Principal Component Analysis (PCA)

Jacob Goldberger

Given orthogonal vector set $b_1, ..., b_k \in R^d$,

$$\hat{x} = \sum_{j=1}^{k} (b_j^T x) b_j$$

is the projection of $x \in R^d$ onto the sub-space spanned by $b_1, ..., b_k$. The low-dimensionality representation of x is

$$y = \begin{pmatrix} b_1^T x \\ \vdots \\ b_k^T x \end{pmatrix} \in R^k.$$

The reconstruction error of a given input set $x_1, ..., x_n \in R^d$ is $\sum_i \|x_i - \hat{x}_i\|^2$. The PCA algorithm finds a linear subspace that induces the minimal reconstruction error (or equivalently that maximizes $\sum_i \|y_i\|^2$).

PCA Algorithm (linear):

- Define $A = \frac{1}{n} \sum_i x_i x_i^T$.
- Let $b_1, ..., b_d \in R^d$ the eigen-vectors of A with eigen-values $\lambda_1 \geq \lambda_2 \geq , ..., \lambda_d \geq 0$.
- The optimal subspace is the one spanned by $b_1, ..., b_k$.

PCA Algorithm (affine):

- Define $\mu = \frac{1}{n} \sum_i x_i$
- Apply linear PCA algorithm on $x_1 - \mu, ..., x_n - \mu$.