MatrixCorrelation

Kristian Hovde Liland[1,2], Tormod Næs[1] & Ulf Geir Indahl[3]

[1] *Nofima – Norwegian Institute of Food, Fisheries and Aquaculture Research, Ås, Norway*
[2] *Norwegian University of Life Sciences, Department of Chemistry, Biotechnology and Food Science, Ås, Norway*
[3] *Norwegian University of Life Sciences, Department of Mathematical Sciences and Technology, Ås, Norway*
E-mail : *kristian.liland@nofima.no*
E-mail : *tormod.naes@nofima.no*
E-mail : *ulf.indahl@nmbu.no*

Abstract
MatrixCorrelation is an R package for comparing data sets. It contains many methods, but emphasizes the use of the Similarity of Matrices Index (SMI). This is a new method that compares stable subspaces from coupled data matrices. Visualization for explorative data analysis and statistical testing of equality are included.
Keywords: R, Similarity of Matrices Index,

1. Introduction

An R package for easy calculation and visualisation of matrix correlations has been created. This can be used to assess the similarity between two data sets having the same objects/samples either measured/recorded under different conditions or using different techniques. The package includes most of the traditional measures used in sensometrics, chemometrics and related areas of science in addition to some of the newer measures proposed lately. Among the former are Ramsey's r1, r2, r3, and r4, Yannai's GCD, and Escoufier's RV. Among the latter, we have included Smilde's RV2, Maye's RVadj, and Indahl's SMI.

2. Similarity of Matrices Index

The similarity of matrices index (SMI) is the recommended measure as it gives a more detailed impression of the similarities of two matrices. An accompanying plot called a diamond plot gives insight into the underlying structures of the matrices and indicates significant differences. SMI can be used as a replacement for the RV coefficient or as an exploratory tool for revealing structural differences. Unlike RV, it is not dominated by the primary direction of variance in the data and can safely be interpreted for data of two or more dimensions. Significance estimations are done with the assumption that the data sets have a common underlying structure/subspace, meaning that tests are used to reveal significant differences between data sets like panels of assessors, spectroscopic instruments or any other techniques resulting in matrices with one row per object/sample.

2.1 Variants

SMI is a framework that uses a compression method, e.g. principal component analysis (PCA), and a regression method, e.g. orthogonal projection (OP, ordinary least squares). Using PCA and OP it is easy to perform permutation testing to assess the significances of the observed similarities with a null hypothesis that two data sets are realizations of the same underlying structure, see Figure 1.

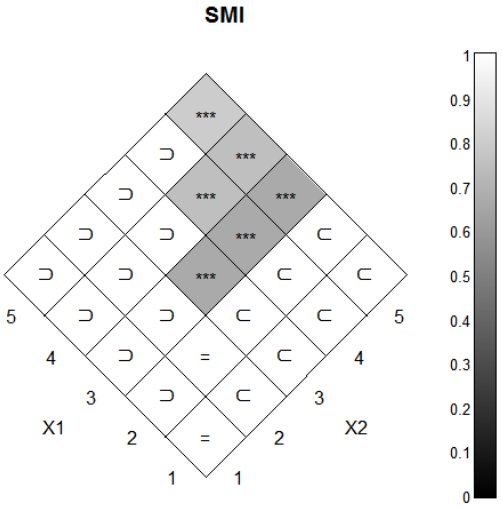

Figure 1: SMI_{OP} for up to 10 component combinations after centering of the matrices X1 (100x300) with random entries from the standard normal distribution compared to X2 (obtained by removing the 3rd SVD component of X1). "=", "⊂" and "⊃" shows that H_0 is not rejected. Stars indicate rejection of H_0 at significance levels: *** = P<0.001, ** = P<0.01, * = P<0.05.

To give different emphasis, one can exchange PCA with another method that generates an orthonormal basis, e.g. partial least squares (PLS). This would steer the subspace towards explanation of a response. The regression method can be exchanged with Procrustes Rotations (PR) for a more rigid measure that compares only through scaling and rotation.

References

Ramsey, 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.

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.

Indahl, U.G., Næs, T. & Liland, K.H. (Submitted) A similarity index for comparing coupled matrices.