Probabilistic Kernel Principal Component Analysis

Zhihua Zhang, Gang Wang, Dit-Yan Yeung and James T. Kwok
Department of Computer Science
The Hong Kong University of Science and Technology
Clear Water Bay, Kowloon, Hong Kong
{zhzhang,wanggang,dyyeung,jamesk}@cs.ust.hk

HKUST-CS04-03, June 2004

Abstract

Kernel principal component analysis (KPCA), as a kernelized version of principal component analysis, is becoming a ubiquitous nonlinear method applied to various data analysis and processing tasks. In this paper, we propose a probabilistic reformulation of KPCA by drawing an equivalence between two notions, that the reproducing kernel follows a *Wishart process* and that there exists a kernel-induced feature space in which the dimensions of the feature vectors are mutually independent *Gaussian processes*. This thus leads to a latent variable model in which maximum likelihood estimation can be used to learn the model parameters and the kernel trick is still applicable. We also present a matrix-variate perspective of probabilistic KPCA and show that the EM algorithm can be applied. Since KPCA essentially performs principal coordinate analysis, its probabilistic version may also be called *probabilistic principal coordinate analysis*.

1 Introduction

A common problem encountered in many machine learning and pattern recognition tasks is the high dimensionality of data, often referred to as the curse of dimensionality problem. Principal component analysis (PCA) [4] is a commonly used linear technique for dimensionality reduction, data compression and visualization. The original derivation of PCA is based on matrix-algebraic theory. Recently, Tipping and Bishop [8] proposed a probabilistic version of PCA, or PPCA. Not only can this probabilistic reformulation reveal the relationship of PCA with other latent variable models such as factor analysis (FA) and independent component analysis (ICA), it can also bring about an approach to various Bayesian treatments of PCA.

Recently, nonlinear dimensionality reduction techniques have become increasingly popular for data visualization and other applications. Kernel PCA (KPCA) [6] is among the most representative techniques proposed. Its original derivation is also based on matrix-algebraic theory. Analogous to the probabilistic reformulation of PCA, we are motivated to devise a probabilistic version of KPCA. As is well-known, in order to implement KPCA in the feature space which is typically of very high or even infinite dimensionality, the so-called *kernel trick* is required so that explicit use of the feature vectors is not needed. As it turns out, it is difficult to follow the line of devising probabilistic PCA in the probabilistic reformulation of KPCA.

Fortunately, our recent work [11] sheds some light on this problem. We treat reproducing kernels as stochastic processes and propose a novel notion called *Wishart processes* for modeling kernels, establishing a new connection between reproducing kernels and Gaussian processes. Specifically, we show that a reproducing kernel follows a Wishart process if and only if there exists a kernel-induced feature space in which the dimensions of the feature vectors are mutually independent Gaussian processes. In this paper, based on this new statistical view, we devise a probabilistic version of KPCA. Although our point of departure is from the feature space, the kernel trick can help us to implement probabilistic KPCA efficiently. Note that KPCA is essentially classical multidimensional scaling (MDS) or called principal coordinate analysis (PCO) [2], as revealed by the equivalence relationship discussed in [9]. Thus, our probabilistic version of KPCA may also be referred to as *probabilistic principal coordinate analysis* (PPCO).

Throughout this paper, matrices and vectors are denoted by boldface uppercase and lowercase letters, respectively. Let $\mathbf{I}_m$ and $\mathbf{1}_m$ denote the $m\times m$ identity matrix and the $m\times 1$ vector of ones, respectively. Let $\mathbf{A} = [a_{ij}]$ be an $m\times n$ matrix. We denote the transpose of $\mathbf{A}$ by $\mathbf{A}'$. Moreover, when $m = n$, we denote the trace of $\mathbf{A}$ by $\mathrm{tr}(\mathbf{A})$, its determinant by $|\mathbf{A}|$, and its inverse (if exists) by $\mathbf{A}^{-1}$. In addition, the Kronecker product of $\mathbf{A}$ and $\mathbf{B}$ is denoted by $\mathbf{A}\otimes\mathbf{B}$. To simplify our presentation, we will employ the notation of [3]. Thus, for an $s\times t$ random matrix $\mathbf{Y}$, $\mathbf{Y}\sim\mathcal{N}_{s,t}(\mathbf{M},\mathbf{A}\otimes\mathbf{B})$ means that $\mathbf{Y}$ follows a matrix-variate normal distribution where the mean matrix is $\mathbf{M}$, the covariance matrix of each column of $\mathbf{Y}$ is $\mathbf{A}$ and the covariance matrix of each row of $\mathbf{Y}$ is $\mathbf{B}$. Also, for an $n\times n$ random matrix $\mathbf{W}$, $\mathbf{W}\sim\mathcal{W}_n(r,\mathbf{\Sigma})$ means that $\mathbf{W}$ follows a *Wishart distribution* with r degrees of freedom and a positive definite parameter matrix $\mathbf{\Sigma}$, while $\mathbf{W}\sim\mathcal{SW}_n(r,\mathbf{\Sigma})$ means that $\mathbf{W}$ follows a *singular Wishart distribution* [7].

2 Some Theoretical Background

Given a reproducing kernel $K : \mathcal{I}\times\mathcal{I}\to\mathbb{R}$, there always exists a corresponding mapping $F : \mathcal{I}\to\mathcal{F}$. In general, we may have infinitely many random variables. In practice, however, we typically observe only a finite number of them which can

be indexed by a finite index set. Here and later, we consider a finite index set $\mathcal{I} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$. As a result, we have an $n \times n$ kernel matrix $\mathbf{K} = [K(\mathbf{x}_i, \mathbf{x}_j)]$.

2.1 Kernel Principal Component Analysis

Suppose that the (finite) dimensionality of the feature space $\mathcal{F}$ is r, with $1 \leq r < \infty$. For an input vector $\mathbf{x} \in \mathcal{I}$, we can express $F(\mathbf{x})$ as an r-dimensional vector $(F_1(\mathbf{x}), \ldots, F_r(\mathbf{x}))'$. Let us define a matrix $\mathbf{F}$ as

$$\mathbf{F} = \begin{bmatrix} F_1(\mathbf{x}_1) & F_2(\mathbf{x}_1) & \ldots & F_r(\mathbf{x}_1) \\ F_1(\mathbf{x}_2) & F_2(\mathbf{x}_2) & \ldots & F_r(\mathbf{x}_2) \\ \vdots & \vdots & \ddots & \vdots \\ F_1(\mathbf{x}_n) & F_2(\mathbf{x}_n) & \ldots & F_r(\mathbf{x}_n) \end{bmatrix}. \tag{1}$$

Apparently, $\mathbf{K} = \mathbf{F}\mathbf{F}'$. Since $F_j(\mathbf{x})$ $(j = 1, \ldots, r)$ represents the jth coordinate of the feature vector $F(\mathbf{x})$, $F_j(\mathbf{x})$ is itself a function from $\mathcal{I}$ to $\mathbb{R}$. Let $\mathbf{f}_i$ and $\mathbf{g}_j$ denote the ith row vector and jth column vector of $\mathbf{F}$, respectively, i.e., $\mathbf{f}_i = (F_1(\mathbf{x}_i), \ldots, F_r(\mathbf{x}_i))'$ $(i = 1, \ldots, n)$ and $\mathbf{g}_j = (F_j(\mathbf{x}_1), \ldots, F_j(\mathbf{x}_n))'$ $(j = 1, \ldots, r)$. We refer to $\mathbf{f}_i$ as the ith feature vector and $\mathbf{g}_j$ as the jth feature dimension.

In the feature space, PCA works with the covariance matrix of the feature vectors defined as

$$\mathbf{C} = \frac{1}{n} \sum_{i=1}^{n} (\mathbf{f}_i - \bar{\mathbf{f}})(\mathbf{f}_i - \bar{\mathbf{f}})', \tag{2}$$

where $\bar{\mathbf{f}} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{f}_i$. Equivalently, we can express (2) in a matrix form as

$$\mathbf{C} = \frac{1}{n} \mathbf{F}'\mathbf{H}\mathbf{F} = \frac{1}{n} \mathbf{F}'\mathbf{H}\mathbf{H}\mathbf{F}, \tag{3}$$

where $\mathbf{H} = \mathbf{I} - \frac{1}{n} \mathbf{1}_n \mathbf{1}_n{}'$. Since the matrix $\mathbf{F}'\mathbf{H}\mathbf{H}\mathbf{F}$ has the same nonzero eigenvalues as $\mathbf{H}\mathbf{F}\mathbf{F}'\mathbf{H} = \mathbf{H}\mathbf{K}\mathbf{H}$, the kernel trick makes KPCA [6] work over $\frac{1}{n}\mathbf{H}\mathbf{K}\mathbf{H}$ instead of $\mathbf{C}$. Thus, KPCA is equivalent to performing PCO in the input space [9].

2.2 Wishart Processes

In our recent work [11], we regard the kernel K itself as a stochastic process and define it as a Wishart process.

Definition 1 *Let $\{K(\mathbf{x}, \mathbf{y}) \,;\, \mathbf{x}, \mathbf{y} \in \mathcal{I}\}$ be a stochastic process parameterized by $\mathbf{x}, \mathbf{y} \in \mathcal{I}$. $\{K(\mathbf{x}, \mathbf{y}) \,;\, \mathbf{x}, \mathbf{y} \in \mathcal{I}\}$ is said to be a* Wishart process *if for any $n \in \mathbb{N}$ and $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\} \subseteq \mathcal{I}$, the $n \times n$ random matrix $\mathbf{K} = [K(\mathbf{x}_i, \mathbf{x}_j)]$ follows a* Wishart distribution *or a* singular Wishart distribution.

Theorem 1 *Let $r < \infty$ and $\mathbf{g}_1, \mathbf{g}_2, \ldots, \mathbf{g}_r$ be independent random vectors normally distributed as $\mathcal{N}(\mathbf{0},\ \mathbf{\Sigma})$. Then we have $\mathbf{K} \sim \mathcal{W}_n(r,\ \mathbf{\Sigma})$ when $r \geq n$ and $\mathbf{K} \sim \mathcal{SW}_n(r,\ \mathbf{\Sigma})$ when $r < n$. Conversely, let $\mathbf{K} \sim \mathcal{W}_n(r,\ \mathbf{\Sigma})$ or $\mathbf{K} \sim \mathcal{SW}_n(r,\ \mathbf{\Sigma})$ with $r < \infty$ being a positive integer. Then, there exist r mutually independent n-dimensional vectors $\{\mathbf{g}_j\}_{j=1}^{r}$ following $\mathcal{N}(\mathbf{0},\ \mathbf{\Sigma})$.*

The important implication of Theorem 1 is that we can define a probabilistic generative model of the kernel matrix $\mathbf{K}$, i.e., $\mathbf{K}$ is a Wishart random matrix $\mathcal{W}_n(r,\ \mathbf{\Sigma})$. From Theorem 1, it is easy to see that $\{K(\mathbf{x}, \mathbf{y}); \mathbf{x}, \mathbf{y} \in \mathcal{I}\}$ is a Wishart process if and only if $\{F_j(\mathbf{x}); \mathbf{x} \in \mathcal{I}\}_{j=1}^{r}$ are r mutually independent Gaussian processes. In other words, there exist r mutually independent n-dimensional vectors $\{\mathbf{g}_j\}_{j=1}^{r}$

following $\mathcal{N}(\mathbf{0},\ \mathbf{\Sigma})$. However, there is no need to assume that the feature vectors $\mathbf{f}_i$ $(i = 1, \ldots, n)$ are mutually independent. This motivates us to develop PPCA in the feature space, which will work with $\mathbf{g}_j$'s instead of $\mathbf{f}_i$'s.

3 Probabilistic Kernel Principal Component Analysis

Suppose that the feature dimension $\mathbf{g} \in \mathbb{R}^n$ is expressed as a linear combination of m basis vectors ($\mathbf{b}_j$'s) in $\mathbb{R}^n$ plus some noise term $\boldsymbol{\epsilon}$:

$$\begin{aligned} \mathbf{g} &= \mathbf{B}\mathbf{w} + u\mathbf{1}_n + \boldsymbol{\epsilon}, & (4)\\ \boldsymbol{\epsilon} &\sim \mathcal{N}(\mathbf{0},\ \mathbf{V}), & (5) \end{aligned}$$

where $\mathbf{w} = (w_1, \ldots, w_m)' \in \mathbb{R}^m$ with $m < n$, $\mathbf{B} = [\mathbf{b}_1, \ldots, \mathbf{b}_m]$ is an $n \times m$ matrix that relates the two sets of variables $\mathbf{g}$ and $\mathbf{w}$, and u is a scalar that allows the model to have non-zero mean.[1] Similar to standard PPCA, the noise variance matrix $\mathbf{V}$ is hyperspherical and the latent variables $w_1, \ldots, w_m$ are independent Gaussians with mean zero and variance n/r, i.e.,

$$\mathbf{V} = n\sigma^2\mathbf{I}_n/r \quad \text{and} \quad \mathbf{w} \sim \mathcal{N}(\mathbf{0},\ n\mathbf{I}_m/r). \tag{6}$$

Note that when $n = r$, our model becomes the standard PPCA over the feature dimensions $\mathbf{g}$'s. The difficulty is that r, the dimensionality of the feature space, is usually unknown. However, we will see that our model can make use of the kernel trick to yield an estimation procedure for the unknown parameters $\mathbf{B}$ and σ^2 that does not depend on r.

From (4) and (5), we can obtain the conditional probability of $\mathbf{g}$ as

$$\mathbf{g} \mid \mathbf{w}, \mathbf{B}, u, \sigma^2 \sim \mathcal{N}(\mathbf{B}\mathbf{w} + u\mathbf{1}_n,\ n\sigma^2\mathbf{I}_n/r),$$

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

$$\mathbf{g} \mid \mathbf{B}, u, \sigma^2 \sim \mathcal{N}(u\mathbf{1}_n,\ n\mathbf{C}/r),$$

with $\mathbf{C} = \mathbf{B}\mathbf{B}' + \sigma^2\mathbf{I}_n$. Now, given r independent $\mathbf{g}_j$'s from $\mathcal{N}(u_j\mathbf{1}_n,\ n\mathbf{C}/r)$, their joint distribution can be given by

$$\begin{aligned} p(\mathbf{g}_1, \ldots, \mathbf{g}_r \mid \mathbf{B}, u_1, \ldots, u_r, \sigma^2) &= \prod_{j=1}^{r} p(\mathbf{g}_j \mid \mathbf{B}, u_j, \sigma^2) \\ &= \Big(\frac{r}{2n\pi}\Big)^{\frac{rn}{2}} e^{-\frac{r}{2}\mathrm{tr}\left(\mathbf{C}^{-1}\mathbf{S}\right)} |\mathbf{C}|^{-\frac{r}{2}}, \qquad (7) \end{aligned}$$

where

$$\mathbf{S} = \frac{1}{n}\sum_{j=1}^{r}(\mathbf{g}_j - u_j\mathbf{1}_n)(\mathbf{g}_j - u_j\mathbf{1}_n)'.$$

Denoting $\mathbf{u} = (u_1, \ldots, u_r)'$ and working with the $n \times r$ data matrix $\mathbf{F} = [\mathbf{g}_1, \ldots, \mathbf{g}_r]$ in the feature space, we can rewrite (7) as

$$p(\mathbf{F} \mid \mathbf{B}, \mathbf{u}, \sigma^2) = \frac{\exp\{-\frac{r}{2n}\mathrm{tr}\big((\mathbf{B}\mathbf{B}' + \sigma^2\mathbf{I}_n)^{-1}(\mathbf{F} - \mathbf{1}_n\mathbf{u}')(\mathbf{F}' - \mathbf{u}\mathbf{1}_n')\big)\}}{(2n\pi)^{\frac{rn}{2}} r^{-\frac{rn}{2}} |\mathbf{B}\mathbf{B}' + \sigma^2\mathbf{I}_n|^{\frac{r}{2}}}. \tag{8}$$

The corresponding log-likelihood function (log posterior probability to be exact) is

$$L = -\frac{r}{2}\log|\mathbf{B}\mathbf{B}' + \sigma^2\mathbf{I}_n| - \frac{r}{2n}\mathrm{tr}\big((\mathbf{B}\mathbf{B}' + \sigma^2\mathbf{I}_n)^{-1}(\mathbf{F} - \mathbf{1}_n\mathbf{u}')(\mathbf{F}' - \mathbf{u}\mathbf{1}_n')\big) \tag{9}$$

[1] Since $\mathbf{g}$ is a feature dimension, its elements should share a common scale. However, for different $\mathbf{g}$'s, the u's should be different. We will see that this setting leads exactly to KPCA.

with some constant terms omitted. As $\mathbf{F}' = [\mathbf{g}_1, \ldots, \mathbf{g}_r]' = [\mathbf{f}_1, \ldots, \mathbf{f}_n]$ $(r \times n)$, it follows immediately from the results of [8] that we can obtain the maximum likelihood (ML) estimate of $\mathbf{u}$ as

$$\widehat{\mathbf{u}} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{f}_i. \tag{10}$$

In this case, we have

$$(\mathbf{F} - \mathbf{1}_n \hat{\mathbf{u}}')(\mathbf{F}' - \hat{\mathbf{u}} \mathbf{1}_n') = \mathbf{H}\mathbf{F}\mathbf{F}'\mathbf{H} = \mathbf{H}\mathbf{K}\mathbf{H}.$$

On the other hand, the gradient of L w.r.t. $\mathbf{B}$ is

$$\frac{\partial L}{\partial \mathbf{B}} = -r(\mathbf{B}\mathbf{B}' + \sigma^2 \mathbf{I}_n)^{-1} \left(\mathbf{I}_n - \frac{1}{n} (\mathbf{F} - \mathbf{1}_n \mathbf{u}')(\mathbf{F}' - \mathbf{u}\mathbf{1}_n')(\mathbf{B}\mathbf{B}' + \sigma^2 \mathbf{I}_n)^{-1} \right) \mathbf{B}.$$

Given $\mathbf{u} = \hat{\mathbf{u}}$ and letting $\frac{\partial L}{\partial \mathbf{B}} = 0$, we have the following stationary equation

$$\frac{1}{n} \mathbf{H}\mathbf{K}\mathbf{H}(\mathbf{B}\mathbf{B}' + \sigma^2 \mathbf{I}_n)^{-1} \mathbf{B} = \mathbf{B}. \tag{11}$$

Since the above equation involves only the kernel matrix $\mathbf{K}$ but not the feature dimensionality r, the kernel trick can help us to work with the results given in [8]. For example, we have the ML estimates of $\mathbf{B}$ and σ^2 as

$$\begin{aligned} \widehat{\sigma}^2 &= \frac{1}{n-m} \sum_{j=m+1}^{n} \lambda_j, \\ \widehat{\mathbf{B}} &= \mathbf{U}(\mathbf{\Lambda} - \widehat{\sigma}^2 \mathbf{I}_n)^{1/2} \mathbf{R}, \end{aligned}$$

where $\mathbf{U}$ is an $n \times m$ orthogonal matrix in which the q column vectors are the principal eigenvectors of $\frac{1}{n}\mathbf{H}\mathbf{K}\mathbf{H}$, $\mathbf{\Lambda}$ is an $m \times m$ diagonal matrix containing the corresponding eigenvalues $\lambda_1, \ldots, \lambda_m$, and $\mathbf{R}$ is an arbitrary $m \times m$ orthogonal matrix.

Note that in our probabilistic KPCA, each row of $\mathbf{B}$ is just our target coordinate in a low-dimensional latent space associated with the corresponding feature vector. This agrees with standard KPCA [6] except for the rotational transformation $\mathbf{R}$. However, this differs from standard PPCA [8] where the target coordinate is the expectation of the latent vector conditioned on the corresponding input vector. Analogous to the relationship between PPCA and FA, we may also devise a probabilistic version of kernel FA by regarding the noise variance matrix $\mathbf{V}$ as a general diagonal matrix.

4 A Matrix-Variate Perspective of Probabilistic KPCA and EM Estimates

In fact, (8) shows that $\mathbf{F} \mid \mathbf{u}, \mathbf{B}, \sigma^2 \sim \mathcal{N}_{n,r}\big(\mathbf{1}_n \mathbf{u}', \; n(\mathbf{B}\mathbf{B}' + \sigma^2 \mathbf{I}_n) \otimes \mathbf{I}_r / r\big)$. This motivates us to reformulate the probabilistic KPCA model discussed above with matrix-variate distributions [3]. Specifically, we assume

$$\mathbf{F} = \mathbf{B}\mathbf{W} + \mathbf{1}_n \mathbf{u}' + \mathbf{E}, \tag{12}$$

where $\mathbf{u}$ is $r \times 1$ and $\mathbf{B}$ is $n \times m$ with $m < n$,

$$\mathbf{W} \sim \mathcal{N}_{m,r}\left(\mathbf{0}, \; n(\mathbf{I}_m \otimes \mathbf{I}_r)/r\right) \quad \text{and} \quad \mathbf{E} \sim \mathcal{N}_{n,r}\left(\mathbf{0}, \; n\sigma^2(\mathbf{I}_n \otimes \mathbf{I}_r)/r\right).$$

Accordingly, we can also obtain

$$\mathbf{F} \mid \mathbf{W} \sim \mathcal{N}_{n,r}\left(\mathbf{BW} + \mathbf{1}_n\mathbf{u}',\ n\sigma^2(\mathbf{I}_n \otimes \mathbf{I}_r)/r\right)$$

and (8). In addition, using Bayes' rule, we can compute the conditional distribution of $\mathbf{W}$ given $\mathbf{F}$. That is, it follows a matrix-variate normal distribution:

$$\mathbf{W} \mid \mathbf{F} \sim \mathcal{N}_{m,r}\big(\mathbf{D}^{-1}\mathbf{B}'(\mathbf{F} - \mathbf{1}_n\mathbf{u}'),\ n\sigma^2(\mathbf{D}^{-1} \otimes \mathbf{I}_r)/r\big), \tag{13}$$

where $\mathbf{D} = \mathbf{B}'\mathbf{B} + \sigma^2\mathbf{I}_m$. Considering $\mathbf{W}$ as the missing data, $\{\mathbf{W}, \mathbf{F}\}$ as the complete data, and $\mathbf{B}$ and σ^2 as the model parameters, we now devise an EM algorithm for our probabilistic KPCA. First, the complete-data log-likelihood is

$$\begin{aligned} L_c &= \log p(\mathbf{W},\ \mathbf{F}) = \log p(\mathbf{F} \mid \mathbf{W}) + \log p(\mathbf{W}) \\ &= -\frac{nr}{2}\log\sigma^2 - \frac{r}{2n}\mathrm{tr}\left(\mathbf{WW}'\right) - \frac{r}{2n\sigma^2}\mathrm{tr}\left((\mathbf{F} - \mathbf{BW} - \mathbf{1}_n\mathbf{u}')(\mathbf{F} - \mathbf{BW} - \mathbf{1}_n\mathbf{u}')'\right), \end{aligned}$$

where we have omitted the terms independent of the model parameters. It is easy to find that $\mathbf{W}$ and $\mathbf{WW}'$ are the complete-data sufficient statistics for $\mathbf{B}$ and σ^2. Given the tth estimates, $\mathbf{B}(t)$ and $\sigma^2(t)$, of $\mathbf{B}$ and σ^2, the E-step computes the expectation of L_c w.r.t. $p(\mathbf{W} \mid \mathbf{B}(t),\ \sigma^2(t))$:

$$\begin{aligned} Q(\mathbf{B}, \sigma^2 \mid \mathbf{B}(t), \sigma^2(t)) &= -\frac{nr}{2}\log\sigma^2 - \frac{r}{2n\sigma^2}\mathrm{tr}\left(\mathbf{HKH}\right) - \frac{r}{2n}\mathrm{tr}\left(\langle\mathbf{WW}'\rangle\right) \\ &\quad -\frac{r}{2n\sigma^2}\mathrm{tr}\left(\langle\mathbf{WW}'\rangle\mathbf{B}'\mathbf{B}\right) + \frac{r}{n\sigma^2}\mathrm{tr}\left(\mathbf{B}\langle\mathbf{W}\rangle(\mathbf{F} - \mathbf{1}_n\mathbf{u}')'\right), \end{aligned}$$

where $\langle\mathbf{W}\rangle = E(\mathbf{W} \mid \mathbf{B}(t), \sigma^2(t))$ and $\langle\mathbf{WW}'\rangle = E(\mathbf{WW}' \mid \mathbf{B}(t), \sigma^2(t))$. Using the properties of matrix-variate normal distributions [3], we have

$$E(\mathbf{W} \mid \mathbf{B}(t),\ \sigma^2(t)) = \mathbf{D}^{-1}(t)\mathbf{B}'(t)(\mathbf{F} - \mathbf{1}_n\mathbf{u}'), \tag{14}$$

$$E(\mathbf{WW}' \mid \mathbf{B}(t),\ \sigma^2(t)) = n\sigma^2(t)\mathbf{D}^{-1}(t) + \mathbf{D}^{-1}(t)\mathbf{B}'(t)\mathbf{HKHB}(t)\mathbf{D}^{-1}(t) \tag{15}$$

The M-step maximizes $Q(\mathbf{B}, \sigma^2 \mid \mathbf{B}(t), \sigma^2(t))$ w.r.t. $\mathbf{B}$ and σ^2, giving new parameter estimates

$$\mathbf{B}(t{+}1) = (\mathbf{F} - \mathbf{1}_n\mathbf{u}')\langle\mathbf{W}'\rangle\left(\langle\mathbf{WW}'\rangle\right)^{-1}, \tag{16}$$

$$\begin{aligned} \sigma^2(t{+}1) = \frac{1}{n^2}\Big(&\mathrm{tr}\big(\mathbf{HKH}\big) + \mathrm{tr}\big(\langle\mathbf{WW}'\rangle\mathbf{B}'(t{+}1)\mathbf{B}(t{+}1)\big) \\ &-2\mathrm{tr}\big(\mathbf{B}(t{+}1)\langle\mathbf{W}\rangle(\mathbf{F} - \mathbf{1}_n\mathbf{u}')'\big)\Big). \end{aligned} \tag{17}$$

Substituting $\langle\mathbf{W}\rangle$ and $\langle\mathbf{WW}'\rangle$ from (14) and (15) into (16) and (17), we can combine the E-step and M-step into one. That is,

$$\mathbf{B}(t{+}1) = \mathbf{HKHB}(t)\big(n\sigma^2(t)\mathbf{I}_m + \mathbf{D}^{-1}(t)\mathbf{B}'(t)\mathbf{HKHB}(t)\big)^{-1}, \tag{18}$$

$$\sigma^2(t{+}1) = \frac{1}{n^2}\mathrm{tr}\Big(\mathbf{HKH} - \mathbf{HKHB}(t)\mathbf{D}^{-1}(t)\mathbf{B}'(t{+}1)\Big). \tag{19}$$

Clearly, this EM algorithm is similar to that of [8] for PPCA. It is important to note that it works with the kernel matrix $\mathbf{K}$ only. Recall that for our probabilistic KPCA, $\mathbf{B}$ is just the data matrix in the latent space, so this iterative procedure is efficient.

5 Discussions

5.1 Probabilistic Principal Coordinate Analysis

As $K(\mathbf{x}_i, \mathbf{x}_j)$ represents the inner product of feature vectors $\mathbf{f}_i$ and $\mathbf{f}_j$, we can directly compute the squared distance a_{ij} between the feature vectors without explicitly using the values of the feature vectors themselves, as

$$a_{ij} = K(\mathbf{x}_i, \mathbf{x}_i) + K(\mathbf{x}_j, \mathbf{x}_j) - 2K(\mathbf{x}_i, \mathbf{x}_j). \tag{20}$$

Let $\mathbf{A} = [a_{ij}]_{n\times n}$. From (20), we have that

$$-\frac{1}{2}\mathbf{HAH} = \mathbf{HKH}$$

and $\mathbf{A}$ is Euclidean [10]. Conversely, given an Euclidean squared-distance matrix $\mathbf{A}$, there corresponds a p.s.d. inner product matrix (i.e., the kernel matrix in this paper). Therefore, our probabilistic model described in Sections 3 and 4 can work with $\mathbf{A}$ instead of $\mathbf{K}$. Based on this perspective, we also refer to our probabilistic KPCA model as probabilistic PCO (PPCO).

Let us investigate further the statistical relationship between PPCA and PPCO. Suppose $\mathbf{F}$ is an $n\times r$ data matrix, of which each row represents an observation and each column represents a variate. Then PPCO is defined by (12), while PPCA is defined as

$$\mathbf{F} = \mathbf{BW} + \mathbf{1}_n\mathbf{u}' + \mathbf{E}, \tag{21}$$

where $\mathbf{u}$ is $r \times 1$ and $\mathbf{W}$ is $m \times r$ with $m < r$,

$$\mathbf{B} \sim \mathcal{N}_{n,m}\left(\mathbf{0},\ \mathbf{I}_n \otimes \mathbf{I}_m\right) \quad \text{and} \quad \mathbf{E} \sim \mathcal{N}_{n,r}\left(\mathbf{0},\ \sigma^2(\mathbf{I}_n \otimes \mathbf{I}_r)\right).$$

Therefore, the duality between them is embodied in that the roles of $\mathbf{B}$ and $\mathbf{W}$ are exchanged.

5.2 Related Work

Lawrence [5] proposed a Gaussian process latent variable model (GPLVM), where he considered to specify a prior distribution of $\mathbf{W}$ and to optimize $\mathbf{B}$. GPLVM is closely related to our probabilistic KPCA model. However, the reason of deriving GPLVM is not so convincing as our model based on Wishart processes. More importantly, GPLVM requires that its used feature vectors have zero mean and the dimensionality of the feature space is specified in advance. These are hard to achieve when kernels are involved. Obviously, our model can avoid these requirements. So, in a way, our method extends GPLVM by relaxing some requirements. Moreover, the "twin kernel PCA" of Lawrence [5] can be obtained immediately by assuming $\mathbf{K} \sim \mathcal{W}_n(r,\ \mathbf{\Sigma})$ as,

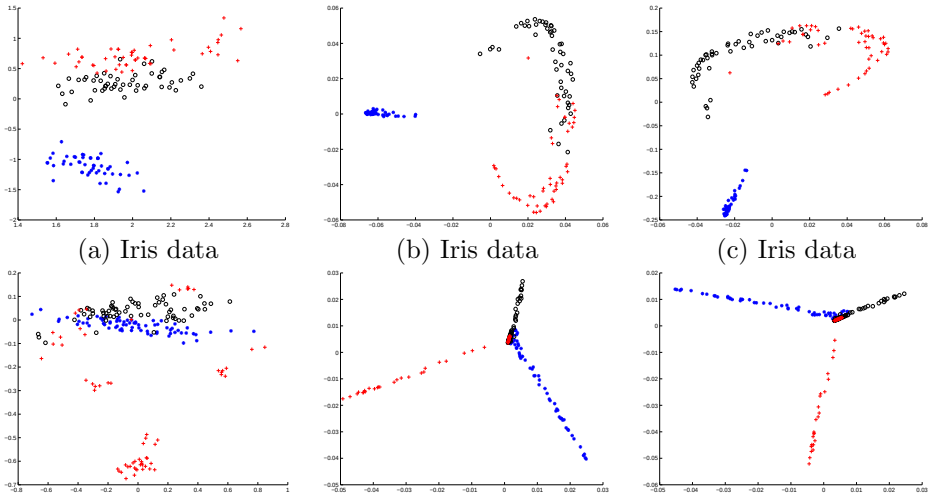
$$p(\mathbf{K} \mid \mathbf{\Sigma}, r) = \frac{1}{C(n,r)}|\mathbf{\Sigma}|^{-r/2}|\mathbf{K}|^{(r-n-1)/2}\exp\left(-\frac{1}{2}\mathrm{tr}(\mathbf{\Sigma}^{-1}\mathbf{K})\right), \tag{22}$$

where $C(n,r) = 2^{rn/2}\pi^{n(n-1)/4}\prod_{j=1}^{n}\Gamma(\frac{r+1-j}{2})$ is a normalization term with $\Gamma(\cdot)$ being the Gamma function. Now, in the "twin kernel PCA", $\mathbf{K}$ is given as a kernel matrix over the high-dimensional space and $\mathbf{\Sigma}$ is defined as a kernel matrix over the low-dimensional latent space. It is necessary to pre-specify the value of r (i.e., D in [5]) for the "twin kernel PCA". However, r is usually unknown for kernel methods. We can handle this problem by assuming $\mathbf{K} \sim \mathcal{W}_n(r,\ \mathbf{\Sigma}/r)$. Furthermore, if $\mathbf{K}$ is singular, we can assume $\mathbf{K} \sim \mathcal{SW}_n(r,\ \mathbf{\Sigma}/r)$ instead.

6 Experiments

We compare the results of applying the linear PCA and probabilistic KPCA to two data sets. The first one is the iris data set and the second one is the multi-phase oil flow data set [1], which consists of 12 features and three classes, *stratified*, *annular* and *homogeneous*, corresponding to the phases of flow in an oil pipeline. For simplicity, here we use only 200 points from the oil flow data.

We adopt the RBF kernel $\mathbf{K} = [\exp(-\|\mathbf{x}_i - \mathbf{x}_j\|^2/\beta)]$ with $\beta = 2.0$ for the iris data and $\beta = 0.2$ for the oil flow data. For probabilistic KPCA, we implement the ML



(a) Iris data (b) Iris data (c) Iris data

(a) Oil flow data (b) Oil flow data (c) Oil flow data

Figure 1: (a) Linear PCA; (b) Probabilistic KPCA with ML; (c) Probabilistic KPCA with EM

estimation with $\mathbf{R} = \mathbf{I}$ and the EM algorithm given in (18) and (19). The EM algorithm is initialized using the standard linear PCA and run for 50 iterations. Figure 1 depicts the two-dimensional latent vectors of the two data sets using the two algorithms. We can see that the two algorithms give essentially the same results which are related by a rotational transformation. This is in line with the theoretical justifications given above.

7 Conclusion

In this paper, we have proposed a probabilistic reformulation of KPCA which can also be called probabilistic PCO. Like PPCA, this opens the door to devising Bayesian methods for KPCA or PCO, such as considering a mixture of probabilistic KPCA models, inferring the dimensionality of the latent space, and constructing a hierarchical model for nonlinear dimensionality reduction. We shall pursue these directions in our future research.

References

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

[2] T. F. Cox and M. A. A. Cox. *Multidimensional Scaling*. Chapman & Hall/CRC, second edition, 2000.

[3] A.K. Gupta and D.K. Nagar. *Matrix Variate Distributions*. Chapman & Hall/CRC, 2000.

[4] I.T. Jolliffe. *Principal component analysis*. Springer, New York, second edition edition, 2002.

[5] N. D. Lawrence. Gaussian process latent variable models for visualisation of high dimensional data. In *Advances in Neural Information Processing Systems 16*, 2004.

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

[7] M. S. Srivastava. Singular Wishart and multivariate Beta distributions. *The Annals of Statistics*, 31(5):1537–1560, 2003.

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

[9] C. K. I. Williams. On a connection between kernel PCA and metric multidimensional scaling. In *Advances in Neural Information Processing Systems 13*, 2001.

[10] Z. Zhang. Learning metrics via discriminant kernels and multidimensional scaling: Toward expected Euclidean representation. In *Proceedings of the 20th International Conference on Machine Learning*, 2003.

[11] Z. Zhang, D.-Y. Yeung, and J. T. Kwok. Wishart processes: A statistical view of reproducing kernels. Technical report, Department of Computer Science, Hong Kong University of Science and Technology, 2004. http://www.cs.ust.hk/∼zhzhang/papers/tr04-01.ps.