sinc(i) Research Institute for Signals, Systems and Computational Intelligence (fich.unl.edu.ar/sinc)
M. Pividori, A. Cernadas, L. A. Haro, F. Carrari, G. Stegmayer & D. H. Milone; "Clustermatch: discovering hidden relations in highly diverse kinds of qualitative and quantitative data without standardization"
Bioinformatics, 2018.

Clustermatch: discovering hidden relations in highly diverse kinds of qualitative and quantitative data without standardization

Milton Pividori [1,2,3*], Andres Cernadas [4], Luis A. de Haro [4],
Fernando Carrari [4,5], Georgina Stegmayer [1] and Diego H. Milone [1]

[1]Research institute for signals, systems and computational intelligence, CONICET-UNL, Santa Fe, 3000, Argentina;
[2]Section of Genetic Medicine, The University of Chicago, Chicago, IL 60637, USA;
[3]Center for Data Intensive Science, The University of Chicago, Chicago, IL 60615, USA;
[4]Institute of Biotechnology, INTA-Castelar, CONICET, Argentina and
[5] Departamento de Botânica, Instituto de Biociências, Universidade de São Paulo, Rua do Matão, 277, São Paulo 05508-090, Brazil.

Abstract

Motivation: Heterogeneous and voluminous data sources are common in modern datasets, particularly in systems biology studies. For instance, in multi-holistic approaches in the fruit biology field, data sources can include a mix of measurements such as morpho-agronomic traits, different kinds of molecules (nucleic acids and metabolites) and consumer preferences. These sources not only have different types of data (quantitative and qualitative), but also large amounts of variables with possibly non-linear relationships among them. An integrative analysis is usually hard to conduct, since it requires several manual standardization steps, with a direct and critical impact on the results obtained. These are important issues in clustering applications, which highlight the need of new methods for uncovering complex relationships in such diverse repositories.
Results: We designed a new method named Clustermatch to easily and efficiently perform data-mining tasks on large and highly heterogeneous datasets. Our approach can derive a similarity measure between any quantitative or qualitative variables by looking on how they influence on the clustering of the biological materials under study. Comparisons with other methods in both simulated and real datasets show that Clustermatch is better suited for finding meaningful relationships in complex datasets.
Availability: Files can be downloaded from
https://sourceforge.net/projects/sourcesinc/files/clustermatch/ and
https://bitbucket.org/sinc-lab/clustermatch/. In addition, a web-demo is available at
http://sinc.unl.edu.ar/web-demo/clustermatch/
Contact: mpividori@sinc.unl.edu.ar
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

In systems biology studies, any given experimental unit can be characterized according to multiple and heterogeneous analyses implying different techniques. For example in crop plants, morpho-agronomic traits are used for general characterization; biochemical composition of fruits obtained from diverse techniques like gas chromatography-mass spectrometry (GC-MS), nuclear magnetic resonance (NMR) and high performance liquid chromatography (HPLC) are employed to quantify fruit soluble and volatile metabolites; tasting panels are used to evaluate consumer preferences like aroma and texture, among many others kinds of measures of different nature. Moreover, these data might be collected not only during one year, but also during several growing seasons.

In all cases, highly-diverse kinds of quantitative measures or variables and qualitative annotations must be related among them in order to discover hidden relations to infer new knowledge. Indeed, pattern analysis on modern genomics data, with multiple sources and high volumes of information, strongly need new integratives techniques and models to better understand how different entities are related to each other (Li *et al.*, 2018). The traditional approach of applying a clustering algorithm, such as k-means, PAM, spectral clustering (Xu and Wunsch, 2009) to integrate heterogeneous variables requires a very complex, manual and time-consuming pre-processing to standardize each particular source of data. The pre-processing has to be done one-by-one, according to each particular

sinc(i) Research Institute for Signals, Systems and Computational Intelligence (fich.unl.edu.ar/sinc)
M. Pividori, A. Cernadas, L. A. Haro, F. Carrari, G. Stegmayer & D. H. Milone; "Clustermatch: discovering hidden relations in highly diverse kinds of qualitative and quantitative data without standardization"
Bioinformatics, 2018.

input data source and measurement technique. This has many important issues to be considered: i) each data source and technology of measurement needs a proper and intrinsic, possibly complex standardization or scaling; ii) this scaling must make all sources comparable to each other in order to use a single clustering method on an integrated input table, otherwise one or several sources would dominate the clustering solutions; iii) and finally but not less important, these tasks tend to be manual (almost hand-crafted), very time-consuming and require, furthermore, very specific knowledge in order to properly choose among the myriad of all possible transformation and normalization methods. For example, does each input data source adjust to some well-known distribution? Is it possible to assume a normal or gaussian distribution? Is it enough to apply mean extraction and divide by standard deviation? Or is it better to transform the data to a logarithmic scale? In the case of qualitative data, such as for example tasting panels, this issue is even worse. How to treat it? How to make it quantitative? And, furthermore, how to relate it with the other quantitative data? For example, relating variables such as the *flavor* of fruits (for instance, with categories like: "frutal", "sweet" and "acid"), its *glucose* contents (with numeric values: 0.69, 0.07 and 0.34), and its *size* ("small", "regular", "large") represents a real challenge today.

There are many proposals in the literature regarding how to relate two variables, typically denominated tests of dependence. Correlation describes a broad class of statistical relationships, including dependence, and may be useful to pinpoint a predictive relationship of interest. The very well-known Pearson's correlation coefficient (Devlin *et al.*, 1975) is the most commonly used correlation method. It detects only linear relationships and is prone to generate potentially misleading values in the presence of outliers and non-linear transformations of the data (Huber, 2011). Yet another common correlation coefficient is the Spearman rank, a nonparametric measure of statistical dependence between two variables defined as the Pearson's correlation between the ranked variables (Spearman, 2010), which assesses whether two variables are monotonically related, even if their relationship is not linear. Because of their simplicity and speed of calculation, both Pearson and Spearman have been the most commonly applied correlation methods in literature. Among the most recently proposed measures of dependence, Distance correlation (DC) (Szkely *et al.*, 2007) was introduced to address deficiencies of Pearson's coefficient, which can easily be zero for dependent variables. DC measures both linear and non-linear associations between two variables, and has been applied in variable selection (Li *et al.*, 2012) and life sciences (Kong *et al.*, 2012). Another recent method is the Maximal Information Coefficient (MIC) (Reshef *et al.*, 2011), a measure of dependence able to capture a wide range of relationships between random variables. Although MIC has gained considerable attention (Nature, 2012; Speed, 2011; Zhang *et al.*, 2014), there were also several discussions about some of its properties (N. Simon, 2011; Kinney and Atwal, 2014; Reshef *et al.*, 2014). One of the main issues resides in the computational cost of MIC's original implementation: a dynamic programming algorithm called ApproxMaxMI that several studies in the literature tried to optimize (Albanese *et al.*, 2013; Zhang *et al.*, 2014; Tang *et al.*, 2014; Chen *et al.*, 2016). Apart from these issues, all the mentioned methods need categorical data to be converted to numerical in order to be applied, which cannot be done in many cases with non-ordinal variables. These are important drawbacks for clustering, where the goal is to find hidden relationships between the variables. Thus, more efficient methods are needed to process large datasets in a reasonable amount of time. Moreover, methods should also have sufficient generality to capture non-trivial relationships in highly heterogeneous data sources.

To the best of our knowledge, up to date there is not a single and simple method available for the easy fusion of such diverse complex data to perform an integrative analysis, without requiring previous manual standardization steps. Thus, we have developed a new type of clustering algorithm for wider use in any application domain, that can analyze the variables but without any standardization. Since similarities between highly diverse variables of interest cannot be directly calculated, we propose to look at how the variables influence the clustering of the biological materials or objects, considering each variable separately. Thus, the first step is obtaining internal partitionings of the objects according to each variable; after that, the following step is to obtain a similarity value between each pair of variables by comparing these internal partitions of the objects. Finally, these similarities are used for clustering the variables.

Clustermatch is a novel method for cluster analysis on variables, providing a unified similarity metric that can compute a score for all combinations of numerical, categorical and ordinal variables. By using the proposed metric we can compute a similarity matrix between all variables, with no previous pre-processing required, and then run a clustering algorithm to derive the final partition of the variables. Our implementation can also smoothly handle cases where a variable was not measured for all objects or in all the experimental conditions evaluated. The main advantage of our proposal is that, while still efficiently detecting complex relationships between variables, it does not require a specific standardization nor transformation of any input data. It enables simple and straightforward integration of categorical as well as different types of numerical variables.

sinc(i) Research Institute for Signals, Systems and Computational Intelligence (fich.unl.edu.ar/sinc)
M. Pividori, A. Cernadas, L. A. Haro, F. Carrari, G. Stegmayer & D. H. Milone; "Clustermatch: discovering hidden relations in highly diverse kinds of qualitative and quantitative data without standardization"
Bioinformatics, 2018.

Table 1: Example of dataset with highly diverse kinds of qualitative and quantitative data. Biological materials or objects are in columns (numbers are sample identifiers) and heterogeneous variables in rows. For each variable the internal clustering of the objects is indicated with different colors.

	549	550	551	715	2523	3806	4750
Agronomics							
Color	intense	regular	muted	intense	muted	muted	regular
Width	97.52	73.31	82.74	49.84	65.30	NA	67.88
Height	77.51	54.28	50.11	51.03	51.67	57.13	NA
...	...	...	...	...	...	...	...
Sensory panels							
Flavour	charact.	sweet	acid	charact.	acid	sweet	sweet
Juiciness	1	3	2	3	2	3	1
Aroma	moldy	floral	herbal	floral	moldy	herbal	floral
...	...	...	...	...	...	...	...
Volatile metabolites							
pinene	0.102	NA	0.114	0.226	0.165	0.042	0.092
cis-3-h.	35.06	53.99	19.03	33.11	16.03	63.04	52.67
hexanal	592.92	198.46	414.85	353.60	834.38	NA	740.00
...	...	...	...	...	...	...	...

2 The Clustermatch method

2.1 Motivation and approach

Let us suppose a large systems biology study involving many measures of a crop plant of biotechnological interest, studied along several years, harvests, seasons and different geographic zones. Our main hypothesis is that if two variables, for example, a metabolite concentration and an agronomic trait, consistently produce a similar clustering of the same biological materials along several repetitions, then a similarity between those variables can be numerically inferred. In Table 1 we present an example showing the structure of a biological data set, where materials (objects) are shown in the columns, and variables from different data sources are in the rows. In this particular example, the analyzed materials are tomato (*Solanum lycopersicum*) fruits harvested from different plants of different germplasm bank accessions identified by their passports (549, 550, 551, etc), and values correspond to measures of a set of morpho-agronomic and biochemical traits from different sources (agronomic traits, sensory panels, etc), resulting in a highly diverse set of variables. The heterogeneity of these data resides not only on possibly different linear or non-linear transformations among continuous measures, but also on the intrinsic nature of the data itself, which could be either quantitative or qualitative (e.g., flavour). It is important to note that the qualitative data could include not only ordinal (e.g., juiciness), but also nominal variables (e.g., aroma and color), where there is no natural order between the categories and thus, they cannot be sorted.

For the example in Table 1, the objects (columns) are partitioned considering each variable (rows) separately. Thus, in the first step one or more partitionings of the objects are obtained for each variable; then, a similarity value between each pair of variables is obtained by comparing the partitions. In Table 1, the internal partition of the objects for each variable is indicated with a different color for each cluster (blue, green and yellow). It can be seen that most of the variable pairs do not match. However, as an example, some of them present a higher degree of similarity: *color*, *flavour* (both categorical variables), and the metabolite *cis-3-hexenal* (numerical variable), where their internal partitions of the objects have been obtained with 3 clusters.

For each categorical variable, objects are grouped together if they share the same category. For example, for *flavour* in Table 1, objects that share the same flavour are grouped together in the same cluster: objects 549 and 715 belong to the “characteristic flavour” cluster (in green), whereas 550, 3806 and 4750 are in the “sweet flavour” cluster (in yellow) and 551 and 2523 are in the “acid flavour” cluster (in blue). Note that this partition with 3 clusters matches only partially to the one calculated according to the *color* variable. That is, the internal partition of the biological materials, according to the categorical variable *color*, does not coincide with the clusters of the objects obtained with the *flavour* variable.

For numerical data, we propose to partition the objects by using a simple and fast one-dimensional clustering algorithm based on quantiles. For *cis-3-hexenal*, it can be seen that the objects are grouped

sinc(i) Research Institute for Signals, Systems and Computational Intelligence (fich.unl.edu.ar/sinc)
M. Pividori, A. Cernadas, L. A. Haro, F. Carrari, G. Stegmayer & D. H. Milone; "Clustermatch: discovering hidden relations in highly diverse kinds of qualitative and quantitative data without standardization"
Bioinformatics, 2018.

Algorithm 1: Clustermatch

```
Input:
    $D$: data matrix of $M \times N$
    $R$: maximum number of internal clusters
    $k$: number of output clusters
Output:
    $\Pi$: partition of $M$ variables into $k$ clusters
1  begin
2      for $i \leftarrow 1$ to $M$ do
3          if $\mathbf{d}_i \in \mathbb{R}^N$ then
4              for $r \leftarrow 2$ to $\min\{R, |\mathbf{d}_i| - 1\}$ do
5                  $\boldsymbol{\rho} \leftarrow (\rho_\ell \mid \Pr(d_{ij} < \rho_\ell) \leq (\ell - 1)/r), \forall \ell \in [1, r+1]$
6                  $\Omega_{ir\ell} \leftarrow \{j \mid \rho_\ell < d_{ij} \leq \rho_{\ell+1}\}, \forall \ell \in [1, r]$
7          else
8              $\mathcal{C}_i \leftarrow \cup_j \{d_{ij}\}$
9              $r \leftarrow |\mathcal{C}_i|$
10             $\Omega_{irc} \leftarrow \{j \mid d_{ij} = \mathcal{C}_{ic}\}, \forall c \in [1, r]$
11
12         $S_{mn} \leftarrow \max_{\forall p,q}\{\mathcal{A}(\Omega_{mp}, \Omega_{nq})\}, \forall m, n \in [1, M], m \neq n$
13     $\Delta_{mm} \leftarrow \sum_n S_{mn}$
14     $L \leftarrow \Delta - S$
15     $\tilde{L} \leftarrow \Delta^{-1/2} L \Delta^{-1/2}$
16     $U \leftarrow k$ largest eigenvectors of $\tilde{L}$
17     $\Pi \leftarrow k$-means on $U$
18 return $\Pi$
```

according to the 3-quantiles (tertiles): 551 and 2523 share the lowest level of this metabolite (thus, they are in the same blue cluster); whereas 549 and 715 have medium values (green cluster), and the rest of the material have the highest values (yellow cluster). Note that now this partition of the biological material according to this metabolite matches perfectly the partitioning of the materials obtained according to *flavour*. This can be interpreted as a strong evidence of a potentially interesting relationship between these two variables (*cis-3-hexenal* and *flavour*). Subsequently, the similarity between this numerical and this categorical variable can be obtained, for example, with a measure widely used to compare partitions in clustering: the adjusted Rand index (ARI) (Hubert and Arabie, 1985; Vinh *et al.*, 2010). This process can be repeated for each pair of variables in the input data and the resulting similarity matrix can be used by any standard clustering algorithm, such as Spectral Clustering (SC) (Shi and Malik, 2000; Ng *et al.*, 2001).

2.2 The Clustermatch algorithm

The Clustermatch algorithm can be seen in detail in Algorithm 1. The input is a data matrix D of size $M \times N$ (M variables and N objects), the maximum number of internal clusters (R) and the final number of clusters to find (k). For each row $\mathbf{d}_i$ (line 2), the algorithm computes a set of internal partitions Ω_{ir} with r clusters according to the row data type. If it is numerical (line 3), the algorithm computes $R - 1$ partitions of the N objects, where each one is obtained by using a set of quantiles ρ (line 5): all objects between two adjacent quantiles belong to the same cluster, where ρ_1 is the minimum value of $\mathbf{d}_i$ and ρ_{r+1} the maximum. Thus, for instance, if $r = 2$ then ρ will contain the minimum value, the median and the maximum value of $\mathbf{d}_i$; this means that the higher half of the N objects will be placed in one cluster and the lower half in another one, producing a data partition with two clusters for $\mathbf{d}_i$. Since the internal clustering method only uses the ordering of the data, handling of ordinal variables is similar than numerical ones. Therefore, if a variable has ordinal values like "bad", "regular" and "good", it should be provided as 0, 1 and 2, respectively. If the data type is categorical (line 7), then only one partition is computed for $\mathbf{d}_i$, which has the clusters defined by the unique values of $\mathbf{d}_i$, considered as the set of categories $\mathcal{C}_i$ (line 8 and 9). Note that, in order to derive a similarity value between two variables, we only need a partition of the objects from each one. This is a very important property of our approach, because it is not limited to compare

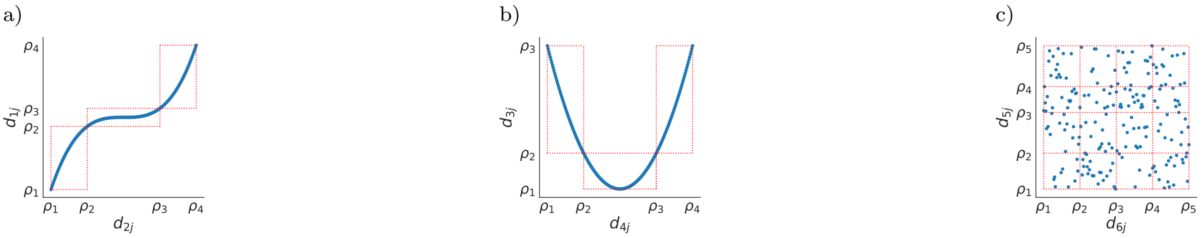


Figure 1: How Clustermatch works when there are different relationships between variables: a) monotonic (ARI=1.0), b) quadratic (ARI=0.57) and c) independent/random (ARI=0.00). The algorithm partitions each numerical variable $\mathbf{d}_i$ using its quantiles ρ, and then computes the similarity between partitions using ARI. Red boxes indicate sets of objects clustered together in two partitions of both variables.

only 1-dimensional variables, but also variables of different dimensions. This would be the case of images, sounds, or the combination of several 1-dimensional variables. The only change needed in Algorithm 1 would be to extend lines 5 and 6 with a call to the appropriate clustering method.

At this point, each variable will have a set of one or more internal partitions associated, Ω_{ir}, with r clusters each. With these, a similarity matrix S for all M variables is derived by comparing their associated partitions (line 12). To compare partitions Ω_i and Ω_j, the ARI is given by

$$\mathcal{A}(\Omega_i, \Omega_j) = \frac{2(n_0 n_1 - n_2 n_3)}{(n_0 + n_2)(n_2 + n_1) + (n_0 + n_3)(n_3 + n_1)}, \tag{1}$$

where n_0 is the number of object pairs that are in the same cluster in both Ω_i and Ω_j, whereas n_1 is the number of pairs in different cluster in both partitions; n_2 (and n_3) are the number of object pairs that were grouped in the same (different) cluster in the first partition, but in a different (same) cluster in the second one. Intuitively, $n_0 + n_1$ is number of object pairs in which both partitions coincide, and $n_2 + n_3$ those in which they disagree. In addition to be a symmetric index, the "adjusted-for-chance" property makes ARI to derive a constant value when both partitions are independently drawn (Hubert and Arabie, 1985), and this also holds in the case of comparing partitions with different number of cluster (Vinh *et al.*, 2010). This is a nice property, given that Clustermatch defines the similarity of variables i and j as the maximum ARI among all possible comparisons of partitions Ω_i and Ω_j (line 12). For two highly related variables i and j, this approach assumes that if pairs of objects are grouped together in the internal partitions of row i, then they should also be grouped together in the internal partitions of row j. If this is the case, then the ARI between them has a value of 1. For independent random partitions it will be always close to 0. Clustermatch, like all clustering algorithms based on a pairwise similarity matrix, builds S under the assumption of variable independence, and then looks for relationships among several variables by analyzing the full similarity matrix in the final clustering stage. Thus, S can be the input of a clustering method such as, for example, spectral clustering (Shi and Malik, 2000) (lines 13-17 in Algorithm 1) which derives the final solution partitioning the M variables into k clusters.

2.3 Clustermatch on different data types

The Clustermatch algorithm can detect hidden relationships by clustering each variable separately, and then computing how much those clusters match. This procedure is exemplified in Figure 1, where numerical data points are shown with blue dots and three different types of relationships are depicted. Each subfigure shows a particular relationship between two variables, where each partition has a number of clusters defined by their quantiles, as shown in Algorithm 1. The red boxes indicate a set of data points clustered together both in $\Omega_{\mathbf{d}_i,p}$ and $\Omega_{\mathbf{d}_j,q}$ partitions (with p and q clusters each, respectively). Figure 1a shows a monotonic relationship between variables $\mathbf{d}_1$ and $\mathbf{d}_2$, where their internal partitions have three clusters. Since these partitions match perfectly, Clustermatch scores this relationship with the maximum value, 1.0. This is the case where both variables produce exactly the same grouping for objects, which is the underlying assumption of Clustermatch to infer a similarity measure. In Figure 1b the relationship is quadratic, and in this case $\Omega_{\mathbf{d}_3,2}$ has two clusters ($p = 2$) whereas $\Omega_{\mathbf{d}_4,3}$ has three ($q = 3$). All objects in cluster $\Omega_{\mathbf{d}_4,3,2}$ (i.e., data points between quantiles ρ_2 and ρ_3) are also clustered together in $\Omega_{\mathbf{d}_3,2,1}$, whereas objects in cluster $\Omega_{\mathbf{d}_3,2,2}$ are clustered in two different groups in partition $\Omega_{\mathbf{d}_4,3}$. In this case, the score is 0.57, which is high enough for Clustermatch to detect this relationship. The third case, in Figure 1c, shows an example where there is no relationship between variables $\mathbf{d}_5$ and $\mathbf{d}_6$. Indeed, these variables do not produce a similar grouping of the objects, since their internal partitions disagree completely on how to cluster objects: each cluster in $\Omega_{\mathbf{d}_5,4}$ contains data points that were grouped in all other clusters

sinc(*i*) Research Institute for Signals, Systems and Computational Intelligence (fich.unl.edu.ar/sinc)
M. Pividori, A. Cernadas, L. A. Haro, F. Carrari, G. Stegmayer & D. H. Milone; "Clustermatch: discovering hidden relations in highly diverse kinds of qualitative and quantitative data without standardization"
Bioinformatics, 2018.

sinc(i) Research Institute for Signals, Systems and Computational Intelligence (fich.unl.edu.ar/sinc)
M. Pividori, A. Cernadas, L. A. Haro, F. Carrari, G. Stegmayer & D. H. Milone; "Clustermatch: discovering hidden relations in highly diverse kinds of qualitative and quantitative data without standardization"
Bioinformatics, 2018.

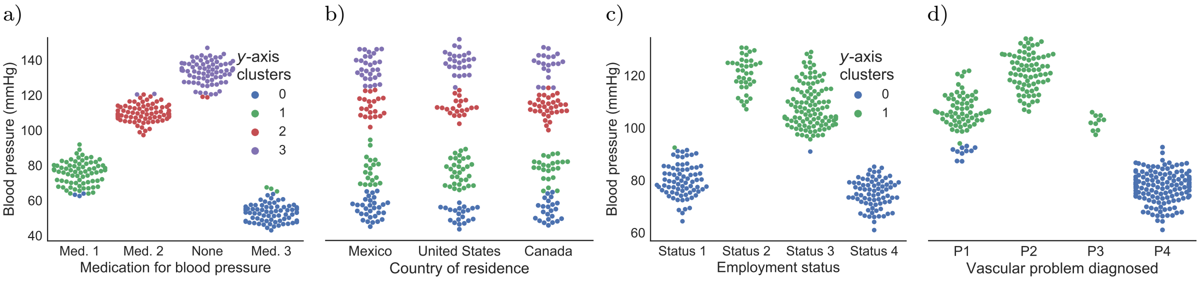


Figure 2: Swarm plots showing different degrees of relationship between a numerical measure (blood pressure in y-axis) and four categorical variables (x-axis): a) represents a strong relationship, b) shows no association, and c) and d) are moderate associations.

in $\Omega_{\mathbf{d_6},4}$. Clustermatch scores this case with the minimum value (0.001), thus correctly detecting no relationship.

Since Clustermatch uses ARI to compute similarity, it is indirectly using a contingency table with the memberships of data points to different clusters in both partitions. As explained above, data points in a numerical array can be converted to a clustering partition by employing quantiles. In the case of categorical data, the categories themselves are considered as clusters, and thus the contingency table can be still obtained and the ARI computed. This approach allows mixing numerical and categorical arrays and it is still able to derive a similarity value.

To exemplify this feature we show here some simulated examples inspired on different heterogeneous variables commonly employed in large biobanks, like the UK Biobank (Bycroft *et al.*, 2017) or the China Kadoorie Biobank[1]. For instance, it could be interesting to automatically find whether a medication for pain relief (categorical variable) is related to pain type (another categorical variable); or whether the reported illness of the father (categorical variable) shows a relationship with blood pressure (numerical), unveiling, for example, an interesting relationship between a progenitor's illness and a particular patient's health measurement. These large cohorts contain hundreds of measures taken for each individual, what makes them an invaluable resource for finding hidden relationships among, for example, blood pressure, cholesterol level, hours of sleep, ethnic background, country of birth, illness of father, type of coffee (decaffeinated, instant or ground coffee) and the current employment status (paid, retired, unable to work because of illness, unemployed, doing voluntary work, etc)[2]. Clustermatch would be particularly useful and well-suited here, since it can process and relate all of these variables automatically, without any user intervention, and with low computational complexity.

Another example can be seen at the left of Figure 2, which shows two possible types of relationships between blood pressure (a numerical variable, in mmHg), and two categorical measures: medication for blood pressure (Figure 2a) and country of residence (Figure 2b). The different colors indicate the internal clustering partition of the numerical variable (blood pressure in y-axis). According to Clustermatch, the first of these categorical variables shows a strong similarity with the numerical measure of blood pressure (in mmHg), while the other categorical variable exhibits no relationship at all. In Figure 2a, every categorical option (medication 1, medication 2, etc) can be distinguished from the other, since each one of them has a characteristic range of blood pressure (visually, each category has almost a unique color). The Clustermatch method assigns an ARI value of 0.91 to this relationship, and this could be an indication that each type of medication seems to have a significantly different impact on the blood pressure. On the other hand, in Figure 2b, there is no relationship between country of residence and blood pressure, since the different categories (countries) are not distinguishable among them (because of the full overlapping). This is reflected by Clustermatch by assigning an ARI very close to zero (0.01).

Between the extreme relationship degrees shown in Figure 2a and 2b, where one categorical variable shows a strong relationship with the numerical variable and the other one does not, at the right of Figure 2 another pair of categorical variables is shown where their association with blood pressure is somehow moderate. Figure 2c shows an interesting relationship with employment status where, differently from Figure 2a, there are two groups of categories where patients exhibit different blood pressure values: patients with Status 2 and 3 have in common high blood pressure, whereas patients with Status 1 and 4 share low blood pressure. The relationship here is not as strong as in Figure 2a: each employment status is not distinguishable from the rest, but they become different if considered in pairs. Thus, Clustermatch assigns here an ARI value of 0.54. In Figure 2d, which

[1] http://www.ckbiobank.org/

[2] Examples of these types of data can be found in: http://biobank.ctsu.ox.ac.uk/showcase/label.cgi

sinc(i) Research Institute for Signals, Systems and Computational Intelligence (fich.unl.edu.ar/sinc)
M. Pividori, A. Cernadas, L. A. Haro, F. Carrari, G. Stegmayer & D. H. Milone; "Clustermatch: discovering hidden relations in highly diverse kinds of qualitative and quantitative data without standardization"
Bioinformatics, 2018.

shows how diagnosed vascular problems relate with blood pressure, three of them (P1, P2 and P3) do not seem to be differentiated by blood pressure, but patients with P4 are indeed very different, where all of them have low blood pressure. Since the group of patients diagnosed with P4 is large and the blue internal cluster matches it almost perfectly, the similarity of vascular problems with blood pressure (ARI ~ 0.63) is stronger than with employment status.

In these examples we show artificial cases where the number of categories is substantially smaller than the number of objects. There could be situations where this might not be the case, that is, the number of categories is large enough to possibly produce non-informative variables (like zip-codes, for instance). In this case, of course, the user might need to merge original categories into higher level categories that will be relevant for the study (like states in the example of zip-codes).

As it has been shown, Clustermatch is able to process different data types by placing data points into a contingency table according to their cluster membership. Then, a similarity measure is derived by using the ARI, thus easily and quickly comparing highly heterogeneous data sets. In clustering applications, several variables are processed simultaneously to find groups of highly related variables. In the following sections, we will show how Clustermatch can generate a similarity matrix where the true structure of the data is more easily detected by the clustering algorithm, even in noisy scenarios.

3 Data and experimental setup

Two types of datasets were employed to test Clustermatch. First, an artificially generated dataset is used to simulate a scenario where objects have highly diverse kinds of variables with different noise levels. The data were generated in three steps. First, we created 100 numerical variables with 1000 objects each, with variables equally distributed in three compact and well-separated gaussian clusters with random means in $(-1, 1)$. This structure is supposed to be very easily found by any conventional clustering algorithm. Secondly, these 100 variables were randomly taken in 10 simulated "data sources" of 10 variables each. Each data source was transformed using one of these functions: x^4, $\ln(|x|)$, 2^x, $100x$, $\ln(|1+x|)$, x^5, $10000x$, $\log_{10}(|x|)$, $0.0001x$ and $\log_2(|x|)$. Finally, a percentage of objects was randomly chosen and replaced by random values in the range of the corresponding transformed source. This noise model is intended to simulate a real scenario where, besides the errors present in the instruments used to measure the variable (batch effects, errors in questionaries, etc), some objects could simply not follow the general trend or could be outliers. The supplementary material contains more details about this numerical data set (Supplementary Figures 1 and 2). Furthermore, we also generated another dataset including both numerical and categorical sources (Supplementary Figures 3 and 4).

Second, we analyzed a real dataset described in previous reports (Asprelli *et al.*, 2017; Cortina *et al.*, 2017, 2018; D'Angelo *et al.*, 2018). Briefly, this dataset was collected from field trials conducted from October to March in open field conditions at the Agronomy School of the National University of Mendoza, Argentina (S33°0.3′; W68°52.2′; 912 meters above sea level), in 3 growing seasons: 2008-09, 2009-10 and 2011-12. Agronomic performance, plant morphology and fruit quality traits were reported elsewhere (Asprelli *et al.*, 2017) and together they constitute the "Agronomic" source. Metabolic traits were measured as described in Cortina *et al.* (2017, 2018); and sensory attributes were recorded in D'Angelo *et al.* (2018).

4 Results

4.1 Noisy and non-linear relationships in numerical data

Five methods were compared by using the simulated data: SC-Pearson, SC-Spearman, SC-DC, SC-MIC and our proposal, Clustermatch. All methods created a similarity matrix using a specific measure, and then SC was applied to derive the final solution with 3 clusters. Each one was run 20 times and the partition obtained compared (using ARI) with the reference partition of the original three gaussians. The results are shown in Figure 3 (for additional results with different clustering algorithms see Supplementary Figures 5 and 6; and for different number of final clusters k, see Supplementary Figures 7 and 8). The x-axis indicates the percentage of noisy objects, from 0 (no noise) up to 55%, since no method found a meaningful solution beyond this point. The y-axis is the average ARI between the reference partition and the solution found by each method. The black lines at the top of each bar are the 95% confidence intervals.

It can be clearly seen that Clustermatch largely outperformed all other methods, including SC-DC and SC-MIC, which are not limited to linear or monotonic relationships as SC-Pearson and SC-Spearman. Clustermatch, SC-MIC and SC-DC found the true structure of the data in low-noise scenarios (0-15%). However, from a noise level of 15%, SC-DC started failing to find a perfectly accurate grouping of the variables, and from this point forward performance decreased constantly, always

sinc(*i*) Research Institute for Signals, Systems and Computational Intelligence (fich.unl.edu.ar/sinc)
M. Pividori, A. Cernadas, L. A. Haro, F. Carrari, G. Stegmayer & D. H. Milone; "Clustermatch: discovering hidden relations in highly diverse kinds of qualitative and quantitative data without standardization"
Bioinformatics, 2018.

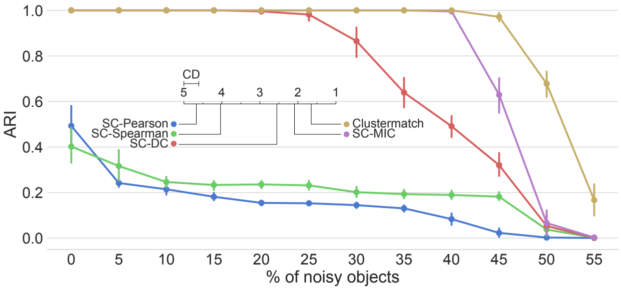


Figure 3: Clustering quality (y-axis) for all methods under different noise levels (x-axis) for one simulated dataset including linear and non-linear sources. A critical difference (CD) diagram with a post-hoc Nemenyi test is also shown.

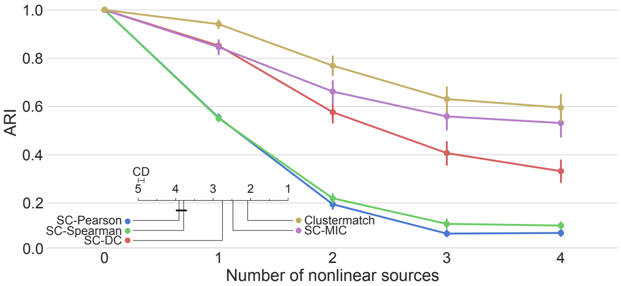


Figure 4: Clustering quality (y-axis) for all methods in five simulated data sets with different amounts of non-linear sources (x-axis). Each dataset included noise levels from 0% to 70%, thus each point is an average over 20 tests.

behind SC-MIC and Clustermatch. In the medium-noise range, 20-35%, SC-MIC and Clustermatch were the only methods able to always find the true partition, proving that they were robust in these noisy scenarios. Furthermore, Clustermatch was clearly the best method also with high noise level: it was able to maintain a very high accuracy where SC-MIC fell abruptly (50% of noisy objects), and it could find a meaningful partition (with ARI $\sim$ 0.70) in scenarios where all the others derived a very poor quality solution (ARI $<$ 0.15). Indeed, from a noise level of 40% to 45% SC-MIC was largely affected, obtaining an ARI of 0.99 and then 0.70, whereas Clustermatch obtained 1.0 and 0.96, respectively. In the case of 50% of noisy objects, it can be seen that while all other methods were significantly affected, Clustermatch was still able to derive a very good quality partition, with an average ARI of 0.70. At the last highest noise configuration, with 55% of noisy objects, all methods dropped performance abruptly (ARI $\sim$ 0.0), while Clustermatch was still able to group correctly almost 20% of variable pairs (ARI $\sim$ 0.18). In addition to ARI, we also measured the final clustering quality using the adjusted mutual information (AMI) index (Vinh *et al.*, 2010), with similar results (Supplementary Figure 9).

These results showed that, with highly diverse data sources (including linear and non-linear relationships) and a varying noise level, only a set of the methods was able to find all different types of relationships: SC-DC, SC-MIC and Clustermatch. SC-Pearson and SC-Spearman were significantly affected when linear and monotonic assumptions did not hold, and they performed very badly under noisy data. When noisy levels increased, Clustermatch was able to find significantly better solutions than SC-DC and SC-MIC. In order to statistically evaluate differences between all tested methods, a critical difference (CD) diagram for post-hoc Nemenyi test (Demšar, 2006) was done. For a CD diagram, a ranking among all methods under evaluation is first obtained (being 1 the best, and 5 is the worst one). Then, the diagram joins with bold lines methods that are not statistically different between them (their average ranks differ by less than the critical difference value). The CD is shown at the center of Figure 3, and it indicates that Clustermatch reached the highest ARI values, being this difference statistically significant from the rest of the methods.

Figure 4 shows how methods were affected with an increasing number of non-linear transformations, running 20 times each case. The x-axis indicates the number of non-linear sources, and each line corresponds to mean ARI. Here, we have used 5 different data sets with an increasing amount of non-linear sources: the first data set contained five linear-only transformations of the original data, whereas the last one contained four non-linear transformations plus only one remaining linear

sinc(i) Research Institute for Signals, Systems and Computational Intelligence (fich.unl.edu.ar/sinc)
M. Pividori, A. Cernadas, L. A. Haro, F. Carrari, G. Stegmayer & D. H. Milone; "Clustermatch: discovering hidden relations in highly diverse kinds of qualitative and quantitative data without standardization"
Bioinformatics, 2018.

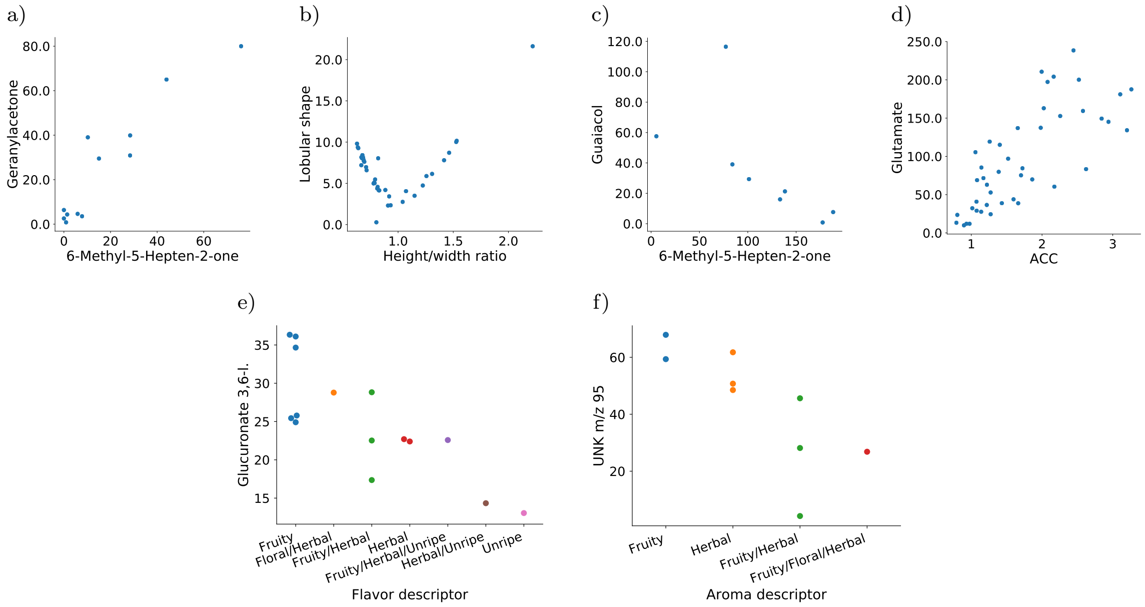


Figure 5: Scatter plots of different pairs of variables from the real tomato dataset showing distinct relationships: a) linear; b) quadratic; c) linear with outlier; d) linear with high dispersion; e) and f) show relationships between categorical and numerical data.

source. In addition, each data set for a combination of transformations was replicated with different noise levels, from 0% to 70%. It can be seen that all methods found the true structure of the data on all-linear sources. However, in front of just a single non-linear transformation (x^2), all methods were affected. Pearson and Spearman suffered the largest drop in performance, meanwhile SC-MIC and SC-DC performed similarly, and Clustermatch was still the best method, with statistically significant differences. Indeed, Clustermatch was the best method in the next data set, where a linear source was replaced by $\log(|x|)$. In the last two cases (with x^4 and $\sin(\pi x)$), Spearman and Pearson were the worst performing methods, and SC-DC was clearly outperformed by SC-MIC and Clustermatch. Overall, Clustermatch was the best performing method in this scenario, which shows how methods were affected when data gradually becomes more complex with the addition of non-linear transformations. The CD diagram in the bottom-left corner of Figure 4 clearly shows that there are statistically significant differences among SC-DC, SC-MIC and Clustermatch. It also shows that the worst methods were SC-Pearson and SC-Spearman.

The supplementary material contains the results for the artificial dataset with mixed numerical and categorical sources (Supplementary Figure 10), where only Clustermatch is able to directly process such mixture of data types. These results are consistent with those presented for the numerical-only dataset (Figure 3). We also performed scalability tests to assess how the methods behave under different number of measured objects/materials (Supplementary Figure 11). For 500,000 objects (for example, this is the number of individuals in the UK Biobank), MIC finished comparing two variables in 32,265 s, DC could not finish after 24 h running, and Clustermatch took just 22 s. This result shows clearly that Clustermatch is very well-suited for large datasets. In summary, we have shown that, regardless of the clustering method employed, the number of final clusters specified by the user, the noise levels and the non-linear transformations present in data, the heterogeneity in the data types, the clustering quality index used, and even the amount of data to be processed, Clustermatch has a very good performance. Furthermore, in most of the cases it is superior to the other methods compared.

4.2 Real and highly diverse data set

In this section we summarized results obtained with Clustermatch using a *bona fide* dataset previously published (Asprelli *et al.*, 2017; Cortina *et al.*, 2017, 2018; D'Angelo *et al.*, 2018) and described above, which is part of a systems biology study comprising a very diverse collection of tomato accessions collected along the Andean Valleys of Argentina. The most adequate number of clusters k to explore has been set according to the consensus index method (Vinh *et al.*, 2010). Very different kinds of variables have been measured and the challenge was to find hidden relationships for hypotheses-building. For example, interesting variables from agronomic descriptors and sensory panels are

sinc(*i*) Research Institute for Signals, Systems and Computational Intelligence (fich.unl.edu.ar/sinc)
M. Pividori, A. Cernadas, L. A. Haro, F. Carrari, G. Stegmayer & D. H. Milone; "Clustermatch: discovering hidden relations in highly diverse kinds of qualitative and quantitative data without standardization"
Bioinformatics, 2018.

categorical (i.e. flavor and aroma descriptors). Therefore, data mining for the identification of new and cryptic relationships between these kind of variables presents a great challenge. To test the reliability and robustness of Clustermatch we first verified well-established linear and non-linear relationships between biochemical or molecular phenotypes (Figure 5). The strongest associations described below were identified by computing a p-value using permutation tests.

Recently, it has been reported that the carotenoid-derived volatile compounds (VOC) *geranylacetone* and *6-methyl-5-hepten-2-one* (MHO) are significantly associated to tomato flavor (Tieman *et al.*, 2017). Our analyses with Clustermatch revealed that the association between these two VOCs is indeed one of the strongest (P¡0.006). Although this relationship is mostly linear (Figure 5a), and could have been detected by traditional methods, this finding is in complete agreement with Tieman's report based on a genome-wide association study (GWAS) using a panel of 398 tomato accessions, which identified seven *loci* co-segregating for the amount of these two VOCs in tomato fruits (Tieman *et al.*, 2017). Thus, the associations that we present here using Clustermatch are also supported by independent genetic analyses. Similarly but with a quadratic relation (Figure 5b), Clustermatch pinpointed a strong, yet logical, relationship between shape (e.g. lobular) and size ratios determinations (e.g. height/width relationship). Note that in an integrated analysis, Clustermatch disentangled these types of non-linear correlations.

These findings prompted us to survey new associations occurring between variables within the highly diverse collection of data. Remarkably, when using all growing seasons, MHO also associated with *guaiacol* (P¡0.02), a phenylpropanoid volatile normally found in tomato fruits whose aroma is often described as "pharmaceutical" or "smoky" (Krumbein and Auerswald, 2018). Interestingly, the *guaiacol*-MHO inverse relationship displays a clear outlier (Figure 5c), corresponding to a breeding line cherry genotype with characteristic purple fruits. This proves one powerful characteristic of Clustermatch, which is the identification of linear relationships even when outliers are present in the dataset and suggests that it can be used in breeding selection programs to identify transgressive genotypes/accessions.

We also noticed another biologically relevant case when we conducted the analysis with a dataset including only one harvesting dataset. A cluster including 9 amino acids was detected, among which 3 of them (*glutamate*, *pyroglutamate* and *gamma-aminobutyric acid* -GABA-) are biochemically connected by two enzymatic reactions in the GABA biosynthetic pathway. Additionally, these amino acids were found grouped together with the ethylene precursor *1-aminocyclopropane-1-carboxylic acid* (ACC) (P¡0.002) and *oxoglutarate* (P¡0.002), being the latter the main precursors of GABA biosynthesis. Although the biological roles of GABA during tomato fruit development is still under debate (Takayama and Ezura, 2015), it is well described as a positive regulator of the *1-aminocyclopropane-1-carboxylate oxidase* expression (Kathiresan *et al.*, 1997), a key enzyme for ethylene biosynthesis. In spite of the above mentioned relations are somehow expected and therefore constitute a test case, the role of GABA as an indirect inductor of fruit ripening appears as a running hypothesis that has to be tested. In particular, the relationship between *glutamate* and ACC (Figure 5d) is a case of large dispersion, which represents a violation of homoscedasticity, an important assumption in linear models. These results demonstrate that Clustermatch consistently identified relevant relationships between variables, and also clustered together several other variables that are part of ethylene signaling pathway, involved in fruit ripening. From an agronomic point of view, this may open up a novel strategy to select varieties for breeding according to metabolic compounds accumulation.

A third hypothesis was inferred using another set of data from the same germplasm collection but harvested in a different growing season; analyses of significant connections revealed a clear relationship between the categorical variables describing consumer's sensory attributes, such as flavor and aroma descriptors, with intermediates of the ascorbate metabolic pathway (e.g., *Glucuronate 3,6-lactone* and *Glucuronate/galacturonate*; P¡0.002) (Figure 5e). Indeed, this cluster is represented by compounds of the ascorbate pathway[3], including *glucuronate*, *gulonate*, *ascorbate* and *dehydroascorbate* linked to the antioxidant capacity of the tomato extracts (measured by *TEAC HS* and *FRAPS* determinations). At the same time, antioxidant capacity was linked with *alpha-terpineol* (P¡0.02), a compound recently defined as an aroma and sourness component of the tomato fruits (D'Angelo *et al.*, 2018). Regarding the ascorbate pathway, it is intriguing to learn about the potential relationship (whether direct or indirect) between this metabolic pathway and the flavor determination in tomato fruits that we identified in this study. We highlight the fact that these relationships between categorical and numerical variables were identified by applying Clustermatch directly, without any need of a special pre-processing step for transforming these data, as other methods would require. In the same line, we observed aroma descriptors strongly associated with an unknown metabolite in the mass spec collection, *UNK m/z 95* (Figure 5f, P¡0.02), which was previously reported to be associated with non-characteristic tomato taste and odour (Cortina *et al.*, 2018). These examples establish relationships supporting the generation of novel hypotheses. Firstly, accumulation of *alpha-terpineol*,

[3] http://www.genome.jp/kegg-bin/show_pathway?map=map00053 &show_description=show

sinc(*i*) Research Institute for Signals, Systems and Computational Intelligence (fich.unl.edu.ar/sinc)
M. Pividori, A. Cernadas, L. A. Haro, F. Carrari, G. Stegmayer & D. H. Milone; "Clustermatch: discovering hidden relations in highly diverse kinds of qualitative and quantitative data without standardization"
Bioinformatics, 2018.

a volatile monoterpenoid alcohol with a wide range of biological applications (Khaleel *et al.*, 2018), is induced in reducing cellular environments. Secondly, the unknown m/z 95 compound is a component of the tomato taste and odour.

A summary of the relevant relationships that led to infer hypotheses for this case of study can be found in Supplementary Table 1. It is important to highlight that this table is not meant to be comprehensive, since a very large number of strong relationships have not been discussed and will be a matter of a future manuscript. Taken together, the hypotheses written above illustrate how Clustermatch consistently succeed to identify well-established relationships between variables, as well as it also revealed novel relationships in a large dataset derived from different sources and measurement methods of a systems biology study, even when non-linear and noisy relationships were present.

5 Conclusions

In this study, we have addressed important issues that characterize modern data sets. We offer a solution to the actual challenge of finding hidden patterns in highly diverse data sets, which include not only several different kinds of quantitative and qualitative variables, but also cases where a large amount of biological materials and experimental conditions are present. Clustermatch is able to efficiently compute a similarity measure between any combination of quantitative and qualitative data, avoiding the need of any preprocessing step and thus easing the application of a clustering algorithm on this complex data. This novel approach is able to seamlessly integrate variables of very different nature, producing a similarity matrix that can be processed by any clustering algorithm. We have shown that, when compared to other state-of-the-art methods, the true structure of the simulated data is accurately detected by Clustermatch, even in the presence of significant amounts of noise and non-linear sources. Our proposal was also tested in a real dataset of tomato accessions with several different data sources available. First of all, Clustermatch was able to process this data without previous preprocessing, even though it included numerical and categorical variables. Secondly, as a validation our method was able to find well-known linear and non-linear relationships between physiological and genetic variables previously reported in other independent studies, even in the presence of outliers.

Funding

This work was supported by Consejo Nacional de Investigaciones Cientificas y Tecnicas (CONICET), Universidad Nacional del Litoral (CAI+D 2016 082), the Agencia Nacional de Promoción Científica y Tecnológica (PICT 2014 2627) and the European Union Horizon 2020 Research and Innovation Programme, Grant Agreement Number 679796.

References

Albanese, D., Filosi, M., Visintainer, R., Riccadonna, S., Jurman, G., and Furlanello, C. (2013). minerva and minepy: a c engine for the mine suite and its r, python and matlab wrappers. *Bioinformatics*, **29**(3), 407–408.

Asprelli, P. D., Sance, M., Insani, E. M., Asis, R., Valle, E. M., Carrari, F., Galmarini, C. R., and Peralta, I. E. (2017). Agronomic performance and fruit nutritional quality of an andean tomato collection. In *Acta Horticulturae*, number 1159, pages 197–204. International Society for Horticultural Science (ISHS), Leuven, Belgium.

Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L. T., Sharp, K., Motyer, A., Vukcevic, D., Delaneau, O., O'Connell, J., Cortes, A., Welsh, S., McVean, G., Leslie, S., Donnelly, P., and Marchini, J. (2017). Genome-wide genetic data on 500,000 uk biobank participants. *bioRxiv*.

Chen, Y., Zeng, Y., Luo, F., and Yuan, Z. (2016). A new algorithm to optimize maximal information coefficient. *PLOS ONE*, **11**(6), 1–13.

Cortina, P., Santiago, A., Sance, M., Peralta, I., Carrari, F., and Asis, R. (2018). Exploring the relationship between volatiles organic compounds and tomato fruit flavor of andean landraces, commercial varieties and an edible wild species. *Metabolomics*. In press.

Cortina, P. R., Asis, R., Peralta, I. E., Asprelli, P. D., and Santiago, A. N. (2017). Determination of volatile organic compounds in andean tomato landraces by headspace solid phase microextraction-gas chromatography-mass spectrometry. *Journal of the Brazilian Chemical Society*, **28**(1), 30–41.

D'Angelo, M., Zanor, M. I., Sance, M., Cortina, P. R., Boggio, S. B., Asprelli, P., Carrari, F., Santiago, A. N., Ass, R., Peralta, I. E., and Valle, E. M. (2018). Contrasting metabolic profiles of tasty tomato fruit of the andean varieties in comparison with commercial ones. *Journal of the Science of Food and Agriculture*.

Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *J. Mach. Learn. Res.*, **7**, 1–30.

sinc(*i*) Research Institute for Signals, Systems and Computational Intelligence (fich.unl.edu.ar/sinc)
M. Pividori, A. Cernadas, L. A. Haro, F. Carrari, G. Stegmayer & D. H. Milone; "Clustermatch: discovering hidden relations in highly diverse kinds of qualitative and quantitative data without standardization"
Bioinformatics, 2018.

Devlin, S. J., Gnanadesikan, R., and Kettenring, J. R. (1975). Robust estimation and outlier detection with correlation coefficients. *Biometrika*, **62**(3), 531–545.

Huber, P. (2011). *International encyclopedia of statistical science*, chapter Robust statistics, page 124851. Springer-Verlag Berlin Heidelberg.

Hubert, L. and Arabie, P. (1985). Comparing partitions. *Journal of Classification*, **2**(1), 193–218.

Kathiresan, A., Tung, P., Chinnappa, C. C., and Reid, D. M. (1997). [gamma]-aminobutyric acid stimulates ethylene biosynthesis in sunflower. *Plant Physiology*, **115**(1), 129–135.

Khaleel, C., Tabanca, N., and Buchbauer, G. (2018). a-terpineol, a natural monoterpene: A review of its biological properties.

Kinney, J. B. and Atwal, G. S. (2014). Equitability, mutual information, and the maximal information coefficient. *Proceedings of the National Academy of Sciences*, **111**(9), 3354–3359.

Kong, J., Klein, B. E. K., Klein, R., Lee, K. E., and Wahba, G. (2012). Using distance correlation and ss-anova to assess associations of familial relationships, lifestyle factors, diseases, and mortality. *Proceedings of the National Academy of Sciences*, **109**(50), 20352–20357.

Krumbein, A. and Auerswald, H. (2018). Characterization of aroma volatiles in tomatoes by sensory analyses. *Nahrung*, **42**(06), 395–399.

Li, R., Zhong, W., and Zhu, L. (2012). Feature screening via distance correlation learning. *Journal of the American Statistical Association*, **107**(499), 1129–1139. PMID: 25249709.

Li, Y., Wu, F.-X., and Ngom, A. (2018). A review on machine learning principles for multi-view biological data integration. *Briefings in Bioinformatics*, **19**(2), 325–340.

N. Simon, R. T. (2011). Comment on detecting novel associations in large data sets by reshef et al. science. *arXiv*, **abs/1401.7645v1**.

Nature (2012). Finding correlations in big data. *Nature Biotechnology*.

Ng, A. Y., Jordan, M. I., and Weiss, Y. (2001). On spectral clustering: Analysis and an algorithm. In *Proceedings of the 14th International Conference on Neural Information Processing Systems: Natural and Synthetic*, NIPS'01, pages 849–856, Cambridge, MA, USA. MIT Press.

Reshef, D. N., Reshef, Y. A., Finucane, H. K., Grossman, S. R., McVean, G., Turnbaugh, P. J., Lander, E. S., Mitzenmacher, M., and Sabeti, P. C. (2011). Detecting novel associations in large data sets. *Science*, **334**(6062), 1518–1524.

Reshef, Y., Reshef, D. N., Sabeti, P. C., and Mitzenmacher, M. (2014). Theoretical foundations of equitability and the maximal information coefficient. *arXiv*, **abs/1408.4908**.

Shi, J. and Malik, J. (2000). Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **22**(8), 888–905.

Spearman, C. (2010). The proof and measurement of association between two things. *International Journal of Epidemiology*, **39**(5), 1137–1150.

Speed, T. (2011). A correlation for the 21st century. *Science*, **334**(6062), 1502–1503.

Szkely, G. J., Rizzo, M. L., and Bakirov, N. K. (2007). Measuring and testing dependence by correlation of distances. *Ann. Statist.*, **35**(6), 2769–2794.

Takayama, M. and Ezura, H. (2015). How and why does tomato accumulate a large amount of gaba in the fruit? *Frontiers in Plant Science*, **6**, 612.

Tang, D., Wang, M., Zheng, W., and Wang, H. (2014). Rapidmic: Rapid computation of the maximal information coefficient. *Evolutionary Bioinformatics*, **10**, EBO.S13121.

Tieman, D., Zhu, G., Resende, M. F. R., Lin, T., Nguyen, C., Bies, D., Rambla, J. L., Beltran, K. S. O., Taylor, M., Zhang, B., Ikeda, H., Liu, Z., Fisher, J., Zemach, I., Monforte, A., Zamir, D., Granell, A., Kirst, M., Huang, S., and Klee, H. (2017). A chemical genetic roadmap to improved tomato flavor. *Science*, **355**(6323), 391–394.

Vinh, N. X., Epps, J., and Bailey, J. (2010). Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *J. Mach. Learn. Res.*, **11**, 2837–2854.

Xu, R. and Wunsch, D. (2009). *Clustering*. Wiley-IEEE Press.

Zhang, Y., Jia, S., Huang, H., Qiu, J., and Zhou, C. (2014). A novel algorithm for the precise calculation of the maximal information coefficient. *Nature Scientific Reports*, **4**, 6662.

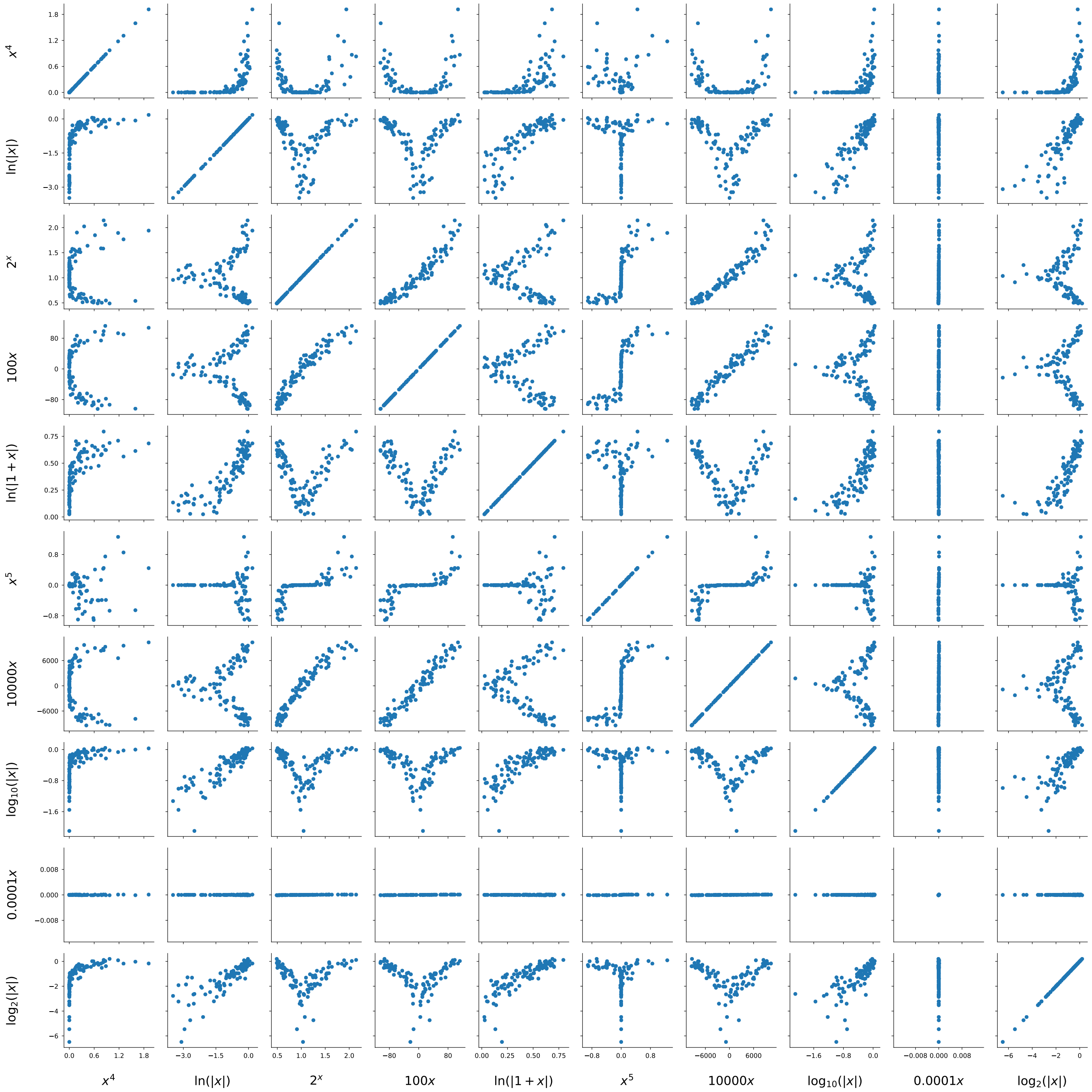
Clustermatch: discovering hidden relations in highly-diverse kinds of qualitative and quantitative data without standardization

Supplementary material

Milton Pividori, Andres Cernadas, Luis de Haro,
Fernando Carrari, Georgina Stegmayer and Diego Milone

sinc(*i*) Research Institute for Signals, Systems and Computational Intelligence (fich.unl.edu.ar/sinc)
M. Pividori, A. Cernadas, L. A. Haro, F. Carrari, G. Stegmayer & D. H. Milone; "Clustermatch: discovering hidden relations in highly diverse kinds of qualitative and quantitative data without standardization"
Bioinformatics, 2018.

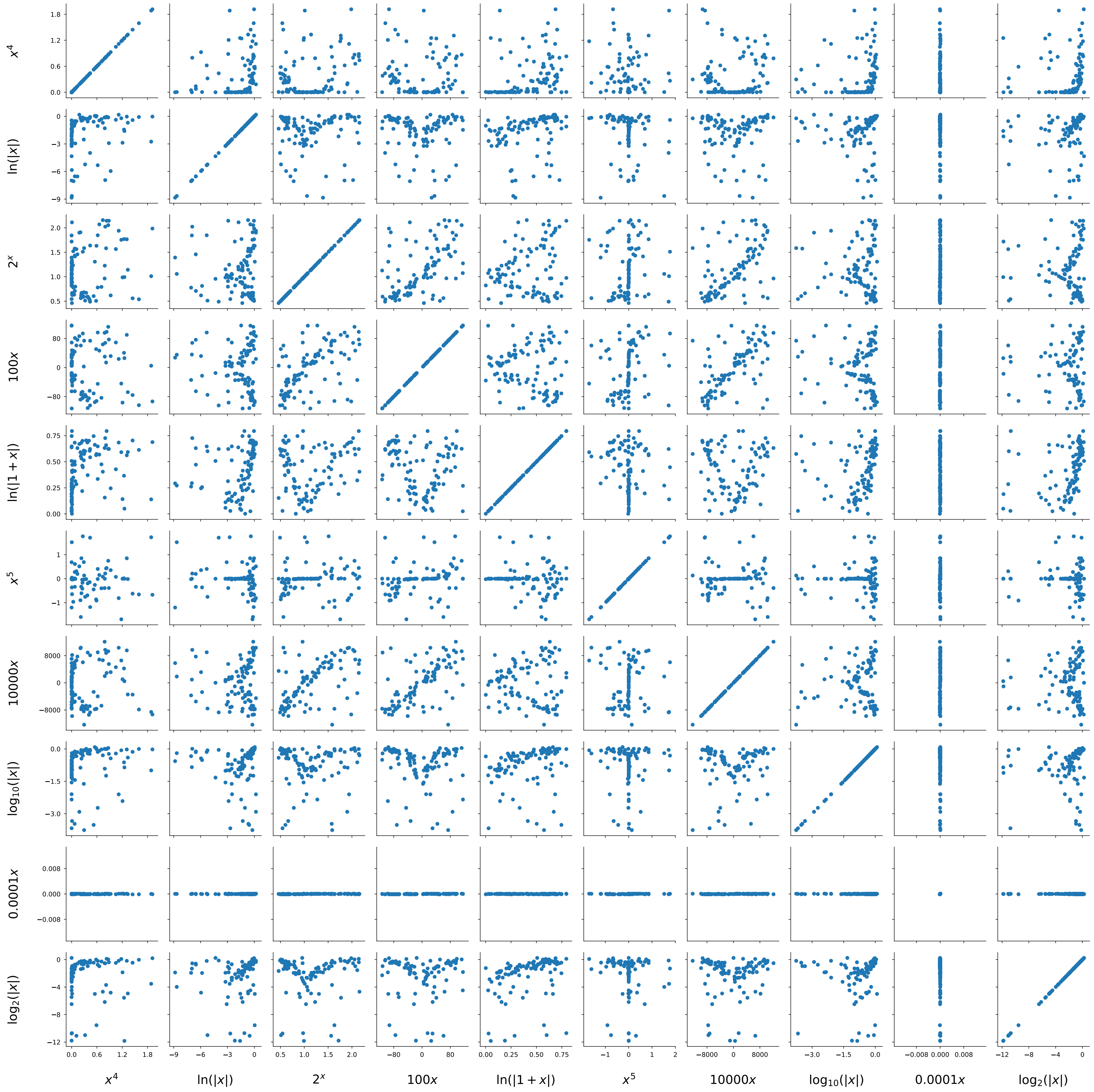
Dataset with numerical sources



Supplementary Figure 1: Relationships in the artificial dataset. The figure includes a set of 10 features from the same cluster, each one from a different source (transformation).

sinc(*i*) Research Institute for Signals, Systems and Computational Intelligence (fich.unl.edu.ar/sinc)
M. Pividori, A. Cernadas, L. A. Haro, F. Carrari, G. Stegmayer & D. H. Milone; "Clustermatch: discovering hidden relations in highly diverse kinds of qualitative and quantitative data without standardization"
Bioinformatics, 2018.

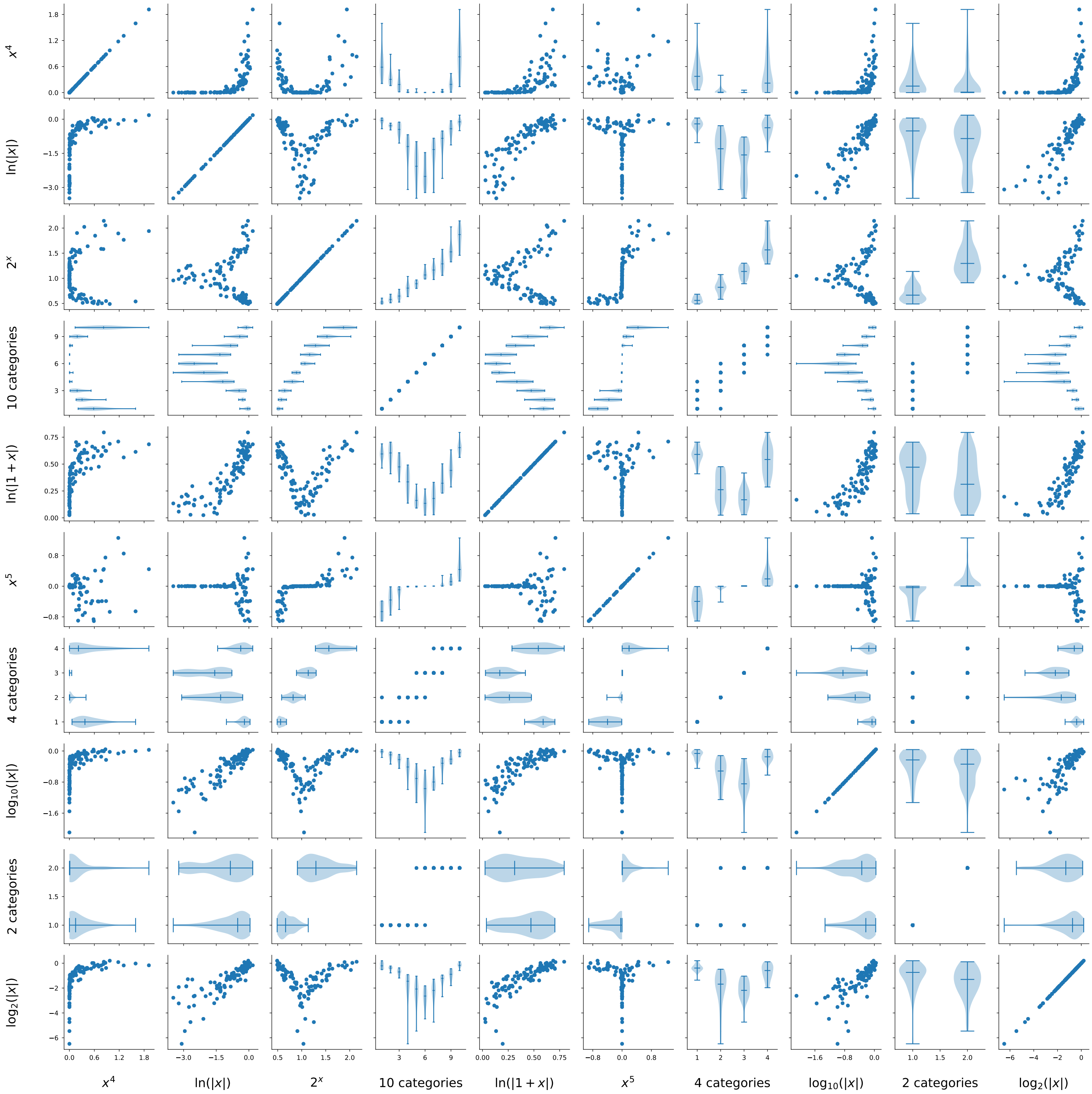
Noisy dataset with numerical sources



Supplementary Figure 2: Relationships in the artificial dataset with 20% of noise. The figure includes a set of 10 features from the same cluster, each one from a different source (transformation).

sinc(*i*) Research Institute for Signals, Systems and Computational Intelligence (fich.unl.edu.ar/sinc)
M. Pividori, A. Cernadas, L. A. Haro, F. Carrari, G. Stegmayer & D. H. Milone; "Clustermatch: discovering hidden relations in highly diverse kinds of qualitative and quantitative data without standardization"
Bioinformatics, 2018.

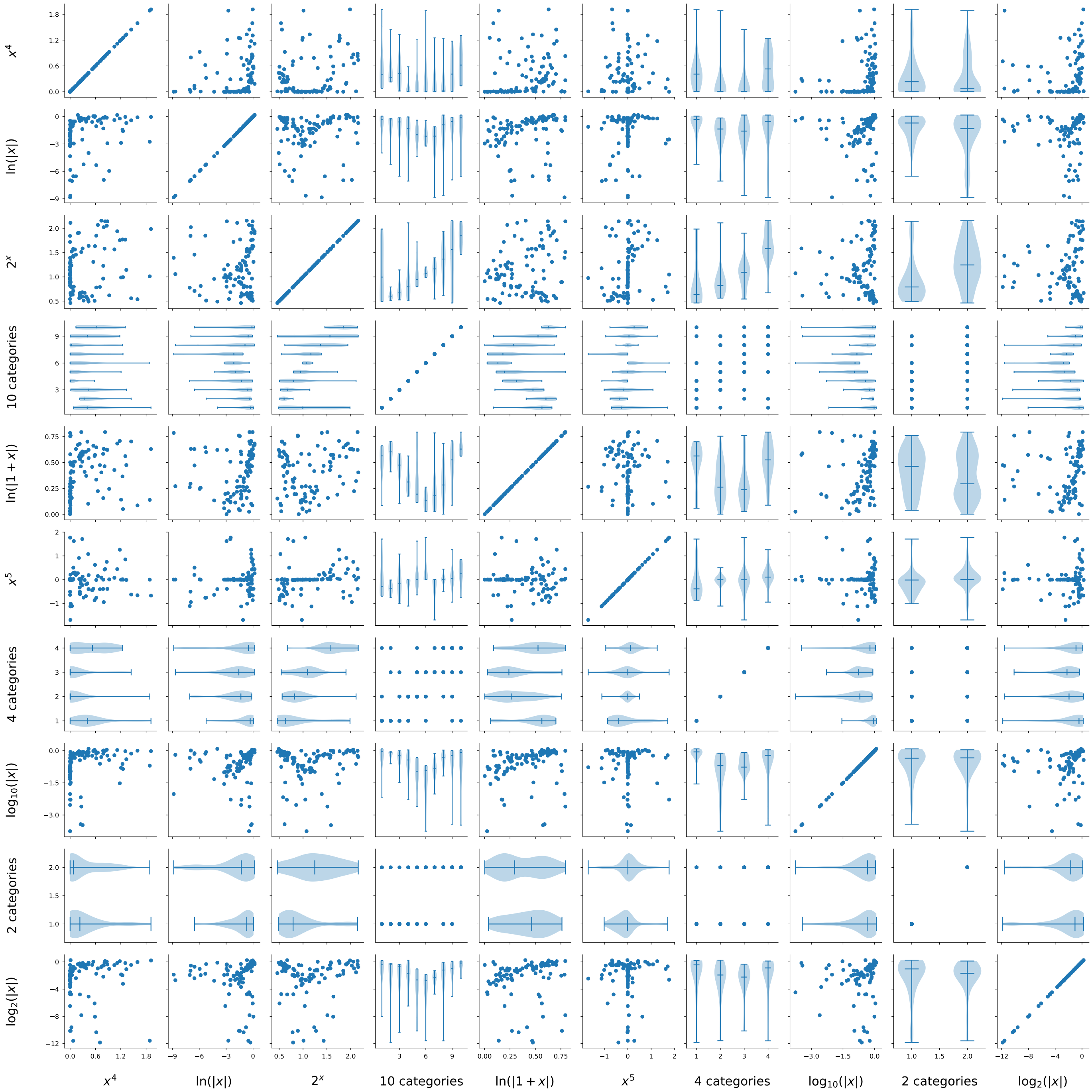
Dataset with numerical and categorical sources



Supplementary Figure 3: Relationships in the artificial dataset with 7 numerical and 3 categorical sources. The figure includes a set of 10 features from the same cluster, each one from a different source (transformation). A scatter plot is shown when both sources are either numerical or categorical, and a violin plot when they differ (numerical vs categorical). Only Clustermatch can process a dataset including numerical and categorical data sources.

sinc(i) Research Institute for Signals, Systems and Computational Intelligence (fich.unl.edu.ar/sinc)
M. Pividori, A. Cernadas, L. A. Haro, F. Carrari, G. Stegmayer & D. H. Milone; "Clustermatch: discovering hidden relations in highly diverse kinds of qualitative and quantitative data without standardization"
Bioinformatics, 2018.

Noisy dataset with numerical and categorical sources

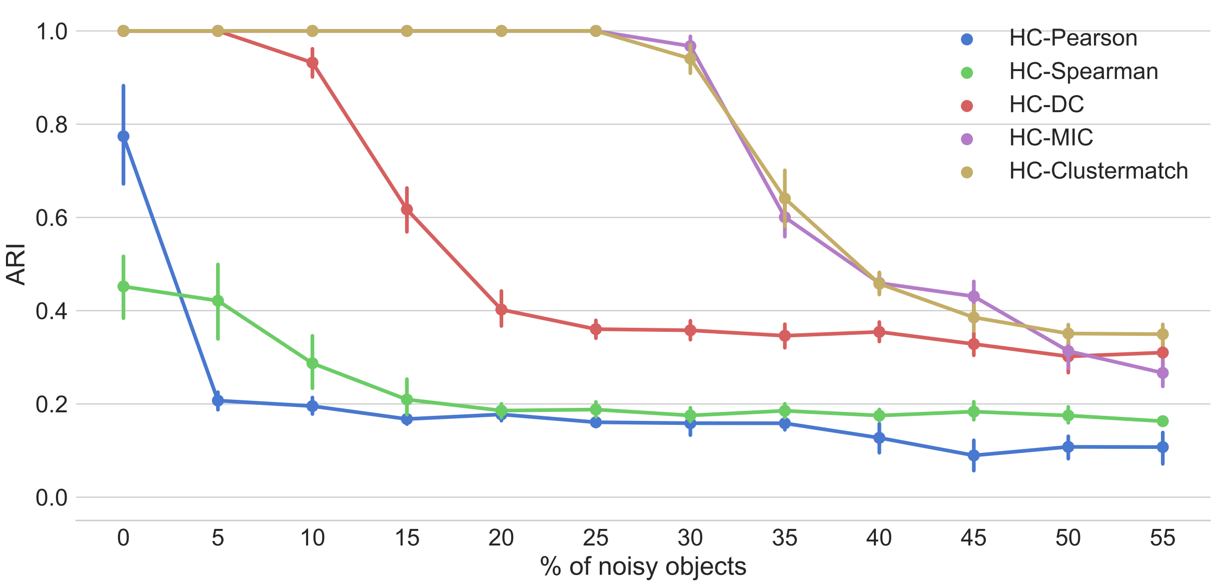


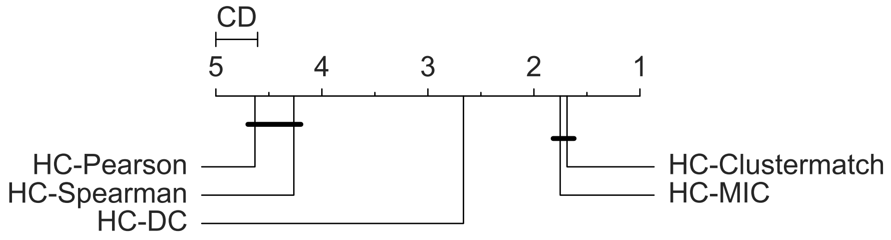
Supplementary Figure 4: Relationships in the artificial dataset with 7 numerical and 3 categorical sources including 20% of noise. The figure includes a set of 10 features from the same cluster, each one from a different source (transformation). A scatter plot is shown when both sources are either numerical or categorical, and a violin plot when they differ (numerical vs categorical). Only Clustermatch can process a dataset including numerical and categorical data sources.

sinc(i) Research Institute for Signals, Systems and Computational Intelligence (fich.unl.edu.ar/sinc)
M. Pividori, A. Cernadas, L. A. Haro, F. Carrari, G. Stegmayer & D. H. Milone; "Clustermatch: discovering hidden relations in highly diverse kinds of qualitative and quantitative data without standardization"
Bioinformatics, 2018.

sinc(i) Research Institute for Signals, Systems and Computational Intelligence (fich.unl.edu.ar/sinc)
M. Pividori, A. Cernadas, L. A. Haro, F. Carrari, G. Stegmayer & D. H. Milone; "Clustermatch: discovering hidden relations in highly diverse kinds of qualitative and quantitative data without standardization"
Bioinformatics, 2018.

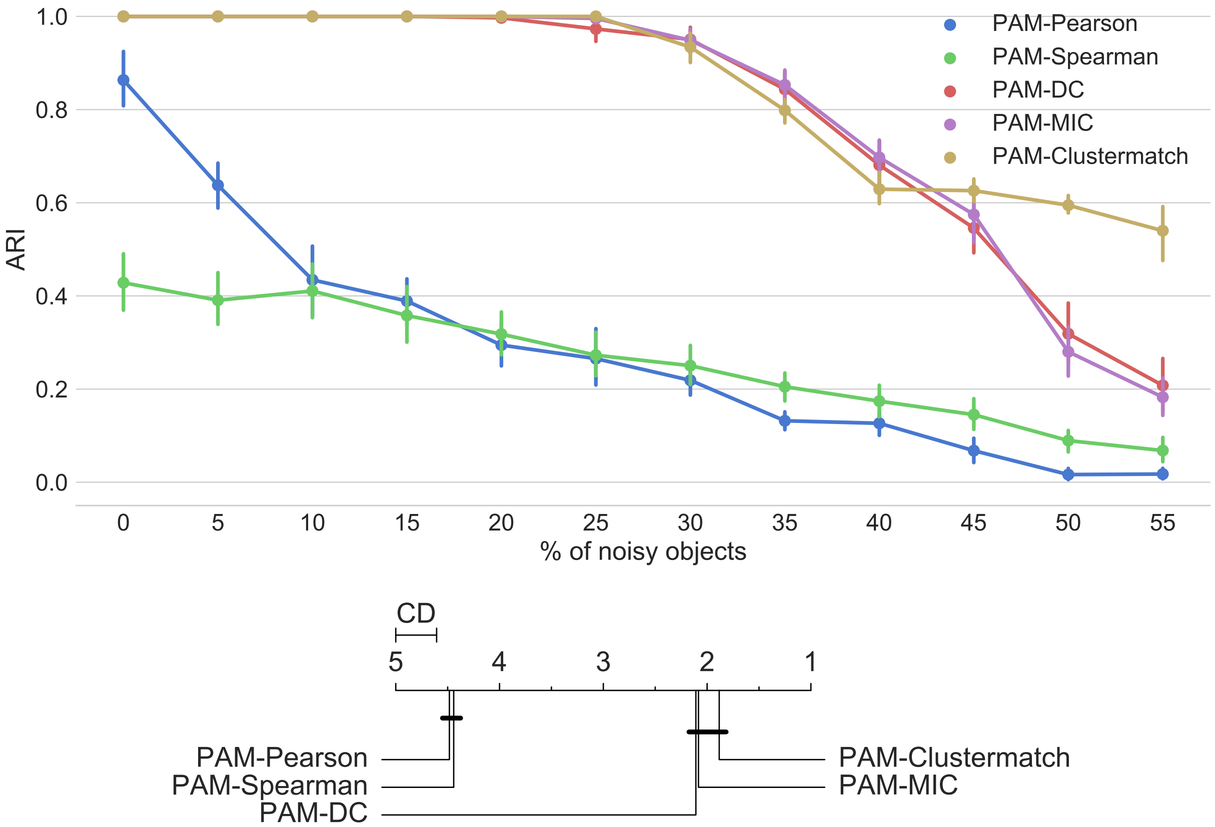
Performance using Hierarchical Clustering (HC)





Supplementary Figure 5: Clustering quality (y-axis) when using a hierarchical clustering (HC) algorithm with average linkage under different noise levels (x-axis) in the artificial dataset with linear and non-linear sources. Results show that HC-MIC and HC-Clustermatch are the best methods for low to medium noise levels.

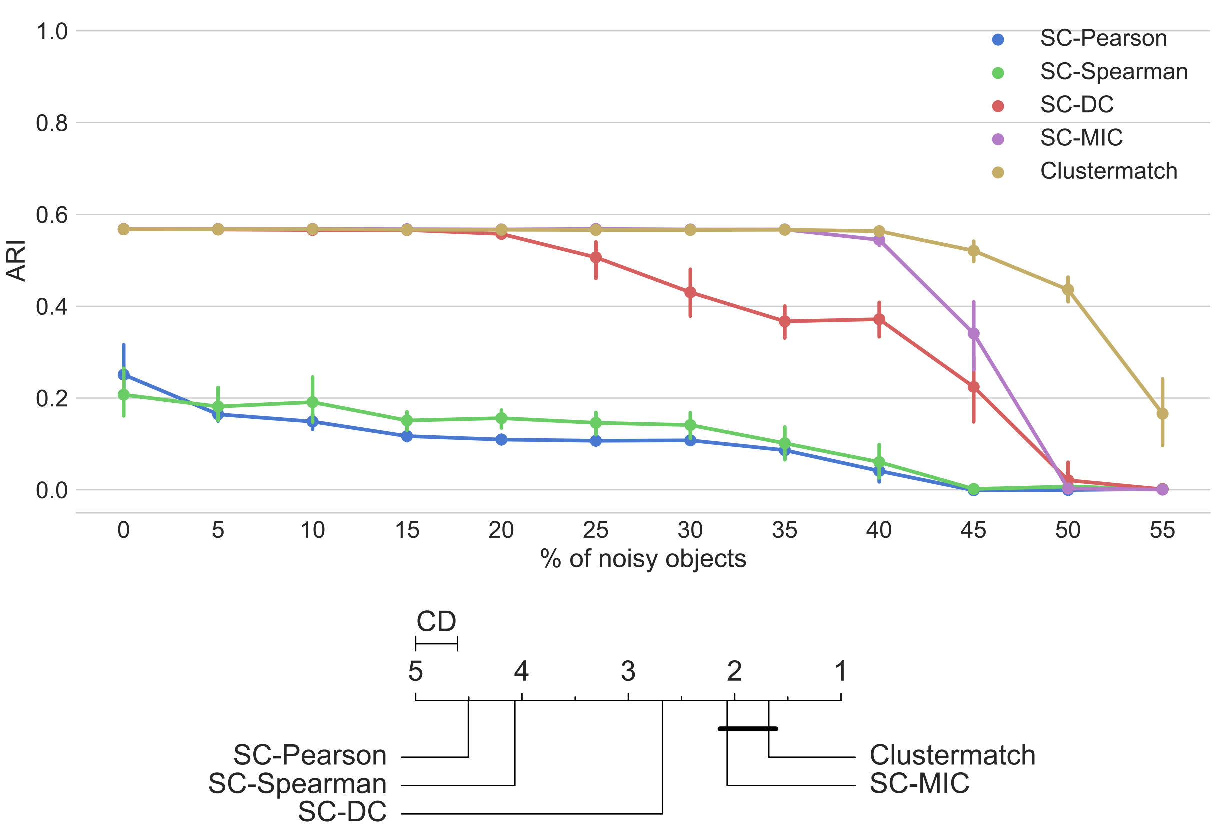
Performance using Partitioning Around Medoids (PAM)



Supplementary Figure 6: Clustering quality (y-axis) when using Partitioning Around Medoids (PAM) algorithm for all methods under different noise levels (x-axis) in the artificial dataset with linear and non-linear sources. Results show that PAM-MIC, PAM-DC and PAM-Clustermatch perform similarly for low to medium noise levels. However, PAM-Clustermatch outperforms all the other methods for highly noisy data.

sinc(i) Research Institute for Signals, Systems and Computational Intelligence (fich.unl.edu.ar/sinc)
M. Pividori, A. Cernadas, L. A. Haro, F. Carrari, G. Stegmayer & D. H. Milone; "Clustermatch: discovering hidden relations in highly diverse kinds of qualitative and quantitative data without standardization"
Bioinformatics, 2018.

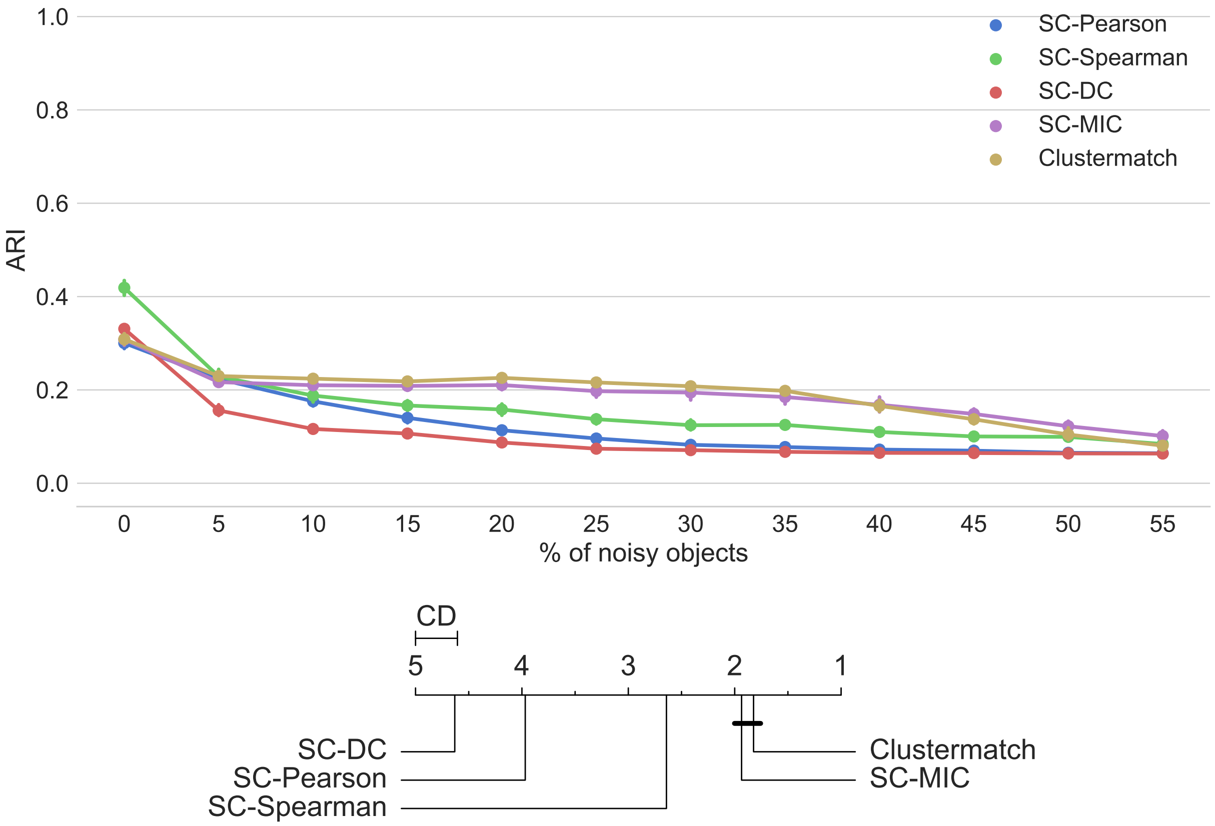
Performance using an extremely low number of final clusters ($k = 2$)



Supplementary Figure 7: Clustering quality (y-axis) when using $k = 2$ (number of final clusters) for all methods under different noise levels (x-axis) in the artificial dataset with linear and non-linear sources. Results show that when the number of clusters is very low, SC-MIC and Clustermatch are the best methods for low to medium noise levels, with Clustermatch performing better for the most noisy levels.

sinc(i) Research Institute for Signals, Systems and Computational Intelligence (fich.unl.edu.ar/sinc)
M. Pividori, A. Cernadas, L. A. Haro, F. Carrari, G. Stegmayer & D. H. Milone; "Clustermatch: discovering hidden relations in highly diverse kinds of qualitative and quantitative data without standardization"
Bioinformatics, 2018.

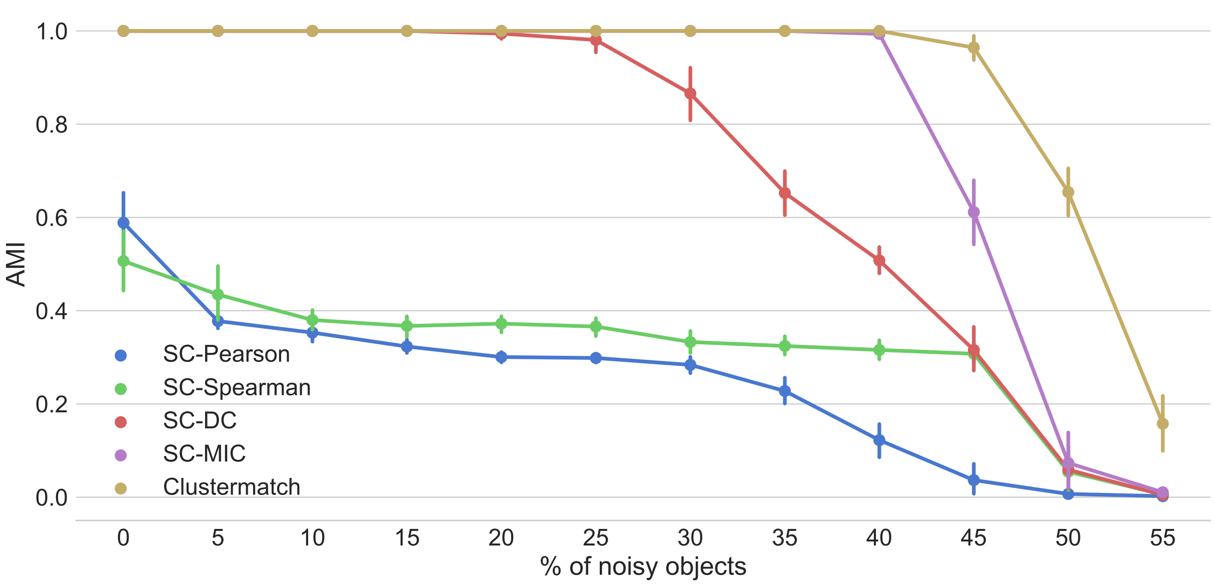
Performance using an extremely high number of final clusters ($k = 18$)

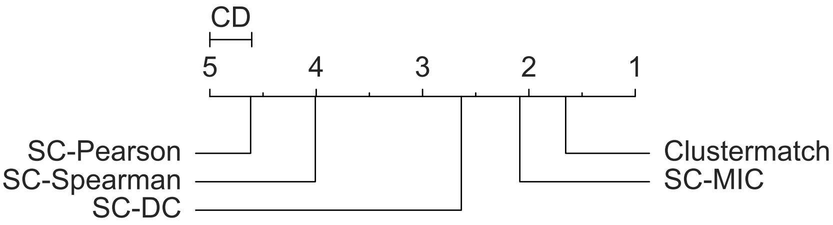


Supplementary Figure 8: Clustering quality (y-axis) when using $k = 18$ (number of final clusters) for all methods under different noise levels (x-axis) in the artificial dataset with linear and non-linear sources. Although in general clustering quality is low for all methods when using an extremely high k for the problem under study, SC-MIC and Clustermatch perform better than the rest in most of the cases (as shown in the CD diagram), with SC-Spearman outperforming SC-DC and SC-Pearson.

sinc(i) Research Institute for Signals, Systems and Computational Intelligence (fich.unl.edu.ar/sinc)
M. Pividori, A. Cernadas, L. A. Haro, F. Carrari, G. Stegmayer & D. H. Milone; "Clustermatch: discovering hidden relations in highly diverse kinds of qualitative and quantitative data without standardization"
Bioinformatics, 2018.

Final clustering evaluation using adjusted mutual information (AMI)



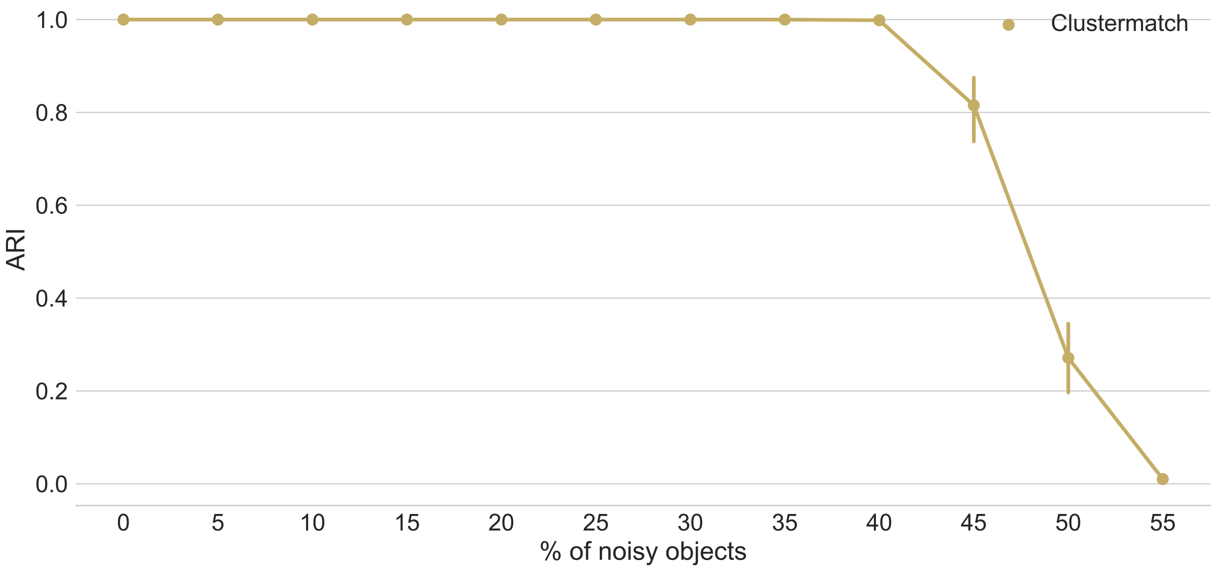


Supplementary Figure 9: Clustering quality (y-axis) measured with the adjusted mutual information (AMI) index (Vinh et al. 2010) for all methods under different noise levels (x-axis) in the artificial dataset with linear and non-linear sources. The top line graph and the CD diagram below show, both, that results are completely consistent with those shown in Figure 3, where the index to assess the clustering quality is the adjusted Rand index (ARI).

sinc(i) Research Institute for Signals, Systems and Computational Intelligence (fich.unl.edu.ar/sinc)
M. Pividori, A. Cernadas, L. A. Haro, F. Carrari, G. Stegmayer & D. H. Milone; "Clustermatch: discovering hidden relations in highly diverse kinds of qualitative and quantitative data without standardization"
Bioinformatics, 2018.

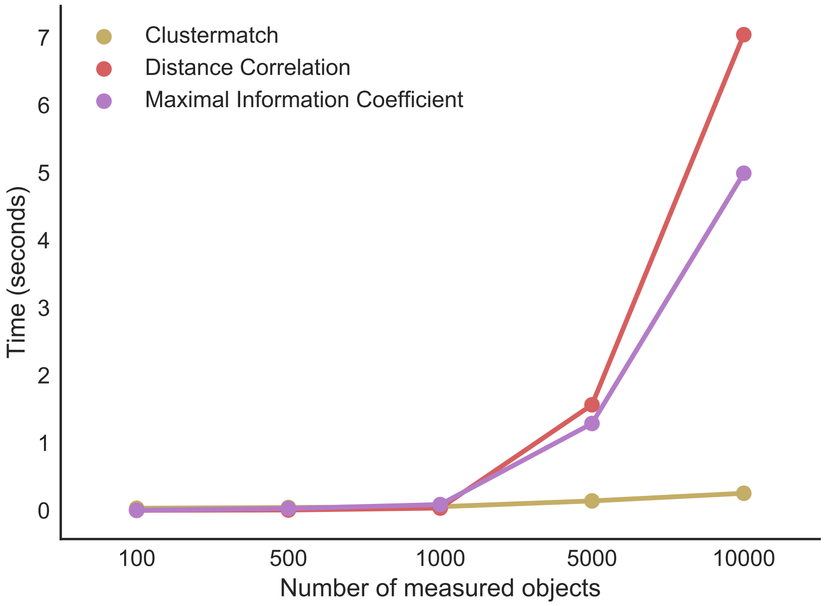
sinc(i) Research Institute for Signals, Systems and Computational Intelligence (fich.unl.edu.ar/sinc)
M. Pividori, A. Cernadas, L. A. Haro, F. Carrari, G. Stegmayer & D. H. Milone; "Clustermatch: discovering hidden relations in highly diverse kinds of qualitative and quantitative data without standardization"
Bioinformatics, 2018.

Performance of Clustermatch on mixed numerical and categorical sources



Supplementary Figure 10: Clustering quality (y-axis) for Clustermatch under different noise levels (x-axis) in the artificial dataset with linear, non-linear and categorical sources (which is exemplified in Supplementary Figures 3 and 4). Results are consistent with those shown in Figure 3, where the artificial dataset is the same but with only numerical data sources. Despite having less information in the categorical variables (compared to the numerical ones), Clustermatch obtains very high quality partitions of the variables until a noise level of 40%. Note that none of the other methods can be applied directly to these mix of numerical and categorical sources.

Time complexity



Supplementary Figure 11: Time complexity for methods with different number of measured objects for each variable. Time reported is the average of 10 repetitions for a pair of variables. For computing the Maximal Information Coefficient we employed `minepy` v1.2.1; for Distance Correlation we used an implementation in Python, `distcorr` (faster than the implementation of the R package `energy`).

sinc(*i*) Research Institute for Signals, Systems and Computational Intelligence (fich.unl.edu.ar/sinc)
M. Pividori, A. Cernadas, L. A. Haro, F. Carrari, G. Stegmayer & D. H. Milone; "Clustermatch: discovering hidden relations in highly diverse kinds of qualitative and quantitative data without standardization"
Bioinformatics, 2018.

Relationships analized from the tomato dataset

Supplementary Table 1: Summary of relationships found in Section 4.2. The related variables are listed in column two, as well as a short description taken from the main text and the studies where these associations were first published.

Relationship number	Related variables	Description	Published in
1	• *geranylacetone* • *6-methyl-5-hepten-2-one* (MHO) • Flavor • *guaiacol*	Carotenoid-derived volatile compounds (VOC) *geranylacetone* and *6-methyl-5-hepten-2-one* (MHO) are significantly associated to tomato flavor.	Tieman et al. 2017
2	• *Lobular shape* • *Height/width ratio*	. . . strong, yet logical, relationship between shape (e.g. lobular) and size ratios determinations (e.g. height/width relationship). Note that in an integrated analysis, Clustermatch disentangled these types of non-linear correlation.	Well-known
3	• *Glutamate* • *Pyroglutamate* • *GABA* • *oxoglutarate* • *1-aminocyclopropane-1-carboxylic acid* (ACC)	*GABA* is well described as a positive regulator of the *1-aminocyclopropane-1-carboxylate* oxidase expression, a key enzyme for ethylene biosynthesis.	This study
4	• Flavor and aroma descriptors • Components of the ascorbate metabolic pathway • Antioxidant capacity (TEAC determinations) • *Alpha-terpineol*	Analyses of significant connections revealed a clear relationship between the categorical variables describing consumer's sensory attributes, such as flavor and aroma descriptors, with intermediates of the ascorbate metabolic pathway (e.g., *Glucuronate 3,6-lactone* and *Glucuronate/galacturonate*; P<0.002). Accumulation of *alpha-terpineol*, a volatile monoterpenoid alcohol with a wide range of biological applications, is induced in reducing cellular environments.	This study
5	• Aroma • Unknown metabolite in the mass spec, *UNK m/z 95*	We observed aroma descriptors strongly associated with an unknown metabolite in the mass spec collection, *UNK m/z 95*, that was previously reported to be associated with non-characteristic tomato taste and odour.	This study