Available online at www.sciencedirect.com
SCIENCE DIRECT®
COMPUTATIONAL STATISTICS & DATA ANALYSIS
ELSEVIER
Computational Statistics & Data Analysis 48 (2005) 5–16
www.elsevier.com/locate/csda

Dimensionality reduction approach to multivariate prediction

Bovas Abraham[a], Giovanni Merola[b,*]

[a] *University of Waterloo, Canada*
[b] *Office for National Statistics, Methodology Group, D2/13, 1 Drummond Gate, London SW1V 2QQ, UK*

Accepted 8 November 2003

Abstract

Dimensionality reduction methods used for prediction can be cast into a general framework by deriving them from a common objective function. Such a function yields continuum of different solutions, including all the known ones. Least-squares and maximum likelihood estimation of the model at the base of dimensionality reduction methods for prediction lead to an additive objective function. By letting this additive function be any convex linear combination of the two addends, another objective function from which a continuum of solutions can be obtained.

Keywords: Dimensionality reduction methods; Prediction; PLS; Reduced rank regression; Principal component regression; Maximum overall redundancy; Multivariate continuum regression

1. Introduction

Dimensionality reduction methods (DRMs) determine a set of orthogonal linear combinations of observed variables, called latent variables (lv's). The use of DRMs in prediction consists of substituting the set of observed explanatory variables in a regression model with fewer lv's. The responses are then predicted through the usual ordinary least-squares (OLS) method. The use of DRMs for prediction is considered heuristic because of the lack of a clear model behind the data and of the lack of optimality of the solutions. In fact, DRMs for prediction seem to succeed in situations were

* Corresponding author. Tel./fax: +44-20-77338253.
E-mail addresses: merolagio@tiscali.co.uk, giovanni.merola@ons.gov.uk (G. Merola).

0167-9473/$ - see front matter
doi:10.1016/j.csda.2003.11.021

the OLS estimates fail to give good predictions. Most of the published applications are in fields in which a large number of explanatory variables are available but the exact nature of the relationship between responses and explanatory variables is not known exactly. That is, fields such as chemometrics (e.g. Gelaldi and Kowalski, 1986), biochemistry (e.g. Schmidli, 1995), sensory analysis (e.g. Tenenhous, 1998) and statistical process control (e.g. Kourti and MacGregor, 1996). In this paper we only consider multivariate prediction, however some DRMs can be also applied for the univariate case.

Different DRMs have been proposed for different purposes; each method obtains the lv's optimizing a different objective function. Because of the lack of a criterion for comparing these methods, it becomes important to relate different DRMs through a common objective function and to have the possibility of deriving alternative intermediate solutions. Burnham et al. (1996, 1999, 2001) have discussed a general framework for multivariate l.v. regression methods.

In the next section we briefly review the most common DRMs and then propose the objective function of multivariate generalization of continuum regression, from which different DRMs can be obtained; lastly in Section 2 we propose another generalized objective function obtained by OLS and maximum likelihood estimation. In Section 3 we present some examples and in the last section we give some concluding remarks.

2. Objective functions of the DRMs used for prediction

Let $\mathbf{X}$ be an $(n \times p)$ matrix containing n rows of independent observations on p explanatory variables and $\mathbf{Y}$ an $(n \times q)$ matrix containing n rows of corresponding observations on q response variables. In what follows it is assumed that the columns of the data-matrices are *autoscaled*, that is centered to zero mean and scaled to unit variance. The lv's $\mathbf{t}_j = \mathbf{X}\mathbf{a}_j, j = 1, \dots, p$ are an ordered sequence of orthogonal linear combinations defined by the p-vectors $\mathbf{a}_i$. We denote matrices with bold upper-case letters and their columns with the corresponding bold lower-case letter. Thus we write $\mathbf{T}_{(d)} = \mathbf{X}\mathbf{A}_{(d)}$ to denote the $(n \times d)$ orthogonal matrix containing d lv's. The use of DRMs for prediction consists of regressing the responses on the first d, $1 \leqslant d \leqslant p$, lv's. Therefore, the fitted response matrix is given by

$$\hat{\mathbf{Y}}_{[d]} = \mathbf{T}_{(d)}(\mathbf{T}'_{(d)}\mathbf{T}_{(d)})^{-1}\mathbf{T}'_{(d)}\mathbf{Y} = \mathbf{X}\mathbf{B}_{[d]}, \tag{1}$$

where the subscript $[d]$ denotes that d l.v.'s were employed and the matrix $\mathbf{B}_{[d]} = \mathbf{A}_{(d)}(\mathbf{T}'_{(d)}\mathbf{T}_{(d)})^{-1}\mathbf{T}'_{(d)}\mathbf{Y}$ is the matrix of regression coefficients obtained with d lv's. When all p lv's are employed, $\mathbf{B}_{[p]}$ are the OLS solutions. To estimate the regression coefficients it is sufficient to estimate $\mathbf{A}_{(d)}$. In all the methods that we consider the solutions with d lv's do not change if further components are added to the model. We now briefly introduce different DRMs, a more thorough discussion can be found, for instance, in Merola and Abraham (2003).