Latent Variable Models for Dimensionality Reduction

Zhihua Zhang
College of Comp. Sci. and Tech.
Zhejiang University
Hangzhou, Zhejiang 310027, China
zhzhang@cs.zju.edu.cn

Michael I. Jordan
Departments of EECS and Statistics
University of California, Berkeley
Berkeley, CA 94720, USA
jordan@cs.berkeley.edu

Abstract

Principal coordinate analysis (PCO), a dual of principal component analysis (PCA), is a classical method for exploratory data analysis. In this paper we provide a probabilistic interpretation of PCO. We show that this interpretation yields a maximum likelihood procedure for estimating the PCO parameters and we also present an iterative expectation-maximization algorithm for obtaining maximum likelihood estimates. Finally, we show that our framework yields a probabilistic formulation of kernel PCA.

1 Introduction

Multidimensional scaling (MDS) (Borg and Groenen, 1997) has been widely applied to data analysis and processing. Like principal component analysis (PCA) (Jolliffe, 2002), MDS is also an important tool for dimensionality reduction and visualization. Given the (dis)similarities between pairs of objects, MDS is concerned with the problem of representing the objects as points in a (usually) Euclidean space so that the distances between the points in the Euclidean space match the original dissimilarities as much as possible.

In terms of the techniques used for the configurations of points, MDS can be categorized into metric and nonmetric methods. Furthermore, metric MDS methods include classical scaling and least squares scaling. Classical scaling is commonly called principal coordinate analysis (PCO). Considering that there exists a duality between PCO and PCA (Gower, 1966), and given our interest in the relationship between MDS and PCA, we prefer to use the term PCO in this paper to refer to classical scaling MDS.

In PCO, the original dissimilarity measure is required to be Euclidean. Equivalently, the inner product function that induces the dissimilarity is positive definite. This inner product function thus defines a similarity measure and it can be referred to as a reproducing kernel.

Thus, Schölkopf et al. (1998) proposed kernel PCA (KPCA) as a nonlinear extension of PCA. There also exists a duality between PCO and KPCA (Williams, 2001). Thus, a nonlinear version of PCO can be devised by using reproducing kernels as similarities.

Different MDS models make use of different techniques to model the configurations of points. For example, conventional PCO employs the spectral decomposition method, while the metric least squares scaling uses an iterative majorization method (Borg and Groenen, 1997). A statistical approach to MDS has been devised maximum likelihood methods can be used to estimate the configurations of points (Ramsay, 1982; Groenen et al., 1995). In addition, Oh and Raftery (2001) proposed a Bayesian method for the configuration. In these statistical treatments, the dissimilarities are generally modeled as following a truncated normal or log-normal distribution.

However, these statistical approaches are not appropriate for PCO. Since the dissimilarities in PCO are Euclidean, the metric inequality (i.e., the triangle inequality) should be satisfied. For the dissimilarities generated from a truncated normal or log-normal distribution, the metric inequality is no longer guaranteed. Thus, a probabilistic formulation is still absent for PCO. In the current paper we attempt to address this gap by showing that PCO may indeed fit into a maximum likelihood estimation framework.

Tipping and Bishop (1999) proposed a probabilistic PCA (PPCA) model in which PCA is reformulated as a normal latent variable model that is closely related to

Appearing in Proceedings of the 12^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2009, Clearwater Beach, Florida, USA. Volume 5 of JMLR: W&CP 5. Copyright 2009 by the authors.

factor analysis (FA) (Bartholomew and Knott, 1999). Owing to the duality between PCO and PCA, it would be desirable to develop such a latent variable model for PCO so that we have a probabilistic formulation of PCO.

Conventional PCO (or KPCA) does not necessarily require that the original objects (called feature vectors in the machine learning literature) are explicitly available. Instead, it only requires that the (dis)similarities are given. However, the original objects themselves are given in PPCA and FA. As a result, it seems difficult to follow the approach taken for PCA, in which a connection to FA is exploited, in developing a latent variable model for PCO.

Recall that since in PCO or KPCA the dissimilarities (or similarities) are Euclidean (or positive definite), there exists a set of feature vectors such that the Euclidean distances (or the inner products) between them are exactly equal to the dissimilarities (or similarities). This motivates us to treat the feature vectors as *virtual* observations. We will show how this treatment allows us to specify normal latent variable models for PCO as well as KPCA. We refer to these interpretations as probabilistic PCO (PPCO) and probabilistic KPCA (PKPCA), respectively.

In PPCO the principal coordinates (the configurations in a low-dimensional Euclidean space) are treated as the model parameters. As a result, we can use maximum likelihood (ML) to estimate the principal coordinates. We shall see that the estimated results agree with those obtained via the spectral decomposition method. Moreover, the latent variable idea allows us to use the expectation maximization (EM) algorithm for PPCO. Importantly, without the explicit usage of the virtual observations themselves, we can still implement ML and EM procedures using only the available (dis)similarities.

In PKPCA the principal components (the orthonormal bases spanning the low-dimensional subspace) are treated as the model parameters, which are also estimated by ML. Our model differs from the PPCA model of Tipping and Bishop (1999) in that they use non-orthonormal principal components (factor loadings) instead of orthonormal principal components. Although this difference seems to be minor, the difference has important consequences in that the solution of our model agrees with that of conventional PCA, but the solution of PPCA of Tipping and Bishop (1999) does not.

The remainder of the paper is organized as follows. Section 2 describes the original formulations of PCO and KPCA. In Section 3 we propose a normal latent variable model for PCO. A direct ML method and an EM algorithm for parameter estimation are also devised. In Section 4 we propose PKPCA and establish the duality between PPCO and PKPCA. Experimental studies and concluding remarks are given in Sections 5 and 6, respectively. All proofs and derivations are omitted, but they are presented in a long version of this paper.

2 PCO and KPCA

Suppose we are given a set of dissimilarities, $\{\delta_{ij};\ i,j=1,\ldots,n\}$, between n objects. Let $\mathbf{\Delta}=[\delta_{ij}^2]$ be the $n\times n$ dissimilarity matrix. We assume that $\mathbf{\Delta}$ is Euclidean. This implies that there exists a set of n points in a Euclidean space, denoted by $\{\mathbf{f}_i : i = 1,\ldots,n\}$, such that

$$\delta_{ij}^2 = (\mathbf{f}_i-\mathbf{f}_j)'(\mathbf{f}_i-\mathbf{f}_j) = \mathbf{f}_i'\mathbf{f}_i + \mathbf{f}_j'\mathbf{f}_j - 2\mathbf{f}_i'\mathbf{f}_j. \tag{1}$$

We thus have

$$-\frac{1}{2}\mathbf{H}\mathbf{\Delta}\mathbf{H} = \mathbf{H}\mathbf{F}\mathbf{F}'\mathbf{H},$$

where $\mathbf{F}=[\mathbf{f}_1,\ldots,\mathbf{f}_n]'$ and $\mathbf{H}=\mathbf{I}_n-\frac{1}{n}\mathbf{1}_n\mathbf{1}_n'$ (a centering matrix). Here and later, $\mathbf{I}_n$ is the $n\times n$ identity matrix and $\mathbf{1}_n$ is the $n\times 1$ vector of 1's. Thus, the assumption that $\mathbf{\Delta}$ is Euclidean is equivalent to the positive semidefiniteness of $-\frac{1}{2}\mathbf{H}\mathbf{\Delta}\mathbf{H}$.

We now establish a connection of $\mathbf{\Delta}$ to the theory of reproducing kernels. Starting with a set of n p-dimensional input vectors, $\{\mathbf{x}_i : i=1,\ldots,n\}\subset\mathcal{X}\subset\mathbb{R}^p$, we define a positive definite function $K:\mathcal{X}\times\mathcal{X}\to\mathbb{R}$, as a kernel function. There are three common kernel functions that are widely used in practice:

(a) Linear kernel: $K(\mathbf{x}_i,\mathbf{x}_j)=\mathbf{x}_i'\mathbf{x}_j$;

(b) Gaussian kernel: $K(\mathbf{x}_i,\mathbf{x}_j)=\exp(-\|\mathbf{x}_i-\mathbf{x}_j\|^2/\theta)$ with $\theta>0$;

(c) Polynomial kernel: $K(\mathbf{x}_i,\mathbf{x}_j)=(\mathbf{x}_i'\mathbf{x}_j+1)^m$.

From the kernel function and the data, we obtain an $n\times n$ *kernel matrix* $\mathbf{K}=[k_{ij}]$ where $k_{ij}=K(\mathbf{x}_i,\mathbf{x}_j)$. Since the kernel matrix $\mathbf{K}$ is positive semidefinite (p.s.d.), it can be regarded as an inner product matrix and induces an Euclidean matrix, which is then defined as the aforementioned $\mathbf{\Delta}$. It is readily seen that

$$\delta_{ij}^2 = k_{ii}+k_{jj}-2k_{ij}. \tag{2}$$

Comparing (2) with (1), we equate k_{ij} with the inner product between $\mathbf{f}_i$ and $\mathbf{f}_j$, i.e., $k_{ij}=\mathbf{f}_i'\mathbf{f}_j$ and $\mathbf{K}=\mathbf{F}\mathbf{F}'$. The vector $\mathbf{f}_i$ is referred to as the *feature vector* corresponding to $\mathbf{x}_i$. Thus, the kernel technology provides us with an approach to the construction of inner product matrices and dissimilarity (distance) matrices.

PCO (or classical multidimensional scaling) was originally used to construct the coordinates for the points, $\{\mathbf{y}_i : i = 1, \ldots, n\}$, in a Euclidean space, such that

$$(\mathbf{y}_i - \mathbf{y}_j)'(\mathbf{y}_i - \mathbf{y}_j) = d_{ij}^2 \approx \delta_{ij}^2 = (\mathbf{f}_i - \mathbf{f}_j)'(\mathbf{f}_i - \mathbf{f}_j). \quad (3)$$

The focus of this paper is dimensionality reduction: letting $\mathbf{y}_i \in \mathbb{R}^q$ and $\mathbf{f}_i \in \mathbb{R}^r$, q should be less than r. Assuming that the centroid of $\mathbf{y}_i$ is at the origin of $\mathbb{R}^q$, from (3) we obtain $-\frac{1}{2}\mathbf{H}\boldsymbol{\Delta}\mathbf{H} \approx \mathbf{Y}\mathbf{Y}'$ where $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_n]'$.

We now consider using a kernel function in the PCO setting. If we use a linear kernel, we obtain standard PCO in which $\mathbf{f}_i = \mathbf{x}_i$. More generally, we obtain a nonlinear dimensionality version of PCO which is based on a nonlinear mapping from $\mathbf{x}_i$ to $\mathbf{f}_i$.

Given either $\boldsymbol{\Delta}$ or $\mathbf{K}$, we henceforth denote $\mathbf{Q} = -\frac{1}{2}\mathbf{H}\boldsymbol{\Delta}\mathbf{H}'$ or $\mathbf{Q} = \mathbf{H}\mathbf{K}\mathbf{H}'$. PCO attempts to find the configurations of the $\mathbf{y}_i$ by applying eigenvalue decomposition to $\mathbf{Q}$. Let $\boldsymbol{\Psi}_q$ be an $n{\times}q$ matrix whose columns are the top q eigenvectors of $\mathbf{Q}$ and $\boldsymbol{\Gamma}_q$ be a $q{\times}q$ diagonal matrix whose diagonal elements are the top q eigenvalues of $\mathbf{Q}$. The solution of PCO is then given by $\mathbf{Y} = \boldsymbol{\Psi}_q{\boldsymbol{\Gamma}_q}^{\frac{1}{2}}\mathbf{S}$ where $\mathbf{S}$ is an arbitrary orthonormal matrix. It is worth noting that the solution satisfies $\mathbf{Y}'\mathbf{1}_n = \mathbf{0}$. In general, we set $\mathbf{S} = \mathbf{I}_q$ so that $\mathbf{Y}'\mathbf{Y} = \boldsymbol{\Gamma}_q$ is diagonal.

In order to explore the nonlinear structure of $\mathbf{x}_i$ in a low-dimensional representation, KPCA (Schölkopf et al., 1998) uses the sample covariance matrix of $\mathbf{f}_i$, which is given by

$$\begin{aligned}\mathbf{R} &= \frac{1}{n}\sum_{i=1}^{n}(\mathbf{f}_i - \bar{\mathbf{f}})(\mathbf{f}_i - \bar{\mathbf{f}})' \\ &= \frac{1}{n}\mathbf{F}'\mathbf{H}\mathbf{F} = \frac{1}{n}\mathbf{F}'\mathbf{H}\mathbf{H}\mathbf{F}\end{aligned}$$

with $\bar{\mathbf{f}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{f}_i$. The goal of KPCA is to find the first q principal components. Typically, $\mathbf{F}$ is not explicitly available. Since $\mathbf{F}'\mathbf{H}\mathbf{H}\mathbf{F}$ has the same nonzero eigenvalues as $\mathbf{H}\mathbf{F}\mathbf{F}'\mathbf{H} = \mathbf{H}\mathbf{K}\mathbf{H} = \mathbf{Q}$, KPCA works with $\mathbf{Q}$ instead. That is, the first q principal components constitute the matrix $\mathbf{S}{\boldsymbol{\Gamma}_q}^{-1/2}{\boldsymbol{\Psi}_q}'\mathbf{H}\mathbf{F}$. The q-dimensional configuration of $\mathbf{F}$ is then $\mathbf{H}\mathbf{Q}\mathbf{H}\boldsymbol{\Psi}_q{\boldsymbol{\Gamma}_q}^{-1/2}\mathbf{S} = \boldsymbol{\Psi}_q{\boldsymbol{\Gamma}_q}^{1/2}\mathbf{S}$, which is computed without the explicit use of $\mathbf{F}$. This shows the duality between KPCA and PCO.

Gower (1966) referred to $-\frac{1}{2}\mathbf{H}\boldsymbol{\Delta}\mathbf{H}$ or $\mathbf{H}\mathbf{K}\mathbf{H}$ as a Q matrix and the covariance matrix $\frac{1}{n}\mathbf{F}'\mathbf{H}\mathbf{F}$ as an R matrix. The corresponding dimensionality reduction methods are called Q and R techniques, respectively. It is clear that PCO employs the Q technique while KPCA employs the R technique. Since the Q and R techniques are dual to each another, there also exists a duality between PCO and KPCA. The difference is that PCO directly computes the low-dimensional configurations, while KPCA computes the bases that span the low-dimensional subspace and the low-dimensional configurations are the projections of the feature vectors onto this subspace.

3 Probabilistic PCO

Before presenting our probabilistic approach to PCO, we first introduce some notation. We use $\mathbf{A}^{+}$ for the Moore-Penrose inverse of $\mathbf{A}$. The Kronecker product of $\mathbf{A}$ and $\mathbf{B}$ is denoted by $\mathbf{A} \otimes \mathbf{B}$. We employ the notation of Gupta and Nagar (2000) for matrix-variate distributions. Thus, for an $s{\times}t$ random matrix $\mathbf{Z}$, $\mathbf{Z} \sim N_{s,t}(\mathbf{M}, \mathbf{A}{\otimes}\mathbf{B})$ means that $\mathbf{Z}$ follows a matrix-variate normal distribution with mean matrix $\mathbf{M}$ $(s{\times}t)$ and covariance matrix $\mathbf{A}{\otimes}\mathbf{B}$, where $\mathbf{A}$ $(s{\times}s)$ and $\mathbf{B}$ $(t{\times}t)$ are p.s.d.

3.1 Normal Latent Variable Model

We attempt to develop PPCO through a normal latent variable model. However, it is not immediately clear how to derive PPCO from PPCA or FA because for PCO, apart from the Q matrix $\mathbf{Q}$, the $\mathbf{f}_i$ and their dimension r are not explicitly available. In this case, we regard the $\mathbf{f}_i$ as virtual observations. Recall that PCO directly calculates the coordinates of the $\mathbf{y}_i$. Thus we treat the $\mathbf{y}_i$ as parameters of the model that need to be estimated. We thus reformulate PCO as a latent variable model in matrix form:

$$\mathbf{F} = \mathbf{Y}\mathbf{W} + \mathbf{1}_n\mathbf{u}' + \boldsymbol{\Upsilon}, \quad (4)$$

where $\mathbf{u}$ is an $r{\times}1$ mean vector, $\mathbf{W}$ is a $q{\times}r$ latent matrix, and $\boldsymbol{\Upsilon}$ is an $n{\times}r$ error matrix. Furthermore, we assume

$$\mathbf{W} \sim N_{q,r}\left(\mathbf{0},\ (\mathbf{I}_q{\otimes}\mathbf{I}_r)/r\right), \boldsymbol{\Upsilon} \sim N_{n,r}\left(\mathbf{0},\ (\lambda\mathbf{I}_n{\otimes}\mathbf{I}_r)/r\right), \quad (5)$$

where $\lambda > 0$. Thus we get a PPCO model where each row of $\mathbf{Y}$ is just the target coordinate associated with $\mathbf{F}$. Now the difficulty is that both r and $\mathbf{F}$ are usually unknown. Fortunately, we will see a linear algebraic manipulation yields an estimation procedure for the unknown parameters, $\mathbf{Y}$ and λ, that does not explicitly depend on r and $\mathbf{F}$.

As described in Section 2, we assume that the columns of $\mathbf{Y}$ are centered to have mean $\mathbf{0}$, i.e., $\mathbf{Y}'\mathbf{1}_n = \mathbf{0}$. Premultiplying (4) by the centering matrix $\mathbf{H}$, we obtain

$$\mathbf{H}\mathbf{F} = \mathbf{H}\mathbf{Y}\mathbf{W} + \mathbf{H}\boldsymbol{\Upsilon}.$$

It is clearly seen that $\mathbf{Y}'\mathbf{H}\mathbf{1}_n = \mathbf{0}$. Thus, we now treat $\mathbf{HY}$ as the low-dimensional representation of $\mathbf{F}$. For notational simplicity, we still use $\mathbf{Y}$ and $\boldsymbol{\Upsilon}$ to denote $\mathbf{HY}$ and $\mathbf{H}\boldsymbol{\Upsilon}$. Thus, we have

$$\mathbf{HF} = \mathbf{YW} + \boldsymbol{\Upsilon} \tag{6}$$

with $\mathbf{Y}'\mathbf{1}_n = \mathbf{0}$ and $\boldsymbol{\Upsilon} \sim N_{n,r}\left(\mathbf{0},\ \lambda(\mathbf{H}\otimes\mathbf{I}_r)/r\right)$. Since $\mathbf{H}$ is singular, the distribution of $\boldsymbol{\Upsilon}$ degenerates to a singular normal distribution (Mardia et al., 1979).

3.2 Maximum Likelihood Estimates

It follows readily from (6) that

$$\mathbf{HF}|\mathbf{W} \sim N_{n,r}\left(\mathbf{YW},\ \lambda(\mathbf{H}\otimes\mathbf{I}_r)/r\right)$$

and so, by integrating out $\mathbf{W}$, we have

$$\mathbf{HF} \sim N_{n,r}\big(\mathbf{0},\ (\mathbf{YY}'+\lambda\mathbf{H})\otimes\mathbf{I}_r/r\big). \tag{7}$$

Using Bayes' rule, we can compute the conditional distribution of $\mathbf{W}$ given $\mathbf{F}$ as

$$\mathbf{W}|\mathbf{F} \sim N_{q,r}\big(\boldsymbol{\Sigma}^{-1}\mathbf{Y}'\mathbf{HF},\ \lambda(\boldsymbol{\Sigma}^{-1}\otimes\mathbf{I}_r)/r\big), \tag{8}$$

where $\boldsymbol{\Sigma} = \lambda\mathbf{I}_q + \mathbf{Y}'\mathbf{Y}$. Now $\mathbf{HF}$ also follows a singular normal distribution and its p.d.f. is given by

$$\frac{(2\pi)^{\frac{(1-n)r}{2}} r^{\frac{nr}{2}}}{\prod_{i=1}^{n-1} \eta_i^{r/2}} \exp\Big[-\frac{r}{2}\mathsf{tr}\big((\mathbf{YY}'+\lambda\mathbf{H})^{+}\mathbf{HFF}'\mathbf{H}\big)\Big]$$

where η_i, $i = 1,\ldots,n-1$, are the nonzero eigenvalues of $\mathbf{YY}'+\lambda\mathbf{H}$. Note that we use the fact that the rank of $\mathbf{YY}'+\lambda\mathbf{H}$ ($= \mathbf{H}(\mathbf{YY}'+\lambda\mathbf{I}_n)\mathbf{H}$) is equal to $n-1$. This is because $\mathbf{YY}'+\lambda\mathbf{I}_n$ is nonsingular and the rank of $\mathbf{H}$ is $n-1$.

Treating $\mathbf{Y}$ and λ as unknown parameters, we now consider their maximum likelihood (ML) estimates. The corresponding log-likelihood function is given by

$$\varphi = -\frac{r}{2}\sum_{i=1}^{n-1}\log\eta_i - \frac{r}{2}\mathsf{tr}\big((\mathbf{YY}'+\lambda\mathbf{H})^{+}\mathbf{HFF}'\mathbf{H}\big) - \frac{(n-1)r}{2}\log(2\pi) + \frac{nr}{2}\log r.$$

It is easily seen that the maximization of φ with respect to (w.r.t.) $\mathbf{Y}$ and λ is equivalent to the minimization of

$$f(\mathbf{Y},\lambda) = \sum_{i=1}^{n-1}\log\eta_i + \mathsf{tr}\big((\mathbf{YY}'+\lambda\mathbf{H})^{+}\mathbf{Q}\big) \tag{9}$$

w.r.t. $\mathbf{Y}$ and λ. Note that f is independent of both $\mathbf{F}$ and r. Thus, given either $\mathbf{K}$ or $\boldsymbol{\Delta}$, we can estimate $\mathbf{Y}$ and λ without the explicit usage of $\mathbf{F}$ and r. In particular, we have

Theorem 1 *Let $\mathbf{Q} = \mathbf{HKH}$ or $\mathbf{Q} = -\frac{1}{2}\mathbf{H}\boldsymbol{\Delta}\mathbf{H}$ be of rank $d \leq n-1$, and let f be a real-valued loss function defined by (9) where $\mathbf{Y} \in \mathbb{R}^{n\times q}$ subject to $q \leq d$ and $\mathbf{Y}'\mathbf{1}_n = \mathbf{0}$. Then, the minimum of f w.r.t. $\mathbf{Y}$ and λ is obtained when*

$$\widehat{\mathbf{Y}} = \boldsymbol{\Psi}_q(\boldsymbol{\Gamma}_q - \hat{\lambda}\mathbf{I}_q)^{1/2}\mathbf{S} \quad \textit{and} \quad \hat{\lambda} = \frac{1}{n-q-1}\sum_{j=q+1}^{n-1}\gamma_j,$$

where $\gamma_1 \geq \cdots \geq \gamma_q \geq \cdots \gamma_{n-1}$ are the eigenvalues of $\mathbf{Q}$, $\mathbf{S}$ is an arbitrary $q\times q$ orthonormal matrix, $\boldsymbol{\Gamma}_q$ is a $q\times q$ diagonal matrix containing the first q principal (largest) eigenvalues γ_i, and $\boldsymbol{\Psi}_q$ is an $n\times q$ orthogonal matrix in which the q column vectors are the principal eigenvectors corresponding to $\boldsymbol{\Gamma}_q$.

Since $\mathbf{Q}\mathbf{1}_n = 0$, we have $\widehat{\mathbf{Y}}'\mathbf{1}_n = \mathbf{0}$. Furthermore, we have that $\widehat{\mathbf{Y}}'\widehat{\mathbf{Y}} = \boldsymbol{\Gamma}_q - \hat{\lambda}\mathbf{I}_q$ when $\mathbf{S} = \mathbf{I}_q$. Thus, the ML estimate of $\mathbf{Y}$ agrees with that obtained via the eigenvalue decomposition method in Section 2. Moreover, if $\hat{\lambda} \to 0$, then the methods are entirely equivalent. It is worth pointing out that if $\gamma_q > \gamma_{q+1}$, $(\widehat{\mathbf{Y}}, \hat{\lambda})$ is a strict local minimum.

3.3 EM Algorithm

Considering $\mathbf{W}$ as the missing data, $\{\mathbf{W},\mathbf{F}\}$ as the complete data, and $\mathbf{Y}$ and λ as the model parameters, we devise an EM algorithm for our PPCO model.

Given the tth estimates $\mathbf{Y}(t)$ and $\lambda(t)$ of $\mathbf{Y}$ and λ, the EM algorithm updates $\mathbf{Y}$ and λ in the following procedure:

$$\mathbf{Y}(t+1) = \mathbf{QY}(t)\big[\lambda(t)\mathbf{I}_q + \boldsymbol{\Sigma}^{-1}(t)\mathbf{Y}'(t)\mathbf{QY}(t)\big]^{-1}, \tag{10}$$

$$\lambda(t+1) = \frac{1}{(n-1)}\big[\mathsf{tr}(\mathbf{Q}) - \mathsf{tr}\left(\mathbf{Y}(t+1)\boldsymbol{\Sigma}^{-1}(t)\mathbf{Y}'(t)\mathbf{Q}\right)\big], \tag{11}$$

where $\boldsymbol{\Sigma}(t) = \lambda(t)\mathbf{I}_q + \mathbf{Y}'(t)\mathbf{Y}(t)$. The derivation of the EM algorithm is omitted. We see that the iterative equations (10) and (11) also work well only with $\mathbf{Q}$. Note that $\mathbf{1}_n'\mathbf{Y}(t) = \mathbf{0}$ is always satisfied in the EM algorithm. It can be proven that if $\lambda(0) > 0$, the λ calculated via (11) is always positive. Moreover, the EM iterates of $\mathbf{Y}$ and λ converge to the direct ML estimates.

Compared with the direct ML estimate, the EM approach is more efficient when n is large, because the former involves the spectral decomposition of an $n\times n$ matrix, whereas the latter involves the inversion of $q\times q$ matrices. Thus, the EM algorithm provides an efficient numerical method for PCO.

4 Probabilistic KPCA

In this section we present our probabilistic KPCA (PKPCA) model and explore its relationship with PPCO. Our point of departure is Tipping and Bishop (1999), who proposed PPCA based on a latent variable model. We extend this approach to KPCA. In particular, PKPCA expresses the feature vector $\mathbf{f} \in \mathbb{R}^r$ as a linear combination of q principal components (say, $\mathbf{w}_j$) plus noise ($\boldsymbol{\epsilon}$):

$$\mathbf{f} = \sum_{j=1}^{q} \mathbf{w}_j y_j + \mathbf{u} + \boldsymbol{\epsilon} = \mathbf{W}'\mathbf{y} + \mathbf{u} + \boldsymbol{\epsilon},$$
$$\boldsymbol{\epsilon} \sim N(\mathbf{0},\ \lambda \mathbf{I}_q),\ \mathbf{w} \sim N(\mathbf{0},\ \mathbf{I}_q),$$

where $\mathbf{y} = (y_1, \ldots, y_q)' \in \mathbb{R}^q$ with $q < \min\{p, r\}$ is the latent vector and $\mathbf{W}' = [\mathbf{w}_1, \ldots, \mathbf{w}_q]$ $(r \times q)$ is an $r \times q$ basis matrix that relates $\mathbf{f}$ and $\mathbf{y}$. Given an input matrix $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n]'$ and its corresponding feature matrix $\mathbf{F}$, we express PKPCA in matrix form as

$$\begin{aligned} \mathbf{F} &= \mathbf{Y}\mathbf{W} + \mathbf{1}_n\mathbf{u}' + \boldsymbol{\Upsilon}, \qquad (12) \\ \mathbf{Y} &\sim N_{n,q}\left(\mathbf{0},\ \mathbf{I}_n \otimes \mathbf{I}_q\right), \boldsymbol{\Upsilon} \sim N_{n,p}\left(\mathbf{0},\ \lambda(\mathbf{I}_n \otimes \mathbf{I}_p)\right). \end{aligned}$$

In this model, $\mathbf{Y}$ is treated as a latent matrix and $\mathbf{W}$ as a matrix of principal components such that $\mathbf{W}\mathbf{W}' = \mathbf{I}_q$. The q-dimensional configuration of $\mathbf{f}$ (or $\mathbf{x}$) is obtained by using the expectation of $\mathbf{y}$ conditioned on $\mathbf{f}$ and $\mathbf{W}$, i.e., $E(\mathbf{y}|\mathbf{f}, \mathbf{W})$. However, in PPCO, $\mathbf{W}$ is treated as a latent matrix and $\mathbf{Y}$ is just the q-dimensional representation of $\mathbf{X}$. This provides a perspective on the duality between PPCA and PPCO. Note that in the PPCA of Tipping and Bishop (1999), the constraint $\mathbf{W}\mathbf{W}' = \mathbf{I}_q$ is not imposed. We will see shortly that this constraint is indeed necessary.

In the linear kernel case, i.e., $\mathbf{f} = \mathbf{x}$, the ML estimate of $\mathbf{W}$ is $\mathbf{U}_q'$ where $\mathbf{U}_q$ is the matrix formed by the top q eigenvectors of $\mathbf{R}$. We thus obtain the configuration of $\mathbf{x}$ in the q-dimensional space as $E(\mathbf{y}|\mathbf{x}, \mathbf{W}) = (\mathbf{W}\mathbf{W}' + \lambda \mathbf{I}_q)^{-1}\mathbf{W}(\mathbf{x} - \mathbf{u}) = \frac{1}{1+\lambda}\mathbf{U}_q'(\mathbf{x} - \mathbf{u})$. When $\lambda \to 0$, the solution of PPCA is the same to that of the conventional PCA. In the PPCA of Tipping and Bishop (1999), however, the ML estimate of $\mathbf{W}$ is $(\boldsymbol{\Gamma}_q - \lambda \mathbf{I}_q)^{\frac{1}{2}}\mathbf{U}_q'$ where $\boldsymbol{\Gamma}_q$ is the diagonal matrix of the q largest eigenvalues of $n\mathbf{R}$. Hence, $E(\mathbf{y}|\mathbf{x}, \mathbf{W}) = (\mathbf{W}\mathbf{W}' + \lambda \mathbf{I}_q)^{-1}\mathbf{W}(\mathbf{x} - \mathbf{u}) = \boldsymbol{\Gamma}^{-1}(\boldsymbol{\Gamma}_q - \lambda \mathbf{I}_q)^{\frac{1}{2}}\mathbf{U}_q'(\mathbf{x} - \mathbf{u})$, which does not converge to the conventional PCA as $\lambda \to 0$.

The ML estimate of $\mathbf{u}$ is $\hat{\mathbf{u}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{f}_i$. After substituting $\hat{\mathbf{u}}$ for $\mathbf{u}$ into the likelihood, the resulting log-likelihood is referred to as the log concentrated likelihood (Magnus and Neudecker, 1999), which is used for the ML estimation of $\mathbf{W}$ and λ. Typically, the feature vectors $\mathbf{f}$ are assumed unavailable. In this case, the ML estimate of $\mathbf{W}$ is given by $\widehat{\mathbf{W}} = \mathbf{S}\boldsymbol{\Gamma}_q{}^{-1/2}\boldsymbol{\Psi}_q{}'\mathbf{H}\mathbf{F}$ where $\boldsymbol{\Gamma}_q$ is the diagonal matrix of the q largest eigenvalues of $\mathbf{Q}$, $\boldsymbol{\Psi}_q$ is the corresponding eigenvector matrix of $\mathbf{Q}$, and $\mathbf{S}$ is a $q \times q$ orthonormal matrix. We omit the derivation. As we can see, $\widehat{\mathbf{W}}$ does not depend on λ and r but on the feature matrix $\mathbf{F}$. Fortunately, since $[\mathbf{y}|\mathbf{f}, \mathbf{W}] \sim N_q(\boldsymbol{\Sigma}^{-1}\mathbf{W}(\mathbf{f} - \mathbf{u}), \lambda\boldsymbol{\Sigma}^{-1})$ where $\boldsymbol{\Sigma} = \mathbf{W}\mathbf{W}' + \lambda \mathbf{I}_q = (1+\lambda)\mathbf{I}_q$, we are able to compute the q-dimensional configuration of $\mathbf{f}$ via the kernel trick. In particular, we have

$$\begin{aligned} E(\mathbf{y}|\mathbf{f}, \mathbf{W}) &= \frac{1}{1+\hat{\lambda}}\boldsymbol{\Gamma}_q{}^{-1/2}\boldsymbol{\Psi}_q{}'\mathbf{H}\mathbf{F}(\mathbf{f} - \hat{\mathbf{u}}) \\ &= \frac{1}{1+\hat{\lambda}}\boldsymbol{\Gamma}_q{}^{-1/2}\boldsymbol{\Psi}_q{}'\mathbf{H}\mathbf{F}\Big(\mathbf{f} - \frac{1}{n}\mathbf{F}'\mathbf{1}_n\Big) \\ &= \frac{1}{1+\hat{\lambda}}\boldsymbol{\Gamma}_q{}^{-1/2}\boldsymbol{\Psi}_q{}'\Big(\mathbf{k} - \frac{1}{n}\mathbf{K}\mathbf{1}_n\Big), \end{aligned}$$

where $\mathbf{k} = (K(\mathbf{x}, \mathbf{x}_1), \ldots, K(\mathbf{x}, \mathbf{x}_n))'$. In matrix form, we have

$$E(\mathbf{Y}|\mathbf{F}, \mathbf{W}) = \frac{1}{1+\hat{\lambda}}\mathbf{Q}\boldsymbol{\Psi}_q\boldsymbol{\Gamma}_q{}^{-\frac{1}{2}} = \frac{1}{1+\hat{\lambda}}\boldsymbol{\Psi}_q\boldsymbol{\Gamma}_q{}^{\frac{1}{2}}.$$

This shows that when $\hat{\lambda} \to 0$, the solution of PKPCA is the same as that of PPCO as well as the conventional KPCA and PCO. Since $\frac{1}{1+\hat{\lambda}}$ is a constant, we use $(1+\hat{\lambda})E(\mathbf{y}|\mathbf{f}, \mathbf{W})$ as the low-dimensional configuration of $\mathbf{f}$. As a result, the solution of PPKCA fully agrees with that of the conventional KPCA as well as the conventional PCO. Moreover, we can ignore the ML estimate of λ for our purpose of dimensionality reduction. Unfortunately, it is not feasible to obtain an EM algorithm for PKPCA when the feature vectors are not explicitly available.

5 Experiments

In this paper our principal focus has been to provide a probabilistic perspective from which to view PCO and KPCA. Although the direct ML estimation approaches to these models give the same solutions as their conventional counterparts, our analysis has also provided an EM algorithm for PPCO, and it is of interest to compare the performance of the EM algorithm with the direct ML method.

As we see in Section 3.3, the EM algorithm is more efficient than the ML method when n is large. Moreover, the solution of the EM algorithm converges to that of the ML estimate. As we know, EM algorithms rely on initial values. We use conventional PCA to initialize the EM algorithm for our PPCO. All algorithms have been implemented in Matlab on a Pentium 4 computer with a 2.00GHz CPU and 1.96GB of RAM.

Table 1: Summary of the benchmark datasets: n—# of samples; p—# of variates; c—# of classes; β—parameter in the Gaussian kernel K.

	Iris	Oil	Letter	Segmen	NIST
n	150	1000	1978	2310	3823
p	4	12	16	19	256
c	3	2	10	7	10
β	2	0.2	100	10000	1000

5.1 Convergence Analysis

Our first experimental analysis is based on two data sets: the iris data set and the multi-phase oil flow data set studied by Bishop and James (1993), which consists of 12 features and three classes, *stratified*, *annular* and *homogeneous*, corresponding to the phases of flow in an oil pipeline.

We adopt the Gaussian kernel $K(\mathbf{x}_i, \mathbf{x}_j) = \frac{1}{n}\exp(-\|\mathbf{x}_i - \mathbf{x}_j\|^2/\beta)$ with $\beta = 2.0$ for the iris data and $\beta = 0.2$ for the oil flow data. For PPCO, we implement the direct ML estimation and the EM algorithm. The EM algorithm is initialized using conventional linear PCA and the maximum number of iterations is 100. Figure 1 depicts the two-dimensional ($q = 2$) principal coordinates of the two data sets using these two algorithms. We can see that the two algorithms give essentially the same results up to a rotation transformation.

5.2 Estimating the Largest Eigenvalue of Q

Let $q = 1$. Since $\mathbf{S} = \pm 1$, we always have $\hat{\mathbf{y}}'\hat{\mathbf{y}} = \gamma_1 - \hat{\lambda}$ where $\hat{\mathbf{y}}$ ($n{\times}1$) and $\hat{\lambda}$ are the ML estimates of PPCO, and γ_1 is the largest eigenvalue of $\mathbf{Q} = \mathbf{HKH}$ (see Theorem 1). Note that we can compute the largest eigenvalue of $\mathbf{Q}$ via $\gamma_1 = \hat{\mathbf{y}}'\hat{\mathbf{y}} + \hat{\lambda}$ and its corresponding principal eigenvector as $\boldsymbol{\psi}_1 = \frac{1}{\sqrt{\gamma_1 - \hat{\lambda}}}\hat{\mathbf{y}}$. Since the solution of the EM algorithm given in (10) and (11) converges to that of the ML estimate, it can be used to iteratively estimate γ_1 and $\boldsymbol{\psi}_1$ by $\gamma_1(t) = \mathbf{y}'(t)\mathbf{y}(t) + \lambda(t)$ and $\boldsymbol{\psi}_1(t) = \frac{1}{\sqrt{\gamma_1(t) - \lambda(t)}}\mathbf{y}(t)$. Interestingly, this EM algorithm bears resemblance to the power method (Golub and Loan, 1996) (see also p. 462 in Anderson (1984)).

Table 2 lists the value of γ_1 and the ML estimates of λ on the iris and oil flow datasets. We also report the EM estimates of λ and $\mathbf{y}$, and hence that of γ_1. From Figure 2, we see that $\gamma_1(t)$ converges to the true value, while $\lambda(t)$ converges to the ML estimate. Moreover, the convergence of λ takes only several iterations.

Table 2: Results on the iris and oil flow datasets: γ_1—largest eigenvalue of $\mathbf{Q}$; $\hat{\lambda}$—ML estimate of λ; $\gamma_1(0)$—initial value of γ_1 in the EM iteration; $\lambda(0)$—initial value of λ in the EM iteration.

	γ_1	$\gamma_1(0)$	$\hat{\lambda}$	$\lambda(0)$
Iris	0.2799	3.7321	0.0029	0.2178
Oil flow	0.0437	0.1004	0.0009	0.0145

Table 3: CPU time (s) of running PPCO with the ML estimate and EM iteration.

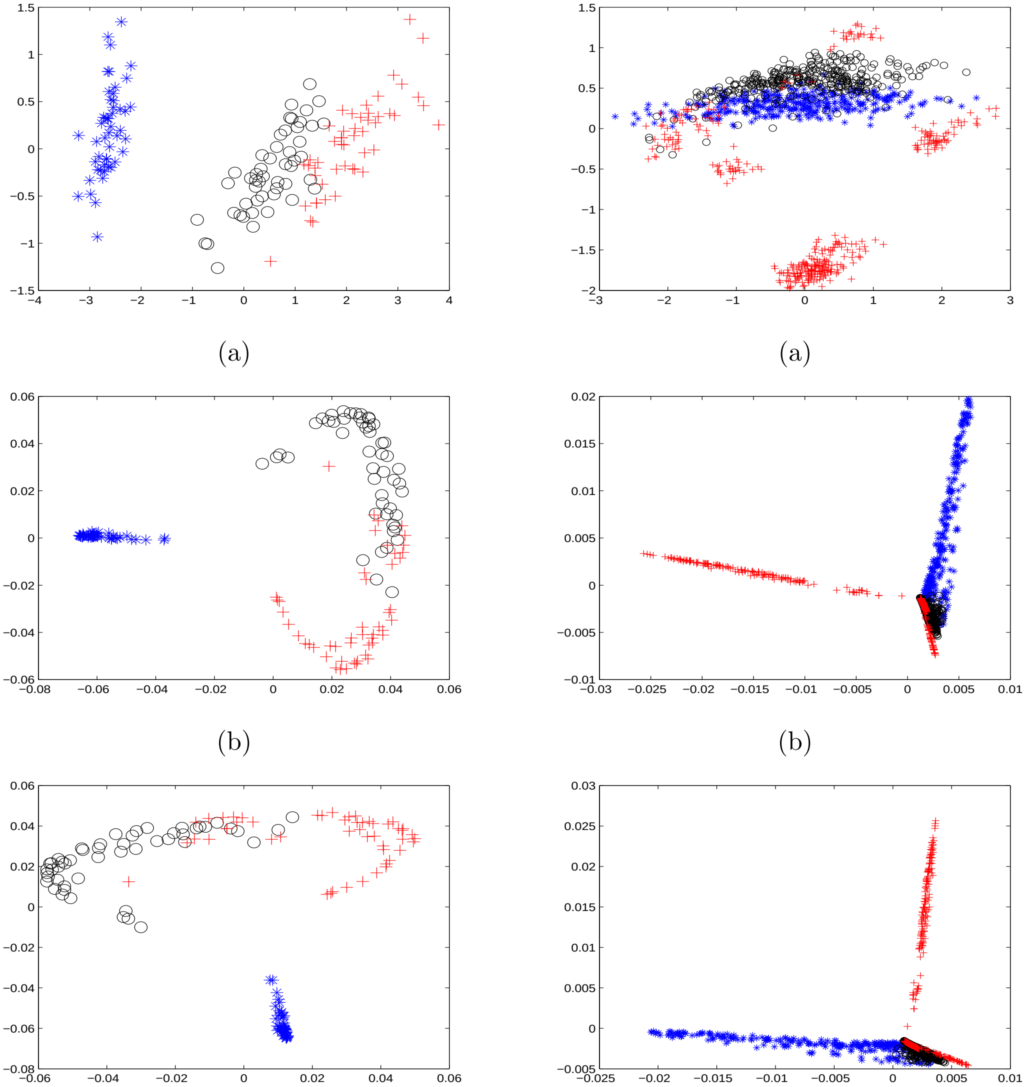
	Iris	Oil	Letter	Segmen	NIST
ML	0.0469	7.6250	59.8438	70.9375	419.0469
EM	0.2344	4.3906	17.7656	23.1875	65.6563

5.3 Performance Analysis

In this subsection we further investigate the performance of the EM algorithm for PPCO with the *iris* and *oil flow* data as well as three publicly available datasets from the UCI machine learning repository (the *NIST* optical handwritten digit data, the *letter* data and the *image segmentation* data). The NIST dataset contains the handwritten digits $0 - 9$, where each instance consists of 16×16 pixels and the digits are treated as classes. The letter dataset consists of images of the letters "A" to "Z." In our experiments we selected the first 10 letters with 195, 199, 182, 207, 203, 210, 226, 196, 188 and 172 instances, respectively. The image segmentation data consists of seven types of images: "brickface," "sky," "foliage," "cement," "window," "path", and "grass." Table 1 gives a summary of these datasets.

In Table 3 we report the CPU time of implementing the direct ML estimate and EM iteration of PCO. The results are based on $q = 2$ and $T = 100$ (the maximum number of iterations of the EM algorithm). The computational complexity of the ML estimate is $O(n^3)$, while that of the EM iteration is $O(Tnq^2)$. We thus see that the EM iteration is more efficient than the ML estimate for large values of n.

Given a kernel K, PKPCA with the ML estimate has the same computational complexity as PPCO with the ML estimate. In the general case it is not possible to devise an EM algorithm for PKPCA because the feature vectors $\mathbf{f}$ are not explicitly available. However, PKPCA directly estimates the principal components other than the principal coordinates. When n is large, in order to reduce computation, we can implement PKPCA on a small-size dataset. This motivates us to devise an initialization method for the EM algorithm for PPCO in the case that both n and p are large.



(a) (a)

(b) (b)

(c) iris data

(c) multi-phase oil flow data

Figure 1: Two-dimensional configurations for: (a) conventional PCA; (b) PPCO with the direct ML estimate; (c) PPCO with the EM estimate.

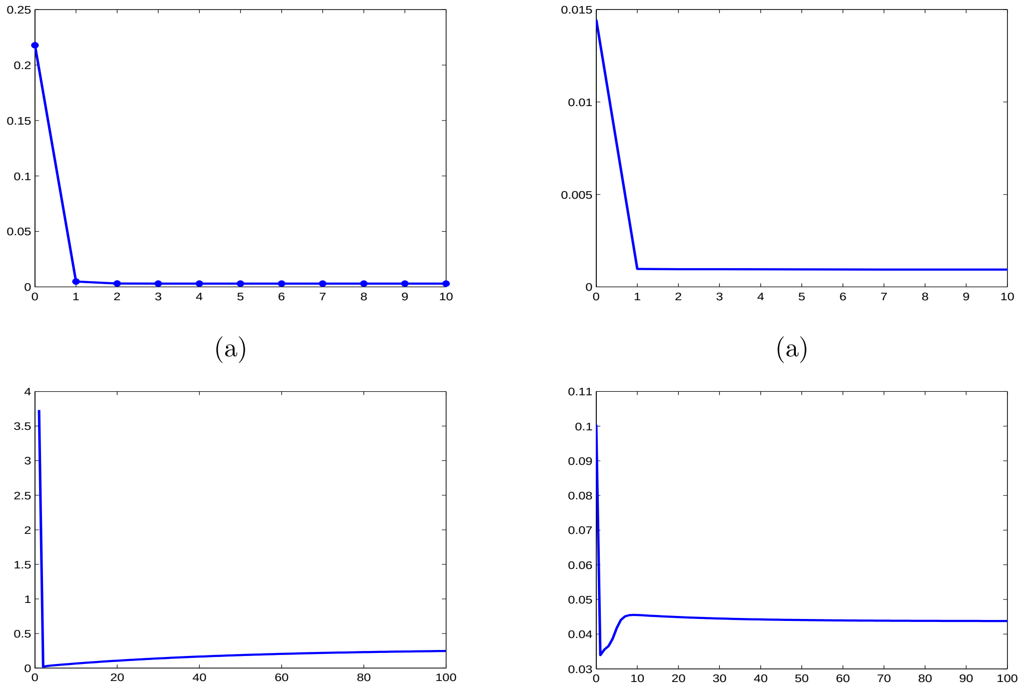
6 Conclusion

In this paper we have studied normal latent variable models for PCO and KPCA. This has yielded algorithms that we refer to as PPCO and PKPCA. Moreover, we have further explored the duality between PCO and KPCA based on their probabilistic formulations. Our work demonstrates that PCO and KPCA can be derived within an ML estimation framework.

These normal latent variable models are closely related to factor analysis. However, we impose some new constraints on the unknown parameter (factor loading) matrix in these models. In particular, we impose the constraint $\mathbf{1}_n'\mathbf{Y} = \mathbf{0}$ on the parameter matrix $\mathbf{Y}$ in PPCO, and the constraint $\mathbf{W}\mathbf{W}' = \mathbf{I}_q$ on the parameter matrix $\mathbf{W}$ in PKPCA. Under these constraints, we have shown that there is still a closed-form solution for the ML estimate of the parameter matrix.

Acknowledgement

Zhihua Zhang has been supported in part by program for Changjiang Scholars and Innovative Research Team in University (IRT0652, PCSIRT), China. Michael Jordan acknowledges support from NSF Grant 0509559 and grants from Google and Microsoft Research.



(a) (a)

(b)
iris data

(b)
multi-phase oil flow data

Figure 2: The largest eigenvalue γ_1 and the noise variance λ vs. the iteration count: (a) the EM estimate of λ; (b) the EM estimate of γ_1.

References

Anderson, T. W. (1984). *An Introduction to Multivariate Statistical Analysis* (Second ed.). New York: John Wiley & Sons.

Bartholomew, D. J. and M. Knott (1999). *Latent Variable Models and Factor Analysis* (second ed.). London: Arnold.

Bishop, C. M. and G. D. James (1993). Analysis of multiphase flows using dual-energy gamma densitometry and neural networks. *Nuclear Instruments and Methods in Physics Research A327*, 580–593.

Borg, I. and P. Groenen (1997). *Modern Multidimensional Scaling.* New York: Springer-Verlag.

Golub, G. H. and C. F. V. Loan (1996). *Matrix Computations* (Third ed.). Baltimore: The Johns Hopkins University Press.

Gower, J. C. (1966). Some distance properties of latent root and vector methods used in multivariate data analysis. *Biometrika 53*, 315–328.

Groenen, P. J. F., R. Mathar, and W. J. Heiser (1995). The majorization approach to multidimensional scaling for Minkowski distance. *Journal of Classification 12*, 3–19.

Gupta, A. and D. Nagar (2000). *Matrix Variate Distributions.* Chapman & Hall/CRC.

Jolliffe, I. (2002). *Principal Component Analysis* (Second Edition ed.). New York: Springer.

Magnus, J. R. and H. Neudecker (1999). *Matrix Calculus with Applications in Statistics and Econometric* (Revised Edition ed.). New York: John Wiley & Sons.

Mardia, K. V., J. T. Kent, and J. M. Bibby (1979). *Multivariate Analysis.* New York: Academic Press.

Oh, M.-H. and A. E. Raftery (2001). Bayesian multidimensional scaling and choice of dimension. *Journal of the American Statistical Association 96*(455), 1031–1044.

Ramsay, J. O. (1982). Some statistical approaches to multidimensional scaling data. *Journal of the Royal Statistical Society Series A 145*, 285–312.

Schölkopf, B., A. Smola, and K.-R. Müller (1998). Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation 10*, 1299–1319.

Tipping, M. E. and C. M. Bishop (1999). Probabilistic principal component analysis. *Journal of the Royal Statistical Society, Series B 61*(3), 611–622.

Williams, C. K. I. (2001). On a connection between kernel PCA and metric multidimensional scaling. In *Advances in Neural Information Processing Systems 13*, Cambridge, MA. MIT Press.