

Dimensionality Reduction

Machine Learning 10-601B

Seyoung Kim

Text document retrieval/labelling

- Represent each document by a high-dimensional vector in the space of words

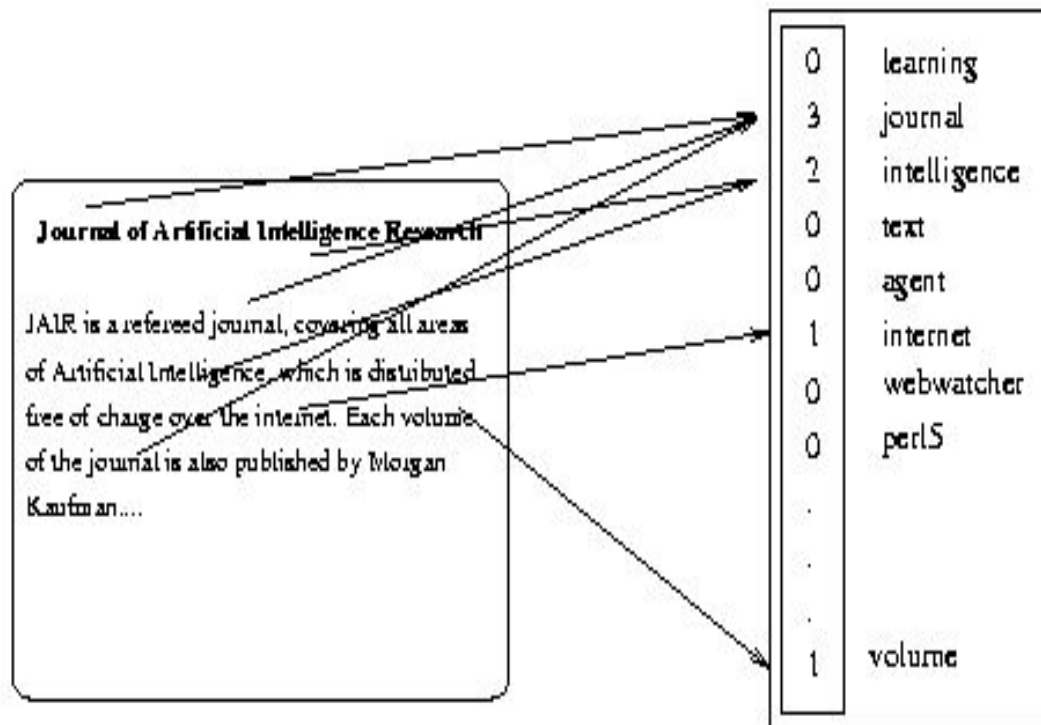


Image retrieval/labelling



$$x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

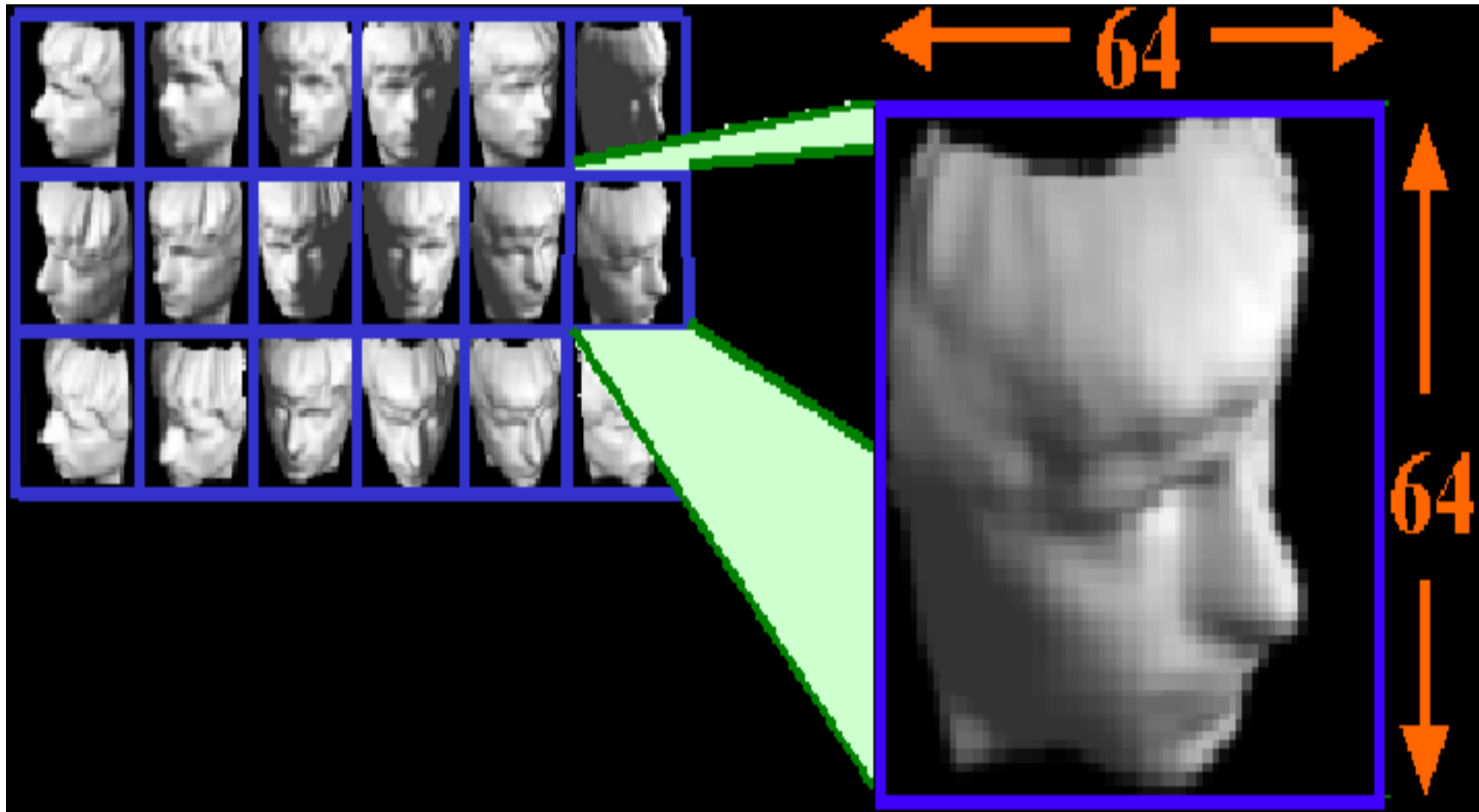
Dimensionality Bottlenecks

- Data dimension
 - Input variables X: **High**
 - 1-5M lexicon token in text documents
 - 1024^2 pixels of a projected image on a IR camera sensor
 - N^2 expansion factor to account for all pairwise correlations
 - 1,000,000 genetic variants in a human's genome
- Information dimension: **Low**
 - Number of free parameters describing probability densities
 - Unsupervised learning $p(X)$
 - Supervised learning $p(Y|X)$: the prediction of Y depends on “information dimension” of X

Intuition: how does your brain store these pictures?

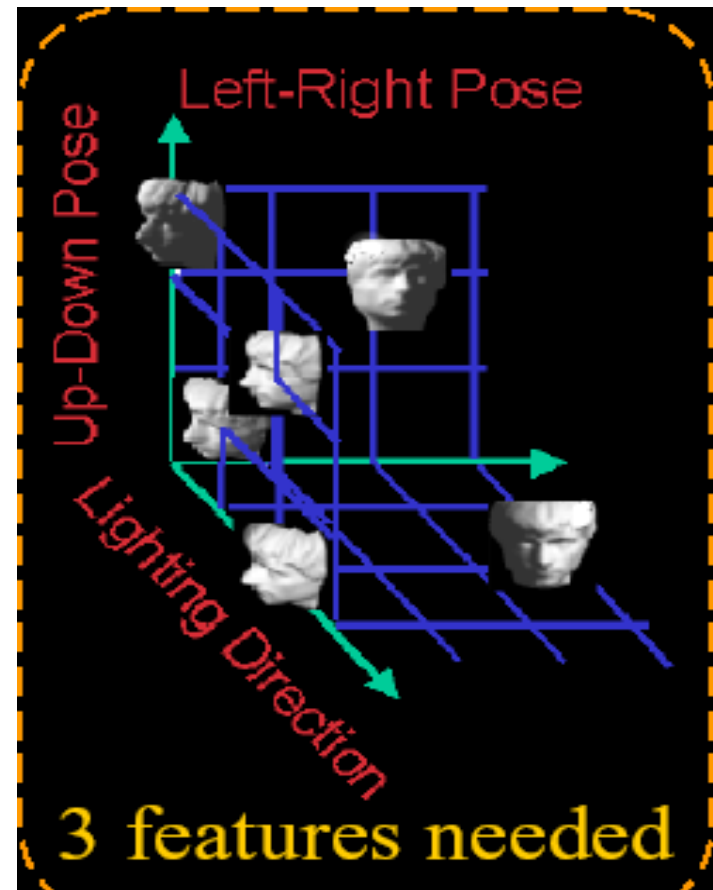


Brain Representation



Brain Representation

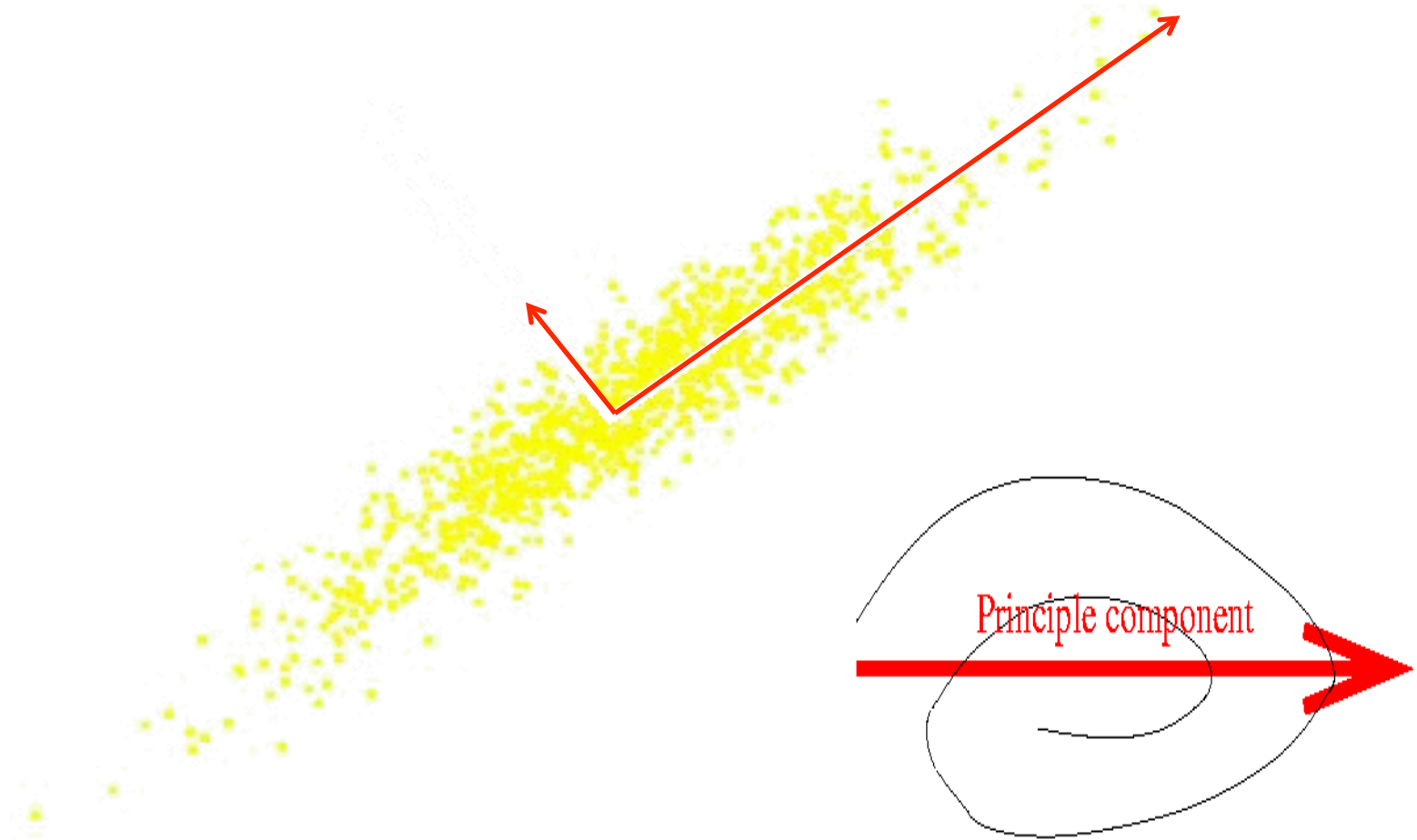
- Every pixel?
 - Or perceptually meaningful structure?
 - Up-down pose
 - Left-right pose
 - Lighting direction
- So, your brain successfully reduced the high-dimensional inputs to an intrinsically 3-dimensional manifold!



Principal Component Analysis

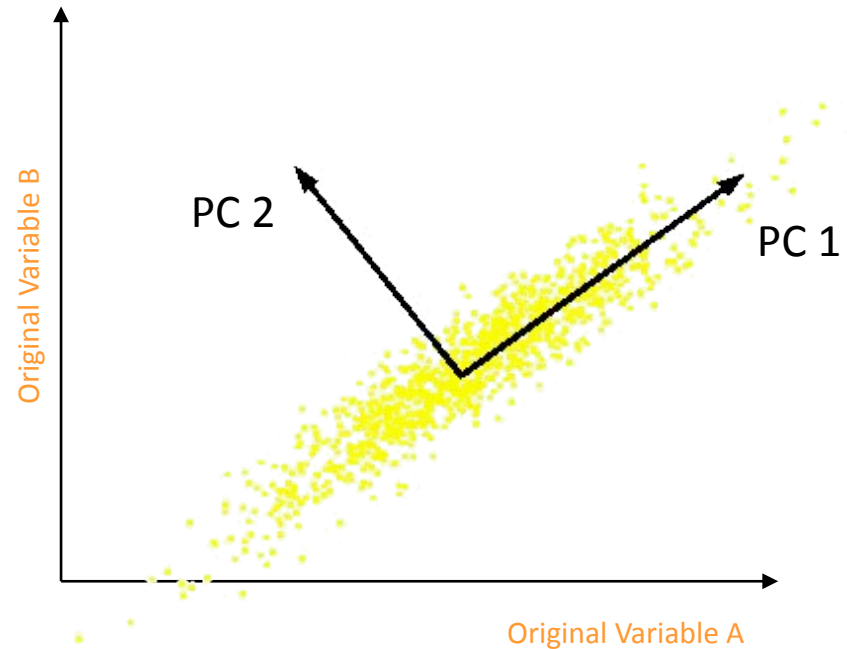
- Areas of variance in data are where items can be best discriminated and key underlying phenomena are observed
- If two items or dimensions are highly correlated or dependent
 - They are likely to represent highly related phenomena
 - We want to combine related variables, and focus on **uncorrelated** or **independent** ones, especially those along which the observations have high variance
- We look for the phenomena underlying the observed covariance/co-dependence in a set of variables
- These phenomena are called “**principal components**”

An example:



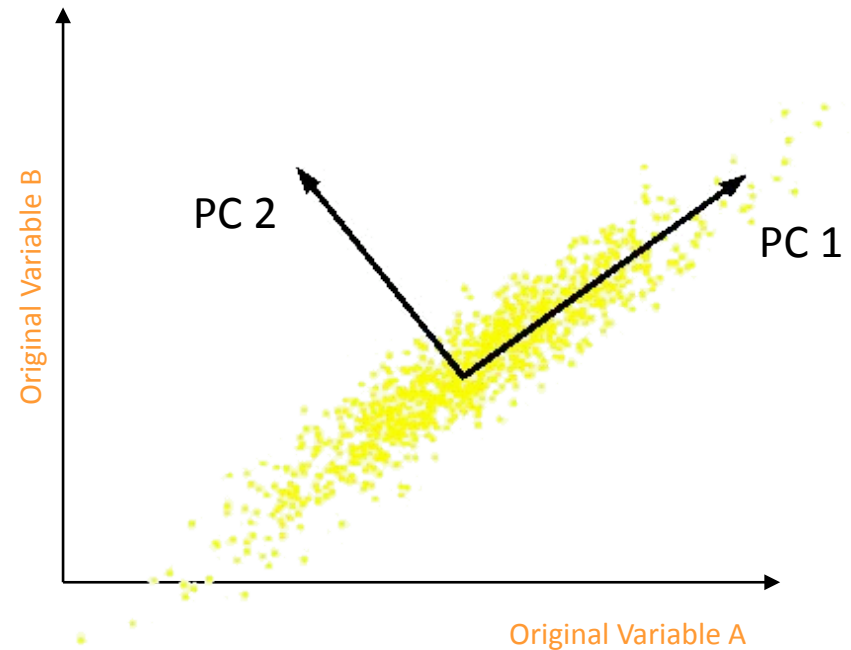
Principal Component Analysis

- The new variables/dimensions
 - Are uncorrelated with one another
 - Orthogonal in original dimension space
 - Capture as much of the original variance in the data as possible
 - Are called Principal Components
 - Are linear combinations of the original ones
- Orthogonal directions of greatest variance in data
- Projections along PC1 discriminate the data most along any one axis



Principal Component Analysis

- First principal component is the direction of greatest variability (covariance) in the data
- Second is the next orthogonal (uncorrelated) direction of greatest variability
 - So first remove all the variability along the first component, and then find the next direction of greatest variability
- And so on ...



Eigen/diagonal Decomposition

- Let $S \in \mathbb{R}^{m \times m}$ be a **square** matrix

- Theorem:** Exists an **eigen decomposition**

$$S = U \Lambda U^{-1} \text{ diagonal}$$

Unique
for
distinct
eigen-
values

(cf. matrix diagonalization theorem)

- Columns of U are **eigenvectors** of S
- Diagonal elements of Λ are **eigenvalues** of S

$$\Lambda = \text{diag}(\lambda_1, \dots, \lambda_m), \quad \lambda_i \geq \lambda_{i+1}$$

Eigenvalues & Eigenvectors

- For symmetric matrices, eigenvectors for distinct eigenvalues are **orthogonal**

$$Sv_1 = \lambda_1 v_1, Sv_2 = \lambda_2 v_2, \text{ and } \lambda_1 \neq \lambda_2 \Rightarrow v_1 \cdot v_2 = 0$$

- All eigenvalues of a real symmetric matrix are **real**.

$$\text{if } |S - \lambda I| = 0 \text{ and } S = S^T \Rightarrow \lambda \in \mathfrak{R}$$

- All eigenvalues of a positive semidefinite matrix are **non-negative**

$$\forall w \in \mathfrak{R}^n, w^T S w \geq 0, \text{ then if } S v = \lambda v \Rightarrow \lambda \geq 0$$

Computing the Components

- Projection of vector $\mathbf{x}$ onto an axis (dimension) $\mathbf{u}$ is $\mathbf{u}^T \mathbf{x}$
- Assume $\mathbf{X}$ is a normalized $n \times p$ data matrix for n samples and p features. Direction of greatest variability is that in which the average square of the projection is greatest:

$$\begin{array}{ll} \text{Maximize} & (\mathbf{1}/n) \mathbf{u}^T \mathbf{X}^T \mathbf{X} \mathbf{u} \\ \text{s.t} & \mathbf{u}^T \mathbf{u} = 1 \end{array}$$

Computing the Components

- Projection of vector $\mathbf{x}$ onto an axis (dimension) $\mathbf{u}$ is $\mathbf{u}^T \mathbf{x}$
- Assume $\mathbf{X}$ is a normalized $n \times p$ data matrix for n samples and p features. Direction of greatest variability is that in which the average square of the projection is greatest:

Maximize $(1/n) \mathbf{u}^T \mathbf{X}^T \mathbf{X} \mathbf{u}$

s.t $\mathbf{u}^T \mathbf{u} = 1$

Construct Lagrangian $(1/n) \mathbf{u}^T \mathbf{X}^T \mathbf{X} \mathbf{u} + \lambda(1 - \mathbf{u}^T \mathbf{u})$

Vector of partial derivatives set to zero

$$1/n \mathbf{X}^T \mathbf{X} \mathbf{u} - \lambda \mathbf{u} = 0$$

Computing the Components

- Projection of vector $\mathbf{x}$ onto an axis (dimension) $\mathbf{u}$ is $\mathbf{u}^T \mathbf{x}$
- Assume $\mathbf{X}$ is a normalized $n \times p$ data matrix for n samples and p features. Direction of greatest variability is that in which the average square of the projection is greatest:

$$\text{Maximize} \quad (1/n) \mathbf{u}^T \mathbf{X}^T \mathbf{X} \mathbf{u}$$

$$\text{s.t} \quad \mathbf{u}^T \mathbf{u} = 1$$

$$\text{Construct Lagrangian} \quad (1/n) \mathbf{u}^T \mathbf{X}^T \mathbf{X} \mathbf{u} + \lambda(1 - \mathbf{u}^T \mathbf{u})$$

Vector of partial derivatives set to zero

$$1/n \mathbf{X}^T \mathbf{X} \mathbf{u} - \lambda \mathbf{u} = 0$$

$$\text{or equivalently } \mathbf{S} \mathbf{u} - \lambda \mathbf{u} = 0 \quad (\mathbf{S} = 1/n \mathbf{X}^T \mathbf{X}: \text{covariance matrix})$$

As $\mathbf{u} \neq \mathbf{0}$ then $\mathbf{u}$ must be an eigenvector of $\mathbf{S}$ with eigenvalue λ

Computing the Components

- Projection of vector $\mathbf{x}$ onto an axis (dimension) $\mathbf{u}$ is $\mathbf{u}^T \mathbf{x}$
- Assume $\mathbf{X}$ is a normalized $n \times p$ data matrix for n samples and p features. Direction of greatest variability is that in which the average square of the projection is greatest:

$$\begin{array}{ll} \text{Maximize} & (1/n) \mathbf{u}^T \mathbf{X}^T \mathbf{X} \mathbf{u} \\ \text{s.t} & \mathbf{u}^T \mathbf{u} = 1 \end{array}$$

Construct Lagrangian $(1/n) \mathbf{u}^T \mathbf{X}^T \mathbf{X} \mathbf{u} - \lambda \mathbf{u}^T \mathbf{u}$

Vector of partial derivatives set to zero

$$1/n \mathbf{X}^T \mathbf{X} \mathbf{u} - \lambda \mathbf{u} = 0$$

or equivalently $\mathbf{S} \mathbf{u} - \lambda \mathbf{u} = 0$ ($\mathbf{S} = 1/n \mathbf{X}^T \mathbf{X}$: covariance matrix)

As $\mathbf{u} \neq \mathbf{0}$ then $\mathbf{u}$ must be an eigenvector of $\mathbf{S}$ with eigenvalue λ

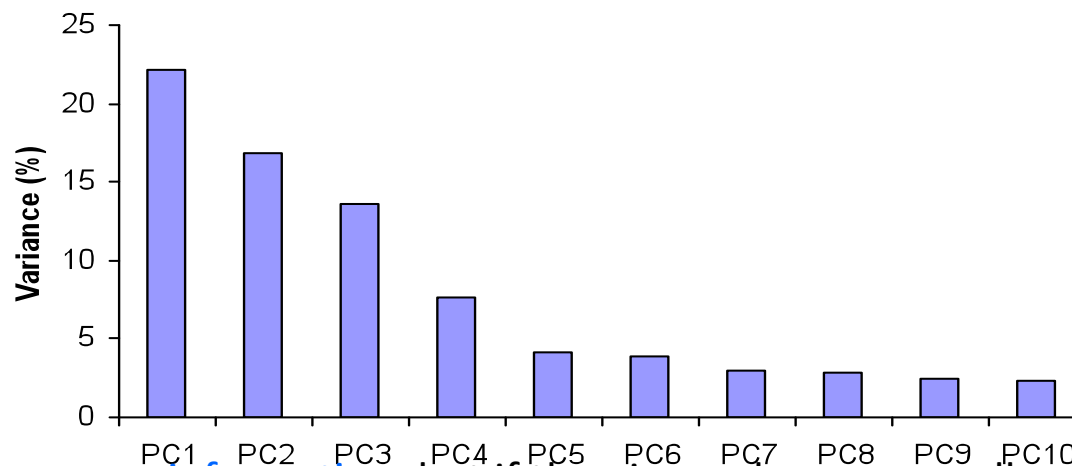
- λ is the **principal eigenvalue** of the **covariance matrix $\mathbf{S}$**
- The eigenvalue denotes the **amount of variability** captured along that dimension

PCs, Variance and Least-Squares

- The first PC retains the greatest amount of variation in the sample
- The k^{th} PC retains the k th greatest fraction of the variation in the sample
- The k^{th} largest eigenvalue of the covariance matrix C is the variance in the sample along the k^{th} PC
- The least-squares view: PCs are a series of linear least squares fits to a sample, each orthogonal to all previous ones (Bishop 12.1.2)

How Many PCs?

- For p original dimensions, sample covariance matrix is $p \times p$, and has up to p eigenvectors. So p PCs.
- Where does dimensionality reduction come from?
Can *ignore the components of lesser significance*.



You do *lose some information*, but if the eigenvalues are small, you don't lose much

- p dimensions in original data
- Calculate p eigenvectors and eigenvalues
- choose only the first q eigenvectors, based on their eigenvalues
- final data set has only q dimensions

Applying PCA to Images

- 361 x 261 pixels, 83781 dimensional data



Reconstructing the Images from 4 PCs

- The principal components are also images



Reconstructing the Images from 4 PCs



Summary:

- Principle
 - Linear projection method to reduce the number of parameters
 - Transfer a set of correlated variables into a new set of uncorrelated variables
 - Map the data into a space of lower dimensionality
 - Form of unsupervised learning
- Properties
 - It can be viewed as a rotation of the existing axes to new positions in the space defined by original variables
 - New axes are orthogonal and represent the directions with maximum variability