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

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

By the PCA theory, the d-dim subspace S that has the largest projected data variance (for the sample points $\phi(x^{(1)}), \ldots, \phi(x^{(n)})$) is

$$S = \operatorname{span}\{v^{(1)}, \cdots, v^{(d)}\} \overset{\text{def of } v^{(i)}}{=} \operatorname{span}\{\frac{1}{\sqrt{\lambda_1}}\Phi u^{(1)}, \cdots, \frac{1}{\sqrt{\lambda_d}}\Phi u^{(d)}\}$$

where $u^{(1)}, \cdots, u^{(d)}$ are unit eigenvectors of $\Phi^T\Phi$ associated to the d largest eigenvalues.

Here
$$\Phi^T\Phi = \begin{pmatrix} | & & | \\ \phi(x^{(1)}) & \cdots & \phi(x^{(n)}) \\ | & & | \end{pmatrix}^T \begin{pmatrix} | & & | \\ \phi(x^{(1)}) & \cdots & \phi(x^{(n)}) \\ | & & | \end{pmatrix}$$

$$= \begin{pmatrix} — \phi(x^{(1)}) — \\ \vdots \\ — \phi(x^{(n)}) — \end{pmatrix} \begin{pmatrix} | & & | \\ \phi(x^{(1)}) & \cdots & \phi(x^{(n)}) \\ | & & | \end{pmatrix}$$

$$\overset{\text{matrix multi.}}{=} \begin{pmatrix} \phi(x^{(1)})^T\phi(x^{(1)}) & \phi(x^{(1)})^T\phi(x^{(2)}) & \cdots & \phi(x^{(1)})^T\phi(x^{(n)}) \\ \phi(x^{(2)})^T\phi(x^{(1)}) & \phi(x^{(2)})^T\phi(x^{(2)}) & \cdots & \phi(x^{(2)})^T\phi(x^{(n)}) \\ \vdots & \vdots & & \vdots \\ \phi(x^{(n)})^T\phi(x^{(1)}) & \phi(x^{(n)})^T\phi(x^{(2)}) & \cdots & \phi(x^{(n)})^T\phi(x^{(n)}) \end{pmatrix}$$

can be constructed if $k(x,y) \overset{\text{def}}{=} \phi^T(x)\phi(y)$

is known. The function $k(x,y)$ is referred to as a <u>kernel function</u>; the matrix $\Phi^T\Phi$ is referred to as a <u>kernel matrix</u> (w.r.t. the features $\phi(x)$), which can be constructed by evaluating the kernel function on the sample pts.

Examples of kernel functions:

(1) $k(x,y) = x^T y$, this corresponds to $\phi(x) = x$, hence kernel PCA reduces to PCA.

(2) $k(x,y) = [x_1^2, \sqrt{2}x_1x_2, x_2^2] \begin{bmatrix} y_1^2 \\ \sqrt{2}y_1y_2 \\ y_2^2 \end{bmatrix}$

$$= (x_1y_1 + x_2y_2)^2 = (x^Ty)^2$$

In general, $k(x,y) = (x^Ty)^n$ is known as the polynomial kernel.

(3) $k(x,y) = e^{-\frac{\|x-y\|^2}{\sigma^2}}$ is known as Gaussian Radial Basis Function (RBF) or Gaussian kernel.

Remark: If $\phi(x^{(1)}), \cdots, \phi(x^{(n)})$ are not centered, instead of using the sample matrix Φ, we use $\widetilde{\Phi} \overset{\text{def}}{=} \Phi H$ where H is the centering matrix. Then

$$\underbrace{\widetilde{K}}_{\text{centered kernel matrix}} = \widetilde{\Phi}^T \widetilde{\Phi} = H^T \Phi^T \Phi H = H \underbrace{K}_{\text{non-centered kernel matrix}} H$$

3.4.4. Algorithm: (Nonlinear / Kernel PCA)