A similarity index for comparing coupled matrices

Ulf G. Indahl[x], Tormod Næs*[+], Kristian Hovde Liland[x*]

x) Faculty of Sciences and Technology, Norwegian University of Life Sciences, N-1432 Ås, Norway

*) Nofima, Oslovegen 1, 1430 Ås

+) Dept. of Food Science, University of Copenhagen.

A similarity index for comparing coupled matrices

Abstract

Application of different multivariate measurement technologies to the same set of samples is an interesting challenge in many fields of applied data analysis. Our proposal is a two-stage similarity index framework for comparing two matrices in this type of situation. The first step is to identify factors (and associated subspaces) of the matrices by methods such as principal component analysis (PCA) or partial least squares (PLS) regression to provide good (low-dimensional) summaries of their information content. Thereafter, statistical significances are assigned to the similarity values obtained at various factor subset combinations by considering orthogonal projections or Procrustes rotations, and how to express the results compactly in corresponding summary plots. Applications of the methodology include the investigation of redundancy in spectroscopic data and the investigation of assessor consistency or -deviations in sensory science. The proposed methodology is implemented in the R-package "MatrixCorrelation" available online from CRAN.

Key words: Similarity index, Canonical Correlation, Significance testing, Orthogonal projections, Procrustes rotations, RV coefficient.

1. Introduction

The problem of comparing pairs of variables/vectors by some type of correlation coefficient is fundamental and well understood (see e.g. Draper and Smith (1998)). In modern science there is, however, also an increasing need for comparing collections of variables (represented by data matrices of multivariate measurements). Interesting situations arise when comparing measurements obtained by different technologies or instruments for a fixed set of (n) samples. Important applications frequently appear in spectroscopy, in the omics areas and when comparing trained sensory assessors for detecting deviating assessments (Tomic et al. (2013)).

The *RV coefficient* -by Robert and Escoufier (1976) is among the most popular methods for comparing matrices in a correlation like style. Smilde et al. (2009) pointed out that the RV-coefficient suffers from an increasing bias (towards 1) when the number of variables (columns) increase compared to the number of samples (rows) in the two matrices. They therefore proposed the RV2-coefficient as a bias reducing (but still biased) alternative to the RV-coefficient. Mayer et al. (2011) pointed at some fundamental problems with the RV2 and proposed a corrected and unbiased alternative to the RV- (and RV2) directly based on the classical *adjusted R-square* statistic. The RV_{gq}-coefficient proposed by El Ghaziri and Qannari (2015) is an alternative unbiased modification of the RV coefficient that also avoids the RV2 shortcomings. Other matrix similarity measures of particular interest for the present study are the *Procrustes similarity index* by Sibson (1978) and the *generalized coefficient of determination (GCD)* by Yanai, (1974).

In the present paper, we propose an alternative similarity index approach for comparing two sets of measurements by considering an associated pair of data matrices. The proposed *similarity of matrices index* (*SMI*) approach is based on the idea of comparing a selection of dominant

subspace combinations derived by appropriate matrix decomposition strategies such as the principal component analysis (PCA) and partial least squares (PLS) regression. We also propose a statistical test of difference/similarity between the matrices associated with the *SMI* calculations. In order to simplify the decision-making part of an *SMI*-based analysis, a so-called *"Diamond plot*" is proposed. Two alternatives for comparing the subspaces will be considered, one that is based on Orthogonal Projections (OP) and one that is based on Procrustes Rotations (PR), see Kendall (1984). Both alternatives correspond to classical choices of linear transformations for comparing subspaces. The particular aspects of similarity considered when calculating the *SMI* depends on i) the *subspace identification method* (such as PCA or PLS) and ii) the type of *regression method* (OP or PR) used in the subsequent comparison. The choice of methods from i) and ii) specifies what will here be called the *context* for comparing the measurements. PCA is the appropriate choice for investigating similarities between the subspaces of dominant and stable variance associated with the two data matrices. PLS is the appropriate alternative when comparing the validated predictive parts of two data matrices with respect to some response variable(s).

Regarding the choice of regression method, we promote the OP as the primary alternative. The PR, we think, should rarely be considered alone, but rather as a valuable supplement to OP in situations where there are particular reasons to reveal if important relationships between the two datasets can be accounted for by scaling and rotations only. A typical field of application, where the PR is of particular interest, is sensory analysis (Amerine et al. (1965)). It can for instance be observed that the data generated by two sensory assessors may describe the same underlying dimensions, but one assessor switches for instance the order of the two first underlying dimensions as compared to the other. We therefore recommend judging the similarity of two

datasets by considering both OP and PR as a good way of capturing the presence of such phenomena.

The paper is organized as follows: In Section 2, we present a brief summary of existing and related methodologies with focus on the RV-coefficient alternatives together with the ideas motivating the SMI approach. Section 3 presents the mathematical definition of the *SMI* and its key properties for both the OP and the PR cases. We then continue by demonstrating an application of the *SMI*-framework to the collection of alternative factor combinations obtained by varying the number of subspace dimensions. This application includes a permutation test for associating statistical significances with the obtained *SMI*-values. Section 4 presents a collection of examples, with both simulated and real datasets, to demonstrate potential applications of the suggested methodology. Finally, we draw our conclusions after discussing the relationships between our proposal and some established alternatives from the literature.

2 Background and motivation

We consider the problem of comparing two different sets of measurement taken on a fixed set of (n) samples. After mean centering of the measured variables, the resulting datasets are typically represented in two matrices $\mathbf{X}_1$ and $\mathbf{X}_2$ of dimensions $n \times m_1$ and $n \times m_2$, respectively. Among the various alternatives for comparing matrices that have been put forward in the literature, there are several interesting methods that are quite closely related to the *RV coefficient* -by Robert and Escoufier (1976). Their original *RV coefficient* is defined as $RV(\mathbf{X}_1, \mathbf{X}_2) = \frac{tr(\mathbf{Y}_1^t \mathbf{Y}_2)}{\sqrt{tr(\mathbf{Y}_1^t \mathbf{Y}_1) tr(\mathbf{Y}_2^t \mathbf{Y}_2)}}$, where $\mathbf{Y}_1 = \mathbf{X}_1 \mathbf{X}_1^t$ and $\mathbf{Y}_2 = \mathbf{X}_2 \mathbf{X}_2^t$, and $tr(\cdot)$ denotes the matrix trace. Besides the original RV coefficient we consider the following methods to be of particular interest for our study: 1) The

modified RV-coefficient by Smilde et al. (2009): $RV2(\mathbf{X}_1, \mathbf{X}_2) = \frac{tr(\mathbf{Y}_1^t \mathbf{Y}_2)}{\sqrt{tr(\mathbf{Y}_1^t \mathbf{Y}_1) tr(\mathbf{Y}_2^t \mathbf{Y}_2)}}$, where $\mathbf{Y}_1 = \mathbf{X}_1\mathbf{X}_1^t - D(\mathbf{X}_1\mathbf{X}_1^t)$, $\mathbf{Y}_2 = \mathbf{X}_2\mathbf{X}_2^t - D(\mathbf{X}_2\mathbf{X}_2^t)$ and $D(\cdot)$ denotes the matrix diagonal. 2) The *adjusted RV-coefficient* by Mayer et al. (2011): $RV_{adj}(\mathbf{X}_1, \mathbf{X}_2) = \frac{p \cdot q \cdot n_c + n_r \cdot tr(\mathbf{C}_{12}^t \mathbf{C}_{12})}{\sqrt{[p \cdot p \cdot n_c + n_r \cdot tr(\mathbf{C}_{11}^t \mathbf{C}_{11})][q \cdot q \cdot n_c + n_r \cdot tr(\mathbf{C}_{22}^t \mathbf{C}_{22})]}}$. Here $\mathbf{C}_{ij}$ is the correlation matrix between $\mathbf{X}_i$ and $\mathbf{X}_j$, p and q are the number of columns in $\mathbf{X}_1$ and $\mathbf{X}_2$, respectively, $n_r = \frac{(n-1)}{(n-2)}$ and $n_c = 1 - n_r$ where n is the number of rows in $\mathbf{X}_1$ (and $\mathbf{X}_2$). 3) The *adjusted RV-coefficient* by Ghaziri & Qannari (2015): $RV_{gq}(\mathbf{X}_1, \mathbf{X}_2) = \frac{RV(\mathbf{X}_1, \mathbf{X}_2) - m_{\mathrm{RV}}}{1 - m_{\mathrm{RV}}}$. Here, $m_{\mathrm{RV}} = \frac{tr(\mathbf{X}_1^t \mathbf{X}_2)}{\sqrt{tr(\mathbf{X}_1^t \mathbf{X}_1) tr(\mathbf{X}_2^t \mathbf{X}_2)}}$ denotes the expected value of the RV coefficient, i.e. the mean RV value for all possible permutations of the rows of one of the matrices. According to the authors, this will correct for random similarities between the two matrices. 4) The *Procrustes similarity index* by Sibson (1978): $(\mathbf{X}_1, \mathbf{X}_2) = \frac{tr(\mathbf{X}_1^t \mathbf{X}_2 \mathbf{H})}{\sqrt{tr(\mathbf{X}_1^t \mathbf{X}_1) tr(\mathbf{X}_2^t \mathbf{X}_2)}}$, where $\mathbf{H}$ is the Procrustes transformation scaling and rotating/reflecting $\mathbf{X}_2$ to minimize the distance $\|\mathbf{X}_1 - \mathbf{X}_2\mathbf{H}\|_F$ with respect to the Frobenius norm. 5) *The generalized coefficient of determination (GCD)* by Yanai, (1974): The *GCD* is originally defined in terms of the projection matrices onto the column spaces of $\mathbf{X}_1$ and $\mathbf{X}_2$. By relatively simple algebraic manipulations it can be shown that the original GCD definition is equivalent to $GCD(\mathbf{X}_1, \mathbf{X}_2) = RV(\mathbf{T}, \mathbf{U})$, where $\mathbf{T}$ and $\mathbf{U}$ are orthogonal bases for the column spaces of $\mathbf{X}_1$ and $\mathbf{X}_2$, respectively.

Note that the definitions given above are chosen to emphasize the relationships between the different methods. The definitions presented in the original references are equivalent, but not necessarily identical to the definitions given above. More measures for comparing matrices can

be found in Ramsay et al (1984). See Section 2.5 for a short summary of the coefficients considered for particular comparison to our own proposals given below.

In spite of their obvious relevance in various situations, application of many well established methods may appear challenging (and in our opinion sometimes confusing) from a practitioner's point of view. In particular, the task of assigning statistical significances to large values (i.e. values close to 1) obtained by the existing coefficients, is not properly dealt with in the literature.

In the present paper, an alternative similarity index approach for comparing two sets of measurements is proposed. As indicated above, we consider the mean centered data matrices $\mathbf{X}_1$ and $\mathbf{X}_2$ that are coupled by the *n* rows typically referring to a joint set of samples.

The underlying assumption throughout our development is that the two data matrices can be decomposed as follows:

$$\begin{aligned} \mathbf{X}_1 &= \mathbf{T}\mathbf{P}_1^{\mathrm{t}} + \mathbf{E}_1, \text{ where } \mathbf{T} = \mathbf{X}_1\mathbf{C}_1 \\ \mathbf{X}_2 &= \mathbf{U}\mathbf{P}_2^{\mathrm{t}} + \mathbf{E}_2, \text{ where } \mathbf{U} = \mathbf{X}_2\mathbf{C}_2. \end{aligned} \tag{1}$$

Here, the matrix products $\mathbf{T}\mathbf{P}_1^{\mathrm{t}}$ and $\mathbf{U}\mathbf{P}_2^{\mathrm{t}}$ correspond to approximations of $\mathbf{X}_1$ and $\mathbf{X}_2$, respectively, representing the relevant structures of interest. The associated residual parts accounting for noise and irrelevant structure are represented by the residual matrices $\mathbf{E}_1$ and $\mathbf{E}_2$. The column vectors of $\mathbf{C}_1$ and $\mathbf{C}_2$ represent the required coefficients to express the **T**- and **U** columns as linear combinations of the $\mathbf{X}_1$- and $\mathbf{X}_2$ columns (variables), respectively. To be consistent with a terminology that is appropriate for both principal component analysis (PCA) and partial least squares (PLS) regression, it is assumed that the matrices **T** and **U** are always orthogonal (with normalized columns), i.e. representing normalized *score vectors*. The corresponding matrices

$\mathbf{P_1} = \mathbf{X_1^t T}$ and $\mathbf{P_2} = \mathbf{X_2^t U}$ are in agreement with the standard chemometrics terminology referred to as *loading* matrices (often considered for interpretation of the components).

The separation of relevant from irrelevant (i.e. $\mathbf{E}_1$ and $\mathbf{E}_2$) structure can be obtained by various approaches depending on the purpose of the analysis. PCA (see Joliffe (2002)) and PLS regression (see Wold et al. (1984)) with normalized scores (see Björck and Indahl (2017)) are the methods emphasized in our applications, but any method for deriving orthogonal matrices **T** and **U** from $\mathbf{X}_1$ and $\mathbf{X}_2$, respectively, will fit into the proposed *similarity of matrices index* (*SMI*) framework.

The new SMI approach is based on the idea of finding linear combinations of the $\mathbf{X}_1$-variables ($\mathbf{X}_1$-factors) that coincide with linear combinations of the $\mathbf{X}_2$-variables ($\mathbf{X}_2$-factors) by matching the two matrix approximations $\mathbf{TP}_1^{\mathrm{t}}$ and $\mathbf{UP}_2^{\mathrm{t}}$ as accurately as possible.

3. Methodology

3.1 The similarity of matrices index framework

Throughout the paper, it is assumed that the orthogonal score matrices **T** and **U** in (1)~~(1)~~ are centered and of dimensions $(n \times p)$ and $(n \times q)$, respectively (i.e. $\mathbf{T}^{\mathrm{t}}\mathbf{T} = \mathbf{I}_p$ and $\mathbf{U}^{\mathrm{t}}\mathbf{U} = \mathbf{I}_q$ are both identity matrices), where $0 < p \leq m_1$ and $0 < q \leq m_2$. The regression coefficient matrices for fitting **U** and **T** according to some regression method *M* of interest (here *M* = OP or *M* = PR) are denoted $\mathbf{B}_{\mathrm{T}}$ and $\mathbf{B}_{\mathrm{U}}$ respectively, with the resulting *fitted values* $\widehat{\mathbf{U}} = \mathbf{TB}_{\mathrm{T}}$ and $\widehat{\mathbf{T}} = \mathbf{UB}_{\mathrm{U}}$. The proportions of explained variance associated with **T** and **U** are given by $\left\|\widehat{\mathbf{T}}\right\|_F^2/p$ and $\left\|\widehat{\mathbf{U}}\right\|_F^2/q$, respectively ($\|\cdot\|_F^2$ denotes the squared Frobenius norm, i.e. $\|\mathbf{A}\|_F^2 = \sum_{i,j}\left|a_{i,j}\right|^2$).

We require that for the regression method *M* of interest, the inequalities $0 \leq \left\|\widehat{\mathbf{T}}\right\|_F^2/p \leq 1$ and $0 \leq \left\|\widehat{\mathbf{U}}\right\|_F^2/q \leq 1$ always hold, and that the maximum value of 1 is obtained if and only if the fitted values $\widehat{\mathbf{T}} = \mathbf{T}$ or $\widehat{\mathbf{U}} = \mathbf{U}$. With reference to the method *M*, the *similarity of matrices index* (*SMI*) of the two matrices **T** and **U** is defined by

$$SMI_M(\mathbf{T}, \mathbf{U}) = max\left(\frac{\|\widehat{\mathbf{T}}\|_F^2}{p}, \frac{\|\widehat{\mathbf{U}}\|_F^2}{q}\right), \qquad (2)$$

i.e. the maximum of the two proportions of explained -variance. In (4) and (5) below it will be seen that taking the maximum in (2) means accounting for as much as possible of the smaller of the two subspaces spanned by **T** and **U**, respectively. Note that if $\mathbf{U} = \widehat{\mathbf{U}} = \mathbf{T}\mathbf{B}_\mathrm{T}$ or $\mathbf{T} = \widehat{\mathbf{T}} = \mathbf{U}\mathbf{B}_\mathrm{U}$, then either $\left\|\widehat{\mathbf{U}}\right\|_F^2/q = 1$ or $\left\|\widehat{\mathbf{T}}\right\|_F^2/p = 1$.

Property 1

$$SMI_M(\mathbf{T}, \mathbf{U}) = max\left(\frac{\|\mathbf{B}_\mathbf{U}\|_F^2}{p}, \frac{\|\mathbf{B}_\mathbf{T}\|_F^2}{q}\right), \qquad (3)$$

which means that knowledge of the regression coefficients $\mathbf{B}_\mathrm{U}$ and $\mathbf{B}_\mathrm{T}$ is sufficient for computing the SMI_M defined in (2).

Proof:

Because $\mathbf{U}^\mathrm{t}\mathbf{U} = \mathbf{I}_q$, $\left\|\widehat{\mathbf{T}}\right\|_F^2 = trace\left(\widehat{\mathbf{T}}^\mathrm{t}\widehat{\mathbf{T}}\right) = trace(\mathbf{B}_\mathbf{U}^\mathrm{t}\mathbf{U}^\mathrm{t}\,\mathbf{U}\mathbf{B}_\mathbf{U}) = trace(\mathbf{B}_\mathrm{U}^\mathrm{t}\mathbf{B}_\mathbf{U}) = \|\mathbf{B}_\mathbf{U}\|_F^2$.

Correspondingly, we obtain $\left\|\widehat{\mathbf{U}}\right\|_F^2 = \|\mathbf{B}_\mathbf{T}\|_F^2$, which proves the Property 1■

3.1.1 The orthogonal projection (*OP*) context

When comparing **T** and **U** in the context of *orthogonal projections* (*M* = *OP*), the associated regression coefficient matrices are particularly simple and closely related, i.e.

Property 2

$$SMI_{OP}(\mathbf{T},\mathbf{U}) = max\left(\frac{\|\mathbf{B_T}\|_F^2}{p}, \frac{\|\mathbf{B_U}\|_F^2}{q}\right) = \frac{\|\mathbf{T}^t\mathbf{U}\|_F^2}{r}, \qquad (4)$$

where $r = min(p,q)$.

Proof:

From our initial assumptions $\mathbf{T}^t\mathbf{T} = \mathbf{I}_p$ and $\mathbf{U}^t\mathbf{U} = \mathbf{I}_q$, we have $\mathbf{B_T} = (\mathbf{T}^t\mathbf{T})^{-1}\mathbf{T}^t\mathbf{U} = \mathbf{T}^t\mathbf{U}$ and $\mathbf{B_U} = (\mathbf{U}^t\mathbf{U})^{-1}\mathbf{U}^t\mathbf{T} = \mathbf{U}^t\mathbf{T} = \mathbf{B_T^t}$. Consequently $\|\mathbf{B_U}\|_F^2 = \|\mathbf{B_T^t}\|_F^2 = \|\mathbf{B_T}\|_F^2 = \|\mathbf{T}^t\mathbf{U}\|_F^2$, and the maximum in equation (3) is clearly obtained by dividing $\|\mathbf{T}^t\mathbf{U}\|_F^2$ with the minimum of p and q ■

In the nontrivial case ($\mathbf{T}^t\mathbf{U} \neq \mathbf{0}$) with $\mathbf{T}^t\mathbf{U} = \mathbf{VSW}^t$ being the compact singular value decomposition (SVD) of the ($p \times q$) matrix $\mathbf{T}^t\mathbf{U}$, it is clear that the associated squared Frobenius norm in the OP context only depends on the nonzero singular values $s_1, s_2, \ldots, s_r$ (where $r = min(p,q)$ if $\mathbf{T}^t\mathbf{U}$ has full rank) and the following property holds:

Property 3

$$SMI_{OP}(\mathbf{T},\mathbf{U}) = \frac{\|\mathbf{S}\|_F^2}{r} = \frac{1}{r}\sum_{k=1}^{r} s_k^2, \qquad (5)$$

where $s_1, s_2, \ldots, s_r$ are the singular values of the $p \times q$ matrix $\mathbf{T}^t\mathbf{U}$ and $r = min(p,q)$.

Proof:

Let the SVD of $\mathbf{T}^{\mathrm{t}}\mathbf{U} = \mathbf{V}\mathbf{S}\mathbf{W}^{\mathrm{t}}$ where the singular values $s_1, s_2, \ldots, s_r$ correspond to the diagonal elements of **S**. Then $\|\mathbf{T}^{\mathrm{t}}\mathbf{U}\|_F^2 = trace((\mathbf{T}^{\mathrm{t}}\mathbf{U})^{\mathrm{t}}(\mathbf{T}^{\mathrm{t}}\mathbf{U})) = trace(\mathbf{W}\mathbf{S}\mathbf{V}^{\mathrm{t}}\mathbf{V}\mathbf{S}\mathbf{W}^{\mathrm{t}})$

$= trace(\mathbf{W}\mathbf{S}^2\mathbf{W}^{\mathrm{t}}) = trace(\mathbf{W}^{\mathrm{t}}\mathbf{W}\,\mathbf{S}^2) = trace(\mathbf{S}^2) = \|\mathbf{S}\|_F^2 = \sum_{k=1}^{r} s_k^2$, and the result therefore follows from equation (4) ■

According to equation (5), the SMI_{OP} is simplified to the average of the squared (non-zero) singular values of $\mathbf{T}^{\mathrm{t}}\mathbf{U}$ in the *OP* context.

By noting that the Frobenius norm is fixed when multiplying $\mathbf{T}^{\mathrm{t}}\mathbf{U}$ from the left and right by the orthogonal matrices **T** and **U**, respectively, the following property also holds:

Property 4

$$SMI_{OP}(\mathbf{T},\mathbf{U}) = \frac{(\mathbf{P_T}\cdot\mathbf{P_U})}{min(\mathbf{P_T}\cdot\mathbf{P_T},\mathbf{P_U}\cdot\mathbf{P_U})}, \qquad (6)$$

where $\mathbf{P_T} = \mathbf{T}\mathbf{T}^{\mathrm{t}}$ and $\mathbf{P_U} = \mathbf{U}\mathbf{U}^{\mathrm{t}}$ are the $n \times n$ projection matrices associated with the subspaces spanned by **T** and **U** respectively, and $(\cdot)$ represent the trace inner product between $n \times n$ matrices.

Proof:

The Frobenius norm being fixed when multiplying $\mathbf{T}^{\mathrm{t}}\mathbf{U}$ from the left and right by the orthogonal matrices **T** and **U**, means that $\|\mathbf{T}^{\mathrm{t}}\mathbf{U}\|_F^2 = \|\mathbf{T}\mathbf{T}^{\mathrm{t}}\mathbf{U}\mathbf{U}^{\mathrm{t}}\|_F^2 = tr(\mathbf{P_T}\mathbf{P_U}) = \mathbf{P_T}\cdot\mathbf{P_U}$. Because $\mathbf{P_T}\cdot\mathbf{P_T} = \|\mathbf{T}^{\mathrm{t}}\mathbf{T}\|_F^2 = \|\mathbf{T}\|_F^2 = p$ and $\mathbf{P_U}\cdot\mathbf{P_U} = \|\mathbf{U}\|_F^2 = q$, equation (6) is obtained by making the obvious substitutions into equation (4) ■

Equation (6) shows that for the centered matrices **T** and **U**, SMI_{OP} is proportional (by multiplication with the scalar $min(p,q)/\sqrt{pq}$) to the correlation between the projection matrices $\mathbf{P_T}$ and $\mathbf{P_U}$ (being considered as n^2 dimensional vectors).

Some comments:

1. The nonzero singular values s_k of $\mathbf{T}^t\mathbf{U}$ coincide with the cosine of the principal angles between the column spaces associated with **T** and **U** or equivalently the associated *canonical correlations* (ρ_k), i.e. $s_k = \rho_k$ ($k = 1,\ldots,r$), see Björck and Golub (1973).
2. Canonical correlations are fixed under rank preserving linear transformations. Hence, for any pair of matrices ($\mathbf{X}_1$, $\mathbf{X}_2$) where the associated column subspace identities $Col(\mathbf{X}_1) = Col(\mathbf{T})$ and $Col(\mathbf{X}_2) = Col(\mathbf{U})$ hold, a canonical analysis of ($\mathbf{X}_1$, $\mathbf{X}_2$) will, according to property 3 (Equation (5)), provide the singular values required for computing $SMI_{OP}(\mathbf{T},\mathbf{U})$.
3. According to Darlington et al. (1973), the remarkable link between the sum of squared canonical correlations and the shared variance between two sets of factors was first proposed by Wrigley and Neuhaus (1955).
4. In the particular situations where **T** and **U** are chosen to span the entire column spaces of $\mathbf{X}_1$ and $\mathbf{X}_2$ respectively, the corresponding projection matrices $\mathbf{P_{X_1}} = \mathbf{P_U}$ and $\mathbf{P_{X_2}} = \mathbf{P_T}$. Yanai (1974) introduced a similarity measure commonly known as *Yanai's generalized coefficient of determination* (*GCD*) by defining $D(\mathbf{X_1},\mathbf{X_2}) = \frac{(\mathbf{P_{X_1}}\cdot\mathbf{P_{X_2}})}{\|\mathbf{P_{X_1}}\|\|\mathbf{P_{X_2}}\|}$. The *GCD* is essentially calculating the correlation between the two projection matrices $\mathbf{P_{X_1}}$ and $\mathbf{P_{X_2}}$. According to our remark after the proof of equation (4), the SMI_{OP} and the *GCD* are

proportional by the scaling factor ($min(p,q)/\sqrt{pq}$), and if $rank(\mathbf{X}_1) = rank(\mathbf{X}_2)$ ($p = q$) the two measures coincide.

3.1.2 The Procrustes Rotation (*PR*) context

In the *OP* context, the associated matrices of regression coefficients (**B**) are derived without any imposed restrictions, and complete similarity ($SMI_{OP}(\mathbf{T},\mathbf{U}) = 1$) occurs if one of the matrices is an exact linear transformation of the other.

Procrustes Rotations, see Gower (1975), represents an interesting closely related alternative for measuring similarities when additional restrictions imposed on the regression coefficients **B** are required. A typical area of application is sensory analysis (Amerine et al. (1965)) where two assessors may perceive the same underlying dimensions, but one assessor reverses, say, the first two dimensions as compared to the other. In such cases it may be particularly useful to consider the SMI_{OP} together with an alternative similarity measure taking the *PR* aspect into account.

To formulate the *PR* context of similarity, we start by considering matrices **T** and **U** of identical size and rank ($p = q$). In particular, the required transformation matrix $\mathbf{B_T}$ is proportional to an orthogonal matrix **R** by some scaling constant (g) so that $\mathbf{B_T} = g\mathbf{R}$. The argument simplifying the *SMI* in the *OP* context (property 3) is valid also for the *PR* context, and extends further (because $p = q = r$ by assumption) into

Property 5

$$SMI_{PR}(\mathbf{T},\mathbf{U}) = max\left(\frac{\|\mathbf{B_T}\|_F^2}{p}, \frac{\|\mathbf{B_U}\|_F^2}{q}\right) = \frac{\|\bar{s}\mathbf{R}\|_F^2}{r} = \bar{s}^2\,\frac{\|\mathbf{R}\|_F^2}{r} = \bar{s}^2, \qquad (7)$$

i.e. the squared average of the associated singular values (canonical correlations).

Proof:

The optimal choice for **R** and the associated scaling constant g to obtain $\mathbf{B_T} = g\mathbf{R}$ is derived from the SVD of $(\mathbf{T}^t\mathbf{U}) = \mathbf{VSW}^t$ by defining

$$\mathbf{R} = \mathbf{VW}^t \text{ and the scalar } g = tr(\mathbf{S})/\|\mathbf{T}\|_F^2 = \bar{s}, \qquad (8)$$

where $\bar{s} = \frac{1}{r}\sum_{k=1}^{r} s_k$ is the average of the non-zero singular values of $\mathbf{T}^t\mathbf{U}$. Correspondingly, $\mathbf{B_U} = \bar{s}\mathbf{R}^t = \mathbf{B}_\mathbf{T}^t$, i.e. the *PR* regression coefficients are derived from the *OP* regression coefficients by interchanging the singular values of $\mathbf{T}^t\mathbf{U}$ by their average■

If the number of columns in **T** and **U** are different ($p \neq q$), and the SVD of $(\mathbf{T}^t\mathbf{U}) = \mathbf{VSW}^t$ with $\mathbf{R} = \mathbf{VW}^t$, we have $\|\mathbf{R}\|_F^2 = \|\mathbf{R}^t\|_F^2 = \min(p, q) = r$ corresponding to the number of non-zero singular values (except for degenerate cases) in **S**. By taking Equation (8) as an alternative definition of SMI_{PR}, we obtain an obvious extension of the *PR* context to the general situation also including matrices **T** and **U** where the number of columns differ ($p \neq q$).

Due to the restrictions imposed on the regression coefficients **B** in the *PR* context, SMI_{OP} obviously dominates SMI_{PR}, and their exact relationship is given by the following property:

Property 6

The difference between SMI_{OP} and SMI_{PR}

$$SMI_{OP}(\mathbf{T}, \mathbf{U}) - SMI_{PR}(\mathbf{T}, \mathbf{U}) = \frac{1}{r}\sum_{k=1}^{r}(s_k - \bar{s})^2 \geq 0, \qquad (9)$$

i.e. it equals the empirical variance of the *r* non-zero singular values (canonical correlations) associated with the matrix product $\mathbf{T}^t\mathbf{U}$.

Proof:

Using the properties 3 and 5 the following is obtained

$$SMI_{OP}(\mathbf{T},\mathbf{U}) - SMI_{PR}(\mathbf{T},\mathbf{U}) = \left(\frac{1}{r}\sum_{k=1}^{r} s_k^2\right) - \bar{s}^2 = \frac{1}{r}\sum_{k=1}^{r}(s_k - \bar{s})^2 \geq 0\blacksquare$$

With reference to the arguments given above for exploring the Procrustes context, it is important to stress that the main interest when using the SMI_{PR} lies in comparing it with the SMI_{OP}. If the two measures result in very different values for a particular data set, that is strong evidence of the information in the two matrices not being satisfactory accounted for by a rotation and scaling only.

3.2 Permutation testing

When there is a strong linear (or rotational) relationship between the measurement variables recorded in $\mathbf{X}_1$ and $\mathbf{X}_2$, one can expect the associated orthogonal matrices **T** and **U** (representing the "stable" and/or "relevant" parts of the structure in $\mathbf{X}_1$ and $\mathbf{X}_2$) to yield large *SMI*-values in the *OP* (or *PR*) context. Analogous to the paired samples t-test the following null hypothesis is formulated:

$\mathbf{H}_0$: "The distributions from which **T** ($n \times p$) and **U** ($n \times q$) have been derived coincide so that in the case where $p \geq q$, the **U**-factors can be expressed as linear (rotated) combinations of the **T**-factors."

Hence, on can reject $\mathbf{H}_0$ and conclude ($\mathbf{H}_1$) that "…the **U**-factors cannot be expressed as linear (rotated) combinations of the **T**-factors…" unless the associated test statistic *SMI*(**T**,**U**) is sufficiently close to 1. Note that for $p \geq q$, *SMI*(**T**,**U**) = 1 if and only if the **U**-columns can be linearly transformed (rotated) into the **T**-columns.

A distribution for *SMI*(**T**,**U**) is not likely to be analytically available, but a procedure for testing $\mathbf{H}_0$ based on random permutations can be justified by considering *the residual similarity of matrices index* defined as $SMI_{res} = 1 - SMI$ (we omit the matrix arguments (**T**,**U**) in the notation from now on). According to the essence of the equations (2) and (4), this definition relates to the corresponding classical ANOVA identity $SS_{res} = SS_{tot} - SS_{reg}$ by multiplying throughout the SMI_{res} with the factor $SS_{tot} = \min(p, q)$. Under the null hypothesis one can expect large *SMI*-values and correspondingly small SMI_{res}-values (unless the number of samples n is close to $min(p, q)$).

The appropriate random sampling of "small" SMI_{res}-values can be obtained by a large number Π (Π = 100000 is used in our examples) of repeated calculations of $SMI_{(perm)} = SMI(\mathbf{T}, \mathbf{U}_{(perm)})$, where $\mathbf{U}_{(perm)}$denotes a permutation of the rows in **U** (by simple symmetry both **T** and **U**, or **T** alone can be permuted in this fashion for the same purpose). By considering the resulting $SMI_{res} = 1 - SMI_{(perm)}$ values as a random sample from the underlying distribution of *SMI*-values consistent with $\mathbf{H}_0$, the $\mathbf{H}_0$ is rejected at the significance level $\alpha > 0$ if the observed *SMI*-value (measured for the original matrices **T** and **U**) is smaller than the empirical $(1 - \alpha)$ percentile of the randomly sampled SMI_{res}-values obtained from the permutation procedure.

By implementing the proposed permutation testing procedure using Π = 100000 random permutations, one can obtain good estimates of the P-values associated with the null distribution by calculating:

$$P = 1 - \frac{\#\left(1 - SMI_{(perm)} \geq SMI\right)}{\Pi} = \frac{\#\left(SMI > 1 - SMI_{(perm)}\right)}{\Pi}.$$

Recall that the score matrices **U** and **T** in our formulas are typically obtained by applying either PCA or PLS to the original data matrices $\mathbf{X}_1$ and $\mathbf{X}_2$. The reason why the proposed permutation

scheme is sound, is that any permutation of the rows in **U** or **T** will also result by applying PCA (or PLS) to the matrix obtained by the identical permutation of the rows in the corresponding $\mathbf{X}_1$ or $\mathbf{X}_2$. Complete PCA- or PLS remodeling from permuted versions of $\mathbf{X}_1$ and $\mathbf{X}_2$ is therefore unnecessary, and this ensures that the permutation part required for sampling from the null distribution can be executed with high efficiency.

Note that when the minimum number of columns $\min(p, q)$ in **T** and **U** is close to the number n of rows in $\mathbf{X}_1$and $\mathbf{X}_2$, the associated *SMI*–value will tend to be large because the columns of both **T** and **U** then are spanning relatively "large" subspaces of the n-dimensional Euclidean space $\boldsymbol{R^n}$. In such cases, it is therefore recommend to avoid using the proposed significance testing. To prevent against possible misuses of the testing procedure in such cases, an alternative suggestion is to consider the following modified P-value estimate

$$P_{mod} = \frac{\#\left(SMI > \max(1 - SMI_{(perm)},\ SMI_{(perm)})\right)}{\Pi}$$

as a more robust alternative for implementations of the *SMI*-framework.

3.3 The stepwise guide to exploring subspace similarities

According to our assumptions, the data matrices $\mathbf{X}_1$ and $\mathbf{X}_2$ are the results of recording two different sets of measurement variables for a common set of samples (followed by column mean centering of both matrices). The following three steps summarize the *SMI*-based data analysis procedure:

Step 1 – the score matrices and variable combinations

According to Equation (1), we compute the score matrices and associated variable combinations

(the coefficient matrices $\mathbf{C}_1$ and $\mathbf{C}_2$) satisfying $\mathbf{T} = \mathbf{X}_1\mathbf{C}_1$ and $\mathbf{U} = \mathbf{X}_2\mathbf{C}_2$. Depending on the purpose of the study, various alternatives may be considered. In the examples shown below, focus is on i) Principal Component Analysis (PCA), which is appropriate for investigating when one suspects that the subspaces of stable and dominant variance associated with the two matrices coincide, ii) Partial Least Squares (PLS) regression, which is appropriate when comparing the validated predictive parts of two data matrices with respect to one or more response variables.

Note that for applications based on PCA, *stability* of the subspaces spanned by **T** and **U** in (1) can be assessed, i.e. by comparing the condition number of the associated loading matrices $\mathbf{P}_1$ and $\mathbf{P}_2$ to some threshold value θ. This is closely related to the consideration of *scree plots* (showing the proportions of variance accounted for by including particular components), see Joliffe (2002) that includes more methods for choosing the appropriate number of PCA components. For PLS, a validation step such as cross-validation (CV) or bootstrapping (Efron and Tibshirani (1993)) may be required for a stable and robust choice of columns to include in **T** and **U**.

Step 2 – the SMI calculations

Equations (4) and (5) represent equivalent alternatives for calculating SMI_{OP}, with (4) as the computationally most efficient alternative. By equation (8), the SMI_{PR} requires an explicit calculation of the singular values of $\mathbf{T}^t\mathbf{U}$. In situations where also consideration of the SMI_{PR} is desired, the associated singular values will also be available for a fast additional computation of SMI_{OP} according to equation (5).

Step 3 – statistical significance and visualization of the results

Let $\mathbf{T}_{(1:p)}$ and $\mathbf{U}_{(1:q)}$ denote the first $p \leq m_1$ and $q \leq m_2$ columns of **T** and **U** respectively, and define $SMI_{OP}^{i,j} = SMI_{OP}\left(\mathbf{T}_{(1:p)}, \mathbf{U}_{(1:q)}\right)$ and $SMI_{PR}^{i,j} = SMI_{PR}\left(\mathbf{T}_{(1:p)}, \mathbf{U}_{(1:q)}\right)$. From the various

possible (*i, j*)-combinations, one can generate a detailed view of the subspace relationships (and associated variable combinations for the two sets of measurements) by considering the *diamond plot* (an example of this plot is shown in Figure 3 below). The diamond plot provides a compact display of the $SMI^{i,j}$-value combinations (shown as grey-level intensities) and their associated statistical significances. The set symbols ("⊃", "⊂" and "=") and significance stars ("*", "**" and "***") in each cell (*i, j*) denote the following relationships:

- No rejection of the null hypothesis associated with $\mathbf{T}_{(1:i)}$ and $\mathbf{U}_{(1:j)}$ is indicated by "=" if $i = j$ and by the subset symbols "⊂" or "⊃" if $i < j$ or $i > j$, respectively. A cell *(i, j)* containing "*", "**" or "***" indicates the significance level for rejection of the null hypothesis (at the 0.05, 0.01 and 0.001 levels, respectively) in the associated comparison.

Practical use of the diamond plot will be illustrated in the examples below.

3.4 Standard criteria for correlation coefficients

By definition, the proposed similarity indices (SMI_{OP} and SMI_{PR}) apply only to pairs of orthogonal matrices **T** and **U** associated with the original data matrices $\mathbf{X}_1$ and $\mathbf{X}_2$. Under this restriction, the proposed similarity indices (SMI_{OP} and SMI_{PR}) relate to the requirements (see Ramsay et al. (1984)) of a matrix correlation measure (r) as follows:

1. $r(a\mathbf{X}_1, \mathbf{X}_2) = r(\mathbf{X}_1, b\mathbf{X}_2) = r(\mathbf{X}_1, \mathbf{X}_2)$ - invariance by scalar multiplication: When the method for extracting the orthogonal score matrices (**T** and **U**) is invariant under scalar multiplications of the original data matrices $\mathbf{X}_1$ and $\mathbf{X}_2$, this property holds for both SMI_{OP} and SMI_{PR}. In particular it holds when **T** and **U** are derived by PCA or PLS.

2. $r(\mathbf{X}_1, \mathbf{X}_2) = r(\mathbf{X}_2, \mathbf{X}_1)$ - symmetry: This property holds for both SMI_{OP} and SMI_{PR} because the non-zero singular values of $\mathbf{T}^t\mathbf{U}$ and $\mathbf{U}^t\mathbf{T}$ are identical.

3. $r(\mathbf{X}, \mathbf{X}) = 1$ - comparison of identical matrices: This holds for both SMI_{OP} and SMI_{PR} because for $\mathbf{U}=\mathbf{T}$, $\mathbf{T}^t\mathbf{U} = \mathbf{T}^t\mathbf{T} = \mathbf{I}$ (the identity matrix) and the associated singular values are all identical to 1.

4. $r(\mathbf{X}_1, \mathbf{X}_2) = 0$ if and only if $\mathbf{X}_1^t\mathbf{X}_2 = 0$ - orthogonality between matrices: This holds for both SMI_{OP} and SMI_{PR} and follows from the fact that only the *0*-matrices have their singular values identical to 0.

In the special case where $\mathbf{X}_1$ and $\mathbf{X}_2$ are vectors, i.e. $\mathbf{X}_1 = \mathbf{x}_1$, $\mathbf{X}_2 = \mathbf{x}_2$ and p=q=1, we have $\mathbf{T} = \mathbf{t} = \mathbf{x_1}/\|\mathbf{x_1}\|$, $\mathbf{U} = \mathbf{u} = \mathbf{x_2}/\|\mathbf{x_2}\|$ and $SMI_{OP}(\mathbf{t}, \mathbf{u}) = (\mathbf{t}^t\mathbf{u})^2 = corr(\mathbf{x_1}, \mathbf{x_2})^2$, i.e. the squared Pearson correlation between $\mathbf{x}_1$ and $\mathbf{x}_2$. Consequently, the proposed permutation testing also gives a valid inference alternative for the squared Pearson correlation, i.e. a possibility of rejecting the null hypothesis, and conclude that two vectors being compared are "not highly correlated" i.e. they do not share a common subspace.

3.5 Relations to indices proposed in the literature

In the literature, in particular the reviews given by Ramsay et al. (1984) and Cramer and Nicewander (1979), there are numerous suggestions of how to define and calculate correlation measures for matrices.

According to Ramsay et al. (1984), the most frequently used among the measures of matrix correlation between two (n×p) matrices $\mathbf{X}_1$ and $\mathbf{X}_2$ is

$$r_1(\mathbf{X}_1, \mathbf{X}_2) = trace(\mathbf{X}_1^t\mathbf{X}_2)/(trace(\mathbf{X}_1^t\mathbf{X}_1)trace(\mathbf{X}_2^t\mathbf{X}_2))^{1/2}. \quad (10)$$

Here, the function r_1 corresponds to the ordinary Pearson correlation function for *np*-dimensional vectors (as obtained by stacking the matrix columns on top of each other). Note that $|r_1|$ satisfies the requirements 1-4 above.

Alternatively, one may suggest measuring the relationship between $\mathbf{X}_1$ and $\mathbf{X}_2$ by applying formula (10) to the associated orthogonal matrices **T**, **U** of $\mathbf{X}_1$ and $\mathbf{X}_2$, i.e.

$$r_1(\mathbf{T},\mathbf{U}) = tr(\mathbf{T}^t\mathbf{U})/(tr(\mathbf{T}^t\mathbf{T})tr(\mathbf{U}^t\mathbf{U}))^{1/2}. \qquad (11)$$

The inherent ambiguity with respect to the choice of directions in the **T**- and **U** basis vectors of Equation (11) makes uncritical applications of the r_1-function problematic, because the diagonal elements in the trace calculation of the numerator may cancel even when **T** and **U** span the same subspace. The following example illustrates the problem:

Example

Consider the orthogonal matrices

$$\mathbf{T} = \begin{bmatrix} 1/\sqrt{2} & -1/\sqrt{6} \\ 0 & 2/\sqrt{6} \\ -1/\sqrt{2} & -1/\sqrt{6} \end{bmatrix}, \ \mathbf{U} = \begin{bmatrix} 1/\sqrt{2} & 1/\sqrt{6} \\ 0 & -2/\sqrt{6} \\ -1/\sqrt{2} & 1/\sqrt{6} \end{bmatrix}.$$

Obviously $SMI_{OP}(\mathbf{T},\mathbf{U}) = SMI_{PR}(\mathbf{T},\mathbf{U}) = 1$, but $r_1(\mathbf{T},\mathbf{U}) = 0$ because the trace $tr(\mathbf{T}^t\mathbf{U}) = tr\begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} = 0$.

It should be noted that the idea of calculating an index for reduced matrices have been suggested earlier (see the paragraph on "*Clipping Transformations"* in Ramsey et al. (1984), page 409). We believe that the geometrical aspects of the proposed *SMI*-framework and the associated

visualization method for the significance testing may ignite renewed interest in the subject both from the applied and theoretical points of view.

The popular RV coefficient of Robert and Escoufier (1976) is mentioned both in the introduction, and in several of the examples presented below. There are alternative equivalent formulations of the RV coefficient, and one of them is based on the definition of the r_1-function in Equation (10):

$$RV(\mathbf{X}_1, \mathbf{X}_2) = r_1(\mathbf{X}_1\mathbf{X}_1^{\mathrm{t}}, \mathbf{X}_2\mathbf{X}_2^{\mathrm{t}}) = r_1(\mathbf{T}\mathbf{S}_1\mathbf{V}_1^{\mathrm{t}}\mathbf{V}_1\mathbf{S}_1\mathbf{T}^{\mathrm{t}}, \mathbf{U}\mathbf{S}_2\mathbf{V}_2^{\mathrm{t}}\mathbf{V}_2\mathbf{S}_2\mathbf{S}_2^2\mathbf{U}^{\mathrm{t}}) = r_1(\mathbf{T}\mathbf{S}_1^2\mathbf{T}^{\mathrm{t}}, \mathbf{U}\mathbf{S}_2^2\mathbf{U}^{\mathrm{t}}). \quad (12)$$

Here $\mathbf{T}$ denotes the left singular vectors- and $\mathbf{S}_1$ the non-zero singular values of $\mathbf{X}_1$, and $\mathbf{U}$ denotes the left singular vectors- and $\mathbf{S}_2$ the non-zero singular values of $\mathbf{X}_2$. $\mathbf{V}_1$ and $\mathbf{V}_2$ denotes the corresponding right singular vectors. The rightmost expression in (12) shows that the squared singular values acts as weights for the various left singular vector directions in their contributions to the RV coefficient. Note that the r_1-canceling problem demonstrated in the above example is avoided for the RV coefficient. This is because the associated trace summations defining r_1 involve positive numbers (squares) only.

By substituting the diagonal singular value matrices $\mathbf{S}_1$ and $\mathbf{S}_2$ in (12) with identity matrices of corresponding size, we obtain a situation where all the singular vectors are treated as equally important and the resulting computation is $RV(\mathbf{T}, \mathbf{U}) = r_1(\mathbf{T}\mathbf{T}^{\mathrm{t}}, \mathbf{U}\mathbf{U}^{\mathrm{t}})$. By recalling (from the introduction) that Yanai's $GCD(\mathbf{X}_1, \mathbf{X}_2) = RV(\mathbf{T}, \mathbf{U})$, it can be concluded that $GCD(\mathbf{X}_1, \mathbf{X}_2) = r_1(\mathbf{T}\mathbf{T}^{\mathrm{t}}, \mathbf{U}\mathbf{U}^{\mathrm{t}})$, i.e. the Pearson correlation between the projection matrices $\mathbf{T}\mathbf{T}^{\mathrm{t}}$ and $\mathbf{U}\mathbf{U}^{\mathrm{t}}$ that indeed corresponds to Yanai's (1974) original definition of the *GCD*. It should be noted that in the original definitions of both $RV(\mathbf{X}_1, \mathbf{X}_2)$ and $GCD(\mathbf{X}_1, \mathbf{X}_2)$, the complete matrices of left singular vectors $\mathbf{T}$ and $\mathbf{U}$ (associated with the respective sets of non-zero singular values) are included in the calculations.

The fundamental idea of the proposed *SMI*-framework is to vary the numbers $p \leq m_1$ and $q \leq m_2$ of included columns for systematic comparison of the reduced matrices $\mathbf{T} = \mathbf{T}_{(1:p)}$ and $\mathbf{U} = \mathbf{U}_{(1:q)}$. Note that an obvious partial version of the *GCD* is obtained when using the indicated reduced versions of **T** and **U**. Within a scaling factor depending on p and q (see comment 4 in Section 3.1.1) the partial *GCD* is related to SMI_{OP} as follows:

From the trace identities

$$tr(\mathbf{U}^t\mathbf{T}\mathbf{T}^t\mathbf{U}) = tr(\mathbf{T}\mathbf{T}^t\mathbf{U}\mathbf{U}^t), \quad tr(\mathbf{T}\mathbf{T}^t) = tr(\mathbf{T}^t\mathbf{T}) = p \quad \text{and}$$

$$tr(\mathbf{U}\mathbf{U}^t) = tr(\mathbf{U}^t\mathbf{U}) = q,$$

and by noting that $min(p,q)/\sqrt{pq} = \sqrt{min\left(\frac{p}{q},\frac{q}{p}\right)}$, the following equations hold

$$GCD(\mathbf{T},\mathbf{U}) = RV(\mathbf{T},\mathbf{U}) = SMI_{OP}(\mathbf{T},\mathbf{U})\sqrt{min\left(\frac{p}{q},\frac{q}{p}\right)}. \qquad (13)$$

It should also be noticed that there is a simple connection between the analogous partial version of the PSI (mentioned in the introduction) and SMI_{PR}. Directly from their respective definitions, it follows that $PSI(\mathbf{T},\mathbf{U})^2 = SMI_{PR}(\mathbf{T},\mathbf{U})$ for the orthogonal matrices **T** and **U** and the Procrustes transformation $\mathbf{H} = g\mathbf{R}$ resulting from equation (14).

As pointed out by Smilde et al. (2009), the RV-coefficient suffers from an increasing bias (towards 1) when the number of variables (columns) increase compared to the number of samples (rows) in the two matrices. They therefore proposed the RV2-coefficient as a bias reducing (but still biased) alternative to the RV-coefficient. Arguing that the main problem of the RV-coefficient is numerator inflation due to the guaranteed positive diagonals of $\mathbf{X}_1\mathbf{X}_1^t$ and $\mathbf{X}_2\mathbf{X}_2^t$, the two diagonals are simply set to 0 in the RV2-coefficient. Mayer et al. (2011) pointed at some

fundamental problems with RV2 and proposed a corrected and unbiased alternative to the RV- (and RV2) directly based on the classical *adjusted R-square* $r^2_{adj}(\mathbf{x},\mathbf{y}) = 1 - \frac{n-1}{n-2}(1 - r^2(\mathbf{x},\mathbf{y}))$ statistic between two *n*-dimensional vectors **x** and **y**. It should be noted that the RV_{gq}-coefficient proposed by El Ghaziri and Qannari (2015) is also unbiased.

Some interesting properties of the alternative RV-coefficients can be illustrated through two simple simulations with random data. We first create a 1000 x 1000 orthogonal matrix and select columns from this to form two matrices spanning from 1 column to 500 columns wide, having no common subspace. Second, we sample standard normal values to fill two 1000 x *p* matrices with randomly overlapping, non-structured subspaces. The results of applying RV, RV2, RV_{adj}, RV_{gq} and SMI_{OP} (using the maximum possible number of components, i.e. equal to GCD) are displayed in **Figure 1**.

For matrices of reasonable dimensions, one would hope to see only 0 coefficients in both simulations. In the case of non-overlapping subspaces, one can observe that RV, RV_{gq} and SMI_{OP} are indeed 0, while RV2 and RV_{adj} decrease to a value of -1 as *p* approaches 500. The latter would imply maximum negative correlation, which is counter intuitive as the spaces spanned by the matrices are orthogonal. In the case of random matrices, SMI_{OP} is the only measure starting at 0, though as expected the proportion of overlap between the subspaces spanned increases linearly until the overlap is complete at $p = 1000$. RV, RV2, RV_{adj} start at 0.57 and increase past 0.9 already at $p = 7$. RV_{gq} starts at -0.7531, but also increases toward 1, though only passing 0.9 as *p* nears 40.

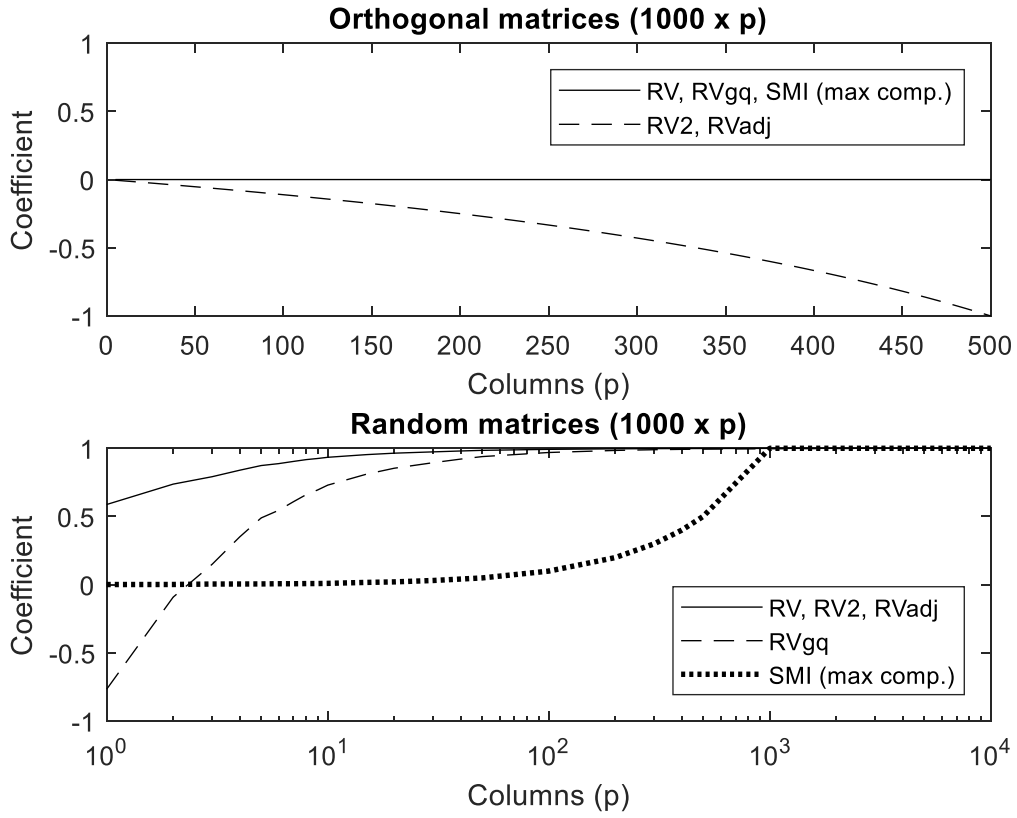

Figure 1 - Matrix correlations of noise matrices of varying number of columns (N=1000). Upper: two orthonormal matrices spanning orthogonal subspaces. Lower: two standard normal random matrices with overlapping subspaces.

4. Examples

In each of the examples, Π = 100,000 random permutations have been used for calculating the reported p-values. For the sake of comparison, the resulting values for both SMI_{OP} and SMI_{PR} (together with some of the other indices) were included in all examples. A complete list of coefficient values for all examples is found in the Supplementary Material. When appropriate, we discuss reasons for consistencies and discrepancies. One of the examples is about prediction, and

PLS has been used for finding the orthogonal matrices **T** and **U** in that one. In the remaining examples, PCA have been used for finding **T** and **U**.

4.1. A simulated example where the RV-coefficient fails

This example illustrates a simple situation with two matrices $\mathbf{X}_1$ and $\mathbf{X}_2$ (see Figure 2, and their numerical values in Appendix Table 1) of size (8×2) and associated orthogonal PCA-score matrices **T** and **U** of identical size. In this situation, the classical RV-coefficient is RV($\mathbf{X}_1$, $\mathbf{X}_2$) = 0.07 and fails to indicate the obvious geometrical relationship in the measurements. On the other hand, both $SMI_{OP}^{2,2}(\mathbf{X}_1, \mathbf{X}_2) = \mathrm{GCD}(\mathbf{X}_1, \mathbf{X}_2) = 0.89$ and $SMI_{PR}^{2,2}(\mathbf{X}_1, \mathbf{X}_2) = 0.89$. The explanation of the disagreement between the RV-coefficient and the *SMI*-values is that the dominant score vector $\mathbf{t}_1$ (the first column of **T**) of $\mathbf{X}_1$ is highly correlated with the second score vector $\mathbf{u}_2$ (the second column of **U**) of $\mathbf{X}_2$ and vice versa, i.e. $corr(\mathbf{t}_1, \mathbf{u}_2) = 0.95$, $corr(\mathbf{t}_2, \mathbf{u}_1) = 0.93$, and $corr(\boldsymbol{t}_1, \boldsymbol{u}_1) = -0.01$. The obvious conclusion based on the RV-coefficient (no relationship between the two datasets) is overwhelmingly inconsistent with the geometrical picture in Figure 2 and the large associated $SMI^{2,2}(\mathbf{X}_1, \mathbf{X}_2)$ -values.

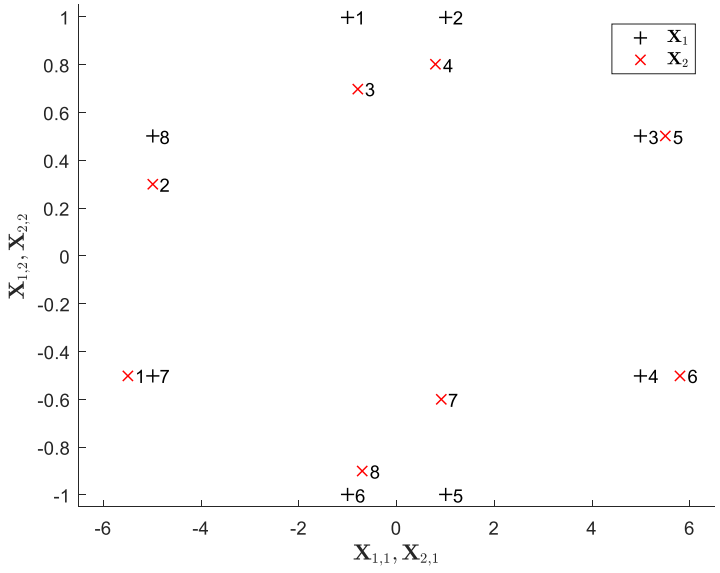

Figure 2 – The two configurations of points essentially differ by rotation and scaling only.

The permutation testing associated with $SMI_{OP}^{1,1}(\mathbf{X}_1, \mathbf{X}_2)$ and $SMI_{PR}^{1,1}(\mathbf{X}_1, \mathbf{X}_2)$ leads to rejection of $\mathbf{H}_0$ at any significance level, indicating significant evidence against collinearity between the two dominant principal components. The other *SMI*-value combinations results in P-values > 0.5 for both *OP* and *PR*, i.e. no significant evidence against $\mathbf{H}_0$ for the associated variable combinations.

It is important to notice that the application of the RV-coefficient after standardization of the columns in $\mathbf{X}_1$ and $\mathbf{X}_2$ yields a completely different result (the new value is 0.89 and equals the $SMI^{2,2}$-values found above in the first two decimal places). This is obviously counterintuitive in perspective of the scale-invariance property of the Pearson correlation calculated between vectors. The same tendencies can be observed for the $\text{RV2}(\mathbf{X}_1, \mathbf{X}_2) = 0.05$ and $\text{RV}_{\text{adj}}(\mathbf{X}_1, \mathbf{X}_2) = 0.06$ for non-standardized data versus $\text{RV2}(\mathbf{Z}_1, \mathbf{Z}_2) = 0.88$ and $\text{RV}_{\text{adj}}(\mathbf{Z}_1, \mathbf{Z}_2) = 0.84$ for the standardized versions $(\mathbf{Z}_1, \mathbf{Z}_2)$ of the $(\mathbf{X}_1, \mathbf{X}_2)$-data, respectively. Finally, using Procrustes rotations, the $\text{PSI}(\mathbf{X}_1, \mathbf{X}_2) = 0.34$ indicate some similarity for the original data, and a relatively large similarity $\text{PSI}(\mathbf{Z}_1, \mathbf{Z}_2) = 0.94$ for the standardized data.

4.2 Two cases with simulated data

The purpose of this example is to illustrate properties of the SMI in some highly structured situations with simulated data.

In the first case, we generate a "wide" matrix $\mathbf{X}_1$ of size (100×300) by random sampling of its entries from the standard normal distribution followed by centering. The associated matrix $\mathbf{X}_2$ is constructed by eliminating the 3rd component from the SVD-expansion of $\mathbf{X}_1$. In the second case, "tall" matrices $\mathbf{X}_1$ and $\mathbf{X}_2$ of size (300×10) where generated according to the same type of random sampling- and elimination of the 3rd SVD-component.

The diamond plots in Figure 3 show the associated $SMI_{OP}^{i,j}(\mathbf{X}_1, \mathbf{X}_2)$ -values for various combinations of PCA-components for both the "wide" and the "tall" cases. Note that the *SMI*-values are large (as can be expected from our construction of $\mathbf{X}_2$) for most combinations. Note that for combinations exposing the eliminated SVD-component (of the $\mathbf{X}_2$ matrices), corresponding reductions in the *SMI*-values appear systematically. From the left plot in Figure 3, note that in the first case with matrices of size 100×300, H_0 is not rejected for the combinations associated with $SMI_{OP}^{8,10}(\mathbf{X}_1, \mathbf{X}_2)$, $SMI_{OP}^{9,9}(\mathbf{X}_1, \mathbf{X}_2)$, $SMI_{OP}^{9,10}(\mathbf{X}_1, \mathbf{X}_2)$ and $SMI_{OP}^{10,10}(\mathbf{X}_1, \mathbf{X}_2)$ (in spite of the eliminated SVD-component in $\mathbf{X}_2$). This observation clearly indicates that the proposed significance testing procedure is conservative.

The pattern formed by the SMI values in Figure 3 is consistent with the explained variance analogy of the SMI coefficient. For $SMI_{OP}^{i,i}(\mathbf{X}_1, \mathbf{X}_2)$ ($i \geq 3$) the resulting SMI-values are 2/3, 3/4, 4/5,…, 9/10 that correspond exactly to the ratios of $\mathbf{X}_2$-dimensions contained in the associated $\mathbf{X}_1$-dimensions, as the 3rd SVD-component of $\mathbf{X}_1$ is absent from $\mathbf{X}_2$.

The associated RV-coefficient values are $RV(\mathbf{X}_1, \mathbf{X}_2) = 0.98$ and $RV2(\mathbf{X}_1, \mathbf{X}_2) = RV_{adj}(\mathbf{X}_1, \mathbf{X}_2) = 0.92$ for the 100×300 matrices, while $PSI(\mathbf{X}_1, \mathbf{X}_2) = 0.99$ and $GCD(\mathbf{X}_1, \mathbf{X}_2) = 1$. For the 300×10 matrices one can observe that $RV(\mathbf{X}_1, \mathbf{X}_2) = RV2(\mathbf{X}_1, \mathbf{X}_2) = RV_{adj}(\mathbf{X}_1, \mathbf{X}_2) = 0.94$, while $PSI(\mathbf{X}_1, \mathbf{X}_2) = 0.94$ and $GCD(\mathbf{X}_1, \mathbf{X}_2) = 0.9$.

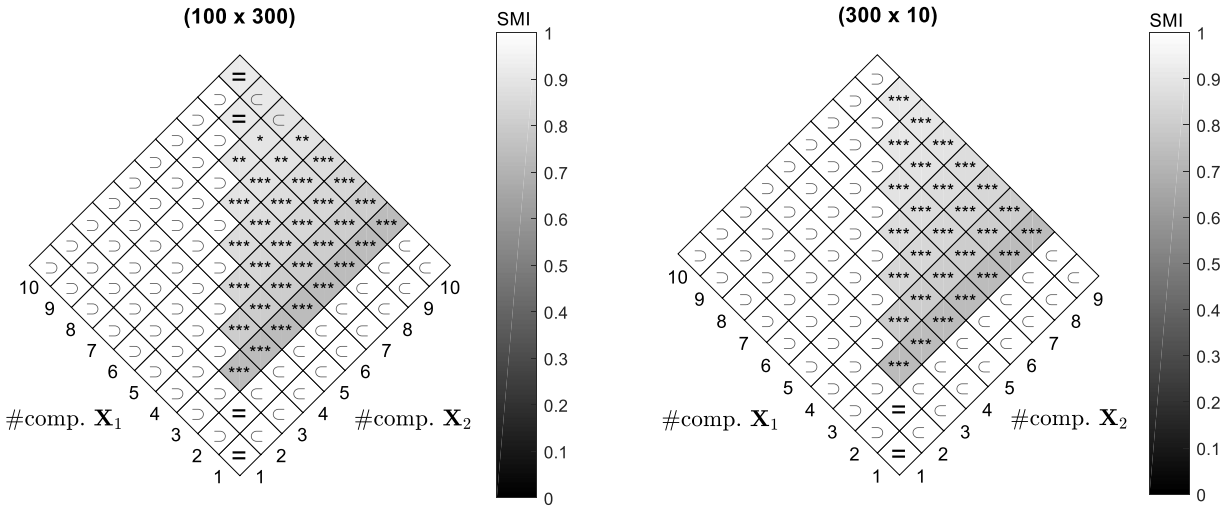

Figure 3 - *SMI*$_{OP}$ including the first 10 component combinations after centering of the matrices X_1 and X_2. The entries of X_1 are drawn randomly from the standard normal distribution (dimensions indicated in the headers). X_2 is obtained by removing the 3rd SVD component from X_1. "=", "⊂" and "⊃" indicate that H_0 is not rejected. The stars indicate rejection of H_0 at different significance levels as follows: *** = P<0.001, ** = P<0.01 and * = P<0.05.

4.3 An example from sensory science

Sensory science is a field where the RV coefficient is often included as a part of the data analysis and –interpretations, see e. g. Tomic et al. (2013). In the example shown here, the data matrices $\mathbf{X}_1$ and $\mathbf{X}_2$ represent the measurements from two sensory labs (doing professional tasting) on a number of candy products (here we can think of each lab as an "instrument" measuring some desired variables in the present context). The two data matrices considered are obtained by averaging the individual assessor score values given on each of the candy products (assessor panel averages). There were six different products (samples), which were all measured three times (3 replicates)

using six different sensory attributes resulting in an (18×6)-panel data matrix. In this particular analysis the products are treated as independent, but as the sample triplicates are expected to have similar attribute values they are grouped together in the permutation testing. The two labs and the associated data sets are parts of a larger study described in Tomic et al. (2010). In this particular type of applications it is of special interest to compare the values of SMI_{OP} and SMI_{PR}. A small difference between these values can be taken as evidence for the extent of agreement between the two panels being accounted for by a possible scaling and rotation of the underlying dimensions, only.

Using the SMI, we find that the dominant PCA-score vectors of each panel data set (accounting for 86% and 83% of the total variance) indicate a large similarity in the dominant PCA-components, i.e. $SMI_{OP}^{1,1}(\mathbf{X}_1, \mathbf{X}_2) = SMI_{PR}^{1,1}(\mathbf{X}_1, \mathbf{X}_2) = 0.93$. However, the correlation between the subsequent pair of components is relatively small (0.30), and by including the second PCA component from both panels, the associated similarities are considerably reduced, i.e. $SMI_{OP}^{2,2} = 0.52$ and $SMI_{PR}^{2,2} = 0.41$. This shows that the two panels have little correspondence in the second subspace dimension and that the difference is even larger when restricting similarity to rotation/scaling. It should also be noticed the PSI and GCD values are moderate, i.e. PSI($\mathbf{X}_1$,$\mathbf{X}_2$) = 0.68 and GCD($\mathbf{X}_1$,$\mathbf{X}_2$) = 0.59. On the other hand, we have RV($\mathbf{X}_1$, $\mathbf{X}_2$) = RV2 ($\mathbf{X}_1$, $\mathbf{X}_2$) = 0.93 and RV_{adj}($\mathbf{X}_1$, $\mathbf{X}_2$) = 0.92. The relatively large values of the RV coefficients are best explained by the strong influences of the most dominating principal component in the two panels (see above), causing the discrepancies between the data matrices along their respective second components to be much less emphasized. Since sensory science data analysis traditionally relies much on the interpretation of two or three components provided that the RV coefficient is sufficiently large, this aspect may indeed imply unfortunate conclusions.

Figure 4 shows the *SMI* values with the factor/subspace combinations traditionally used for interpretation of the sensory data (indicated by the thicker frame). Figure 5 shows the observed and fitted score values. Clearly, the two panels strongly agree on the most dominant dimension in the two datasets. The agreement when including the second dimension, however, appears as much vaguer, and the comparison by SMI_{PR} indicates a significant mismatch between the 2-dimensional representations of the panels. The mismatch for SMI_{OP} is also illustrated by Figure 5 showing a larger difference between the observed $\mathbf{T}_{(1:2)}$ and *OP*-predicted sample score-values $\hat{\mathbf{T}}_{(1:2)} = \mathbf{U}_{(1:2)}\mathbf{B}_{\mathbf{U}(1:2)}$ in the second dimension (the vertical axis). The statistics of the significance testing shows that the P-values associated with $SMI_{OP}^{2,2}(\mathbf{X}_1, \mathbf{X}_2)(= 0.52)$ and $SMI_{PR}^{2,2}(\mathbf{X}_1, \mathbf{X}_2)(= 0.41)$ are 0.06 and 0.0000, respectively. Consequently, at the $\alpha = 0.05$ significance level, $\mathbf{H}_0$ is rejected in the *PR*-context and nearly rejected in the *OP*-context for the associated *SMI*-values.

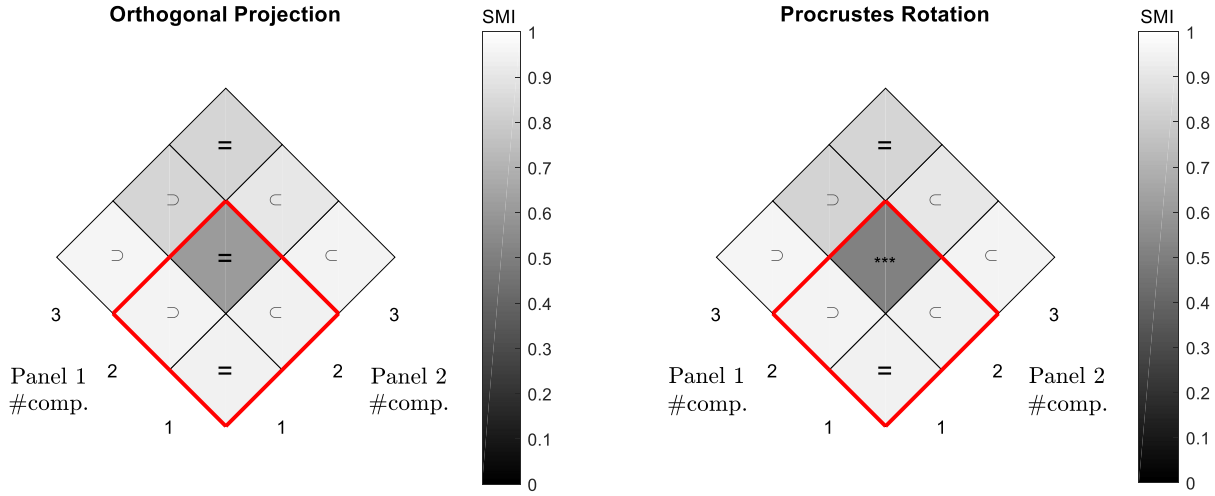

Figure 4 - SMI_{OP} and SMI_{PR} for all combinations of up to three PCA-components from Panel 1 and Panel 2. The red square (thicker line) indicates the dimensions most popular for

interpretations in sensory analyses. The "=", "⊂" and "⊃" indicate that H_0 is not rejected. Triple stars (***) indicate rejection of H_0 at the significance level P<0.001.

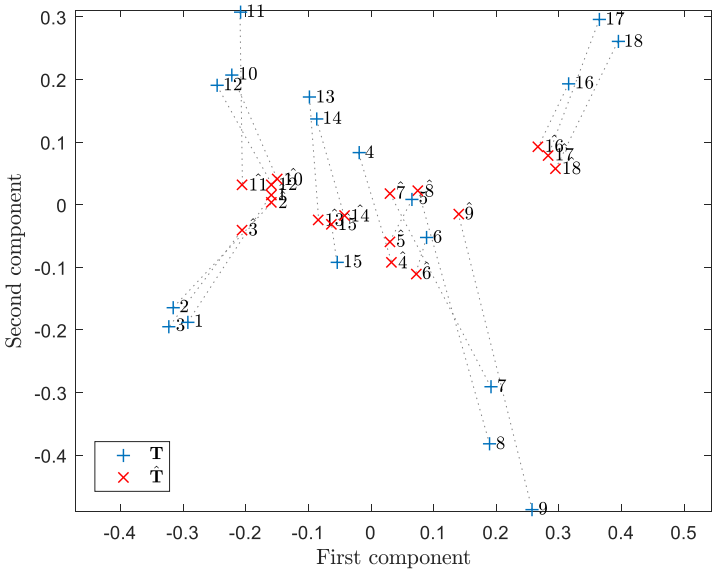

Figure 5 - Plot of the first two components of $T_{(1:2)}$ (+) and the predictions $\widehat{T}_{(1:2)}$(×). Dashed lines connect the scores and the predicted scores to indicate the level of mismatch in both components.

4.4 An example based on predictive PLS spaces from spectroscopic datasets

In this example, we consider the subspaces and associated factors obtained by two measurement technologies through predictive modelling by the use of both

- PLS-regression with several responses extracting (with respect to predictions) the orthogonal matrices **V** and **W** from $\mathbf{X}_1$ and $\mathbf{X}_2$, respectively.
- PCA extracting (with respect to variance content) the orthogonal matrices **T** and **U** from $\mathbf{X}_1$ and $\mathbf{X}_2$, respectively.

Using PLS for identifying the subspaces to compare, means that the SMI will measure what the two matrices $\mathbf{X}_1$ and $\mathbf{X}_2$ have in common with regard to the modelling of some particular response(s).

Spectroscopic measurements by Raman shifts (Gardiner and Graves (1989)) and NIR wavelengths (Stark et al. (1986)) are both highly multivariate measurement technologies that are here used for prediction of two polyunsaturated fatty acid (PUFA) **Y**-responses (standardized) from the same set of (n = 69) samples.

A leave-one-out cross-validation approach was used for choosing the appropriate dimensions and associated factors for each dataset separately. With the Raman ($\mathbf{X}_1$) data of size (69 x 1096), p = 6 PLS components (**V**) were required in order to give the best possible predictions by means of cross validation. With the NIR ($\mathbf{X}_2$) data of size (69 x 301), selection of q = 10 components (**W**) was indicated as the best cross validated choice. The associated $SMI_{OP}^{6,10}(\mathbf{V},\mathbf{W}) = 0.67$, indicate that the NIR predictive space accounts for a substantial proportion of the variation in the Raman predictive space. The other matrix correlation measures are a bit lower when using the same 6 and 10 dimensional subspaces as input: $SMI_{PR}^{6,10}(\mathbf{V},\mathbf{W}) = 0.59$, RV/GCD($\mathbf{V}_{1:6},\mathbf{W}_{1:10}$) = 0.52, RV2($\mathbf{V}_{1:6},\mathbf{W}_{1:10}$) = 0.47, $\text{RV}_{\text{adj}}(\mathbf{V}_{1:6},\mathbf{W}_{1:10})$ = 0.46, PSI($\mathbf{V}_{1:6},\mathbf{W}_{1:10}$) = 0.60 and GCD($\mathbf{V}_{1:6},\mathbf{W}_{1:10}$) = 0.52.

Figure 6 shows the diamond plots of the $SMI_{OP}^{i,j}$ by including all subspace combinations (and associated factors) up to 10 dimensions for both the NIR and Raman data based on PCA (left) and PLS (right). Note that the diamond plot similarity pattern is quite consistent for the two subspace selection alternatives (a closer inspection shows that the SMI_{OP} is slightly larger on average for the PLS subspaces, $SMI_{OP}^{6,10}(\mathbf{T},\mathbf{U}) = 0.53$). The cells marked with "⊂" represent *SMI*-values sufficiently large for not rejecting $\mathbf{H}_0$. The correspondence between data similarity (PCA) and

predictive similarity (PLS) is comparable for the other matrix correlation measures when applied to the PCA-scores: RV/GCD($\mathbf{T}_{1:6}, \mathbf{U}_{1:10}$) = 0.41, RV2($\mathbf{T}_{1:6}, \mathbf{U}_{1:10}$) = 0.35, RV_{mod}($\mathbf{T}_{1:6}, \mathbf{U}_{1:10}$) = 0.34 and PSI($\mathbf{T}_{1:6}, \mathbf{U}_{1:10}$) = 0.51.

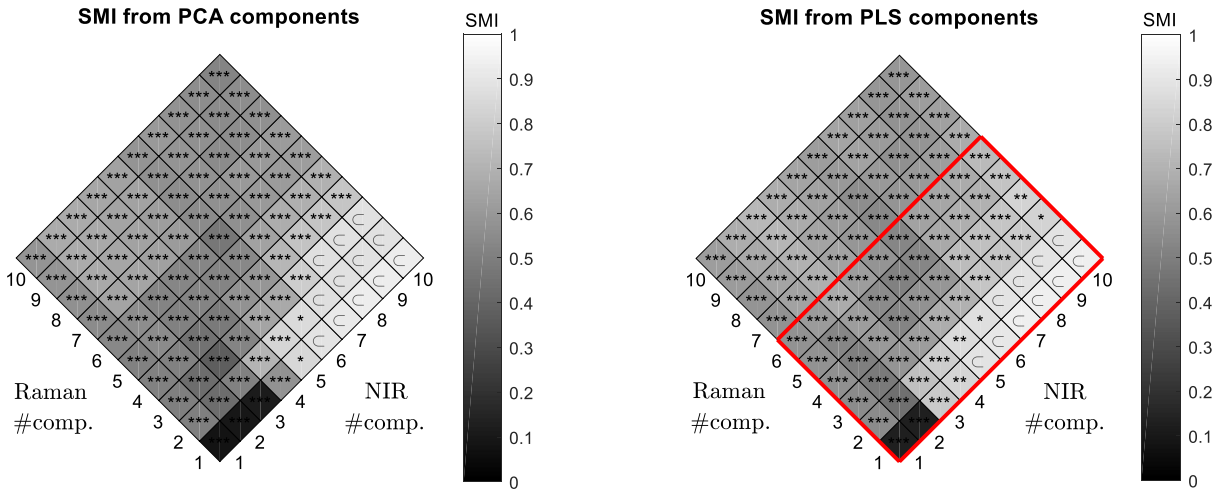

Figure 6 - The left plot shows *SMI*$_{OP}$ for all component combinations of PCA on NIR and Raman spectra. The right plot shows the SMIs components found by PLS regression. The red rectangle shows all component combinations up to the numbers found to be optimal for prediction by cross-validation for NIR and Raman, respectively. "=", "⊂" and "⊃" shows that H$_0$ is not rejected. Stars indicate rejection of H$_0$ at the significance levels: *** = P<0.001, ** = P<0.01, * = P<0.05.

4.5 Example on spectroscopic datasets to compare non-overlapping subsets of variables

This example demonstrates an application of the *SMI*-framework to assess subspace (and associated factor) similarities based on subsets of variables obtained by NIR-measurements. By vertically splitting the NIR-data matrix from Example 4 into four equally sized blocks ($\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3, \mathbf{X}_4$), we consider the variables corresponding with the two middle blocks $\mathbf{X}_2$ and $\mathbf{X}_3$

(associated with the wavelengths ranging from 1550 to 1698nm and 1700 to 1848nm, respectively). The diamond plot in Figure 7 shows the $SMI_{OP}^{i,j}(\mathbf{X}_2, \mathbf{X}_3)$-values for up to 10 PCA-component combinations. Because NIR measurements often give strong correlations along large bands of wavelengths, one can expect several of the *SMI*-values to be relatively large. It should be noted that for the subspaces (and associated factors) of equal dimension, rejections of $\mathbf{H}_0$ are obtained. In particular, we obtain $SMI_{OP}^{7,7}(\mathbf{X}_2, \mathbf{X}_3) = 0.80$ with a corresponding P-value less than 0.001. By including extra components for either one of the subspaces (and associated set of factors), however, the corresponding P-values increase and the associated null hypotheses are no longer rejected. Thus, the diamond plot shows that much of the same information is present in both matrices. The discrepancy along the diagonal in Figure 7 is most likely related to the very first pair of PCA-scores that are not very similar ($SMI_{OP}^{1,1}(\mathbf{X}_2, \mathbf{X}_3) = 0.45 \Leftrightarrow corr(\boldsymbol{t}_1, \boldsymbol{u}_1) = 0.67$). Finally, note that RV($\mathbf{X}_2$, $\mathbf{X}_3$) = 0.60, a relatively small value (compared to the entire collection of SMI_{OP}-values indicated in Figure 7) not revealing the evidently strong relationships between the two data blocks, due to the low correspondence in the first principal component. The same conclusion can be drawn from $\text{RV2}(\mathbf{X}_2, \mathbf{X}_3) = 0.59$ and $\text{RV}_{\text{adj}}(\mathbf{X}_2, \mathbf{X}_3) = 0.60$. Because of the large number of variables, $\text{GCD}(\mathbf{X}_2, \mathbf{X}_3) = 1$, while $\text{PSI}(\mathbf{X}_2, \mathbf{X}_3) = 0.95$, as it rotates to principal components of higher correspondence.

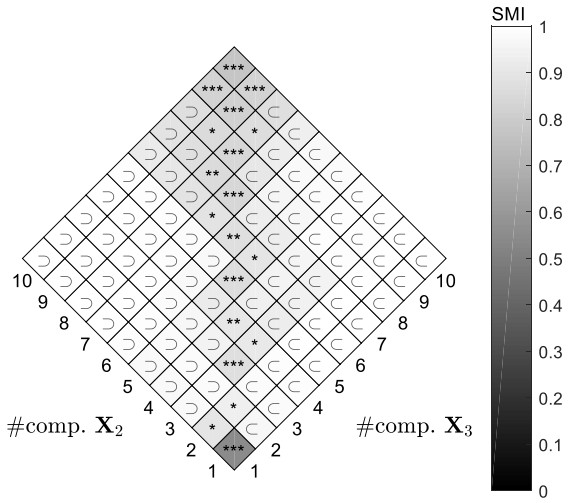

Figure 7 - SMI_{OP} for all combinations of up to 10 components from wavelength numbers 76:150 and wavelength numbers 151:225. "=", "⊂" and "⊃" shows that H_0 is not rejected. Stars indicate rejection of H_0 at the significance levels: *** = P<0.001, ** = P<0.01, * = P<0.05.

5. Discussion

The proposed two-step *SMI*-framework for comparing matrices (that may be associated with different measurement technologies) goes as follows: First, one identifies the stable subspaces accounting for the relevant directions of variability for the two data matrices. Thereafter, the associated subspaces are compared with respect to either an orthogonal projection (*OP*) or a Procrustes rotation (*PR*). Based on the singular values of $\mathbf{T}^t\mathbf{U}$, both the *OP* and the *PR* can be calculated to expose the nature and level of similarity between the matrices considered.

When interest lies in investigating whether the samples in two different matrices have more or less the same configuration, but the actual components included in the SMI-measurements have

different explained variances in the two matrices, the similarities as measured by SMI_{PR} and SMI_{OP} can be used as criterion (see Section 3.1.2).

Note that different subspace estimation procedures can be used, depending on the scope of the study. In most cases PCA is the natural alternative, but in situation where the focus is mainly on the subspaces providing particular predictive information, using PLS for the subspace identification is a more appropriate alternative.

The suggested *SMI*-framework for assessing similarities has been related to several established methods (the various RV coefficients, Yanai's GCD and the PSI), and some interesting advantages were demonstrated: By concentrating on comparing subspaces, we were able to introduce statistical testing of the hypothesis assuming *equality of a set of factors associated with the compared datasets*. In contrast, traditional applications of the RV coefficient and classical applications of canonical analysis focus on testing the hypothesis of *absence of relationship between the matrices*. In most real cases, where the investment in extra resources and new technology for collecting more data is preferred only when proven profitable, this type of statistical testing is the answer to a less interesting question.

Compared to the RV coefficient (Equation (12)), which depends on the squared singular values of the data matrices $\mathbf{X}_1$ and $\mathbf{X}_2$, the *SMI*-framework considers only the stable structural elements without taking the size of the singular values into further account. From a geometrical point of view, the weighting of dimensions in the RV coefficient represents a bias towards the dimensions associated with the larger singular values.

In Section 3.5 we illustrated some further benefits of using the SMI instead of the different RV coefficients using simulated data. When generating orthogonal matrix columns (variables)

spanning non-overlapping subspaces (Figure 1, upper part) , one could see that the values of RV2 and RV_{adj} were decreasing monotonously towards -1 when the dimensionality of the non-overlapping subspaces associated with the two matrices were increasing.

When the entries of the two matrices were randomly drawn from the standard normal distribution, their columns are typically spanning overlapping subspaces. As is shown in the lower part of Figure 1, all of the alternative RV coefficients yielded large matrix correlation values that were quickly (after including 10-100 columns) increasing towards 1. The properties exposed here should be kept in mind when drawing conclusions based on the different RV coefficients, especially when the matrices to be compared have a large number of columns compared to the number of rows. Note that the SMI_{OP} did not expose any of the counterintuitive properties observed for the various RV-alternatives. First of all, it does not increase towards the value 1 before the number of subspace dimensions become close to the number of variables. This is a necessary consequence of the fact that the *n*-dimensional columns of the two matrices both are spanning an increasingly large part of the full Euclidean space ($\mathbf{R}^n$). Furthermore, SMI_{OP} is also straight forward to use for cases where the original (raw) matrices are wide (containing more columns than rows), as the subspaces interesting for comparisons often have much lower dimensions.

6. Conclusions

Based on the example applications and the arguments given above, we claim that the *SMI*-framework (and Yanai's GCD) offers a way of measuring matrix similarity which is more in line with a common sense understanding of matrix similarity than the various RV coefficients. The *SMI*-framework equipped with the *diamond plot* that visualizes both the *SMI*-values and corresponding significance, is also better facilitated for doing ordinary statistical inference. The entire *SMI*-approach has been introduced as an explorative framework for investigation of

various similarities of subspace combinations. We believe the diamond plots may be recognized as particularly useful for quickly recognizing the "broader" geometrical picture of relationships present in the datasets subject to comparison.

Acknowledgments: This work was financially supported by the Research Council of Norway (pro. nr. 239070) and the Norwegian Levy on Agricultural Products.

References:

Amerine, M. A., Pangborn, R. M. and Roessler, E. B. (1965), Principles of sensory evaluation of food, New York: Academic press.

Björck, A. and Golub, G. H. (1973), "Numerical methods for computing angles between linear subspaces," *Mathematics of Computation*, 27, 123, 579–594.

Björck, A. and Indahl, U. G. (2017), "Fast and stable Partial Least Squares modelling: A benchmark study with theoretical comments," *Journal of Chemometrics*, DOI: 10.1002/cem.289.

Cramer, E. and Nicewander, A. (1979), "Some symmetric, invariant measures of multivariate association," *Psychometrika,* 44, 43-54

Darlington, R. B., Weinberg, S. L. and Walberg, H. J. (1973), "Canonical Variate Analysis and Related Techniques," *Review of Educational Research*, 43, 4, 433-454.

Draper, N. R. and Smith, H. (1998), Applied Regression Analysis, New York: Wiley-Interscience.

Efron, B. and Tibshirani, R. (1993), An Introduction to the Bootstrap, Chapman & Hall/CRC.

El Ghaziri, A. and Qannari E. M (2015), "Measures of association between two datasets; Application to sensory data," *Food Quality and Preference*, 40, 116-124.

Gardiner, D. J. and Graves, P. R. (1989), Practical Raman spectroscopy, Berlin: Springer-Verlag.

Gower, J. C. (1975), "Generalised Procrustes analysis," *Psychometrica.* 40, 1, 33-51.

Hastie, T., Tibshirani, R. and Friedman, J. H. (2009), The Elements of Statistical Learning: Prediction, Inference and Data Mining. Second Edition, New York: Springer-Verlag.

Jolliffe, I. (2002), Principal component analysis, New York: Springer-Verlag.

Kendall, D. G. (1984), "Shape Manifolds, Procrustean Metrics, and Complex Projective Spaces," *Bull. London Math. Soc.* 16, 2, 81-121.

Mayer, C.D., Lorent, J. and Horgan G.W. (2011), "Exploratory Analysis of Multiple Omics Datasets Using the Adjusted RV Coefficient," *Statistical Applications in Genetics and Molecular Biology*, 10, 1, Article 14.

Ramsay, J. O., ten Berge, J. and Styan, G. P. H (1984), "Matrix correlation," *Psychometrika*, 49, 3, 403-423.

Robert, P. and Escoufier, Y. (1976), "A Unifying Tool for Linear Multivariate Statistical Methods: The *RV*-Coefficient," *Applied Statistics*, 25, 3, 257–265.

Sibson, R. (1978). "Studies in the robustness of multidimensional scaling: Procrustes Statistics," *Journal of Royal Statistical Society*, Series B, 40, 234–238.

Smilde, A. K., Kiers, H. A. L., Bijlsma, S., Rubingh, C. M. and Erk, M. J. (2009), "Matrix correlations for high-dimensional data: the modified RV coefficient," *Bioinformatics*. 25, 401-405.

Stark, E., Luchter, K. and Margoshes, M. (1986), "Near-Infrared Analysis (NIRA): A Technology for Quantitative and Qualitative Analysis," *Applied Spectroscopy Reviews*, 22, 4, 335-399.

Tomic, O., Luciano, G., Nilsen, A., Hyldig, G., Lorensen, K., Næs, T. (2010), "Analysing sensory panel performance in a proficiency test using the PanelCheck software," *European Food Research and Technology*. 230**. 3,** 497-511.

Tomic, O. Forde, C., Delahunty, C. and Næs, T. (2013), "Performance Indices in Descriptive Sensory Analysis - a complimentary screening tool for assessor and panel performance," *Food Quality and preference,* 28, 122-133.

Wold, S. Ruhe, A. Wold, H. and Dunn, W. J. (1984), "The collinearity problem in linear regression. The partial least squares (PLS) approach to generalized inverses,"
SIAM J. Sci. Statist. Comput., 5, 735–743.

Wrigley C and Neuhaus J. E. (1955), "The matching of two sets of factors," *American Psychologist,* 10, 418-419.

Yanai, H. (1974), "Unification of various techniques of multivariate analysis by means of generalized coefficient of determination (GCD)," *Kodo Keiryogaku (The Japanese Journal of Behaviormetrics)* 1, 46-54.

Appendix

Table 1: Coordinates for the two sets of points in the first example in Section 4.1.

$\mathbf{X}_1$		$\mathbf{X}_2$	
-1,0	1,0	-5,5	-0,5
1,0	1,0	-5,0	0,3
5,0	0,5	-0,8	0,7
5,0	-0,5	0,8	0,8
1,0	-1,0	5,5	0,5
-1,0	-1,0	5,8	-0,5
-5,0	-0,5	0,9	-0,6
-5,0	0,5	-0,7	-0,9