BIOINFORMATICS ORIGINAL PAPER

Vol. 25 no. 3 2009, pages 401–405
doi:10.1093/bioinformatics/btn634

Data and text mining

Matrix correlations for high-dimensional data: the modified RV-coefficient

A. K. Smilde[1,*], H. A. L. Kiers[2], S. Bijlsma[3], C. M. Rubingh[3] and M. J. van Erk[3]

[1]Biosystems Data Analysis, Swammerdam Institute for Life Sciences, University of Amsterdam, Nieuwe Achtergracht 166, 1018 WV Amsterdam, [2]Heymans Institute, University of Groningen, Groningen and [3]TNO Quality of Life, Utrechtseweg 48, 3704 HE Zeist, The Netherlands

Received on July 21, 2008; revised on October 30, 2008; accepted on December 06, 2008

Advance Access publication December 10, 2008

Associate Editor: John Quackenbush

ABSTRACT

Motivation: Modern functional genomics generates high-dimensional datasets. It is often convenient to have a single simple number characterizing the relationship between pairs of such high-dimensional datasets in a comprehensive way. Matrix correlations are such numbers and are appealing since they can be interpreted in the same way as Pearson's correlations familiar to biologists. The high-dimensionality of functional genomics data is, however, problematic for existing matrix correlations. The motivation of this article is 2-fold: (i) we introduce the idea of matrix correlations to the bioinformatics community and (ii) we give an improvement of the most promising matrix correlation coefficient (the RV-coefficient) circumventing the problems of high-dimensional data.

Results: The modified RV-coefficient can be used in high-dimensional data analysis studies as an easy measure of common information of two datasets. This is shown by theoretical arguments, simulations and applications to two real-life examples from functional genomics, i.e. a transcriptomics and metabolomics example.

Availability: The Matlab m-files of the methods presented can be downloaded from http://www.bdagroup.nl.

Contact: a.k.smilde@uva.nl

1 INTRODUCTION

Functional genomics research generates high-dimensional data, e.g. transcriptomics, proteomics or metabolomics data. The central characteristic of these types of data is the low sample-to-variable ratio. Transcriptomics (or gene-expression) data typically has thousands of variables and the number of samples is in the order of tens to hundred. Similar characteristics hold for proteomics and metabolomics data. Often multiple datasets are available (i.e. measured) on the same samples of the biological system. This calls for data fusion methods: methods that are able to extract the mutual information from all datasets simultaneously (Alter *et al.*, 2003).

A first useful step in such a data fusion strategy is to probe the similarity between pairs of datasets in a simple and comprehensive way (Smilde *et al.*, 2005b). Matrix correlations can be used for this purpose. These correlations take values between zero and one, defining a scale of similarity between two matrices. This scale can be interpreted in much the same way as the absolute value of the Pearson correlation coefficient known to biologists. Hence, its use in functional genomics data fusion can be straightforward.

Matrix correlations have already a long history in multivariate analysis (Robert and Escoufier, 1976; Yanai, 1974). A comprehensive overview is given in Ramsay *et al.* (1984). For this article, we focus our attention on the RV-coefficient as a typical example of a matrix correlation already in use in metabolomics (Smilde *et al.*, 2005b). While using the RV-coefficient in a transcriptomics study, we ran into problems: the RV-coefficient gave high values in almost all cases. This pointed to trivial results. We explain this trivial result (i.e. the break-down of the RV-coefficient for high-dimensional data) and give a solution to circumvent this unwanted behavior.

2 METHODS

2.1 Matrix correlations

The idea of a matrix correlation is to provide a measure of the similarity of matrices. We start our explanation with matrices $\mathbf{X}(I\times J)$ and $\mathbf{Y}(I\times J)$ sharing the row-mode. The latter means that different types of measurements, e.g. transcriptomics and metabolomics, are performed on the same physical samples (the requirement that both matrices have an equal number of columns will be relaxed later). The mapping $r:\mathbf{R}^{IJ}\times\mathbf{R}^{IJ}\longrightarrow[0,1]$ is called a correlation function if for all non-zero scalars a and b for $\mathbf{X}$ and $\mathbf{Y}$ not both zero holds that

$$\begin{aligned}&C1: r(a\mathbf{X},\mathbf{Y})=r(\mathbf{X},b\mathbf{Y})=r(\mathbf{X},\mathbf{Y}) \qquad (1)\\ &C2: r(\mathbf{X},\mathbf{Y})=r(\mathbf{Y},\mathbf{X})\\ &C3: r(\mathbf{X},\mathbf{Y})=1 \textit{ if } \mathbf{X}=b\mathbf{Y}\\ &C4: r(\mathbf{X},\mathbf{Y})=0 \textit{ iff } \mathbf{X}'\mathbf{Y}=0\end{aligned}$$

where *iff* is the abbreviation of *if and only if* (Ramsay *et al.*, 1984). Matrices can be similar in a variety of ways; this means that rule C3 can be changed, e.g. $\mathbf{X}$ and $\mathbf{Y}$ can have a correlation of one if they only differ by an orthogonal rotation ($\mathbf{X}=\mathbf{YQ}$ with $\mathbf{Q}'\mathbf{Q}=\mathbf{I}$). In that case, the arrangement of the I points (rows) of $\mathbf{X}$ and those of $\mathbf{Y}$ is essentially equal apart from the rotation. An example of a matrix correlation satisfying C1 to C4 is the absolute value of

$$r_{in}(\mathbf{X},\mathbf{Y})=\frac{tr(\mathbf{X}'\mathbf{Y})}{\sqrt{tr(\mathbf{X}'\mathbf{X})tr(\mathbf{Y}'\mathbf{Y})}} \qquad (2)$$

which is based on the inner product of two matrices.

[*]To whom correspondence should be addressed.

A commonly used matrix correlation which allows for a different number of columns in $\mathbf{X}(I\times J_1)$ and $\mathbf{Y}(I\times J_2)$ is the RV-coefficient (Robert and Escoufier, 1976):

$$RV(\mathbf{X},\mathbf{Y})=\frac{tr(\mathbf{XX}'\mathbf{YY}')}{\sqrt{tr[(\mathbf{XX}')^2]tr[(\mathbf{YY}')^2]}} \tag{3}$$

which is an orientation independent measure, i.e. rotations of the two matrices do not affect the RV-coefficient (it satisfies $C1$, $C2$ and $C4$, as well as a relaxed version of $C3$; Appendix). This is usually a desirable property since in many functional genomics applications similarities of the configuration of the samples generated by the two matrices is of interest and not their specific orientation. Stated otherwise, the relationships between the samples are of interest not their absolute positions in space.

The RV-coefficient can also be written using the singular value decomposition (SVD) of both $\mathbf{X}$ and $\mathbf{Y}$:

$$\begin{aligned}\mathbf{X}&=\mathbf{U}_1\mathbf{D}_1\mathbf{V}_1'=\mathbf{T}_1\mathbf{V}_1'\\ \mathbf{Y}&=\mathbf{U}_2\mathbf{D}_2\mathbf{V}_2'=\mathbf{T}_2\mathbf{V}_2'\end{aligned} \tag{4}$$

where $\mathbf{U}_1$ and $\mathbf{U}_2$ are $I\times I$ orthogonal matrices; $\mathbf{D}_1$ and $\mathbf{D}_2$ are $I\times I$ diagonal matrices with the singular values of $\mathbf{X}$ and $\mathbf{Y}$, respectively, on their diagonals; $\mathbf{V}_1(J_1\times I)$ and $\mathbf{V}_2(J_2\times I)$ are column-orthogonal matrices ($\mathbf{V}_1'\mathbf{V}_1=\mathbf{V}_2'\mathbf{V}_2=\mathbf{I}$). Then it holds that (Ramsay *et al.*, 1984)

$$RV(\mathbf{X},\mathbf{Y})=r(\mathbf{U}_1\mathbf{D}_1^2\mathbf{U}_1',\mathbf{U}_2\mathbf{D}_2^2\mathbf{U}_2'), \tag{5}$$

which can easily be verified by substitution and links the two matrix correlation coefficients r and RV. Equation (5) shows that directions in $\mathbf{X}$ and $\mathbf{Y}$ with more importance (i.e. with high singular values) are given more importance in calculating the RV-coefficients. This property is valuable in high-dimensional data because the interest is usually in communality between important dimensions of the matrices.

Alternative expressions for the RV are

$$RV(\mathbf{X},\mathbf{Y})=\frac{ssq(\mathbf{Y}'\mathbf{X})}{\sqrt{ssq(\mathbf{XX}')\times ssq(\mathbf{YY}')}} \tag{6}$$

or

$$RV(\mathbf{X},\mathbf{Y})=\frac{Vec(\mathbf{XX}')'Vec(\mathbf{YY}')}{\sqrt{Vec(\mathbf{XX}')'Vec(\mathbf{XX}')\times Vec(\mathbf{YY}')'Vec(\mathbf{YY}')}} \tag{7}$$

where ssq means sum-of-squares (the sum of squares of all elements of the corresponding matrix) and $Vec(\mathbf{X})$ is the symbol for the vectorized version of $\mathbf{X}$ (see Appendix). The similarity with the Pearson correlation between two vectors $\mathbf{x}$ and $\mathbf{y}$ becomes clear when writing the latter as

$$r_P(\mathbf{x},\mathbf{y})=\frac{\sum_{i=1}^{i=I}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\left[\sum_{i=1}^{i=I}(x_i-\bar{x})^2\right]\left[\sum_{i=1}^{i=I}(y_i-\bar{y})^2\right]}} \tag{8}$$

where x_i, y_i and $\bar{x}$, $\bar{y}$ are the typical elements and means, respectively, of the vectors $\mathbf{x}$, $\mathbf{y}$. Rewriting (8) gives

$$r_P(\mathbf{x},\mathbf{y})=\frac{\tilde{\mathbf{x}}'\tilde{\mathbf{y}}}{\sqrt{\tilde{\mathbf{x}}'\tilde{\mathbf{x}}\times\tilde{\mathbf{y}}'\tilde{\mathbf{y}}}} \tag{9}$$

where $\tilde{\mathbf{x}}$ is the column-centered version of $\mathbf{x}$ and likewise for $\tilde{\mathbf{y}}$. Since $Vec(\mathbf{XX}')$ and $Vec(\mathbf{YY}')$ in (7) play the same roles as $\tilde{\mathbf{x}}$ and $\tilde{\mathbf{y}}$ it is clear that the RV-coefficient can be interpreted as a correlation coefficient. The interpretation of RV as an association measure becomes even more evident when the uncentered correlation is used, or Tucker's congruence coefficient (Lorenzo-Seva and Ten Berge, 2006)

$$r_T(\mathbf{x},\mathbf{y})=\frac{\mathbf{x}'\mathbf{y}}{\sqrt{\mathbf{x}'\mathbf{x}\times\mathbf{y}'\mathbf{y}}} \tag{10}$$

which shows that the RV-coefficient bears also similarities with Tucker's congruence coefficient.

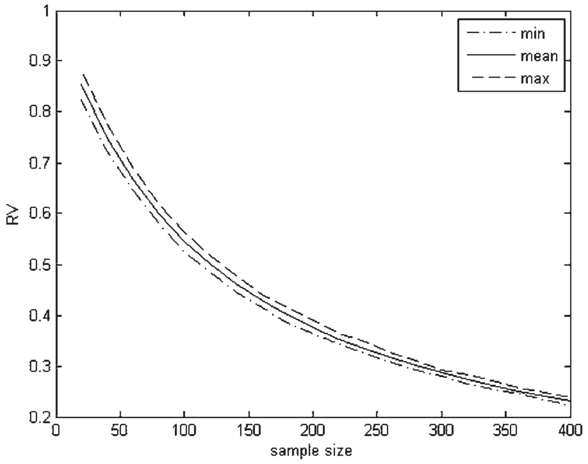

Fig. 1. Simulation of RV-coefficients with different numbers of samples. Plotted is the mean, minimum and maximum for 100 repeats each of the RV as a function of the number of samples.

The RV-coefficient is only independent of a rotation or an overall scaling of the matrices ($RV(\mathbf{X},\mathbf{Y})=RV(\alpha\mathbf{XQ}_1,\beta\mathbf{YQ}_2)$ for non-zero α and β and orthogonal $\mathbf{Q}_1$ and $\mathbf{Q}_2$). All other preprocessing operations are influencing the RV coefficient, e.g. centering has a profound effect similar to the difference between centered and uncentered correlations. Hence, the user has to make a choice regarding the preprocessing and, thus, the metric in which to compare the matrices. Recommendations to this end are available in the literature (Bro and Smilde, 2003; van den Berg *et al.*, 2006).

2.2 Problems with the RV-coefficient

While investigating two gene-expression dataset of sizes 5 × 130 and 5 × 113 from a functional genomics experiment (Kleeman *et al.*, 2007) high RV-coefficients were found (values between 0.5 and 0.99). These could neither directly be understood from the underlying biology nor from independent calculations more extensively investigating the similarities between the two datasets. Hence, a small set of initial simulations was performed where random matrices of the same sizes were generated and RV-coefficients calculated. Despite the fact that the matrices were drawn from random numbers (N(0,1)) the RV-coefficient was always high. Increasing the sample size of the random matrices to 100 samples showed that the RV-coefficient depends on the sample size.

In an extensive set of simulations with random numbers, the samples sizes were systematically increased while the other sizes (130 and 113) remained the same. The result is shown in Figure 1 and shows the problematic behavior. The behavior of the RV-coefficient was also investigated for moderate and strongly unequal sizes of the matrices. Figure 2 shows that the problematic behavior is already visible at much lower numbers of variables (simulations performed similarly as the ones of Fig. 1).

The reason for the unwanted behavior is as follows. According to (7), the RV-coefficient can be written as

$$RV(\mathbf{X},\mathbf{Y})=\mathbf{a}'\mathbf{b} \tag{11}$$

with

$$\begin{aligned}\mathbf{a}&=\frac{Vec(\mathbf{XX}')}{(ssq(\mathbf{XX}')^{1/2}}\\ \mathbf{b}&=\frac{Vec(\mathbf{YY}')}{(ssq(\mathbf{YY}')^{1/2}}.\end{aligned} \tag{12}$$

Now, suppose $\mathbf{X}$ and $\mathbf{Y}$ are fully random matrices, with elements drawn from standard normal distributions. When J_1 is large, $\mathbf{XX}'$ can be expected to have diagonal elements close to J_1 and off-diagonal elements close to 0.

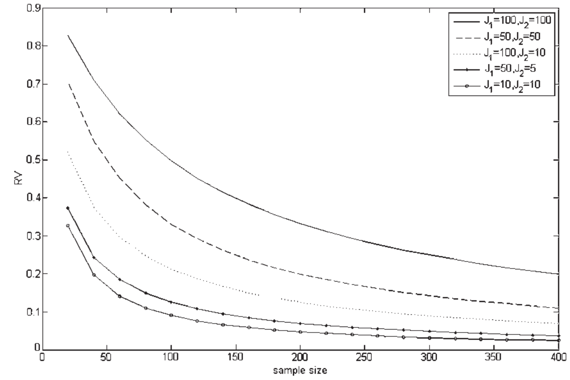

Fig. 2. Simulation of RV-coefficient with different numbers of samples and variables. Plotted are the means of the RV-coefficient from 100 repeats.

Specifically, we can write

$$[\mathbf{XX}']_{ii} = \mathbf{x}_i'\mathbf{x}_i = J_1 + \epsilon_{ii}^x \tag{13}$$

$$[\mathbf{XX}']_{ij} = \mathbf{x}_i'\mathbf{x}_j = \epsilon_{ij}^x, i \neq j,$$

where $\mathbf{x}_i'$ denotes the i-th row of $\mathbf{X}$; the values ϵ_{ii}^x $(i = 1,\dots,I)$ can be considered as random draws from a distribution with zero mean and (the same) standard deviations σ_x. Likewise, the values ϵ_{ij}^x $(i,j = 1,\dots,I, i \neq j)$ can be considered as random draws from a distribution with zero mean and (the same) standard deviations τ_x. The reasons for these distributional properties are as follows. Because $\mathbf{x}_i$ has random elements from a standard normal distribution, $E(\mathbf{x}_i'\mathbf{x}_i)$, i.e. the expected value of a sum of J_1 squares of such values, equals J_1; the variation across realizations of $\mathbf{x}_i'\mathbf{x}_i$ is the same for all i, because the elements of the vectors $\mathbf{x}_i$ are drawn from the same distributions. Likewise, because $\mathbf{x}_i$ and $\mathbf{x}_j$ have independent random elements, $E(\mathbf{x}_i'\mathbf{x}_j) = 0$, and the variation across realizations of $\mathbf{x}_i'\mathbf{x}_j$ is the same for all i,j because the elements of all vectors $\mathbf{x}_i$ and $\mathbf{x}_j$ are drawn from the same distributions. When J_1 is large, σ_x and τ_x can be expected to be small compared with J_1. Indeed, when the elements of $\mathbf{X}$ are drawn from standard normal distributions, it holds that $\sigma_x = \sqrt{2J_1}$ and $\tau_x = \sqrt{J_1}$, as follows from general results on stochastic theory for sums and products (Mood *et al.*, 1974).

Using the above distributional results, we can give approximate descriptions of the normalized vectors **a** and **b**. The numerator of **a** is the vector with I values of $(J_1 + \epsilon_{ii}^x)$ and $I(I-1)$ values ϵ_{ij}^x. Furthermore, we can approximate the denominator in **a** as

$$(ssq(\mathbf{XX}'))^{1/2} \approx (IJ_1^2 + I\sigma_x^2 + I(I-1)\tau_x^2)^{1/2}, \tag{14}$$

which can be explained as follows. The squares of the I diagonal elements of $\mathbf{XX}'$ sum to $\sum_i (J_1 + \epsilon_{ii}^x)^2 = \sum_i J_1^2 + 2\sum_i J_1\epsilon_{ii}^x + \sum_i (\epsilon_{ii}^x)^2$. Now using that $\sum_i J_1^2 = IJ_1^2$, $\sum_i \epsilon_{ii}^x \approx 0$, and that $\sum_i (\epsilon_{ii}^x)^2 \approx I\sigma_x^2$, we get $IJ_1^2 + I\sigma_x^2$ as approximation of the sum of the squared diagonal values. Furthermore, the sum of squared off-diagonal values is $\sum_{i \neq j} (\epsilon_{ij}^x)^2 \approx I(I-1)\tau_x^2$, which completes the explanation. Analogously, the numerator of **b** is the vector with I values of $(J_2 + \epsilon_{ii}^y)$ and $I(I-1)$ values ϵ_{ij}^y, and the denominator in **b** can be approximated as

$$(ssq(\mathbf{YY}'))^{1/2} \approx (IJ_2^2 + I\sigma_y^2 + I(I-1)\tau_y^2)^{1/2}, \tag{15}$$

Now, the RV-coefficient can be approximated as

$$RV(\mathbf{X},\mathbf{Y}) = \mathbf{a}'\mathbf{b} = \tag{16}$$

$$\approx \frac{(IJ_1J_2)}{(IJ_1^2 + I\sigma_x^2 + I(I-1)\tau_x^2)^{1/2}(IJ_2^2 + I\sigma_y^2 + I(I-1)\tau_y^2)^{1/2}}$$

$$= \frac{(J_1J_2)}{(J_1^2 + \sigma_x^2 + (I-1)\tau_x^2)^{1/2}(J_2^2 + \sigma_y^2 + (I-1)\tau_y^2)^{1/2}},$$

taking into account that in the numerator of $\mathbf{a}'\mathbf{b}$ all terms including the random values ϵ_{ii}^x and ϵ_{ii}^y can be expected to roughly cancel. We can further simplify this expression by using the theoretical values for σ_x, τ_x, σ_y and τ_y to obtain

$$RV(\mathbf{X},\mathbf{Y}) \approx \frac{(J_1J_2)}{(J_1^2 + 2J_1 + (I-1)J_1)^{1/2}(J_2^2 + 2J_2 + (I-1)J_2)^{1/2}} \tag{17}$$

$$= \frac{(J_1J_2)}{(J_1^2 + (I+1)J_1)^{1/2}(J_2^2 + (I+1)J_2)^{1/2}}.$$

Table 1. Comparison of derived and simulated RV-values

I	Derived RV-values	Mean simulated RV-values
20	0.852	0.855
40	0.747	0.748
60	0.665	0.666
80	0.599	0.599
100	0.545	0.545
200	0.376	0.376
300	0.287	0.286
400	0.232	0.232

From (16) and (17) it can be seen that the value of RV for random data matrices depends on I: for small I, the RV is close to 1, whereas, as I increases the denominator increases and the value approaches zero. The accuracy of these approximations depends on I but as has been verified in simulations these approximations are typically quite good. Specifically, approximations by the above approach and mean values for RV over 100 random trials yielded the results as given in Table 1 showing that the average values of RV coefficients are very well approximated by the computation of RV according to (17). Apart from rounding errors, there is a close correspondence between the approximations [i.e. (17)] and the means of the simulated RV-values. These results show again that the RV-value is artificially high for small I. Interestingly, the RV-value for the limiting case, i.e. $I = 1$, leads to an RV-value of 1, as is easily verified as follows. In the case of $I = 1$, the matrices $\mathbf{XX}'$ and $\mathbf{YY}'$ reduce to single numbers, and normalizing these trivially leads to setting these numbers equal to 1, so that the RV-value (i.e. their product) also equals 1.

2.3 The modified RV-coefficient

2.3.1 Definition A solution to the problem of the RV-coefficient presents itself by considering the nature of the problem: the numerator of (17) does not tend to zero for random numbers and large J_1 and J_2. This can be traced back to the diagonal of the matrices $\mathbf{XX}'$ and $\mathbf{YY}'$. Indeed, if these diagonal elements are ignored (or, equivalently, set to zero), then the problem disappears since, e.g. $Vec(\mathbf{XX}')$ would be a vector of values randomly varying around zero. After using again (7) this would result in an RV-coefficient of nearly zero, as should be the case for the two random matrices. This is then also exactly our proposal for RV_2, namely instead of using $\mathbf{XX}'$ use $[\mathbf{XX}' - diag(\mathbf{XX}')] = \widetilde{\mathbf{XX}'}$, where $diag(\mathbf{XX}')$ is a matrix containing only the diagonal elements of $\mathbf{XX}'$ on its diagonal, and zero's elsewhere. Using the analogous definition for $\widetilde{\mathbf{YY}'}$ we get

$$RV_2(\mathbf{X},\mathbf{Y}) = \frac{Vec(\widetilde{\mathbf{XX}'})'Vec(\widetilde{\mathbf{YY}'})}{\sqrt{Vec(\widetilde{\mathbf{XX}'})'Vec(\widetilde{\mathbf{XX}'}) \times Vec(\widetilde{\mathbf{YY}'})'Vec(\widetilde{\mathbf{YY}'})}}. \tag{18}$$

Stated otherwise, ignoring the diagonal elements of $\mathbf{XX}'$ and $\mathbf{YY}'$ gives a new vector **a** with I values of 0 and $I(I-1)$ values ϵ_{ij}^x. This solves the problem because the numerator of (17) when using RV_2 then becomes zero.

2.3.2 Properties The RV_2 has different properties than the original RV. The most striking one is that RV_2 can become negative. Suppose for example that

$$\mathbf{XX}' = \begin{pmatrix} 1 & -0.2 \\ -0.2 & 1 \end{pmatrix}, \tag{19}$$

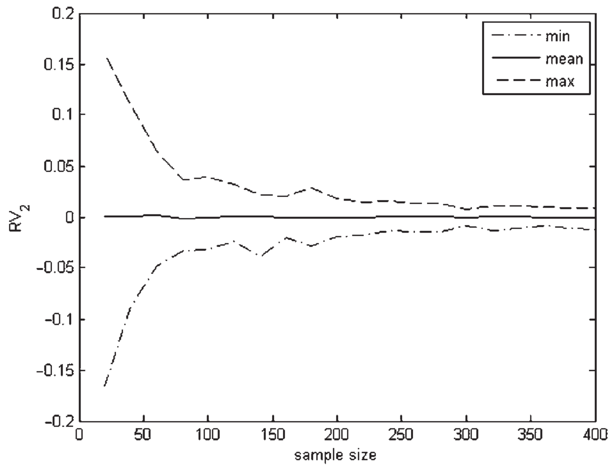

Fig. 3. Simulation of the modified RV-coefficient (RV_2) with different numbers of samples ($J_1=130, J_2=113$). Plotted is the mean, minimum and maximum for 100 repeats each of the RV_2 as a function of the number of samples.

and

$$\mathbf{YY}'=\begin{pmatrix} 1 & 0.2 \\ 0.2 & 1 \end{pmatrix}, \tag{20}$$

then using (18) gives $RV_2(\mathbf{X},\mathbf{Y})=-1$. If instead

$$\mathbf{XX}'=\begin{pmatrix} 1 & 0.2 \\ 0.2 & 1 \end{pmatrix}, \tag{21}$$

then $RV_2(\mathbf{X},\mathbf{Y})=1$. The interpretation of $RV_2=-1$ is that the association between the rows of **X** is proportional to the association between the rows of **Y** but with a negative sign (equivalent to a negative Pearson correlation).

The RV_2 depends only on the cross-products $\mathbf{XX}'$ and $\mathbf{YY}'$, thus the RV_2 is also orientation independent. The RV_2 has values in-between -1 and 1. This follows immediately from the Cauchy–Schwarz inequality applied to the vectors in (18).

3 EXAMPLES

3.1 Simulated examples

Two simulation examples will be used to illustrate the working of the modified RV-coefficient. The first example addresses the (too) large values of the original RV-coefficient. This example follows closely the gene-expression dataset in which the problem was initially encountered. Two datasets **X** of size ($I\times 130$) and **Y** of size ($I\times 113$) were generated 100 times with standard normal distributed numbers. The number of samples was increased from 20 to 400 with steps of 20. For each simulation run, the modified RV-coefficient was calculated. The results are shown in Figure 3.

The second example shows the working of the RV-coefficients for the case that the amount of overlap between **X** and **Y** gradually increases. Two matrices **X** and **Y** were simulated both of size (10×100) with random numbers drawn from a N(0,1) distribution and this was repeated 100 times. Gradually, columns of **Y** are exchanged with those of **X** in steps of 10%, 20%,.... Hence, the amount of overlap between **X** and **Y** increases. Figure 4 shows the original and modified RV-coefficient. Indeed, the modified RV-coefficient gradually increases whereas the original RV-coefficient already has high values from the start.

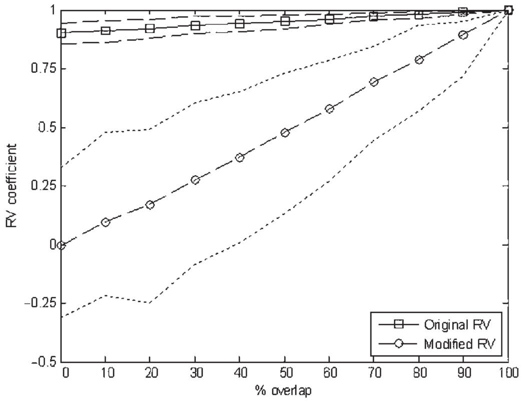

Fig. 4. Simulation of the original and modified RV-coefficient with different amounts of overlap. Plotted are the mean, minimum and maximum for 100 repeats RV's as a function of the amount of overlap.

Table 2. The original and modified RV-values of the gene-expression data (for abbreviations: see text)

Case	Original RV	Modified RV
Ct, VA	0.84	0.57
Low, VA	0.94	0.87
High, VA	0.91	0.85
Ct, IC	0.52	0.27
Low, IC	0.55	0.18
High, IC	0.83	0.78

Summarizing, Figures 3 and 4 show that the modified-RV coefficient has the desired behavior: (i) on average it equals zero for not related matrices, (ii) for larger sample sizes its variability decreases and (iii) it increases with an increasing amount of overlap.

3.2 Gene-expression example

The gene-expression example is taken from the paper of Kleeman *et al.* (2007). In short, gene-expression profiles were measured in livers of female ApoE*3L transgenic mice (E3L mice) from three diet groups ($n=5$ mice per group): control diet (Ct, no cholesterol), low cholesterol diet (Low) and high cholesterol diet (High). Diets were consumed for 10 weeks. RNA from livers was analyzed using Affymetrix whole-genome mouse array MOE430-2.0. Subsets of genes used for matrix correlation were selected based on functional annotation of genes in biological processes cholesterol metabolism (C, J = 71) and inflammation (I, J = 66), vascular development (V, J = 69) and amino-acid metabolism (A, J = 91). Interest focussed on comparison within the different treatment groups, therefore, the RV-coefficients were calculated for the VA blocks and the IC blocks of gene-expressions within each treatment group. All gene-expression values were expressed as deviations from the average Ct group ones without further preprocessing.

Table 2 shows that the modified RV-coefficient is always lower that the original RV-coefficient as it should be. The modified RV-coefficients are more reasonable from a biological perspective. All four selected biological processes were enriched in the selection of genes differentially expressed in response to cholesterol feeding. From these, the more pronounced response was found on cholesterol metabolism (both in response to low and high cholesterol) and inflammation (high cholesterol) (Kleeman *et al.*, 2007). The dose-dependent gene-expression responses are likely to result in increased correlation between the matrices. Contrary, it is not reasonable to have high correlations between the groups of genes for the control animals (Ct). Hence, the values of the modified RV-coefficient for the control groups are more reasonable than the original RV-coefficients. Also from a statistical point of view, the modified RV-coefficients are better than the original ones. This will be explained for the numbers in the first row of Table 2. Consensus-principal component analysis (CPCA) is an alternative method to probe similarities between matrices (Smilde *et al.*, 2003). Using PCA on the V and A matrices individually gives explained variances of 70.0% and 66.9% for two principal components, respectively. When using CPCA, the two CPCA components explain 57.4% and 64.9% in each block, respectively. The drop in explained variances per matrix (especially for the V block) means that there is some overlap between the matrices but also differences. This agrees nicely with the much lower value of 0.57 instead of 0.84. Note that CPCA is used here only to judge the performance of the modified RV-coefficient. This method does not give an alternative *measure* to the RV-coefficient but shows qualitatively the same behavior as the modified RV-coefficient supporting the credibility of the latter.

3.3 Metabolomics example

The modified RV-coefficient was applied to a metabolomics dataset (Smilde *et al.*, 2005a). Metabolites were measured in *Escherichia coli* as a model system. The metabolites were measured using two analytical chemical methods, namely gas-chromatography-mass spectrometry (GC-MS) and liquid-chromatography-mass spectrometry (LC-MS). This generated two datasets with dimensions 28×12553 (GC-MS) and 28×2532 (LC-MS) which clearly fit into the framework of our modified RV-coefficient. The original RV-coefficient was 0.79 and the modified RV-coefficient was 0.71. Hence, the difference was not large in this case. The CPCA analysis performed in the original publication (Smilde *et al.*, 2005a) showed that both matrices had overlap, but also a substantial non-overlapping part. Although both types of RV-coefficients did not differ much in this case, the example is shown to illustrate that the modified RV-coefficient gives also a reasonable value in this case.

In Smilde *et al.* (2005a), a truncation was used prior to calculating the RV-value, i.e. the RV-value was calculated using the first principal components of both matrices. Simulations (results not shown) have pointed out that this approach suffers from the same problems as the RV itself and this approach is therefore not recommended.

4 CONCLUSION

It is often convenient to obtain insight into the relationships between blocks of functional genomics data e.g. as a first step in a data fusion strategy. The modified RV-coefficient is a matrix correlation giving such an insight with a single number between −1 and 1. This number can easily be calculated and interpreted in the same way as an ordinary correlation coefficient. The modified RV-coefficient is theoretically motivated and tested with simulations and real data. The results show that this correlation coefficient is reliable and can be used in bioinformatics practice. This coefficient can also easily be combined with permutation testing for assessing significance (Kazi-Aoual *et al.*, 1995) or with bootstrapping to obtain confidence intervals but that is beyond the scope of this article.

Conflict of Interest: none declared.

REFERENCES

Alter,O. *et al.* (2003) Generalized singular value decomposition for comparative analysis of genome-scale expression data sets of two different organisms. *Proc. Natl Acad. Sci. USA*, **100**, 3351–3356.

Bro,R. and Smilde,A. (2003) Centering and scaling in component analysis. *J. Chemometr.*, **17**, 16–33.

Kazi-Aoual,F. *et al.* (1995) Refined approximations to permutation tests for multivariate inference. *Comput. Statist. Data Anal.*, **20**, 643–656.

Kleeman,R. *et al.* (2007) Increased dietary cholesterol-induced atherosclerosis is associated with liver inflammation: Identification of novel regulatory pathways and transcriptional regulators involved in switch from metabolic adaptation to inflammatory state. *Gen. Biol.*, **8**, R200.

Lorenzo-Seva,U. and Ten Berge,J.M.F. (2006) Tucker's congruence coefficient as a meaningful index of factor similarity. *Methodology*, **2**, 57–64.

Mood,A.M. *et al.* (1974) *Introduction to the Theory of Statistics*. McGraw-Hill Kogakusha Ltd, Tokyo.

Ramsay,J.O. *et al.* (1984) Matrix correlation. *Psychometrika*, **49**, 403–423.

Robert,P. and Escoufier,Y. (1976) A unifying tool for linear multivariate statistical methods: the rv-coefficient. *Appl. Stat.*, **25**, 257–265.

Smilde,A.K. *et al.* (2003) A framework for sequential multiblock component methods. *J. Chemometr.*, **17**, 323–337.

Smilde,A.K. *et al.* (2005a) Anova-simultaneous component analysis (asca): a new tool for analyzing designed metabolomics data. *Bioinformatics*, **21**, 3043–3048.

Smilde,A.K. *et al.* (2005b) Fusion of mass-spectrometry based metabolomics data. *Anal. Chem.*, **77**, 6729–6736.

van den Berg,R. *et al.* (2006) Centering, scaling, and transformations: improving the biological information content of metabolomics data. *BMC Genomics*, **7**, 142.

Yanai,H. (1974) Unification of various techniques of multivariate analysis by means of generalized coefficient of determination (gcd). *J. Behaviormetr.*, **1**, 45–54.

A APPENDIX

A.1 Notation

$\mathbf{x}$ (vector)	bold lowercase
$\mathbf{X}$ (matrix)	bold uppercase
$i=1,\ldots,I$	object index
$j=1,\ldots,J$	variable index
$r=1,\ldots,R$	principal component index

A.2 RV and orientations

Changing the orientation of the sample configuration of $\mathbf{X}$ and $\mathbf{Y}$ can be formalized by using arbitrary orthogonal matrices $\mathbf{Q}_1$ and $\mathbf{Q}_2$ to rotate $\mathbf{X}$ and $\mathbf{Y}$, respectively. Upon defining $\widetilde{\mathbf{X}}=\mathbf{X}\mathbf{Q}_1$ and $\widetilde{\mathbf{Y}}=\mathbf{Y}\mathbf{Q}_2$ and observing that the RV-coefficient can be written to depend only on products $\mathbf{X}\mathbf{X}'$, (see (7)) it holds that $\mathbf{X}\mathbf{Q}_1(\mathbf{X}\mathbf{Q}_1)'=\mathbf{X}\mathbf{Q}_1\mathbf{Q}_1'\mathbf{X}'=\mathbf{X}\mathbf{X}'$ due to the orthogonality property of $\mathbf{Q}_1$ (and similarly for $\mathbf{Y}$). Hence, $RV(\widetilde{\mathbf{X}},\widetilde{\mathbf{Y}})=RV(\mathbf{X},\mathbf{Y})$.

The $I\times I$ matrices $\mathbf{X}\mathbf{X}'$ are called configuration matrices and describe the configuration of the I points (rows of $\mathbf{X}$) in their respective row-spaces. The RV-coefficient only measures differences in configurations ($\mathbf{X}\mathbf{X}'$) and not orientations ($\mathbf{Q}_j$).

A.3 RV in Vec notation

The equality that $tr(\mathbf{A}'\mathbf{B})=Vec(\mathbf{A})'Vec(\mathbf{B})$ can easily been proven by writing both $tr(\mathbf{A}'\mathbf{B})$ and $Vec(\mathbf{A})'Vec(\mathbf{B})$ in terms of the elements of the respective matrices. Using this equality in (3) gives (7).