# PCA

We use notations from the following reference:

Goodfellow-et-al-2016,
Deep Learning,
Ian Goodfellow and Yoshua Bengio and Aaron Courville,
MIT press,
http://www.deeplearningbook.org
year = 2016

it's a convention; an hypothesis; in PCA

$D \in \mathbb{R}^{n \times \ell}$ decode matrix  (H) $D^T D = I_\ell$

$c \in \mathbb{R}^{\ell}$ code of a line

matrix dimensions provide valuable information for checking calculations

$X \in \mathbb{R}^{m \times n}$ data

Reconstruction function: $r : \mathbb{R}^n \to \mathbb{R}^n$
$x \mapsto DD^T x$

Our aim is to minimize $\displaystyle\sum_{j=1}^{n} \sum_{i=1}^{m} \left( x_j^{(i)} - (DD^T x^{(i)})_j \right)^2 = (*)$

⚠ Be careful: $(DD^T x^{(i)})_j \neq DD^T x_j^{(i)}$

$$(*) = \sum_j^n \sum_i^m X_{ij}^2 - 2X_{ij}(DD^T X_{i,:}^T)_j + (DD^T X_{i,:}^T)_j^2$$

So we are looking for:

$$D^* = \underset{\substack{D \\ \text{s.t. } D^T D = I_\ell}}{\operatorname{argmin}} \; -2 \sum_j^n \sum_i^m X_{ij} (DD^T X_{i,:}^T)_j + \sum_j^n \sum_i^m (DD^T X_{i,:}^T)_j^2$$

always explicit matrix products

$$① \sum_j^n \sum_i^m X_{ij} (DD^T X_{i,:}^T)_j = \sum_j^n \sum_i^m X_{ij} \sum_{k,l} D_{jk} D_{lk} X_{il}$$

put the X's in front because we'll differentiate /D later

$$= \sum_j^n \sum_i^m \sum_{k,l} X_{ij} X_{il} D_{jk} D_{lk}$$

organize indexes to prepare the product: $j \to i \to l \to k \to j$

$(X^T)_{ji} X_{il} D_{lk} (D^T)_{kj}$

$$= \sum_j^n (X^T X DD^T)_{jj}$$

use definition of matrix trace!

$$= Tr(X^T X DD^T)$$

② Exercise: I let you do the same with

$$\sum_j^n \sum_i^m (DD^T X_{i,:}^T)_j^2$$

We obtain $Tr(X^T X D \underbrace{D^T D}_{(H)\ = I_l} D^T) = Tr(X^T X DD^T)$

$\to$ So $\displaystyle D^* = \underset{\substack{D \\ s.t.\ D^T D = I_l}}{\operatorname{argmin}} - Tr(X^T X DD^T)$

$\displaystyle = \underset{D^T D = I_l}{\operatorname{argmax}} Tr(X^T X DD^T)$

We write the lagrangian $\Lambda \in \mathbb{R}^{l \times l}$

$$\mathcal{L}(D, \Lambda) = Tr(X^T X DD^T) - Tr(\Lambda(D^T D - I_l))$$

It is easy to see that in the formula $D^T D = I_l$, there are $l^2$ constraints by only $\frac{l(l+1)}{2}$ useful constraints because the formula is entirely symmetric.

Therefore, we could decide that $\Lambda$ is upper triangular, but it seems more interesting to choose $\Lambda$ symmetric: $\Lambda^T = \Lambda$

It's a choice.

Now let's compute $D_D\mathcal{L}(D)(H)$

$$\begin{aligned}
\mathcal{L}(D+H, \Lambda) &= Tr\left(X^T X (D+H)(D+H)^T\right) - Tr\left(\Lambda\left((D+H)^T(D+H) - I_\ell\right)\right) \\
&= \mathcal{L}(D,\Lambda) + Tr(X^T X H D^T) + Tr(X^T X D H^T) \\
&\quad - Tr(\Lambda H^T D) - Tr(\Lambda D^T H) + o(tr(H)) \\
&= \mathcal{L}(D,\Lambda) + Tr(D^T X^T X H) + Tr(H D^T X^T X) \\
&\quad - Tr(D^T H \Lambda^T) - Tr(\Lambda D^T H) + o(tr(H)) \\
&= \mathcal{L}(D,\Lambda) + Tr(D^T X^T X H) + Tr(D^T X^T X H) \\
&\quad - Tr(\Lambda^T D^T H) - Tr(\Lambda D^T H) + o(tr(H))
\end{aligned}$$

(KKT)

So
$$D^{*T} X^T X = \Lambda^T D^{*T}$$
$$X^T X D^* = D^* \Lambda$$
$$\Lambda = D^{*T} X^T X D^*$$

- Let $\lambda$ be an eigenvalue of $X^T X$ and $v$ an eigenvector for this eigenvalue

$$D^{*T} X^T X v = \Lambda^T D^{*T} v \quad \text{so} \quad \lambda D^{*T} v = \underbrace{\Lambda^T}_{=\Lambda} D^{*T} v$$

So $v \in \operatorname{Ker} D^{*T}$ or $\lambda \in Sp(\Lambda)$ $\qquad$ $= \{0\}$ because $D^{*T} D^* = I_\ell$

- Conversely $\tilde{\lambda}, \tilde{v}$ eigenvalue / eigenvector of $\Lambda \Rightarrow \tilde{v} \in \operatorname{Ker} D^*$ or $\tilde{\lambda} \in Sp(X^T X)$

In the meantime, we see that we are solving

$$D^* = \underset{\substack{\Lambda D^T = D^T X^T X \\ D^T D = I_\ell}}{\operatorname{argmax}} Tr(\Lambda) \qquad \text{with } Sp(\Lambda) \subset Sp(X^T X)$$

But the trace is the sum of the eigenvalues

Therefore, one best $\Lambda$ is the diagonal matrix of the first $l$ biggest eigenvalues of $X^TX$

not necessarily unique

achieves

$$Tr(\Lambda) = \max_{D^TD = I_l} Tr(X^TXDD^T)$$

Let's write the eigenvalue equation for $X^TX$ :

To remind $Q \in \mathbb{R}^{n \times n}$
$Q^TQ = I_n$ ] orthogonal matrix

$$\underset{\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_n}{Diag(\lambda_1, \dots, \lambda_n)} = Q X^TX Q^T$$

Q orthogonal matrix $\in \mathbb{R}^{n \times n}$ such that $Q^T_{i,:}$ is a unit vector corresponding to $\lambda_i$

Finally we just have to write $D^{*T}$ as the first $l$ rows of Q.