Chemometrics and Intelligent Laboratory Systems 162 (2017) 160–171

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

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

Selection of non-zero loadings in sparse principal component analysis

Shriram Gajjar[a], Murat Kulahci[b,c], Ahmet Palazoglu[a,*]

[a] Department of Chemical Engineering, University of California, Davis, CA 95616, USA
[b] Department of Applied Mathematics and Computer Science, Technical University of Denmark, Lyngby, Denmark
[c] Department of Business Administration, Technology and Social Sciences, Luleå University of Technology, Luleå, Sweden

ARTICLE INFO

Keywords:
Sparse Principal Component Analysis (SPCA)
Principal Component Analysis (PCA)
Genetic algorithm
Pitprops data
Tennessee Eastman process

ABSTRACT

Principal component analysis (PCA) is a widely accepted procedure for summarizing data through dimensional reduction. In PCA, the selection of the appropriate number of components and the interpretation of those components have been the key challenging features. Sparse principal component analysis (SPCA) is a relatively recent technique proposed for producing principal components with sparse loadings via the variance-sparsity trade-off. Although several techniques for deriving sparse loadings have been offered, no detailed guidelines for choosing the penalty parameters to obtain a desired level of sparsity are provided. In this paper, we propose the use of a genetic algorithm (GA) to select the number of non-zero loadings (NNZL) in each principal component while using SPCA. The proposed approach considerably improves the interpretability of principal components and addresses the difficulty in the selection of NNZL in SPCA. Furthermore, we compare the performance of PCA and SPCA in uncovering the underlying latent structure of the data. The key features of the methodology are assessed through a synthetic example, pitprops data and a comparative study of the benchmark Tennessee Eastman process.

1. Introduction

With the recent advances in communication technologies and the emergence of smart factories, large volumes of data are routinely collected and stored at high sampling rates. Such a dramatic increase in data volume and frequency requires developing new statistical methods and visualizations to gain insights from multivariate data to generate actionable information for optimizing and troubleshooting process operations. Historically, multivariate statistical analysis techniques have been applied in a wide range of fields including genomics, financial econometrics, signal processing, and various industrial processes [1–5]. Principal component analysis (PCA) is among the most commonly and widely used multivariate techniques with various applications ranging from facial recognition to data dimension reduction to clustering.

PCA reduces the dimensionality of the dataset while capturing as much variability as possible. In other words, PCA captures the essential information (variance) in m variables of the original data set in l retained principal components (PCs). In most conventional settings, data are high dimensional but the underlying signal has a low-dimensional structure. Thus, l is often much smaller than m. Mathematically, finding such PCs reduces to an eigenvalue/eigenvector problem. The PCs obtained from PCA represent the eigenvectors of the sample covariance or the sample correlation matrix of the original dataset. This means that PCA successively maximizes the variance by finding PCs that have the following properties: they are linear combinations of the original m variables, ordered according to their variance magnitudes, uncorrelated and the vectors of their coefficients, also called component loadings, are orthogonal. Such constraints on the derivation of PCs have advantages and disadvantages. For the former, preservation of the distance between data points, having a diagonal covariance matrix and uncorrelated components are the most obvious choices. On the other hand, to satisfy these constraints, most PCs have non-zero loadings for all original variables. This, in turn, makes the interpretation of PCs challenging when the dimension m is large.

A number of researchers proposed approaches to address the interpretation concerns in PCA [6–10]. Rotation of PCs is a common practice wherein the rotated components are easier to interpret without any loss of information. Jolliffe [8] described several normalization techniques for the rotation of PCs that are helpful for interpreting the individual components. The rotated components can be either pairwise uncorrelated or orthogonal, depending upon the normalization chosen for the loadings prior to the rotation. In addition, different normalization criteria can lead to different quantitative results. Moreover, in conventional PCA, each component captures as much variance as possible. Thus, the first PC captures the maximum variance and the

* Corresponding author.
E-mail address: anpalazoglu@ucdavis.edu (A. Palazoglu).

http://dx.doi.org/10.1016/j.chemolab.2017.01.018
Received 6 October 2016; Received in revised form 20 January 2017; Accepted 23 January 2017
Available online 31 January 2017

variance captured by remaining components decreases monotonically.

Other techniques proposed in the literature to improve interpretability of the PCs do so by imposing additional constraints. In such approaches, the sparsity in PC loadings is obtained at the expense of explained variance. The simplified component technique (SCoT) is one such approach wherein a penalty function is introduced to obtain the required sparsity of PC loadings [11]. Each successive components obtained using SCoT can be constrained to be orthogonal to, or uncorrelated with, one another to obtain the desired sparsity. Jolliffe and Uddin [11] demonstrated that SCoT outperformed rotated PCA in terms of the varimax criterion [12]. However, SCoT suffers from having many local optima and the choice of the penalty function is problem specific; there is no single penalty value that would work for all cases.

Jolliffe et al. [13] proposed the Simplified Component Technique – LASSO (SCoTLASS) which adds a "least absolute shrinkage and selection operator" (LASSO) constraint to SCoT. In SCoTLASS, an extra constraint is introduced in the form of a bound on the sum of absolute values of loadings in that component. This constraint shrinks some of the loadings on the components to be exactly zero which makes it more favorable for variable selection. While SCoTLASS has clear advantages over rotated PCA and SCoT, the introduction of the additional constraint requires a decision on a tuning parameter (t) that limits the search space for an optimal solution. Jolliffe et al. [13] showed that for values of t larger than the square root of the number of variables ($\sqrt{m}$), a PCA solution is obtained and as t gets smaller than $\sqrt{m}$, the number of non-zero loadings (NNZL) on each component also decreases. Again, the increased sparsity in the loading structure is obtained by a decrease in variance explained. In addition, the value of t also affects the correlation between the PCs and there is no satisfactory rule for selecting t. Thus, the choice of t is crucial and has to be studied subjectively to obtain a suitable sparsity-variance trade-off.

There are several other methodologies proposed in the literature to obtain sparse loadings [4,13–19]. Trendafilov [20] and Jolliffe et al. [21] offered reviews of main approaches and recent developments for improving the interpretation of results obtained from PCA. The technique that will be used and discussed in this paper is the one proposed by Zou et al. [4] who obtained sparse loadings by reformulating PCA as a regression problem and imposing LASSO (elastic net) constraints on the L_1 norm of the regression coefficients (sparse loadings). This methodology, known as sparse principal component analysis (SPCA), has several advantages such as it efficiently solves the optimization problem with a cost of a single least squares fit, can be applied in the case when m is much larger than sample size and the desired NNZL can be specified for each component. The SPCA algorithm will be discussed in detail in the preliminaries section.

Specifying NNZL for each SPC is a numerically hard combinatorial problem [22]. Some examples of such problems are the travelling salesman, the knapsack problem, cloud deployment options for supporting migration of software to the cloud or financial applications such as constrained portfolio selection [23,24]. They could be solved by general search heuristic procedures like simple enumeration that requires large computational times and is impractical if the problem dimension is large. To tackle such hard problems, Evolutionary Algorithms (EA) based on the principle of evolution were introduced in the past decade [25]. Genetic algorithms (GA) represent a widely used type of EA for constrained and unconstrained optimization [24,26–28] and are unbiased adaptive heuristic search algorithms [29]. In this paper, we propose the use of genoud (GENetic Optimization Using Derivatives) function that effectively combines EA methods with a derivative based (quasi-Newton) method to solve difficult optimization problems [30,31]. The function genoud can be used to solve problems for which derivatives do not exist, that are nonlinear or perhaps even discontinuous in the parameters of the function to be optimized.

Inspired by the challenges mentioned above, we propose the use of a genetic algorithm to specify the number of NNZL on each principal component for the SPCA.

The paper is organized as follows: the next section briefly introduces PCA, SPCA and genetic optimization concepts for the sake of completeness, followed by the introduction of the case studies based on the synthetic example, pitprops data and Tennessee Eastman benchmark process simulation. The results for selecting NNZL for each SPC are discussed next. Subsequently, the results obtained from the SPCA on the case studies are compared with the conventional PCA. Finally, the conclusions and directions for future work are presented.

2. Preliminaries

2.1. Principal Component Analysis (PCA)

PCA is the eigenvector decomposition of the covariance or the correlation matrix obtained from data matrix $X \in \mathbf{R}^{n\times m}$ that contains n observations of m process variables and is already scaled to zero mean and unit variance, into a transformed subspace of reduced dimension. The sample covariance matrix of X is defined as:

$$\mathrm{cov}(X) = \Sigma = \frac{X^T X}{n-1} \tag{1}$$

The decomposition is then expressed as follows:

$$X = TP^T = \widetilde{X} + E \tag{2}$$

where $T \in \mathbf{R}^{n\times m}$ and $P \in \mathbf{R}^{m\times m}$ are the score matrix and the loading matrix, respectively. The matrices $\widetilde{X}$ and E represent the estimation of X and the residual part of the PCA model, respectively, and are defined as follows:

$$\widetilde{X} = T_l P_l^T \triangleq \sum_{i=1}^{l} t_i p_i^T \tag{3}$$

$$E = T_{m-l} P_{m-l}^T \triangleq \sum_{i=l+1}^{m} t_i p_i^T \tag{4}$$

The PC projection reduces the original set of variables to l PCs where l must be equal or less than the smaller dimension of X. The decomposition assumes that PC loadings are orthonormal ($p_i^T p_j = 0 \ for \ i \neq j, \ p_i^T p_j = 1 \ for \ i = j$) and PC scores are orthogonal ($t_i^T t_j = 0 \ for \ i \neq j$). The p_i are the eigenvectors of the covariance,Σ or correlation matrix, given by:

$$\mathrm{cov}(X) p_i = \Lambda_i p_i \tag{5}$$

where Λ_i is the eigenvalue associated with the eigenvector p_i. The loadings (p_i) contain information on how the variables relate to each other whereas t_i vectors are scores that contain information on how the samples relate to each other.

As mentioned earlier, the dimension of the data is set to be equal to the number of PCs. The number of PCs is often reduced to a set of size l, where $1 \leq l \leq m$. The optimal number of components is chosen such that the model captures the variation in the dataset and not the noise. The trade-off here is that the closer the value of l is to m the better the PCA model will fit the data since more information is then retained, while the closer l is to 1, the simpler the model. However, to make the analysis easier, we only want to retain components capturing most of the variation in the dataset. Several techniques to determine l, i.e., the number of "meaningful" components, have been proposed in the literature [6,32]. An overview of the techniques can also be found in Cangelosi and Goriely [33]. In this work, we use the simple rule that the first l PCs retain 85% *CPV*.

The percentage of explained variance (*PEV*) is the fraction of explained variance by each PC and the cumulative percent variation is a measurement of the percent variation captured by the first l ordered PCs given by: