Using Total Correlation to Discover Related Clusters of Clinical Chemistry Parameters

Tamás Ferenci
Physiological Controls Group
Óbuda University
Budapest, Hungary
Email: ferenci.tamas@nik.uni-obuda.hu

Levente Kovács
Physiological Controls Group
Óbuda University
Budapest, Hungary
Email: kovacs.levente@nik.uni-obuda.hu

***Abstract*—Clinical chemistry tests are widely used in medical diagnosis. Physicians typically interpret them in a univariate sense, by comparing each parameter to a reference interval, however, their correlation structure may also be interesting, as it can shed light on common physiologic or pathological mechanisms. The correlation analysis of such parameters is hindered by two problems: the relationships between the variables are sometimes non-linear and of unknown functional form, and the number of such variables is high, making the use of classical tools infeasible. This paper presents a novel approach to address both problems. It uses an information theory-based measure called total correlation to quantify the dependence between clinical chemistry variables, as total correlation can detect any dependence between the variables, non-linear or even non-monotone ones as well, hence it is completely insensitive to the actual nature of the relationship. Another advantage is that is can quantify dependence not only between pairs of variables, but between larger groups of variables as well. By the virtue of this fact, a novel approach is presented that can handle the high dimensionality of clinical chemistry parameters. The approach is implemented and illustrated on a real-life database from the representative US public health survey NHANES.**

I. Introduction

Clinical chemistry is the subspecialty of clinical pathology that focuses on the analysis of various body fluids [1], [2]. Perhaps the most important part is the analysis of the constituents of blood; these tests include haematology tests, renal and liver function tests, determination of various electrolyte concentrations etc. Such ”blood tests” are now routine diagnostic tools that are used by many clinical fields to come up with and verify diagnoses.

Physicians typically use tests results in a univariate sense (i.e. they interpret them separately) and the most abundant interpretation is to compare them to limits – so-called reference intervals [3] – that indicate the range which can be considered ”normal” and outside which some pathology can be suspected to be present.

Despite this, the multivariate structure of the test results might also be of interest. Stochastically connected parameters can indicate common mechanism behind the physiological regulation of the measured quantities, or the fact that the tests are affected by similar pathological processes. This phenomenon, and also its association with states like obesity has been discussed in the literature [4].

However, the multivariate interpretation of laboratory results is hindered by two problems. First and foremost, the number of tests, even the commonly used ones, is huge (and increasing). With several dozen – or more – variables, the classical methods of investigating correlations can not be applied or are infeasible. This can be considered as an example of the ”curse of dimensionality” [5]. Second, the clinical chemistry parameters are themselves often not normally distributed, their connections can be non-linear (or perhaps even non-monotone).

This paper presents a novel approach to address these issues. The problem of non-linear (or non-monotone) relationships is addressed by the application of a dependence-measure that is insensitive to the nature of the relationship, while the high dimensionality is addressed by a clustering algorithm that reduces the search space by rearranging it is a search tree and pruning unnecessary parts as early as possible.

II. Materials and Methods

First, different metrics to measure the dependence of variables will be presented, culminating in the introduction of the metric that will be used in our approach. The estimation of this metric from sample will be discussed in more detail afterwards. Next, the other important element of the approach is presented, the clustering algorithm. The approach will be illustrated on a real-life example, this is presented next, together with the details of the software implementation of the approach.

A. Measuring Dependence of Variables

The issue of quantitatively measuring the dependence (stochastic connection, relationship) of variables dates back to the late 19th century [6]. We will now confine our discussion to continuous variables, i.e. the question of correlation.

1) Bivariate Case: First the theory of the bivariate case was devised, initially focusing on capturing linear relationships (”Pearson” or linear or product-moment correlation [7]). While this was of huge importance for a myriad of applications (linear regression, for instance), it is unable to capture non-linear connections which poses a problem in certain circumstances. To circumvent this, several alternative measures were developed later, the most famous being the Spearman-ρ [8] and the Kendall-τ [9]. These coefficients – both being rank correlation [10] – are able to capture arbitrary monotone (but no longer necessarily linear) relationship between the variables.

978-1-4799-5996-9/14/$31.00 ©2014 IEEE

These are, however, still unable to capture non-monotonic connections. When one is willing to assume a given functional form on the relationship, the estimation of its strength is relatively straightforward, but to capture this without any presumption, i.e. completely non-parametrically, is a much more complicated issue. One classical solution is the application of mutual information. To introduce this, first note that the *entropy* [11] (or information entropy) of a discrete random variable X concentrated to the real values $\{x_1, x_2, \ldots, x_k\}$ with probability mass function $p(x) = \mathbb{P}(X = x)$ is

$$H(X) = \mathbb{E}\left[-\log p(X)\right] = -\sum_{i=1}^{k} p(x_i)\log p(x_i). \tag{1}$$

This definition founds its roots in thermodynamics, and is related to the uncertainty in the outcome of the variable, or its information content [12]. Entropy can be analogously defined for multivariable case [11], i.e. for a p-dimensional random vector $\mathbf{X}$, called *joint entropy*:

$$\begin{aligned} H\left(X_1, X_2, \ldots, X_p\right) &= \mathbb{E}\left[-\log p\left(X_1, X_2, \ldots, X_p\right)\right] \\ &= -\sum_{i_1=1}^{k_1} \cdots \sum_{i_p=1}^{k_p} p\left(x_{i_1}, \ldots, x_{i_p}\right) \log p\left(x_{i_1}, \ldots, x_{i_p}\right). \end{aligned} \tag{2}$$

(The entropy does not depend on the values the random variable can take, only on its probabilities. Also note that the joint entropy is not fundamentally different from the univariate one: it can be considered as a univariate entropy for a variable defined on the product space.) Considering the "information content" view of the entropy, $H(X, Y)$ is the information in the joint distribution (dealing now with the bivariate case), that is, it includes every information known on the distribution of these two variables, while the sum of the informations that are known marginally, i.e $H(X) + H(Y)$, includes both the information that is "contained" only in either of the variables, and twice the information that is "contained" in both. Thus, it is quite logical to define the quantity

$$I(X, Y) = H(X) + H(Y) - H(X, Y), \tag{3}$$

which is called *mutual information* [11]. As it follows from the above reasoning, this can be used to measure how the two variable is connected, without *any* presumption on the nature of their relationship: we can measure their linear, arbitrary monotone, and even non-monotone relationships. (This, of course, does not mean that such measures are universally superior to the more traditional correlation coefficients: apart from interpretability, these might also be inferior in inductive statistical sense when being estimated from a sample.) Note that mutual information is the Kullback–Leibler divergence [13] between the variable's actual joint distribution, and the joint distribution obtained when they are presumed to be independent, i.e. the product of their marginal distributions: $I(X, Y) = D_{\mathrm{KL}}\left(p(x, y) \| p(x) p(y)\right)$.

Mutual information is not the only measure devised to capture the correlation of two variable universally (in the above sense). Another notable example is the *distance correlation* [14], [15] that was only recently introduced.

2) p-variate Case: All the above measures characterize the dependence between two variables, we might however be also be interested in characterizing the dependence between $p > 2$ variables. The traditional correlation has such no generalization (multiple correlation [16], that can be defined using the classical terms, is not symmetrical, i.e. requires the declaration of a variable to be dependent on the others, hence it is unfit to measure the overall dependence of the variables in which the role of the variables is obviously symmetrical).

Mutual information can, in contrast, be extended to $p > 2$ variables, but this generalization is not straightforward. Several alternatives have been described [17], [18], including interaction information [19], total correlation and dual total correlation [20]. Each grabs different aspect of mutual information that is generalized to higher dimensions.

Total correlation will be now employed for the further studies as a multivariate measure of the – universal – dependence of variables. *Total correlation* [21] is defined as

$$C\left(X_1, X_2, \ldots, X_p\right) = \left[\sum_{i=1}^{p} H(X_i)\right] - H\left(X_1, X_2, \ldots, X_p\right). \tag{4}$$

It is immediately obvious that this definition generalizes (3), and it can also be shown that

$$\begin{aligned} &C\left(X_1, X_2, \ldots, X_p\right) \\ &= D_{\mathrm{KL}}\left(p\left(x_1, x_2, \ldots, x_p\right) \| p(x_1) p(x_2) \ldots p\left(x_p\right)\right). \end{aligned} \tag{5}$$

stands as well, i.e. it is still the Kullback–Leibler divergence between the actual joint distribution of the variables and the joint distribution obtained when they are presumed to be independent.

Note that total correlation measures every dependence "buried" within the connections of the variables (also including every possible interaction). These may be decomposed, as it has has been discussed in the literature [22], [21] but is not elaborated in more detail here.

B. Estimating Total Correlation from Sample

Even the estimation of entropy (from a sample) is not a trivial issue. Considering now continuous variables, there is a definition for entropy holding for continuous variables as well, called differential entropy [11]. It is the straightforward generalization of the discrete definition, that is, for a continuous random variable X with probability density function f, the differential entropy is

$$H(X) = -\int_{\mathrm{supp}(f)} f(x) \log f(x)\,\mathrm{d}x. \tag{6}$$

There is a great deal of approaches to estimate this quantity from a sample [23]. Now a method will be used which first discretizes the continuous variable, and then applies a James–Stein-type shrinkage estimator [24]. This estimator is demonstrated to be effective in a number of scenarios with $p = 1000$ variables even with less than 100 sample [24].

The discretization was performed by equal width binning, with 4 bins for each variable (i.e. the range of the given variable was divided to four segments of equal length). Higher number of bins offers a more faithful representation of the original variable's distribution, but it also quickly increases the dimensionality of the problem. (Note that after discretization, the dimension of the entropy estimation problem for p variables will be b^p, where b is the number of bins. That is, the dimensionality is increasing exponentially with the number of bins.)

After the marginal and joint entropies are estimated with the above method, the total correlation is calculated directly, using the definition.

C. Clustering Algorithm

Three possible clustering algorithms will be presented and investigated that use total correlation to measure the dependence between a set of variables.

1) Traditional agglomerative hierarchical clustering: Agglomerative hierarchical clustering [25] defines distances between clusters using the distances between the objects to be clustered (which the algorithm receives as an input) and merges the nearest clusters starting from a situation in which every object is considered to be a separate cluster until all is merged to the same cluster. The clusters that are merged, together with the distance of merging can be visualized on a very illustrative diagram called dendrogram [25].

The major point of hierarchical clustering is that it directly reduces the definition of cluster-distance to object-distance (by taking, for instance, the minimal or maximal or average distance between the objects of a cluster as the cluster's distance) without attempting to calculate a distance using the cluster as a whole. In other words, it always uses only pairwise distances, hence, it is unable to utilize total correlation's possibility to describe to correlation of $p > 2$ variables.

2) Greedy clustering using total correlation: To address the above issue, it is logical to define the similarity of two clusters as the total correlation after merging them, but otherwise follow the logic of hierarchical clustering [26]. That is: initially we consider every object to be a separate cluster, and merge those two objects/clusters in every step which results in the highest total correlation of the merged cluster from every possible merge. It may be called greedy clustering as it retains the greedy nature [27] of hierarchical clustering.

At this point it worth noting that total correlation has a monotonicity property, that is, the total correlation of a set of variables can not be smaller than the total correlation of any subset of them. The proof is simple by applying the chain rule for entropy [11]:

$$\begin{aligned}
C\left(X_1,\ldots,X_p\right) &= \left[\sum_{i=1}^{p} H\left(X_i\right)\right] - H\left(X_1,\ldots,X_p\right) \\
&= \left[\sum_{i=1}^{p-1} H\left(X_i\right)\right] - H\left(X_p|X_1,\ldots,X_{p-1}\right) \\
&+ H\left(X_1,\ldots,X_{p-1}\right) + H\left(X_p\right) = C\left(X_1,\ldots,X_{p-1}\right) \\
&+ \Big[H\left(X_p\right) - H\left(X_p|X_1,\ldots,X_{p-1}\right)\Big] \\
&\geq C\left(X_1,\ldots,X_{p-1}\right),
\end{aligned} \tag{7}$$

as the conditional entropy can not be greater than the unconditional entropy [11].

3) APRIORI-style clustering with threshold: Another possible algorithm is to relax the requirement of greedy cluster-growing (i.e. to select ”the” best merge), and rather to set a threshold – for the total correlation – above which we accept every cluster as an ”interesting” group of objects. Itself this definition is not useful due to the aforementioned monotonicity: it implies an upward closure property (if a set meets the criteria, every superset will also meet the threshold), but this can be addressed by requiring that only *minimally connected* groups are found, i.e. clusters that themselves meet the threshold for total correlation, but none of their subsets does.

This can be best imagined on a usual lattice that contains every possible subset of the variables, ordered by inclusion. Simply requiring a minimum total correlation would mean that if a node meets the criteria, every connected node above it will also meet the threshold. The minimally connected requirement however means that in such case, only the lowest group (node) will be found to be a cluster. The tree above can be *pruned.*

This dictates the following algorithm: first, two-element cluster are checked against the threshold. Those that meet it, are returned as ”interesting” clusters and the trees connected them are pruned (in effect, variables included in these clusters are removed). Then, with the remaining variables, three-element clusters are checked against the threshold. Once again, those that meet it, are returned as results, they are removed, and four-element clusters are checked within the remaining variable, and so on.

This search tree pruning, bottom-up logic is close to the very-well known APRIORI algorithm [28]. It is interesting to note that there exists an algorithm operating similarly to the one described here and also using entropy, called ENCLUS [29]. ENCLUS is a so-called subspace clustering algorithm [30], hence it is somewhat different from the approach discussed above: it also has criteria for the cluster's entropy to be low enough (as it represents ”interestingness” in subspace clustering context). As we are now interested in the correlation structure, this criteria is not needed.

D. Illustrative Example

To illustrate the above approach, data from the representative United States survey called National Health and Nutrition Examination Survey (NHANES) will be used. NHANES is now a continuous public health program, with results published in biannual cycles [31]. It is a nation-wide survey aimed to be representative for the whole civilian non-institutionalized US population, by employing a complex, stratified multi-stage probability sampling plan. The amount of collected data is tremendous (although sometimes varying from cycle to cycle), including demographic data, physical examination, collection of clinical chemistry parameters, and a thorough questionnaire concentrating on anamnesis and lifestyle.

In this example, data from the 2011-2012 cycle will be used, as these are the latest that are currently available [32]. To achieve more homogeneity, the database will be filtered to males, aged $>$ 18 years.

To account for the survey design of the NHANES, weighting has to be used. As per the analytic guidelines, the weights of the smallest analysis subpopulation have to be used, which were the so-called MEC weights in this case (as variables that were measured in the mobile examination center were also included in the analysis), named `WTMEC2YR`.

The following clinical chemistry parameters will be used in the example:

- Standard Biochemistry Profile (`BIOPRO_G` data file).
- Complete Blood Count with 5-Part Differential in Whole Blood (`CBC_G` data file).
- Glycohemoglobin (`GHB_G` data file).
- HDL-Cholesterol (`HDL_G` data file).

Together, they include 39 variables; they will be addressed by abbreviations, internationally used ones, wherever possible.

Demographic data were extracted from the Demographic Variables and Sample Weights (`DEMO_G`) data file.

Every subject with missing value was removed (resulting in a database without missing values); the final sample size was $n = 2480$.

E. Implementation

The approach described in this paper was implemented under `R` statistical program package, version 3.1.0 [33] using library `entropy` version 1.2.0 [34]. The script is available from the corresponding author on request.

III. Results and Discussion

Pairwise dependences between the variables of the database are shown on Fig. 1. As a comparison, the – traditional – correlation matrix of the database is also shown (Fig. 1a). The pairwise total correlations (mutual informations) are shown on Fig. 1b. Note that this latter can not be considered to be a valid correlation matrix, as the ”self-total correlation” need not be unity, and no normalizing was applied to enforce this. To achieve comparability of the two matrices Fig. 1a depicts the absolute values of the correlations. Both matrix is visualized with a heat map, using logarithmic colouring which was necessary as the correlations span several magnitudes.

These measures can be used to perform a traditional hierarchical clustering. Using Ward’s method as linkage criteria [35], [36], and the usual Lance–Williams algorithm [37], we obtain the dendrograms shown on Fig. 2. Like noted, this algorithm makes no actual use of the true strength of total correlation (that is, the ability to characterize the dependence between $p > 2$ variables), rather, it simply extends the traditional clustering (based on linear correlations) to be able to handle possibly non-linear relationships between the variables as well, without any further extension.

Note that we have used the matrix $\left[1-\log\left|c_{ij}\right|\right]_{i,j=1}^{n}$ in both cases to perform the clustering. The logarithm was used to aid the algorithm, as the (absolute) correlations spanned several magnitudes.

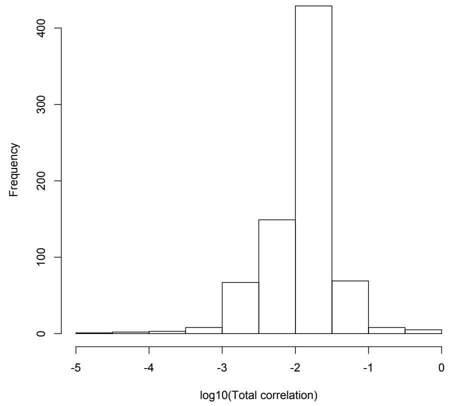


Fig. 3. Histogram of the pairwise total correlations between the variables of the database, on logarithmic scale.

While the dendrograms are globally different, many similar structure can be observed: GHB-GLU, MCV-MCH+MCHC-RDW+RBC-HGB-HCT, ANC-ABC, BUN-SCR are examples for structures that remain unchanged indicating no strong presence of non-linear relationships.

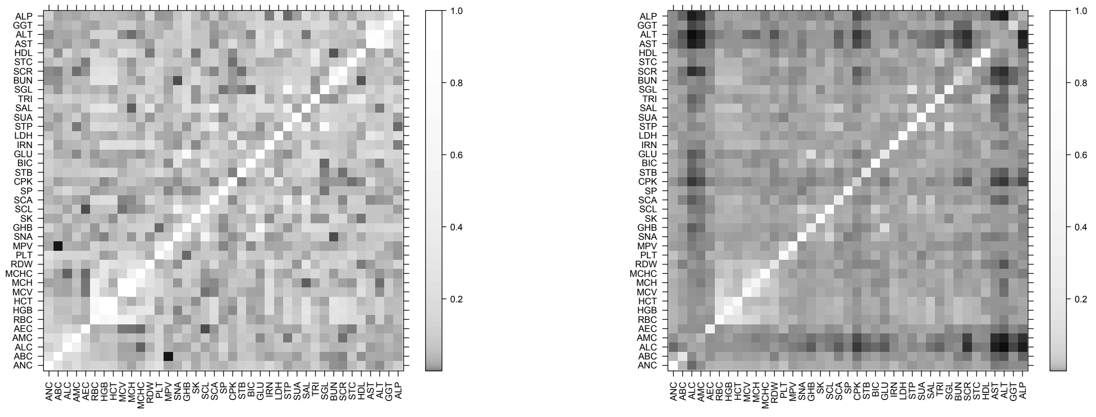
Next, we applied the method that do takes the multivariate dependences – through total correlation – into account. The greedy clustering however actually results in a single cluster being grown larger and larger. Namely, the merges in the first eight steps are the following:

$$
\begin{aligned}
&\{\mathrm{HGB},\mathrm{HCT}\} \to \{\mathrm{HGB},\mathrm{HCT},\mathrm{RBC}\} \to\\
&\to \{\mathrm{HGB},\mathrm{HCT},\mathrm{RBC},\mathrm{MCH}\} \to\\
&\to \{\mathrm{HGB},\mathrm{HCT},\mathrm{RBC},\mathrm{MCH},\mathrm{MCV}\} \to\\
&\to \{\mathrm{HGB},\mathrm{HCT},\mathrm{RBC},\mathrm{MCH},\mathrm{MCV},\mathrm{MCHC}\} \to\\
&\to \{\mathrm{HGB},\mathrm{HCT},\mathrm{RBC},\mathrm{MCH},\mathrm{MCV},\mathrm{MCHC},\mathrm{IRN}\} \to\\
&\to \{\mathrm{HGB},\mathrm{HCT},\mathrm{RBC},\mathrm{MCH},\mathrm{MCV},\mathrm{MCHC},\mathrm{IRN},\mathrm{ANC}\}\,.
\end{aligned}
$$

This can be readily explained by the monotonicity of the total correlation already discussed: clusters with high number of variables have an advantage when their total correlation is compared to smaller clusters (simply due to the inflation of total correlation), and after a point, this can not be ”defeated” by smaller clusters, hence it will be necessity that a single cluster is grown until it includes every variable. This phenomenon severely limits the applicability of this algorithm (it is rather ordinating the variables, then performing a true clustering).

Finally, we performed the APRIORI-style, bottom-up clustering with pruning. The histogram of the pairwise total correlations (on a logarithmic scale) is shown on Fig. 3.

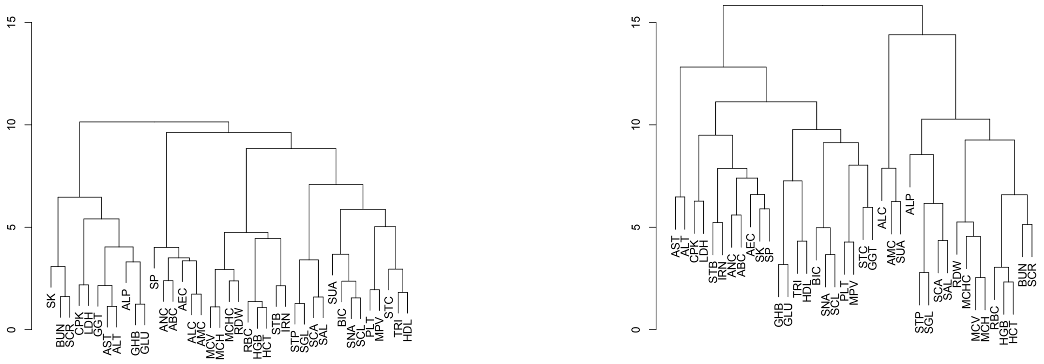
Based on this, we have chosen a threshold of $\varepsilon = 0.1$, which resulted in 6 clusters labelled as interesting, all including two variables of course. To obtain minimally connected clusters, the variables in these clusters were removed and the algorithm was re-run, this time using triples. Using the same threshold again, 1 three-element cluster is obtained and then, repeating the procedure, 1 four-element cluster. These results, i.e. the ”interesting” groups of clinical chemistry parameters



(a) Absolute value of linear correlation

(b) Total correlation

Fig. 1. Pairwise dependences of the variables in the database, using absolute value of linear correlation coefficient (a) and total correlation (b). Note the logarithmic colouring.



(a) Log absolute value of linear correlation

(b) Log of total correlation

Fig. 2. Dendrograms for the hierarchical clustering of variables using linear correlation coefficient (a) and total correlation (b). Note that the logarithms of the absolute values of correlations were used in both cases.

TABLE I. Minimally connected clusters obtained with the presented novel clustering method

Cluster size	Cluster members	Total correlation
2	$\{\mathrm{RBC}, \mathrm{HGB}\}$	0.1538
2	$\{\mathrm{RBC}, \mathrm{HCT}\}$	0.1502
2	$\{\mathrm{HGB}, \mathrm{HCT}\}$	0.2650
2	$\{\mathrm{MCV}, \mathrm{MCH}\}$	0.2129
2	$\{\mathrm{GHB}, \mathrm{GLU}\}$	0.1125
2	$\{\mathrm{STP}, \mathrm{SGL}\}$	0.1681
3	$\{\mathrm{SNA}, \mathrm{SCL}, \mathrm{BIC}\}$	0.1593
4	$\{\mathrm{ANC}, \mathrm{PLT}, \mathrm{MPV}, \mathrm{SK}\}$	0.1001

obtained with the presented clustering algorithm are shown on Table I.

IV. Conclusion

Total correlation can be effectively employed to detect connected groups among variables of a dataset, even in the possible presence of non-linear (or even non-monotone) relationships. The presented algorithm works well for high-dimensionality datasets as well, as it was demonstrated on a real-life example of clinical chemistry parameters. In addition to the elaboration of the details, and a thorough validation, several improvement – such as the discovery of non-minimally connected clusters – of this approach can be imagined, which might be worthy of further research.

Acknowledgement

Levente Kovács is supported by the János Bolyai Research Scholarship of the Hungarian Academy of Sciences.

References

[1] M. L. Bishop, E. P. Fody, and L. E. Schoeff, *Clinical Chemistry: Principles, Techniques, and Correlations*. Lippincott Williams & Wilkins, 2013.

[2] C. A. Burtis, E. R. Ashwood, and D. E. Bruns, *Tietz textbook of clinical chemistry and molecular diagnostics*. Elsevier Health Sciences, 2012.

[3] G. L. Horowitz, *Defining, establishing, and verifying reference intervals in the clinical laboratory: Approved guideline*. Clinical and Laboratory Standards Institute, 2010.

[4] T. Ferenci, "Two applications of biostatistics in the analysis of pathophysiological processes," Ph.D. dissertation, Obudai Egyetem, 2013.

[5] T. Hastie, R. Tibshirani, J. Friedman, T. Hastie, J. Friedman, and R. Tibshirani, *The elements of statistical learning*. Springer, 2009.

[6] J. Lee Rodgers and W. A. Nicewander, "Thirteen ways to look at the correlation coefficient," *The American Statistician*, vol. 42, no. 1, pp. 59–66, 1988.

[7] K. Pearson, "Notes on the history of correlation," *Biometrika*, pp. 25–45, 1920.

[8] C. Spearman, "The proof and measurement of association between two things," *The American Journal of Psychology*, vol. 15, no. 1, pp. 72–101, 1904.

[9] M. G. Kendall, "A new measure of rank correlation," *Biometrika*, pp. 81–93, 1938.

[10] M. Kendall and J. Gibbons, *Rank Correlation Methods*. Edward Arnold, 1990.

[11] T. M. Cover and J. A. Thomas, *Elements of information theory*. John Wiley & Sons, 2012.

[12] A. Greven, G. Keller, and G. Warnecke, *Entropy*. Princeton University Press, 2003.

[13] S. Kullback and R. A. Leibler, "On information and sufficiency," *The Annals of Mathematical Statistics*, pp. 79–86, 1951.

[14] G. J. Székely, M. L. Rizzo, N. K. Bakirov *et al.*, "Measuring and testing dependence by correlation of distances," *The Annals of Statistics*, vol. 35, no. 6, pp. 2769–2794, 2007.

[15] G. J. Székely, M. L. Rizzo *et al.*, "Brownian distance covariance," *The annals of applied statistics*, vol. 3, no. 4, pp. 1236–1265, 2009.

[16] J. Cohen, *Applied Multiple Regression/correlation Analysis for the Behavioral Sciences*. Routledge, 2003.

[17] N. Timme, W. Alford, B. Flecker, and J. M. Beggs, "Multivariate information measures: an experimentalist's perspective," *arXiv preprint arXiv:1111.6857*, 2011, 1111.6857.

[18] T. Van de Cruys, "Two multivariate generalizations of pointwise mutual information," in *Proceedings of the workshop on distributional semantics and compositionality*. Association for Computational Linguistics, 2011, pp. 16–20.

[19] W. J. McGill, "Multivariate information transmission," *Psychometrika*, vol. 19, no. 2, pp. 97–116, 1954.

[20] T. S. Han, "Nonnegative entropy measures of multivariate symmetric correlations," *Information and Control*, vol. 36, no. 2, pp. 133–156, 1978.

[21] S. Watanabe, "Information theoretical analysis of multivariate correlation," *IBM Journal of research and development*, vol. 4, no. 1, pp. 66–82, 1960.

[22] W. R. Garner, *Uncertainty and structure as psychological concepts*. Wiley, 1962.

[23] J. Beirlant, E. J. Dudewicz, L. Györfi, and E. C. Van der Meulen, "Nonparametric entropy estimation: An overview," *International Journal of Mathematical and Statistical Sciences*, vol. 6, no. 1, pp. 17–39, 1997.

[24] J. Hausser and K. Strimmer, "Entropy inference and the james-stein estimator, with application to nonlinear gene association networks," *The Journal of Machine Learning Research*, vol. 10, pp. 1469–1484, 2009.

[25] B. Everitt, S. Landau, M. Leese, and D. Stahl, *Cluster Analysis*. Wiley, 2011.

[26] S. Ondáš, J. Juhár, M. Pleva, M. Lojka, E. Kiktová, M. Sulír, A. Čižmár, and R. Holcer, "Speech technologies for advanced applications in service robotics," *Acta Polytechnica Hungarica*, vol. 10, no. 5, 2013.

[27] T. Cormen, C. Leiserson, R. Rivest, and C. Stein, *Introduction To Algorithms*. MIT Press, 2001.

[28] R. Agrawal, R. Srikant *et al.*, "Fast algorithms for mining association rules," in *Proc. 20th int. conf. very large data bases, VLDB*, vol. 1215, 1994, pp. 487–499.

[29] C.-H. Cheng, A. W. Fu, and Y. Zhang, "Entropy-based subspace clustering for mining numerical data," in *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 1999, pp. 84–93.

[30] L. Parsons, E. Haque, and H. Liu, "Subspace clustering for high dimensional data: a review," *ACM SIGKDD Explorations Newsletter*, vol. 6, no. 1, pp. 90–105, 2004.

[31] Centers for Disease Control and Prevention, National Center for Health Statistics, "National Health and Nutrition Examination Survey," http://www.cdc.gov/nchs/nhanes.htm, 2013, [Online; accessed 20. 06. 2014.]. [Online]. Available: http://www.cdc.gov/nchs/nhanes.htm

[32] ——, "National Health and Nutrition Examination Survey, NHANES 2011-2012," http://wwwn.cdc.gov/nchs/nhanes/search/nhanes11_12.aspx, 2013, [Online; accessed 20. 06. 2014.]. [Online]. Available: http://wwwn.cdc.gov/nchs/nhanes/search/nhanes11_12.aspx

[33] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2014. [Online]. Available: http://www.R-project.org/

[34] J. Hausser and K. Strimmer, *entropy: Estimation of Entropy, Mutual Information and Related Quantities*, 2013, R package version 1.2.0. [Online]. Available: http://CRAN.R-project.org/package=entropy

[35] J. H. Ward Jr, "Hierarchical grouping to optimize an objective function," *Journal of the American statistical association*, vol. 58, no. 301, pp. 236–244, 1963.

[36] F. Murtagh and P. Legendre, "Ward's hierarchical agglomerative clustering method: Which algorithms implement ward's criterion?" *arXiv preprint arXiv:1111.6285*, 2013, 1111.6285.

[37] G. Lance and W. Williams, "A general theory of classificatory sorting strategies 1. Hierarchical systems," *The Computer Journal*, vol. 9, no. 4, pp. 373–380, 1967.