Principal Component Analysis: Basic Idea

Ira A. Fulton Schools of Engineering
Arizona State University

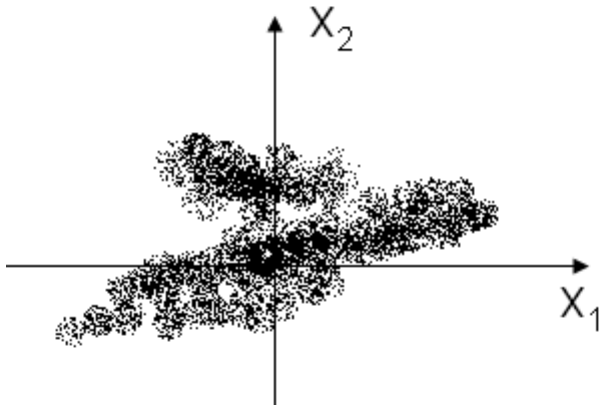
Objective

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).

$\mathbf{e}_1$ is a d-dimensional vector with unit norm.

Find $\mathbf{e}_1$

Let's compute the variance of the projected data on a given direction **e**.

– The n projected samples are given as, for $i = 1, \ldots, n$,

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

– 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 $\mathbf{e}_1$ (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 $\mathbf{x}_\mathrm{i}$

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

k-th component of **e**

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

– C is the sample covariance matrix.

Find $\mathbf{e}_1$ (cont'd)

➔ To find $\mathbf{e}_1$, 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.

$$\text{maximize} \quad \mathrm{F}(\mathbf{e}) = \mathbf{e}^t C \mathbf{e} - \lambda(\mathbf{e}^t \mathbf{e} - 1)$$

Lagrange multiplier

Find $\mathbf{e}_1$ (cont'd)

Set the partial derivative to 0, we have

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

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

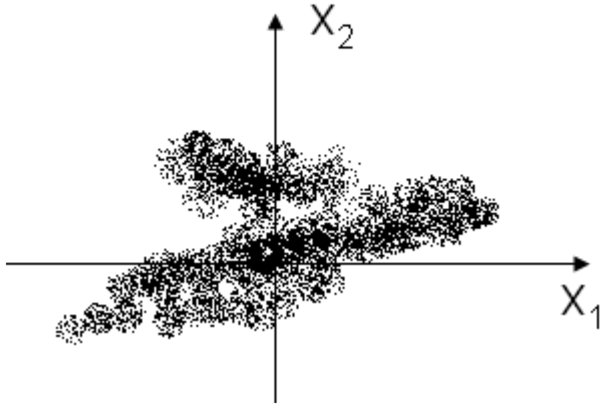
$\rightarrow$ The solution is an eigenvector of C, with eigenvalue λ, which is also the variance under $\mathbf{e}$:

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

$\rightarrow$ We should set $\mathbf{e}_1$ 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 $\mathbf{e}_1$ should be:

  → It should be the eigenvector of corresponding to the largest eigenvalue λ_1, of the covariance matrix.