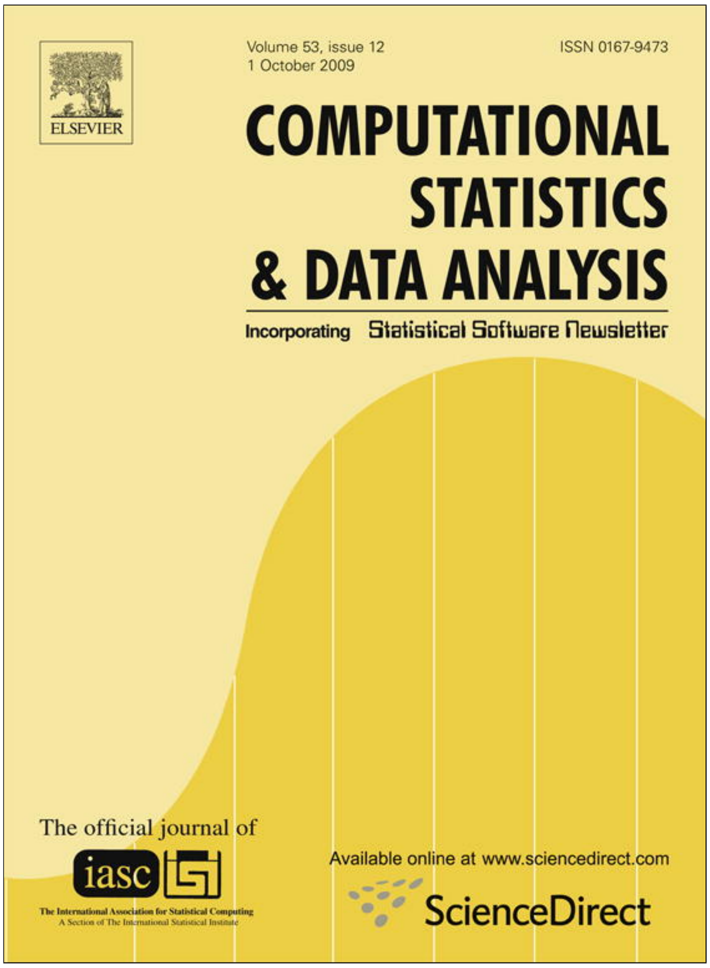

Volume 53, issue 12
1 October 2009

ISSN 0167-9473

COMPUTATIONAL STATISTICS & DATA ANALYSIS

Incorporating Statistical Software Newsletter

The official journal of iasc
The International Association for Statistical Computing
A Section of The International Statistical Institute

Available online at www.sciencedirect.com

ScienceDirect

ELSEVIER

Computational Statistics and Data Analysis 53 (2009) 3987–3998

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Partition clustering of high dimensional low sample size data based on p-values

George von Borries[a], Haiyan Wang[b,*]

[a] Departamento de Estatística, IE, Universidade de Brasília, 70910-900, DF, Brazil

[b] Department of Statistics, Kansas State University, 66506-0802, KS, USA

ARTICLE INFO

Article history:
Received 7 February 2009
Received in revised form 26 April 2009
Accepted 19 June 2009
Available online 26 June 2009

ABSTRACT

Clustering techniques play an important role in analyzing high dimensional data that is common in high-throughput screening such as microarray and mass spectrometry data. Effective use of the high dimensionality and some replications can help to increase clustering accuracy and stability. In this article a new partitioning algorithm with a robust distance measure is introduced to cluster variables in high dimensional low sample size (HDLSS) data that contain a large number of independent variables with a small number of replications per variable. The proposed clustering algorithm, PPCLUST, considers data from a mixture distribution and uses p-values from nonparametric rank tests of homogeneous distribution as a measure of similarity to separate the mixture components. PPCLUST is able to efficiently cluster a large number of variables in the presence of very few replications. Inherited from the robustness of rank procedure, the new algorithm is robust to outliers and invariant to monotone transformations of data. Numerical studies and an application to microarray gene expression data for colorectal cancer study are discussed.

Published by Elsevier B.V.

1. Introduction

Mining in high dimensional low sample size (HDLSS) data is an active research topic due to the advance in data collection technologies that allow the obtaining of information from a large number of variables (for example, genes, proteins) at the same time. Contradictory to the requirement of plenty of replications as demanded by traditional methods, the number of replications for such data is often limited due to time or cost constraint. For example, a medium-sized microarray study often contains information from thousands of genes with no more than a hundred samples for each gene. An important task is to investigate and identify disease response genes using the post-genome data. This can provide target for drug development in public health and give the focus for genetic alteration to yield disease resistant crops.

Statistical methods for such purposes are mainly in three categories. One category is through the analysis of individual gene and then apply false discovery rate (FDR) control (Benjamini and Hochberg, 1995; Efron, 2007) to adjust for multiple comparison issues. A large volume of work in the literature falls in this category. Even though FDR is meant to improve the identification of true positives, it still leads to conservative results in genomic applications (Storey and Tibshirani, 2003). This is especially true in the case of small sample sizes since the test statistics calculated from small replications are lack of power for nonparametric methods and are sensitive to deviations from assumptions for parametric methods. As a result, when only a small amount of useful information exists among a large amount of noises, the limitation of these methods prevails. A second category of methods is referred to as gene set enrichment that considers a set of genes selected based on biological knowledge from pathway information or literature mining to increase power (Subramanian et al., 2005; Efron and Tibshirani, 2007). Unfortunately, pathway or gene ontology information is not known for all genomes and so gene set enrichment

[*] Corresponding author.
E-mail addresses: gborries@unb.br (G. von Borries), hwang@ksu.edu (H. Wang).

0167-9473/$ – see front matter. Published by Elsevier B.V.
doi:10.1016/j.csda.2009.06.012

Table 1
High dimensional data layout, where $a \to \infty$ and $n_i \geq 2$.

Factor level	Distribution	Observations				Sample size
1	$F_1(x)$	X_{11}	X_{12}	$\dots$	X_{1n_1}	n_1
2	$F_2(x)$	X_{21}	X_{22}	$\dots$	X_{2n_2}	n_2
⋮	⋮	⋮	⋮	⋮	⋮	
a	$F_a(x)$	X_{r1}	X_{r2}	$\dots$	X_{rn_r}	n_a

methods may not be applicable. A third category is through clustering to identify groups of differentially expressed genes (Fraley, 1998; Alon et al., 1999; Notterman et al., 2001; Yeung and Ruzzo, 2001; Jiang et al., 2004; Huttenhower et al., 2007; Fu and Medico, 2007). Clustering based methods are more flexible. However, non-probabilistic distance measures and corresponding clusters obtained can lead to difficulty in interpretation. Further, most algorithms are sensitive to monotone transformations and produce different results when applied to different transformations of data. In addition to the above mentioned problems, most available methods require a user to pre-specify the number of clusters. This is difficult and could produce misleading results when incorrect number of clusters are specified. Mixture model based clustering (MCLUST) developed by Fraley and Raftery (2006) can automatically estimate the number of clusters using Bayesian Information Criterion. However, this algorithm relies heavily on normality assumption and may produce poor clustering accuracy when the data are heavily skewed. Further, as pointed out by the authors, MCLUST is not recommended to apply to HDLSS data directly due to its dependence on the covariance matrix estimation.

We propose to approach the problem from a combination of clustering and gene set enrichment idea without having to rely on known biological information. Specifically, we assume that at least two replications are available for each variable (gene) to start with. All the variables and their observations together can be viewed as originated from high dimensional mixtures of distributions, where each unique distribution defines a cluster. We then introduce a new partitional algorithm using a robust measure of similarity to cluster the large number of variables. The robust similarity measure evolves from p-values obtained from the rank test of no nonparametric effect of groups (Wang and Akritas, 2004) specially developed for the HDLSS structure. The new algorithm can automatically determine the number of clusters and are invariant to monotone transformations of data. Numerical studies show that the proposed algorithm has high clustering accuracy and stability. Additionally, the algorithm is fast and do not show memory allocation problems observed in some algorithms when the number of variables in the study is very high (10 000 or more variables).

2. Review of the nonparametric test for homogeneous distribution

Suppose we have observations from a mixture of unknown distributions. Let a cluster be all the observations generated from the same distribution. Differences among clusters can be reflected in many ways such as different mean values or different variances. In this article, the problem of clustering on observations is proceeded as a problem of detecting a significant difference on the distribution of the observations from each distribution. Let X_{ij} denote the jth observation from the ith variable (or factor), where $\{X_{ij}, 1 \leq j \leq n_i\}$ are independent observations from some unknown distribution $F_i(x)$, $i = 1, 2, \dots, a$. The observed data can be viewed as a matrix with elements X_{ij}. Each row represents the level of a factor, and each column represents an observation (replication), as is shown in Table 1.

We first test to see if these observations are from the same distribution, i.e., we test the hypothesis

$$H_0 : F_1(x) = \cdots = F_a(x). \tag{1}$$

The Kruskal–Wallis test can be used when the number of distributions is small. However, the test is not valid in a high dimensional setting since the inference is based on large sample size and small number of distributions. We also do not recommend to use traditional ANOVA F-test as the error terms in ANOVA model need to be i.i.d. Gaussian with a constant variance. Akritas and Arnold (2000) showed that the ANOVA F-test is robust to departure from homoscedasticity when there are a large number of factors, but it is not asymptotically valid for unbalanced data with small sample sizes even under homoscedasticity. Later, Akritas and Papadatos (2004) considered test procedures for unbalanced and/or heteroscedastic situations when the number of factors tends to infinity. However, their tests are based on original observations that are not invariant to monotone transformation of data. To overcome all these limitations, Wang and Akritas (2004) considered a nonparametric rank test of the null hypothesis of equality of distribution functions for each factor level when the number of factors is large and the number of replications is either small (referred as HDLSS data here) or large. We use the p-value from testing the hypothesis in (1) using the test statistic in Wang and Akritas (2004) as the measure of similarity among groups.

Let R_{ij} represent the (mid-)rank of observation X_{ij} in the set of all $n_1 + n_2 + \cdots + n_a$ observations. Then under H_0, all observations are i.i.d. realizations of a common distribution. So these (mid-)ranks are discrete uniformly distributed random numbers between 1 and $\sum_{i=1}^{a} n_i$ for continuous data. Let $\overline{R}_{i.} = n_i^{-1} \sum_{j=1}^{n_i} R_{ij}$ be the mean rank of observations for the ith factor level and $\widetilde{R}_{..} = a^{-1} \sum_{i=1}^{a} \overline{R}_{i.}$ be the overall unweighted mean of ranks from all factor levels. Define the test statistic,

$$F_R = \frac{MST_R}{MSE_R} \tag{2}$$

where MST_R is the unweighted mean square error due to factor levels calculated over (mid-)ranks:

$$MST_R = \frac{1}{a-1}\sum_{i=1}^{a}(\bar{R}_{i.} - \widetilde{R}_{..})^2, \tag{3}$$

and MSE_R is the pooled estimate of the sample variance, also obtained over (mid-)ranks:

$$MSE_R = \frac{1}{a}\sum_{i=1}^{a}\frac{1}{n_i}S^2_{R,i}, \tag{4}$$

where $S^2_{R,i}$ is the sample variance calculated using (mid-)ranks of observations from the ith factor level.

The asymptotic distribution of $\sqrt{a}(F_R - 1)$ under H_0, as $a \to \infty$, is given in Wang and Akritas (2004). For convenience of further discussion, we restate the theorem below.

Theorem 1. *Let $F_i(x)$ be arbitrary cumulative distribution functions and $H(x) = (\sum_{i=1}^{a} n_i)^{-1}\sum_i n_iF_i(x)$ be the average cumulative distribution function. Assume that the observations are independent. Define $\sigma_i^2 = \mathrm{Var}(H(X_{ij}))$ and*

$$\upsilon_2^2 = \frac{1}{a}\sum_{i=1}^{a}\frac{1}{n_i}\sigma_i^2 > 0, \qquad \tau_2 = \frac{1}{a}\sum_{i=1}^{a}\frac{2\sigma_i^4}{n_i(n_i-1)}. \tag{5}$$

Then under $H_0 : F_1(x) = \cdots = F_a(x)$, the limit of τ_2/υ_2^4 exists as $a \to \infty$. Further,

$$\sqrt{a}(F_R - 1) \xrightarrow{d} N(0, \lim_{a\to\infty}\tau_2/\upsilon_2^4), \tag{6}$$

regardless of n_i stay bounded or go to ∞, provided that $\max_i\{n_i\}/\min_i\{n_i\} = O(1)$ for $n_i \geq 2$.

The statistic $\sqrt{a}(F_R - 1)$ compared to the normal critical values can be used to obtain an approximate p-value to give a sample evidence of the homogeneity of the distributions. A large p-value indicates that the given sample does not provide evidence to conclude that the factor levels being tested have different distributions. In such case, we cluster these factors levels into the same group. In contrast, a small p-value gives evidence against H_0 indicating that at least two distributions are different. The use of the hypothesis testing results from (6) to obtain similarity measure allows flexible modeling and robust clustering at the same time. With this general setup, the data collected can be balanced or unbalanced and the user does not have to worry about normality or skewness of the data. Heteroscedastic variances are naturally incorporated. This is important as gene regulations are very complicated and the variations of the expression data from different genes can be dramatically different. In addition, the results hold for small or large sample sizes. In particular, allowing reliable inference with the sample sizes as small as two can lead to significant reduction in cost for considering the number of arrays required.

Before we apply the results of (6) in clustering, we first evaluate its performance. The estimated type I error and power were not studied in Wang and Akritas (2004). We report our simulation results in the next section.

2.1. Type I error and power estimate when the number of variables is large

Table 2 reports the Type I error estimate using the asymptotic distribution of the test statistic in (6) at significance levels 0.10, 0.05 and 0.01. For performance of other nonparametric tests in such a setting, one can see Akritas and Papadatos (2004). In the simulations the number of random variables, a, takes values 1000, 2000 and 4000, and the number of observations per variable is set to be 4. The simulations are based on 2000 runs and observations were generated from normal, lognormal, exponential, and Cauchy distributions. The Jackknife bias corrected estimator (Pawitan, 2001) of σ_i^4 were used in the estimation of the asymptotic variances.

The Type I error rates reported in Table 2 are close to the true α levels, indicating that the test statistic $\sqrt{a}(F_r - 1)$ performs well in testing the hypothesis in (1) regardless of whether the distribution is symmetric (normal), skewed (lognormal, exponential), or heavy tailed (Cauchy).

To study the power of the test described in Section 2, we generated data for 2000 random variables from mixture distributions with four observations per variable. Normal, lognormal, exponential, and Cauchy distributions are considered to evaluate robustness of the test. For all cases except the exponential distribution, observations for 95% of the variables are generated with the distribution having location parameter 0 and scale parameter 1, and the remaining 5% of the random variables have location parameter d ranging from 0 to 2.25. The achieved power at significance level $\alpha = 0.05$ is given in Fig. 1. The test appears to be very powerful in detecting small proportion of differences in all cases considered.

3. Partition clustering algorithm based on *p*-values

The p-values obtained from the test in Section 2 can serve as a similarity measure in a clustering algorithm with high dimensional data. In this section, we introduce a partition algorithm, PPCLUST (p-values based partitional clustering), to iteratively conduct nonparametric hypothesis testing and partition the random variables into subgroups whenever the similarity is below a certain threshold. That is, a group of variables is partitioned into two smaller groups when the test

Table 2
Type I error estimate. The test has accurate size regardless of the distribution being symmetric, skewed or heavy-tailed.

Distribution	Number of Factor levels	Nominal level		
		0.10	0.05	0.01
		Type I error		
Normal(0,1)	1000	0.0965	0.0500	0.0130
	2000	0.1060	0.0620	0.0110
	4000	0.1075	0.0535	0.0135
Lognormal(0,1)	1000	0.1040	0.0605	0.0155
	2000	0.1150	0.0610	0.0140
	4000	0.1140	0.0620	0.0170
Exponential(1)	1000	0.1160	0.0625	0.0130
	2000	0.1100	0.0585	0.0130
	4000	0.1135	0.0605	0.0155
Cauchy(0,1)	1000	0.1105	0.0545	0.0165
	2000	0.1005	0.0555	0.0150
	4000	0.1100	0.0625	0.0105

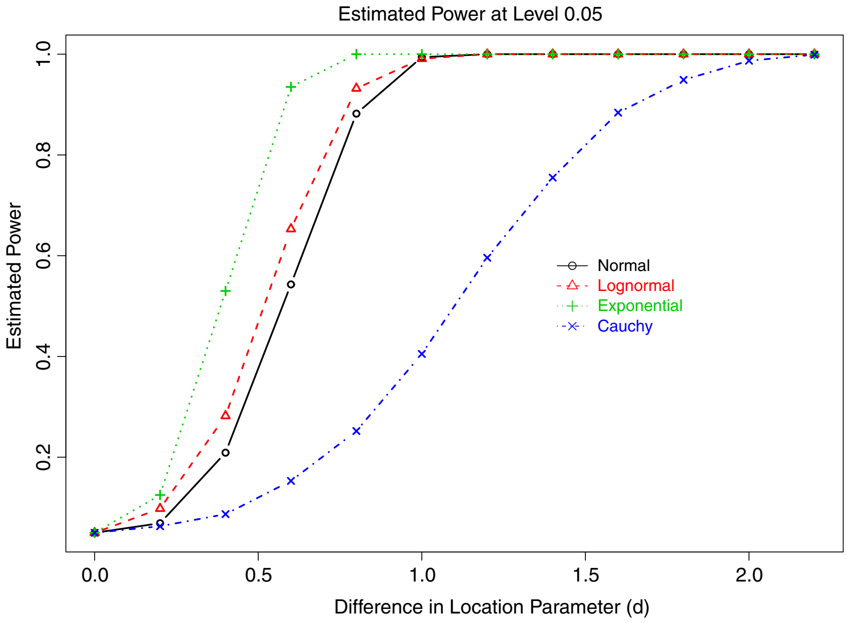

Fig. 1. Achieved power for HDLSS data with $\alpha = 0.05$, considering shifted differences in mean (d) in a group of 100 factor levels in a total of 2000 factor levels and data generated from four distributions: Normal(0, 1) (continuous line in blue), Lognormal(0, 1) (dashed line in black), Exponential(1) (dotted line in red) and Cauchy(0, 1) (dotted-dashed line in green).

of identical distribution in (1) is rejected and the group remains intact if the test is not rejected. When H_0 is rejected, smaller groups are created for further testing. The algorithm continues until when there are no groups with similarity measures below the threshold.

3.1. The algorithm

For $g \geq 1$, let $g - 1$ be the number of groups identified such that all the variables within each group have identical distribution. PPCLUST is described below in 9 steps. Throughout the algorithm, the subset of data to be tested are always stored as in Table 1 with each row representing a random sample from the same variable.

1. Let D_1 denote the matrix of observations from all variables as in Table 1. Each row contains observations from the same variable. The number of rows in D_1 is denoted as $n_f(D_1)$. Set $g = 1$.
2. Calculate the (mid-)rank of all the observations in D_1 and store them in D_{1R} in the same format as in Table 1.
3. Calculate the median (mid-)rank for each variable (i.e. each row) in D_{1R}.
4. Sort the variables (i.e. rows) in D_1 according to the median ranks from Step 3.
5. Conduct the test to evaluate if the variables (rows) in D_1 have identical distribution.
   5.1. If H_0 is not rejected, report all the variables in D_1 as a single group. Go to Step 9.
   5.2. If H_0 is rejected: continue to Step 6.
6. Take the first half of the number (rounded to integer) of variables from consecutive rows of D_1 and denote the data in this subset including corresponding observations as D_2. Let $n_f(D_2)$ be the number of variables in D_2.

```
7. Conduct the test to evaluate if the variables (rows) in $D_2$ have identical distribution.
  7.1 If $H_0$ is not rejected:
      7.1.1. Assign the variables of $D_2$ and corresponding observations to group $g$.
      7.1.2. Assign $g + 1$ to $g$.
      7.1.3. Remove the variables in $D_2$ and corresponding observations from $D_1$.
      7.1.4. If $n_f(D_1) = 0$, then go to Step 9.
      7.1.5. If $n_f(D_1) \geq 1$, then do steps A and B below:
             A. Test to see if each variable in $D_1$ belongs to the newly assigned group by testing the corresponding
                hypothesis that all involved random variables have the same distribution. Remove the variable and its
                observations from $D_1$ when $H_0$ is not rejected and put them into the newly assigned group.
             B. Let $D_2$ be the set that contains the remaining variables and their observations in $D_1$ and go to Step 8.
  7.2. If $H_0$ is rejected:
      7.2.1. Take the first half of the number (rounded to integer) of variables from $D_2$ and denote the data from this subset
             with corresponding observations as $D_3$.
      7.2.2. Assign all the variables and corresponding observations that are not in $D_3$ to $D_1$.
      7.2.3. Let $D_2 = D_3$ and delete $D_3$.
      7.2.4. Go to Step 8.
8. If $n_f(D_2) = 1$, then perform Steps 8.1-8.5; otherwise, return to Step 7.
  8.1. Allocate the variable and corresponding observations in $D_2$ to group 0.
  8.2. Remove the variable in $D_2$ and corresponding observations from $D_1$.
  8.3. If $n_f(D_1) = 0$ then go to Step 9.
  8.4. If $n_f(D_1) = 1$, then let $D_2 = D_1$ and return to Step 8.
  8.5. If $n_f(D_1) > 1$, then let $D_2 = D_1$ and go to Step 7.
9. Stop the clustering and report the groups identified.
```

Remark. For Step 3, please note that each variable has multiple i.i.d. observations. The sorting is only done to the variables, not to the observations. The observations from each variable remain unordered so that they are still independent and identically distributed. For the same set of variables to be tested with given i.i.d observations from each variable, the test statistic defined in (2) and the asymptotic variance calculated in (5) using Jackknife bias corrected estimator of $\widehat{\sigma}_i^4$ remain unchanged no matter we sort the variables or not. Therefore, the sorting has no effect on the test. However, it provides computational advantage for the clustering by putting similar variables in nearby groups without altering the basic requirement of Theorem 1.

For Steps 6 and 7.2.1, an alternative way to partition the variables is to split between two rows that has the largest gap in their median ranks. This can potentially increase the speed of clustering if the distributions underlying different clusters are well separated. However, the advantage is not significant if the underlying distributions have substantial overlap as in the numerical study in Section 4.

Step 7 basically repeatedly partition and group the variables until no further partition is possible.

Step 8 basically put the random variables that cannot be clustered to any of the identified groups into a group labeled as 0. Therefore, the random variables with group label 0 are not necessarily similar (or dissimilar). Instead, they are judged to belong to none of the identified groups. In other words, the random variables in group 0 resulted in a rejection of H_0 when tested with random variables of any other identified group.

By the end of the algorithm, $g - 1$ is the total number of different groups. A group labeled with a lower number in the output contain random variables with lower median observation values than those groups labeled with higher numbers. For example, if the data are the ratios of gene expressions under a treatment and a control, a group labeled with a lower number may contain down-regulated genes and a group labeled with a higher number may contain up-regulated genes. Intermediate groups contain genes not differentially expressed. In addition to the up or down regulations, the genes from different groups are significantly different as a result of the hypothesis testing.

3.2. About the significance level to use

Note that to determine if all the variables in a group have identical distribution, Theorem 1 only applies when the number of variables (rows) is large. As the partition proceeds, the number of variables in the group to be tested will reduce. The left panel of Table 3 gives the type I error estimate when the number of variables is no more than 500 when each variable contains two replications (under four different distributions). This and Table 2 together indicate that the test in Theorem 1 is liberal when the number of variables is no more than 50. To remedy this, we suggest to use small significance level in determining whether to reject a test. We recommend to take the upper bound of all significance levels, α, such that smaller levels yield similar clustering results. If a significance level used leads to too many small clusters, it indicates that the level is not small enough and the clustering results obtained is not reliable. This is because the test does not have acceptable type I error for small number of variables with small sample sizes. In such case, even smaller significance levels need to be considered.

We choose not to use Kruskal–Wallis test in that this test requires large sample sizes and small number of variables. Our numerical results show that this test is very conservative when the number of variables is large and the sample sizes are small (see the right panel of Table 3 for type I error estimate). For example, in a simulation we generated 15 random variables with scale parameter 1 from normal, lognormal, and Cauchy distributions. Ten of them have location parameter 0.5 and the

Table 3
Type I error estimate at level $\alpha = 0.05$ for the test in Theorem 1 (left panel) and Kruskal–Wallis test (right panel) under four distributions when the number of variables is below 500. Each variable has 2 replications. The test in Theorem 1 is liberal when the number of variables is no more than 50 and the Kruskal–Wallis test is conservative for all the cases considered. All distributions have location parameter 0 and scale parameter 1 (the Unif is for uniform distribution on (0, 1)).

Test in Theorem 1					Kruskal–Wallis test				
n_f	Unif	Normal	Lognormal	Cauchy	n_f	Unif	Normal	Lognormal	Cauchy
7	0.158	0.157	0.152	0.160	5	0.000	0.000	0.000	0.000
15	0.106	0.096	0.102	0.119	10	0.002	0.002	0.001	0.002
30	0.077	0.086	0.080	0.079	15	0.005	0.004	0.002	0.002
50	0.066	0.067	0.072	0.073	20	0.005	0.007	0.004	0.004
75	0.061	0.054	0.054	0.062	25	0.006	0.003	0.004	0.003
100	0.056	0.052	0.048	0.056	30	0.007	0.003	0.007	0.005
150	0.053	0.060	0.051	0.051	50	0.006	0.005	0.004	0.004
200	0.052	0.051	0.052	0.048	70	0.007	0.009	0.005	0.009
250	0.048	0.050	0.052	0.051	100	0.006	0.010	0.007	0.008
300	0.048	0.058	0.052	0.046	120	0.011	0.005	0.007	0.005
350	0.051	0.046	0.058	0.051	150	0.009	0.013	0.007	0.006
400	0.052	0.048	0.049	0.049					
450	0.049	0.051	0.052	0.044					
500	0.048	0.046	0.046	0.048					

remaining 5 variables have location parameter 0. Two replications were generated for each variable. The estimated power at level 0.05 from the Kruskal–Wallis test for these distributions are 0.006, 0.006, and 0.004 respectively. So Kruskal–Wallis test is not sensitive enough to detect heterogeneous distributions to partition the data.

3.3. *Advantage of PPCLUST compared to traditional clustering algorithms*

The robust similarity measure and the clustering mechanism entails the following advantage of PPCLUST.

1. **Invariance to monotone transformations:** The use of overall ranks of the observations in the test statistic leads to similarity measure that is invariant to monotone transformation of data and this in turn makes PPCLUST to have such property. Many clustering algorithms produce different results before and after monotone transformations of the data due to the fact that such transformations change the similarity matrices used in clustering. PPCLUST does not have this drawback so that a user does not need to explore appropriate transformations of data to satisfy some model assumptions. This is particularly useful since choosing appropriate transformations for HDLSS data is a difficult question itself.
2. **Automatic specification of the number of groups:** PPCLUST does not require the number of clusters to be specified in advance. It will determine the number of clusters automatically by specification of a significance level as the threshold to be compared with the p-values for testing the hypothesis of identical distribution. Estimating the number of mixture components is itself a popular research topic that is often computationally extensive. In low dimensional case, it has been a nuisance and difficult for a user to choose the number of clusters even though the clustering results may be visualized. In high dimensional case, effective visualization tools are not available to aid a user. So it is even harder to specify the number of clusters for a real dataset. PPCLUST produces this information directly.
   The specification of a significance level is not as intrusive as the specification of the number of groups, which is one of the objectives of clustering analysis. In fact, the significance level can be used as a guidance in finding the number of groups in a real data set. For example, decreasing the significance level in PPCLUST will decrease the number of groups found because it decreases the Type I error committed by the test. The use of different significance levels can serve as a fine tuning parameter in revealing the total number of different groups G where the algorithm tends to stabilize, i.e, find G that is more common to different α levels. This can be used as an indication of the true number of groups in the data. We remark that lowering the significance level too much will also decrease the power of the test in finding new and small groups. The delicate balance can be achieved in the same way as how we handle the type I and type II error in regular hypothesis testing.
3. **Less concern for multiple comparison problems in HDLSS data:** Reducing false discoveries while striving to maintain the power to identify true discoveries is one of the challenges for HDLSS data analysis (Storey, 2002; Sabatti, 2006; Qiu and Yakovlev, 2006; Strimmer, 2008). This is less of a concern in PPCLUST since the test is applied to groups of variables instead of on a one-by-one basis.
4. **PPCLUST favors HDLSS for asymptotic distribution of the test statistics while other algorithms often need prior dimension reduction before being applied to high dimensional data.** In high dimensional studies it is common to apply some dimension reduction technique such as principal components analysis before clustering data (Johnson, 1998). Some studies do not recommend the use of PCA before clustering except in very special situations (Yeung and Ruzzo, 2001). Simulations in Yeung and Ruzzo show that clustering principal components instead of original data produce different results on many algorithms using different similarity metrics. PPCLUST does not require previous dimension reduction to the analysis. Instead, PPCLUST takes advantage of the high dimensionality to provide power to produce reliable similarity measure. This is specially appealing when only very small number of replications are available.

5. **Flexible to work with unbalanced data with small sample sizes:** the algorithm works with both balanced or unbalanced data. The only requirement is that the number of replications per variable is at least 2. There is no need that all variables have the same number of replications. Unbalanced data is common in studies of microarray gene expression data and some algorithms require balanced data. Solutions like elimination from the study of factor levels with incomplete information or imputation of data can hide or seriously compromise the result of the study.
6. **PPCLUST produces fast solution for computationally costly problems as the computational complexity is O(log2(N)).** Note that traditional clustering algorithms need to do optimization at each stage to find the optimal partition of the data based on a criteria. As the number of variables increases, the optimization cost becomes a major concern for exhaustive search. Genetic algorithms are often used to speed up the search. Instead of searching for the optimal solution at each stage, PPCLUST relies on statistical evidence obtained from hypothesis testing to judge whether a group of variables is from the same distribution or not. As long as the null hypothesis is not rejected, the members are not significantly different and therefore a group is formed. In other words, PPCLUST only need the similarity measure from hypothesis testing and eliminates the optimization process. With the similarity measure being obtained through a single test of hypothesis, the computational burden is dramatically reduced to $O(\log_2(N))$ as opposed to $O(N\log_2(N))$, the best time complexity case for hierarchical clustering. This is confirmed from our simulations (see Section 4), where it takes PPCLUST less than a minute to complete the clustering of a data set containing up to 7000 random variables with sample sizes ranged from 5 to 20 per variable using PC machine running Windows XP with Intel Pentium M processor, 1.6 GHz, and 1 Gb of RAM memory.

4. Numerical comparison for clustering of HDLSS data

In this section, we compare PPCLUST with some benchmark algorithms on simulated data.

To evaluate the similarity between two clustering partitions, Rand (1971) proposed the Rand index that gives the fraction of all pairs that are correctly put in the same cluster or correctly put in separate clusters. However, the expected value of the Rand index of two random partitions does not take a constant value. Hubert and Arabie (1985) considered the adjusted Rand index (ARI) which is centered at zero and has maximum value of 1 achieved when the two partitions are identical up to renumbering of the subsets. Milligan and Cooper (1986) compared multiple indices for measuring agreement between two partitions in clustering analysis with different numbers of clusters, and they recommended the ARI as the index of choice. We adopt the ARI to compare the performance of these algorithms in clustering consistency compared to the truth as is known from data generation.

Study I: Clustering for symmetric data

In the following simulations we generated high dimensional data from mixture distributions having mixture components similar to the gene expression data from a colorectal cancer study (Notterman et al., 2001), which contain several large groups having overall distribution of a t-distribution with 15 degrees of freedom shifted by some location parameter μ and stretched by a scale parameter σ. Specifically, observations for 4000 random variables were generated according to the following scheme:

- Group 1: 300 random variables from $0.25 \times t_{15} - 0.5$.
- Group 2: 200 random variables from $0.25 \times t_{15} - 0.2$.
- Group 3: 2500 random variables from $0.25 \times t_{15}$.
- Group 4: 800 random variables from $0.25 \times t_{15} + 0.5$.
- Group 5: 200 random variables from $0.25 \times t_{15} + 1$.

The densities of these five distributions have substantial overlap. Five observations were generated for each random variable. PPCLUST using significance level $\alpha = 10^{-8}$ and the following 10 benchmark clustering algorithms are applied to the generated data:

- Partitional Algorithms: K-means, partitioning around medoids (PAM), clustering large applications (CLARA) with Euclidean metric, Self-Organizing Maps (SOM) with dimension 5×1.
- Hierarchical Algorithms: hierarchical clustering (HCLUST) with Ward's agglomeration method, agglomerative nesting (AGNES), divisive analysis clustering (DIANA) with Euclidean metric, hierarchical clustering by minimum energy distance with Euclidean norm $\|x - y\|$).
- Fuzzy Algorithm: fuzzy clustering (FANNY).
- Model Based Algorithm: mixture model based clustering (MCLUST) with automatic choice of best model through Bayesian Information Criterion.

For details of each algorithm, one can see McQueen (1967), Kaufman and Rousseeuw (1990), Kohonen (1989), Székely and Rizzo (2005) and Fraley and Raftery (2006).

In all algorithms that need pre-specification of the number of clusters, we set the number to be 5, the true number of groups. It should be noted that this information is often not known in real practice which contribute to additional uncertainty for their clustering performance. *R* software (version 2.4.1) with packages energy, mclust, cluster, and SOM were used in the simulation. PPCLUST was written in SAS$^{©}$ macro language (version 9.3.1), and the ARI was calculated using both R and SAS. For each algorithm in R, we use the default setting except that we supply the number of clusters with the true number

Table 4
Mean and standard deviations (std) of adjusted Rand index for all algorithms over 200 simulated datasets. Different sample sizes are considered. The groups are generated from symmetric distributions (Study I).

Adjusted Rand index								
	Sample sizes							
	5		10		15		20	
Algorithm	Mean	Std.	Mean	Std.	Mean	Std.	Mean	Std.
PPCLUST	**0.8600**	**0.0104**	**0.9205**	**0.0110**	**0.9541**	0.0084	**0.9688**	0.0059
PAM	0.4709	0.0323	0.5342	0.0166	0.5352	0.0135	0.5346	0.0121
K-means	0.4652	0.0416	0.4938	0.1585	0.5476	0.1697	0.5407	0.1362
Energy	0.4846	0.0469	0.6456	0.1149	0.8562	0.0851	0.9184	0.0262
Mclust	**0.6184**	0.1163	**0.8826**	0.0894	**0.9425**	**0.0056**	**0.9613**	**0.0043**
Clara	0.4219	0.0927	0.5315	0.0656	0.5516	0.0565	0.5527	0.0560
Diana	0.5662	0.1477	0.6182	0.1387	0.6401	0.1151	0.7145	0.1052
HCLUST	0.4846	0.0469	0.6456	0.1149	0.8559	0.0852	0.9184	0.0262
Agnes	0.4963	0.0557	0.6915	0.1193	0.8551	0.0841	0.9150	0.0332
Fanny	0.3820	0.1368	0.6122	0.0431	0.6067	0.0378	0.5856	0.0516
SOM	0.2920	**0.0048**	0.3750	**0.0057**	0.4280	**0.0059**	0.4622	**0.0045**

of groups. For example, by default, the algorithm of Hartigan and Wong (1979) is used for K-means. In addition, with the specified number of clusters in the K-means algorithm, a random set of (distinct) rows of the data is automatically chosen as the initial centers. The random selection for the centers and rows is the standard initialization method used in R. It has been confirmed empirically to have better performance than other initialization methods (Bradley and Fayyad, 1998; Pena et al., 1999).

To evaluate the stability of the clustering performance, we repeat the data generation 200 times and apply above algorithms on these 200 data sets. In order to verify the performance of PPCLUST under different sample sizes, the complete simulation study was repeated considering also samples of sizes 10, 15, and 20.

The average and standard deviation of the ARI reflect the clustering accuracy and stability respectively. They are reported in Table 4 for all the algorithms applied to the 200 data sets with different sample sizes. The best two mean ARIs and standard deviations are highlighted.

From Table 4, it can be seen that as the number of replications increases, the clustering accuracy increases for all algorithms. PPCLUST has the best clustering accuracy for all sample sizes considered. In addition, PPCLUST is also the most stable algorithm for small sample size (5) among all 11 algorithms since the ARI of PPCLUST has the smallest standard deviation for sample size 5. The standard deviation of the ARI for PPCLUST stays almost the same for sample sizes 5, 10 and 15. MCLUST has comparable average ARI to PPCLUST for sample sizes 15 and 20, but has significantly worse performance than PPCLUST for small sample sizes in both clustering accuracy and stability. SOM showed consistent stability but with very low clustering accuracy as the average ARI for SOM is less than 0.5 for all sample sizes. Algorithms Energy, HCLUST, and Agnes, are competitive to PPCLUST for samples of size 15 or higher, but those algorithms are not as stable as PPCLUST and MCLUST. Diana and Fanny showed the lowest stability among all algorithms and should not be used with HDLSS data. Fig. 2 gives a graphical summary of the performance of these algorithms through boxplots. Overall, PPCLUST has the best clustering performance in terms of both accuracy and stability. For larger samples, MCLUST is a good alternative to PPCLUST.

Study II: Clustering for skewed data

In a second study, the data generated for study I are transformed using the function $e^{4(x+1)}$, where x is an observation generated in study I. The resulting distribution of the data is close to a lognormal distribution but with more extreme points since x was generated from t-distribution instead of a normal distribution. The resulting distributions are heavily skewed. There is still a significant amount of overlap among the densities.

Table 5 and Fig. 3 summarize the clustering performance of these algorithms on the transformed data. PPCLUST is considerably better than all other algorithms in all sample size situations. Clara has the worst results and PAM is the best algorithm among the other algorithms, but never had average ARI greater than 0.65. PPCLUST applied to the transformed data yields identical results to those before the transformation because it is invariant to monotone transformations.

The simulations and all calculations were performed using Windows XP with Intel Pentium M processor, 1.6 GHz, and 1 Gb of RAM memory. The processing time for PPCLUST is consistently less than 1 min for each run of data sets with 4000 random variables and PAM is the only faster algorithm. MCLUST, the closer competitor to PPCLUST, showed processing times at least 3 times higher than PPCLUST.

5. Application

Clustering of genes using expression data can identify genes that are differentially expressed under different conditions. Such genes may be responsible for disease progression or responsive to treatment. Identification of such genes can aid in biomarker identification for drug development. Additionally, the differentially expressed genes can be used to classify

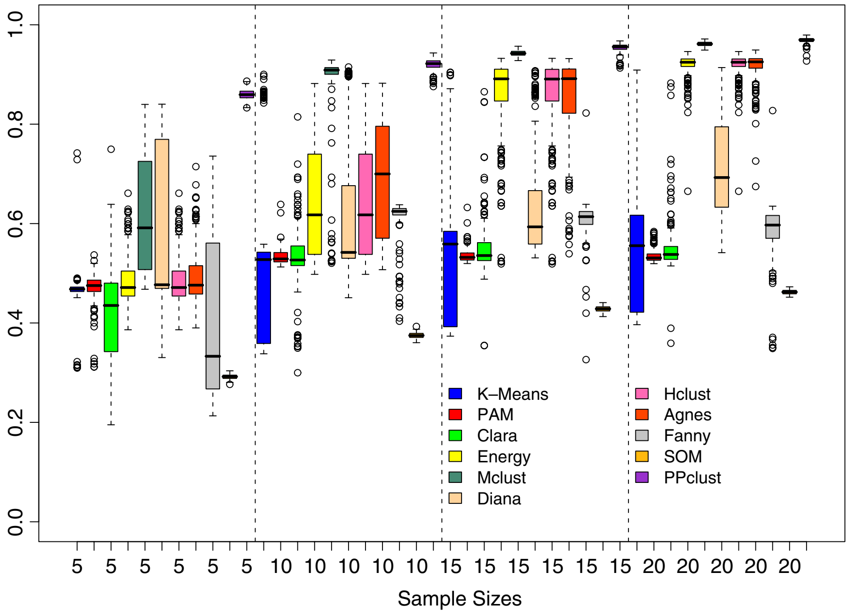

Fig. 2. Boxplots of adjusted Rand index for PPCLUST and 10 other algorithms on symmetric data based on 200 simulated datasets with different sample sizes (Study I).

Table 5
Mean and standard deviations (std) of adjusted Rand index for all algorithms over 200 simulated datasets with skewed distribution. Different sample sizes are considered. The data are generated from heavily skewed distributions as described in Study II.

Adjusted Rand index								
	Sample Sizes							
	5		10		15		20	
Algorithm	Mean	Std.	Mean	Std.	Mean	Std.	Mean	Std.
PPCLUST	**0.8600**	**0.0104**	**0.9205**	0.0110	**0.9541**	**0.0084**	**0.9688**	0.0059
PAM	**0.5396**	0.0257	**0.6073**	0.0218	**0.6347**	0.0178	**0.6454**	0.0377
K-means	0.1321	0.0192	0.2670	0.1524	0.5194	0.1164	0.5700	0.1410
Energy	0.4923	0.0370	0.5905	0.0184	0.6213	0.0114	0.6318	0.0097
Mclust	0.3495	0.0547	0.4126	0.0372	0.4656	0.0813	0.5028	0.1040
Clara	0.3860	0.1515	0.3696	0.0213	0.3256	0.2134	0.3390	0.2171
Diana	0.0130	0.0079	0.0066	**0.0037**	0.0045	**0.0019**	0.0039	**0.0013**
HCLUST	0.4923	0.0370	0.5905	0.0184	0.6211	0.0114	0.6318	0.0097
Agnes	0.1420	0.0320	0.1553	0.0181	0.1580	0.0052	0.1579	**0.0036**
Fanny	0.2990	0.0250	0.4236	0.0375	0.4716	0.0457	0.5221	0.0235
SOM	0.4766	**0.0112**	0.5412	**0.0083**	0.5775	0.0087	0.5999	0.0060

patient disease status. For example, using all genes from the whole genome can lead to inefficiency in classifying tumor patients as no inference can deal with high dimensional prediction without imposing strong assumptions. Instead, using only the genes found to be differentially expressed from the clustering algorithm can significantly reduce the complexity of the classification problem. That is, results from the clustering can serve as a dimension reduction tool for classification. These studies would allow to improve treatments by identification of targets for therapy in many diseases.

In this section, we apply PPCLUST to data from Notterman et al. (2001) study about transcriptional gene expression profiles of colorectal cancer. Heatmaps are used to visualize the results of PPCLUST.

Clustering genes in colorectal cancer

Colon and rectal cancer have many features in common and for this reason both are often referred to as colorectal cancer. This cancer begins in most cases as a growth of tissue, called polyp, inside the wall of the colon or rectum. If the cells of a tumor (adenomas) acquire the ability to invade and spread into the intestine and other areas, a malignant tumor develops (carcinoma or adenocarcinoma). Understanding how change in DNA causes cells of the colon and rectum to become cancerous could guide scientists in the development of new drugs, treatments and actions during early stages of the disease.

In Notterman et al. (2001) study, normal tissues were paired with the two types of tumors, adenoma and adenocarcinoma. The data [1] consist of mRNA expression patterns probed in 4 colon adenoma tissues, 18 adenocarcinoma and 22 paired normal

[1] Available in microarray.princeton.edu/oncology.

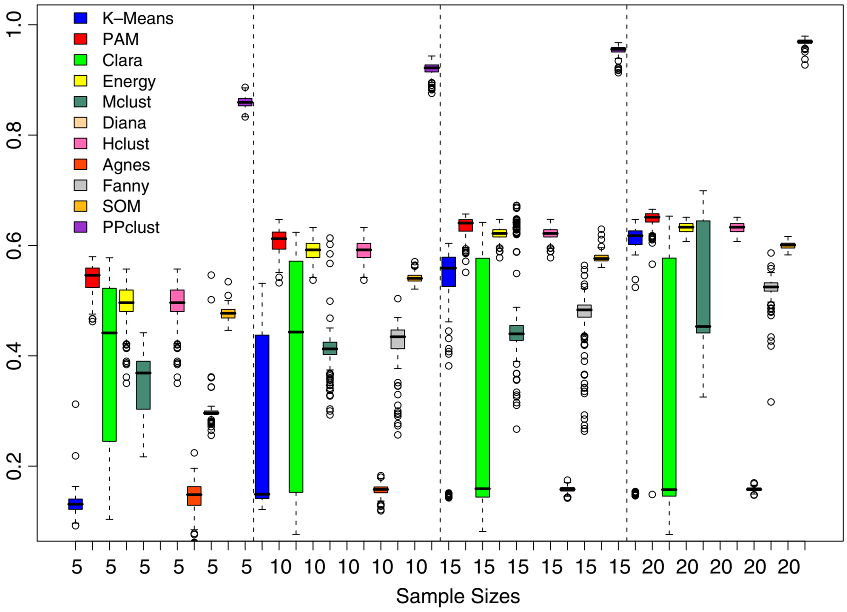

Fig. 3. Boxplots of adjusted Rand index for PPCLUST and 10 other algorithms on heavily skewed data (Study II) based on 200 simulated datasets under different sample sizes.

colon samples. In their study, a two-way hierarchical clustering algorithm was used to show that genome-wide expression profiling may permit a molecular classification of the three different types of tissues. Here instead of clustering on the tissues, we apply PPCLUST to cluster genes.

Since some of the genes in the original data were observed more than once, the median of expression levels of duplicated genes in each database (adenoma and paired normal tissues database, and adenocarcinoma and paired normal tissues database) was calculated. Then similar transformations as described in Notterman et al.'s (2001) study were performed prior to the application of PPCLUST in the composite database, i.e., the following steps were applied to each dataset:

- Deletion of expression levels ≤ 0;
- Calculation of the logarithm of the expression level;
- Deletion of genes having more than 25% of their values missing. In Notterman the percentage cutoff was 15% resulting in a smaller sample.

Two data sets are obtained, one with 4 adenoma and paired normal tissues for 4175 genes and the other one with 18 adenocarcinomas and paired normal tissues for 4234 genes. Only 1038 genes are common to both data sets. The existence of paired data allows the application of PPCLUST to the difference in gene expression levels of cancer (adenoma or adenocarcinoma) and normal tissues. The idea is that genes not related to the disease should not have significant changes in expression levels for cancer and normal tissues. However, genes that have significant changes in expression level can be identified through a clustering algorithm.

PPCLUST with a few significance levels is applied to the data. For significance levels greater than 1×10^{-10}, the clustering resulted in too many small clusters of genes; for significance levels much smaller than 1×10^{-10}, the main structure of groups obtained stays stable. So we use 1×10^{-10} as our significance level. Fig. 4 presents the heatmap of original differences in expression levels of adenoma and normal tissues in paired samples and the heatmap with genes ordered by the groups to which they were allocated. There is a concentration of zero to positive expression levels for this data with no clear existence of any gene groups. After applying PPCLUST, genes were clustered into 6 groups with 38 (0.91%), 316 (7.57%), 9 (0.22%), 3573 (85.58%), 221 (5.29%), and 15 (0.36%) genes, respectively. The first three groups contains genes that are significantly down regulated and the last two groups consist of genes that are significantly up regulated for adenomas compared to normal tissues. There is also a set of 3 (0.07%) genes that cannot be clustered with any other gene. The largest group is formed mostly by genes that had no significant difference in their expression levels between adenomas and normal tissues.

We also applied PPCLUST to the difference in expression levels of adenocarcinoma and normal tissues. In this case, 7 groups are obtained with only 4 (0.09%) genes not assigned to any group. The number of genes in each group are 91 (2.15%), 774 (18.28%), 9 (0.21%), 2673 (63.13%), 5 (0.12%), 655 (15.47), and 23 (0.54%). The heatmaps before and after clustering are given in Fig. 5.

Among the 1038 genes that are present in both data sets, the membership assignment for comparing adenoma versus normal and adenocarcinoma versus normal tissues are tabulated in Table 6. Among these genes, 558 of them had no significant change in expression for both adenoma and adenocarcinoma tissues. For the other genes that are not significantly differentially expressed in adenoma tissues, usually, there is no significant change of expression levels in carcinoma tissues. Similarly, genes that are significantly down regulated in adenoma tissues tend to be also down regulated in carcinoma

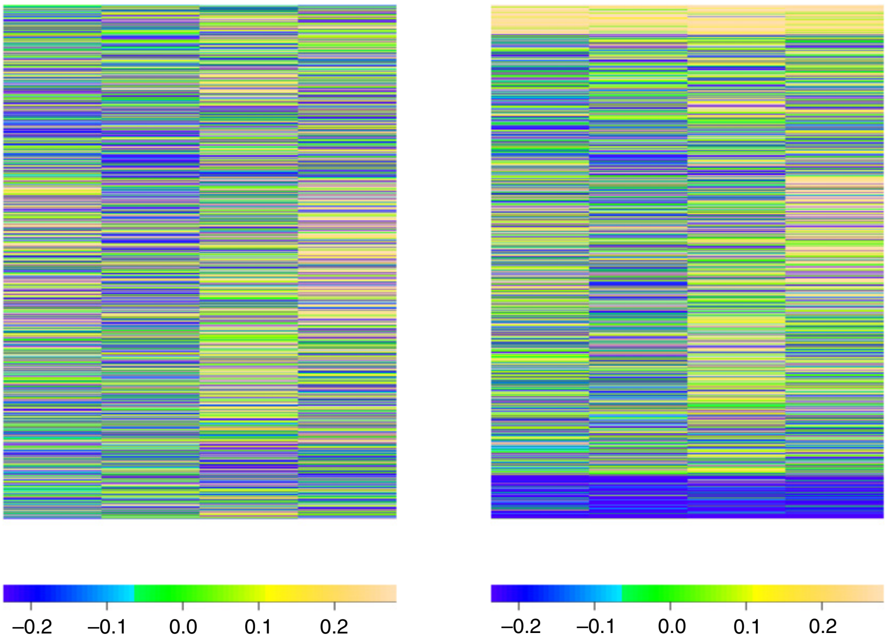

Fig. 4. Heatmaps for Adenoma–Normal Tissues before and after grouping.

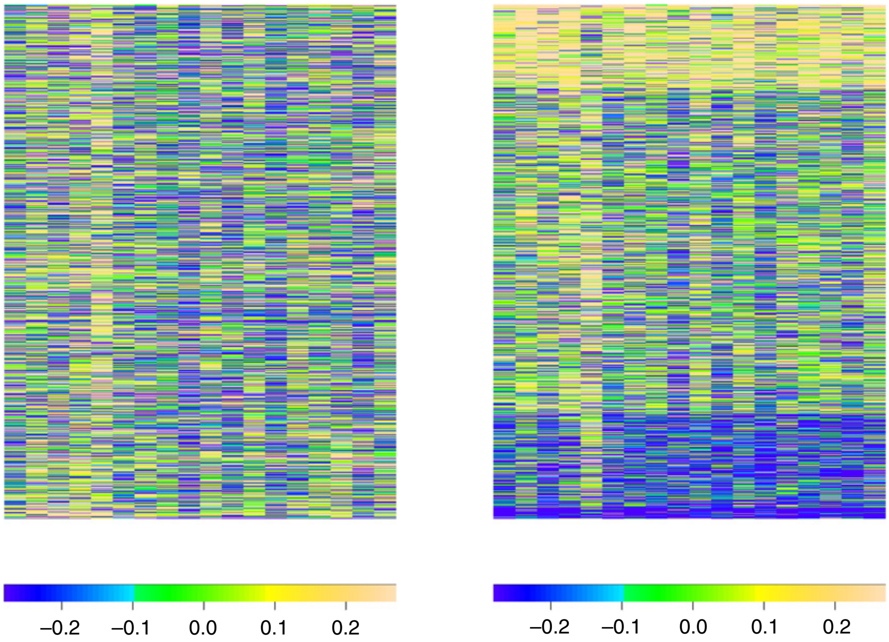

Fig. 5. Heatmaps for Adenocarcinoma–Normal Tissues before and after clustering.

Table 6
Distribution of 1038 genes present in both adenoma and adenocarcinoma tissue types. Genes in group 0 are not grouped by PPCLUST, and genes in group 4 are those genes that are differentially expressed in neither tissue type.

Adenoma groups	Adenocarcinoma groups							
	0	1	2	3	4	5	6	7
0	0	0	1	0	1	0	0	0
1	0	2	4	0	3	0	0	0
2	0	10	41	0	38	0	4	0
4	0	12	148	2	558	1	152	3
5	1	0	6	0	24	0	24	2
6	0	0	0	0	0	0	1	0

tissues. The same pattern is also observed for significantly up-regulated genes. Only 10 genes had opposite expression levels in both types of tissues.

Comparing the heatmaps obtained before and after clustering in both tissue types reveals that in carcinoma tissues the clustering of genes is more evident than in adenoma tissues. This is due to the larger differences in expression levels of carcinoma related genes. Results from the clustering of the gene expression data in the colorectal cancer study suggest target genes to molecular biologists for further lab experiments.

6. Conclusion

In this article, we proposed a novel computational algorithm, PPCLUST, for effectively clustering a large number of random variables with small number of replications per variable. The availability of replications allows us to use p-values from a (mid-)rank test of homogeneous distribution developed by Wang and Akritas (2004) as similarity measures to determine if a group need to be partitioned. Since no optimization is necessary, the computational cost is dramatically reduced compared to commonly used algorithms applied to a large number of variables. In addition, PPCLUST has the advantage that it is invariant to monotone transformations of data and can automatically determine the number of clusters with a specified significance level. In our simulation studies, PPCLUST outperformed 10 other benchmark algorithms commonly used in the microarray literature when considering clustering accuracy, stability and speed. The superior performance of PPCLUST on high dimensional data with small sample sizes make it a useful tool in such data that arise from many disciplines.

Acknowledgements

We are grateful to the two referees and the Editor for their helpful comments that improved the presentation of this manuscript. We would also like to acknowledge SAS Institute Brazil for the use of SAS through academic agreement with University of Brasilia.

References

Akritas, M.G., Arnold, S., 2000. Asymptotics for analysis of variance when the number of levels is large. Journal of The American Statistical Association 95, 212–226.

Akritas, M.G., Papadatos, N., 2004. Heteroscedastic one-way ANOVA and lack-of-fit tests. Journal of The American Statistical Association 99, 368–382.

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

Benjamini, Y., Hochberg, Y., 1995. Controlling the false discovery rate: A practical and powerful approach to multiple testing. JRSSB 57, 289–300.

Bradley, P.S., Fayyad, U.M., 1998. Refining initial points for K-means clustering. In: Proceedings of the Fifteenth International Conference on Machine Learning. Morgan kaufmann publishers, Inc., San Francisco, CA, pp. 91–99.

Efron, B., 2007. Correlation and large-scale simultaneous significance testing. Journal of the American Statistical Association 102, 93–103.

Efron, B., Tibshirani, R., 2007. On testing the significance of sets of genes. Annals of Applied Statistics 1, 107–129.

Fraley, C., 1998. Algorithms for model-based Gaussian hierarchical clustering. SIAM 20.

Fraley, C., Raftery, A.E., 2006. MCLUST version 3.0: An R package for normal mixture modeling and model-based clustering, Technical Report, University of Washington.

Fu, L., Medico, E., 2007. Flame, a novel fuzzy clustering method for the analysis of DNA microarray data. BMC Bioinformatics 8.

Hartigan, J.A., Wong, M.A., 1979. A K-means clustering algorithm. Applied Statistics 28, 100–108.

Hubert, L., Arabie, P., 1985. Comparing partitions. Journal of Classification 2, 193–218.

Huttenhower, C., Flamholz, A.I., Landis, J.N., Sahi, S., Myers, C.L., Olszewski, K.L., Hibbs, M.A., Siemens, N.O., Troyanskaya, O.G., Coller, H.A., 2007. Nearest neighbor networks: Clustering expression data based on gene neighborhoods. BMC Bioinformatics 8.

Jiang, D., Tang, C., Zhang, A., 2004. Cluster analysis for gene expression data: A survey. IEEE Transactions on Knowledge and Data Engineering 16, 1370–1386.

Johnson, D.E., 1998. Applied Multivariate Methods for Data Analysis. Duxbury.

Kaufman, L., Rousseeuw, P.J., 1990. Finding Groups in Data: An Introduction to Cluster Analysis. Wiley Interscience.

Kohonen, T., 1989. Self-organization and Associative Memory. Springer.

McQueen, J.B., 1967. Some methods for classification and analysis of multivariate observations. In: Proceedings of Fifth Berkeley Symposium on Mathematical Statistics and Probability.

Milligan, G.W., Cooper, M.C., 1986. A study of the comparability of external criteria for hierarchical cluster analysis. Multivariate Behavioral Research 21, 441–458.

Notterman, D.A., Alon, U., Sierk, A.J., Levine, A.J., 2001. Transcriptional gene expression profiles of colorectal adenoma, adenocarcinoma, and normal tissue examined by oligonucleotide arrays. Cancer Research 61, 3124–3130.

Pawitan, Y., 2001. In all likelihood: Statistical modeling and inference using likelihood, Oxford.

Pena, J.M., Lozano, J.A., Larranaga, P., 1999. An empirical comparison of four initialization methods for the K-Means algorithm. Pattern Recognition Letters 20, 1027–1040.

Qiu, X., Yakovlev, A., 2006. Some comments on instability of false discovery rate estimation. Journal of Bioinformatics and Computational Biology 4, 2006.

Rand, W.M., 1971. Objective criteria for the evaluation of clustering methods. JASA 36, 846–850.

Sabatti, C., 2006. False discovery rate and multiple comparison procedures. In: DNA Microarrays and Related Genomics Techniques: Design, Analysis, and Interpretation of Experiments. Chapman & Hall/CRC, pp. 289–304.

Storey, J., 2002. A direct approach to false discovery rates. Journal of the Royal Statistical Society B 64 (3), 479–498.

Storey, J.D., Tibshirani, R., 2003. Statistical significance for genomewide studies. Proceedings of the National Academy of Sciences USA 16, 9440–9445.

Strimmer, K., 2008. A unified approach to false discovery rate estimation. BMC Bioinformatics 9, 303.

Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., Mesirov, J.P., 2005. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences USA 43, 15545–15550.

Székely, G.J., Rizzo, M.L., 2005. Hierarchical clustering via joint between-within distances: Extending Ward's minimum variance method. Journal of Classification 22, 151–183.

Wang, H., Akritas, M.G., 2004. Rank tests for ANOVA with large number of factor levels. Journal of Nonparametric Statistics 16, 563–589.

Yeung, K.Y., Ruzzo, W.L., 2001. Principal component analysis for clustering gene expression data. Bioinformatics 9, 763–774.