Heterofusion

Smilde, Age K.; Song, Yipeng; Westerhuis, Johan; Kiers, Henk A. L.; Aben, Nanne; Wessels, Lodewyk F. A.

Published in:
Journal of Chemometrics

DOI:
10.1002/cem.3200

Publication date:
2020

Received: 9 May 2019 | Revised: 8 November 2019 | Accepted: 12 November 2019

DOI: 10.1002/cem.3200

SPECIAL ISSUE - RESEARCH ARTICLE

WILEY Journal of CHEMOMETRICS

Heterofusion: Fusing genomics data of different measurement scales

Age K. Smilde[1] | **Yipeng Song**[1] | **Johan A. Westerhuis**[1] | **Henk A. L. Kiers**[2] | **Nanne Aben**[3] | **Lodewyk F. A. Wessels**[3]

[1]Biosystems Data Analysis, Swammerdam Institute for Life Sciences, University of Amsterdam, Amsterdam, The Netherlands

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

[3]Oncode Institute, Netherlands Cancer Institute, Amsterdam, The Netherlands

Correspondence
Age K. Smilde, Biosystems Data Analysis, Swammerdam Institute for Life Sciences, University of Amsterdam, Amsterdam, The Netherlands.
Email: a.k.smilde@uva.nl

Abstract

In systems biology, it is becoming increasingly common to measure biochemical entities at different levels of the same biological system. Hence, data fusion problems are abundant in the life sciences. With the availability of a multitude of measuring techniques, one of the central problems is the heterogeneity of the data. In this paper, we discuss a specific form of heterogeneity, namely, that of measurements obtained at different measurement scales, such as binary, ordinal, interval, and ratio-scaled variables. Three generic fusion approaches are presented of which two are new to the systems biology community. The methods are presented, put in context, and illustrated with a real-life genomics example.

KEYWORDS

data fusion, data integration, heterogeneous data, low-level fusion, measurement scales

1 | INTRODUCTION

1.1 | General

With the availability of comprehensive measurements collected in multiple related data sets in the life sciences, the need for a simultaneous analysis of such data to arrive at a global view on the system under study is of increasing importance. There are many ways to perform such a simultaneous analysis, and these go also under very different names in different areas of data analysis: data fusion, data integration, global analysis, and multiset or multiblock analysis to name a few. We will use the term data fusion in this paper.

Data fusion plays an especially important role in the life sciences; eg, in genomics, it is not uncommon to measure gene expression (array data or RNA-sequencing [RNAseq] data), methylation of DNA, and copy number variation (CNV). Sometimes, also proteomics and metabolomics measurements are available. All these examples serve to show that having methods in place to integrate these data is not a luxury anymore.

1.2 | Types of data fusion

Without trying to build a rigorous taxonomy of data fusion, it is worthwhile to distinguish several distinctions in data fusion. The first distinction is between model-based and exploratory data fusion. The former uses background knowledge in the form of models to fuse the data, one example being genome-scale models in biotechnology.[1] The latter does not rely on models, since these are not available or poorly known, and thus uses empirical modeling to explore the data. In this paper, we will focus on exploratory data fusion.

The next distinction is between low-, medium-, and high-level fusion.[2] In low-level fusion, the data sets are combined at the lowest level, that is, at the level of the (preprocessed) measurements. In medium-level fusion, each separate data set is first summarized, eg, by using a dimension reduction method or through variable selection. The reduced data sets are subsequently subjected to the fusion. In high-level fusion, each data set is used for prediction or classification of an outcome, and the prediction or classification results are then combined, eg, by using majority voting.[3] In machine learning, this is known as early, intermediate, and late integration. All these types of fusions have advantages and disadvantages, which are beyond the scope of this paper. In this paper, we will focus on low- and medium-level fusion.

The final characteristic of data fusion relevant for this paper is heterogeneity of the data sets to be fused. Different types of heterogeneity can be distinguished. The first one is the type of data, such as metabolomics, proteomics, and RNAseq data in genomics. Clearly, these data relate to different parts of the biological system. The second one is the type of measurement scale in which the data are measured that are going to be fused. In genomics, an example is a data set where gene expressions are available and mutation data in the processed form of single-nucleotide variants (SNVs). The latter are binary data, and gene expression is quantitative data. They are clearly measured at a different scale. Ideally, data fusion methods should consider the scale of such measurements, and this will be the topic of this paper.

1.3 | Types of measurement scales

The history of measurement scales goes back a long time. A seminal paper drawing attention to this issue appeared in the 40 ties.[4] Since then, numerous papers, reports, and books have appeared.[5-10] The measuring process assigns numbers to aspects of objects (an empirical system), eg, weights of persons. Hence, measurements can be regarded as a mapping from the empirical system to numbers, and scales are properties of these mappings.

In measurement theory, there are two fundamental theorems[6]: the representation theorem and the uniqueness theorem. The representation theorem asserts the axioms to be imposed on an empirical system to allow for a homomorphism of that system to a set of numerical values. Such a homomorphism into the set of real numbers is called a scale and thus represents the empirical system. A scale possesses uniqueness properties: We can measure the weight of persons in kilograms or in grams, but if one person weighs twice as much as another person, this ratio holds true regardless the measurement unit. Hence, weight is a so-called ratio-scaled variable, and this ratio is unique. The transformation of measuring in grams instead of kilograms is called a permissible transformation since it does not change the ratio of two weights. For a ratio-scaled variable, only similarity transformations are permissible; ie, $\tilde{x} = \alpha x; \alpha > 0$ where x is the variable on the original scale and $\tilde{x}$ is the variable on the transformed scale. This is because

$$\frac{\tilde{x_i}}{\tilde{x_j}} = \frac{\alpha x_i}{\alpha x_j} = \frac{x_i}{x_j}. \tag{1}$$

Note that this coincides with the intuition that the unit of measurement is immaterial.

The next level of scale is the interval-scaled measurement. The typical example of such a scale is degrees Celsius, and the permissible transformation is affine; ie, $\tilde{x} = \alpha x + \beta; \alpha > 0$. In that case, the ratio of two intervals is unique because

$$\frac{\tilde{x_i} - \tilde{x_j}}{\tilde{x_k} - \tilde{x_l}} = \frac{(\alpha x_i + \beta) - (\alpha x_j + \beta)}{(\alpha x_k + \beta) - (\alpha x_l + \beta)} = \frac{\alpha(x_i - x_j)}{\alpha(x_k - x_l)} = \frac{x_i - x_j}{x_k - x_l}. \tag{2}$$

Stated differently, the zero point and the unit are arbitrary on this scale.

Ordinal-scaled variables represent the next level of measurements. Typical examples are scales of agreement in surveys: strongly disagree, disagree, neutral, agree, and strongly agree. There is a rank order in these answers but no relationship in terms of ratios or intervals. The permissible transformation of an ordinal scale is a monotonic increasing transformation since such transformations keep the order of the original scale intact.

Nominal-scaled variables are next on the list. These variables are used to encode categories and are sometimes also called categorical. Typical examples are gender, race, brands of cars, and the like. The only permissible transformation for a nominal-scaled variable is the one-to-one mapping. A special case of a nominal-scaled variable is the binary (0/1) scale. Binary data can have different meanings; they can be used as categories (eg, gender) and are then nominal-scale variables. They can also be two points on a higher-level scale, such as absence and presence (eg, for methylation data).

The above four scales are the most used ones, but others exist.[6,10] Counts, eg, have a fixed unit and are therefore sometimes called absolute-scaled variables.[9] Another scale is the one for which the power transformation is permissible; ie, $\tilde{x} = \alpha x^{\beta}; \alpha, \beta > 0$, which is called a log-interval scale because a logarithmic transformation of such a scale results in

an interval scale. An example is density.[6] Sometimes, the scales are lumped in quantitative (ie, ratio and interval) and qualitative (ordinal and nominal) data.

An interesting aspect of measurement scales is to what extent meaningful statistics can be derived from such scales (see table 1 in Stevens[4]). A prototypical example is using a mean of a sample of nominal-scaled variables, which is generally regarded as being meaningless. This has also provoked a lot of discussion,[5,11] and there are nice counterexamples of apparently meaningless statistics that still convey information about the empirical system.[12] As always, the world is not black or white.

1.4 | Motivating example

Examples of fusing data of different measurement scales are abundant in modern life science research. We will first give a short description of modern measurements in genomics that will illustrate this. In a sample extracted from biological systems (eg, cells), it is possible to measure the mRNA molecules. This is done nowadays with RNAseq techniques, and in essence, the mRNAs are counts per volume, hence a concentration. Epigenetics concerns, among others, the methylation of some of the sites of a DNA molecule and is in essence a binary variable (yes/no methylated at a given location of the DNA). Another feature in genetics is whether a location of the DNA is mutated, a phenomenon called SNVs, which is also binary. Lastly, there are CNVs of genes on the genome, which is in essence a (limited) number of counts and sometimes expressed as copy number abberations (CNAs) with a binary coding (no: normal number of copies, yes: aberrant number of copies). If we move to the field of metabolomics and proteomics, then most of the measurements are relative intensities and in some cases—when calibration lines have been made—concentrations, which are ratio scaled.

The above exposition clearly shows that if we want to fuse different types of genomics data or fuse genomics data with metabolomics and/or proteomics, then there is a problem of different measurement scales. This problem is aggravated by the fact that some of these data are very high dimensional. Single-nucleotide polymorphism (SNP) and methylation data can contain 100.000 features or variables, and RNAseq data have usually around 20.000 genes. Shotgun proteomics data (based on liquid chromatography—mass spectrometry [LC-MS] or liquid chromatography with tandem mass spectrometry [LC-MS/MS]) can also easily contain 100.000 features. Hence, in many cases, dimension reduction has to take place, asking for methods to deal properly with the corresponding measurement scale. For some of the methods to be discussed in this paper, there are already examples in the literature. There are examples of the use of the parametric approach using latent variables [13,14] and also of the optimal scaling approach.[15,16] For the third approach to be discussed, we have not found examples yet in the life sciences. We will come back to these examples in Section 4.

1.5 | Goal of the paper

In this paper, we describe low- and mid-level fusion ideas of data of different measurement scales. We will restrict ourselves to data sharing the object mode. Mid-level fusion first selects variables and then is subjected to the methods described below. These methods can be applied in different fields of science, but we will illustrate them by using a genomics example.

We think this paper is needed since the different methods originate from different fields of data analysis, psychometrics, and bioinformatics with limited cross talk between those fields; we will try to fill this gap. Moreover, there are relationships between the methods, and this might help in selecting the proper method for a particular application. Hence, we will also discuss the properties of the different methods.

We will select and discuss methods that provide coordinates of the objects that can be used for plotting and visualizing the relationships between the objects. Moreover, we think it is also worthwhile to consider methods that generate importance values for the variables in the different blocks since this will facilitate interpretation of the results in substantive terms.

2 | THEORY

2.1 | Three basic ideas

We will describe three basic ideas that can be used for fusion of data of different measurement scales on a conceptual level. A more detailed explanation is given in following sections. One of these methods is parametric and thus depends on distributions.[13] The other two methods are nonparametric and based on concepts of representation matrices[17,18] and optimal scaling. [19]

The first idea[17] is illustrated in Figure 1. Supposing we have three blocks of data, the first block ($\mathbf{X}_1$) contains ratio-scaled data, the second block ($\mathbf{X}_2$) binary data, and the third block ($\mathbf{X}_3$) categorical data with each of the J_3 variables having four categories (labeled A, B, C, and D). Each variable in each block is represented by an $I \times I$ representation matrix (to be

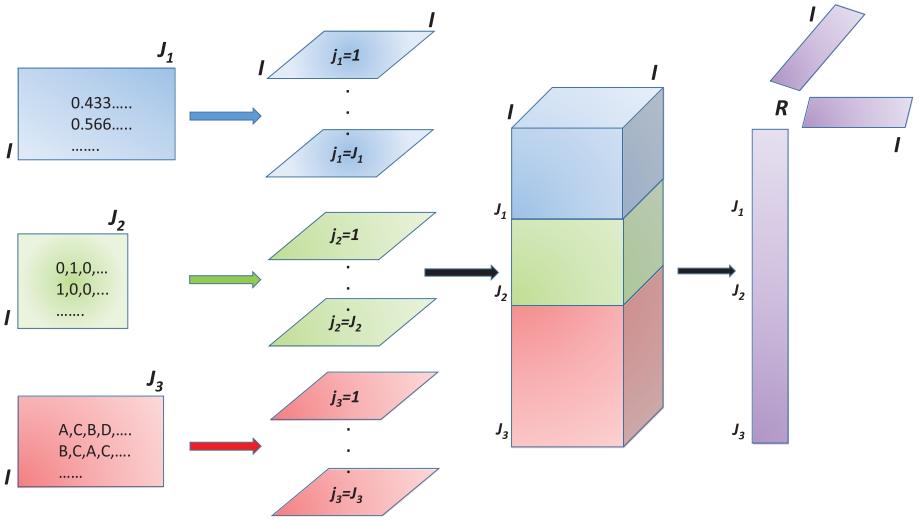

FIGURE 1 Heterofusion using representation matrices (see text)

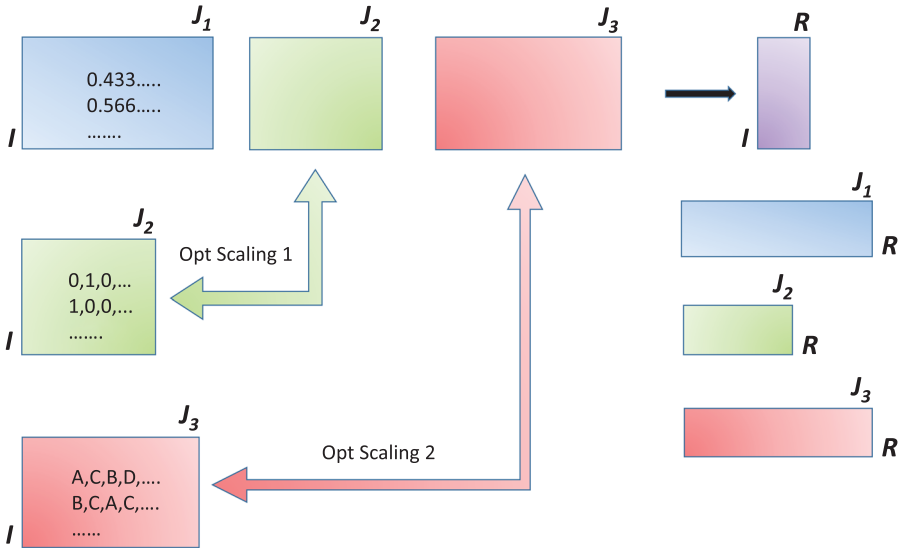

FIGURE 2 Heterofusion using optimal scaling (see text)

explained later). Then, these representation matrices can be stacked, and the resulting three-way array can be analyzed by a suitable three-way method using R components giving coordinates for the objects and weights for the variables.

The second idea[19,20] is illustrated in Figure 2. The original matrices are subjected to optimal scaling, and the fusion problem is solved as one global optimization problem (to be explained later). The idea of optimal scaling goes back already to R. Fisher and nice introductions are available.[21] For the first block, the variables remain the same, but for the second and third blocks, these variables are (optimally) transformed. Using optimal scaling, the three blocks are made comparable and are analyzed simultaneously by a multiblock method (eg, simultaneous component analysis or consensus principal component analysis [PCA]) giving R coordinates for the objects (the $I \times R$ matrix) and loadings (the $J_1 \times R$, $J_2 \times R$, and $J_3 \times R$ matrices) for the transformed variables.

The third idea relies on the explicit use of the R latent variables collected in $\mathbf{Z}$ (see Figure 3).[13] These latent variables are then thought to generate the manifest variables in the different blocks using different distributions. For the ratio-scaled block, a regression model is assumed based on the normal distribution and with parameters $\boldsymbol{\alpha_{j1}}$ and $\boldsymbol{\beta_{j1}}$. For the binary block, a logit or probit model is assumed with parameters $\boldsymbol{\alpha_{j2}}$ and $\boldsymbol{\beta_{j2}}$. The final—categorical—block is modeled by a multilogit model with parameters $\boldsymbol{\alpha_{j3c}}$ and $\boldsymbol{\beta_{j3c}}$ where $c = A, B, C, D$.

We will use the following conventions for notations. A vector $\mathbf{x}$ is a bold lowercase, and a matrix ($\mathbf{X}$), a bold uppercase. Running indices will be used for samples (i=1, … ,I) with I is the number of samples; we will use likewise the indices k=1, … ,K for the data blocks; variables within a data block are indexed by j_k=1, … ,J_k, and we will use r=1, … , R as an index for latent variables or components.

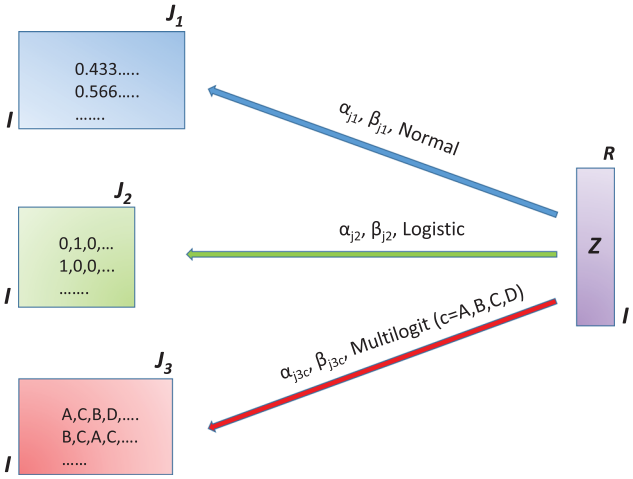

FIGURE 3 Heterofusion using parametric models (see text)

2.2 | Representation matrices approaches

2.2.1 | Representation matrices

Idea of representation matrices

Suppose we have a data matrix $\mathbf{X}(I \times J)$ with columns $\mathbf{x}_j$ containing the measurements of the objects on variable j. Such measurements can be a ratio-scaled value but can also be a binary value, a categorical value, or an ordinal-scaled value. A representation operator works on this vector and produces a representation matrix, which serves as a building block to calculate associations between variables and to analyze several variables simultaneously.[17,18] Such a representation matrix can be a vector $(I \times 1)$, a rectangular matrix $(I \times R; R < I)$, or a square matrix $(I \times I)$. Let $\mathbf{S}_j$ and $\mathbf{S}_k$ be the representation matrices for variables j and k, respectively. Then, a general equation of the association between variables j and k is

$$q_{jk} = \frac{2tr(\mathbf{S}_j^T \mathbf{S}_k)}{tr(\mathbf{S}_j^T \mathbf{S}_j) + tr(\mathbf{S}_k^T \mathbf{S}_k)}, \tag{3}$$

where the symbol tr is used to indicate the trace of a matrix. In most cases that follow below, the representation matrices are standardized (centered and scaled to length one*), and in these cases, Equation (3) simplifies to

$$\tilde{q}_{jk} = tr(\mathbf{S}_j^T \mathbf{S}_k), \tag{4}$$

since both $tr(\mathbf{S}_j^T \mathbf{S}_j)$ and $tr(\mathbf{S}_k^T \mathbf{S}_k)$ are one. As will be shown in the following, Equation (4) can generate the familiar associations such as the Pearson correlation or the Spearman correlation. An extensive description of all kinds of representation matrices is beyond the scope of this paper; we will discuss the most relevant ones for the problem of heterofusion. The idea of representation matrices† goes back to the work of Janson and Vegelius[22] and Zegers.[18] Examples of different representation matrices are given in Appendix A.

Representation matrices for ratio- and interval-scaled values

For ratio- and interval-scaled values, two types of representation matrices can be defined: vectors and square matrices. If $\mathbf{x}_j$ represents the raw measurements of the objects on variable j, then the vector quantification can be this vector itself (ie, $\mathbf{s}_j = \mathbf{x}_j$) or a standardized version of it. When the latter is used in Equation (4), Pearson's R value is obtained. In standard multivariate analysis, this is by far the most used representation matrix.

There is also another possibility for ratio- and interval-scaled values, namely, square representation matrices. Two examples are the following. Define

$$\tilde{\mathbf{S}}_j = (\mathbf{x}_j \mathbf{1}^T - \mathbf{1}\mathbf{x}_j^T), \tag{5}$$

where $\mathbf{1}$ is an $I \times 1$ column of ones. This $\tilde{\mathbf{S}}_j$ generates a skew-symmetric matrix enumerating all differences between the object scores of variable j (for an example, see A.2). Hence, distances between objects are obtained per variable, and

*An alternative is scaling to variance 1, but this only differs with the same constant for each variable.
†Their original name was quantification matrices, but that name has also been used differently. Hence, our choice to rename such matrices.

these distance matrices can be subjected to an INdividual Differences SCALing (INDSCAL) model.[17] Upon standardizing $\tilde{\mathbf{S}}_j$ by $\mathbf{S}_j = (tr\tilde{\mathbf{S}}_j^T\tilde{\mathbf{S}}_j)^{-1/2}\tilde{\mathbf{S}}_j$ and using this $\mathbf{S}_j$ (and a similarly defined $\mathbf{S}_k$) in Equation (4) gives again Pearson's R value. Another example is using $\mathbf{S}_j = \mathbf{s}_j\mathbf{s}_j^T$ where $\mathbf{s}_j$ is the standardized version of $\mathbf{x}_j$. Using this $\mathbf{S}_j$ (and a similarly defined $\mathbf{S}_k$) in Equation (4) gives Pearson's R^2 value. Such representation matrices correspond to the blue-squared matrices in Figure 1 and are the basis of kernel and multidimensional scaling methods. Note that it is also possible to calculate distance matrices between the I samples of $\mathbf{X}$ across all variables resulting in an $I\times I$ distance matrix. The crucial difference with the representation approach is that this distance matrix is calculated across all variables and thus, variable-specific information is lost.

Representation matrices for ordinal-scaled values

When the data are ordinal scaled, then the vector of readings can be encoded in terms of rank-orders $\mathbf{r}_j(I\times 1)$. For the earlier example of strongly disagree, disagree, neutral, agree, and strongly agree such a ranking may be encoded as 1 (strongly disagree) to 5 (strongly agree). Then, again, as in the ratio-scaled variables, representation can be done using the vector $\mathbf{r}_j$ or their standardized version. In the latter case, applying Equation (4) to this version gives Spearman's rank-order correlation coefficient. Another representation is by using (the raw) $\mathbf{r}_j$ in Equation (5) instead of $\mathbf{x}_j$, and this generates Spearman's rank-order correlation coefficient after using Equation (3).

Representation matrices for nominal-scaled values

We will discuss the representation matrices for nominal-scaled variables separately for binary data and categorical data. We first discuss representation matrices for categorical data. We have to distinguish two situations: one in which all categorical variables have the same number of categories and the situation that this is not the case. Since the latter is more general and encountered more often, we will restrict ourselves to this case. Then, only square representation matrices are available. These are based on indicator matrices.[17-19] If variable $\mathbf{x}_j$ has four categories (A, B, C, and D), then this can be encoded in the rectangular matrix $\mathbf{G}_j(I\times 4)$ where each column $\mathbf{g}_{jk}$ in $\mathbf{G}_j$ represents a category and each row an object. This matrix has only zeros or ones, where g_{ijk} is one, if and only if object i belongs to the category represented by k. The representation matrix can now be built using the products $\mathbf{G}_j\mathbf{G}_j^T(I\times I)$. There are very many versions of such square representation matrices based on indicator matrices, and some of them give rise to a known correlation, eg,

$$\mathbf{J}\mathbf{G}_j\mathbf{D}_j^{-1}\mathbf{G}_j^T\mathbf{J}, \tag{6}$$

where $\mathbf{J}(I\times I)$ is the centering operator $(\mathbf{I}-\mathbf{1}\mathbf{1}^T/I)$ with $\mathbf{I}$ being the unity matrix and $\mathbf{1}$ a column of ones and $\mathbf{D}_j(C_j\times C_j)$ is a diagonal matrix containing the marginal frequencies of categories $1,\ \ldots,C_j$ for variable j. The corresponding correlation coefficient is the so-called T^2 coefficient.[23] These representation matrices correspond to the red-square matrices in Figure 1. Examples are given in Appendix A.3.

For binary data (if all variables are binary), then rectangular representation matrices are possible. This comes down to the same idea as above, namely, to consider the binary variables as representing two categories. This results then in representation matrices $\mathbf{G}_j$ of sizes $(I\times 2)$. When fusing with other types of variables is the goal, then a squared type of representation is needed such as in Equation (6) and visualized in Figure 1 (green matrices). Examples are given in Appendix A.4.

2.2.2 | Data fusion using representation matrices

To illustrate how to use representation matrices, we will work with four data matrices, each on a different measurement scale and that share the same set of I samples. The first matrix $\mathbf{X}_1(I\times J_1)$ contains binary data; the second matrix $\mathbf{X}_2(I\times J_2)$ contains ordinal-scaled data; the third $\mathbf{X}_3(I\times J_3)$ contains nominal data; and the last matrix $\mathbf{X}_4(I\times J_4)$ contains ratio- or interval-scaled data.

The representation matrices $\mathbf{S}_j$ can now be used in a three-way model for symmetric data. The basic model for a single data block is the INDSCAL model:

$$\min_{\mathbf{Z},\mathbf{A}_j}\sum_{j=1}^{J}||\mathbf{S}_j-\mathbf{Z}\mathbf{A}_j\mathbf{Z}^T||^2, \tag{7}$$

where $\mathbf{A}_j$ is the diagonal matrix with the jth row of the loadings $\mathbf{A}(J\times R)$ on its diagonal and the matrix $\mathbf{Z}(I\times R)$ contains the object scores. The loadings $\mathbf{A}(J\times R)$ are nonnegative to ensure the fitted part of the model $(\mathbf{Z}\mathbf{A}_j\mathbf{Z}^T)$ to be positive (semi-) definite. If the additional constraint that $\mathbf{Z}^T\mathbf{Z}=\mathbf{I}$ is used, then the model is called INDORT (INDscal with ORThogonal constraints).[17]

The INDORT method can now be generalized to analyze simultaneously all blocks by simply stacking all similarity matrices on top of each other (see Figure 1):

$$\min_{\mathbf{Z},\mathbf{A}_{j_k}} \sum_{k=1}^{4} \sum_{j_k=1}^{J_k} ||\mathbf{S}_{j_k} - \mathbf{Z}\mathbf{A}_{j_k}\mathbf{Z}^T||^2, \tag{8}$$

where $\mathbf{A}_{j_k}$ is the diagonal matrix with the j_kth row of the loadings $\mathbf{A}_k(J_k \times R)$ on its diagonal and the matrix $\mathbf{Z}(I \times R)$ contains the object scores. This model is called IDIOMIX for obvious reasons.[17]

2.3 | Optimal scaling approaches

There are many ways to explain optimal scaling; we will follow the exposition given by Michailidis and de Leeuw.[20] Suppose that the matrix $\mathbf{X}(I \times J)$ contains J categorical variables not necessarily with the same number of categories. Each variable $\mathbf{x}_j$ can now be encoded with indicator matrix $\mathbf{G}_j(I \times L_j)$ where L_j is the number of categories for variable j as discussed before. The idea of optimal scaling is to find objects scores $\mathbf{Z}(I \times R)$ and category quantification matrices $\mathbf{Y}_j(L_j \times R; j = 1, \ldots, J)$ such that the following problem is solved[20]:

$$\min_{\mathbf{Z},\mathbf{Y}_j} \sum_{j=1}^{J} ||\mathbf{Z} - \mathbf{G}_j\mathbf{Y}_j||^2, \tag{9}$$

under the constraints that $(1/I)\mathbf{Z}^T\mathbf{Z} = \mathbf{I}$ and these scores are centered around 0 (to avoid trivial solutions of Equation 9). This method—including the alternating optimization method to solve Equation (9)—is called homogeneity analysis or HOMALS for short.[19] The rows of $\mathbf{Z}$ give a low-dimensional representation of the objects, and the matrices $\mathbf{Y}_j(j = 1, \ldots, J)$ give the optimal quantifications of the categorical variables. Note that these matrices $\mathbf{Y}_j(j = 1, \ldots, J)$ are not loadings; they give quantifications for the categorical variables, which are different for the R components, namely, $\mathbf{y}_{jr}(L_j \times 1; r = 1, \ldots, R)$ where $\mathbf{y}_{jr}$ is the rth column of $\mathbf{Y}_j$.

Upon restricting the rank of $\mathbf{Y}_j(j = 1, \ldots, J)$ to be 1, we arrive at nonlinear PCA (PRINCALS).[19,20] Then, Equation (9) can be rewritten as

$$\min_{\mathbf{Z},\mathbf{y}_j,\mathbf{a}_j} \sum_{j=1}^{J} ||\mathbf{Z} - \mathbf{G}_j\mathbf{y}_j\mathbf{a}^T{}_j||^2, \tag{10}$$

with the same constraints on $\mathbf{Z}$ as before (ie, $(1/I)\mathbf{Z}^T\mathbf{Z} = \mathbf{I}$). As an identification constraint for $\mathbf{y}_j$ and $\mathbf{a}_j$, we impose $\mathbf{y}_j^T\mathbf{G}_j^T\mathbf{G}_j\mathbf{y}_j = I$. Now, the vectors $\mathbf{a}_j(R \times 1)$ are the loadings, and $\mathbf{y}_j(L_j \times 1)$ contain the quantifications, which are the same for all R dimensions of the solution. The relationship between (linear) PCA and nonlinear PCA becomes clear when rewriting Equation (10) (following Gifi,[19] pp. 167-168) as

$$\begin{aligned} &\min_{\mathbf{Z},\mathbf{y}_j,\mathbf{a}_j} \sum_{j=1}^{J} ||\mathbf{Z} - \mathbf{G}_j\mathbf{y}_j\mathbf{a}_j^T||^2 = \\ &\min_{\mathbf{Z},\mathbf{y}_j,\mathbf{a}_j} \sum_j tr(\mathbf{Z}^T\mathbf{Z}) - 2\sum_j tr(\mathbf{Z}^T\mathbf{G}_j\mathbf{y}_j\mathbf{a}_j^T) + \sum_j tr(\mathbf{a}_j\mathbf{y}_j^T\mathbf{G}_j^T\mathbf{G}_j\mathbf{y}_j\mathbf{a}_j^T) = \\ &\min_{\mathbf{Z},\mathbf{y}_j,\mathbf{a}_j} IJtr\mathbf{I} - 2\sum_j tr(\mathbf{Z}^T\mathbf{G}_j\mathbf{y}_j\mathbf{a}_j^T) + I\sum_j tr(\mathbf{a}_j\mathbf{a}_j^T), \end{aligned} \tag{11}$$

using the constraints on $\mathbf{Z}$ and $\mathbf{y}_j$. The function in Equation (2.3) differs only a constant from the function

$$g(\mathbf{Z}, \mathbf{y}_j, \mathbf{a}_j) = \sum_j ||\mathbf{G}_j\mathbf{y}_j - \mathbf{Z}\mathbf{a}_j||^2, \tag{12}$$

as follows from rewriting $g(\mathbf{Z}, \mathbf{y}_j, \mathbf{a}_j)$ using the constraints on $\mathbf{Z}$ and $\mathbf{y}_j$:

$$\begin{aligned} &g(\mathbf{Z}, \mathbf{y}_j, \mathbf{a}_j) = \\ &\sum_j \mathbf{y}_j^T \mathbf{G}_j^T \mathbf{G}_j \mathbf{y}_j - 2 \sum_j tr(\mathbf{a}_j^T \mathbf{Z}^T \mathbf{G}_j \mathbf{y}_j) + \sum_j tr(\mathbf{a}_j^T \mathbf{Z}^T \mathbf{Z} \mathbf{a}_j) = \\ &IJ - 2 \sum_j tr(\mathbf{Z}^T \mathbf{G}_j \mathbf{y}_j \mathbf{a}_j^T) + I \sum_j tr(\mathbf{a}_j^T \mathbf{a}_j). \end{aligned} \tag{13}$$

Thus, it has been shown that problem Equation (10) subject to the constraints $(1/I)\mathbf{Z}^T\mathbf{Z} = \mathbf{I}$ and $\mathbf{y}_j^T \mathbf{G}_j^T \mathbf{G}_j \mathbf{y}_j = I$ is equivalent to the problem

$$\begin{aligned} &\min_{\mathbf{Z}, \mathbf{y}_j, \mathbf{a}_j} \sum_{j=1}^{J} \|\mathbf{G}_j \mathbf{y}_j - \mathbf{Z}\mathbf{a}_j\|^2 = \\ &\min_{\mathbf{Z}, \mathbf{y}_j, \mathbf{a}_j} \|[\mathbf{G}_1 \mathbf{y}_1 | \ldots | \mathbf{G}_J \mathbf{y}_J] - \mathbf{Z}\mathbf{A}^T\|^2 = \\ &\min_{\mathbf{Z}, \mathbf{y}_j, \mathbf{a}_j} \|\mathbf{X}^* - \mathbf{Z}\mathbf{A}^T\|^2, \end{aligned} \tag{14}$$

where $\mathbf{A}$ has rows $\mathbf{a}_j^T$ and $[\mathbf{G}_1\mathbf{y}_1 | \ldots | \mathbf{G}_J\mathbf{y}_J]$ is written as $\mathbf{X}^*$ where the superscript "*" represents the optimal scaled data and this is seen to be the (nonlinear) analog of ordinary PCA.[19]

The nature of the measurement scale can now be incorporated by allowing the quantifications to be free for nominal-scale data and monotonic for ordinal-scaled data. The latter quantification ensures the order in the ordinal-scaled data. Framed in terms of Equation (2.3), this becomes

$$x^*_{ij} > x^*_{kj} \quad if \quad x_{ij} > x_{kj}, \tag{15}$$

where x^*_{ij} and x^*_{kj} are elements of $\mathbf{X}^*$ and x_{ij} and x_{kj} are elements of $\mathbf{X}$. Ties in the original data can be treated in different ways depending on whether the underlying measurements can be considered continuous or discrete,[24-26] but this is beyond the scope of this paper.

There are close similarities between optimal scaling and multiple correspondence analysis.[17,20] Binary data represent a special case. When considered as categorical data, nonlinear PCA using optimal scaling is the same as performing a (linear) PCA on the standardized binary data; for a proof, see Appendix A. [17,27]

2.3.1 | Data fusion using optimal scaling matrices

There are different ways to use optimal scaling for fusing data. One method generalizes (generalized) canonical correlation analysis (OVERALS[28]), and the other method generalizes simultaneous component analysis (SCA) (MORALS[21]). Experiences with generalized canonical correlations show that this method tends to overfit for high-dimensional data. An attempt to overcome this problem is by introducing sparsity constraints,[29] but it is not trivial to combine this with optimal scaling. Hence, we chose to use the extension of SCA. Note that SCA was originally developed for analyzing multiple data sets sharing the same set of variables,[30] but it can likewise be formulated for multiple data sets having the sampling mode in common.[31] Using the latter interpretation of SCA leads to the following approach.

We take the same data matrices as in Section 2.2.2, and upon writing $\mathbf{X}^* = [\mathbf{X}^*{}_1 | \mathbf{X}^*{}_2 | \mathbf{X}^*{}_3 | \mathbf{X}^*{}_4]$, the problem becomes

$$\min_{Par} ||\mathbf{X}^* - \mathbf{Z}\mathbf{A}^T||^2 = \min_{Par} ||[\mathbf{X}^*{}_1 | \mathbf{X}^*{}_2 | \mathbf{X}^*{}_3 | \mathbf{X}^*{}_4] - \mathbf{Z}[\mathbf{A}^T{}_1 | \mathbf{A}^T{}_2 | \mathbf{A}^T{}_3 | \mathbf{A}^T{}_4]||^2, \tag{16}$$

with an obvious partition of the loading matrix $\mathbf{A}$ and where the term *Par* stands for all parameters. Apart from the scores $\mathbf{Z}$ and loadings $\mathbf{A}$, these are the following. For the ratio, interval-scaled block, there are no extra parameters since the original scale is used (ie, $\mathbf{X}^*{}_1 = \mathbf{X}_1$). The second–ordinal scaled—block puts restrictions on $\mathbf{X}^*$ following the restrictions of Equation (15) estimation of which can be done with monotonic regression methods.[32] The third (nominal) block has underlying indicator matrices $\mathbf{G}$ and associated quantifications $\mathbf{y}$, and loadings $\mathbf{A}_3$ obey the rules of Equation (10). Finally, the binary block $\mathbf{X}_4^*$ is simply the standardized version of $\mathbf{X}_4$, and this ensures an optimal scaling as mentioned above. Note that this way of fusing data assumes an identity link function[33] and is thus an extension of methods like consensus

PCA and SCA. We will call this method OS-SCA in the sequel. There is no differentiation between common and distinct components [34]

2.4 | Parametric approaches

A different class of methods has its roots in factor analysis and can be summarized as follows (see Figure 3). The basic idea is that a set of (shared) latent variables is responsible for the variation in all the blocks[13,14,35] and, subsequently, models are built for the individual blocks based on those shared latent variables. We will describe the generalized simultaneous component analysis (GSCA) method[36] in more detail since that is the method used in this paper. If $\mathbf{X}_1$ is the binary data matrix, then we assume that there is a low-dimensional deterministic structure $\Theta_\mathbf{1}(I\times J_1)$ underlying $\mathbf{X}_1$ and the elements of $\mathbf{X}_1$ follow a Bernoulli distribution with parameters $\phi(\theta_{1ij})$, thus $x_{1ij} \sim B(\phi(\theta_{1ij}))$. The function $\phi(.)$ can be taken as the logit link $\phi(\theta) = (1 + exp(-\theta))^{-1}$, and x_{1ij}, θ_{1ij} are the ijth elements of $\mathbf{X}_1$ and $\Theta_\mathbf{1}$, respectively. The $\Theta_\mathbf{1}$ is now assumed to be equal to $\mathbf{1}\boldsymbol{\mu}_1^T + \mathbf{Z}\mathbf{A}_1$, where μ_1 represents the offset term, $\mathbf{Z}$ the common scores, and $\mathbf{A}_1$ the loadings for the binary data.

The quantitative measurements $\mathbf{X}_2$ are assumed to follow the model $\mathbf{X}_2 = \mathbf{1}\boldsymbol{\mu}_2^T + \mathbf{Z}\mathbf{A}_2 + \mathbf{E}$ where the elements e_{ij} of $\mathbf{E}$ are normally distributed with mean 0 and variance σ^2. The matrix $\mathbf{A}_2$ contains the loadings of the quantitative data set; $\mathbf{Z}$ are again the common scores; and the constraints $\mathbf{Z}^T\mathbf{Z} = I\mathbf{I}_R$ and $\mathbf{1}^T\mathbf{Z} = 0$ are imposed for identifiability. The shared information between $\mathbf{X}_1$ and $\mathbf{X}_2$ is assumed to be represented fully by the common latent variables $\mathbf{Z}$. Thus, $\mathbf{X}_1$ and $\mathbf{X}_2$ are stochastically independent given these latent variables, and the negative log-likelihoods of both parts can be summed:

$$
\begin{aligned}
f_1(\Theta_1) &= -\sum_i^I \sum_j^{J_1} [x_{1ij} log(\phi(\theta_{1ij})) + (1 - x_{1ij}) log(1 - \phi(\theta_{1ij})], \\
f_2(\Theta_2, \sigma^2) &= \frac{1}{2\sigma^2} \|\mathbf{X}_2 - \Theta_2\|_F^2 + \frac{1}{2} log(2\pi\sigma^2), \\
f(\Theta_1, \Theta_2, \sigma^2) &= f_1(\Theta_1) + f_2(\Theta_2, \sigma^2),
\end{aligned}
\tag{17}
$$

and minimized simultaneously. This requires some extra constraints; details are given elsewhere.[36]

3 | PRACTICAL ISSUES AND EXAMPLES

3.1 | Genomics example

The genomics example is from the field of cancer research, and the data are obtained from the Genomics in Drugs Sensitivity in Cancer from the Sanger Institute (http://www.cancerrxgene.org/). Briefly, this repository consists of measurements performed on cell lines pertaining to different types of cancer. We used the CNA and gene-expression data of the cell lines related to breast cancer (BRCA), lung cancer (LUAD), and skin cancer (SKCM). After the samples, which had values for all these types of cancer, are selected, we filtered the gene-expression data by selecting the 1000 variables with the highest variance across the samples. The CNA data contain amplifications and losses of DNA regions as compared with the average copy numbers in the population. Both amplifications and losses are encoded as ones indicating deviances. The zeros in the CNA data indicate a normal diploid copy number. This provides us with $I = 160$ samples; $J_1 = 410$ binary values for the CNA data; and $J_2 = 1000$ variables for the gene-expression data.

For the representation approach, we built a three-way array of size $160\times160\times(410+1000)$ and performed an IDIOMIX analysis. For the binary part, this array contains the slabs $\mathbf{S}_j$ according to Equation (6), and for the gene-expression part, the slabs $\mathbf{S}_j$ are defined by the outer products of the samples in the gene-expression data after autoscaling the columns of that data. The optimal scaling results are obtained by autoscaling both raw data sets and subsequently perform an OS-SCA on the concatenated data $\mathbf{X}_{sc} = [\mathbf{X}_{1sc}|\mathbf{X}_{2sc}]$. The final way of fusing the two data sets is by using the GSCA model where the autoscaled genomics data are fused with the logit-transformed binary data as explained in Section 2.4.

The amounts of explained variations are shown in Table 1, which contains a lot of information and should be interpreted with care. First, for IDIOMIX, OS-SCA, and the quantitative part of the GSCA model, the explained variation is calculated using sums of squares. This is not the case for the binary part of GSCA (for details, see Song et al[36]). Second, IDIOMIX on the one hand and OS-SCA and GSCA on the other hand are very different types of models; ie, they use the data directly (OS-SCA, GSCA) or indirectly (IDIOMIX) so a simple comparison of explained sums of squares between these types of models is difficult. The final column of the table reports the amounts of explained variation of a regular PCA on the (autoscaled) gene-expression data.

TABLE 1 Variances explained by the various methods (expressed as %)

Method	IDIOMIX			OS-SCA			GSCA			PCA
Data Type	Binary	Quant	Total	Binary	Quant	Total	Binary	Quant	Total	Quant
SC1	0.06	9.32	6.60	13.65	22.15	16.14	63.58	22.11	23.64	22.15
SC2	0.03	2.38	1.69	5.66	10.17	7.48	15.72	10.19	10.39	10.17
SC3	3.73	0.01	1.10	4.99	4.52	4.35	6.17	4.48	4.54	4.52
Cum	3.82	11.70	9.39	24.29	36.84	27.96	85.47	36.77	38.58	36.84

Abbreviations: GSCA, generalized SCA; PCA, principal component analysis; SC, simultaneous component; SCA, simultaneous component analysis.

The first observation to make regarding the values in Table 1 is that the amounts of explained variations of the PCA model of the gene-expression data is closely followed by the amounts of explained variations in the gene-expression simultaneous components for OS-SCA and GSCA. This means that the data fusion is dominated by the gene-expression block. This is confirmed by plotting the scores of PC1 and PC2 of the PCA on gene expression against the SC scores 1 and 2 of OS-SCA and GSCA: These are almost on a straight line (plot not shown). Although the explained variances for IDIOMIX are much lower, the same observation is valid for IDIOMIX: Also, for this method, the first two SC scores resemble the ones of a PCA on the gene expression almost perfectly. This dominance of the gene-expression block in the data fusion as reflected in the first two components cannot completely be explained by the differences in block sizes (1000 variables for gene expression and 410 variables for the CNA block) but is also due to dominant intrinsic systematic patterns in the gene-expression data.

To get a feeling for what is represented in the first two SCs (that are virtually identical across the three methods), we show the scores for the GSCA method on SC1 and SC2 in Figure 4. The scores show a clear separation in cancer types

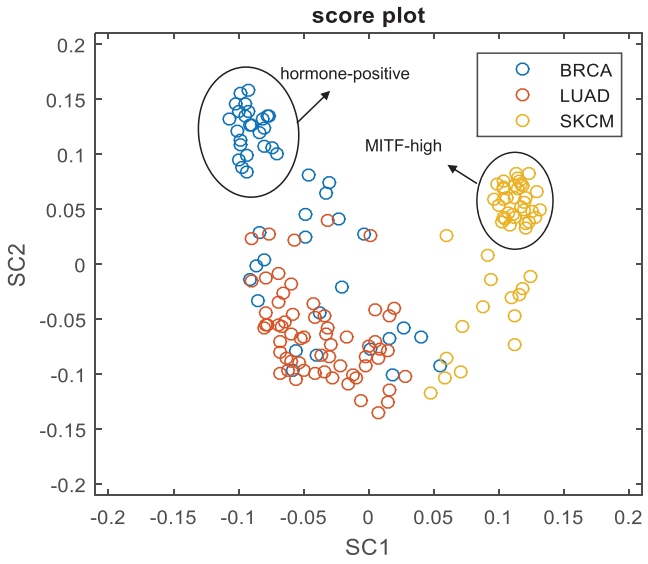

FIGURE 4 Scores on SC1 and SC2 for the GSCA model (see text)

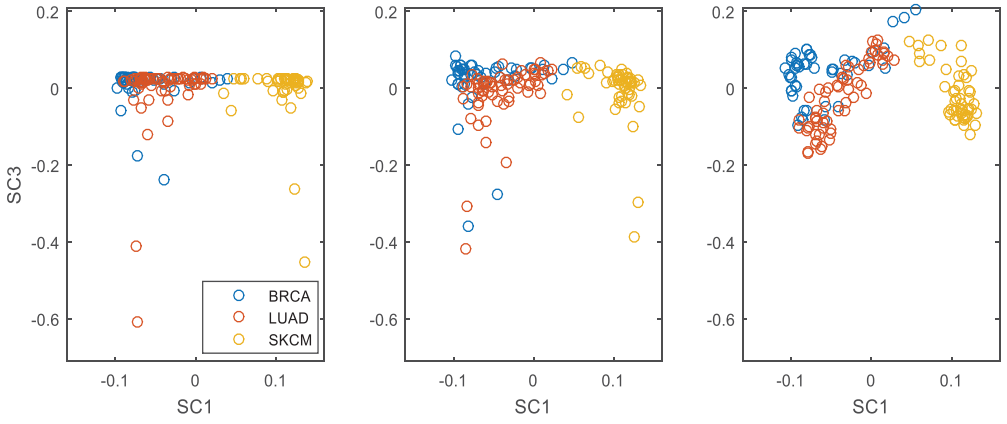

FIGURE 5 Score plots of SC1 versus SC3 of in the genomics example using all models (see text). Left: IDIOMIX; middle: OS-SCA; right: GSCA

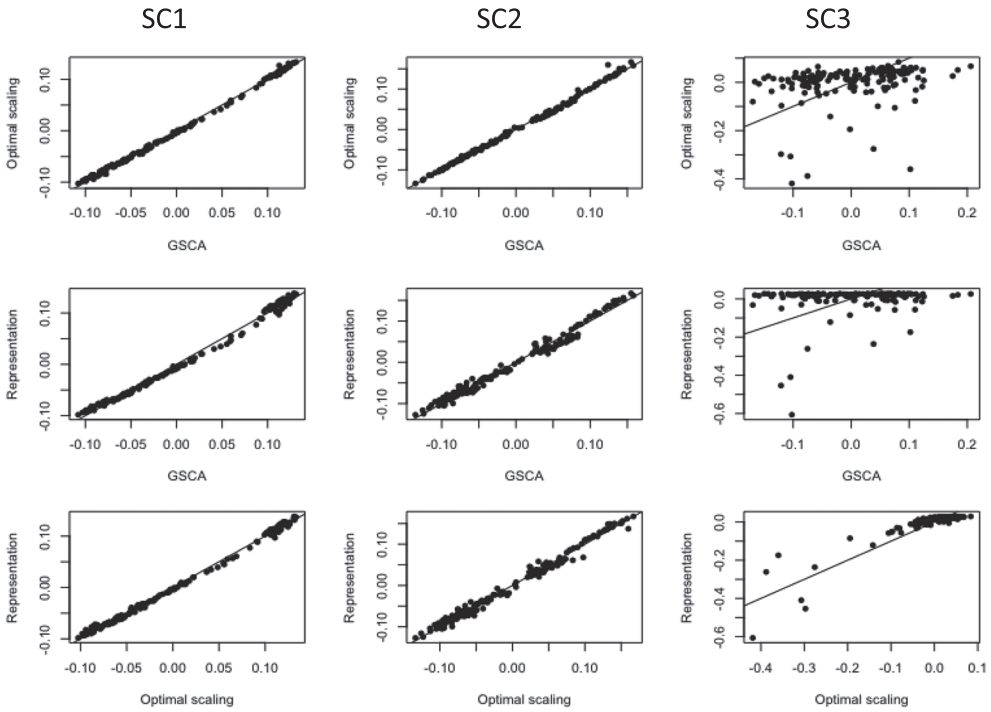

FIGURE 6 Score plots of SC1-SC3 for all fusion methods. Optimal scaling is OS-SCA; representation is IDIOMIX

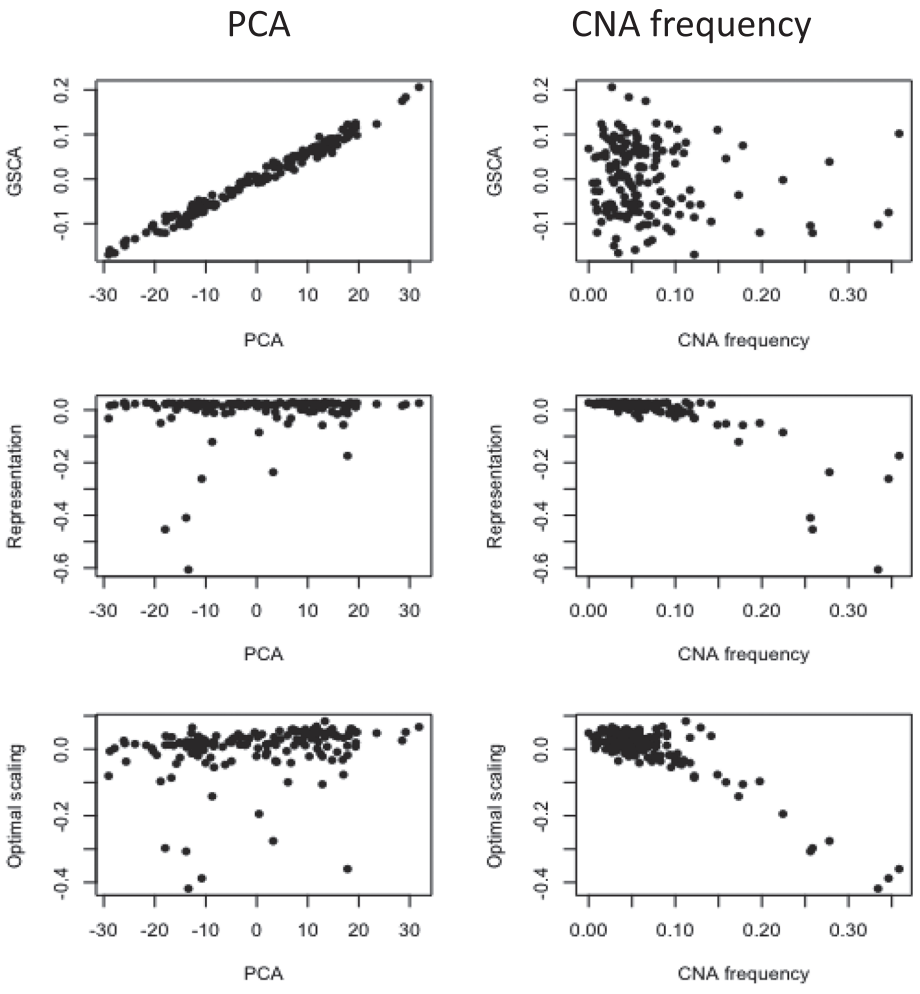

FIGURE 7 Left panels: score plots of PC3 of principal component analysis (PCA) on gene-expression data (*x*-axis) compared with the SC3 from the fusion results (*y*-axis); optimal scaling is OS-SCA, and representation is IDIOMIX. Right panels: scores of SC3 of all methods compared with copy number abberation (CNA) frequency

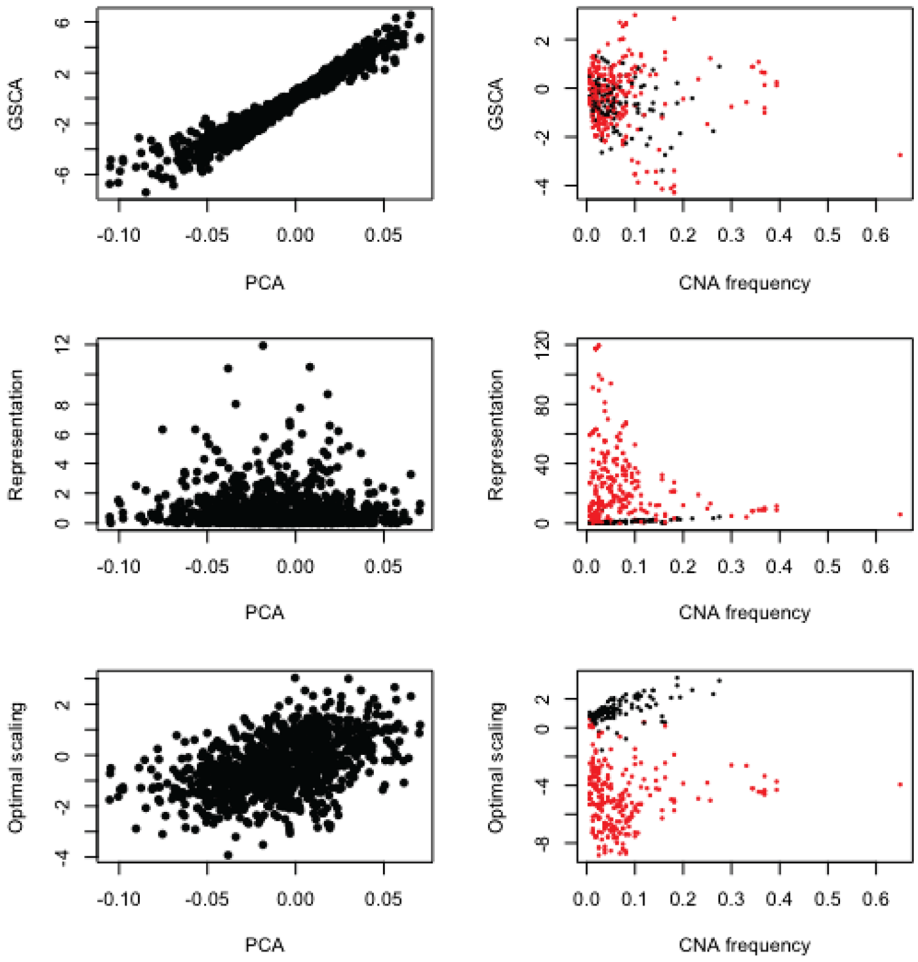

FIGURE 8 Left panels: PC3 loadings of PCA on gene-expression data (*x*-axis) compared with the SC3 loadings from the fusion results (*y*-axis); optimal scaling is OS-SCA, and representation is IDIOMIX. Right panels: loadings of SC3 of all fusion methods, amplifications (black) and losses (red)

with specific subclusters for hormone-positive breast cancer (within the BRCA group) and MITF high melanoma (in the SKCM group) (for a more elaborate interpretation, see Song et al[36]).

Whereas the three approaches give similar results for the first two simultaneous components, qualitative differences can be seen in SC3. This is especially apparent in Table 1 where the third component for IDIOMIX is now dominated by the CNA data. This is visualized in Figure 5, which shows the score plots of SC1 versus SC3 for all methods, which are clearly different. To further confirm this, the scores of the different methods for the three different components were plotted against each other (see Figure 6), and this confirms that indeed, the first two SCs are very similar for all methods, but that SC3 shows differences where GSCA is the most deviating. To shed some light on this deviating behavior, we plotted the scores of a PCA on the gene-expression data against the SC scores of the fusion methods for the third component, see Figure 7. The left panels in this figure show that SC3 from GSCA is very similar to the PC3 of a PCA on the gene-expression data alone (see also again Table 1). The same does not hold true for the other methods. The CNA values are available for each sample, and thus, the scores on the fusion SC3 can be plotted against the frequency at which such an aberration occurs (number of ones divided by the total). From the right panels of Figure 7, it then becomes clear that SC3 of IDIOMIX and OS-SCA are mostly picking up the differences in frequencies, contrary to the GSCA-SC3 scores.

A similar comparison can be made for the loadings (see Figure 8). The left panels show the PC3 loadings from gene expression using PCA and the fusion methods. In the right panels, the fusion loadings are plotted against CNA frequencies (now across DNA positions), and those show no correlation. As explained earlier, the aberrations can either be amplifications or losses. The amplification frequencies (black dots) are clearly correlated with the IDIOMIX and OS-SCA loadings, and such frequencies are biologically not the most interesting ones.

To interpret the GSCA loadings, these loadings were subjected to a gene set enrichment analysis (GSEA). This resulted in a highly significant enrichment for epithelial-mesenchymal transition (EMT), a process undergone by tumor cells frequently associated with invasion of surrounding tissues and subsequent metastases (see Table A2 in Appendix). The largest positive loading on GSCA-PC3 for the gene expression is ZEB1, a transcription factor associated with EMT. A plot

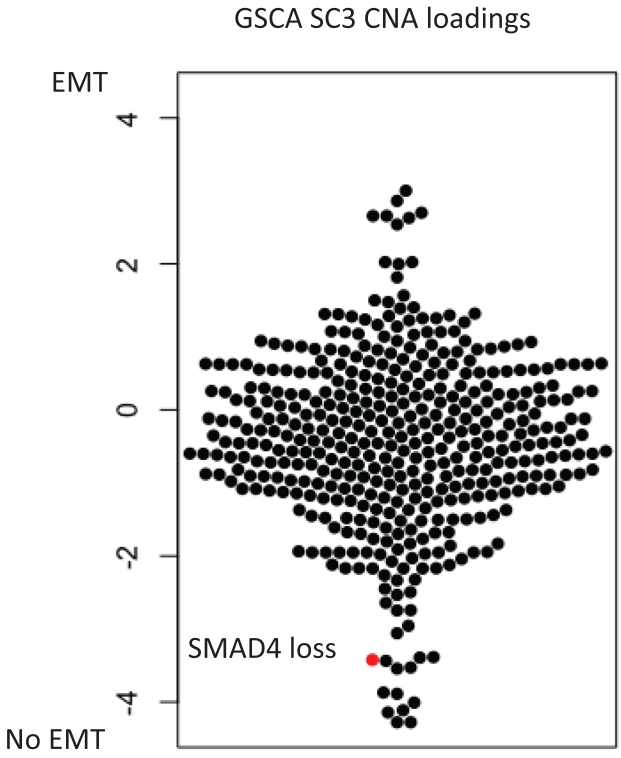

FIGURE 9 SC3 loadings of GSCA. In red: SMAD4loss (see text)

of the loadings of the CNA data is shown in Figure 9, and one of the loadings identifies SMAD4 loss as an important factor. SMAD4 is required for TGF-β–driven EMT, which confirms the finding that the GSCA gene-expression loadings are strongly enriched for EMT. [37]

Summarizing, IDIOMIX and OS-SCA are very similar for the whole analysis. For the first two SCs, also, the GSCA resembles the other approaches. The difference of GSCA is in the third SC. It seems that GSCA is focusing more on the gene-expression data, whereas IDIOMIX and OS-SCA pick up specific aspects of the CNA data in this third SC. The results of the GSCA-SC3 are biologically relevant; this is less the case for SC3 of the other approaches. It may be that GSCA is focusing more on the common variation between the two data sets and is less influenced by the distinctive parts.[34] This needs further exploration in a follow-up paper.

4 | DISCUSSION

In this paper, we have described and compared three methods of fusing data of different measurement scales. We used the example of quantitative and binary data, but all methods can also deal with ordinal data. For the example, it appears that IDIOMIX and OS-SCA give very similar results, whereas GSCA is different. One of the reasons may be that the methods deal differently with common and distinct parts of the data.

All methods have meta-parameters; that is, prior choices have to be made. For IDIOMIX, this is the type of representation to select; for OS-SCA, it is the type of restrictions to apply; for GSCA, it is the distribution to assume for the separate data sets. All methods also require selecting the complexity of the models, ie, the number of components. The selection of all these meta-parameters will, in practice, be made based on a mixture of domain knowledge and validation, such as cross validation or scree tests for selecting model complexity.

We hesitate in giving recommendations regarding which method to use for a particular application. First, the example of this paper concerns an exploratory study for which it is always difficult to judge the relative merits of the methods. Second, the cultural background of the investigator plays a role. In data analysis and chemometrics, the culture is to avoid distributional assumptions and have a more data analytic approach, thus resulting in a preference for IDIOMIX or OS-SCA. In statistics and, to some extent, in bioinformatics, there is more a tendency to go for parametric modes, hence GSCA in our context. Third, these methods have not yet been used to a large extent by researchers; hence, experience on their behavior upon which a recommendation can be based is lacking.

In terms of ease of use, we have a slight preference for IDIOMIX. Once the representation matrices are built, standard three-way analysis software can be used to fit the models. There is also software available for OS-SCA and GSCA, but this software is more difficult to implement.

There remain open issues to be investigated. Some of the more prominent ones is to understand the behavior of the methods regarding common, distinct, and local components in fusing data sets. Little has been done in this field regarding data of different measurement scales.

ORCID

Age K. Smilde https://orcid.org/0000-0002-3052-4644
Yipeng Song https://orcid.org/0000-0002-7180-2647
Johan A. Westerhuis https://orcid.org/0000-0002-6747-9779

REFERENCES

1. Zimmermann M, Kogadeeva M, Gengenbacher M, et al. Integration of metabolomics and transcriptomics reveals a complex diet of mycobacterium tuberculosis during early macrophage infection. *Msystems*. 2017;2(4):UNSP e00057-17.
2. Steinmetz V, Sevila F, Bellon-Maurel V. A methodology for sensor fusion design: application to fruit quality assessment. *J Agri Eng Res*. 1999;74:21-31.
3. Doeswijk TG, Smilde AK, Hageman JA, Westerhuis JA, van Eeuwijk FA. On the increase of predictive performance with high-level data fusion. *Analytica Chimica Acta*. 2011;705(1-2):41-47.
4. Stevens SS. On the theory of scales of measurement. *Science*. 1946;103(2684):677-680.
5. Hand DJ. Statistics and the theory of measurement. *J R Stat Soc Ser A-Stat Soc*. 1996;159:445-473.
6. Krantz DH, Luce RD, Suppes P, Tversky A. *Foundations of Measurement (Volume I)*. London: Dover; 1971.
7. Luce RD, Narens L. Measurement scales on the continuum. *Science*. 1987;236(4808):1527-1532.
8. Narens L. On the scales of measurement. *J Math Psychol*. 1981;24(3):249-275.
9. Narens L, Luce RD. Measurement—the theory of numerical assignments. *Psychol Bull*. March 1986;99(2):166-180.
10. Suppes P, Zinnes JL. Basic measurement theory. Technical report, Psychology Series, Stanford University. Institute for Mathematical Studies in the Social Sciences. Palo Alto, CA; 1962.
11. Adams EW, Fagot RF, Robinson RE. A theory of appropriate statistics. *Psychometrika*. 1965;30(2):99-127.
12. Michell J. Measurement scales and statistics—a clash of paradigms. *Psychol Bull*. 1986;100(3):398-407.
13. Mo QX, Wang SJ, Seshan VE, et al. Pattern discovery and cancer gene identification in integrated cancer genomic data. *Proc Nat Acad Sci Unit States Am*. 2013;110(11):4245-4250.
14. Shen RL, Olshen AB, Ladanyi M. Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. *Bioinformatics*. 2009;25(22):2906-2912.
15. van Wietmarschen HA, Dai WD, van der Kooij AJ, et al. Characterization of rheumatoid arthritis subtypes using symptom profiles, clinical chemistry and metabolomics measurements. *PLoS ONE*. 2012;7(9):e44331.
16. van Wietmarschen HA, Reijmers TH, van der Kooij AJ, et al. Sub-typing of rheumatic diseases based on a systems diagnosis questionnaire. *PLoS ONE*. 2011;6(9):e24846.
17. Kiers HAL. *Three-Way Methods for the Analysis of Qualitative and Quantitative Two-Way Data*. Leiden: DSWO Press; 1989.
18. Zegers FE. *A General Family of Association Coefficients*. Groningen/Haren: Boomker; 1986.
19. Gifi A. *Nonlinear Multivariate Analysis*. Chichester: John Wiley & Sons; 1990.
20. Michailidis G, de Leeuw J. The Gifi system of descritpive multivariate analysis. *Stat Sci*. 1998;13(4):307-336.
21. Young FW. Quantitative-analysis of qualitative data. *Psychometrika*. 1981;46(4):357-388.
22. Janson S, Vegelius J. Correlation-coefficients for more than one scale type. *Multivariate Behavioral Res*. 1982;17(2):271-284.
23. Tschuprow AA. *Principles of the Mathematical Theory of Correlation*. London: William Hodge; 1939.
24. De Leeuw J, Young FW, Takane Y. Additive structure in qualitative data: an alternating least squares method with optimal scaling features. *Psychometrika*. 1976;41(4):471-503.
25. Takane Y, Young FW, De Leeuw J. Nonmetric individual differences multidimensional scaling: an alternating least squares method with optimal scaling features. *Psychometrika*. 1977;42(1):7-67.
26. Young FW, Takane Y, De Leeuw J. The principal components of mixed measurement level multivariate data: an alternating least squares method with optimal scaling features. *Psychometrika*. 1978;43(2):279-281.
27. De Leeuw J. Canonical analysis of categorical data. *Ph.D. Thesis*: University of Leiden; 1973.
28. Van der Burg E, De Leeuw J, Verdegaal R. Homogeneity analysis with k sets of variables: an alternating least squares method with optimal scaling features. *Psychometrika*. 1988;53(2):177-197.
29. Waaijenborg S, Hamer PCVDW, Zwinderman AH. Quantifying the association between gene expressions and DNA-markers by penalized canonical correlation analysis. *Stat Appl Gen Molecul Biol*. 2008;7(1):3.
30. Ten Berge JMF, Kiers HAL, van der Stel V. Simultaneous component analysis. *Stat Applicata 4, 4, p 377-392*. 1992;4(4):377-392.

31. 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.
32. De Leeuw J. Correctness of Kruskal's algorithms for monotone regression with ties. *Psychometrika*. 1977;42(1):141-144.
33. Van Mechelen I, Smilde AK. A generic linked-mode decomposition model for data fusion. *Chemometrics Intel Lab Syst*. 2010;104:83-94.
34. Smilde AK, Måge I, Naes T, et al. Common and distinct components in data fusion. *J Chemometrics*. 2017;31:e2900.
35. Curtis C, Shah SP, Chin SF, et al. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature*. 2012;486(7403):346-352.
36. Song Y, Aben N, Wessels LFA, Westerhuis JA, Groenen PJF, Smilde AK. Generalized simultaneous component analysis. arXiv, (1807.04982v2); 2018.
37. Tian M, Schiemann WP. The TGF-β paradox in human cancer: an update; 2009.
38. Liberzon A, Birger C, Thorvaldsdóttir H, Ghandi M, Mesirov J, Tamayo P. The molecular signatures database hallmark gene set collection. *Cell Systems*. 2015;1(6):417-425.

How to cite this article: Smilde AK, Song Y, Westerhuis JA, Kiers HAL, Aben N, Wessels LFA. Heterofusion: Fusing genomics data of different measurement scales. *Journal of Chemometrics*. 2021;35:e3200. https://doi.org/10.1002/cem.3200

A.1 | Optimal scaling of binary data equals analyzing standardized data

The fact that optimal scaling of binary data equals the analysis of standardized data can be shown as follows. Suppose that a binary vector has n_0 values of 0, n_1 values of 1, and n values in total and a is the optimal scaled value for the zeros and b for the ones. In optimal scaling, the optimal scaled variables need to get some kind of normalization. A common set of choices (see Gifi[19]) is to make sure that the scaled values have mean 0 and variance 1. This leads to the following two equations:

$$\begin{aligned} n_0 a + n_1 b &= 0, \\ n_0 a^2 + n_1 b^2 &= n, \end{aligned} \tag{A1}$$

and these equations can be solved for a and b since n_0, n_1, and n are known. This gives two values for a: one positive and one negative. The values of b follow automatically with the opposite sign. Hence, both solutions are practically equal. The values a and b are the optimal scaling constants for the binary data.

A.2 | Examples of representation matrices for ratio- and interval-scaled data

We will illustrate some ideas of representation matrices using a small example of an (4×2) matrix $\mathbf{X} = [\mathbf{x}_1 | \mathbf{x}_2]$:

$$\mathbf{X} = \begin{pmatrix} 2 & 9 \\ 4 & 9 \\ 6 & 10 \\ 8 & 12 \end{pmatrix}, \tag{A2}$$

and the standardized version of this is

$$\mathbf{X}_s = \begin{pmatrix} -0.671 & -0.408 \\ -0.224 & -0.408 \\ 0.224 & 0 \\ 0.671 & 0.816 \end{pmatrix}, \tag{A3}$$

where indeed, $\mathbf{x}^T{}_{s1}\mathbf{x}_{s1} = 1$, $\mathbf{x}^T{}_{s2}\mathbf{x}_{s2} = 1$, and $\mathbf{x}^T{}_{s1}\mathbf{x}_{s2} = 0.913$, the latter being the correlation between $\mathbf{x}_1$ and $\mathbf{x}_2$. The square representation using Equation (5) on $\mathbf{x}_1$ gives

$$\tilde{\mathbf{S}}_1 = \begin{pmatrix} 0 & -2 & -4 & -6 \\ 2 & 0 & -2 & -4 \\ 4 & 2 & 0 & -2 \\ 6 & 4 & 2 & 0 \end{pmatrix}, \tag{A4}$$

which is skew-symmetric ($\tilde{\mathbf{S}}_1^T = -\tilde{\mathbf{S}}_1$) and contains all the differences between the elements of $\mathbf{x}_1$. The standardized version of $\tilde{\mathbf{S}}_1$ is

$$\mathbf{S}_1 = \begin{pmatrix} 0 & -0.158 & -0.316 & -0.474 \\ 0.158 & 0 & -0.158 & -0.316 \\ 0.316 & 0.158 & 0 & -0.158 \\ 0.474 & 0.316 & 0.158 & 0 \end{pmatrix}, \tag{A5}$$

and a similar matrix can be made for $\mathbf{x}_2$. Then, using Equation (4) on the pairs $(\mathbf{S}_1, \mathbf{S}_1)$ and $(\mathbf{S}_2, \mathbf{S}_2)$ gives a value of 1, and on the pair $(\mathbf{S}_1, \mathbf{S}_2)$ gives 0.913, which is the Pearson's correlation again.

Alternative square representations of $\mathbf{x}_{s1}$ and $\mathbf{x}_{s2}$ are

$$\mathbf{S}_{A1} = \mathbf{x}_{s1}\mathbf{x}^T{}_{s1} = \begin{pmatrix} 0.45 & 0.15 & -0.15 & -0.45 \\ 0.15 & 0.05 & -0.05 & -0.15 \\ -0.15 & -0.05 & 0.05 & 0.15 \\ -0.45 & -0.15 & 0.15 & 0.45 \end{pmatrix}, \tag{A6}$$

and

$$\mathbf{S}_{A2} = \mathbf{x}_{s2}\mathbf{x}^T{}_{s2} = \begin{pmatrix} 0.167 & 0.167 & 0 & -0.333 \\ 0.167 & 0.167 & 0 & -0.333 \\ 0 & 0 & 0 & 0 \\ -0.333 & -0.333 & 0 & 0.667 \end{pmatrix}, \tag{A7}$$

and using Equation (4) on $\mathbf{S}_{A1}$ and $\mathbf{S}_{A2}$ gives 0.833, which is the squared Pearson's correlation between the original variables.

A.3 | Examples of representation matrices for nominal data

We will illustrate some ideas on representing nominal data using two categorical variables $\mathbf{x}_1$ and $\mathbf{x}_2$. The first variable contains four categories encoded as A, B, C, and D and reads $\mathbf{x}_1 = (A, B, A, C, D, C, B, D)^T$; the second variable has three categories encoded as I, II, and III and reads $\mathbf{x}_2 = (I, II, II, I, III, III, I, II)^T$ where the roman capitals are used to show that the two variables encode different types of categories. The indicator matrices are now

$$\mathbf{G}_1 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \tag{A8}$$

and

$$\mathbf{G}_2 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}, \tag{A9}$$

and a special feature of this kind of data becomes present, namely, that some objects have exactly the same rows in $\mathbf{G}_1$ (and similarly in $\mathbf{G}_2$). Moreover, the matrices show closure ($\mathbf{G}^T{}_1\mathbf{1} = \mathbf{1}, \mathbf{G}^T{}_2\mathbf{1} = \mathbf{1}$). The marginal frequencies are collected in

$$\mathbf{D}_1 = \begin{pmatrix} 2 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 2 \end{pmatrix} = \mathbf{G}_1^T\mathbf{G}_1, \tag{A10}$$

and

$$\mathbf{D}_2 = \begin{pmatrix} 3 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 2 \end{pmatrix} = \mathbf{G}_2^T\mathbf{G}_2, \tag{A11}$$

with obvious properties.

Simple representations of these variables are now

$$\mathbf{S}_{1s} = \begin{pmatrix} 1&0&1&0&0&0&0&0 \\ 0&1&0&0&0&0&1&0 \\ 1&0&1&0&0&0&0&0 \\ 0&0&0&1&0&1&0&0 \\ 1&0&1&0&0&0&0&0 \\ 0&0&0&0&1&0&0&1 \\ 0&0&0&1&0&1&0&0 \\ 0&0&0&0&1&0&0&1 \end{pmatrix} = \mathbf{G}_1\mathbf{G}_1^T, \tag{A12}$$

and

$$\mathbf{S}_{2s} = \begin{pmatrix} 1&0&0&1&0&0&1&0 \\ 0&1&1&0&0&0&0&1 \\ 0&1&1&0&0&0&0&1 \\ 1&0&0&1&0&0&1&0 \\ 0&0&0&0&1&1&0&0 \\ 0&0&0&0&1&1&0&0 \\ 1&0&0&1&0&0&1&0 \\ 0&1&1&0&0&0&0&1 \end{pmatrix} = \mathbf{G}_2\mathbf{G}_2^T, \tag{A13}$$

and these representations are encoding, which objects have equal categories in the variables. The more complex representations (according to Equation 6) are now

$$\mathbf{S}_{1c} = \begin{pmatrix} 0.375 & -0.125 & 0.375 & -0.125 & -0.125 & -0.125 & -0.125 & -0.125 \\ -0.125 & 0.375 & -0.125 & -0.125 & -0.125 & -0.125 & 0.375 & -0.125 \\ 0.375 & -0.125 & 0.375 & -0.125 & -0.125 & -0.125 & -0.125 & -0.125 \\ -0.125 & -0.125 & -0.125 & 0.375 & -0.125 & 0.375 & -0.125 & -0.125 \\ -0.125 & -0.125 & -0.125 & -0.125 & 0.375 & -0.125 & -0.125 & 0.375 \\ -0.125 & -0.125 & -0.125 & 0.375 & -0.125 & 0.375 & -0.125 & -0.125 \\ -0.125 & 0.375 & -0.125 & -0.125 & -0.125 & -0.125 & 0.375 & -0.125 \\ -0.125 & -0.125 & -0.125 & -0.125 & 0.375 & -0.125 & -0.125 & 0.375 \end{pmatrix}, \tag{A14}$$

and

$$\mathbf{S}_{1c} = \begin{pmatrix} 0.375 & -0.125 & 0.375 & -0.125 & -0.125 & -0.125 & -0.125 & -0.125 \\ -0.125 & 0.375 & -0.125 & -0.125 & -0.125 & -0.125 & 0.375 & -0.125 \\ 0.375 & -0.125 & 0.375 & -0.125 & -0.125 & -0.125 & -0.125 & -0.125 \\ -0.125 & -0.125 & -0.125 & 0.375 & -0.125 & 0.375 & -0.125 & -0.125 \\ -0.125 & -0.125 & -0.125 & -0.125 & 0.375 & -0.125 & -0.125 & 0.375 \\ -0.125 & -0.125 & -0.125 & 0.375 & -0.125 & 0.375 & -0.125 & -0.125 \\ -0.125 & 0.375 & -0.125 & -0.125 & -0.125 & -0.125 & 0.375 & -0.125 \\ -0.125 & -0.125 & -0.125 & -0.125 & 0.375 & -0.125 & -0.125 & 0.375 \end{pmatrix}, \tag{A14}$$

and

$$\mathbf{S}_{2c} = \begin{pmatrix} 0.208 & -0.125 & -0.125 & 0.208 & -0.125 & -0.125 & 0.208 & -0.125 \\ -0.125 & 0.208 & 0.208 & -0.125 & -0.125 & -0.125 & -0.125 & 0.208 \\ -0.125 & 0.208 & 0.208 & -0.125 & -0.125 & -0.125 & -0.125 & 0.208 \\ 0.208 & -0.125 & -0.125 & 0.208 & -0.125 & -0.125 & 0.208 & -0.125 \\ -0.125 & -0.125 & -0.125 & -0.125 & 0.375 & 0.375 & -0.125 & -0.125 \\ -0.125 & -0.125 & -0.125 & -0.125 & 0.375 & 0.375 & -0.125 & -0.125 \\ 0.208 & -0.125 & -0.125 & 0.208 & -0.125 & -0.125 & 0.208 & -0.125 \\ -0.125 & 0.208 & 0.208 & -0.125 & -0.125 & -0.125 & -0.125 & 0.208 \end{pmatrix}, \tag{A15}$$

which are indeed double centered and standardized. Using Equation (4) on the matrices $\mathbf{S}_{1c}$ and $\mathbf{S}_{2c}$ gives the correlation coefficient $T^2 = 0.5$.

TABLE A1 Calculation of the ϕ coefficient (the values of the example between brackets)

	$x_2 = 0$	$x_2 = 1$	Total
$x_1 = 1$	$n_{11}(1)$	$n_{10}(2)$	$n_{1.}(3)$
$x_1 = 0$	$n_{01}(4)$	$n_{00}(1)$	$n_{0.}(5)$
Total	$n_{.1}(5)$	$n_{.0}(3)$	$n(8)$

TABLE A2 GSEA results. Legend: Size is size of the gene set; ES is enrichment score; NES is normalized enrichment score; Nom-p is nominal p-value; FDR-q is FDR q-value; EMT is Epithelial Mesenchymal Transition; UV-R-DN is UV Response DNA; IL2-STAT5-S is IL2-STAT5-Signaling; KRAS-S-UP is KRAS-Signaling-UP. The value 0 means < 0.0001

Process	Size	ES	NES	Nom-p	FDR-q
EMT	87	0.61	3.06	0	0
UV-R-DN	35	0.60	2.47	0	0
IL2-STAT5-S	24	-0.57	-1.76	0.003	0.040
APOPTOSIS	31	-0.47	-1.53	0.027	0.219
KRAS-S-UP	56	-0.42	-1.51	0.017	0.167

A.4 | Examples of representation matrices for binary data

As an example for binary data, we will use a simple data set consisting of two binary variables $\mathbf{x}_1$ and $\mathbf{x}_2$, which are columns of

$$\mathbf{X} = \begin{pmatrix} 0 & 1 \\ 0 & 1 \\ 1 & 0 \\ 1 & 1 \\ 0 & 1 \\ 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{pmatrix}, \tag{A16}$$

with indicator matrices

$$\mathbf{G}_1 = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 1 & 0 \end{pmatrix}, \tag{A17}$$

and

$$\mathbf{G}_2 = \begin{pmatrix} 0 & 1 \\ 0 & 1 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 1 & 0 \\ 0 & 1 \\ 1 & 0 \end{pmatrix}. \tag{A18}$$

A correlation measure between binary variables is the ϕ coefficient, which is defined as

$$\frac{n_{11}n_{00} - n_{10}n_{01}}{\sqrt{n_{1.}n_{0.}n_{.0}n_{.1}}}, \tag{A19}$$

where the values n are shown in Table A1. For the example, this ϕ coefficient equals −0.4667, which is also equivalent to the Pearson correlation between $\mathbf{x}_1$ and $\mathbf{x}_2$.

There are two alternative square representations of $\mathbf{x}_1$ and $\mathbf{x}_2$. The first uses Equation (6) based on the indicator matrices, and the results are

$$\mathbf{S}_{1c} = \begin{pmatrix} 0.075 & 0.075 & -0.125 & -0.125 & 0.075 & -0.125 & 0.075 & 0.075 \\ 0.075 & 0.075 & -0.125 & -0.125 & 0.075 & -0.125 & 0.075 & 0.075 \\ -0.125 & -0.125 & 0.208 & 0.208 & -0.125 & 0.208 & -0.125 & -0.125 \\ -0.125 & -0.125 & 0.208 & 0.208 & -0.125 & 0.208 & -0.125 & -0.125 \\ 0.075 & 0.075 & -0.125 & -0.125 & 0.075 & -0.125 & 0.075 & 0.075 \\ -0.125 & -0.125 & 0.208 & 0.208 & -0.125 & 0.208 & -0.125 & -0.125 \\ 0.075 & 0.075 & -0.125 & -0.125 & 0.075 & -0.125 & 0.075 & 0.075 \\ 0.075 & 0.075 & -0.125 & -0.125 & 0.075 & -0.125 & 0.075 & 0.075 \end{pmatrix}, \tag{A20}$$

and

$$\mathbf{S}_{2c} = \begin{pmatrix} 0.075 & 0.075 & -0.125 & 0.075 & 0.075 & -0.125 & 0.075 & -0.125 \\ 0.075 & 0.075 & -0.125 & 0.075 & 0.075 & -0.125 & 0.075 & -0.125 \\ -0.125 & -0.125 & 0.208 & -0.125 & -0.125 & 0.208 & -0.125 & 0.208 \\ 0.075 & 0.075 & -0.125 & 0.075 & 0.075 & -0.125 & 0.075 & -0.125 \\ 0.075 & 0.075 & -0.125 & 0.075 & 0.075 & -0.125 & 0.075 & -0.125 \\ -0.125 & -0.125 & 0.208 & -0.125 & -0.125 & 0.208 & -0.125 & 0.208 \\ 0.075 & 0.075 & -0.125 & 0.075 & 0.075 & -0.125 & 0.075 & -0.125 \\ -0.125 & -0.125 & 0.208 & -0.125 & -0.125 & 0.208 & -0.125 & 0.208 \end{pmatrix}, \tag{A21}$$

and when these are used in Equation (4), the result is 0.2178, which is the square of the ϕ coefficient.

The other representations are based on the standardized x variables $\mathbf{z}_1$ and $\mathbf{z}_2$ (with $\mathbf{z}^T{}_1\mathbf{z}_2 = -0.4667$). It now holds that $\mathbf{J}\mathbf{G}_j\mathbf{D}_j^{-1}\mathbf{G}_j^T\mathbf{J} = \mathbf{z}_j\mathbf{z}^T{}_j$, and hence, both representations coincide.

A.5 | GSEA results

Gene set enrichment analysis (GSEA) was performed by using the Hallmark gene sets from MSigDB.[38] The GSEA-preranked results are summarized in Table A2 and show that the EMT process has the lowest P value.

A.6 | Software

The software to estimate the parameters in the different methods can be downloaded from www.bdagroup.nl.