

Partial ordering of dissimilarity measures for gene clustering

F. Marta L. Di Lascio and Alberto Roverato

Abstract Clustering methods are widely used in the analysis of microarray data for their ability to discover co-regulated genes. In a previous work we introduced, in a hierarchical clustering context, a theoretical framework for the comparison of dissimilarity measures on the basis of their ability to identify functional modules consisting of a transcription factor and the associated target genes. In this paper we extend these results by including in the analysis a set of dissimilarity measures based on the “1 – absolute value of correlation coefficient” proximity between genes. We show that such theoretical framework allows one to obtain a partial ordering of the considered measures in which we identify three minimal elements that are then compared on the basis of both simulated and real data.

Key words: Bayesian networks, cluster analysis, Gaussian DAG model, microarray data, sparse canonical correlation, transcription module, Wilks’ Λ .

1 Introduction

Clustering methods have been widely used for the analysis of expression data and they have been applied to both the sampling observation (i.e. experimental conditions) and the variables (i.e. genes); see [7], Chapter 7.

We focus on gene clustering by means of agglomerative hierarchical procedures (see [7]). The application of hierarchical clustering requires first the specification of a proximity measure giving the dissimilarity between pair of genes and, then, a linkage rule as a measure of dissimilarity between groups of genes. A variety of these measures is available (see [7]) but there is no stan-

F. Marta L. Di Lascio e-mail: francesca.dilascio@unibo.it

Alberto Roverato e-mail: alberto.roverato@unibo.it

Department of Statistics, via Belle Arti 41, Università di Bologna, Bologna, Italy.

dard way to compare their different performance and this makes it difficult to compare objectively clustering methods.

One of the main goals of clustering genes is the identification of groups of genes which belong to a same functional module. Roverato and Di Lascio [6] proposed a criterium for the comparison of dissimilarity measures on the basis of their ability to identify transcription modules. We refer to a transcription module as a set of genes made up of a transcription factor, that is a gene which serves to regulate other genes, and the associated set of target genes; see [6] for further discussion on this point. The above criteria is used in [6] to provide a total ordering of a set of dissimilarity measures. All these measures are based on the same proximity between genes that is “1 – squared correlation coefficient”. However, also the “1 – absolute value of correlation coefficient” is a commonly used proximity measure in the context of gene clustering (see [7]). For this reason, it is of interest to extend the analysis in [6] by including a set of dissimilarity measures based on the latter proximity. We show that the notion of strict dominance induces a strict partial ordering in this extended set of dissimilarity measures. The “smaller” elements have better properties than “larger” ones and we show that this poset has three minimal elements, which are then compared on the basis of both simulated and real data.

The paper is organized as follows. In Section 2 the notation and the conceptual background for this paper is given. Section 3 shows the comparison between different dissimilarity measures and their partial ordering. Section 4 reports the applications to simulated and real data and, finally, Section 5 contains a brief discussion.

2 Notation and background

We denote by $\mathbf{tf}$ a transcription factor and by V a set of target genes of $\mathbf{tf}$ under a specific experimental condition. A *transcription module* for $\mathbf{tf}$ is consequently defined as $V^{\mathbf{tf}} := \{\mathbf{tf}\} \cup V$ and we use numerical labels for target genes, $V = \{1, 2, \dots, p\}$. The transcription module as a mathematical object is formalized through a random vector $X_{V^{\mathbf{tf}}}$ indexed by $V^{\mathbf{tf}}$.

For a general introduction to cluster analysis the reader is referred to [2]. *Agglomerative hierarchical clustering* generates a hierarchy of nested clusters by starting with each gene forming a cluster and by merging, at every step of the procedure, exactly one pair of clusters. The algorithm stops when the last two clusters are merged to form $V^{\mathbf{tf}}$. In order to formally deal with hierarchical clustering algorithms, for a subset $U \subseteq V$ we denote by $\tilde{\mathcal{P}}(U^{\mathbf{tf}})$ the set of all partitions $\mathcal{P} = \{A^{\mathbf{tf}}, B_1, \dots, B_r\}$ of $U^{\mathbf{tf}}$ where $r \in \{0, 1, \dots, |U|\}$ and $|U|$ denotes the cardinality of U . Assume that the set of genes $V^{\mathbf{tf}}$ is submitted to an agglomerative hierarchical clustering algorithm. Then, every step of the algorithm starts from a partition $\mathcal{P}_r = \{A^{\mathbf{tf}}, B_1, \dots, B_r\} \in \tilde{\mathcal{P}}(V^{\mathbf{tf}})$ and merges two elements of $\mathcal{P}_r$ producing a new partition $\mathcal{P}_{r-1} \in \tilde{\mathcal{P}}(V^{\mathbf{tf}})$.

Roverato and Di Lascio [6] proposed a comparison between dissimilarity measures in the context of transcription modules identification. They introduced two criteria to theoretically assess the adequacy and to compare dissimilarity measures. They firstly introduced the concepts of *adequateness* and *strong adequateness* of a dissimilarity measure (see Definitions 1 and 2 in [6]) relying on the standpoint that a dissimilarity measure applied to a set of genes forming a transcription module $V^{\mathbf{tf}}$ should give a central role to the $\mathbf{tf}$ adding, at every step of the algorithm, target genes to it. Then they defined the concept of *strong dominance* of a measure δ_1 with respect to a second measure δ_2 , denoted by $\delta_1 \ll \delta_2$ (see Definitions 3 and 4 and Proposition 1 in [6]) that can be used to compare and possibly ranking different measures.

In the context of microarray data analysis the correlation-based proximity measures are advised [7] so that [6] focused on the “1 – squared correlation coefficient” as proximity measure and on the most commonly used linkage rules: the *complete* linkage d_c , the *single* linkage d_s and the *average* linkage d_a . Furthermore, [6] proposed the use of two dissimilarity measures, λ and λ' (see Definition 5 in [6]), based on the Wilks’ Λ statistic [5].

The comparison between dissimilarity measures is performed in [6] by assuming that the distribution of the vector $X_{V^{\mathbf{tf}}}$ of the random variables associated with the genes in the transcription module belongs to a Gaussian directed acyclic graphical model that they called *balanced DAG Transcription Module* (B-DAG-TM) model. They showed that under this model the entries of the correlation matrix $R = \text{cor}(X_{V^{\mathbf{tf}}})$ are given by $\text{cor}(X_{\mathbf{tf}}, X_i) = \text{sgn}(\beta_i) \times \rho$ and $\text{cor}(X_i, X_j) = \text{sgn}(\beta_i) \times \text{sgn}(\beta_j) \times \rho^2$ where $\text{sgn}(\beta_i) = +1$ if the transcription factor $\mathbf{tf}$ is an activator of the gene X_i whereas $\text{sgn}(\beta_i) = -1$ if $\mathbf{tf}$ is a repressor of X_i . Furthermore, ρ is the absolute correlation coefficient between $\mathbf{tf}$ and any target gene X_i for $i \in V$.

3 Comparison of dissimilarity measures under the B-DAG-TM model

Under the B-DAG-TM model, [6] compared the performance of λ , λ' with that of d_c , d_s and d_a linkage rules based on the $(1 - \rho^2)$ as proximity measure. They shown that all but d_c are strongly adequate and $\lambda \ll d_s$, $\lambda \ll d_a$ and, when $|V| \geq 3$, $d_s \ll d_a$ so that, in this case, $\lambda \ll d_s \ll d_a$.

We here enrich these results by comparing λ and λ' dissimilarities with the complete d_c^* , the single d_s^* and the average d_a^* linkage rules based on $1 - |\rho|$ as proximity measure. In order to carry out this comparison we need first to compute d_c^* , d_s^* and d_a^* on an arbitrary partition $\mathcal{P} = \{A^{\mathbf{tf}}, B_1, \dots, B_r\} \in \tilde{\mathcal{P}}(U)$ where U is an arbitrary subset of V . Hereafter, $B_i, B_j \in \mathcal{P}$ are two arbitrary elements of this partition with $B_i, B_j \neq A^{\mathbf{tf}}$. Under the linkage rules considered the entries of the dissimilarity matrix are as follows:

$$d_c^*(B_i, B_j) = 1 - \rho^2 \quad \text{and} \quad d_c^*(A^{\text{tf}}, B_i) = 1 - \rho^2 \quad (1)$$

$$d_s^*(B_i, B_j) = 1 - \rho^2 \quad \text{and} \quad d_s^*(A^{\text{tf}}, B_i) = 1 - |\rho| \quad (2)$$

$$d_a^*(B_i, B_j) = 1 - \rho^2 \quad \text{and} \quad d_a^*(A^{\text{tf}}, B_i) = (1 - \rho^2) - \frac{|\rho| - \rho^2}{|A^{\text{tf}}|}. \quad (3)$$

It is straightforward to see that the dissimilarity matrix based on the complete linkage (1) is constant so that d_c^* is not adequate for V^{tf} . For the other measures, instead, we give the following result.

Proposition 1 *Let the probability distribution of $X_{V^{\text{tf}}}$ belong to the B-DAG-TM model. Then d_a^* and d_s^* are strongly adequate for V^{tf} .*

Proof. This is a direct application of Definition 2 in [6] to eqn. (2) and (3).

We can now give the main result of this paper concerning the relationship of dominance between the considered dissimilarity measures.

Theorem 2 *Let the probability distribution of $X_{V^{\text{tf}}}$ belong to the B-DAG-TM model. Then, when $|V| \geq 3$, it holds that*

$$\begin{array}{ccc} \lambda' = \lambda & \ll & d_s \ll d_a \\ & \Downarrow & \Downarrow \\ & d_s^* & \ll d_a^* \end{array}$$

and, furthermore, $\lambda \not\ll d_s^*$ and $d_s^* \not\ll \lambda$.

Proof. For $U \subseteq V$ let $\mathcal{P} = \{A^{\text{tf}}, B_1, \dots, B_r\}$ be a partition of U^{tf} with $r \geq 2$ and let $B_i, B_j \in \mathcal{P}$ be an arbitrary pair of sets of target genes with $B_i, B_j \neq A^{\text{tf}}$. Note that the assumptions of the B-DAG-TM model imply that both $\rho \neq 0$ and $\rho \neq \pm 1$.

The fact that $\lambda = \lambda'$ and $\lambda \ll d_s \ll d_a$ are proved in Proposition 5 and Theorem 7 of [6].

The proof that $d_s^* \ll d_s$ follows immediately by applying Definition 4 in [6]; indeed, it is trivially true that

$$\frac{d_s^*(A^{\text{tf}}, B_{opt}^s)}{d_s^*(B_i, B_j)} = \frac{1 - \rho}{1 - \rho^2} < \frac{d_s(A^{\text{tf}}, B_{opt}^a)}{d_s(B_i, B_j)} = \frac{1 - \rho^2}{1 - \rho^4}.$$

See [6] for the definition of B_{opt} .

Now we show that $d_s^* \ll d_a^*$. We have to prove that

$$\frac{d_s^*(A^{\text{tf}}, B_{opt}^s)}{d_s^*(B_i, B_j)} = \frac{1 - \rho}{1 - \rho^2} \leq \frac{d_a^*(A^{\text{tf}}, B_{opt}^a)}{d_a^*(B_i, B_j)} = \frac{(1 - \rho) \left\{ 1 + \left(\frac{|A^{\text{tf}}| - 1}{|A^{\text{tf}}|} \right) \rho \right\}}{1 - \rho^2}$$

that is, that $1 \leq 1 + \left(\frac{|A^{\text{tf}}|-1}{|A^{\text{tf}}|}\right)\rho$ which is trivially true with strict inequality when A is non empty.

The proof that $d_a^* \ll d_a$ is quickly shown by recalling again Definition 4 in [6] whereby it results that

$$\frac{d_a^*(A^{\text{tf}}, B_{opt}^a)}{d_a^*(B_i, B_j)} = \frac{1 + \left(\frac{|A^{\text{tf}}|-1}{|A^{\text{tf}}|}\right)\rho}{1 + \rho} < \frac{d_a(A^{\text{tf}}, B_{opt}^a)}{d_a(B_i, B_j)} = \frac{1 + \left(\frac{|A^{\text{tf}}|-1}{|A^{\text{tf}}|}\right)\rho^2}{1 + \rho^2}$$

that is true because $\left(\frac{|A^{\text{tf}}|-1}{|A^{\text{tf}}|}\right) < 1$ when A is nonempty.

By exploiting Proposition 1 in [6] and the latter two results, we may also conclude that $d_s^* \ll d_a$.

In the end, in order to prove that $\lambda \not\ll d_s^*$ and that $d_s^* \not\ll \lambda$ we provide a counterfactual example for each one of the two cases. Set $|B_j| = 5$ and $\rho = 0.5$; for $|B_i| = 6$ and $|B_{i'}| = 3$ we have that

$$\frac{\lambda(A^{\text{tf}}, B_{opt}^\lambda)}{\lambda(B_i, B_j)} > \frac{d_s^*(A^{\text{tf}}, B_{opt}^s)}{d_s^*(B_i, B_j)} \quad \text{and} \quad \frac{\lambda(A^{\text{tf}}, B_{opt}^\lambda)}{\lambda(B_{i'}, B_j)} < \frac{d_s^*(A^{\text{tf}}, B_{opt}^s)}{d_s^*(B_{i'}, B_j)}$$

respectively so that the proof is complete. $\square$

The set of considered dissimilarity measures with the strict partial order ' $\ll$ ' is a poset. The strict ordering comes directly from the definition of dominance since it is a binary relation which is irreflexive (a dissimilarity measure cannot strictly dominate itself) and transitive (see Proposition 1 in [6]) and, therefore, asymmetric. Theorem 2 clarifies the ordering structure of the considered dissimilarity measures. We can thus depict the results of this section by means of the following relational structure

$$\{\lambda', \lambda, d_s^*\} \prec \{d_s, d_a^*\} \prec \{d_a\} \prec \{d_c, d_c^*\}.$$

Dissimilarity measures belonging to the same set are incomparable between each other. The symbol $\prec$ is a binary relationship: for every measure δ in the set on the right hand side of $\prec$ there is at least one measure in the set on the left hand side of $\prec$ that should be preferred to δ on either the concepts of strong adequateness or strong dominance. This relational structure implicitly gives the *strict partial ordering* and the set $\{\lambda', \lambda, d_s^*\}$ contains the minimal elements of the ordering that are empirically compared in the next section.

4 Experimental results

Here we use simulated and real data to compare the behavior of λ , λ' and d_s^* .

Simulations are conducted by randomly generating data from the B-DAG-TM model for different values of the sample size n , the dimension of the transcription module $|V^{\text{tf}}|$ and the signal-to-noise ratio η and for two different settings: one for the case of $|V^{\text{tf}}| < n$ and one for the case of $|V^{\text{tf}}| > n$. The dissimilarity measures are computed on the sample correlation matrix $\hat{R}$. We remark that d_s^* can be always estimated even when the number of genes in the clusters exceeds n , case in which, instead, the computation of λ and λ' becomes impracticable. For this reason, in the case of $|V^{\text{tf}}| > n$, as soon as the dimension of the biggest cluster may be formed is greater than $n/2$, we compute λ and λ' by exploiting their relationship with the canonical correlation analysis (see eqn. (5) in [6]) and the method for their sparse estimation developed by [8] (for technical details see [6]). We remark that λ and λ' are identical under the B-DAG-TM model but the computation of λ' requires the estimation of an additional parameter and, for this reason, λ and λ' differ in the applications. Figure 1 shows the results for $n = 25, 50$ and $|V^{\text{tf}}| = 20$

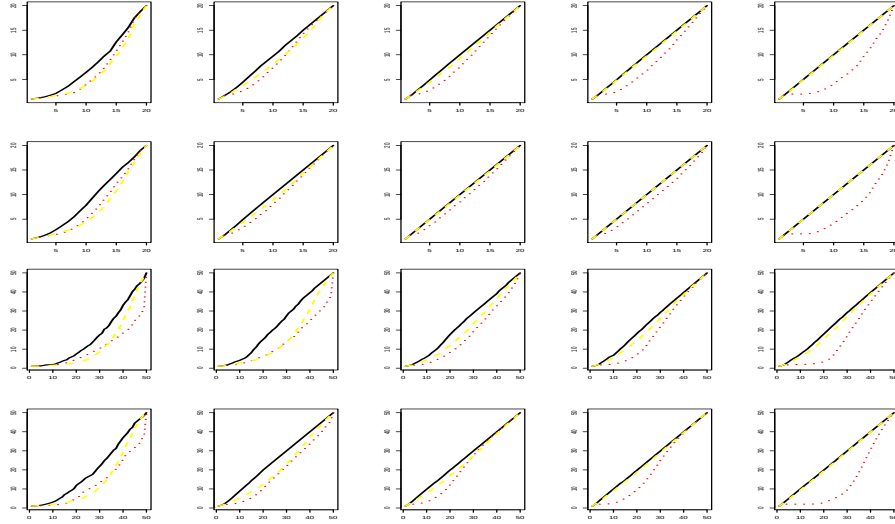


Fig. 1 Simulation results for $|V^{\text{tf}}| = 20$, $n = 25, 50$ (first two lines), $|V^{\text{tf}}| = 50$, $n = 10, 25$ (third and fourth line) and $\rho = (0.1, 0.3, 0.5, 0.7, 0.9)$ (by cols). x -axes: number of iteration of the clustering algorithm, y -axes: average of $|A^{\text{tf}}|$. Black solid line for λ , red dotted line for λ' and yellow dashed line for d_s .

(first two lines) and for $n = 10, 25$ and $|V^{\text{tf}}| = 50$ (second and third line). The correlation ρ we consider takes values in $(0.1, 0.3, 0.5, 0.7, 0.9)$. For each experimental setting 100 replications are performed and the agglomerative hierarchical clustering algorithm is applied to the variables (genes) by using the three different measures selected. The ability of the dissimilarity measures to recognize the transcription module A^{tf} at every step of the procedure is

shown by computing at every step of the algorithm the dimension of A^{tf} that is expected to increase uniformly for strongly adequate measures. The curves in the plots show the average value of $|A^{\text{tf}}|$ across the 100 replications. In both the simulation settings, the performance of the clustering procedures appears to be, on the one hand, quite independent from the dimension of the transcription module $|V^{\text{tf}}|$ and the sample size n . On the other hand, an important role is played by ρ . Indeed, as soon as ρ increases, the performance of λ and d_s^* improves revealing themselves able to identify **tf** in the early steps of the procedure and to give a central role to **tf** merging it to target genes step by step. λ' , instead, appears to work better for non-extreme values of ρ . Notice that in Figure 1 the black solid line for λ and the yellow dashed line for d_s^* are overlapping for high values of ρ ; however, λ outperforms d_s^* for low and medium value of ρ in all the considered scenarios. Moreover, it is interesting to highlight that when the values of λ and λ' are estimated by sparse canonical correlations their performance does not seem to be affected and λ is still uniformly better than the other measures.

Now we apply the hierarchical clustering procedures to real microarray data coming from the work of [4] in which they study the transcriptional response of inhibition of the *Escherichia coli* gyrase. This data set is made up of 16 arrays and its NCBI Gene Expression Omnibus [1] accession number is GSE4408. From the Regulon database [3] we selected 4 transcription factors and the sets of target genes at present known to be regulated by them: CytR with 11 target genes for the case of $|V^{\text{tf}}| < n$ and OxyR with 16 target genes, MarA with 25 target genes and LexA with 38 target genes for the case of $|V^{\text{tf}}| > n$. We apply the hierarchical clustering algorithm as on the simulated data and Figure 2 shows the behavior of the measures compared. The behavior of the dissimilarity measures on the real data is consistent with

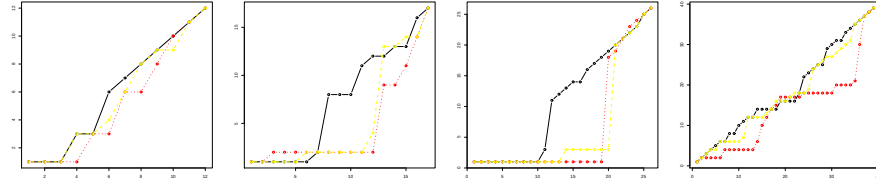


Fig. 2 Applications to real data. From left to right: $|V^{\text{tf}}| < n$: CytR, 11 target genes; $|V^{\text{tf}}| > n$: OxyR, 16 target genes; MarA, 25 target genes; LexA, 38 target genes. x -axes: number of iteration of the clustering algorithm, y -axes: $|A^{\text{tf}}|$. Black solid line for λ , red dotted line for λ' and yellow dashed line for d_s^* .

the results of the simulations, under the assumption that data come from the B-DAG-TM model. Indeed, λ shows a better performance than the two other dissimilarity measures both when $|V^{\text{tf}}| < n$ and when $|V^{\text{tf}}| > n$.

5 Discussion

Starting from the work of [6], we have approached the problem of choosing a dissimilarity measure useful for the identification of transcription modules in a hierarchical clustering context. We have compared different dissimilarity measures on the basis of the criteria introduced by [6] enriching the analysis they have performed and providing a strict partial ordering of all the measures considered. We have considered $1 - |\rho|$ proximity between genes that is perhaps the most commonly used proximity measure for gene clustering.

Theoretical results as well as experimental ones, suggest that λ should be preferred to the other measures irrespectively from the proximity measure and the dimension of the transcription module. Indeed, λ outperforms d_s^* both when d_s^* is advantaged in the first steps of the hierarchical procedure being based on a smaller proximity ($1 - |\rho| < 1 - \rho^2$) and when $|V^{\text{tf}}| > n$ although λ is computed by employing its sparse estimation.

Acknowledgements Partial funding was provided by MIUR (the Italian Ministry for Research), [PRIN-2007AYHZWC] “Statistical methods for learning in clinical research”. Financial support to the first author has also been provided by the MIUR-FISR grant number 2982/Ric (Mitica).

References

1. T. Barrett, D.B. Troup, S.E. Wilhite, P. Ledoux, D. Rudnev, C. Evangelista, I.F. Kim, A. Soboleva, M. Tomashevsky, and R. Edgar. NCBI GEO: mining tens of millions of expression profiles—database and tools update. *Nucl. Acids Res.*, 35(Database issue):D760, 2007.
2. B.S. Everitt, S. Landau, and M. Leese. *Cluster analysis*. Hodder Arnold, London., 4th edition, 2001.
3. S. Gama-Castro, V. Jimenez-Jacinto, M. Peralta-Gil, A. Santos-Zavaleta, M.I. Penaloza-Spinola, B. Contreras-Moreira, J. Segura-Salazar, L. Muniz-Rascado, I. Martinez-Flores, H. Salgado, et al. RegulonDB (version 6.0): gene regulation model of Escherichia coli K-12 beyond transcription, active (experimental) annotated promoters and Textpresso navigation. *Nucl. Acids Res.*, 2007.
4. K.S. Jeong, Y. Xie, H. Hiasa, and A.B. Khodursky. Analysis of pleiotropic transcriptional profiles: A case study of DNA gyrase inhibition. *PLoS Genetics*, 2(9):1464–1476, 2006.
5. A.C. Rencher. *Methods of multivariate analysis*. Wiley, New York, 1995.
6. A. Roverato and F.M.L. Di Lascio. Wilks’ λ dissimilarity measures for gene clustering: an approach based on the identification of transcription modules. *Submitted*, <http://www2.stat.unibo.it/roverato/roverato-dilascio-wilks-lambda.pdf>, 2010.
7. E. Wit and J.D. McClure. *Statistics for microarrays: design, analysis, and inference*. Wiley, 2004.
8. D.M. Witten, R. Tibshirani, and T. Hastie. A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Bio-statistics*, 10(3):515–534, 2009.