Separating common from distinctive variation

van der Kloet, F.M.; Sebastián-León, P.; Conesa, A.; Smilde, A.K.; Westerhuis, J.A.

DOI
10.1186/s12859-016-1037-2

Publication date
2016

Document Version
Final published version

Published in
BMC Bioinformatics

van der Kloet et al. BMC Bioinformatics 2016, 17(Suppl 5):195
DOI 10.1186/s12859-016-1037-2

BMC Bioinformatics

RESEARCH Open Access

Separating common from distinctive variation

Frans M. van der Kloet[1], Patricia Sebastián-León[2], Ana Conesa[2], Age K. Smilde[1] and Johan A. Westerhuis[1*]

From Statistical Methods for Omics Data Integration and Analysis 2014
Heraklion, Crete, Greece. 10-12 November 2014

Abstract

Background: Joint and individual variation explained (JIVE), distinct and common simultaneous component analysis (DISCO) and O2-PLS, a two-block (X-Y) latent variable regression method with an integral OSC filter can all be used for the integrated analysis of multiple data sets and decompose them in three terms: a low(er)-rank approximation capturing common variation across data sets, low(er)-rank approximations for structured variation distinctive for each data set, and residual noise. In this paper these three methods are compared with respect to their mathematical properties and their respective ways of defining common and distinctive variation.

Results: The methods are all applied on simulated data and mRNA and miRNA data-sets from GlioBlastoma Multiform (GBM) brain tumors to examine their overlap and differences. When the common variation is abundant, all methods are able to find the correct solution. With real data however, complexities in the data are treated differently by the three methods.

Conclusions: All three methods have their own approach to estimate common and distinctive variation with their specific strength and weaknesses. Due to their orthogonality properties and their used algorithms their view on the data is slightly different. By assuming orthogonality between common and distinctive, true natural or biological phenomena that may not be orthogonal at all might be misinterpreted.

Keywords: Integrated analysis, Multiple data-sets, JIVE, DISCO, O2-PLS

Background

To understand and ultimately control any kind of process, albeit biological, chemical or sociological, it is necessary to collect data that functions as a proxy for these processes. Subsequent statistical data analysis on these data should reveal the relevant information to that process. For hypothesis testing such an approach of theory and measuring can be relatively straightforward especially if the analytical instruments are designed specifically for that purpose. In lack of such hypotheses and using generic but readily available analytical instruments, obvious data structures are rarely observed and extensive data analysis and interpretation are necessary (e.g. untargeted analysis [1], data-mining [2]). To make the data-analysis even more complex, the number of observations (I) is usually much smaller than the number of variables (J) (e.g. transcriptomics data) which prevents the use of classical regression models. Data-analysis and interpretation of the huge number of variables is possible when the number of variables can be summarized in fewer factors or latent variables [3]. For this purpose methods such as factor analysis (FA) [4] or principal component analysis (PCA) [4] were developed.

In functional genomics research it becomes more and more common that multiple platforms are used to explore the variation in samples for a given study. This leads to multiple sets of data with the same objects but different features. Data integration and/or data fusion methods can then be applied to improve the understanding of the differences between the samples. A new

* Correspondence: j.a.westerhuis@uva.nl
[1]Biosystems Data Analysis, Swammerdam Institute for Life Sciences, University of Amsterdam, Science Park 904, 1098 XH, Amsterdam, The Netherlands
Full list of author information is available at the end of the article

group of low level data fusion methods has recently been introduced that are able to separate the variation in all data-sets.

To investigate if the same latent processes underlie the different data-sets, component analysis can be very useful [5]. The construct of latent variables has properties that enable the integrated analysis of multiple data sets with a shared mode (e.g. same objects or variables). With shared variation across multiple data-sets a higher degree of interpretation is achieved and co-relations between variables across the data-sets become (more) apparent. Methods such as generalised SVD (GSVD), latent variable multivariate regression (LVMR), simultaneous component analysis (SCA) and canonical correlation analysis (CCA) have been used successfully in earlier studies [6–9]. Most of these methods or applications of these methods (i.e. CCA) focuses on the common/shared variation across the data-sets only. The interpretation of data however is not only improved by focussing on what is common but likely as important are those parts that are different from each other. These parts could include for example, measurement errors or other process and/or platform specific variations that would be distinctive for each data-set.

The concept of common and distinctive variation is visualized in Fig. 1a and b in which two different situations of overlapping data-sets ($\mathbf{X}_1(I \times J_1)$ and $\mathbf{X}_2(I \times J_2)$) are shown. The two data-sets are linked via common objects (I) but have different variables (J_1 and J_2). The areas of the circles are proportional to the total amount of variation in each data-set. The overlapping parts are tagged as $\mathbf{C}_1$ $(I \times J_1)$ and $\mathbf{C}_2$ $(I \times J_2)$ and describe shared (column) spaces for both data-sets. The spaces are not the same but are related (e.g. $\mathbf{C}_1 = \mathbf{C}_2\mathbf{W}_{2\rightarrow 1} + \mathbf{E}_1$ and $\mathbf{C}_2 = \mathbf{C}_1\mathbf{W}_{1\rightarrow 2} + \mathbf{E}_2$, in which the $\mathbf{W}$'s are the respective weight matrices). Whether or not the residuals $\mathbf{E}_1$ and $\mathbf{E}_2$ are truly zero, depends on the specific method. The distinctive parts $\mathbf{D}_1$ $(I \times J_1)$ and $\mathbf{D}_2$ $(I \times J_2)$ describe the variation specific for each data-set and the remainders are indicated by $\mathbf{E}_1$ $(I \times J_1)$ and $\mathbf{E}_2$ $(I \times J_2)$. In most methods the common parts are built up from the same latent components.

Figure 1a visualizes $\mathbf{C}_1$ and $\mathbf{C}_2$ as the intersection of the two data-sets. The common parts do not necessarily have to explain a similar amount of variation in each of the sets. The schematic in Fig. 1b demonstrates the situation in which the overlap of the two matrices is proportionally the same for data-set 2 (as in example A) but not for data-set 1.

Attempts have been made to capture both common and distinctive sources of variation across data-sets using GSVD [10], but it has been shown that GSVD does not yield an optimal approximation of the original data in a limited number of components [11]. Alternatives specifically designed for this purpose have been developed and complement the set of low level data fusion methods. In this paper we compare three implementations of such methods (JIVE [12, 13], DISCO-SCA [14, 15] and O2-PLS [16, 17]) with respect to their mathematical properties, interpretability, ease of use and overall performance using simulated and real data-sets. The different approaches to separate common from distinctive variation and the implications on (biological) interpretation are compared. For demonstration purposes we use mRNA and miRNA data from GlioBlastoma Multiform cells available at The Cancer Genome Atlas (TCGA) website [12, 18] as well as simulated data to identify the specific properties of the methods. We will only focus on the integrated analysis of two data-sets that are linked by their common objects. We assume that the data-sets are column-centered. A list of abbreviations and definitions is included in the Appendix.

Methods

From a general point of view Joint and Individual Variation Explained (JIVE), DIStinct and COmmon simultaneous component analysis (DISCO) and the 2 block latent variable regression with an orthogonal filtering

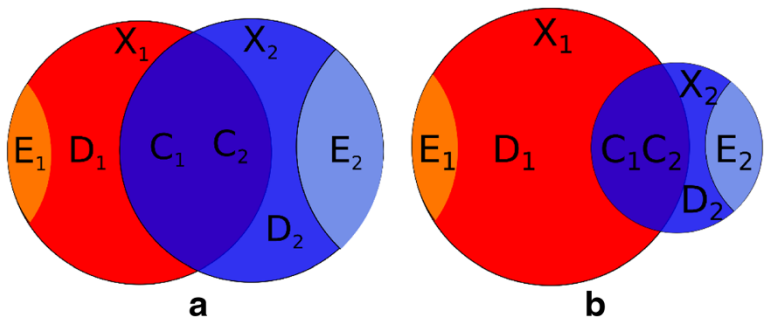

Fig. 1 Schematic overview of common and distinctive parts for two data-sets. **a**: two data-sets with equal total variance and **b**: two data-sets with different total variance

step (O2-PLS) all use a model in which the overlap of two (or more) data-sets is defined as common. The part that is not common is separated into a systematic part called distinctive while the nonsystematic part is called residual. The sum of the common part, the distinctive part and the residual error adds up to the original data-set. The generic decomposition of the two data-sets ($\mathbf{X}_1$ ($I \times J_1$) and $\mathbf{X}_2$ ($I \times J_2$)) in their respective common and distinctive parts for all three methods can be viewed as:

$$\begin{aligned} \mathbf{X}_1 &= \mathbf{C}_1 + \mathbf{D}_1 + \mathbf{E}_1 \\ \mathbf{X}_2 &= \mathbf{C}_2 + \mathbf{D}_2 + \mathbf{E}_2 \end{aligned} \quad (1)$$

In which $\mathbf{C}_1(I \times J_1)$ and $\mathbf{C}_2(I \times J_2)$ refer to the common parts, $\mathbf{D}_1(I \times J_1)$ and $\mathbf{D}_2(I \times J_2)$ to the distinctive parts and $\mathbf{E}_1(I \times J_1)$ and $\mathbf{E}_2(I \times J_2)$ to the residual error for both data-sets.

In their respective papers [10, 11, 14] the various authors use different terms that seem to have similar meaning like distinctive, systemic and individual, common and joint etc. For clarity purposes throughout this document we use **common** for combined or joint variation across data sets and **distinctive** for variation specific to each data set. Because the decomposition itself is different for each method, the interpretation of what is common and what is distinctive however, should be placed in the context of the method that is used. We will address the aspects of the different methods in terms of approximations of real data, orthogonalities, explained variance and we will discuss the complexity of proper model selection.

Algorithms

To compare the three different algorithms it is useful to first briefly reiterate through the different key steps of each method. For the specific implementation the reader is referred to the original papers but for convenience the algorithms are included in the Appendix. The Matlab [19] source code is available for download. Throughout this document the objects ($i = 1..I$) are the rows of the matrices ($I \times J$) and the variables correspond to the columns ($j = 1..J$). A full list of used symbols and dimensions of the different matrices can be found in the Appendix.

DISCO

After concatenation of the two matrices, $\mathbf{X}(I \times J) = [\mathbf{X}_1(I \times J_1)|\mathbf{X}_2(I \times J_2)]$, with $J = J_1 + J_2$), DISCO starts with an SCA routine on the concatenated matrix $\mathbf{X}$. This is followed by an orthogonal rotation step of the SCA scores and loadings towards an optimal user-defined target loading matrix $\mathbf{P}^*$ (i.e. a matrix in which each component is either distinctive for a specific data-set or common for any data-set). As an example, for two data-sets, $\mathbf{X}_1$ ($I \times 2$) and $\mathbf{X}_2$ ($I \times 3$), with one common component ($c_c = 1$) and one distinctive component for each data-set ($c_1 = c_2 = 1$), the total number of components c_t for the whole model is 3.

$$\mathbf{X} = [\mathbf{X}_1|\mathbf{X}_2]$$

$$\mathbf{X} = \mathbf{U}_{(c_t)}\mathbf{S}_{(c_t)}V^t_{(c_t)}$$

$$\mathbf{T}_{sca} = \mathbf{U}_{(c_t)}$$

$$\mathbf{P}_{sca} = \mathbf{V}_{(c_t)}\mathbf{S}_{(c_t)}$$

$$\hat{\mathbf{X}} = \mathbf{T}_{sca}\mathbf{P}^{\mathrm{t}}_{sca}$$

And $\mathbf{P}^*$ is:

$$\mathbf{P}^* = \begin{bmatrix} 1 & 0 & 1 \\ 1 & 0 & 1 \\ 0 & 1 & 1 \\ 0 & 1 & 1 \\ 0 & 1 & 1 \end{bmatrix}$$

In $\mathbf{P}^*$, the zeros are a hard constraint while the ones are not restricted and can be any value. The first two rows relate to the (two) variables in the first data-set, the last 3 rows relate to the variables for the second data-set. The first column relates to the first distinctive component (for data-set 1). The second column is reserved for the distinctive component for the second data-set and the third column is the loading for the common component in both data-sets. Through orthogonal rotation the best rotation matrix ($\mathbf{B}_{opt}$ ($c_t \times c_t$)) to rotate the $\mathbf{P}_{\mathrm{sca}}$ loadings ($\mathbf{P}_r$) towards the target loadings $\mathbf{P}^*$ is found by minimizing the squared sum of the 0 entries in the $\mathbf{P}_r$ matrix. To do just that a weight matrix ($\mathbf{W} = 1 - \mathbf{P}^*$) is used, in which all the 1 entries are set to 0 and the 0 entries to 1:

$$\mathbf{B}_{opt} \underset{min}{\rightarrow} \sum \left(\mathbf{W} \circ (\mathbf{P}_{\mathrm{sca}}\mathbf{B})\right)^2 \text{ s.t. } \mathbf{B}^{\mathbf{t}}\mathbf{B} = \mathbf{I}$$

$\mathbf{B}_{opt}$ is used to calculate the final rotated scores and loadings ($\mathbf{T}_r = \mathbf{T}_{sca}\mathbf{B}_{opt}$ and $\mathbf{P}_r = \mathbf{P}_{sca}\mathbf{B}_{opt}$). Consequently the smallest distance criterion is based only on the 0 entries (in $\mathbf{P}^*$) and thus on the distinctive components only. A perfect separation of the distinctive components is often not achieved; the positions where $\mathbf{P}^*$ is 0 are not exactly 0 in $\mathbf{P}_r$. Furthermore, the common variation is forced to be orthogonal to these distinctive parts which clearly could lead to sub-optimal estimations of this common variation. The effects of the orthogonality constraints are discussed later. The final decomposition of the DISCO algorithm is:

$$\begin{aligned} \mathbf{X}_1 &= \mathbf{C}_1 + \mathbf{D}_1 + \mathbf{E}_1 = \mathbf{T}_c\mathbf{P}^{\mathrm{t}}_{c_1} + \mathbf{T}_{d_1}\mathbf{P}^{\mathrm{t}}_{d_1} + \mathbf{E}_1 \\ \mathbf{X}_2 &= \mathbf{C}_2 + \mathbf{D}_2 + \mathbf{E}_2 = \mathbf{T}_c\mathbf{P}^{\mathrm{t}}_{c_2} + \mathbf{T}_{d_2}\mathbf{P}^{\mathrm{t}}_{d_2} + \mathbf{E}_2 \end{aligned} \quad (2)$$

The common scores ($\mathbf{T}_c$) for both data-sets are the same and are obtained by optimizing on the distinctive components.

JIVE

The JIVE algorithm is also based on an SCA of the concatenated data-sets ($\mathbf{X}$). The common parts for both data-sets ($\mathbf{C_k}$) are estimated simultaneously, $\mathbf{C} = [\mathbf{C}_1|\mathbf{C}_2] = \mathbf{T}_{sca}\mathbf{P}^{\mathrm{t}}_{sca}$ ($I \times J$), but now with only the number of common components (c_c) and not all the components (c_t) like in DISCO. The distinctive parts ($\mathbf{D}_1$ and $\mathbf{D}_2$) are estimated separately and iteratively based on an orthogonal residual ($\mathbf{R_k} - \mathbf{T}_{sca}\mathbf{T}^{\mathrm{t}}_{sca}\mathbf{R}_k$) matrix with c_k distinctive components. Using the same example as before;

$$\mathbf{X} = \mathbf{U}_{(c_c)}\mathbf{S}_{(c_c)}\mathbf{V}^{t}_{(c_c)}$$

$$\mathbf{T}_{sca} = \mathbf{U}_{(c_c)}$$

$$\mathbf{P}_{sca} = \mathbf{V}_{(c_c)}\mathbf{S}_{(c_c)}$$

$$\mathbf{C_k} = \mathbf{T}_{sca}\mathbf{P}^{\mathrm{t}}_{sca}$$

$$\mathbf{R_k} = \mathbf{X_k} - \mathbf{C_k}$$

$$\mathbf{R_k} - \mathbf{T}_{sca}\mathbf{T}^{\mathrm{t}}_{sca}\mathbf{R}_k = \mathbf{U}_{\boldsymbol{d_k}(\boldsymbol{c_k})}\mathbf{S}_{\boldsymbol{d_k}(\boldsymbol{c_k})}\mathbf{V}^{\boldsymbol{t}}_{\boldsymbol{d_k}(\boldsymbol{c_k})}$$

$$\mathbf{D_k} = \mathbf{U}_{d_k(\boldsymbol{c_k})}\mathbf{S}_{d_k(\boldsymbol{c_k})}\mathbf{V}^{\mathrm{t}}_{d_k(\boldsymbol{c_k})}$$

$$\mathbf{X} = \mathbf{X} - [\mathbf{D}_1|\mathbf{D}_2]$$

The steps are repeated until convergence of the combined common and distinctive matrices ($\mathbf{C} + \mathbf{D}$). By using the iterative and alternate optimization of the common and distinctinve parts, the orthogonality between the two distinctive parts that does exist in DISCO is no longer enforced. The resulting fit should be able to accommodate more types of data (e.g. the data has to conform to less criteria) than DISCO. Similar to DISCO the common parts are estimated from an SCA on both data-sets simultaneously and like DISCO there is no guarantee that both blocks take part in the common loadings $\mathbf{P}_{sca}$. As a consequence, the optimal solution could for example be one where $\mathbf{P}_{sca}(=[\mathbf{P}_1|\mathbf{P}_2])$ only has values for $\mathbf{P}_1$ and not $\mathbf{P}_2$ which hardly can be considered common.

The resulting decompostion (Eq. 3) in scores and loadings is exactly the same as for DISCO:

$$\begin{aligned}\mathbf{X}_1 &= \mathbf{C}_1 + \mathbf{D}_1 + \mathbf{E}_1 = \mathbf{T}_c\mathbf{P}^{\mathrm{t}}_{c_1} + \mathbf{T}_{d_1}\mathbf{P}^{\mathrm{t}}_{d_1} + \mathbf{E}_1 \\ \mathbf{X}_2 &= \mathbf{C}_2 + \mathbf{D}_2 + \mathbf{E}_2 = \mathbf{T}_c\mathbf{P}^{\mathrm{t}}_{c_2} + \mathbf{T}_{d_2}\mathbf{P}^{\mathrm{t}}_{d_2} + \mathbf{E}_2\end{aligned} \quad (3)$$

The common scores ($\mathbf{T}_c$) for both data-sets are the same. Because SCA is a least squares method and the common parts are determined first, those variables with much variance are likely to end up in the common parts. Because JIVE is an iterative solution the initial guesses for common and distinctive parts can change considerably during these iterations (see Additional file 1). If however, the distinctive variation is larger than the (combined) common variation these iterations will not prevent the method to mis-identify the common components.

O2-PLS

In contrast to DISCO and JIVE, that use an SCA on the concatenated data-sets, O2-PLS starts with an SVD on the covariance matrix ($\mathbf{X}^{\mathrm{t}}_1\mathbf{X}_2$ ($J_1 \times J_2$)) for an analysis of the common variation. Similar to JIVE, the common components are estimated first and from the orthogonal remainder to $\mathbf{P}_{c_k}$ ($\mathbf{R}^{\mathrm{t}}_{\boldsymbol{k}}\mathbf{T}_{c_k}$), per data-set. The distinctive component is estimated per component. When all distinctive components are removed from the data the common scores are updated. Using the same matrices $\mathbf{X}_1$ and $\mathbf{X}_2$;

$$\mathbf{X}^{\mathrm{t}}_2\mathbf{X}_1 = \mathbf{P}_{c_1(c_c)}\mathbf{D}_{(c_c)}\mathbf{P}^{\boldsymbol{t}}_{c_2(c_c)}$$

Deflate $\mathbf{X}_k$ per component:

$$\mathbf{T}_{c_k} = \mathbf{X}_k\mathbf{P}_{c_k}$$

$$\mathbf{R}_k = \mathbf{X}_k - \mathbf{T}_{c_k}\mathbf{P}^{\mathrm{t}}_{c_k}$$

$$\mathbf{R}^{\mathrm{t}}_k\mathbf{T}_{c_k} = \mathbf{u}_{d_k(1)}\mathbf{s}_{\boldsymbol{d_k}(1)}\mathbf{v}^{\boldsymbol{t}}_{d_k(1)}$$

$$\mathbf{t}_{d_{k,l}} = \mathbf{X}_k\mathbf{u}_{d_k}$$

$$\mathbf{p}_{d_{k,l}} = \left(\mathbf{t}^{\mathrm{t}}_{d_{k,l}}\mathbf{t}_{d_{k,l}}\right)^{-1}\mathbf{X}_k{}^{\mathrm{t}}\mathbf{t}_{d_{k,l}}$$

$$\mathbf{X}_k = \mathbf{X}_k - \mathbf{t}_{d_{k,l}}\mathbf{p}^{\mathrm{t}}_{d_{k,l}}$$

The choice of a covariance matrix seems appropriate since we are interested in co-varying variables across the data-sets. In case of orthogonal blocks where no common variation exists, the covariation matrix would be 0 and no common variation can be estimated. Similar to JIVE, the distinctive parts are calculated orthogonal to the common part for every data-set individually. Because the common parts are estimates from the individual blocks (not the concatenation) the algorithm itself is less restrictive than JIVE. With different common scores per data-set the decomposition of Eq. 1 in scores and loadings is almost similar to Eqs. 2 and 3;

$$\begin{aligned}\mathbf{X}_1 &= \mathbf{C}_1 + \mathbf{D}_1 + \mathbf{E}_1 = \mathbf{T}_{c_1}\mathbf{P}^{\mathrm{t}}_{c_1} + \mathbf{T}_{d_1}\mathbf{P}^{\mathrm{t}}_{d_1} + \mathbf{E}_1 \\ \mathbf{X}_2 &= \mathbf{C}_2 + \mathbf{D}_2 + \mathbf{E}_2 = \mathbf{T}_{c_2}\mathbf{P}^{\mathrm{t}}_{c_2} + \mathbf{T}_{d_2}\mathbf{P}^{\mathrm{t}}_{d_2} + \mathbf{E}_2\end{aligned} \quad (4)$$

As a post-processing step the common scores can be combined and by means of a regression model [20], for example an SCA of the combined common parts, global common scores can be calculated (i.e. $\mathbf{T}_c$ invariant for a block) so Eq. 4 would be exactly Eqs. 2 and 3 [21]. This would however also require recalculation of $\mathbf{P}_{c_1}$ and $\mathbf{P}_{c_2}$.

Orthogonalities

The similarity between the three methods is large in terms of scores and loadings that are created in accordance with the algorithms. The methods however are different in terms of constraints that are applied during

the decompositions which leads to different orthogonality properties and consequently different independence of the different common and distinctive parts.

The similarity between DISCO and JIVE is a consequence of the use of SCA in both methods. Because the final step in DISCO involves an orthogonal rotation of scores and loadings, the orthogonality between all the rotated scores and loadings remains. This rotation also forces orthogonality between the separate terms: $\mathbf{C}_1\mathbf{D}_1^t = 0$, $\mathbf{C}_1\mathbf{D}_2^t = 0$, $\mathbf{D}_1\mathbf{D}_2^t = 0$, $\mathbf{C}_2\mathbf{D}_1^t = 0$ and $\mathbf{C}_2\mathbf{D}_2^t = 0$. The error terms ($\mathbf{E}_1$ and $\mathbf{E}_2$) are orthogonal to each respective common part and distinctive part only. Orthogonality between the distinctive and common part per data-set in JIVE is enforced by estimation of the distinct components orthogonally to the common scores ($\mathbf{T}_{sca}\left(\mathbf{I}-\mathbf{T}_{sca}\mathbf{T}_{sca}^t\right)\mathbf{R}_k = \mathbf{U}_{\boldsymbol{d}_k(\boldsymbol{c}_k)}\mathbf{S}_{\boldsymbol{d}_k(\boldsymbol{c}_k)}\mathbf{V}_{\boldsymbol{d}_k(\boldsymbol{c}_k)}^{\boldsymbol{t}}$). There is no restriction for orthogonality between the distinctive parts of the different data-sets. Because the distinctive parts are calculated as the final step, the error matrix ($\mathbf{E}_k$) is orthogonal to the distinctive part but not to the common part.

The decomposition in scores and loadings using the O2-PLS algorithm (Eq. 4) is similar to those obtained when using JIVE or DISCO (Eqs. 2 and 3). The significant difference in terms of orthogonality follows from the fact that there is room for the common parts (i.e. $\mathbf{C}_1$ and $\mathbf{C}_2$) to have different loadings **and** scores. The common scores for each block ($\mathbf{T}_{c_1}$ and $\mathbf{T}_{c_2}$) themselves are expected to have a high correlation because the SVD was applied on the covariance matrix of the two matrices. The distinctive parts are estimated under the restriction that they are orthogonal to the common part per data-set. As a consequence the common parts per data-set share no variance with the distinctive parts. The distinctive parts themselves are not orthogonal to the common parts of the other data-set although the correlations are very small. Similar to JIVE the residuals ($\mathbf{E}_1$ and $\mathbf{E}_2$) in O2-PLS are found to be orthogonal only to the distinctive parts that are calculated as a final step.

A summary of the different orthogonality constraints for the three algorithms can be found in Table 1. It is clear that DISCO is the most strict and O2-PLS the most lenient regarding orthogonality properties. The different constraints that each algorithm imposes will affect the decomposition in different scores and loadings. What is designated as common and what is distinctive per method depends on these constraints. In DISCO the common part is defined as what is orthogonal to the distinctive parts while in JIVE this is the reverse i.e., what is distinctive is what is orthogonal to what is common. From a semantical point of view this seems equivalent but mathematically can generate very different results. These constraints will therefore be of importance when interpreting the data and consequently also for the application of the method. Orthogonality properties make it easier to come to a clear definition of these terms. Furthermore, orthogonality properties make the estimation of the separate parts easier.

Table 1 Summary table of all orthogonalities constraints for the three algorithms

	DISCO	JIVE	O2-PLS
Orthogonalities ($k \neq l$)			
$\boldsymbol{C}_k^t\boldsymbol{D}_k$	0	0	0
$\boldsymbol{E}_k^t\boldsymbol{C}_k$	0	≠0	≠0
$\boldsymbol{E}_k^t\boldsymbol{D}_k$	0	0	0
$\boldsymbol{C}_k^t\boldsymbol{D}_l$	0	0	≠0
$\boldsymbol{D}_k^t\boldsymbol{D}_l$	0	≠0	≠0
$\boldsymbol{E}_k^t\boldsymbol{C}_l$	0	≠0	≠0
$\boldsymbol{E}_k^t\boldsymbol{D}_l$	0	≠0	≠0
Characteristics			
Fusion	$[\mathbf{X}_1\|\mathbf{X}_2]$	$[\mathbf{X}_1\|\mathbf{X}_2]$	$\mathbf{X}_2^t\mathbf{X}_1$
First step	Distinctive	Common	Common
Optimization	Distinctive	Common + Distinctive	Common/Distinctive

(0: orthogonal, ≠0: no forced orthogonality)

The orthogonality constraints between allmost all parts in DISCO enforce that all underlying sources of variation can be split up in orthogonal parts, even the distinctive parts. From a mathematical viewpont this is a perfect separation but in biological phenomena such behavior will be rare. The solution therefore might be easier to find but it makes the interpretation more difficult. In JIVE the orthogonality constraint between the distinctive parts is removed and consequently is expected to be better suitable for biological data. With the single restriction of the distinctive parts to be orthogonal to the common part, O2-PLS is expected to suit most data-sets. The flexibility of O2-PLS is advantageous for fitting the best common and distinctive parts but might come at the expense of more loosely coupled common parts. Furthermore, the distinctive parts in O2-PLS are referred to as orthogonal to the counter common parts (e.g. $\boldsymbol{C}_k^t\boldsymbol{D}_l = 0$) and therefore do not optimally describe the total variation in the residual block ($\mathbf{R}_k$) which would limit the interpretation of these distinctive parts. The fact that we did not fully observe $\boldsymbol{C}_k^t\boldsymbol{D}_l = 0$ but still find some small residuals originates from the updated scores ($\mathbf{T}_{c_k} = \mathbf{X}_k\mathbf{P}_{c_k}$) after deflation in the algorithm.

Explained variances

The orthogonalities discussed above imply, because of the centering, uncorrelated structure between the distinctive and common parts. A closer look at the algorithms reveals an additional layer of complexity. This is

especially true for DISCO and JIVE where the SVD is taken from the concatenated matrix **X**. The simultaneous decomposition in DISCO:

$$\mathbf{X} = [\mathbf{X}_1|\mathbf{X}_2]$$

$$\hat{\mathbf{X}} = \mathbf{T}\mathbf{P}^{\mathrm{t}} = \mathbf{T}\mathbf{B}\mathbf{B}^{\mathrm{t}}\mathbf{P}^{\mathrm{t}} (\mathbf{T}_{rot} = \mathbf{T}\mathbf{B},\ \mathbf{P}_{rot} = \mathbf{P}\mathbf{B})$$

$$\hat{\mathbf{X}} = \mathbf{T}_{rot}\mathbf{P}_{rot}^{\mathrm{t}}$$

$$[\mathbf{X}_1|\mathbf{X}_2] = \mathbf{T}_{rot}\mathbf{P}_{rot}^{\mathrm{t}} + [\mathbf{E}_1|\mathbf{E}_2]$$

$$[\mathbf{X}_1|\mathbf{X}_2] = [\mathbf{C}_1|\mathbf{C}_2] + [\mathbf{D}_1|\mathbf{D}_2] + [\mathbf{E}_1|\mathbf{E}_2] = \mathbf{C} + \mathbf{D} + \mathbf{E}$$

decomposes the concatenated data-sets together in orthogonal **combined** parts. The explained variances of the separate parts of the **combined** model add up:

$$\|\mathbf{X}\|^2 = \|\mathbf{C}\|^2 + \|\mathbf{D}\|^2 + \|\mathbf{E}\|^2 = \|\mathbf{C} + \mathbf{D} + \mathbf{E}\|^2 \quad (5)$$

$\| \mathbf{E} \|^2$ is minimal for a given total number of components (c_t). The best $\mathbf{P}_{rot}$ however, is an approximation of $\mathbf{P}^*$ and because of orthogonality constraints, situations can occur where the rotation is not perfect. In such cases the elements set to zero in the original target matrix are different from zero in $\mathbf{P}_{rot}$. The exact estimation of $\mathbf{X}_k$ is:

$$\mathbf{X}_k = \mathbf{T}_c\mathbf{P}_{c_k}^{\mathrm{t}} + \mathbf{T}_{d_k}\mathbf{P}_{d_k}^{\mathrm{t}} + \mathbf{T}_{d_{\neq k}}\mathbf{P}_{d_k}^{\mathrm{t}} + \mathbf{E}_k \quad (6)$$

The cross-over ($\mathbf{T}_{d_{\neq k}}\mathbf{P}_{d_k}^{\mathrm{t}}$) part of the original $\mathbf{X}_k$, the variation in $\mathbf{X}_k$ that is explained by the distinctive components of the other data sets, is minimized during the DISCO iterations and is indicative for the influence both data-sets have on each others individual loadings and thus affect direct interpretation. The size of the cross-over part depends on the data and the number of distinctive components reserved for the other data-sets. The model selection procedure is based on minimization of this cross-over content.

Contrary to DISCO, not all parts in both JIVE and O2-PLS are orthogonal (see Table 1). Equation 5 does not hold and should be reduced, per data-set, to:

$$\|\mathbf{C}_k\|^2 + \|\mathbf{D}_k\|^2 = \|\mathbf{C}_k + \mathbf{D}_k\|^2 \quad (7)$$

The residual $\mathbf{E}_k$ is not orthogonal to the common part $\mathbf{C}_k$ which indicates that the final solution found for $\mathbf{E}_k$ could still hold some information from $\mathbf{C}_k$. To find the correct value for $\mathbf{E}_k$ type III partial explained sum of squares for residuals should be applied by projecting $\mathbf{E}_k$ on $\mathbf{C}_k$ and only consider orthogonal parts of residual [22].

Interpretation

Even though the fusion methods have separated common from distinctive variation the interpretation of the results can be hampered or sometimes even prohibited by the fact that the data-sets themselves do not conform to the appropriate criteria. The most apparent critereon is the link between the samples across the different data-sets. If the different data-sets for example contain technical replicates, the fusion can only be performed on the averages of the technical replicates as the technical replicates of different data sets are not directly related. Secondly, in order to give equal chance to all data sets to be represented in the model, large blocks should not be favoured just because of their size. Therefore after variable scaling, a block scaling is usually applied such that the sum of squares of all blocks is equal. This block scaling however lowers the influence of the individual variables if the data-set consists of many variables and thus could be the cause of under-estimation.

Common variation can be thought of as variation that is related between data-sets. Because there is no mandatory contribution of both data-sets to the common parts when using JIVE or DISCO the results should always be validated for a shared variation between the data-sets. Second, for blocks where I is larger than J_k the rank of data-set $\mathbf{X}_k$ is bounded by the number of variables. The selection of the common score $\mathbf{T}_c$ from the concatenated matrix $\mathbf{X}$ defines a direction in the I^{th}-dimensional columnspace that may be outside the J_k-dimensional subspace in R^{I} defined by $\mathbf{X}_k$. $\mathbf{C}_k$, which is built from $\mathbf{T}_c$ will therefore also be outside the J_{k} dimensional subspace defined by $\mathbf{X}_k$. Thus there will be variation in $\mathbf{C}_k$ which is not in $\mathbf{X}_k$. When scores $\mathbf{T}_{d_k}$ for the distinctive part $\mathbf{D}_k$ are calculated, they are forced to be orthogonal to $\mathbf{T}_c$, but not forced to be in the columnspace of $\mathbf{X}_k$. This means that also the distinctive part $\mathbf{D}_k$ may not be in the columnspace of $\mathbf{X}_k$. Because of this, the interpretation of the loadings from $\mathbf{C}_k$ and $\mathbf{D}_k$ can go wrong, as they may represent variation that is not in $\mathbf{X}_k$.

To check whether the distinctive and common parts are still in the column space of the original matrix of the separate data-sets, the projections of $\mathbf{C}_k$ and $\mathbf{D}_k$ on $\mathbf{X}_k$ can be determined via:

$$\hat{\mathbf{C}}_k = \mathbf{X}_k\mathbf{X}_k^{+}\mathbf{C}_k \quad (8)$$

The residual (i.e. $\| \mathbf{C}_k - \hat{\mathbf{C}}_k \|^2$ or $\|\mathbf{D}_k - \hat{\mathbf{D}}_k\|^2$) is zero for a perfect projection and different from zero if $\mathbf{C}_k$ or $\mathbf{D}_k$ is not within the column space of $\mathbf{X}_k$.

The common and distinct parts of O2-PLS are based on an SVD of the covariance matrix of $\mathbf{X}_1$ and

$\mathbf{X}_2$ ($[\mathbf{P}_{c_1}\mathbf{D}, \mathbf{P}_{c_2}] = svd(\mathbf{X}_2^t \boldsymbol{X}_1, c_c)$). The SVD decomposes the covariance matrix in orthogonal contributions. $\mathbf{P}_{c_1}$ is expressed in terms of variables of $\mathbf{X}_1$ and $\mathbf{P}_{c_2}$ in terms of variables of $\mathbf{X}_2$. The subsequent steps in the algorithm only affect the individual blocks. Consequently, no variation from one data-set is introduced into the other and projection issues like in JIVE and DISCO do not occur. If the post-processing step is performed to calculate global common scores, variation from other data-sets is introduced and also in this case the projection errors need to be evaluated.

The issue that the common scores of multiple data sets may not be in the column space of each data set separately, and the problems this brings was already discussed earlier for multiblock PLS models [23, 24]. In the latter paper the common score was called the super score. It was shown that deflation of information from the separate blocks using the super score leads to introduction of variation that was never present in the block. When information which is not present in the data set is subtracted from that dataset, it is actually (negatively) introduced.

Model selection

Both orthogonalities and explained variances on touch the heart of exactly what is common and what is distinctive. The three methods are all different in this respect. All three methods however, can only decompose the data-sets if the optimal number of common and distinctive components for the final model are known. It is important that the selected model is appropriate for the data-sets that are analysed and each method has its own strategy of selecting the appropriate model.

Model selection in DISCO is a two step process. In the first step the total number of components (c_t) is selected based on proportion of variance accounted for by the simultaneous components for each individual data block. The second step finds the "best" performing model from all possible combinations of common (c_c) and distinctive components (c_{d_k}) by minimizing the cross-over parts of each data-set.

In JIVE the configuration of the model is based on the analysis of permuted versions of the original matrix. For the common components complete rows of each data-set are permuted. This removes the link between the objects from the different data-sets, but does not remove the correlation structure inside each block. The eigenvalues for a large number of permuted matrices are determined. The number of common components is defined as that number where the eigenvalues of the original matrix ($\mathbf{X}$) are (still) larger than the permuted ones (with a certain α). For the distinct components per data set $\mathbf{X}_k$, the rows of each variable in that data-set are permuted to disturb the variable object relationship. Again the eigenvalues of the original data set are compared to the eigenvalues of the permuted data sets to find the optimal number of distinct components for each $\mathbf{X}_k$. These setting are used as input for a new start of the estimation of the number of components. This process is repeated until convergence of the number of common and distinctive components.

The model selection of O2-PLS as described in the papers [16, 17] is not clear about exactly which procedures to follow. We have adopted the strategy of first selecting the number of common components based on the covariance matrix followed by an estimation of the number of individual components per data-set using PCA cross validation after the common parts have been removed from the data-sets using an OPLS approach.

Experimental

To test the three methods in different conditions we use simulated data. We will keep the model itself small with only 1 common component and 1 (or 2) individual component(s) per data-set. To generate the data we use the score and loading structure from Eqs. 2 and 3.

$$\mathbf{X}_1 = \mathbf{T}_c\mathbf{P}_{c_1}^t + \mathbf{T}_{d_1}\mathbf{P}_{d_1}^t + \mathbf{E}_1$$
$$\mathbf{X}_2 = \mathbf{T}_c\mathbf{P}_{c_2}^t + \mathbf{T}_{d_2}\mathbf{P}_{d_2}^t + \mathbf{E}_2$$

The scores $\mathbf{T}_c$, $\mathbf{T}_{d_1}$ and $\mathbf{T}_{d_2}$ are drawn from a standard normal distribution in such a way that they are orthogonal to each other. Then each scores vector was scaled to length 1. The error terms $\mathbf{E}_1$ and $\mathbf{E}_2$ are based on pseudo numbers drawn from a standard normal distribution. The data-sets have 70 observations each ($I = 70$) and $\mathbf{X}_1$ contains 100 variables ($J_1 = 100$) and $\mathbf{X}_2$ 50 variables ($J_2 = 50$). The data of each data-set is column-centered and the variance of each block is scaled to unit variance. In our examples we have chosen a set of spectral loadings for illustrative purposes. In functional genomics data-sets e.g. transcriptomics or metabolomics data a similar situation can be envisioned when in functional groups the features are expected to highly correlate. The latent components then describe structured variation of the functional groups over the objects.

The three methods will be evaluated using the model settings that were suggested by the original model estimation procedure of each method respectively and if different from the actual model, with the real model settings as well. Two different scenarios are evaluated in

which two different situations are simulated for the two data-sets:

1. Scenario 1, abundant variation in common loadings, almost orthogonal loadings
2. Scenario 2, low abundant variation in common loadings, almost orthogonal loadings

Figure 2 shows the loadings that are used to generate the data of the two blocks for both scenarios. The contributions of the distinctive and common parts for the different scenarios are listed in Additional file 1: Table S1 and Table 2 (Scenario 1: ($0.66^{c1}/0.28^{d1}$ and $0.85^{c2}/0.13^{d2}$), scenario 2: ($0.11^{c1}/0.88^{d1}$ and $0.62^{c2}/0.36^{d2}$)). The first scenario should give insight in the performance of the methods under conditions well suited to find the common variation. The second scenario should reveal issues for data that is more realistic like for example, the detection and removal of batch effects.

The three methods will also be applied to experimental data from GlioBlastoma Multiform (GBM) brain tumors available at The Cancer Genome Atlas (TCGA). The mRNA (234 × 23293) and miRNA (234 × 534) data-sets describe the messenger RNA's and small RNA's profiles of 234 subjects that suffer from different kinds of brain tumors. The same data was already analysed by JIVE in its original paper [12]. Here we use it for comparison of JIVE and the other two methods.

Results

Scenario 1, abundant common variance, almost orthogonal loadings

The data sets in the 1st scenario did not lead to any problems. All three methods properly select the model of common and distinctive components (i.e. 1 common, and 1 distinctive component for each data-set). The results of DISCO, JIVE and O2-PLS almost exactly match the simulated scores and loadings, which from a mathematical point of view is also expected (see Appendix, "Observations on JIVE, SCA and covariance"). The loadings are plotted in Additional file 1: Figure S3. The correlation of the fitted scores with the original scores is 1 for all methods.

Additional file 1: Table S1 summarizes the explained variances for the fitted results by the different models. The different methods decompose the two data-sets into the same common and distinctive parts. As discussed earlier, the errors for JIVE and O2-PLS are not orthogonal to the common parts and therefore cannot be calculated as the difference of $\mathbf{X}_k$ and the common and distinctive variance combined ($\mathbf{C}_k + \mathbf{D}_k$). In this case however, the data was fabricated with orthogonal common and distinctive scores and we were able to

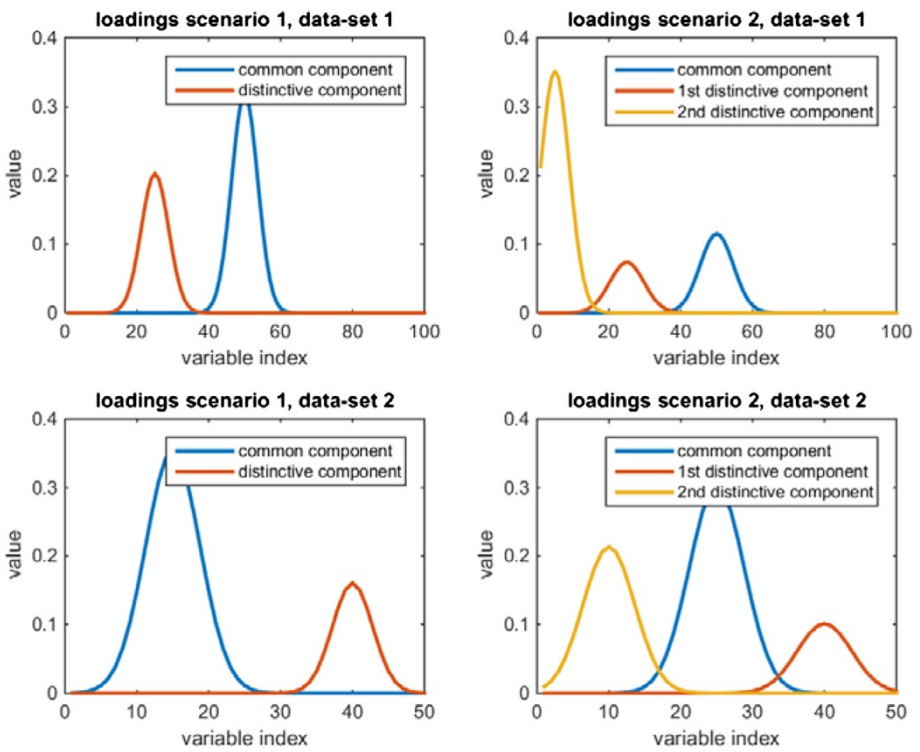

Fig. 2 The loadings that were used to generate the data with for both scenarios

Table 2 Summary table of explained variances by the different methods in the second scenario using the real model settings (1,2,2)

Data-set	Part	Real	DISCO	JIVE	O2-PLS
1	Common	0.11	0.11	0.83	0.11
1	Distinctive	0.88	0.88	0.16	0.88
1	Error	0.01	0.01	0.01	0.01
2	Common	0.62	0.62	0.00	0.62
2	Distinctive	0.36	0.36	0.91	0.32
2	Error	0.02	0.02	0.09	0.06

calculated the error as the difference. Furthermore $\|\mathbf{C}_k\mathbf{C}_k^+\mathbf{E}_k\|^2 \ll \|\mathbf{E}_k\|^2$ which implies that the projection of $\mathbf{E}_k$ on $\mathbf{C}_k$ is very small indeed.

Scenario 2, low abundant common variance, almost orthogonal loadings

In the second scenario the model was made more complex with less abundant common variance and more distinctive components per data-set. The difference between the methods already becomes apparent in the model selection. Additional file 1: Table S2 shows the estimated number of component models for the different methods. Each of the three methods selects a different 'best' model. With the O2-PLS cross-validation the 'real' model is selected. Both JIVE and DISCO select 0 common components.

For completeness, the loading plots and score assessments of the decompositions of JIVE and DISCO with the suggested model settings are included in the Additional file 1. The estimated common and distinctive loadings for the methods with the real model settings (1,2,2) are shown in Fig. 3.

The DISCO results with the 'real' model settings show a perfect decomposition in loadings and scores for both data-sets. The JIVE results show that all three components of the first data-set are fitted perfectly but that the common component is identified incorrectly; the component with the largest variance is identified as common. Because of the orthogonality restriction of $\mathbf{C}_1\mathbf{D}_1^t = 0$ and $\mathbf{C}_1\mathbf{D}_2^t = 0.$, the real common component in data-set 2 cannot be selected anymore which results in a score vector of zero (the blue line). The two remaining distinctive components are used to fit the two loadings with the largest variation.

In JIVE the first step is to select the allocated number of common components. At this stage this selection is only determined by the largest variance, regardless whether or not this is 'real'. If this selected part happens to be the distinctive part, the 'real' distinctive part is designated as common variance. In these cases the JIVE algorithm is not able to classify it as common, even after all the iterations. This behavior is investigated further by generating different data-sets with increasing variation in the common component. For each data-set the JIVE decomposition is run and the proper identification of the common and distinctive components is recorded (see Additional file 1). Only when the total common variation is larger than the variation of the largest distinctive component, the proper common component is identified.

The O2-PLS method suggested the real model complexity and the decomposition in loadings and scores show a good fit to the original data. The loading profiles show a good fit for the first data-set but for the second data-set the smallest individual component is underestimated. This is also reflected in the amount of explained distinctive variation for the second data-set. Table 2 summarizes the explained variation for the fitted blocks by the different models. All methods steer towards a maximum amount of explained variation. Again, the residuals were determined as differences with the original data because the data was generated with orthogonal scores and $\|\mathbf{C}_k\mathbf{C}_k^+\mathbf{E}_k\|^2 \ll \|\mathbf{E}_k\|^2$.

GlioBlastoma

The mRNA and miRNA measurements of Glioblastoma cells were used in the JIVE paper to introduce the method. We use the data to compare JIVE to DISCO and O2-PLS. We adopted the model settings that were found by the permutation approach (i.e. 5 common components, 33 distinctive components for mRNA and 13 for miRNA). For completeness the optimal number of components for the models was estimated again with each model selection method and the results are shown in Additional file 1: Table S4. The data were mean centered for each feature and each data-set was normalized to unit sum of squares. The data concerns different types of brain tumor cells.

As an example the O2-PLS score plots for both mRNA and miRNA for the common and distinctive parts are presented in Fig. 4. The common part shows a much clearer separation between the groups than the distinctive parts. The explained common and distinctive variation of the methods are listed in Table 3. With the exact same model settings, the JIVE method is able to explain approximately 5 % more of combined distinctive and common variation than DISCO and O2-PLS ($\|\mathbf{C}_k\mathbf{C}_k^+\mathbf{E}_k\|^2 \ll \|\mathbf{E}_k\|^2$ for both data-sets). In comparison to DISCO and O2-PLS, JIVE describes less common variation but more distinctive variation. This phenomenon can possibly be accounted for by the iterative behavior of JIVE. By iteratively estimating the common and distinctive parts from only a selected part of the variation in the data, the common part seems less affected by over fitting. This phenomenon is further discussed in the Additional file 1.

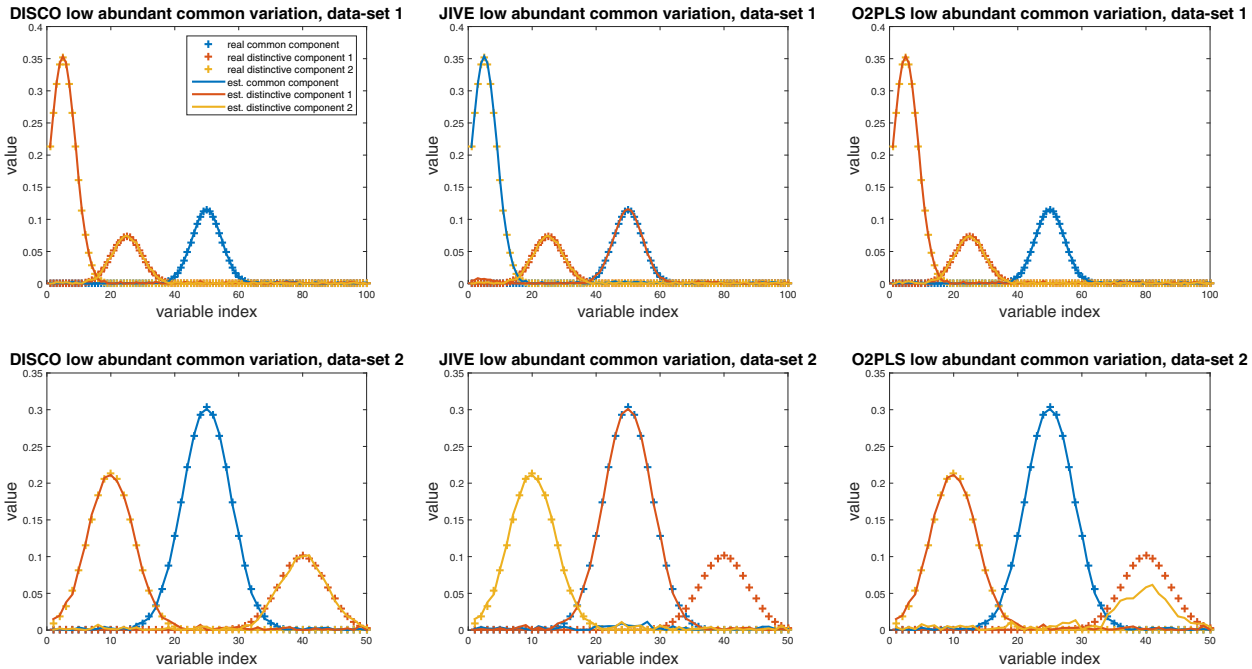

Fig. 3 Real and fitted common and distinctive parts for each methods when there is an low abundant common variation (scenario 2). The real common component is show with a blue cross, the first and second distinctive component with an orange and yellow cross. The fits by the methods are identified by a line with in their respective color

To study the overlap in the three methods, the percentages explained variation in the common part and in the distinctive part per gene are plotted against each other in Fig. 5 for mRNA and miRNA. On the left side the results of the common parts are given. The explained variation for the genes in the common part using O2-PLS and DISCO are strongly correlating. The explained variation for the genes using JIVE is clearly different. The common part in JIVE describes a lower amount of explained variation than the other methods. The distinctive part (on the right-hand side) shows the same phenomenon. Again the explained variation for the distinctive part is similar using O2-PLS and DISCO, while JIVE now describes a higher amount of explained variation. The figures on the diagonal show the distribution in explained variation for each of the 3 methods. This is very similar for the three methods. What is striking however is the difference in distribution of explained variation between the common and distinctive parts. In the common part, most genes are hardly explained while a low number of genes is highly explained. For the distinctive part no such preference is observed and a normal like distribution of explained variation is obtained.

For the miRNA, the situation is similar to the mRNA data. Again JIVE has a lower explained variation for each miRNA in the common part and a higher explained variation in the distinctive part compared to DISCO and O2-PLS. The distribution of the explained variation of the distinctive part is clearly different than for the mRNA. For miRNA, still many features are not well described. This could be related to a lower amount of systematic variation in the miRNA's and consequently, lower correlation between the different miRNA's. Therefore, each component only describes few miRNA features.

One explanation for DISCO is that orthogonality restrictions prohibit optimal fitting and as a result the cross over variation (i.e. the variation for miRNA explained by the distinctive score for the mRNA) is significant. For miRNA this was 13 % and for mRNA this was 4 % of the total variance. This amount of cross over variation is much larger for miRNA than mRNA because the 33 distinctive mRNA components all add to the cross over variation of miRNA compared to only 13 components vice versa.

In the O2-PLS method the initial common scores ($\mathbf{T}_{c_k}$) are estimated from the initial loadings ($\mathbf{P}_{\mathrm{k}}$) and original data ($\mathbf{X}_k$). The distinctive components are removed from the remainder $\mathbf{R}_k$ $\left(\mathbf{R}_k = \mathbf{X}_k - \mathbf{T}_{c_k}\mathbf{P}_{c_k}^{\mathrm{t}}\right)$ and $\mathbf{X}_k$ is updated. However, in the final step (step 12 in the O2-PLS algorithm see Appendix), the common part is recalculated from the updated $\mathbf{X}_k$. This recalculation

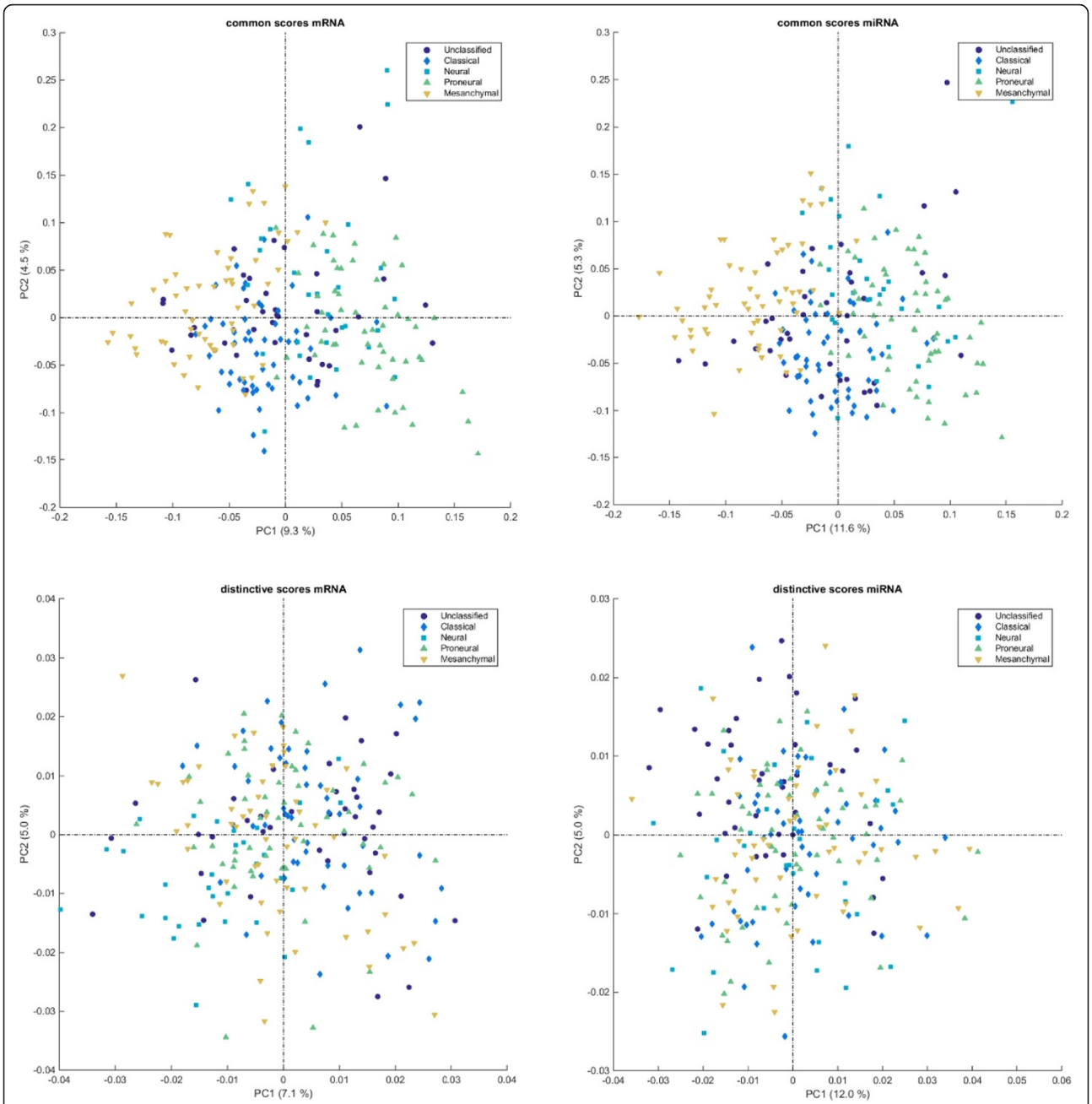

Fig. 4 Score plots of the common (*top row*) and distinctive (*bottom row*) parts of the mRNA data-set (*left column*) and miRNA data-set (*right column*) after O2-PLS decomposition

gives a lower amount of variation for the common part than before $\mathbf{X}_k$ was deflated with distinct components. This variation can neither be described by the distinctive nor common part of the model anymore. Large discrepancies indicate that the estimation of the initial common part contained larger amounts of orthogonal variation. After $\mathbf{T}_{c_k}$ has been re-estimated, the distinctive part is not recalculated anymore. Perhaps more total variance could have been accounted for if O2-PLS would have used an iterative procedure like JIVE, which is fully iterative.

The score plots of the common and distinctive parts for the different methods all reveal a better separation of the classes in the common part of the miRNA data-set. To indicate the quality of class separation we adopted the standardized subtype within sums of squares (SWISS) from the original JIVE paper. This represents the variability within subtypes

Table 3 Summary table of fitted explained variation by the different methods using the real mRNA and miRNA data-sets

Data-set	Part	DISCO	JIVE	O2-PLS
mRNA	Common	0.22	0.15	0.20
mRNA	Distinctive	0.45	0.57	0.48
mRNA	Total	0.68	0.72	0.68
miRNA	Common	0.33	0.26	0.29
miRNA	Distinctive	0.40	0.49	0.40
miRNA	Total	0.73	0.75	0.70

(across all rows) as a proportion of total variability. A lower score indicates better class separation. Table 4 shows the SWISS scores for both data sets using all three methods. The SWISS score for the common parts is compared to the SWISS scores of a 5 component PCA solution of both sets to see whether the removal of the distinctive information would provide a better set of common scores compared to the normal PCA scores. For O2-PLS we see a slight improvement to a SWISS of 0.65, while the JIVE SWISS score is worse (0.74). We see that the distinctive parts of the data have lost their discriminative power. Note that the SWISS for the common parts for both data sets is exactly the same for DISCO and JIVE as the common scores are the same for those methods.

The high correspondence in explained variation for each mRNA and miRNA feature between DISCO and O2-PLS is corroborated by their scores. Table 5 shows the RV matrix correlation [25] between the scores of the different methods. Again a high correlation between the O2-PLS and DISCO scores are observed for the common part. For the distinctive part this cannot be observed.

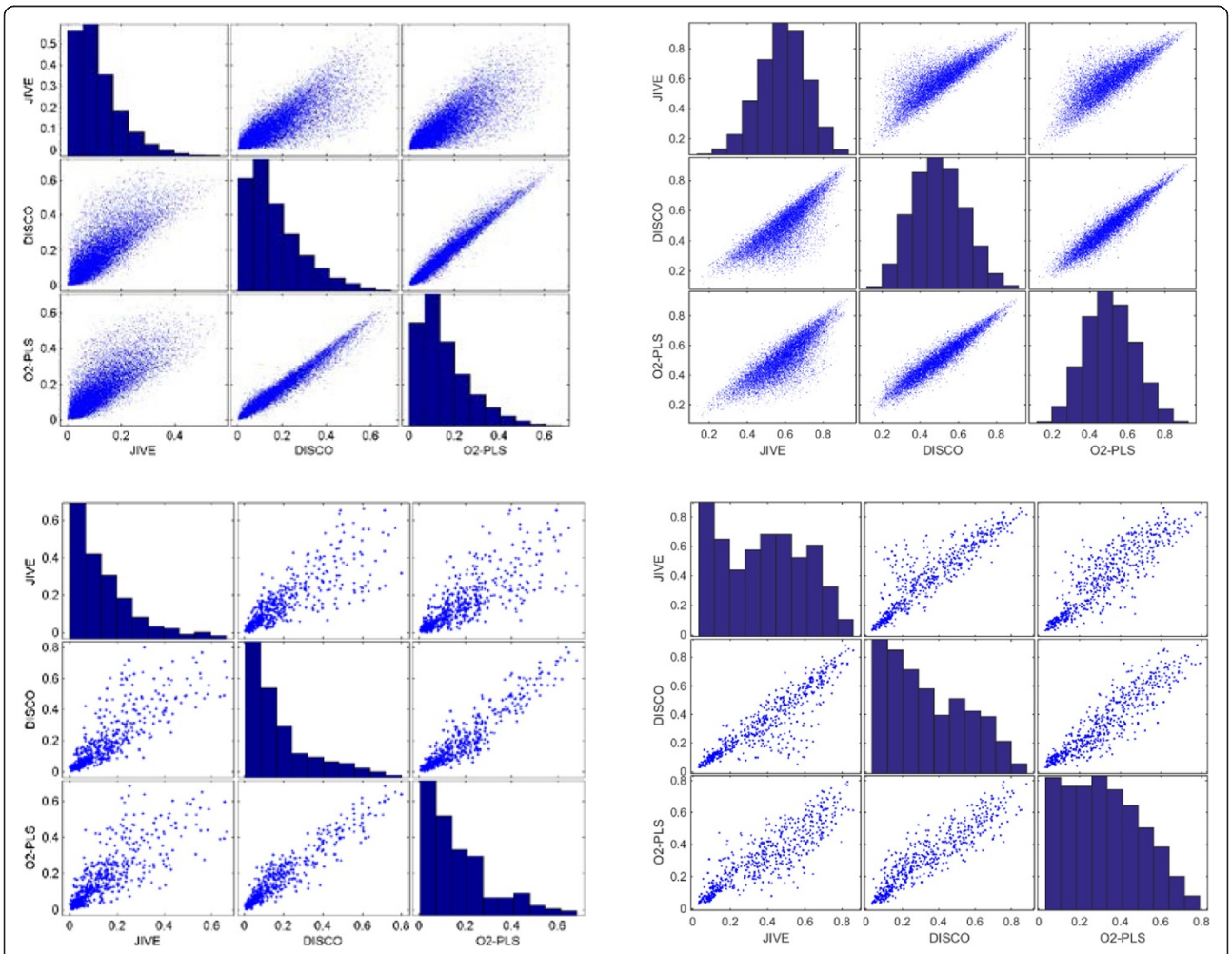

Fig. 5 Scatter plots of % variance of original variable explained in common parts (*left*) and % variance of original variable explained in distinctive parts (*right*) of the mRNA data-set on the top row and of the miRNA on the bottom row between the the different algorithms. DISCO and O2-PLS look very similar. JIVE shows more genes of which more variance is used in the distinctive part which coincides with the increased amount of distinct variance explained by JIVE in comparison to DISCO and O2-PLS

Table 4 Summary of the SWISS scores for the common and distinctive parts identified by the different models during the analysis of the mRNA/miRNA GlioBlastoma data

	Common (mRNA/miRNA)	Distinctive mRNA	Distinctive miRNA
REAL (5 PC's)	0.66/0.79		
DISCO	0.67	0.94	0.99
JIVE	0.74	0.92	0.94
O2-PLS	0.65/0.66	0.97	0.93

Discussion and conlusions

The three methods discussed in this paper to separate common from distinct information all use different approaches, which lead to slightly different models of the data. What is exactly common variation and what is distinctive depends on the different orthogonality constraints applied and the algorithms used to estimate these different parts. When the common variation is abundant, all methods are able to find the correct solution. With real data however, complexities in the data are treated differently by the three methods.

Due to fewer orthogonality constraints that are imposed by JIVE and O2-PLS, there is more freedom to select the scores and loadings for the two data-sets. This freedom is not present in DISCO which has the most severe orthogonality restrictions. In the two scenarios shown in this paper, all scores and loadings were chosen orthogonal. Therefore DISCO was able to find the correct scores and loadings while JIVE and O2-PLS found variations thereof that still obayed their orthogonality assumptions. In case of less abundant common variation, both JIVE and DISCO failed to detect the proper amount of common components which can be understood from the methods themselves. Not knowing the real model however can give rise to unexpected results while decomposing the data in common and distinctive components.

Table 5 RV modified coefficients of the common and distinctive scores for GlioBlastoma data-sets

Data-set	Part	Method	O2-PLS	DISCO	JIVE
mRNA	Common	O2-PLS	X	0.77/0.67	0.42/0.41
mRNA	Common	DISCO	0.77/0.67	X	0.58
mRNA	Common	JIVE	0.42/0.41	0.58	X
mRNA	Distinctive	O2-PLS	X	0.53	0.58
mRNA	Distinctive	DISCO	0.53	X	0.68
mRNA	Distinctive	JIVE	0.58	0.68	X
miRNA	Distinctive	O2-PLS	X	0.56	0.55
miRNA	Distinctive	DISCO	0.56	X	0.74
miRNA	Distinctive	JIVE	0.55	0.74	X

Even with the optimal model settings selected the JIVE method is the most susceptible to identifying the wrong common components. Due to the SCA of the concatenated matrix JIVE has problems finding common components especially when they are smaller than a distinctive component in one of the blocks. If the common and distinctive variation is approximately of the same magnitude, JIVE is able to properly identify them due to its iterative nature. JIVE re-estimates the common and distinctive parts until they converge, while O2-PLS, which only once re-estimates the common part once, seems to be stuck in a sub optimal solution for the distinctive part.

When small data sets with a low number of features ($J_k < I$) are used, these data sets may not be well represented by the common scores in JIVE, and even worse, the common scores present information that is not even present in these blocks. This may lead to misinterpretation of both common scores and distinctive scores of such a block [24]. The O2-PLS algorithm is the most flexible one and allows the separate and distinctive parts to be determined using block scores instead of super scores. This way no information is transferred from one data-set to the other. The distinctive parts however, are also limited by orthogonality constraints and therefore have a biased interpretability.

In the real data example the three methods all selected a smaller number of common than distinct components. In contrast to the simulations, O2-PLS suggested a smaller number of common components than JIVE and DISCO. This could possibly indicate an over estimation of the number of common components by DISCO and JIVE. It was shown that the lack of structure in the raw miRNA data-set has been replaced by an apparent structure in the common part. The combination of the data-sets has revealed a subset of miRNA's that mathematically can be linked to the mRNA's by all three methods. Because the methods are not supervised, the appearing structure gives rise to further biological interpretation of not only the common parts but also the distinctive parts. In situations like these, DISCO, JIVE and O2-PLS can be considered to act as pre-processing steps (i.e. filtering steps).

In summary, all three methods have their own approach to estimate common and distinctive variation with their specific strength and weaknesses. Due to their orthogonality properties and their used algorithms their view on the data is slightly different. By assuming orthogonality between common and distinctive, true natural or biological phenomena that may not be orthogonal at all might be misinterpreted.

Appendix

List of used symbols

K	total number of datasets, $k = 1..K$
I	number of rows (objects)
J_k	number of columns (variables) for matrix k
J	total number of variables ($\sum_1^K J_k$)
c_c	number of components for common part
c_k	number of components for distinctive part of matrix k
c_t	total number of components ($(\sum_{k=1}^{K} c_k) + c_c$)
$\mathbf{X}_k$	data matrix ($I \times J_k$)
$\mathbf{X}$	concatenated data matrix $[\mathbf{X}_k] \ldots |\mathbf{X}_K](I \times J)$
$\mathbf{C}_k$	common part of matrix k ($I \times J_k$)
$\mathbf{C}$	concatenated common parts $[\mathbf{C}_k] \ldots |\mathbf{C}_K](I \times J)$
$\mathbf{D}_k$	distinctive part of matrix k ($I \times J_k$)
$\mathbf{D}$	concatenated distinctive parts $[\mathbf{D}_k] \ldots |\mathbf{D}_K]$ ($I \times J$)
$\mathbf{E}_k$	the residual error of matrix k ($I \times J_k$)
$\mathbf{E}$	concatenated residual errors $[\mathbf{E}_k] \ldots |\mathbf{E}_K]$ ($I \times J$)
$\mathbf{T}_{sca}$	scores of SCA model (corresponds to objects) ($I \times J_t$)
$\mathbf{P}_{sca}$	loadings of SCA model (corresponds to variables) ($J \times c_t$)
$\mathbf{P}^*$	rotation target loading in DISCO model ($J \times c_t$)
$\mathbf{B}$	rotation matrix in DISCO ($c_t \times c_t$)
$\mathbf{W}$	weight matrix (used in DISCO) to penalize rotation matrix ($J \times c_t$)
$\mathbf{T}_c$	common scores (SCA and JIVE) ($I \times c_c$)
$\mathbf{P}_c$	common loadings (JIVE) ($I \times c_c$)
$\mathbf{T}_{c_k}$	common scores (O2-PLS) for matrix k ($I \times c_c$)
$\mathbf{P}_{c_k}$	common loadings for matrix k ($J_k \times c_c$)
$\mathbf{T}_{d_k}$	distinctive scores for matrix k ($I \times c_k$)
$\mathbf{P}_{d_k}$	distinctive loadings for matrix k ($J_k \times c_k$)
$\circ$	Hadamard (element-wise) matrix product

Algorithms

DISCO

1. Define a target loading matrix ($\mathbf{P}^*$) of zeros and ones based on the model that was defined by the common and distinctive components (c_c, c_1, and c_2);

$$\mathbf{P}^* = \begin{bmatrix} 1^{J_1 \times c_1} & 0^{J_1 \times c_2} & 1^{J_1 \times c_c} \\ 0^{J_2 \times c_1} & 1^{J_2 \times c_2} & 1^{J_2 \times c_c} \end{bmatrix}$$

2. Define the weight matrix as $\mathbf{W} = 1 - \mathbf{P}^*$, where $\mathbf{1}$ is a matrix of ones.
3. $\mathbf{X} = [\mathbf{X}_1|\mathbf{X}_2]$
4. $\mathbf{X} = \mathbf{U}_{(c_t)}\mathbf{S}_{(c_t)}V^t_{(c_t)}$
5. $\mathbf{T}_{sca} = \mathbf{U}_{(c_t)}$
6. $\mathbf{P}_{sca} = \mathbf{V}_{(c_t)}\mathbf{S}_{(c_t)}$
7. Randomly initialize $\mathbf{B}^0$ subject to $\mathbf{B}^{0t}\,\mathbf{B}^0 = \mathbf{I} = \mathbf{B}^0\mathbf{B}^{0t}$
8. Intialize iteration index $l = 1$
9. $\mathbf{Y} = \mathbf{P}_{sca}\mathbf{B}^{l-1} + \mathbf{W} \circ (\mathbf{P}^* - \mathbf{P}_{sca}\mathbf{B}^{l-1})$
10. $[\mathbf{U}_r, \mathbf{S}_r, \mathbf{V}_r] = svd(\mathbf{Y}^t\,\mathbf{P}_{sca})$
11. $\mathbf{B}^l = \mathbf{V}_r\mathbf{U}_r^t$
12. Compute $h(\mathbf{B}^l) = \| \mathbf{W} \circ (\mathbf{P}_{sca}\mathbf{B}^l - \mathbf{P}^*) \|^2$
13. Repeat step **9-12** until $h(\mathbf{B}^l) - h(\mathbf{B}^{l-1}) < \tau$ or $l > l_{max}$

After convergence and because $\mathbf{B}$ is subject to $\mathbf{B}^t\,\mathbf{B} = \mathbf{I}$ the rotated scores and loadings can be calculated via $\mathbf{T}_r = \mathbf{T}_{\mathbf{sca}}\mathbf{B}$ and $\mathbf{P}_r = \mathbf{P}_{\mathbf{sca}}\mathbf{B}$. The common and individual scores ($\mathbf{T}_c$ and $\mathbf{T}_{d_i}$) and loadings ($\mathbf{P}_{c_i}$ and $\mathbf{P}_{d_i}$) are separated according to the target matrix ($\mathbf{P}^*$). The loadings can then be determined as specific subsets of the rotated loadings to calculate the terms from Eq. 1:

$$\mathbf{C}_1 = \mathbf{T}_c\mathbf{P}_{c_1}^t,\ \mathbf{C}_2 = \mathbf{T}_c\mathbf{P}_{c_2}^t,\ \mathbf{D}_1 = \mathbf{T}_{d_1}\mathbf{P}_{d_1}^t \text{and } \mathbf{D}_2 = \mathbf{T}_{d_2}\mathbf{P}_{d_2}^t$$

JIVE

1. $\mathbf{X} = [\mathbf{X}_1|\mathbf{X}_2]$
2. $\mathbf{X} = \mathbf{U}_{(c_c)}\mathbf{S}_{(c_c)}\mathbf{V}^t_{(c_c)}$
3. $\mathbf{P}_c = \mathbf{V}_c\mathbf{S}_c$
4. $\mathbf{C} = \mathbf{T}_c\mathbf{P}_c^t$
5. $\mathbf{R}_k = \mathbf{X}_k - \mathbf{C}_k$
6. $\mathbf{R}_k - \mathbf{T}_{c_k}\mathbf{T}_{c_k}^t\mathbf{R}_k = \mathbf{U}_{d_k(c_k)}\mathbf{S}_{d_k(c_k)}\mathbf{V}^t_{d_k(c_k)}$
7. $\mathbf{T}_{d_k} = \mathbf{U}_{d_k}$
8. $\mathbf{P}_{d_k} = \mathbf{V}_{d_k}\mathbf{S}_{d_k}$
9. $\mathbf{D}_k = \mathbf{T}_{d_k}\mathbf{P}_{d_k}^t$
10. $\mathbf{X} = \mathbf{X} - [\mathbf{D}_1|\mathbf{D}_2]$
11. Repeat steps **2** through **11** until convergence of $\mathbf{C} + \mathbf{D}$, where $\mathbf{C} = [\mathbf{C}_1|\mathbf{C}_2]$ and $\mathbf{D} = [\mathbf{D}_1|\mathbf{D}_2]$

O2-PLS

Slightly different implementations were published which leaves room for possible different interpretations. In our implementation we followed the pseudo code described by Löfstedt et al ([20, 26, 27]) and made sure that our O2-PLS results corresponded to the 2 data-set OnPLS results.

1. $\mathbf{X}_2^t\boldsymbol{X}_1 = \mathbf{P}_{c_1(c_c)}\mathbf{D}_{(c_c)}\mathbf{P}^t_{c_2(c_c)}$
2. Intialize iteration index $l = 1$
3. $\mathbf{T}_{c_k} = \mathbf{X}_k\mathbf{P}_k$
4. $\mathbf{R}_k = \mathbf{X}_k - \mathbf{T}_{c_k}\mathbf{P}_{c_k}^t$
5. $\mathbf{R}_k^t\mathbf{T}_{c_k} = \mathbf{u}_{d_k(1)}\mathbf{s}_{d_k(1)}\mathbf{v}^t_{d_k(1)}$
6. $\mathbf{t}_{d_k,l} = \mathbf{X}_k\mathbf{u}_{d_k(1)}$
7. $\mathbf{p}_{d_k,l} = \left(\mathbf{t}_{d_k,l}^t\mathbf{t}_{d_k,l}\right)^{-1}\mathbf{X}_k{}^t\mathbf{t}_{d_k,l}$
8. $\mathbf{X}_k = \mathbf{X}_k - \mathbf{t}_{d_k,l}\mathbf{p}_{d_k,l}^t$
9. Repeat steps 3 through 8 for the number of distinctive components per data-set ($l = 1..c_k$) and both data-sets ($k = 1..2$).
10. $\mathbf{T}_{d_k} = \left[\mathbf{t}_{d_k,1}|\mathbf{t}_{d_k,2}|\cdots|\mathbf{t}_{d_k,c_k}\right]$
11. $\mathbf{P}_{d_k} = \left[\mathbf{p}_{d_k,1}|\mathbf{p}_{d_k,2}|\cdots|\mathbf{p}_{d_k,c_k}\right]$
12. $\mathbf{T}_{c_k} = \mathbf{X}_k\mathbf{P}_{c_k}$

After these steps have been performed the elements from Eq. 1 can be calculated via

$$\mathbf{C}_1 = \mathbf{T}_{c_1}\mathbf{P}_{c_1}^{\mathrm{t}},\ \mathbf{C}_2 = \mathbf{T}_{c_2}\mathbf{P}_{c_2}^{\mathrm{t}},\ \mathbf{D}_1 = \mathbf{T}_{d_1}\mathbf{P}_{d_1}^{\mathrm{t}} \text{ and } \mathbf{D}_2 = \mathbf{T}_{d_2}\mathbf{P}_{d_2}^{\mathrm{t}} \quad (9)$$

Observations on JIVE, SCA and covariance

If there is no distinctive information SCA and O2-PLS give the same result:

1. $\mathbf{X}_1 = \mathbf{T}_c\mathbf{P}_1^t + \mathbf{E}_1$ and $\mathbf{X}_2 = \mathbf{T}_c\mathbf{P}_2^t + \mathbf{E}_2$

Without noise these equations reduce to:

2. $\mathbf{X}_1 = \mathbf{T}_c\mathbf{P}_1^t$ and $\mathbf{X}_2 = \mathbf{T}_c\mathbf{P}_2^t$
3. $\mathbf{X}_2^t\mathbf{X}_1 = (\mathbf{T}_c\mathbf{P}_2^t)^t(\mathbf{T}_c\mathbf{P}_1^t) = \mathbf{P}_2(\mathbf{T}_c^t\mathbf{T}_c)\mathbf{P}_1^t$

$\mathbf{T}_c$ can be chosen such that $\mathbf{T}_c^t\mathbf{T}_c = \mathbf{I}$ so consequently:

4. $\mathbf{X}_2^t\mathbf{X}_1 = \mathbf{P}_2\mathbf{P}_1^t$

The analysis of the covariance matrix therefor will generate the same result if and only if there is no distinctive variation. This principle likely can be extended (no proof given) to those cases where the common variation is larger than the distinctive variation.

If there is a distinctive part it can be shown that an svd on the covariance matrix is less susceptible to larger distinctive parts and will better identify the common variation than the SCA approach used in JIVE.

1. $\mathbf{X}_1 = \mathbf{C}_1 + \mathbf{D}_1$ and $\mathbf{X}_2 = \mathbf{C}_2 + \mathbf{D}_2$
2. $\mathbf{X}_1^t\mathbf{X}_2 = (\mathbf{C}_1 + \mathbf{D}_1)^t(\mathbf{C}_2 + \mathbf{D}_2) = \mathbf{C}_1^t\mathbf{C}_2 + \mathbf{D}_1^t\mathbf{D}_2$

Because the common variation is expected to occur in both data-sets their cross product is expected to be larger than the crosspoduct of the distinctive parts that by definition should show less correlation between the subjects in both data-sets. The crossproducts of the distinctive and common parts can be neglected in comparison because of the orthogonality constraints (in O2-PLS).

Additional file

Additional file 1: Supplementary data. This file contains supplementary Tables S1-S4 and Figures S1-S7. (DOCX 735 kb)

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

FK and PBL contributed to the software development in Matlab (and R). FK performed all calculations and analyses. FK and JW wrote the manuscript. AC focused on the biological interpretation and AS on the statistical interpretation. All authors read and approved the final manuscript.

Acknowledgements

This work was funded from STATegra the Seventh Framework Programme [FP7/2007-2013] under grant agreement №306000.

Declarations

Publication charges for this work were funded by the STATegra the Seventh Framework Programme [FP7/2007-2013] under grant agreement №306000." This article has been published as part of BMC Bioinformatics Volume 17 Supplement 5, 2016: Selected articles from Statistical Methods for Omics Data Integration and Analysis 2014. The full contents of the supplement are available online at http://bmcbioinformatics.biomedcentral.com/articles/supplements/volume-17-supplement-5.

Author details

[1]Biosystems Data Analysis, Swammerdam Institute for Life Sciences, University of Amsterdam, Science Park 904, 1098 XH, Amsterdam, The Netherlands. [2]Computational Genomics Program, Centro de Investigaciones Príncipe Felipe, Valencia, Spain.

Published: 6 June 2016

References

1. Garg N, Kapono CA, Lim YW, Koyama N, Vermeij MJA, Conrad D, Rohwer F, Dorrestein PC. Mass spectral similarity for untargeted metabolomics data analysis of complex mixtures. Int J Mass Spectrom. 2014;377(MS 1960 to now):719–27.
2. Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning. 2nd edition. New York: Springer; 2008.
3. Bollen KA. Latent Variables In Psychology And The Social Sciences. Annu Rev Psychol. 2002;53:605–34.
4. Joliffe I, Morgan B. Principal component analysis and exploratory factor analysis. Stat Methods Med Res. 1992;1:69–95.
5. De Roover K, Ceulemans E, Timmerman ME, Vansteelandt K, Stouten J, Onghena P. Clusterwise simultaneous component analysis for analyzing structural differences in multivariate multiblock data. Psychol Methods. 2012;17:100–19.
6. Tan CS, Salim A, Ploner A, Lehtiö J, Chia KS, Pawitan Y. Correlating gene and protein expression data using Correlated Factor Analysis. BMC Bioinformatics. 2009;10:272.
7. Xiao X, Moreno-Moral A, Rotival M, Bottolo L, Petretto E. Multi-tissue analysis of co-expression networks by higher-order generalized singular value decomposition identifies functionally coherent transcriptional modules. PLoS Genet. 2014;10:e1004006.
8. Berger JA, Hautaniemi S, Mitra SK, Astola J. Jointly analyzing gene expression and copy number data in breast cancer using data reduction models. IEEE/ACM Trans Comput Biol Bioinforma. 2006;3:2–16.
9. Ponnapalli SP, Saunders MA, van Loan CF, Alter O. A higher-order generalized singular value decomposition for comparison of global mRNA expression from multiple organisms. PLoS One. 2011;6:e28072.
10. Alter O, Brown PO, Botstein D. Generalized singular value decomposition for comparative analysis of genome-scale expression data sets of two different organisms. Proc Natl Acad Sci. 2003;100:3351–6.
11. Van Deun K, Van Mechelen I, Thorrez L, Schouteden M, De Moor B. DISCO-SCA and Properly Applied GSVD as Swinging Methods to Find Common and Distinctive Processes. PLoS One. 2012;7:e37840.
12. Lock EF, Hoadley KA, Nobel AB. Supplement Joint and individual variation explained (JIVE) for integrated analysis of multiple data types. Ann Appl Stat. 2013;7(Supplement):1–11.
13. Lock EF, Hoadley KA, Marron JS, Nobel AB. Joint and individual variation explained (JIVE) for integrated analysis of multiple data types. Ann Appl Stat. 2013;7:523–42.
14. Schouteden M, Van Deun K, Wilderjans TF, Van Mechelen I. Performing DISCO-SCA to search for distinctive and common information in linked data. Behav Res Methods. 2013;46:576–87.
15. Van Deun K, Smilde AK, Thorrez L, Kiers HAL, Van Mechelen I. Identifying common and distinctive processes underlying multiset data. Chemom Intell Lab Syst. 2013;129:40–51.
16. Trygg J, Wold S. O2-PLS, a two-block (X-Y) latent variable regression (LVR) method with an integral OSC filter. J Chemom. 2003;17:53–64.
17. Trygg J. O2-PLS for qualitative and quantitative analysis in multivariate calibration. J Chemom. 2002;16:283–93.

18. The Cancer Genome Atlas [http://cancergenome.nih.gov]
19. Mathworks Inc. Matlab. 2013.
20. Löfstedt T, Trygg J. OnPLS-a novel multiblock method for the modelling of predictive and orthogonal variation. J Chemom. 2011;25:441–55.
21. Bylesjö M, Eriksson D, Kusano M, Moritz T, Trygg J. Data integration in plant biology: the O2PLS method for combined modeling of transcript and metabolite data. Plant J. 2007;52:1181–91.
22. Stanimirova I, Michalik K, Drzazga Z, Trzeciak H, Wentzell PD, Walczak B. Interpretation of analysis of variance models using principal component analysis to assess the effect of a maternal anticancer treatment on the mineralization of rat bones. Anal Chim Acta. 2011;689:1–7.
23. Hassani S, Hanafi M, Qannari EM, Kohler A. Deflation strategies for multi-block principal component analysis revisited. Chemom Intell Lab Syst. 2013;120:154–68.
24. Westerhuis JA, Smilde AK. Deflation in multiblock PLS. J Chemom. 2001;15(June 2000):485–93.
25. Smilde AK, Kiers HAL, Bijlsma S, Rubingh CM, van Erk MJ. Matrix correlations for high-dimensional data: the modified RV-coefficient. Bioinformatics. 2009;25:401–5.
26. Löfstedt T. OnPLS : Orthogonal projections to latent structures in multiblock and path model data analysis. Phd Thesis. Umeå universitet; 2012.
27. Löfstedt T, Hoffman D, Trygg J. Global, local and unique decompositions in OnPLS for multiblock data analysis. Anal Chim Acta. 2013;791(June 2012):13–24.