A Robust Clustering Method and Visualization Tool Based on Data Depth.

Rebecka Jörnsten, Yehuda Vardi, Cun-Hui Zhang.

Abstract. We present a robust clustering method based on a modified Weiszfeld algorithm for the multivariate median, and associated data depth. The multivariate medians are used to represent the clusters, while the induced *relative* L_1*-depths* are used to identify outliers and to select the number of clusters. We develop a cluster validation and visualization tool based on the within-cluster data depths, and the cluster data depths with respect to competing clusters. We apply our method to high-dimensional gene expression data, and several simulated data sets. Our method successfully identifies the number of clusters in noisy data sets, and generates accurate cluster assignments.

1. Introduction.

In this paper we present a robust clustering method, and a novel cluster validation tool based on data depth. We consider a clustering method robust if it is not be affected by small perturbations of the data, i.e. noisy observations, or by the inclusion of unrelated variables, i.e. clustering in a higher dimension than necessary. The K-means and hierarchical clustering methods, often used in practice, are not robust by the above definition. Hierarchical clustering is affected by perturbations since it is agglomerative, and K-means is notoriously sensitive to outliers. Another popular algorithm is PAM (partition around "mediods"), which approximates a K-median clustering and is thus less sensitive to outliers than K-means. The PAM algorithm restricts the medians to belong to the set of observations, so that pairwise distances are sufficient input. However, this restriction is hazardous with noisy, high-dimensional data. For cluster validation and selection of the number of clusters in a data set, many internal and external indices have been proposed. Most internal indices are based on the within and between cluster distances. An example of such an index is the silhouette width commonly used in conjunction with PAM ([7]). However, distance based indices are sensitive to the inclusion of unrelated variables. Moreover, these indices can be dominated by high variance clusters which can lead to under-fitting (section 4). Some internal indices, such as the gap-statistic, are calibrated against a null (single cluster) ([12]). The calibrating requires simulations under the null, which is often assumed uniform or unimodal. Recent studies indicate that using the gap-statistic often leads to

Key words and phrases. Clustering, Robust, Data depth.
This research was supported in part by grants from the NSF, NSA, and DARPA.

over-fitting ([2]). External indices require test data sets on which to validate the clustering. In practice such test sets are usually not available, and cross-validation (sub-sampling) is used instead. This in turn leads to problems of a different nature, such as fewer observations on which to cluster ([2], [13]).

There is a renewed interest in clustering in the statistical community, a reason being the abundance of high-dimensional gene expression data. Gene expression data are comprised of measured gene expression levels, under various experimental conditions. The "activities" of the genes are measured simultaneously for a large collection of genes, where the expression levels of a gene across conditions is called the *gene profile*. Conditions, hereafter referred to as *samples*, may correspond to for example a tissue type, or a type of cancer. In such experiments, class discovery (sample clustering) is an important task. Though pathologists may classify cancers in a certain way, genetic markers can identify subclasses or a novel division of types of cancer. It is believed that such subclasses may inform us about the progression of the disease, and be used to select an appropriate treatment regime. Sample clustering can thus provide a deeper understanding of diseases such as cancer. In gene expression experiments, the number of samples are on the order of hundreds, whereas the number genes studied are often on the order of tens of thousands. We usually want to follow up on class discovery with building a sample class predictive model, using all or some selected gene profiles. Prior to building predictive models, some form of dimension reduction is often applied. One approach is to use principal components to create so-called "super-genes", i.e. weighted averages of gene profiles ([15]). The super-genes are then used as explanatories in the predictive models. Gene-clustering is another form of dimension reduction, with the cluster representatives used as explanatories in the predictive models. Selecting the number of gene-clusters in a data set is crucial since we may otherwise generate reduced dimension data sets with poor sample class predictive properties. When the sample classes are known, we can cluster genes in a supervised fashion ([4],[6],[11]). When the sample classes are unknown, or treated as such, the unsupervised clustering of genes may identify new sample classes. In addition, gene-clustering is an important task in itself. Genes with similar profiles across samples may comprise a genetic pathway for a disease. Hierarchical methods are frequently used for gene-clustering method, though PAM has recently gained much popularity ([2], [8]).

We introduce a K-median clustering algorithm which iterates between the modified Weiszfeld algorithm of Vardi and Zhang ([14]), and a Nearest-Neighbor allocation scheme with simulation annealing. We select the number of clusters in a data set using *relative* L_1*-depth* (ReD) induced by the L_1-median with respect to individual clusters. Each observation x_i has a within-cluster data depth D_i^w with respect to the cluster it belongs to, and between-cluster data depths $\widetilde{D}_i^b$ to the competing clusters. We select the number of clusters K in a data set that maximizes the average of $\text{ReD}_i = D_i^w - \widetilde{D}_i^b$. ReD is an internal cluster validation index that is *independent of the scales* of individual clusters, and is thus not dominated by high variance clusters. ReD focuses on centrality and separation of observations rather

than spread, and is therefore a useful tool for outlier identification. We demonstrate on real and simulated examples that ReD is superior to the scale-dependent silhouette width as a cluster selection criterion. Our K-median clustering and ReD selection are robust with respect to the noise level, and the inclusion of unrelated variables. It generates more accurate cluster assignments than PAM, and provides a useful visualization tool for high-dimensional data.
The paper is organized as follows. In section 2 we present our K-median algorithm and ReD. In section 3 we apply our methods to the acute Leukemia data set of Golub et al ([3]). We present results from a simulation study in section 4. Section 5 concludes the paper with a summary and ideas for future work.

2. Robust Clustering and Cluster Validation by Data Depth.

We get a K-median clustering by iterating between a Nearest-Neighbor allocation scheme (Euclidean distance) and median computations via the modified Weiszfeld algorithm. The Nearest-Neighbor allocation is self-explanatory. Here we give a brief description of the modified Weiszfeld algorithm. Let us denote by $x_i, i = 1, \cdots, N$ the distinct observations in R^p which we wish to cluster. With each observation x_i we associate a multiplicity η_i. If the data set has no tie, $\eta_i = 1, \forall i$. A cluster assignment is represented by a partition $I(1), \ldots, I(K)$ of $\{1, \ldots, N\}$, where $I(k)$ is the set of labels of those x_i in cluster k. Given a cluster assignment, the centroid of the k-th cluster, denoted by $y_0(k)$, is the multivariate L_1-median defined by

$$y_0(k) = \arg\min C(y|k),\ C(y|k) = C\big(y\big|I(k)\big) = \textstyle\sum_{i\in I(k)} \eta_i \|x_i - y\|, \quad \forall k,$$

where $\|x - y\|$ is the Euclidean distance from y to x. Vardi and Zhang's modified Weiszfeld algorithm converges quickly and monotonically to $y_0(k)$ in individual clusters from any starting point y by iterating the following step;

$$y \leftarrow \max\big\{0, 1 - \eta(y|k)/r(y|k)\big\}\tilde{T}(y|k) + \min\big\{1, \eta(y|k)/r(y|k)\big\}y,$$

$$\text{where } \tilde{T}(y|k) = \Big\{\textstyle\sum_{i\in I(k), x_i\neq y} \eta_i x_i/\|x_i - y\|\Big\}\Big/\Big\{\sum_{i\in I(k), x_i\neq y} \eta_i/\|x_i - y\|\Big\}.$$

Here, $r(y|k) = \big\|\big\{\sum_{i\in I(k), x_i\neq y} \eta_i (x_i - y)/\|x_i - y\|\big\}\big\|$, $\eta(y|k) = \eta_i$ if $y = x_i$ and $i \in I(k)$, and 0 otherwise (convention: $0/0 = 0$).
Let us fix the number of clusters to K. We want to find the L_1-optimal allocation among the K clusters and the optimal cluster representatives. Equivalently, we minimize the sum of L_1-distances from the observations to their nearest cluster representatives, since

$$\min_{I(1),\cdots,I(K)} \sum_{k=1}^{K} \min_{y} C\big(y\big|I(k)\big) = \min_{y_0(1),\cdots,y_0(K)} \sum_{i=1}^{N} \eta_i \min_{1\le k\le K} \|x_i - y_0(k)\|.$$

Though the modified Weiszfeld algorithm and the Nearest-Neighbor allocation provide solutions to the inside minimization problems above, we cannot guarantee the convergence of the iteration between the two to the global K-median solution. To prevent convergence to a local minimum we use a standard simulated annealing approach. Thus, for given centroids $y_0(1), \ldots, y_0(K)$, an observation x_i may *not* be allocated to the cluster $I(k_i^*)$ with the centroid $y_0(k_i^*)$ to which it is nearest, where $k_i^* = \arg\min_k \|x_i - y_0(k)\|$. The probability of such a "mis-allocation" depends

on the cost-distances $\Delta_i(k,l) = \|x_i - y_0\| - \|x_i - y_0(k)\|$. For $\Delta_i(k_i^*, l) > 0$, the probability of mis-allocation is set at the level $P\{i \in I(\ell)\} = \exp(-\Delta_i(k,l)/T)$, where T is the current "temperature". We set the initial temperature to $T_0 > 0$ and update with $T \leftarrow \alpha T$ at every m-th iteration between the Nearest-Neighbor and Weiszfeld, where $\alpha \in (0,1)$ is the "cooling-rate". The best choice for (T_0, α) depends on the dimension of the problem p, the number of clusters K, and the number of observations N. If (T_0, α) is selected appropriately, the above scheme generates the global K-median. We iterate until convergence where T becomes numerical zero and the sum of L_1-distances stabilizes.

Vardi and Zhang ([14]) introduced the L_1-data depth induced from the multivariate L_1-median. ReD is based on a natural extension of the L_1-data depth with respect to individual clusters. Given a cluster assignment and $z \in R^p$, let

$$\bar{e}(z|k) = \textstyle\sum_{i \in I(k), x_i \neq z} \eta_i e_i(z) / \sum_{j \in I(k)} \eta_j,$$

where $e_i(z) = (x_i - z)/\|x_i - z\|$ is the unit vector from z to observation x_i. Thus $\bar{e}(z|k)$, also called the spatial rank function ([9]), is the average of the unit vectors from z to all observations in the k-th cluster. The data depth of point z with respect to the k-th cluster is defined as

$$D(z|k) = 1 - \max\big[0, \|\bar{e}(z|k)\| - f(z|k)\big],$$

where $f(z|k) = \eta(z)/(\sum_{i \in I(k)} \eta_i)$ with $\eta(z) = \sum_{i=1}^N \eta_i I\{z = x_i\}$. The statistical interpretation of the cluster data depth is that $1 - D(z|k)$ is the minimum additional weight needed at z in order to make *it* the multivariate L_1-median for the data set $\{z\} \cup \{x_i, i \in I(k)\}$ ([14]). Now, each x_i is associated with cluster data depths $D(x_i|1), \ldots, D(x_i|K)$. We propose to select the number of clusters K based on the cluster data depths $D(x_i|1), \ldots, D(x_i|K)$. We make the cluster data depths directly comparable by normalizing with respect to cluster size. This is necessary since the data depths are on average higher in small clusters (e.g. equal to 1 for a cluster of size 1). We normalize the cluster data depths with

$$D(z|k) \leftarrow D(z|k)\big[\textstyle\sum_{i \in I(k)} \eta_i\big] \big/ \big[\sum_{i \in I(k)} \eta_i D(x_i|k)\big], \quad \forall z \in R^p,$$

so that the average normalized cluster data depth is 1 for each cluster, Hereafter, all data depths are assumed to be normalized. As mentioned in the introduction, ReD is the difference between the within- and between-cluster data depths. We define the within-cluster data depth of observations $x_i, i \in I(k)$ as $D_i^w = D(x_i|k)$. For $i \in I(k)$, we order the remaining $K-1$ cluster data depths, $D(x_i|l)$, $l \neq k$, by

$$D\big(x_i|l_1(i)\big) \prec \cdots \prec D\big(x_i|l_{K-1}(i)\big)$$

according to $\|x_i - y_0(l_1)\| \leq \cdots \leq \|x_i - y_0(l_{K-1})\|$, and define the between-cluster data depth of observation $x_i \in I(k)$ as $D_i^b = D(x_i|l_1(i))$. An observation x_i is well-clustered if $D_i^w \gg D_i^b$, since D_i^b is the depth of x_i with respect to the *nearest* competing cluster. We introduce the ReD selection statistic

$$\text{ReD}(K) = \sum_{i=1}^N \eta_i \text{ReD}_i \Big/ \sum_{i=1}^N \eta_i, \quad \text{ReD}_i = D_i^w - D_i^b. \tag{2.1}$$

We select the number of clusters K that maximizes $\text{ReD}(K)$. The ReD selection statistic is similar to the average silhouette width, where the silhouette width of observation i is defined as follows. For $z \in R^p$, let $\bar{d}(z|k) = \sum_{i \in I(k), x_i \neq z} \eta_i \|z -$

$x_i\|/\sum_{i\in I(k),x_i\neq z}\eta_i$ be the average distance between z and observations not equal to z in the k-th cluster. For x_i, $i\in I(k)$, we compute the *silhouette width*

$$\mathrm{sil}_i=(b_i-a_i)/\max(a_i,b_i),\quad a_i=\bar{d}(x_i|k),\ b_i=\min_{l\neq k}\bar{d}(x_i|\ell).$$

The silhouette method selects the number of clusters K that maximizes the average silhouette width $\sum_i\eta_i\mathrm{sil}_i/\sum_i\eta_i$. The silhouette widths are *not* normalized with respect the size of the clusters, or the scales within individual clusters. Thus, if the data is drawn from a distribution with different within-cluster variances, the silhouettes tend to be larger for the tight clusters (e.g. Fig. 4), and small clusters (Fig. 2). This means that the silhouette width may fail to identify outliers in small clusters and may even lead to under-fitting (section 4). In contrast, since the cluster data depths are independent of cluster scales, selecting K using ReD will not be driven by high variance clusters.

Both the average silhouette width sil_i and ReD_i in (2.1) ignore competing clusters other than the nearest, and observations that fall between many clusters are not detected. However, such observations may be crucial for selecting K. We therefore introduce the generalized relative L_1-depths and their average

$$\mathrm{ReD}^{[m]}(K)=\sum_{i=1}^{N}\eta_i\mathrm{ReD}_i^{[m]}\Big/\sum_{i=1}^{N}\eta_i,\quad \mathrm{ReD}_i^{[m]}=D_i^w-D_i^{[m]},\tag{2.2}$$

based on the first m tiers of cluster data depths of competing clusters. Again, we select the number of clusters K by maximizing $\mathrm{ReD}^{[m]}(K)$. It remains to define the generalized between-cluster depth $D_i^{[m]}$. Let $D^{(m)}(i)=D(x_i|l_m(i))$ be the depth of x_i with respect to the m-th nearest competing cluster according to the ordering of $D(x_i|l)$ given earlier. The m-th tier depths are $D^{(m)}(i)$, $i\leq N$. For $m=1$, we define $D_i^{[1]}=D_i^b$, so that (2.2) provide the same ReDs as (2.1). Now, we consider $m=2$, with the basic idea to replace small tier one between-cluster depths $D_i^{[1]}$ with large tier two depths $D_i^{(2)}$. Let these depths be ordered by $D^{(2)}(i_1^{(2)})\geq\cdots\geq D^{(2)}(i_N^{(2)})$ and $D^{[1]}(i_1^{[1]})\leq\cdots\leq D^{[1]}(i_N^{[1]})$, where $D^{[1]}(i)=D_i^{[1]}$. Define j^* as the largest j satisfying $D^{(2)}(i_j^{(2)})>D^{[1]}(i_j^{[1]})$. If the sets of labels $\{i_1^{(2)},\ldots,i_{j^*}^{(2)}\}$ and $\{i_1^{[1]},\ldots,i_{j^*}^{[1]}\}$ are disjoint, we simply define

$$D^{[2]}_{i_j^{(2)}}=D^{[1]}_{i_j^{(2)}}+D^{(2)}(i_j^{(2)}),\quad D^{[2]}_{i_j^{[1]}}=0,\quad j=1,\ldots,j^*,$$

and $D_i^{[2]}=D_i^{[1]}$ otherwise. The rational is to penalize those x_i with large $D_i^{(2)}$ and ignore those with small $D_i^{[1]}$. If the sets of labels overlap, the above direct approach may replace $D_i^{(1)}$ by $D_i^{(2)}$ at certain data points x_i. This is unreasonable since for any observation x_i the depths with respect to nearer clusters are always more important. To avoid such pitiful situations, we introduce an algorithm which iterates the following steps with the initialization $\Omega=\{1,\ldots,N\}$:

$$\begin{aligned}&i_1\leftarrow\arg\max_{i\in\Omega}D_i^{(2)};\quad \Omega\leftarrow\Omega\setminus\{i_1\};\quad i_2\leftarrow\arg\min_{i\in\Omega}D_i^{(1)}>0;\\ &\text{if}\quad D_{i_1}^{(2)}\leq D_{i_2}^{(1)},\qquad \Omega\leftarrow\emptyset;\\ &\text{else}\qquad \Omega\leftarrow\Omega\setminus\{i_2\},\ D_{i_1}^{[1]}\leftarrow D_{i_1}^{(1)}+D_{i_1}^{(2)},\ D_{i_2}^{[1]}\leftarrow 0.\end{aligned}$$

Here $D_i^{(2)}=D^{(2)}(i)$ and we set $D_i^{[2]}\leftarrow D_i^{[1]}$ at the end of the iteration when

$\Omega = \emptyset$. What's happening is that in each iteration we add the largest remaining $D^{(2)}_{i_1}$ and remove the smallest remaining $D^{(1)}_{i_2}$, $i_2 \neq i_1$, as long as $D^{(2)}_{i_1} > D^{(1)}_{i_2}$, and we update Ω by removing labels as soon as used. The (combined) between-cluster depth $D^b_i = D^{[2]}_i$ recognizes if observation x_i falls between two competing clusters. The description of ReD for general m is similar and omitted. In practice we are rarely able to add any third or higher tier depths before the index set $\{i_2\}$ is empty, such that $\text{ReD}_i = \text{ReD}^{[2]}_i$ and $D^b_i = D^{[2]}_i$ are used.

The within- and between-cluster data depths can be used for cluster validation by visual inspection. An example with K=3 clusters is shown in Figure 1. The top panel displays sorted within-cluster data depths for each cluster, with corresponding between-cluster data depths in the lower panel. Just below the x-axis are the first tier data depths, and stacked below are the second tier depths, colored by cluster. A lot of information is contained in these plots. A well-defined cluster has a smoothly decaying within-cluster data depth profile. A drop in within-cluster data depth indicates an elongated cluster with poorly defined center, or the presence of outliers. Color patterns in the between-cluster data depths gives information about cluster boundaries. Observations that are "deep" with respect to several competing clusters, and have low within-cluster data depths are suspect.

3. Application to Gene Expression Data.

We apply the K-median clustering and ReD selection to the gene expression data set presented in Golub et al ([3]). The data set comes from a study of gene expressions in two types of leukemia: acute lymphoblastic leukemia (ALL), and acute myeloid leukemia (AML). The study was made on 25 cases of AML, and 47 cases of ALL, i.e. $n = 72$ samples total. The ALL cases were comprised by 38 B-cell type (ALL-B) and 9 T-cell type (ALL-T) samples. Gene expressions were measured for $p = 6,817$ genes simultaneously. We reduce the number of genes to $3,571$ by excluding genes with low signal-to-noise ratio (max/min<5 or max-min<500, over 72 samples), and standardize such that each sample is centered at mean 0 with variance 1. Standardization prevent single experiments from dominating the analysis. Most genes are not "active" and we select only the 100 most variable genes (across samples) for further analysis. The reduced data set thus consists of $K = 3$ classes of samples, in $p = 100$ dimensions.

A single application of PAM with $K = 3$ produces the silhouette width plot in Figure 2 (b). The ALL-B sample 17 is mis-allocated to the ALL-T cluster. Still, sil_{17} is relatively high (last observation in Fig. 2 (b)). This is because sample 17 has been mis-allocated to the small ALL-T cluster, and the silhouette widths are not normalized. We also cluster the leukemia data set with PAM for other values of K, and find that the average silhouette width is maximized for $K = 3$. We generate a K-median clustering with $K = 3$, and display the data depths and ReDs in Figures 1 and 2 (a). Figure 1 shows the within and between-cluster data depths in the same plot, where the dashed lines are a mirror plot of the within-cluster data depths. We can conclude from this plot that there are several observations that fall

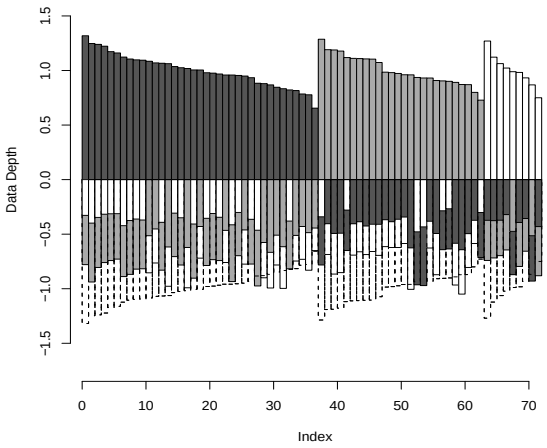
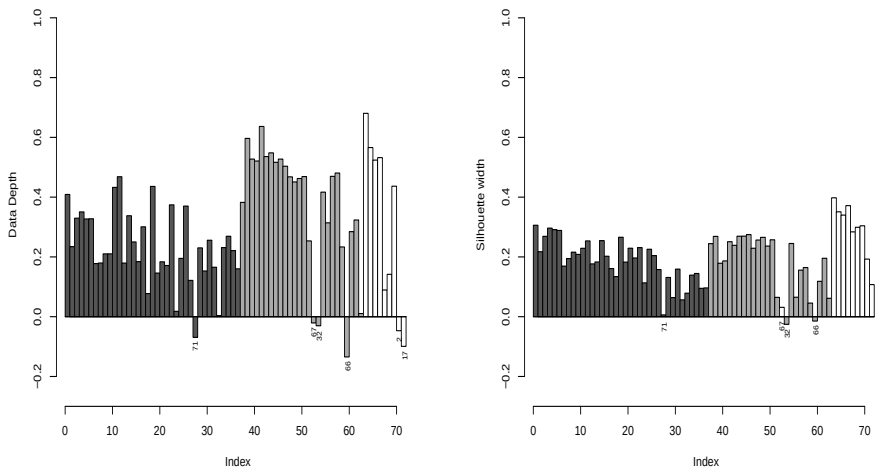


FIGURE 1. Within (top) and Between (bottom) Cluster Data Depths for the acute leukemia data.

between the clusters. These observations have stacked between data depths that exceed the within-cluster data depths (cross the dashed line). Two observations are mis-allocated with the K-median, sample 17 and sample 67. Sample 67 is an ALL-T sample that is incorrectly allocated with the AML samples. ReDs for 17 and 67 are the among the smallest for the 72 samples (Fig. 2 (a)). In [1] a comparison study of discriminant and class-predictive methods was conducted on the leukemia data set. The samples with low ReDs are the same samples that were difficult to classify, where, for example, the cross-validated error rate of sample 67 was found in [1] to be 76 %. Low silhouette widths are sometimes also indicative of a low class-predictive strength, but the overlap is erratic (e.g. sample 17). The PAM sample clustering changes if we re-standardize after selecting the 100 most variable genes. There are now 3 errors (samples 17, 66 and 67). The K-median clustering is not affected by re-standardization.

We now turn to gene clustering. Subclasses of samples, such as AML and ALL, cluster the genes differently. This means that different sets of genes are "active" under different conditions. On the leukemia data set, two clusters of genes were selected within each sample class with both the silhouette width and ReD. Both the K-median and PAM find clusters of up- and down-regulated genes. However, the gene clusters are *not* identical, e.g. some genes are up-regulated within the AML sample class, but not in other sample classes, and it is therefore unlikely that a single set of up- and down-regulated genes can distinguish between all three sample classes (AML, ALL-B, ALL-T). We find that the K-median clusters are less heterogeneous than the clusters generated by PAM, but omit these results

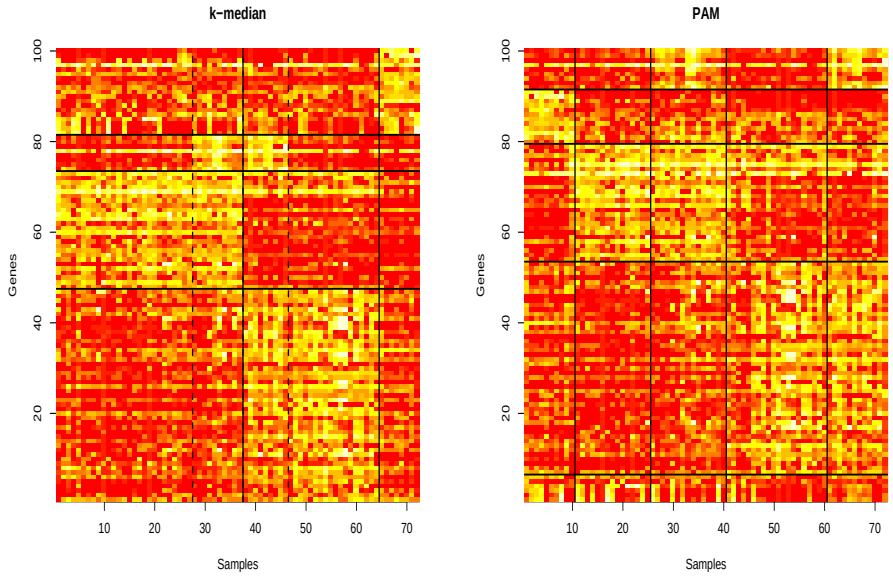


(a) (b)

FIGURE 2. Acute Leukemia data. (a) K-median and ReDs. Errors: 17, 67. Small ReDs correspond to observations that are difficult to classify ([1]). (b) PAM and silhouette widths. Error: 17. Low silhouette widths may mark observations that are difficult to classify, but many "difficult" observations are missed (e.g. 17).

here in order to conserve space.

We can also generate *global* gene-clusters using all the samples. It has been noted that global gene-clusters often have poor sample class-predictive properties, in either a supervised or unsupervised setting ([8]). This is a motivation for separate gene-clustering within identified classes. However, a two-way clustering is often more illustrative. One reason why globally generated gene-clusters may lack sample-predictive power is that the subsets of genes activated by the different conditions are difficult to identify with a global approach. Another reason may be that PAM is not sufficiently robust. If the identified gene clusters are heterogeneous, the corresponding mediods are poor cluster representatives. To investigate this, we generate global gene clusters using PAM and the K-median, and compare their sample clustering properties. The silhouette width selects 5 gene clusters, whereas ReD selects 4. The identified gene clusters are shown in Fig. 3. We then use the 5 PAM mediods to cluster the samples, and similarly the 4 multivariate medians. We thus cluster the 72 samples in 5 and 4 dimensions respectively. Using the 5 PAM mediods we select 5 sample clusters using PAM and the silhouette width. These 5 clusters are not a subdivision of the AML or ALL classes, and in Figure 3 (b) we see that the PAM clusters blocks are far from homogeneous. Generating more gene



(a) (b)

FIGURE 3. Global gene clusters. (a) K-median and ReD. The 4 gene clusters generate a subdivision of the AML and ALL-B classes (dashed lines). (b) PAM and silhouette width. The blocks are less homogeneous than the K-median blocks, and the sample clusters do not correspond to subdivision of the known classes.

clusters with PAM does not help. We need more than 60 clusters before PAM can identify the AML, ALL-B and ALL-T classes, and no recognizable subdivisions are seen when more than 3 sample clusters are generated. The reason behind the poor performance of PAM is that the clustering is driven by a few extreme observations and variables. The results are quite different when we use the 4 gene-clusters generated by the K-median. ReD selects 5 sample clusters using the 4 dimensional feature vectors. In Figure 3 (a) the gene and sample clusters are shown. The blocks are much more homogeneous than the PAM clusters. The 5 sample clusters correspond to a subdivision of the AML and ALL-B clusters from the three K-median clusters based on all genes, and are indicated by dashed vertical lines in the plot. If we use the unclustered genes to generate a 5 sample clustering, we get a similar subdivision, but with 4 more samples "mis-allocated". We can thus think of K-median gene-clustering as *denoising*, allowing us to discover structure in the data that is otherwise occluded. When we use the 4 K-median gene clusters to generate 3 sample clusters we do not get the AML, ALL-B, ALL-T classes (nor does PAM with the 5 gene-mediods). The known sample classes in fact corresponds to a *local*

minimum of the sum of L_1-distances. This local minimum has a higher average ReD than the global minimum, which suggests that a ReD *penalty* may be used to generate tighter and better clusters (section 5).

To make the comparison complete we also use 4 and 5 PAM gene-clusters in conjunction with a K-median sample clustering, and vice versa. We find that the PAM gene clusters are too noisy to be useful for sample clustering. Using PAM to cluster the samples with the K-median gene-clusters does not generate a subdivision of the known sample classes. We conclude that using the PAM approximation to the K-median clustering is inadvisable on this noisy data set.

4. A Simulation Study.

We investigate the robustness of the K-median clustering and ReD selection via a simulation study. We first compare the noise level susceptibility of PAM and the K-median in terms of clustering accuracy on the leukemia data set. We also study how the silhouette width and ReD selection schemes are affected by the noise level.

SNR	$\hat{K}$		Errors	
	Data depth	Silhouette	K-median	PAM
10	3.00(50/50)	3.00(50/50)	2.16	2.68
4	3.02(49/50)	2.94(47/50)	2.68	3.56
2	3.18(40/50)	2.60(20/50)	3.36	5.92
1.25	3.44(20/50)	2.28(14/50)	4.56	9.16

TABLE 1. Acute leukemia data + noise. ReD is more robust than the silhouette width, and K-median more accurate than PAM.

We simulate 50 data sets with i.i.d. normal noise added to the leukemia data. The signal to noise ratio (SNR) is defined as the ratio of the variance of the gene expression data to the noise variance. In Table 1 we present results for SNR 10, 4, 2 and 1.25. The two right columns of Table 1 show the average number of errors generated by PAM and the K-median with $K = 3$. In the noiseless case PAM results in 1 error, and the K-median in 2 errors. At SNR 10 and 4, the K-median generates an identical clustering as in the noiseless case most of the time (average number of errors 2.16 and 2.68). In a few of the simulated data sets sample 66 was also misallocated. Sample 66 was recognized in [1] as another observation that is difficult classify. At SNR 10 and 4, PAM results in on average 2.68 and 3.56 errors. Two of the errors are always sample 17 and 67, whereas the third and fourth errors vary between simulation. At SNR 2 and 1.25 the K-median generates on average 3.36 and 4.56 errors, and PAM on average 5.92 and 9.16 errors. Comparisons at lower SNR are not meaningful since this corresponds to a noise level equal to the signal level. We now turn to the problem of selecting the number of clusters. In the two left columns of Table 1 we show the average number of selected clusters $\hat{K}$ using ReD and the silhouette width, and the proportion of times the correct number of

clusters are selected. At SNR 10 and 4, ReD selects the correct number of clusters for almost all simulated data sets (50/50 and 49/50). The silhouette width selects the correct number of clusters at SNR 10, but under-fits the data 3/50 times at SNR 4. At SNR 2 and 1.25 ReD over-fits (average $\hat{K}$ 3.18 and 3.44), whereas the silhouette width under-fits (average $\hat{K}$ 2.60 and 2.28). However, ReD selects the true number of clusters more frequently than the silhouette width does.

We also conduct 4 simulation studies with multivariate normal data. We generate clustered data with $K = 3$ clusters, in 3 or 8 dimensions. The four simulation set-ups are as follows. The cluster sizes are chosen as 25, 50 and 25. Model 1 is a data set with linearly dependent mean vectors, $\mu_1 = (0,0,0)$, $\mu_2 = (0,5,-5), \mu_3 = (0,-5,5)$. The clusters are independent and have the same diagonal covariance matrix, Diag$(.25,1,1)$. In Model 2 we use orthogonal mean vectors, $\mu_1 = (5,0,0)$, $\mu_2 = (0,5,-5), \mu_3 = (0,5,5)$. The clusters are independent, with diagonal covariance matrix Diag$(3,3,3)$. Model 3 has the same mean structure as Model 2, but also include 5 unrelated variables with means $\mu_j = (0,0,0)$, $j = 4,\cdot,8$. The clusters are independent with diagonal covariance matrix, and variance 1.5 in all 8 dimensions. Model 4 investigates a situation with differing within-cluster variances. We use the following mean structure, $\mu_1 = (5,0,0)$, $\mu_2 = (0,8,8), \mu_3 = (0,5,-5)$. The clusters are independent. The first cluster has diagonal covariance matrix with variances $(4,4,4)$, i.e. a highly variable cluster. The second cluster is negatively correlated in the second and third dimension (correlation -0.5), with variances $(9,9,9)$. The second cluster is thus elongated. The covariance structure of the third cluster is diagonal with variances $(1,1,1)$, i.e. a tight cluster. Cluster 2 is clearly separated from the other clusters, and clusters 1 and 3 are close.

We simulate 50 data sets from the models, and show the results in Table 2. Model 1 is a pathological example with linearly dependent mean vectors. ReD fails in this scenario, even though the noise level is low. The reason for this is that the data depth is directional, and does not recognize cluster separability on a line. The silhouette width selects the true model 50/50 times, whereas ReD under-fits the data selecting two clusters the majority of the times (30/50). Model 2 generates noisy data. ReD selects the true model 47/50 times. The silhouette width is more noise sensitive, as we saw in the simulation using the leukemia data, and selects the true model only 37/50 times. If we use $K = 3$ PAM results in 3.77 errors on average, and the K-median 3.08 errors on average. In Model 3 we drop the noise level, but include 5 unrelated variables. Using the K-median and ReD we select the true model 49/50 times, and get 1.33 errors on average. The silhouette width is more sensitive to the inclusion of unrelated variables, selecting the true model 43/50 times, and PAM gives 1.74 errors on average. With Model 4 we generate clusters with different within-cluster variances. Cluster 1 is highly variable, whereas cluster 3 is tight, and cluster 2 is a clearly separated highly variable cluster. The silhouette width is dominated by the highly variable clusters and under-fits by joining clusters 1 and 3. The true model is only selected 10/50 times. 5/50 times

a 3 cluster model is selected, but this model corresponds to a local L_1 distance minimum where cluster 2 is split and clusters 1 and 3 joined. This generates 47-50 "errors", inflating the average number of errors using PAM to 3.37. If these cases are discarded the average error drops to 1.39. ReD selects the true model 38/50 times, and the K-median gives 1.33 errors on average. In Figure 4 (b) the silhouette width for a realization under Model 4 is shown. Cluster 3 is the tight cluster and exhibits overall large silhouette widths. Cluster 1 is the high variance cluster close to cluster 3. The silhouette widths are low for this cluster since its nearest cluster, to which it is being compared, is the tight cluster. ReD does not depend on scale, and is not affected by the difference in within-cluster variance between clusters 1 and 3 (Fig. 4 (a)). Both the silhouette width and ReD selected 3 cluster for this data set, and resulted in the same single error. An observation from cluster 1 was incorrectly allocated to the tight cluster 3. The data depth clearly identifies this error with a negative ReD. The silhouette width does not recognize this error since it falls in the tight cluster. A negative silhouette width is instead seen for an observation that was correctly allocated to cluster 1 (Fig. 4 (b)).

We conclude from this set of simulations that ReD is more robust than the silhouette width for selection when the data is noisy, unrelated variables are included, and when the within-cluster variances differ. However, ReD fails in the pathological scenario when the clusters means are linearly dependent.

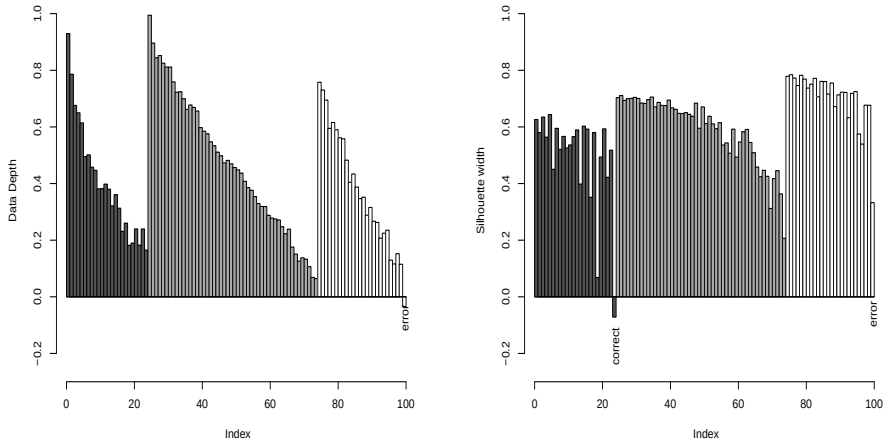
Model	$\hat{K}$		Errors	
	Data depth	Silhouette	K-median	PAM
1	2.40(20/50)	3.00(50/50)	0.12	0.04
2	2.98(47/50)	2.74(37/50)	3.08	3.77
3	3.02(49/50)	2.86(43/50)	1.30	1.74
4	3.00(38/50)	2.36(15/50)	1.33	3.37

TABLE 2. Simulation Study: ReD fails when cluster means are linearly dependent (1). ReD outperforms silhouette width when data is noisy (2), and when unrelated variables are included (3). The silhouette width fails when within-cluster variances differ (4).

5. Conclusions.

We have demonstrated that our K-median algorithm is more robust than the popular approximation PAM, and that using PAM may not suffice in some real data analysis situations. We found that ReD is a robust selection statistic for the number of clusters in a data set. It is not sensitive with respect to the noise level, differing within-cluster variances, and the inclusion of unrelated variables. We demonstrated with a pathological example that ReD cannot select the number of clusters in a data set if the cluster means are linearly dependent.

We only compared PAM to the K-median clustering, and the silhouette width to ReD selection in this paper. We found that our methods improve on PAM and



(a) (b)

FIGURE 4. The silhouette width (b) is dominated by the high variance cluster 1, and fails to detect the mis-allocated sample, and a correctly allocated sample has a large negative silhouette width. The mis-allocated sample is clearly identified by ReD (a).

the silhouette width, and we intend to follow up with a more extensive study, comparing other internal indices and validation schemes. We believe that the K-median clustering and associated ReD are useful tools for the clustering of genes and samples in gene expression data. We limited the present study to the acute leukemia data set of Golub et al, and future work will involve analysis of several diverse gene expression data sets. We found that local minima of the L_1-distance cost function may correspond to sensible gene and sample clusterings. These local minima had much higher ReD than the global minimum. We will investigate whether including a ReD penalty can improve on the K-median clustering scheme, perhaps generating gene-clusters that have better sample class predictive properties.

References

[1] S. Dudoit, J. Fridlyand, T. Speed. *Comparision of discrimination methods for the classification of tumors using gene expression data.* Journal of the American Statistical Association, **97** (2002), 77-87.

[2] S. Dudoit, J. Fridlyand. *Application of resampling methods to estimate the number of clusters and to improve the accuracy of a clustering method.* Technical report 600 (2001), Department of Statistics, UC Berkeley.

[3] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, and E. S.

Lander. *Molecular classification of cancer: class discovery and class prediction by gene expression monitoring.* Science, **286** (1999), 531-537

[4] T. Hastie, R. Tibshirani, M. B. Eisen, A. Alizadeh, R. Levy, L. Straudt, W. C. Chang, D. Botstein, P. Brown. *Gene shaving as a method for identifying distinct sets of genes with similar expression patterns.* Genome Biology, **1(2)** (2000), 1-21

[5] T. Hastie, R. Tibshirani, D. Botstein, P. Brown. *Supervised Harvesting of Expression Trees.* Technical report (2000), Department of Statistics, Stanford University.

[6] R. Jornsten *Data compression and its statistical implications: with an application to the analysis of microarray images.*, PhD Thesis (2001), Department of Statistics, UC Berkeley.

[7] L. Kaufman, and P. J. Rousseeuw. *Finding Groups in Data: An introduction to cluster analysis.* (1990) Wiley, New York.

[8] K. Pollard, M van der Laan. *Statistical inference for simultaneous clustering of gene expression data.* Technical report (2001), Department of Biostatistics, UC Berkeley.

[9] J. Möttönen, and H. Oja. J. Nonparametric Statistics, **5** (1995), 201-203.

[10] A. Owen, and L. Lazzeroni. *The plaid model.* Technical report (2000), Department of Statistics, Stanford University.

[11] D. Rocke, D. Nguyen. *Tumor classification by partial least squares using microarray gene expression data.* Bioinformatics, **18(1)** (2002), 39-50.

[12] R. Tibshirani, G. Walther, and T. Hastie. *Estimating the number of clusters in a dataset via the gap statistic.* Technical report (2000), Stanford University, Department of Biostatistics.

[13] R. Tibshirani, G. Walther, D. Botstein, and P. Brown. *Cluster validation by prediction strength* Technical report (2001), Stanford University, Department of Biostatistics.

[14] Y. Vardi, and C-H. Zhang. *The multivariate L_1-median and associated data depth.* Proceedings of the National Academy of Sciences, **97** (2000), 1423-1426.

[15] M. West, J. R. Nevins, J. R. Marks, R. Spang, C. Blanchette, H. Zuzan. *DNA microarray data analysis and regression modeling for genetic expression profiling.* Preprint (2001), Department of Statistics (Duke Univ).

Department of Statistics Rutgers University, PISCATAWAY, NJ, 08854, USA
E-mail address: `rebecka,cunhui,vardi@stat.rutgers.edu`