Published in final edited form as:
Comput Stat Data Anal. 2009 October 1; 53(12): 4290–4300.

A permutation test for determining significance of clusters with applications to spatial and gene expression data

P.J. Park[a], **J. Manjourides**[b], **M. Bonetti**[c], and **M. Pagano**[b,*]

P.J. Park: peterpark@harvard.edu; J. Manjourides: jmanjour@hsph.harvard.edu; M. Bonetti: marco.bonetti@unibocconi.it

[a] Harvard Medical School, Boston, Massachusetts, USA

[b] Department of Biostatistics, Harvard School of Public Health Boston, Massachusetts, USA

[c] Department of Decision Sciences, Bocconi University, Milan, Italy

Abstract

Hierarchical clustering is a common procedure for identifying structure in a data set, and this is frequently used for organizing genomic data. Although more advanced clustering algorithms are available, the simplicity and visual appeal of hierarchical clustering has made it ubiquitous in gene expression data analysis. Hence, even minor improvements in this framework would have significant impact. There is currently no simple and systematic way of assessing and displaying the significance of various clusters in a resulting dendrogram without making certain distributional assumptions or ignoring gene-specific variances. In this work, we introduce a permutation test based on comparing the within-cluster structure of the observed data with those of sample datasets obtained by permuting the cluster membership. We carry out this test at each node of the dendrogram using a statistic derived from the singular value decomposition of variance matrices. The p-values thus obtained provide insight into the significance of each cluster division. Given these values, one can also modify the dendrogram by combining non-significant branches. By adjusting the cut-off level of significance for branches, one can produce dendrograms with a desired level of detail for ease of interpretation. We demonstrate the usefulness of this approach by applying it to illustrative data sets.

Keywords

cluster analysis; microarray data; dendrogram; within-cluster variance; spatial data

1. Introduction

Clustering and classification plays an important role in many branches of science [1,2]. Assuming that we can define distances between objects, clustering is done on the basis of some dissimilarity measure. Specifically, we have a collection of n objects in p dimensional space, and we have a distance measure defined on this space. We are interested in identifying the cluster structure, if any, inherent in the data. A popular approach in cluster analysis is hierarchical clustering. In this agglomerative method, one starts with n objects, each as its own cluster, and combines the “closest” clusters into a larger one. This is done successively until a single cluster emerges. Euclidean distance is perhaps the most commonly used distance, but the choice of distance measure depends on the type of data and the particular features in which one is interested. Pearson’s correlation coefficient is often

*Corresponding author: pagano@hsph.harvard.edu (M. Pagano).

used, for example in genetics and time series, although it is not a “distance measure” in the strict sense.

While a hierarchical clustering algorithm can be implemented easily, interpretation of the resulting dendrograms is more difficult. It is well known that they suffer from many undesirable properties, such as non-uniqueness and inversion [3], as well as sensitivity to outliers and small perturbations in the data [4]. Also, the ordering of the leaves of a dendrogram contains some arbitrariness and can be misleading, although some efforts have been made towards a better arrangement algorithm [5]. Thus, the interpretation of a given dendrogram requires caution. Most importantly, by the nature of the algorithm, hierarchical clustering will always yield clusters even when no clear cluster structure exists. It is therefore very helpful to assign significance to the apparent clustering in order to determine whether the dendrogram reliably describes the true structure of the data. While there are methods for determining when to cut branches of the dendrogram, these rules generally rely on a prespecified branch height, with all connected branches below considered to be separate clusters. Dynamic tree cutting [6] has been developed as a more flexible method for cluster identification, but there is no measure of significance associated with the resultant dendrogram.

Despite its many weaknesses, hierarchical clustering is pervasive in genomic data analysis, especially for gene expression data from microarrays. Its intuitive methodology, ease of implementation and simple, attractive visualization in two dimensions have made it popular with biologists and clinicians. A large number of more complicated and statistically sound algorithms from multivariate statistics and machine learning have been developed and compared [7]. One reasonable approach has been based on the idea of consensus clustering [8,9], where multiple clustering runs are made with perturbed data and these are averaged to give the final result. Various methods have been used to generate perturbed data and to summarize results, including parametric bootstrap [10] and adding Gaussian noise [11,8]. Machine learning algorithms for clustering include those based on genetic algorithms [12] (see [13] for an example), neural networks, and Bayesian model selection [14] (see [15] for an example). Gene clustering methods have also been extended to situations involving repeated measure data (see [16] for an example). Yet, the same standard method first used for expression data several years ago [17] continues to be used pervasively in the current literature.

Given the continued popularity of the simple hierarchical method, we introduce a method that will facilitate the interpretation of the result in that framework. We propose a permutation test that measures the significance of each division in a dendrogram. The method is based on the observation that if the clustering algorithm correctly finds the clusters, the within-cluster scatter will be relatively small. We compute a statistic that summarizes these variances and obtain a p-value based on this statistic. While the “tightness” of certain clusters can be sometimes discerned by visual inspection of the dendrogram, i.e., by comparing the lengths of different branches, this is not always reliable. With the permutation test, we are able to assign a level of significance for the division into two branches at each node in a systematic way, based on the original data. The method we introduce does not assume any parametric form for the distribution of the data, nor do we preselect the number of clusters in the data as is done, for example, in the Hierarchical Ordered Partitioning and Collapsing Hybrid (HOPACH) approach [18]. The method we propose is a simple way of annotating the dendrogram to indicate a degree of confidence in each node. We deal with a single tree from one clustering algorithm in this work, but it can also be applied to a consensus tree or to other situations.

2. Methods

After performing hierarchical clustering and obtaining a dendrogram, one has below each node a subset of the data divided into two clusters. We compute a statistic that measures the "goodness" of this apparent clustering. This procedure is similar, in spirit, to the multiresponse permutation procedure (MRPP) [19,20], which uses a permutation procedure on between-object distances to evaluate the hypothesis of nonrandom clustering. We propose that this statistic be based on the within-cluster variance as described below.

Measuring Clustering Quality

We use the pooled within-cluster variance matrix to measure the goodness of clustering [21]. This is similar to the idea behind hypothesis testing in Multivariate Analysis of Variance (MANOVA), which involves decomposing the total variance into between-cluster and within-cluster components and comparing their relative sizes [22]. We compute the between, within, and total variance matrices, respectively, for g groups with n_l elements in group l ($l = 1, \ldots, g$) as follows:

$$\mathbf{B} = \sum_{l=1}^{g} n_l (\bar{\mathbf{x}}_l - \bar{\mathbf{x}})(\bar{\mathbf{x}}_l - \bar{\mathbf{x}})',$$

$$\mathbf{W} = \sum_{l=1}^{g} \sum_{j=1}^{n_l} (\mathbf{x}_{lj} - \bar{\mathbf{x}}_l)(\mathbf{x}_{lj} - \bar{\mathbf{x}}_l)',$$

$$\mathbf{T} = \mathbf{B} + \mathbf{W} = \sum_{l=1}^{g} \sum_{j=1}^{n_l} (\mathbf{x}_{lj} - \bar{\mathbf{x}})(\mathbf{x}_{lj} - \bar{\mathbf{x}})'.$$

Here $\mathbf{x}_{lj}$ is the element j in group l, $\bar{\mathbf{x}}_l$ is the average of the elements in group l, and $\bar{\mathbf{x}}$ is the average over all the elements. We say that the difference between groups exists if the within-cluster variance is sufficiently small in some metric when compared to the between-cluster variance. There are several formal statistics to assess this equality of multivariate means, they include Wilks' lambda statistic, $\lambda = |\mathbf{W}|=|\mathbf{B} + \mathbf{W}| = 1=|\mathbf{B}\mathbf{W}^{-1} + \mathbf{I}|$; the Lawley-Hotelling Trace, $\mathrm{tr}(\mathbf{B}\mathbf{W}^{-1})$; the Pillai Trace, $\mathrm{tr}(\mathbf{B}(\mathbf{B}+\mathbf{W})^{-1})$; and Roy's largest root, the maximum eigenvalue of $\mathbf{B}\mathbf{W}^{-1}$. If one assumes normality, then as sample sizes increase, these tests become equivalent, and p-values can be obtained from the appropriate asymptotic distributions. Wilks' lambda is the most common statistic, as it is equivalent to the likelihood ratio test under the assumption of normality. Advantages and disadvantages of various statistics are described, for example, in [23].

The major disadvantage of using these asymptotic results, however, is that they require the assumption of normality. While this may be valid in some cases, there is usually no *a priori* reason to assume it to be true. In particular, it does not seem to apply to microarray data, which provides the original motivation for this work. Some have noted that the gene expression levels might be close to a log-normal distribution (marginally), but this assumption is difficult to justify in many cases [24]. Without normality, asymptotic distributions do not seem to hold readily, and thus testing hypotheses is not feasible. This is of course even more difficult for small samples, and we cannot standardize the statistic to obtain a p-value.

In general, **B** and **W** offer an attractive way of measuring goodness of clusters and they should be used together to compute a single measure. In the permutation test we propose below, however, we note that $\mathbf{T} = \mathbf{W} + \mathbf{B}$ always remains constant across all permutations. Therefore, it is sufficient to consider only how **W** varies for our statistic. A slightly different measure of cluster quality is the Mann-Whitney U-test [25,26] and it appears, in our experience, to give comparable results.

Using the singular values

We now discuss some choices regarding the use of **W**. In the univariate case, the group sum-of-squares is used as a natural criterion in the analysis of variance (ANOVA). The simplest generalization of this sum-of-squares to the multivariate case is the trace, $\mathrm{tr}(\mathbf{W})=\sum_{i=1}^{n} W_{ii}$, where the within-group sum-of-squares is computed over all the variables. In the context of hierarchical clustering, this idea is related to that of the Ward method [27], in which the pair of clusters that minimizes the sum-of-squares is combined among all possible pairs at each step of the algorithm.

While the trace criterion is often used, there are some problems associated with it. One problem is that it implicitly tries to produce spherical clusters in p dimensions. Thus, when this is used as a clustering measure, this can result in an incorrect set of clusters. For example, it will fail to identify ellipsoidal clusters that are very close to each other [1]. Another problem with the trace is that it is scale-dependent. When the raw data are standardized differently, different clusters will result by this criterion.

Some prefer the determinant, $|\mathbf{W}|$. However, it still makes the implicit assumption that all the clusters have the same shape, but this requirement is less stringent than that of the trace. The determinant criterion is scale-independent and is not affected by normalization of the data. For more discussion of the use of the determinant instead of the trace, see [23]. There are other advantages to using $|\mathbf{W}|$, such that highly correlated components are not given excessive weight [28].

While the determinant is preferred in general, a problem occurs when the number of components is greater than the number of samples ($p > n$). This is the case when clustering the samples in microarray data, which has a distinguishing characteristic that the number of genes typically far exceeds the number of samples. There are several ways of screening genes to assess their potential relevance and filter them to obtain a smaller set of genes, but in most case p is still much larger than n and, as a result, **W** is singular.

A solution in this case is to compute the quantity $\sum_{i=1}^{r}\sigma_i$, where r is the rank of **W** ($\mathrm{rank}(\mathbf{W}) = r < \min(n, p)$) and σ_i are the (non-zero) singular values of **W**. Recall that by singular value decomposition, $\mathbf{W}_{n\times p} = \mathbf{U}_{n\times n}\,\mathbf{D}_{n\times p}\,\mathbf{V}_{p\times p}$, where **U** and $\mathbf{V}^T$ are orthonormal ($\mathbf{U}\mathbf{U}^T = \mathbf{U}^T\mathbf{U} = \mathbf{I}_{n\times n}$ and $\mathbf{V}\mathbf{V}^T = \mathbf{V}^T\mathbf{V} = \mathbf{I}_{p\times p}$), and the entries of the diagonal matrix **D** are the singular values $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_r \geq 0 = \cdots = 0$.

So when the matrix is 1) singular because $p > n$; or 2) $p < n$ after filtering but still nearly singular due to highly co-expressed genes, we can use the product of the singular values in place of the determinant. In general, deciding on the right threshold for counting a singular value as zero is not simple; but our threshold based on the product of the largest singular value and the machine precision appears to work well. For computing the singular values, fast and stable algorithms are available. If the number of genes is very large or the computation is limited by the memory size, we can resort to the trace, which is not affected by the problem of a singular matrix. In our applications below, the trace and the product of singular values give similar results, with the trace generally giving slightly larger p-values.

Implementation of the the permutation approach in this setting is simple. Following Pitman [29], in spirit, at each node of the dendrogram we compare the value of the statistics to the values of the statistic we would have obtained if the group memberships had been randomly assigned. In each random sample, the same data points are used and the number of elements in each of the two groups remains the same, but different sets of elements are assigned to each group. We perform this membership permutation many times, and compute a p-value

by counting the proportion of times that the same or a more extreme value of the statistic is obtained in the random samples. For small n, we can do a complete enumeration; in general, we can only sample from the permutation distribution.

An illustrative example

Figure 1(a) shows a two-dimensional sample dataset with a clearly defined cluster structure. Two sets of ten data points, numbered {1{10} and {11{20}, were generated from two multivariate normal distributions with equal identity covariance matrices but with means 4 units apart in each dimension. In Figure 1(b), we performed a hierarchical clustering using Euclidean distance and average linkage. The p-values shown are from the permutation test based on $|\mathbf{W}|$, using 1000 samples at each node. One can see the presence of two distinct clusters, as evidenced by the low p-value of <0.001 at the top division. Following the left branch, the p-value is 0.010, which suggests that the left cluster that consists of {6,1,5,3,4} is distinct from the right cluster containing {9,8,10,2,7}. There is no clearly defined structure past that point if we are to use the typical $p = 0.05$ value. Note that without these numbers, it is difficult to determine which nodes may be more significant just from the dendrogram. For example, the node joining {12,18} and {16,20,14,11,13} is similar in appearance to the one joining {6,1,5,3,4} and {9,8,10,2,7}. But the first node has a larger p-value associated with it than the second one. These and other similar observations are consistent with the structure of the data shown in Figure 1(a). Note also that for low dimensional data, such as this example, one may resort to examining the original data. But, this option is generally not available, especially for high dimensional data.

In Figure 2(a), we list the p-values computed on the same data using the $\mathrm{tr}(\mathbf{W})$ criterion. The values are mostly similar to those obtained previously, except near the bottom of the dendrogram where the number of objects within each group is too small to give meaningful results. One reason for the dissimilarity is the different assumptions underlying each criterion. For example, because the trace tends to be smaller for spherical structures, the p-values may be inflated slightly if there are adjacent elongated structures correctly captured by the hierarchical clustering. In that case, some cluster-membership permutations may result in more spherical clusters with a smaller trace even though the clustering is incorrect.

After identifying significant nodes, one can proceed to combine the non-significant branches, as is done in [26]. Using a cut-off p-value of 0.05, we revise the dendrogram shown in Figure 1(b) and display it in Figure 2(b). There are only two nodes with p-values less than 0.05 and so there are only three clusters left. Looking at the data in Figure 1(a), we see that the new dendrogram is a good summary of the data. The cut-off p-value can be chosen to reflect the desired level of detail in the dendrogram. Since the repeated tests on subsets of the same data are not independent, we can use a “correction” as explained in Westfall and Young[30]. Since we are more intent on displaying the properties of the dendrogram trimming algorithm, we chose to use the simpler Bonferroni correction idea as a guide in selecting an initial cut-off value, e.g., $0.05/(n - 1)$, since there are $n - 1$ nodes.

3. Applications

Spatial data

In the analysis of spatial data, one is often interested in determining if there is any clustering of cases. This clustering can represent any number of interesting aspects, such as excess cases of disease [32], point sources of exposure [33] or the existence of health disparities within a community [34]. Consider the spatial data collected by Alt and Vach [31] who map a medieval burial site in Neresheim, Baden-Wurttemberg, Germany. The authors address an anthropologic question regarding burial patterns; specifically to determine if members of the same family tend to be buried in close proximity as a familial unit. To answer this question,

the authors examined skeletal remains of 152 grave sites for evidence of a known hereditary genetic feature: missing or reduced wisdom teeth. The locations of the 30 grave sites with dental feature (referred as cases) are presented in Figure 3(a). In this example, if the proposed hypothesis is correct, and familial units are buried together, then we would expect to see several clusters distributed throughout the grave site. In fact, each (extended) family might be considered its own cluster.

We perform hierarchical clustering on the spatial locations of the cases using Euclidean distances and average linkage. Figure 3(b) shows the resulting dendrogram, from which it is nearly impossible to determine which groupings are important. We calculate p-values for each node from the permutation test based on $|\mathbf{W}|$, using 1000 simulations at each node, which are also shown in Figure 3(b). These p-values suggest 4 distinct clusters, where the node significance is less than Bonferoni corrected p-value, $p = 0.001724$ among the cases. These clusters are composed of the the cases numbered {15,23,24,30}, {4,5,7,25,26,27,28,29}, {1,2,3,6,8,21}, and {9,10,11,12,13,14,16,17,18,19,20,22}.

Once we have calculated the p-values for the nodes, we once again, combine the non-significant branches to simplify the dendrogram. We use a Bonferoni p-value cutoff of $0.05/(n - 1) = 0.00172$, where n is the number of nodes, to determine significance, and we display the resulting dendrogram in Figure 4(a).

We then turn our attention to examining the locations of the non-affected graves, the controls. If the hypothesis of familial burials is correct, we would expect to see many small clusters, similar in size, to the clusters of the cases. We perform the same analysis on the controls that we applied to the cases and the resultant dendrogram is presented in Figure 4(b). As expected, we see evidence of many small clusters within the spatial distribution of the non-affected graves.

Application to microarray data

We apply our technique to three microarray datasets. The first two are for clustering patient samples and the third is for clustering genes. The advantage of the technique becomes more apparent as the size of the dataset grows larger, but we concentrate on clustering samples rather than genes for illustration due to space constraints. We first consider RNA data studied in [35], which was used to demonstrate how cigarette smoking status effects the the human airway epithelial cell transcriptomethe based on gene expression levels. Among the 72 patients, there are 34 patients identified as ‘current smokers’ and 23 identified as ‘never smokers’. The remaining 15 subjects are identified as ‘former smokers’, which we do not include in our classification, for the sake of clarity. Those patients numbered {1–34} are ‘current smokers’ and those numbered {35–57} are ‘never smokers’. From the more than 8000 genes, we select 848 genes, highly correlated to smoking status, for classification purposes. There are still more dimensions than objects, however, and we use the trace criterion for the permutation test. The result is plotted in Figure 5(a). After some normalizations to the data, the hierarchical clustering gives a good result, with clearly defined clustering between never smokers and current smokers.

In Figure 5(b), we use the cut-off p-value of $.05/(n - 1) = 0.000893$ (Bonferroni correction) to eliminate the nodes that are not significant. This pruning of the dendrogram provides a much simpler and more informative view of the data.

In Figure 6(a), we apply the same technique to a dataset on embryonal tumors of the central nervous system [36]. This dataset consists of several types of heterogeneous tumors about which little is known, including medulloblastomas, malignant gliomas, renal/extrarenal rhabdoid tumors, supratenotorial primitive neuroectodermal tumors, as well as normal

cerebella. A main conclusion of the paper is molecular profiles from those of other embryonal CNS tumors[36]. For normalization of the data, we have tried to follow the steps described in the supplemental information of the article, with the additional step of using variance filter to reduce the number of genes to 100. Different numbers of genes gave similar results.

There are more disease categories here than in the epithelial data, but the significance-adjusted dendrogram appears to capture the important part of the dendrogram very well. As was the case in [36], PNET samples are dispersed among different clusters; NCER (normal) samples are separated from the rest at the top branch; most other samples are clustered with those of the same type. The modified dendrogram presents a clear and concise description for easier interpretation.

For both the epithelial and the CNS tumor data, using the trace criterion appears to inflate the p-values slightly, causing more branching in the tree. In the second case, for example, it results in an additional subdivision of the MD branch. In both cases, the Bonferroni corrected p-value of 0.05 appears to give a clear result. For Figures 5(a) and 6, we used 10000 permutations at each node.

We can also apply the permutation test with genes as objects and samples (experiments) as variables, although the size of the dendrogram makes it harder to examine the details in that case. In Figure 7, we cluster the genes using the compendium data of *S. cerevisiae* experiments [17]. We eliminate those genes with too many missing values and carry out the missing value imputation using k nearest neighbors for the rest. Because the dimension of $\mathbf{W}$ is equal to the number of experiments, the computational cost for clustering genes is low; however, we selected only the 500 genes with the largest variations due to the space available for visualization. For clustering genes, the $\mathbf{W}$ matrix is not singular, unless the experiments are nearly identical. Figure 7(a) and 7(b) show the clustering result before and after the modification by the permutation test, again using 10000 samples and p-value cut-off of $0.05/(n-1)$. As before, the simpler structure facilitates the examination and interpretation of the co-expressed genes. Further work would involve annotation of the main clusters identified using Gene Ontology and other annotation databases.

4. Discussion

We have introduced a method to annotate dendrograms with a statistic that supplies additional information on the significance of each node. The length of the branches supplies some information already, but it is not a reliable measure. It is clear from the examples we provide that the p-values may have little to do with the branch lengths. When the number of remaining objects in a branch is small, the p-value should be interpreted with some caution since the number of permutations possible may be small; however, this is not a problem because the low branches are likely to be non-significant and we are generally interested in branches higher up.

Motivated by techniques from multivariate statistics, we work with the within-cluster variance matrix as a summary measurement from which we deduce “goodness” of clusters, using determinant for non-singular matrices, or its equivalent in the singular case by the use of singular values. We could imagine using one of the statistics directly as a global optimization function in producing clusters at the beginning, rather than using it as a post-hoc evaluation tool. However, finding a globally optimal solution requires searching in an exponentially large space and is a computationally intractable problem unless the number of objects is very small.

Because of how dendrograms are created, when one performs the permutation test described, one expects to find significance. Rejecting the null hypothesis of no clustering should be expected, but failing to reject it is surprising and of interest.

Most methods for finding some type of significance measure for a dendrogram require a reference distribution or a model from which random datasets are generated. These are then compared to the original data through some statistic, or by seeking repeated occurrences of same elements in a cluster. The simplest method, for instance, may be to sample from a uniform distribution for each variable, from the range of that variable found in the original data. A more sophisticated but computationally heavy variation is to sample uniformly from the convex hull computed from the data. Advantages of using such uniform reference distributions are not clear, however, particularly in high dimensional situations. Other null distributions include randomizing the dissimilarity matrix [37] and adding normally distributed errors to the data [38,10,8]. Perturbing the data with noise can be reasonable when one has a good idea of errors associated with each variable. For gene expression, however, the quantities that are needed are *gene-specific* variances, which cannot be obtained except in relatively large studies with enough replicates.

We have seen that we can create a new dendrogram by merging branches when the p-values are less than some cut-off value. The original dendrogram with a bifurcation at each node is replaced with a “smoother” one in which many similar elements may be joined at a single node. The cut-off p-value is a parameter that can be set according to the desired level of detail in the dendrogram, and as we see in our examples, this procedure can summarize the structure of the data in a concise and informative way. The null distribution that we adopt by switching the membership labels is simple and intuitive, and does not require making any unnecessary distributional assumptions.

While annotation of the dendrogram can provide additional information, its usefulness is related to the effectiveness of the clustering algorithm. The p-values can give some indication as to the relative compactness of clusters obtained, but not what the correct clustering should be. Weakness of the algorithm, such as visually appealing but inherently suboptimal leaf arrangement in hierarchical clustering, cannot be overcome with the annotation. It also should be stressed that the p-values we compute at each node are *conditional* on the clustering at the previous level. If the clustering fails to captures the structure correctly at the top of the dendrogram, for example, inferences on the structure of the data using the subsequent p-values should be interpreted in relation to the first splitting and its corresponding p-value.

Acknowledgments

This work was supported in part by grants ST32-AI07358 for P. J. P., T32 AI007535-10 for J. M., and R01-EB006195 for M. P. and M. B. from the National Institutes of Health.

References

1. Everit, B. Cluster Analysis. 3. Halsted Press; New York: 1993.
2. Gordon, A. Classification. Chapman & Hall/CRC; 1999.
3. Morgan B, Ray A. Non-uniqueness and inversions in cluster analysis. Applied statistics 1995;44(1): 117–134.
4. Cheng R, Milligan G. Measuring the influence of individual data points in a cluster analysis. Journal of classification 1996;13(2):315–335.
5. Bar-Joseph Z, Gifford DK, Jaakkola TS. Fast optimal leaf ordering for hierarchical clustering. Bioinformatics 2001;17(Suppl 1):S22–S29. [PubMed: 11472989]

6. Langfelder P, Zhang B, Horvath S. Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R. Bioinformatics 2008;24(5):719–720. [PubMed: 18024473]
7. Gibbons, FD.; Roth, FP. Judging the quality of gene expression-based clustering methods using gene annotation; Genome Res. 2002. p. 1574-1581.URL http://dx.doi.org/10.1101/gr.397002
8. McShane L, Radmacher M, Freidlin B, Yu R, Li M, Simon R. Methods for assessing reproducibility of clustering patterns observed in analyses of microarray data. Bioinformatics 2002;18(11):1462–1469. [PubMed: 12424117]
9. Monti S, Tamayo P, Mesirov J, Golub T. Consensus Clustering: A Resampling-Based Method for Class Discovery and Visualization of Gene Expression Microarray Data. Machine Learning 2003;52(1):91–118.
10. Zhang K, Zhao H. Assessing reliability of gene clusters from gene expression data. Functional & Integrative Genomics 2000;1(3):156–173. [PubMed: 11793234]
11. Kerr MK, Churchill GA. Bootstrapping cluster analysis: Assessing the reliability of conclusions from microarray experiments. Proc Natl Acad Sci USA 2001;98:8961–8965. [PubMed: 11470909]
12. Falkenauer, E. Genetic Algorithms and Grouping Problems. John Wiley & Sons, Inc; New York, NY, USA: 1998.
13. Di Gesu V, Giancarlo R, Lo Bosco G, Raimondi A, Scaturro D. GenClust: a genetic algorithm for clustering gene expression data. BMC Bioinformatics 2005;6:289. [PubMed: 16336639]
14. MacKay, D. Information Theory, Inference and Learning Algorithms. Cambridge University Press; 2003.
15. Grotkjaer T, Winther O, Regenberg B, Nielsen J, Hansen L. Robust multi-scale clustering of large DNA microarray datasets with the consensus algorithm. Bioinformatics 2006;22(1):58–67. [PubMed: 16257984]
16. Yeung KY, Medvedovic M, Bumgarner RE. Clustering gene-expression data with repeated measurements. Genome Biology 2003;4(5):R34. [PubMed: 12734014]
17. Eisen M, Spellman P, Brown P, Botstein D. Cluster analysis and display of genome-wide expression patterns. Proceedings of the National Academy of Sciences 1998;95(25):14863–14868.
18. van der Laan MJ, Pollard K. A new algorithm for hybrid hierarchical clustering with visualization and the bootstrap. Journal of Statistical Planning and Inference 2003;117(2):275–303.
19. Mielke P, Berry K, Johnson E. Multi-response permutation procedures for a priori classifications. Comm Statistit A 1976;5:1409–1424.
20. Good, P. Permutation, Parametric, and Bootstrap Tests of Hypotheses. Springer-Verlag New York, Inc; Secaucus, NJ, USA: 2005.
21. Wilks S. Certain generalizations in the analysis of variance. Biometrika 1932;24(3–4):471–494.
22. Wald A, Wolfowitz J. Statistical tests based on permutations of the observations. Ann Math Statist 1944;15:358–372.
23. Johnson, R.; Wichern, D. Applied multivariate statistical analysis. Englewood Cliffs, New Jersey: Prentice-Hall, Inc; 1982.
24. Strimmer K. Modeling gene expression measurement error: a quasi-likelihood approach. BMC Bioinformatics 2003;4:10. [PubMed: 12659637]
25. Mann H, Whitney D. On a test of whether one of two random variables is stochastically larger than the other. Annals of Mathematical Statistics 1947;18(1):50–60.
26. Gordon A. Identifying genuine clusters in a classification. Computational Statistics and Data Analysis 1994;18(5):561–581.
27. Ward J. Hierarchical grouping to optimize an objective function. Journal of the American Statistical Association 1963;58(301):236–244.
28. Marriott F. Practical problems in a method of cluster analysis. Biometrics 1971;27(3):501–14. [PubMed: 5116570]
29. Pitman E. Significance tests which may be applied to sample from any population. Royal Statistical Society Suplement 1937;4:119–130. 225–232.
30. Westfall, P.; Young, S. Resampling-based multiple testing: Examples and methods for p-value adjustment. Wiley-Interscience; 1993.

31. Alt K, Vach W. Odontologic kinship analysis in skeletal remains: concepts, methods, and results. Forensic Sci Int 1995;74(1–2):99–113. [PubMed: 7665137]

32. Bonetti, M.; Pagano, M. The interpoint distance distribution as a descriptor of point patterns, with an application to spatial disease clustering; Stat Med. 2005. p. 753-773.URL http://dx.doi.org/10.1002/sim.1947

33. Meselson M, Guillemin J, Hugh-Jones M, Langmuir A, Popova I, Shelokov A, Yampolskaya O. The sverdlovsk anthrax outbreak of 1979. Science 1994;266:1202. [PubMed: 7973702]

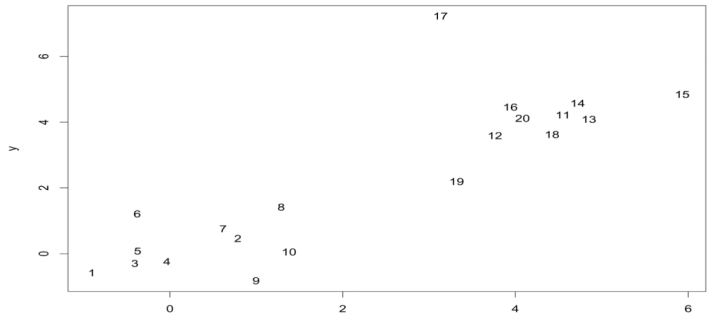
34. Schulz A, Williams D, Israel B, Lempert L. Racial and Spatial Relations as Fundamental Determinants of Health in Detroit. Milbank Quarterly 2002;80(4):677. [PubMed: 12532644]

35. Spira A, Beane J, Shah V, Liu G, Schembri F, Yang X, Palma J, Brody J. Effects of cigarette smoke on the human airway epithelial cell transcriptome. Proceedings of the National Academy of Sciences 2004;101(27):10143–10148.

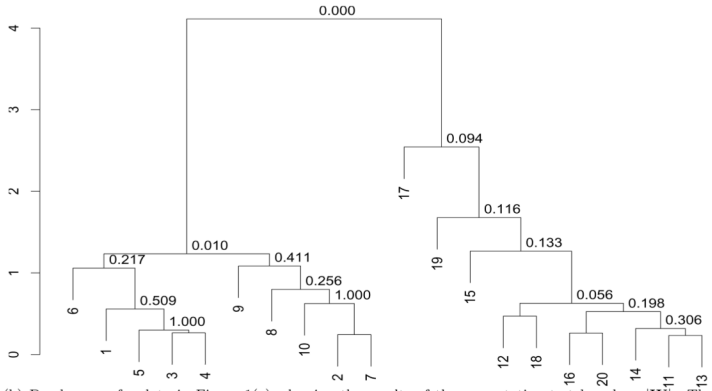
36. Pomeroy S, Tamayo P, Gaasenbeek M, Sturla L, Angelo M, McLaughlin M, Kim J, Goumnerova L, Black P, Lau C, et al. Prediction of central nervous system embryonal tumour outcome based on gene expression. Nature 2002;415(6870):436–442. [PubMed: 11807556]

37. Ling R. A probability theory of cluster analysis. Journal of the American Statistical Association 1973;68(341):159–164.

38. Bittner M, Meitzer P, Chen Y, Jiang Y, Seftor E, Hendrix M, Radmacher M, Simon R, Yakhini Z, Ben-Dor A. Molecular classification of cutaneous malignant melanoma by gene expression profiling. Nature 2000:536–540. [PubMed: 10952317]

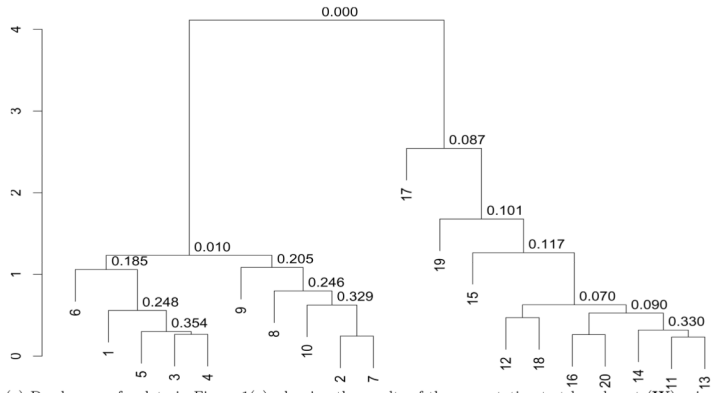


(a) Sample data: two clusters of size 10 (labeled {1-10} and {11-20}, respectively) are generated, each from a bivariate normal distribution with identity covariance matrix; the two means are $\boldsymbol{\mu}_1 = (0, 0)$ and $\boldsymbol{\mu}_2 = (4, 4)$.

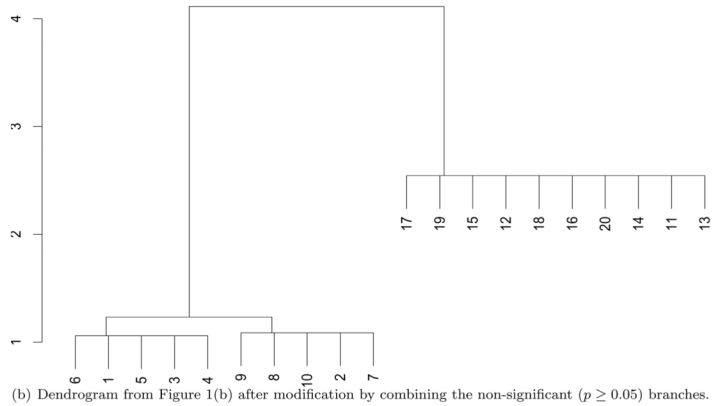


(b) Dendrogram for data in Figure 1(a), showing the results of the permutation test based on $|\mathbf{W}|$. The p-values are computed with 1000 membership-permuted datasets at each node.

Figure 1. An application of the permutation test to illustrative data

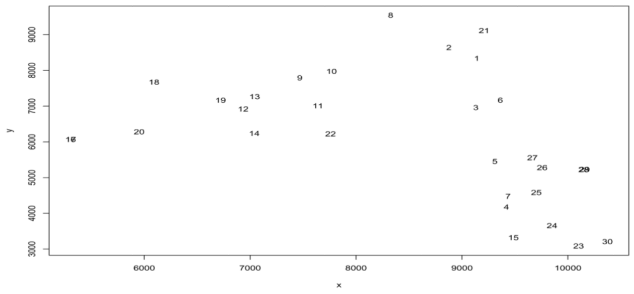


(a) Dendrogram for data in Figure 1(a), showing the results of the permutation test based on tr(**W**) using 1000 samples. The p-values are comparable to those obtained using |**W**|.

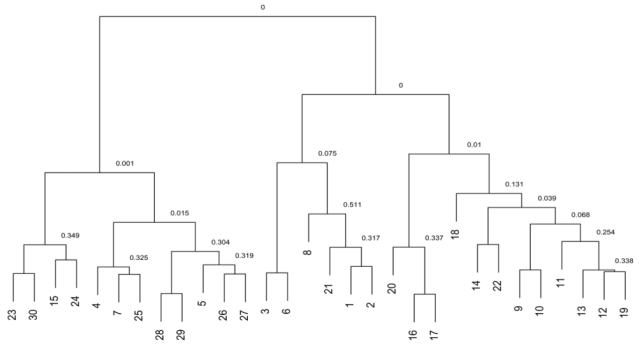


(b) Dendrogram from Figure 1(b) after modification by combining the non-significant ($p \geq 0.05$) branches.

Figure 2. A continuation of the application of the permutation test to illustrative data



(a) Cases from the medieval burial site data collected by Alt and Vach [31]. Note that there are two cases located at approximately (5500,6000), labeled 16 and 17 and two cases located at approximately (11000,5500) labeled 28 and 29.



(b) Dendrogram for data in Figure 3(a), showing the results of the permutation test based on $|\mathbf{W}|$. The p-values are computed with 1000 membership-permuted data sets at each node.

Figure 3. An application of the permutation test to a spatial data set

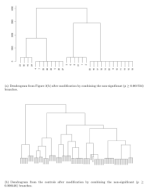
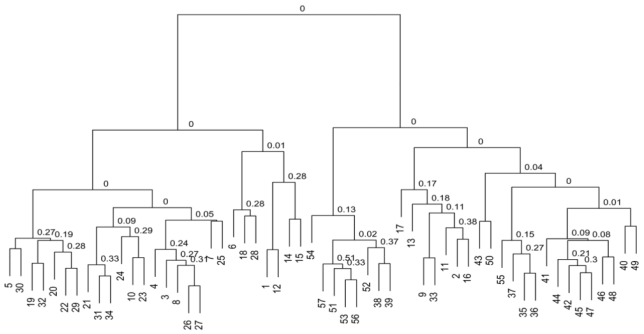
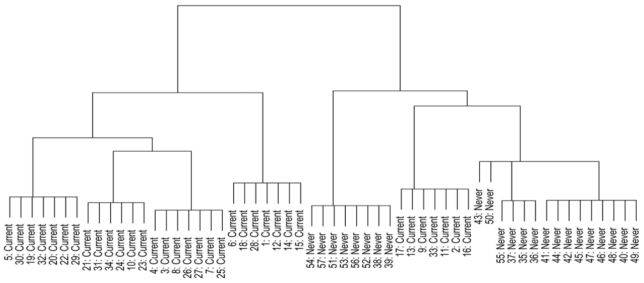


Figure 4. A continuation of the spatial data set application

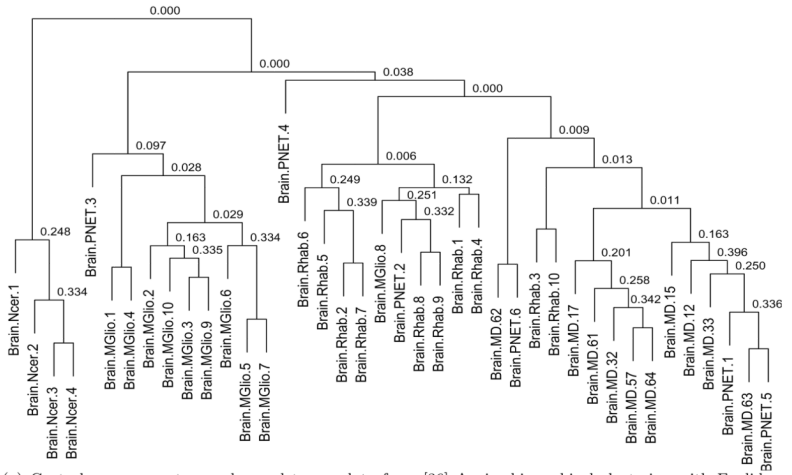


(a) Epithelial cell data from [35]. Patients {1–34} are *current smokers* and patients {35–57} are *never smokers*. Hierarchical clustering with Euclidean distance and average linkage was used. The p-values are based on the $tr(\mathbf{W})$ criterion

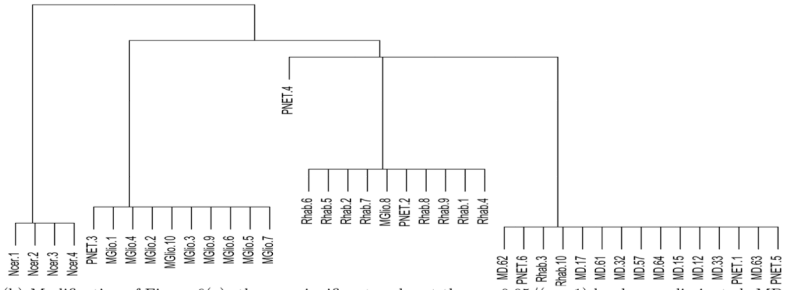


(b) Modification of Figure 5(a): the non-significant nodes at the $p = 0.000893$ level were eliminated.

Figure 5. Microarray data

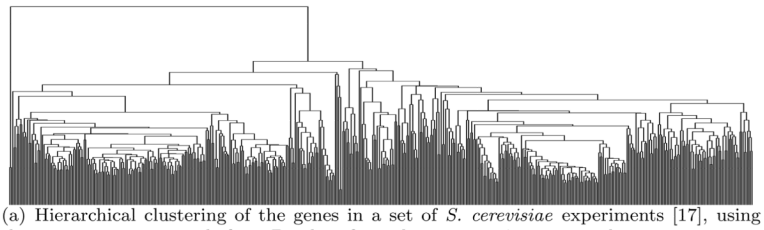


(a) Central nervous system embryonal tumor data from [36] Again, hierarchical clustering with Euclidean distance and average linkage was used and the p-values were computed based on 10000 iterations using the product of the singular values.

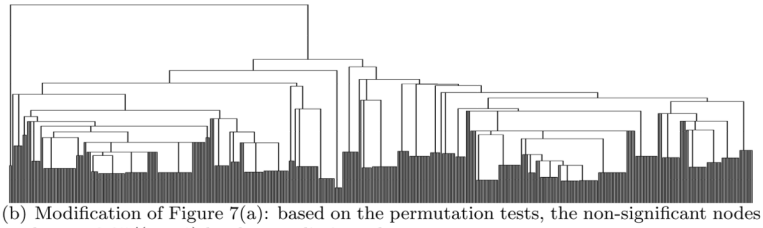


(b) Modification of Figure 6(a): the non-significant nodes at the $p = 0.05/(n-1)$ level were eliminated. MD: medulloblastoma; MGlio, malignant glioma; Ncer, normal cerebella; Rhab, renal/extrarenal rhabdoid tumors; PNET, supratenotorial primitive neuroectodermal tumor. The samples cluster along tissue types in most cases, except for PNET.

Figure 6. Application of the permutation test to CNS embryonal tumor data



(a) Hierarchical clustering of the genes in a set of *S. cerevisiae* experiments [17], using the same parameters as before. P-values from the permutation tests and gene names were deleted due to space limitations.



(b) Modification of Figure 7(a): based on the permutation tests, the non-significant nodes at the $p = 0.05/(n-1)$ level were eliminated.

Figure 7. Permutation test applied to genes in *S. cerevisiae* experiments