a mixture of factor analyzers. More precisely, they select a gene on the basis of a likelihood ratio statistic for testing one versus two components in the mixture model. In the second step of their algorithm, they cluster the samples by fitting a two-component mixture of factor analyzers.

Although McLachlan et al. (2002) approach sounds as a good possibility to cluster samples in a high-dimensional space, there are two main limitations. Firstly, the parametric assumptions about clusters distributions can be restrictive (Li et al., 2007); for example, two Gaussian random variables can result in a single mode (one cluster) or even a two component multivariate Gaussian mixture can lead to more than two modes (Li et al., 2007). Moreover, it needs pre-specification of the number of the mixture components. This, from an unsupervised perspective, which assumes that the true number of clusters is unknown, represents a serious limitation.

To be consistent with microarray applications, we will use hereafter the typical microarray terminology: we will denote with "genes" the p variables and with "samples" the n observations. Nonetheless, it should be clear that the approach proposed here is not limited to microarray data, but, in principle, it could be applied to every application with "large p, small n".

Our approach can be summarized as follows: (i) cluster samples p times using the

univariate distribution of each gene and select for the subsequent analyses the genes, p' say, that recognize two or more groups in the data; (ii) reduce dimensionality from p' to p'' with some data reduction technique; (iii) apply *pdfClust* algorithm in the p''-dimensional space.

As for step (i), i.e., *Gene selection*, we consider a gene relevant if its values in one category (healthy, say) are different from the ones in the other category (unhealthy, say) or categories. From another point of view, this means that the samples representing the healthy subjects are separated from the unhealthy ones, or, more simply, the samples are in different clusters. In this way, it seems reasonable to apply a cluster method to each gene, and retain as relevant those genes for which the method recognizes different clusters. In a nonparametric framework, we can apply to each gene *pdfClust*, taking advantage of the self-detection number of clusters feature. We select only the genes for whose the method recognizes two or more clusters.

As for step (ii), i.e., *Dimensionality reduction*, considered if the selected genes are still too many, we propose to keep the first principal components, as in Azzalini & Torelli (2007). The principal component analysis is a very simple procedure which reduces the dimension of a data set of a large number of interrelated variables, preserving as much as possible of the data set variation. Since it has no requirements about the data distribution, it is consistent with our non-parametric strategy. Actually, there are no guarantees that principal components preserve the cluster structure in the reduction of original dimension of data, as shown by Menardi (2006). Alternative methods have been proposed, based upon the idea of reduction pursuit (Friedman & Tukey, 1974). Menardi (2006) suggests to use the critical window width (Silverman, 1981) as projection index.

3 Simulation models

In this Section, we evaluate our proposal by mean of simulated data. For simulating data with structure similar to that of real microarray experiments, we use two schemes, i.e., the Gamma-Gamma model (Kendziorski et al., 2003) and the Normal-Uniform model (Garrett & Parmigiani, 2003). In Appendix B is reported a complementary simulation experiment involving multivariate normal distributions, to evaluate the behaviour of *pdfClust* in a multidimensional space, designed to evaluate how the increasing dimension and the different correlation structures influence its performances.

3.1 Gamma-Gamma (GG) model

The samples are assumed to be independently generated from Gamma distributions with a constant shape parameter α and gene-specific random scales λ_i, $i = 1, \ldots, p$; λ_i is assumed to have a Gamma distribution with shape hyperparameter α_0 and scale hyperparameter ν. The genes are generated to be either "equally expressed" (i.e. one group) or "differentially expressed" (i.e. two groups) among the samples. We generated $n = 100$ samples and $p = 2,000$ genes, each with probability 0.05 of being differentially expressed. We fixed parameter values as suggested by Chiogna et al. (2009). We applied the algorithm stated in Section 2 to the data matrix obtained,

selecting a number of relevant genes and using the first three principal components as input for the *pdfClust* algorithm. We repeated this procedure $B = 5,000$ times.

3.2 Normal-Uniform (NU) model

The model deals with k-class classification of samples, for general k. It is based on a mixture of Normal and Uniform distributions. We exploit the model to simulate gene expressions for a three-class problem, similar to that of the leukaemia data (see Section 4.2).

Let us denote with x_{ji} the measured intensity of gene j in sample i, $j = 1, \dots, p$, $i = 1, \dots, n$. We define three categories from which x_{ji} can arise and use e_{ji} to represent them: (i) $e_{ji} = -1$, i.e., gene j has abnormally low expression in sample i (down-regulation); (ii) $e_{ji} = 0$, i.e., gene j has normal expression in sample i; (iii) $e_{ji} = 1$, i.e., gene j has abnormally high expression in sample i (up-regulation). For each gene j,

$$x_{ji}|(e_{ji} = e) \sim f_{e,j}, \quad e \in \{-1, 0, 1\}.$$

Following Garrett & Parmigiani (2003), we use a Uniform distribution for $f_{-1,j}$ and $f_{1,j}$ and a Normal distribution for $f_{0,j}$. More specifically,

$$\begin{aligned} f_{-1,j} &= \mathcal{U}(-\kappa_j + \alpha_i + \mu_j, \alpha_i + \mu_j), \\ f_{0,j} &= \mathcal{N}(\alpha_i + \mu_j, \sigma_j), \\ f_{1,j} &= \mathcal{U}(\alpha_i + \mu_j, \alpha_i + \mu_j + \kappa_j), \end{aligned}$$

where μ_j represents the gene-effect and α_i the sample-effect for the normal expression level (see Garrett & Parmigiani, 2003, for details). The Authors justify the choice of distributions arguing that, for normally expressed genes, the differences in observed values are due mainly to noise introduced in the experimental stage, while the Uniform distribution may reflect the failure of a biological mechanism that controls the expression level. We simulated data from the model in a hierarchical framework, with the following initial parameter values:

$$\begin{aligned} \mu_j &\sim \mathcal{N}(7.5, 1.5), \\ \sigma_j^{-1} &\sim \mathcal{G}(2, 1), \\ \alpha_i &\sim \mathcal{N}(0, 1), \\ \kappa_j &\sim \mathcal{E}(1) + 7\sigma_j, \end{aligned}$$

where $\mathcal{G}$ denotes the Gamma and $\mathcal{E}$ the Exponential distribution. We simulated $B = 5,000$ datasets of $n = 100$ samples, $p = 1,000$ genes and $m = 3$ classes defined as follows: *class 1* consists of 40 samples with 150 up-regulated and 50 down-regulated genes; *class 2* consists of 40 samples with 50 down-regulated genes; *class 3* consists of 20 samples with neither up- nor down-regulated genes. Note that classes 2 and 3 are "close" to each other with respect to class 1.

4 Real data

Along with simulations, we consider two benchmarking real datasets, studied before by several Authors (Alon et al., 1999; Chow et al., 2001; Dudoit et al., 2002; Getz

et al., 2000; Golub et al., 1999; McLachlan et al., 2002), to which we will refer as the colon data and the leukaemia data.

4.1 Colon data

Alon et al. (1999) used Affymetrix oligonucleotide arrays to measure the expression of 6,500 human genes in 40 tumor and 22 normal colon tissue samples. They focused on the subset of 2,000 genes with highest minimal intensity across the samples: the raw expression values of these 2,000 genes comprise our dataset. Following McLachlan et al. (2002) notation, we named 1-40 the tumor samples and 41-62 the normal ones. Before clustering the tissues, we pre-processed the raw intensities taking the logarithm and applying the quantile normalization (Bolstad et al., 2003), which is a standard choice for single-channel microarray technology.

4.2 Leukaemia data

Golub et al. (1999) studied the gene expression of two types of acute leukaemias, acute lymphoblastic leukaemia (ALL) and acute myeloid leukaemia (AML). Gene expression levels were measured using Affymetrix oligonucleotide arrays containing 6,817 human genes. The dataset comprises 47 cases of ALL (38 B-cell and 9 T-cell) and 25 cases of AML. The classification of samples is more difficult in this example than in colon data because it is much harder to classify between subclasses of the same plasticity than to distinguish between healthy and cancer tissues. Moreover, we have a typical hierarchical structure, since B-cell and T-cell are subclasses of the ALL class, harder to separate than AML from ALL. Following Dudoit et al. (2002), three preprocessing steps are applied to the intensity matrix: (a) thresholding, floor of 100 and ceiling of 16,000; (b) filtering, exclusion of genes with max/min$\leq$ 5 or (max $-$ min) $\leq$ 500; (c) base 10 log transformation. This procedure left us with 3,892 genes.

5 Results and discussion

5.1 Evaluation criteria

Both in simulated and in real data, we evaluated the performances of the methods by calculating the error rate (proportion of misclassified samples, ER), the sensitivity (SE) and specificity (SP). Moreover, in the simulation studies, we recorded the frequency of each method in finding the correct number of clusters (CC), and we evaluated the performance of the methods to select discriminant genes, considering the error rate in the classification of relevant genes, knowing *a priori* which genes have been generated to have different values among the groups.

Since class 2 and 3 of the Normal-Uniform model have been simulated to be close to each other, in this model we consider also the number of times in which each method is able to recognize two clusters (class 1 versus class 2-3) or three clusters.

5.2 Simulated data

5.2.1 GG model

Table 1 shows that both *pdfClust* and *Mclust* provide results surprisingly good in correct cluster recognition, low error rate and high sensitivity/specificity: this could be explained by an extreme distance between the two groups in the original p-dimensional space.

More interesting is the very different behaviour in the choice of the relevant genes: *pdfClust* is very good in recognizing them, with a very low error rate (about 8%), while *Mclust* shows a very high error rate (about 78%).

5.2.2 NU model

As expected, Table 2 shows that in this model both *pdfClust* and *Mclust* lead to higher classification errors than in GG model. Also in the gene selection step, both methods have difficulties in finding the relevant genes. Nevertheless, *pdfClust* outperforms *Mclust* according to the gene selection error rate ("RG" row in Table 2).

Mclust is able to recognize three clusters in the 39% and two clusters in the 34% of the simulations; *pdfClust* recognizes three clusters in the 19% and two clusters in the 47% of the simulations. On the other hand, the mean error rate of the final classification is 0.135 for *pdfClust* while for *Mclust* is 0.227.

5.3 Real data

5.3.1 Colon data

As described in Section 2, we analyzed the dataset, following three steps: (i) gene selection, (ii) dimensionality reduction, (iii) clustering of samples. Namely, the first step of the procedure consists in applying the cluster algorithm to the univariate distribution of each gene. The genes that show two or more clusters are considered for the further steps.

In the first step, *pdfClust* algorithm was able to recognize 84 genes, which discriminate data in two or more groups. We proceeded by considering the first three principal components of this reduced data-matrix. The procedure found three clusters, summarized in Table 3, which clearly correspond to biologically meaningful groups. The first cluster consists of tumor tissues (with 3 misclassified), while clusters 2 and 3 comprise normal tissues (with 5 misclassified). It is worth noting that six out of the eight misallocated samples (tumor tissues 30, 33 and 36 and normal tissues 48, 58 and 60) are found to be misclassified in several previous analyses, including McLachlan et al. (2002); Chow et al. (2001). As stated for instance in Chow et al. (2001), these six samples are likely to be wrongly labeled. Furthermore, Getz et al. (2000) reported that there was a change in the protocol during the experiments: tumor samples 1-11 and normal samples 41-51 were collected within the first protocol, while tumor samples 12-40 and normal samples 52-62 were collected within the second. Although for the tumor samples our approach did not recognize

any difference between the protocols, cluster 2 and cluster 3 split normal tissues in two groups according to the protocols.

In order to compare our approach to *Mclust*, we carried out a procedure analogous to the one described in Section 2, but using the normal-mixture model both in step (i) and (iii). In the first step, *Mclust* was able to find 369 discriminant genes. We considered the first three principal components of this sub-space for clustering. The procedure found two clusters, with a rather high missclassification error (see Table 4). We also applied a *k-means* algorithm to the entire dataset. The results of the three approaches are shown in Table 4. It can be seen that *k-means*, exploited in the original p-dimensional space, does not perform well. Moreover, *pdfClust* outperforms (in terms of error rate) *Mclust*, if one considers cluster 2 and 3 together as the normal samples.

As stated before, McLachlan et al. (2002) studied the same microarray dataset. They selected 446 relevant genes, achieving clusters that seem to recognize the change of protocol in the data structure, but fail to recognize the normal/tumor differences (McLachlan et al., 2002). Nevertheless, they achieved results slightly better than ours (ER$= 0.1$) considering a particular subspace: they clustered genes in 20 groups and considered only the second group (consisting of 24 genes) to cluster data (McLachlan et al., 2002). Although this approach leads to good results in this example, it seems difficult to reproduce the procedure in an unsupervised setting.

5.3.2 Leukaemia data

As stated in Dudoit et al. (2002), the Leukaemia dataset presents two different problems: an easier one, consisting in separating ALL from AML (two-class problem, hereafter) and an harder one, consisting in recognizing also the differences in B-cell and T-cell subclasses (three-class problem).

Again, we considered the strategy described in Section 2. In the variable selection step, *pdfClust* recognizes 313 discriminant genes. Note that the higher number of genes selected with respect to Colon data is consistent with the higher difficulty of the problem. We proceeded by considering the first three principal components of this subspace. *PdfClust* found two clusters, which clearly represent ALL and AML samples, with 4 AML samples classified as ALL and 5 ALL samples classified as AML, leading to a missclassification error rate of 0.125 (Table 5): *pdfClust* is able to solve the two-class problem, but it misses the three-class problem.

In the first step, *Mclust* failed to select relevant genes, recognizing as discriminant among the groups 3,119 out of 3,892 genes. Based on the first three principal components of the subspace spanned by these genes, *Mclust* clustered samples in four groups. We could interpret the merged cluster 1-2 as the ALL B-cell class, and cluster 4 as the AML class, while cluster 3 interpretation is less clear (Table 5). Although *Mclust* is able to find more than two clusters, it fails to distinguish between B-cell and T-cell classes, leading to hardly interpretable clusters.

Leukaemia dataset has been studied by McLachlan et al. (2002) as well. The Authors found 2,015 relevant genes after the variable selection step. For the two-class problem, their results were very good (only one sample misallocated), but they failed to solve the three-class problem.

It should be noted that, unlike our algorithm, the procedure used in McLachlan et al. (2002) needs the prior specification of the number of clusters, which is not desirable in an unsupervised learning, especially in cancer tissues classification, where one of the main goal is to find new subclasses of tumors.

6 Conclusions

Model-based approaches to clustering of data have received increasing attention in the last few years, as they provide a sound mathematical-based method.

Here, we have discussed a nonparametric density estimation-based algorithm for clustering microarray expression data. Our approach has shown promising results both in simulated data and in two real applications, with surprisingly good computational performances.

In our simulation experiments, we have found that *pdfClust* has performances comparable to those of *Mclust* even in a Gaussian setting. Moreover, the gene selection step is much more effective using *pdfClust* than using *Mclust* both in simulated and in real datasets. Here, "effective" means good results in terms of both dimension reduction (e.g. in Leukaemia data *pdfClust* selected 313 genes versus 3,119 of *Mclust*) and of correct selection (e.g. in GG model the gene selection error rate are 0.08 and 0.77, respectively).

Due to its nonparametric nature, *pdfClust* needs more observations (samples) than *Mclust* to perform well: in our experience, however, $n \geq 50$ is sufficient to have good performances. Fortunately, cancer microarray experiments often contain about 100 samples, and this number is expected to grow (Dudoit et al., 2002). Thus, the sample size should not be considered as an issue.

Although we have used here principal components in order to reduce the space dimension, future effort could be done in trying different approaches to this problem, as the reduction pursuit by Friedman & Tukey (1974) or the principal curves by Hastie & Stuetzle (1989). Nonetheless, in our case the principal component analysis gives good results (comparable to that obtain after a projection pursuit, Hyvarinen & Oja, 2000) and provides a low dimensional dataset on which directly apply *pdfClust*.

All the statistical analyses and simulations have been performed with R (R Development Core Team, 2009). Packages used: affy, deldir, fastICA, MASS, mclust, sm, snow, spdep.

The datasets used are both freely available as Bioconductor (Gentleman et al., 2004) packages ("colonCA" for Colon data and "golubEsets" for Leukaemia data).

Table 1: Simulation results for *pdfClust* and *Mclust* in GG model: correct number of cluster (CC), sensitivity (SE), specificity (SP), error rate (ER) and error rate in the selection of relevant genes (RG).

	PC		MC	
	mean	se	mean	se
SE	0.9877	0.0007	0.9991	0.0001
SP	0.9866	0.0008	0.9985	0.0004
ER	0.0128	0.0006	0.0012	0.0002
RG	0.0837	0.0061	0.7787	0.0093
CC	0.77		0.84	

Table 2: Simulation results for *pdfClust* and *Mclust* in NU model: rate of two clusters identification (CC2), rate of three clusters identification (CC3), error rate (ER) and error rate in the selection of relevant genes (RG).

	PC		MC	
	mean	se	mean	se
ER	0.135	0.004	0.227	0.005
RG	0.433	0.041	0.616	0.077
CC2	0.47		0.34	
CC3	0.19		0.39	

Table 3: Clusters found after *pdfClust* procedure in Colon data; tumor samples are labeled 1-40, normal samples 41-62; misallocated samples are shown in bold.

Cluster 1	1-6,8-19,21-29,31,32,34,35,37-40,**48,58,60**
Cluster 2	**7**,41-47,49-52
Cluster 3	**20,30,33,36**,53-57,59,61,62

Table 4: Confusion matrices for *pdfClust* (PC), *Mclust* (MC) and *k-means* (KM) with error rates (ER) for Colon data; in the first column "1" corresponds to tumor samples and "2" to normal.

	PC		MC		KM	
Real	1	2-3	1	2	1	2
1	35	5	29	11	23	17
2	3	19	12	10	6	16
ER:	0.13		0.37		0.37	

Table 5: Confusion matrices for *pdfClust* (PC), *Mclust* (MC) and *k-means* (KM) for Leukaemia data.

	PC		MC				KM		
Real	1	2	1	2	3	4	1	2	3
ALL B-cell	37	1	9	20	9	0	15	0	23
ALL T-cell	5	4	0	0	7	2	7	2	0
AML	4	21	0	2	1	22	1	23	1

Appendix

A pdfClust: an overview

In the literature, nonparametric cluster analyses based on mode identification have already been presented. See Li et al. (2007), Fraley & Raftery (2002), Banfield & Raftery (1993), Li & Zha (2006) and Banerjee et al. (2005). *pdfClust* (Azzalini & Torelli, 2007) starts from a quite simple idea, introduced by Hartigan (1975), who stated:

> Clusters may be thought of as regions of high density separated from other such regions by regions of low density.

These regions are achieved by "cutting" the density function computed out of observations by a level c, that varies through the algorithm.

More formally, consider a p-dimensional space, $\mathcal{X} \subseteq \mathbb{R}^p$. Let $x_1, \ldots, x_n$ be a vector of p-dimensional observations, $x_i \in \mathcal{X}$, for $i = 1, \ldots, n$. Starting from this vector, using a method of nonparametric density estimation, we can obtain $\hat{f}(x), x \in \mathcal{X}$, i.e. the empirical version of the density $f(x)$.

There is not a specific method for the nonparametric density estimation related to *pdfClust*, since the only restriction is that $\hat{f}(x_i) < +\infty$ for all $i = 1, \ldots, n$. This restriction is not discriminating, because almost all estimation techniques satisfy it. Following Azzalini & Torelli (2007), we choose a kernel method with Gaussian kernel and constant smoothing parameter $h = (h_1, \ldots, h_p)^\top$, with $h_j = \left(\frac{4}{(p+2)n}\right)^{1/(p+4)} s_j$, $j = 1, \ldots, p$, where s_j is the estimated standard deviation of the j-th variable. This choice is related to the minimization of the asymptotic integrated mean square error (Azzalini & Torelli, 2007), but tuned investigations should be done in order to achieve an optimal h for the specific problem. Empirically, Azzalini & Torelli (2007) realize that it is often advantageous to slightly shrink h toward zero, suggesting a shrinkage factor of 3/4.

Cutting the computed $\hat{f}(x)$ at a level $c \in [0; \max \hat{f}]$, they obtain m subspaces $\mathcal{M}_k$, $k = 1, \ldots, m$, of the sample space $\mathcal{X}$. Dropping the observations not belonging to $\cup_{k=1}^{m} \mathcal{M}_k$, they select only those observations x_i such that $\hat{f}(x_i) > c$. The observations belonging to the same $\mathcal{M}_k$ are connected by the Delaunay triangulation (see, e.g., de Berg et al., 2008) to form the "cluster cores". Finally, the unallocated observations are allocated by a classification method, based on nonparametric density estimation too: if x_0 is the unallocated observation, the estimated density $\hat{f}_k(x_0)$ based on the data already assigned to group k is computed, and x_0 is assigned to the group with highest ratio $\hat{f}_k(x_0) / \max_{l \neq k} \hat{f}_l(x_0)$. Finally, it is important to notice that *pdfClust* selects by itself the number of the clusters.

B Multivariate Normal Simulation

B.1 Settings

As we said, we want to check if the performance of *pdfClust* is comparable with that of the parametric method (McLachlan et al., 2002) in a multidimensional space

characterized by a complicated correlation structure. The simplest way to simulate this kind of space is by mean of multivariate normal variables.

Setting 1. We simulate $n_1 = 50$ samples from a multivariate normal distribution $N_p(\mu_1, \Sigma)$ and $n_2 = 50$ samples from a multivariate normal distribution $N_p(\mu_2, \Sigma)$, where

$$\mu_1 \sim N_p(\mathbf{7}, I_p), \qquad \mu_2 = \mu_1 + \eta, \text{ where } \eta \sim N_p(\mathbf{2}, I_p),$$

where I_p indicates the identity matrix of order p, and $\mathbf{7}$, $\mathbf{2}$ represents p-dimensional vectors of 7 and 2, respectively. The covariance matrix Σ has diagonal elements equal to 1 and the remaining elements randomly chosen uniformly between 0.6 and 0.9. The choice of the values of μ_1 and μ_2 are driven by real microarray datasets, which often (in the log scale) are centered around these values. Moreover, the "log-fold-change" (difference in logarithm) of 2 is a standard value in detecting relevant genes.

We replicate this simulation $B = 10,000$ times, applying on each replication *pdfClust*, *k-means* and *Mclust*. For each algorithm, we consider the specificity (SP), the sensitivity (SE), the error rate (ER) and the percentage of times in whose the algorithms find the two clusters (CC). We performed the analyses for increasing dimensions of the sample space, p, from 2 to 5.

Setting 2. Here, we are interested in seeing how the three methods work when the differences between the two groups are in only one dimension. In particular, we construct μ_1 in the same way as before, while

$$\mu_2 = \mu_1 + \eta', \text{ where } \eta' = (\eta_1, 0, \ldots, 0)^\top,$$

where $\eta_1 \sim N(2, 1)$. In this way, although the space dimension p grows, the differences useful to cluster the groups belong only to one dimension.

Setting 3. As stated before, microarray data are characterized by a rather complicated correlation structure. Indeed, the genes are often positively or negatively correlated (co-regulated genes). Thus, we fixed $p = 2$ and explored the behavior of the methods with different values of the correlation ρ, namely $\rho \in \{0.8, -0.8, 0\}$, considering three different scenarios: i) both genes are responsible for the differences in the two groups; ii) one gene is responsible for the differences in the two groups; iii) no differences between the two groups.

B.2 Results

The results of the first setting are summarized in Table 6 and in Figure 1. Panel (a) of Figure 1 shows the error rates of the three methods in the first setting. As we expected, the more p increases, the more the groups are well-separated, and hence, the more the error rate decreases. The same pattern holds for sensitivity and specificity (Figure 1 panels (b) and (c)). *PdfClust* seems to overperform *Mclust* in recognizing the two clusters in the population for each considered p. *Mclust* provides better sensitivity, specificity and error rate, but this could be a consequence of the fact that *pdfClust* analyzes also the cases where the clusters are closer to each other and it is more difficult to correctly allocate the samples. Surprisingly, with $p = 2$

both algorithms fail to recognize the data structure, finding two clusters respectively in about the 20% and the 1% of the simulations. We exploited with more details the case of $p = 2$, with various correlations, in Setting 3.

As we move towards the second setting (Table 7 and Figure 2), we find a similar situation, with *pdfClust* performing better in recognizing the corrected clusters, while *Mclust* provides better sensitivity, specificity and error rate. As in the previous setting, it is worth noting that the worse values of *pdfClust* are probably due to the effort of the method in clustering groups very close to each other: the closeness between the values of sensitivity, specificity and error rate of *pdfClust* and *k-means*, that is constrained to build two groups, seems to support our consideration. As we expected, all methods behave worse as the number of "noise" dimensions increases.

In setting 3, we try to investigate more carefully the influence of different correlations in the performance of the methods. We focus on $p = 2$. Table 8 shows the performance in finding the correct number of clusters of *pdfClust* and *Mclust* algorithms (since the number of cluster is to be specified in *k-means*, this analysis is pointless for this algorithm). The first part of Table 8 refers to a situation in which the data are generated from a unique population; both algorithms, in general, correctly recognize that there are no clusters in the data; nevertheless, *Mclust* outperforms *pdfClust* with strong correlations. As in previous settings, we find again the same behaviour, with *pdfClust* being more effective in finding different clusters when they are close to each other (it performs better than *Mclust* when there is only one significant dimension) while *Mclust* works better when facing strong differences: notice that with negative correlation, when the distance between groups is more evident (see Figure 3), it performs better than *pdfClust*.

Interestingly, in all settings *pdfClust* performances are fairly comparable with those of *Mclust* (if not better), although the former does not assume normality of data.

Table 6: sensitivity (SE), specificity (SP), error rate (ER) and frequency in finding correct number of clustering (CC) for *pdfClust* (PC), *Mclust* (MC), *k-means* (KM). Setting 1.

	PC		KM		MC	
p=2	mean	se	mean	se	mean	se
SE	0.7249	0.0028	0.7309	0.0009	0.6847	0.0212
SP	0.7237	0.0028	0.7331	0.0009	0.6583	0.0221
ER	0.2757	0.0011	0.2680	0.0004	0.3285	0.0084
CC	0.23		1		0.01	
p=3	mean	se	mean	se	mean	se
SE	0.7526	0.0020	0.7997	0.0008	0.9162	0.0015
SP	0.7546	0.0019	0.7992	0.0008	0.9173	0.0013
ER	0.2464	0.0009	0.2006	0.0005	0.0832	0.0009
CC	0.44		1		0.24	
p=4	mean	se	mean	se	mean	se
SE	0.8663	0.0012	0.8658	0.0008	0.9516	0.0006
SP	0.8685	0.0012	0.8658	0.0008	0.9531	0.0006
ER	0.1326	0.0008	0.1342	0.0006	0.0477	0.0004
CC	0.70		1		0.52	
p=5	mean	se	mean	se	mean	se
SE	0.8990	0.0008	0.9091	0.0005	0.9509	0.0006
SP	0.9017	0.0008	0.9100	0.0005	0.9517	0.0006
ER	0.0996	0.0004	0.0904	0.0003	0.0487	0.0004
CC	0.80		1		0.50	

Table 7: sensitivity (SE), specificity (SP), error rate (ER) and frequency in finding correct number of clustering (CC) for *pdfClust* (PC), *Mclust* (MC), *k-means* (KM). Setting 2.

	PC		KM		MC	
p=2	mean	se	mean	se	mean	se
SE	0.6866	0.0026	0.7130	0.0009	0.9089	0.0016
SP	0.6915	0.0026	0.7128	0.0009	0.9112	0.0015
ER	0.3110	0.0013	0.2871	0.0005	0.0900	0.0011
CC	0.28		1		0.34	
p=3	mean	se	mean	se	mean	se
SE	0.5946	0.0032	0.6114	0.0009	0.6165	0.0132
SP	0.5870	0.0033	0.6119	0.0009	0.6298	0.0129
ER	0.4092	0.0013	0.3883	0.0005	0.3768	0.0059
CC	0.18		1		0.02	
p=4	mean	se	mean	se	mean	se
SE	0.5685	0.0028	0.5684	0.0009	0.7636	0.0203
SP	0.5524	0.0029	0.5666	0.0009	0.7749	0.0198
ER	0.4395	0.0010	0.4325	0.0004	0.2307	0.0146
CC	0.18		1		0.01	
p=5	mean	se	mean	se	mean	se
SE	0.5674	0.0023	0.5811	0.0010	0.6714	0.0900
SP	0.5651	0.0024	0.5768	0.0010	0.7886	0.0484
ER	0.4338	0.0014	0.4210	0.0007	0.2700	0.0053
CC	0.13		1		0.01	

Table 8: Frequency of correct clusters (with standard errors in parenthesis) in different settings for *pdfClust* (PC), *Mclust* (MC), *k-means* (KM). 'diff' states the number of components of the mean vector different between the groups and ρ the correlation between the components.

diff	ρ	PC	MC
0	0.8	0.815 (0.004)	0.985 (0.001)
0	-0.8	0.819 (0.004)	0.985 (0.001)
0	0	0.921 (0.003)	0.989 (0.001)
1	0.8	0.186 (0.004)	0.015 (0.001)
1	-0.8	0.589 (0.005)	0.720 (0.004)
1	0	0.222 (0.004)	0.188 (0.004)
2	0.8	0.612 (0.005)	0.161 (0.004)
2	-0.8	0.741 (0.004)	0.819 (0.004)
2	0	0.503 (0.005)	0.558 (0.005)

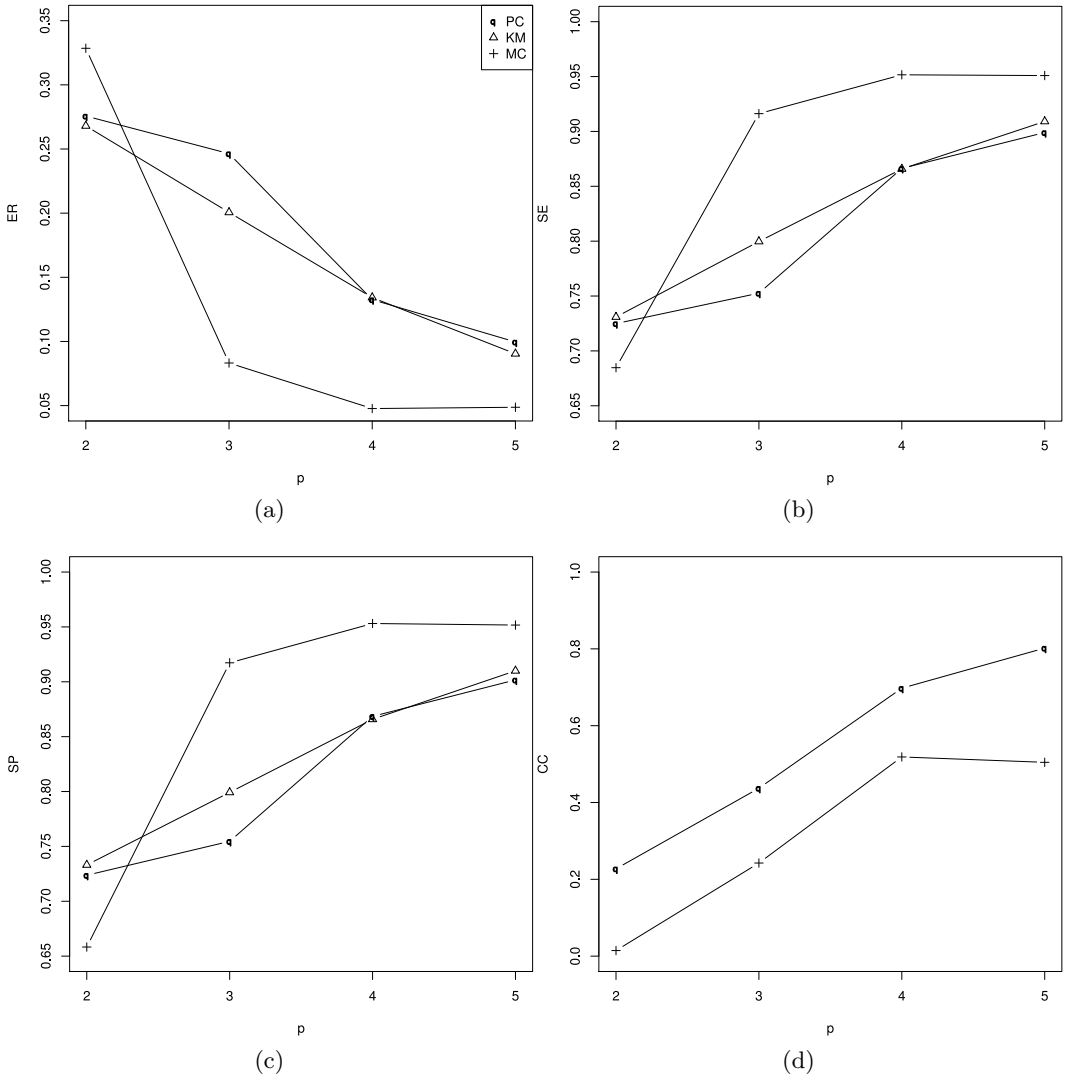

Figure 1: Results of the simulation, Setting 1: 100 samples and growing values of the dimension (p). Legend: circle *pdfClust*, triangles *k-means*, crosses *Mclust*. (a) Error rates (ER); (b) Sensitivity (SE); (c) Specificity (SP); (d) frequencies of correct number of clusters (CC) found.

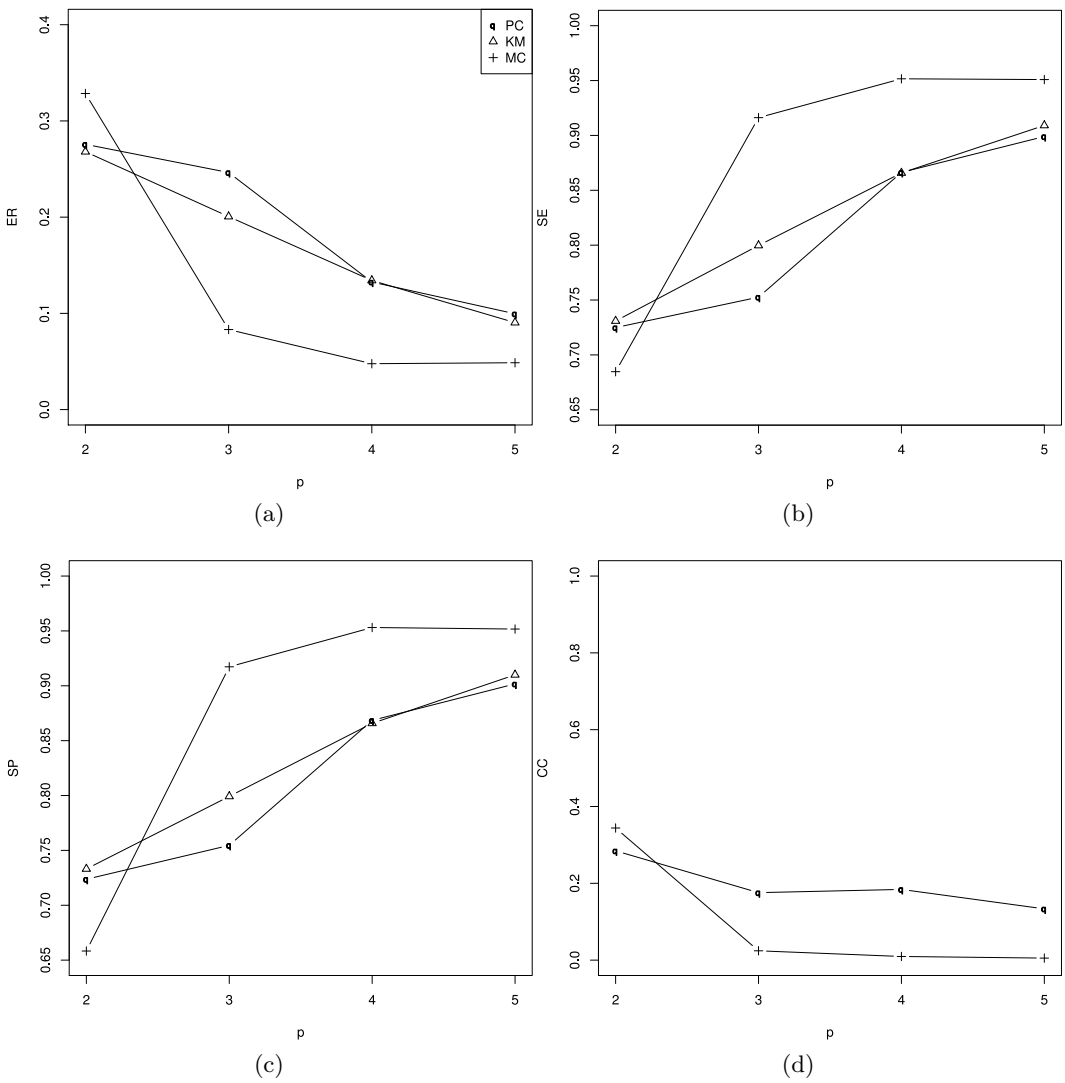

Figure 2: Results of the simulation, Setting 2: 100 samples and growing values of the dimension (p) and groups defined only by the first dimension. Legend: circle *pdfClust*, triangles *k-means*, crosses *Mclust*. (a) Error rates (ER); (b) Sensitivity (SE); (c) Specificity (SP); (d) frequencies of correct number of clusters (CC) found.

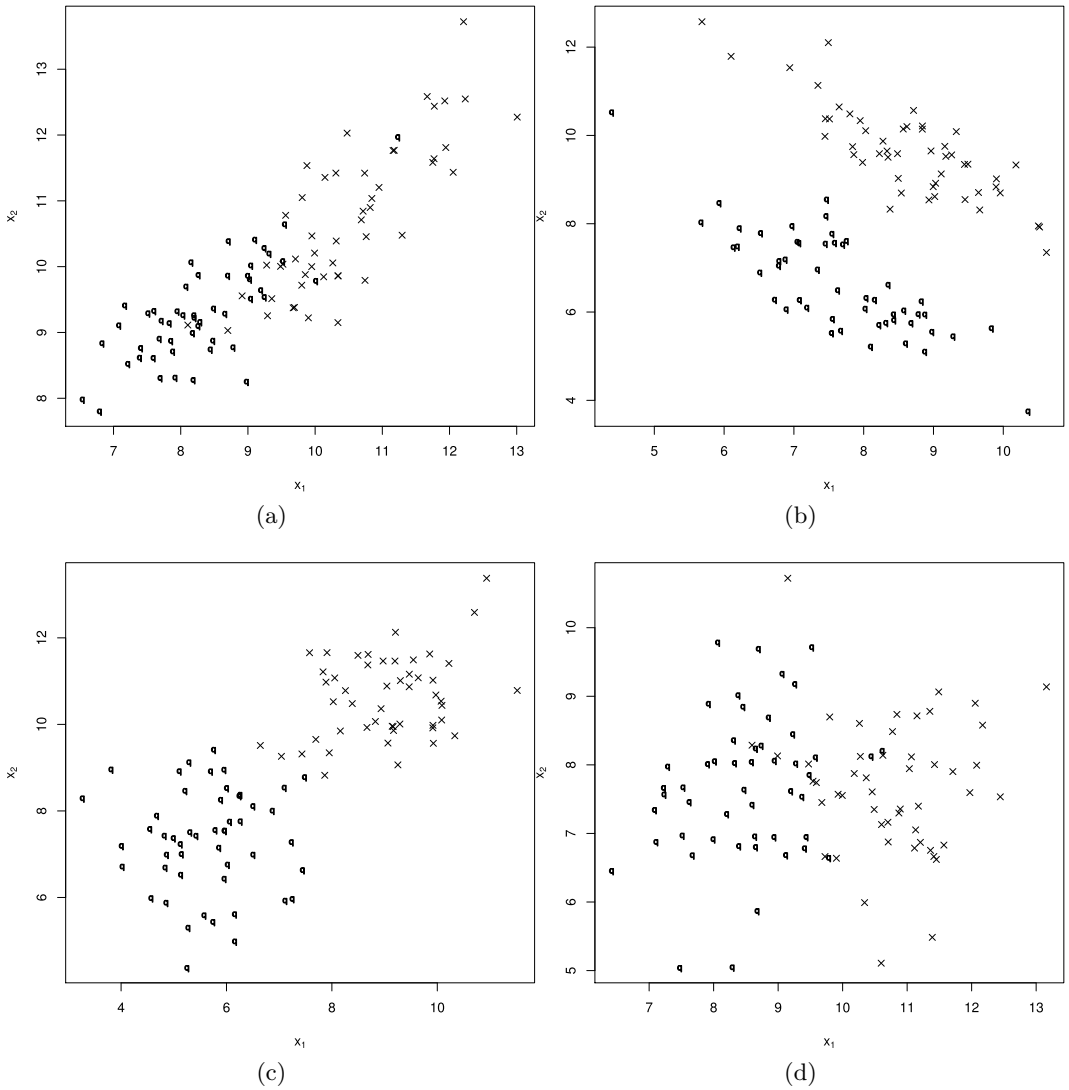

Figure 3: Some simulated data from Setting 3, varying the correlation (ρ) and the number (k) of genes responsible for the clusters. Circles: first cluster, crosses: second cluster. (a) $\rho = 0.8$ and $k = 2$; (b) $\rho = -0.8$ and $k = 2$; (c) $\rho = 0$ and $k = 2$; (d) $\rho = 0$ and $k = 1$.

References

Alon, U., Barkai, N., Notterman, D. A., Gish, K., Ybarra, S., Mack, D. & Levine, A. J. (1999). Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proceedings of the National Academy of Sciences of the United States of America* 96 6745–6750.

Azzalini, A. & Torelli, N. (2007). Clustering via nonparametric density estimation. *Statistics and Computing* 17 71–80.

Banerjee, A., Dhillon, I. S., Ghosh, J. & Sra, S. (2005). Clustering on the unit hypersphere using von mises-fisher distributions. *Journal of Machine Learning Research* 6 1345–1382.

Banfield, J. D. & Raftery, A. E. (1993). Model-based gaussian and non-gaussian clustering. *Biometrics* 49 803–821.

Barber, C. B., Dobkin, D. P. & Huhdanpaa, H. (2006). The quickhull algorithm for convex hulls. *ACM Trans. Math. Software* 22 469–483.

Bolstad, B., Irizarry, R., Astrand, M. & Speed, T. (2003). A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* 19 185–193.

Chiogna, M., Massa, M. S., Risso, D. & Romualdi, C. (2009). A comparison on effects of normalisations in the detection of differentially expressed genes. *BMC Bioinformatics* 10 61.

Chow, M. L., Moler, E. J. & Mian, I. S. (2001). Identifying marker genes in transcription profiling data using a mixture of feature relevance experts. *Physiol. Genomics* 5 99–111.

de Berg, M., Cheong, O., van Kreveld, M. & Overmars, M. (2008). *Computational Geometry: Algorithms and Applications*. Springer.

Dudoit, S., Fridlyand, J. & 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.

Fraley, C. & Raftery, A. E. (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association* 97 611–631.

Fraley, C. & Raftery, A. E. (2006). MCLUST version 3 for R: model mixture modeling and model-based clustering. Tech. rep., no. 504, Department of Statistics, Univerisity of Washington.

Friedman, J. H. & Tukey, J. W. (1974). A projection pursuit algorithm for exploratory data analysis. *IEEE Trans. Comput.* 23 881–890.

Garrett, E. S. & Parmigiani, G. (2003). POE: statistical methods for qualitative analysis of gene expression. In G. Parmigiani, E. S. Garrett, R. A. Irizarry & S. L. Zeger, eds., *The analysis of gene expression data*. Springer, 362–387.

Gentleman, R. C., Carey, V. J., Bates, D. M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J., Hornik, K., Hothorn, T., Huber, W., Iacus, S., Irizarry, R., Leisch, F., Li, C., Maechler, M., Rossini, A. J., Sawitzki, G., Smith, C., Smyth, G., Tierney, L., Yang, J. Y. H. & Zhang, J. (2004). Bioconductor: Open software development for computational biology and bioinformatics. *Genome Biology* 5 R80.

Getz, G., Levine, E. & Domany, E. (2000). Coupled two-way clustering analysis of gene microarray data. *Proceedings of the National Academy of Sciences of the United States of America* 97 12079–12084.

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., Bloomfield, C. D. & Lander, E. S. (1999). Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science* 286 531–537.

Hartigan, J. A. (1975). *Clustering Algorithms*. John Wiley & Sons.

Hastie, T. & Stuetzle, W. (1989). Principal curves. *Journal of the American Statistical Association* 84 502–516.

Hyvarinen, A. & Oja, E. (2000). Independent component analysis: algorithms and applications. *Neural Networks* 13 411–430.

Kendziorski, C., Newton, M. A., Lan, H. & Gould, M. N. (2003). On parametric empirical bayes methods for comparing multiple groups using replicated gene expression profiles. *Statistics in Medicine* 22 3899–3914.

Kerr, G., Ruskin, H., Crane, M. & Doolan, P. (2008). Techniques for clustering gene expression data. *Computers in Biology and Medicine* 38 283 – 293.

Li, J., Ray, S. & Lindsay, B. G. (2007). A nonparametric statistical approach to clustering via mode identification. *Journal of Machine Learning Research* 8 1687–1723.

Li, J. & Zha, H. (2006). Two-way poisson mixture models for simultaneous document classification and word clustering. *Computational Statistics & Data Analysis* 50 163–180.

McLachlan, G. J., Bean, R. W. & Peel, D. (2002). A mixture model-based approach to the clustering of microarray expression data. *Bioinformatics* 18 413–422.

Menardi, G. (2006). *Un metodo di raggruppamento basato sulla stima di densità: alcuni approffondimenti*. Ph.D. thesis, Department of Statistics, Univerisity of Padova.

R Development Core Team (2009). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL `http://www.R-project.org`.

Silverman, B. W. (1981). Using kernel density estimates to investigate multimodality. *Journal of the Royal Statistical Society, Series B: Methodological* 43 97–99.

Acknowledgements

We are grateful to Prof. Adelchi Azzalini for the valuable advice and suggestions. We also would like to thank Dr. Giovanna Menardi, who kindly provided us with an updated version of the *pdfClust* routines and Prof. Monica Chiogna for the useful comments

Working Paper Series
Department of Statistical Sciences, University of Padua

You may order paper copies of the working papers by emailing wp@stat.unipd.it
Most of the working papers can also be found at the following url: http://wp.stat.unipd.it

Department of Statistical Sciences
University of Padua
Italy