**Goals:**

- Review of some linear algebra
- Deriving Principal Component Analysis

**Idea:** Given some data set \( \{x_1, x_2, \ldots, x_p \} \subseteq \mathbb{R}^d \), 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
Let \( X = \{ x_1, \ldots, x_p \} \) and define the sample mean

\[
\bar{x}_p = \frac{1}{p} \sum_{i=1}^{p} x_i
\]

and sample covariance

\[
S_p = \frac{1}{p-1} \sum_{i=1}^{p} (x_i - \bar{x}_p)(x_i - \bar{x}_p)^T
\]

Homework: Prove that \( \bar{x}_p \) & \( S_p \) are unbiased estimators for the mean & variance of the distribution from which each \( x_i \) is drawn independently.
Now, assuming \( V, V_2, \ldots, V_k \) form an orthonormal basis for a \( k \)-dim. subspace, we want to minimize

\[
\min_{\mu, V} \sum_{i=1}^{P} \left\| x_i - (\mu + V\beta_i) \right\|^2
\]

subject to \( V^T V = I \) and \( \beta \in \mathbb{R}^k \), vector of coefficients.

(II) Optimal \( \mu \):

\[
\nabla_{\mu} f(\mu) = 0 \Rightarrow
\]

\[
\nabla_{\mu} \sum_{i=1}^{P} (x_i - \mu - V\beta_i)^T (x_i - \mu - V\beta_i) = 0
\]

So

\[
\sum_{i=1}^{P} (x_i - \mu_{opt} - V\beta_i) = 0
\]

\& \quad \mu_{opt} = \frac{1}{n} \sum_{i=1}^{P} x_i + \frac{1}{n} \sum_{i=1}^{P} V\beta_i
\]

assume. WLOG that \( \sum \beta_i = 0 \)

\[\implies \quad \mu_{opt} = \frac{1}{n} \sum_{i=1}^{P} x_i \quad \text{sample mean} \]
(II) Optimal $\beta_i$, $i=1, \ldots, p$

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

We can solve for each $\beta_i$ separately (obj. function is separable)

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

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

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

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

(III) Optimal $V$:

$$\textbf{Want} \quad \min_{V} \sum_{i=1}^{p} \| x_i - \mu_{opt} - V\beta_{i_{opt}} \|^2$$

$$\Leftrightarrow \min_{V} \sum_{i=1}^{p} \| x_i - \mu_{opt} - V V^T (x_i - \mu_{opt}) \|^2$$

$$\Leftrightarrow \min \| (I - V V^T) (x_i - \mu_{opt}) \|^2$$

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

ind. of $V$
so we want to maximize (over $V, V^T V = I$) 

$$
\sum_{i=1}^{P} (x_i - \mu_{opt})^T V V^T (x_i - \mu_{opt})
$$

$$
= \sum_{i=1}^{P} \text{trace} \left( (x_i - \mu_{opt})^T V V^T (x_i - \mu_{opt}) \right)
$$

(by trace property: $\text{trace}(X) = \sum_{i} X_{ii}$)

$$
= \sum_{i=1}^{P} \text{trace} \left( V^T (x_i - \mu_{opt})^T (x_i - \mu_{opt})^T V \right)
$$

(by cyclicity of trace: $\text{trace}(AXB) = \text{trace}(BAX)$)

$$
= \text{trace} \left( \sum_{i=1}^{P} (V^T (x_i - \mu_{opt})^T (x_i - \mu_{opt})^T V \right)
$$

(by linearity of trace: $\text{trace}(A + B) = \text{trace}(A) + \text{trace}(B)$)

$$
= \text{trace} \left( V^T \sum_{i=1}^{P} (x_i - \mu_{opt})^T (x_i - \mu_{opt})^T V \right)
$$

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

$$
= (P-1) \text{trace} (V^T \Sigma_V V)
$$
So we want to maximize

$$\max \quad \text{Tr} \left( V^T \Sigma_p V \right) \quad \text{subject to} \quad V^T V = I, \quad V \in \mathbb{R}^{d \times k}$$

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

Why? Linear algebra

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

$$\Rightarrow \quad \text{want } V = \left[ V_1 \mid V_2 \mid \ldots \mid V_k \right], \quad V^T V = I$$

such that \( \left\{ \left( \begin{array}{c} v^T_i x_i \\ \vdots \\ v^T_k x_i \end{array} \right) \right\}^T \) has maximal variance.

\[
\left( \begin{array}{c} v^T_1 x_i \\ \vdots \\ v^T_k x_i \end{array} \right)_{i=1}^n \in \mathbb{R}^k
\]
\begin{align*}
&\text{(3) Want max } \max_{V: V^TV = I} \sum_{i=1}^{p} \| V^T x_i - \frac{1}{p} \sum_{i'=1}^{p} V^T x_{i'} \|_2^2 \\
&\quad = \sum_{i=1}^{p} \| V^T (x_i - \mu_P) \|_2^2 \\
&\quad = \text{Tr} (V^T \Sigma_P V) \\
&\text{as before}
\end{align*}

---

**Computational Complexity:**

Cost of computing \( \text{PCA via } k \) leading eigenvectors of \( \Sigma_P \) via SVD:

\[
\text{cost of Computing } \Sigma_P \text{ via SVD}
\]

\[= O\left( d^2 p + kd p \right) \]

can be reduced via randomized algorithms

Choice of \( k \) is for dimensionality
reduction is usually done heuristically.

- Let \( \lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_k \geq \ldots \geq 0 \)

be the eigenvalues of \( \Sigma_p \).

select \( k \) so that \( \frac{\sum_{i=1}^{k} \lambda_i}{\text{Tr} (\Sigma_p)} \)

is smallest

is "large enough".

- Or pick \( k \) at the "elbow"

of a scree plot.

\[
\begin{array}{c}
\lambda_1 \\
\lambda_2 \\
\vdots
\end{array}
\]

\[
\begin{array}{c}
1 \\
2 \\
\ldots
\end{array}
\]
Extensions to Classical PCA:

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

- Non-negative matrix factorization
  \[
  \min_{L,R} \| X - LR \|_F
  \quad \text{subject to}
  \begin{align*}
  L & \geq 0 \\
  R & \geq 0
  \end{align*}
  \]

- Multilinear PCA (X is a tensor)

- Robust PCA (\( X = L + S \))
  \[
  \begin{align*}
  \min_{L,S} & \| L \|_* + \lambda \| S \|_1 \\
  \text{s.t.} & \quad L + S = M
  \end{align*}
  \]
  - \( \ell_1 \)-norm of vectorized \( S \)
  - Nuclear norm

- Many others
Properties / Drawbacks:

PCA is data dependent, so if an encoder performs dimensionality reduction via PCA, i.e., computes the map
\[ \mathbb{R}^d \times \mathbb{R}^d \times \cdots \times \mathbb{R}^d \to \mathbb{R}^k, \]
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.

\[ \Rightarrow \text{Discrete cosine transform, wavelets, etc...} \]