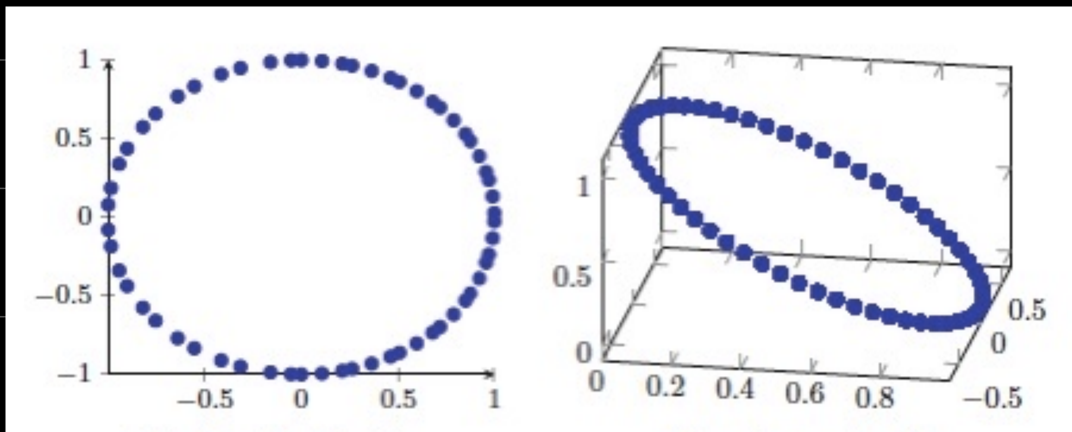
3.4 Nonlinear / Kernel PCA

3.4.1 Motivation

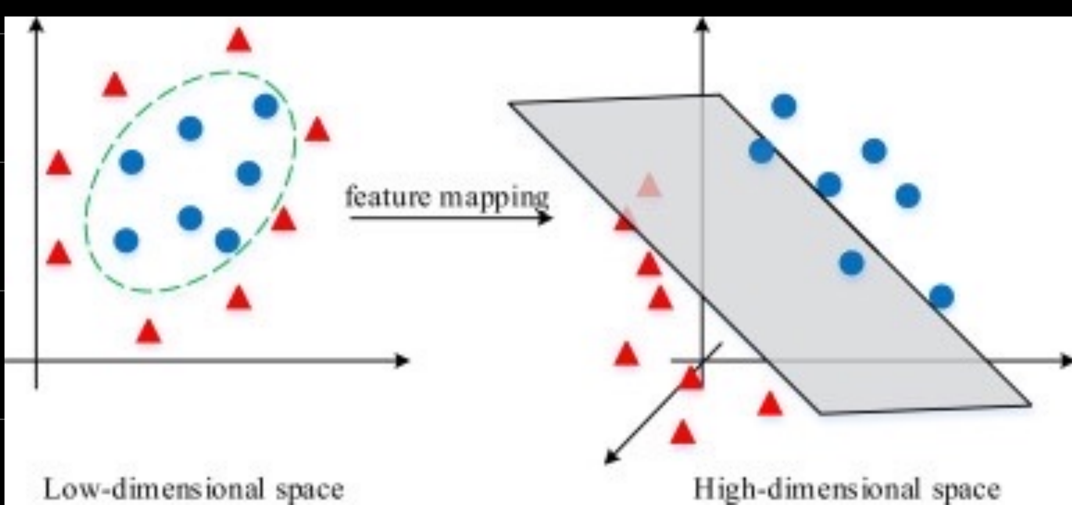


(a) $x_1^2 + x_2^2 - 1 = 0$ (b) $z_1 + z_3 - 1 = 0$

$$\phi : \mathbb{R}^2 \to \mathbb{R}^3$$

$$(x_1, x_2) \mapsto (z_1, z_2, z_3) = (x_1^2, \sqrt{2}x_1x_2, x_2^2)$$

Observation: (1) nonlinear change of features can turn nonlinear eqns into linear eqns.



Observation: (2) in some cases, increasing the number of features is helpful.

Nonlinear/Kernel PCA: a method to embed sample pts into a high dimensional space before applying PCA such that the structure of the sample becomes linear.

Setting: Given sample pts $x^{(1)}, \cdots, x^{(n)} \in \mathbb{R}^p$ consider a nonlinear embedding map ϕ, known as the feature map.

$$\phi : \mathbb{R}^p \to \mathbb{R}^D \qquad p \ll D$$

$$\underset{\text{ith sample pt}}{x^{(i)}} = \underbrace{(x_1^{(i)}, \cdots, x_p^{(i)})^T}_{\text{old features}} \mapsto \underbrace{\phi(x^{(i)})}_{\text{image of the ith sample pt}} = \underbrace{(\phi_1(x^{(i)}), \cdots, \phi_D(x^{(i)}))^T}_{\text{new features}}$$

In the higher dim space $\mathbb{R}^D$, the sample pts become $\phi(x^{(1)}), \dots, \phi(x^{(n)}) \in \mathbb{R}^D$.
and the sample matrix becomes

$$\Phi \overset{\text{def}}{=} \begin{pmatrix} | & & | \\ \phi(x^{(1)}) & \cdots & \phi(x^{(n)}) \\ | & & | \end{pmatrix}_{D \times n}$$

In nonlinear / kernel PCA, we do:

Step 1: center $\phi(x^{(1)}), \dots, \phi(x^{(n)})$.

If $\mu \overset{\text{def}}{=} \frac{1}{n} \sum_{i=1}^{n} \phi(x^{(i)}) \neq 0$, replace them by $\phi(x^{(1)}) - \mu, \dots, \phi(x^{(n)}) - \mu$.

This corresponds to replace Φ by ΦH where $H \overset{\text{def}}{=} I - \frac{1}{n} \mathbb{1}\mathbb{1}^T$ is the centering matrix.

Step 2: apply PCA to $\phi(x^{(1)}), \dots, \phi(x^{(n)})$.

i.e., find the eigenvectors associated to the largest eigenvalues of $\Phi\Phi^T$

However, $\underset{D\times n}{\Phi}\underset{n\times D}{\Phi^T} \in \mathbb{R}^{D\times D}$, it can be computationally expensive to get its EVD if D is large.

3.4.2. Math Prep.

Let $A \in \mathbb{R}^{m\times n}$.

> **Lemma:** If v is an eigenvector of A^TA associated to $\lambda > 0$, then
> $$u \overset{\text{def}}{=} Av$$
> is an eigenvector of AA^T associated to the same λ. Moreover, $\|u\| = \sqrt{\lambda}\,\|v\|$.

Proof: We have $A^TAv = \lambda v$, then

$$AA^Tu \overset{\text{def of } u}{=} A(A^TAv) \overset{v \text{ is eigenvector}}{=} A(\lambda v) = \lambda Av \overset{\text{def of } u}{=} \lambda u$$

which means u is an eigenvector of AA^T associated to λ. Moreover,

$$\overset{\text{def of } u}{\|u\|^2} = \|Av\|^2 = (Av)^T(Av) = v^T(A^TAv) \overset{v \text{ is an eigenvector}}{=} v^T(\lambda v)$$

$$= \lambda \|v\|^2 \quad \blacksquare$$

3.4.3. Theory.

If $n << D$, the lemma says that the eigenvectors of $\Phi\Phi^T \in \mathbb{R}^{D\times D}$ can be obtained from the eigenvectors of $\Phi^T\Phi \in \mathbb{R}^{n\times n}$.

We proceed in this way to reduce the computational load.

Let
$$\left(\Phi^T\Phi\right)_{n\times n} = U\Lambda U^T = \begin{pmatrix} | & & | \\ u^{(1)} & \cdots & u^{(n)} \\ | & & | \end{pmatrix} \begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{pmatrix} \begin{pmatrix} — u^{(1)} — \\ \vdots \\ — u^{(n)} — \end{pmatrix}$$

where $\lambda_1 \geq \cdots \geq \lambda_n \geq 0$ and $u^{(i)}$ is a unit eigenvector of $\Phi^T\Phi$ associated to λ_i.

By the lemma, $\Phi u^{(i)}$ is an eigenvector of $\Phi\Phi^T$ associated to λ_i, and

$$\|\Phi u^{(i)}\| = \sqrt{\lambda_i}\, \|u^{(i)}\| = \sqrt{\lambda_i}.$$

Thus $v^{(i)} \stackrel{\text{def}}{=} \frac{1}{\sqrt{\lambda_i}} \Phi u^{(i)}$ is a unit eigenvector of $\Phi\Phi^T$ associated to λ_i.