Available online at www.sciencedirect.com
SCIENCE DIRECT®
Journal of Statistical Planning and Inference 136 (2006) 2630–2639
journal of statistical planning and inference
www.elsevier.com/locate/jspi

A confidence interval for the number of principal components

Pinyuen Chen*

Department of Mathematics, Syracuse University, 215 Carnegie Hall, Syracuse, NY 13244-1150, USA

Received 9 January 2004; accepted 11 October 2004
Available online 23 November 2004

Abstract

This paper proposes a confidence interval for the number of important principal components in principal component analysis. An important principal component is defined as a principal component whose value is close to the value of the largest principal component. More specifically, a principal component λ_i is called important if λ_i/λ_1 is sufficiently close to 1 where λ_1 is the largest eigenvalue. A distance measure for closeness will be defined under the framework of ranking and selection theory. A confidence interval for the number of important principal components will be proposed using a stepwise selection procedure. The proposed interval, which is asymptotic in nature, includes the true important components with a specified confidence. Numerical examples are given to illustrate our procedure.

MSC: primary 62H25; secondary 62F07; 62F25; 62H10

Keywords: Asymptotic distribution; Confidence limit; Covariance matrix; Eigenvalue; Principal component analysis

1. Introduction

Principal component analysis was introduced by Pearson (1901) as a tool of fitting planes to a system of points in space and later put forward by Hotelling (1933) in the analysis of a psychological measurement. Bartlett (1950, 1951a,b, 1954) and Lawley (1956) have made

* Tel.: +315 443 1577; fax: +315 443 1475.
E-mail address: pinchen@syr.edu (P. Chen).

0378-3758/$ - see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.jspi.2004.10.016

significant contributions in estimating and testing of the number of components in this area. Bartlett (1950) also used examples to illustrate the application of principal component analysis in education and psychology. A thorough investigation of the asymptotic theory for principal component analysis was made by Anderson (1963). More recently, principal component analysis has been applied to financial economics. We refer to Campbell et al. (1997, Chapter 6), Hasbrouck and Seppi (2001), and Johnson and Wichern (1997, Chapter 8), for details.

While the majority of previous work in principal components has been focused on hypothesis testing and estimation of the number of components, confidence interval estimation was studied by Anderson (1963) (Section 3) for the smallest eigenvalue. In this paper, we are interested in the confidence interval of the number of principal components instead of their values and we accomplish this by following an asymptotic approach. Huang and Tseng (1992) used a ranking and selection approach to select the number of components in principal component analysis and used the ratios of two sums of eigenvalues to define the preference zone in the parameter space. The ratio in their study as well as in typical principal component analysis represents the proportion of total population variance due to individual component. Since the distribution of the plug-in statistics of those ratios depends on the unknown parameters in a complex manner, majorization theory was used in their article to determine the minimum value of the probability of a correct estimation. In defining the preference zone, we use instead the ratios of individual eigenvalues to the largest eigenvalue, which are invariant under a general group of transformations. (See Section 8.3 in Muirhead, 1982.) The distribution of our test statistics, which are the plug-in estimates of the ratios used in defining the preference zone, depends on unknown parameters in a simple manner. Therefore, we can determine the minimum value of the probability of a correct estimation from the underlying distribution. No comparison is attempted between Huang and Tseng's (1992) study and our study because the two different distance measures lead to different interpretations of the "important components" as described in this section.

This paper is organized as follows. In Section 2, we formally define the selection formulation for the problem, and construct the upper and the lower limits for our confidence interval. In Section 3, we prove some properties of our procedure. In Section 4, we demonstrate in two examples how the procedure constants can be estimated through simulation. We also use the data in the examples to illustrate our procedure.

2. Formulation and procedure

Let $X_1, X_2, \ldots, X_n$ be n-independent and identically distributed multivariate normal random p-vectors with mean vector μ and covariance matrix Σ, and $S=\sum_{i=1}^{n}(X_i-\bar{X})(X_i-X_i)'/(n-1)$ be the sample covariance of the sample. Let the eigenvalues of Σ be denoted by $\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_p$ and the ordered eigenvalues of S be denoted by $\ell_1 \geqslant \ell_2 \geqslant \cdots \geqslant \ell_p$. It is known that the variance of the ith principal component is λ_i and its maximum likelihood estimate is ℓ_i. We are interested in the number of principal components whose variances are significantly larger than the variances of the remaining principal components. Define $\Omega=\{(\lambda_1, \lambda_2, \ldots, \lambda_p)|\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_p > 0\}$, which is the parameter space for the ordered eigenvalues of Σ. For $1 < i \leqslant p$ and $0 < \delta^* \leqslant 1$, let $\Omega_{g,i}$ and $\Omega_{b,i}$, be two disjoint subsets of