Available online at www.sciencedirect.com
SCIENCE DIRECT®

Computational Statistics & Data Analysis 50 (2006) 2529 – 2535

COMPUTATIONAL STATISTICS & DATA ANALYSIS

www.elsevier.com/locate/csda

A simultaneous CONCOR algorithm for the analysis of two partitioned matrices

Roger Lafosse[a], Jos M.F. ten Berge[b,*]

[a] *Université Paul Sabatier, 118 Rte de Narbonne, 31062 Toulouse Cedex 4, France*
[b] *University of Groningen, Netherlands*

Received 3 January 2005; received in revised form 8 March 2005; accepted 1 April 2005
Available online 27 April 2005

Abstract

A standard approach to derive underlying components from two or more data matrices, holding data from the same individuals or objects, is the (generalized) canonical correlation analysis. This technique finds components (canonical variates) with maximal sums of correlations between them. The components do not necessarily explain much variance in the matrices they were derived from. This observation has given rise to alternative techniques, which maximize the sum of covariances between the components, subject to orthonormality constraints on the weight matrices applied to generate the components from the data matrices. However, a method called ConcorGM, maximizing the sum of squared covariances, has also been proposed. It has the additional feature that it allows for the analysis of two sets of data matrices which may, but need not coincide. The ConcorGM algorithm maximizes the sum of squared covariances successively, by first finding the best single-component solution, and repeating that process in the respective residual spaces. An algorithm for maximizing the sum of squared covariances simultaneously is offered.

Keywords: Canonical correlation; Generalized Maxbet analysis; Concor; Partitioned data matrix analysis

In standard (generalized) canonical correlation problems, we have two (or more) data matrices $\mathbf{X}_1, \ldots, \mathbf{X}_m$, centered column wise, each with p rows pertaining to the same individuals or objects, and we seek linear combinations (canonical variates) for each matrix

* Corresponding author. Tel.: +31 50 3636349; fax: +31 50 3636304.
E-mail addresses: Roger.Lafosse@lsp.ups-tlse.fr (R. Lafosse), j.m.f.ten.berge@rug.nl (J.M.F. ten Berge).

0167-9473/$ - see front matter
doi:10.1016/j.csda.2005.04.002

such that these canonical variates have the highest possible sum of correlations between them. It is well known that the canonical variates capitalize on shared variance between the matrices, possibly explaining very little variance within the matrices they were derived from. For this reason, alternatives have been proposed which maximize the sum of covariances rather than the sum of correlations between the canonical variates (the variances of canonical variates included), with orthonormality constraints on the matrices of weights that define the canonical variates. For instance, Van de Geer (1984) has proposed Maxbet, which maximizes

$$f\left(\mathbf{U}_1, \ldots, \mathbf{U}_k\right) = \Sigma_{i,j} \operatorname{tr}\left(\mathbf{U}_i'\mathbf{X}_i'\mathbf{X}_j\mathbf{U}_j\right), \tag{1}$$

subject to $\mathbf{U}_i'\mathbf{U}_i = \mathbf{I}_r$, for a fixed number r of canonical variates. A computational method for Maxbet can be found in Ten Berge (1988), also see Hanafi and Ten Berge (2003). This method can be further generalized by applying it to two different sets of matrices, centered column wise, which we shall refer to as $\mathbf{X}_1, \ldots, \mathbf{X}_m$ and $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$, measured on the same p individuals or objects, to optimize

$$f\left(\mathbf{U}_1, \ldots, \mathbf{U}_m, \mathbf{V}_1, \ldots, \mathbf{V}_n\right) = \Sigma_{j,i} \operatorname{tr}\left(\mathbf{V}_j'\mathbf{Y}_j'\mathbf{X}_i\mathbf{U}_i\right), \tag{2}$$

subject to $\mathbf{U}_i'\mathbf{U}_i = \mathbf{V}_j'\mathbf{V}_j = \mathbf{I}_r$. Maximizing (2) offers a method of analyzing relationships between two partitioned matrices $\mathbf{X} = [\mathbf{X}_1|\ldots|\mathbf{X}_m]$ and $\mathbf{Y} = [\mathbf{Y}_1|\ldots|\mathbf{Y}_n]$. Kissita et al. (2004) have proposed a modification of this generalized Maxbet function, called ConcorGM, by replacing the plain sum of covariances by the corresponding sum of squared covariances. Specifically, they proposed maximizing

$$f\left(\mathbf{U}_1, \ldots, \mathbf{U}_m, \mathbf{V}_1, \ldots, \mathbf{V}_n\right) = \Sigma_{j,i} \left\|\operatorname{diag}\left(\mathbf{V}_j'\mathbf{Y}_j'\mathbf{X}_i\mathbf{U}_i\right)\right\|^2, \tag{3}$$

subject to the same constraints as (2). This attributes higher importance to large covariances, at the expense of the smaller ones. The special case $n = 1$ of (3) is the function maximized by the so-called Concor method of Lafosse and Hanafi (1997), also see Chessel and Hanafi (1996).

The ConcorGM method is aimed at finding components from each of the data matrices such that each set of $n + m$ components is as concordant as possible, in the sense that the sum of squares of the $n \times m$ between-covariances is a maximum. Potential applications of ConcorGM can be found in various disciplines. For instance, in the measurement of personality, we may have a set of matrices $\mathbf{X}_1, \ldots, \mathbf{X}_m$, holding self-report measures of individuals on various personality questionnaire items, each matrix pertaining to one of m particular situations (situations of conflict, of prosperity, and so on). In addition, the same individuals may have been assessed by a parent and by a friend, giving rise to matrices $\mathbf{Y}_1$ and $\mathbf{Y}_2$, respectively. ConcorGM finds components for each situation and each assessor, such that links between assessors and situations emerge. For instance, it may appear that a component of "agreeableness", when assessed by a friend, is much stronger related to self-reported agreeableness for situations of conflict than the parent's assessment.

It is important to note that the algorithm developed by Kissita et al. (2004) called ConcorGM, maximizes the function (3) *successively*. That is, after finding an optimal set of first columns of $\mathbf{U}_1, \ldots, \mathbf{U}_m$, $\mathbf{V}_1, \ldots, \mathbf{V}_n$, these columns are fixed and the second columns are found in the respective residual spaces, orthogonal to the first columns. This approach has