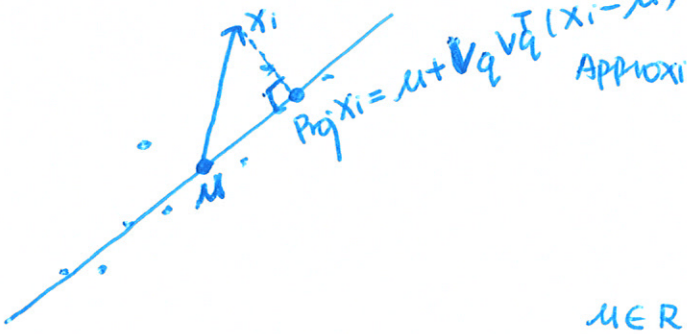
Principal Component Analysis



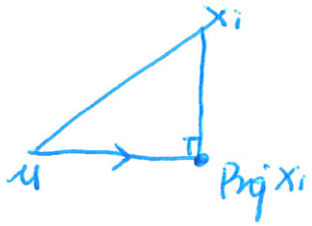
Approximate data $x_1, \cdots, x_N \in R^p$ by the plane

$$f(\lambda) = \mu + V_q \lambda$$

$$\mu \in R^p \qquad V_q \in R^{p \times q}$$

$$\min_{\text{plane}} \sum_{i=1}^{N} \| x_i - \text{Proj}\ x_i \|^2.$$

$$= \min_{\mu, V_q} \sum_{i=1}^{N} \| x_i - \mu - V_q V_q^T (x_i - \mu) \|^2$$



$$= \min_{\mu, V_q} \sum_{i=1}^{N} \|x_i - \mu\|^2 - \| V_q V_q^T (x_i - \mu) \|^2.$$

Step 1: minimize $\sum_{i=1}^{N} \|x_i - \mu\|^2 \Rightarrow \mu = \frac{1}{N} \sum_{i=1}^{N} x_i$ Mean.

Step 2. Find V_q to maximize $\sum_{i=1}^{N} \| V_q V_q^T (x_i - \mu) \|^2$.

Notice $\| V_q V_q^T (x_i - \mu) \|^2 = \| V_q^T (x_i - \mu) \|^2$

$$= \text{Trace}\left[V_q^T (x_i - \mu)(x_i - \mu)^T V_q \right]$$

$$\max_{V_q} \sum_{i=1}^{N} \text{Trace}\left[V_q^T (x_i - \mu)(x_i - \mu)^T V_q \right]$$

$$\Rightarrow \max_{V_q} \text{Trace}\left[V_q^T \left(\sum_{i=1}^{N} (x_i - \mu)(x_i - \mu)^T \right) V_q \right]$$

Let $A = \sum_{i=1}^{N} (x_i - \mu)(x_i - \mu)^T$ $\quad A^T = A$.

$\max_{V_q} \text{Trace}\left(V_q^T A V_q\right) = \lambda_1 + \lambda_2 + \cdots + \lambda_q$.

top q eigenvalues of A.

if $V_q = \left[\vec{u}_1 \; \vec{u}_2 \; \cdots \; \vec{u}_q\right]$

the eigenvectors associated with $\lambda_1 \; \lambda_2 \cdots \lambda_q$.

Equivalently: Let $X = \left[x_1 - \mu \;\; x_2 - \mu \;\; \cdots \;\; x_N - \mu\right]$

SVD of X: $X = U D V^T$ → right singular vectors.

↓ left singular vectors

$V_q = \left[\vec{u}_1 \cdots \vec{u}_q\right]$

the left singular vectors associated with the largest q singular values.