Randomized maps for assessing the reliability of patients clusters in DNA microarray data analyses

Alberto Bertoni [a] Giorgio Valentini [a]

[a] *DSI, Dipartimento di Scienze dell' Informazione,*
Università degli Studi di Milano,
Via Comelico 39, Milano, Italy

Full address of the corresponding author:
Giorgio Valentini,
DSI, Dipartimento di Scienze dell' Informazione,
Università degli Studi di Milano,
Via Comelico 39, Milano, Italy
E - mail: valentini@dsi.unimi.it,
Tel: +39 02 50316225,
Fax: +39 02 50316.

Preprint submitted to Artificial Intelligence in Medicine *22 March 2006*

Abstract

Objective:

Clustering algorithms may be applied to the analysis of DNA microarray data to identify novel subgroups that may lead to new taxonomies of diseases defined at bio-molecular level. A major problem related to the identification of biologically meaningful clusters is the assessment of their reliability, since clustering algorithms may find clusters even if no structure is present.

Methodology:

Recently, methods based on random "perturbations" of the data, such as bootstrapping, noise injections techniques and random subspace methods have been applied to the problem of cluster validity estimation. In this framework, we propose stability measures that exploits the high dimensionality of DNA microarray data and the redundancy of information stored in microarray chips. To this end we randomly project the original gene expression data into lower dimensional subspaces, approximately preserving the distance between the examples according to the Johnson-Lindenstrauss (JL) theory. The stability of the clusters discovered in the original high dimensional space is estimated by comparing them with the clusters discovered in randomly projected lower dimensional subspaces. The proposed cluster-stability measures may be applied to validate and to quantitatively assess the reliability of the clusters obtained by a large class of clustering algorithms.

Results and conclusion:

We tested the effectiveness of our approach with high dimensional synthetic data, whose distribution is a priori known, showing that the stability measures based on randomized maps correctly predict the number of clusters and the reliability of each individual cluster. Then we showed how to apply the proposed measures to the analysis of DNA microarray data, whose underlying distribution is unknown. We evaluated the validity of clusters discovered by hierarchical clustering algorithms in diffuse large B-cell lymphoma (DLBCL) and malignant melanoma patients, showing that the proposed reliability measures can support bio-medical researchers in the identification of stable clusters of patients and in the discovery of new subtypes of diseases characterized at bio-molecular level.

Key words: Gene expression data clustering, assessment of cluster stability, cluster reliability, random subspace, random projections, DNA microarrays.

1 Introduction

Profiling of tissue samples using DNA microarray data is widely applied to discover molecular fingerprints that characterize human diseases [1–3]. In particular clustering algorithms have been widely applied to discover and define subtypes of diseases on molecular basis [4–8]. Indeed unsupervised learning

methods, exploiting the overall gene expression profile of a patient, may research and discover subclasses of pathologies that cannot be detected with traditional biochemical, histopathological and clinical criteria [9, 10]. The discovery of diseases subtypes defined by gene expression data may lead to more refined predictions than classical clinical correlates in terms of correct diagnosis, survival, disease-free survival and disease recurrence [11–13]. Moreover the definition of subtypes of diseases on molecular basis may help to develop therapies targeted to the bio-molecular characteristics of patients [14], and to design automatic classification methods for supporting diagnostic procedures [15–18].

Cluster analysis has been used for investigating structure in microarray data, such as the search of new tumor taxonomies [19]. It provides a way for validating groups of patients according to prior biological knowledge or to discover new "natural groups" inside the data. Unfortunately, clustering algorithms always find structure in the data, even when no structure is present instead. Hence we need methods for assessing the validity of the discovered clusters to test the existence of biologically meaningful clusters. The discovered clusters depend on the clustering algorithm, the initial condition, the parameters of the algorithm, the distance or correlation measure applied to the data and other clustering and data-dependent factors [20].

In particular, for a given data set different clustering algorithms may provide very different partitions or clusterings of the data. For instance, in hierarchical clustering [9] it is not obvious by simply looking at the dendrogram which are the significant clusters; the choice of the clusters depends on the particular "cut" of the dendrogram. With dendrograms resulting from the analysis of tumor specimens, bio-medical knowledge is fundamental to select a proper "cut", but also in this case an "objective" and data-driven assessment of the reliability of the clusters may be useful to support bio-medical decisions. Even when clusters and cluster boundaries are univocally defined by the clustering algorithms, such as in K-means [21], or in Self-Organizing-Maps [22], the number of clusters must be chosen a priori and the results depend on the initial conditions. Other methods based on a Bayesian paradigm that combines a priori knowledge with observational data, can automatically select the "optimal" number of clusters, but their accuracy is decremented when small samples are used, due to their asymptotic assumptions [23, 24]. Several other clustering approaches, such as bi-clustering methods [25, 26] have been proposed for the analysis of gene expression data (see e.g. [10] and [27] for an overview), but in all cases the problem of the reliability and validity of the discovered clusters remains open.

Two of the main concerns with gene expression clustering analysis are the estimate of the number of clusters in a dataset, and the stability of the individual clusters [28]. Indeed in many cases we have no sufficient biological

knowledge to "a priori" evaluate both the number of clusters (e.g. the number of biologically distinct tumor classes), as well as the validity of the discovered clusters (e.g. the reliability of new discovered tumor classes). Note that this is an intrinsically "ill-posed" problem, since in unsupervised learning we lack an external objective criterion, that is we have not an equivalent of a priori known class label as in supervised learning, and hence the evaluation of the validity/reliability of the discovered classes becomes elusive and difficult.

Most of the works focused on the estimate of the number of clusters in gene expression data [27, 29–32], while the problem of stability of each individual cluster has been less investigated. Nevertheless, the stability and reliability of the obtained clusters is crucial to assess the confidence and the significance of a bio-medical discovery [33, 34].

Some recent approaches to estimate the reliability of the discovered clusters are based on the concept of the stability with respect to perturbations [33–35]. In the context of gene expression data, that are usually characterized by relatively high level of noise [36], stability can be considered an important property: how much the characteristics and composition of the discovered clusters hold when perturbation such as added noise, resampling or random projections are introduced? Can we design stability measures to assess the reliability of the discovered clusters?

Our approach proposes to estimate the reliability of individual clusters exploiting the redundancy inherent to microarray gene chips. Indeed the number of genes in a chip is usually much larger than the number of samples, and we may reasonably expect that using subsets of genes to perform clustering of tissues, we may obtain meaningful clusters of data. To this end we apply multiple random projections to the DNA microarray samples, reducing the high dimensionality of the original data. The main idea behind our approach consists in evaluating the stability of the clusters discovered in the original high dimensional space comparing them with the clusters discovered in randomly projected lower dimensional subspaces. Our concept of reliability is tied to the concept of stability: we consider reliable a cluster if it is stable, that is if that cluster is maintained in the projected space without too large changes. To properly evaluate the reliability of the clusters, the random projections should not induce too large modifications of the distances between the examples in the projected space. To this end, we use the concept of random projections with bounded metric distortions, according to the Johnson-Lindenstrauss (JL) theory [37].

The proposed method is related to the Smolkin and Gosh [34] approach based on an unsupervised version of the random subspace method [38]. We extend the unsupervised random subspace approach to more general random projections, in the framework of random embeddings between euclidean spaces, and

we propose cluster stability measures based on similarity between randomly projected data.

In the next section we present a brief introduction to randomized embeddings in euclidean spaces, focusing on random projections obeying the *JL* lemma. In Sect. 3 we compare the theoretical and empirical distortion induced by randomized embeddings in gene expression data, in order to get insights into the better strategy to reduce the dimensionality in high dimensional DNA microarray data while approximately preserving the distances between the examples. In Sect. 4 we present our approach to the estimate of cluster stability based on random projections, and in Sect. 5 we compare our method with other related approaches presented in the literature. In Sect. 6 we show how to apply the proposed stability measures to several gene expression data sets in order to evaluate the reliability of the discovered clusters of patients. The discussion (Sect. 7) and the conclusions (Sect. 8) end the paper.

2 Randomized embeddings and dimensionality reduction in euclidean spaces.

Our goal consists in using multiple instances of the data obtained by projections from the original to lower dimensional subspaces to assess the reliability of patients clusters. In order to maintain the characteristic of the examples (patients) in the original space we would like to preserve the similarities (in terms of euclidean distances) as well as possible in the projected low dimensional space. Unfortunately there are no deterministic maps that can in general satisfy this property. However, using a stochastic approach, we may deal with this problem, and we can obtain dimensionality reduction by mapping points from a high to a low-dimensional space, approximately preserving some characteristics, i.e. the distances between points. In this context randomized embeddings with low distortion represent a key concept.

2.1 Randomized embeddings with low distortion.

A *randomized embedding* between L_2 normed metric spaces with distortion $1+\epsilon$, with $\epsilon > 0$ and failure probability P is a distribution probability over mappings $\mu : \mathbb{R}^d \to \mathbb{R}^{d'}$, with $d' > d$, such that for every pair $p, q \in \mathbb{R}^d$, the following property holds with probability $1 - P$:

$$\frac{1}{1+\epsilon} \leq \frac{||\mu(p) - \mu(q)||_2}{||p - q||_2} \leq 1 + \epsilon \tag{1}$$

The ratio of the distances between points in the embedded and original metric space in eq. 1 is defined as *distortion*. Randomized embeddings have been successfully applied both to combinatorial optimization and data compression [39]. The main result on randomized embedding is due to Johnson and Lindenstrauss [37], who proved the existence of a randomized embedding $\mu : \mathbb{R}^d \rightarrow \mathbb{R}^{d'}$ with distortion $1 + \epsilon$ and failure probability $e^{\Omega(-d'\epsilon^2)}$, for every $0 < \epsilon < 1/2$. As a consequence, for a fixed data set $S \subset \mathbb{R}^d$, with $|S| = n$, by union bound, it holds:

$$Prob\left(\forall p, q \in S, \quad \frac{1}{1+\epsilon} \leq \frac{||\mu(p) - \mu(q)||_2}{||p - q||_2} \leq 1 + \epsilon\right) \geq 1 - n^2 e^{\Omega(-d'\epsilon^2)} \quad (2)$$

Hence, by choosing d' such that $n^2 e^{\Omega(-d'\epsilon^2)} < 1/2$, it is proved the following property:
Johnson-Lindenstrauss (JL) lemma: Given a set S with $|S| = n$ there exists a $1 + \epsilon$-distortion embedding into $\mathbb{R}^{d'}$ with $d' = c \log n/\epsilon^2$, where c is a suitable constant.

Note that surprisingly the dimensionality d' of the projected subspace by which we can obtain a limited distortion does not depend on the dimension of the original space d, but only on the cardinality of the available data and on the magnitude ϵ of the desired distortion. In this way we may obtain projections with limited distortion even for very high dimensional data, such as DNA microarray data usually are.

2.2 Randomized maps.

The embedding exhibited in [37] consists in random projections from $\mathbb{R}^d$ into $\mathbb{R}^{d'}$, represented by matrices $d' \times d$ with random orthonormal vectors. Similar results may be obtained by using simpler maps, represented through random $d' \times d$ matrices $P = 1/\sqrt{d'}(r_{ij})$, where r_{ij} are random variables such that:

$$E[r_{ij}] = 0, \qquad Var[r_{ij}] = 1$$

This kind of embeddings may be obtained through randomized linear combinations of the variables in the original space (see below). Strictly speaking, these are not projections, but for sake of simplicity, we call random projections even this kind of embeddings. For sake of simplicity, we call random projections even this kind of embeddings. The randomized maps may be represented through $d' \times d$ matrices R such that the columns of the "compressed" data set represented by a $d' \times n$ matrix $D_R = RD$ have approximately the same distance as the examples in the original space stored in the $d \times n$ matrix D. Examples of random projections are the following:

(1) *Plus-Minus-One (PMO)* random projections: represented by $d' \times d$ matrices $R = 1/\sqrt{d'}(r_{ij})$, where r_{ij} are uniformly chosen in $\{-1, 1\}$, such that $Prob(r_{ij} = 1) = Prob(r_{ij} = -1) = 1/2$. In this case the *JL lemma* holds with $c \simeq 4$.

(2) *Achlioptas* random projections [40]: represented by $d' \times d$ matrices $P = 1/\sqrt{d'}(r_{ij})$, where r_{ij} are chosen in $\{-\sqrt{3}, 0, \sqrt{3}\}$, such that $Prob(r_{ij} = 0) = 2/3$, $Prob(r_{ij} = \sqrt{3}) = Prob(r_{ij} = -\sqrt{3}) = 1/6$. In this case also we have $E[r_{ij}] = 0$ and $Var[r_{ij}] = 1$ and the *JL lemma* holds.

(3) *Normal* random projections [41]: this *JL lemma* compliant randomized map is represented by a $d' \times d$ matrix $R = 1/\sqrt{d'}(r_{ij})$, where r_{ij} are distributed according to a gaussian with 0 mean and unit variance.

(4) *Random Subspace (RS)* [38]: represented by $d' \times d$ matrices $R = \sqrt{d/d'}(r_{ij})$, where r_{ij} are uniformly chosen with entries in $\{0, 1\}$, and with exactly one 1 per row and at most one 1 per column. It is worth noting that, in this case, the "compressed" data set $D_R = RD$ can be quickly computed in time $\mathcal{O}(nd')$, independently from d. Unfortunately, *RS* does not satisfy the *JL lemma*.

Using the above randomized maps (with the exception of *RS* projections), the *JL lemma* guarantees that the "compressed" examples of the data set represented by the matrix $D_R = RD$ have approximately the same distance (up to a distortion $1 + \epsilon$) of the corresponding examples in the original space, represented by the columns of the matrix D, as long as $d' \leq c \log n/\epsilon^2$.

3 Distortion induced by random projections of gene expression data.

Clustering algorithms are in general sensitive to distortions, since they are usually based on distance/dissimilarity measures between objects. Projections from high to lower dimensional spaces may induce relevant distortions, that depend both on the characteristics of the projection and on the distribution of the data. In this section we present experiments with randomized maps to understand in which conditions and to which extent random projections may induce distortions in DNA microarray data. In particular, we compared the theoretical distortions bounds predicted by the *JL lemma* with the empirical bounds estimated on multiple randomly projected gene expression data, using different types of randomized maps.

Given a data set $D \subset \mathbb{R}^d$ and a map $\mu : \mathbb{R}^d \to \mathbb{R}^{d'}$, for $x, y \in D$ the *distortion* $dist_\mu(x, y)$ is defined:

$$dist_\mu(x, y) = \frac{||\mu(x) - \mu(y)||_2}{||x - y||_2} \tag{3}$$

Of course, $dist_\mu(x, y) = 1$ means that no distortion is introduced. In our experiments we evaluated the *maximum, minimum* and *average distortion* of μ on D:

$$\begin{aligned} max.dist_\mu(D) &= \max_{x,y \in D} dist_\mu(x, y) \\ min.dist_\mu(D) &= \min_{x,y \in D} dist_\mu(x, y) \\ ave.dist_\mu(D) &= \frac{1}{|D|(|D|-1)} \sum_{x,y \in D, x \neq y} dist_\mu(x, y) \end{aligned} \tag{4}$$

We estimated, given a DNA microarray data set D and a randomized map μ, the expectation of the random variables $max.dist_\mu(D), min.dist_\mu(D)$ and $ave.dist_\mu(D)$ (eq. 4).

We considered 4 DNA microarray data sets referred to different tumor specimens: *Lung tumor* [42], *Melanoma* [43], *DLBCL* [11], *Primary-metastasis* [2]. For each data set we performed *Achlioptas*, *Normal*, *PMO* and *RS* random projections (Sect. 2.2). We experimentally evaluated the expectation of the random variables $max.dist_\mu(D), min.dist_\mu(D)$ and $ave.dist_\mu(D)$ (eq. 4) averaging their values over 50 repeated random projections and we compared the results with the estimated theoretical distortion predicted by the *JL lemma*. For each data set we performed random projections into subspace whose dimensions correspond to predicted distortions $1 + \epsilon$, with $\epsilon \in [0.1, 0.5]$. We estimated also the empirical distribution of the pairwise distances between example in the original and projected data, to provide a visual clue of the distortions induced by the random projections. All the code to implement the experiments have been written in *R* language and it is available from the authors.

[Fig. 1 about here.]

[Fig. 2 about here.]

In Fig. 1 and 2 are reported the results relative to *Lung tumor* and *Melanoma* data sets. In abscissa are represented the dimensions of the projected subspace and in ordinate the corresponding distortion. Continuous lines represent the bounds of the maximum and minimum distortion according to the *JL* lemma; dashed lines represent the empirical average maximum and minimum distortion computed and averaged over 50 random projections. The pairs of dotted lines just above and below the dashed lines represent the confidence interval at 99 % confidence level. The dash-dotted line represents the average distortion. The circles on the continuous lines represent the distortions theoretically estimated for ϵ ranging from 0.5 to 0.1 at 0.05 step intervals. The dashed and

dot-dashed lines represent the corresponding estimated empirical values (the lines are simply computed by linear interpolation between points). We recall that distortion equal to 1 means no distortion (eq.1).

For the *Achlioptas*, *Normal* and *PMO* random projections, the empirical bounds of the maximum and minimum distortions are largely inside the theoretical bounds predicted by the *JL lemma* (Fig. 1 and 2, (a), (b) and (c)). On the contrary, with *RS* random projections in both cases the maximum and minimum distortions are largely outside the theoretical bounds (Fig. 1 and 2, (d)). Note also that these results are significant at 99 % confidence level, as the dashed lines of the confidence intervals do not intersect the continuous lines of the theoretical bounds. In any case the average empirical distortion is very close to 1 (that is we have no distortion on the average), even if with *RS* projections for large values of *epsilon* (that corresponds to very low dimensional subspaces) the average empirical distortions moves slightly from 1. Similar results are obtained also with the *DLBCL* and *Primary-metastasis* data sets (data not shown).

[Fig. 3 about here.]

[Fig. 4 about here.]

The density distribution of the pairwise distances between samples in the original and randomly projected space can get some additional insights into the characteristics of the distortions induced by the different randomized maps. More precisely, for each data set we computed the euclidean distance of a sample from each other sample, repeating this procedure for all the samples of the data set, for both the original and the projected data. It is worth noting that equal density empirical distributions does not necessarily mean that no distortion is induced in the projected space. Fig.3 and 4 substantially confirm previous results. Indeed at $\epsilon = 0.1$ distortion with *Achlioptas*, *Normal* and *PMO* random projections the empirical distribution of the pairwise distances of the *Lung tumor* and *Melanoma* DNA microarray data are quite similar in both the original and projected data (first three graphs of the first row of Fig. 3 and 4), while for the *RS* projection a certain discrepancy can be observed (fourth graph of the first row of Fig. 3 and 4), especially in the *Melanoma* data set (Fig. 4). For larger distortions (e.g. $\epsilon = 0.5$, second row of Fig. 3 and 4) we can observe, as expected, larger discrepancies between distributions. In this case also the differences between pairwise distances distributions in the original and projected space are significantly more relevant for *RS* projections with respect to the other types of randomized maps. Similar results are obtained also with the *DLBCL* and *Primary-metastasis* data sets (data not shown).

4 Cluster stability measures based on randomized embeddings

The main goal of the proposed stability measures is to assess the stability of the clusters discovered by a suitable clustering algorithm. To this end we exploit the redundancy of features (genes) that characterize DNA microarray data: for each patient we dispose of thousands of gene expression measures, and we know that the expression levels of many genes are correlated. Indeed several works showed that subsets of coordinately expressed genes (*expression signatures*) characterize the functional status of a patient. A gene expression signature is characterized by either the cell type in which its component genes were expressed, or the biological process in which its component genes are known to function [4]. For instance, expression signatures has been discovered and analyzed in gene expression profiles of malignancies [1–4, 6, 11, 44].

Exploiting this redundancy we quantitatively evaluated the stability of the discovered clusters by using multiple random projections from the original high dimensional data to lower dimensional subspaces. Comparing the clusters obtained by using multiple instances of the randomly projected data with the clusters obtained in the original high dimensional gene space, we measure if and at which extent the individual clusters are maintained in the projected subspaces. In other words we can expect that some distance-based properties, such as similarity between examples and membership to the same cluster are maintained in stable and reliable clusters across multiple random projections of the data, as long as these projections do not induce too much distortion in the data. Our experimental results in the previous section just showed that several random projections (e.g. *Achlioptas*, *Normal* and *PMO*) do not introduce relevant distortions in gene expression data, according to the Johnson and Lindenstrauss theory. Moreover the empirical distortions appear to be significantly lower than that predicted by the theoretical bounds (Sect. 3).

For these reasons it seems to be reasonable to evaluate cluster stability through multiple random projections in the context of gene expression data analysis. Note that if we consider projected data as "perturbed" data, we may set random projections in the framework of data perturbations methods (such as bootstrapping or data noise-injections) for assessing cluster stability [35] (see Sect. 5 for more details).

The proposed procedures to measure cluster reliability are divided into several steps. At first, multiple random random projections of the data are generated, choosing a subspace dimension in concordance with the *JL lemma*. Then each instance of projected data is given as input to a clustering algorithm, and the resulting clusters are compared with that obtained in the original high dimensional space. The stability measure of an individual cluster is computed by counting how many pairs of elements of the cluster in the original space are

preserved in the cluster of the projected space; singleton clusters are considered apart. An overall measure of stability of the entire clustering is derived by the stability measures of the individual cluster by simply averaging between them, and a measure of "cluster membership" for each example, based on pairwise membership to the same cluster in the projected space is also proposed.

In the following subsections we describe more in detail the proposed stability measures.

4.1 Similarity matrix

In this section we introduce the pairwise *similarity matrices* between examples [28, 45], since they are used to compute the stability measures proposed in this paper. The similarity matrix is a sort of distributed memory of the clusters, by which memberships of pairs of examples to the same cluster are stored.

Consider a data set $X = \{x_1, x_2, \ldots, x_n\}$, where $x_i \in \mathbb{R}^d,\ (1 \leq i \leq n)$; a subset $A \subseteq \{1, 2, \ldots, n\}$ univocally individuates a subset of examples $\{x_j | j \in A\} \subseteq X$. The data set X may be represented as a $d \times n$ matrix D, where columns correspond to the examples, and rows correspond to the "components" of the examples $x \in X$. A *k-clustering* C of X is a list $C =< A_1, A_2, \ldots, A_k >$, with $A_i \subseteq \{1, 2, \ldots, n\}$ and such that $\bigcup A_i = \{1, \ldots, n\}$. A *clustering algorithm* $\mathcal{C}(X, k)$ is a procedure that, having as input a data set X and an integer k, outputs a k-clustering on the basis of the distances $||x_i - x_j||$, $(1 \leq i, j \leq n)$.

We can associate a $n \times n$ similarity matrix M to a k-clustering; the elements M_{ij} of M are defined as:

$$M_{ij} = \sum_{s=1}^{k} \chi_{A_s}[i] \cdot \chi_{A_s}[j] \tag{5}$$

where $i, j \in \{1, 2, \ldots, n\}$, and $\chi_{A_s} \in \{0, 1\}^n$ is the characteristic vector of $A_s \subseteq \{1, 2, \ldots, n\}$, i.e. $\chi_{A_s}[i] = 1$ if $i \in A_s$, otherwise $\chi_{A_s}[i] = 0$. If the k-clustering identifies a partition, then $M_{ij} \in \{0, 1\}$: in other words, M_{ij} denotes if elements i and j belong to the same cluster. Consider also a random projection $\mu : \mathbb{R}^d \to \mathbb{R}^{d'}$ that verifies the *JL* lemma (i.e. *PMO*, see Sect. 2.2).

Then we can compute a *cumulative similarity matrix* M^C, using the following algorithm:

(1) Generate t independent projections $\mu_r : \mathbb{R}^d \to \mathbb{R}^{d'},\ 1 \leq r \leq t$, such that $d' = 4\,\frac{\log|X| + \log t}{\epsilon^2}$

(2) Apply $\mathcal{C}$ to the new projected data $\mu_r(X)$, obtaining a set of clusterings, for $1 \leq r \leq t$:

$$\mathcal{C}(\mu_r(X)) =< B_1^r, \ldots, B_k^r >,\ B_s^r \subset \mu_r(X), 1 \leq s \leq k \tag{6}$$

where B_s^r is the s^{th} cluster of the r^{th} clustering.

(3) Set the elements M_{ij}^C of the cumulative similarity matrix:

$$M_{ij}^C = \frac{1}{t} \sum_{s=1}^{k} \sum_{r=1}^{t} \chi_{B_s^r}[i] \cdot \chi_{B_s^r}[j] \tag{7}$$

where $\chi_{B_s^r}$ is the characteristic vector for the cluster B_s^r.

Since the elements M_{ij}^C measure the occurrences of the examples $\mu_r(x), \mu_r(y) \in \mu_r(X)$ in the same clusters B_s^r for $1 \leq r \leq t$, then M^C represents how much pairs of projected examples belong to the same cluster across all the t repeated projections. If the clustering defines a partition, it is easy to see that $0 \leq M_{ij} \leq 1$, for each $x_i, x_j \in X$.

With respect to the algorithm above we may observe:
Remark 1. Since the failure probability is $e^{\Omega(-d'\epsilon^2)}$, similarly to eq.2 in Sect. 2, by union bound we have:

$$P\left(\forall x, y \in X, \quad \frac{1}{1+\epsilon} \leq \frac{||\mu_r(y) - \mu_r(x)||_2}{||x-y||_2} \leq 1+\epsilon\right) \geq 1 - t|X|^2 e^{\Omega(-d'\epsilon^2)}$$

Therefore for $d' \simeq \mathcal{O}\left(\frac{\log|X| + \log t}{\epsilon^2}\right)$, we obtain with high probability that all the projections preserve the distances between the elements in X up to a distortion $1+\epsilon$.
Remark 2. Singleton clusters, that is clusters composed by a single element are considered apart: if B_s^r is composed by a single element i the element M_{ii} of the similarity matrix is incremented by one. As a consequence the diagonal elements of the cumulative similarity matrix M_{ii}^C represent the frequency by which the singleton cluster $\{i\}$ occurs across the t clusterings.
Remark 3. A fuzzy similarity matrix may be obtained simply substituting in eq. 7 the characteristic function with a membership function and the algebraic product with a suitable *t-norm*. In this way fuzzy or possibilistic clustering approaches may also be applied [46, 47]. In this case if we would maintain the property by which $0 \leq M_{ij} \leq 1$, for each $x_i, x_j \in X$, we need to normalize eq. 7 by $1/k$.

4.2 Stability indices

Using the similarity matrix computed as described in Sect. 4.1, we may easily compute a set of stability indices that may be useful to evaluate the validity

and the reliability of the obtained clusters, as well as the confidence of the assignments of each example to each cluster.

Indeed, using the previously computed M^C similarity matrix, we may introduce the following *stability index* s for a cluster A:

$$s(A) = \frac{1}{|A|(|A|-1)} \sum_{(i,j)\in A\times A, i\neq j} M^C_{ij} \tag{8}$$

The index $s(A)$ estimates the stability of a cluster A by measuring how much the projections of the pairs $(i,j) \in A \times A$ occur together in the same cluster in the projected subspaces.

Let be $< B^r_1, \ldots, B^r_k >$, $B^r_p \subset \mu_r(X), 1 \leq p \leq k, 1 \leq r \leq t$ the clusters obtained in the embedded spaces $\mu_r(X)$, using t random projections. To get some insights into the way the proposed stability index s works, consider the limit cases $s(A) = 1$ and $s(A) = 0$:

(1) Suppose $s(A) = 1$.
This is true if and only if for all r there is a p such that $A \subseteq B^r_p$: the cluster A is present inside some cluster across all clusterings performed in the projected spaces. This fact highlights the stability of A.
(2) Suppose $s(A) = 0$.
This is true if and only if for all r and all p it holds $A \cap B^r_p = \emptyset$: the cluster A is disjoint from all clusters across all clusterings performed in the projected spaces. In this case the stability index points to the instability of cluster A.

In all the remaining cases, the stability index has values between 0 and 1: low values indicate no reliable clusters, high values denote stable clusters. The proposed stability index for individual clusters shows the desirable property by which more similar is a cluster to the clusters obtained in the multiple randomly projected subspaces, larger will be its value and viceversa.

An overall measure of the stability of the clustering in the original space may be obtained averaging between the stability indices:

$$S(k) = \frac{1}{k} \sum_{r=1}^{k} s(A_r) \tag{9}$$

By this measure we may select the "optimal" number of clusters k, using the stability indices s of the individual clusters. In this case also we have that $0 \leq S(k) \leq 1$, where k is the number of clusters.

The *Assignment-Confidence (AC)* index estimates the confidence of the as-

signment of an example i to a cluster A:

$$AC(i, A) = \frac{1}{|A| - 1} \sum_{j \in A, j \neq i} M_{ij}^{C} \tag{10}$$

Using a set of realizations of a given randomized projection, the AC-index represents the frequency by which the example i appears with the other elements of the cluster A across multiple clusterings on the randomly projected subspaces.

5 Related work

The idea of using stability to assess the reliability of clusters and to define meaningful partitions is not new. For instance the stability of hierarchical clustering [48] as well as of more general clustering methods [49] have been addressed, but with respect to the overall partition, while less work has been dedicated to the evaluation of the stability of the individual clusters. Several methods proposed to evaluate the "natural" number of clusters, ranging from strategies that attempt to maximize measures of cluster compactness [50] to jackknife and resampling-based approaches [27, 30, 31].

Focusing on the problem primary addressed by our work, that is the estimate of the reliability of individual gene expression data clusters, Smolkin and Gosh [34] introduced unsupervised random subspace methods to evaluate the stability of clusters. They generated multiple instances of the original data by projecting the original high dimensional gene expression space into lower dimensional subspaces, choosing for each instance a randomly selected set of features (genes). We extended their approach to more general randomized maps, where *RS* is only a particular case (Sect. 2). We showed also that in some cases projecting gene expression data using random subspace, we may introduce relevant distortions into the data, while using *PMO*, *Achlioptas* and *Normal* projections no relevant distortions are induced (Sect. 3). Of course, if the random subspace projection induces large distortions, the stability measures based on it become unreliable; for this reason we strongly suggest using randomized projections that approximately preserve distances between examples. Moreover in [34] no principled approaches to define the subspace dimension of the projected data is proposed. We introduced the concept of controlled distortion in the framework of *JL* theory in order to properly choose the dimension of the projected data. The stability indices proposed by Smolkin and Gosh used the clusters found in the projected space as a whole to evaluate the reliability of the clusters in the original space. In particular if $< A_1, \ldots, A_K >$ are the K clusters found in the original space and $< A_1^m, \ldots, A_K^m >$ the K clusters found in the projected space, they proposed a sensitivity index p_i for

the cluster A_i:

$$p_i = \frac{1}{s} \sum_{m=1}^{s} \sum_{j=1}^{K} I(A_i \subset A_j^m) \tag{11}$$

where s is the number of repeated random subspace projections and $I(x) = 1$ if x is true, otherwise $I(x) = 0$. Even this measure if useful to evaluate the reliability of the clusters, note that a cluster A_i is considered "stable" only if it is completely included inside a cluster A_j^m obtained in the projected space. Our stability measure consider also as "partially" reliable a cluster A_i if $A_i \cap A_j^m \neq \emptyset$, and A_i is more reliable as long as $A_i \cap A_j^m$ is larger. In other words we consider all the conditions from $A_i \cap A_j^m = \emptyset$ (no inclusion), through $A_i \cap A_j^m \neq \emptyset$ (intersection) to $A_i \subset A_j^m$ (inclusion): in these conditions our stability index increases monotonically from 0 to 1, whereas Smolkin and Gosh considered only the extreme case when $A_i \subset A_j^m$, that corresponds to $s(A_i) = 1$. In this way we may capture more subtle relationships between data, obtaining more refined stability indices.

Bittner et al. [43] estimated cluster stability in melanoma patients introducing random perturbations to the data sets by adding gaussian noise to the data. A similar approach has been proposed by McShane et al. [33]. They added random noise independently across the genes to generate multiple randomly perturbed data, estimating the error variance from the experimental data. These approaches, even if related to ours, require to add independent noise to the data and the relatively ad hoc choice of using the overall experimental variance for data perturbation. On the contrary our method requires only to randomly project genes, without assumptions about independently distributed noise across the data. Moreover, as observed in [34], the assumption of independently distributed noise across genes, even if practical, is not fully biologically plausible. Anyway, in [33] an index similar to our stability index is proposed, even if no similarity matrix is explicitly introduced. The main difference between our and McShane et al. approach consists in the way multiple clusterings are generated. Note that our stability index does not take into account differences between A_i and A_j^m when $A_i \subset A_j^m$: in these cases the A_i cluster is "fully supported" by our stability index, even if there are examples in A_j^m that are not included in A_i. This is in most cases reasonable, as we would confirm the stability of the clusters in the original space (that is cluster of patients sharing similar expression profile), even if these may be part of larger clusters. To explicitly consider these cases we may apply the D index proposed in [33]:

$$D(A_i) = Omissions(A_i) + Additions(A_i)$$

By this we can explicitly consider examples included in A_j^m but not included in A_i (additions). In particular if B_b is the cluster with the best matching with

the cluster A_i in the original space, that is:

$$B_b(A_i) = \arg\max_j \sum_{k=1}^{n_i} \sum_{h=1}^{n_j} I(a_i^k = b_j^h) \tag{12}$$

where a_i^k is the k^{th} element of the cluster A_i in the original space, with $|A_i| = n_i$, and b_j^h is the h^{th} element of the cluster B_j in the transformed space, with $|B_j| = n_j$ we may explicitly consider examples that are included in A_i but not in B_b (omissions), and, viceversa, examples included in B_b but not in A_i (additions).

Other approaches to estimate the cluster stability are based on resampling techniques, e.g. by bootstrapping or subsampling examples from the data. Bhattacharjee et al. proposed bootstrap-based methods to estimate the stability of subclasses of adenocarcinomas discovered in cancer lung patients. [42]. Kerr and Churchill applied analysis of variance (ANOVA) models and bootstrap techniques to evaluate the reliability of discovered clusters, taking into account different sources of variation of the data [32]; this approach is well-suited for time series experiments and it is applied with the Chu et al. clustering algorithm [51], but it is not explained how it can be applied to other clustering algorithms. Using a resampling-based scheme Monti et al. [35] proposed stability measures similar to that proposed in this work. They compute a connectivity matrix (corresponding to our similarity matrix) using multiple bootstrap replicates of the original data, while we used multiple random projections obeying the *JL* lemma. They need to consider only the examples extracted by bootstrapping techniques (as they are a subset of the overall data), while our proposed index does not suffer this limitation, since we projected gene expression values using the overall data set. This can be a critical problem with gene expression data as they are usually characterized by small cardinality and very high dimensionality.

6 Experimental analysis of cluster reliability

In this section we show how to apply the proposed stability measures to the analysis of the reliability of patients clusters using high dimensional DNA microarray data.

Anyway, an experimental evaluation of the accuracy and reliability of the proposed stability measures would require to a priori know the "correct" subdivision of patients in clusters. Unfortunately, in many cases the cluster partition of patients is not known in advance: on the contrary, the clustering algorithms are just applied to discover the clusters structure of the analyzed data. For these reasons, to show the effectiveness of the proposed stability measures, we

used high-dimensional synthetic data, whose distribution, composition and clusters subdivision is a priori completely known. Then we applied the stability measures to the analysis of reliability of clusters patients, considering two cases of gene expression analysis of malignancies from the literature: diffuse large B-cell (*DLBCL*) lymphoma [11], and *Melanoma* [43] patients.

Our stability measures can be applied to the analysis of general distance-based clustering algorithms, such as k-means [52] or Self-Organizing-Maps [22], but we used hierarchical agglomerative clustering algorithms [53, 54].

In our experiments we used hierarchical agglomerative clustering algorithms, using as dissimilarity function the euclidean distance. This choice is due to the fact that hierarchical agglomerative clustering has been widely applied in functional genomics for exploratory DNA microarray data analysis [4, 9]. Indeed, even if these algorithms are not in general "better" than others, they provide multi-resolution views of the data, and an appealing visual clue of the obtained clusters. Moreover they do not automatically determine the number of clusters, but provide a hierarchical structure of nested clusters: in this way the bio-medical scientists may choose a proper "cut" in the dendrogram to obtain biologically meaningful clusters. From another standpoint it can be very difficult and somehow arbitrary to choose a proper cut or a cluster as "meaningful" with respect to others, especially when no sufficient bio-medical knowledge on the data is available. This is usually true with DNA microarray data, and for these reasons the stability measures proposed in this paper may help researchers to better evaluate the reliability of the discovered clusters.

In all the experiments described in this section we computed the stability indices s (eq. 8) for all the clusters found by the clustering algorithm, considering several $1+\epsilon$ distortions induced by different types of random projections. Here we reported the results obtained with *PMO* projections, while the results obtained with *Achlioptas* and *Normal* random projections are reported only for the *melanoma* patients, since they are quite similar to that obtained with *PMO*. We computed also the average stability index $S(k)$ (eq. 9) for different number k of clusters, to evaluate also the reliability of the overall clustering.

We developed the *clusterv* package [55] written in R language to implement the generator of high-dimensional synthetic data, the random projections described in Sect.3 and the stability measures described in Sect. 4. This package is downloadable for research and teaching purposes from its web-site: `http://homes.dsi.unimi.it/~valenti/SW/clusterv` (Accessed: 19 March 2006).

6.1 Experimental evaluation of the stability indices with synthetic data

In this section we used high-dimensional synthetic data, whose samples are distributed according to known distributions, to evaluate the effectiveness and the reliability of the proposed stability measures. We analyzed the results of the Ward's hierarchical agglomerative clustering algorithm [54], using as dissimilarity function the euclidean distance.

6.1.1 Data sets

We experimented with 2 different sample generators, whose samples are distributed according to different mixtures of high dimensional gaussian distributions.

Sample1 is a generator for 5000-dimensional data sets composed by 3 clusters. The elements of each cluster are distributed according to a spherical gaussian with standard deviation equal to 3. The first cluster is centered in the null vector $\mathbf{0}$. The other two clusters are centered in $0.5\mathbf{e}$ and $-0.5\mathbf{e}$, where $\mathbf{e}$ is a vector with all components equal to 1.

Sample2 is a a generator for 6000-dimensional data sets composed by 5 clusters of data normally distributed. The diagonal of the covariance matrix for all the classes has its element equal to 1 (first 1000 elements) and equal to 2 (last 5000 elements). The first 1000 variables of the five clusters are respectively centered in $\mathbf{0}$, $\mathbf{e}$, $-\mathbf{e}$, $5\mathbf{e}$, $-5\mathbf{e}$. The remaining 5000 variables are centered in 0 for all clusters.

For each generator, we considered 30 different random samples each respectively composed by 30 and 50 examples (that is, 10 examples per class).

6.1.2 Results

Tab.1 summarizes the results with *sample1*. Note that the numbers at the leaves of the dendrogram of Fig. 5 are the labels that identify the different examples and they correspond to the members of clusters of the second column of Tab 1.

[Table 1 about here.]

[Fig. 5 about here.]

The maximum of the average stability index $S(k)$ is reached when the dendrogram (Fig. 5) is cut at 3 clusters level, and the corresponding stability indices

s are equal to 1 for each of the 3 clusters. We achieve the same values for the stability indices independently of the selected ϵ value that in turn defines the subspace dimension. Two clusters are judged quite reliable too, especially the first cluster that corresponds to one of the "true" cluster, while the second cluster that derives from the merging of the other two "true" cluster, shows a lower stability index (Tab.1). Both the average and the individual stability indices are lower when different number of clusters are selected, showing that the proposed stability measures correctly detect 3 clusters, identifying them as highly reliable, according to the fact that *sample1* generates three well-defined and separated clusters. For instance with 5 clusters the overall stability index S is lower, as well as the stability indices s of the individual clusters: only for the third cluster s is sightly larger, since this cluster corresponds to one (the second) of the three original "true" clusters (Tab 1). With 10 clusters we generate an unnatural fragmentation of the clusters, and the corresponding stability indices are significantly much lower (Fig. 5 and Tab.1).

[Table 2 about here.]

[Fig. 6 about here.]

Sample2 is composed by 5 clusters, two well-separated, while for the other three the corresponding underlying gaussian distributions largely overlap. The stability indices correctly predict largely separated as well as less reliable clusters. Indeed the stability indices are high for the 2 well separated clusters, while for the other clusters that come from partially overlapped gaussian distributions, the stability indices are significantly lower (Tab. 2, see the case with $N = 5$ clusters). Note that the maximum of the overall stability index is for $N = 3$ clusters, as well as the reliability of the corresponding individual clusters. This is nor surprising, as the three "overlapped clusters" are identified by the hierarchical clustering algorithm as a single cluster (Cl.3, Cl.4 and Cl.5 in Fig. 6), since the underlying gaussian distributions from which the data are drawn largely overlap. Anyway, note the first two clusters for 3, 4 and 5 clusters are predicted as very reliable ($s = 1$), according to the fact that their underlying distributions are largely separated and without relevant overlaps (Tab. 2 and Fig. 6).

[Fig. 7 about here.]

In order to evaluate the effectiveness and the reliability of the *Assignment-Confidence (AC)* index (eq. 10), we applied them to the analysis of the synthetic data sets (Sect. 6.1.1), since, in this case we know in advance the distribution and composition of the clusters. In this way we may establish if the clustering algorithm correctly classifies the examples, and, separately analyzing the distributions of the AC values for right and wrong predictions, we can evaluate their reliability.

Fig. 7 shows the distributions of the AC index values over the 30 realizations of each synthetic data set, with separate boxplots for right and wrong predictions. The results show that AC indices are significantly higher for right predictions, supporting their reliability.

6.2 Experimental analysis of the reliability of patients clusters

In this section we apply the stability measures based on random projections to the analysis of the reliability of patients clusters, using DNA microarray data sets previously published by several authors. In particular we analyze the reliability of gene expression clustering of malignancies, considering diffuse large B-cell (*DLBCL*) lymphoma [11] and *Melanoma* [43] specimens.

For the pre-processing of DNA microarray data we wrote R scripts using the Biobase and genefilter packages of the Bioconductor library [56], and we implemented the same pre-processing and normalization steps described by the authors that published the original DNA microarray data [11, 43].

6.2.1 DLBCL patients

We applied the proposed stability indices to a set of gene expression tumor specimens from 58 diffuse large B-cell lymphoma (DLBCL) and 19 follicular lymphoma (FL) patients [11]. For each patient, expression levels of 7129 genes or EST sequences are provided from Affymetrix HU6800 oligonucleotide arrays. Raw data have been pre-processed and re-scaled according to the procedures described in [11]. In particular we processed the raw expression values in Affymetrix's scaled average difference units generated by Affymetrix's GeneChip software using a "windowsizing" procedure for ceiling all the values above the 16000 units and to set a lower threshold at 20 units to minimize fluorescence saturation of the scanner and noise effects. Then the expression values underwent a double filter to exclude all genes with a fold-change less than 3-fold variation or with absolute variation less than 100 units. The 6286 genes that passed the double filter have been normalized with respect to their mean and standard deviation.

In this analysis we applied *PMO* projections to estimate the stability indices of the clusters computed using the hierarchical clustering Ward's method (Tab. 3 and Fig. 8). The tables in Tab. 3 as well as the following tables are structured in two parts: above, the overall stability indices S (eq. 9) are reported for different values ϵ of the distortion (Sect. 2.1), considering different numbers $N.$ of clusters; below, the individual stability indices s (eq. 8) for different values of ϵ and different numbers $N.$ of clusters are reported. Note that in the first column of Tab. 3 the clusters are labeled with numbers, and these

number assignments correspond to left-to-right clusters in the dendrogram of Fig. 8.

In Fig. 8 is represented the dendrogram of the hierarchical clustering, with the dotted lines that highlight the clusters obtained at different cut levels. Note that leaves labeled with D refer to *DLBCL* patients, while F-labeled leaves refer to *FL* patients. Looking at the results obtained with *PMO* projections (Tab. 3), the average S index is slightly larger when the hierarchical clustering dendrogram is cut at 2 clusters level (Fig. 8), but comparable values (even if lower) are also registered with $3, 4$ and 5 clusters. In this case indeed the clusters are not clearly delineated. For instance, considering a cut at 4 clusters level, the first cluster (with a high s stability index equal to 0.9882) is composed by homogeneous FL patients (Fig. 8), the second (less reliable $s = 0.7637$) is composed by both DLBCL and FL patients, while the third (more reliable $s = 0.9800$) is composed only by DLBCL patients, as well as the slightly less reliable ($s = 0.9016$) fourth cluster. If we split the data in 10 or more clusters we note a significant decrement of both the s indices and the average S index: this fact suggests that no significant structure can be observed in small-sized clusters (data not shown).

These results are congruent with the bio-medical characteristics of the data. Indeed even if nodal tumor specimens are subdivided into 2 groups (DLBCL and FL), Alizadeh et al. [4] discovered subclasses among DLBCL patients, confirmed also by the supervised analysis of the data [12]; our results show that the DLBCL subclusters 3 and 4 with 4 clusters and 3, 4 and 5 with 5 clusters are judged reliable by the proposed stability indices (Tab. 3). Moreover Shipp et al. [11] highlighted that FL patients frequently evolve over time and acquire the morphologic and clinical features of DLBCLs: this is confirmed by the high reliability of cluster 1 both with $N = 2$ and $N = 3$ clusters that groups together some DLBCL and FL patients.

[Table 3 about here.]

[Fig. 8 about here.]

6.2.2 Melanoma patients

In this subsection we study the reliability of the clusters obtained in melanoma patients, using a cDNA microarray data set of 38 examples, including 31 melanomas and 7 controls [43]. The 8150 cDNAs represent 6971 unique genes in the melanoma array used in the experiments. For this dataset we directly downloaded the ratio expression levels of the just filtered 3613 genes from the web site associated with the Bittner et al. paper. According to [43], to avoid distortions of the data resulting from ratios where the signal in one channel (Cy5 or Cy3) is large, and the signal in the other channel is undetectable,

we truncated ratios higher than 50 or lower than 0.02. We restricted our experiments to only the 31 melanoma examples to better verify the reliability of the "tightly clustered" set of 19 specimens found in [43]. We present here the stability indices computed using not only *PMO* projections, but also *Achlioptas* and *Normal* projections, to compare the results obtained with the different randomized maps described in Sect. 2.2.

The overall stability index estimates as $N = 4$ the optimal number of clusters, if we disregard the case with $N = 2$ and $N = 3$ clusters, characterized by the presence of singleton clusters (Tab. 4 and Fig. 9). Indeed also the stability indices for all the individual clusters strongly support their reliability (Tab. 4 and 5). With $N = 4$ clusters the first two clusters are singletons, while the third is a big cluster with 23 examples, including the 19-members melanoma subclass found out in [43]; the fourth very stable cluster groups together the remaining 6 examples. To find the same 19-members Bittner's cluster we need to choose $N = 9$ clusters: the fifth cluster exactly corresponds to it. However the stability index of this cluster is quite low ($s \simeq 0.4$ with *PMO*, *Achlioptas* and *Normal* projections), as well as the overall stability index for $N = 9$. Bittner et al. provided an overall stability measure for the clustering ($WADP_k$) that is based on perturbation of the original data by adding random noise (see Sect. 5): their results support clusterings with $N \leq 9$ clusters, but they do not provide an individual cluster stability measure. Note that the application of the Ben-Hur et al. method [30], based on bootstrapping techniques to estimate the "natural" number of clusters, found $N = 4$ clusters as the most reliable number of estimated clusters in the data.

Summarizing, our results support the existence of a highly reliable cluster of melanoma patients, composed by the 19 examples found by Bittner et al. plus other 3-4 examples [1] (Tab. 4, 5 and Fig. 9). Comparing the results obtained with different random projections, we can observe that *PMO*, *Achlioptas* and *Normal* projections provide very similar results. In all cases 2, 3 and 4 clusters are identified as very stable, while a number of subclasses larger than 5 are considered unreliable, showing that subclasses of melanoma patients cannot be found inside the big cluster composed by 23 patients.

[Table 4 about here.]

[Table 5 about here.]

[Fig. 9 about here.]

[1] in particular the Bittner's case no. UACC-1529, TC-FO27, HA-A and UACC-1097.

6.2.3 Experimental comparison with other stability-based methods

In this subsection we experimentally compare our approach with other related stability based methods. In particular, we compare stability measures obtained by *PMO* projections with stability-based methods based respectively on noise-injections into the original data [33] and on random subspace [34] (see Sect. 5 for more details about these methods). For the comparison, we chose the *melanoma* data set [43] because it was studied by many researchers [43, 57, 58], and on this basis we can formulate well-founded hypotheses about its characteristics. In particular we assume that no more than 4 clusters can be found in the data; a big stable cluster collecting about 20 patients is present in the data and no more than 4 subclasses of melanoma patients may be considered reliable. In all cases the hierarchical clustering algorithm with average linkage has been applied and the results are summarized in Tab. 6.

[Table 6 about here.]

We can observe that the noise-injection based method fails to detect the correct number of clusters. Indeed the overall R-index identifies as highly reliable 7 clusters. The Smolkin and Gosh method based on random subspace does not provide a technique to estimate the number of clusters. As suggested by the authors, the model explorer algorithm [30] has been applied to estimate the correct number of cluster. The model explorer algorithm is specifically designed to estimate only the number of cluster (no estimation of the reliability of each individual cluster is provided) and it exploits the overall distribution of the similarity measures to asses the stability of the clustering. Both model explorer and our method based on *PMO* random projections (see Sect. 2.2) predict the number of 4 clusters as highly reliable. Moreover, reflecting the fact that clusters of melanoma patients show a hierarchical structure, both model explorer and our overall stability index (eq. 9) identify as highly reliable 2 and 3 clusters too.

Considering the reliability of each individual cluster, the noise-injection method recognizes as highly reliable 6 of the 7 clusters (Tab. 6), whereas we know that no more than 4 clusters of melanoma patients can be considered reliable. On the other hand the random subspace method considers reliable only 2 of the four clusters, while our method correctly identified as highly reliable all the 4 clusters of melanoma patients. Moreover, the Smolkin and Gosh method does not provide any quantitative suggestion about the dimension of the projected subspace. As a consequence, if we use e.g. 85% of genes the third cluster (that is the big cluster with about 20 patients) is judged reliable (stability score ~ 0.8), otherwise if we use 25% of genes the same cluster is considered unreliable (stability score ~ 0.5). With our approach using $\epsilon \leq 0.2$ the stability index for the third cluster is above 0.9 (that is highly reliable). These results confirm that random subspace projections may produce unreliable results, as

they may introduce too large distortions in the data, as we showed in Sect. 3.

Anyway even if the results show that our proposed method works better than others on the considered data set, we warn the readers that this direct comparison needs to be considered with caution, because we necessarily made assumptions about the "real" number of clusters and the "real" reliability of the discovered clusters, mainly on the basis of the a-priori biological knowledge about the problem. This is necessary, because the comparison of the performance of different methods for assessing the reliability of the discovered clusters requires that we know in advance which is the "real" number of clusters and what is the "real" reliability of the clusters obtained with a given clustering algorithm. Unfortunately in real DNA microarray data all these items are unknown and, as a consequence we need to make assumptions about the structure of the data. In other words, from a computational standpoint, this a particularly ill-posed problem, and in real DNA microarray data analyses we need to strictly integrate bio-medical knowledge and artificial intelligence methods to get insights into the real structure of the data. In particular, methods to assess the reliability of the discovered clusters may be used to suggest hypotheses about the real structure of the data, but these hypotheses need to be biologically validated. On the other hand these methods may be used to assess the reliability of bio-medical hypotheses about e.g. the structure of diagnostic or prognostic classes of malignancies or other diseases on the basis of the bio-molecular characteristics of the patients.

7 Discussion

The experiments with high dimensional synthetic data tested the theoretical applicability of the the proposed measures for assessing the reliability of discovered clusters, when the underlying distribution of the data is a priori known (Sect. 6.1). In particular, with *sample1* the overall stability index S achieves its maximum when the "true" number of clusters is considered, and the three "true" corresponding clusters obtain the maximum of the individual stability index s; with *sample2* the largest stability is achieved for the two well-defined and separated clusters, while for the other less reliable clusters the stability index is lower. Hence, the results with high dimensional synthetic data show the feasibility of the proposed stability measures to estimate the reliability of the discovered clusters.

The results with real gene expression data (Sect.6.2), showed also how to apply the proposed reliability measures to the exploratory analysis and discovery of gene expression patterns in DNA microarray data. In particular, the analysis of the stability of individual clusters through their reproducibility across multiple random projections greatly improves understanding of data structure. Indeed,

the capability of identifying within the discovered clusters those that can be considered reliable or not, may get insights into the underlying characteristics of patients clusters (Sect.6.2).

For instance, the proposed individual stability indices support the hypothesis of DLBCL subclasses within DLBCL patients (Sect. 6.2.1). On the contrary they do not directly support the Bittner's thesis about the reliability of the 19-member subclass inside melanoma patients [43], but suggest a larger more reliable and reproducible melanoma subclass composed by 23 specimens (Sect. 6.2.2).

Moreover, the proposed stability measures may help to discover multi-level structures in the data, e.g. a main cluster and subclusters discovered inside the main cluster itself, if both are supported by large values of the stability indices.

The need for individual cluster reliability comes also from other very common situations that arise from the analysis of clusters patients. For instance, even if an overall cluster stability measure can assess the "natural" number of clusters in the data, some clusters inside the obtained clustering can be more reliable than other, and some clusters may be not reliable at all: in the experimental part of this paper several situation of this kind have been depicted (Sect. 6). Moreover, less common situations in which only one cluster is present in the data and the other clusters are only noisy samples around the "true" cluster are very difficult to detect with a method searching for an optimal number of clusters only: we cannot distinguish the situations when only one cluster or no clusters are present. In all these cases we need a stability measure for each individual cluster found by the clustering algorithm.

Considering the effectiveness of the proposed cluster stability measures, we observe that *PMO*, *Achlioptas* and *Normal* random projections should be in general preferred to *RS* projections(Sect. 2.2), for both theoretical and experimental reasons.

Indeed the stability measures based on random projections are effective if a not too large distortion is induced into the projected data. In Sect. 2 we outlined that the Johnson-Lindenstrauss lemma provides bounds to the distortions induced by randomized maps, if suitable conditions are satisfied. *PMO*, *Achlioptas* and *Normal* random projections satisfy the *JL lemma*, and hence we can expect that a bounded distortion will be introduced into the projected space if a sufficiently high subspace dimension is chosen. On the contrary, this does not in general hold for random subspace projections, since *RS* does not obey the *JL lemma*. Our experiments for the evaluation of the distortions induced by randomized maps into lower dimensional subspaces confirm the theoretical results, showing that with DNA microarray data *RS* projections

may introduce relevant distortions into the data (Sect. 3). The same experiments show that *PMO*, *Achlioptas* and *Normal* are "well-behaved" projections, in the sense that the distortions of the projected data are largely inside the *JL* theoretical bounds, and in all cases significantly lower than that obtained with *RS* projections (Sect. 3). The stability measures evaluated with the *JL* compliant random projections are comparable in all data sets we analyzed in our experiments, while stability measures estimated with *RS* projections in some cases significantly differ from the others, especially when large ϵ values (corresponding to low dimensional subspaces) are chosen . Moreover the variance of the stability indices with respect to the multiple instances of the projected data is significantly larger when *RS* with respect to *PMO*, *Achlioptas* and *Normal* projections are used (data not shown).

Considering randomized maps that obey the *JL lemma*, we have no suggestions about the preferential choice of a particular one, since no significant differences have been observed in stability indices computed with *PMO*, *Achlioptas* and *Normal* projections.

We proposed a principled way to choose the subspace dimension of the embedded space. Indeed we proposed to use a dimension related to the distortion predicted according to the *JL lemma* (Sect. 4). Note that with *PMO*, *Achlioptas* and *Normal* projections we could in practice use also lower subspace dimensions than that predicted through the *JL* lemma, since our experimental results show that the empirical distortion bounds measured in gene expression data are significantly lower than the predicted theoretical bounds (Sect. 3), with corresponding savings in time and space computational resources. In practice our experimental results show that subspace dimensions corresponding to distortions $1 + \epsilon \leq 1.2$ are largely sufficient to reasonably compute the proposed stability indices.

Our work focused on stability indices to evaluate the reliability of individual clusters, but an overall stability measure of the clustering has been simply derived by averaging the reliability of the single clusters (Sect. 4.2). In spite of its straightforwardness, the proposed measure has revealed useful for analyzing the structure of patients clusters, as shown by our experiments. Nevertheless, if the main goal is to estimate the "natural" or "optimal" number of clusters we suggest to use also other more principled global measures based on distribution of some property of the data, such as measures based on distribution of pairwise similarity between clusterings of subsamples of a dataset [30].

As observed in Sect.4, our work may be set in the framework of cluster stability evaluation through multiple random "perturbation" of the data, where for perturbation we mean bootstrap samples drawn from the data [32, 35], or random noise injection into the data [33] or random projections into lower dimensional subspaces [34].

The basic idea behind this general approach consists in randomly perturbing in a suitable way the original data and then in using multiple instances of the perturbed data to assess the reliability of the clusters discovered with a suitable clustering algorithm. It is worth noting that all the methods that evaluate the reliability of the discovered clusters through the estimate of their stability need to properly modify and somehow distort the available data. Indeed using bootstrapping techniques only a subset of the data is available for each clustering, and adding noise we necessarily modify the original characteristics of the data. Random projections introduce distortion into the the original data too, but our approach permits on one hand to control the distortion, on the other hand to automatically find an "optimal" dimension for the projected subspaces.

A natural question that may arise from these considerations is: which of these general approaches appear to be best suited for the analysis of patients clusters in DNA microarray data? Experimental results provided in Sect. 6.2.3, with the warnings and cautions due to the characteristics of the problem, show that our approach is competitive with other proposed stability-based methods.

Our methods based on randomized maps are well-suited to the characteristics of DNA microarray data: indeed the low cardinality of the examples, the very large number of features (genes) involved in microarray chips, the redundancy of information stored in the spots of microarrays (as just discussed in Sect. 4) are all characteristics in favour of our approach. On the contrary using bootstrapping techniques to obtain smaller samples from just small samples of patients should induce more randomness in the estimate of cluster stability. This approach appears to be more well-suited to evaluate the cluster stability of genes, since significantly larger samples are available in this case [32]. Injecting noise into the data to obtain multiple instance of perturbed data poses difficult statistical problems for evaluating what kind and which magnitude of noise should be added to the data [34].

All the perturbation-based methods need to properly select a parameter to control the amount of perturbation of the data: resampled-based methods need to select the "optimal" fraction of the data to be subsampled; noise-injection-based methods needs to choice the amount of noise to be introduced; random subspace and random projections-based methods needs to select the proper dimension of the projected data. Anyway only our approach provides a theoretically motivated method to automatically find an "optimal" value for the perturbation parameter (see Sect. 4).

Another problem common to all perturbation-based methods is the dependence of the reliability measures on a specific cluster algorithm. Indeed also in our case, even if we may use any clustering algorithm that applies euclidean distances to analyze the data, the stability results will depend on the

particular choice of the clustering algorithm. It is well-known that different clustering algorithms may provide also very different clusters, and in this way not only we may have different estimates of cluster stability, but also different clusters found by the clustering algorithms. These are ineliminable problems arising from the fact that different clustering algorithms can discover different views and characteristics of the data. A possible seminal research line could be to integrate several different clustering algorithms in the evaluation of stability measures computed with multiple instances of projected data obtained through randomized maps, in order to explicitly consider the different results provided by different clustering algorithms.

Another related open issue is represented by data normalization. Even if in our experiments we did not observe many differences in the results when hierarchical clustering algorithms have been applied to normalized and non-normalized data, in [35] is reported that with self-organizing-map algorithms in some cases the results do not agree.

8 Conclusions

The problem of assessing the reliability of clusters patients obtained by clustering algorithms is crucial to estimate the significance of subclasses of diseases detectable at bio-molecular level, and more in general to support bio-medical discovery of patterns in gene expression data.

We proposed stability indices based on random projections with low metric distortion to measure the reliability of the clusters discovered by a suitable clustering algorithm. The proposed approach can be applied with any distance-based clustering algorithm and it does not require assumptions about the distribution of the data. Experiments with diffuse large B-cell Lymphoma and melanoma DNA microarray data showed how to apply the stability measures to the analysis of the reliability of patients clusters identified by hierarchical clustering algorithms. Moreover our experiments show that the average stability index may also be useful to estimate the most likely number of clusters in gene expression data.

Both theory on randomized maps between euclidean spaces and our experimental results show that random subspace projections, even if useful to analyze the stability of patients clusters, are less effective than Plus-Minus-One, Achlioptas and Normal random projections in assessing the reliability of the discovered clusters. For these reasons we strongly suggest to apply *JL lemma* compliant randomized maps to study the structure and the reliability of clusters patients.

We suggest also to choose the dimension of the projected subspace according to the *JL lemma*, to avoid too large distortions that could introduce noise in the estimate of the stability indices.

Acknowledgment

This work has been developed in the context of *CIMAINA* Center of Excellence and it has been partially funded by the italian COFIN project *Linguaggi formali ed automi: metodi, modelli ed applicazioni*. We would like to thank the anonymous reviewers for their comments and suggestions.

References

[1] Q. Ye, L. Qin, M. Forgues, P. He, J. Kim, A. Peng, R. Simon, Y. Li, A. Robles, Y. Chen, Z. Ma, Z. Wu, S. Ye, Y. Liu, Z. Tang, and X. Wang. Predicting hepatitis b virus-positive metastatic hepatocellular carcinomas using gene expression profiling and supervised machine learning. *Nature Medicine*, 9(4):416–423, 2003.

[2] S. Ramaswamy, K. Ross, E. Lander, and T. Golub. A molecular signature of metastasis in primary solid tumors. *Nature Genetics*, 33:49–54, 2003.

[3] Y. Yu, J. Khan, C. Khanna, L. Helman, P. Meltzer, and G. Merlino. Expression profiling identifies the cytoskeletal organizer ezrin and the developmental homoprotein Six-1 as key metastatic regulators. *Nature Medicine*, 10(2):175–181, 2004.

[4] A. Alizadeh, M.B. Eisen, R.E. Davis, C. Ma, I.S. Lossos, A. Rosenwald, J.C. Boldrick, H. Sabet, T. Tran, X. Yu, J.I. Powell, L. Yang, G.E. Marti, T. Moore, J. Hudson, L. Lu, D.B. Lewis, R. Tibshirani, G. Sherlock, W.C. Chan, T.C. Greiner, D.D. Weisenburger, J.O. Armitage, R. Warnke, R. Levy, W. Wilson, M.R. Grever, J.C. Byrd, D. Botstein, P.O. Brown, and L.M. Staudt. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*, 403:503–511, 2000.

[5] A. Rosenwald, R. Wright, W. Chan, J.M. Connors, R.I. Campo, E.and Fisher, R.D. Gascoyne, H.K. Muller-Hermelink, E.B. Smeland, J. Giltnane, E. Hurt, H. Zhao, L. Averett, L. Yang, W. Wilson, E. Jaffe, R. Simon, R.D. Klausner, J. Powell, P.L. Duffey, D.L. Longo, T.C. Greiner, D. Weisenburger, W.G. Sanger, B. Dave, J. Lynch, J. Vose, J. Armitage, E. Montserrat, A. Lopez-Guillermo, T. Grogan, T. Miller, M. LeBlanc, G. Ott, S. Kvaloy, J. Delabie, H. Holte, P. Krajci, T. Stokke, and L.M. Staudt. The use of molecular profiling to predict survival after chemoterapy for diffuse large-b-cell lymphoma. *New England Journal of Medicine*, 346(25):1937–1947, 2002.

[6] L. Dyrskjøt, T. Thykjaer, M. Kruhøffer, J. Jensen, N. Marcussen, S. Hamilton-Dutoit, H. Wolf, and T. Ørntoft. Identifying distinct classes of bladder carcinoma using microarrays. *Nature Genetics*, 33(jan.):90–96, 2003.

[7] D. Kotska and R. Spang. Finding disease specific alterations in the co–expression of genes. *Bioinformatics*, 20(suppl.1):i194–i199, 2004.

[8] J. Lapointe, C. Li, J.P. Higgins, M. van de Rijn, E. Bair, K. Montgomery, M. Ferrari, L. Egevad, W. Rayford, U. Bergerheim, P. Ekman, A.M. DeMarzo, R. Tibshirani, D. Botstein, P.O. Brown, J.D. Brooks, and J.R. Pollack. Gene expression profiling identifies clinically relevant subtypes of prostate cancer. *Proceedings of the National Academy of Sciences of the United States of America*, 101(3):811–816, 2004.

[9] M.B. Eisen, P.T. Spellman, P.O. Brown, and D. Botstein. Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences of the United States of America*, 95(25): 14863–14868, 1998.

[10] R. Shamir and R. Sharan. Algorithmic approaches to clustering gene expression data. In T. Jiang, Y. Xu, and M.Q. Zhang, editors, *Current Topics in Computational Biology*. MIT Press, Boston, USA, 269–300, 2001.

[11] M. Shipp, K. Ross, P. Tamayo, A. Weng, J. Kutok, R. Aguiar, M. Gaasenbeek, M. Angelo, M. Reich, G. Pinkus, T. Ray, M. Koval, K. Last, A. Norton, T. Lister, J. Mesirov, D. Neuberg, E. Lander, J. Aster, and T. Golub. Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. *Nature Medicine*, 8(1):68–74, 2002.

[12] G. Valentini. Gene expression data analysis of human lymphoma using support vector machines and output coding ensembles. *Artificial Intelligence in Medicine*, 26(3):283–306, 2002.

[13] S. Ramaswamy, P. Tamayo, R. Rifkin, S. Mukherjee, C.H. Yeang, M. Angelo, C. Ladd, M. Reich, E. Latulippe, J.P. Mesirov, T. Poggio, W. Gerald, M. Loda, E.S. Lander, and T.R. Golub. Multiclass cancer diagnosis using tumor gene expression signatures. *Proceedings of the National Academy of Sciences of the United States of America*, 98(26):15149–15154, 2001.

[14] B.J. Druker, M. Talpaz, D.J. Resta, B. Peng, E. Buchdunger, J.M. Ford, N.B. Lydon, H. Kantarjian, R. Capdeville, S. Ohno-Jones, and C.L. Sawyers. Efficacy and safety of a specific inhibitor of the bcr-abl tyrosine kinase in chronic myeloid leukemia. *New England Journal of Medicine*, 344(14):1031–1037, 2001.

[15] T.S. Furey, N. Cristianini, N. Duffy, D. Bednarski, M. Schummer, and D. Haussler. Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics*, 16 (10):906–914, 2000.

[16] J. Khan, J.S. Wei, M. Ringner, L.H. Saal, M. Ladanyi, F. Westermann, F. Berthold, M. Schwab, C.R. Antonescu, C. Peterson, and P.S. Meltzer.

Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nature Medicine*, 7(6):673–679, 2001.

[17] C. Yeang, S. Ramaswamy, P. Tamayo, S. Mukherjee, R.M. Rifkin, M. Angelo, M. Reich, E. Lander, J. Mesirov, and T. Golub. Molecular classification of multiple tumor types. In *ISMB 2001, Proceedings of the 9th International Conference on Intelligent Systems for Molecular Biology*, Copenaghen, Denmark, 2001. Special issue of *Bioinformatics*, vol.17, suppl.1, pages 316–322, Oxford University Press.

[18] W.L. Tung and C. Quek. GenSo-FDSS: a neural-fuzzy decision support system for pediatric all cancer subtype identification using gene expression data. *Artificial Intelligence in Medicine*, 33(1):61–88, 2005.

[19] A. Alizadeh, D.T. Ross, C.M. Perou, and M. van de Rijn. Towards a novel classification of human malignancies based on gene expression. *Journal of Pathology*, 195(1):41–52, 2001.

[20] A.K. Jain, M.N. Murty, and P.J. Flynn. Data Clustering: a Review. *Association for Computing Machinery Computing Surveys*, 31(3):264–323, 1999.

[21] T. Hastie, R. Tibshirani, and R. Friedman. *The Elements of Statistical Learning*. Springer, New York, 2001.

[22] T. Kohonen. The self-organizing map. *Neurocomputing*, 21:1–6, 1998.

[23] D.M. Chickering and D. Heckerman. Efficient approximation for the marginal likelihood of bayesian networks with hidden variables. *Machine Learning*, 29:181–212, 1997.

[24] K.Y. Yeung, C. Fraley, A. Murua, A.E. Raftery, and W.L. Ruzzo. Model-based clustering and data transformations for gene expression data. *Bioinformatics*, 17(10):977–987, 2001.

[25] Y. Cheng and G.M. Church. Biclustering of expression data. In R. Altman, T. Bailey, P. Bourne, M. Gribskov, T Lengauer, I Shindyalov, L. Ten Eyck, and H. Weissig editors, *Proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology*, pages 93–103. AAAI Press, Menlo Park, California, USA, 2000.

[26] S.C. Madeira and A.L. Oliveira. Biclustering algorithms for biological data analysis: A survey. *IEEE/Association for Computing Machinery Transactions on Computational Biology and Bioinformatics*, 1(1):24–45, 2004.

[27] S. Dudoit and J. Fridlyand. A prediction-based resampling method for estimating the number of clusters in a dataset. *Genome Biology*, 3(7): 1–21, 2002.

[28] S. Dudoit and J. Fridlyand. Bagging to improve the accuracy of a clustering procedure. *Bioinformatics*, 19(9):1090–1099, 2003.

[29] A. Ben-Dor, R. Shamir, and Z. Yakhini. Clustering gene expression patterns. *Journal of Computational Biology*, 6(3):281–297, 1999.

[30] A. Ben-Hur, A. Ellisseeff, and I. Guyon. A stability based method for discovering structure in clustered data. In R.B. Altman, A.K. Dunker,

L. Hunter, T. Klein, and K. Lauderdale, editors, *Pacific Symposium on Biocomputing*, volume 7, pages 6–17, Lihue, Hawaii, USA, 2002. World Scientific.

[31] K.Y. Yeung, D.R. Haynor, and W.L. Ruzzo. Validating clustering for gene expression data. *Bioinformatics*, 17(4):309–318, 2001.

[32] M.K. Kerr and G.A. Curchill. Bootstrapping cluster analysis: assessing the reliability of conclusions from microarray experiments. *Proceedings of the National Academy of Sciences of the United States of America*, 98: 8961–8965, 2001.

[33] L.M. McShane, D. Radmacher, B. Freidlin, R. Yu, M.C. Li, and R. Simon. Method for assessing reproducibility of clustering patterns observed in analyses of microarray data. *Bioinformatics*, 18(11):1462–1469, 2002.

[34] M. Smolkin and D. Gosh. Cluster stability scores for microarray data in cancer studies. *BMC Bioinformatics*, 36(4), 2003.

[35] S. Monti, P. Tamayo, J. Mesirov, and T. Golub. Consensus Clustering: A Resampling-based Method for Class Discovery and Visualization of Gene Expression Microarray Data. *Machine Learning*, 52:91–118, 2003.

[36] J.J. Chen, R. Delongchamp, C. Tsai, H. Hsueh, F. Sisatare, K. Thompson, V. Deasi, and J. Fuscoe. Analysis of variance components in gene expression data. *Bioinformatics*, 20(9):1436–1446, 2004.

[37] W.B. Johnson and J. Lindenstrauss. Extensions of Lipshitz mapping into Hilbert space. In *Conference in modern analysis and probability*, volume 26 of *Contemporary Mathematics*, pages 189–206. American Mathematical Society, 1984.

[38] T.K. Ho. The random subspace method for constructing decision forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(8): 832–844, 1998.

[39] P. Indyk. Algorithmic Applications of Low-Distortion Geometric Embeddings. In *Proceedings of the 42nd IEEE symposium on Foundations of Computer Science*, pages 10–33, Washington DC, USA, 2001. IEEE Computer Society.

[40] D. Achlioptas. Database-friendly random projections. In P. Buneman, editor, *Proc. Association for Computing Machinery Symp. on the Principles of Database Systems*, Contemporary Mathematics, pages 274–281, New York, NY, USA, 2001. Association for Computing Machinery Press.

[41] E. Bingham and H. Mannila. Random projection in dimensionality reduction: Applications to image and text data. In R. Kohavi and B. Masand editors: *Proc. of Knowledge Discovery and Data Mining 2001*, San Francisco, CA, USA, 2001. Association for Computing Machinery Press.

[42] A. Bhattacharjee, W.G. Richards, J. Staunton, C. Li, S. Monti, P. Vasa, C. Ladd, J. Beheshti, R. Bueno, M. Gillette, M. Loda, G. Weber, E.J. Mark, E.S. Lander, W. Wong, B.E. Johnson, T.R. Golub, D.J. Sugarbaker, and M. Meyerson. Classification of human lung carcinoma by mRNA expression profiling reveals distinct adenocarcinoma subclasses. *Proceedings of the National Academy of Sciences of the United States of*

America, 98(24):13790–13795, 2001.

[43] M. Bittner, P. Meltzer, Y. Chen, Y. Jiang, E. Seftor, M. Hendrix, M. Radmacher, R. Simon, Z. Yakhini, A. Ben-Dor, N. Sampas, E. Dougherty, E. Wang, F. Marincola, C. Gooden, J. Lueders, A. Glatfelter, P. Pollock, J. Carpten, E. Gillanders, D. Leja, K. Dietrich, C. Beaudry, M. Berens, D. Alberts, and V. Sondak. Molecular classification of malignant melanoma by gene expression profiling. *Nature*, 406:536–540, 2000.

[44] A. Martoglio, J. Miskin, S. Smith, and D. MacKay. A decomposition model to track gene expression signatures: preview on observer-independent classification of ovarian cancer. *Bioinformatics*, 18(12):1617–1624, 2002.

[45] S. Hadjitodorov, L. Kuncheva, and L. Todorova. Moderate Diversity for Better Cluster Ensembles. *Information Fusion*. Published on-line: `http://www.sciencedirect.com` (Accessed: 23 March 2006), 2005.

[46] C. Windischberger, M. Barth, C. Lamm, L. Schroeder, H. Bauer, R.C. Gur, and E. Moser. Fuzzy cluster analysis of high-field functional MRI data. *Artificial Intelligence in Medicine*, 29(3):203–223, 2003.

[47] F. Masulli and A. Schenone. A fuzzy clustering based segmentation system as support to diagnosis in medical imaging. *Artificial Intelligence in Medicine*, 16:129–147, 1999.

[48] S.P. Smith and R. Dubes. Stability of a hierarchical clustering. *Pattern Recognition*, 12:177–187, 1980.

[49] E. Levine and E. Domany. Resampling method for unsupervised estimation of cluster validity. *Neural Computation*, 13(11):2573–2593, 2001.

[50] G.W. Milligan and M.G. Cooper. An examination of procedures for determining the number of clusters in a data set. *Psychometrika*, 50:159–179, 1985.

[51] S. Chu, J. DeRisi, M. Eisen, J. Mulholland, J. Botstein, and P.O. Brown. *Science*, 282:699–705, 1998.

[52] J. McQueen. Some methods for classification and analysis of multivariate observations. In L.M. LeCam and J. Neyman, editors, *Proceedings of the Fifth Berkeley Symposium of Mathematical Statistics and Probability*, pages 281–297. University of California Press, 1967.

[53] B. King. Step-wise clustering procedures. *Journal of the American Statistical Association*, 69:86–101, 1967.

[54] J.H. Ward. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58:236–244, 1963.

[55] G. Valentini. Clusterv: a tool for assessing the reliability of clusters discovered in DNA microarray data. *Bioinformatics*, 22(3):369–370, 2006.

[56] R. Gentleman, V. Carey, W. Huber, R. Irizarry, and S. Dudoit. *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*. Springer, New York, 2005.

[57] D. Ghosh and A.M. Chinnaiyan. Mixture modelling of gene expression data from microarray experiments. *Bioinformatics*, 18(2):275–286, 2002.

[58] P. Brafford and M. Herlyn. Gene expression profil-

ing of melanoma cells: searching the haystack. *Journal of Translational Medicine*, 3, 2005. Published on-line: `http://www.pubmedcentral.nih.gov/tocrender.fcgi?iid=18084` (Accessed: 23 March 2006).

List of Figures

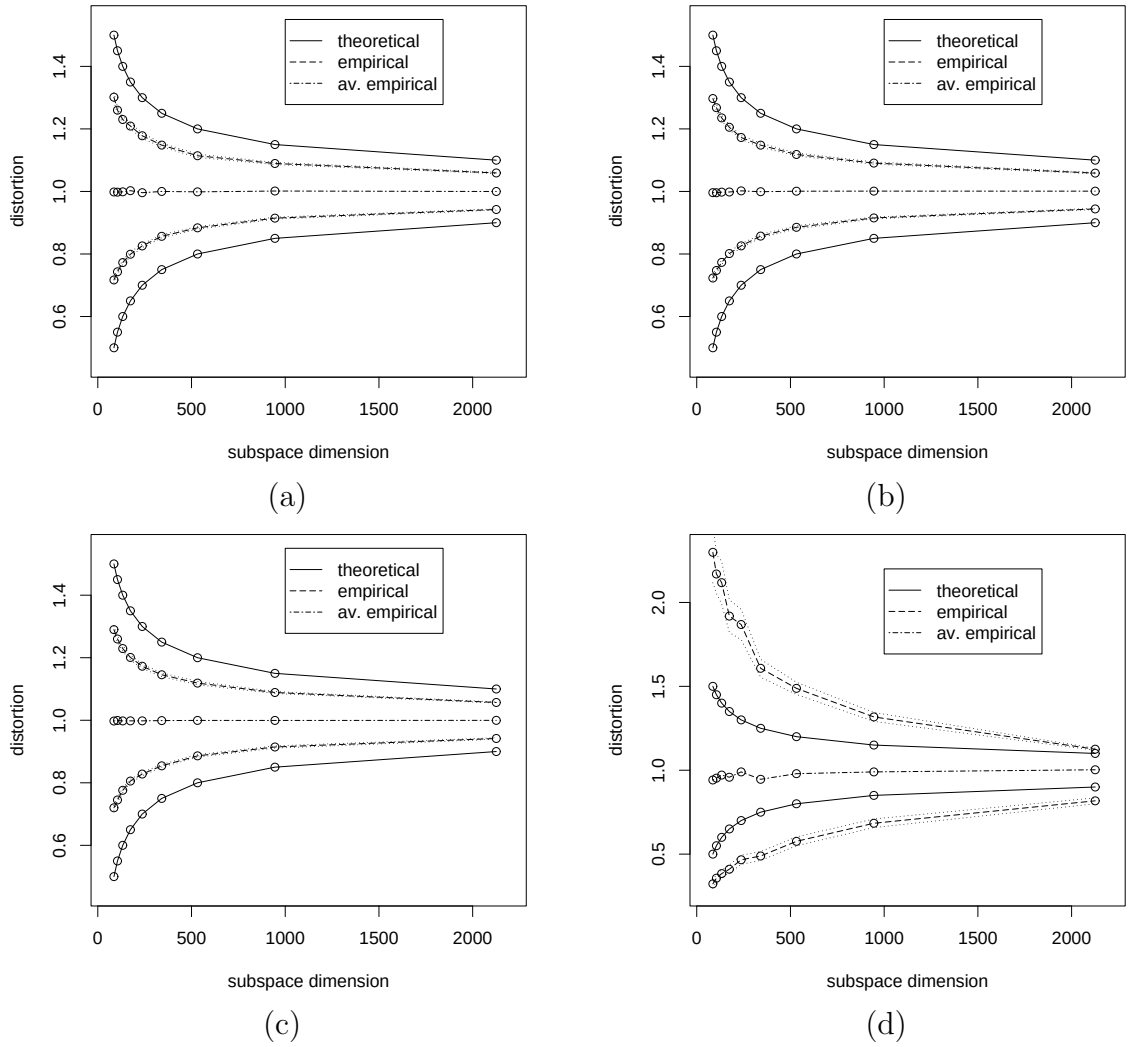

Fig. 1. Comparing theoretical and empirical distortion with *Lung tumor* samples using (a) *Achlioptas* (b) *Normal* (c) *PMO* and (d) *RS* projections.

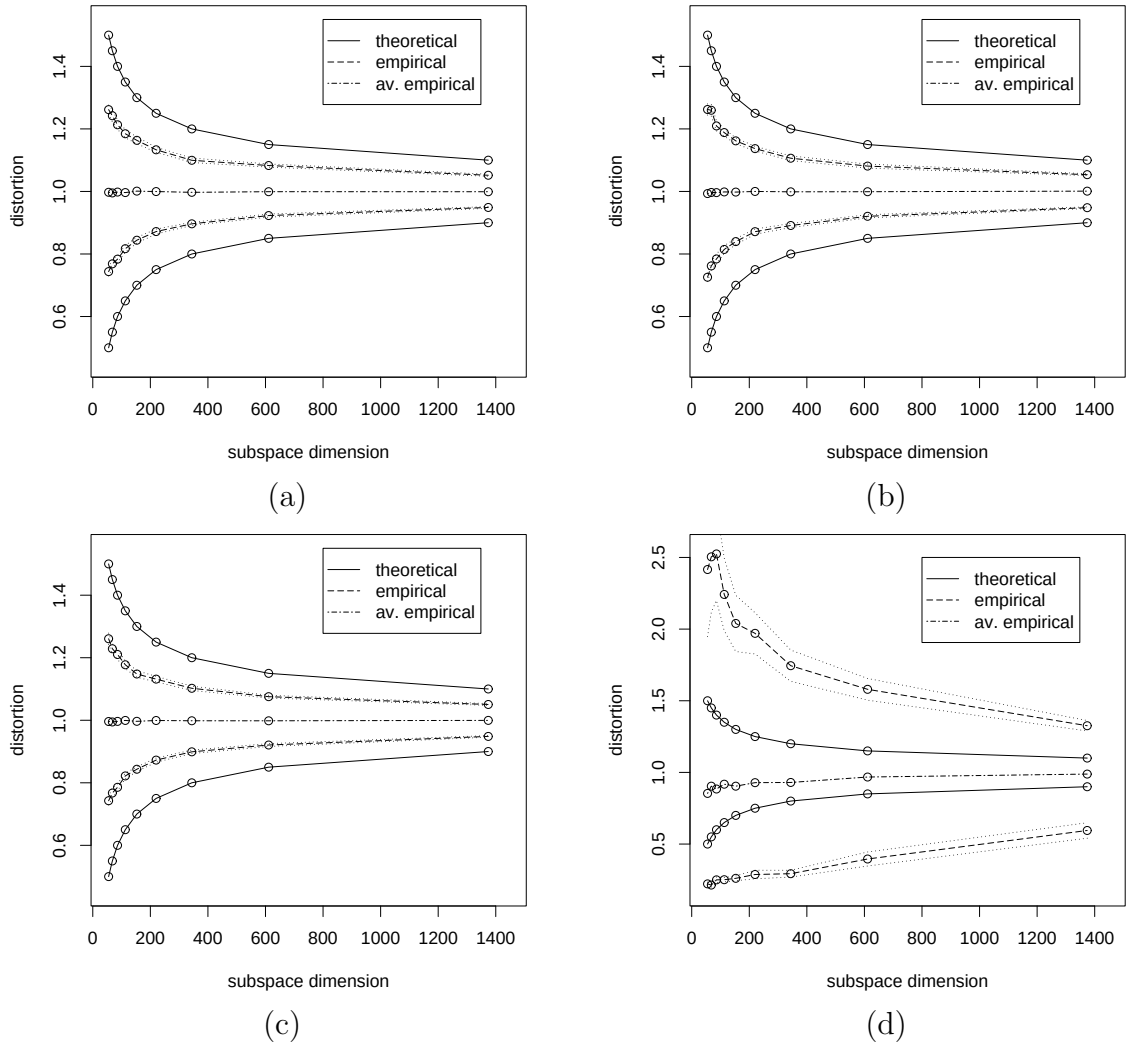

Fig. 2. Comparing theoretical and empirical distortion with *Melanoma* samples using (a) *Achlioptas* (b) *Normal* (c) *PMO* and (d) *RS* projections.

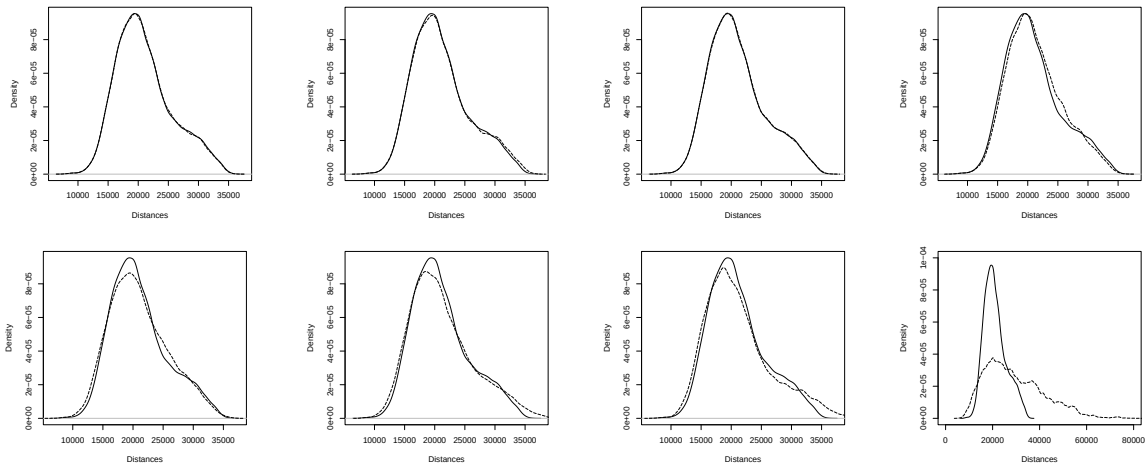

Fig. 3. *Lung tumor* DNA microarray data. Empirical distribution of the pairwise distances between examples in original and projected data. The continuous line represents the distribution in the original 3312-dimensional space, the dashed line the distribution in the projected space. First row of figures: projection to a 2126-dimensional space ($\epsilon = 0.1$); second row: projection to an 86-dimensional space ($\epsilon = 0.5$). From left to right in both rows are respectively represented *Achlioptas*, *Normal*, *PMO* and *RS* projections.

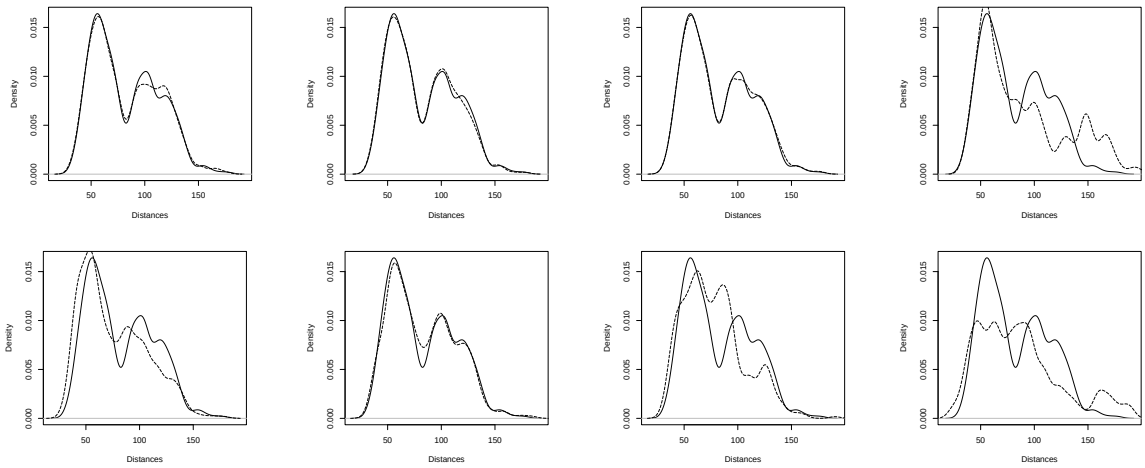

Fig. 4. *Melanoma* DNA microarray data. Empirical distribution of the pairwise distances between examples in original and projected data. The continuous line represents the distribution in the original 3613-dimensional space, the dashed line the distribution in the projected space. First row of figures: projection to a 1374-dimensional space ($\epsilon = 0.1$); second row: projection to a 55-dimensional space ($\epsilon = 0.5$). From left to right in both rows are respectively represented *Achlioptas*, *Normal*, *PMO* and *RS* projections.

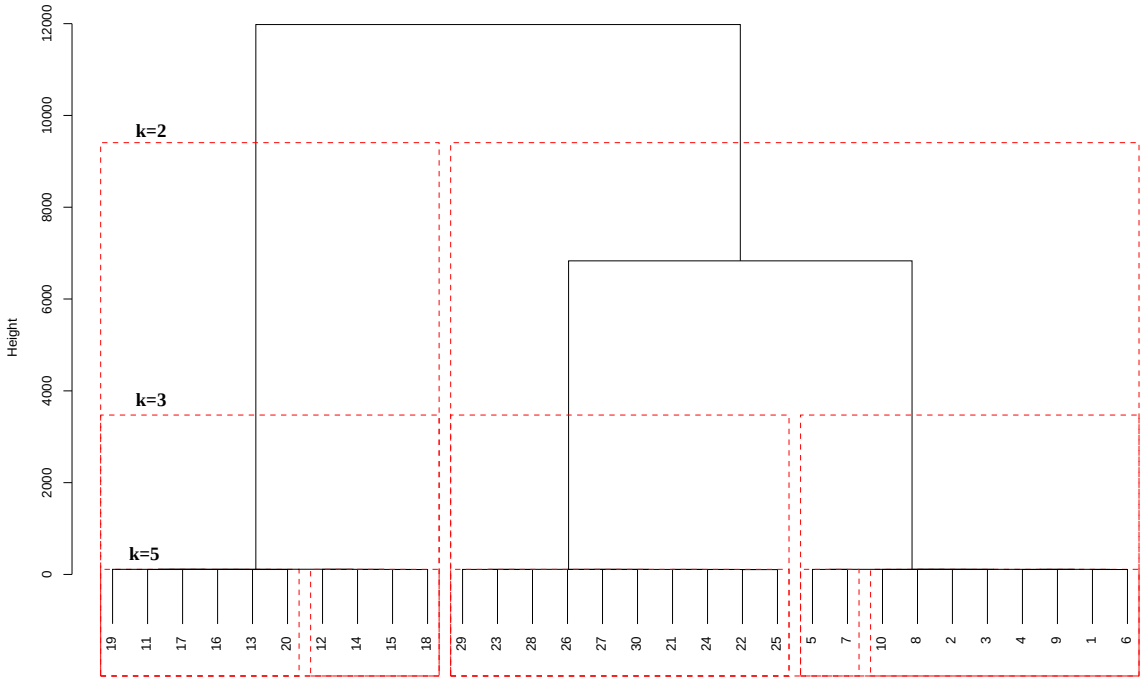

Fig. 5. Hierarchical clustering of *sample1* examples (Ward method). Gray dotted lines cut the dendrogram such that exactly k clusters are produced, for $k = 2, 3, 5$.

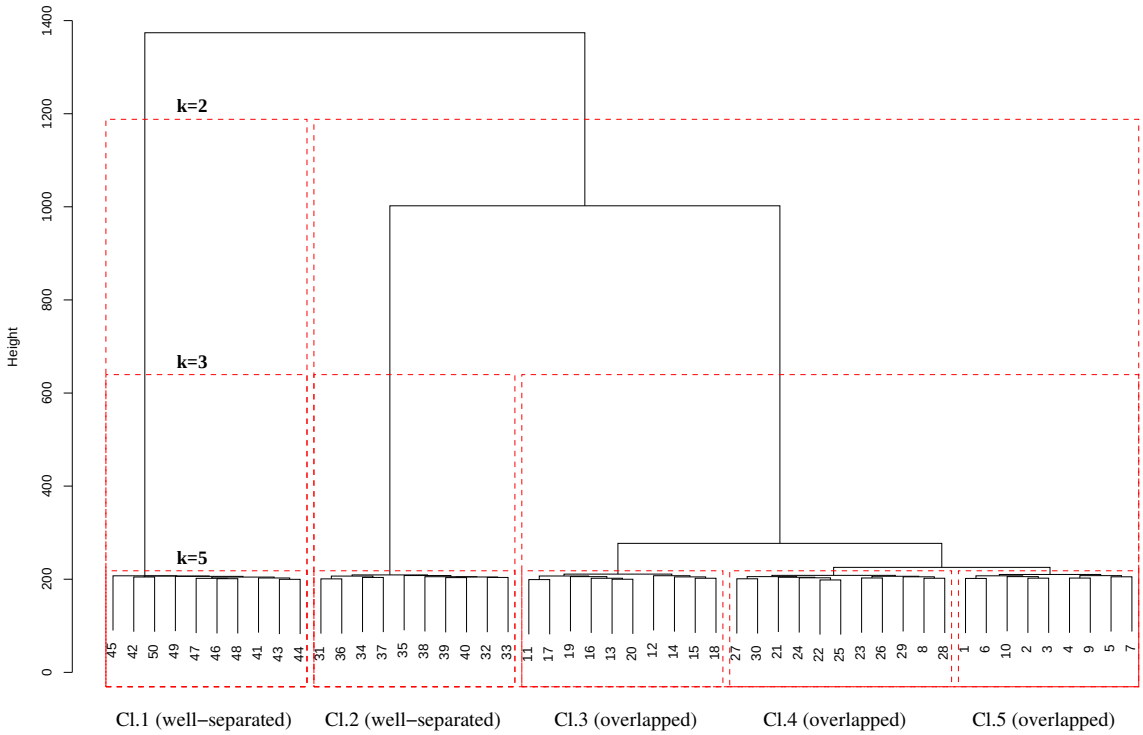

Fig. 6. Hierarchical clustering of *sample2* examples (Ward method). Gray dotted lines cut the dendrogram such that exactly k clusters are produced, for $k = 2, 3, 5$. The two clusters on the left are identified as stable, while the remaining three are evaluated as less table (see Table 2).

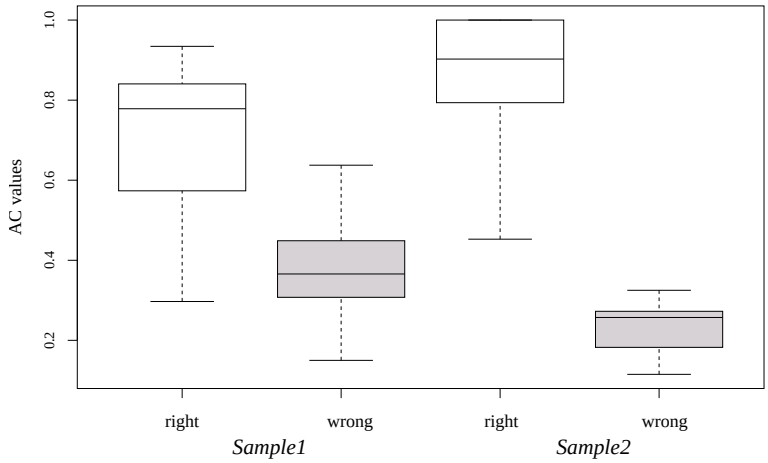

Fig. 7. Boxplots of the AC indices distributions with the synthetic data *Sample1* and *Sample2*. White boxes refer to correct predictions, gray boxes to wrong predictions. The tick line inside the boxes represents the median value, the bottom and the top of the boxes represent respectively the first and third quartile.

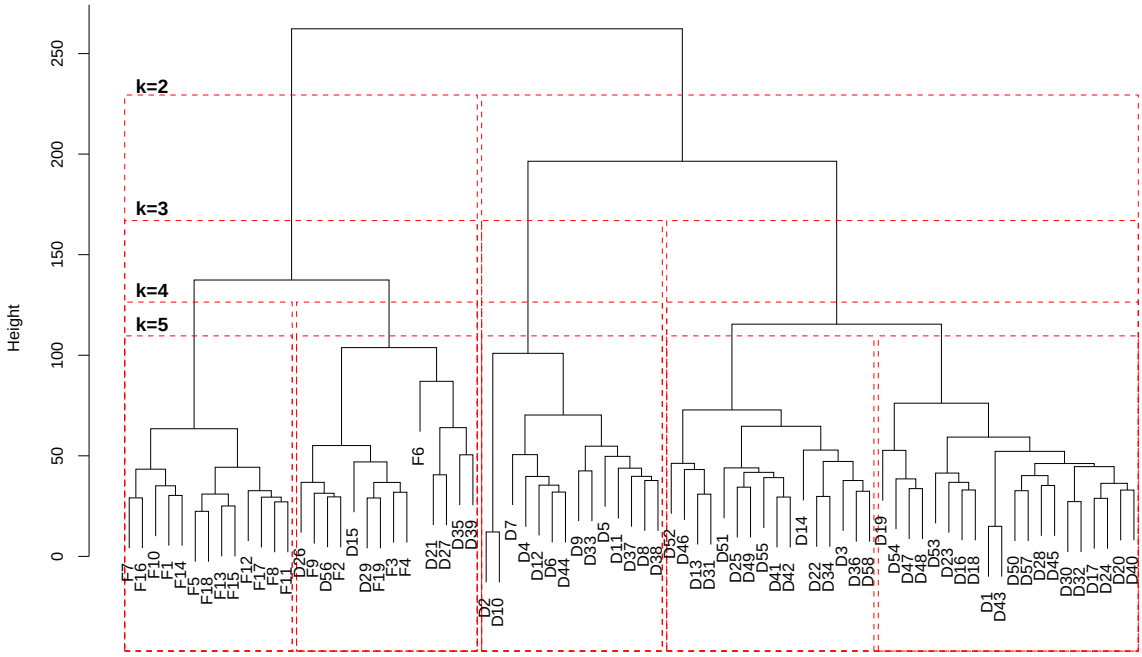

Fig. 8. Hierarchical clustering of *DLBCL-FL* examples (Ward method). Leaves labeled with "D" refer to DLBCL patients, while "F" to FL patients. Gray dotted lines cut the dendrogram such that exactly k clusters are produced, for $k = 2, 3, 4, 5$.

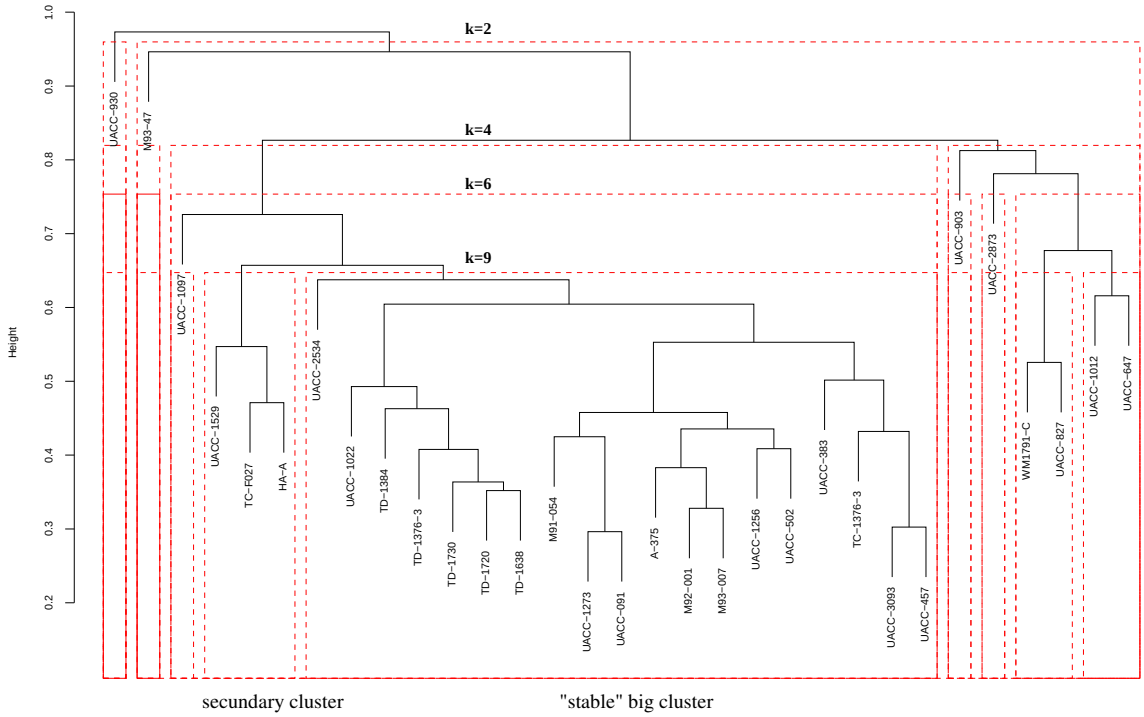

Fig. 9. Hierarchical clustering of *Melanoma* samples (Average linkage method with 1- Pearson dissimilarity measure). Gray dotted lines cut the dendrogram such that exactly k clusters are produced, for $k = 2, 4, 6, 9$. We pointed out the big "stable" cluster discovered by Bittner. See Table 4 and 5 for the the corresponding stability indices.

List of Tables

Table 1
Sample1: Estimate of cluster stability.

Clusters	**Members of Clusters**	**Stability index s**				
		$\epsilon = 0.5$	$\epsilon = 0.4$	$\epsilon = 0.3$	$\epsilon = 0.2$	$\epsilon = 0.1$
2 clusters		$S = 0.8631$	$S = 0.8684$	$S = 0.8684$	$S = 0.9157$	$S = 0.9421$
1	11-20	1.0000	1.0000	1.0000	1.0000	1.0000
2	1-10,21-30	0.7263	0.7368	0.7368	0.8314	0.8842
3 clusters		$S = 1.0000$	$S = 1.0000$	$S = 1.0000$	$S = 1.0000$	$S = 1.0000$
1	11-20	1.0000	1.0000	1.0000	1.0000	1.0000
2	21-30	1.0000	1.0000	1.0000	1.0000	1.0000
3	1-10	1.0000	1.0000	1.0000	1.0000	1.0000
5 clusters		$S = 0.7059$	$S = 0.6843$	$S = 0.7044$	$S = 0.7004$	$S = 0.7472$
1	11,13,16,17,19,20	0.6973	0.7346	0.7293	0.6506	0.7560
2	12,14,15,18	0.6666	0.7066	0.6866	0.6466	0.7133
3	21-30	0.7155	0.7582	0.7448	0.7591	0.8364
4	5,7	0.7600	0.5600	0.6800	0.7400	0.7800
5	1-4,6,8-10	0.6900	0.6621	0.6814	0.7057	0.6507
10 clusters		$S = 0.3093$	$S = 0.3043$	$S = 0.2651$	$S = 0.3286$	$S = 0.3936$
1	19	0.0600	0.1200	0.0600	0.2000	0.2400
2	11,13,16,17,20	0.4260	0.3520	0.2900	0.3360	0.4560
3	12	0.1400	0.1600	0.1600	0.2000	0.1400
4	14,15,18	0.4066	0.3533	0.3200	0.3800	0.4200
5	23,28,29	0.3733	0.3000	0.2866	0.3600	0.4200
6	21,22,24-27,30	0.3276	0.3419	0.3285	0.3866	0.3933
7	5,7	0.3600	0.2800	0.3000	0.3600	0.3800
8	2,3,8,10	0.3000	0.3366	0.3066	0.3433	0.3866
9	4,9	0.3400	0.4000	0.2600	0.4200	0.5000
10	1,6	0.3600	0.4000	0.3400	0.3000	0.6000

Table 2
Sample2: Estimate of cluster stability.

N.	**Overall stability index S**				
	$\epsilon = 0.5$	$\epsilon = 0.4$	$\epsilon = 0.3$	$\epsilon = 0.2$	$\epsilon = 0.1$
2	0.8846	0.8959	0.8846	0.9115	0.9153
3	0.9872	0.9964	1.0000	1.0000	1.0000
4	0.7991	0.8103	0.8393	0.8886	0.9159
5	0.6445	0.6739	0.7083	0.7620	0.8412
6	0.5443	0.5819	0.5956	0.6707	0.7652
8	0.4382	0.4709	0.4860	0.5316	0.6270
10	0.3073	0.3423	0.3366	0.3812	0.4843
20	0.1694	0.1895	0.1911	0.2183	0.3339

N.	**Cl.**	**Stability index s**				
		$\epsilon = 0.5$	$\epsilon = 0.4$	$\epsilon = 0.3$	$\epsilon = 0.2$	$\epsilon = 0.1$
2	1	1.0000	1.0000	1.0000	1.0000	1.0000
	2	0.7693	0.7918	0.7692	0.8230	0.8307
3	1	0.9960	0.9920	1.0000	1.0000	1.0000
	2	0.9893	1.0000	1.0000	1.0000	1.0000
	3	0.9765	0.9973	1.0000	1.0000	1.0000
4	1	0.9960	0.9920	1.0000	1.0000	1.0000
	2	0.9711	0.9893	1.0000	1.0000	1.0000
	3	0.6315	0.6782	0.6991	0.9088	0.9760
	4	0.5981	0.5818	0.6582	0.6456	0.6877
5	1	0.9662	0.9533	0.9928	0.9906	1.0000
	2	0.9511	0.9800	0.9782	0.9817	1.0000
	3	0.4555	0.4973	0.5560	0.7351	0.8422
	4	0.4378	0.5505	0.5501	0.6672	0.7596
	5	0.4122	0.3883	0.4644	0.4355	0.6044
10	1	0.5662	0.5284	0.5364	0.6342	0.6457
	2	0.5033	0.6133	0.5866	0.6033	0.6500
	3	0.0200	0.0200	0.0000	0.0400	0.0800
	4	0.5480	0.6560	0.5920	0.5660	0.6240
	5	0.2600	0.2866	0.2973	0.3573	0.5013
	6	0.2833	0.3700	0.3300	0.3766	0.5800
	7	0.2853	0.3066	0.3253	0.4133	0.6240
	8	0.2100	0.2500	0.2660	0.3240	0.3620
	9	0.2140	0.1960	0.2060	0.2340	0.3960
	10	0.1833	0.1966	0.2266	0.2633	0.3800

Table 3
DLBCL-FL: Estimate of cluster stability with PMO projections.

N.		**Overall stability index S**				
		$\epsilon = 0.5$	$\epsilon = 0.4$	$\epsilon = 0.3$	$\epsilon = 0.2$	$\epsilon = 0.1$
2		0.7744	0.8097	0.8290	0.8736	0.9523
3		0.7363	0.7698	0.8086	0.8454	0.9433
4		0.6896	0.7465	0.8169	0.8266	0.9084
5		0.6556	0.7031	0.7816	0.7991	0.8724
N.	**Cl.**	**Stability index s**				
		$\epsilon = 0.5$	$\epsilon = 0.4$	$\epsilon = 0.3$	$\epsilon = 0.2$	$\epsilon = 0.1$
2	1	0.8353	0.8853	0.9340	0.9446	0.9797
	2	0.7136	0.7340	0.7241	0.8026	0.9249
3	1	0.6394	0.6578	0.7651	0.8445	0.9478
	2	0.8602	0.9158	0.9389	0.9435	0.9800
	3	0.7092	0.7358	0.7218	0.7483	0.9022
4	1	0.8010	0.8133	0.8943	0.9676	0.9882
	2	0.5907	0.6778	0.7624	0.6720	0.7637
	3	0.7720	0.8501	0.9178	0.9208	0.9800
	4	0.5945	0.6450	0.6932	0.7458	0.9016
5	1	0.7858	0.7779	0.8887	0.9646	0.9882
	2	0.5103	0.6241	0.7017	0.6516	0.7367
	3	0.7164	0.7461	0.8079	0.8136	0.9272
	4	0.5986	0.5970	0.6830	0.6953	0.8550
	5	0.6668	0.7705	0.8268	0.8703	0.8548

Table 4
Melanoma: Estimate of cluster stability using PMO projections

N.	**Overall stability index S**				
	$\epsilon = 0.5$	$\epsilon = 0.4$	$\epsilon = 0.3$	$\epsilon = 0.2$	$\epsilon = 0.1$
2	0.8186	0.8613	0.9040	0.9253	1.0000
3	0.8946	0.8752	0.9129	0.9786	1.0000
4	0.8907	0.9266	0.9618	0.9728	0.9782
5	0.7010	0.7306	0.7384	0.7430	0.7316
6	0.5800	0.5950	0.5942	0.5929	0.5930
7	0.4789	0.4947	0.4950	0.4930	0.4969
8	0.4998	0.5272	0.5357	0.5407	0.5481
9	0.4977	0.5098	0.5149	0.5138	0.5187
10	0.5049	0.5245	0.5370	0.5389	0.5378
12	0.3993	0.3795	0.3998	0.3812	0.3910

N.	**Cl.**	**Stability index s**				
		$\epsilon = 0.5$	$\epsilon = 0.4$	$\epsilon = 0.3$	$\epsilon = 0.2$	$\epsilon = 0.1$
2	1	0.6600	0.7400	0.8200	0.8600	1.0000
	2	0.9773	0.9826	0.9880	0.9906	1.0000
3	1	0.9600	0.9600	0.9200	1.0000	1.0000
	2	0.7600	0.7000	0.8400	0.9400	1.0000
	3	0.9639	0.9658	0.9789	0.9958	1.0000
4	1	0.9600	1.0000	0.9800	1.0000	1.0000
	2	0.8000	0.8800	0.9800	0.9800	1.0000
	3	0.8098	0.8265	0.8875	0.9113	0.9130
	4	0.9933	1.0000	1.0000	1.0000	1.0000
5	1	0.9600	1.0000	1.0000	1.0000	1.0000
	2	0.9200	0.9800	0.9800	1.0000	1.0000
	3	0.6534	0.6733	0.7124	0.7152	0.6580
	4	0.0000	0.0000	0.0000	0.0000	0.0000
	5	0.9720	1.0000	1.0000	1.0000	1.0000
6	1	1.0000	1.0000	1.0000	1.0000	1.0000
	2	0.9800	1.0000	1.0000	1.0000	1.0000
	3	0.5635	0.5802	0.5657	0.5577	0.5584
	4	0.0000	0.0000	0.0000	0.0000	0.0000
	5	0.0000	0.0000	0.0000	0.0000	0.0000
	6	0.9366	0.9900	1.0000	1.0000	1.0000
9	1	1.0000	1.0000	1.0000	1.0000	1.0000
	2	1.0000	1.0000	1.0000	1.0000	1.0000
	3	0.0000	0.0000	0.0000	0.0000	0.0000
	4	0.6066	0.5200	0.4933	0.4733	0.3466
	5	0.3732	0.3888	0.3810	0.3914	0.4023
	6	0.0000	0.0000	0.0000	0.0000	0.0000
	7	0.0000	0.0000	0.0000	0.0000	0.0000
	8	0.6600	0.7400	0.8000	0.7800	0.9400
	9	0.8400	0.9400	0.9600	0.9800	0.9800

Table 5
Melanoma: Estimate of cluster stability. Left: Achlioptas projections; Right: "Normal" projections.

N.	**Overall stability index S**				
	$\epsilon = 0.5$	$\epsilon = 0.4$	$\epsilon = 0.3$	$\epsilon = 0.2$	$\epsilon = 0.1$
2	0.8186	0.8613	0.9040	0.9253	1.0000
3	0.8946	0.8752	0.9129	0.9786	1.0000
4	0.8907	0.9266	0.9618	0.9728	0.9782
5	0.7010	0.7306	0.7384	0.7430	0.7316
6	0.5800	0.5950	0.5942	0.5929	0.5930
7	0.4789	0.4947	0.4950	0.4930	0.4969
8	0.4998	0.5272	0.5357	0.5407	0.5481
9	0.4977	0.5098	0.5149	0.5138	0.5187
10	0.5049	0.5245	0.5370	0.5389	0.5378
12	0.3993	0.3795	0.3998	0.3812	0.3910

N.	**Cl.**	**Stability index s**				
		$\epsilon = 0.5$	$\epsilon = 0.4$	$\epsilon = 0.3$	$\epsilon = 0.2$	$\epsilon = 0.1$
2	1	0.6600	0.7400	0.8200	0.8600	1.0000
	2	0.9773	0.9826	0.9880	0.9906	1.0000
3	1	0.9600	0.9600	0.9200	1.0000	1.0000
	2	0.7600	0.7000	0.8400	0.9400	1.0000
	3	0.9639	0.9658	0.9789	0.9958	1.0000
4	1	0.9600	1.0000	0.9800	1.0000	1.0000
	2	0.8000	0.8800	0.9800	0.9800	1.0000
	3	0.8098	0.8265	0.8875	0.9113	0.9130
	4	0.9933	1.0000	1.0000	1.0000	1.0000
5	1	0.9600	1.0000	1.0000	1.0000	1.0000
	2	0.9200	0.9800	0.9800	1.0000	1.0000
	3	0.6534	0.6733	0.7124	0.7152	0.6580
	4	0.0000	0.0000	0.0000	0.0000	0.0000
	5	0.9720	1.0000	1.0000	1.0000	1.0000
6	1	1.0000	1.0000	1.0000	1.0000	1.0000
	2	0.9800	1.0000	1.0000	1.0000	1.0000
	3	0.5635	0.5802	0.5657	0.5577	0.5584
	4	0.0000	0.0000	0.0000	0.0000	0.0000
	5	0.0000	0.0000	0.0000	0.0000	0.0000
	6	0.9366	0.9900	1.0000	1.0000	1.0000
9	1	1.0000	1.0000	1.0000	1.0000	1.0000
	2	1.0000	1.0000	1.0000	1.0000	1.0000
	3	0.0000	0.0000	0.0000	0.0000	0.0000
	4	0.6066	0.5200	0.4933	0.4733	0.3466
	5	0.3732	0.3888	0.3810	0.3914	0.4023
	6	0.0000	0.0000	0.0000	0.0000	0.0000
	7	0.0000	0.0000	0.0000	0.0000	0.0000
	8	0.6600	0.7400	0.8000	0.7800	0.9400
	9	0.8400	0.9400	0.9600	0.9800	0.9800

N.	**Overall stability index S**				
	$\epsilon = 0.5$	$\epsilon = 0.4$	$\epsilon = 0.3$	$\epsilon = 0.2$	$\epsilon = 0.1$
2	0.7866	0.8720	0.9360	0.9573	1.0000
3	0.8114	0.8896	0.9496	0.9786	1.0000
4	0.8988	0.9347	0.9628	0.9769	0.9782
5	0.7122	0.7264	0.7341	0.7369	0.7314
6	0.5867	0.5892	0.5950	0.5919	0.5938
7	0.4814	0.4879	0.4965	0.4961	0.4958
8	0.5202	0.5248	0.5358	0.5410	0.5452
9	0.4985	0.5004	0.5070	0.5161	0.5208
10	0.5133	0.5212	0.5312	0.5513	0.5427
12	0.3839	0.3798	0.3874	0.4004	0.3997

N.	**Cl.**	**Stability index s**				
		$\epsilon = 0.5$	$\epsilon = 0.4$	$\epsilon = 0.3$	$\epsilon = 0.2$	$\epsilon = 0.1$
2	1	0.6000	0.7600	0.8800	0.9200	1.0000
	2	0.9733	0.9840	0.9920	0.9946	1.0000
3	1	0.8600	0.9600	0.9800	1.0000	1.0000
	2	0.6200	0.7400	0.8800	0.9400	1.0000
	3	0.9543	0.9690	0.9890	0.9958	1.0000
4	1	0.9400	0.9800	1.0000	1.0000	1.0000
	2	0.8200	0.9000	0.9800	1.0000	1.0000
	3	0.8488	0.8582	0.8713	0.9070	0.9130
	4	0.9867	1.0000	1.0000	1.0000	1.0000
5	1	0.9800	1.0000	1.0000	1.0000	1.0000
	2	0.9200	0.9600	1.0000	1.0000	1.0000
	3	0.6773	0.6842	0.6705	0.6849	0.6571
	4	0.0000	0.0000	0.0000	0.0000	0.0000
	5	0.9840	0.9880	1.0000	1.0000	1.0000
6	1	1.0000	1.0000	1.0000	1.0000	1.0000
	2	0.9600	1.0000	1.0000	1.0000	1.0000
	3	0.6007	0.5754	0.5703	0.5519	0.5633
	4	0.0000	0.0000	0.0000	0.0000	0.0000
	5	0.0000	0.0000	0.0000	0.0000	0.0000
	6	0.9600	0.9600	1.0000	1.0000	1.0000
9	1	1.0000	1.0000	1.0000	1.0000	1.0000
	2	0.9800	1.0000	1.0000	1.0000	1.0000
	3	0.0000	0.0000	0.0000	0.0000	0.0000
	4	0.5200	0.5333	0.6400	0.4533	0.3733
	5	0.3871	0.3902	0.3831	0.3922	0.3939
	6	0.0000	0.0000	0.0000	0.0000	0.0000
	7	0.0000	0.0000	0.0000	0.0000	0.0000
	8	0.6400	0.6400	0.6200	0.8400	0.9200
	9	0.9600	0.9400	0.9200	0.9600	1.0000

Table 6
Comparison of different stability-based methods with the *melanoma* data set

Perturbation by noise-injection (Mc Shane et al.)							
7 clusters	Overall R-index = 0.993						
Cluster:	1	2	3	4	5	6	7
R-index:	0.92	0.12	1.00	1.00	0.99	0.98	0.91
Perturbation by random subspace (Smolkin and Gosh)							
4 clusters							
Cluster:	1		2		3		4
stability:	1.00		0.28		0.83		0.34
Perturbation by random projections (PMO)							
4 clusters	Overall S-index = 0.978						
Cluster:	1		2		3		4
s-index:	1.00		1.00		0.91		1.00