On triangular Inequalities of correlation-based distances for gene expression profiles

Jiaxing Chen[1 [0000−0001−5795−6722]], Yen Kaow Ng[2 [0000−0003−1556−9438]], Lu Lin[1 [0000−0002−7160−5184]], Yiqi Jiang[1 [0000−0003−4950−937X]], and Shuaicheng Li[1 [0000−0001−6246−6349]]

[1] Department of Computer Science, City University of Hong Kong, Tat Chee Avenue, Kowloon, Hong Kong SAR
shuaicli@gmail.com
[2] Department of Computer Science, Faculty of Information and Communication Technology, Universiti Tunku Abdul Rahman, Kampar, Malaysia

Abstract. Various distance functions for evaluating the differences between gene expression profiles have been proposed in the past. Such a function would output a low value if the profiles are strongly correlated—either negatively or positively—and vice versa. One popular distance function is the absolute correlation distance, $d_a = 1 - |\rho|$, where ρ is some similarity measures, such as Pearson or Spearman correlation. However, absolute correlation distance fails to fulfill the triangular inequality, which would have guaranteed better performance at vector quantization, allowed fast data localization, as well as sped up data clustering. In this work, we propose $d_r = \sqrt{1 - |\rho|}$ as an alternative. We prove that d_r satisfies the triangular equality when ρ represents Pearson correlation, Spearman correlation, or Cosine similarity. We empirically compared d_r with d_a in gene clustering and sample clustering experiment, using real biological data. The two distances performed similarly in both gene cluster and sample cluster in hierarchical cluster and PAM cluster. However, d_r demonstrated more robust clustering. According to bootstrap experiment, the number of times where d_r generated more robust sample pair partition is significantly (p-value < 0.05) larger. This advantage in robustness is also supported by the class "dissolved" event.

Keywords: Correlation · distance · triangular inequality · cluster · gene expression analysis.

1 Introduction

In biological data analysis we are frequently required to evaluate how similar two genetic expression profiles are. For example, when identifying gene expression patterns across different conditions, when clustering genes of similar functions [6,10], when detecting the gene temporal profile of relevant functional categories by time-series data clustering [8], when measuring similarity between genes in microbial community [5], and when inferring gene regulatory network [20].

Several distance functions are currently used to evaluate this similarity—the most prominent one being the absolute correlation distance. The function regards positive correlation and negative correlation equally, giving a value of zero to highly correlated profiles (whether positively or negatively correlated), and a value of one to uncorrelated profiles. More precisely, the absolute correlation distance is defined as $d_a = 1 - |\rho|$, where ρ can be Pearson correlation, Spearman correlation, uncentered Pearson correlation (which is equivalent to Cosine similarity), or Kendell's correlation. Profiles which are highly correlated have $\rho = 1$ or $\rho = -1$, and hence resulting in $d_a = 0$; profiles which are unrelated have $\rho = 0$, hence resulting in $d_a = 1$. The absolute correlation distance is widely used, for example, in measuring the co-expression similarity between the profiles of genes in WGCNA [18], clustering of gene expressions [4], and in defining the abundance similarity between OTUs in microbiome area [5]. However, in spite of its widespread usage, it has been noted that most variants of the measure, with the exception of the absolute Kendall's correlation, suffer from the drawback of not satisfying the triangular inequality [8, 12].

A distance measure d which (1) satisfies the triangular inequality and (2) has $d(x, y) = 0$ when $x = y$, is called a *metric* [1, 21]. Researchers have observed that the performance of vector quantization improves when the measure used is a metric [23]. A measure which fulfills triangular inequality would allow faster data localization as well as speed up data clustering [1, 22]. Many clustering algorithms, such as k-means [7] and DBSCAN [17], can exploit triangular inequality to achieve better performance. For instance, a distance calculation can be skipped as soon as it is found to exceed lower or upper bounds estimated through triangular inequality [7]. The same strategies cannot be applied on distance measures that violate triangular inequality without compromising the quality of the clustering [1].

Variants of the absolute correlation distance are not the only distance measure used in gene expression analysis that violate triangular inequality. Prasad *et al.* [24] compiled a list of distance measures for analysis on gene expression profiles. Many of the measures in the list do not fulfill triangular inequality. These include the Harmonically summed euclidean distance, Bray-Curtis distance, Pearson correlation distance, absolute Pearson correlation distance, uncentered correlation distance, absolute uncentered correlation distance, Pearson linear dissimilarity, Spearman correlation distance, absolute Spearman rank correlation, and the Cosine distance.

In this work, we propose an alternative d_r to the absolute correlation distance, defined as $d_r = \sqrt{1 - |\rho|}$, where ρ can be Pearson correlations, Spearman correlations, or uncentered Pearson correlation (or Cosine similarity). We show that d_r, unlike d_a, satisfies the triangular equality for all of these correlations.

We compared the performance of d_r to d_a in biological data clustering. The clustering method includes hierarchical clustering and PAM (partitioning around medoids) [16]. For ρ we used Pearson correlation, Spearman correlation, and Cosine similarity. As data we used 16 normalized time-series datasets and cancer samples cluster in 35 expression datasets. Performances for the sample cluster

tests were evaluated with adjusted Rand index (ARI) [25], while those for the gene cluster tests were evaluated with functional analysis.

Our result shows the two distance measures led to identical hierarchical cluster partition in complete linkage and single linkage, but different in average linkage. In the gene cluster experiment, d_r outperformed d_a in 10, 9, and 10 datasets among 16 datasets for average linkage hierarchical cluster, and 9, 12, 7 for PAM experiment. In sample cluster experiment, d_a and d_r obtained the same ARI in at least 27 datasets among all 35 sample cluster dataset. The two distances have comparable performances in real gene cluster and sample cluster, although the clustering performed with d_r are more robust than those with d_a. When tested with multiple bootstrap test, d_r outperformed d_a at robustness. d_r led to more robust clusters than d_a in both hierarchical cluster, when considering internal nodes, and PAM cluster when any of the correlations is used as ρ. For PAM clustering with Pearson correlation used as ρ, in more than 34 datasets, d_r generated significantly (p-vlaue < 0.05) more robust sample pair partition than d_a. Similar results were obtained when ρ is Spearman correlation and Cosine similarity. The robustness of d_r is also supported by statistics on the time a class "dissolved".

We also compare d_r to other variants of d_a where ρ is squared [27], that is, $d_s = \sqrt{1-\rho^2}$. Our results showed d_r to have better performance at clustering.

2 Method

2.1 Prove triangular inequality of the transformation on Pearson correlation

The original absolute correlation distance $d_a = 1 - |\rho|$ dissatisfy triangular inequality. We propose a new measure d_r, as

$$d_r(X,Y) = \sqrt{1-|\rho(X,Y)|}$$

where X and Y are expression profiles, and ρ can be any one of Pearson correlation coefficient, Spearman correlation, or uncentered Pearson correlation.

We first show that d_r is a metric. Take X=$(x_1, x_2, ..., x_n)$, Y=$(y_1, y_2, ..., y_n)$ and Z=$(z_1, z_2, ..., z_n)$, then the triangular inequality can be written as

$$d_r(X,Y) + d_r(Y,Z) \geq d_r(X,Z) \tag{1}$$

This demonstrates that d_r satisfies the triangular inequality, when ρ is Pearson correlation coefficient, Spearman correlation or the uncentered Pearson correlation, according to the details in the supplementary material.

2.2 Evaluation

To compare our modified absolute Pearson correlation distance $d_r(X,Y) = \sqrt{1-|\rho(X,Y)|}$ to the original absolute Pearson correlation distance $d_a(X,Y) =$

$1 - |\rho(X, Y)|$, we performed clustering experiment on real microarray datasets, including 16 gene time-series profile datasets [15] and 35 datasets for clustering of cancer samples [26]. The clustering algorithms for the test include hierarchical clustering and PAM. The input of a clustering task is a distance matrix and the output is a partitioning which gives the clusters. We performed clustering by both gene and sample.

For the sample clusters, we selected the number of clusters, k, according to benchmark. We evaluated the clustering result by examining how consistent the clusters are with the benchmark by ARI [25]. A greater ARI value indicates higher concordance between the cluster partition and the benchmark partition. Given a partition u and a reference partition v,

$$ARI = \frac{a - \frac{(a+b)(a+c)}{(a+b+c+d)}}{\frac{(a+b)(a+c)}{2} - \frac{(a+b)(a+c)}{(a+b+c+d)}}, \tag{2}$$

where a refers to the total number of sample pairs belonging to the same cluster in both u and v, b refers to the total number of sample pairs in the same cluster in u but in different clusters in v, c is the total number of sample pairs that are in different clusters in u but in the same clusters in v, and d refers to the total number of sample pairs that are in different clusters in both u and v.

For the gene clusters, we evaluated clustering performance by gene functional analysis [14]. The number of clusters was determined according to Calinski-Harabasz Index (CH_{index}) [2] as follows. The CH_{index} is given as

$$CH_{index} = \frac{SS_B}{SS_W} \times \frac{N-k}{k-1}, \tag{3}$$

where k is the number of clusters, and N is the total number of samples, SS_W is the overall within-cluster variance, SS_B is the overall between-cluster variance. A higher CH_{index} value implies a better solution. We used the value of k which corresponds to the peak or at least an abrupt elbow on the line-plot of CH_{index} value.

After obtaining the clusters, we performed GO enrichment for each generated cluster with R package [3, 9, 13]. For each cluster generated by d_a, we got a set of significant GO terms with p-value < 0.05, denoted as r1. Similarly, for cluster generated by d_r, we got a set of significant GO term r2. After that, for two result list r1 and r2, we counted the number of times that the GO term of r1 has smaller p-value than that of r2, denoted as $\neq (r1 < r2)$, and the number of times that GO term of r2 has smaller p-value than it of r1, denoted as $\neq (r2 < r1)$. Then we calculated

$$comparison(r1, r2) = \log(\frac{\neq (r1 < r2)}{\neq (r2 < r1)}). \tag{4}$$

Positive values of $comparison(r1, r2)$ imply that r1 is better than r2, and negative values imply the opposite. So the negative values mean d_r wins d_a in this

dataset. If we change the order of the results under comparison (r1,r2) or (r2,r1), it will only change the sign of the result, but not its absolute value.

2.3 Robustness test

To test the robustness of cluster with different distance measures, we performed bootstrap experiments on the 35 microarray datasets in clustering cancer samples, and investigated the "dissolved" [11] event for class given by different cluster processes. For each dataset, we first obtained an original partition p_o from the original dataset. Then, for each dataset, suppose there are n samples in total, we bootstraped 100 times. For each time, we randomly selected n samples with replicate from the original dataset, and performed clustering on the resampled data to get a resulting partition p_i. We compared p_i with p_o. Denote c_{oj} as class j in p_o, c_{ik} as class k in p_i. For each i, we calculated the Jaccard similarity, J_{ijk}, between each c_{oj} and all c_{ik}. Then we calculated $J_{ij} = \max(J_{ijk})$. After repeating 100 times, for each c_{oj} in p_o, we obtained 100 similarity values, respectively denoted J_{ij} for each of the bootstraps. If $J_{ij} < 0.5$, we take c_{oj} as having "dissolved" in bootstrap i. We counted the number of times the class c_{oj} dissolved in 100 bootstrap. If this frequency is larger than 40, we regard c_{oj} as being dissolved in the experiment. We repeated the bootstrap process for multiple iterations. We tested the robustness of the class by comparing the times it dissolved in multiple iterations. Finally, we compared the performance of d_a and d_r by comparing the robustness of the classes they generated.

We also investigated sample pairs for robustness. We selected sample pairs that are clustered together to see whether they are consistently clustered together across multiple runs, in which case, the result for the sample pair is robust. Similarly we examined sample pairs that are not clustered together to see if they are consistently placed in different classes. For each sample pair i and j in one dataset, we counted the number of times n_1 they are sampled together in 100 bootstraps, the number of times n_2 they are clustered in the same class, and the number of times n_3 they are clustered in different classes. If $n_2 > n_3$, then this pair is decided as consistently clustered, otherwise they are not consistently clustered. For each sample pair, we calculated the ratios n_2/n_1 as well as the median value $m_{together}$ for all the non-zero ratio values. This is repeated for n_3/n_1 and their median, $m_{notTogether}$. Then we calculated $\upsilon = m_{together} * m_{notTogether}$. A larger υ implies a more robust clustering. We recorded this as a "win" event for d_r if $\upsilon_r > \upsilon_a$. For 35 files, we got a list of υ for d_r and d_a. We did Wilcox test for the list of υ_r and υ_a with alternative hypothesis as true location shift is not equal to 0.To see wether υ for d_r is significantly larger than υ for d_a.

3 Results

3.1 Performance in gene cluster

First, we evaluated the performance of $d_r = \sqrt{1 - |\rho|}$ and the $d_a = 1 - |\rho|$ on gene clustering. As data we used 16 time-series profile datasets with normalization

from a previous work [15]. For each dataset we calculated the distances d_r and d_a for pairwise gene profiles, resulting in the distance matrices M_r and M_a. Then, we applied hierarchical clustering and PAM for each distance matrix, estimating the number of cluster k by CH_{index}. Since the data sets do not have a reference partition for genes, we evaluated the performance with biological functional analysis [14]. The clustering result with a higher scored GO term is considered as the better solution. For the hierarchical clusters, we tested three modes, namely complete linkage, single linkage, and average linkage. Clustering using either d_r or d_a led to identical dendrograms in complete linkage hierarchical clustering as well as in single linkage.

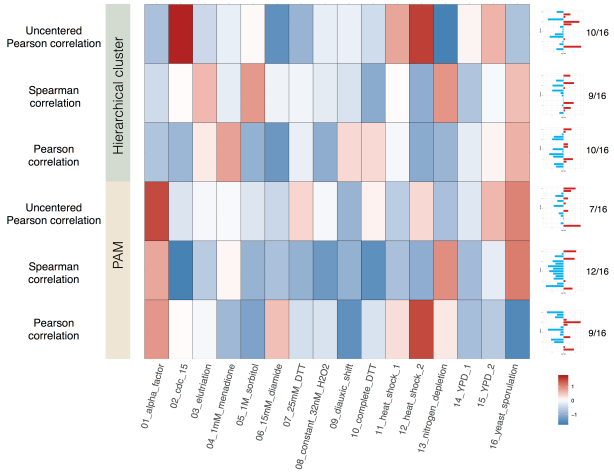

Fig. 1. Result of comparing d_r and d_a in gene clustering. Each column corresponds to one time-series profile dataset. Each row corresponds to one comparison between d_r and d_a while ρ is different correlation in certain clustering method. Color refers to the value of $comparison(r1, r2)$. Negative value implies that d_r is better than d_a, and positive values implies the opposite. For each comparison combination, there is a barplot on the right side on the corresponding row. The x-axis of the barplot refers to $comparison(r1, r2)$.

For hierarchical clustering with average linkage, d_r outperformed d_a in 10, 9, and 10 datasets among 16 datasets when ρ is any of Pearson correlation, Spearman correlation, and Cosine similarity respectively (see Fig. 1). In PAM experiments, d_r outperformed d_a in 9, 12, 7. The two measures outperformed each other for nearly equal number of times.

3.2 Performance in sample cluster

To compare the performance of d_a and d_r in sample clusters, we used 35 datasets from a previous work [26]. The samples in each dataset is assigned a label such as disease or healthy. We applied normalization to each dataset by scaling each gene to the standard normal distribution. We then performed hierarchical clustering and PAM, with the number of clusters k set as the number of the unique labels in each dataset. We evaluated the performance by ARI [25], which measures the consistency between cluster partition and benchmark labels.

For hierarchical cluster, the complete and single mode resulted in identical dendrograms. When ρ is Pearson correlation, the hierarchical cluster in average mode using both d_a and d_r resulted in similar ARI across all 31 datasets (see Fig. 2). When ρ is Spearman correlation or uncentered Pearson correlation, both d_a and d_r resulted in the same ARI in at least 27 datasets, for both methods of clustering. For those datasets with different ARI in comparison, the number of times d_r outperforms d_a is close to the number of times d_a outperforms d_r. As an example, in PAM d_r outperformed d_a 5 times while d_a outperformed d_r 3 times when ρ is Spearman correlation. These results show that they have comparably good performance in our sample cluster experiment.

Fig. 2. Result of comparing d_r and d_a in sample clustering. For each subfigure, x-axis refers to different datasets, y-axis refers to the ARI value. A larger ARI implies a better partitioning.

3.3 Detailed analysis with an example

To see how d_r and d_a lead to different cluster results, we used one dataset as an example and observed the clustering process. We used the 18-th dataset [19] in the sample cluster experiment, performing hierarchical clustering, using Pearson correlation as ρ.

In the beginning, the distance matrices M_r and M_a calculated according to d_r and d_a are the same in rank, in the sense that if we sort the values in M_r and M_a increasingly, the two lists will have the same order. For hierarchical clustering with the complete linkage and single linkage, d_r and d_a led to the same resultant dendrogram because they only take maximum or minimum distance value when calculating the distance between cluster, thus introducing no new value of distance during the entire clustering process. For hierarchical cluster with average linkage, the same two samples are merged at the first step, thus the smallest distances are combined into one cluster. Since an average distance is computed of the newly generated cluster, a difference in rank emerges. In Fig. 3A, the circle network shows the pairs which are different in the ranks of the distance sets generated by d_a and d_r in step 2 to step 6. In this dataset, d_a and d_r led to the same ARI even though the resultant dendrograms are different in structure. The dendrogram for d_a is shown in Fig. 3B and that for d_r in Fig. 3C. The difference between the two dendrograms is colored in red. Fig. 3D shows the distribution of the ranks which are different.

From step 2 to step 52, there exist different ranks in the distance of pairs in two distance experiments. However, the pairs of rank 1 are the same, showing that both d_a and d_r led to the same two samples being merged into a new cluster. In step 53, the pair of rank 1 started to differ, showing that different samples in two distance experiments have been selected. This difference is reflected in the resultant dendrogram. As shown in Fig. 3B and Fig. 3C, for d_a, c_{42} and "$PT102_2$" have been merged, while for d_r, c_{42} and c_{51} have been merged (c represents the internal node in the dendrogram).

In the sample cluster experiment, due to scarcity in the number of pairs (the maximum number of samples in a single dataset is 248 among datasets in this sample cluster experiments), the difference in ARI only occurred in 4 out of 35 datasets. In gene cluster experiment, the boosted number of pairs enlarged the differences in the dendrogram, hence the partition is different in all 16 time-series datasets.

3.4 Robustness test

We compared the methods' robustness with bootstrap experiments in clustering cancer samples on 35 microarray datasets. This is done by examining the number of sample pairs that are consistently clustered across 20 iterations. In each iteration, we resampled 100 times for each dataset. For PAM, d_r displayed more robust clustering than d_a. Fig. 4A, B, C and D are for comparing d_r and d_a through PAM clustering using Pearson correlation as ρ. Fig. 4A shows the number of times d_r achieved a win over 20 iterations in each dataset. d_r achieved more win in 34 datasets among 35 datasets (see Fig. 4B). Fig. 4C shows the box plot for v over 20 iterations and 35 datasets.

Fig. 4D shows the results where we evaluated robustness through the number of times a class is "dissolved". The number of classes dissolved through d_a is larger than it in d_r in all 20 iterations. Hence, d_r led to more robust clustering results, consistent with our earlier results in Fig. 4A, B, C. Similar results are

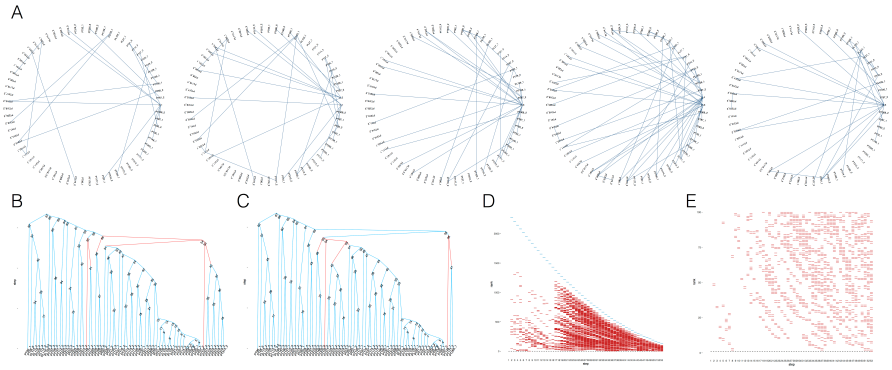

Fig. 3. An example for average linkage hierarchical cluster with d_a and d_r. A. The circle networks show the pairs where the distance is different in the ranks generated by d_a and d_r in step 2 to step 6. Nodes in network refer to samples for clustering. Edges refer to the distance where two sample are different in rank in d_a and d_r. c_1 refers to the the class generated in step 1, c_2 refers to the class generated in step 2. B. Dendrogram for d_a. C. Dendrogram for d_r. The difference between the two dendrograms is colored in red. D. Distribution of ranks which are different in d_a and d_r. E. Zoom in for the top 100 rank for Figure 3D.

obtained when ρ is Spearman correlation and Cosine similarity, as shown in Fig. 4E and Fig. 4F.

For the hierarchical cluster, we examined all the internal nodes for the number of times those class dissolved for each dataset. Fig. 4G shows the number of datasets where d_r achieved a win. Fig. 4H shows the comparison according to each dataset. Both figures show that d_r achieved a win for more times than d_a. Across 20 iterations, the average number of times when d_r wins is larger than the time d_a wins. In summary, the use of d_r resulted in more robust clustering than d_a in both hierarchical and PAM clustering.

4 Discussion

Failure in satisfying the triangular inequality is a severe problem in absolute correlation distance. We show how frequently this violation occurs in the 35 sample cluster datasets [26] (see Fig.S1 in supplementary material). The distributions differ across datasets, with fewer violations after normalization. The number of violations also appear to decrease during the merge process in average linkage hierarchical cluster. More violations (of up to 40%) appeared in the 16 gene cluster dataset, as shown in Fig.S1D.

Besides, we also compare d_r to squared correlation distance. In [27], two variants of the absolute correlation were proposed, namely $d_o = \sqrt{\frac{1}{2}(1-\rho)}$ and $d_s = \sqrt{(1-\rho^2)}$, where Pearson correlations is used as ρ. These efforts would

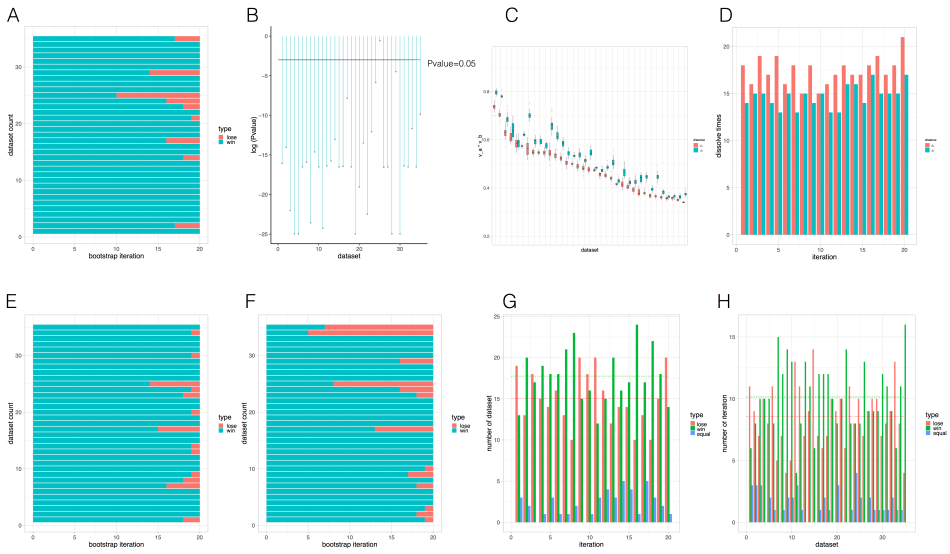

Fig. 4. Result for robustness test on d_a and d_r. A,B,C,D. Results obtained using Pearson correlation as ρ on PAM. A. The number of times d_r win over 20 iterations in each dataset. Each row corresponds to one dataset. B. p-values in testing the difference between the number of times d_r wins in all 35 datasets. Each point corresponds to one dataset. C. Each box represents one v value over 20 iterations per dataset. We compared the box plot for d_a and d_r in each dataset. The datasets in C have been reordered to fit the decrease of y value to show the trend more clearly. D. The number of classes "dissolved" in d_a and d_r across all 20 iterations. E. Result for Spearman correlation as ρ in PAM clustering. F. Result for Uncentered Pearson correlation as ρ in PAM clustering. G, H Results for Pearson correlation as ρ in hierarchical clustering, considering all internal nodes as classes, G. Result for comparing d_a and d_r by the number of times classes "dissolved" in 35 datasets over 20 iterations. The number of times d_r win, lose, or is equal to d_a. The green horizontal line represents the average number across all the iterations where d_r wins. The red horizontal line represent the average number across all the iterations where d_r lose. H. Result for comparing d_a and d_r per dataset.

result in metric distances. The first variant, d_o, has a range of 0 to $\sqrt{2}$, which results in inconsistencies with d_a, thus limiting its use. The squared correlation distance, d_s, on the other hand, is analytically less sensitive than d_r in responding to changes in ρ. This observation is confirmed by our empirical tests using hierarchical clustering (see Fig.S2 in supplementary material). In the tests, d_r-based clustering outperformed d_s in 15 datasets, while losing out to d_s in only 8.

5 Conclusion

The absolute correlation distance $d_a = 1 - |\rho|$ is widely used in biological data clustering in spite of its shortcoming of not satisfying the triangular inequality. In this paper we proposed an alternative, d_r, that does. Our comparison of d_r and d_a on gene clustering using 16 normalized time-series datasets and sample cluster in 35 expression datasets shows that the two distance measures led to identical clusters in hierarchical clustering with complete linkage and single linkage. The two distances have comparable performances in both gene cluster and sample cluster, using both hierarchical as well as PAM cluster, although d_r-based clustering led to more robust clustering. The robustness of d_r-based clustering is also supported by evaluation based on the number of times that a class "dissolved".

References

1. Baraty, S., Simovici, D.A., Zara, C.: The impact of triangular inequality violations on medoid-based clustering. In: International Symposium on Methodologies for Intelligent Systems. pp. 280–289. Springer (2011)
2. Caliński, T., Harabasz, J.: A dendrite method for cluster analysis. Communications in Statistics-theory and Methods **3**(1), 1–27 (1974)
3. Carlson, M., Falcon, S., Pages, H., Li, N.: org. hs. eg. db: Genome wide annotation for human. R package version 3.3 (2013)
4. Datta, S., Datta, S.: Comparisons and validation of statistical clustering techniques for microarray gene expression data. Bioinformatics **19**(4), 459–466 (2003)
5. Deng, Y., Jiang, Y.H., Yang, Y., He, Z., Luo, F., Zhou, J.: Molecular ecological network analyses. BMC bioinformatics **13**(1), 113 (2012)
6. Eisen, M.B., Spellman, P.T., Brown, P.O., Botstein, D.: Cluster analysis and display of genome-wide expression patterns. Proceedings of the National Academy of Sciences **95**(25), 14863–14868 (1998)
7. Elkan, C.: Using the triangle inequality to accelerate k-means. In: Proceedings of the 20th International Conference on Machine Learning (ICML-03). pp. 147–153 (2003)
8. Ernst, J., Nau, G.J., Bar-Joseph, Z.: Clustering short time series gene expression data. Bioinformatics **21**(suppl_1), i159–i168 (2005)
9. Falcon, S., Gentleman, R.: Using gostats to test gene lists for go term association. Bioinformatics **23**(2), 257–258 (2006)
10. Hardin, J., Mitani, A., Hicks, L., VanKoten, B.: A robust measure of correlation between two genes on a microarray. BMC bioinformatics **8**(1), 220 (2007)

11. Hennig, C., et al.: Dissolution point and isolation robustness: robustness criteria for general cluster analysis methods. Journal of multivariate analysis **99**(6), 1154–1176 (2008)
12. ttnphns (https://stats.stackexchange.com/users/3277/ttnphns): Is triangle inequality fulfilled for these correlation-based distances? Cross Validated, https://stats.stackexchange.com/q/135231, uRL:https://stats.stackexchange.com/q/135231 (version: 2017-04-13)
13. Huber, W., Carey, V.J., Gentleman, R., Anders, S., Carlson, M., Carvalho, B.S., Bravo, H.C., Davis, S., Gatto, L., Girke, T., et al.: Orchestrating high-throughput genomic analysis with bioconductor. Nature methods **12**(2), 115 (2015)
14. Jaskowiak, P.A., Campello, R.J., Costa, I.G.: On the selection of appropriate distances for gene expression data clustering. In: BMC bioinformatics. vol. 15, p. S2. BioMed Central (2014)
15. Jaskowiak, P.A., Campello, R.J., Costa Filho, I.G.: Proximity measures for clustering gene expression microarray data: a validation methodology and a comparative analysis. IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB) **10**(4), 845–857 (2013)
16. Kaufman, L., Rousseeuw, P.J.: Finding groups in data: an introduction to cluster analysis, vol. 344. John Wiley & Sons (2009)
17. Kryszkiewicz, M., Lasek, P.: Ti-dbscan: Clustering with dbscan by means of the triangle inequality. In: International Conference on Rough Sets and Current Trends in Computing. pp. 60–69. Springer (2010)
18. Langfelder, P., Horvath, S.: Wgcna: an r package for weighted correlation network analysis. BMC bioinformatics **9**(1), 559 (2008)
19. Lapointe, J., Li, C., Higgins, J.P., Van De Rijn, M., Bair, E., Montgomery, K., Ferrari, M., Egevad, L., Rayford, W., Bergerheim, U., et al.: Gene expression profiling identifies clinically relevant subtypes of prostate cancer. Proceedings of the National Academy of Sciences **101**(3), 811–816 (2004)
20. Margolin, A.A., Nemenman, I., Basso, K., Wiggins, C., Stolovitzky, G., Dalla Favera, R., Califano, A.: Aracne: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. In: BMC bioinformatics. vol. 7, p. S7. BioMed Central (2006)
21. McCune, B., Grace, J.B., Urban, D.L.: Analysis of ecological communities, vol. 28. MjM software design Gleneden Beach, OR (2002)
22. Moore, A.W.: The anchors hierarchy: Using the triangle inequality to survive high dimensional data. In: Proceedings of the Sixteenth conference on Uncertainty in artificial intelligence. pp. 397–405. Morgan Kaufmann Publishers Inc. (2000)
23. Pan, J.S., McInnes, F.R., Jack, M.A.: Fast clustering algorithms for vector quantization. Pattern Recognition **29**(3), 511–518 (1996)
24. Prasad, T.V., Babu, R.P., Ahson, S.I.: Gedas-gene expression data analysis suite. Bioinformation **1**(3), 83 (2006)
25. Santos, J.M., Embrechts, M.: On the use of the adjusted rand index as a metric for evaluating supervised classification. In: International Conference on Artificial Neural Networks. pp. 175–184. Springer (2009)
26. de Souto, M.C., Costa, I.G., de Araujo, D.S., Ludermir, T.B., Schliep, A.: Clustering cancer gene expression data: a comparative study. BMC bioinformatics **9**(1), 497 (2008)
27. Van Dongen, S., Enright, A.J.: Metric distances derived from cosine similarity and pearson and spearman correlations. arXiv preprint arXiv:1208.3145 (2012)

Supplementary material

Supplementary figures

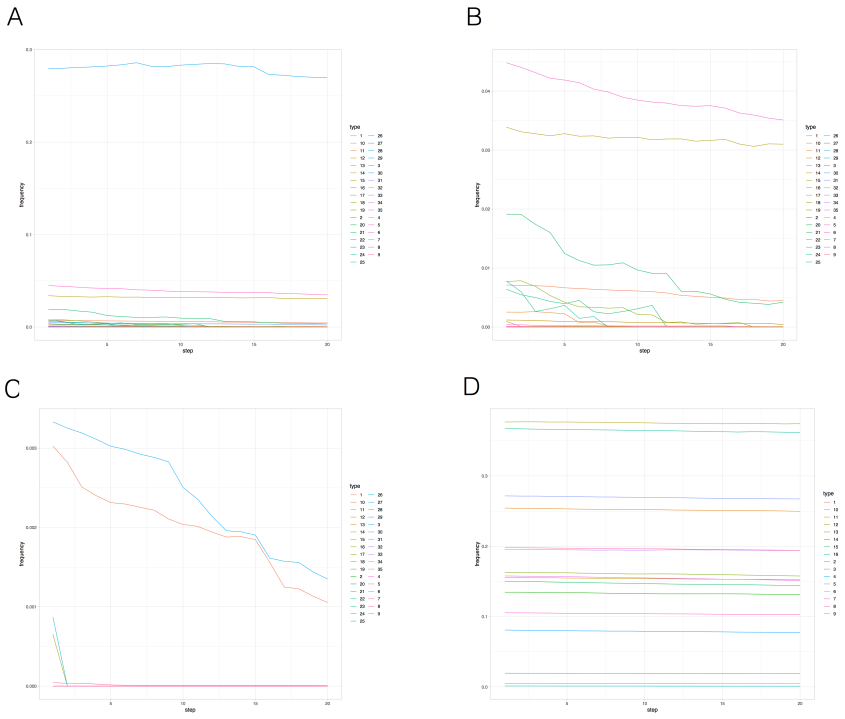

Fig. S1. Percentage of pairs which dissatisfies triangular inequality under d_a. If the distance between any pair of points fails to observe the triangular inequality for some third point, we consider the pair to have failed triangular inequality; the percentage of pairs that do not satisfy triangular inequality is shown. A. Percentage of pairs which dissatisfies triangular inequality, from step 2 to step 20 in hierarchical clustering, within the 35 sample cluster datasets without normalization. B. Detailed view of A. C. Percentage of pairs for the dataset with normalization, with scaling, for each gene. D. Percentage of pairs in hierarchical clustering in gene clustering.

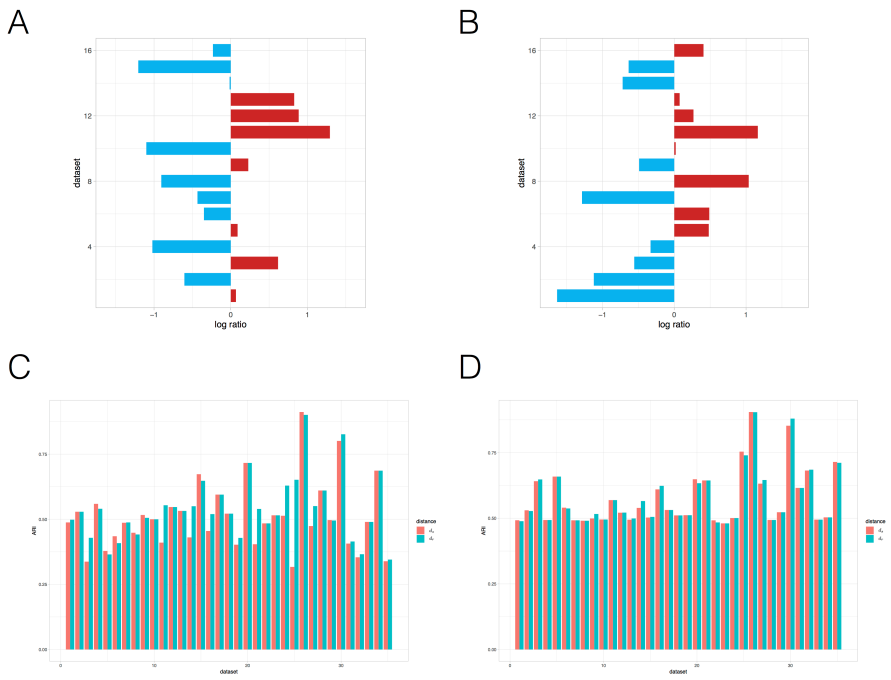

Fig. S2. Result for comparing d_r and d_s. Pearson correlation is used as ρ. A. Comparison on gene clustering using hierarchical cluster. X-axis refers to the value of $comparison(r1, r2)$. Negative value implies that d_r is better than d_a, while positive value implies that d_s is better. B. Comparison on gene clustering using PAM. C. Comparison on sample clustering using hierarchical cluster. Y-axis refers to ARI. D. Comparison on sample clustering using PAM.

Proof of d_r fulfilling the triangular inequality for Pearson correlation as ρ

We define the distance of X and Y by $d_r(X,Y) = \sqrt{1-|\rho(X,Y)|}$, where ρ is the Pearson correlation coefficient.

By the triangular inequality of distance in n-dimensional Euclidean space, $d_r(X,Y) + d_r(Y,Z) \geq d_r(X,Z)$. Take $X = (x_1, x_2, ..., x_n)$, $Y = (y_1, y_2, ..., y_n)$ and $Z = z_1, z_2, ..., z_n)$ such that

$$\begin{array}{c} \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} y_i^2 = \sum_{i=1}^{n} z_i^2 = 1 \\ \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i = \sum_{i=1}^{n} z_i = 0 \end{array} \tag{5}$$

Then the triangular inequality

$$\sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} + \sqrt{\sum_{i=1}^{n}(y_i - z_i)^2} \geq \sqrt{\sum_{i=1}^{n}(x_i - z_i)^2} \tag{6}$$

can be rewritten as

$$\sqrt{1 - \sum_{i=1}^{n} x_i y_i} + \sqrt{1 - \sum_{i=1}^{n} y_i z_i} \geq \sqrt{1 - \sum_{i=1}^{n} x_i z_i} \tag{7}$$

For data from a sample, the Pearson correlation coefficient can be calculated as follows

$$\rho_{XY} = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{(n-1)s_X s_Y} \tag{8}$$

Since Pearson correlation coefficient is invariant under linear transformation, which means $\rho_{\tilde{X}\tilde{Y}}$=$\rho_{XY}$ with $\tilde{X} = a(X - \bar{x})$ and $\tilde{Y} = b(Y - \bar{y})$ satisfying

$$\begin{array}{c} \sum_{i=1}^{n} \tilde{x}_i^2 = \sum_{i=1}^{n} \tilde{y}_i^2 = 1 \\ \sum_{i=1}^{n} \tilde{x}_i = \sum_{i=1}^{n} \tilde{y}_i = 0 \end{array} \tag{9}$$

where $\bar{x}$ and $\bar{y}$ are the sample means of X and Y, it can be rewritten as

$$\rho_{XY} = \rho_{\tilde{X}\tilde{Y}} = \sum_{i=1}^{n} \tilde{x}_i \tilde{y}_i \tag{10}$$

Without loss of generality, we assume that the samples are normalized (i.e. satisfying Equation 5).

Therefore, we have the modified Pearson distance

$$d_r(X,Y) = \sqrt{1 - |\rho_{XY}|} = \sqrt{1 - \left|\sum_{i=1}^{n} x_i y_i\right|} \tag{11}$$

To prove the triangular inequality of d_r, we divide this into eight cases by the signs of ρ.

Case I When $\rho_{XY} \geq 0$, $\rho_{YZ} \geq 0$, $\rho_{XZ} \geq 0$,

$$\begin{aligned} & d_r(X,Y) + d_r(Y,Z) - d_r(X,Z) \\ =& \sqrt{1 - \sum_{i=1}^{n} x_i y_i} + \sqrt{1 - \sum_{i=1}^{n} y_i z_i} - \sqrt{1 - \sum_{i=1}^{n} x_i z_i} \geq 0 \end{aligned} \tag{12}$$

by (7)

Case II When $\rho_{XY} \geq 0$, $\rho_{YZ} < 0$, $\rho_{XZ} < 0$, take $c_i = -z_i$.

$$\begin{aligned} & d_r(X,Y) + d_r(Y,Z) - d_r(X,Z) \\ =& \sqrt{1 - \sum_{i=1}^{n} x_i y_i} + \sqrt{1 + \sum_{i=1}^{n} y_i z_i} - \sqrt{1 + \sum_{i=1}^{n} x_i z_i} \\ =& \sqrt{1 - \sum_{i=1}^{n} x_i y_i} + \sqrt{1 - \sum_{i=1}^{n} y_i c_i} - \sqrt{1 - \sum_{i=1}^{n} x_i c_i} \\ \geq & 0 \end{aligned} \tag{13}$$

by (7)

Case III The case when $\rho_{XY} < 0$, $\rho_{YZ} \geq 0$, $\rho_{XZ} < 0$ is equivalent to case II.

$$d_r(X,Y) + d_r(Y,Z) - d_r(X,Z) \geq 0 \tag{14}$$

holds

Case IV When $\rho_{XY} < 0$, $\rho_{YZ} < 0$, $\rho_{XZ} \geq 0$, take $b_i = -y_i$.

$$\begin{aligned} & d_r(X,Y) + d_r(Y,Z) - d_r(X,Z) \\ =& \sqrt{1 + \sum_{i=1}^{n} x_i y_i} + \sqrt{1 + \sum_{i=1}^{n} y_i z_i} - \sqrt{1 - \sum_{i=1}^{n} x_i z_i} \\ =& \sqrt{1 - \sum_{i=1}^{n} x_i b_i} + \sqrt{1 - \sum_{i=1}^{n} b_i z_i} - \sqrt{1 - \sum_{i=1}^{n} x_i z_i} \\ \geq & 0 \end{aligned} \tag{15}$$

by (7)

Case V When $\rho_{XY} < 0$, $\rho_{YZ} < 0$, $\rho_{XZ} < 0$

$$\begin{aligned} & d_r(X,Y) + d_r(Y,Z) - d_r(X,Z) \\ =& \sqrt{1 + \sum_{i=1}^{n} x_i y_i} + \sqrt{1 + \sum_{i=1}^{n} y_i z_i} - \sqrt{1 + \sum_{i=1}^{n} x_i z_i} \end{aligned} \tag{16}$$

Take $b_i = -y_i$. Therefore we have

$$\begin{array}{l} \sum_{i=1}^{n} x_i b_i > 0 \\ \sum_{i=1}^{n} b_i z_i > 0 \\ \sum_{i=1}^{n} x_i z_i < 0 \end{array} \tag{17}$$

$$\begin{aligned} & d_r(X,Y) + d_r(Y,Z) - d_r(X,Z) \\ =& \sqrt{1 + \sum_{i=1}^{n} x_i y_i} + \sqrt{1 + \sum_{i=1}^{n} y_i z_i} - \sqrt{1 + \sum_{i=1}^{n} x_i z_i} \\ =& \sqrt{1 - \sum_{i=1}^{n} x_i b_i} + \sqrt{1 - \sum_{i=1}^{n} b_i z_i} - \sqrt{1 + \sum_{i=1}^{n} x_i z_i} \\ >& \sqrt{1 - \sum_{i=1}^{n} x_i b_i} + \sqrt{1 - \sum_{i=1}^{n} b_i z_i} - \sqrt{1 - \sum_{i=1}^{n} x_i z_i} \\ \geq & 0 \end{aligned} \tag{18}$$

by (7)

Case VI When $\rho_{XY} < 0$, $\rho_{YZ} \geq 0$, $\rho_{XZ} \geq 0$,

$$\begin{aligned} & d_r(X,Y) + d_r(Y,Z) - d_r(X,Z) \\ =& \sqrt{1 + \sum_{i=1}^{n} x_i y_i} + \sqrt{1 - \sum_{i=1}^{n} y_i z_i} - \sqrt{1 - \sum_{i=1}^{n} x_i z_i} \end{aligned} \tag{19}$$

Take $a_i = -x_i$,

$$\begin{aligned} & \sqrt{1 + \sum_{i=1}^{n} x_i y_i} + \sqrt{1 - \sum_{i=1}^{n} y_i z_i} - \sqrt{1 - \sum_{i=1}^{n} x_i z_i} \\ =& \sqrt{1 - \sum_{i=1}^{n} a_i y_i} + \sqrt{1 - \sum_{i=1}^{n} y_i z_i} - \sqrt{1 + \sum_{i=1}^{n} x_i z_i} \\ > & 0 \end{aligned} \tag{20}$$

by (18)

Case VII The case when $\rho_{XY} < 0$, $\rho_{YZ} \geq 0$, $\rho_{XZ} \geq 0$, is equivalent to case VI.

$$d_r(X,Y) + d_r(Y,Z) - d_r(X,Z) > 0 \tag{21}$$

still holds.

Case VIII When $\rho_{XY} \geq 0$, $\rho_{YZ} \geq 0$, $\rho_{XZ} < 0$,

$$\begin{aligned} & d_r(X,Y) + d_r(Y,Z) - d_r(X,Z) \\ =& \sqrt{1 - \sum_{i=1}^{n} x_i y_i} + \sqrt{1 - \sum_{i=1}^{n} y_i z_i} - \sqrt{1 + \sum_{i=1}^{n} x_i z_i} \\ \geq & 0 \end{aligned} \tag{22}$$

by (18)

$$d_r(X,Y) + d_r(Y,Z) \geq d_r(X,Z) \tag{23}$$

holds for any X, Y and Z.

Proof of d_r fulfilling the triangular inequality for Spearman correlation as ρ

We define the distance of X and Y by $d_r(X,Y) = 1 - \sqrt{|\rho(X,Y)|}$, where ρ is the Spearman correlation.

$d_r(X,Y) = 1 - \sqrt{|\rho(X,Y)|}$, when ρ is the Spearman correlation, can be regarded as a special case of $d_r(X,Y) = \sqrt{1 - |\rho(X,Y)|}$, when ρ is Pearson correlation, where X=$(x_1, x_2, ..., x_n)$, Y=$(y_1, y_2, ..., y_n)$ and $x_1, x_2, ..., x_n$, $y_1, y_2, ..., y_n$ are integers. Since the inequality holds for the case of Pearson correlation, the inequality holds here.

Proof of d_r fulfilling the triangular inequality for uncentered Pearson correlation as ρ

We define the distance of X and Y by $d_r(X,Y) = \sqrt{1 - |\rho(X,Y)|}$, where ρ is the uncentered Pearson correlation coefficient.

Take $X = (x_1, x_2, ..., x_n)$, $Y = (y_1, y_2, ..., y_n)$ and $Z = (z_1, z_2, ..., z_n)$.

Then the triangular inequality

$$d_r(X,Y) + d_r(Y,Z) \geq d_r(X,Z) \tag{24}$$

For any data from a sample, the uncentered Pearson correlation coefficient can be calculated as follows

$$\rho_{XY} = \frac{1}{n} \sum_{i=1}^{n} (\frac{x_i}{\sigma_x^{(o)}})(\frac{y_i}{\sigma_y^{(o)}}) \tag{25}$$

where $\sigma_x^{(o)} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} x_i^2}$, $\sigma_y^{(o)} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} y_i^2}$.

ρ_{XY} can be written as cosine similarity,

$$cos\theta = \frac{X \cdot Y}{|X||Y|} = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2}\sqrt{\sum_{i=1}^{n} y_i^2}}, \tag{26}$$

where θ is the angle between X and Y. Suppose X, Y, and Z are on the same plane. Let α denote the angle between X and Y, β denote the angle between Y and Z, such that the angle between X and Z is $\alpha + \beta$. To prove the triangular inequality of Γ, we divide this into multiple cases according to the range of α and β (sign of $cos\alpha$ and $cos\beta$). Suppose $0 \le \alpha \le \pi$, $0 \le \beta \le \pi$,

Case I $0 \le \alpha \le \frac{\pi}{2}$, $0 \le \beta \le \frac{\pi}{2}$, $0 \le \alpha + \beta \le \frac{\pi}{2}$,

$$\begin{aligned} & d_r(X,Y) + d_r(Y,Z) - d_r(X,Z) \\ &= \sqrt{1-cos\alpha} + \sqrt{1-cos\beta} - \sqrt{1-cos(\alpha+\beta)} \\ &= \sqrt{2}(sin\frac{\alpha}{2}(1-cos\frac{\beta}{2}) + sin\frac{\beta}{2}(1-cos\frac{\alpha}{2})) \\ &\ge 0 \end{aligned} \tag{27}$$

by $1 - cos\frac{\beta}{2} \ge 0$, $1 - cos\frac{\alpha}{2} \ge 0$.

Case II $0 \le \alpha \le \frac{\pi}{2}$, $0 \le \beta \le \frac{\pi}{2}$, $\frac{\pi}{2} \le \alpha + \beta \le \pi$, $\frac{\pi}{4} \le \frac{\alpha+\beta}{2} \le \frac{\pi}{2}$

$$\begin{aligned} & d_r(X,Y) + d_r(Y,Z) - d_r(X,Z) \\ &= \sqrt{1-cos\alpha} + \sqrt{1-cos\beta} - \sqrt{1+cos(\alpha+\beta)} \\ &= \sqrt{2}(sin\frac{\alpha}{2} + sin\frac{\beta}{2} - cos\frac{\alpha+\beta}{2}) \\ &\ge 0 \end{aligned} \tag{28}$$

can be written as

$$\begin{aligned} sin\frac{\alpha}{2} + sin\frac{\beta}{2} &\ge cos\frac{\alpha+\beta}{2} \\ sin^2\frac{\alpha}{2} + sin^2\frac{\beta}{2} + 2sin\frac{\alpha}{2}sin\frac{\beta}{2} &\ge 1 - sin^2\frac{\alpha+\beta}{2} \\ -cos(\alpha+\beta) + 2sin\frac{\alpha}{2}sin\frac{\beta}{2}(1-cos\frac{\alpha+\beta}{2}) &\ge 0 \end{aligned} \tag{29}$$

holds for $1 - cos\frac{\alpha+\beta}{2} \ge 0$, $sin\frac{\alpha}{2}sin\frac{\beta}{2} \ge 0$ and $cos(\alpha+\beta) \le 0$

Case III $0 \le \alpha \le \frac{\pi}{2}$, $\frac{\pi}{2} \le \beta \le \pi$, $\frac{\pi}{2} \le \alpha + \beta \le \frac{3\pi}{2}$

$$\begin{aligned} & d_r(X,Y) + d_r(Y,Z) - d_r(X,Z) \\ & = \sqrt{1-cos\alpha} + \sqrt{1+cos\beta} - \sqrt{1+cos(\alpha+\beta)} \\ & = \sqrt{2}(sin\frac{\alpha}{2}(1+sin\frac{\beta}{2}) - cos\frac{\beta}{2}(cos\frac{\alpha}{2}-1)) \\ & \geq 0 \end{aligned} \tag{30}$$

by $sin\frac{\alpha}{2} \geq 0$, $cos\frac{\beta}{2} \geq 0$, $cos\frac{\alpha}{2} - 1 \leq 0$, $1 + sin\frac{\beta}{2} \geq 0$

Case IV $\frac{\pi}{2} \leq \alpha \leq \pi$, $0 \leq \beta \leq \frac{\pi}{2}$, $\frac{\pi}{2} \leq \alpha+\beta \leq \frac{3\pi}{2}$ and $\frac{\pi}{4} \leq \frac{\alpha}{2} \leq \frac{\pi}{2}$

$$\begin{aligned} & d_r(X,Y) + d_r(Y,Z) - d_r(X,Z) \\ & = \sqrt{1+cos\alpha} + \sqrt{1-cos\beta} - \sqrt{1+cos(\alpha+\beta)} \\ & = \sqrt{2}(sin\frac{\beta}{2}(1+sin\frac{\alpha}{2}) - cos\frac{\alpha}{2}(cos\frac{\beta}{2}-1)) \\ & \geq 0 \end{aligned} \tag{31}$$

for $sin\frac{\beta}{2} \geq 0$, $1 + sin\frac{\alpha}{2} \geq 0$, $cos\frac{\alpha}{2} \geq 0$, $cos\frac{\beta}{2} - 1 \leq 0$.

Case V $\frac{\pi}{2} \leq \alpha \leq \pi$, $\frac{\pi}{2} \leq \beta \leq \pi$, $\frac{\pi}{2} \leq \alpha+\beta \leq 2\pi$ and $cos(\alpha+\beta) > 0$ and $\frac{\pi}{4} \leq \frac{\alpha}{2} \leq \frac{\pi}{2}$, $\frac{\pi}{4} \leq \frac{\beta}{2} \leq \frac{\pi}{2}$

$$\begin{aligned} & d_r(X,Y) + d_r(Y,Z) - d_r(X,Z) \\ & = \sqrt{1+cos\alpha} + \sqrt{1+cos\beta} - \sqrt{1-cos(\alpha+\beta)} \\ & = \sqrt{2}(cos\frac{\alpha}{2}(1-sin\frac{\beta}{2}) + cos\frac{\beta}{2}(1-sin\frac{\alpha}{2})) \\ & \geq 0 \end{aligned} \tag{32}$$

by $cos\frac{\alpha}{2} \geq 0$, $1 - sin\frac{\beta}{2} \geq 0$, $cos\frac{\beta}{2} \geq 0$, $1 - sin\frac{\alpha}{2} \geq 0$

Case VI $\frac{\pi}{2} \leq \alpha \leq \pi$, $\frac{\pi}{2} \leq \beta \leq \pi$, $\frac{\pi}{2} \leq \alpha+\beta \leq 2\pi$ and $cos(\alpha+\beta) < 0$, and $\frac{\pi}{4} \leq \frac{\alpha}{2} \leq \frac{\pi}{2}$, $\frac{\pi}{4} \leq \frac{\beta}{2} \leq \frac{\pi}{2}$

$$\begin{aligned} & d_r(X,Y) + d_r(Y,Z) - d_r(X,Z) \\ & = \sqrt{1+cos\alpha} + \sqrt{1+cos\beta} - \sqrt{1+cos(\alpha+\beta)} \\ & = \sqrt{2}sin\frac{\alpha}{2}sin\frac{\beta}{2}(\frac{cos\frac{\alpha}{2}+cos\frac{\beta}{2}}{sin\frac{\alpha}{2}sin\frac{\beta}{2}} - \frac{cos\frac{\alpha}{2}cos\frac{\beta}{2}}{sin\frac{\alpha}{2}sin\frac{\beta}{2}} + 1) \\ & \geq 0 \end{aligned} \tag{33}$$

for $cos\frac{\alpha}{2} \geq 0$, $cos\frac{\beta}{2} \geq 0$, $sin\frac{\alpha}{2} \geq 0$, $sin\frac{\beta}{2} \geq 0$
$cos\frac{\alpha}{2} \leq sin\frac{\alpha}{2}$, $cos\frac{\beta}{2} \leq sin\frac{\beta}{2}$ and $\frac{cos\frac{\alpha}{2}}{sin\frac{\alpha}{2}} \leq 1$, $\frac{cos\frac{\beta}{2}}{sin\frac{\beta}{2}} \leq 1$