Available online at www.sciencedirect.com

ScienceDirect

Computational Statistics & Data Analysis 51 (2006) 918 – 930

COMPUTATIONAL STATISTICS & DATA ANALYSIS

www.elsevier.com/locate/csda

Automatic dimensionality selection from the scree plot via the use of profile likelihood

Mu Zhu*, Ali Ghodsi

Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, Ont., Canada N2L 3G1

Received 14 January 2005; received in revised form 21 September 2005; accepted 21 September 2005
Available online 25 October 2005

Abstract

Most dimension reduction techniques produce ordered coordinates so that only the first few coordinates need be considered in subsequent analyses. The choice of how many coordinates to use is often made with a visual heuristic, i.e., by making a scree plot and looking for a "big gap" or an "elbow." In this article, we present a simple and automatic procedure to accomplish this goal by maximizing a simple profile likelihood function. We give a wide variety of both simulated and real examples.

Keywords: Data compression; Denoising; Isomap; Latent semantic indexing; Manifold learning; Principal component analysis (PCA); Resampling methods; Singular value decomposition (SVD)

1. Introduction

In modern applications, we often encounter high-dimensional data. More often than not, the intrinsic dimensionality of the data is much lower, i.e., even though the data are lying in a high-dimensional space, only a few dimensions are actually important for the analysis. Various dimension reduction techniques are available. Suppose $\mathbf{x} \in \mathbb{R}^p$ is a p-dimensional vector for some relatively large p. Most of these dimension reduction techniques will produce ordered mappings from $\mathbb{R}^p$ to $\mathbb{R}$, say $\alpha_1(\mathbf{x}), \alpha_2(\mathbf{x}), \ldots, \alpha_m(\mathbf{x})$ where $m \leqslant p$, such that $\alpha_1(\mathbf{x})$ is the most important coordinate, $\alpha_2(\mathbf{x})$ is the next most important coordinate, and so on. Associated with each mapping is a measure of its relative importance, say $d_1 \geqslant d_2 \geqslant \cdots \geqslant d_m$, so that the resulting coordinates can be ordered in a meaningful way. Dimension reduction is then achieved by selecting only the top few coordinates.

The problem that we shall focus on in this article is that of deciding how many coordinates should be retained. Ideally, if the intrinsic dimensionality is, say 3, we would like to see $d_1, d_2, d_3 > 0$ and $d_j = 0$ for all $j > 3$. However, this seldom happens because the data are often noisy.

To review some of the commonly used techniques for making such a decision, it is convenient to focus on the case of principal component analysis (PCA), which is perhaps the best-known dimension reduction technique of all.

* Corresponding author. Tel.: +1 519 888 4567x6987; fax: +1 519 746 1875.
E-mail addresses: m3zhu@uwaterloo.ca (Mu Zhu), aghodsib@uwaterloo.ca (Ali Ghodsi).

0167-9473/$ - see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2005.09.010

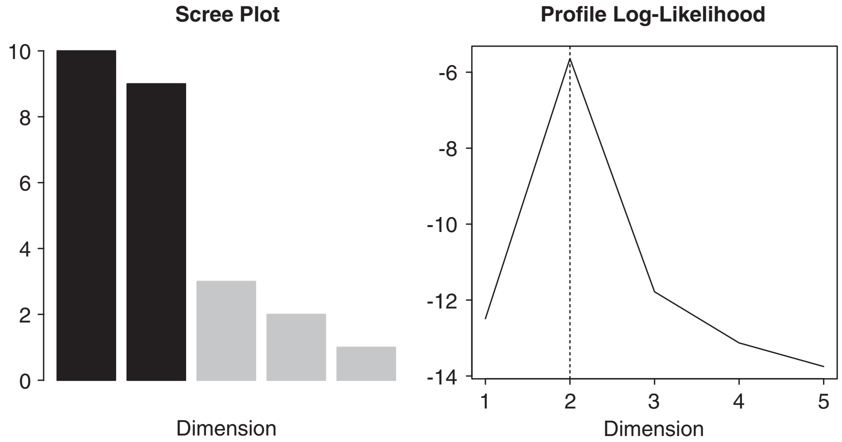

Fig. 1. A first example. Suppose the eigenvalues are 10, 9, 3, 2, 1; there exists a "big gap" between the second and third eigenvalues.

Given data $\mathbf{x} \in \mathbb{R}^p$, the principal components are defined by unit vectors $\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \ldots, \boldsymbol{\alpha}_p \in \mathbb{R}^p$ such that

$$\operatorname{Var}\left(\boldsymbol{\alpha}_1^{\mathrm{T}}\mathbf{x}\right) \geqslant \operatorname{Var}\left(\boldsymbol{\alpha}_2^{\mathrm{T}}\mathbf{x}\right) \geqslant \cdots \geqslant \operatorname{Var}\left(\boldsymbol{\alpha}_p^{\mathrm{T}}\mathbf{x}\right)$$

and $\operatorname{Cov}\left(\boldsymbol{\alpha}_i^{\mathrm{T}}\mathbf{x}, \boldsymbol{\alpha}_j^{\mathrm{T}}\mathbf{x}\right) = 0$ for all $i \neq j$.

Therefore, PCA produces mappings from $\mathbb{R}^p$ to $\mathbb{R}$ that are simply projections, i.e., $\alpha_j(\mathbf{x}) = \boldsymbol{\alpha}_j^{\mathrm{T}}\mathbf{x}$, and the importance of mapping j is simply measured by the marginal variance along the projection, i.e., $d_j = \operatorname{Var}\left(\boldsymbol{\alpha}_j^{\mathrm{T}}\mathbf{x}\right)$. Mardia et al. (1979) give a nice overview of PCA as well as other related multivariate techniques. Let $\mathbf{S}$ be the sample variance–covariance matrix of the data; the principal components $\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \ldots, \boldsymbol{\alpha}_p$ are simply the eigenvectors of $\mathbf{S}$, and $d_1 \geqslant d_2 \geqslant \cdots \geqslant d_p$ are simply the (ordered) eigenvalues of $\mathbf{S}$.

In order to determine how many principal components are needed, a number of approaches are available; Jolliffe (2002, Chapter 6) gave perhaps the most comprehensive and up-to-date summary on the status of the current practices. These methods can be roughly classified into the following categories although within each category there are still various minor variations:

(1) *Percent variance*: Find the smallest number of components to capture a certain percentage of the total variance, i.e., retain the top q components where q is the smallest integer between 1 and p such that

$$\frac{d_1 + d_2 + \cdots + d_q}{d_1 + d_2 + \cdots + d_p} \geqslant \gamma,$$

where γ is a pre-determined level, say, 80% or 90%.

(2) *Scree plot*: Plot the eigenvalues $d_1, d_2, \ldots, d_p$ in descending order (often called a scree plot) and look for a "big gap" or an "elbow" in such a graph. For example, as illustrated in the left panel of Fig. 1, there is a "big gap" between the second and third eigenvalues, so the first two principal components are retained and the rest, discarded.

(3) *Sequential tests*: Sequentially conduct a series of formal hypothesis tests to determine whether the small eigenvalues are equal. For $j = 1, 2, \ldots, p-1$, consider a series of null hypotheses:

$$\mathrm{H}_{0,j} : d_p = d_{p-1} = \cdots = d_{p-j}.$$

We start by testing $\mathrm{H}_{0,1}$, $\mathrm{H}_{0,2}$ and so on until a null hypothesis is first rejected. Suppose $\mathrm{H}_{0,q}$ is the first rejected null hypothesis, then the first $p - q$ components are retained.

(4) *Resampling*: Estimate the null distribution of each d_j by resampling the data repeatedly, e.g., via permutation or the bootstrap. Retain component j if d_j exceeds the 95 percentile of the corresponding null distribution and discard component j if otherwise.