Computational Statistics and Data Analysis 53 (2009) 3855–3863

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Principal component regression for data containing outliers and missing elements

Sven Serneels[a], Tim Verdonck[b,*]

[a] LS Services and Consultancy, Edegem, Belgium

[b] Agoras Group, Department of Mathematics and Computer Science, University of Antwerp, Belgium

ARTICLE INFO

Article history:
Received 25 January 2008
Received in revised form 9 April 2009
Accepted 16 April 2009
Available online 23 April 2009

ABSTRACT

A methodology is presented to construct an expectation robust algorithm for principal component regression. The presented method is the first multivariate regression method which can resist outliers and which can cope with missing elements in the data simultaneously. Simulations and an example illustrate the good statistical properties of the method.

1. Introduction

In many fields of science, for instance bio-informatics, chemometrics and environmetrics, a frequently occurring problem is to estimate a regression relation between a predictand matrix $\boldsymbol{Y} \in \mathfrak{R}^{n\times q}$ (or $\boldsymbol{y}$ if the predictand is univariate; most applications of PCR concern univariate regression) and a predictor data matrix $\boldsymbol{X} \in \mathfrak{R}^{n\times p}$. At the normal distribution, the optimal solution to this problem is to use the least squares estimator. However, in these fields of science, the predictor data matrix $\boldsymbol{X}$ contains correlated columns, or its number of variables exceeds (sometimes vastly) the number of cases, such that the least squares regression estimator cannot be computed. There are plenty of ways to tackle this problem so as to obtain reliable regression estimates for such data have been proposed in the literature. One of these is principal component regression (Jolliffe, 1986). Principal component regression (PCR) is based on the following idea: instead of using the original high dimensional regressor block, use the principal component scores ($\boldsymbol{T}$) of this block as regressors. Because the principal components ($\boldsymbol{P}$) are by definition uncorrelated, the problem of multicollinearity in $\boldsymbol{X}$ is countered. Moreover, the principal components $\boldsymbol{p}_i$ are defined according to the following maximisation criterion:

$$\boldsymbol{p}_i = \underset{\boldsymbol{a}\in\mathfrak{R}^p}{\operatorname{argmax}} \{\operatorname{var}(\boldsymbol{Xa})\} \quad i = 1, \ldots, p \tag{1a}$$

under the constraints that

$$\|\boldsymbol{p}_i\| = 1 \quad \text{and} \quad \operatorname{cov}(\boldsymbol{Xp}_i, \boldsymbol{Xp}_j) = 0 \quad \text{for } j < i. \tag{1b}$$

The principal components are directions in space along which the maximum of variance is observed in that subspace of the p-dimensional space not yet spanned by the previous principal components. If ordered according to a decrease in captured variance, it is viable to assume that the first few principal components contain the majority of the information from $\boldsymbol{X}$. Thus, a low number (h) of principal components can suffice in order to capture the most information from $\boldsymbol{X}$ and thus to build a satisfactory regression model. In practice, virtually always $h \ll \min(n, p)$. Hence, using principal components as regressors, one also counters the problem of high dimensionality of the data.

[*] Corresponding address: Agoras Group, Department of Mathematics and Computer Science, University of Antwerp, Middelheimcampus, M.G.320b, Middelheimlaan 1, 2020 Antwerp, Belgium. Tel.: +32 32653896.
E-mail addresses: sven.serneels@telenet.be (S. Serneels), tim.verdonck@ua.ac.be (T. Verdonck).

0167-9473/$ – see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2009.04.008

Principal component regression is a powerful regression tool, popular in several fields of application, like chemometrics. However, the applicability of PCR is based on the assumption of normality. The principal components are defined according to the maximisation of classical variance (the squared classical standard deviation) (1a), which is an optimal estimator of scale at the normal distribution. The same remark holds for the least squares regression estimator used after estimation of the principal components. For data deviating from normality, such as a normal contaminated with outliers, principal component regression may become unreliable (depending on the type of non-normality). Especially in the case when outliers are present in the data, principal component regression yields erroneous results. For these reasons, several robust alternatives for PCR have been proposed (Walczak and Massart, 1995; Hubert and Verboven, 2003; Zhang et al., 2003; Serneels et al., 2005). The pros and cons of these methods will be discussed in Section 3.

Another problem arising in practice is that data are sometimes incomplete. Data can be missing due to different reasons. In what follows we will assume that the reason why a data point is missing is not related to its actual value, i.e. the data are missing at random (MAR) in the sense of Rubin (1976). Principal component regression, nor its robust counterparts, can deal with incomplete data sets. For classical principal components regression, a straightforward way to extend it to incomplete data would be to use a PCA estimator which can deal with missing data and then perform regression on the estimated principal components. There are several ways in which missing values can be imputed; up till now the only one that has been extended to PCA for high dimensional data, is up to our knowledge, the expectation-maximisation (EM) approach. An algorithm for PCA which can use incomplete data by virtue of EM, has been proposed by Walczak and Massart (2001).

Because the method proposed by Walczak and Massart (2001) is limited to classical PCA, which implies that it is non-robust, it may yield unreliable results if, apart from the missing elements, the data contain outliers as well. Roughly speaking, this problem can be quite easily circumvented by using either a robust PCA algorithm or a robust estimator for the variance–covariance matrix within the EM algorithm. This combination is referred to as expectation robust PCA (ER-PCA). A preliminary version ER-PCA, based on the spherical PCA estimator (Locantore et al., 1998), has first been proposed by Stanimirova et al. (2007). A more elaborate study introducing and comparing the two different approaches (using a robust PCA plug-in or based on an ER covariance matrix) as well as comparing different plug-in estimators into each of both approaches, has recently shown that by using ER-PCA, the beneficial properties of robust PCA for data containing outliers carry through to incomplete data (Serneels and Verdonck, 2008).

In what follows, the methodology of ER-PCA will be extended to principal component regression. Again, several robust estimators can be plugged in, each having different properties. In a simulation study, the properties of the different resulting ER-PCR methods will be evaluated. In an example, the applicability of the method will be illustrated.

2. The ER-PCR algorithm

As stated in the Introduction, the ER algorithm is a robustification of the EM algorithm, by means of insertion of a robust plug-in estimator (Little and Smith, 1987; Cheng and Victoria-Feser, 2002). Both the EM and ER algorithms are thus very similar; the main idea behind them is the following. Each iteration of the EM/ER algorithms consists of two steps: at first, the missing elements are filled in according to what they are expected to equal, based on the model from the previous iteration. This is called the Expectation or E-step. Secondly, the model is estimated from the data in which the missing elements have been filled in. This second step is referred to as the maximisation step or M-step for classical estimators and as the robust step or R-step for robust estimators. The iterative procedure stops upon convergence on the estimated missing elements.

For PCR both $\boldsymbol{X}$ and the dependent variables $\boldsymbol{Y}$, may contain missing elements. PCR fits into the latent variable regression model (Burnham et al., 1999), i.e.

$$\begin{cases} \boldsymbol{X} = \boldsymbol{T}\boldsymbol{P}^{\mathrm{T}} + \boldsymbol{E} \\ \boldsymbol{Y} = \boldsymbol{T}\boldsymbol{Q}^{\mathrm{T}} + \boldsymbol{F} \end{cases} \tag{2}$$

where the rows of $\boldsymbol{E}$ and $\boldsymbol{F}$ are assumed to be random errors. In PCR, the entities $\boldsymbol{T}$ and $\boldsymbol{P}$ are estimated as the score and loading matrices resulting from PCA, whereas an estimate $\hat{\boldsymbol{Q}}$ is obtained using multivariate regression. For robust PCR, analogous robust estimates are found by using robust PCA and robust multivariate regression instead. In practice, not all possible $\boldsymbol{p}_i$ in $\boldsymbol{P}$ are used, but only the first h $\boldsymbol{p}_i$ corresponding to an optimum in the bias-variance trade-off. The matrices which only contain the first h scores, loadings, etc., will notationally be distinguished from their complete counterparts by the use of an index h.

The optimal number of components is often selected as that h for which the Root Mean Squared Error (RMSE) or its cross-validated (RMSECV) version is minimal. A robust version of this measure, denoted R-RMSE, is proposed by Hubert and Verboven (2003). Furthermore Engelen and Hubert (2005) also proposed a robust value for the cross-validated version (R-RMSECV) and introduced the robust component selection (RCS) by combining the R-RMSE and R-RMSECV.If there is no prior information about the optimal number of components, we suggest to use all variables to fill in the data and apply the robust criteria to this complete data set in order to select the optimal complexity h.

Notably, estimators for the latent variables regression model (2) also yield an estimate for the vector of regression coefficients of the linear model

$$\boldsymbol{Y} = \mathfrak{B}_0 + \boldsymbol{X}\mathfrak{B} + \boldsymbol{R}. \tag{3}$$