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

Marios Mattheakis and Pavlos Protopapas

CS109A Introduction to Data Science

Pavlos Protopapas and Kevin Rader

Outline

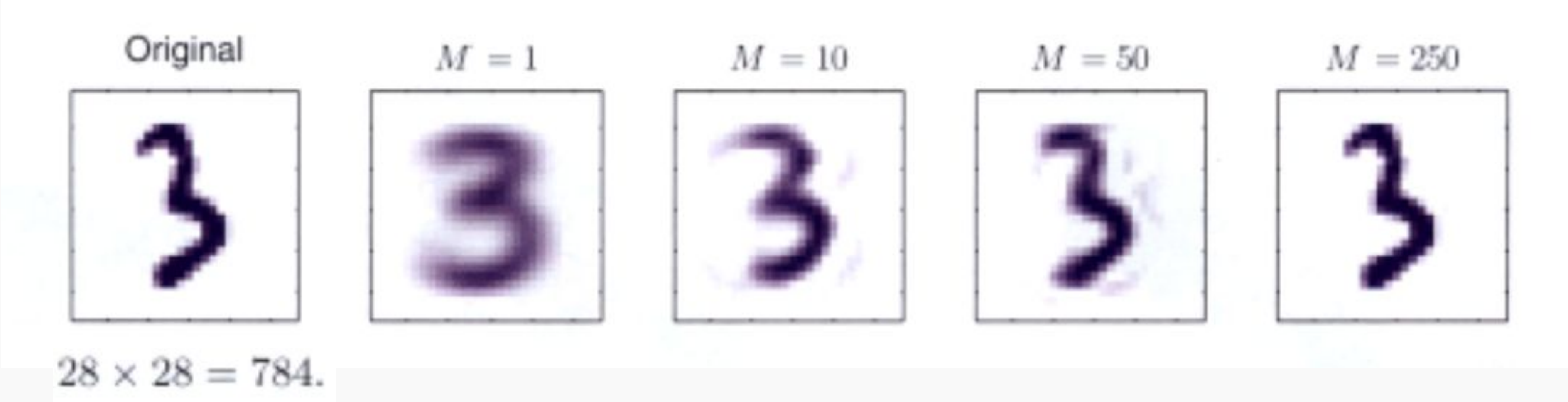
1. Introduction:
   a. Why Dimensionality Reduction?
   b. Linear Algebra (Recap).
   c. Statistics (Recap).

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.

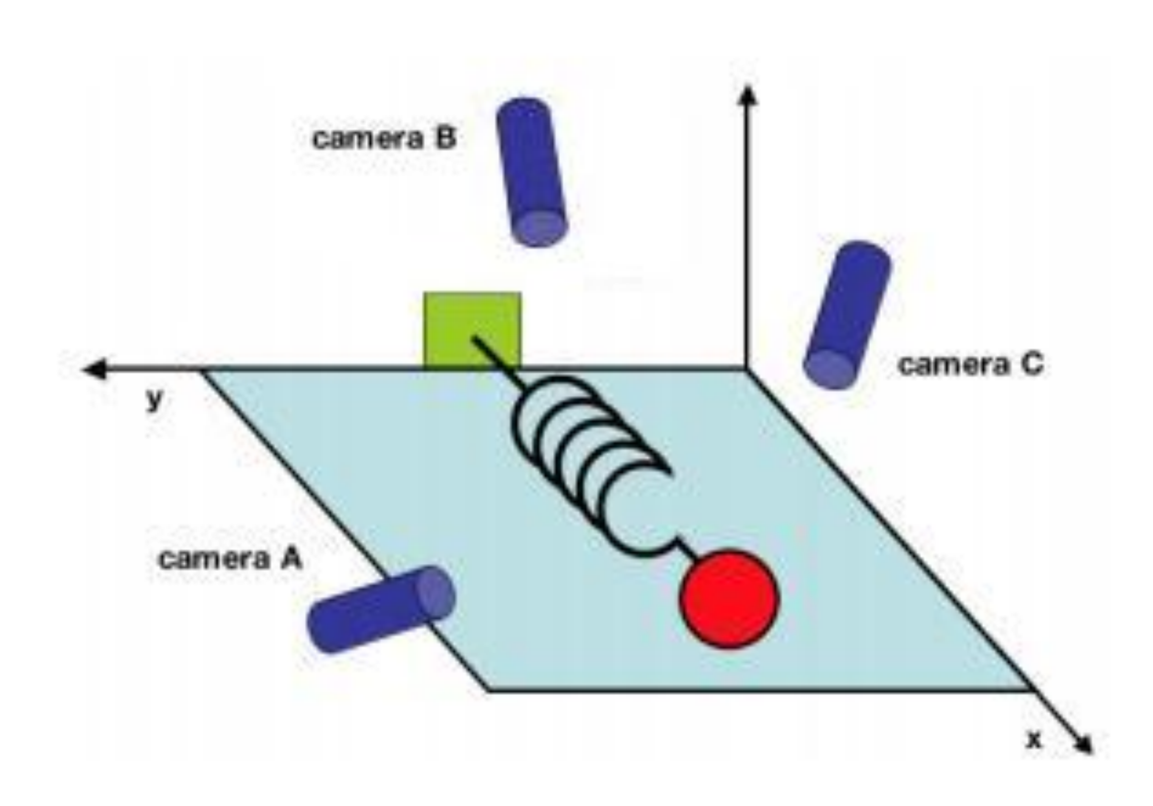
To find a more meaningful basis to express our data filtering the noise and revealing the hidden structure.



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

A simple example taken by Physics

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



LASSO regression model:

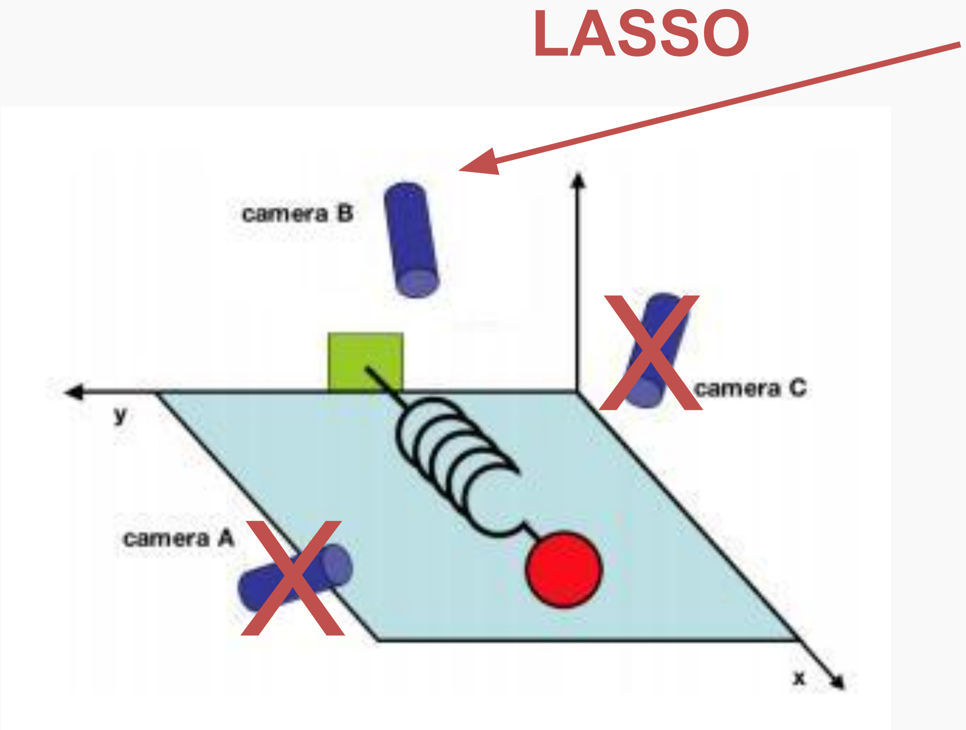
$$Y = \beta_A x_A + \beta_B x_B + \beta_C x_C$$

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

Suppose a design (or data) matrix consists of *n* observations and *p* predictors, hence:

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

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

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

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

and the associate linearly independent 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 consist an *orthonormal basis.*

Spectrum and Eigen-decomposition

Spectrum:

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

Unitary Matrix:

$$U = \begin{pmatrix} u_{11} & u_{21} & \cdots & u_{p1} \\ u_{12} & u_{22} & \cdots & u_{p2} \\ \vdots & \vdots & \ddots & \vdots \\ u_{1p} & u_{2p} & \cdots & u_{pp} \end{pmatrix}$$

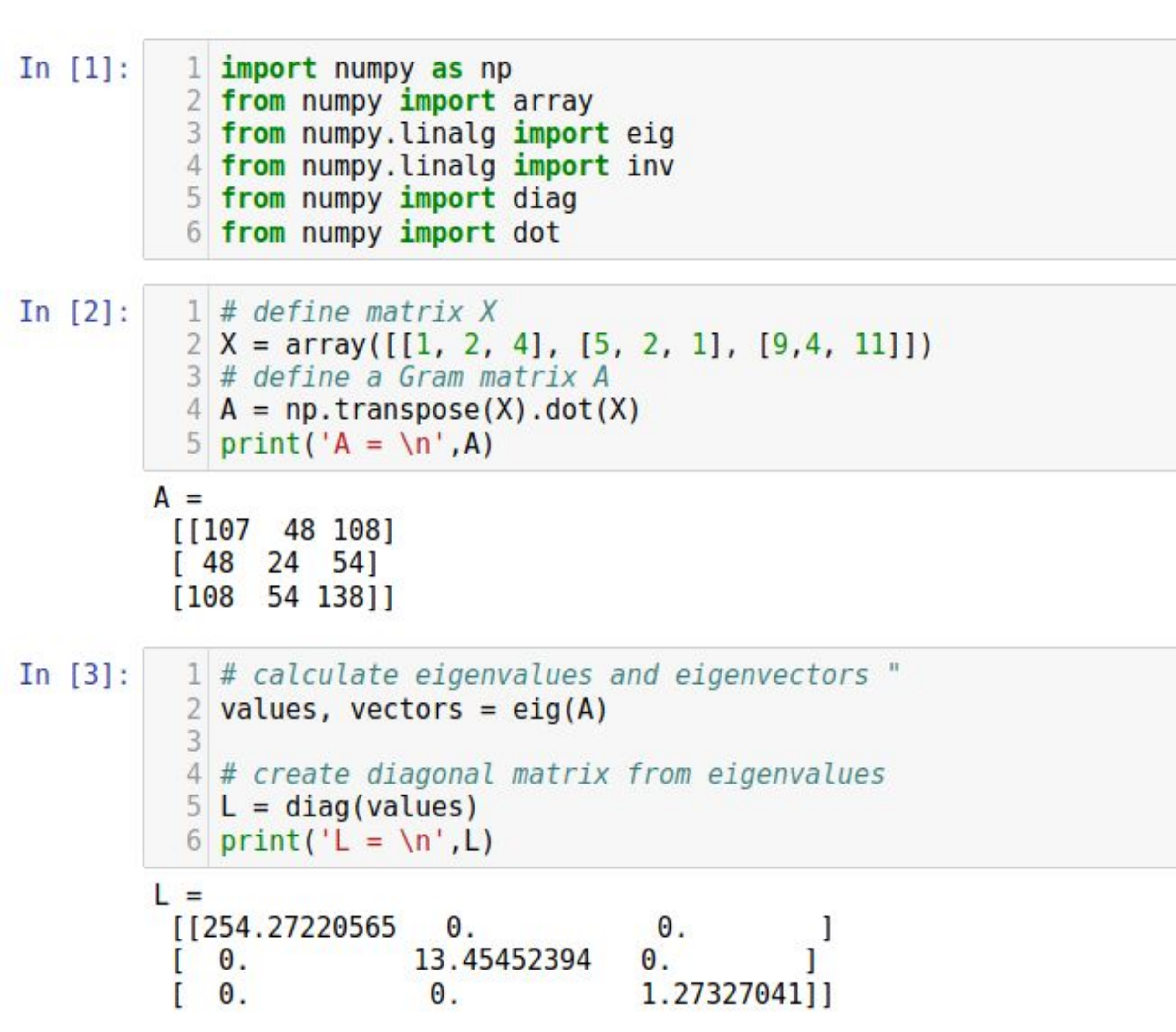
$$(U^{-1} = U^T)$$

$$(U^T U = \mathbf{I})$$

Eigen-decomposition:

$$A = U \Lambda U^T$$

Numerical verification of decomposition property



In [1]:

```
import numpy as np
from numpy import array
from numpy.linalg import eig
from numpy.linalg import inv
from numpy import diag
from numpy import dot
```

In [2]:

```
# define matrix X
X = array([[1, 2, 4], [5, 2, 1], [9,4, 11]])
# define a Gram matrix A
A = np.transpose(X).dot(X)
print('A = \n',A)
```

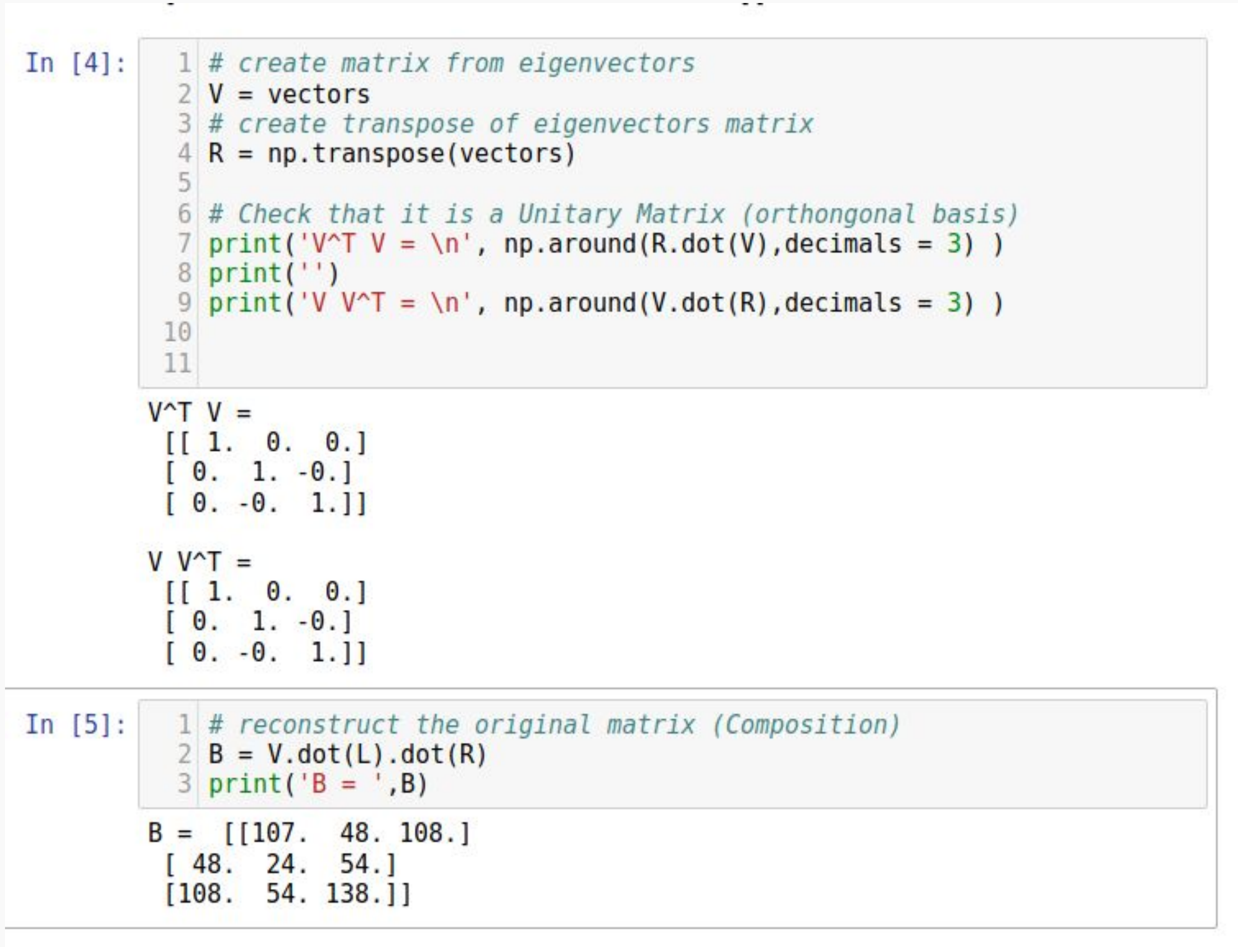
```
A =
 [[107  48 108]
 [ 48  24  54]
 [108  54 138]]
```

In [3]:

```
# calculate eigenvalues and eigenvectors "
values, vectors = eig(A)

# create diagonal matrix from eigenvalues
L = diag(values)
print('L = \n',L)
```

```
L =
 [[254.27220565   0.            0.        ]
 [  0.          13.45452394   0.        ]
 [  0.           0.            1.27327041]]
```



In [4]:

```
# create matrix from eigenvectors
V = vectors
# create transpose of eigenvectors matrix
R = np.transpose(vectors)

# Check that it is a Unitary Matrix (orthongonal basis)
print('V^T V = \n', np.around(R.dot(V),decimals = 3) )
print('')
print('V V^T = \n', np.around(V.dot(R),decimals = 3) )


```

```
V^T V =
 [[ 1.  0.  0.]
 [ 0.  1. -0.]
 [ 0. -0.  1.]]

V V^T =
 [[ 1.  0.  0.]
 [ 0.  1. -0.]
 [ 0. -0.  1.]]
```

In [5]:

```
# reconstruct the original matrix (Composition)
B = V.dot(L).dot(R)
print('B = ',B)
```

```
B =  [[107.  48. 108.]
 [ 48.  24.  54.]
 [108.  54. 138.]]
```

Real & Positive Eigenvalues: Gram Matrix

- The eigenvalues of $X^T X$ are positive and real numbers:

$$\begin{aligned} X^T X u &= \lambda u \\ u^T X^T X u &= u^T \lambda u \\ (Xu)^T (Xu) &= \lambda u^T u \\ ||Xu||^2 &= \lambda ||u||^2 \\ &\Rightarrow \lambda > 0 \end{aligned}$$

Similar for $X X^T$

➢ Hence, $X^T X$ and $X X^T$ are **Gram** matrices.

Same eigenvalues

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

$$
\begin{gathered}
X^T X u = \lambda u \\
X X^T X u = X \lambda u \\
X X^T (X u) = \lambda (X u) \\
X X^T \tilde{u} = \lambda \tilde{u}
\end{gathered}
$$

Same eigenvalues.

Transformed eigenvectors:

$$\tilde{u} = X u$$

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

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

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

$$\mathrm{Tr}(BC) = \sum_{i}^{m} (BC)_{ii} = \sum_{i}^{m} \sum_{j}^{n} B_{ij} C_{ji}$$

$$\sum_{i}^{m} \sum_{j}^{n} C_{ji} B_{ij} = \sum_{j}^{n} (CB)_{jj} = \mathrm{Tr}(CB)$$

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

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

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

Statistics (Recap)

Centered Model Matrix

Suppose 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)

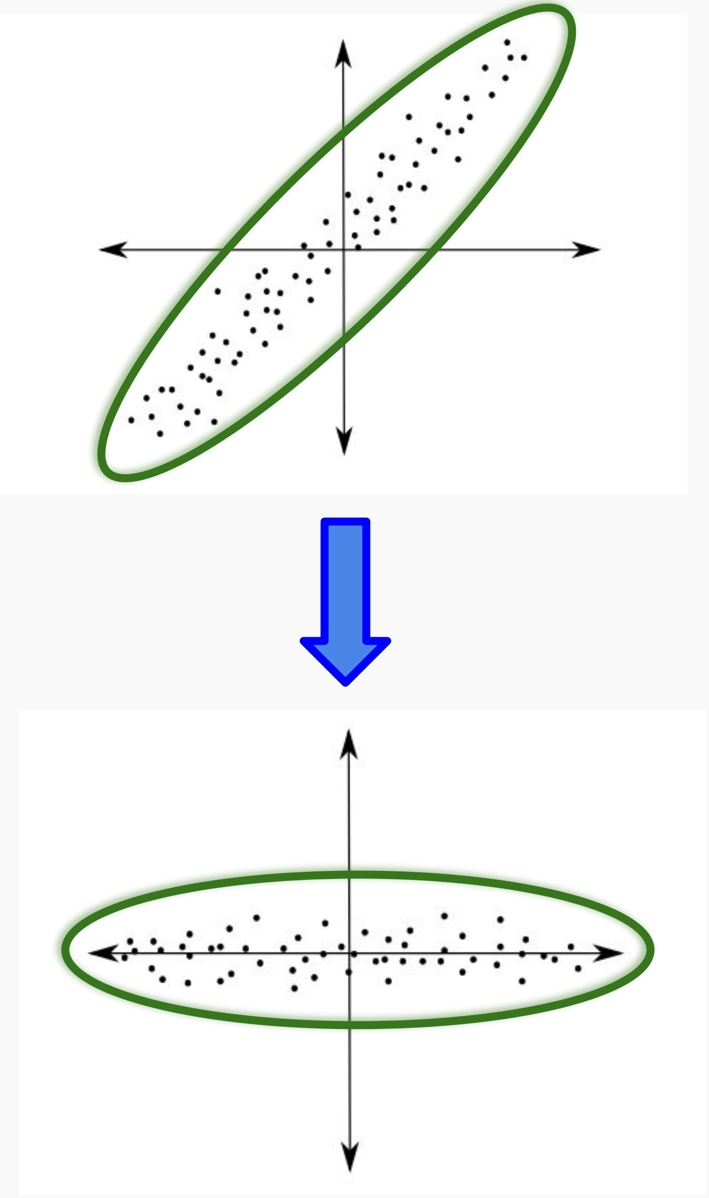
PCA

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

Since $X^T X$ is a Gram matrix, S will be a Gram matrix too, hence:

$$Sv_i = \lambda_i v_i$$

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

The eigenvalues are sorted in Λ 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:

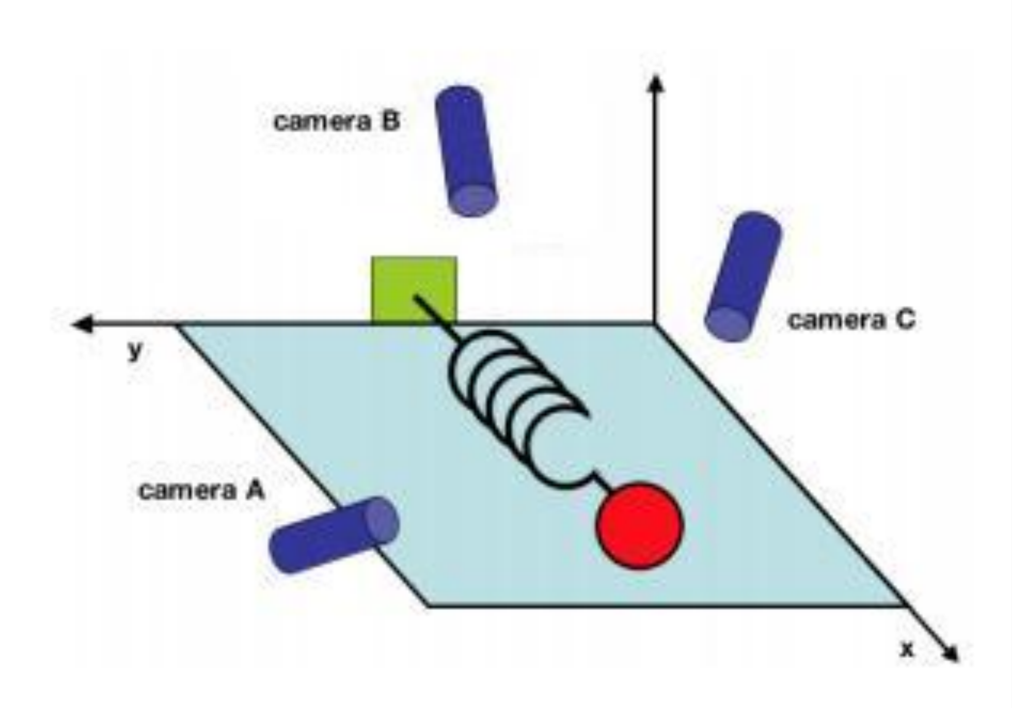
$$\mathrm{Tr}(S) = \sum_{j=1}^{p} S_{jj} = \frac{1}{n-1} \sum_{j=1}^{p} \sum_{i=1}^{n} (x_{ij} - \hat{\mu}_j)^2 = \sum_{i=1}^{p} \lambda_i$$

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, the λ_i indicates the "importance" of the i^th principal component.

Back to spring-mass example



PCA finds:

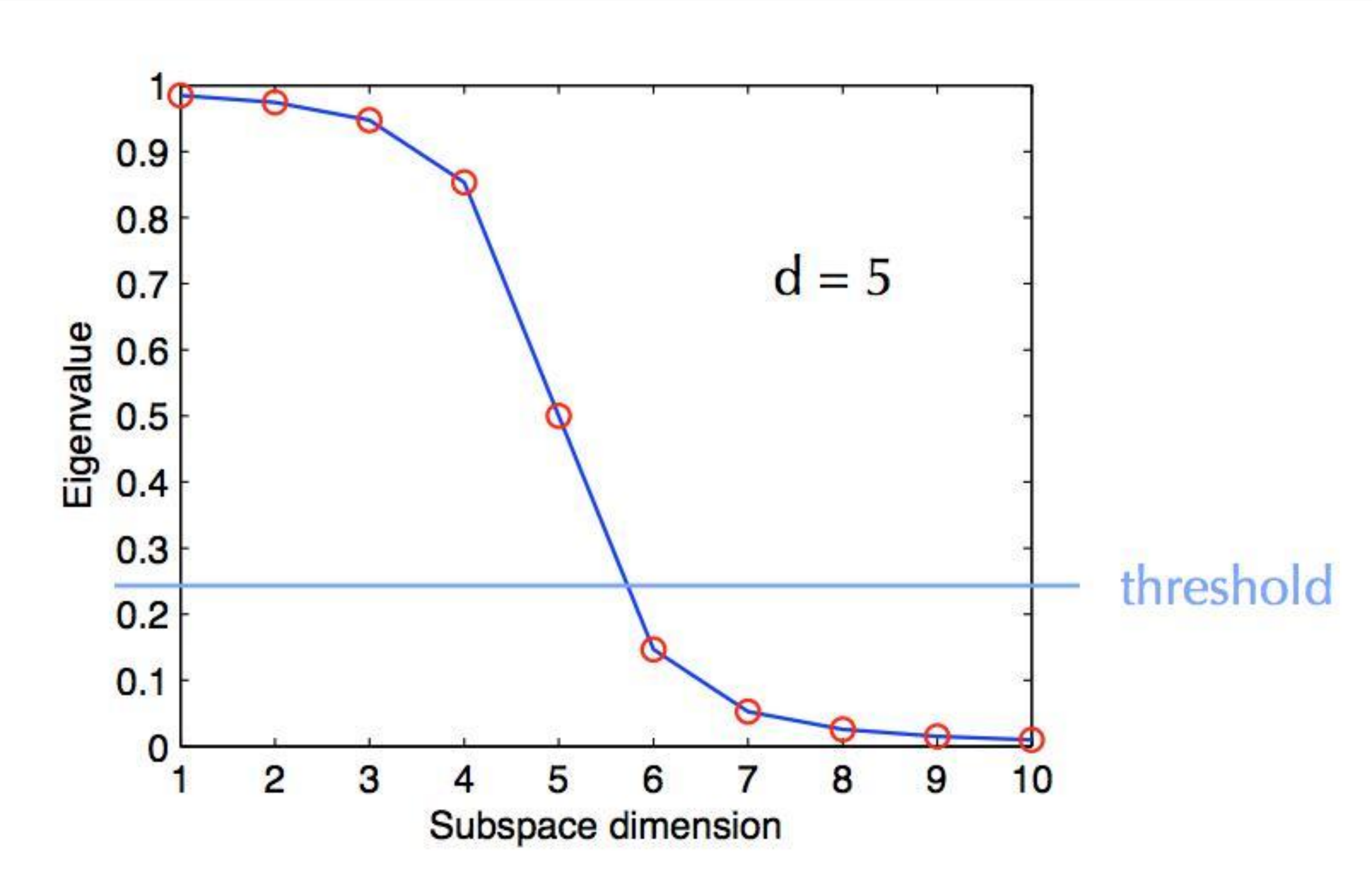
$$\lambda_1 / \sum_j \lambda_j \simeq 1$$

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.

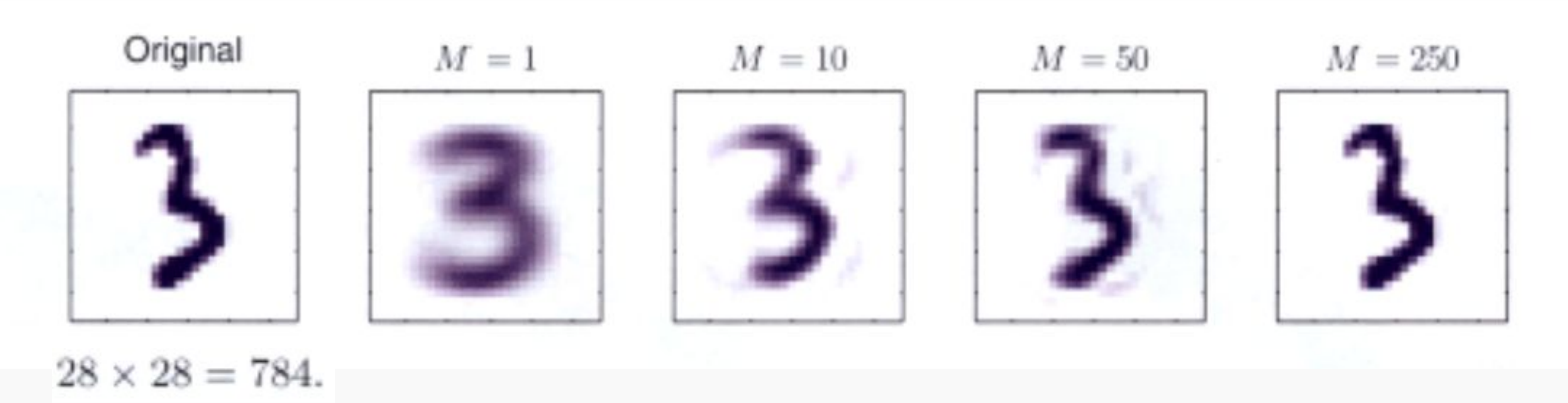
PCA Dimensionality Reduction

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 in analyst's discretion.



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

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 true only if the predictors are drawn by a multivariable Normal distribution, but approximately works for many situations.

When a predictor is heavily deviate from Normal distribution, an appropriate nonlinear transformation may solve this problem.

High Variance indicates importance

The eigenvalue λ_i is measures the "importance" of the ith principal component.

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

Principal Components are orthogonal

PCA assumes that the *intrinsic dimensions* are orthogonal allowing us to use linear algebra techniques.

When this assumption fails, we need to assume non-orthogonal components which are non compatible with PCA.

Linear Change of Basis

PCA assumes that data lie on a lower dimensional linear manifold. So, a linear transformation yields an orthonormal basis.

When the data lie on a nonlinear manifold in the predictor space, then linear methods are doomed to fail.

Kernel PCA for Nonlinear Dimensionality Reduction

Applying a nonlinear map Φ (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:

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

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.

Advanced Section 4: Dimensionality Reduction, PCA

Thank you

Office hours for Adv. Sec.

Monday 6:00-7:30 pm

Tuesday 6:30-8:00 pm