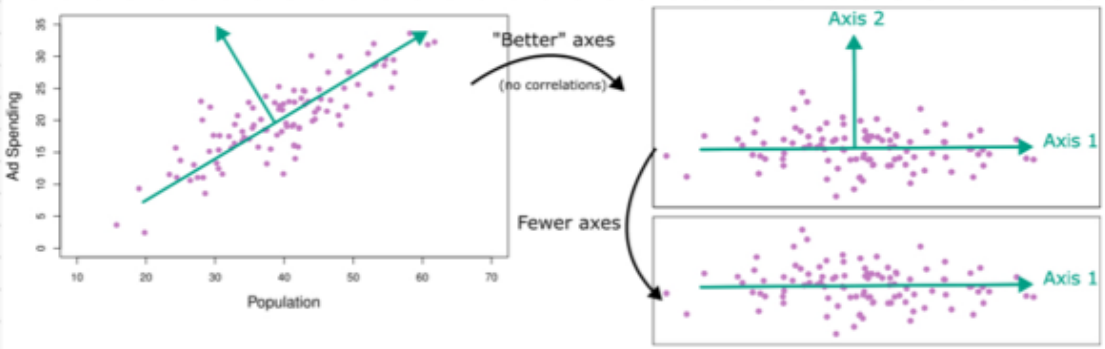
Lecture 10 - 01/10/24

Dimensionality Reduction & Visualization

- Normally our datasets are high dimensional



Principal Component Analysis

- Wish to represent data in a useful way for patterns
- Wanna see if data can be reduced to lower dimensions

PCA is an unsupervised method for such thing

- This means → no information about any class labels is used.
- PCA → finds a pattern
- PCA → looks for direction with greatest variance

* Finding new Basis!

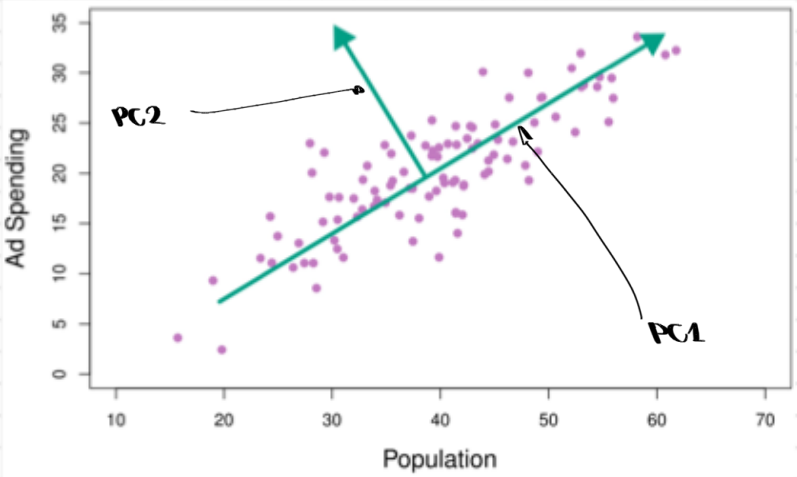
We want to build an orthogonal basis where new basis vectors are chosen to explain direction of the greatest variance.

* Project to lower dimension

First K principal components span a K-dimensional subspace that may be seen as the "best" K-dimensional view of data.

- Consider a set of p features $X_1, \ldots, X_p$ each a real-valued random var.

PCA
↳ gives a new set of p features, principle components, each a linear comb of original p



Here the PC1 score of data point i is:

$$z_{i1} = 0.839(\text{pop}_i - \overline{\text{pop}}) + 0.544(\text{ad}_i - \overline{\text{ad}}).$$

- It points in (0.839, 0.544) in original
- New basis vector is centered at the column means for the data

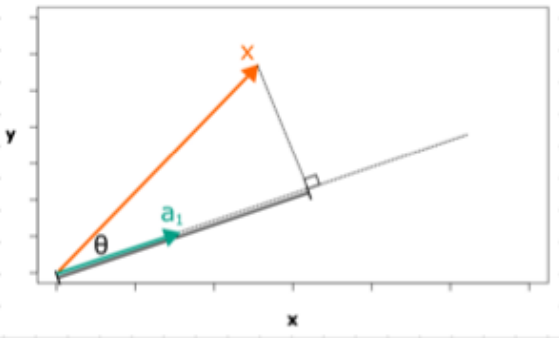
The PC2 is instead orthogonal:

$$z_{i2} = -0.544(\text{pop}_i - \overline{\text{pop}}) + 0.839(\text{ad}_i - \overline{\text{ad}})$$

First Principal Component

- Imagine that we have n data points, each with p features
  - Data point i is: $x_i = \begin{pmatrix} x_{i1} \\ \vdots \\ x_{ip} \end{pmatrix}$ → sample mean $\bar{x} = \sum_{i=1}^{n} x_i / n$
- Wanna find direction of maximal variance
  - We can use Linear Algebra to solve this
  - * Say that x_i represents one data point
  - * Say that a_1 is some other vector
  - * Do the dot product
    - projects one vector onto another

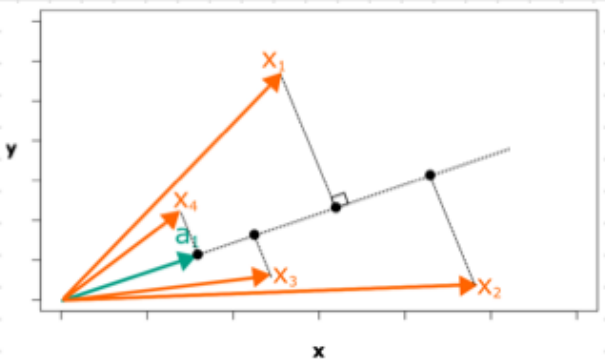
$$x_i \cdot a_1 = \|x_i\| \|a_1\| \cos\theta$$



  - * We can now call our a_1 our first PC

$$a_1 = \begin{pmatrix} a_{11} \\ \vdots \\ a_{1p} \end{pmatrix}$$

  - * If a_1 points to max variance, $a_1^* x$ vary a lot
    - we wanna maximize $\mathrm{Var}(a_1 X)$



Maximization

$\mathrm{Var}(a_1^* X) = a_1^* S a_1$ where $S = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(x_i - \bar{x})^*$ = Covariance matrix

Maximize this to constrain $a_1^* a_1 = 1$

We can use Lagrange multipliers

$$F(a_1, \lambda_1) = a_1^* S a_1 - \lambda_1 (a_1^* a_1 - 1)$$

- function we want to maximize

take partial derivative wrt a_1:

$$\frac{\partial F}{\partial a_1} = 2 S a_1 - 2 \lambda_1 a_1 \xrightarrow{=0} S a_1 - \lambda_1 a_1 = 0$$

$$S a_1 = \lambda_1 a_1$$

Eigenvector corresponding to eigenvalue λ_1 in decomp.

We also need to max F in λ_1:

- Rewrite F using a_1 satisfying $S a_1 = \lambda_1 a_1$

$$\begin{aligned} F(a_1, \lambda_1) &= a_1^T S a_1 - \lambda_1 (a_1^T a_1 - 1) \\ &= a_1^T \lambda_1 a_1 - \lambda_1 a_1^T a_1 + \lambda_1 \\ &= \lambda_1. \end{aligned}$$

We choose it to be the largest eigenvalue

FINDING THE SECOND & SUBSEQUENT COMPONENT!

MAXIMIZE:

$$\text{VAR}(a_2^T X) = a_2^T S a_2 \xrightarrow{\text{CONSTRAINTS}} a_2^T a_2 = 1 \;\&\; a_2^T a_2 = 0$$

AS BEFORE USE LAGRANGE MULTIPLIERS:

$$F(a_2, \lambda_2, M) = a_2^T S a_2 - \lambda_2(a_2^T a_2 - 1) - M(a_1^T a_2)$$

FROM HERE PARTIAL DERIVATIVES, MAX, ETC...

EIGENDECOMPOSITION

- WITH PCA WE CONSIDER EIGENDECOMPOSITION OF S:
$$S = A \Lambda A^T$$
- WE CHOOSE Λ TO BE THE DIAGONAL MATRIX WITH EIGENVALUES $\lambda_1 \geq \dots \geq \lambda_p \geq 0$
- THE ORTHOGONAL $p \times p$ MATRIX $A = [a_1, \dots, a_p]$ HAS EIGENVECTORS AS ITS COLUMNS
- THIS MEANS THAT:
  - PCA CHOOSES EIGENDECOMPOSITION THAT ORDERS EIGENVECTORS ACCORDING TO DEC EIGVAL
  - THE k^{TH} PRINCIPAL COMPONENT IS:
$$a_k^T X = a_{k1} X_1 + a_{k2} X_2 + \dots + a_{kp} X_p$$
- COEFFICIENTS OF PRINCIPLE COMPONENTS = LOADINGS.
  - VECTOR OF LOADINGS = EIGENVECTOR a_k
- PRINCIPAL COMPONENT SCORES = OBS VAL OF $a_k^T X$

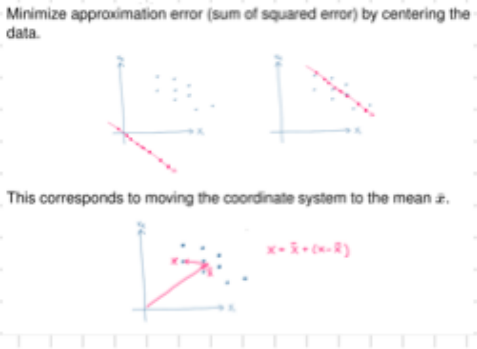
!!!!!

- THE PCA ROTATES YOUR DATA AND PROJECTS IT TO $K \leq p$ DIMENSIONS
- PCA CREATES UNCORRELATED FEATURES THAT ARE LINEAR COMB OF EXISTING p.
- VARIANCE OF K'TH PRINCIPAL COMPONENT IS THE K'TH LARGEST EIGENVALUE → $\text{VAR}(a_k^T X) = \lambda_k$
- EACH PC IS A VECTOR OF LEN p AS ORIGINAL DATA.
- IF USE ALL p PC'S CAN RECONSTRUCT DATA FROM X:

$$X = \sum_{i=1}^{p} (a_i^T x) a_i$$

($a_i^T x$: i'TH PC SCORE FOR X, MEANING COORDINATE IN PC SPACE)

- DO THE SAME AS ABOVE BUT WITH FIRST m PC'S & WILL GET AN APPROXIMATION



Mean and first four eigenvectors visualised:

Reconstructed images for increasing number of principal components

How to choose # of PC's?

- Many heuristic to do a good choice
- In supervised analysis, we can treat it as a tuning parameter

Lets talk about Variance!

- Total Sample Variance:

$$\text{Var}(X_1) + \dots + \text{Var}(X_p) = \text{Var}(z_1^* X) + \dots + \text{Var}(z_p^* X) = \lambda_1 + \dots + \lambda_p$$

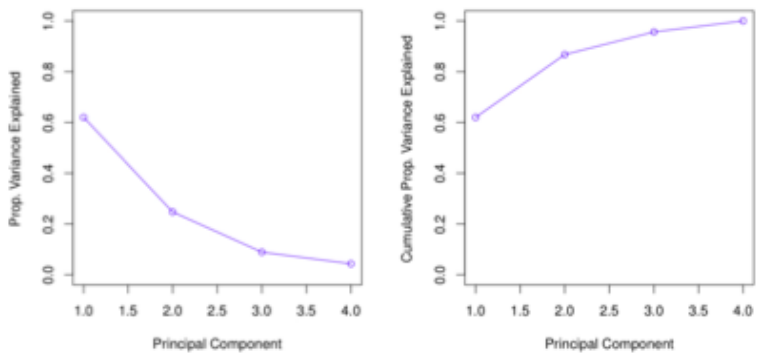
- Proportion of Variance explained by k^{th} PC:

$$\frac{\lambda_k}{\lambda_1 + \dots + \lambda_p}$$

Note

Total Var is the same for the original p's & p^*'s given that trace of $S = P \Lambda P^{-1}$ is the same as trace of Λ

So how do we then decide # of components??



Left: How much Var by each PC

Right: Cumulative of left-hand

Centering & Standardization

- A variable may be centered by:

$$\tilde{X} = X - EX$$

- Can then be scaled to unit variance:

$$\tilde{X} = \frac{X}{\sqrt{\text{Var } X}} = (\text{Var } X)^{-1/2} X$$

- We can have the choice of Standardise features commuwise:
  - All features have unit var & correlation remains
- Another choice might be to Sphere the data:
  - All features have unit Var & no correlation