Homework #2

Yongjin Shin
20090488
Industrial Management Engineering, POSTECH
dreimer@postech.ac.kr

October 24, 2018

Problem 1. Detailed derivation of PPCA

Solution Description

PCA can be expressed as the maximum likelihood solution of a probabilistic latent variable model. This is known as probabilistic PCA which represents a constrained form of the Gaussian distribution in which the number of free parameters can be restricted while still allowing the model to capture the dominant correlations in a data set. We can derive an EM algorithm for PCA that is computationally efficient.

Firstly, let's introduce an explicit latent variable **s** corresponding to the principal-component subspace. Then we can define a Gaussian prior distribution $p(s)$ over the latent variable, together with a Gaussian conditional distribution $p(x|s)$ for the observed variable **x** conditioned on the value of the latent variable. Specifically, the prior distribution over z is given by a zero-mean and unit-covariance Gaussian

$$p(s) = \mathcal{N}(0, I) \tag{1}$$

Similarly, let's define the conditional distribution of the observed variable x, conditioned on the value of the latent variable z, is again Gaussian, of the form

$$p(x|s) = \mathcal{N}(x|Ws + \mu, \sigma^2 I) \tag{2}$$

in which the mean of x is a general linear function of z governed by the $D \times M$ matrix W. Factor Analysis[a.k.a. FA] and Probabilistic PCA[a.k.a. PPCA] only differ from this covariance $\sigma^2 I$. If we just choose any diagonal matrix Σ, then this would be FA case. However, we only derive PPCA case for the simplicity.

Then, in a generative viewpoint, the D-dimensional observed variable x is defined by a linear transformation of the M-dimensional latent variable s plus additive Gaussian noise (i.e. ϵ follows $\mathcal{N}(0, \sigma^2 I)$) , so that

$$x = Ws + \mu + \epsilon \tag{3}$$

Suppose we wish to determine the values of the parameters W, μ and σ^2 using ML. Then, we need the marginal distribution p(x) such that,

$$p(x) = \int p(x, s)dz = \int p(x|s)p(s)ds \tag{4}$$

Because of the linearity, we can find statistics as follows:

$$\mathbb{E}[x] = \mathbb{E}[Ws + \mu + \epsilon] = \mathbb{E}[Ws] + \mathbb{E}[\mu] + \mathbb{E}[\epsilon] = \mu \tag{5}$$

And if we consider that **s** has a unit covariance and independent with ϵ, so that **s** and ϵ are uncorrelated. If we keep in mind that the noise has $\sigma^2 I$, then

$$cov[x] = \mathbb{E}[(Ws + \epsilon)(Ws + \epsilon)^T] = \mathbb{E}[(Wss^T W^T) + (\epsilon s^T W^T) + (Ws\epsilon) + (\epsilon\epsilon^T)] = WW^T + \sigma^2 I \tag{6}$$

Thus we can find that the marginal distribution $p(x)$ is again Gaussian, and is given by

$$p(x) = \mathcal{N}(x|\mu, WW^T + \sigma^2 I) \tag{7}$$

Let's call this covariance as $C = WW^T + \sigma^2 I$ for the furthur derivations.

From now on, we can derive the EM algorithm for probabilistic PCA by following the general framework for EM. Thus we write down the complete-data log likelihood and take its expectation with respect to the posterior distribution of the latent distribution evaluated using 'old' parameter values. Maximization of this expected complete-data log likelihood then yields the 'new' parameter values.

Therefore, we need to find the posterior distribution of the latent distribution. If we know the joint distribution $p(x, z)$, then we can easily find out the conditional distribution $p(z|x)$ by using *equation (352), (353) and (354) in [3]*. At this moment, we only need to know $cov[x, s]$ and $cov[s, x]$.

$$cov[x, s] = \mathbb{E}[(Ws + \epsilon)s^T] = \mathbb{E}[Wss^T] + \mathbb{E}[\epsilon] = W \tag{8}$$

$$cov[s, x] = \mathbb{E}[s(Ws + \epsilon)^T] = \mathbb{E}[ss^T W^T] + \mathbb{E}[s\epsilon^T] = W^T \tag{9}$$

Thus,

$$p\Big(\begin{bmatrix} s \\ x \end{bmatrix}\Big) = \mathcal{N}\Big(\begin{bmatrix} s \\ x \end{bmatrix} | \begin{bmatrix} 0 \\ \mu \end{bmatrix}, \begin{bmatrix} I & W^T \\ W & WW^T + \sigma^2 I(= C) \end{bmatrix}\Big) \tag{10}$$

Deploying the equation commented above (equation (352), (353) and (354) in [3]),

$$p(s|x) = \mathcal{N}(W^T C^{-1}(x - \mu), I - W^T C^{-1} W) \tag{11}$$

Note that if we consider more efficient algorithm, then we can use *matrix inversion lemma*. By the *equation (157) in [3]*,

$$\begin{aligned} C^{-1} &= (WW^T + \sigma^2 I)^{-1} \\ &= \sigma^{-2} I - \sigma^{-2} IW(I + W^T \frac{1}{\sigma^{-2}} IW)^{-1} W^T \frac{1}{\sigma^{-2}} I \\ &= \sigma^{-2} I - \sigma^{-2} W(W^T W + \sigma^2 I)^{-1} W^T \end{aligned} \tag{12}$$

In this representation, we only need to inverse $W^T W$ which shape is $\mathbb{M} \times \mathbb{M}$, instead of WW^T which shape is $\mathbb{D} \times \mathbb{D}$ ($\mathbb{M} < \mathbb{D}$, $\mathbb{D}, \mathbb{M} \in \mathcal{R}$). Therefore it must be much better to use this C^{-1} representation. Then the equation (11) can be represented like follows (Define $M = W^T W + \sigma^2 I$):

$$\begin{aligned} p(s|x) &= \mathcal{N}(W^T C^{-1}(x - \mu), I - W^T C^{-1} W) \\ &= \mathcal{N}(W^T(\sigma^{-2} I - \sigma^{-2} W M^{-1} W^T)(x - \mu), I - W^T(\sigma^{-2} I - \sigma^{-2} W M^{-1} W^T)W) \\ &= \mathcal{N}(s|M^{-1} W^T(x - \mu), \sigma^2 M^{-1}) \end{aligned} \tag{13}$$

The mean value $M^{-1} W^T x$ was achieved as follows:

$$\begin{aligned} W^T(\sigma^{-2} I - \sigma^{-2} W M^{-1} W^T) &= W^T(\sigma^{-2} I - \sigma^{-2} W(\sigma^2 I + W^T W)^{-1} W^T) \\ &= \sigma^{-2} W^T - \sigma^{-2} W^T W(\sigma^2 I + W^T W)^{-1} W^T \\ &= \big(\sigma^{-2} I - \sigma^{-2} W^T W(\sigma^2 I + W^T W)^{-1}\big) W^T \\ &= \big(\sigma^{-2}(\sigma^2 I + W^T W) - \sigma^{-2} W^T W\big)(\sigma^2 I + W^T W)^{-1} W^T \\ &= (\sigma^2 I + W^T W)^{-1} W^T \\ &= M^{-1} W^T \end{aligned} \tag{14}$$

Also, $\sigma^2 M^{-1}$ was derived as follows:

$$
\begin{aligned}
I - W^T(\sigma^{-2}I - \sigma^{-2}WM^{-1}W^T)W &= I - W^T(\sigma^{-2}I - \sigma^{-2}W(\sigma^2 + W^TW)^{-1}W^T)W \\
&= I - \sigma^{-2}W^T + \sigma^{-2}W^TW(\sigma^2 + W^TW)^{-1}W^TW \\
&= \sigma^{-2}\big[\sigma^2 I - W^TW + W^TW(\sigma^2 I + W^TW)^{-1}W^TW\big] \\
&= \sigma^{-2}\big[\sigma^2 I + \big(-(\sigma^2 I + W^TW) + W^TW\big)(\sigma^2 I + W^TW)^{-1}W^TW\big] \\
&= I - (\sigma^2 I + W^TW)^{-1}W^TW \\
&= (\sigma^2 I + W^TW)^{-1}\big((\sigma^2 I + W^TW) - W^TW\big) \\
&= (\sigma^2 I + W^TW)^{-1} \times \sigma^{-2}I \\
&= \sigma^2(\sigma^2 I + W^TW)^{-1} \\
&= \sigma^{-2}M^{-1}
\end{aligned}
\tag{15}
$$

Since data points are assumed independent, the complete-data log likelihood function takes the form

$$
\log p(X, S|W, \mu, \sigma^2) = \sum_{n=1}^{N} \big(\log p(x_n|s_n) + \log p(s_n)\big) \tag{16}
$$

Making use of the expressions (1) and (2) for the latent and conditional distributions given above, but simplifying and ignoring the constants, we get

$$
\log p(X, S|W, \mu, \sigma^2) = -\sum_{n=1}^{N} \big[\frac{D}{2}\log 2\pi\sigma^2 + \frac{1}{2\sigma^2}\|x_n - \mu\|^2 - \frac{1}{\sigma^2}s_n^TW^T(x_n - \mu) + \frac{1}{2\sigma^2}Tr(s_ns_n^TW^TW) + \frac{1}{2}Tr(s_ns_n^T)\big] \tag{17}
$$

Here, we know that if A is a column vector, then $A^TA = Tr(AA^T)$. Thus 2nd, 4th and 5th terms are the product of column vectors so that we can use Trace. Especially in the 4th term, we can use a *trace trick*, which is that $Tr(ABC) = Tr(BCA) = Tr(CAB)$referenced in *equation (15) in [3].* Thus $Tr(s_n^TW^TWs_n) = Tr(W^TWs_ns_n^T) = Tr(s_ns_n^TW^TW)$. *(Note that since PPCA assume the covariance of $p(x|s)$ is $\sigma^2 I$, so that this log likelihood function is slightly different from FA.)* And taking the expectation with respect to the posterior distribution over the latent variables, we obtain

$$
\mathbb{E}[\log p(X, S|W, \sigma^2)] = -\sum_{n=1}^{N} \big[\frac{D}{2}\log 2\pi\sigma^2 + \frac{1}{2\sigma^2}\|x_n - \mu\|^2 - \frac{1}{\sigma^2}\mathbb{E}(s_n)^TW^T(x_n - \mu) + \frac{1}{2\sigma^2}Tr(\mathbb{E}(s_ns_n)^TW^TW) + \frac{1}{2}Tr(\mathbb{E}(s_ns_n^T))\big] \tag{18}
$$

Note that this depends on the posterior distribution only through the sufficient statistics of the Gaussian. Thus in the E step, we use the old parameter values to evaluate

$$
\mathbb{E}(s_n) = M^{-1}W^T(x_n - \mu) \tag{19}
$$

But, we already know that the exact maximum likelihood solution for μ is given by the sample mean $\bar{x}$ defined by $\bar{x} = \frac{1}{N}\sum_{n=1}^{N} x_n$, thus

$$
\mathbb{E}(s_n) = M^{-1}W^T(x_n - \bar{x}) \tag{20}
$$

$$
\mathbb{E}(s_ns_n^T) = \sigma^2 M^{-1} + \mathbb{E}(s_n)\mathbb{E}(s_n)^T \tag{21}
$$

which follow directly from the posterior distribution (13). In (21), we used the fact that

$$
\begin{aligned}
cov(s_n) &= \mathbb{E}((s_n - \mathbb{E}(s_n))(s_n - \mathbb{E}(s_n))^T) \\
&= \mathbb{E}(s_ns_n^T - s_n\mathbb{E}(s_n)^T - \mathbb{E}(s_n)s_n^T + \mathbb{E}(s_n)\mathbb{E}(s_n)^T) \\
&= \mathbb{E}(s_ns_n^T) - \mathbb{E}(s_n)\mathbb{E}(s_n)^T
\end{aligned}
\tag{22}
$$

Thus $\mathbb{E}(s_n s_n^T) = cov(s_n) + \mathbb{E}(s_n)\mathbb{E}(s_n)^T$.

In the M step, we maximize with respect to W and σ^2, keeping the posterior statistics fixed. We can utilize the *equation (112) and (71) in reference [3]* as follows:

$$\begin{aligned}\frac{\partial \mathbb{E}(\mathcal{L})}{\partial W} &= -\sum_{n=1}^{N}\big[\frac{-1}{\sigma^2}(x_n-\bar{x})\mathbb{E}(s_n)^T + \frac{W}{2\sigma^2}(\mathbb{E}(s_n s_n^T) + \mathbb{E}(s_n s_n^T)^T)\big] \\ &= \sum_{n=1}^{N}\big[\frac{1}{\sigma^2}(x_n-\bar{x})\mathbb{E}(s_n)^T - \frac{W}{\sigma^2}(\mathbb{E}(s_n s_n^T)\big] \\ &= 0\end{aligned} \tag{23}$$

Thus,

$$W_{new} = \big[\sum_{n=1}^{N}(x_n-\bar{x})\mathbb{E}(s_n)^T\big]\big[\sum_{n=1}^{N}\mathbb{E}(s_n s_n^T)\big]^{-1} \tag{24}$$

Maximization with respect to σ^2 is straightforward.

$$\begin{aligned}\frac{\partial \mathbb{E}(\mathcal{L})}{\partial \sigma^2} &= -\sum_{n=1}^{N}\big[D\sigma^{-1} - \sigma^{-3}\|x_n-\bar{x}\|^2 + 2\sigma^{-3}\mathbb{E}(s_n)^T W^T(x_n-\bar{x}) - \sigma^{-3}Tr(\mathbb{E}(s_n s_n^T)W^T W)\big] \\ &= -\sigma^{-3}\sum_{n=1}^{N}\big[D\sigma^2 - \|x_n-\bar{x}\|^2 + 2\mathbb{E}(s_n)^T W^T(x_n-\bar{x}) - Tr(\mathbb{E}(s_n s_n^T)W^T W)\big] \\ &= 0\end{aligned} \tag{25}$$

Thus, $ND\sigma^2 = \sum_{n=1}^{N}\big[\|x_n-\bar{x}\|^2 - 2\mathbb{E}(z_n)^T W^T(x_n-\bar{x}) + Tr(\mathbb{E}(z_n z_n^T)W^T W)\big]$, therefore

$$\sigma^2_{new} = \frac{1}{ND}\sum_{n=1}^{N}\big[\|x_n-\bar{x}\|^2 - 2\mathbb{E}(s_n)^T W_{new}^T(x_n-\bar{x}) + Tr(\mathbb{E}(s_n s_n^T)W_{new}^T W_{new})\big] \tag{26}$$

Some points need to be mentioned that there are a few benefits of PPCA over PCA. First of all, PPCA can handle missing data by treating missed data as latent variable in E step. And regarding the computation cost, it is not necessary to compute the D x D covariance to achieve eigen-decomposition. Especially when K is small, we have less expensive work (i.e. inverting K x K). Another noticeable benefit is that it give us a fully Bayesian treatment so that we can infer the number of K (which makes our life tired). And the last one is that it is quite easy to plug in PPCA as part of more complex problems. One of these is Mixtures of PPCA which we will derive in the problem (2). By doing this, we can handle nonlinear dimensional reduction or even subspace clustering.

Problem 2. Detailed derivation of mixtures of PPCA

Solution Description

The association of a probability model with PCA offers the tempting prospect of being able to model complex data structures with a combination of local PCA models through the mechanism of a mixture of PPCA. This formulation would permit all of the model parameters to be determined from maximum likelihood, where both the appropriate partitioning of the data and the determination of the respective principal axes occur automatically as the likelihood is maximized.

The corresponding generative model for the mixture case now requires the random choice of a mixture component according to the proportions π_k, followed by sampling from the X and ϵ distributions and applying equation (1) as in the single model case, taking care to use the appropriate parameters μ_k, W_k and σ_k^2. We can develop an iterative EM algorithm for optimization of all of the model parameters, $\{\pi_k, \mu_k, W_k, \sigma_k^2\}$.

Like how we defined in problem 1, we also take the same distribution of $p(s_n)$, however from now on we have to consider another latent variable z_n which distribution follows:

$$p(z_n) = \prod_{k=1}^{K} \pi_k^{z_{k,n}} \tag{27}$$

Thus,

$$p(x_n) = \sum_{k=1}^{K} \pi_k p(x_n | z_{k,n} = 1) \tag{28}$$

In this representation, we deploy one hot encoding as usual such that $z_n \in \{0,1\}^K$. Besides, $p(x_n)$ is same as what we derived in the problem 1 (equation 7), but each cluster has different μ_k, W_k, σ_k^2. Therefore,

$$p(x_n | z_{k,n} = 1) = \mathcal{N}(x_n | \mu_k, C_k) \quad where \quad C_k = \sigma_k^2 I + W_k W_k^T \tag{29}$$

Next we can verify the distribution over observed variables conditioned on two latent variables easily as follows:

$$p(x_n | s_n, z_n) = \prod_{k=1}^{K} \mathcal{N}(x_n | W_k s_n + \mu_k, \sigma_k^2 I)^{z_{k,n}} \tag{30}$$

For the E step, the expected complete-data log-likelihood is calculated as (Note that in the following equations, $<>$ denotes the expectation with respect to the posterior distributions of both z_n and s_n and terms independent of the model parameters have been omitted.)

$$\begin{aligned}
\mathbb{E}(L) =& < \sum_{n=1}^{N} \log p(x_n, s_n, z_n) > \\
=& < \sum_{n=1}^{N} (\log p(x_n | s_n, z_n) + \log p(s_n) + \log p(z_n) > \\
=& \sum_{n=1}^{N} < \sum_{k=1}^{K} z_{k,n} \log \mathcal{N}(W_k s_n + \mu_k, \sigma_k^2 I) + \log \mathcal{N}(0, I) + \sum_{k=1}^{K} z_{k,n} \log \pi_k > \\
=& -\sum_{n=1}^{N} \sum_{k=1}^{K} < z_{k,n} > [\frac{D}{2} \log 2\pi\sigma_k^2 + \frac{1}{2\sigma_k^2} \|x_n - \mu_k\|^2 - \frac{1}{\sigma_k^2} < s_n^T > W_k^T (x_n - \mu_k) + \\
& \frac{1}{2\sigma^2} Tr(< s_n s_n^T > W_k^T W_k) + \frac{1}{2} Tr(< s_n s_n^T >) - \log \pi_k]
\end{aligned} \tag{31}$$

Since only when $z_{k,n} = 1$, the terms which related with $z_{k,n}$ can exist and that will be only once. Thus we can take out $z_{k,n}$ out of the whole equation and then calculate expectation. Meanwhile, in terms of the expectation of $z_{k,n}$ we can derive it as follows:

$$\mathbb{E}[z_{k,n}] = p(z_{k,n} = 1 | x_n) * 1 + p(z_{k,n} = 0 | x_n) * 0 = p(z_{k,n} = 1 | x_n) \tag{32}$$

If we call this $\mathbb{E}[z_{k,n}]$ as $r_{k,n}$, then using the result of (28). (29):

$$r_{k,n} = p(z_{k,n} = 1|x_n) = \frac{p(z_{k,n} = 1, x_n)}{p(x_n)} = \frac{\pi_k \mathcal{N}(\mu_k, C_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mu_j, C_j)} \tag{33}$$

Also, notice that in the previous problem, we have already derived the posterior distribution of $p(s_n|x_n)$ thus we can utilize the result of it again:

$$p(s_n|x_n, z_{k,n} = 1) = \mathcal{N}(s_n|M_k^{-1}W_k(x_n - \mu_k), \sigma_k^2 M_k^{-1}) \quad where \quad M_k = W_k^T W_k + \sigma_k^2 I \tag{34}$$

Therefore,

$$\mathbb{E}(s_n|x_n, z_{k,n} = 1) = M_k^{-1}W_k^T(x_n - \mu_k) \tag{35}$$

$$\mathbb{E}(s_n s_n^T|x_n, z_{k,n} = 1) = \sigma_k^2 M_k^{-1} + \mathbb{E}(s_n|x_n, z_{k,n} = 1)\mathbb{E}(s_n|x_n, z_{k,n} = 1)^T \tag{36}$$

where $M_k = W_k^T W_k + \sigma_k^2 I$.

Now, we can update the parameters $\{\pi_k, W_k, \mu_k, \sigma_k^2\}$ with using the results of E step such that $r_{k,n}, \mathbb{E}(s_n|x_n, z_{k,n} = 1), \mathbb{E}(s_n s_n^T|x_n, z_{k,n} = 1)$. This will be quite redundant process with how we did in the problem 1 except considering $r_{k,n}$ and updating μ_k , π_k. First of all W_k, μ_k can be updated by using (23), (24) and (25), (26) respectively. (note that we are going to abuse the notation, so that $\mathbb{E}(s_n|x_n, z_{k,n} = 1)$ will be represented simply by $\mathbb{E}(s_{k,n})$.

$$\begin{aligned}\frac{\partial \mathbb{E}(\mathcal{L})}{\partial W_k} &= -\sum_{n=1}^{N} \mathbb{E}(z_{k,n})\big[\frac{-1}{\sigma^2}(x_n - \mu_k)\mathbb{E}(s_{k,n})^T + \frac{W_k}{2\sigma^2}(\mathbb{E}(s_{k,n}s_{k,n}^T) + \mathbb{E}(s_{k,n}s_{k,n}^T)^T)\big] \\ &= \sum_{n=1}^{N} r_{k,n}\big[\frac{1}{\sigma_k^2}(x_n - \mu_k)\mathbb{E}(s_{k,n})^T - \frac{W_k}{\sigma^2}(\mathbb{E}(s_{k,n}s_{k,n}^T)\big] = 0\end{aligned} \tag{37}$$

Thus,

$$W_{k,new} = \big(\sum_{n=1}^{N} r_{k,n}(x_n - \mu_k)\mathbb{E}(s_{k,n})^T\big)\big(\sum_{n=1}^{N} r_{k,n}\mathbb{E}(s_{k,n}s_{k,n}^T)\big)^{-1} \tag{38}$$

Next,

$$\begin{aligned}\frac{\partial \mathbb{E}(\mathcal{L})}{\partial \sigma_k^2} &= -\sum_{n=1}^{N} \mathbb{E}(z_{k,n})\big[D\sigma^{-1} - \sigma_k^{-3}\|x_n - \mu_k\|^2 + 2\sigma^{-3}\mathbb{E}(s_{k,n})^T W_k^T(x_n - \mu_k) - \sigma_k^{-3}Tr(\mathbb{E}(s_{k,n}s_{k,n}^T)W_k^T W_k)\big] \\ &= -\sigma_k^{-3}\sum_{n=1}^{N} r_{k,n}\big[D\sigma_k^2 - \|x_n - \mu_k\|^2 + 2\mathbb{E}(s_{k,n})^T W_k^T(x_n - \mu_k) - Tr(\mathbb{E}(s_{k,n}s_{k,n}^T)W_k^T W_k)\big] = 0\end{aligned} \tag{39}$$

Thus,

$$\sigma_{k,new}^2 = \frac{\sum_{n=1}^{N} r_{k,n}\big[\|x_n - \mu_k\|^2 - 2\mathbb{E}(s_{k,n})^T W_{k,new}^T(x_n - \mu_k) + Tr(\mathbb{E}(s_{k,n}s_{k,n}^T)W_{k,new}^T W_{k,new})\big]}{D\sum_{n=1}^{N} r_{k,n}} \tag{40}$$

In the case of μ_k, we have to use the fact that $\frac{\partial A^T A}{\partial A} = 2A$. Then apply this in (31) as follows:

$$\begin{aligned}\frac{\partial \mathbb{E}(\mathcal{L})}{\partial \mu_k} &= -\sum_{n=1}^{N} \mathbb{E}(z_{k,n})\big[\frac{1}{\sigma_k^2}(\mu_k - x_n) + \frac{1}{\sigma_k^2}W_k\mathbb{E}(s_{k,n})\big] \\ &= -\sum_{n=1}^{N} r_{k,n}(\mu_k - x_n + W_k\mathbb{E}(s_{k,n})) = 0\end{aligned} \tag{41}$$

Therefore,

$$\mu_{k,new} = \frac{\sum_{n=1}^{N} \left[r_{k,n}(x_n - W_k \mathbb{E}(s_{k,n})) \right]}{\sum_{n=1}^{N} r_{k,n}} \tag{42}$$

The maximization with respect to π_k must take account of the constraint that $\sum_{k=1}^{K} \pi_k = 1$. This can be achieved with the use of a Lagrange multiplier λ and maximizing $\mathbb{E}(\mathcal{L}') = \mathbb{E}(\mathcal{L}_c) + \lambda(\sum_{k=1}^{K} \pi_k - 1)$. Then,

$$\begin{aligned} \frac{\partial \mathbb{E}(\mathcal{L}')}{\partial \pi_k} &= \sum_{n=1}^{N} \frac{\mathbb{E}(z_{k,n})}{\pi_k} - \lambda \\ &= \sum_{n=1}^{N} \frac{r_{k,n}}{\pi_k} - \lambda = 0 \end{aligned} \tag{43}$$

Also,

$$\frac{\partial \mathbb{E}(\mathcal{L}')}{\partial \lambda} = \sum_{k=1}^{K} \pi_k - 1 = 0 \tag{44}$$

Thus use this result to sum up on the both sides of (43), then

$$\lambda = \sum_{k=1}^{K} \sum_{n=1}^{N} r_{k,n} = N \tag{45}$$

Therefore,

$$\pi_{k,new} = \frac{1}{N} \sum_{n=1}^{N} r_{k,n} \tag{46}$$

Problem 3. Implement the algorithm on mixture of PPCA and run it on at least two different datasets. You can have a look at Tipping and Bishop's experiments.

Solution Description

Algorithm 1: MPPCA

input : Dataset $\mathcal{D} = \{x_1, ..., x_N\}$

output: $\pi_k, \mu_k, \Sigma_k^2, W_{n,k}$ s.t $k = \{1, ..., K\}$

Choose an initial setting for the parameters $\pi^{old}, W^{old}, \sigma^2, \mu$

while *NOT converging OR iteration* $<$ *max_iteration* **do**

$r_{k,n} = \frac{\pi_k \mathcal{N}(\mu_k, C_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mu_j, C_j)}$

$\mathbb{E}(s_n | x_n, z_{k,n} = 1) = M_k^{-1} W_k^T (x_n - \mu_k)$

$\mathbb{E}(s_n s_n^T | x_n, z_{k,n} = 1) = \sigma_k^2 M_k^{-1} + \mathbb{E}(s_n | x_n, z_{k,n} = 1)\mathbb{E}(s_n | x_n, z_{k,n} = 1)^T$

$W_{k,new} = \left(\sum_{n=1}^{N} r_{k,n}(x_n - \mu_k)\mathbb{E}(s_{k,n})^T \right) \left(\sum_{n=1}^{N} r_{k,n} \mathbb{E}(s_{k,n} s_{k,n}^T) \right)^{-1}$

$\sigma_{k,new}^2 = \frac{\sum_{n=1}^{N} r_{k,n} \left[\|x_n - \mu_k\|^2 - 2\mathbb{E}(s_{k,n})^T W_{k,new}^T (x_n - \mu_k) + Tr(\mathbb{E}(s_{k,n} s_{k,n}^T) W_{k,new}^T W_{k,new}) \right]}{D \sum_{n=1}^{N} r_{k,n}}$

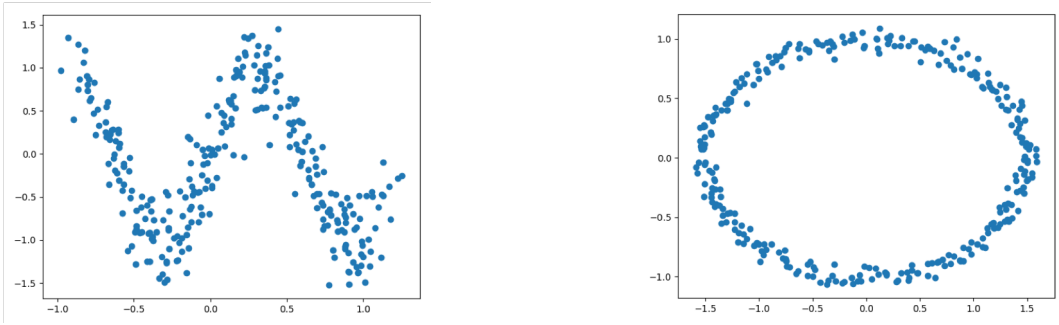
$\mu_{k,new} = \frac{\sum_{n=1}^{N} \left[r_{k,n}(x_n - W_k \mathbb{E}(s_{k,n})) \right]}{\sum_{n=1}^{N} r_{k,n}}$

$\pi_{k,new} = \frac{1}{N} \sum_{n=1}^{N} r_{k,n}$

end

Return

Two 300-point data sets with added gaussian noise were handled by this MPPCA algorithm using six mixture components. For the sake of visualization, data was only generated in $\mathbb{R}^2$ so that the dimension of corresponding latent variable should have less than 2 therefore, our model only have 1 dimension of latent variables, but still it could handle the non-linear dimensional reduction problem. Two toy data set looks as follows:



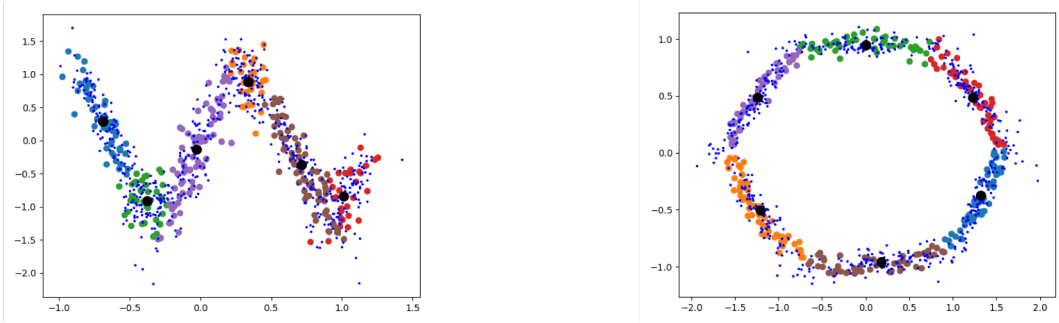
(a) Sine data

(b) Eclipse data

Figure 1: Toy Dataset

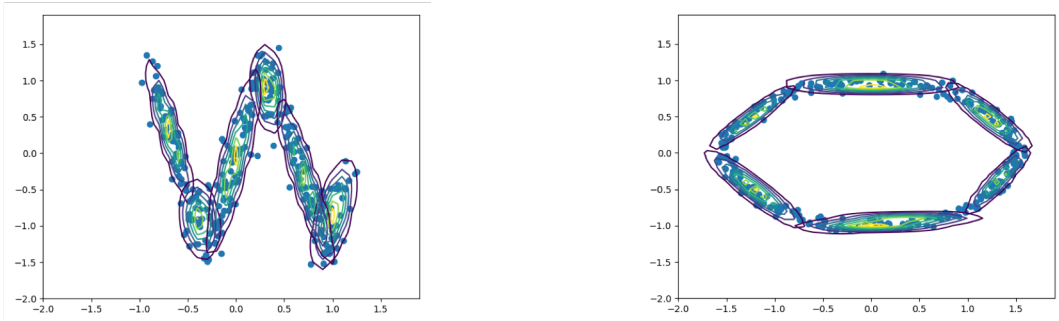
The result of Mixture of PPCA are shown in figure 2 and figure 3. They are quite well clustered with 6 components if we consider the result of Kmeans in Figure(b). And also we can find that each component reduce its dimension into 1. As I already mentioned, MPPCA can divide nonlinear problem into several components and solves it as like linear problem (a.k.a divide and conquer).

Since we have discussed that PPCA can give us some clue to infer the number of K. I guess that if we have n number of components, but m number of components out of n have dominant prior probability(P_i), then do we really need other components which have very small prior probability? In figure 3(a), the smaller ones have 11% of prior probability. Thus, after getting rid of two smallest components, so that if we use $K = 4$, then we have nearly 4 components of nearly similar (25%) prior probability. Of course, if we just discard the smaller ones, that might issue a problem of perverting the given data.



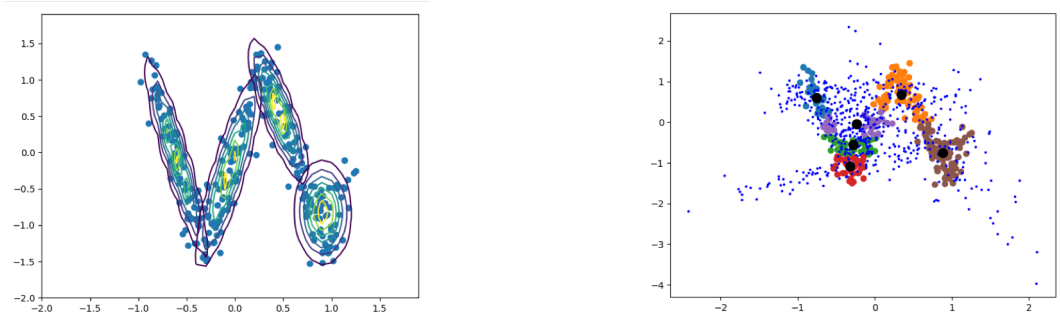
(a) Sine data (b) Eclipse data

Figure 2: Result of MPPCA - 1



(a) Sine data (b) Eclipse data

Figure 3: Result of MPPCA - 1



(a) Sine data with 4 components (b) Kmeans

Figure 4: Other Experiments

References

[1] Christopher M. Bishop. *Pattern Recognition and Machine Learning.*

[2] Seungjin Choi. *Lecture Note.*

[3] Kaare Brandt Petersen. *The Matrix Cookbook.*

[4] Michael E. Tipping, Christopher M. Bishop. *Mixtures of Probabilistic Principal Component Analyzers.*