Available online at www.sciencedirect.com
ScienceDirect
Computational Statistics & Data Analysis 51 (2007) 4101 – 4123
COMPUTATIONAL STATISTICS & DATA ANALYSIS
www.elsevier.com/locate/csda

On Bayesian principal component analysis

Václav Šmídl[a,*,1], Anthony Quinn[b]

[a] *Institute of Information Theory and Automation, Pod Vodárenskou věží 4, Prague 8, Czech Republic*
[b] *Trinity College Dublin, Ireland*

Received 31 October 2005; received in revised form 17 January 2007; accepted 18 January 2007
Available online 1 February 2007

Abstract

A complete Bayesian framework for principal component analysis (PCA) is proposed. Previous model-based approaches to PCA were often based upon a factor analysis model with isotropic Gaussian noise. In contrast to PCA, these approaches do not impose orthogonality constraints. A new model with orthogonality restrictions is proposed. Its approximate Bayesian solution using the variational approximation and results from directional statistics is developed. The Bayesian solution provides two notable results in relation to PCA. The first is uncertainty bounds on principal components (PCs), and the second is an explicit distribution on the number of relevant PCs. The posterior distribution of the PCs is found to be of the von-Mises–Fisher type. This distribution and its associated hypergeometric function, ${}_0F_1$, are studied. Numerical reductions are revealed, leading to a stable and efficient orthogonal variational PCA (OVPCA) algorithm. OVPCA provides the required inferences. Its performance is illustrated in simulation, and for a sequence of medical scintigraphic images.

Keywords: Principal component analysis (PCA); Variational bayes (VB); Orthogonal variational PCA (OVPCA); von-Mises–Fisher distribution; Medical scintigraphic imagery

1. Introduction

Principal component analysis (PCA) is one of the classical data analysis tools for dimensionality reduction. It is used in many application areas including data compression, de-noising, pattern recognition, shape analysis and spectral analysis. For an overview of its use, see Jolliffe (2002).

PCA is often used as a black box numerical tool, because of its mature numerical implementation and ease of use. The correspondence between principal components (PCs) and the eigenvectors of a symmetric matrix makes it intuitively appealing. However, further investigation of PCs—namely the choice of an optimal number of relevant PCs and confidence intervals on the PCs—is complicated. Many approximate solutions are available, both formal and *ad hoc* (Anderson, 1971; Jolliffe, 2002, Chapter 6 and Section 3.6). Simple *ad hoc* criteria are used in applications such as data compression and de-noising, where the number of PCs is restricted by an available bit rate or computational

* Corresponding author. Tel.: +420 26605 2420; fax: +420 26605 2068.
E-mail address: smidl@utia.cas.cz (V. Šmídl).
[1] Supported by MŠMT Grant 1M0572.

0167-9473/$ - see front matter
doi:10.1016/j.csda.2007.01.011

cost, rather than by statistical relevance. However, formal solutions are required in applications where uncertainty of estimates is an important part of the problem. A typical example is spectral analysis (Kay, 1993) or functional analysis of dynamic image data (Buvat et al., 1998).

Traditionally, probability distributions for PCs were derived using sampling theory (Jolliffe, 2002), but these results are mostly asymptotic. Recently, the problem was addressed using Bayesian methodology (Bishop, 1999), by invoking the factor analysis (FA) model with isotropic Gaussian noise. However, the FA model does not impose orthogonality restrictions, and so the posterior inferences are not identical to PCA. Moreover, rotational ambiguity in the FA model presents a computational difficulty that must be overcome by means of regularizing priors.

The original concept of PCA (Section 2) as well as probabilistic models which yield a maximum likelihood (ML) solution identical to that of PCA (Section 3) are reviewed. The ML solution does not provide an estimate of rank nor uncertainty bounds for the model estimates. Therefore, the problem is reformulated using the Bayesian methodology (Section 4). A variational approximation of the posterior distribution is investigated, and a numerically efficient algorithm—orthogonal variational PCA (OVPCA)—for estimation of an approximating posterior distribution is presented (Section 5). Further analysis yields both the distribution of rank as well as uncertainty bounds on PCs (Section 6). The performance of the method is illustrated using a simple simulation study. A contemporary application in medical imaging is also presented (Section 7). The posterior distribution of PCs is found to be in the form of the von-Mises–Fisher distribution. This is a matrix normal distribution with orthogonality restrictions on the matrix argument. Hence, the support of the distribution is confined to a Cartesian product of a set of hyperballs, known as the Stiefel manifold. For convenience, the key results for this distributional class are reviewed in Appendix A. The numerical efficiency of the OVPCA technique depends on the approach taken to evaluating the involved hypergeometric function of matrix argument, ${}_0F_1$. A novel approximation of this function—yielding results of acceptable accuracy and computational cost—is presented in Appendix B.

The following notational conventions are used:

$A \in \Re^{n\times m}$; $\boldsymbol{a}_i$	matrix of dimensions $n \times m$, generally denoted by a capital letter; its ith column, using bold-faced letter.
$A = U_A L_A V_A'$	singular value decomposition (SVD) of matrix $A \in \Re^{n\times m}$, where $U_A \in \Re^{n\times q}$, $L_A \in \Re^{q\times q}$, $V_A \in \Re^{m\times q}$, $q = \min(n, m)$. The SVD is expressed in short form, i.e. in terms of the only guaranteed-non-zero part of L_A, namely the upper-left $q \times q$ diagonal submatrix.
$U_{;r}$, $U_{A;r}$	operator selecting the first r columns of matrix U, U_A, respectively.
$L_{;r,r}$, $L_{A;r,r}$	operator selecting the $r \times r$ upper-left subblock of matrix L, L_A, respectively.
$\boldsymbol{l}_{;r}$, $\boldsymbol{l}_{A;r}$	operator extracting upper length-r subvector of vector $\boldsymbol{l}$, $\boldsymbol{l}_A$, respectively.
$f(x\|\theta)$	probability density function (pdf) of continuous random variable x, conditioned by known θ.
$\widehat{\theta}$	maximizer of $f(x\|\theta)$, with latter taken as a function of θ (the ML estimate).

2. Principal component analysis (PCA)

PCA is a widely used tool (Jolliffe, 2002) for representation of data sets. Consider a set of n p-dimensional data vectors, $\boldsymbol{d}_i$, from data space $\mathscr{D}$:

$$D = [\boldsymbol{d}_1, \ldots, \boldsymbol{d}_n], \quad \boldsymbol{d}_i \in \mathscr{D} = \Re^p.$$

For simplicity, the sample mean vector $\langle \boldsymbol{d} \rangle_n = (1/n)\sum_{i=1}^n \boldsymbol{d}_i$ is assumed to be zero. If $\langle \boldsymbol{d} \rangle_n \neq \mathbf{0}$, it can be subtracted from the raw data in a preprocessing step. Explicit modelling of the sample mean is discussed in Section 7.3. We also assume, without loss of generality, that $p \leqslant n$.

Let $\mathscr{P}_r$ be the orthogonal projection operator from $\mathscr{D}$ into the r-dimensional subspace, $\mathscr{A}_r$, with orthonormal basis $W_r = [\boldsymbol{w}_1, \ldots, \boldsymbol{w}_r] \in \Re^{p\times r}$:

$$\mathscr{P}_r : \mathscr{D} \to \mathscr{A}_r,$$

$$\boldsymbol{d}_i \to \boldsymbol{m}_i.$$