Journal of Inequalities and Applications
a SpringerOpen Journal

RESEARCH Open Access

Comparison of hierarchical cluster analysis methods by cophenetic correlation

Sinan Saraçli[1*], Nurhan Doğan[2] and İsmet Doğan[2]

*Correspondence:
ssaracli@aku.edu.tr
[1]Department of Statistics, Faculty of Arts and Sciences, Afyon Kocatepe University, Afyonkarahisar, 03200, Turkey
Full list of author information is available at the end of the article

Abstract

Purpose: This study proposes the best clustering method(s) for different distance measures under two different conditions using the cophenetic correlation coefficient.

Methods: In the first one, the data has multivariate standard normal distribution without outliers for $n = 10, 50, 100$ and the second one is with outliers (5%) for $n = 10, 50, 100$. The proposed method is applied to simulated multivariate normal data *via* MATLAB software.

Results: According the results of simulation the Average (especially for $n = 10$) and Centroid (especially for $n = 50$ and $n = 100$) methods are recommended at both conditions.

Conclusions: This study hopes to contribute to literature for making better decisions on selection of appropriate cluster methods by using subgroup sizes, variable numbers, subgroup means and variances.

Keywords: cophenetic correlation; hierarchical clustering methods; distance measures

1 Introduction

Classification, in its widest sense, has to do with forms of the relatedness and with the organization and display of the relations in a useful manner. The items to be studied could be anything: people, bacteria, religions, books, *etc.* The attributes in each case would be those features of the items that are of interest for the purpose of the study [1]. Classifications are generally pictured in the form of hierarchical trees, also called a dendrogram. A dendrogram is the graphical representation of an ultrametric (= cophenetic) matrix; so dendrograms can be compared to one another by comparing their cophenetic matrices [2].

Cluster Analysis (CA), Principal Components Analysis (PCA) and Discriminant Analysis (DA) are three of the primary methods of modern multivariate analysis. Because of its utility, clustering has emerged as one of the leading methods of multivariate analysis [3].

Cluster analysis is a multivariate statistical technique which was originally developed for biological classification. Biologists Robert Soka1 and Peter Sneath published their seminal text '*Principles of Numerical Taxonomy*' in 1963. Sokal and Sneath demonstrated that cluster analysis could be utilized to efficiently classification a data set which contained all relevant characteristics of an organism. When the organisms had been classified based on these characteristics, it could be determined in which way they differed, and if they belonged to different species. In this way, Sokal and Sneath asserted, researchers could trace the path of evolution from one species to another [4].

In this study for clustering, two measures of cluster 'goodness' or quality are used. One type of measure allows us to compare different sets of clusters without reference to external knowledge and is called an internal quality which is used as a measure of 'overall similarity' based on the pairwise similarity of documents in a cluster. The other type of measures allows evaluating how well the clustering is working by comparing the groups produced by the clustering techniques to known classes. This type of measure is called an external quality measure, which is not scope of this study [5].

The joining or tree clustering method uses the dissimilarities (similarities) or distances (Euclidean distance, squared Euclidean distance, city-block (Manhattan) distance, Chebychev distance, power distance, Mahalanobis distance, *etc.*) between objects when forming the clusters. Similarities are a set of rules that serve as criteria for grouping or separating items. These distances (similarities) can be based on a single dimension or multiple dimensions, with each dimension representing a rule or condition for grouping objects. The joining algorithm does not 'care' whether the distances that are 'fed' to it are actual real distances, or some other derived measure of distance that is more meaningful to the researcher; and it is up to the researcher to select the right method for his/her specific application [6].

The next step is to identify how one can find the natural clusters among items characterized by many attributes. A number of cluster analysis procedures (single linkage (nearest neighbor), Complete linkage (furthest neighbor), Unweighted pair-group average (UPGMA), Weighted pair-group average (WPGMA), Unweighted pair-group centroid (UPGMC), Weighted pair-group centroid (median), Ward's method, *etc.*) are available; many of these begin with an n-dimensional space in which each entity is represented by a single point. The dimensions in the space represent the characteristics upon which the entities are to be compared. Similarity between entities can be measured by: (1) the correlation of entities' scores on the dimensions (cophenetic correlation) or (2) the distance between points in the space (points closest to each other are most similar) [7, 8].

Suppose that the original data $\{X_i\}$ have been modeled using a cluster method to produce a dendrogram $\{T_i\}$; that is, a simplified model in which data that are 'close' have been grouped into a hierarchical tree. Define the following distance measures. $x(i,j) = |X_i - X_j|$, the ordinary Euclidean distance between the ith and jth observations. $t(i,j)$ = the dendrogrammatic distance between the model points T_i and T_j. This distance is the height of the node at which these two points are first joined together. Then, letting x be the average of the $x(i,j)$, and letting t be the average of the $t(i,j)$, the cophenetic correlation coefficient c is defined as in (1) [9].

$$c = \frac{\sum_{i<j}(x(i,j) - x)(t(i,j) - t)}{\sqrt{[\sum_{i<j}(x(i,j) - x)^2][\sum_{i<j}(t(i,j) - t)^2]}}. \tag{1}$$

Since its introduction by Sokal and Rohlf [10], the cophenetic correlation coefficient has been widely used in numerical phenetic studies, both as a measure of degree of fit of a classification to a set of data and as a criterion for evaluating the efficiency of various clustering techniques [11]. In statistics, and especially in biostatistics, cophenetic correlation (more precisely, the cophenetic correlation coefficient) is a measure of how faithfully a dendrogram preserves the pairwise distances between the original unmodeled data points. Although it has been most widely applied in the field of biostatistics (typically to

assess cluster-based models of DNA sequences, or other taxonomic models), it can also be used in other fields of inquiry where raw data tend to occur in clumps, or clusters. This coefficient has also been proposed for use as a test for nested clusters [12].

The problem of comparing classifications with numerical methods is not new; the first effective numerical method known to us is the 'cophenetic correlation' technique of Sokal and Rohlf [10]. Beginning with the development of cophenetic correlations methods for comparison of dendrograms have recently been the object of strong interest. Baker [13] investigated the impact of observational errors on the dendrograms produced by the complete linkage and single linkage hierarchical grouping techniques. The goodness of fit of the dendrograms was measured by means of the Goodman-Kruskal gamma coefficient. The gamma coefficients indicated that the single linkage grouping technique was more sensitive to the type of data errors employed than the complete linkage technique. Hubert [14] compared two rank orderings of the object pairs. He tested hypothesis that the given set of proximity values have been assigned randomly by referring the Goodman-Kruskal rank correlation γ statistic to an approximate permutation distribution. Kuiper and Fisher [15] compared six hierarchical clustering procedures (single linkage, complete linkage, median, average linkage, centroid and Ward's method) for multivariate normal data, assuming that the true number of clusters was known. The authors used the Rand index, which gives a proportion of correct groupings, to compare the clustering methods. In their study for clusters of equal sizes, Ward's method and complete linkage method, with very unequal cluster sizes centroid and average linkage method found best, respectively. Blashfield [16] compared four types of hierarchical clustering methods (single linkage, complete linkage, average linkage and Ward's method) for accuracy in recovery of original population clusters. He used Cohen's statistic to measure the accuracy of the clustering methods. According to his results, Ward's method performed significantly better than the other clustering procedures and average linkage gave relatively poor results. According to Milligan [17], complete linkage and Ward's method reacted badly when outliers were introduced into the simulated data.

Hands and Everitt [18] compared five hierarchical clustering techniques (single linkage, complete linkage, average, centroid, and Ward's method) on multivariate binary data. They found that Ward's method was the best overall than other hierarchical methods. Yao [19] discussed six classical clustering algorithms: k-means, SOM, EM-based clustering, classification EM clustering, fuzzy k-means, leader clustering and different combination scenarios of these algorithms. He used a count of cluster categories, classification accuracy and cluster entropy. Ferreira and Hitchcock [20] compared the performance of four major hierarchical methods (single linkage, complete linkage, average linkage and Ward's method) for clustering functional data. They used the Rand index to compare the performance of each clustering method. According to their study, Ward's method was usually the best, while average linkage performed best in some special situations, in particular, when the number of clusters is over specified. Milligan and Cooper [21] used four agglomerative hierarchical clustering methods to generate partition solutions and formed one factor in the overall design. These were the single link, complete link, group average (UPGMA) and Ward's minimum variance methods. As a result, they found that the single link technique was least effective while the group average and Ward's methods gave the best overall recovery.

Consider the studies in the literature and the importance of using the most convenient cluster method under different conditions (sample size, variables number and distance measures), a detailed simulation study is undertaken. This study gives more insight into the functioning of the cluster method under different conditions. The purpose of this research is to investigate the best clustering method under different conditions.

2 Method

In this study, seven cluster analysis methods are compared by the cophenetic correlation coefficient computed according to different clustering methods with a sample size ($n = 10$, $n = 50$ and $n = 100$), variables number ($x = 3$, $x = 5$ and $x = 10$) and distance measures *via* a simulation study. The simulation program is developed in a MATLAB software development environment by the authors. We have 567 different simulation scenarios and $100{,}000/n$ replications for each scenario. The performance is monitored by two different conditions that are mentioned in Table 1 and Table 2 with 7 cluster methods, 9 distance measures by cophenetic correlation coefficient in various settings of subgroup means, variances, sample size and variable numbers simultaneously.

For 567 different simulation scenarios, the data was derived from multivariate normal distribution for $\mu = 0$, $\delta^2 = 1$ with and without outliers, respectively. The data set for outliers is obtained according to Dixon's [22] 'Outlier Model' like $(N - r) \sim N(0,1) + r \sim N(0,5)$. In this study, $r = [0,5 + 0,1 * N]$ means that while 95% of the data set does not include any outliers, 5% of the data set includes outliers.

3 Results and discussion

All numerical results, obtained by running the simulation program, are given in Table 1 and Table 2. According to Table 1 and Table 2, the average method gives the best results at all measures and at all variable numbers for both distributions with sample size $n = 10$. Moreover, increasing the sample size to $n = 50$ and $n = 100$ favors the complete, weighted, and centroid methods for all measures. However, the cophenetic correlation coefficient for the Mahalanobis measure cannot be calculated in both distributions when there are 10 variables with sample size $n = 10$, whereas there is not any meaningful explanation for this unexpected result, we still could not find the main reason for this situation, but the same result is obtained for more than three times run of the simulation program.

4 Conclusion

In general, researchers especially nonstatisticians use cluster analysis methods and distance measures in different conditions. In addition, they choose to use the most famous cluster analysis methods and distance measures, which are available in statistical packages, without evaluating the validity of different conditions. When the different conditions are considered, drawn inferences are dubious, and may lead the decision-makers to incorrect decisions. It is noted that, with respect to the selection of a distance measures, the researcher must be aware that their choice can often significantly affect the results of the clustering. For example, some distance measures are inappropriate when different conditions of the variables are not met. On this point, the determination of the correct distance measures to use under various cases is the main motivation of researchers working on this subject to determine which distance measures should be used in case of different conditions.

Table 1 The cophenetic correlation coefficient values for $\mu = 0$, $\sigma^2 = 1$ (without outliers)

Distance measure	Clustering method	Cophenetic correlation coefficient								
		$x = 3$			$x = 5$			$x = 10$		
		$n = 10$	$n = 50$	$n = 100$	$n = 10$	$n = 50$	$n = 100$	$n = 10$	$n = 50$	$n = 100$
Euclidean	Average	***0.7552***	0.6358	0.6009	***0.7255***	0.6017	0.5753	***0.6922***	0.5728	0.5339
	Centroid	0.6927	***0.6393***	***0.6038***	0.6463	***0.6028***	***0.5557***	0.5829	***0.5759***	***0.5605***
	Complete	0.6858	0.4940	0.4313	0.6397	0.4135	0.3386	0.5920	0.3439	0.2640
	Median	0.7309	0.5600	0.4797	0.6897	0.5412	0.4605	0.6220	0.5338	0.4989
	Single	0.7034	0.6033	0.5947	0.6633	0.5663	0.5577	0.6157	0.5332	0.5295
	Ward	0.6832	0.4956	0.4422	0.6384	0.4124	0.3496	0.5895	0.3367	0.2664
	Weighted	0.7367	0.5428	0.4692	0.7064	0.5024	0.4105	0.6749	0.4791	0.3941
Squared Euclidean	Average	***0.7381***	0.6352	0.5978	***0.7090***	0.5974	0.5546	***0.6792***	0.5700	0.5315
	Centroid	0.7285	***0.6402***	***0.6056***	0.6874	***0.6085***	***0.5765***	0.6254	***0.5831***	***0.5632***
	Complete	0.6565	0.4826	0.4271	0.6128	0.4014	0.3325	0.5659	0.3318	0.2586
	Median	0.7102	0.5538	0.4745	0.6679	0.5386	0.4553	0.6013	0.5344	0.5039
	Single	0.6874	0.6038	0.5950	0.6453	0.5656	0.5578	0.6009	0.5324	0.5301
	Ward	0.6533	0.4830	0.4350	0.6093	0.3995	0.3421	0.5661	0.3264	0.2589
	Weighted	0.7191	0.5351	0.4615	0.6904	0.4978	0.4054	0.6623	0.4768	0.3969
Mahalanobis	Average	***0.6957***	0.6276	0.5950	***0.6176***	0.5865	0.5469	***NaN***	0.5591	0.5284
	Centroid	0.6749	***0.6325***	***0.6010***	0.5864	***0.5993***	***0.5723***	***NaN***	***0.5789***	***0.5630***
	Complete	0.5627	0.4578	0.4121	0.3958	0.3625	0.3097	***NaN***	0.2711	0.2270
	Median	0.6587	0.5426	0.4719	0.5762	0.5320	0.4538	***NaN***	0.5385	0.5106
	Single	0.6565	0.6032	0.5956	0.5642	0.5655	0.5589	***NaN***	0.5293	0.5317
	Ward	0.5541	0.4621	0.4235	0.3876	0.3539	0.3147	***NaN***	0.2495	0.2155
	Weighted	0.6760	0.5202	0.4549	0.6078	0.4843	0.3963	***NaN***	0.4869	0.4013
Cityblock	Average	***0.7427***	0.6228	0.5844	***0.7120***	0.5810	0.5349	***0.6787***	0.5484	0.5076
	Centroid	0.7372	***0.6280***	***0.5944***	0.6983	***0.5902***	***0.5557***	0.6343	***0.5579***	***0.5368***
	Complete	0.6716	0.4780	0.4229	0.6247	0.3976	0.3272	0.5771	0.3285	0.2502
	Median	0.7194	0.5460	0.4710	0.6799	0.5233	0.4455	0.6129	0.5117	0.4774
	Single	0.6876	0.5839	0.5756	0.6443	0.5410	0.5324	0.5975	0.5037	0.5009
	Ward	0.6728	0.4869	0.4348	0.6281	0.4052	0.3427	0.5823	0.3322	0.2596
	Weighted	0.7244	0.5290	0.4546	0.6936	0.4878	0.3969	0.6615	0.4601	0.3824
Minkowski	Average	***0.7552***	0.6393	0.6009	***0.7255***	0.6017	0.5557	***0.6922***	0.5728	0.5339
	Centroid	0.7497	***0.6447***	***0.6087***	0.7099	***0.6124***	***0.5791***	0.6450	***0.5833***	***0.5629***
	Complete	0.6858	0.4940	0.4313	0.6397	0.4135	0.3386	0.5920	0.3439	0.2640
	Median	0.7309	0.5600	0.4797	0.6897	0.5412	0.4605	0.6220	0.5338	0.4989
	Single	0.7034	0.6033	0.5947	0.6633	0.5663	0.5577	0.6157	0.5332	0.5295
	Ward	0.6832	0.4956	0.4422	0.6384	0.4124	0.3496	0.5895	0.3367	0.2664
	Weighted	0.7367	0.5428	0.4692	0.7064	0.5024	0.4105	0.6749	0.4791	0.3941
Cosine	Average	***0.7590***	***0.6277***	***0.5839***	***0.6994***	***0.5143***	***0.4524***	***0.6441***	***0.4152***	***0.3413***
	Centroid	0.7518	0.6045	0.5478	0.6866	0.4711	0.3894	0.6097	0.3242	0.2312
	Complete	0.7230	0.5782	0.5320	0.6501	0.4518	0.3929	0.5808	0.3428	0.2700
	Median	0.7340	0.5530	0.5040	0.6681	0.4273	0.3546	0.5911	0.3004	0.2146
	Single	0.6931	0.4695	0.3898	0.6034	0.3070	0.2202	0.5211	0.2017	0.1275
	Ward	0.7381	0.6142	0.5730	0.6716	0.4979	0.4418	0.6083	0.3974	0.3280
	Weighted	0.7433	0.5786	0.5336	0.6854	0.4711	0.4093	0.6311	0.3834	0.3069
Correlation	Average	***0.8217***	***0.7470***	***0.7226***	***0.7229***	***0.5581***	***0.5065***	***0.6507***	***0.4268***	***0.3542***
	Centroid	0.8169	0.7358	0.7037	0.7124	0.5245	0.4594	0.6212	0.3441	0.2518
	Complete	0.7979	0.7083	0.6775	0.6791	0.4987	0.4514	0.5893	0.3545	0.2855
	Median	0.7982	0.6886	0.6621	0.6952	0.4754	0.4128	0.6028	0.3177	0.2324
	Single	0.7939	0.6715	0.6285	0.6389	0.3624	0.2781	0.5307	0.2124	0.1355
	Ward	0.8069	0.7388	0.7155	0.6974	0.5429	0.4938	0.6170	0.4099	0.3426
	Weighted	0.8052	0.7061	0.6816	0.7075	0.5105	0.4534	0.6379	0.3916	0.3185
Spearman	Average	***0.8207***	0.7600	0.7441	***0.7240***	***0.5636***	***0.5163***	***0.6487***	***0.4274***	***0.3567***
	Centroid	0.8116	0.7413	0.7199	0.7132	0.5336	0.4715	0.6184	0.3452	0.2511
	Complete	0.8094	***0.7788***	***0.7780***	0.6815	0.5337	0.4966	0.5884	0.3572	0.2896
	Median	0.7903	0.7162	0.7114	0.6958	0.4818	0.4235	0.5990	0.3175	0.2329
	Single	0.6854	0.6737	0.6738	0.6397	0.4010	0.1830	0.5287	0.2198	0.1427
	Ward	0.7892	0.7358	0.7275	0.6973	0.5488	0.5030	0.6145	0.4093	0.3434
	Weighted	0.8118	***0.7788***	***0.7780***	0.7086	0.5154	0.4672	0.6357	0.3916	0.3194

Table 1 (*Continued*)

Distance measure	Clustering method	Cophenetic correlation coefficient								
		$x = 3$			$x = 5$			$x = 10$		
		$n = 10$	$n = 50$	$n = 100$	$n = 10$	$n = 50$	$n = 100$	$n = 10$	$n = 50$	$n = 100$
Chebychev	Average	***0.7375***	0.6183	0.5804	***0.6933***	0.5595	0.5141	***0.6448***	0.4958	0.4523
	Centroid	0.7317	***0.6241***	***0.5870***	0.6811	***0.5693***	***0.5334***	0.6100	***0.5067***	***0.4805***
	Complete	0.6647	0.4780	0.4235	0.6035	0.3824	0.3164	0.5423	0.2962	0.2281
	Median	0.7140	0.5431	0.4630	0.6625	0.5036	0.4287	0.5928	0.4628	0.4249
	Single	0.6833	0.5792	0.5695	0.6223	0.5199	0.5084	0.5536	0.4468	0.4405
	Ward	0.6680	0.4852	0.4317	0.6128	0.3949	0.3341	0.5595	0.3140	0.2423
	Weighted	0.7189	0.5255	0.4494	0.6759	0.4734	0.3878	0.6294	0.4249	0.3525

Table 2 The cophenetic correlation coefficient values for $\mu = 0$, $\sigma^2 = 1$ (with outliers)

Distance measure	Clustering method	Cophenetic correlation coefficient								
		$x = 3$			$x = 5$			$x = 10$		
		$n = 10$	$n = 50$	$n = 100$	$n = 10$	$n = 50$	$n = 100$	$n = 10$	$n = 50$	$n = 100$
Euclidean	Average	***0.8478***	***0.8065***	0.7848	***0.8280***	***0.7818***	0.7629	***0.8102***	***0.7647***	0.7488
	Centroid	0.8188	0.8061	***0.7875***	0.7872	0.7816	***0.7704***	0.7484	0.7638	***0.7606***
	Complete	0.8095	0.7273	0.7006	0.7808	0.6865	0.6551	0.7535	0.6494	0.6136
	Median	0.8342	0.7644	0.7262	0.8073	0.7509	0.7182	0.7661	0.7432	0.7311
	Single	0.8168	0.7836	0.7774	0.7903	0.7582	0.7578	0.7653	0.7400	0.7426
	Ward	0.8064	0.7278	0.7050	0.7801	0.6869	0.6606	0.7531	0.6464	0.6161
	Weighted	0.8382	0.7551	0.7185	0.8182	0.7352	0.6918	0.8006	0.7197	0.6835
Squared Euclidean	Average	***0.8434***	0.8088	0.7859	***0.8239***	0.7837	0.7636	***0.8087***	0.7663	0.7490
	Centroid	0.8386	***0.8107***	***0.7900***	0.8123	***0.7880***	***0.7730***	0.7768	***0.7695***	***0.7629***
	Complete	0.8022	0.7331	0.7027	0.7724	0.6937	0.6580	0.7505	0.6550	0.6166
	Median	0.8289	0.7652	0.7275	0.8017	0.7525	0.7177	0.7637	0.7417	0.7313
	Single	0.8142	0.7838	0.7774	0.7865	0.7584	0.7575	0.7633	0.7397	0.7424
	Ward	0.7996	0.7337	0.7070	0.7725	0.6929	0.6645	0.7506	0.6532	0.6191
	Weighted	0.8333	0.7592	0.7197	0.8141	0.7358	0.6956	0.7997	0.7192	0.6824
Mahalanobis	Average	***0.8103***	***0.8565***	0.8315	***0.6965***	0.8239	0.8053	***NaN***	0.7705	0.7782
	Centroid	0.7976	0.8570	***0.8333***	0.6701	***0.8276***	***0.8113***	***NaN***	***0.7770***	***0.7882***
	Complete	0.6966	0.8051	0.7787	0.4529	0.7575	0.7380	***NaN***	0.6480	0.6848
	Median	0.7895	0.8219	0.7843	0.6695	0.7875	0.7600	***NaN***	0.7453	0.7309
	Single	0.7908	0.8220	0.8030	0.6633	0.7841	0.7680	***NaN***	0.7510	0.7515
	Ward	0.7313	0.8018	0.7755	0.5497	0.7523	0.7350	***NaN***	0.6442	0.6839
	Weighted	0.7980	0.8181	0.7848	0.6899	0.7824	0.7531	***NaN***	0.7230	0.7173
Cityblock	Average	***0.8404***	0.7982	0.7767	***0.8206***	0.7707	0.7518	***0.8027***	0.7522	0.7352
	Centroid	0.8367	***0.7997***	***0.7805***	0.8113	***0.7741***	***0.7611***	0.7739	***0.7552***	***0.7482***
	Complete	0.7995	0.7226	0.6935	0.7727	0.6761	0.6484	0.7464	0.6418	0.6059
	Median	0.8267	0.7566	0.7196	0.8018	0.7412	0.7074	0.7623	0.7306	0.7174
	Single	0.8077	0.7737	0.7676	0.7804	0.7448	0.7447	0.7533	0.7248	0.7269
	Ward	0.8004	0.7243	0.7003	0.7744	0.6817	0.6568	0.7493	0.6446	0.6131
	Weighted	0.8305	0.7507	0.7114	0.8111	0.7241	0.6867	0.7936	0.7102	0.6718
Minkowski	Average	***0.8478***	0.8065	0.7848	***0.8280***	0.7818	0.7629	***0.8102***	0.7647	0.7488
	Centroid	0.8441	***0.8088***	***0.7883***	0.8179	***0.7860***	***0.7721***	0.7797	***0.7695***	***0.7628***
	Complete	0.8095	0.7273	0.7006	0.7808	0.6865	0.6551	0.7535	0.6494	0.6136
	Median	0.8342	0.7644	0.7262	0.8073	0.7509	0.7182	0.7661	0.7432	0.7311
	Single	0.8168	0.7836	0.7774	0.7903	0.7582	0.7578	0.7653	0.7400	0.7426
	Ward	0.8064	0.7278	0.7050	0.7801	0.6869	0.6606	0.7531	0.6464	0.6161
	Weighted	0.8382	0.7551	0.7185	0.8182	0.7352	0.6918	0.8006	0.7197	0.6835
Cosine	Average	***0.7689***	***0.6484***	***0.6138***	***0.7107***	***0.5463***	***0.4946***	***0.6596***	***0.4549***	***0.3908***
	Centroid	0.7618	0.6285	0.5866	0.6982	0.5106	0.4462	0.6270	0.3717	0.2936
	Complete	0.7320	0.5960	0.5568	0.6604	0.4773	0.4226	0.5977	0.3743	0.3064
	Median	0.7438	0.5702	0.5265	0.6802	0.4521	0.3862	0.6058	0.3286	0.2517
	Single	0.7082	0.4942	0.4287	0.6182	0.3484	0.2651	0.5429	0.2451	0.1693
	Ward	0.7467	0.6325	0.5991	0.6808	0.5240	0.4733	0.6213	0.4232	0.3521
	Weighted	0.7534	0.5961	0.5531	0.6961	0.4958	0.4296	0.6454	0.4104	0.3378

Table 2 (*Continued*)

Distance measure	Clustering method	Cophenetic correlation coefficient								
		$x = 3$			$x = 5$			$x = 10$		
		$n = 10$	$n = 50$	$n = 100$	$n = 10$	$n = 50$	$n = 100$	$n = 10$	$n = 50$	$n = 100$
Correlation	Average	***0.8214***	***0.7466***	***0.7239***	***0.7213***	***0.5584***	***0.5060***	***0.6508***	***0.4269***	***0.3550***
	Centroid	0.8171	0.7330	0.7064	0.7117	0.5248	0.4587	0.6210	0.3434	0.2560
	Complete	0.7978	0.7102	0.6794	0.6770	0.4992	0.4457	0.5887	0.3561	0.2874
	Median	0.7984	0.6852	0.6596	0.6935	0.4756	0.4128	0.6027	0.3150	0.2328
	Single	0.7931	0.6647	0.6319	0.6374	0.3624	0.2748	0.5293	0.2157	0.1354
	Ward	0.8069	0.7378	0.7162	0.6955	0.5431	0.4940	0.6161	0.4096	0.3428
	Weighted	0.8052	0.7017	0.6843	0.7059	0.5106	0.4557	0.6373	0.3914	0.3191
Spearman	Average	***0.8198***	0.7583	0.7455	***0.7233***	***0.5638***	***0.5159***	***0.6505***	***0.4267***	***0.3567***
	Centroid	0.8113	0.7396	0.7212	0.7133	0.5336	0.4729	0.6199	0.3429	0.2537
	Complete	0.8090	***0.7788***	***0.7762***	0.6802	0.5341	0.4983	0.5909	0.3558	0.2887
	Median	0.7887	0.7136	0.7140	0.6960	0.4819	0.4238	0.6021	0.3126	0.2304
	Single	0.6861	0.6736	0.6742	0.6404	0.4008	0.1821	0.5302	0.2190	0.1404
	Ward	0.7881	0.7364	0.7273	0.6963	0.5488	0.5037	0.6170	0.4081	0.3425
	Weighted	0.8112	***0.7788***	***0.7762***	0.7076	0.5157	0.4687	0.6366	0.3926	0.3197
Chebychev	Average	***0.8373***	0.7945	0.7740	***0.8094***	0.7588	0.7398	***0.7824***	0.7246	0.7059
	Centroid	0.8338	***0.7966***	***0.7774***	0.8008	***0.7627***	***0.7483***	0.7601	***0.7280***	***0.7181***
	Complete	0.7965	0.7182	0.6953	0.7581	0.6707	0.6424	0.7262	0.6226	0.5951
	Median	0.8244	0.7546	0.7195	0.7913	0.7315	0.6996	0.7510	0.7061	0.6907
	Single	0.8044	0.7700	0.7640	0.7663	0.7329	0.7324	0.7289	0.6940	0.6951
	Ward	0.7978	0.7216	0.6996	0.7647	0.6761	0.6495	0.7351	0.6345	0.6038
	Weighted	0.8279	0.7470	0.7135	0.8007	0.7171	0.6783	0.7745	0.6895	0.6557

One may conclude that the results of this study, which is similar to findings of Johnson and Wichern [23], indicate the data set with outliers have higher cophenetic correlation values than the data set without outliers.

This study hopes to contribute to literature for making better decisions on selection of appropriate cluster methods by using subgroup sizes, variable numbers, subgroup means and variances.

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
SS has made intellectual contributions in order to carry out this study and also has carried out the simulation study. ND has determined the research design as well as has coordinated the whole process. İD has made theoretical contributions and has performed statistical analysis of the study. All authors read and approved the final manuscript.

Author details
[1]Department of Statistics, Faculty of Arts and Sciences, Afyon Kocatepe University, Afyonkarahisar, 03200, Turkey.
[2]Department of Biostatistics, Faculty of Medicine, Afyon Kocatepe University, Afyonkarahisar, 03200, Turkey.

Acknowledgements
Dedicated to Professor Hari M Srivastava.
The authors would like to thank Rıdvan ÜNAL for support of technical help. He is a lecturer at the Afyon Kocatepe University, Faculty of Science, Department of Physics, Afyonkarahisar/Turkey.

Received: 31 December 2012 Accepted: 10 April 2013 Published: 23 April 2013

References

1. Carmichael, JW, George, JA, Julius, RS: Finding natural clusters. Syst. Zool. **17**(2), 144-150 (1968)
2. Lapointe, FJ, Legendre, P: Comparison tests for dendrograms: a comparative evaluation. J. Classif. **12**, 265-282 (1995)
3. Kettenring, JR: The practice of cluster analysis. J. Classif. **23**, 3-30 (2006)
4. Gunnarsson, J: Portfolio-Based Segmentation and Consumer Behaviour: Empirical Evidence and Methodological Issues. Ph.D. Dissertation, Stockholm School of Economics, The Economic Research Institute, p. 274 (1999)
5. Steinbach, M, Karypis, G, Kumar, V: A comparison of document clustering techniques. Text mining workshop. In: Proc. of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2000), Boston, MA, pp. 20-23 (2000)
6. Hill, T, Lewicki, P: STATISTICS: Methods and Applications. StatSoft, Tulsa (2007)

7. Lessig, VP: Comparing cluster analyses with cophenetic correlation. J. Mark. Res. **9**(1), 82-84 (1972)
8. Sneath, HA, Sokal, RR: Numerical Taxonomy: The Principles and Practice of Numerical Classification, p. 573. Freeman, San Francisco (1973)
9. Mathworks statistics toolbox: http://www.mathworks.com/help/stats/cophenet.html (2012)
10. Sokal, RR, Rohlf, FJ: The comparison of dendrograms by objective methods. Taxon **11**, 33-40 (1962)
11. Farris, JS: On the cophenetic correlation coefficient. Syst. Zool. **18**(3), 279-285 (1969)
12. Rohlf, FJ, David, LF: Test for hierarchical structure in random data sets. Syst. Zool. **17**, 407-412 (1968)
13. Baker, FB: Stability of two hierarchical grouping techniques - case I: sensitivity to data errors. J. Am. Stat. Assoc. **69**, 440-445 (1974)
14. Hubert, L: Approximate evaluation techniques for the single-link and complete-link hierarchical clustering procedures. J. Am. Stat. Assoc. **69**, 698-704 (1974)
15. Kuiper, FK, Fisher, LA: A Monte Carlo comparison of six clustering procedures. Biometrics **31**, 777-783 (1975)
16. Blashfield, RK: Mixture model tests of cluster analysis: accuracy of four agglomerative hierarchical methods. Psychol. Bull. **83**, 377-388 (1976)
17. Milligan, GW: An examination of the effect of six types of error perturbation on fifteen clustering algorithms. Psychometrika **45**, 325-342 (1980)
18. Hands, S, Everitt, B: A Monte Carlo study of the recovery of cluster structure in binary data by hierarchical clustering techniques. Multivar. Behav. Res. **22**, 235-243 (1987)
19. Yao, KB: A comparison of clustering methods for unsupervised anomaly detection in network traffic. Ph.D. Thesis, University of Copenhagen (2006)
20. Ferreira, L, Hitchcock, DB: A comparison of hierarchical methods for clustering functional data. Commun. Stat., Simul. Comput., **38**, 1925-1949 (2009)
21. Milligan, GW, Cooper, MC: A study of standardization of variables in cluster analysis. J. Classif. **5**, 181-204 (1988)
22. Dixon, WJ: Analysis of extreme value. Ann. Math. Stat. **21**, 488-506 (1950)
23. Johnson, RA, Wichern, DW: Applied Multivariate Statistical Analysis, 5th edn. Prentice Hall, New York (2002)

doi:10.1186/1029-242X-2013-203
Cite this article as: Saraçli et al.: **Comparison of hierarchical cluster analysis methods by cophenetic correlation.** *Journal of Inequalities and Applications* 2013 **2013**:203.