Received: 1 November 2016 | Revised: 30 January 2017 | Accepted: 14 March 2017

DOI: 10.1002/cem.2900

RESEARCH ARTICLE

WILEY Journal of CHEMOMETRICS

# Common and distinct components in data fusion

Age K. Smilde[1] | Ingrid Måge[2] | Tormod Næs[2] | Thomas Hankemeier[3] | Mirjam Anne Lips[4] | Henk A. L. Kiers[5] | Ervim Acar[6] | Rasmus Bro[6]

[1]Biosystems Data Analysis, Faculty of Sciences, University of Amsterdam, Science Park 904 Amsterdam, 1090 GE, The Netherlands

[2]Nofima, Ås, Norway

[3]Lacdr, Leiden University, Leiden, The Netherlands

[4]Department of Endocrinology and Metabolism, Leiden University Medical Center, Leiden, The Netherlands

[5]Heymans Institute, University of Groningen, Groningen, The Netherlands

[6]Department of Food Science, University of Copenhagen, Copenhagen, Denmark

**Correspondence**
Biosystems Data Analysis, Faculty of Sciences, University of Amsterdam, Science Park 904, PO Box 94215, 1090 GE, Amsterdam, the Netherlands.
Email: a.k.smilde@uva.nl

In many areas of science, multiple sets of data are collected pertaining to the same system. Examples are food products that are characterized by different sets of variables, bioprocesses that are online sampled with different instruments, or biological systems of which different genomic measurements are obtained. Data fusion is concerned with analyzing such sets of data simultaneously to arrive at a global view of the system under study. One of the upcoming areas of data fusion is exploring whether the data sets have something in common or not. This gives insight into common and distinct variation in each data set, thereby facilitating understanding of the relationships between the data sets. Unfortunately, research on methods to distinguish common and distinct components is fragmented, both in terminology and in methods: There is no common ground that hampers comparing methods and understanding their relative merits. This paper provides a unifying framework for this subfield of data fusion by using rigorous arguments from linear algebra. The most frequently used methods for distinguishing common and distinct components are explained in this framework, and some practical examples are given of these methods in the areas of medical biology and food science.

**KEYWORDS**
DISCO, GSVD, JIVE, O2PLS

## 1 | INTRODUCTION AND MOTIVATION

### 1.1 | Data fusion

Simultaneous analysis of several data blocks has been proposed a long time ago,[1,2] but today we can see a renewed interest fueled by the strongly increasing needs in many sciences. A number of different methods have been put forward,[3-7] all of them with a common interest of either understanding relations better or obtaining better prediction results. The methodologies are known under different names in different disciplines, important examples being data fusion, data integration, multiblock analysis, multiset analysis, and multimode analysis (for definitions, see the work of Van Mechelen and Smilde[8] and Lahat et al[9]). Some of the methods are rather straightforward generalizations of standard methods for 1 or 2 data sets such as concatenated principal components analysis (PCA) and partial least squares (PLS) regression,[10] while others are explicitly developed for handling multiblock data focusing on a number of concepts unique for such applications. In the latter group, one can find methods such as sequentially and parallel orthogonalized PLSs,[11] distinct and common simultaneous component analysis (DISCO-SCA, or DISCO for short),[12] Orthogonal 2-block PLS (O2PLS),[4] and generalized singular value decomposition (GSVD).[13]

This paper will focus on one particular aspect that appears crucial in data fusion, namely, the distinction between common and distinct information in the blocks. The main aim is to provide motivation and concrete definitions of the concepts that are still lacking in the literature and to discuss how these definitions relate to the most well-known methods in the multiblock area such as DISCO, joint and individual

  

variances explained (JIVE), and On-PLS. A conceptual approach with a focus on understanding will be taken without any ambition of making a full theoretical and empirical comparison between the methods. Main attention will be given to interchangeable data blocks sharing the row mode, which usually consists of samples or subjects; thus, multiblock predictive or supervised methods such as sequentially and parallel orthogonalized PLSs[11] are not discussed. Selected methods will be illustrated by real data sets. The main aim of the examples is to illustrate typical applications and how the concepts of common and distinct components can shed light on relations between multiblock data sets. To our best knowledge, there are only a limited number of papers discussing and comparing several methods for distinguishing common and distinct variation.[5,14-16] Our paper differs in several aspects: (1) we present a general mathematical framework and (2) we present more properties of the methods.

## 1.2 | Motivating examples

### 1.2.1 | Food science

In food product development, we are typically interested in understanding how product formulations (eg, ingredients) of a set of product prototypes are related to the descriptive sensory properties of the product and also possibly to the consumer liking of the product. A typical situation is in new product development where the developer wants to understand how two important sensory modalities such as smell and taste are affected by the ingredients used. It is crucial for further product optimization to know how this happens, for instance, whether the smell and taste have a joint source of variability and/or what is influencing only one of them.

### 1.2.2 | Biology

An important class of health problems is Type II diabetes mellitus (DM2). Consider measurements performed on DM2 patients using a metabolomics platform (eg, liquid chromatography–mass spectrometry), clinical measurements (such as insulin resistance, fasting glucose levels, and blood pressure), and lifestyle variables. Then these measurements will have common and distinctive parts. The common part between the metabolomics and clinical measurements may reflect the relation between branched amino acids and insulin resistance[17]; there may also be common parts between the lifestyle variables and the clinical measurements, such as exercise and blood pressure. Some of the metabolites, such as bile acids, may not be directly related to insulin resistance and lifestyle and will, hence, be distinct. Since all measurements pertain to the same system (DM2), it is worthwhile exploring and understanding the complete data set in a holistic way.

### 1.2.3 | General idea

The above two examples show common features, which are summarized in Figure 1. Knowledge is required of a complex system (first and upper layers; eg, DM2). Measurements are performed on this system, resulting in 3 blocks of data: $\mathbf{X}_1$, $\mathbf{X}_2$, and $\mathbf{X}_3$ (second layer; eg, metabolomics, clinical, and lifestyle measurements in the DM2 example; smell, taste, and consumer liking in the food science example).

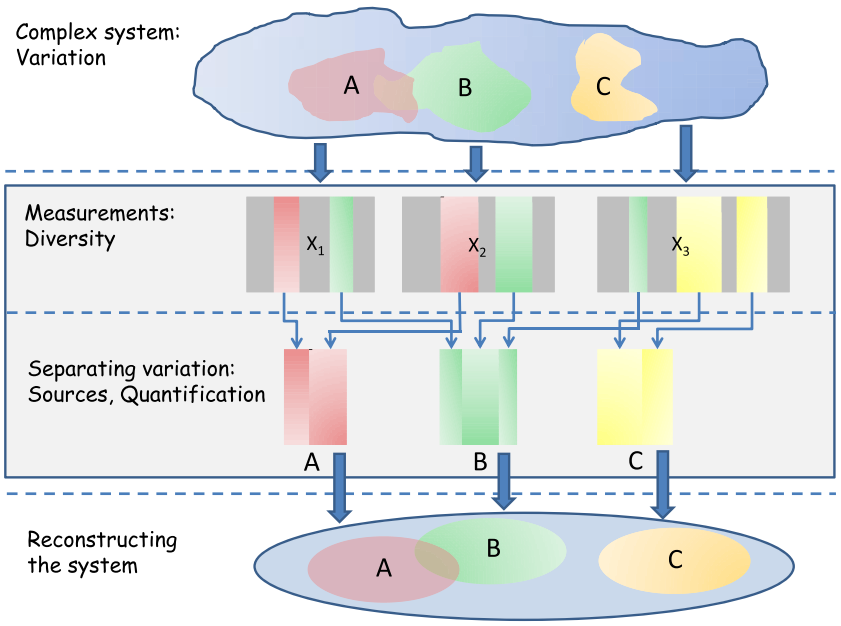


**FIGURE 1** Measurements are performed on a complex system probing parts A, B, and C of that system. The resulting data blocks $\mathbf{X}_1$, $\mathbf{X}_2$, and $\mathbf{X}_3$ contain mixed variation, which has to be separated in sources (red/green is common; yellow is distinct; grey is irrelevant variation and noise). These quantified sources are then used to reconstruct the system. This paper concerns mainly the box within the blue lines

These measurements are preferably collected in such a way that *diversity* is increased.[9,18] Although diverse and information-rich data are obtained, the problem is that the data blocks contain partly overlapping contributions of parts A, B, and C of the system (eg, A is insulin/glucose/amino acid metabolism and B reflects cardiovascular complications in the DM2 example; sweetness [A] in the case of taste and smell in the food science example) and also irrelevant variation and noise. The idea behind finding common and distinct variation in the three data blocks is to separate and quantify the different sources of variation that are spread across all data blocks (third layer). Interpreting the different sources of variation will then lead to a reconstruction of the system (fourth and bottom layers; eg, the etiology of DM2). In our paper, we will mainly describe moving from the second to third layers (the boxed part) and will only touch upon moving from the third to fourth layers. In Section 4, we will present some real-life examples, which were already introduced above.

## 2 | GENERAL MATHEMATICAL FRAMEWORK

For the definition of the basic concepts, we will start with 2 data matrices or blocks $\mathbf{X}_1$ of size $(I \times J_1)$ and $\mathbf{X}_2$ of size $(I \times J_2)$ and afterwards discuss how these concepts can be extended to 3 or more blocks of data. It is assumed that the 2 matrices share the first mode (the $I$ mode[19]) usually representing samples or objects and the data have been column centered throughout. Note that the two data sets may have different numbers of columns, usually representing variables, which means that they may (and often will) contain different types of measurements.

This section will be devoted to precise definitions of common and distinct components for the blocks in the data set. All these definitions are inspired by and related to previous definitions, but the main aim here is to make the definitions precise and unambiguous and therefore better suited for comparing methodologies. The definitions will be made in terms of subspaces, but later on we will expand to discuss the same concepts in terms of components that are basis vectors, chosen in one way or another, for the subspaces.

The mathematical framework represents the idealized situation of noiseless data. In practice, of course, this never happens. Hence, in later sections, we are also going to discuss which kind of compromises and choices have to be made in real-life situations. In that context, we also discuss several existing methods for finding common and distinct subspaces as used in the psychometrics, bioinformatics, chemometrics, computer science, data analysis, and statistics literature.

### 2.1 | Description of the framework

#### 2.1.1 | The two-block case

The two spaces spanned by the columns of $\mathbf{X}_1$ and $\mathbf{X}_2$ ($R[\mathbf{X}_1]$ and $R[\mathbf{X}_2]$) are located in the same $I$-dimensional column space $R^I$ (see Figure 2 for an illustration in 3-dimensional space). Each variable is a vector in this coordinate system, indicating the level of that variable for each sample (row). These variables are not explicitly shown in this figure but will lie within the space indicated by the blue and green column spaces.

If the 2 column spaces intersect nontrivially (the 0 is always shared), then the intersection space is called the common space. In Figure 2, there is only 1 common direction (ie, the common space is 1-dimensional), but there can be more or none. The common subspace, representing everything that is in common between $R(\mathbf{X}_1)$ and $R(\mathbf{X}_2)$, will be called $R(\mathbf{X}_{12C})$, where the subscript C stands for "common." Note that $R(\mathbf{X}_{12C}) \subseteqq R(\mathbf{X}_1)$ and $R(\mathbf{X}_{12C}) \subseteqq R(\mathbf{X}_2)$. The common part of the 2 blocks will in most cases not span the whole of $R(\mathbf{X}_1)$ and $R(\mathbf{X}_2)$. Some definitions regarding the rest of these spaces are therefore needed. As will be discussed later, it is useful to distinguish between different ways of representing these subspaces, depending on choices regarding orthogonality. In all cases, these subspaces representing the rest after identification of the common part will be called "distinct" subspaces. The requirement is that the space spanned by the columns in a block $\mathbf{X}_k (k = 1, 2)$ is a direct sum of the common space and the distinct space within that block. Hence, these 2 parts within a block are linearly independent (2 subspaces are linearly independent if no vector in 1 subspace can be

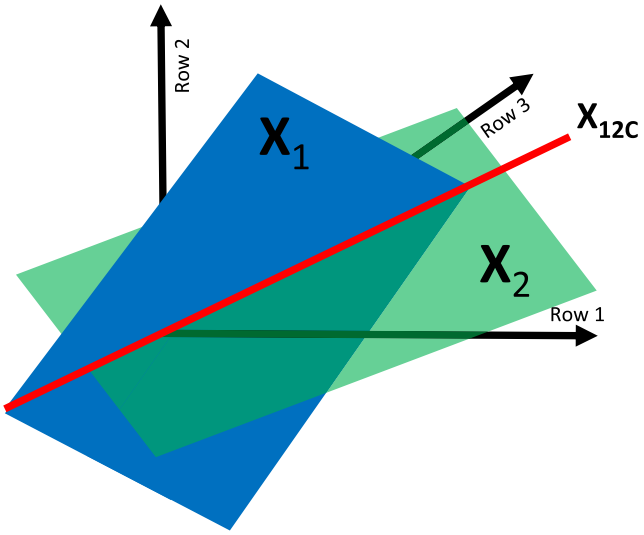


**FIGURE 2** The $I$-dimensional space having $R(\mathbf{X}_1)$ (blue) and $R(\mathbf{X}_2)$ (green) as subspaces. Only 3 axes of this $I$-dimensional space are drawn. The red line $\mathbf{X}_{12C}$ represents the common subspace. For the sake of illustration, the dimensions of both column spaces are equal (2). This is not necessarily always the case

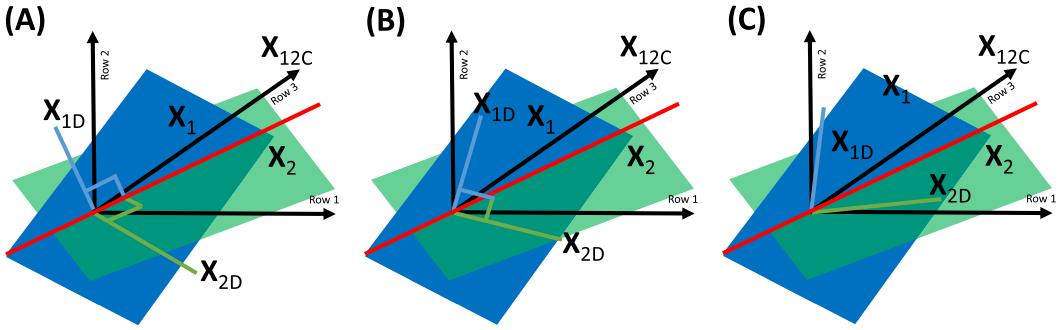


**FIGURE 3** See also the legend in Figure 2. The distinct subspaces (1-dimensional in this case) are spanned by $\mathbf{X}_{1D}$ and $\mathbf{X}_{2D}$ for $R(\mathbf{X}_1)$ and $R(\mathbf{X}_2)$, respectively. A, Both distinct subspaces are chosen orthogonal to the common subspace. B, Both distinct subspaces are chosen mutually orthogonal. C, No orthogonality

written as a linear combination of the vectors of the other and vice versa).

These subspaces are called $R(\mathbf{X}_{1D})$ and $R(\mathbf{X}_{2D})$ where the subscript D stands for "distinct." The choice of whether or not to choose orthogonality depends on the application and/or the model. In Figure 3, 3 possibilities are shown (namely, making the distinct subspaces orthogonal to the common subspace, making the distinct subspaces orthogonal to each other, or imposing no orthogonality at all). In general, it is not possible to combine the orthogonality properties of Figure 3A and B.

What we have accomplished now is decomposing $R(\mathbf{X}_1)$ and $R(\mathbf{X}_2)$ into direct sums of spaces:

$$\begin{aligned} R(\mathbf{X}_1) &= R(\mathbf{X}_{12C}) \oplus R(\mathbf{X}_{1D}) \\ R(\mathbf{X}_2) &= R(\mathbf{X}_{12C}) \oplus R(\mathbf{X}_{2D}) \end{aligned} \tag{1}$$

because $R(\mathbf{X}_{12C}) \cap R(\mathbf{X}_{1D}) = \{0\}$ and $R(\mathbf{X}_{12C}) \cap R(\mathbf{X}_{2D}) = \{0\}$.[20] Note that this implies that $R(\mathbf{X}_{1D}) \cap R(\mathbf{X}_{2D}) = \{0\}$ since $R(\mathbf{X}_{12C})$ represents everything that is in common by $R(\mathbf{X}_1)$ and $R(\mathbf{X}_2)$. Hence, it also holds that

$$\begin{aligned} \dim R(\mathbf{X}_1) &= \dim R(\mathbf{X}_{12C}) + \dim R(\mathbf{X}_{1D}) \\ \dim R(\mathbf{X}_2) &= \dim R(\mathbf{X}_{12C}) + \dim R(\mathbf{X}_{2D}) \end{aligned} \tag{2}$$

If the distinct orthogonal to common option is chosen (see Figure 3A), then additionally it holds that $R(\mathbf{X}_{12C}) \perp R(\mathbf{X}_{1D})$ and $R(\mathbf{X}_{12C}) \perp R(\mathbf{X}_{2D})$. Note that for this case, given the common space, the decomposition is unique because then $R(\mathbf{X}_{1D})$ is the orthogonal complement of $R(\mathbf{X}_{12C})$ within $R(\mathbf{X}_1)$ and likewise for $R(\mathbf{X}_{2D})$ (but not necessarily the basis within the subspaces if these have dimensions higher than 1). In the nonorthogonal case, the distinct part can be defined by any set of linearly independent vectors that are in the original spaces, but not in the common space. For a thorough description of direct sums of spaces, see the work of Yanai et al.[21]

We can take it one step further by also decomposing both the distinct subspaces $R(\mathbf{X}_{1D})$ and $R(\mathbf{X}_{2D})$ in 2 parts:

$$\begin{aligned} R(\mathbf{X}_{1D}) &= R(\mathbf{X}_{1DO}) \oplus R(\mathbf{X}_{1DNO}) \\ R(\mathbf{X}_{2D}) &= R(\mathbf{X}_{2DO}) \oplus R(\mathbf{X}_{2DNO}) \end{aligned} \tag{3}$$

where $R(\mathbf{X}_{1DO})$ is the "distinct orthogonal" (DO) part and the other part will be called "distinct nonorthogonal" (DNO), where $R(\mathbf{X}_{1DO}) \perp R(\mathbf{X}_{2DO})$ and $R(\mathbf{X}_{1DNO})$ is the remaining part of $R(\mathbf{X}_{1D})$ after removing $R(\mathbf{X}_{1DO})$ and likewise for $R(\mathbf{X}_{2DNO})$. Again, by the definition of direct sum, we have $R(\mathbf{X}_{1DO}) \cap R(\mathbf{X}_{1DNO}) = \{0\}$ and $R(\mathbf{X}_{2DO}) \cap R(\mathbf{X}_{2DNO}) = \{0\}$. The argument for the split of Equation 3 is that one may be interested in looking at the parts of the blocks that have no correlation with (parts of) each other at all. Note that such an additional split can only be performed when the dimensions of the subspaces allow so, eg, in Figure 3, both distinct subspaces have only dimension 1 and thus cannot be decomposed further. Depending on the dimensions of the distinct subspaces and their relative positioning in space, different possibilities can be distinguished. A choice has to be made by the user and is application dependent. A summary of alternatives is presented in the Appendix A.1, but one example is the following. If $\mathbf{X}_1$ and $\mathbf{X}_2$ contain measurements of 2 instruments, then choosing $R(\mathbf{X}_{2DO})$ orthogonal to the whole of $R(\mathbf{X}_1)$ can be interpreted as the unique contribution of instrument 2 relative to instrument 1 or, stated differently, what is the gain by adding instrument 2?

Summarizing, we arrive at the following direct sum decompositions of the column spaces of $\mathbf{X}_1$ and $\mathbf{X}_2$:

$$\begin{aligned} R(\mathbf{X}_1) &= R(\mathbf{X}_{12C}) \oplus R(\mathbf{X}_{1D}) = R(\mathbf{X}_{12C}) \oplus R(\mathbf{X}_{1DO}) \oplus R(\mathbf{X}_{1DNO}), \\ R(\mathbf{X}_2) &= R(\mathbf{X}_{12C}) \oplus R(\mathbf{X}_{2D}) = R(\mathbf{X}_{12C}) \oplus R(\mathbf{X}_{2DO}) \oplus R(\mathbf{X}_{2DNO}), \end{aligned} \tag{4}$$

representing our general definition of the basic concepts of common (C), distinct (D), distinct orthogonal (DO), and distinct nonorthogonal (DNO) subspaces. This decomposition is unique, meaning that when the decomposition of Equation 4 is chosen, then every vector in $R(\mathbf{X}_1)$ can be written uniquely as a sum of 3 vectors in the 3 different subspaces $R(\mathbf{X}_{12C})$, $R(\mathbf{X}_{1DO})$ and $R(\mathbf{X}_{1DNO})$ and likewise for $R(\mathbf{X}_2)$, if the dimensions allow so.

The decomposition of Equation 4 gives also a breakdown of the dimensions of the separate subspaces:

$$\begin{aligned}\dim R(\mathbf{X}_1) &= \dim R(\mathbf{X}_{12C}) + \dim R(\mathbf{X}_{1D}) \\ &= \dim R(\mathbf{X}_{12C}) + \dim R(\mathbf{X}_{1DO}) + \dim R(\mathbf{X}_{1DNO}) \\ \dim R(\mathbf{X}_2) &= \dim R(\mathbf{X}_{12C}) + \dim R(\mathbf{X}_{2D}) \\ &= \dim R(\mathbf{X}_{12C}) + \dim R(\mathbf{X}_{2DO}) + \dim R(\mathbf{X}_{2DNO}).\end{aligned} \tag{5}$$

### 2.1.2 | Generalizations to three blocks

The generalization to 3 blocks of data goes as follows. Consider the sets $\mathbf{X}_1(I \times J_1)$, $\mathbf{X}_2(I \times J_2)$ ,and $\mathbf{X}_3(I \times J_3)$. We can define again a part that is in common between all 3 column spaces, $R(\mathbf{X}_{123C})$ with obvious notation. Next, we can define a part in common between $R(\mathbf{X}_1)$ and $R(\mathbf{X}_2)$, which is not intersecting with $R(\mathbf{X}_3)$, $R(\mathbf{X}_{12C})$, and likewise, we can define $R(\mathbf{X}_{13C})$ and $R(\mathbf{X}_{23C})$. The complete part of $R(\mathbf{X}_1)$, which is shared with the other blocks, can then be written as $R(\mathbf{X}_{123C}) \oplus R(\mathbf{X}_{12C}) \oplus R(\mathbf{X}_{13C})$ with the properties that $R(\mathbf{X}_{123C}) \cap R(\mathbf{X}_{12C}) = \{0\}$, $R(\mathbf{X}_{123C}) \cap R(\mathbf{X}_{13C}) = \{0\}$ and $R(\mathbf{X}_{12C}) \cap R(\mathbf{X}_{13C}) = \{0\}$.

The distinct part of $R(\mathbf{X}_1)$ can again be defined as the part of $R(\mathbf{X}_1)$ linearly independent of $R(\mathbf{X}_{123C}) \oplus R(\mathbf{X}_{12C}) \oplus R(\mathbf{X}_{13C})$. This leads to the following decomposition:

$$R(\mathbf{X}_1) = R(\mathbf{X}_{123C}) \oplus R(\mathbf{X}_{12C}) \oplus R(\mathbf{X}_{13C}) \oplus R(\mathbf{X}_{1D}), \tag{6}$$

and the distinct part $R(\mathbf{X}_{1D})$ can again be broken down into several parts. The first part may be chosen to be the subspace of $R(\mathbf{X}_{1D})$ orthogonal to $R(\mathbf{X}_2) \cup R(\mathbf{X}_3)$ with obvious notation $R(\mathbf{X}_{1DO23})$. Then there is a part orthogonal to only $R(\mathbf{X}_2)$, $R(\mathbf{X}_{1DO2})$, and a part only orthogonal to only $R(\mathbf{X}_3)$, $R(\mathbf{X}_{1DO3})$, where again $R(\mathbf{X}_{1DO23}) \cap R(\mathbf{X}_{1DO2}) = \{0\}$ and $R(\mathbf{X}_{1DO23}) \cap R(\mathbf{X}_{1DO3}) = \{0\}$. Hence, the full decomposition of $R(\mathbf{X}_1)$ becomes

$$\begin{aligned}R(\mathbf{X}_1) = R(\mathbf{X}_{123C}) \oplus R(\mathbf{X}_{12C}) \oplus R(\mathbf{X}_{13C}) \oplus R(\mathbf{X}_{1DO23}) \\ \oplus R(\mathbf{X}_{1DO2}) \oplus R(\mathbf{X}_{1DO3}) \oplus R(\mathbf{X}_{1DNO}),\end{aligned} \tag{7}$$

which represents the most elaborate decomposition of $R(\mathbf{X}_1)$ if all dimensions allow so with different possibilities for orthogonalities. Because of the direct sum properties, the dimensions add up in the same way as in Equations 2 and 5. Similar decompositions can be made for $R(\mathbf{X}_2)$ and $R(\mathbf{X}_3)$. Schematically, the decomposition of Equation 7 is shown in Figure 4.

Equations 4, 6, and 7 show an increasing degree of complexity. We give here the full decompositions to be complete, but it is important to mention that in most practical cases, one is not interested in all these subspaces, making the actual practical decomposition simpler. This is even more so in cases with more than 3 blocks.

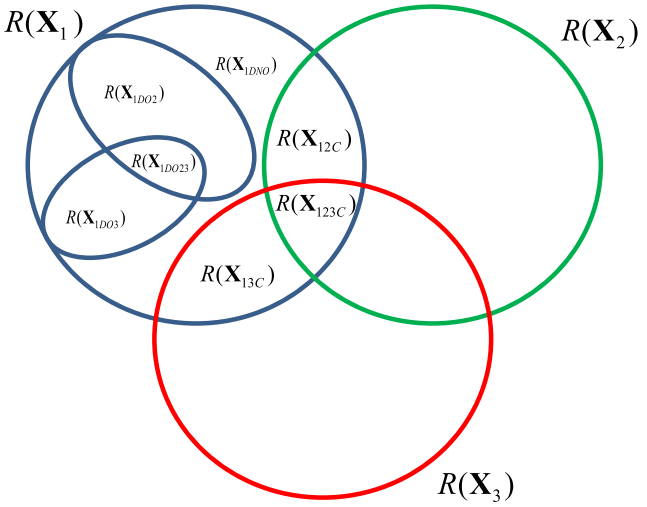


**FIGURE 4** The decomposition of $R(\mathbf{X}_1)$ in the 3-block situation

## 2.2 | Theoretical considerations

In practice, data always contain noise, and therefore, we cannot expect to find a decomposition that satisfies all the idealistic requirements described above. Also, which decomposition to make under which constraints depends very much on the type of application. Before showing how various types of already existing methods try to solve this challenge, we will here discuss some of the major issues that have to be taken into account. These issues represent choices that have to be made regarding the nature of the common and distinct components, diagnostic tools such as explained sum of squares (SS), scaling of the variables, and the data sets.

### 2.2.1 | Fundamentally different choices of common components

Of particular importance here is the concept of common variation because it can be considered as a starting point of the decomposition. Since practical implementations are usually based on extracting components or basis vectors for the different spaces, most of the following discussions will be related to components rather than to general vector spaces as was the case above.

In noisy data, the situation as shown in Figure 2 does not usually hold: There is no common space in mathematical terms (an intersection) because the column spaces have changed because of the noise. There are two fundamentally different categories of approaches, and these are present in the methods that are discussed in Section 3. In the first category, a common component is found as the best compromise solution between the 2 column spaces: Vector $\mathbf{X}_{12C}$ in Figure 5A (although it is customary to use a bold lowercase character for a vector, we keep the notation using a matrix to stress the fact that we are generally discussing subspaces). This vector is neither in the column space of $\mathbf{X}_1$ nor in the column space

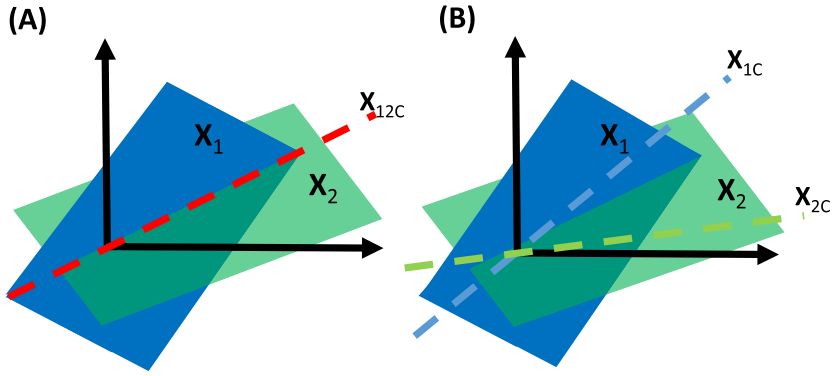


**FIGURE 5** The common subspace under noisy conditions. It can be chosen as a compromise in-between $R(\mathbf{X}_1)$ and $R(\mathbf{X}_2)$ but not a part of either of those (red dashed line, A) or as parts of their respective column spaces but unequal to each other (blue and green lines, B)

of $\mathbf{X}_2$. In the second category, a different choice is made. The common component is estimated separately in each column space. Hence, rather than 1 common component, 2 separate ones are found but generally in a manner that seeks them to be as similar as possible (although $\mathbf{X}_{1C} \neq \mathbf{X}_{2C}$; Figure 5B), and thus, they can be seen as representing a common component. Both choices are made in the methods to be discussed, and both approaches have their pros and cons. Note the change in notation of the common parts to emphasize this difference.

### 2.2.2 | Sums of squares and explained variation

When it comes to assessing the importance of a subspace in the decomposition, there are at least 2 aspects that have to be taken into account: the dimension of the subspaces identified and variances explained by those subspaces in the original data. The former relates to how many linearly independent components are estimated to form the subspace. The latter relates to the contributions of the subspaces to the total variation in a block. If orthogonality is used in the decomposition when defining the distinct space (Figure 3A), it is easy to show that the total SS for a block can be split in 1 contribution from the common space ($R[\mathbf{X}_{12C}]$) and 1 for the orthogonal distinct contribution ($R[\mathbf{X}_{1D}]$):

$$\|\mathbf{X}_1\|^2 = \|\mathbf{X}_{12C}\|^2 + \|\mathbf{X}_{1D}\|^2, \tag{8}$$

where we use the symbol $\|\cdot\|^2$ to indicate the squared Frobenius norm of a matrix. An analogous equation can be written for $\mathbf{X}_2$. If orthogonality is not imposed between the common and distinct parts (Figure 3B), a decomposition of SS is still possible, but the interpretation of the last term is different. In that case, it is simply defined as the additional variation that is explained by adding the distinct components, ie, as $\|\mathbf{X}_1\|^2 = \|\mathbf{X}_{12C}\|^2 + \|\tilde{\mathbf{X}}_{1D}\|^2$, where $\tilde{\mathbf{X}}_{1D}$ is the part of $R(\mathbf{X}_{1D})$ orthogonal to $R(\mathbf{X}_{12C})$. This is sometimes called extra SS (ESS; see also the work of Peres-Neto et al[22]). Note that the order in which the terms are calculated in Equation 8 matters in the nonorthogonal case. For the orthogonal case, the 2 interpretations coincide.

The issue of variance explained by common components in a block is visualized in Figure 6 for the second category of methods. The column vectors making up the column spaces of $\mathbf{X}_1$ and $\mathbf{X}_2$ are explicitly drawn in the figure. In Figure 6A, all these column vectors (ie, variables) are close to the common components within each block. Hence, the common components are representative of their respective column spaces: They are embedded well and explain a high amount of variation in each block. This is not the case for Figure 6B:

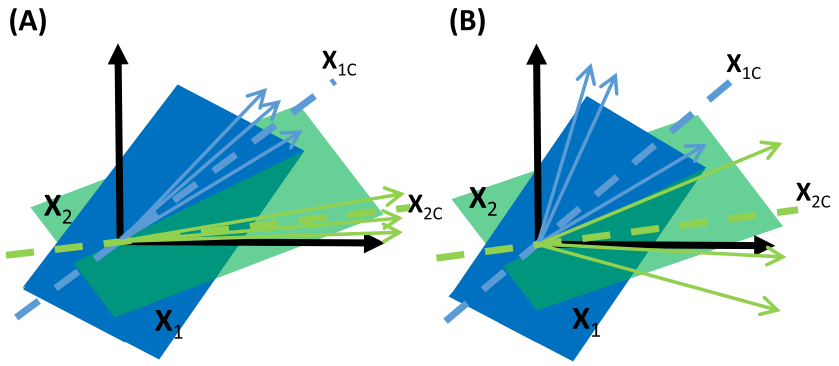


**FIGURE 6** A, Well-embedded common component. B, Poorly embedded common direction

The common component $\mathbf{X}_{1C}$ is not well embedded in $\mathbf{X}_1$. Usually, explaining within-block variation and having between-block correlation cannot be achieved simultaneously and a good account of this trade-off is given elsewhere.[2]

### 2.2.3 | High-dimensional data

High-dimensional data need some extra considerations. This type of data is abundant in modern scientific fields such as genomics, eg, when considering gene expression data where the number of genes (variables) is much larger than the number of samples. In our framework, there is now *necessarily* a common subspace simply due to the dimensions. For instance, if $\mathbf{X}_1$ has size $20 \times 1000$ and $\mathbf{X}_2$ has size $20 \times 10000$ both of rank 20, then they trivially share the same 20-dimensional column space, which is thus $R(\mathbf{X}_{12C})$. In such cases, calculating for instance canonical correlations is problematic and some type of regularization is necessary. Without such regularization, the chances are in most cases high that only uninteresting, trivial, and noisy components are identified. One way of trying to solve the problem with many variables and few objects is to use PCA for each block separately, in this way reducing both the noise and the dimensionality (see the work of Van den Berg et al[23] and Mage et al[24]). Whether this approach is preferable depends on a number of properties (ranks of the different subspaces, noise characteristics, etc).

### 2.2.4 | Scaling issues

Another general aspect that is important to discuss is the scaling of the blocks and variables within blocks. We will refer to this as between-block scaling and within-block scaling. One example of the latter is known as autoscaling in chemometrics, standardization in psychometrics, and normalization in statistics. We assumed already centered data, and autoscaling on top of that also divides every column of a matrix by its standard deviation. Hence, the data are analyzed in correlation mode. The between-block scaling is related to the total variation of a block. It is often natural to do some type of overall scaling of the blocks to avoid too much dominance of one of the blocks. For instance, in cases where one of the blocks has only a few variables and another one has many variables, the joint approach could put almost all emphasis on trying to model the larger data set. This may lead to solutions where one is not modeling the joint variation, but merely within-block variability, which is clearly not the intention in data fusion. A possible way to counter this is to divide each block by the Frobenius norm prior to analysis. General guidelines for centering and scaling are available,[25,26] and there is also literature on scaling in multiblock data analysis.[27-30]

## 3 | HOW ESTABLISHED METHODS RELATE TO THE DEFINITIONS

In this section, we will discuss how a number of already existing methods aiming for identifying common and distinct components are related to the definitions given in Section 2. We will discuss these methods mostly by using 2 blocks of data and more than 2 blocks if clarity allows us to do so (we will index the blocks by $k = 1, \ldots, K$). Tables 1 and 2 summarize properties of these discussed methods. A summary of the models underlying the presented methods is given in the column "Model" of Table 1. It appears that for the 2-block case the general model is as follows:

$$\begin{aligned} \mathbf{X}_1 &= \mathbf{X}_{1C} + \mathbf{X}_{1D} + \mathbf{E}_1, \\ \mathbf{X}_2 &= \mathbf{X}_{2C} + \mathbf{X}_{2D} + \mathbf{E}_2, \end{aligned} \tag{9}$$

with different properties of the matrices $\mathbf{X}_{kC}$, $\mathbf{X}_{kD}$, and $\mathbf{E}_k$. Some methods do not estimate $\mathbf{X}_{kD}$, and some methods do not distinguish between common and distinct components. Different choices are made regarding the positioning of the column spaces of $\mathbf{X}_{kC}$ and $\mathbf{X}_{kD}$ (see Table 2 under "Subspace properties"). Also, different (although sometimes implicit) choices are made regarding orthogonality (see Table 2 under "Orthogonality") resulting in differences in (E)SS. However, for none of the methods did a rigorous direct sum decomposition as in Equations 4 and 7 hold. The cases for more than 2 blocks show an even wider variety of possibilities. The choice of orthogonality constraints depends on the application.

The methods to be discussed originate from different fields of science and thus use different notations. We will try to harmonize this by using as much as possible a uniform notation based on the familiar PCA model:

$$\mathbf{X} = \mathbf{X}\mathbf{W}\mathbf{P}^T + \mathbf{E} = \mathbf{T}\mathbf{P}^T + \mathbf{E}, \tag{10}$$

where the matrix of weights $\mathbf{W}$ defines linear combinations of the columns of $\mathbf{X}$, generating scores $\mathbf{T}$ and loadings $\mathbf{P}$, which are the regression coefficients of $\mathbf{X}$ on $\mathbf{T}$. In PCA, the matrix $\mathbf{W}$ will be identical to $\mathbf{P}$, but this is not necessarily so for all methods. To arrive at a consistent terminology for all the methods to be discussed, we will use the terms and corresponding symbols weights, scores, and loadings in the following.

### 3.1 | Simultaneous component analysis, generalized canonical correlation, and a compromise

The two most different ways of defining common variability are probably PCA on the concatenated matrix $[\mathbf{X}_1|\cdots|\mathbf{X}_K]$, which focuses on explaining the simultaneous variation in all blocks, and canonical correlation or its generalized form (see below), which only focuses on correlation between the blocks.

**TABLE 1** Fundamental aspects of the fusion methods

| Methods | Model | Uniqueness |
|---|---|---|
| SCA | $\mathbf{X}_k = \mathbf{T}\mathbf{P}_k^T + \mathbf{E}_k$ or | |
| | $\mathbf{X}_k = \tilde{\mathbf{T}}_k\mathbf{P}_k^T + \mathbf{F}_k$ $(\tilde{\mathbf{T}}_k = \mathbf{X}_k\mathbf{X}_k^+\mathbf{T})$ | $R(\mathbf{T})$ is unique |
| GCA | $\mathbf{X}_k = \mathbf{X}_k\mathbf{W}_k\mathbf{P}_k^T + \mathbf{E}_k = \mathbf{T}_k\mathbf{P}_k^T + \mathbf{E}_k = \mathbf{X}_{kC} + \mathbf{E}_k$ | $R(\mathbf{T})$ is unique |
| O2PLS | $\mathbf{X}_1 = \mathbf{T}_{1C}\mathbf{W}_1^T + \mathbf{T}_{1D}\mathbf{P}_{1D}^T + \mathbf{E}_1 = \mathbf{X}_{1C} + \mathbf{X}_{1D} + \mathbf{E}_1$ | |
| | $\mathbf{X}_2 = \mathbf{T}_{2C}\mathbf{W}_2^T + \mathbf{T}_{2D}\mathbf{P}_{2D}^T + \mathbf{E}_2 = \mathbf{X}_{2C} + \mathbf{X}_{2D} + \mathbf{E}_2$ | ?? |
| DISCO | $\mathbf{X}_1 = \mathbf{T}_1\mathbf{P}_{11}^T + \mathbf{T}_2\mathbf{P}_{12}^T + \mathbf{T}_3\mathbf{P}_{13}^T + \mathbf{E}_1 = \mathbf{X}_{1C} + \mathbf{X}_{1DO} + \mathbf{X}_{1DNO} + \mathbf{E}_1$ | |
| | $\mathbf{X}_2 = \mathbf{T}_1\mathbf{P}_{21}^T + \mathbf{T}_2\mathbf{P}_{22}^T + \mathbf{T}_3\mathbf{P}_{23}^T + \mathbf{E}_2 = \mathbf{X}_{2C} + \mathbf{X}_{2DNO} + \mathbf{X}_{2DO} + \mathbf{E}_2$ | $R(\mathbf{T})$ is unique |
| GSVD | $\mathbf{X}_1 = \mathbf{T}\mathbf{D}_1\mathbf{V}_1^T + \mathbf{E}_1 = \mathbf{T}_1\mathbf{D}_{11}\mathbf{V}_{11}^T + \mathbf{T}_2\mathbf{D}_{12}\mathbf{V}_{12}^T + \mathbf{T}_3\mathbf{D}_{13}\mathbf{V}_{13}^T + \mathbf{E}_1 =$ | |
| | $\mathbf{X}_{1DO} + \mathbf{X}_{1C} + \mathbf{X}_{1DNO} + \mathbf{E}_1$ | |
| | $\mathbf{X}_2 = \mathbf{T}\mathbf{D}_2\mathbf{V}_2^T + \mathbf{E}_2 = \mathbf{T}_1\mathbf{D}_{21}\mathbf{V}_{21}^T + \mathbf{T}_2\mathbf{D}_{22}\mathbf{V}_{22}^T + \mathbf{T}_3\mathbf{D}_{23}\mathbf{V}_{23}^T + \mathbf{E}_2 =$ | |
| | $\mathbf{X}_{2DNO} + \mathbf{X}_{2C} + \mathbf{X}_{2DO} + \mathbf{E}_2$ | ?? |
| JIVE | $\mathbf{X}_k = \mathbf{T}(\mathbf{P}_{kC}^T) + \mathbf{T}_k(\mathbf{P}_{kD})^T + \mathbf{E}_k = \mathbf{X}_{kC} + \mathbf{X}_{kD} + \mathbf{E}_k$ | Subspaces unique |
| SRDF | $\mathbf{X}_k = \mathbf{T}\mathbf{D}_k\mathbf{V}_k^T + \mathbf{E}_k = \mathbf{X}_{kC} + \mathbf{X}_{kD} + \mathbf{E}_k$ | ?? |

Abbreviations: C, common; D, distinct; DISCO, distinct and common simultaneous component analysis; GCA, generalized canonical correlation analysis; JIVE, joint and individual variances explained; O2PLS, orthogonal 2-block PLS; SCA, simultaneous component analysis. Colors: brown is mixed subspace; green is common subspace; red/blue are distinct subspaces.

**TABLE 2** Some properties of the fusion methods

| Methods | Subspace Properties | Orthogonality |
|---|---|---|
| SCA | $R(\mathbf{T}) \subseteq R[\mathbf{X}_1\|\cdots\|\mathbf{X}_K]$; $R(\mathbf{T}) \subsetneq R[\mathbf{X}_k]$ ; $R(\tilde{\mathbf{T}}_k) \subseteq R(\mathbf{X}_k)$ | $\mathbf{T}^T\mathbf{T} = \mathbf{I}$ |
| GCA | $R(\mathbf{T}) \subseteq R[\mathbf{X}_1\|\cdots\|\mathbf{X}_K]$; $R(\mathbf{T}) \subsetneq R[\mathbf{X}_k]$; $R(\tilde{\mathbf{T}}_k) \subseteq R(\mathbf{X}_k)$ | $\mathbf{T}^T\mathbf{T} = \mathbf{I}$ |
| O2PLS | $R(\mathbf{X}_{kC}) \subseteq R(\mathbf{X}_k)$; $R(\mathbf{X}_{kD}) \subseteq R(\mathbf{X}_k)$ | $\mathbf{X}_{kC}^T\mathbf{X}_{kD} = 0$; $\mathbf{E}_k^T\mathbf{X}_{kD} = 0$; $\mathbf{E}_k^T\mathbf{X}_{kC} \neq 0$; |
| | | $\mathbf{X}_{kC}^T\mathbf{X}_{k'D} \neq 0 (k \neq k')$; $\mathbf{X}_{kD}^T\mathbf{X}_{k'D} \neq 0 (k \neq k')$ |
| DISCO | $R(\mathbf{T}) \subseteq R[\mathbf{X}_1\|\cdots\|\mathbf{X}_K] R(\mathbf{T}_l) \subsetneq R(\mathbf{X}_k); l = 1,2,3; \forall k$ | All orthogonal except: |
| | | $(\mathbf{X}_{1DO})^T\mathbf{X}_{2DNO} \neq 0$; |
| | | $(\mathbf{X}_{2DO})^T\mathbf{X}_{1DNO} \neq 0$ |
| GSVD | $R(\mathbf{T}) \subseteq R[\mathbf{X}_1\|\mathbf{X}_2]$; $R(\mathbf{T}_l) \subsetneq R(\mathbf{X}_k)$; $l = 1,2,3; k = 1,2$ | $\mathbf{T}^T\mathbf{T} \neq 0$; $\mathbf{X}_{1DO}$, $\mathbf{X}_{1DNO}$ and $\mathbf{X}_{1C}$ mutually orthogonal; |
| | | $\mathbf{X}_{2DO}$, $\mathbf{X}_{2DNO}$ and $\mathbf{X}_{2C}$ mutually orthogonal |
| JIVE | $R(\mathbf{T}) \subsetneq R(\mathbf{X}_k)$ ; $R(\mathbf{T}_k) \subseteq R(\mathbf{X}_k)$ ; $R(\mathbf{T}) \subseteq R[\mathbf{X}_1\|\cdots\|\mathbf{X}_K]$ | $(\mathbf{X}_{k'C})^T\mathbf{X}_{kD} = 0; \forall k, k'$; $(\mathbf{X}_{k'D})^T\mathbf{X}_{kD} \neq 0 (k \neq k')$ |
| SRDF | ?? | No orthogonality |

Abbreviations: DISCO, distinct and common simultaneous component analysis; GCA, generalized canonical correlation analysis; JIVE, joint and individual variances explained; O2PLS, orthogonal 2-block PLS; SCA, simultaneous component analysis.

### 3.1.1 | Simultaneous component analysis

The optimization criterion for simultaneous component analysis (SCA) is

$$\min_{(\mathbf{T},\mathbf{P}_k)} \sum_{k=1}^{K} \|\mathbf{X}_k - \mathbf{T}\mathbf{P}_k^T\|^2 \tag{11}$$

where the simultaneous components are represented by $\mathbf{T}(I \times R)$ and the loadings $\mathbf{P}_k(J_k \times R)$ measure how these components are related to the original data. This model is known under different names: SUM-PCA in chemometrics,[31] SCA-P in psychometrics,[32] and Tucker1 in 3-way analysis.[33] The underlying idea of using this model is that $\mathbf{T}$ represents as much as possible the variation in all data blocks simultaneously. Hence, a model of each block can be written as

$$\mathbf{X}_k = \mathbf{T}\mathbf{P}_k^T + \mathbf{E}_k; k = 1, \ldots, K, \tag{12}$$

and several properties of SCA are described in Tables 1 and 2. The optimization problem of Equation 11 is stated as a least squares problem but can also be formulated as the problem of finding the eigenvectors of

$$\mathbf{Z}_{\text{SCA}} = \sum_{k=1}^{K} \mathbf{X}_k\mathbf{X}_k^T \tag{13}$$

and selecting the $R$ eigenvectors belonging to the $R$ largest eigenvalues. Alternatively, the components $\mathbf{T}$ can be found using the singular value decomposition (SVD) of $[\mathbf{X}_1|\cdots|\mathbf{X}_K]$ and choosing the $R$ left singular vectors corresponding to the $R$ largest singular values (ie, a PCA on the concatenated matrix $[\mathbf{X}_1|\cdots|\mathbf{X}_K]$). Hence, the components $\mathbf{T}$ are in the column space of $[\mathbf{X}_1|\cdots|\mathbf{X}_K]$ and not necessarily in the column spaces of any of the individual matrices $\mathbf{X}_k$. The matrix $\mathbf{T}$ represents both common and distinct vari-

ations according to the definitions given above, and hence, the model is not separating common and distinct sources of variation like in the general model of Equation 9. This is emphasized in Table 1 by using the brown color. Moreover, the term *simultaneous* component analysis suggests a focus on common components, which is not the case. Nevertheless, we present SCA here since it is much used in multiset analysis and a starting point of other methods. Note that the least squares property does not hold per data block, but only across all blocks simultaneously. However, given $\mathbf{T}$, Equation 12 is a least squares model for the set of all $\mathbf{X}_k$.

Without loss of generality, the simultaneous components, $\mathbf{T}$, can be chosen to be orthogonal owing to the rotational freedom of the model. The subspace spanned by $\mathbf{T}$ is unique like in ordinary PCA. The residuals $\mathbf{E}_k$ are orthogonal to the model part of $\mathbf{X}_k$ (which is $\mathbf{T}\mathbf{P}_k^T$), and thus a breakdown of SS can be calculated. Note, however, that because $\mathbf{T}$ is not necessarily in the range of $\mathbf{X}_k$, neither is $\mathbf{E}_k$. Simultaneous component analysis is sensitive to between- and within-block scaling.

There is also a history of sequential methods in chemometrics, and these methods are known under different names and versions (hierarchical PCA, consensus PCA, and multiblock PCA). Because of their sequential nature, it is sometimes difficult to assess their properties, but some results exist.[10,31] Simultaneous component analysis has been used in several areas of science and is a special case of a much broader method in data mining called collective matrix factorization.[34] In metabolomics and process chemometrics, it is used in conjunction with multilevel data analysis and as a step after an initial analysis of variance.[35-37] It is also used in spectroscopy[38-41] and in sensory science.[42-44]

### 3.1.2 | Generalized canonical correlation analysis

The goal of generalized canonical correlation analysis (GCA) is to identify linear combinations of the blocks, $\mathbf{X}_k\mathbf{W}_k$, which fit as well as possible to a set of orthogonal common components $\mathbf{T}$. This is done by minimizing the criterion

$$\min_{(\mathbf{T},\mathbf{W}_k)} \sum_{k=1}^{K} \|\mathbf{X}_k\mathbf{W}_k - \mathbf{T}\|^2 \tag{14}$$

with respect to $\mathbf{T}(\mathbf{T}^T\mathbf{T} = \mathbf{I})$ and $\mathbf{W}_k(k = 1, \ldots, K)$.[45] The number of columns in $\mathbf{T}$, $A$, must be smaller than or equal to the number of columns in the $\mathbf{X}_k$ with the smallest number of columns. If the number of samples, $I$, is smaller than all $J_k(k = 1, \ldots, K)$, then $A = I$ is the maximum number of components. Note that the same solution can be obtained by maximizing a sum of correlations between linear combinations of the X blocks, which is the typical formulation for the situation with only 2 X blocks.[46] In that case, this is usually referred to as canonical correlation analysis. In practice, the actual solution $\mathbf{T}$ is found as the eigenvectors of the matrix

$$\mathbf{Z}_{\mathrm{GCA}} = \sum_{k=1}^{K} \mathbf{X}_k\left(\mathbf{X}_k^T\mathbf{X}_k\right)^{+}\mathbf{X}_k^T, \tag{15}$$

where the + means the Moore-Penrose (pseudo)inverse. The $\mathbf{W}_k$'s can then be found by regressing $\mathbf{T}$ on $\mathbf{X}_k$: $\mathbf{W}_k = \mathbf{X}_k^{+}\mathbf{T}$.

If there are $A$ common components in the X blocks according to the definition given in Section 2.1, the criterion in Equation 14 will exactly be equal to 0. In the 2-block case, the common components will correspond to components with a canonical correlation equal to 1.

The solution $\mathbf{T}$ in Equation 14 is not necessarily within any of the column spaces of the $\mathbf{X}_k$'s, but it is in the column space of $[\mathbf{X}_1|\cdots|\mathbf{X}_K]$ (for a proof, see the Appendix). Although this is not the goal of GCA, when needed, a model of $\mathbf{X}_k$ can be obtained by regressing $\mathbf{X}_k$ on $\mathbf{T}_k = \mathbf{X}_k\mathbf{W}_k$ giving loadings $\mathbf{P}_k$ from which also explained variances can be calculated (see Table 1).

Since GCA only concentrates on correlation and gives no emphasis on within-block variability (thereby potentially poorly embedded and hence unstable), several methods have been developed for balancing the two aspects. One particular solution is obtained by defining a continuum of solutions between SCA and GCA using a ridge regression type of formulation joining Equations 13 and 15 in one single formula.[43] Enhancing stability of the GCA components can also be obtained by using PCA on the individual data blocks before using GCA[23] or by regularization.[47] The solution using PCA as a first step will be called PCA-GCA in the example section. The GCA presented above only finds common components, which is indicated in Table 1 using the green color.

It is possible to also obtain distinct components using PCA-GCA. This can be done by regressing each block on its own common components. The residuals from these regressions represent 2 distinct subspaces $R(\mathbf{X}_{1\mathrm{D}})$ and $R(\mathbf{X}_{2\mathrm{D}})$, which can subsequently be subjected to a PCA for each subspace. Note that in this case $R(\mathbf{X}_{1\mathrm{D}})$ is orthogonal to $R(\mathbf{X}_{1\mathrm{C}})$ and likewise $R(\mathbf{X}_{2\mathrm{D}})$ is orthogonal to $R(\mathbf{X}_{2\mathrm{C}})$, but $R(\mathbf{X}_{1\mathrm{D}})$ is not necessarily orthogonal to $R(\mathbf{X}_{2\mathrm{D}})$. Hence, we are in the situation of Figure 3A ,and the resulting models fit in the general model of Equation 9. Generalized canonical correlation analysis does not depend on within-block and between-block scaling, and thus, the distinct subspaces also do not depend on that. However, performing a PCA on the distinct subspaces depends of course on the within scaling of the distinctive matrices. Examples of the use of GCA can be found, eg, in sensory science.[43,45] Also, in signal processing, GCA-type methods are used[48] based on the work in biometrics.[1]

## 3.2 | O2PLS (Orthogonal 2-block PLS)

There seem to be 3 different implementations of O2PLS.[4,49,50] The last implementation is a generalization of O2PLS to OnPLS (for more than 2 blocks). The O2(n)PLS methods are

usually described in terms of iterative algorithms rather than through formal definitions of well-defined criteria, which makes their properties difficult to assess. We describe the implementation of Lofstedt.[51]

The starting point for O2PLS is the SVD of the covariance matrix $\mathbf{X}_2^T\mathbf{X}_1$,

$$\mathbf{U}\mathbf{S}\mathbf{V}^T = \mathbf{X}_2^T\mathbf{X}_1, \tag{16}$$

and collecting the $R$ singular vectors corresponding to the $R$ largest singular values of Equation 16 in $\mathbf{U}_R$ (left-singular vectors) and $\mathbf{V}_R$ (right-singular vectors), respectively, as weights for the (preliminary) common components. This SVD is known as the product SVD and is a member of a broad class of generalizations of the ordinary SVD,[52,53] and Equation 16 is actually also the first step of Bookstein's version of PLS.[54] Define $\mathbf{F}_1 = \mathbf{X}_1 - \mathbf{X}_1\mathbf{V}_R\mathbf{V}_R^T$ and $\mathbf{F}_2 = \mathbf{X}_2 - \mathbf{X}_2\mathbf{U}_R\mathbf{U}_R^T$; then because of the truncation to $R$ components, the (preliminary) common components $\tilde{\mathbf{T}}_{1C} = \mathbf{X}_1\mathbf{V}_R$ still share some variation with $\mathbf{F}_1$ and likewise $\mathbf{X}_2\mathbf{U}_R$ with $\mathbf{F}_2$. This part can be calculated by solving

$$\max_{\|\mathbf{z}_1\|=1} \|\tilde{\mathbf{T}}_{1C}^T\mathbf{F}_1\mathbf{z}_1\|^2, \tag{17}$$

which maximizes the shared variation of $\tilde{\mathbf{T}}_{1C}$ and $\mathbf{F}_1$ in 1 (orthogonal) component (indexed by $l = 1, \ldots, L$). A deflation procedure then subsequently regresses $\mathbf{X}_1$ on this component $\mathbf{F}_1\mathbf{z}_1$ and gives the residuals $\mathbf{X}_{res1}$. A similar procedure can be used for $\mathbf{X}_2$, and the number of orthogonal components has to be chosen (or found). Then (final) common components between the deflated matrices can be extracted one by one by using the MAXDIFF criterion[55]:

$$\max_{\mathbf{w}_1,\mathbf{w}_2} \mathrm{tr}(\mathbf{w}_1^T\mathbf{X}_{res2}^T\mathbf{X}_{res1}\mathbf{w}_2) = \mathrm{tr}(\mathbf{t}_{2C}^T\mathbf{t}_{1C}); \mathbf{W}_k^T\mathbf{W}_k = \mathbf{I}(k = 1, 2), \tag{18}$$

where the matrices $\mathbf{T}_{1C}$ and $\mathbf{T}_{2C}$ collect the vectors $\mathbf{t}_{1C}$ and $\mathbf{t}_{2C}$, respectively, and the matrix $\mathbf{W}_k$, $k = 1, 2$, contains the weights for the different dimensions. This will result in the following models for $\mathbf{X}_1$ and $\mathbf{X}_2$:

$$\begin{aligned} \mathbf{X}_1 &= \mathbf{T}_{1C}\mathbf{W}_1^T + \mathbf{T}_{1D}\mathbf{P}_{1D}^T + \mathbf{E}_1 = \mathbf{X}_{1C} + \mathbf{X}_{1D} + \mathbf{E}_1, \\ \mathbf{X}_2 &= \mathbf{T}_{2C}\mathbf{W}_2^T + \mathbf{T}_{2D}\mathbf{P}_{2D}^T + \mathbf{E}_2 = \mathbf{X}_{2C} + \mathbf{X}_{2D} + \mathbf{E}_2, \end{aligned} \tag{19}$$

where $\mathbf{T}_{1D}$ collects the orthogonal components $\mathbf{F}_1\mathbf{z}_l$ (hence the name O[rthogonal]2PLS) and $\mathbf{P}_{1D}$ its loadings, and likewise for $\mathbf{T}_{2D}$ and $\mathbf{P}_{2D}$. The orthogonality properties between the different matrices are shown in Table 2 (see the work of Van der Kloet et al[15]). This means that O2PLS takes the viewpoint of Figure 3A (apart from the fact that each block has its own common component). Equation 19 shows that also O2PLS fits in our framework and obeys the general model of Equation 9 (also indicated in Table 1), but calculating explained variances is hampered by the orthogonality properties. Note again that we changed the notation of the common parts to emphasize that $R(\mathbf{X}_{1C}) \neq R(\mathbf{X}_{2C})$. O2PLS is within block scale dependent but between block scale independent.

The O2PLS method has been used among others in spectroscopy[56-59] and in the plant sciences,[60,61] and its extension to more than 2 blocks (OnPLS) has been used in genomics.[62,63] The latter paper also describes an implementation of the multiblock problem as shown in Figure 4, showing the complexity of such a decomposition. There is an interesting relationship of O2PLS with Procrustes analysis, as explained in the Appendix.

### 3.3 | DIstinct and COmmon SCA (DISCO)

Also the DISCO method[5,12] can be posed in terms of our framework. The first step in DISCO is to solve an SCA problem to find scores $\tilde{\mathbf{T}}(I \times R)$ and loadings $\tilde{\mathbf{P}}([J_1 + J_2] \times R)$ of the concatenated matrix $[\mathbf{X}_1|\mathbf{X}_2]$. The loading matrix $\tilde{\mathbf{P}}$ can be partitioned in $\tilde{\mathbf{P}}_1(J_1 \times R)$ and $\tilde{\mathbf{P}}_2(J_2 \times R)$. Subsequently, the matrix $\tilde{\mathbf{P}}$ is orthogonally rotated to a simple structure reflecting distinct and common components. For the sake of illustration, assume that $R = 3$; there are 1 common and 2 distinct components (1 for each block). Then $\tilde{\mathbf{P}}$ is orthogonally rotated to a structure $\mathbf{P}_{\text{target}}$ according to

$$\min_{\mathbf{Q}^T\mathbf{Q}=\mathbf{I}} \left\|\mathbf{V} \odot (\tilde{\mathbf{P}}\mathbf{Q} - \mathbf{P}_{\text{target}})\right\|^2, \tag{20}$$

with $\mathbf{V}$ a matrix of 0's and 1's selecting the elements across which the minimization occurs, the symbol $\odot$ indicates the Hadamard or elementwise product and

$$\mathbf{P}_{\text{target}} = \begin{vmatrix} x & 0 & x \\ x & 0 & x \\ x & 0 & x \\ 0 & x & x \\ 0 & x & x \\ 0 & x & x \\ 0 & x & x \end{vmatrix}, \tag{21}$$

where the symbol $x$ means an arbitrary value not necessarily 0 and $\mathbf{P} = \tilde{\mathbf{P}}\mathbf{Q} = [\mathbf{P}_1^T|\mathbf{P}_2^T]^T$. This will result in the first component being distinct for $\mathbf{X}_1$, the second component being distinct for $\mathbf{X}_2$, and the third component will be the common one. After finding the optimal $\mathbf{Q}$, the scores $\tilde{\mathbf{T}}$ are counter-rotated, resulting in $\mathbf{T} = \tilde{\mathbf{T}}\mathbf{Q} = [\mathbf{t}_1\mathbf{t}_2\mathbf{t}_3]$, and the following decomposition is obtained:

$$\begin{aligned} \mathbf{X}_1 &= \mathbf{T}\mathbf{P}_1^T = \mathbf{t}_1\mathbf{p}_{11}^T + \mathbf{t}_2\mathbf{p}_{12}^T + \mathbf{t}_3\mathbf{p}_{13}^T + \mathbf{E}_1, \\ \mathbf{X}_2 &= \mathbf{T}\mathbf{P}_2^T = \mathbf{t}_1\mathbf{p}_{21}^T + \mathbf{t}_2\mathbf{p}_{22}^T + \mathbf{t}_3\mathbf{p}_{23}^T + \mathbf{E}_2, \end{aligned} \tag{22}$$

where $\mathbf{p}_{11}$ gives loadings for the distinct component for $\mathbf{X}_1$, $\mathbf{p}_{22}$ for the distinct component for $\mathbf{X}_2$, and $\mathbf{p}_{13}$ and $\mathbf{p}_{23}$ for the common component. If $\mathbf{p}_{12}$ is not close to 0, then there is a distinct nonorthogonal part in the decomposition of $\mathbf{X}_1$ (the red $\mathbf{X}_{1DNO}$ in Table 1). Hence, DISCO fits in our general model of Equation 9, which is also shown in Table 1.

**TABLE 3** Overview of distinct and common simultaneous component analysis models for the medical biology example

| SCA Comp | Increasing Fit Values 1 | | 2 | | 3 | | 4 | | ExplVar, % |
|---|---|---|---|---|---|---|---|---|---|
| | C-AL | 0.13 | C-AL | 0.15 | C-AO | 0.16 | C-ALO | 0.20 | 53 |
| 3 | C-AL | | C-AO | | C-ALO | | C-ALO | | |
| | C-ALO | | C-LO | | C-ALO | | C-ALO | | |
| | C-ALO | 0.19 | D-A | 0.20 | D-A | 0.23 | C-AL | 0.24 | 58 |
| 4 | C-ALO | | C-AL | | D-O | | C-AL | | |
| | C-ALO | | C-AL | | C-ALO | | C-AL | | |
| | C-ALO | | C-ALO | | C-ALO | | C-ALO | | |
| | D-A | 0.24 | C-AL | 0.26 | D-A | 0.28 | D-A | 0.28 | 63 |
| | D-O | | C-AO | | D-O | | D-O | | |
| 5 | C-AL | | C-ALO | | C-AL | | C-LO | | |
| | C-AL | | C-ALO | | C-ALO | | C-ALO | | |
| | C-ALO | | C-ALO | | C-ALO | | C-ALO | | |

Abbreviation: SCA, simultaneous component analysis.

The table shows models for the 4 lowest-fit values for rotations based on 3 to 5 SCA components. Components are labeled as common (C) or distinct (D). The colored components are subspaces that are the same across several models, and the correlation between these subspaces are given. The framed model is selected for further interpretation

The SCA solution has orthogonal columns in $\tilde{\mathbf{T}}$ and rotates orthogonally afterwards; thus, these columns remain orthogonal. Hence, $\mathbf{T}$ is orthogonal, and it holds that $\mathbf{X}_{1DNO}$ is orthogonal to both $\mathbf{X}_{1DO}$ and $\mathbf{X}_{1C}$, but it is clearly not orthogonal to $\mathbf{X}_{2DO} = \mathbf{t}_2\mathbf{p}_{22}^T$ (see Table 3 and Figure 3A). Minimizing the sum of squared elements of $\mathbf{p}_{12}$ and $\mathbf{p}_{21}$ is exactly what the abovementioned rotation tries to do, thereby minimizing the sizes of these distinct nonorthogonal parts and defining those parts as being distinct nonorthogonal (see $\mathbf{P}_{\text{target}}$ in Equation 21). Thus, this is yet another implementation of the general decomposition scheme where the vectors can be matrices when more than 1 common and distinct components are present. Contrary to O2PLS, the common parts in DISCO span the same column space. Because DISCO starts with an SCA and subsequently uses a rotation, both $\mathbf{T}$ and $\tilde{\mathbf{T}}$ are in the column space of the combined $[\mathbf{X}_1|\mathbf{X}_2]$ rather than the individual parts. Because of the orthogonality of the score matrix $\mathbf{T}$, explained variances can be calculated based on Equation 22. DISCO is within and between block scale dependent and has been used in metabolomics[5] and in gene expression analysis.[14]

## 3.4 | Generalized SVD

A method used in gene expression data to separate common from distinct components is the GSVD,[3] which is also a generalization of the SVD known as the quotient SVD.[52] The mathematics of the GSVD dates back already some time.[64,65] The original GSVD is used for fusion of data sharing the same columns, but this problem can be transposed to our situation. The original GSVD is a matrix decomposition method and does not have least squares properties. To repair its sensitivity to noise, we follow the implementation of the adapted GSVD, which comes down to first filtering the data with an SCA step.[5] For the 2-block case, the model is

$$\begin{aligned}\mathbf{X}_1 &= \hat{\mathbf{X}}_1 + \mathbf{E}_1 = \mathbf{T}\mathbf{D}_1\mathbf{V}_1^T + \mathbf{E}_1,\\ \mathbf{X}_2 &= \hat{\mathbf{X}}_2 + \mathbf{E}_2 = \mathbf{T}\mathbf{D}_2\mathbf{V}_2^T + \mathbf{E}_2,\end{aligned} \tag{23}$$

with $\hat{\mathbf{X}}_k$ as the filtered data, $\mathbf{V}_k^T\mathbf{V}_k = \mathbf{I}$ $(k = 1, 2)$, $\mathbf{D}_k$ $(k = 1, 2)$ as the diagonal such that $\mathbf{D}_1^2 + \mathbf{D}_2^2 = \mathbf{I}$, and $\mathbf{T}$ as a full-rank matrix but not necessarily orthogonal. Because of the latter constraint, it is possible to divide the generalized singular values (the elements of $\mathbf{D}_k[k = 1, 2]$) in 3 groups: if $d_{1R}^2 \approx 1$, the corresponding component is distinctive for $\mathbf{X}_1$; if $d_{2R}^2 \approx 1$, the corresponding component is distinctive for $\mathbf{X}_2$; and if $d_{1R}^2 \approx d_{2R}^2$, the corresponding component is common. Obviously, there is a certain amount of arbitrariness in these choices. Once such a choice is made, Equation 23 can be written as

$$\begin{aligned}\mathbf{X}_1 &= \mathbf{T}_1\mathbf{D}_{11}\mathbf{V}_{11}^T + \mathbf{T}_2\mathbf{D}_{12}\mathbf{V}_{12}^T + \mathbf{T}_3\mathbf{D}_{13}\mathbf{V}_{13}^T + \mathbf{E}_1,\\ \mathbf{X}_2 &= \mathbf{T}_1\mathbf{D}_{21}\mathbf{V}_{21}^T + \mathbf{T}_2\mathbf{D}_{22}\mathbf{V}_{22}^T + \mathbf{T}_3\mathbf{D}_{23}\mathbf{V}_{23}^T + \mathbf{E}_2,\end{aligned} \tag{24}$$

which fits our framework and Equation 9. Upon assuming that $\mathbf{T}_1\mathbf{D}_{11}\mathbf{V}_{11}^T = \mathbf{X}_{1DO}$, $\mathbf{T}_2\mathbf{D}_{22}\mathbf{V}_{22}^T = \mathbf{X}_{2DO}$ and $\mathbf{T}_3\mathbf{D}_{13}\mathbf{V}_{13}^T, \mathbf{T}_3\mathbf{D}_{23}\mathbf{V}_{23}^T$ are the common components, then $\mathbf{T}_2\mathbf{D}_{12}\mathbf{V}_{12}^T = \mathbf{X}_{1DNO}$ and $\mathbf{T}_1\mathbf{D}_{21}\mathbf{V}_{21}^T = \mathbf{X}_{2DNO}$. Because of the orthogonality of both $\mathbf{V}_1$ and $\mathbf{V}_2$, it holds that $\mathbf{X}_{1DO}$, $\mathbf{X}_{1DNO}$, and $\mathbf{X}_{1C}$ are mutually orthogonal, and likewise for block $\mathbf{X}_2$. However, $\mathbf{X}_{1DNO}$ is not orthogonal to $\mathbf{X}_{2DO}$ and similarly $\mathbf{X}_{2DNO}$ is not orthogonal to $\mathbf{X}_{1DO}$. This is the same as for DISCO and is again the situation of Figure 3A. Generalized SVD is within and between block scale dependent and has been used in gene expression analysis[3] and has been extended for more than 2 blocks in different ways.[66,67]

### 3.5 | Joint and individual variances explained

The method of JIVE[6] goes as follows. For 2 blocks, it derives directly a decomposition according to

$$\mathbf{X}_1 = \mathbf{T}_C\mathbf{P}_{1C}^T + \mathbf{T}_{1D}\mathbf{P}_{1D}^T + \mathbf{E}_1 = \mathbf{X}_{1C} + \mathbf{X}_{1D} + \mathbf{E}_1,$$
$$\mathbf{X}_2 = \mathbf{T}_C\mathbf{P}_{2C}^T + \mathbf{T}_{2D}\mathbf{P}_{2D}^T + \mathbf{E}_2 = \mathbf{X}_{2C} + \mathbf{X}_{2D} + \mathbf{E}_2,$$

which fits directly in our framework and Equation 9. Note that we use the notation $\mathbf{T}_C$ to stress that the common scores are the same. In estimating this decomposition, we use the following constraints:

$$\mathbf{X}_{1C}^T\mathbf{X}_{1D} = 0; \mathbf{X}_{1C}^T\mathbf{X}_{2D} = 0; \mathbf{X}_{2C}^T\mathbf{X}_{1D} = 0; \mathbf{X}_{2C}^T\mathbf{X}_{2D} = 0 \quad (25)$$

and, thus, the distinct part in a block is orthogonal to the common parts in all blocks, but the distinct parts in different blocks are not necessarily orthogonal. This is again an implementation of Figure 3A. The (low) ranks of all common and distinct matrices involved are determined by permutation tests. Joint and individual variances explained is within and between block scale dependent and has been applied in gene expression analysis.[6]

### 3.6 | Structure-revealing data fusion

In structure-revealing data fusion,[7] an approach is chosen based on penalties. The method is developed for fusing 2-way and 3-way arrays but can equally well be used for fusing 2-way arrays. The starting point is the model

$$\begin{aligned}\mathbf{X}_1 &= \mathbf{T}\mathbf{D}_1\mathbf{V}_1^T + \mathbf{E}_1 = \mathbf{T}\mathbf{P}_1^T + \mathbf{E}_1,\\ \mathbf{X}_2 &= \mathbf{T}\mathbf{D}_2\mathbf{V}_2^T + \mathbf{E}_2 = \mathbf{T}\mathbf{P}_2^T + \mathbf{E}_2,\end{aligned} \quad (26)$$

where the matrices $\mathbf{D}_1$ and $\mathbf{D}_2$ are diagonal and the diagonals of $\mathbf{T}^T\mathbf{T}$, $\mathbf{V}_1^T\mathbf{V}_1$, and $\mathbf{V}_2^T\mathbf{V}_2$ consist of 1's (ie, the columns of $\mathbf{T}$, $\mathbf{V}_1$, and $\mathbf{V}_2$ have length 1). The components are now estimated under an $L_1$ penalty[68]:

$$\min_{\mathbf{T},\mathbf{V}_1,\mathbf{V}_2,\mathbf{D}_1,\mathbf{D}_2} \left\|\mathbf{X}_1 - \mathbf{T}\mathbf{D}_1\mathbf{V}_1^T\right\|^2 + \left\|\mathbf{X}_2 - \mathbf{T}\mathbf{D}_2\mathbf{V}_2^T\right\|^2 + \lambda(\|\mathrm{diag}(\mathbf{D}_1)\|_1 + \|\mathrm{diag}(\mathbf{D}_2)\|_1) \quad (27)$$

where $\lambda \geqslant 0$ is the penalty parameter to be set by the user, the symbol $\|\cdot\|_1$ represents the $L_1$-norm, and $\mathrm{diag}(\mathbf{D}_k)$ is the vector carrying the diagonal of $\mathbf{D}_k$. Increasing the penalty value $\lambda \geqslant 0$ will force more elements in $\mathbf{D}_1$ and $\mathbf{D}_2$ to become 0. From the patterns of these 0's, the common and distinct components are defined, and thus, this method also obeys Equation 9. This type of approach—albeit in fusing 3- and 2-way data—has been used in metabolomics.[7,16] Structure-revealing data fusion is within and between block scale dependent.

A special class of structure-revealing data fusion methods is the multivariate curve resolution method used in chemometrics.[19] This class of methods performs data fusion mostly using hard constraints on the parameters based on chemical information. There are very many applications of this method in different fields of chemistry.

## 4 | EXAMPLES

To illustrate some of the methods falling under our framework and their relationships, we will show some real data examples that were already introduced shortly in Section 1. The 2 data sets are from medical biology and food science. The aim of these examples is to emphasize the added value of distinguishing between common and distinctive subspaces, and to show how it can be implemented and interpreted by existing methodologies. All data will be analyzed by the methods PCA-GCA (see Section 3.1.2) and DISCO (see Section 3.3). These methods are selected because they represent different orthogonality constraints (referring to Figure 3A and B, respectively). They also represent different choices of common components as discussed in Section 2.2: PCA-GCA estimates separate common components for each data block, while DISCO estimates a best compromise, which is in the column space of the concatenated data blocks.

For all methods, the first step is to decide the dimensionalities of the subspaces. This is not a trivial task, and different strategies exist for the different methods. The strategies for PCA-GCA and DISCO will be explained briefly in the examples below, but a thorough discussion of the model selection is not within the scope of this paper. Once the dimensions are decided, it is straightforward to estimate basis vectors (or components) for each of the subspaces.

### 4.1 | Sensory example

The sensory example focuses on one of the typical aspects of a product development process: The product developer is interested in understanding how well 2 important modalities of the descriptive sensory profile relate to the ingredients in the recipe. A typical issue of interest for being able to optimize product quality is whether the recipe influences both smell and taste and in which way this happens. In particular, one is interested in knowing what aspects of smell and taste that are common and what is unique in the 2 sensory profiles.

This example consists of descriptive sensory attributes of flavored water samples and is a subset of a larger data set.[24] The 18 water samples are created according to a full-factorial experimental design with 2 flavor types (A and B), 3 flavor doses (0.2, 0.5, and 0.8 g/L), and 3 sugar levels (20, 40, and 60 g/L). A trained sensory panel consisting of 11 assessors evaluated the samples first by smelling (9 descriptors) and then by tasting (14 descriptors), using an intensity scale from 1 to 9. Two data blocks (smell and taste) were constructed by

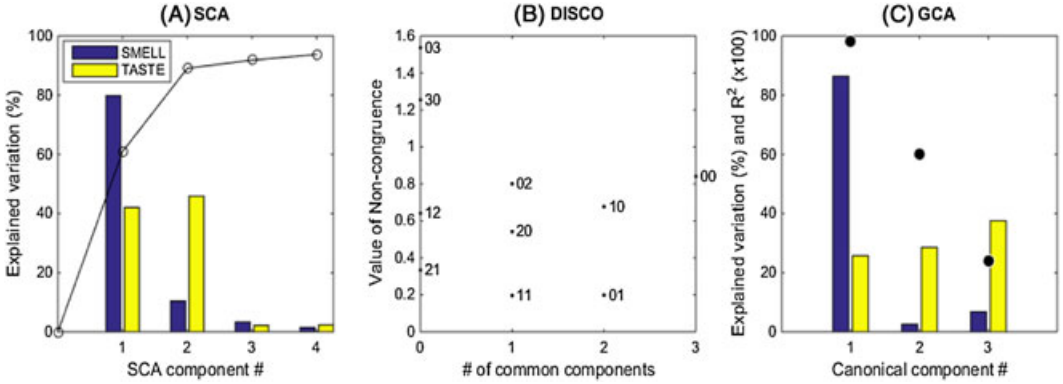


**FIGURE 7** Plots for selecting numbers of components for the sensory example. A, Simultaneous component analysis (SCA). The curve represents cumulative explained variance for the concatenated data blocks. The bars show how much variance each component explains in the individual blocks. B, Distinct and common SCA (DISCO). Each point represents the noncongruence value for a given target (model). The plot includes all possible combinations of common and distinct components based on a total rank of 3. The horizontal axis represents the number of common components, and the numbers in the plot represent the number of distinct components for smell and taste. C, Principal components analysis–generalized canonical correlation analysis (GCA): Black dots represent the canonical correlation coefficient (×100), and the bars show how much variance the canonical components explain in each block

averaging across assessors. The blocks were mean centered and block scaled to SS 1 prior to analysis.

A crucial aspect of the decomposition is to decide the dimensions of the common and distinct subspaces. For DISCO, this is a 2-step process: First, the number of SCA components is selected. This number represents the sum of the dimensions of all subspaces, ie, $\dim R(\mathbf{X}_{12C}) + \dim R(\mathbf{X}_{1D}) + \dim R(\mathbf{X}_{2D})$. Then, the most appropriate target matrix $\mathbf{P}_{\text{target}}$ (Equation 21) is sought by evaluating the noncongruence value (Equation 20) for all possible allocations of common and distinct components. Since there are 3 independent design factors in this experiment (flavor type, flavor dose, and sugar level), we choose to keep 3 SCA components even if the third component explains very little variance (Figure 7A). The lowest noncongruence value is approximately equal for models with 1 and 2 common components (Figure 7B), but after a closer inspection of the scores, we choose the model with 1 common component and 1 distinct component per block.

The DISCO decomposition for this data set is then as follows:

$$\begin{aligned}\mathbf{X}_S &= \mathbf{X}_{C,S}(70.0\%) + \mathbf{X}_{DNO,T}(3.4\%) + \mathbf{X}_{D,S}(9.4\%) + \mathbf{E}_S,\\ \mathbf{X}_T &= \mathbf{X}_{C,T}(26.8\%) + \mathbf{X}_{D,T}(55.5\%) + \mathbf{X}_{DNO,S}(1.9\%) + \mathbf{E}_T,\end{aligned} \tag{28}$$

where each subspace is of dimension 1; S and T stand for smell and taste, respectively, and between brackets is the amount of explained variation in the block. For real data, the noncongruence value is never 0, meaning that the 0's in the target matrix are not exactly 0 in the rotated loadings. This constitutes the so-called distinct nonorthogonal subspaces, represented by $\mathbf{X}_{DNO,T}$ and $\mathbf{X}_{DNO,S}$ in Equation 28. Note that both these subspaces are very small in this case and probably consist of noise only.

For PCA-GCA, the dimension selection is also a stepwise procedure: First, an appropriate number of principal components are selected for each data block, corresponding to $\dim R(\mathbf{X}_1)$ and $\dim R(\mathbf{X}_2)$ in Equation 2. Next, the correlation coefficients and explained variances from GCA are evaluated to decide the number of common components, $\dim R(\mathbf{X}_{12C})$. The number of distinct components is then given as the difference between $\dim R(\mathbf{X}_k)$ and $\dim R(\mathbf{X}_{12C})$. In this example, we choose to keep 3 components for each block, following the same argument as for DISCO (3 design factors). Figure 7C shows that the canonical correlation together with the explained variances clearly suggests 1 common component (correlation = 0.98), which means that the distinct subspaces are 2-dimensional. The distinct subspaces can be split into an orthogonal part and a nonorthogonal part as for DISCO, but that is not done here. The decomposition from PCA-GCA is then as follows:

$$\begin{aligned}\mathbf{X}_S &= \mathbf{X}_{C,S}(86.4\%) + \mathbf{X}_{D,S}(9.3\%) + \mathbf{E}_S,\\ \mathbf{X}_T &= \mathbf{X}_{C,T}(25.7\%) + \mathbf{X}_{D,T}(66.2\%) + \mathbf{E}_T,\end{aligned} \tag{29}$$

where the common part has dimensionality 1 and both distinct parts have dimensionality 2.

The subspaces found by PCA-GCA and DISCO are very similar. The correlation between the common DISCO component ($\mathbf{T}_{12C}$) and the common PCA-GCA components ($\mathbf{T}_{1C}$ and $\mathbf{T}_{2C}$) is 0.98 for both blocks. The correlation between the distinct (orthogonal) smell component from DISCO ($\mathbf{T}_{1DO}$) and the first distinct smell component from PCA-GCA (first column of $\mathbf{T}_{2D}$) is 0.74. The corresponding number for the distinct taste components is 0.99. Figure 8 shows biplots from

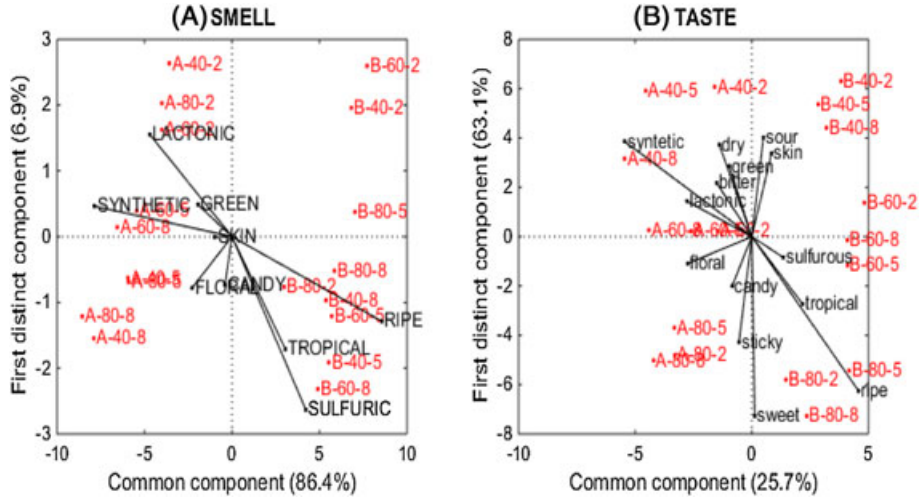


**FIGURE 8** Biplots from principal components analysis–generalized canonical correlation analysis, showing the variables as vectors and the samples as points. The samples are labeled according to the design factors flavor type (A and B), sugar level (40, 60, and 80), and flavor dose (2, 5, and 8). The plots show the common component (horizontal) against the first distinct component for each of the 2 blocks

PCA-GCA for each of the 2 blocks. It is clear that the common component distinguishes between flavor types (A and B). This component explains 86% of the smell variation and 26% of the taste variation. As a validation of the commonness, note that the sensory attributes that span this subspace are the same both for smelling and tasting: synthetic/lactonic/oral for flavor type A versus ripe/tropical/sulfurous for flavor type B. The first distinctive smell component explains 7% of the variation and is related to the flavor dose, showing that the lowest dose tends to give a more lactonic smell. The first distinctive taste component explains 63% of the variation and describes differences in sugar level. The attributes that span this component are sweet/ripe versus sour/synthetic/skin/dry.

This example shows that both methods are able to separate common and distinct subspaces in a similar way. The subspaces that explain a large proportion of the variance (common and distinct taste) are practically equal for both methods (correlations $>$ 0.98), while there is less agreement regarding the weaker distinct smell component (correlation = 0.74).

## 4.2 | Medical biology example

The data set is a subset of a larger study on the effects of gastric bypass surgery on obese and diabetic subjects.[69] Here, we focus on 14 obese patients with DM2 who underwent gastric bypass surgery. Blood samples were taken 4 weeks before and 3 weeks after surgery, and on each occasion, samples were taken both before and after a meal. The blood samples were then analyzed on multiple analytical platforms for the determination of amines, lipids, and oxylipins. The 3 data blocks amines (A), lipids (L), and oxylipins (O) consist of 14 subjects $\times$ 4 samples = 56 rows and 34, 243, and 32 variables, respectively. All variables in all 3 blocks were square root transformed, to obtain more evenly distributed data. Individual differences between subjects were removed by subtracting each subjects' average profile. All variables were then scaled to unit variance. The blocks were also scaled to unit norm prior to SCA, to normalize scale differences between blocks.

Selecting the dimensions of the subspaces is more complicated when the numbers of blocks increase. In this 3-block example, we need to decide the dimensions of 7 subspaces: $\mathbf{X}_{123C}$, $\mathbf{X}_{12C}$, $\mathbf{X}_{13C}$, $\mathbf{X}_{23C}$, $\mathbf{X}_{1D}$, $\mathbf{X}_{2D}$, and $\mathbf{X}_{3D}$. For DISCO, we start by deciding the sum of all the dimensions, ie, the number of SCA components. Explained variance as a function of components for SCA is given in Figure 9A. The curve of cumulative variance does not have a clear bend, which makes it hard to decide the cutoff between structure and noise. To allocate the common and distinct components, we need to fix the number of SCA components and then compare the fit values of Equation 20. The computations are time-consuming, as there are, eg, 462 possible target matrices for the 5-component model. To illustrate the complexity in selecting the dimensions for the subspaces, we have calculated all possible rotations for models with 3 to 5 SCA components, and the results for the 4 best-fit values are given in Table 3. The values are very similar, making it hard to conclude which rotation gives the best fit. Looking further into the actual rotated score vectors, we discover that many of the models agree on some of the subspaces. These are marked with colors in Table 3. We choose to interpret the 5-component model with fit value 0.24 (the best 5-component model), since this model includes all the agreed-upon subspaces. The model contains 1 component that is common across all 3 blocks, 2 components common for A and L, and 1 distinct component from both A and O. The decomposition of each block is illustrated by pie charts in Figure 10A to C. Notice that there is a substantial contribution

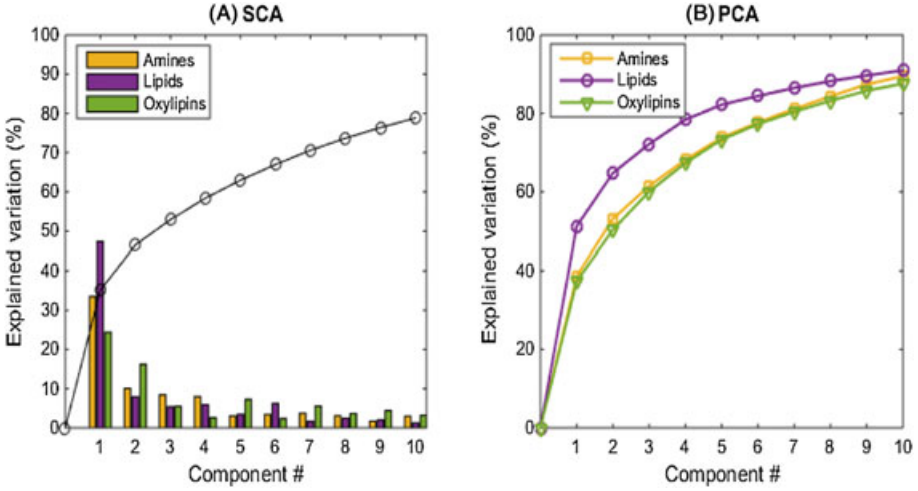


**FIGURE 9** A, Explained variances for simultaneous component analysis. The bars represent variances within each block, and the curve represent cumulative explained variance in all blocks combined. B, Explained variances for principal components analysis on each block separately

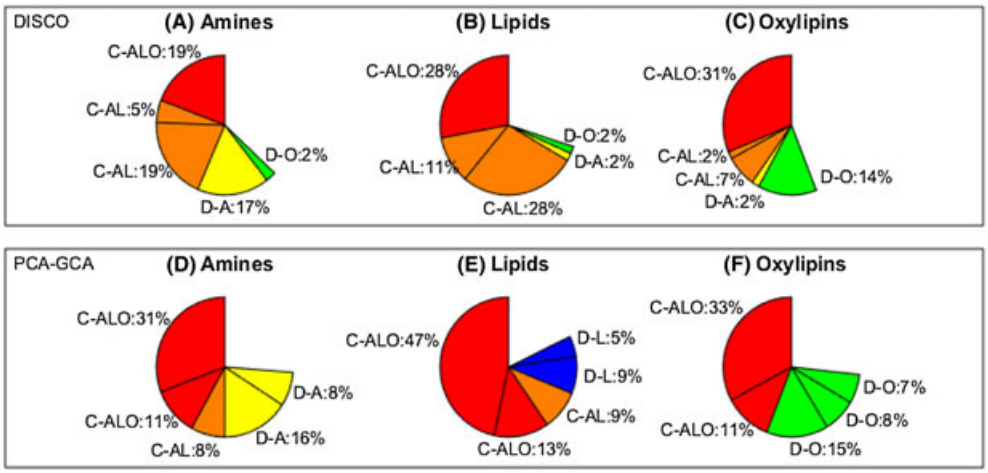


**FIGURE 10** A-C, Decomposition by distinct and common simultaneous component analysis (DISCO) for blocks A, L, and O respectively. D-F, Corresponding decomposition by principal components analysis–generalized canonical correlation analysis. Each segment represents a component (dimension)

of one of the C-AL components also in the O block (7%), which implies that this component could perhaps also be regarded as common across all 3 blocks.

In PCA-GCA, the number of principal components needs to be set for each block separately before performing GCA. Explained variations for the 3 PCA models are shown in Figure 9B. As for SCA, it is not clear how many components to keep for each block. To investigate how the choice affects the GCA, we ran GCA on all combinations of 5 to 8 components from each block (64 combinations in total). The canonical correlation coefficient for cases with more than 2 blocks is defined as the average correlation between all pairs of components from different blocks. Using 0.7 as correlation threshold for commonness in the GCA, we found that 85% of the models had 2 common components across all blocks and 1 common component across A and L. The model based on 5 components for each block is illustrated in Figure 10D to F. This model is slightly different from the DISCO model, as it contains 2 components that are common across all 3 blocks. A closer investigation of the components revealed that the second common component across all 3 blocks is very similar to the one of the C-AL DISCO component mentioned above, which explained 7% of the variation in O. This illustrates the complexity of splitting common and distinct components in noisy and complex data.

To interpret the different subspaces, we plot the scores and loadings from the DISCO model. Figure 11 shows the 1-dimensional subspace that is common for all 3 blocks (C-ALO), which accounts for 19%, 28%, and 31% of the variation in A, L, and O, respectively. The scores are shown in the top panel of Figure 11. It is clear that the component contains information related to both surgery and meal; the scores are increasing after surgery and decreasing after the meal. The variables spanning this dimension in each of the 3 blocks

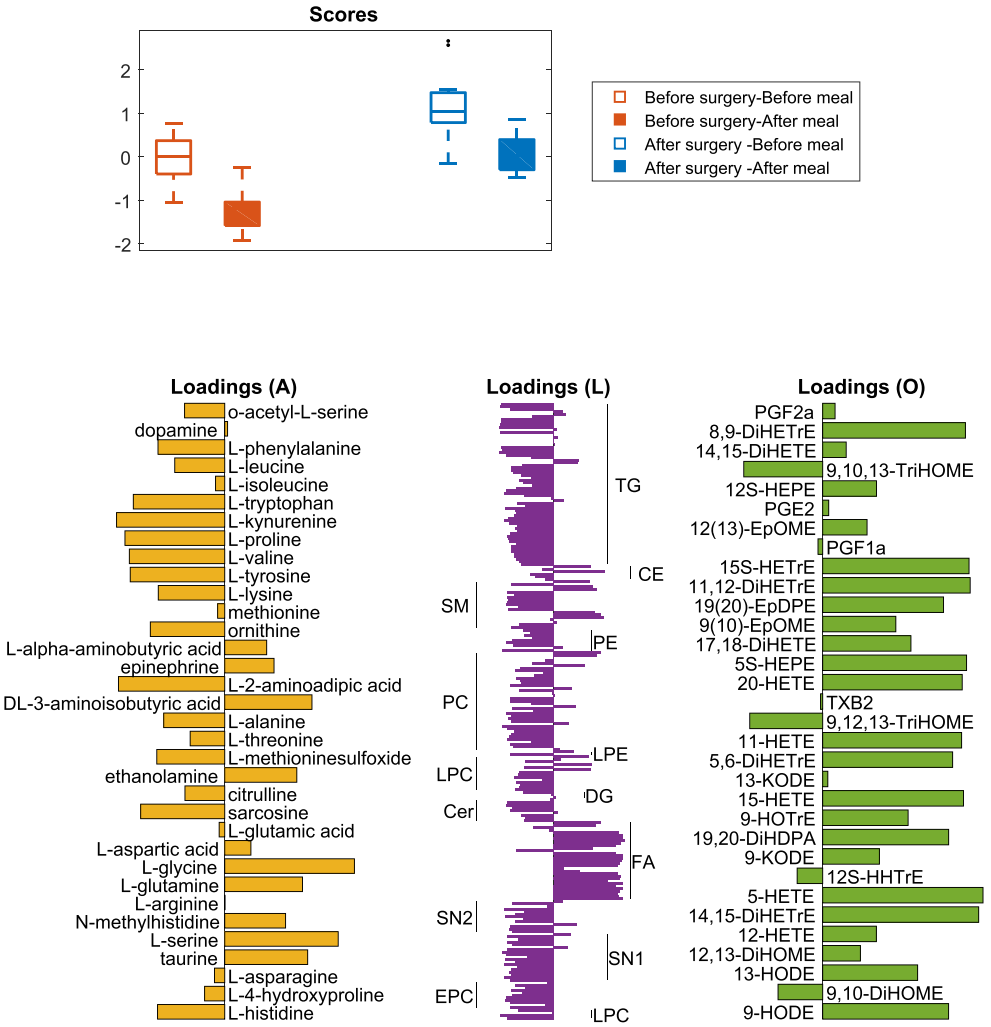


**FIGURE 11** Scores and loadings for the 1-dimensional distinct and common simultaneous component analysis subspace common between all 3 blocks

are shown in the bar plots of Figure 10 (bottom). The most striking observation is that the branched-chain amino acid leucine, valine, and L-2-aminoadipic acid (closely related to branched-chain amino acids) are downregulated after surgery, which confirms earlier findings.[69] There is more in common between amines and lipids than oxylipids; both amines and lipids are involved in central carbon and energy metabolism, and therefore, they may show higher correlation among some amino acids and some lipid groups (as reflected by common subspace).

The 2-dimensional subspace common between A and L is shown in Figure 12. These 2 components together account for 24% and 39% in the A and L blocks, respectively, and they even explain 9% in the O block. Here, also, we see groupings according to both surgery and meal, especially in the vertical dimension. Note that the 2 groups that were overlapping in the C-ALO component ("before surgery–before meal" versus "after surgery–after meal") are completely separated in this subspace. Plots of the distinct components (not shown) did not reveal clear patterns related to the factors treatment and meal. Hence, all effects are seen in the common parts, meaning that a large part of the metabolism is affected simultaneously by these 2 factors.

## 4.3 | Summary of the examples

A general conclusion from both examples is that the concepts of common and distinct components are useful for understanding the relation between data sets. They also show that different methods, although they seek to reveal common and distinct information, do not necessarily extract exactly the same spaces. This may be seen as advantageous since it may highlight different aspects of the data, but it also blurs the relation between the concise mathematical definitions and the actual implementations. This points to a need for a more direct bridge between the concise definitions and actual methodology and also for more insight into how to implement them the best possible way.

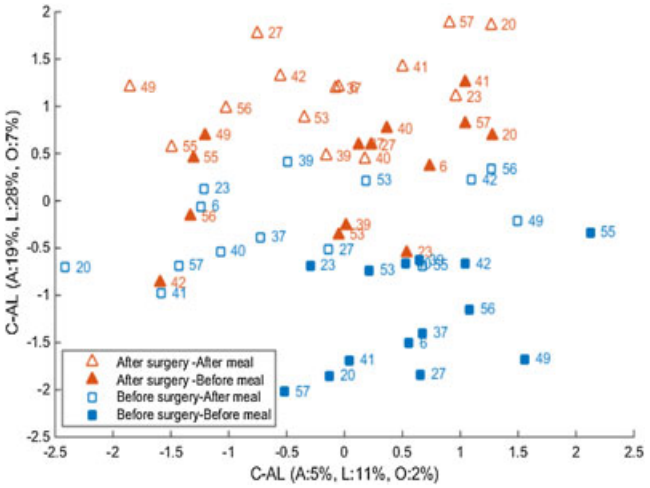


**FIGURE 12** Scores for the 2-dimensional distinct and common simultaneous component analysis subspace common between the amine and lipid blocks

# 5 | DISCUSSION

## 5.1 | Finding common and distinct subspaces

There is an interesting difference in the way the various methods find common and distinct subspaces. Some methods work clearly in the column spaces of the matrices involved (GCA and JIVE), some methods work through the row spaces, (DISCO and O2PLS), and some methods work in both types of spaces simultaneously (GSVD and structure-revealing data fusion). What consequences this has for the interpretation of the results of the different models is an open question.

## 5.2 | Open issues and future work

There are obviously many open issues in this field of research. We have only briefly touched upon the issue of explained variances, but there are many nontrivial aspects that need attention. Also, the problem of interpretation, that is, moving from layers 3 to 4 in Figure 1, needs attention. This is a very important issue because interpretation is one of the raisons d'être for data fusion methods. Our framework primarily considers the column space of the data matrices, but interpretation is done mostly in the row space. How to investigate this depends on the scope of the analysis and the type of data available, and it is not possible to set up a completely general procedure. There are, however, some general tools that can be useful: One important possibility is to simply project the original data blocks onto the estimated subspaces. For instance, for $R(\mathbf{X}_{12C})$, one simply regresses $\mathbf{X}_1$ onto a suitable basis for the space $R(\mathbf{X}_{12C})$. In this way, one obtains information about how the original data are related to the basis for each subspace. Also, moving to analyzing more than 2 blocks simultaneously is not trivial. Many choices have to be made, and no clear guidelines exist on how to perform this. Model selection becomes an even more important issue then, and possibly Bayesian factor analysis methods with automated model selection can be of use in this context.[70]

## ACKNOWLEDGEMENTS

We thank Frans van der Kloet and Johan Westerhuis (both from Biosystems Data Analysis, University of Amsterdam) for stimulating discussions.

## REFERENCES

1. Kettenring JR. Canonical analysis of several sets of variables. *Biometrika*. 1971;58:433–460.
2. Van de Geer JP. Linear relations among *k* sets of variables. *Psychometrika*. 1984;49(1):79–94.
3. Alter O, Brown PO, Botstein D. Generalized singular value decomposition for comparative analysis of genome-scale expression data sets of two different organisms. *P Natl Acad Sci USA*. 2003;100:3351–3356.
4. Trygg J, Wold S. O2-PLS, a two-block (*x*-*y*) latent variable regression (LVR) method with an integral OSC filter. *J Chemometr*. 2003;17(1):53–64.
5. Van Deun K, Van Mechelen L, Thorrez I, et al. DISCO-SCA and properly applied GSVD as swinging methods to find common and distinctive processes. *PLoS One*. 2012;7:5.
6. 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(1):523–542.
7. Acar E, Papalexakis EE, Gurdeniz G, et al. Structure-revealing data fusion. *BMC Bioinformatics*. 2014;15:239.
8. Van Mechelen I, Smilde AK. A generic linked-mode decomposition model for data fusion. *Chemometr Intell Lab*. 2010;104:83–94.
9. Lahat D, Adali T, Jutten C. Multimodal data fusion: an overview of methods, challenges and prospects. *Proc IEEE*. 2015;103(9):1449–1477.
10. Westerhuis JA, Kourti T, MacGregor JF. Analysis of multiblock and hierarchical PCA and PLS models. *J Chemometr*. 1998;12(5):301–321.
11. Naes T, Tomic O, Afseth NK, Segtnan V, Mage I. Multi-block regression based on combinations of orthogonalisation, PLS-regression and canonical correlation analysis. *Chemometr Intell Lab*. 2013;124:32–42.
12. Schouteden M, Van Deun K, Pattyn S, Van Mechelen I. SCA with rotation to distinguish common and distinctive information in linked data. *Behav Res Methods*. 2013;45(3):822–833.
13. Golub GH, Van Loan C. *Matrix Computations*. 3rd ed. Baltimore, MD: John Hopkins University Press; 1996.
14. Van Deun K, Smilde AK, Thorrez L, Kiers HAL, Van Mechelen I. Identifying common and distinctive processes underlying multiset data. *Chemometr Intell Lab*. 2013;129:40–51.
15. Van der Kloet FM, Sebastian-Leon P, Conesa A, Smilde AK, Westerhuis JA. Separating common from distinct variation. *BMC Bioinformatics*. 2016;17(5):195.
16. Acar E, Bro R, Smilde AK. Data fusion in metabolomics using coupled matrix and tensor factorizations. *Proc IEEE*. 2015;103(9):1602–1620.
17. Lynch CJ, Adams SH. Branched-chain amino acids in metabolic signalling and insulin resistance. *Nat Rev Endocrinol*. 2014;10:723–736.

18. Sidiropoulos N, Bro R. On communication diversity for blind identifiability and uniqueness of low-rank decompositions of *N*-way arrays. *Proc Int Conf Acoust Speech and Signal Process*. 2000;5:2449–2452.
19. Tauler R, Smilde AK, Kowalski BR. Selectivity, local rank, three-way data analysis and ambiguity in multivariate curve resolution. *J Chemometr*. 1995;9(1):31–58.
20. Schott JR. *Matrix Analysis for Statistics*. New York, NY: Wiley and Sons; 1997.
21. Yanai H, Takeuchi K, Takane Y. *Statistics for Social and Behavioral Sciences*. New York, NY: Springer; 2011.
22. Peres-Neto PR, Legendre P, Dray S, Borcard D. Variation partitioning of species data matrices: estimation and comparison of fractions. *Ecology*. 2006;87(10):2614–2625.
23. Van den Berg RA, Rubingh CM, Westerhuis JA, van der Werf MJ, Smilde AK. Metabolomics data exploration guided by prior knowledge. *Anal Chim Acta*. 2009;651(2):173–181.
24. Mage I, Menichelli E, Naes T. Preference mapping by PO-PLS: separating common and unique information in several data blocks. *Food Qual Prefer*. 2012;24(1):8–16.
25. Bro R, Smilde AK. Centering and scaling in component analysis. *J Chemometr*. 2003;17:16–33.
26. Van den Berg RA, Hoefsloot HCJ, Westerhuis JA, Smilde AK, van der Werf MJ. Centering, scaling, and transformations: improving the biological information content of metabolomics data. *BMC Genomics*. 2006;7(142).
27. Van Deun K, Smilde AK, van der Werf MJ, Kiers HAL, Van Mechelen I. A structured overview of simultaneous component based data integration. *BMC Bioinformatics*. 2009;10:246.
28. Simsekli U, Ermis B, Cemgil AT, Acar E. Optimal weight learning for coupled tensor factorization with mixed divergences. *Proc. 21st. Eur. Signal Process. Conf.* Marrakech, Morocco; 2013.
29. Wilderjans TF, Ceulemans E, van Mechelen I, van den Berg RA. Simultaneous analysis of coupled data matrices subject to different amounts of noise. *British J Math Stat Psychol*. 2011;64: 277–290.
30. Timmerman ME, Hoefsloot HCJ, Smilde AK, Ceulemans E. Scaling in ASCA. *Metabolomics*. 2015;accepted.
31. Smilde AK, Westerhuis JA, de Jong S. A framework for sequential multiblock component methods. *J Chemometr*. 2003;17(6):323–337.
32. Timmerman ME, Kiers HAL. Four simultaneous component models of multivariate times series from more than one subject to model intraindividual and interindividual differences. *Psychometrika*. 2003;86:105–122.
33. Smilde AK, Bro R, Geladi P. *Multi-way Analysis: Applications in the Chemical Sciences*. Chichester, UK: John Wiley & Sons Inc.; 2004.
34. Singh A, Gordon GJ. Relational learning via collective matrix factorization. *Knowledge Discovery and Data Mining (KDD)*, Las Vegas; 2008.
35. Smilde AK, Jansen JJ, Hoefsloot HCJ, Lamers RJAN, van der Greef J, Timmerman ME. ANOVA–simultaneous component analysis (ASCA): a new tool for analyzing designed metabolomics data. *Bioinformatics*. 2005;21(13):3043–3048.
36. Jansen JJ, Hoefsloot HCJ, van der Greef J, Timmerman ME, Smilde AK. Multilevel component analysis of time-resolved metabolic fingerprinting data. *Anal Chim Acta*. 2005;530(2):173–183.
37. de Noord OE, Theobald EH. Multilevel component analysis and multilevel PLS of chemical process data. *J Chemometr*. 2005;19(5-7):301–307.
38. Bevilacqua M, Bucci R, Materazzi S, Marini F. Application of near infrared (NIR) spectroscopy coupled to chemometrics for dried egg–pasta characterization and egg content quantification. *Food Chem*. 2013734;140(4):726.
39. Tao Y, Wu D, Sun DW, et al. Quantitative and predictive study of the evolution of wine quality parameters during high hydrostatic pressure processing. *Innov Food Sci Emerg*. 201390;20:81.
40. Tomassini A, Vitalone A, Marini F, et al. H-1 NMR-based urinary metabolic profiling reveals changes in nicotinamide pathway intermediates due to postnatal stress model in rat. *J Proteome Res*. 20145859;13(12):5848.
41. Shan RF, Zhao Y, Fan ML, Liu XW, Cai WS, Shao XG. Multilevel analysis of temperature dependent near-infrared spectra. *Talanta*. 2015;131(170):174.
42. Pages J. Collection and analysis of perceived product inter-distances using multiple factor analysis: application to the study of 10 white wines from the Loire Valley. *Food Qual Prefer*. 2005;16(7):642–649.
43. Dahl T, Naes T. A bridge between Tucker-1 and Carroll's generalized canonical analysis. *Comput Stat Data An*. 2006;50(11):3086–3098.
44. Bro R, Qannari EM, Kiers HAL, Naes T, Frøst MB. Multi-way models for sensory profiling data. *J Chemom*. 2008;22:36–45.
45. van der Burg E, Dijksterhuis G. Generalized canonical analysis of individual sensory profiles and instrumental data.. In: Naes T, Risvik E, eds. *Multivariate Analysis of Data in Sensory Science*. Amsterdam: Elsevier; 1996.
46. Hotelling H. Relations between two sets of variates. *Biometrika*. 1936;28:321–377.
47. Tenenhaus A, Tenenhaus M. Regularized generalized canonical correlation analysis for multiblock or multigroup data analysis. *Eur J Oper Res*. 2014;238(2):391–403.
48. Correa N, Adali T, Calhou VD. Canonical correlation analysis for data fusion and group inference: examining applications of medical imaging data. *IEEE Signal Proc Mag*. 2010;27:39–50.
49. Trygg J. O2-PLS for qualitative and quantitative analysis in multivariate calibration. *J Chemometr*. 2002;16(6):283–293.
50. Lofstedt T, Trygg J. OnPLS—a novel multiblock method for the modelling of predictive and orthogonal variation. *J Chemometr*. 2011;25(8):441–455.
51. Lofstedt T. OnPLS. *Ph.D. Thesis*, Umea University, Umea, Sweden, 2012.
52. De Moor B, Zha H. A tree of generalizations of the ordinary singular value decomposition. *Linear Algebra Appl*. 1991;147:469–500.
53. De Moor B. On the structure and geometry of the product singular value decomposition. *Linear Algebra Appl*. 1992;168:95–136.
54. Bookstein FL. Partial least squares: a dose response model for measurement in the behavioral and brain sciences. *Psycoloquy*. 1994;5(23):1.
55. Hanafi M, Kiers HAL. Analysis of *k* sets of data, with differential emphasis on agreement between and within sets. *Comput Stat Data An*. 2006;51(3):1491–1508.
56. Mattarucchi E, Stocchero M, Moreno-Rojas JM, Giordano G, Reniero F, Guillou C. Authentication of trappist beers by LC-MS fingerprints and multivariate data analysis. *J Agr Food Chem*. 2010;58(23):12089–12095.
57. Consonni R, Cagliani LR, Stocchero M, Porretta S. Evaluation of the production year in Italian and Chinese tomato paste for geographical determination using O2PLS models. *J Agr Food Chem*. 2010;58(13):7520, 7525.
58. Kirwan GM, Hancock T, Hassell K, et al. Nuclear magnetic resonance metabonomic profiling using tO2PLS. *Anal Chim Acta*. 201340;781:33.
59. Petrakis EA, Cagliani LR, Polissiou MG, Consonni R. Evaluation of saffron (Crocus sativus L.) adulteration with plant

adulterants by H-1 NMR metabolite fingerprinting. *Food Chem*. 2015;173(890):896.

60. Bylesjo 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(6):1181–1191.
61. Szymanski J, Brotman Y, Willmitzer L, Cuadros-Inostroza A. Linking gene expression and membrane lipid composition of Arabidopsis. *Plant Cell*. 2014;26(3):915–928.
62. Srivastava V, Obudulu O, Bygdell J, et al. OnPLS integration of transcriptomic, proteomic and metabolomic data shows multi-level oxidative stress responses in the cambium of transgenic hipi-superoxide dismutase Populus plants. *BMC Genomics*. 2013;14:893.
63. Lofstedt T, Hoffman D, Trygg J. Global, local and unique decompositions in OnPLS for multiblock data analysis. *Anal Chim Acta*. 2013;791:13–24.
64. Van Loan CF. Generalizing the singular value decomposition. *SIAM J Numer Anal*. 1976;13:76–83.
65. Paige CC, Saunders MA. Towards a generalized singular value decomposition. *SIAM J Numer Anal*. 1981;18(3):398–405.
66. De Lathauwer L. An extension of the generalized SVD for more than two matrices. Internal Report 09-206, Leuven, Belgium, ESAT-SISTA, KU Leuven; 2009.
67. 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(12):e28072.
68. Tibshirani R. Regression shrinkage and selection via the lasso: a retrospective. *J Roy Stat Soc B*. 2011;73:273–282.
69. Lips MA, Van Klinken JB, Van Harmelen V, et al. Roux-en-Y gastric bypass surgery, but not calorie restriction, reduces plasma branched-chain amino acids in obese women independent of weight loss or the presence of type 2 diabetes mellitus. *Diabetes Care*. 2014;37(12):3150–3156.
70. Ray P, Zheng LL, Lucas J, Carin L. Bayesian joint analysis of heterogeneous genomics data. *Bioinformatics*. 2014;30(10):1370–1376.
71. Schonemann PH. A generalized solution of the orthogonal procrustes problem. *Psychometrika*. 1966;31:1–10.
72. Ten Berge JMF. Orthogonal procrustes rotation for 2 or more matrices. *Psychometrika*. 1977;42(2):267–276.

**How to cite this article:** Smilde AK, Måge I, Næs T, et al. Common and distinct components in data fusion. *Journal of Chemometrics*. 2017;31:e2900. https://doi.org/10.1002/cem.2900

## APPENDIX A

## A.1 | Possibilities of orthogonal decompositions

### A.1.1 | Choices for distinct orthogonal (DO) spaces

There are several possibilities for choosing orthogonality in the decompositions of Equation 4. These will be outlined and explained below. We will focus attention on $R(X_{1DO})$, but analogous results hold for $R(X_{2DO})$; we will consider $R(X_{1DNO})$ as a "rest" term and not consider this space explicitly. The first level to discuss possibilities is regarding the status of $R(X_{1D})$. The alternatives are as follows:

A0: no orthogonality restrictions for $R(X_{1D})$;
A1: $R(X_{1D}) \perp R(X_{12C})$ (see Figure 3A);
A2: $R(X_{1D}) \perp R(X_{2D})$ (see Figure 3B); and
A3: $R(X_{1D}) \perp R(X_2)$, which implies A1 and A2

and, as said earlier, alternative A3 is not always possible. The status of $R(X_{1DO})$ is nested in alternatives A0 to A3, since $R(X_{1DO})$ is a part of $R(X_{1D})$.

The alternatives under A0 are as follows:

A01: $R(X_{1DO}) \perp R(X_{2DO})$;
A02: $R(X_{1DO}) \perp R(X_{2D})$; and
A03: $R(X_{1DO}) \perp R(X_2)$,

which shows increasing degrees of orthogonality.

The alternatives under A1 are as follows:

A10: $R(X_{1DO}) \perp R(X_{12C})$, which follows from A1;
A11: $R(X_{1DO}) \perp R(X_{12C})$ and $R(X_{1DO}) \perp R(X_{2DO})$; and
A12: $R(X_{1DO}) \perp R(X_{12C})$ and $R(X_{1DO}) \perp R(X_{2D})$, which implies $R(X_{1DO}) \perp R(X_2)$ and is the same as alternative A03,

which shows again increasing degrees of orthogonality.

The alternatives under A2 are as follows:

A20: $R(X_{1DO}) \perp R(X_{2D})$, which follows from A2 and is the same as alternative A02;
A21: $R(X_{1DO}) \perp R(X_{2D})$ and $R(X_{1DO}) \perp R(X_{12C})$, which is again the same as alternative A03,

and under alternative A3, there is only 1 option, namely, $R(X_{1DO}) \perp R(X_2)$, which is again the same as option A03. In conclusion, for the 2-block case, there are 5 different alternatives to select $R(X_{1DO})$: A01, A02, A03, A10, or A11. Whether these alternatives are available for a specific application depends on the dimensions and positioning of the subspaces. An example of this and how to analyze such situations is presented in Section A.1.2.

### A.1.2 | A specific example

As an example of using rigorous linear algebra results to explore possibilities for (non)orthogonal decompositions consider the example of 2 distinct subspaces both of dimension 2 (see Section 2.1.1). It can be proven that if $R(\mathbf{X}_{1D})$ and $R(\mathbf{X}_{2D})$ are both 2-dimensional and not orthogonal, then for every vector $\mathbf{x}$ in $R(\mathbf{X}_{1D})$ there is exactly 1 vector $\mathbf{y}$ in $R(\mathbf{X}_{2D})$ orthogonal to $\mathbf{x}$. This goes as follows. Suppose that $\mathbf{A}$ and $\mathbf{B}$ are both orthogonal matrices serving as bases for $R(\mathbf{X}_{1D})$ and $R(\mathbf{X}_{2D})$, respectively. Assume also that $r(\mathbf{A}^T\mathbf{B}) = 2$ where $r(.)$ means the rank of a matrix. This implies the following:

- $R(\mathbf{X}_{1D})$ is not orthogonal to $R(\mathbf{X}_{2D})$; otherwise, $\mathbf{A}^T\mathbf{B} = 0$.

- $R(\mathbf{X}_{1D})$ does not contain a vector orthogonal to the whole of $R(\mathbf{X}_{2D})$ (this vector $\mathbf{g}$ could be written as $\mathbf{g} = \mathbf{Ah}$ and then $\mathbf{h}^T\mathbf{A}^T\mathbf{B} = 0$, which contradicts $r(\mathbf{A}^T\mathbf{B}) = 2$).
- $R(\mathbf{X}_{2D})$ does not contain a vector orthogonal to the whole of $R(\mathbf{X}_{1D})$ (analogously as above).

Now there is for any nonzero vector $\mathbf{x} \in R(\mathbf{X}_{1D})$ exactly 1 nonzero vector $\mathbf{y} \in R(\mathbf{X}_{2D})$ such that $\mathbf{x}^T\mathbf{y} = 0$.

*Proof.* Write $\mathbf{x} = \mathbf{Au}$, $\mathbf{y} = \mathbf{Bv}$ and $\mathbf{B} = \mathbf{AU}+\mathbf{A}^{\perp}\mathbf{V}$. Find a vector $\mathbf{v}$ such that $\mathbf{x}^T\mathbf{y} = 0$, or, such that $\mathbf{u}^T\mathbf{A}^T(\mathbf{AU} + \mathbf{A}^{\perp}\mathbf{V})\mathbf{v} = 0$ which equals $\mathbf{u}^T\mathbf{Uv} = 0$ because of the orthogonality of $\mathbf{A}$ and the definition of the orthogonal complement $\mathbf{A}^{\perp}$. Then $\mathbf{U} = \mathbf{A}^T\mathbf{B}$ which follows by premultiplying $\mathbf{B} = \mathbf{AU} + \mathbf{A}^{\perp}\mathbf{V}$ with $\mathbf{A}^T$; by defining $\mathbf{t} = [t_1|t_2]^T = \mathbf{U}^T\mathbf{u}$ which is a unique nonzero vector (because $r(\mathbf{U} = \mathbf{A}^T\mathbf{B}) = 2$) it follows that $\mathbf{v} = [t_2|-t_1]^T$ is the vector which makes $\mathbf{u}^T\mathbf{Uv} = 0$. Hence, there is exactly 1 vector $\mathbf{y} = \mathbf{Bv}$ in $R(\mathbf{X}_{2D})$ which is orthogonal to $\mathbf{x}$. □

## A.2 | GCA proof

In the main text it was stated that the solution $\mathbf{T}$ in Equation 14 is in the column space of $[\mathbf{X}_1|\cdots|\mathbf{X}_K]$. This will be proven now for the 2-block situation for simplicity, but is easily generalized to the more than 2-block situation. Equation 15 can also be written as

$$\sum_{k=1}^{2}\mathbf{X}_k\left(\mathbf{X}_k^T\mathbf{X}_k\right)^{+}\mathbf{X}_k^T = [\mathbf{X}_1|\mathbf{X}_2]\left[\mathbf{X}_1(\mathbf{X}_1^T\mathbf{X}_1)^{+}|\mathbf{X}_2(\mathbf{X}_2^T\mathbf{X}_2)^{+}\right]^T = \mathbf{TST}^T \tag{A1}$$

where the full eigenvalue decomposition (ie, $\mathbf{S} > 0$) is used. Postmultiplying both sides of Equation A1 by $\mathbf{TS}^{-1}$ gives now

$$[\mathbf{X}_1|\mathbf{X}_2][\mathbf{X}_1(\mathbf{X}_1^T\mathbf{X}_1)^{+}|\mathbf{X}_2(\mathbf{X}_2^T\mathbf{X}_2)^{+}]^T\mathbf{TS}^{-1} = \mathbf{T} \tag{A2}$$

or

$$[\mathbf{X}_1|\mathbf{X}_2]\mathbf{Q} = \mathbf{T}, \tag{A3}$$

which shows that $\mathbf{T}$ (and also its first columns when only those are used) is in the range of $[\mathbf{X}_1|\mathbf{X}_2]$. This argument is easily extended to more than 2 blocks.

## A.3 | Relationship between O2PLS and procrustes analysis

An interesting relationship exists between O2PLS and Procrustes Analysis. The Procrustes problem can be stated as

$$\min_{\mathbf{R}^T\mathbf{R}=\mathbf{I}} \|\mathbf{X}_2\mathbf{R} - \mathbf{X}_1\|^2 \tag{A4}$$

and the solution of this problem is $\mathbf{R} = \mathbf{UV}^T$ where $\mathbf{U}$ and $\mathbf{V}$ are from the SVD of $\mathbf{X}_2^T\mathbf{X}_1 = \mathbf{USV}^T$.[71] Then post-multiplying both $\mathbf{X}_2\mathbf{R}$ and $\mathbf{X}_1$ with $\mathbf{V}$ gives $\mathbf{X}_2\mathbf{RV} = \mathbf{X}_2\mathbf{UV}^T\mathbf{V} = \mathbf{X}_2\mathbf{U}$ and $\mathbf{X}_1\mathbf{V}$ which are the same quantities as obtained for O2PLS. Note that the Procrustes problem of Equation A4 is equivalent to

$$\min_{\mathbf{R}_k^T\mathbf{R}_k=\mathbf{I}} \|\mathbf{X}_2\mathbf{R}_2 - \mathbf{X}_1\mathbf{R}_1\|^2 \tag{A5}$$

which is the symmetric formulation of the problem with solution $\mathbf{R}_2 = \mathbf{U}$ and $\mathbf{R}_1 = \mathbf{V}$.[72]