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

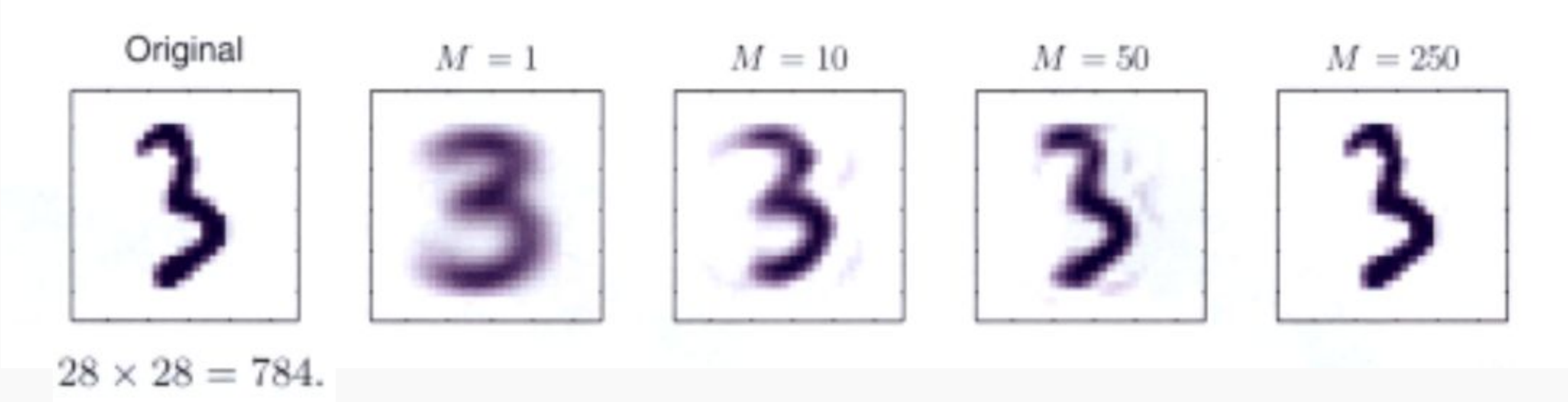
1. Introduction:
   a. Why Dimensionality Reduction?
   b. Linear Algebra (Recap)
   c. Statistics (Recap)

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.

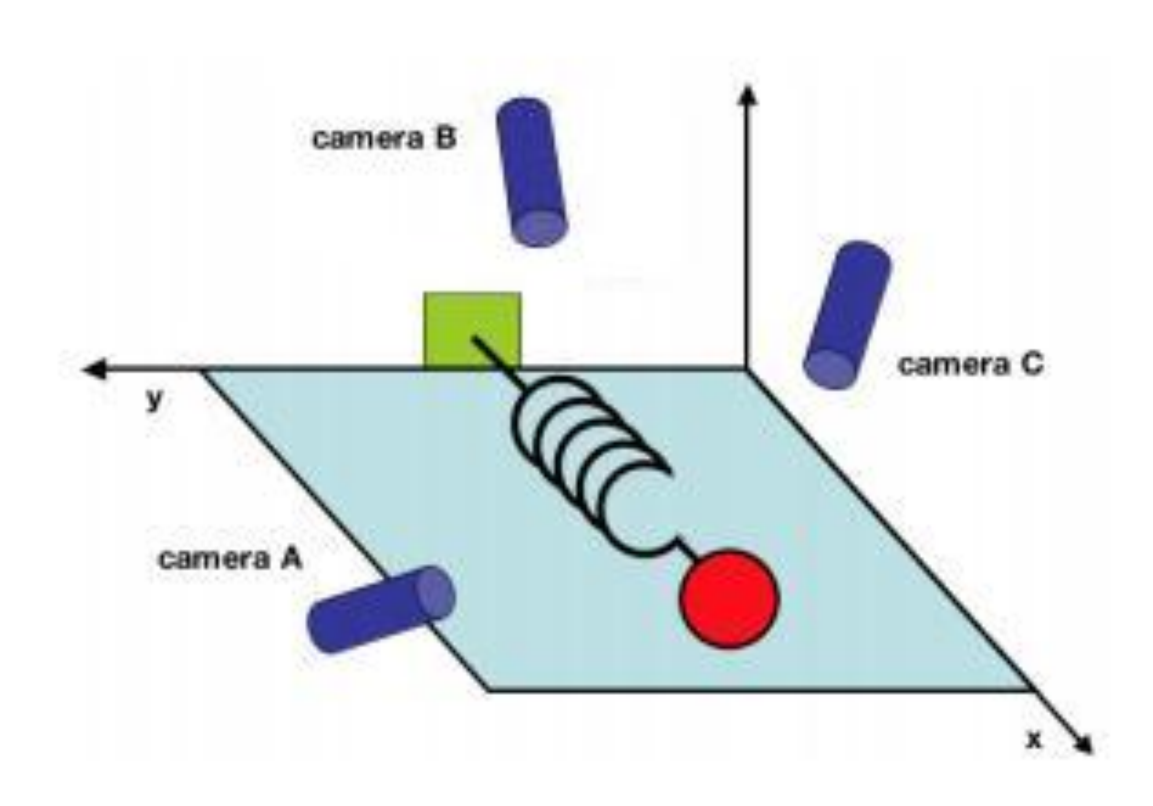
To find a more meaningful basis to express our data filtering the noise and revealing the hidden structure



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 force Y that spring exerts on the wall.



LASSO regression model:

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

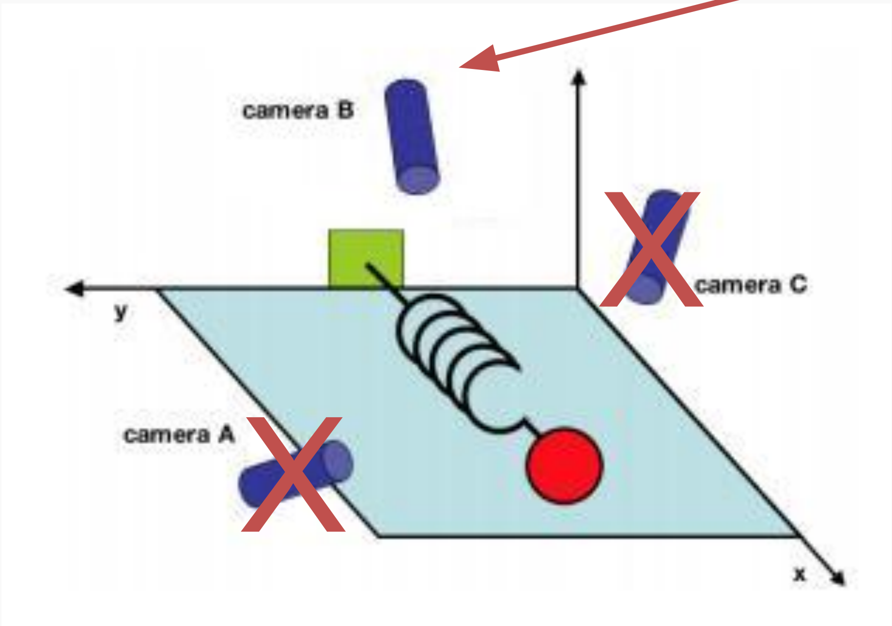
LASSO variable selection:

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

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

Principal Component Analysis versus LASSO

LASSO



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 that 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 an $n \times p$ arbitrary matrix of real numbers $X \in \mathbb{R}^{n \times p}$

- Then the $X^T X$ and $X X^T$ are symmetric matrices.

Symmetric property: $A^T = A$

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

Eigenvalues and Eigenvectors

Suppose a real, symmetric matrix: $A \in \mathbb{R}^{p \times p}$

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

and the associate linear 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:

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

Real & Positive Eigenvalues: Gram Matrix

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

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

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

Same eigenvalues

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

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

Same eigenvalues: λ_i

Modified eigenvectors: $\tilde{u}_i = Xu$

The sum of eigenvalues of $X^T X$ is equal to the 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.

$$\mathrm{Tr}(X^T X) = \mathrm{Tr}(U\Lambda U^T) = \mathrm{Tr}(U^T U\Lambda) = \mathrm{Tr}(\Lambda)$$

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

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

$$\tilde{X} = \left(X_1 - \hat{\mu}_1, \ldots, X_p - \hat{\mu}_p\right)$$

Sample Covariance Matrix

Consider the Covariance matrix

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

Inspecting the terms:

- Th 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

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

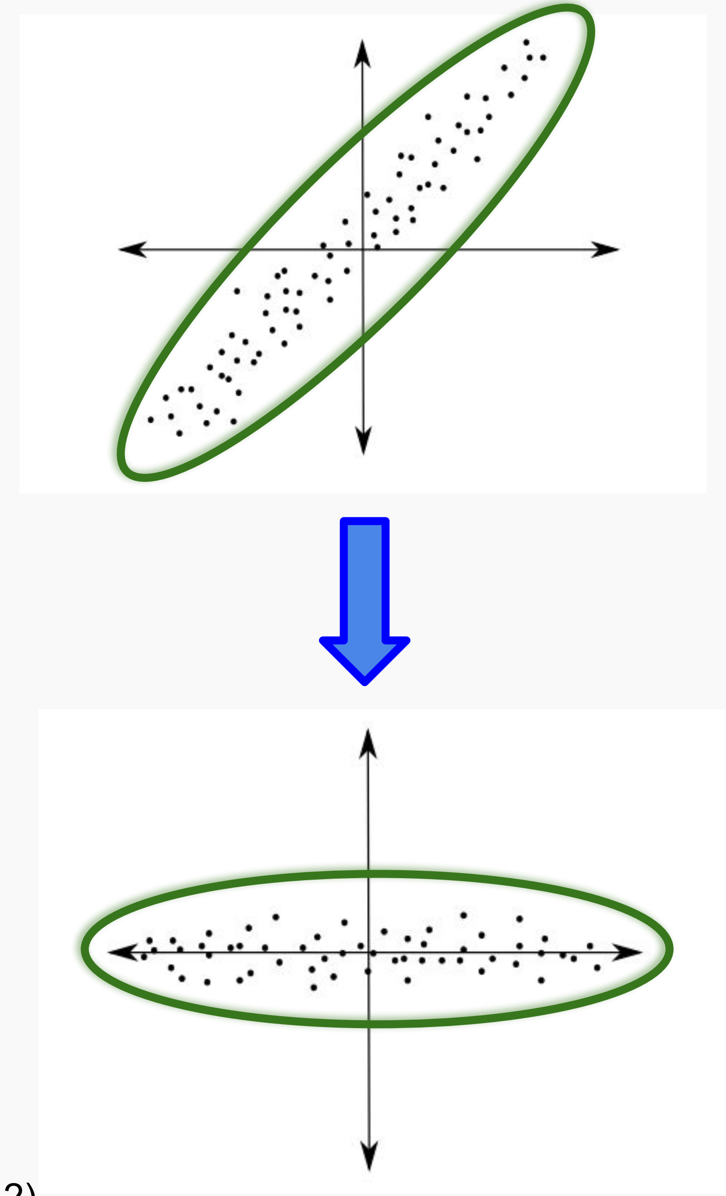
Principal Components Analysis (PCA)

PCA

PCA tries to fit an n-dimensional **ellipsoid** to the data.

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

Now, the data with the greatest variance lie on the first coordinate (first principal component) and so on.



J. Jauregui (2012)

PCA foundation

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

$$S v_i = \lambda_i v_i$$

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

The eigenvalues are sorted in the Λ

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

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:

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

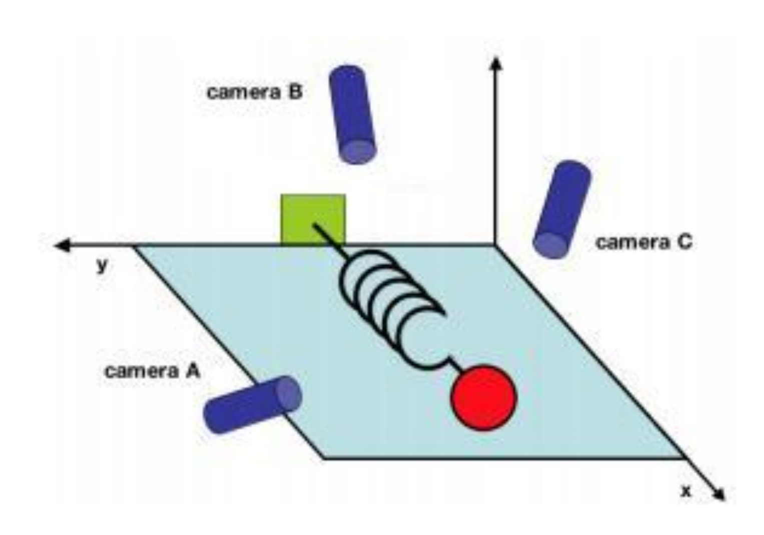
The fraction of the total sample variance that corresponds to v_i

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

so, the λ_i indicates the "importance" of the ith principal component

Back to spring-mass example

The principal comp. v_i denote directions in $\mathbb{R}^n$ that are “natural” for the data and linear combinations of the original coordinates.



For the spring-mass example:

$$v_1 = (0.2, 0.9, 0.4)$$

with

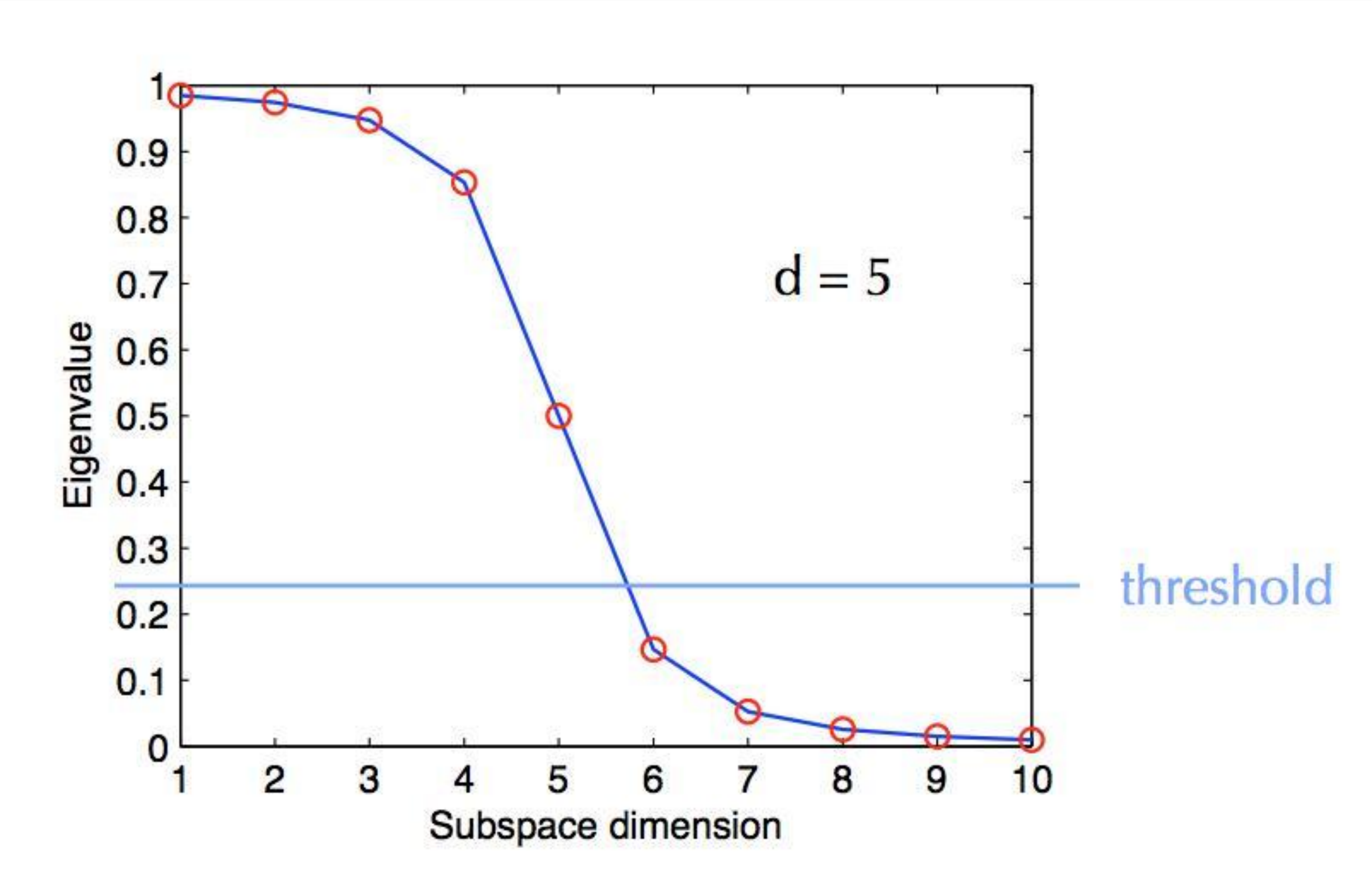
$$\lambda_1 / \sum_j \lambda_j \simeq 1$$

as it represents the x-axis

Hence, PCA indicates that there may be fewer variables, represented by v_i , that are responsible for the variability of the response

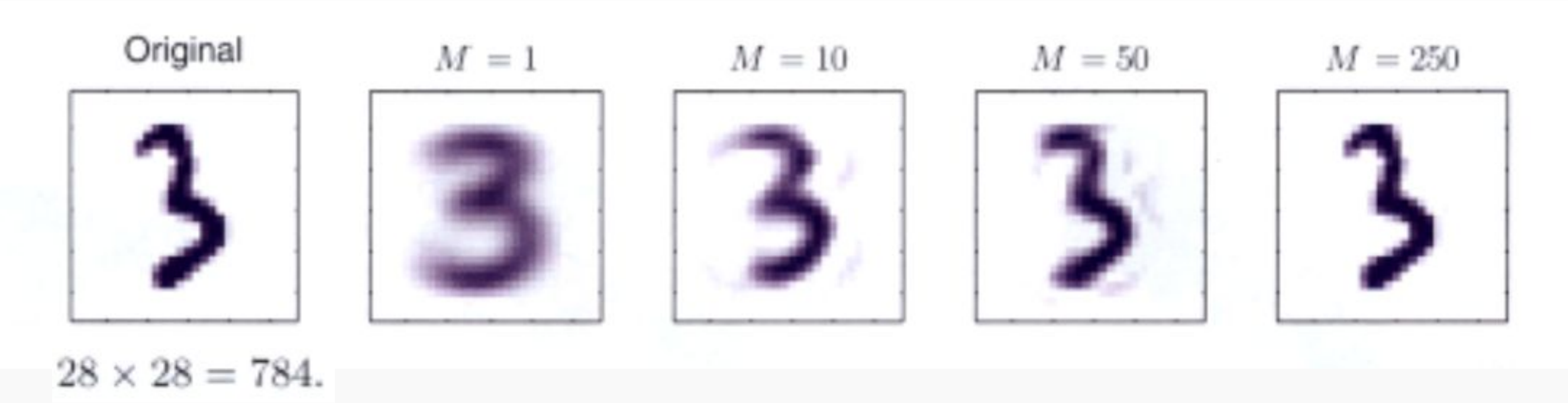
PCA Dimensionality Reduction

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 of the predictor may solve this problem.

High Variance indicates importance

The eigenvalue λ_i is measures the "importance" of the i^{th} principal component.

This is intuitively reasonable, that lower variability components describe lesse the data, but it is not always true and needs to be checked.

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 consists of a change of basis from Euclidean basis, where we measure the predictors, to an orthonormal basis of eigenvectors of $X^T X$

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.

Thank you

Office hours are:

Monday 6:00-7:30 pm

Tuesday 6:30-8:00 pm