

Lecture 15

03/15/2022

- The Curse of Dimensionality

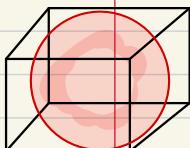
- * Imagine you want a car, and you're very picky
 - ↳ for each choice, there's only 50% chance of finding a car at the dealership.

$$\begin{aligned}\Rightarrow P_{\text{car}} &= P_{\text{red}} \times P_{\text{good mileage}} \times P_{\text{leather seats}} \times P_{\text{sunroof}} \\ &= (0.5)^4 \\ &= 0.0625 \quad \} \text{only } \sim 6\% !\end{aligned}$$

- * Another way of seeing this ...

- ↳ imagine designing a suite of simulations where you want to cover the full space of initial conditions over many dimensions.

- ⇒ don't distribute over unit hypersphere of initial conditions!



$$f_D = \frac{V_D(r)}{(2r)^D} = \frac{\pi^{D/2}}{D \cdot 2^{D-1} \Gamma(\frac{D}{2})}$$

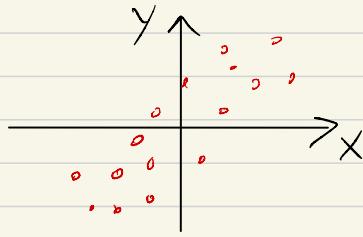
$\rightarrow 0$ as $D \uparrow$

 PHYSICALLY... PCA gives you the directions along which data varies, and ranks those directions by their eigen-value.

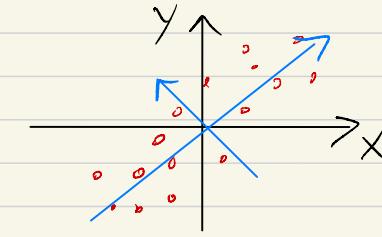
- Principal Component Analysis (PCA)

↳ find linear combinations of features that describe directions of maximal variance.

e.g



X, Y are measured features



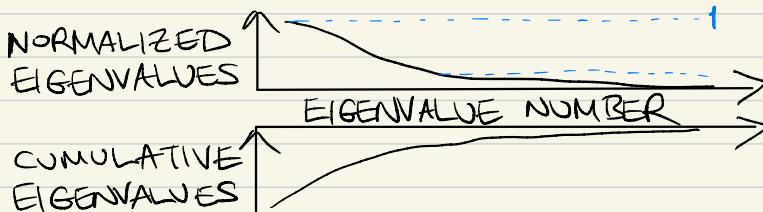
New axes are linear combinations

* EIGEN-ANALYSIS \Rightarrow principal components are eigen-vectors.

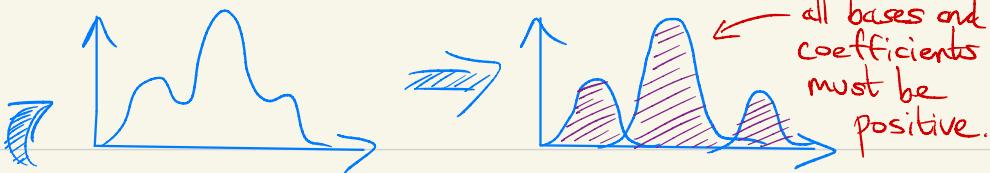
* Great for taking data with huge numbers of features, and finding the most important combinations of features.

↳ DATA COMPRESSION

* Find a new (smaller) basis that better describes the data.



↳ SCREE PLOTS tell you how many important principal components there are.



- Non-negative Matrix Factorization (NMF)

↳ what if you know that your basis functions and co-efficients have to be positive?

e.g. amplitudes of GMMs to describe a PDF.
→ PDFs have to be ≥ 0 .

$\Rightarrow X$ has all +ve elements

$$\Rightarrow X = WY$$

minimize $\|X - WY\|^2$

to find the basis matrix, W , and +ve coefficients, Y .

- Independent Component Analysis (ICA)

↳ separate signals components that are statistically-independent.

e.g. isolating different voices at a party.

* ICA relies on the **CENTRAL LIMIT THEOREM**
i.e. adding a bunch of independent random signals should produce noise with **GAUSSIAN STATISTICS**.

$\Rightarrow \therefore$ let's separate out the signals to be as individually non-Gaussian as possible.