Principal Component Analysis

1 How it works

1.1 Just a little more theory to better understand

Let $X \in \mathcal{M}_{n,p}(\mathbb{R})$ be a matrix of our dataset, where n is the number of samples, each composed of p parameters. We want to find a matrix

$$R \in \mathcal{M}_{n,k}(\mathbb{R}), \quad k < p.$$

The first step is to compute $X^\top X$. Indeed, if we take a random direction $w \in \mathbb{R}^p$ such that $\|w\| = 1$, then the data projected on this axis is given by Xw. Because we are looking for the direction that keeps the most information, which means maximizing the variance of the projection, we are basically looking for

$$\begin{aligned}\mathrm{Var}(Xw) &= \mathbb{E}[(Xw)^2] - \mathbb{E}[Xw]^2 \\ &= \frac{1}{n}\|Xw\|^2 \\ &\propto w^\top X^\top X w.\end{aligned}$$

If necessary, we define

$$\tilde{X}_{(i)} = X_{(i)} - \mathrm{mean}(X_{(i)}),$$

so we can assume that $\mathbb{E}[X] = 0$. This is in fact the very first thing to do before computing $X^\top X$.

In order to find our first direction, we need to solve

$$\arg\max_{\|w\|=1} \left\{ w^\top X^\top X w \right\} = \arg\max \left\{ \frac{w^\top X^\top X w}{w^\top w} \right\}.$$

We recognize the Raylight quotient. It is known that for a matrix such as $X^\top X$, this quantity is maximized for the largest eigenvalue

$$\lambda_{\max} \in \mathrm{Sp}(X^\top X),$$

which occurs when $w = v_{\max}$, with $v_{\max}$ the eigenvector associated with $\lambda_{\max}$.

Thus, we found our first direction $w_1 = v_{\max}$. Our intuition tells us that we now have to find the second vector by using the exact same idea, but on a new matrix $\tilde{X}$ where all the information given by w_1 has been removed:

$$\begin{aligned}\tilde{X}_k &= X - \sum_{s=1}^{k-1} X w_s w_s^\top \\ &= X \left(I - \sum_{s=1}^{k-1} w_s w_s^\top \right) \\ &= X \left(I - P_{k-1} \right)\end{aligned}$$

Where P_{k-1} is the orthogonal projector on $\mathrm{Span}(v_1, \ldots, v_{k-1})$. So

$$
\begin{aligned}
i < k &\Rightarrow \tilde{X}_k v_i = 0 \\
i \geq k &\Rightarrow \tilde{X}_k v_i = X v_i = \lambda_i v_i
\end{aligned}
$$

This means that $\mathrm{Sp}(\tilde{X}_k) = \{0, \ldots, 0, \lambda_k, \ldots, \lambda_p\}$ so looking for the direction on the unit circle that minimize the varience we end up with the same Rayleigh quotient as before. Assuming that $0 \leq \lambda_k \leq \cdots \leq \lambda_p$ we get $w_k = v_k$

$$R = (X w_i)_i = (X v_i)_i$$