CS 540 Introduction to Artificial Intelligence
Statistics & Math Review

Yudong Chen
University of Wisconsin-Madison

Sep 21, 2021

Announcements

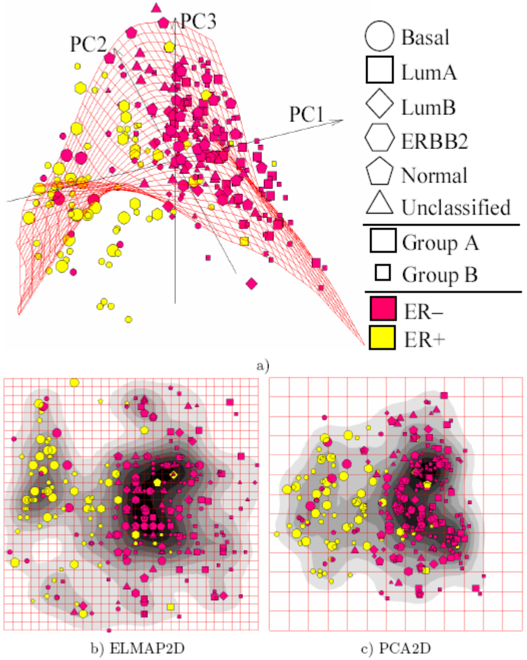
- **Homeworks**:
  - HW2 due Tuesday---get started early!
- Class roadmap:

Tuesday, Sep 14	Probability
Thursday, Sep 16	Linear Algebra and PCA
Tuesday, Sep 21	**Statistics and Math Review**
Thursday, Sep 23	Introduction to Logic
Tuesday, Sep 28	Natural Language Processing

Fundamentals

Outline

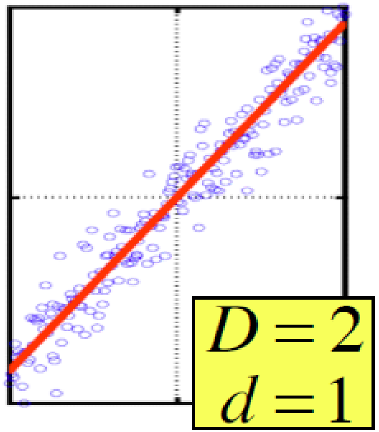
- Finish last lecture: **PCA**
- Review of probability
- Statistics: sampling & estimation

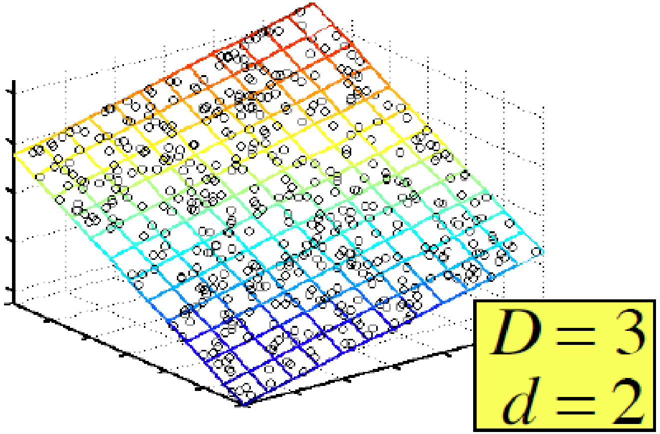


Wikipedia

Principal Components Analysis (**PCA**)

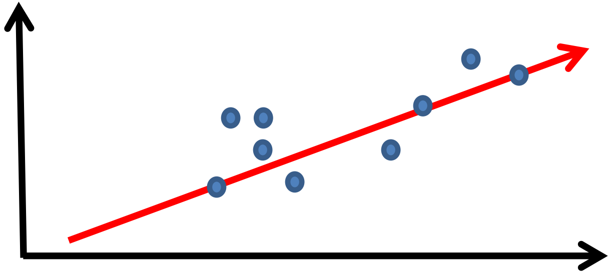
- A type of dimensionality reduction approach
  - For when data is **approximately lower dimensional**
- Goal: find a low-dimensional subspace
  - Will project to this subspace; want to minimize loss of information





Principal Components Analysis (**PCA**)

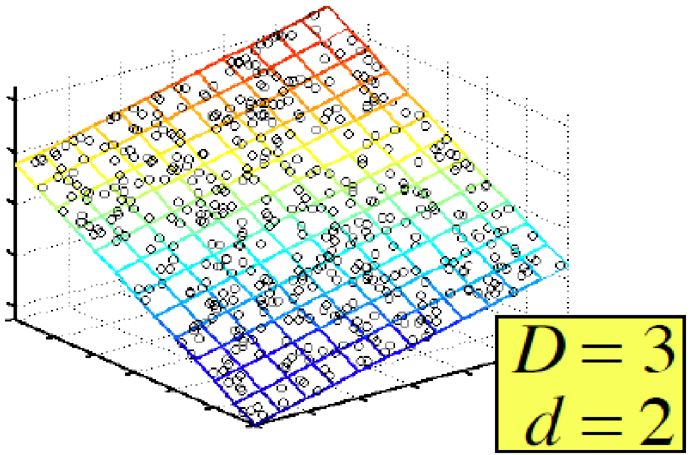
- From 2D to 1D:
  - Find a $v_1 \in \mathbb{R}^d$ so that we maximize "variability"



  - New representations are along this vector (1D!)

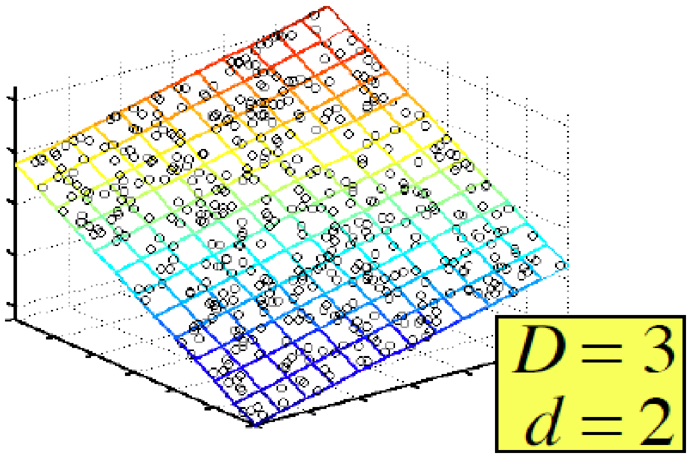
Principal Components Analysis (**PCA**)

- From *d* dimensions to *r* dimensions:
  - Sequentially get orthogonal vectors $v_1, v_2, \ldots, v_r \in \mathbb{R}^d$
  - Maximize variability when projecting to them
  - The vectors are the **principal components**



PCA Setup

- **Inputs**
  - Data: $x_1, x_2, \ldots, x_n, \;\; x_i \in \mathbb{R}^d$
  - Can arrange into $X \in \mathbb{R}^{n \times d}$
  - **Centered**! $\frac{1}{n}\sum_{i=1}^{n} x_i = 0$
- **Outputs**
  - $v_1, v_2, \ldots, v_r \in \mathbb{R}^d$
    (principle components, orthogonal)



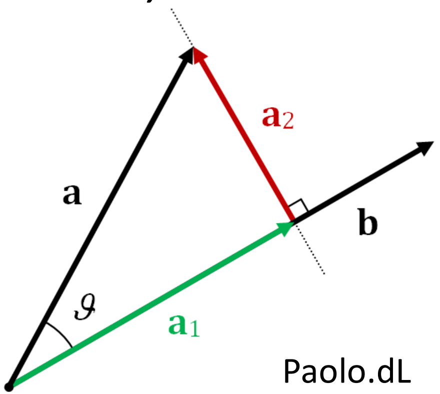
PCA Setup

- Want directions (unit vectors) so that projecting data maximizes variance
  - What's projection? To project *a* onto unit vector *b*,

$$\langle a, b \rangle b$$

$\langle a, b \rangle$ ↑ Length

b ← Direction



Paolo.dL

  - Variance of projection:

$$\sum_{i=1}^{n} \langle x_i, v \rangle^2 = \|Xv\|^2$$

PCA First Step

- First component:

$$v_1 = \arg\max_{\|v\|=1} \sum_{i=1}^{n} \langle v, x_i \rangle^2$$

$$= \arg\max_{\|v\|=1} \|Xv\|^2$$

PCA: k^{th} step

- Once we have *k-1* components, compute

$$\hat{X}_k = X - \sum_{i=1}^{k-1} X v_i v_i^T$$

Deflation

- Then do the same thing

$$v_k = \arg\max_{\|v\|=1} \|\hat{X}_k w\|^2$$

- Deflation ensures v_k is orthogonal to $v_1, \ldots, v_{k-1}$

PCA: Connection to Eigenvectors

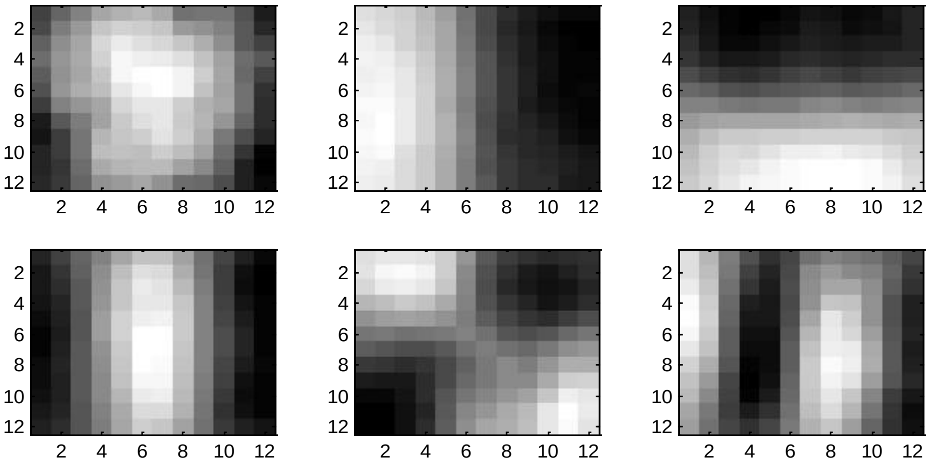
- v_k is the k^{th} eigenvector of $\frac{1}{n}X^TX$
  - Proof: linear algebra! (omitted)
- $\frac{1}{n}X^TX \in \mathbb{R}^{d\times d}$ is sample covariance matrix of data
  - When data is centered (has 0 mean)
- Therefore, PCA can be done via eigendecomposition of sample covariance

Application: Image Compression

- Original image:
- Divide into 12x12 patches
- Each patch is a 144-D vector x_i
- Want to reduce to 6-D

Application: Image Compression

- 6 most important components (as an image)



Application: Image Compression

- Project to 6D,

Compressed

Original

Break & Quiz

Q 1.1: What is the projection of $[1\ 2]^T$ onto $[0\ 1]^T$?

- A. $[1\ 2]^T$
- B. $[-1\ 1]^T$
- C. $[0\ 0]^T$
- D. $[0\ 2]^T$

Break & Quiz

Q 1.2: We wish to run PCA on 10-dimensional data in order to produce *r*-dimensional representations. Which is the most accurate (least loss of information)?

- A. $r = 3$
- B. $r = 9$
- C. $r = 10$
- D. $r = 20$

Probability Review: Outcomes & Events

- Outcomes: possible results of an **experiment**
- **Events:** subsets of outcomes we're interested in

Ex: $\Omega = \underbrace{\{1, 2, 3, 4, 5, 6\}}_{\text{outcomes}}$

$$\mathcal{F} = \underbrace{\{\emptyset, \{1\}, \{2\}, \ldots, \{1, 2\}, \ldots, \Omega\}}_{\text{events}}$$

Review: Probability Distribution

- We have outcomes and events.
- Now assign probabilities For $E \in \mathcal{F},\ P(E) \in [0,1]$

Back to our example:

$$\mathcal{F} = \underbrace{\{\emptyset, \{1,3,5\}, \{2,4,6\}, \Omega\}}_{\text{events}}$$

$$P(\{1,3,5\}) = 0.2, P(\{2,4,6\}) = 0.8$$

Review: Random Variables

- Map outcomes to real values $X : \Omega \to \mathbb{R}$
- Probabilities for a random variable:

$$P(X = 3) := P(\{\omega : X(\omega) = 3\})$$

- Cumulative Distribution Function (CDF)

$$F_X(x) := P(X \leq x)$$

Review: Random Variables

- Back to our example: $\mathcal{F} = \{\emptyset, \{1,3,5\}, \{2,4,6\}, \Omega\}$

  $P(\{1,3,5\}) = 0.2, P(\{2,4,6\}) = 0.8$

- Consider random variable: $X(\omega) = \begin{cases} 1, & \omega = 1,3,5 \\ 0, & \omega = 2,4,6 \end{cases}$

- $P(X = 1) = P(\{\omega: X(\omega) = 1\}) = P(\{1,3,5\}) = 0.2$
- $P(X = 0) = 0.8$
- CDF $F_X(x)$?

Review: Expectation & Variance

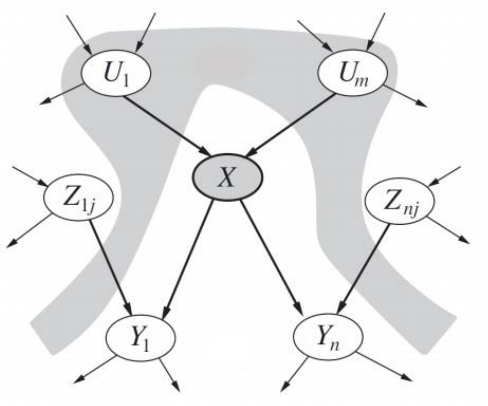
- Expectation: $E[X] = \sum_a a \times P(x = a)$
  - The “average”

- Variance: $Var[X] = E[(X - E[X])^2]$
  - A measure of spread

Review: Conditional Probability

- For when we know something,

$$P(X = a|Y = b) = \frac{P(X = a, Y = b)}{P(Y = b)}$$



Credit: **Devin Soni**

- Conditional independence

$$P(X, Y|Z) = P(X|Z)P(Y|Z)$$

Review: Bayes Rule

- Bayes rule:

Likelihood Prior

Posterior

$$P(H|E_1, E_2, \ldots, E_n) = \frac{P(E_1, \ldots, E_n|H)P(H)}{P(E_1, E_2, \ldots, E_n)}$$

- Assuming **conditional independence:**

$$P(H|E_1, E_2, \ldots, E_n) = \frac{P(E_1|H)P(E_2|H)\cdots, P(E_n|H)P(H)}{P(E_1, E_2, \ldots, E_n)}$$

Review: Classification

$$P(H|E_1, E_2, \ldots, E_n) = \frac{P(E_1|H)P(E_2|H)\cdots, P(E_n|H)P(H)}{P(E_1, E_2, \ldots, E_n)}$$

- Called **Naïve Bayes Classifier**
  - HW2: applied to document classification

- *H*: some class we'd like to infer from evidence $E_1, \ldots, E_n$
  - Estimate prior $P(H)$ from data
  - Estimate likelihood $P(E_i|H)$ from data
  - How?

Samples and Estimation

- Usually, we don't know the distribution P
  - Instead, we see a bunch of samples

- Typical statistics problem: **estimate parameters** from samples
  - Estimate probability $P(H)$
  - Estimate the mean $E[X]$
  - Estimate parameters $P_\theta(X)$

Samples and Estimation

- Typical statistics problem: **estimate parameters** from samples
  - Estimate probability $P(H)$
  - Estimate the mean $E[X]$
  - Estimate parameters $P_\theta(X)$

- Example: Bernoulli with parameter p
  - $p = E[X] = P(X = 1)$

Examples: Sample Mean

- Bernoulli with parameter/mean p
- See samples $x_1, x_2, \ldots, x_n$
  - Estimate mean with **sample mean**

$$\hat{\mathbb{E}}[X] = \frac{1}{n}\sum_{i=1}^{n} x_i$$

  - Counting heads

Break & Quiz

Q 2.1: You see samples of X given by [0,1,1,2,2,0,1,2]. Empirically estimate $E[X^2]$

A. 9/8
B. 15/8
C. 1.5
D. There aren't enough samples to estimate $E[X^2]$

Break & Quiz

Q 2.2: You are empirically estimating *P(X)* for some random variable *X* that takes on 100 values. You see 50 samples. How many of your *P(X=a)* estimates might be 0?

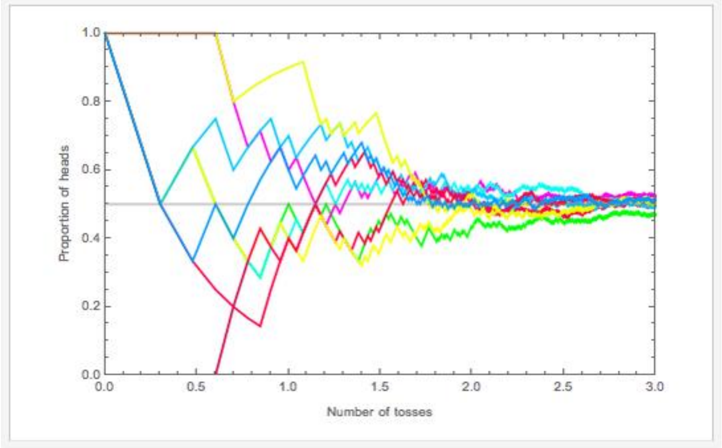
A. None.
B. Between 5 and 50, exclusive.
C. Between 50 and 100, inclusive.
D. Between 50 and 99, inclusive.

Estimation Theory

- Is sample mean is a good estimate of true mean?
  - Law of large numbers: $\widehat{\mathbb{E}}[X] \xrightarrow{n\to\infty} \mathbb{E}[X]$
  - Central limit theorem: limit distribution of $\widehat{\mathbb{E}}[X]$
  - Concentration inequalities

$$P(|\mathbb{E}[X] - \hat{\mathbb{E}}[X]| \geq t) \leq \exp(-2nt^2)$$

- Covered in advanced ML/stat courses



Wolfram Demo