Nonmetric ANOVA: a generic framework for analysis of variance on dissimilarity measures

Alina Malyutina\textsuperscript{a}, Jing Tang\textsuperscript{a1}, and Ali Amiryousefi\textsuperscript{a1}

\textsuperscript{a}Research Program in Systems Oncology, Faculty of Medicine, PO Box 63, FI-00014, University of Helsinki, Finland.

This manuscript was compiled on November 25, 2021

Classic Analysis of Variance (ANOVA; cA) tests the explanatory power of a partitioning on a set of objects. Nonparametric ANOVA (npA) extends to a case where instead of the object values themselves, their mutual distances are available. While considerably widening the applicability of the cA, the npA does not provide a statistical framework for the cases where the mutual dissimilarity measurements between objects are nonmetric. Based on the central limit theorem (CLT), we introduce nonmetric ANOVA (nmA) as an extension of the cA and npA models where metric properties (identity, symmetry, and subadditivity) are relaxed. Our model allows any dissimilarity measures to be defined between objects where a distinctiveness of a specific partitioning imposed on those are of interest. This derivation accommodates an ANOVA-like framework of judgment, indicative of significant dispersion of the partitioned outputs in nonmetric space. We present a statistic which under the null hypothesis of no differences between the mean of the imposed partitioning, follows an exact $F$-distribution allowing to obtain the consequential $p$-value. Three biological examples are provided and the performance of our method in relation to the cA and npA is discussed.

ANOVA | dissimilarity | partitioning | hypothesis testing | metric function

Classical analysis of variance (ANOVA; cA) (1), stands as a renown generic statistical tool for testing mean similarity of different groups (2, 3). Since its introduction, cA has been developed to take into account various aspects of the experimental design and properties of observed variables from unbalanced sampling (4) to non-Gaussian residuals by Kruskal-Wallis type of nonparametric testing (5–7). A special case where the reference point to which the measurement of objects is not available but still one is able to collect all the pairwise distances between them, caught considerable attention by a series of work (8, 9). In this scenario, the emphasis has been put on the use of distances between responses rather than their actual values (9), as well as different measures for quantifying the distances (10), and a permutative method for forming a pseudo-$F$ hypothesis testing statistic (hence nonparametric ANOVA; npA) based on the distances of the responses (11). In npA, the original sum of squared distances of responses from their means was replaced by the mean sum of square distances between responses. This was shown to lead into a variance decomposition that equates the case where the observation vectors themselves are considered (S3.2). While the quantification of most experimental variables could be measured with metric functions, there are cases for which the use of these functions are not more than a mere coercion. For example, a data transformation of non-metric outputs to Euclidean mean (12) or by incorporating a deduced metrics from a principal coordinates analysis on the dissimilarity matrix (11), while permits the npA model to operate, is leading to invalid judgments due to compressing and deforming the data beyond sufficiency. The failure to appropriately acknowledge the data inherency, is at the core of the misjudgment based upon the deficient data coercion. Network graphs rising in multiple biological studies for example (13–15), could exhibit the nonsymmetrical measurements between the nodes, lack the edge between two, or even contain a self-looped nodes; all of which being cases where the metric properties are negated (S1). Each of these nonmetric scenarios should then be represented with a dissimilarity matrix between the nodes (rather than being metric-forced) and so demand a statistical framework that could incorporate this input as it is.

Here, based on the central limit theorem (CLT) and harnessing the natural relationship between statistical distributions (16), we generalize the cA method to a nonmetric case by introducing a nonmetric ANOVA (nmA). This development allows the direct integration of the nonmetric data input and unlike npA, provides a parametric $p$-value based on the theoretical distribution of an $F$-statistic that under null hypothesis of means similarity between partitioning classes follows the exact $F$-distribution. This enables the hypothesis testing for the experimental settings with nonmetric outputs. As we demonstrate with three diverse biological case studies, the nmA provides a statistical testing solution that allows the integration of the unmodulated dissimilarities input into a test statistic, downstream testing mechanics, and a resultant $p$-value.

**Significance Statement**

The Nonmetric Analysis of Variance (nmANOVA) conveys a framework that allows a compatible type of ANOVA for the cases where the proper metric measurements between objects are either lost, unknown or however inaccessible. While classic ANOVA is based on the measurements of the data from a base datum, the nmANOVA is formulated on the dissimilarity outputs (not necessarily metric) defined between all objects. As the main goal of ANOVA in providing a statistical test for assessing the significance of a considered partitioning on the data, the nmANOVA is yielding a paralleled scheme of inference with 1) accommodating the outcomes dissimilarities into within and between groups statistics, 2) assessing their respective divergence with a parametric distribution, and 3) providing a resultant $p$-value indicative of evidences for rejecting the null hypothesis.

AA conceived the study and developed the model, AA and AM implemented the methods, JT provided the funding, AM, JT, and AA refined the model and wrote the paper. The authors declare no competing interest.

\textsuperscript{1}To whom correspondence should be addressed. E-mail: jing.tang@helsinki.fi; ali.amiryousefi@helsinki.fi
nmA tests differences of dissimilarities

In many practical cases the level of dissimilarities expressed between items does not follow the metric properties (17). Based on triangle equality, the simultaneous affirmation of all derived equations presented in 10 is necessary for a set of dissimilarities to be defined as metric (18). The negation of each to all of these equations (and hence invalidating the metric definitions) is extensively classified (19). For example, the term semimetric denotes scenarios where the triangle inequality fails to hold in general, while a pseudometric refers to cases where the forward condition of the identity property fails (20); the case where two distinct objects have a zero distance. Referring to the network analogy, this equates the cases with two distinct nodes with a complete connectivity such that the information in either of nodes is surely available in the other. Metametric on the other hand, is an opposite case where the distance between two identical items can be more than zero; a self-looped node that might loose the information received (21), and lastly, quasimetric refers to cases where the symmetry condition is distorted such that the distance from objects A to B is not necessarily the same as distance from B to A (22); such as inequality of the edges and hence unbalanced information flow between two nodes in a directed network. These violations of the metric properties in general provide a ground for assessing cases where the measurements do not exhibit metric properties (23). Next to the classic invention of nonmetric methods like Kruskal multidimensional rank test (24) and Mantel permutation tests (25), the prevalence of the nonmetric data has led to considerable advancement in other statistical domains as a proportionate response for the existing demand. For example, decision trees finding based on nonmetric data provides a robust tree optimization, and partial least square and generalized linear models, offer solutions for the regression fitting for nonmetric outputs (26–29). In alignment to these advancement and reference to the partitioning scenarios, here we extend the cA model to nmA. This extension allows for example, obtaining the p-value for each of the bifurcating nodes on the decision tree and assessing the multipartition of a generic network. Also as a meaningful add to the nonmetric generalized linear models, nmA permits a similar hypothesis testing between objects in different partitions. As such, given the dissimilarity measurements, the resultant p-value of this method could also be used as a yardstick for finding the most distinct partitioning underlying a set of objects. (30).

The \( F_{nm} \) formulated in (2) is encapsulating the ratio of divergence of the mean of between to within groups. Under null hypothesis of no difference between the mean values of the partitioning groups, the sum of normalized means of dissimilarity values between each partition would be the same as their within counterparts (S2). To the degree of deflection from this ground truth, the observed \( F_{o} \) would be inflated and so the \( \Pr(\frac{F_{o} - g, g - 1}{F_{0}}) \) quantity would be the assigned p-value to indicate whether this deflection is of any significance with a designated \( \alpha \)-level threshold. Table 1 summarizes the essentials of this model with respect to the metric ones.

General properties of nmA

nmA is sensitive to groups differences. To assess nmA performance in different situations, we conduct series of simulations from Gaussian and uniform distributions which cover various scenarios via adjusting distribution parameters. While literally any other parametric distribution could be harnessed in here, this choice of models was encouraged as at the constant variance level, the Gaussian distribution is known to have the minimum negentropy (31, 32); hence the most chaotic sequences of values are expected. The uniform distribution on the other hand is chosen as it is usually taken as the state of uninformative in the Bayesian scenarios (33).

Firstly, we are interested in the performance of the nmA when the null hypothesis is true. Under this condition for any parametric test, the p-values should follow the \( U(0,1) \) distribution (34). In our case, using the same sampling distribution for generation of the dissimilarities with given partition imposed on them led to the consistent conclusion (Figure 1A).

Secondly, moving toward the alternative hypothesis, we investigated the sensitivity of the model with respect to the marked changes in the means of partitioning groups as a measure of deviation from the null hypothesis. The incremental shifting of the values of the between groups with respect to the within groups has led to the expected inflation of the \( F_{o} \) statistics in nmA model (Figure 1B). Despite sensitivity of nmA to the location parameter, the scale changes were not leading to the same significant effect as increasing the standard deviation of between to within groups does not lead to the marked changes in the nmA judgment denoting the robustness of the method with respect to scale dimension of the inputs (Figure 1C).

In contrary to its conventional use, nmA can assist in selecting the most similar sub clustering of the objects aggregated in a dissimilarity matrix. This application of the method was assessed by tailoring the apparent differentiation between the groups and allowing the model to retrieve that (Figure 2A). To do that, we first simulate a dissimilarity matrix which represents 4 groups of different sizes - 13, 21, 34, 55. Then we split the matrix into 3 groups via joining the first two clusters, or 5 groups via splitting the last partitioning into 2 groups of sizes 21 and 34. We fix within and between group dissimilarities distributions to \( N(0,1) \) and \( N(5,1) \) respectively. It is important to note that we select the cluster sizes on purpose, as they represent a sub-sequence from a Fibonacci sequence allowing minimum inclusion of the between elements to within when joining or splitting the clusters. Fixing theumber of clusters to 3 as described, we obtain clusters of 34, 34 and 55 sizes. This will not only provide relatively similar sizes for the clusters but also guarantee the smallest influence on the between original cluster 1 and cluster 2 dissimilarity values. More precisely,
we get \(2^4 \times 21\) between cluster 1 and cluster 2 dissimilarity values coming from \(N(5, 1)\) distribution mixed with \(13^2 + 21^2\) dissimilarity values from \(N(0, 1)\). Thus, organising clusters this way leads to the smallest possible influence of higher-valued \((N(5, 1))\) dissimilarities to the newly-obtained within group matrix. Similarly, via splitting the cluster of size 55 into 2 clusters of 21 and 34, we provide the minimum values from \(N(0, 1)\) to appear in between group dissimilarities. However, the \(p\)-values that correspond to the 5 group partitioning are very close to the initial partitioning \(p\)-values, but their difference starts to grow with the increase of elements rearrangement (Figure 2A). The \(p\)-value closeness in the beginning can be explained via subtle changes in between group matrices with the strategy we selected for partitioning. Overall, progressive cluster rearrangements were reflective of growing \(p\)-values to the level that no more differences between imposing groups were existing. The resulting \(p\)-values have indicated that no partition can achieve as low \(p\)-values as those that we got with the initial most differentiated clustering (Figure 2A).

**Asymptotic sampling improves accuracy.** To assess the convergence of \(p\)-values with increasing number of the proportional samplings (S1), we perform simulations where we collect a sequence of \(p\)-values which correspond to the same dissimilarity matrix but with a varying number of proportional samplings. Every \(p\)-value in a sequence is obtained with one more proportional sampling added to the one obtained for a previous \(p\)-value. For these simulations we select 3 groups of sizes 10, 20 and 30 and sample dissimilarity matrix from \(U(0, 1)\) for within group matrices and \(U(2, 4)\) for between group matrices. Thus, we introduce a difference between the groups via increasing mean and standard deviation, and expect \(p\)-values to reflect it. To check the asymptotic sampling behaviour, we apply nmA to the matrix and increase proportional sampling number from 1 to \(2 \times 10^6\). As Figure 2B shows, \(p\)-values capture the significant difference between the groups and converge to their mean value. This example demonstrates that one can improve the accuracy of a \(p\)-value via increasing the number of proportional samplings while even a small number of proportional samplings could be conductive of a significant difference.

Table 1. Summary of different ANOVA models

| Method          | ANOVA            | Nonparametric | Nonmetric |
|-----------------|------------------|---------------|-----------|
| Specification   | Responses        | Distances     | Dissimilarities |
| Input           | \(SS_H/N-g\) - 1 | \(SS_W/N-g\) - g | \(\Delta N \times N - 1\) |
| Test Statistic  | \(F_{g-1,N-g} > F_{a}\) | \(D_{N \times N} - 1\) | \(SNM_{g-1} / g - 1\) |
| \(p\)-value     | \(P\left(F_{g-1,N-g} > F_{a}\right)\) | \(\sum_{i=1}^{N} \left| F_{i} - F_{a}\right|\) | \(P\left(F_{g-1,N-g} > F_{a}\right)\) |

* The high level similarity of nmA to CA is its closed form distribution of the Test Statistic leading to the respective \(p\)-value, while the matrix structure of the Input is common between npA and nmA, albeit former based on the Distances (D), and Dissimilarities (\(\Delta\)) the latter (S1-S3).

Fig. 1. The consistency of the \(F_{nm}\) under different scenarios. A) Uniform distribution of the \(p\)-values. Uniform distribution of 1,000 \(p\)-values for three equal-sized groups of 10 when the whole dissimilarity matrix is coming from \(U(0, 1)\) distribution. The results of two-sided Kolmogorov-Smirnov test are highlighted on top. The same behavior was observed with simulation from normal distribution with 4 groups of 10 each and 1,000 replication (\(p\)-value = 0.635). B) The effect of progressively changing the mean of the between group matrices distribution on the \(p\)-values. We use the previous group dissimilarity matrix as our starting point and continuously change one of the uniform distribution parameters via adjusting \(\psi\) factor (as an incremental adjustment of the \(a\) and \(b\) in \(U(a, b)\)). With every step \(i\) from \(1, 10\), we change the between group matrix distributions to be \(U(0 + \psi_{i}, 1 + \psi_{i})\) and obtain the corresponding \(p\)-value. At the same time, we track the effect of the proportion of between group data being affected via \(\psi\) parameter. The three scenarios are visualized (from left to right): when only group 1 and group 2 pairwise between group matrices are adjusted; when group 1 and group 2 but also group 1 and group 3 dissimilarity matrices are simultaneously adjusted; when all 6 possible between group matrices are simultaneously adjusted. The red dots represent each \(p\)-value with corresponding \(\psi\) and the mean values of them are connected with the solid blue lines. The red dashed line corresponds to the \(p\)-value equal to 0.05. C) The effect of progressively changing the standard deviation of the between group matrices distribution on the \(p\)-values. While in previous case the difference in between group distributions was triggered via changing the mean of uniform distribution, here we fix the mean but change standard deviation via adjusting the between group distribution according to \(U(0 - \psi_{i}, 1 + \psi_{i})\). The adjustment is performed for the same three scenarios as previously. As indicated with the lower values (of the mean and standard deviation) and the higher gradient of descent (solid blue lines), nmA is more sensitive with the dislocation of the mean as the first central moment of the values than the dispersal standard deviation parameter.
nmA outperforms cA and npA in nonmetric cases. We also consider the comparative performance of cA, npA and nmA. We note that while npA reduce to cA with the Euclidean inputs (9), the relation of the nmA to either cA or npA is not canonical. So any type application of the nmA method on the metric data would not be more than a coercion and hence the results would not be comparable nor conductive of any judgment. Here however, we were interested in examining some of these main data modulation that at least structurally formats the date to be a transferable input across different models. The results showed the almost identical sensitivity of the npA and cA models under two different metric measures while the application of nmA on the modulated data led to a more conservative threshold for the rejection of the null hypothesis (Figure 2C). The expected congruency and high sensitivity of the cA and npA models are mainly due to the correct application of these models on the very data that are within the space of legit inputs of these models. However, nmA being abused on the metric data, our result underlies the rather more conservative sensitivity with regard to the null hypothesis, such that significant judgment for the rejection of it is always supported with the cA and npA while not necessarily the other way around (Figure 2C).

Tailoring the inputs perfectly for the nmA model and this time coercing the data into metric, fit for the npA model, we observed an almost inverse phenomenon (Figure 2D). With the nonmetric data, nmA is more sensitive than the npA with respect to rejection of the null hypothesis. We believe the same reason of inadequate compression of the inputs to form a proper input candidate of the npA model is again the main cause of this discrepancy. In effect, this is confirming the merit of each models in the defined sphere of their underlying assumptions with regards to the input data.

nmA provides solutions for real datasets

Phylogenetic bipartition assessment with BLAST bit scores.

For our first example, we consider a set of 78 angiosperms - flowering plants that bear their seeds in fruits (35). The data is split into two groups: 26 monocots (one seed leaf plants) and 52 eudicots (two seed leaf plants) (Table S3), leading to a total of (78^2) 6084 possible pairings. For each pair, we collect basic local alignment search tool (BLAST) output bit scores of the fundamental chloroplast conserved ribosomal protein S8, rps8 gene. Since BLAST bit score is a semimetric measure (36), the cA and npA are not applicable, as neither a base value for a single gene is available (cA requisite) nor the bit scores for angiosperms represent a metric (npA requisite). A higher bit score is a marker of genetic homogeneity between the species, therefore inverted bit scores could be used as dissimilarity measurement of their phylogeny. Aggregating these values for all the possible pairs of the species led to the nonmetric matrix of dissimilarities (Δ_m,78; SD1). Considering the partitioning of species into monocots and eudicots imposed on the above matrix, we quantify the significance of this bipartition via applying nmA. Our method indicates that the difference between the two species groups is significant with a p-value of 0.044 (F_e = 257.87), which is in accordance with prior findings in plant physiology (37, 38). Indeed, the separation is noticeable on Figure 2E, where we apply multidimensional scaling (MDS) to the data matrix (39). This result suggests the use of nmA for the phylogenetic bifurcating structure assessment that is often based on nonmetric similarity matrices. The higher order indices of nucleotide sequence similarities (e.g. bit score, E-value (40)) are obtained from an underlying nonmetric nucleotide substitution matrices. In fact, all the nucleotide substitution models that allows for nonsymmetric nucleotide substitution rates (F81, HKY85, TN93, and GTR) will essentially result in nonmetric similarity measure between sequences (41). As the metric conditions can be also negated in the case of the codon or amino acids substitution models (such as BLOSUM or PAM), the nmA use can be extended for downstream assessment of partitionings constructed on these models outputs as well (42–45).

Drug combination sensitivity scoring for cell lines. Combination sensitivity scores (CSS) have been introduced to provide information about drug combination efficacy applied to cancer cells (46). After drugs sensitivity estimation, it is usually of interest to see whether particular drug groups lead to similar sensitivity patterns. The directional nature of the CSS scores as to which drug combination is a baseline, renders the CSS quasimetric. Consequently, nmA can be applied to any partitioning of drugs to evaluate its significance. In this example, we apply nmA to assess significance of drug clusters obtained from drug combination experiments. We download relative inhibition (RI) scores (which are sensitivity scores in single drug experiments) and CSS from DrugComb portal (47). We filter data to come from O’Neil drug combination study as it contains the most complete dataset (48). The data provides RI and CSS for 38 unique drugs and 39 cancer cell lines (SD2). For each cell line, we obtain dissimilarity matrices for the 38 drugs, based on the correlations of their CSS1 or CSS2 values (Figure S3). Afterwards, we apply hierarchical clustering on the obtained matrices and fix the number of clusters to 2 aiming to find the most distinctive drug partitioning. To avoid cases when clustering is governed by a small number of drugs (e.g. due to drugs exceptional sensitivity patterns), we apply nmA to dissimilarity matrices that have at least three drugs in each partition. As a result, we identify 7 cell lines out of 39 for which the obtained partitioning is significant (Table 2, Figure S4). For example, the 38 drugs are clearly separated into two clusters for EFM192B cell line (Figure S3), which can be explained by their mechanisms of actions as the majority of the first group compounds are targeted drugs, meanwhile the second group mostly consists of conventional chemother-apy drugs; antineoplastic agents, mitotic and topoisomerase inhibitors (Table S4). The same trend can be seen for the remaining 6 cell lines (Figure S4), which supports the use of nmA as a tool for assessing the significance of the proposed clustering.

| Cell line | N drugs | F-score | p-value |
|-----------|---------|---------|---------|
| VCAP      | 28,10   | 57358   | 0.003   |
| OVCAR3    | 32,6    | 38103   | 0.004   |
| CAOV3     | 22,16   | 2337    | 0.015   |
| HT144     | 19,19   | 777     | 0.025   |
| ES2       | 25,13   | 558     | 0.03    |
| EFM192B   | 24,14   | 436     | 0.034   |
| HCT116    | 25,13   | 220     | 0.048   |

Table 2. Nonmetric ANOVA model specifications that shows significance (p < 0.05) for the CSS data
Kullback-Leibler divergence for gene expression data. For our next example we consider bladder cancer gene expression data obtained for 57 patient samples that were separated into 5 groups: 12 samples representing superficial transitional cell carcinoma with surrounding carcinoma in situ (sTCC+CIS) and 16 samples without surrounding carcinoma (sTCC-CIS), 12 samples within muscle invasive carcinomas (mTCC), 8 samples of normal bladder cells from healthy patients (Normal), and 9 histologically normal samples (Biopsy) close to the carcinoma in situ location (49, 50). We aim to apply nmA to check if the difference between the sample groups is supported with the gene expression-based dissimilarity matrix. To capture the full distributional level of proximity between the gene expressions, we used the Kullback-Leibler (KL) divergence (51) rooted in information theory. This choice is preferred over commonly known point-based methods of differential gene expression such as t-test as its providing a more comprehensive level of information regarding the dispersion of the expressions. We first create a dissimilarity matrix via applying KL divergence scoring on the bladder cancer gene expression data (N = 22,283). We apply unsupervised hierarchical clustering to the dissimilarity matrix (Figure S5) and principal component analysis (PCA) to the raw gene expression counts (Figure 2F) to check if KL-based clustering is in accordance with PCA grouping. It can be seen that KL-based dissimilarity matrix separates 3 carcinoma groups, namely mTCC, sTCC-CIS, and sTCC+CIS, and Normal or Biopsy groups even better than PCA. Therefore, nmA distinguishes both Normal and Biopsy group as significant separation from 3 carcinoma groups (Table S5). At the same time, neither any bipartitions within 3 carcinoma groups nor Normal versus Biopsy partitions are found to be significant (Table S5). The initial partitioning of the bladder gene expression data into 5 groups is significant with p-value = 0.041. Similarly to cA, nmA test provides a significant p-value if at least one distinct bipartition is present in the data.

Discussion

We introduced a generic type of nonmetric ANOVA method that is based on the dissimilarities between a set of objects. These dissimilarities are not needed to be defined with a metric function nor produced systematically. As long as a numeric value representing the closeness of a set of items subject to different partitions are given, the nmA could be applied for...
with reference to their gene expressions proximity. The same where a specific multipartitism with an underlying dissimilarity matrix is given. As pointed in the last example, the nmA can provide a statistical testing for the patient stratification with reference to their gene expressions proximity. The same method of quantification could be applied for labeling the level of distinctiveness of cells in the UMAP or t-SNE plots based on the single cell differential RNA expressions measured with nonmetric methods such as KL divergence (53, 54).

The performance of the nmA to exiting methods including cA and npA also needs to be properly acknowledged. While the cA is solely constructed on the response values and supported by parametric assumptions such as normality of the residuals (7), the npA is formulated on the distances between the responses, free from parametric assumption but still confined to the metric conditions. Similar to npA, nmA is constructed on the dissimilarity measures between responses and devoid of any metric or parametric assumptions. Based on the CLT, the closed distribution of the test statistic for nmA is obtained, allowing a parametric derivation of the p-value. While proportional sampling could provide a more robust outputs for the nmA, we note that the comparative analysis across the discussed models is still primarily biased as lack of a totally objective method in transforming the data to fit perfectly to each model (Table 1). While these transformations are creeping the obtained results and hence, rendering the comparative task ineffective, the retrieval of the responses themselves (as needed for cA) is impossible even with the availability of the dissimilarities or distances between objects (Figure 2D, 1).

We also note the limitation of the nmA in face of growing number of partitions in relation to the number of observations. This is leading to quadratic explosion of the between samples values with regard to within. While the degree of the freedom division in $F_{MV}$ formulation is accommodating for this, the different rate of convergence of the underlying means applied in the CLT is creeping the results toward a non-centrality of the within mean square error leading to a more false negative judgments toward rejecting the null hypothesis (S1). Similar phenomenon could be observed in the case of unbalanced partition sizes where the results would be primarily affected by the biggest partition. In such cases a weighted sampling can be a solution, however in extreme unbalanced scenarios this may not be tractable. Furthermore, the curse of dimensionality aroused in multivariate ANOVA and the unbalanced sampling are also challenging scenarios for cA that need to be properly addressed specifically in each experiment (55, 56) and likely so, nmA could be both developed and adopted in similar scenarios. Taken together, we provide nmA as an attractive method for a variety of scientific questions which involve the assessment of a given partitioning from nonmetric dissimilarity measurements.

**Data availability**

The R code for implementing the cA, npA, and nmA as well as the supplementary data for the case studies are available from github.com/AmiryousefiLab/nmAANOVA.

**Funding**

This work has been supported by the European Research Council (ERC) starting grant agreement (No. 716063) and Academy of Finland grant (No. 317680).

1. RA Fisher, Studies in crop variation. I. an examination of the yield of dressed grain from broadbalk. J. Agric. Sci. 11, 107 (1921).

2. G Sklirr,er, Validation of rapid assessment of damage in urban mangrove forests and relationships with multispectral and various models. J. Mar. Biol. Assoc. United Kingdom 76, 701 (1996).

3. A Rosenblad, Applied multivariate statistics for the social sciences, fifth edition by James p. stevens. Int. Stat. Rev. 77, 476–476 (2009).

4. RG Shaw, T Mitchell-Olts, Anova for unbalanced data: An overview. Ecology 74, 1638 (1993).

5. WH Kruskal, WA Warranty, Use of ranks in one-criterion variance analysis. J. Am. Stat. Assoc. 47, 583–621 (1952).

6. BM Katz, M McSweeney, A multivariate kruskal-walls test with post hoc procedures. Multivar. Behav. Res. 15, 281–297 (1990).

7. DC Montgomery, Design and Analysis of Experiments. (John Wiley & Sons), (2012).

8. BH McAlindon, MJ Anderson, Fitting multivariate models to community data: A comment on distance based redundancy analysis. Ecology 82, 290–297 (2001).

9. MJ Anderson, A new method for non-parametric multivariate analysis of variance. Austral. Ecol. 26, 32–46 (2001).

10. MJ Anderson, KE Ellersieck, BH McAlinda, Multivariate dispersion as a measure of beta diversity. Ecol. Lett. 9, 683–693 (2006).

11. MJ Anderson, Distance-based tests for homogeneity of multivariate dispersions. Biometrics 62, 545–562 (2005).

12. P Legendre, ED Gallagher, Ecologically meaningful transformations for ordination of species data. Oecologia 129, 271–280 (2001).

13. R Mill, et al., Network motifs: simple building blocks of complex networks. Science 298, 824–827 (2002).

14. M Santtilni, AL Bababasi, Predicting perturbation patterns from the topology of biological networks. PloS. Natl. Acad. Sci. 115, E6375–E6383 (2018).

15. M Jatni, Y Wang, A Amryouso, J Tang, Unsupervised learning and multipartite network models: a promising approach for understanding traditional medicine. Front. pharmacology 11, 1319 (2020).

16. G Castella, RL Berger, Statistical inference. (Cengage Learning), (2021).

17. A Tversky, Features of similarity. Psychol. Rev. 84, 327–352 (1977).

18. JC Gower, P Legendre, Metric and euclidean properties of dissimilarity coefficients. J. Classif. 3, 5–48 (1986).

19. S Rolewicz, Functional Analysis and Control Theory (Springer Science), (2021).

20. D Burago, Y Burago, S Ivanov, A Course in Metric Geometry. (American Mathematical Society), (2001).

21. J Väisälä, Gromov hyperbolic spaces. Expo. Math. 23, 187–231 (2005).

22. Q Xia, The geodesic problem in quasimetric spaces. J. Geom. Analysis 19, 452–479 (2009).

23. P Gromov, Non-metric statistical analyses: Some metric alternatives. J. Stat. Plan. Inference 13, 377–387 (1986).

24. JB Kruskal, Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. Psychometrika 29, 1–27 (1964).

25. M Mantel, The detection of disease clustering and a generalized regression approach. Cancer research 27, 209–220 (1967).

26. PM Meer, D Mintz, A Rosenfeld, DY Kim, Robust regression methods for computer vision: A review. Int. J. Comput. Vis. 5, 59–70 (1991).

27. G Russolillo, Non-metric partial least squares. Electron. J. Stat. 6, 1641–1669 (2012).

28. NR Langley, B Dudzik, A Cloutier, A decision tree for nonmetric sex assessment from the skull. J. Forensic sciences 63, 31–37 (2018).

29. RG Lehnen, GG Koch, A general linear approach to the analysis of nonmetric data: Applications for political science. Am. J. Polit. Sci. 283–313 (1974).

30. G Matthews, R Matthews, W Landis, Nonmetric clustering: New approaches for ecological data. In Proceedings of the Tenth Conference on Artificial Intelligence for Applications. (IEEE), pp. 482–483 (1994).

31. J Borradaire, The regrettor principle of information. J. Appl. Phys. 112, 1165–1163 (1953).

32. A Hyvärinen, E Oja, Independent component analysis: algorithms and applications. Neural networks 13, 411–430 (2000).

33. A Gelman, JB Carlin, HS Stern, DB Rubin, Bayesian data analysis. (Chapman and Hall/CRC), (1996).

34. JO Berger, T Sellke, Testing a point null hypothesis: The irreconcillability of p values and evidence. J. Am. statistical Assoc. 82, 112–122 (1987).

35. RMS Dahlgren, A revised system of classification of the angiosperms. Bot. J. Linnean Soc. 90, 91–124 (1984).

36. AE Hirsh, HFreiberger, Protein dispensability and rate of evolution. Nature 411, 1046–1049 (2001).

37. DW Taylor, LHokey, Flowering Plant Origin, Evolution & Phylogeny. (Springer Science, Business Media), (1996).

38. MW Chase, An update of the angiosperm phylogeny group classification for the orders and families of flowering plants: APG II. Bot. J. Linnean Soc. 141, 399–436 (2003).

39. T Cox, M Cox, Multidimensional Scaling, Second Edition. (Chapman and Hall/CRC), (2000).
\( F_{\alpha\beta} \) as the outcome of the dissimilarity function \( \delta(y_{\alpha}, y_{\beta}) \) where \( y \) is either a scalar or a vector of the interested response. Collecting all the pairwise dissimilarities in a square matrix \( \Delta_{N \times N} = \{\delta_{\alpha\beta}\} \) forms the dissimilarity matrix between object as the counterpart of diagonal matrix in npA \((S3.2)\). For each \( j' = 1, 2, \ldots, g \) and \( j \neq j' \), let’s denote \( \delta_{jj} \) and \( \delta_{jj'} \) as the mean dissimilarity measures of the diagonal and off-diagonal sub-matrices of the \( \Delta \) and \( \delta_{jj'} \)prop as the mean of a portion of dissimilarity measures in all diagonal matrices, which comprises the same fraction of diagonal matrices as the number of elements in the current sub-matrix in all off-diagonal sub-matrices. It is important to note that \( \delta_{jj'} \)prop represents the means of proportional fractions that contain elements which do not overlap for any combination of \( j \) and \( j' \). The latter is needed to ensure independence of the sum elements in equation 2. Under the null hypothesis of similarity of all partition means, the \( F_{\alpha\beta} \) as presented below is following the above calculation.

\[
F_{\alpha\beta} = \frac{SNM_{\alpha\beta}/g}{SNM_{W}/(g-1)} \sim H_0 \ F(x)_{g^2-g-1},
\]

where \( SNM_{\alpha\beta} \) and \( SNM_{W} \) are the sums of squared normalized means of dissimilarity values between and within partitioning groups defined as;

\[
SNM_{\alpha\beta} = \sum_{j=1}^{g} (n_j (\delta_{jj} - \delta_{W}))^2
\]

\[
SNM_{W} = \sum_{j=1}^{g} \sum_{j' \neq j} (\delta_{jj'} - \delta_{jj'} \text{prop})^2
\]

where \( \delta_{W} \) is the overall mean value of all the dissimilarity values in the partitioning groups \( \sum_{\alpha} \sum_{j} \theta_{\alpha j} \sigma_{\alpha j} / \sum_{\alpha} \sum_{j} \sigma_{\alpha j} \)

\( \epsilon \) is an indicator function defined in 13. \( N_{off} \) is the number of elements in all off-diagonal sub-matrices and \( N_{diag} \) is the number of elements in all diagonal sub-matrices.

### S2. Proof of \( F_{\alpha\beta} \) distribution. Based on the central limit theorem the mean of \( \bar{\delta}_{\alpha\beta} \) unknown independently distributed \( \delta_{\alpha\beta} \) with population mean \( \mu \) and variance \( \sigma^2 \) is normally distributed hence;

\[
\forall j \delta_{jj} \sim N(\mu, \sigma^2/n^2),
\]

Normalization of these leads to,

\[
\forall j Z_j = \frac{n_j (\delta_{jj} - \mu)}{\sigma} \sim N(0,1),
\]

Now under null hypothesis of the sameness of the competing groups mean with each other \( H_0 : \forall j, j' \in \{1, \ldots, g\}, \mu_j = \mu_j' = \mu \) and hence \( \delta_{11} = \delta_{12} = \delta_{22} = \ldots = \delta_{gg} = \delta_W \), the sum of squared values of the above would be distributed as the \( \chi^2(g-1) \) as,

\[
\chi^W = \sum_{j=1}^{g} (n_j (\delta_{jj} - \delta_W))^2 \sim H_0 \ \chi^2(g-1),
\]

This is true since the sums of squared \( g \) standard normal random variables is Chi-squared distributed.

Following the same logic for dissimilarities between groups we have,

\[
\chi^B = \sum_{j=1}^{g} \sum_{j' \neq j} (\delta_{jj'} - \bar{\delta}_{jj'} \text{prop})^2 \sim H_0 \ \chi^2(g^2-g-1),
\]

where in that tacitly we have used the equation of the mean dissimilarities between groups to be equal with that one as the within groups, and for the same reason albeit unlike \( \chi^W \), we are not losing an extra degree of freedom in \( \chi^B \). Again with noticing that the division of two Chi-squared distributed divided by their corresponding degrees of freedom is \( F \)-distributed we have

\[
F_{\alpha\beta} = \frac{\chi^B / (g^2-g)}{\chi^W / (g-1)} \sim H_0 \ F(g^2-g-1),
\]

or with rearranging of the values in the above we have,

\[
F_{\alpha\beta} = \frac{1}{\frac{1}{\chi^B} \sum_{j=1}^{g} \sum_{j' \neq j} (\delta_{jj'} - \bar{\delta}_{jj'} \text{prop})^2 / g(g-1)}
\]

\[
\frac{1}{\frac{1}{\chi^W} \sum_{j=1}^{g} (n_j (\delta_{jj} - \delta_W))^2 / (g-1)} .
\]

that with crossing out the \( 1/\sigma^2 \) and \( g-1 \) the (2) is obtained. Note that the \( F_{\alpha\beta} \) is capturing the amount of noncentrality incurred by degree of deflection of the between mean dissimilarities in contrast to the within. To the extend of this deflection, \( F_{\alpha\beta} \) would be proportionally inflated and if such observed value \( F_{\alpha\beta} \) is more than certain degree \( P(F_{\alpha\beta} > F_0) \), it stands as the existence of evidences for significant differences between the competing groups.
S3. Essential notes on metric ANOVA. Consider N data points, realizations of a random variable Y, and indexed by \( \alpha, \beta, \theta \in \{1, \ldots, N\} \). We further define a metric function \( d : S \to [0, \infty) \) as a set of all metric spaces spanned by a metric \( d(S, S) : S \times S \to [0, \infty) \) on its domain set \( S \). Note that the functions in \( S \) are not necessarily surjective. The metric function defines distances between data points, \( (y_\alpha, y_\beta) \). An \( N \times N \) distance matrix \( D \) expresses all pairwise distances between data points. The distances in \( D \) follow a metric if and only if for all \( \alpha, \beta, \theta \in \{1, \ldots, N\} \) the triplets \( (\alpha, \beta, \theta) \) fulfill the triangle inequality \( d(\alpha, \beta) \leq d(\alpha, \theta) + d(\theta, \beta) \) (18). From swapping the arguments two more equality arise that the collectively with the triangle inequality are defined as metric conditions as:

1. Identity: \( d(\alpha, \alpha) = 0 \) \( \iff \) \( \alpha = \beta \),
2. Symmetry: \( d(\alpha, \beta) = d(\beta, \alpha) \),
3. Subadditivity: \( d(\alpha, \beta) \leq d(\alpha, \gamma) + d(\gamma, \beta) \).

\[ \sum_{j=1}^{N} \sum_{i=1}^{N} d_{ij}^2 = \sum_{j=1}^{N} \sum_{i=1}^{N} d_{ij}^2 \leq \sum_{j=1}^{N} \sum_{i=1}^{N} d_{ij}^2 + \sum_{j=1}^{N} \sum_{i=1}^{N} d_{ij}^2 \]

In contrast to partitioning in terms of the proportion of variance explained by using the decomposition of sums of squares which partitions the total sums of squares \( S_S \) to between and within groups sums of squares, \( S_B \) and \( S_W \), respectively (7):

\[ S_S = S_B + S_W \quad \text{(11)} \]

In cA the distances are Euclidean. For the sake of simplicity and without loss of generality, we only use one-dimensional response variable in the following, that is, \( p = 1 \). In this setup the elements of Equation (11) are defined as:

\[ S_S = \sum_{j=1}^{N} \sum_{i=1}^{N} (y_\alpha - \bar{y})^2 \]
\[ S_B = \sum_{j=1}^{N} \sum_{i=1}^{N} (y_j - \bar{y})^2 \]
\[ S_W = \sum_{j=1}^{N} \sum_{i=1}^{N} (y_\alpha - \bar{y})^2 \]

where \( \bar{y} \) and \( \bar{y}_j \) are overall mean of the responses and mean of responses in the \( j \)th group, respectively.

2. Nonparametric ANOVA (npA). npA considers pairwise distance matrices between objects (9). The distance matrix is given by a metric function \( d \) on the responses such that \( d : Y \times Y \to [0, \infty) \), where \( Y \) is a realization of a \( p \) dimensional random variable, given by \( y = (y_1, y_2, \ldots, y_p) \in Y \) for all \( i \) and \( j \). We are interested in distances that besides their domain, form an element of \( S \). All the outcomes of the distance metric function can be collected into a pairwise distance matrix \( D \). We refer to the values of test statistic based on the observed outcomes of the experiments as \( F_0 \). This value then will be compared against the null \( F \) distribution to obtain the \( P \)-value, \( P = \Pr(F_g \geq F_0) \). Rejection of the null hypothesis indicates the presence of at least one group with mean value differing from the others. With the similar indication for the \( F \)-value, in npA a pseudo permutative \( F \)-statistic is defined as:

\[ pF_\pi = \frac{S_B^\pi / \pi - 1}{S_W^\pi / \pi} \sim F_\pi \quad \text{pseudoF}, \quad \text{(15)} \]

Indexing each values of \( pF_\pi \) obtained from \( \pi \) permutation as \( pF_\pi^{\pi} \) the \( P \)-value can be obtained as \( P_\pi = \pi \sum \frac{1_{\{F_\pi^{\pi} > F_0\}}(x)}{\pi} \) where \( 1_{\{A\}}(x) \) is an indicator function (that is 1 if \( x \in A \) and 0 otherwise), and \( F_\pi \) is the observed pseudo \( F \)-statistic.

S4. Data. The R code for implementing the npA algorithm as well as the supplementary data SD1, SD2, and SD3 are available from https://github.com/AmiryousefLab/nmAANOVA.
Fig. 3. Significant bipartitioning of CSS-based dissimilarity matrix for EFM192B cell line. The dissimilarity matrix is achieved by first determining the quasimetric $38 \times 38$ CSS-based matrix with rows and columns corresponding to the drugs. At the same time, we fill the diagonal values with RI scores, as they bring additional information about the drugs’ sensitivity. Then obtaining two directional correlation-based matrices; we subtract the absolute values of a correlation matrix applied to either CSS-based matrix or to its transpose version, from 1. After that, we create a joint matrix via filling upper and lower triangular parts of the new matrix with the upper triangles of the obtained symmetric matrices. Thus, we take into account correlation-based drug similarities as nonsymmetric. We also fix diagonal values of the matrix to zero given that the row and column drug is the same. The obtained EFM192B specific bipartitioning, which is marked by the two black lines on the dissimilarity matrix based heatmap, resulted in $p$-value = 0.034.
In order to obtain a dissimilarity matrix for the data, we apply KL divergence represented via heatmaps, has been marked as significant by nmA.

Fig. 4. Significant bipartitioning of CSS-based dissimilarity matrices for 6 cell lines. The marked bipartitioning of the cell line specific dissimilarity matrices which are represented via heatmaps, has been marked as significant by nmA.

Fig. 5. KL divergence-based dissimilarity matrix for the bladder gene expression data. In order to obtain a dissimilarity matrix for the data, we apply KL divergence scoring. We first transform sample specific gene expression distributions into probability distributions and then obtain KL divergence values for each distribution pair. Since KL divergence calculation is dependent on the order of the two distributions, it is non symmetric. We visualize the obtained dissimilarity matrix via a heatmap of its logarithmic values. We also apply unsupervised hierarchical clustering to show relation between the samples. We find out that dissimilarity matrix separates 3 carcinoma groups (mTCC, sTCC-CIS and sTCC+CIS) and Normal or Biopsy samples. These separations have been marked as significant by nmA with $p$-values listed in Table 5.
Table 3. Set of 78 angiosperms partitioned into 52 eudicots and 26 monocots that was used to form a semimetric matrix of dissimilarities based on the inverse values of the bit scores of rps8 chloroplast gene for all the possible pairwise comparisons

| Eudicots                        | Monocots                          |
|---------------------------------|-----------------------------------|
| Aethionema grandiflorum         | Acidosasa purpurea                |
| Ageratina adenophora            | Acorus calamus                    |
| Anthriscus cerefolium           | Agrostis stolonifera              |
| Arabidopsis thaliana            | Anomochloa marantoiidea           |
| Arabis hirsuta                  | Bambusa emeiensis                 |
| Arbutus unedo                   | Brachypodium distachyon           |
| Atropa belladonna                | Ferrocalamus rimosiaquinus        |
| Barbara verna                   | Festuca arundinacea               |
| Betula pendula                  | Hordeum vulgare                   |
| Brassica napus                  | Indocalamus longiauritus          |
| Capsella bursapastoris          | Leersia tisserantii               |
| Carica papaya                   | Lemma minor                       |
| Castaneamollissima              | Lolium perenne                    |
| Citrus sinensis                 | Oryza meridionalis                |
| Coffea arabica                  | Panicum virgatum                  |
| Crucihimalaya wallichii         | Phalaenopsis aphrodite            |
| Cucumis melo                    | Phyllostachys nigra               |
| Daucus carota                   | Rhyynchoryza subulata             |
| Eleutherococcus senticosus      | Saccharum hybrid                  |
| Eucalyptus grandis              | Sorghum bicolor                   |
| Fagopyrum esculentum            | Spirodela polyrhiza               |
| Fragaria vesca                  | Triticum aestivum                 |
| Glycine max                     | Typha latifolia                   |
| Gossypium raimondii             | Wolffiella australiana            |
| Guizotia abyssinica             | Wolfiella lingulata               |
| Helianthus annuus               | Zea mays                          |
| Hevea brasiliensis              |                                   |
| Ipomoea purpurea                |                                   |
| Jaconbae vulgaris               |                                   |
| Lactuca sativa                  |                                   |
| Lepidium virginicum             |                                   |
| Lobularia maritima              |                                   |
| Lotus japonicus                 |                                   |
| Manihot esculenta               |                                   |
| Medicago truncatula             |                                   |
| Millettia pinnata               |                                   |
| Morus indica                    |                                   |
| Nasturtium officinale           |                                   |
| Nelumbo nucifera                |                                   |
| Nicotiana tomentosiformis       |                                   |
| Oenothera elata                 |                                   |
| Olea europaea                   |                                   |
| Olivarabidopsis pumila          |                                   |
| Panax ginseng                   |                                   |
| Parthenium argentatum           |                                   |
| Populus trichocarpa             |                                   |
| Prunus persica                  |                                   |
| Pyrus pyrifolia                 |                                   |
| Silene noctiflora               |                                   |
| Solanum lycopersicum            |                                   |
| Spinacia oleracea               |                                   |
| Theobroma cacao                 |                                   |
Table 4. Mechanism of action/target and class information for the drugs used in CSS example

| Drug               | Mechanism/Targets                  | Class                               |
|--------------------|------------------------------------|-------------------------------------|
| 391210-10-9        | MEK inhibitor                       | Kinase inhibitor                    |
| 5-FLUOROURACIL     | Antineoplastic antimetabolite       | Conventional chemotherapy           |
| 891494-63-6        | CHK1 inhibitor                      | Kinase inhibitor                    |
| 915019-65-7        | PI3K and mTOR inhibitor             | Kinase inhibitor                    |
| BORTEZOMIB         | Proteasome inhibitor                | Protease/proteasome inhibitor       |
| CARBOPLATINUM      | Platinum-based antineoplastic agent | Conventional chemotherapy           |
| CYCLOPHOSPHAMIDE   | Alkylating agent                    | Conventional chemotherapy           |
| DASATINIB          | Broad kinase inhibitor              | Kinase inhibitor                    |
| DEBOROLOMUS        | mTOR inhibitor                      | Kinase inhibitor                    |
| DEXAMETHASONE      | Glucocorticoid, immunomodulatory    | Immunomodulatory                    |
| DINALCRICLIB       | CDK inhibitor                       | Kinase inhibitor                    |
| DOXORUBICIN        | Topoisomerase II inhibitor          | Conventional chemotherapy           |
| ELOXATIN (TN)      | Platinum-based antineoplastic agent | Conventional chemotherapy           |
| ERLOTINIB          | EGFR inhibitor                      | Kinase inhibitor                    |
| ETOPOSIDE          | Topoisomerase II inhibitor          | Conventional chemotherapy           |
| GELDANAMYCIN       | HSP90 inhibitor                     | Conventional chemotherapy           |
| GEMCITABINE        | Antimetabolite                      | Conventional chemotherapy           |
| L-778123 FREE BASE | FPTase and GGPTase-I inhibitor      | Conventional chemotherapy           |
| LAPATINIB          | HER2, EGFR inhibitor                | Kinase inhibitor                    |
| METFORMIN          | AMPK activator                      | Metabolic modifier                  |
| METHOTREXATE       | Antimetabolite                      | Metabolic modifier                  |
| MITOMYCIN C        | Antineoplastic anatibiotic          | Conventional chemotherapy           |
| MK-1775            | WEE1 inhibitor                      | Kinase inhibitor                    |
| MK-2206            | AKT inhibitor                       | Kinase inhibitor                    |
| MK-4541            | Androgen receptor modulator         | Hormone therapy                     |
| MK-5108            | Aurora A inhibitor                  | Kinase inhibitor                    |
| MRK003             | Gamma secretase inhibitor           | Gamma secretase inhibitor           |
| NIRAPARIB          | PARP inhibitor                      | Differentiating/ epigenetic modifier|
| PACLITAXEL         | Mitotic inhibitor, taxane microtubule stabilizer | Conventional chemotherapy |
| SN-38              | Topoisomerase I inhibitor           | Conventional chemotherapy           |
| SORAFENIB          | Broad kinase inhibitor              | Kinase inhibitor                    |
| SUNITINIB          | Broad kinase inhibitor              | Kinase inhibitor                    |
| TEMOZOLOMIDE       | Alkylating agent                    | Conventional chemotherapy           |
| TOPOTECAN          | Topoisomerase I inhibitor           | Conventional chemotherapy           |
| VELIPARIB          | PARP inhibitor                      | Differentiating/ epigenetic modifier|
| VINBLASTINE        | Mitotic inhibitor, vinca alkaloid   | Conventional chemotherapy           |
| VINORELBINE        | Mitotic inhibitor, vinca alkaloid   | Conventional chemotherapy           |
| VORINOSTAT         | HDAC inhibitor                      | Differentiating/ epigenetic modifier|

Table 5. Nonmetric ANOVA model specifications for the bladder gene expression data

| N partitions | Partitions                          | F  | p-value |
|--------------|-------------------------------------|----|---------|
| 5            | Normal | Biopsy | mTCC | sTCC-CIS | sTCC+CIS | 7  | 0.041 |
| 2            | Biopsy | mTCC, sTCC-CIS, sTCC+CIS | 2738 | 0.014 |
| 2            | Normal | mTCC, sTCC-CIS, sTCC+CIS | 187 | 0.05   |
| 2            | mTCC   | sTCC+CIS | 1.73 | 0.47   |
| 2            | mTCC   | sTCC-CIS | 0.3  | 0.79   |
| 2            | sTCC-CIS | sTCC+CIS | 0.43 | 0.43   |