Dimensionality Reduction

Introduction

Ira A. Fulton Schools of Engineering
Arizona State University

Objective

Objective

Illustrate the need for dimensionality reduction

What is Dimensionality Reduction?

- **We have N data points in a high-dimensional space,**
  - e.g., in the order of tens of thousands of dimensions.

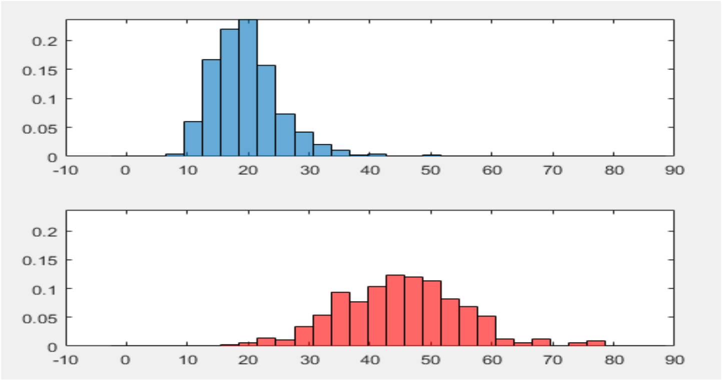
- **We want to project them into some low-dimensional space,**
  - e.g., in the order of tens of dimensions.

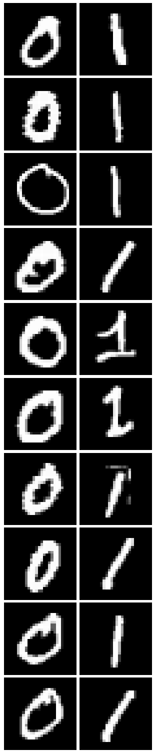
- **Why dimensionality reduction?**
  - A key technique to mitigate *curse of dimensionality*

The Curse of Dimensionality

- Consider histogram as a density estimator.

- Exponentially more samples would be needed in higher-dimensional spaces for the same “resolution”.





Many Techniques for Dimensionality Reduction

- **Many ways for going from a higher-dimensional space to a lower-dimensional space.**
  - Feature Selection achieves this by keeping only a subset of the original features/dimension.
- **There are many other techniques, employing a feature mapping/projection approach.**
  - New features are generated (instead of selecting only from the original features).
  - The underlying assumptions and/or goals of the techniques are often different.

Examples of Feature Mapping

- Linear discriminant analysis (LDA)
- Independent component analysis (ICA)
- Non-negative matrix factorization (NMF)
- Auto-encoder
- Self-organizing maps
- Principal component analysis (and its variants)

Dimensionality Reduction

Principal Component Analysis: Basic Idea

Ira A. Fulton Schools of Engineering
Arizona State University

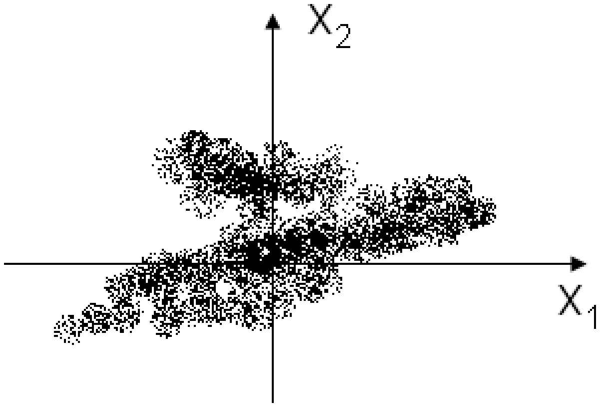
Objective

Objective

Illustrate the basic idea of Principal Component Analysis

Principal Component Analysis: Basic Idea

- **Look at a simple 2-D to 1-D example: we want to use a single feature to describe the 2-D samples**



- **Consider these possibilities**
  - Naïve: randomly discard one dimension
  - Better: discard the less-descriptive one (x_2 in the figure)
  - Much better: project the data to a most-descriptive direction and use the projections.

How to Formulate this Idea?

- "Most descriptive" ≈ Largest "variance"
- So the problem is to find the direction of the largest variance.

Problem

Given n samples $D=\{\mathbf{x}_1, \mathbf{x}_2,\ldots, \mathbf{x}_n\}$ in d-dimensional space, find a direction $\mathbf{e}_1$, such that the projection of D onto $\mathbf{e}_1$ gives the largest variance (compared with any other direction).

e_1 is a d-dimensional vector with unit norm.

Find e_1

- **Let's compute the variance of the projected data on a given direction e.**
  - The *n* projected samples are given as, for *i* = 1, …, *n*,

$$y_i = \mathbf{x}_i \cdot \mathbf{e}$$

  - The mean of the projections:

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i \cdot \mathbf{e} = \bar{\mathbf{x}} \cdot \mathbf{e}$$

  - Thus the (sample) variance of the projections:

$$\sigma^2(\mathbf{e}) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2 = \frac{1}{n}\sum_{i=1}^{n}[(\mathbf{x}_i - \bar{\mathbf{x}})\cdot \mathbf{e}]^2$$

n vs *n-1*

Find e_1 (cont'd)

- Expand the previous expression

$$\sigma^2(\mathbf{e}) = \sum_{j=1}^{d}\sum_{k=1}^{d} e_j e_k [\frac{1}{n}\sum_{i=1}^{n}(x_{i,j}-\bar{x}_{i,j})(x_{i,k}-\bar{x}_{i,k})]$$

k-th component of $\mathbf{x}_i$

$$=\sum_{j=1}^{d}\sum_{k=1}^{d} e_j e_k C_{jk} = \mathbf{e}^t C\mathbf{e}$$

k-th component of $\mathbf{e}$

(j,k)-th element of the matrix C

- C is the sample covariance matrix.

Find e_1 (cont'd)

- To find e_1, we can do

$$\mathbf{e}_1 = \arg\max_{\mathbf{e}} \sigma^2(\mathbf{e}) \quad \underline{\text{subject to } \|\mathbf{e}\| = 1}$$

what if without this constraint?

- Constrained maximization: use Lagrange multiplier method.

$$\text{maximize} \quad F(\mathbf{e}) = \mathbf{e}^t C\mathbf{e} - \underset{\substack{\uparrow \\ \text{Lagrange multiplier}}}{\lambda}(\mathbf{e}^t\mathbf{e} - 1)$$

Find e_1 (cont'd)

- Set the partial derivative to 0, we have

$$\partial F / \partial \mathbf{e} = 2C\mathbf{e} - 2\lambda\mathbf{e} = 0$$

$$\rightarrow \quad C\mathbf{e} = \lambda\mathbf{e}$$

- The solution is an eigenvector of *C*, with eigenvalue *λ*, which is also the variance under e:

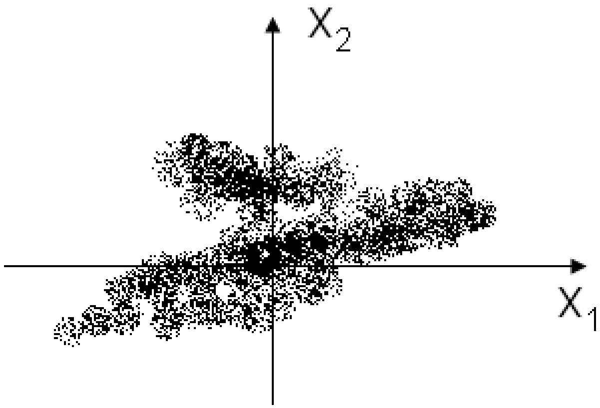
$$\sigma^2(\mathbf{e}) = \mathbf{e}^t C\mathbf{e} = \lambda$$

- We should set e_1 to be the eigenvector corresponding to the largest eigenvalue λ_1.

Recap of the Key Idea

- **We want to project the given data samples to certain direction so that the variance is maximized, compared with any other direction.**

- **We figured out what this optimal direction e_1 should be:**
  - It should be the eigenvector of corresponding to the largest eigenvalue λ_1, of the covariance matrix.



Dimensionality Reduction

Principal Component Analysis: The Algorithm & Important Properties

Ira A. Fulton Schools of Engineering
Arizona State University

Objective

Objective

Implement the PCA algorithm

Objective

Discuss some important properties of PCA

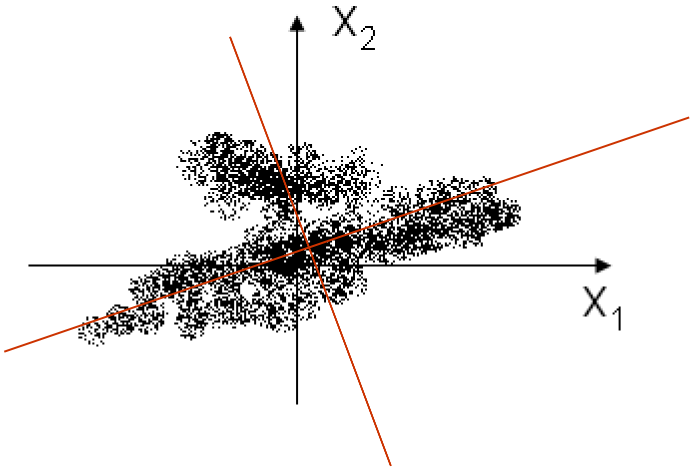
Principal Components

- **We found $\mathbf{e}_1$, which gives the direction of the largest variance after projection**
  - The first **principal component.**

- **The process can be continued in the subspace orthogonal to $\mathbf{e}_1$, and so on and so forth.**
  - Obtaining other principal components: $\mathbf{e}_2$, $\mathbf{e}_3$, etc., corresponding to other eigenvectors of C, ordered by the corresponding eigenvalues λ_i

Principal Components (cont'd)

- The principal components are orthogonal to each other → $\{e_i\}$ forms an orthonormal basis in the *d*-dimensional space.



- The total variance is given by the sum of the variances of the projections.

$$\sigma^2 = \sum_{j=1}^{d} \lambda_j$$

How Many Principal Components to Keep?

- To reduce dimensions, we will need to keep only *d'* << *d* projections.
- We can measure how much of the total variance a *d'*-dimensional subspace captures, by the ratio

$$\sum_{j=1}^{d'} \lambda_j \Big/ \sum_{j=1}^{d} \lambda_j$$

- Variance may be related to the "energy" of a signal: how accurately we want to represent the data.
  - The ratio can be used to guide in choosing a proper *d'* for desired accuracy.

The PCA Algorithm

1. Compute the dxd sample covariance matrix C

2. Find the eigenvalues and corresponding eigenvectors of C

3. Project the original data onto the space spanned by the eigenvectors
   - The projection may be done onto a d'-dimensional subspace spanned by the first d' eigenvectors (ordered by the eigenvalue in descending order)
     - d' is determined by the desired accuracy

Important Properties of PCA

- **PCA represents the data in a new space, in which the components of the data is ordered by their “significance”.**
  - Dimension reduction can be done by simply discarding less significant dimensions.
- **Linearity assumption ➔ extensions exist**
- **“Variance ≈ Importance” is meaningful only under large *signal-to-noise ratio***

PCA as Feature Mapping

- **When we use only *d'* dimensions from PCA (with original dimension *d* > *d'*), this may look like feature section.**
  - But in general they are different approaches.
- **PCA**
  - Unsupervised (in general)
  - Generates new features (linear combination of original ones)
- **Feature Selection**
  - Supervised (in general)
  - Selects a few original features (e.g., for better classification)

Can PCA Help Classification?

- **Can we do better classification in a lower-dimensional space from *d'* principal components given by PCA?**
  - Not necessarily.

- **LDA may be better posed for such a task.**