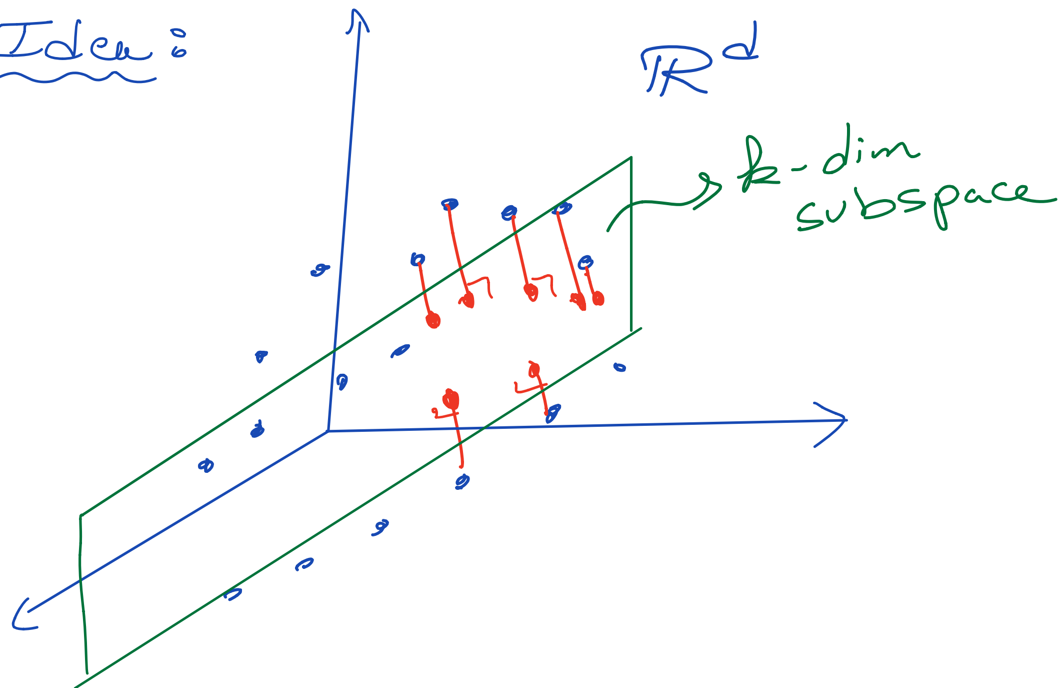
Goals:

- Review of some linear algebra
- Deriving Principal Component Analysis.

Idea: Given some data set $\{x_1, x_2, \dots, x_p\} \subseteq \mathbb{R}^d$, want to find a linear projection to $k < p$ dimensions while preserving key properties of the data set.

Why?
- Visualization ($k = 2$ or 3)
- "Interpretation"

Idea:



Let $X = \{x_1, \ldots, x_p\}$ and define the sample mean

$$\mu_p = \frac{1}{p} \sum_{i=1}^{p} x_i$$

and sample covariance

$$\Sigma_p = \frac{1}{p-1} \sum_{i=1}^{p} (x_i - \mu_p)(x_i - \mu_p)^T$$

Homework: Prove that μ_p & Σ_p are <u>unbiased</u> estimators for the mean & variance of the distribution from which each x_i is drawn independently.

Now, assuming $v_1, v_2, \ldots, v_k$ form an orthonormal basis for a k-dim. subspace, we want to minimize

$\in \mathbb{R}^{d\times k}$ $\in \mathbb{R}^k$, vector of coefficients

$$\min_{\substack{\mu, V \\ V^TV = I}} \sum_{i=1}^{p} \| x_i - (\mu + V\beta_i) \|_2^2$$

affine shift $\in \mathbb{R}^d$ $V = [V_1 | V_2 | \cdots V_k]$, $V_i \in \mathbb{R}^d$

(I) Optimal μ:

$$\nabla_\mu (\qquad) = 0 \Rightarrow$$

$$\nabla_\mu \sum_{i=1}^{p} (x_i - \mu_{opt} - V\beta_i)^T (x_i - \mu_{opt} - V\beta_i) = 0$$

so $$\sum_{i=1}^{p} (x_i - \mu_{opt} - V\beta_i) = 0$$

& $$\mu_{opt} = \frac{1}{n} \sum_{i=1}^{p} x_i + \frac{1}{n} V \underbrace{\sum_{i=1}^{p} \beta_i}$$

assume, WLOG that $\sum \beta_i = 0$

$$\Rightarrow \mu_{opt} = \frac{1}{n} \sum_{i=1}^{p} x_i$$

sample mean

(II) Optimal β_i, $i=1,\dots,P$

$$\min_{\beta_i} \sum_{i=1}^{P} \| x_i - \mu_{opt} - V\beta_i \|_2^2$$

We can solve for each β_j separately (obj. function is separable)

$\Rightarrow$ want to minimize $\| x_i - \mu_{opt} - V\beta_i \|_2^2$

$$= (x_i - \mu_{opt} - V\beta_i)^T (x_i - \mu_{opt} - V\beta_i)$$

$$\nabla_{\beta_i} = 0 \Rightarrow V^T(x_i - \mu_{opt} - V\beta_i) = 0 \quad \text{(chain rule)}$$

$$\Rightarrow \beta_i = V^T(x_i - \mu_{opt})$$

(III) Optimal V:

want
$$\min_{V,\, V^TV=I} \sum_{i=1}^{P} \| x_i - \mu_{opt} - V\beta_i^{opt} \|_2^2$$

$$\Leftrightarrow \min_{V,\, V^TV=I} \sum_{i=1}^{P} \underbrace{\| x_i - \mu_{opt} - VV^T(x_i - \mu_{opt}) \|_2^2}_{\underbrace{\| (I - VV^T)(x_i - \mu_{opt}) \|_2^2}}$$

$$= \sum_{i=1}^{P} \underbrace{\| x_i - \mu_{opt} \|^2}_{\text{ind. of } V} - (x_i - \mu_{opt})^T VV^T (x_i - \mu_{opt})$$

So we want to maximize (over $V, V^TV = I$)

$$\sum_{i=1}^{p} (x_i - \mu_{opt})^T VV^T (x_i - \mu_{op})$$

$$= \sum_{i=1}^{p} Tr\left((x_i - \mu_{opt})^T VV^T (x_i - \mu_{op})\right)$$

↳ trace : $\text{trace}(X) = \sum_i X_{ii}$

$$= \sum_{i=1}^{p} Tr\left(V^T (x_i - \mu_{opt})(x_i - \mu_{opt})^T V \right)$$

↳ cyclicity of trace : $Tr(AXB) = Tr(BAX)$

$$= Tr \sum_{i=1}^{p} \left(V^T (x_i - \mu_{opt})(x_i - \mu_{opt})^T V \right)$$

↳ linearity of trace : $\left(Tr(A+B) = Tr(A) + Tr(B) \right)$

$$= Tr\ V^T \underbrace{\sum_{i=1}^{p} (x_i - \mu_{opt})(x_i - \mu_{opt})^T}_{= (p-1)\Sigma_p} V$$

$$= (p-1)\ Tr\left(V^T \Sigma_p V \right)$$

So we want to maximize

$$\max_{\substack{V^TV=I \\ V\in\mathbb{R}^{d\times k}}} \mathrm{Tr}\left(V^T \Sigma_p V\right)$$

$\Rightarrow$ $\boxed{V = \text{matrix of } k \text{ leading eigenvectors of } \Sigma_p}$

Why? Linear algebra

Alternative Interpretation: PCA finds the directions that "preserve the most variance."

$\Leftrightarrow$ want $V = \left[v_1 \mid v_2 \mid \cdots \mid v_k \right]$, $V^TV = I$

such that $\left\{ \underbrace{\begin{pmatrix} v_1^T x_i \\ \vdots \\ v_k^T x_i \end{pmatrix}}_{\in \mathbb{R}^k} \right\}_{i=1}^{P}$ has maximal variance

$\{V^T x_k\} =$

$\iff$ want $\displaystyle\max_{\substack{V:V^TV=I \\ \in \mathbb{R}^{d\times k}}} \underbrace{\sum_{i=1}^{p} \| V^T x_i - \frac{1}{p}\sum_{j=1}^{p} V^T x_j \|_2^2}$

$$= \sum_{i=1}^{p} \| V^T (x_i - \mu_p) \|_2^2$$

$$= \mathrm{Tr}\left(V^T \Sigma_p V \right)$$

as before.

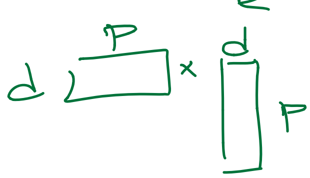
~ x ~

Computational Complexity:

Cost of computing PCA via k leading eigenvectors of Σ_p via SVD =

cost of computing Σ_p + cost of SVD

$$= O\big(\underbrace{d^2 p} + \underbrace{kdp}_{\text{can be reduced via randomized algorithms}} \big)$$



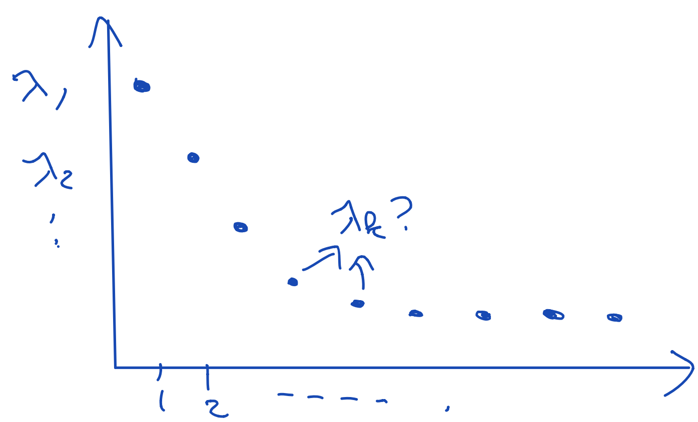
Choice of k: for dimensionality

reduction is usually done heuristically.

- Let $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_k \geq \cdots \geq 0$ be the eigenvalues of Σ_p, select smallest k so that $\sum_{i=1}^{k} \lambda_i / Tr(\Sigma_p)$ is "large enough".

$$Tr(\Sigma_p) = \sum_{i=1}^{d} \lambda_i$$

- Or pick k at the "elbow" of a scree plot.



Extensions to Classical PCA:

- Nonlinear dimensionality reduction
  e.g.: • kernel PCA (may discuss towards end of course)
  • principal manifold embeddings

- Non-negative matrix factorization
  want
  $$\min_{\substack{L,R \\ L_{ij} \geq 0 \\ R_{ij} \geq 0}} \| X - LR \|_F$$

- Multilinear PCA (X is a tensor)

- Robust PCA ($X = \underbrace{L}_{\text{Low-rank}} + \underbrace{S}_{\text{sparse}}$)

$$\min \; \underbrace{\|L\|_*}_{\text{nuclear norm}} + \lambda \underbrace{\|S\|_1}_{\ell_1\text{-norm of vectorized } S}$$
s.t. $L + S = M$

- many others

Properties / Drawbacks:

PCA is data dependent, so if an encoder performs dimensionality reduction via PCA, i.e., computes the map

$$\mathbb{R}^d \supset X = [x_1 | x_2 | \text{---} | x_p] \longmapsto V^T X$$

$k \times d$
PCA matrix

a decoder would need to know V to recover an approximation of any x_i.

But V depends on X, hence x_i.

A solution to this is to use a fixed transformation that takes

data-independent

into account general properties of the signal model, but not the data itself.

$\Longrightarrow$ Discrete cosine transform, wavelets, etc....