Epidemiology Biostatistics and Public Health - 2013, Volume 10, Number 2

Wilks' dissimilarity for gene clustering: computational issues

F. MARTA L. DI LASCIO[1], ALBERTO ROVERATO[2]

Abstract

Clustering methods are widely used in the analysis of gene expression data for their ability to uncover coordinated expression profiles. One important goal of clustering is to discover co–regulated genes because it has been postulated that co–regulation implies a similar function. In the context of agglomerative hierarchical clustering, we introduced a dissimilarity measure based on the Wilks' Λ statistic that we called the *Wilks' dissimilarity* and showed its usefulness in the identification of transcription modules. In this paper, we discuss the ability of the Wilks' dissimilarity to identify clusters of co–expressed genes by providing an example where the most commonly used dissimilarity measures fail. Furthermore, we carry out a set of simulations aimed to investigate the use of a sparse canonical correlation technique in the estimation of the Wilks' dissimilarity and provide guidelines for its use.

Keywords: Wilks' Λ; Hierarchical clustering algorithm; Gene clustering; Canonical correlations; Sparse estimation

DOI: 10.2427/8761

1 INTRODUCTION

Clustering methods have been widely used in the organization of expression data since the early paper by [1]. They are now considered important tools for the analysis of expression profiles and are applied to both the sampling units and the variables, that in this case are genes [2].

In this paper we focus on gene clustering by means of agglomerative hierarchical procedures [see 3, Sections 7.3]. Clustering genes can be useful for many purposes, but it is mainly motivated by the fact that coordinated expression (*co–expression*) patterns are postulated to imply a similar function. Clustering can thus be useful to deduce functions of unknown genes from known genes with co–expression patterns as well as to identify groups of genes which belong to a same functional module.

The application of hierarchical clustering requires the specification of a measure of dissimilarity between sets of genes, and this is typically obtained by specifying first an appropriate proximity measure between pairs of genes, and then a linkage rule which gives the dissimilarity between two sets of genes as a function of the pairwise proximities of genes in the sets. The notion of co–expression as similarity of genes requires that an appropriate proximity is defined

[1] ***Correspondig Author***, School of Economics and Management, Free University of Bozen-Bolzano, piazza Università 1, 39100, Bolzano, Italy. *e-mail*: marta.dilascio@unibz.it

[2] Department of Statistical Sciences, via Belle Arti 41, 40126 Bologna, Italy.*e-mail*: alberto.roverato@unibo.it

and [3, Section 7.3.1] suggests the use of correlation–based proximities when co–regulation of genes is of concern.

Roverato and Di Lascio introduced in [4] a dissimilarity measure based on the Wilks' Λ statistic that they called the *Wilks' dissimilarity* and showed its usefulness in a context where clustering is performed in order to identify transcription modules. The Wilks' dissimilarity amounts to use the Pearson correlation coefficient as a proximity between genes, but differs from other measures based on the same proximity for the linkage rule. The estimation of the Wilks' dissimilarity is unfeasible when the number of genes in the cluster exceeds the sample size, and in this case they computed it by exploiting the method for sparse canonical correlations introduced by [5].

In this paper, we show the usefulness of the Wilks' dissimilarity by providing an example where the most commonly used dissimilarity measures fail to group clusters of co–regulated genes. Furthermore, we carry out a set of simulations aimed to investigate the use of a sparse canonical correlation technique in the estimation of the Wilks' dissimilarity and provide guidelines for its use.

The paper is organized as follows. In Section 2 we give the notation and the background on hierarchical clustering required for this paper. Section 3 is devoted to the presentation of the theoretical properties of the Wilks' dissimilarity whereas the discussion about its computation and the simulations are presented in Section 4. Finally, Section 5 contains a brief discussion.

2 NOTATION AND BACKGROUND

In this section we review the theory related to hierarchical clustering as required for this paper. The reader is referred to [6] for a general introduction to cluster analysis.

Cluster analysis or *clustering* is the assignment of a set of objects into subsets (called *clusters*) so that objects in the same cluster are similar in some sense. Here, we denote by $V = \{1, \ldots, p\}$ the set of objects, that is, of genes. *Hierarchical clustering* generates a hierarchy of nested clusters and in the *agglomerative* approach, the procedure starts with each gene forming a cluster and, at every next step it moves up in the hierarchy by merging exactly one pair of clusters. The algorithm stops when the last two clusters are merged to form V. The decision on which clusters should be combined is based on a *dissimilarity measure* between sets of genes. In most methods of hierarchical clustering, this is achieved by means of a *proximity measure*, possibly a metric, between pairs of genes [see 6, Section 4.2.2] and a *linkage rule* which specifies the dissimilarity of sets as a function of the pairwise proximities of elements in the sets.

An inappropriate proximity measure for the problem under investigation can lead to misleading conclusions, and [3, Section 7.3] states that the meaning of "proximity" of gene expression profiles is different from that of other kinds of objects and that, in this case, correlation–based proximity measures should be preferred to other proximities such as the Euclidean and the Manhattan distances. Hence, typically the dissimilarity between a pair of genes $i, j \in V$ is taken to be proportional to $1 - |\rho_{ij}|$ or, equivalently, to $1 - \rho_{ij}^2$ where ρ_{ij} denotes the Pearson correlation between two genes i and j.

Among the most commonly used linkage rules between two sets there are: the *complete* linkage, the *single* linkage and the *average* linkage. Under the complete linkage the dissimilarity between clusters is the maximum dissimilarity between the genes in the two clusters, the single linkage uses the minimum dissimilarity between genes in the two clusters whereas the average linkage uses the average of all dissimilarities between genes in the two clusters. A different linkage rule will result in a different output. The single linkage rule tends to produce long chains of objects whereas the complete linkage rule tends to produce compact, spherical clusters. The

average linkage is a compromise between these two extremes [see 6, p. 62].

In order to formally deal with hierarchical clustering algorithms, for a set V we denote by $\widetilde{\mathcal{P}}(V)$ the set of all partitions $\mathcal{P} = \{B_1, \ldots, B_r\}$ of V with $1 \leq r \leq p$. Assume that the set of genes V is submitted to an agglomerative hierarchical clustering algorithm. Then, every step of the algorithm starts from a partition $\mathcal{P} = \{B_1, \ldots, B_r\} \in \widetilde{\mathcal{P}}(V)$ and merges two elements of $\mathcal{P}$ producing a new partition in $\widetilde{\mathcal{P}}(V)$. A dissimilarity measure for V is a function $\delta \equiv \delta(Q, P)$ defined on all the pairs of subsets $Q, P \subseteq V$. Dissimilarity measures are nonnegative, $\delta(Q, P) \geq 0$, symmetric, $\delta(Q, P) = \delta(P, Q)$, and $\delta(Q, P) = 0$ for $P = Q$. Furthermore, whenever we write $\delta(P, Q)$ the sets P and Q are meant to be nonempty, $P, Q \neq \emptyset$, and disjoint, $P \cap Q = \emptyset$. The expression profile for the genes in V is the realization of a random vector X_V, and we only consider dissimilarity measures that are computed as a function of the correlation matrix $\mathbf{R}_{VV} = \{\rho_{ij}\}$ of $\mathbf{X}_V$ so that $\delta(Q, P)$ is a shorthand for $\delta(Q, P|\, \mathbf{R}_{VV})$.

In practical applications, a random sample $\mathbf{X}_V^{(1)}, \ldots, \mathbf{X}_V^{(n)}$ from $\mathbf{X}_V$ is available and we denote by $\mathbb{X} \equiv \mathbb{X}_V$ the corresponding $n \times p$ data matrix. In this case, we write $\widehat{\delta}(P, Q)$ to denote the estimate of $\delta(Q, P)$ and note that $\widehat{\delta}$ may often be obtained by plugging an estimate $\widehat{\mathbf{R}}_{VV}$ of $\mathbf{R}_{VV}$ in $\delta(Q, P)$, that is, $\widehat{\delta}(P, Q) = \delta(P, Q|\, \widehat{\mathbf{R}}_{VV})$ but this is not always possible as, for instance, for the case of the Wilks' dissimilarity when $p > n$, as described in the forthcoming sections.

3 THE WILKS' DISSIMILARITY MEASURE

Roverato and Di Lascio introduced in [4] the following correlation–based dissimilarity measure for gene clustering

Definition 1 *For a random vector $\mathbf{X}_V$ with correlation matrix R_{VV} the* Wilks dissimilarity *is defined as*

$$\lambda(P, Q) = \frac{|\mathbf{R}_{P\cup Q, P\cup Q}|}{|\mathbf{R}_{PP}|\,|\mathbf{R}_{QQ}|} \tag{1}$$

for every $P, Q \subseteq V$ such that $P, Q \neq \emptyset$ and $P \cap Q = \emptyset$.

The Wilks' dissimilarity $\lambda(P, Q)$ is directly obtained from the Wilks' Λ statistic

$$\widehat{\Lambda}(\mathbb{X}_P, \mathbb{X}_Q) = \frac{\left|\widehat{\mathbf{R}}_{P\cup Q, P\cup Q}\right|}{\left|\widehat{\mathbf{R}}_{PP}\right|\left|\widehat{\mathbf{R}}_{QQ}\right|} \tag{2}$$

given, among others, in [7, eqn. 7.28].

Note that, the Wilks' dissimilarity is not given by first providing a dissimilarity measure between genes and then by giving a linkage rule. Nevertheless, for every pair of genes $i, j \in V$ it holds that $\lambda(i, j) = 1 - \rho_{ij}^2$, where the latter is the usual correlation–based dissimilarity between genes. Furthermore, we recall that the use of the Wilks' Λ in gene clustering is also possible with a wider class of dissimilarities between genes. More specifically, Roverato and Di Lascio introduced in [4] a generalized version of the Wilks' dissimilarity that can be specified in a flexible way starting from any dissimilarity between genes $i, j \in V$ that can be written as a function of $|\rho_{ij}|$. Hence, Definition 1 introduces, in an implicit way, a linkage rule, and the novelty of the Wilks' dissimilarity stands on such linkage rule rather than on the dissimilarity between genes on which it is based. We refer to [4] for a more comprehensive description of the properties of the Wilks' dissimilarity.

The rest of this section is devoted to the role played by the Wilks' dissimilarity in gene clustering when the aim of the procedure is the identification of clusters of co–regulated genes. This is done by comparing the behavior of the Wilks' dissimilarity with that of the dissimilarity measures commonly used in this context, that is, the complete, single, and average linkage rules applied to the $1 - \rho_{ij}^2$ dissimilarity between pairs of genes.

Consider the case where the task of the clustering procedure is to group co–expressed genes. In this case it seems natural to choose a dissimilarity measure that gives zero dissimilarity to a pair of clusters when there exists a "perfect" co–expression relationship between them. The following example describes a simple case of perfect co–regulation where the Wilks' dissimilarity is shown to behave as expected, whereas the dissimilarities based on the single, complete and average linkage provide misleading results.

Example 1 *Let $P, Q \subseteq V$ be a pair of clusters such that $\mathbf{X}_Q$ is q-dimensional whereas $Y \equiv \mathbf{X}_P$ is a single random variable, namely, $|P| = 1$. Furthermore, without loss of generality we set $Q = \{1, \ldots, q\}$. We assume that there is a perfect linear co–expression relationship between the genes in Q and the gene in P, formally, $Y = \sum_{i=1}^{q} X_i$. To simplify the computations, we assume that the vector $\mathbf{X}_Q$ has been standardized so that $\mathrm{var}(X_i) = 1$ for every $i = 1, \ldots, q$ and, furthermore, that the correlation between variables in $\mathbf{X}_Q$ is constant, i.e., $\mathrm{cor}(X_i, X_j) = \rho$ for every $i, j \in Q$ with $i \neq j$. It is worth recalling that this statistical model is well-defined for $-\frac{1}{q-1} < \rho < 1$. In this framework, it is not difficult to show that*

$$\mathrm{cor}(Y, X_i) = \sqrt{\frac{1 + (q-1)\rho}{q}} \quad \textit{for every} \quad i = 1, \ldots, q. \tag{3}$$

If we denote by $d_c(\cdot,\cdot)$, $d_s(\cdot,\cdot)$ and $d_a(\cdot,\cdot)$ the dissimilarity measures corresponding to the complete, single and average linkage respectively applied to the $1 - \mathrm{cor}(Y, X_i)^2$ dissimilarity between the genes, then if follows from (3) that

$$d_c(P, Q) = d_s(P, Q) = d_a(P, Q) = \frac{(q-1)(1-\rho)}{q}. \tag{4}$$

Hence, even in the case of what we deem to be the most simple instance of perfect co-expression relationship between genes, the complete, single and average linkage fail to produce a zero dissimilarity. In fact, such linkage rules lead to a dissimilarity equal to zero if and only if $q = 1$, that is when the multivariate relationship between $\mathbf{X}_Q$ and $\mathbf{X}_P$ boils down to a bivariate relationship. In particular, note that, for any fixed value of ρ, equation (4) is an increasing function of q so that, in this case, the larger the number of genes in Q the larger is the dissimilarity assigned to P and Q by the complete, single and average linkage rules which is an unexpected, and clearly misleading, behavior. The Wilks' dissimilarity, on the contrary, behaves as expected because, from the perfect linear relationship between $\mathbf{X}_Q$ and $\mathbf{X}_P$ it follows that $|\mathbf{R}_{P\cup Q, P\cup Q}| = 0$ and therefore $\lambda(P, Q) = 0$, as required.

The above example is somehow surprising. The average linkage rule is perhaps the most common of the linkage methods considered here and it is popularly known as UPGMA (Unweighted Pair Group Method with Arithmetic Mean) but, nevertheless, it fails to recognize a straightforward instance of co–regulation. This can be explained by noticing that the main difference between the Wilks' dissimilarity and the other considered linkage rules stands on the use they make of the information contained in the correlation matrix. More specifically, for two clusters P and Q, the complete, single and average linkage are based on the information provided

by the set of bivariate distributions of $\mathbf{X}_{\{i,j\}}$ for every $i \in P$ and $j \in Q$, whereas the Wilks' dissimilarity fully exploits the information on the linear association between P and Q provided by the multivariate distribution of $\mathbf{X}_{P\cup Q}$. The perfect co-regulation relationship considered in the example above is very simple but, nevertheless, it cannot be detected by any dissimilarity measure that is a function of $\mathbf{R}_{P,Q}$ only.

4 ESTIMATION OF THE WILKS' DISSIMILARITY WHEN p IS LARGER THAN n

In the previous section, we have shown that the Wilks' dissimilarity represents an appealing correlation–based dissimilarity measure. However, its use is difficult in practice when the number of variables, i.e. of genes, exceeds the available sample size n, as in the case of microarray data. In the Gaussian case, $\widehat{\mathbf{R}}_{P\cup Q,P\cup Q}$ has full rank, with probability one, if and only if $n > (|P|+|Q|)$ [8] and therefore, the computation of $\widehat{\lambda}(P,Q) = \lambda(P,Q|\widehat{\mathbf{R}}_{P\cup Q,P\cup Q})$ makes sense whenever $(|P|+|Q|)$ is smaller than the sample size n, but it is not obvious how to proceed for larger clusters.

It can be shown [see 7, eqn. 7.29] that $\lambda(P,Q)$ is the determinant of the correlation matrix of the canonical variables of $\mathbf{X}_P$ and $\mathbf{X}_Q$ [9] and, more specifically, if ϱ_k, for $k = 1,\ldots,h$, are the canonical correlations between $\mathbf{X}_P$ and $\mathbf{X}_Q$ it holds that

$$\lambda(P,Q) = \prod_{k=1}^{h}(1-\varrho_k^2). \tag{5}$$

Hence, an alternative way to compute λ is through the canonical correlations between the two clusters P and Q. The computation of canonical correlations via maximum likelihood [9] involves finding two vectors $\mathbf{u}$ and $\mathbf{v}$ such that the correlation between $\mathbb{X}_P\mathbf{u}$ and $\mathbb{X}_Q\mathbf{v}$ is maximum, that is, consists in solving the following

$$\max_{\mathbf{u},\mathbf{v}} \mathbf{u}^T\mathbb{X}_P^T\mathbb{X}_Q\mathbf{v} \quad \text{subject to} \quad \mathbf{u}^T\mathbb{X}_P^T\mathbb{X}_P\mathbf{u} \le 1 \ \text{ and } \ \mathbf{v}^T\mathbb{X}_Q^T\mathbb{X}_Q\mathbf{v} \le 1. \tag{6}$$

The maximization in (6) has a closed–form solution for $\mathbf{u}$ and $\mathbf{v}$ with probability one only if $n > (|P|+|Q|)$. Otherwise, in order to compute the canonical correlations, additional assumptions, such as sparsity, have to be posed on the correlation structure of variables.

In recent years there has been a substantial amount of work on high–dimensional and computationally tractable methods for sparse covariance matrix estimation and related techniques, such as canonical correlation analysis; see, among others, [10], [11] and [5]. In order to obtain an estimate of λ within the 'small n, large p' framework we exploit the connection between λ and the canonical correlations in (5) and apply the method for sparse canonical correlation analysis developed by [5]. Following this approach it is possible to compute, under the assumption of sparsity, a smaller number of canonical correlations by imposing two additional penalties P_1 and P_2 in (6). These are convex penalty functions which can take on different forms. Here they take on the lasso penalty form so that P_1 and P_2 are the L_1–norm of the vectors $\mathbf{u}$ and $\mathbf{v}$: $P_1(\mathbf{u}) = \sum_{i=1}^n |u_i|$ and $P_2(\mathbf{v}) = \sum_{i=1}^n |v_i|$. In high–dimensional problems, it has been shown that treating the covariance matrix as diagonal can yield good results [12, e.g.] so that substituting the identity matrix $\mathbf{I}$ for both $\mathbb{X}_P^T\mathbb{X}_P$ and $\mathbb{X}_Q^T\mathbb{X}_Q$ in (6) a sparse version of the canonical correlations can be obtained by solving the following problem

$$\max_{\mathbf{u},\mathbf{v}} \mathbf{u}^T\mathbb{X}_P^T\mathbb{X}_Q\mathbf{v} \quad \text{subject to} \quad ||\mathbf{u}||_2^2 \le 1\ ,\ ||\mathbf{v}||_2^2 \le 1\ ,\ P_1(\mathbf{u}) \le c_1\ ,\ P_2(\mathbf{v}) \le c_2 \tag{7}$$

where $||\mathbf{u}||_g$ denotes the L_g–norm of the vector $\mathbf{u}$, i.e. $(\sum_{i=1} u_i^g)^{\frac{1}{g}}$, and c_1 and c_2 are restricted to the ranges $1 \leq c_1 \leq \sqrt{n}$ and $1 \leq c_2 \leq \sqrt{|P|}$, respectively, in order to do active the constraints [see 5].

The use we make of the sparse canonical correlations computed from (7) is different from the original one proposed by [5]. Beyond the matter of choosing the value for the constraints c_1 and c_2, the use of sparse canonical correlations for the estimation of λ is not straightforward. We address here two main questions. Firstly, we are forced to use sparse canonical correlations whenever $(|P| + |Q|) \geq n$, but it is well–known that, as pointed out by [13], the eigenstructure of a covariance matrix tends to be systematically distorted by the sample covariance matrix unless $n \gg (|P| + |Q|)$ [14]. Therefore, it is not clear whether the sparse canonical correlations should be preferred to the traditional maximum likelihood estimation even when $(|P| + |Q|)$ is large but, nevertheless, smaller than n. Secondly, when computing λ by (5) it is clear that the smallest canonical correlations, whose values are close to zero, have small impact on the value of λ. It is therefore of interest investigate the behavior of a modified version of (5) that only involves a smaller number, $h' < h$, of canonical correlations. Our aim is to study these two issues by comparing the performance of the estimate of λ obtained through the standard maximum likelihood procedure, denoted by $\widehat{\lambda}_{\text{st}}$, and the estimate of λ obtained through the sparse procedure, denoted by $\widehat{\lambda}_{\text{sp}}$. This is carried out by means of a simulation study.

4.1 A MONTE CARLO STUDY

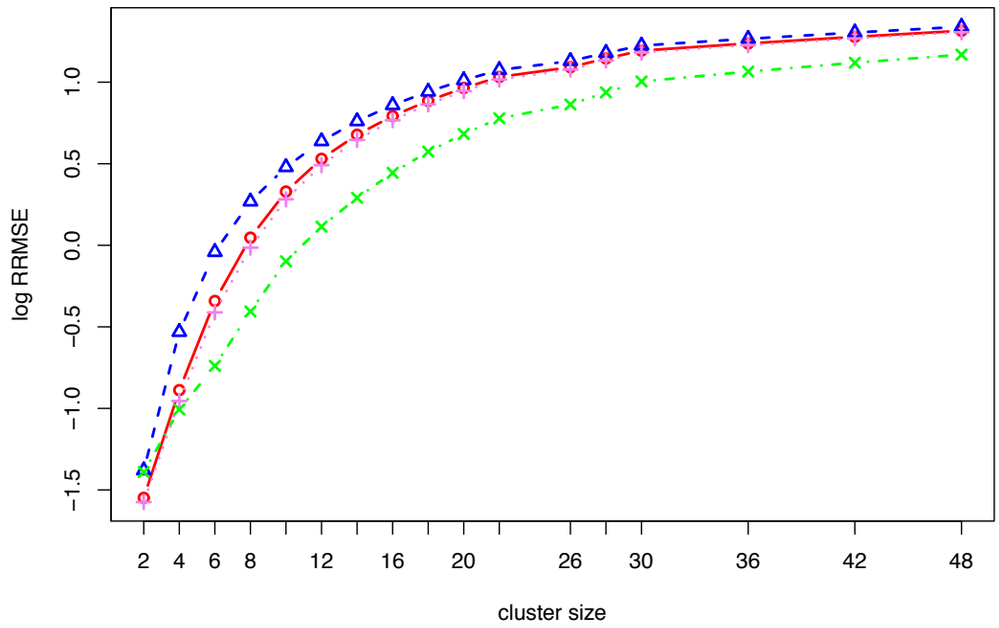
We perform a Monte Carlo study randomly generating data from a multivariate Gaussian distribution. We first compute the maximum likelihood estimate of the canonical correlations as well as their sparse estimate as described in the previous section. We obtain, in this way, two different estimates of the canonical correlations and for each one of them we compute two further different estimates of λ: $\widehat{\lambda}_{\text{st1}}$ and $\widehat{\lambda}_{\text{sp1}}$ are computed by applying (5) to the maximum likelihood and sparse estimates, respectively, while $\widehat{\lambda}_{\text{st2}}$ and $\widehat{\lambda}_{\text{sp2}}$ are obtained by involving in the computation of (5) the $h' = \min(|P|, |Q|, 5)$ largest canonical correlations estimated through the two considered methods.

We simulate two scenarios, one in which the true correlation structure of the clusters is randomly generated and one in which it is constant with $\mathbf{R}_{P\cup Q, P\cup Q} = \{\rho\}$ and $\rho = 0.5$. In both these two scenarios, without loss of generality, we set $|P| = |Q|$ and we allow the dimension of the two clusters P and Q to be compared to vary from 2 to 48 and the sample sizes n in $(50, 250)$. In all the simulations we have that $n > p$ in such a way that the computation of λ through the standard canonical correlations is feasible. For each experimental setting we perform $B = 100$ replications and we compare the four estimates of λ with its true value by computing, for each cluster dimension considered, the *relative root mean squared error* (RRMSE) defined as

$$\text{RRMSE} = \sqrt{\frac{1}{B}\sum_{b=1}^{B}\left(\frac{\widehat{\lambda}_{\star b} - \lambda}{\lambda}\right)^2} \tag{8}$$

where $\widehat{\lambda}_\star$ is one of the four estimates considered. Note that we compute the standard canonical correlations by means of the `CANCOR` function of the R package `STATS` while we obtain the sparse canonical correlations via the `CCA` function of the R package `PMA` [15] with the default values for the parameters of lasso penalty functions. The results are shown in Figures 1 and 2 in which the logarithm of the relative root mean squared error is plotted against the dimension of the two clusters compared. Figure 1 gives the results of simulations for $n = 50$ (Figure on the

Figure 1. Simulation results for $n = 50, 250$ (top and bottom) and $\mathbf{R}_{P\cup Q, P\cup Q} = \{\rho\}$ with $\rho = 0.5$. $x-$axes: dimension of clusters, $y-$axes: log relative root mean squared error of the estimates. Solid (red) line for $\widehat{\lambda}_{\mathrm{st1}}$, dashed (blue) line for $\widehat{\lambda}_{\mathrm{sp1}}$, dotted (violet) line for $\widehat{\lambda}_{\mathrm{st2}}$, dot–dashed (green) line for $\widehat{\lambda}_{\mathrm{sp2}}$. This figure appears in color in the electronic version of this article.



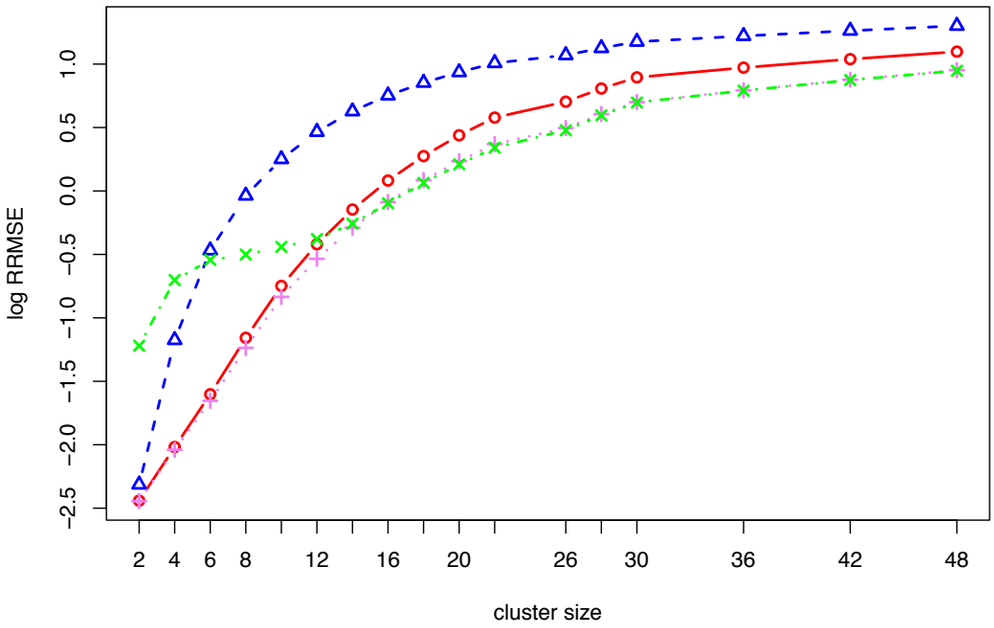
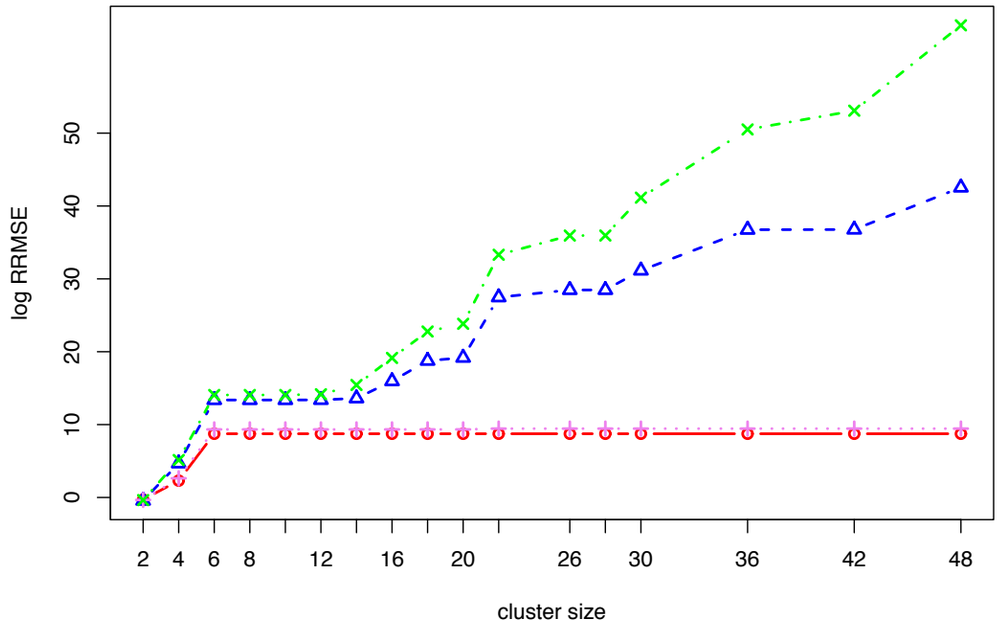
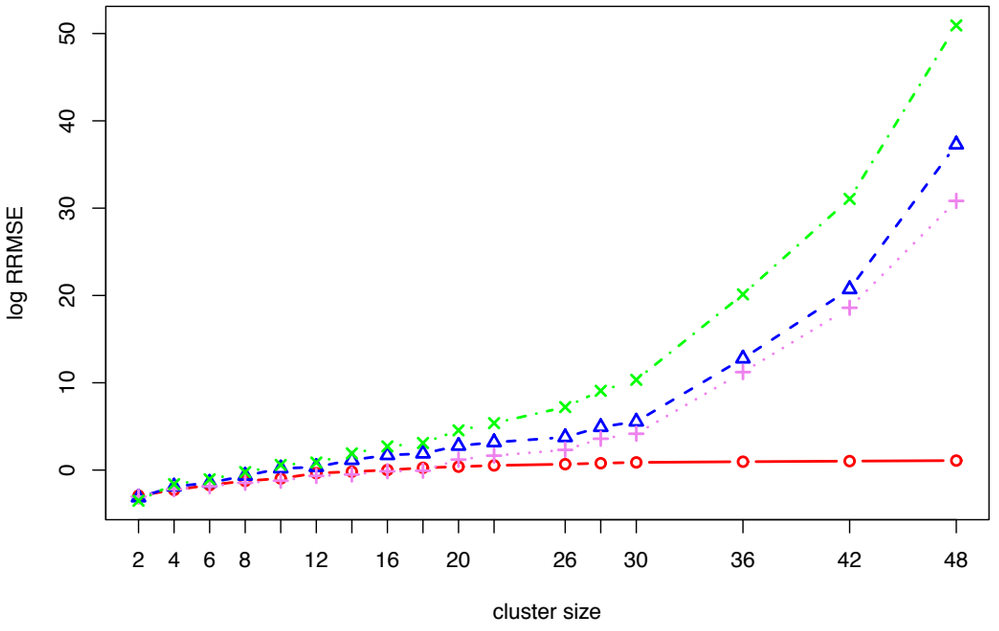


Figure 2. Simulation results for $n = 50, 250$ (top and bottom) and randomly generated correlation structures. $x-$axes: dimension of clusters, $y-$axes: log relative root mean squared error of the estimates. Solid (red) line for $\widehat{\lambda}_{\mathrm{st1}}$, dashed (blue) line for $\widehat{\lambda}_{\mathrm{sp1}}$, dotted (violet) line for $\widehat{\lambda}_{\mathrm{st2}}$, dot–dashed (green) line for $\widehat{\lambda}_{\mathrm{sp2}}$. This figure appears in color in the electronic version of this article.





small sample size, the performance of the estimates for λ based on the standard estimation of the canonical correlations appears to be worse than $\widehat{\lambda}_{\text{sp2}}$ but better than $\widehat{\lambda}_{\text{sp1}}$, irrespectively of the number of canonical correlations involved in its computation. Hence, the number of canonical correlations involved in the computation of λ seems to not affect the performance of the standard estimates of λ whereas plays an important role when its sparse version is of concern. Indeed, $\widehat{\lambda}_{\text{sp}}$ shows a relative root mean squared error uniformly higher than that of $\widehat{\lambda}_{\text{st1}}$ and $\widehat{\lambda}_{\text{st2}}$ only when all the sparse canonical correlations estimated are involved in its computation. When n is large (Figure 1, bottom panel), the standard estimates outperform both $\widehat{\lambda}_{\text{sp1}}$ and $\widehat{\lambda}_{\text{sp2}}$ for small clusters but, as soon as the dimension of clusters increases, the RRMSE of $\widehat{\lambda}_{\text{st2}}$ tends to coincide with that of $\widehat{\lambda}_{\text{sp2}}$. Here, the number of canonical correlations involved in the computation of λ appears to have an effect on both the estimation methods considered. By this set of simulations we may conclude that (i) $\widehat{\lambda}_{\text{sp2}}$ appears to be the best estimate of λ when the sample size is small with respect to the dimension of the clusters, (ii) $\widehat{\lambda}_{\text{st2}}$ is the most appropriate estimate of λ when the sample size is large, (iii) the biggest error arises with $\widehat{\lambda}_{\text{sp1}}$. Finally, note that, as expected, the RRMSE of the four estimates considered increases according to the dimension of the clusters compared, that is, according to the closeness of the cluster dimension to the sample size. Figure 2 gives the results of the second set of simulations for $n = 50$ (Figure on the top panel) and $n = 250$ (Figure on the bottom panel). In both the scenarios the performance of $\widehat{\lambda}_{\text{st}}$ outperforms the other one. When n is large, the RRMSE of $\widehat{\lambda}_{\text{st1}}$ is uniformly lowest whereas that of $\widehat{\lambda}_{\text{st2}}$ increases with the dimension of the clusters compared converging to that of $\widehat{\lambda}_{\text{sp1}}$ and $\widehat{\lambda}_{\text{sp2}}$. In this simulation setting the sample size appears to have poor influence in the comparison between the two estimation methods considered for all but $\widehat{\lambda}_{\text{st2}}$.

From the simulations performed, we may conclude that

(i) when the correlation structure of the clusters is constant, *few* canonical correlations should be involved in the computation of λ and their sparse version is advised as soon as n is close to the dimension of the clusters to be compared;

(ii) when the correlation structure of the clusters is not constant, *all* the canonical correlations computed between the two sets compared should be used to estimate λ and their maximum likelihood estimate is the most appropriate independently from the sample size.

5 DISCUSSION

In this paper we have shown interesting features of λ, a dissimilarity measure based on the Wilks' Λ statistic and recently introduced by [4] in the context of agglomerative hierarchical clustering of genes. In particular, we have shown that the most commonly used dissimilarity measures can fail to identify very basic, linear, co-expression relationships between genes whereas the Wilks' dissimilarity behaves consistently in such situations.

There are, however, some difficulties in the estimation of λ. Unlike the complete, single and average linkage whose computation is always feasible, the computation of the Wilks' dissimilarity requires that the sample version of the correlation matrix of the two clusters compared, $\widehat{\mathbf{R}}_{P\cup Q,P\cup Q}$, has full rank. Consequently, when the number of genes in the clusters exceeds the sample size the computation of λ becomes unpracticable. A possible solution to this problem is based on the indirect estimation of λ by exploiting its connection with canonical correlation and, more specifically, by exploiting existing methods for the estimation of sparse canonical correlations. We have focused on the method developed by [5] and we have carried out a set of simulations so as to better understand the use of this method for the computation of λ and, eventually, provide guidelines for its use.

ACKNOWLEDGMENTS: Work performed under the Italian Ministry for Education, University, and Research protocol 2007AYHZWC "Statistical methods for learning in clinical research."

References

[1] Eisen M, Spellman P, Brown P, Bolstein, D. Cluster analysis and display of genome-wide axpression patterns. Proccedings of the National Academy of Sciences of the USE 1998; 95(25): 148–63.

[2] Chipman H, Hastie T, Tibshirani R. Clustering microarray data. In Statistical analysis of gene expression microarray data, T. Speed, Ed. Chapman and Hall/CRC, 2003; 159–200.

[3] Wit E, McClure J. Statistics for microarrays: design, analysis, and inference. Wiley, 2004.

[4] Roverato A, Di Lascio FML. Wilks'λ dissimilarity measures for gene clustering: an approach based on the identification of transcription modules. Biometrics 2011; 67(4): 1236-1248.

[5] Witten D, Tibshirani R, Hastie T. A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. Biostatistics 2009; 10(3): 515–534.

[6] Everitt B, Landau S, Leese M. Cluster analysis, 4th ed. Hodder Arnold, London, 2001.

[7] Rencher, A. Methods of multivariate analysis. Wiley, New York, 1995.

[8] Dystra R. Establishing the positive definiteness of the sample covariance matrix. The Annals of Mathematical Statistics 1970; 41(6): 2153–2154.

[9] Hotelling H. Relations between two sets of variates. Biometrika 1936; 28(3): 321–377.

[10] Lykou A, Whittaker J. Sparse cca using a lasso with positivity constraints. Computational Statistics and Data Analysis 2009;

[11] Parkhomenko E, Tritchler D, Beyene J. Sparse canonical correlation analysis with application to genomic data integration. Statistical Applications in Genetics and Molecular Biology 2009. 8(1), DOI: 10.2202/1544-6115.1406.

[12] Dudoit S, Fridlyand J, Speed T. Comparison of discrimination methods for the classification of tumors using gene expression data. Journal of the American Statistical Association 2001; 96: 1151–1160.

[13] Dempster A. Elements of continuous multivariate analysis. Addison–Wsley, Reading, MA, 1969.

[14] Yang R, Berger J. Estimation of a covariance matrix using the reference prior. The Annals of Statistics 1994; 22(3): 1195–1211.

[15] Witten D, Tibshirani R, Gross R. PMA: Penalized Multivariate Analysis, 2009. R package version 1.0.5.