Chemometrics and Intelligent Laboratory Systems 149 (2015) 99–116

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

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

Determining the number of components in principal components analysis: A comparison of statistical, crossvalidation and approximated methods

Edoardo Saccenti [a,*], José Camacho [b,*]

[a] Laboratory of Systems and Synthetic Biology, Wageningen University and Research Centre, Dreijenplein 10, 6703 HB Wageningen, The Netherlands
[b] Network Engineering and Security Group, Signal Theory, Networking and Communications Department, University of Granada, C/ Periodista Daniel Saucedo Aranda s/n 18071, Granada, Spain

ARTICLE INFO

Article history:
Received 26 June 2015
Received in revised form 11 October 2015
Accepted 14 October 2015
Available online 26 October 2015

Keywords:
Covariance matrix
Tracy–Widom distribution
Dimensionality assessment
Random matrix theory
Eigenanalysis

ABSTRACT

Principal component analysis is one of the most commonly used multivariate tools to describe and summarize data. Determining the optimal number of components in a principal component model is a fundamental problem in many fields of application. In this paper, we compare the performance of several methods developed for this task in different areas of research. We consider statistical methods based on results from random matrix theory (Tracy–Widom and Kritchman–Nadler testing procedures), cross-validation methods (namely the well-characterized element wise *k*-fold algorithm, *ekf*, and its corrected version *cekf*) and methods based on numerical approximation (SACV and GCV). The performance of these methods is assessed on both simulated and real life data sets. In both cases, differential behavior of the considered methods is observed, for which we propose theoretical explanations.

1. Introduction

Multivariate statistical models are widely used in many fields of research to handle data sets with a very large number of variables and (possibly) of observations. Principal component analysis (PCA) [1,2] is one of the most commonly used multivariate tools to describe and summarize large omics data sets by finding the subspace in the space of the original variables where the data most vary [3]. In PCA the possibly correlated original variables are converted into sets of linearly uncorrelated variables called principal components which are in number less or equal than the number of original variables.

The PCA model follows the expression

$$\mathbf{X} = \mathbf{T}_K \mathbf{P}_K^{\mathrm{T}} + \mathbf{E}_K \tag{1}$$

where $\mathbf{X}$ is a $n \times p$ data matrix, $\mathbf{T}_K$ is the $n \times K$ scores matrix containing the projection of the observations onto the K-dimensional space defined by the first K principal components, $\mathbf{P}_K$ is the $p \times K$ matrix of the loadings, containing the linear combination of the original variables represented by each principal component, and $\mathbf{E}_K$ is the $n \times p$ matrix of the residuals.

Determining the optimal number of components K that best fit the data is a fundamental task in multivariate analysis and, as noted by several authors [4,5], it is an ill-posed problem when formulated without specifying for which purpose PCA is used. Generally speaking, one can refer to the optimal number of components implying that the model describes systematic variation in the data but not the noise [4], but this can be different depending whether the application is, for instance, in process monitoring or data compression. Camacho and Ferrer [5] recently proposed a taxonomy for PCA applications, depending on what the interest is focused on: 1) the (accurate approximation of the) observed variables such as in data compression or dimensionality reduction, 2) understanding and interpretation of latent variables and 3) the distribution in latent variables and residuals. In this paper we place ourselves in the situation described in 1: the interest relies in using PCA to extract information which is embedded in a high-dimensional space and describing it with a limited number of components, a problem typical in modern functional genomics, econometrics, signal theory and image processing.

A great deal of attention has been dedicated to this problem and a plethora of methods has been proposed, mostly by the chemometrics, psychometrics and statistics communities (for a review see for instance [6] and references therein).

Jolliffe [7] and Jackson [8] outlined a taxonomy of the criteria proposed to find the optimum number of components in PCA, distinguishing three broad categories:

1. *Ad-hoc* rules, like Cattel's scree test [9], the indicator function or the embedded error [10].
2. Statistical tests, like Bartlett's test for the first component [11], the sphericity test [12] or the Malinowski's *F*-test [13].

* Corresponding authors.
E-mail addresses: esaccenti@gmail.com (E. Saccenti), josecamacho@ugr.es (J. Camacho).

http://dx.doi.org/10.1016/j.chemolab.2015.10.006
0169-7439/

3. Computational criteria, like cross-validation (CV), bootstrapping and permutation like Horn's parallel analysis [14] or the SVD based methods proposed by Dray [15].

The array of available methods for dimensionality assessment is constantly increasing. For instance, the CHull approach [16] for model selection can be added to the first category: it detects a model with an optimal balance between a (large) model fit and (low) number of parameter and can be applied to indicate the number of principal components [17]. Similarly, Josse and Husson [18] proposed new methods based on the numerical approximation of the CV procedure that add to the third category.

The idea of comparing methods for determining the number of principal components is certainly not new, and several studies presented comparative investigations [8,19–23]. However, these comparative studies did not consider recently developed statistical tools based on results from random matrix theory (RMT). Moreover, the performance of the latter has never been compared with state-of-the-art implementations of the cross-validation procedure or numerical approximation. For this reason it seems timely to review and perform an in-depth assessment of cross-validation, approximated and statistical methods through a large comparative study.

For this task we made use of 5 simulation schemes, corresponding to more than 12,000 different simulated data sets accounting for different data structure, data distribution and homo- and heteroscedastic noise. In addition, we made use of 8 real life chemometrics data sets (mostly NIR spectroscopy data, among which some well-known benchmark data sets) to investigate the behavior of the methods on experimental data. As the problem of determining the number of components is not limited to chemometrics, we additionally considered 12 data sets stemming from disciplines where chemometrics tools are routinely applied to model and extract information, such as metabolomics (5 data sets), proteomics (1 data set), and other (functional genomics, computational linguistics, *etc*..., 4 data sets).

The paper is organized as follows. Section 2 offers a brief overview of past works related to the problem of dimensionality assessment in PCA; Section 3 is dedicated to the illustration of methods based on random matrix theory, cross-validation and approximation of the cross-validation for determining the number of components in PCA. To make the paper self-contained, we provide a theoretical background on which to base the discussion and interpretation of the results. Section 4 gives the description of the data sets used for comparison of the different methods and Section 5 is dedicated to the software used. Section 6 offers a discussion of the results. We end with some final considerations in Section 7 where we also suggest some guidelines for the practitioners.

2. Related work

Until recently, the statistical tools to attack the problem of determining the number of components in PCA consisted mainly in methods developed in the field of psychometrics (like Bartlett's test for the first component [11], the sphericity test [12], the Kaiser–Guttman's 1 rule [24]) or chemometrics (like Malinowski's *F*-test [13] and the Faber–Kowalski test [25]). All these methods suffer from the drawback of being of limited applicability, either because restricted to the first component, derived under assumption rarely met in real practice or lacking a solid statistical background as they rely on approximated distributions from which deliberating on data dimensionality.Most statistical methods are based on eigenanalysis and attempt to distinguish between eigenvalues of the sample covariance matrix associated to systematic variation and eigenvalues due to noise.

Only recently, results from RMT provided a solid and firm statistical foundation to correctly describe the distributional properties of the behavior of noise eigenvalues. The main result was the finding that the so-called Tracy–Widom distribution is the limiting distribution of the largest eigenvalue(s) of random sample covariance matrices. This finding, expressed by Johnstone's theorem [26], opened the way to a long-sought inferential dimensionality assessment in PCA.

Like any other statistical test, RMT methods are based on assumptions like certain distributional properties of the data or precise structures of the covariance model under which the null model for noise eigenvalues is derived. Although these assumptions can be mild, they cannot be always met or verifiable in practice. For this reason, it is of interest to benchmark and compare the performance of such methods together with that of methods that do not require distributional assumptions.

A class of methods fulfilling these requirements is that of cross-validation (CV) methods. In contrast with statistical methods that focus on the eigenvalues of the sample covariance matrix, CV methods try to reproduce the error estimation procedure when applying a model on new/independent data [27–29].

The idea of applying a cross-validation method to identify the dimension that best describes the systematic variation in data dates back to the seminal paper by Wold [30]. In general, in the cross-validation the data in $\mathbf{X}$ is partitioned in G groups and at each step a PCA model is fitted using G-1 groups. Then the data in the left out group is predicted using the model. A criterion of goodness of fit is then defined and the procedure is repeated for 1, 2,… components. The optimal number of components is then estimated by inspecting the shape of the goodness of fit curve. Several cross-validation strategies have been proposed: in this study we will focus on the so called element wise k-fold *ekf* cross validation (see Section 3.2 for more details), a technique that has been found to perform very well in comparative studies [4] and that has recently received a lot of attention at both the theoretical and applicative levels [5,31].

Disposing distributional properties comes at a cost: cross-validation methods are time consuming and although nowadays calculation power is not a limiting factor, it may not be practical to use them when (extremely) large data sets are considered, like, for instance, in modern functional genomics. Also, CV methods are not universally suited to determine the optimal number of components in all situations encountered in data analysis and modeling. The cross-validation approach considered in this paper is based on the prediction error, meaning that a specific piece of data is not used to compute its own prediction. Strictly speaking, the prediction error is only suited suitable to select the number of components when the goal of the PCA model is to perform predictions, for instance in missing data imputation. Still, this approach might be useful in other contexts as a heuristic indicator.

To overcome the problem of computational time, Josse and Husson [18] recently proposed to approximate the cross-validation procedure by an original interpretation of PCA as a smoothing operator. In such a way they provide two different approximations of another well-known cross-validation method, the so called expectation–maximization, an approach also found to perform well in [4]. Nonetheless, we will show that these approximated methods carry on some characteristics of the original cross-validation algorithm.

3. Methods for determining the number of principal components

In this Section, the methods under comparison are introduced. Methods were proposed in different research areas. As a result, their definition to the problem of selecting the number of components, including the considered model of noise, varies. The input of the methods reflects these differences. Thus, some methods take as input the matrix of data while others operate over the eigenvalues. Here, we propose a taxonomy based on the class of input, which in turn is determined by the model of noise under consideration.

A data matrix $\mathbf{X}$ is characterized by both the distribution of the observations and the relationships among the variables (covariances). As shown in [32], very different data samples can lead to the same sample covariance matrix: the covariance matrix can be seen as a summary of the data where the information about the distribution of the