Computational Statistics and Data Analysis 57 (2013) 95–111

Contents lists available at SciVerse ScienceDirect

Computational Statistics and Data Analysis

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

Exact methods for variable selection in principal component analysis: Guide functions and pre-selection

Joaquín Pacheco *, Silvia Casado, Santiago Porras

Departamento de Economía Aplicada, Universidad de Burgos, Spain

ARTICLE INFO

Article history:
Received 16 November 2011
Received in revised form 5 June 2012
Accepted 13 June 2012
Available online 18 June 2012

Keywords:
PCA
Variable selection
Branch & Bound methods
Guide functions
Filters

ABSTRACT

A variable selection problem is analysed for use in Principal Component Analysis (PCA). In this case, the set of original variables is divided into disjoint groups. The problem resides in the selection of variables, but with the restriction that the set of variables that is selected should contain at least one variable from each group. The objective function under consideration is the sum of the first eigenvalues of the correlation matrix of the subset of selected variables. This problem, with no known prior references, has two further difficulties, in addition to that of the variable selection problem: the evaluation of the objective function and the restriction that the subset of selected variables should also contain elements from all of the groups. Two Branch & Bound methods are proposed to obtain exact solutions that incorporate two strategies: the first one is the use of "fast" guide functions as alternatives to the objective function; the second one is the preselection of variables that help to comply with the latter restriction. From the computational tests, it is seen that both strategies are very efficient and achieve significant reductions in calculation times.

1. Introduction

Data sets with large numbers of variables are processed in many disciplines such as economics, sociology, engineering, medicine and biology, among others. The researcher has to process a lot of data classified by a large number of variables, which are often difficult to summarize or interpret. One useful approach involves the reduction of data dimensionality, while trying to preserve as much of the original information as possible. A common way of doing this is through Principal Component Analysis (PCA).

PCA is widely applied in data mining to investigate data structures. Its purpose is to construct components, each of which contains a maximal amount of variation with respect to the data that are unexplained by any other components. Standard results guarantee that retaining the k Principal Components (PCs) with the largest associated variance produces the k-subset of linear combinations of the n original variables which, according to various criteria, represents the best approximation of the original variables (see, for example, Jolliffe, 2002). The user therefore expects that the information in the data can be summarized into a few principal components. Once the principal components have been determined, all further analysis can be carried out on them instead of on the original data, as they summarize the relevant information. Thus, PCA is frequently

* Correspondence to: Fac C. Económicas y Empresariales, Plaza Infanta Elena s/n, BURGOS 09001, Spain. Tel.: +34 947 25 90 21; fax: +34 947 25 80 13.
E-mail addresses: jpacheco@ubu.es (J. Pacheco), scasado@ubu.es (S. Casado), sporras@ubu.es (S. Porras).

0167-9473/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2012.06.014

considered the first step of a statistical data analysis that aims at data compression: decreasing the dimensionality of the data, but losing as little information as possible.

While PCA is highly effective at reduced-dimensional representation, it does not provide a real reduction of dimensionality in terms of the original variables, since all n original variables are required to define a single Principal Component (PC). As McCabe (1984) has stated, "interpretation of the results and possible subsequent data collection and analysis still involve all of the variables". Moreover, Cadima and Jolliffe (1995, 2001) have shown that a PC can provide a misleading measure of variable importance, in terms of preserving information, because it is based on the assumption that the resultant of a linear combination (PC) is dominated by the vectors (variables) with large magnitude coefficients in that linear combination (high PC loadings). This assumption ignores the influence of the magnitude of each vector (the standard deviation of each variable) and the relative positions of the vectors (the pattern of correlations between the variables). One way of achieving a simple interpretation is to reduce the number of variables, that is, to look for a subset of the n variables that approximate, as far as possible, the k retained PCs. We consider the combinatorial problem of identifying, for any arbitrary integer p ($k \leq p \leq n$), a p-variable subset which is optimal with respect to a given criterion.

In most cases, the inclusion of all the variables in a statistical analysis is, at best, unnecessary and, at worst, a serious impediment to the correct interpretation of the data. From a computational point of view, variable selection is a Nondeterministic Polynomial Time-Hard or NP-Hard problem (Kohavi, 1995); and (Cotta et al., 2004): there is therefore no guarantee of finding the optimum solution. This means that when the size of the problem is large, finding an optimum solution is, in practice, unfeasible. Two different methodological approaches have been developed for variable selection problems: optimal or exact techniques (enumerative techniques), which are able to guarantee an optimal solution, but which are only applicable to small-sized sets; and heuristic techniques, which are able to find good solutions (although unable to guarantee the optimum) within a reasonable amount of time.

The problem of selecting a subset of variables from a larger candidate pool abounds in areas such as multiple linear regression (Furnival and Wilson, 1974; Miller, 2002; Gatu and Kontoghiorghes, 2006; Hofmann et al., 2007; Gatu et al., 2007), logistic regression (Pacheco et al., 2009), polynomial regression (Peixoto, 1987; Brusco et al., 2009b; Brusco and Steinley, 2010), factor analysis (Kano and Harada, 2000; Hogarty et al., 2004), cluster analysis (Brusco and Cradit, 2001; Steinley and Brusco, 2008; Krzanowski and Hand, 2009), and discriminant analysis (McCabe, 1975; McKay and Campbell, 1982a,b; Pacheco et al., 2006).

Specifically, the problem of variable selection in PCA has been investigated by Jolliffe (1972), Jolliffe (1973), Robert and Escoufier (1976), McCabe (1984), Bonifas et al. (1984), Krzanowski (1987a), Falguerolles and Jmel (1993), Mori et al. (1994), Jolliffe (2002), Duarte Silva (2002), Cadima et al. (2004), Mori et al. (2007) and Brusco et al. (2009a) among others. These studies sought to obtain PCs based on a subset of variables, in such a way that they retain as much information as possible in comparison to PCs that are based on all the variables. To address this problem, it is necessary to tackle two secondary problems: (1) the establishment of an objective criterion that can measure the quality or fitness of every subset of variables; and (2) the development of solution procedures for finding optimal, or at least near-optimal, subsets based on these criteria. The methods proposed by Jolliffe (1972, 1973) consider PC loadings, and those of McCabe (1984) and Falguerolles and Jmel (1993) use a partial covariance matrix to select a subset of variables which, as far as possible, maintains information on all variables. Robert and Escoufier (1976) and Bonifas et al. (1984) used the RV-coefficient, and Krzanowski (1987a,b) used Procrustes analysis to evaluate closeness between the configuration of PC computations based on selected variables and the configuration based on all of the variables. Tanaka and Mori (1997) discussed a method called "modified PCA" (MPCA) to derive PCs which were computed by using only a selected subset of variables but which represented all of the variables, including those that were not selected. Since MPCA naturally includes variable selection procedures in its analysis, its criteria can be used directly to detect a reasonable subset of variables. Further criteria may be considered based, for example, on the influence analysis of variables and on predictive residuals, using the concepts reported in Tanaka and Mori (1997), and in Krzanowski (1987a,b), respectively; for more details see Iizuka et al. (2003).

Thus, the existence of several methods and criteria is one of the typical features of variable selection in multivariate methods without external variables such as PCA (here the term "external variable" is used as a variable to be predicted or explained using the information derived from other variables). Moreover, the existing methods and criteria often provide different results (selected subsets of variables), which is regarded as a further typical feature. This occurs because each criterion has its own reasonable variable selection method, despite its purpose of selecting variables; we cannot say, therefore, that one is better than any other. These features are not observed in multivariate methods with external variable(s), such as multiple regression analysis (Mori et al., 2007).

In practical applications of variable selection, it is desirable to have a computational environment where those who want to select variables can apply a suitable method for their own selection purposes without difficulties and/or can try various methods and choose the best method by comparing results.

In many studies, the initial variables are divided into previously selected groups. In these cases it is required, or at least recommended to use variables from all groups considered. This happens, for example, in the construction of composite indicators that are used in several areas (economy, society, quality of life, nature, technology, etc.). They are used as measure of the evolution of regions or countries in such areas. The composite indicators should try to cover all points of view of