Available online at www.sciencedirect.com
SCIENCE DIRECT®
ELSEVIER
Computational Statistics & Data Analysis 48 (2005) 47–62
COMPUTATIONAL STATISTICS & DATA ANALYSIS
www.elsevier.com/locate/csda

Combined Procrustes analysis and PLSR for internal and external mapping of data from multiple sources

Garmt Dijksterhuis*, Harald Martens, Magni Martens

Sensory Science group, Department of Dairy and Food Science, KVL, Royal Veterinary and Agricultural University, Copenhagen, Denmark

Received 13 September 2003; accepted 19 September 2003

Abstract

Generalised Procrustes analysis (GPA) is a method for producing a group average from rotated versions of a set of individual data matrices followed by bi-linear approximation of this group average for graphical inspection. Partial Least Squares Regression (PLSR) is a method for relating one data matrix to another data matrix, via bi-linear low-rank regression modelling. The merger of these methods proposed aims to produce an average (e.g. a sensory group panel average), which balances an "intersubjective", internal consensus between the individual assessors' data against an "objective" external correspondence between the sensory data and other types of data on the same samples (e.g. design information, chemical or physical measurements or consumer data). Several ways of merging GPA with PLSR are possible, of which one is selected and applied. The proposed "GP–PLSR" method is compared to a conventional GPA followed by an independent PLSR, using a data set about milk samples assessed by a group of sensory judges with respect to a set of sensory descriptor terms, and also characterised by experimental design information about the samples. The GP–PLSR gave a more design-relevant group average than traditional GPA. The proposed algorithm was tested under artificially increased noise levels.

Keywords: Generalised Procrustes analysis; Partial Least-Squares regression; GPA; PLS; PLSR Double criterion; Internal and external mapping; K-sets analysis

* Corresponding author. Department of Marketing and Marketing Science, Department of Economics, University of Groningen and Agrotechnology and Food Innovations BV, Wageningen University and Research Centre, The Netherlands. Tel.: +31-317-47-5268; fax: +31-317-47-5347.

E-mail address: garmt@tref.nl (G. Dijksterhuis).

doi:10.1016/j.csda.2003.09.013

1. Introduction

'Generalised Procrustes-Partial Least Squares Regression' GP–PLSR, as we coined it, is a combination of Generalised Procrustes matching of data matrices (GPA; Gower, 1975; Dijksterhuis and Gower, 1991/2) and Partial Least Squares Regression (PLSR; Wold et al., 1983). The combination is such that the Procrustes rotations inherent in the GPA are guided by an external data set $\mathbf{Y}$ by means of incorporating a PLSR step in the iterative GPA process. We will denote the, K, individual data sets by $\mathbf{X}_k$, $k = 1, \ldots, K$, and the GPA group average by $\bar{\mathbf{X}}$. The goal of our method is to obtain a group average which is representative of two things:

1. The agreement among the K data sets $\mathbf{X}_k$.
2. The structure in $\mathbf{Y}$ as far as it accords with $\bar{\mathbf{X}}$.

The above mentioned 1 can be referred to as a 'democratic process' leading to a group average $\bar{\mathbf{X}}$ which is fairly representing all individual $\mathbf{X}_k$. The second point refers to side-conditions that also must be taken into account, defined by the external data in $\mathbf{Y}$. It can so happen that an individual data set $\mathbf{X}_j$, $j \in \{1, \ldots, K\}$ is off, with respect to the other sets $k \neq j$ in $\mathbf{X}_k$ but in accordance with the external information in $\mathbf{Y}$. In that case this $\mathbf{X}_j$ should get the opportunity to make itself heard. This will be accomplished through the PLSR relating $\bar{\mathbf{X}}$ to $\mathbf{Y}$. Analogously, when the majority of sets $\mathbf{X}_k$, $k = 1, \ldots, K$, are in strong agreement, they will together mainly shape $\bar{\mathbf{X}}$, and the effect of the PLSR relating $\bar{\mathbf{X}}$ to $\mathbf{Y}$ should be outweighed by them.

This is only a general description of the ideas behind the method. Below we will discuss some alternative implementations and some problems connected to the reduction of the rank of $\bar{\mathbf{X}}$ through the PLSR.

2. GPA and PLSR methods

2.1. The GPA modelling

The GPA modelling has two stages: The first iterative step defines a common group average $\bar{\mathbf{X}}$ from all the K individual input data matrices $\mathbf{X}_k$, $k = 1, 2, \ldots, K$. As part of this process, the algorithm rotates each of these input data matrices $\mathbf{X}_k$ towards the common group average $\bar{\mathbf{X}}$ by its estimated rotation matrix $\mathbf{H}_k$. The second GPA step extracts the principal components from the group average $\bar{\mathbf{X}}$ and visualises this for human interpretation.

The GPA algorithm is presented in Fig. 1. It is based on the algorithm given in Kristof and Wingersky (1971) and in Gower (1975).

Inside the first step's outer iteration loop there is an inner loop over the K sets. The heart of the first step in the algorithm is the two-set Procrustes rotation (cf. Schönemann, 1966), situated in this inner loop. Each individual set $\mathbf{X}_k$ is rotated to a current version of the group average $\bar{\mathbf{X}}$, resulting in K transformation matrices $\mathbf{H}_k$. These matrices are orthonormal rotation matrices, i.e. $\mathbf{H}_k'\mathbf{H}_k = \mathbf{H}_k\mathbf{H}_k' = \mathbf{I}$, and are