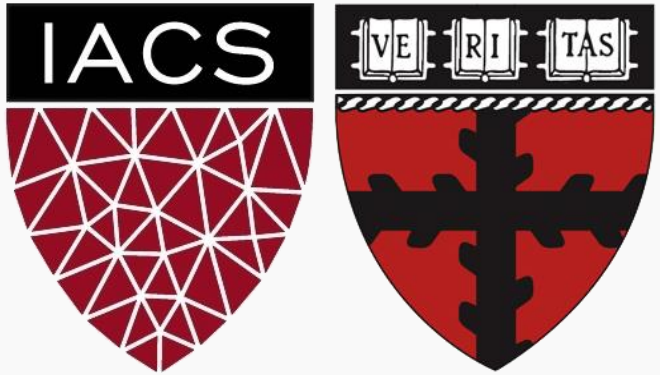
# Advanced Section #4: Methods of Dimensionality Reduction: Principal Component Analysis (PCA)

Cedric Flamant

CS109A Introduction to Data Science

Pavlos Protopapas, Kevin Rader, and Chris Tanner



# Outline

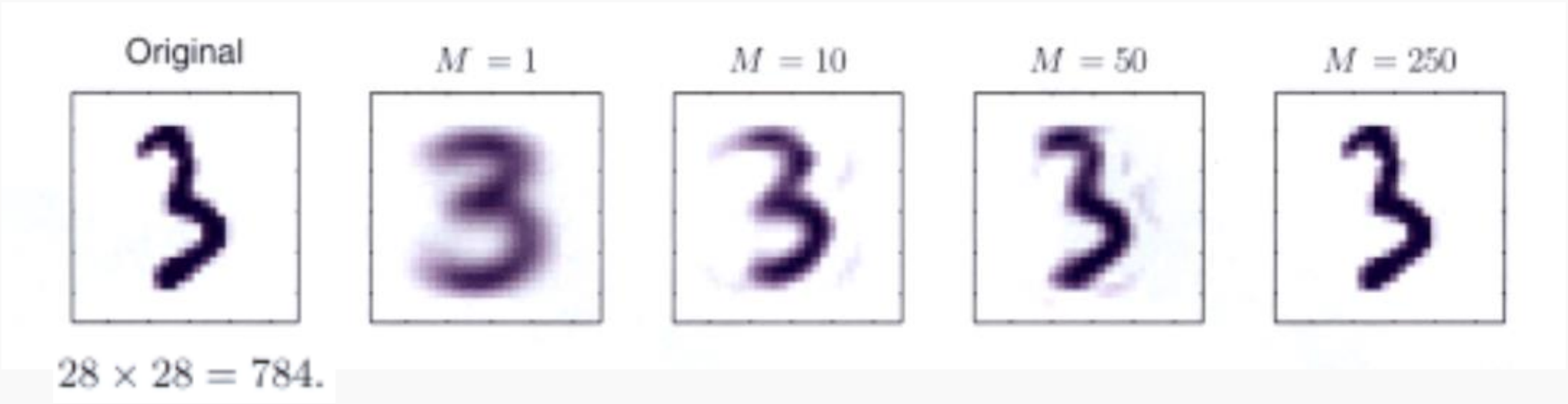
1. Introduction:
   a. Why Dimensionality Reduction?
   b. Linear Algebra (Recap).
   c. Statistics (Recap).

2. Principal Component Analysis:
   a. Foundation.
   b. Assumptions & Limitations.
   c. Kernel PCA for nonlinear dimensionality reduction.

# Dimensionality Reduction, why?

A process of reducing the number of predictor variables under consideration.

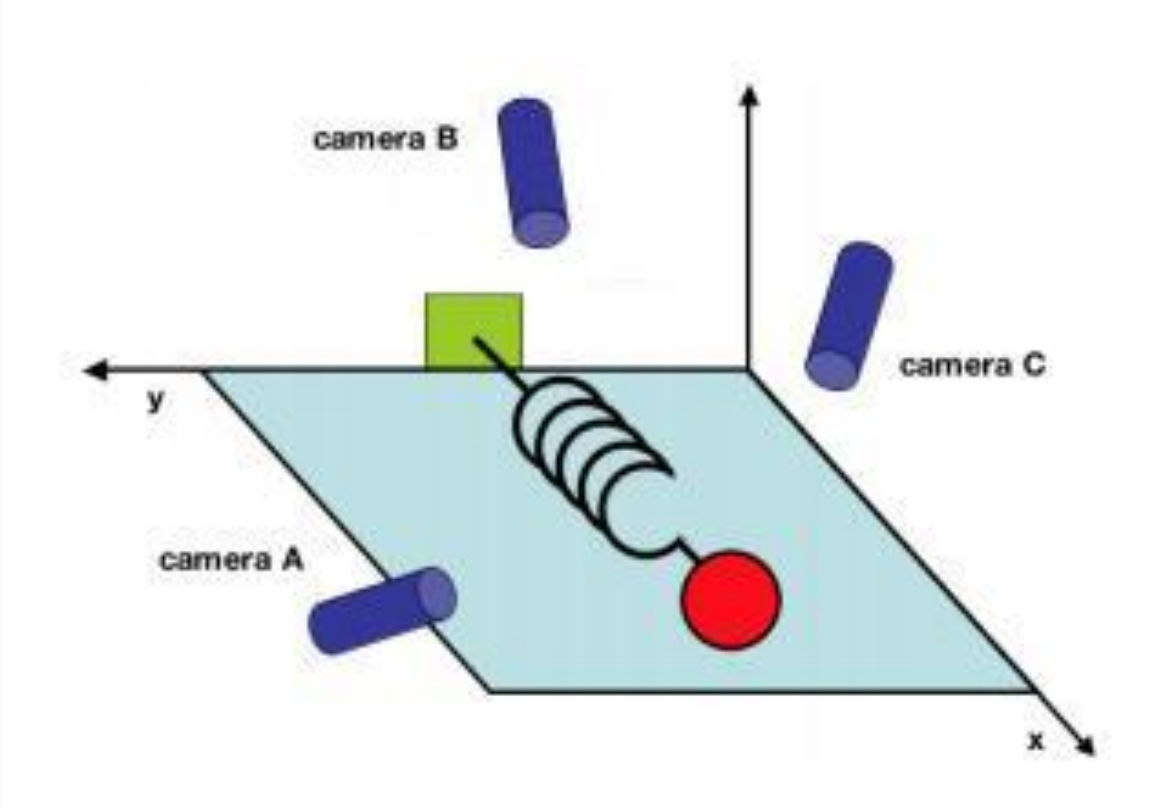
To find a more meaningful basis to express our data filtering the noise and revealing the hidden structure.




C. Bishop, *Pattern Recognition and Machine Learning*, Springer (2008).

# A simple example taken from Physics

Consider an ideal spring-mass system oscillating along x.
Seeking the pressure *Y* that spring exerts on the wall.




LASSO regression model:

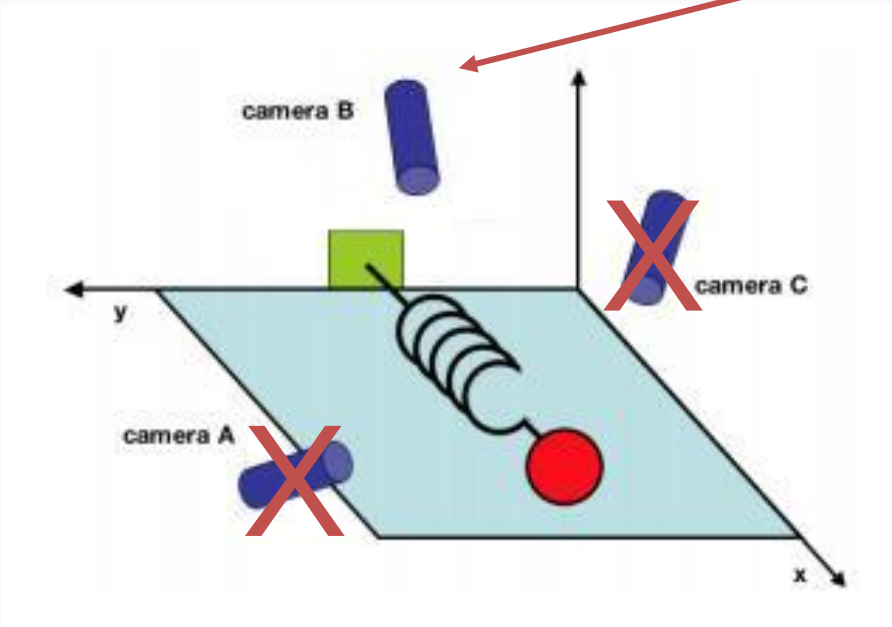
$$Y = \beta_A x_A + \beta_B x_B + \beta_C x_C$$

LASSO variable selection:

$$\hat{\beta}_A = \hat{\beta}_C = 0$$

J. Shlens, *A Tutorial on Principal Component Analysis*, (2003).

# Principal Component Analysis versus LASSO




LASSO simply selects one of the arbitrary directions, *scientifically unsatisfactory.*

We want to use all the measurements to situate the position of mass.

We want to find a lower-dimensional manifold of predictors on which data lie.

✓ **Principal Component Analysis (PCA):**
A powerful *Statistical* tool for analyzing data sets and is formulated in the context of *Linear Algebra*.

# Linear Algebra (Recap)

# Symmetric matrices

Consider a design (or data) matrix consists of n observations and p predictors:

$$X \in \mathbb{R}^{n \times p}$$

Then $X^T X$ is a symmetric matrix.

Symmetric: $A^T = A$

Using that : $(BC)^T = C^T B^T$

$$(X^T X)^T = X^T (X^T)^T = X^T X$$

Similar for $XX^T$

# Eigenvalues and Eigenvectors

For a real and symmetric matrix: e.g. $X^T X = A \in \mathbb{R}^{p \times p}$

There exists a unique set of real eigenvalues: $\{\lambda_1, \ldots, \lambda_p\}$

and the associated eigenvectors: $\{u_1, \ldots, u_p\}$

$$Au_i = \lambda_i u_i \qquad (\lambda_i \in \mathbb{R})$$

such that:

$$u_i^T u_j = \delta_{ij} \qquad \text{(orthogonal)}$$

$$||u_i||^2 = 1 \qquad \text{(normalized)}$$

➢ Hence, they form an *orthonormal basis.*

# Spectrum and Eigen-decomposition

Spectrum:

$$\Lambda = \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_p \end{pmatrix}$$

Orthogonal Matrix:

$$Q = \begin{pmatrix} u_{11} & u_{21} & \cdots & u_{p1} \\ u_{12} & u_{22} & \cdots & u_{p2} \\ \vdots & \vdots & \ddots & \vdots \\ u_{1p} & u_{2p} & \cdots & u_{pp} \end{pmatrix} \quad (Q^{-1} = Q^T)$$

$$(Q^T Q = Q Q^T = I)$$

Eigen-decomposition:

$$A = Q \Lambda Q^T$$

# Real & Positive Eigenvalues: Gram Matrix

- The eigenvalues of $X^T X$ are non-negative real numbers:

$$
\begin{aligned}
X^T X u &= \lambda u \\
u^T X^T X u &= u^T \lambda u \\
(Xu)^T (Xu) &= \lambda u^T u \\
||Xu||^2 &= \lambda ||u||^2
\end{aligned}
$$

$$\Rightarrow \lambda \geq 0$$

Similar for $XX^T$

- Hence, $X^T X$ and $XX^T$ are **positive-semidefinite** .

# Same eigenvalues

- $X^T X$ and $XX^T$ share the same eigenvalues:

$$X^T X u = \lambda u$$
$$XX^T X u = X\lambda u$$
$$XX^T (Xu) = \lambda (Xu)$$
$$XX^T \tilde{u} = \lambda \tilde{u}$$

Same eigenvalues.

Transformed eigenvectors:

$$\tilde{u} = Xu$$

# The sum of eigenvalues of $X^T X$ is equal to its trace

- Cyclic Property of Trace: $\text{Tr}(BC) = \text{Tr}(CB)$

  Suppose the matrices: $B_{m\times n}$ & $C_{n\times m}$

$$\text{Tr}(BC) = \sum_i^m (BC)_{ii} = \sum_i^m \sum_j^n B_{ij} C_{ji}$$

$$\sum_i^m \sum_j^n C_{ji} B_{ij} = \sum_j^n (CB)_{jj} = \text{Tr}(CB)$$

- The trace of a Gram matrix is the sum of its eigenvalues.

$$\text{Tr}(\underbrace{X^T X}_{p\times p}) = \text{Tr}(U\Lambda U^T) = \text{Tr}(U^T U\Lambda) = \text{Tr}(\Lambda)$$

$$\Rightarrow \text{Tr}(X^T X) = \sum_{i=1}^p \lambda_i$$

# Statistics (Recap)

# Centered Model Matrix

Consider the model (data) matrix $X \in \mathbb{R}^{n \times p}$

We make the predictors *centered* (each column has zero expectation) by subtracting the sample mean:

$$\hat{\mu}_j = \frac{1}{n} \sum_{i=1}^{n} x_{ij}$$

Centered Model Matrix:

$$\tilde{X} = (\vec{x}_1 - \hat{\mu}_1, \ldots, \vec{x}_p - \hat{\mu}_p)$$

# Sample Covariance Matrix

Consider the Covariance matrix:

$$S = \frac{1}{n-1}\tilde{X}^T\tilde{X}$$

Inspecting the terms:

➢ The diagonal terms are the sample variances:

$$S_{jj} = \frac{1}{n-1}\sum_{i=1}^{n}(x_{ij} - \hat{\mu}_j)^2$$

➢ The non-diagonal terms are the sample covariances:

$$S_{jk} = \frac{1}{n-1}\sum_{i=1}^{n}(x_{ij} - \hat{\mu}_j)(x_{ik} - \hat{\mu}_k) \qquad (j \neq k)$$

# Principal Components Analysis (PCA)

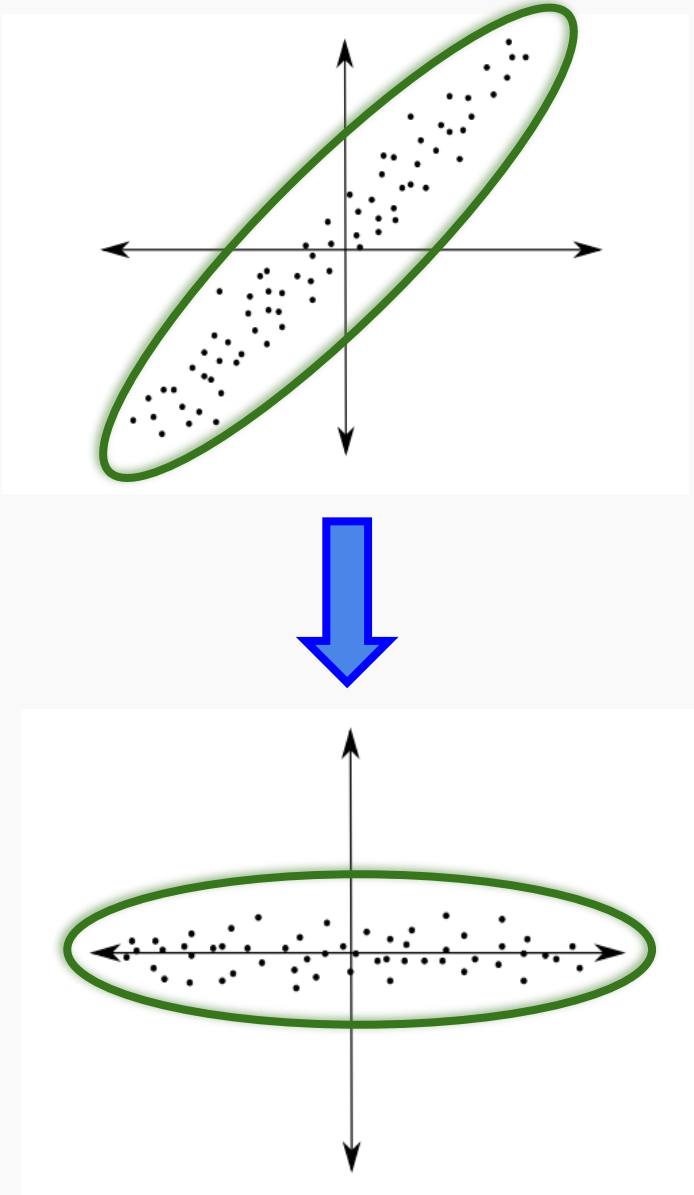
# PCA

PCA tries to fit an **ellipsoid** to the data.

PCA is a **linear transformation** that transforms data to a new coordinate system.

The data with the greatest variance lie on the first axis (first principal component) and so on.

PCA reduces the dimensions by throwing away the low variance principal components.



J. Jauregui (2012)

# PCA foundation

$$S = \frac{1}{n-1}\tilde{X}^T\tilde{X}$$

Note that the covariance matrix is symmetric, so it permits an orthonormal eigenbasis:

$$Sv_i = \lambda_i v_i$$

$$S = V\Lambda V^T$$

The eigenvalues can be sorted in $\Lambda$ as:

$$\lambda_1 > \lambda_2 > \ldots > \lambda_p$$

The eigenvector $v_i$ is called the ith **principal component** of $S$

# Measure the importance of the principal components

The **total sample variance** of the predictors:

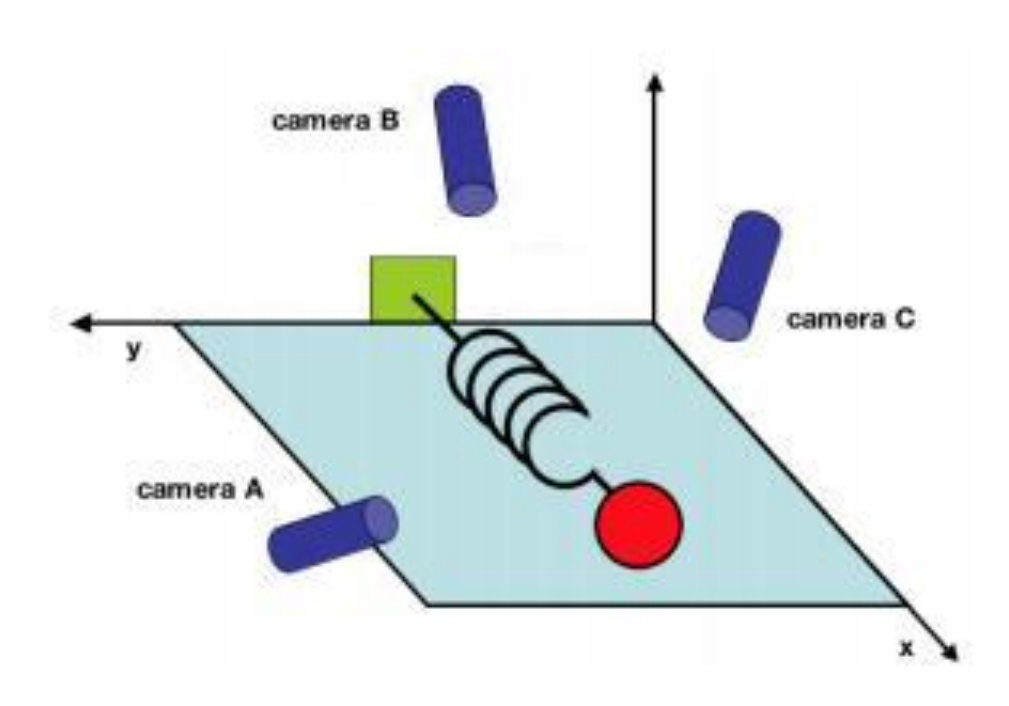
$$\mathrm{Tr}(S) = \sum_{j=1}^{p} S_{jj} = \frac{1}{n-1} \sum_{j=1}^{p} \sum_{i=1}^{n} (x_{ij} - \hat{\mu}_j)^2 = \sum_{i=1}^{p} \lambda_i$$

The fraction of the total sample variance that corresponds to $v_i$ :

$$\frac{\lambda_i}{\sum_{j=1}^{p} \lambda_j} = \frac{\lambda_i}{\mathrm{Tr}(S)}$$

so, $\lambda_i$ indicates the "importance" of the ith principal component.

# Back to spring-mass example




PCA finds:

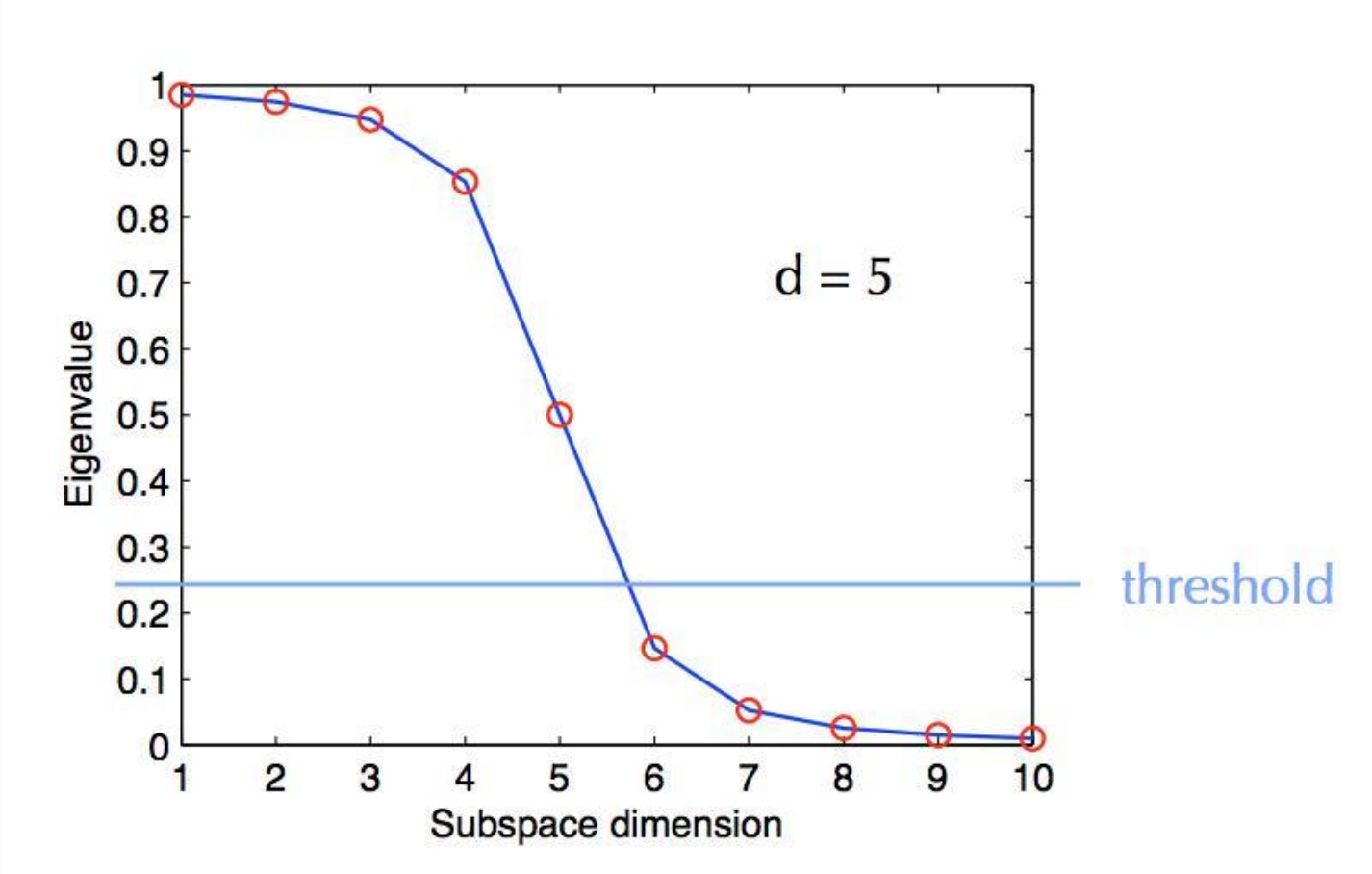
$$\lambda_1 / \sum_j \lambda_j \simeq 1$$

revealing the one-degree of freedom.

Hence, PCA indicates that there may be fewer variables that are essentially responsible for the variability of the response.

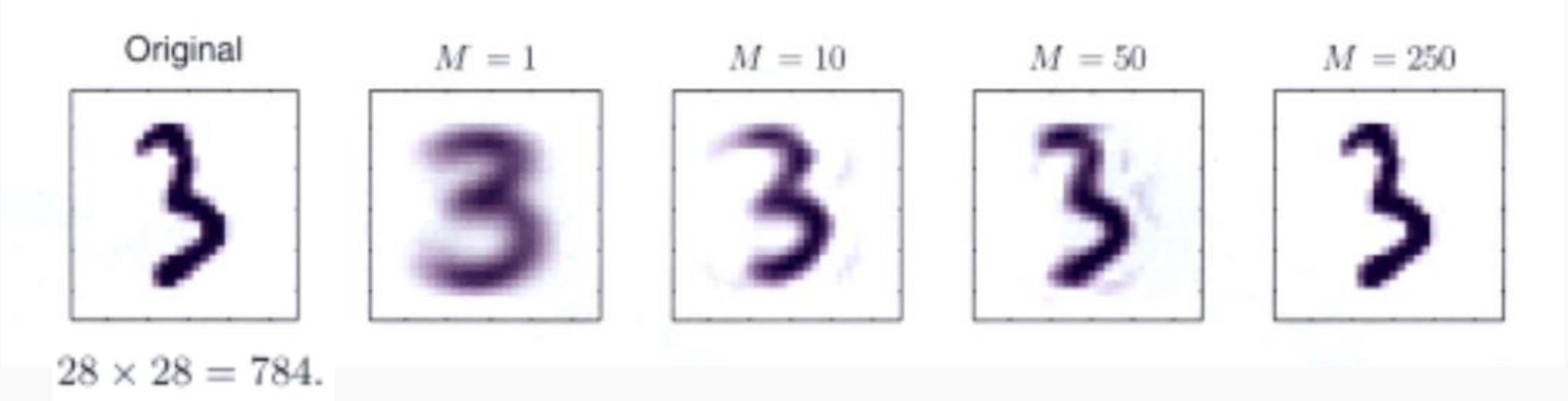
# PCA Dimensionality Reduction

The Spectrum represents the dimensionality reduction by PCA.



# PCA Dimensionality Reduction

There is no rule in how many eigenvalues to keep, but it is generally clear and left to the analyst's discretion.




C. Bishop, *Pattern Recognition and Machine Learning*, Springer (2008).

# PCA Dimensionality Reduction

An example on leaves (thanks to Chris Rycroft, AM205)

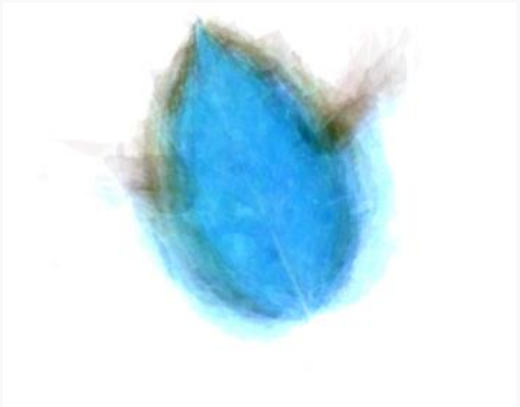
# PCA Dimensionality Reduction

The average leaf

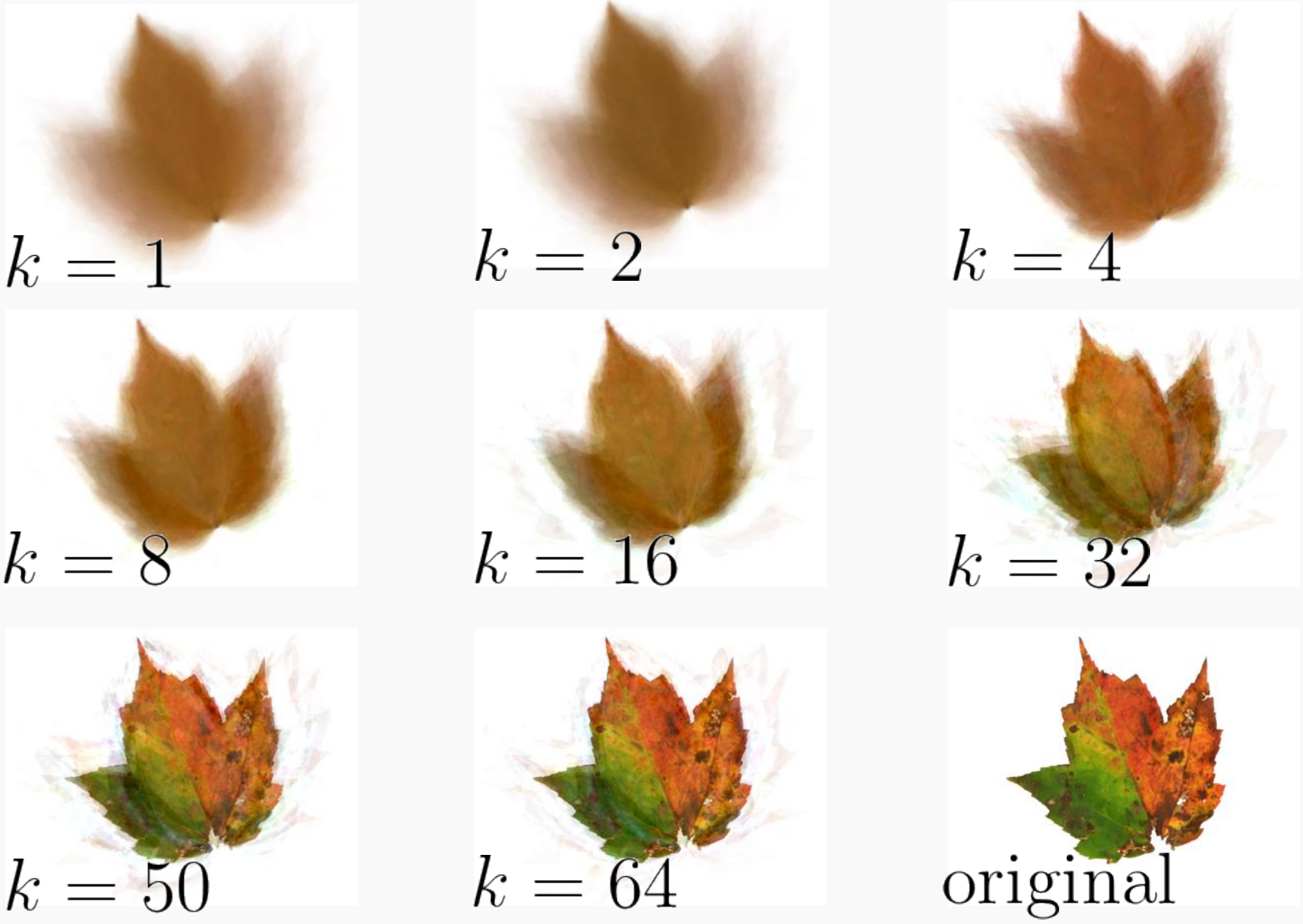
(Why do we need this again?)

# PCA Dimensionality Reduction

First three principal components

| | $v_1$ | $v_2$ | $v_3$ |
|---|---|---|---|
| positive |  |  |  |
| negative |  |  |  |

# PCA Dimensionality Reduction – Keeping up to k Components



# Assumptions of PCA

Although PCA is a powerful tool for dimension reduction, it is based on some strong assumptions.

The assumptions are reasonable, but they must be checked in practice before drawing conclusions from PCA.

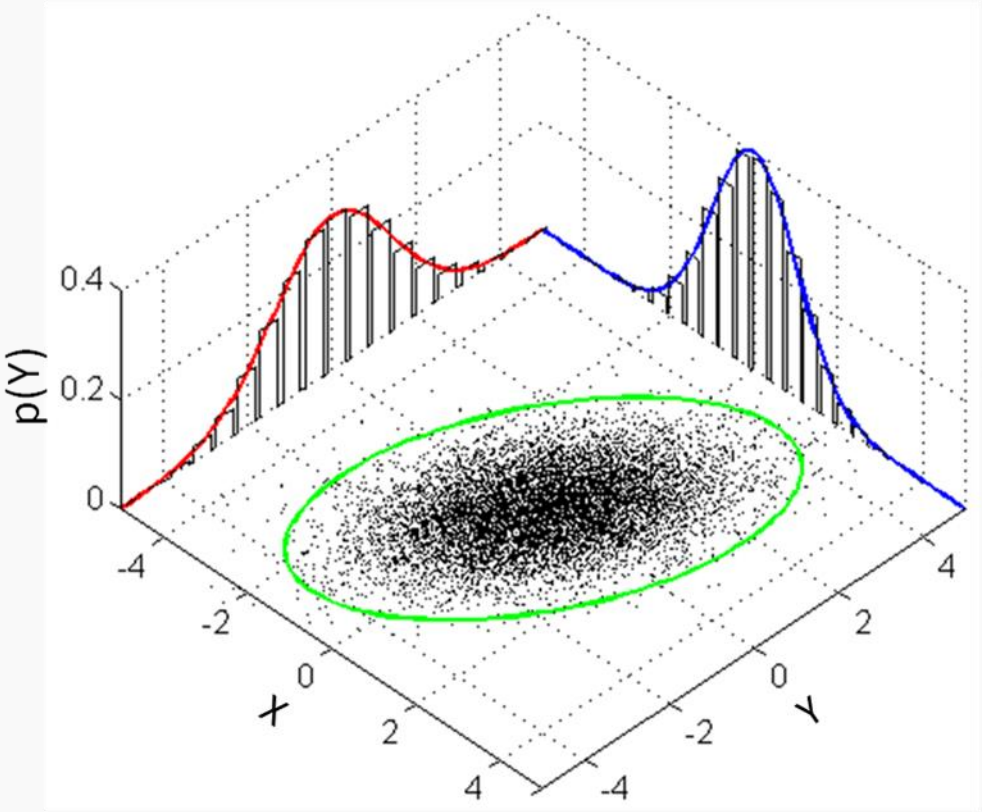
When PCA assumptions fail, we need to use other Linear or Nonlinear dimension reduction methods.

# Mean/Variance are sufficient

In applying PCA, we assume that means and covariance matrix are sufficient for describing the distributions of the predictors.

This is only exactly true if the predictors are drawn from a multivariable Normal distribution, but works approximately for many situations.

When a predictor deviates heavily from being Normally distributed, an appropriate nonlinear transformation may solve this problem.




Wikipedia – multivariate normal distribution

# High Variance indicates importance

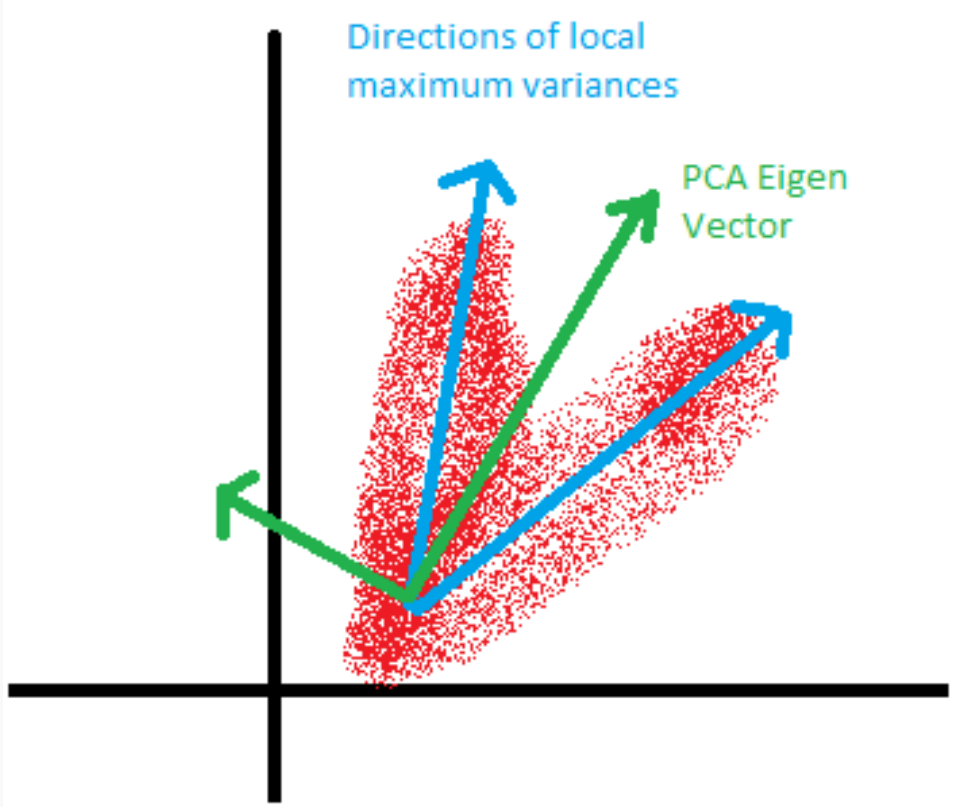
Assumption: The eigenvalue $\lambda_i$ is measures the "importance" of the i$^{th}$ principal component.

It is intuitively reasonable that lower variability components describe the data less, but it is not always true.

# Principal Components are orthogonal

PCA assumes that the *intrinsic dimensions* are orthogonal.

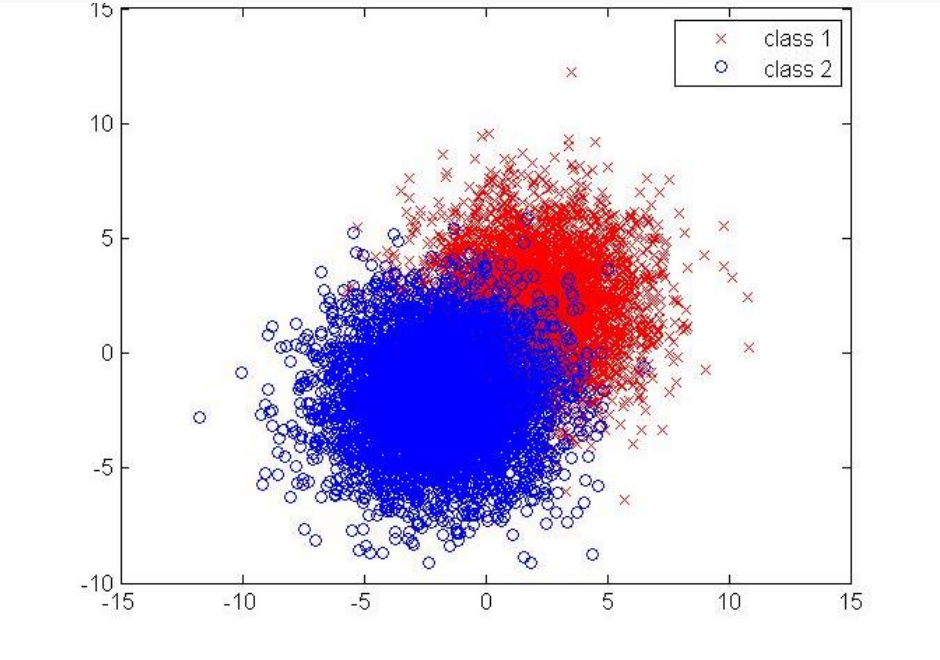
When this assumption fails, we need to assume non-orthogonal components which are not compatible with PCA.




Balaji Pitchai Kannu (on Quora)

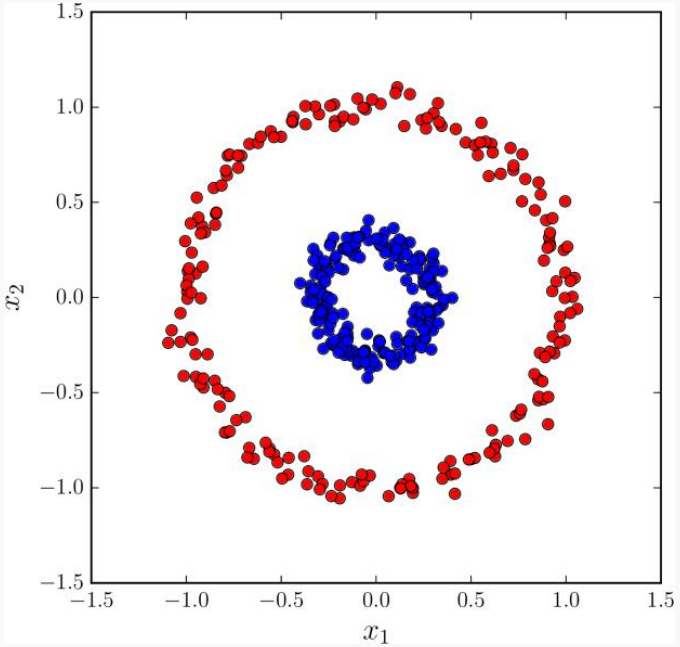
# Linear Change of Basis

PCA assumes that data lie on a lower dimensional linear manifold.




projectrhea.org

vs




Alexsei Tiulpin

When the data lie on a nonlinear manifold in the predictor space, then linear methods are likely to be ineffective.

# Kernel PCA for Nonlinear Dimensionality Reduction

Applying a nonlinear map $\Phi$ (called *feature map*) on data yields PCA kernel:

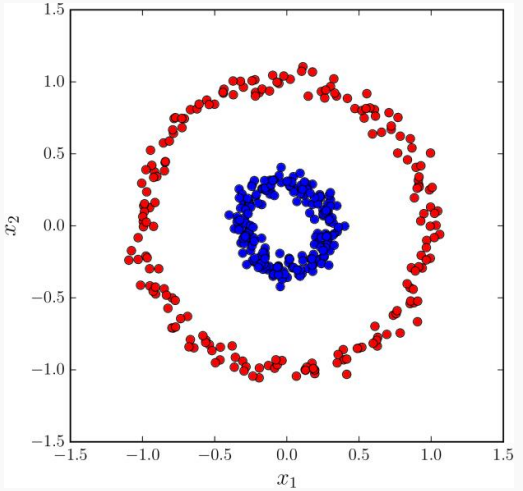
$$K = \Phi(X)^T \Phi(X)$$

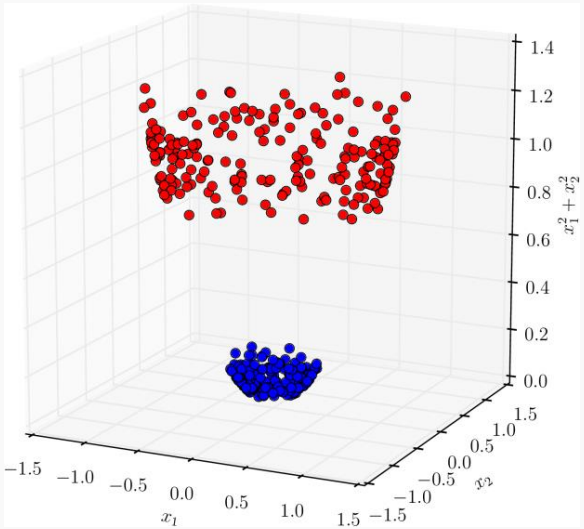
Centered nonlinear representation:

$$\tilde{\Phi}(X) = \Phi(X) - E[\Phi(X)]$$

Apply PCA to the modified Kernel:

$$\tilde{K} = \tilde{\Phi}(X)^T \tilde{\Phi}(X)$$




Alexsei Tiulpin

# Summary

- **Dimensionality Reduction Methods**
  1. A process of reducing the number of predictor variables under consideration.
  2. To find a more meaningful basis to express our data filtering the noise and revealing the hidden structure.

- **Principal Component Analysis**
  1. A powerful *Statistical* tool for analyzing data sets and is formulated in the context of *Linear Algebra*.
  2. Spectral decomposition: We reduce the dimension of predictors by reducing the number of principal components and their eigenvalues.
  3. PCA is based on strong assumptions that we need to check.
  4. Kernel PCA for nonlinear dimensionality reduction.

# Thank you