Dimensionality Reduction

Via Principal Component Analysis

Probability and Statistics for Data Science

Carlos Fernandez-Granda

NYU COURANT INSTITUTE OF MATHEMATICAL SCIENCES

NYU DATA SCIENCE

These slides are based on the book Probability and Statistics for Data Science by Carlos Fernandez-Granda, available for purchase here. A free preprint, videos, code, slides and solutions to exercises are available at https://www.ps4ds.net

Motivation

Data with a large number of features can be difficult to analyze/process

Solution: Reduce dimensionality while preserving as much information as possible

Important preprocessing step in many applications

Linear dimensionality reduction

We model data as samples from d-dimensional random vector $\tilde{x}$

$\tilde{x}$ has zero mean (otherwise we center by subtracting mean)

For any orthonormal basis $b_1, \ldots, b_d$

$$\tilde{x} = \sum_{i=1}^{d} \tilde{a}[i] b_i \qquad \tilde{a}[i] := b_i^T \tilde{x}$$

$\tilde{a}[1], \tilde{a}[2], \ldots, \tilde{a}[d]$ is an equivalent representation of $\tilde{x}$

Idea: Use only first $k < d$ coefficients

$$\operatorname{approx}_{b_1,\ldots,b_k}(\tilde{x}) := \sum_{i=1}^{k} \tilde{a}[i] b_i \qquad \tilde{a}[i] := b_i^T \tilde{x}$$

How do we choose $b_1, \ldots, b_k$?

$$\underbrace{\sum_{i=1}^{d} \tilde{a}[i] b_i}_{\tilde{x}} = \underbrace{\sum_{i=1}^{k} \tilde{a}[i] b_i}_{\underset{b_1,\ldots,b_k}{\text{approx}}(\tilde{x})} + \underbrace{\sum_{i=k+1}^{d} \tilde{a}[i] b_i}_{\text{error}}$$

$$||\tilde{x}||_2^2 = \left|\left|\sum_{i=1}^{k} \tilde{a}[i] b_i\right|\right|_2^2 + ||\text{error}||_2^2$$

$$||\text{error}||_2^2 = ||\tilde{x}||_2^2 - \left|\left|\sum_{i=1}^{k} \tilde{a}[i] b_i\right|\right|_2^2$$

$$= ||\tilde{x}||_2^2 - \sum_{i=1}^{k} \tilde{a}[i]^2$$

$$= ||\tilde{x}||_2^2 - \sum_{i=1}^{k} \left(b_i^T \tilde{x}\right)^2$$

Mean ℓ_2-norm error

$$
\begin{aligned}
\mathrm{E}\left[||\text{error}||_2^2\right] &= \mathrm{E}\left[||\tilde{x}||_2^2\right] - \sum_{i=1}^{k} \mathrm{E}\left[\left(b_i^T \tilde{x}\right)^2\right] \\
&= \mathrm{E}\left[||\tilde{x}||_2^2\right] - \sum_{i=1}^{k} \mathrm{Var}\left[b_i^T \tilde{x}\right]
\end{aligned}
$$

What $b_1, \ldots, b_k$ maximize directional variance?

Principal directions

Let $u_1, \ldots, u_d$ be the eigenvectors of covariance matrix $\Sigma_{\tilde{x}}$

$$u_1 = \arg \max_{||a||_2=1} \mathrm{Var}[a^T \tilde{x}]$$

$$u_k = \arg \max_{||a||_2=1, a \perp u_1, \ldots, u_{k-1}} \mathrm{Var}[a^T \tilde{x}] \qquad 2 \leq k \leq d$$

Principal component analysis

Let u_1, ..., u_d be the eigenvectors of covariance matrix $\Sigma_{\tilde{x}}$

$$\{u_1, \ldots, u_k\} = \arg \min_{\substack{\{b_1,\ldots,b_k\} \\ ||b_i||_2=1, 1\leq i \leq k \\ b_i \perp b_j, i \neq j}} \mathrm{E}\left[\left|\left|\tilde{x} - \underset{b_1,\ldots,b_k}{\text{approx}}(\tilde{x})\right|\right|_2^2\right]$$

The optimal linear k-dimensional approximation is

$$\underset{u_1,\ldots,u_k}{\text{approx}}(\tilde{x}) = \sum_{i=1}^{k} \tilde{w}_i u_i \qquad \tilde{w}_i := u_i^T \tilde{x}$$

Wheat seeds

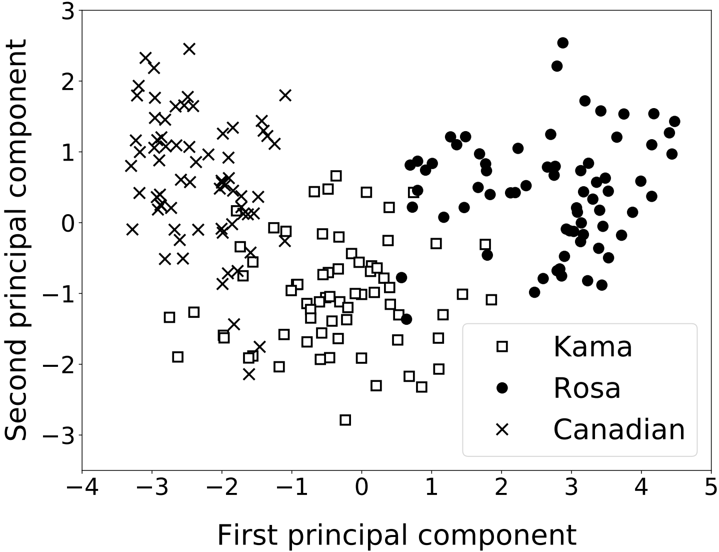
3 varieties: Kama, Rosa and Canadian

Features:

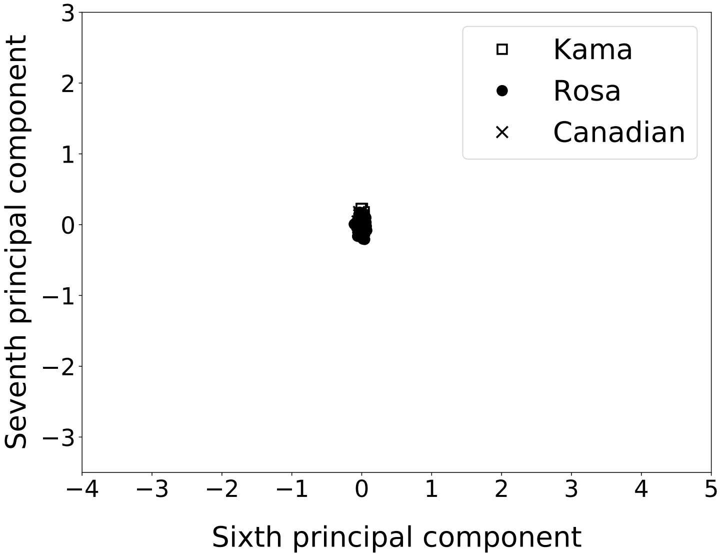
- Area
- Perimeter
- Compactness
- Length of kernel
- Width of kernel
- Asymmetry coefficient
- Length of kernel groove

Challenge: How to visualize the data in two dimensions?

Two first principal components



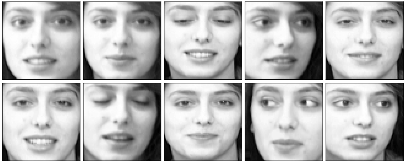
Two last principal components



Faces

64×64 images from 40 subjects

Vectorized images interpreted as vectors in $\mathbb{R}^{4096}$



Sample mean

Principal directions

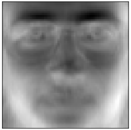
u_1

18.8

u_2

11.1

u_3



6.30

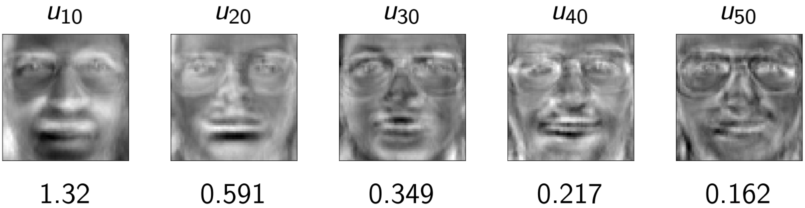
u_4

3.95

u_5

2.86

Principal directions



Principal directions

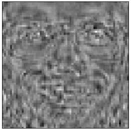
u_{100}

0.061

u_{200}

0.019

u_{250}

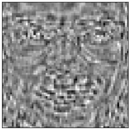


0.011

u_{300}

0.008

u_{350}

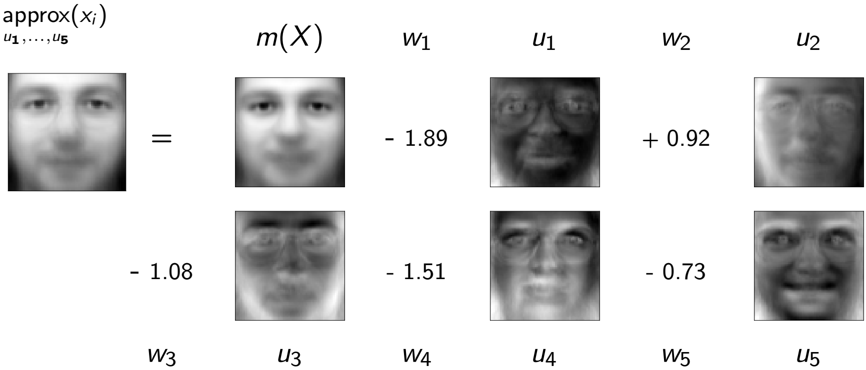


0.004

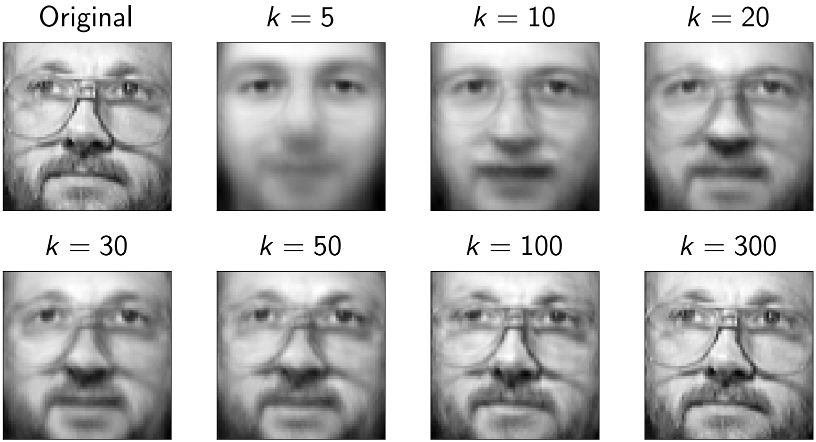
Faces

$$\underset{u_1,\ldots,u_5}{\text{approx}}(x_i) := m(X) + \sum_{j=1}^{5} w_j[i]u_j \qquad w_j[i] := u_j^T \operatorname{ct}(x_i)$$

$k = 5$


$\text{approx}_{u_1,\ldots,u_5}(x_i)$
$m(X)$
w_1
u_1
w_2
u_2
=
- 1.89
+ 0.92
- 1.08
- 1.51
- 0.73
w_3
u_3
w_4
u_4
w_5
u_5

Approximation



Face recognition

Goal: Identify person

Training set: $\{x_1, y_1\}, \ldots, \{x_n, y_n\}$

Nearest-neighbor classification

$$i^* := \arg\min_{1 \le i \le n} ||x_{\text{test}} - x_i||_2$$

Cost: $\mathcal{O}(nd)$ to classify new point

Classification in reduced space

Compute sample mean and k first principal directions $u_1, \ldots, u_k$ from training data

For each test data point x_{test}

1. Center using training sample mean to obtain $\text{ct}\left(x_{\text{test}}\right)$
2. Compute k principal components

$$w_{\text{test}}[i] := u_i^T \, \text{ct}\left(x_{\text{test}}\right)$$

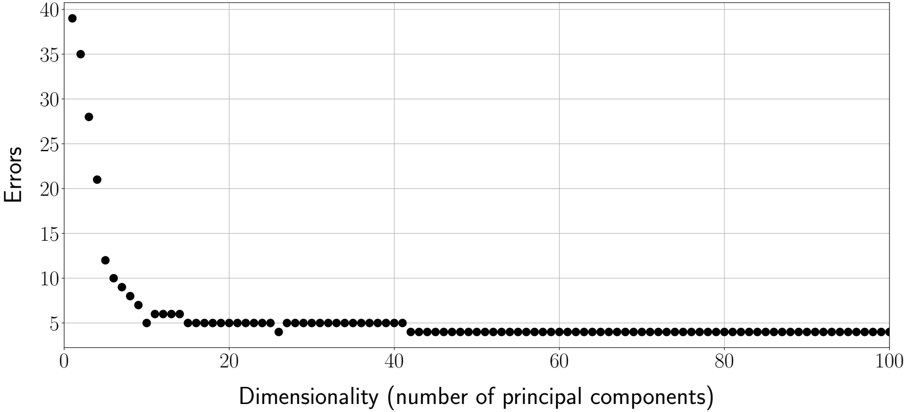
3. Compare to principal components of training data

$$i^*_{[k]} := \arg \min_{1 \leq i \leq n} \left|\left| w_{\text{test}} - w_{[1:k]}[i] \right|\right|_2$$

Cost reduced to $\mathcal{O}\left(nk\right)$

Computing eigendecomposition is costly, but is done only once

Performance



$k := 42$

x_{test}

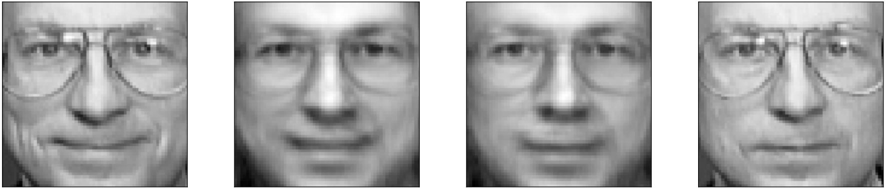
$\underset{u_1,\ldots,u_k}{\text{approx}}(x_{\text{test}})$

$\underset{u_1,\ldots,u_k}{\text{approx}}\left(x_{i^*_{[k]}}\right)$

$x_{i^*_{[k]}}$

$k := 42$

x_{test} $\quad$ $\underset{u_1,\ldots,u_k}{\operatorname{approx}}(x_{\text{test}})$ $\quad$ $\underset{u_1,\ldots,u_k}{\operatorname{approx}}\left(x_{i^*_{[k]}}\right)$ $\quad$ $x_{i^*_{[k]}}$



$k := 42$

x_{test} $\quad$ $\underset{u_1,\ldots,u_k}{\text{approx}}(x_{\text{test}})$ $\quad$ $\underset{u_1,\ldots,u_k}{\text{approx}}\left(x_{i^*_{[k]}}\right)$ $\quad$ $x_{i^*_{[k]}}$

$k := 42$

x_{test}

$\underset{u_{\mathbf{1}},\ldots,u_k}{\text{approx}}(x_{\text{test}})$

$\underset{u_{\mathbf{1}},\ldots,u_k}{\text{approx}}\left(x_{i^*_{[k]}}\right)$

$x_{i^*_{[k]}}$

Optimality

Let $u_1, \ldots, u_d$ be the eigenvectors of covariance matrix $\Sigma_{\tilde{x}}$

$$\{u_1, \ldots, u_k\} = \arg \min_{\substack{\{b_1,\ldots,b_k\} \\ ||b_i||_2=1, 1\leq i\leq k \\ b_i \perp b_j, i\neq j}} \mathrm{E}\left[\left|\left|\tilde{x} - \underset{b_1,\ldots,b_k}{\mathrm{approx}}(\tilde{x})\right|\right|_2^2\right]$$

The optimal linear k-dimensional approximation is

$$\underset{u_1,\ldots,u_k}{\mathrm{approx}}(\tilde{x}) := \sum_{i=1}^{k} \tilde{w}_i u_i \qquad \tilde{w}_i := u_i^T \tilde{x}$$

Proof of optimality

How do we prove this?

$$\{u_1, \ldots, u_k\} = \arg \min_{\substack{\{b_1,\ldots,b_k\} \\ ||b_i||_2=1, 1\leq i\leq k \\ b_i \perp b_j, i\neq j}} \mathrm{E}\left[\left|\left|\tilde{x} - \underset{b_1,\ldots,b_k}{\text{approx}}(\tilde{x})\right|\right|_2^2\right]$$

Sum of directional variances

$$\mathrm{E}\left[\left|\left|\tilde{x} - \underset{b_1,\ldots,b_k}{\text{approx}}(\tilde{x})\right|\right|_2^2\right] = \mathrm{E}\left[||\tilde{x}||_2^2\right] - \sum_{i=1}^{k} \mathrm{Var}\left[b_i^T \tilde{x}\right]$$

We prove that principal directions are optimal by induction on k

$k = 1$

By the spectral theorem

$$u_1 = \arg \max_{||b||_2=1} \mathrm{Var}[b^T \tilde{x}]$$

Induction step

We need to show that

$$\{u_1, \ldots, u_{k-1}\} = \arg \max_{\substack{\{b_1,\ldots,b_{k-1}\} \\ ||b_i||_2=1, 1\leq i\leq k-1 \\ b_i \perp b_j, i\neq j}} \sum_{i=1}^{k-1} \mathrm{Var}\left[b_i^T \tilde{x}\right]$$

implies

$$\{u_1, \ldots, u_k\} = \arg \max_{\substack{\{b_1,\ldots,b_k\} \\ ||b_i||_2=1, 1\leq i\leq k \\ b_i \perp b_j, i\neq j}} \sum_{i=1}^{k} \mathrm{Var}\left[b_i^T \tilde{x}\right]$$

Sum of variances

Fix arbitrary set of k orthonormal vectors b_1, ..., b_k

$$\begin{aligned}
\sum_{i=1}^{k} \mathrm{Var}\left[b_i^T \tilde{x}\right] &= \sum_{i=1}^{k} \mathrm{E}\left[\left(b_i^T \tilde{x}\right)^2\right] \\
&= \mathrm{E}\left[\sum_{i=1}^{k} \left(b_i^T \tilde{x}\right)^2\right] \\
&= \mathrm{E}\left[\left|\left|\sum_{i=1}^{k} b_i^T \tilde{x} b_i\right|\right|_2^2\right] \\
&= \mathrm{E}\left[||\mathcal{P}_{\mathcal{S}}\, \tilde{x}||_2^2\right] \qquad \mathcal{S} := \mathrm{span}(b_1, \ldots, b_k)
\end{aligned}$$

Key insight

For any orthonormal basis $a_1, \ldots, a_k$ of $\mathcal{S}$

$$\mathcal{P}_{\mathcal{S}}\,\tilde{x} := \sum_{i=1}^{k} b_i^T \tilde{x} b_i$$
$$= \sum_{i=1}^{k} a_i^T \tilde{x} a_i$$

We need to choose wisely!

Orthogonal vector

$\mathcal{S}$ has dimension k

There is at least one vector $a_\perp \in \mathcal{S}$ orthogonal to $u_1, \ldots, u_{k-1}$

Can we have

$$\mathrm{Var}[u_k^T \tilde{x}] < \mathrm{Var}[a_\perp^T \tilde{x}] \quad ?$$

No!

$$u_k = \arg \max_{||a||_2=1, a \perp u_1, \ldots, u_{k-1}} \mathrm{Var}[a^T \tilde{x}]$$

Wise choice

Set $a_k := a_\perp$

Choose $a_1, \ldots, a_{k-1}$ so $a_1, \ldots, a_k$ span $\mathcal{S}$

$$\mathrm{Var}[u_k^T \tilde{x}] \geq \mathrm{Var}[a_\perp^T \tilde{x}] = \mathrm{Var}[a_k^T \tilde{x}]$$

$$\sum_{i=1}^{k-1} \mathrm{Var}\left[u_i^T \tilde{x}\right] \geq \sum_{i=1}^{k-1} \mathrm{Var}\left[a_i^T \tilde{x}\right] \qquad \text{by induction hypothesis}$$

$$\sum_{i=1}^{k} \mathrm{Var}\left[u_i^T \tilde{x}\right] \geq \sum_{i=1}^{k} \mathrm{Var}\left[a_i^T \tilde{x}\right]$$

$$= \mathrm{E}\left[||\mathcal{P}_{\mathcal{S}}\, \tilde{x}||_2^2 \right] = \sum_{i=1}^{k} \mathrm{Var}\left[b_i^T \tilde{x}\right]$$

What have we learned?

Definition of dimensionality reduction

How to perform linear dimensionality reduction via PCA

Proof of optimality for mean ℓ_2-norm error