Dimensionality Reduction

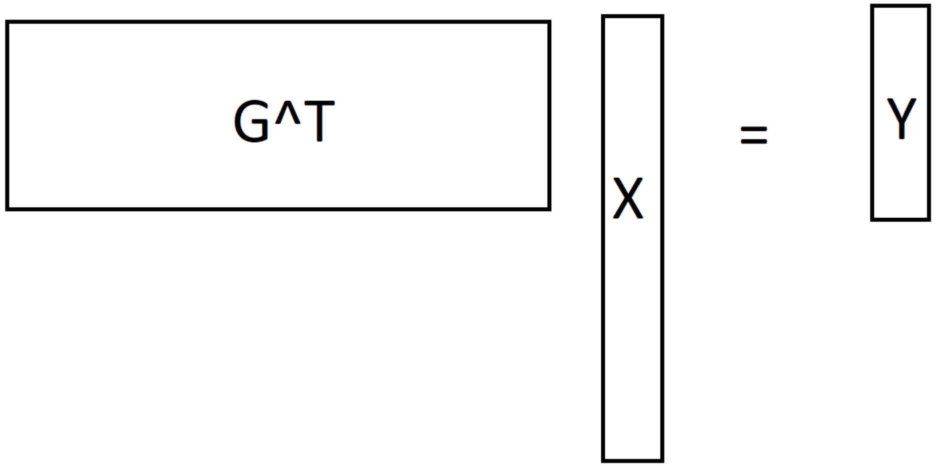
आई आई टी हैदराबाद
IIT Hyderabad

ML Problems

	Supervised Learning	Unsupervised Learning
Discrete	classification or categorization	clustering
Continuous	regression	dimensionality reduction

What is Dimensionality Reduction

- Refers to the mapping of the original high-dimensional data onto a lower-dimensional space.
- Criterion for feature reduction can be different for different problems.
  - Unsupervised setting: minimize the information loss
  - Supervised setting: maximize the class discrimination
- Given a set of data points of p variables the linear transformation (projection)



Dimensionality Reduction vs Feature Selection

- **Feature reduction**
  - All original features are used
  - The transformed features are linear combinations of the original features (in case of linear DR).

- **Feature selection**
  - Only a subset of the original features are used.

Why DR?

- Most machine learning techniques may not be effective for high-dimensional data
  - Curse of Dimensionality
  - Query accuracy and efficiency degrade rapidly as the dimension increases
  - Lower space and time complexity
  - Visualization, Data compression, Noise/irrelevant feature removal
- The intrinsic dimension may be small
  - For example, the number of genes responsible for a certain type of disease may be small

High-dimensional data are strange

- Consider the hypersphere of radius r on a space of dimension d

$$\mathcal{S} = \left\{ \mathbf{x} \middle| \sum_{i=1}^{d} x_i^2 \leq r^2 \right\}$$

- Its volume is

$$V_d(r) = \frac{r^d \pi^{\frac{d}{2}}}{\Gamma\left(\frac{d}{2} + 1\right)}$$

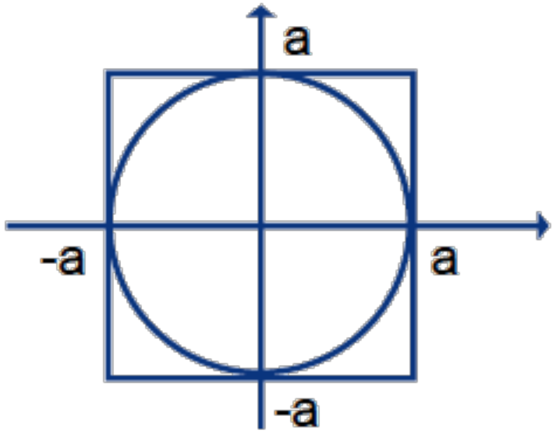
- Where Γ(n) is the Gamma function

$$\Gamma(n) = \int_0^{\infty} e^{-x} x^{n-1} dx$$

Source: N Vasconcelos

Hypercube vs Hypersphere

- Consider the hyper-cube $[-a,a]^d$ and the inscribed hyper-sphere, i.e.

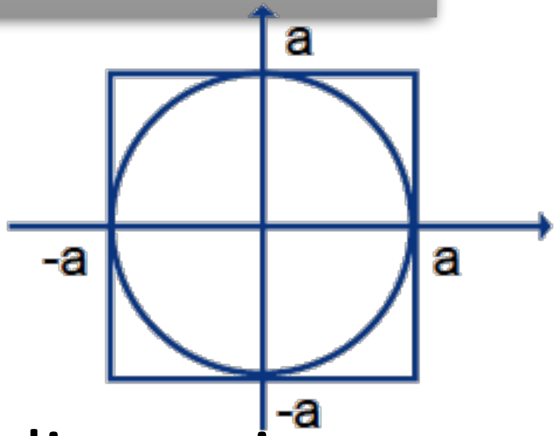


- What does your intuition tell you about the relative sizes of these two objects?
  - Volume of sphere ~= volume of cube
  - Volume of sphere >> volume of cube
  - Volume of sphere << volume of cube

Source: N Vasconcelos

Hypercube vs Hypersphere

- Let's compute the answer



$$f_d = \frac{Vol(sphere)}{Vol(cube)} = \frac{\frac{a^d \pi^{\frac{d}{2}}}{\Gamma\left(\frac{d}{2}+1\right)}}{(2a)^d} = \frac{\pi^{\frac{d}{2}}}{2^d \Gamma\left(\frac{d}{2}+1\right)}$$

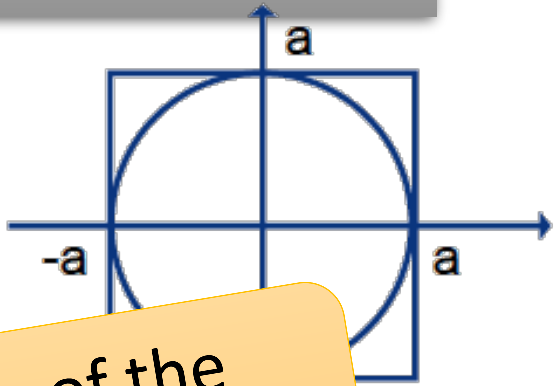
- Sequence that does not depend on a, just on the dimension d!

d	1	2	3	4	5	6	7
f_d	1	.785	.524	.308	.164	.08	.037

- It goes to zero, and goes to zero fast!

Source: N Vasconcelos

Hypercube vs Hypersphere



- Let's compute the answer

$$f_d = \frac{Vol(sphere)}{Vol(cube)} = \frac{\frac{a^d \pi^{\frac{d}{2}}}{\Gamma\left(\frac{d}{2}+1\right)}}{(2a)^d} = \frac{\pi^{\frac{d}{2}}}{2^d \Gamma\left(\frac{d}{2}+1\right)}$$

- Sequence that does not depe[illegible] d!

As the dimension of the space increases, the volume of the sphere is much smaller (infinitesimal) than that of the cube!

[illegible]	[illegible]	[illegible]	[illegible]	5	6	7
[illegible]	.785	.524	.308	.164	.08	.037

- It goes to zero, and goes to zero fast!

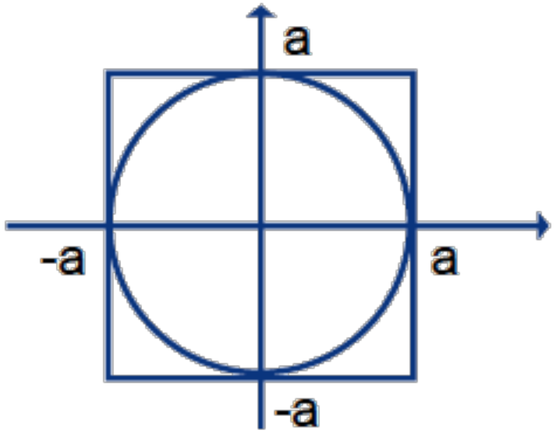
Hypercube vs Hypersphere

Actually not very surprising

1. d = 1

Volume is the same

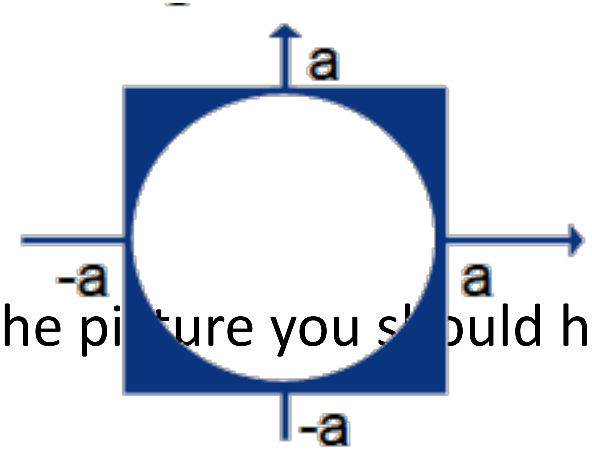
2. d = 2



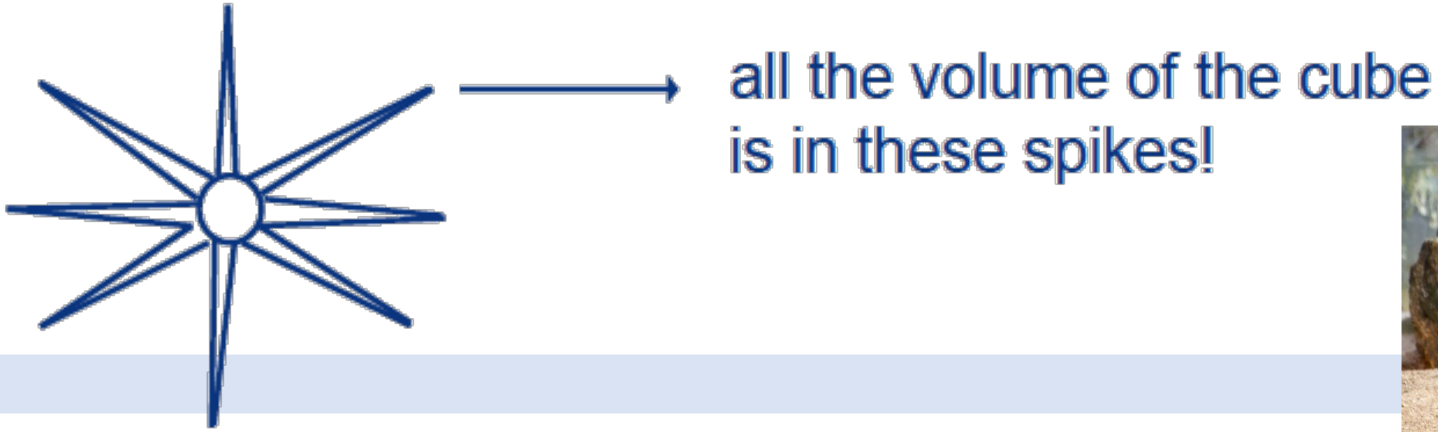
Volume of sphere is already smaller

Hypercube vs Hypersphere

- As the dimension increases the volume of the shaded corners becomes larger



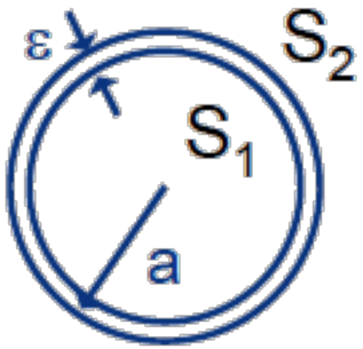
- In high dimensions the picture you should have in mind is



Source: N Vasconcelos

The Curse of Dimensionality

- Consider the crust of unit hypersphere of thickness ε
- Let's compute the ratio of volumes

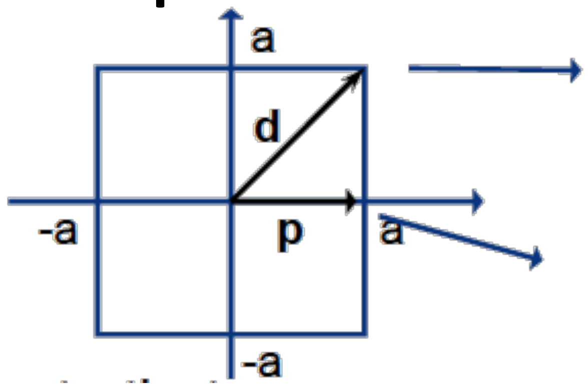


$$\frac{Vol(S_1)}{Vol(S_2)} = \frac{\frac{(a-\epsilon)^d \pi^{\frac{d}{2}}}{\Gamma\left(\frac{d}{2}+1\right)}}{\frac{a^d \pi^{\frac{d}{2}}}{\Gamma\left(\frac{d}{2}+1\right)}} = \frac{a^d \left(1-\frac{\epsilon}{a}\right)^d}{a^d} = \left(1-\frac{\epsilon}{a}\right)^d$$

- No matter how small ε is, ratio goes to zero as d increases i. e. **"all the volume is in the crust!"**

We can check mathematically

- Consider **d** and **p**



$$\mathbf{d} = (a, a, \ldots, a) \in \mathcal{R}^d$$

$$\mathbf{p} = (a, 0, \ldots, 0) \in \mathcal{R}^d$$

- Note that

$$\frac{||d||^2}{||p||^2} = \frac{da^2}{a^2} = d \rightarrow \infty \qquad cos\theta = \frac{\mathbf{d}^T\mathbf{p}}{\sqrt{||d||^2||p||^2}} \qquad \lim_{d\rightarrow\infty} \frac{\text{dist}_{\max} - \text{dist}_{\min}}{\text{dist}_{\min}} \rightarrow 0$$

$$= \frac{a^2}{\sqrt{da^2a^2}} = \frac{1}{\sqrt{d}} \rightarrow 0$$

- **d** orthogonal to **p** as **d**
- increases and infinitely larger!!!

Source: N Vasconcelos

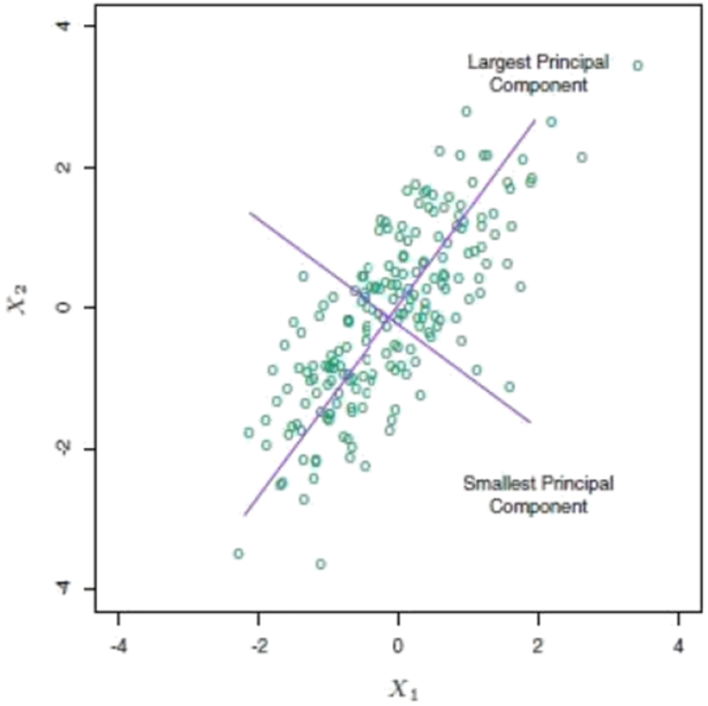
DR Methods

- **Linear**
  - Principal Component Analysis (PCA)
  - Linear Discriminant Analysis (LDA)
  - Canonical Correlation Analysis (CCA)

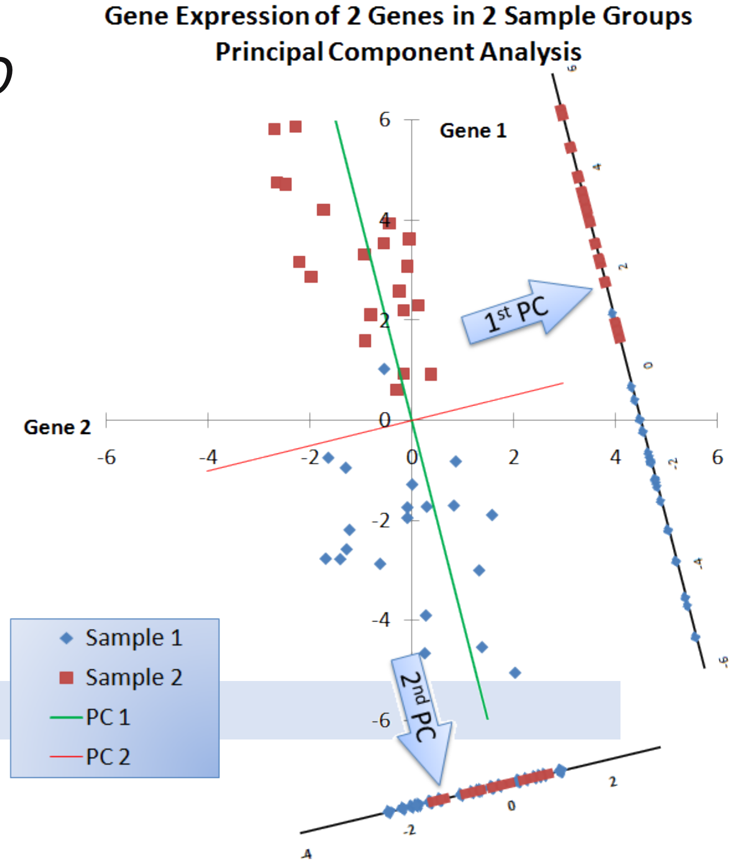
- **Non-linear**
  - Nonlinear feature reduction using kernels
  - Manifold learning

Principal Component Analysis (PCA)

- Find a low-dimensional space such that when $\boldsymbol{x}$ is projected there, information loss is minimized.
- The projection of $\boldsymbol{x}$ on the direction of $\boldsymbol{w}$ is: $z = \boldsymbol{w}^T\boldsymbol{x}$
- Find $\boldsymbol{w}$ such that Var(z) is maximized,
- Our goal is to project the data onto a space having dimensionality $M < D$



riance



Principal Component Analysis (PCA)

- Find a low-dimensional space such that when $\boldsymbol{x}$ is projected there, information loss is minimized.
- The projection of $\boldsymbol{x}$ on the direction of $\boldsymbol{w}$ is: $z = \boldsymbol{w}^T\boldsymbol{x}$
- Find $\boldsymbol{w}$ such that Var(z) is maximized

$$\begin{aligned}
\mathrm{Var}(z) &= \mathrm{Var}(\boldsymbol{w}^T\boldsymbol{x}) = \mathrm{E}[(\boldsymbol{w}^T\boldsymbol{x} - \boldsymbol{w}^T\boldsymbol{\mu})^2] \\
&= \mathrm{E}[(\boldsymbol{w}^T\boldsymbol{x} - \boldsymbol{w}^T\boldsymbol{\mu})(\boldsymbol{w}^T\boldsymbol{x} - \boldsymbol{w}^T\boldsymbol{\mu})] \\
&= \mathrm{E}[\boldsymbol{w}^T(\boldsymbol{x} - \boldsymbol{\mu})(\boldsymbol{x} - \boldsymbol{\mu})^T\boldsymbol{w}] \\
&= \boldsymbol{w}^T\,\mathrm{E}[(\boldsymbol{x} - \boldsymbol{\mu})(\boldsymbol{x} - \boldsymbol{\mu})^T]\boldsymbol{w} = \boldsymbol{w}^T\boldsymbol{\Sigma}\,\boldsymbol{w}
\end{aligned}$$

where $\mathrm{Var}(\boldsymbol{x}) = \mathrm{E}[(\boldsymbol{x} - \boldsymbol{\mu})(\boldsymbol{x} - \boldsymbol{\mu})^T] = \boldsymbol{\Sigma}$

PCA

- Maximize Var(z) subject to $||\boldsymbol{w}||=1$

Using Lagrange multipliers

$\boldsymbol{\Sigma} \boldsymbol{w}_1 = \alpha \boldsymbol{w}_1$ that is, $\boldsymbol{w}_1$ is an eigenvector of $\boldsymbol{\Sigma}$

Choose the one with the largest eigenvalue for Var(z) to be max

- Second principal component: Max Var(z_2), s.t., $||\boldsymbol{w}_2||=1$ and orthogonal to $\boldsymbol{w}_1$

How?

$\boldsymbol{\Sigma}\ \boldsymbol{w}_2 = \alpha\ \boldsymbol{w}_2$ that is, $\boldsymbol{w}_2$ is another eigenvector of $\boldsymbol{\Sigma}$

and so on.

Computational cost : O(MD^2) for M principal components

आई आई टी हैदराबाद
IIT Hyderabad

Minimum-error formulation

- introduce a complete orthonormal set of *D*-dimensional basis vectors {**u***i*}

$$\mathbf{u}_i^{\mathrm{T}}\mathbf{u}_j = \delta_{ij}.$$

$$\mathbf{x}_n = \sum_{i=1}^{D} \alpha_{ni}\mathbf{u}_i$$

$$\mathbf{x}_n = \sum_{i=1}^{D} \left(\mathbf{x}_n^{\mathrm{T}}\mathbf{u}_i\right)\mathbf{u}_i.$$

Minimum-error formulation

- introduce a complete orthonormal set of *D*-dimensional basis vectors $\{\mathbf{u}_i\}$

$$\mathbf{x}_n = \sum_{i=1}^{D} \left(\mathbf{x}_n^{\mathrm{T}} \mathbf{u}_i\right) \mathbf{u}_i .$$

- Approximate this data point using a representation involving a restricted number $M < D$ of variables corresponding to a projection onto a lower-dimensional subspace.

$$\widetilde{\mathbf{x}}_n = \sum_{i=1}^{M} z_{ni} \mathbf{u}_i + \sum_{i=M+1}^{D} b_i \mathbf{u}_i$$

- Minimize the distortion introduced by the reduction in dimensionality.

$$J = \frac{1}{N} \sum_{n=1}^{N} \|\mathbf{x}_n - \widetilde{\mathbf{x}}_n\|^2 .$$

Minimum-error formulation

- Minimize the distortion introduced by the reduction in dimensionality.

$$J = \frac{1}{N} \sum_{n=1}^{N} \|\mathbf{x}_n - \widetilde{\mathbf{x}}_n\|^2.$$

$$\mathbf{x}_n = \sum_{i=1}^{D} \left(\mathbf{x}_n^{\mathrm{T}} \mathbf{u}_i\right) \mathbf{u}_i.$$

- derivative with respect to *znj*

$$z_{nj} = \mathbf{x}_n^{\mathrm{T}} \mathbf{u}_j$$

- Derivative wrt *bi*

$$b_j = \overline{\mathbf{x}}^{\mathrm{T}} \mathbf{u}_j$$

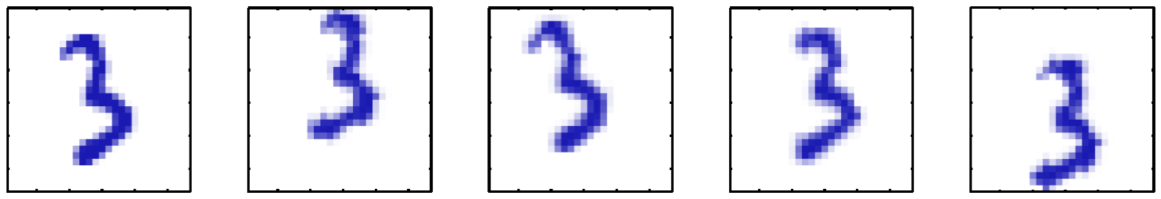
- Displacement vector lies in the space orthogonal to the principal subspace

$$\mathbf{x}_n - \widetilde{\mathbf{x}}_n = \sum_{i=M+1}^{D} \left\{(\mathbf{x}_n - \overline{\mathbf{x}})^{\mathrm{T}} \mathbf{u}_i\right\} \mathbf{u}_i$$

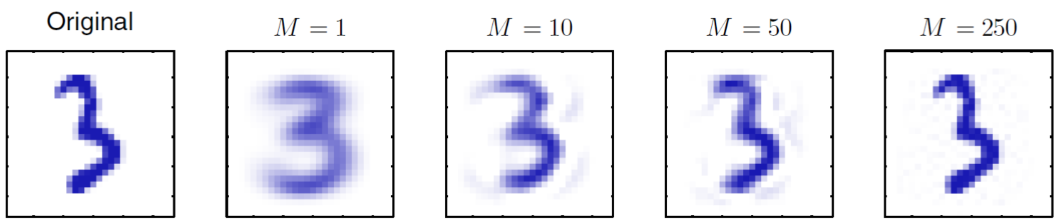
- distortion measure *J*

$$J = \frac{1}{N} \sum_{n=1}^{N} \sum_{i=M+1}^{D} \left(\mathbf{x}_n^{\mathrm{T}} \mathbf{u}_i - \overline{\mathbf{x}}^{\mathrm{T}} \mathbf{u}_i\right)^2 = \sum_{i=M+1}^{D} \mathbf{u}_i^{\mathrm{T}} \mathbf{S} \mathbf{u}_i.$$

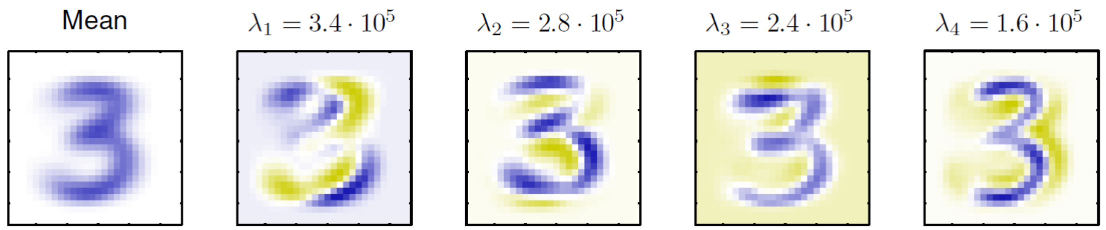
PCA



A synthetic data set obtained by taking one of the off-line digit images and creating multiple copies in each of which the digit has undergone a random displacement and rotation within some larger image field. The resulting images each have $100 \times 100 = 10,000$



An original example from the off-line digits data set together with its PCA reconstructions obtained by retaining M principal components for various values of M. As M increases the reconstruction becomes more accurate and would become perfect when $M = D = 28 \times 28 = 784$.



The mean vector $\overline{\mathbf{x}}$ along with the first four PCA eigenvectors $\mathbf{u}_1, \ldots, \mathbf{u}_4$ for the off-line digits data set, together with the corresponding eigenvalues.

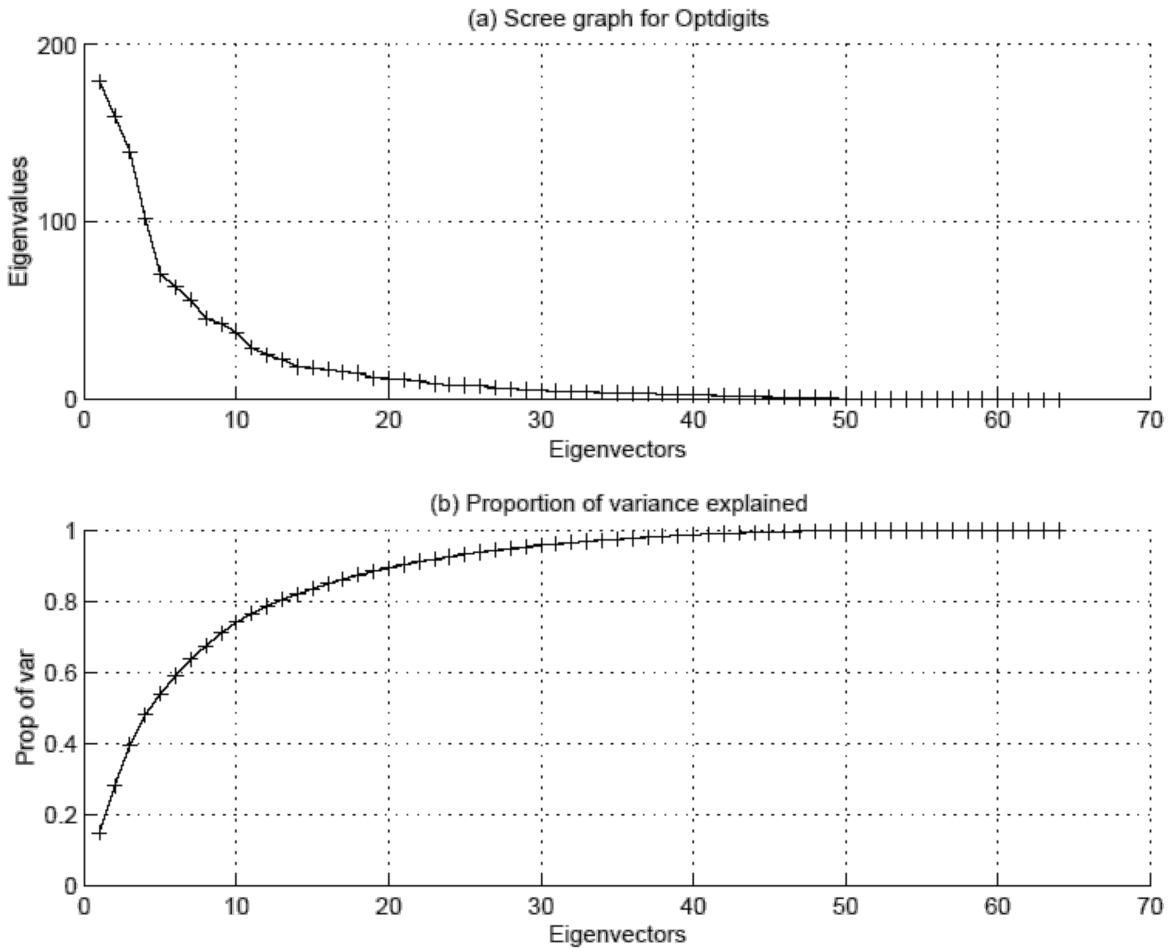
How to choose k

- Proportion of Variance (PoV) explained

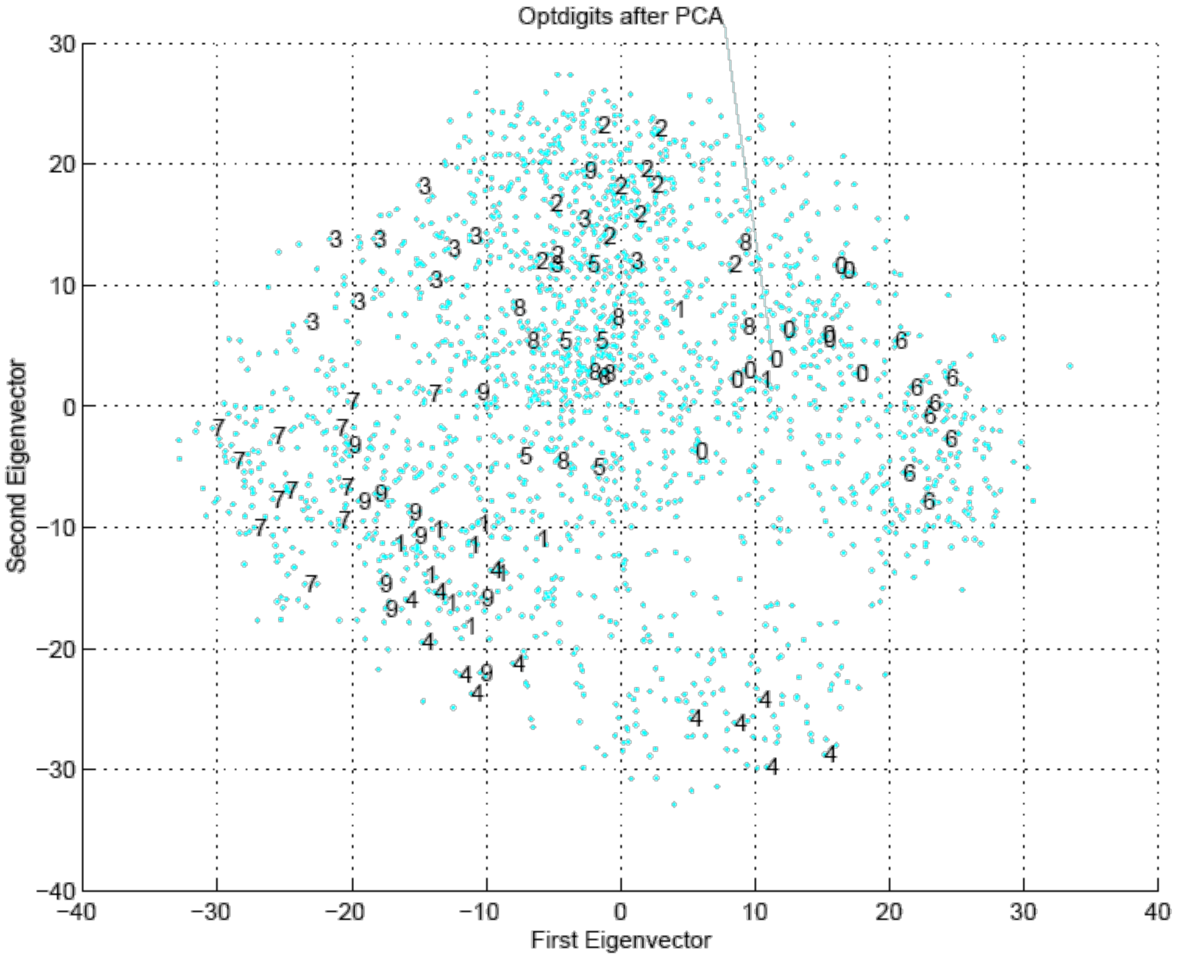
  when λ_i are sorted in descending order

- Typically, stop at PoV>0.9
- See graph plots of PoV vs *k*, stop at "elbow"

Illustration



Example



Optdigits after PCA
Second Eigenvector
First Eigenvector
30
20
10
0
−10
−20
−30
−40
−40 −30 −20 −10 0 10 20 30 40