Principal Component Analysis (PCA)

PCA finds projection directions for a multi-dimensional random vector such that each pair of projections are uncorrelated and each projection has maximal variance subject to these constraints.

1^{st} Principal Component

$(E\{x\}=0)$

Suppose $x \in \mathbb{R}^n$ is a zero-mean random vector with covariance $E\{xx^T\} = \Sigma$.

Find vector $w \in \mathbb{R}^n$ such that $y = w^T x$ has maximal variance, while keeping $\|w\|_2^2 = 1$.

$$\max_w \ Var(y) \quad \text{s.t.} \quad w^T w = 1$$

Equivalently, $\max_w \ E\{(y - E[y])^2\}$ s.t. $w^T w - 1 = 0$

Note that $E\{y\} = E\{w^T x\} = w^T E\{x\} = w^T 0 = 0$.

$\therefore \ Var(y) = E\{y^2\} = E\{(w^T x)^2\} = E[w^T x x^T w] = w^T \Sigma w$.

Therefore, we need to solve $\max_w \ w^T \Sigma w$ s.t. $w^T w - 1 = 0$

The Lagrangian is $\mathcal{L}(w, \lambda) = w^T \Sigma w - \lambda (w^T w - 1)$

At the optimal solution (of this equality constrained optimization problem), the gradient of the Lagrangian should be zero.

$$\frac{\partial \mathcal{L}(w,\lambda)}{\partial w} = 2\Sigma w - 2\lambda w = 0 \iff \boxed{\Sigma w = \lambda w}$$

$$\frac{\partial \mathcal{L}(w,\lambda)}{\partial \lambda} = -(w^T w - 1) = 0 \iff \boxed{w^T w = 1}$$

Clearly, w must be an eigenvector of Σ.

Noticing that $w^T \Sigma w = w^T(\lambda w) = \lambda w^T w = \lambda$, the optimal w is the eigenvector of Σ that corresponds to the largest eigenvalue of Σ.

Let $\Sigma = \lambda_1 \vec{q}_1 \vec{q}_1^T + \dots + \lambda_n \vec{q}_n \vec{q}_n^T$ be the spectral decomposition of Σ such that $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_n$ and $\vec{q}_i^T \vec{q}_j = \begin{cases} 1 & \text{if } i=j \\ 0 & \text{if } i \neq j \end{cases} \triangleq \delta_{ij}$ (Kronecker-δ).

Then the optimal solution is $w = q_1$ (dropping $\rightarrow$)*.

* for simplicity of notation

The 1st PC is $y_1 = q_1^T x$.

Note that $\mathrm{Var}(y_1) = q_1^T \Sigma q_1 = q_1^T(\lambda_1 q_1) = \lambda_1$, which is the largest eigenvalue of Σ.

2^{nd} Principal Component

Given $x \in \mathbb{R}^n$ with $E[x]=0$, $E[xx^T]=\Sigma$
and $y_1 = q_1^T x$

find $y = w^T x$ such that $\mathrm{var}(y)$ is max.
subject to y and y_1 being uncorrelated,
and $w^T w = 1$.

y and y_1 are uncorrelated when ~~$E[y]$~~

$$E[y y_1] = \underbrace{E[y]}_{=0} \underbrace{E[y_1]}_{=0}$$

$$\parallel$$

$$E[w^T x x^T q_1] = 0$$

i.e. y and y_1 are uncorrelated iff $w^T \Sigma q_1 = 0$

but $\Sigma q_1 = \lambda_1 q_1$, so $\lambda_1 w^T q_1 = 0$ or $w^T q_1 = 0$

After this simplification, we get the following equivalent optimization problem to solve.

$$\max_w \; w^T \Sigma w \quad \text{s.t.} \quad \left.\begin{aligned} w^T w - 1 &= 0 \\ w^T q_1 &= 0 \end{aligned}\right\} \text{2 constraints}$$

$$\mathcal{L}(w, \gamma_1, \gamma_2) = w^T \Sigma w - \gamma_1 (w^T w - 1) - \gamma_2 w^T q_1$$

Lagrangian Lagrange multipliers

Taking the gradient ... then equating to zero:

$$\frac{\partial \mathcal{L}}{\partial w} = 2\Sigma w - 2\gamma_1 w - \gamma_2 q_1 = 0 \qquad (\star)$$

$$\left.\begin{aligned} \frac{\partial \mathcal{L}}{\partial \gamma_1} &= -(w^T w - 1) = 0 \iff w^T w = 1 \\ \frac{\partial \mathcal{L}}{\partial \gamma_2} &= -w^T q_1 = 0 \iff w^T q_1 = 0 \end{aligned}\right\} \text{constraints}$$

Multiply $(\star)$ from left with w^T:

$$2w^T\Sigma w - 2\gamma_1 \underbrace{w^T w}_{=1} - \cancel{\gamma_2 \underbrace{w^T q_1}_{=0}} = \underbrace{\cancel{w^T 0}}_{0}$$

$$\Rightarrow 2w^T\Sigma w - 2\gamma_1 = 0 \Rightarrow \boxed{\gamma_1 = w^T \Sigma w}$$

Multiply $(\star)$ from left with q_1^T:

$$2\underbrace{q_1^T \Sigma w}_{=\lambda_1 \underbrace{q_1^T w}_{=0}} - 2\gamma_1 \underbrace{q_1^T w}_{=0} - \gamma_2 \underbrace{q_1^T q_1}_{=1} = \underbrace{\cancel{q_1^T 0}}_{0}$$

$$\Rightarrow \boxed{\gamma_2 = 0}$$

Substituting these back into $(\star)$:

$$2\Sigma w - 2(w^T\Sigma w)w = 0 \Rightarrow \Sigma w = \underbrace{(w^T \Sigma w)}_{\text{eigenvalue of } \Sigma} w$$

eigenvector of Σ.

At the solution, w is an eigenvector of Σ, but we know it is orthogonal to q_1 (recall $w^T q_1 = 0$). Also the eigenvalue corresponding to w is $w^T \underbrace{\Sigma w}_{\lambda_i w} = \lambda_i \underbrace{w^T w}_{=1} = \lambda_i$

where λ_i is one of the eigenvalues of Σ.

Since w must be $\perp$ to q_1 and maximize the variance of $w^T x$ (which is λ_i), w must be the eigenvector corresponding to λ_2.

$w = q_2$ (unit length second eigenvector)

2nd PC: $y_2 = q_2^T x$

Proceeding in this fashion, we can define and determine the 3rd, 4th, ..., nth PCs as follows:

$$y_3 = q_3^T x$$
$$y_4 = q_4^T x$$
$$\vdots$$
$$y_n = q_n^T x$$

In matrix form

$$\underbrace{\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}}_{y} = \underbrace{\begin{bmatrix} q_1^T \\ \vdots \\ q_n^T \end{bmatrix}}_{Q^T} x$$

In summary, for $x \in \mathbb{R}^n$ with $E[x]=0$, $E[xx^T]=\Sigma$

the principal components are

$$y = Q^T x$$

where $\Sigma = Q \Lambda Q^T$ with

$$Q = [q_1 \cdots q_n] \quad \text{and} \quad \Lambda = \begin{bmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_n \end{bmatrix}$$

orthogonal matrix ... diagonal matrix

in which $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n \geq 0$

Since Q is an orthogonal matrix, $y = Q^T x$ is, in general, a combination of the following 3 operations:

(1) rotation

(2) coordinate axis flip (negative sign)

(3) permutation

We can think of PCA as a coordinate rotation (and sign change, and permutations) that orders the dimensions in decreasing variance ordering.

Furthermore, note that $\text{trace}(\Sigma) = \lambda_1 + \lambda_2 + \dots + \lambda_n$ is the "total variance in x" and the first m PCs have a total variance of $\lambda_1 + \lambda_2 + \dots + \lambda_m$ where $m \leq n$. Therefore, the fraction of variance captured by the first m PCs is

$$0 \leq \frac{\lambda_1 + \dots + \lambda_m}{\text{trace}(\Sigma)} \leq 1.$$

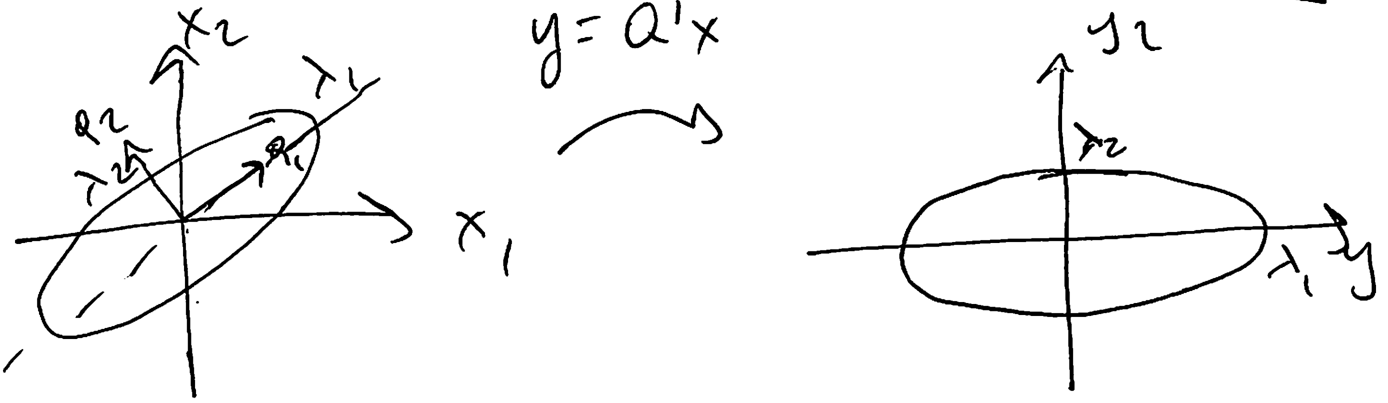
Finally, the matrix $W = [q_1 \cdots q_m] \in \mathbb{R}^{n \times m}$ can be shown to be an optimal solution for the following "data compression" problem:

$$\min_{W \in \mathbb{R}^{n \times m}} E\left[\| x - \underbrace{WW^T x}_{\hat{x}\text{: approximation of } x \text{ as in autoencoders}} \|_2^2 \right]$$

<u>Proof</u>: Involves showing that for all $W \in \mathbb{R}^{n \times m}$ the objective value is greater than or equal to what we achieve with $[q_1 \cdots q_m]$.

$\therefore$ Projecting the data to the first $m \leq n$ PCs achieves minimum MSE reconstruction performance.

Visualization of PCA using a 2D Gaussian



The largest variance direction in x (in this case x_1) becomes y_1. The second largest orthogonal direction in x becomes y_2...

The example above uses a 2D Gaussian for illustration, but the outcome is only dependent on the assumption that

Mean $E[x] = 0$

Covariance matrix $E[x x^T] = \Sigma$

so it is more general.