Applied Linear Algebra for Data Analysis

Application: Dimensionality reduction and PCA

Sivakumar Balasubramanian

Department of Bioengineering
Christian Medical College, Bagayam
Vellore 632002

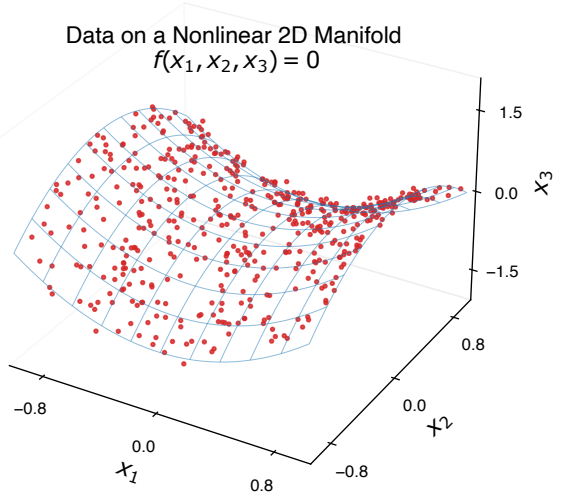
High dimensional data

- High dimensional data is encounted in in many applications, e.g. imaging, genomics, neural time series, wearable sensors, etc.
- Digital health care is a prime example of high dimensional with heterogeneous variables.
- This data can be organized in a rectangular form – a matrix – $\mathbf{D} \in \mathbb{R}^{m \times n}$ with m samples and n features (variables, dimensions, etc.) .
- Some examples:
  - A set of grayscale images $\{I_j\}_{j=1}^{m}$ of size $p \times q$ pixels. This set can be put in a matrix $\mathbf{D} \in \mathbb{R}^{m \times pq}$.
  - Neural firing rates of n neurons recorded for m time points.
  - Digital health data of m patients with n variables.
  - ...

Data often lies in a low dimensional manifold

Although, n can be large the data often lies in a low dimensional manifold with a lower dimension k.

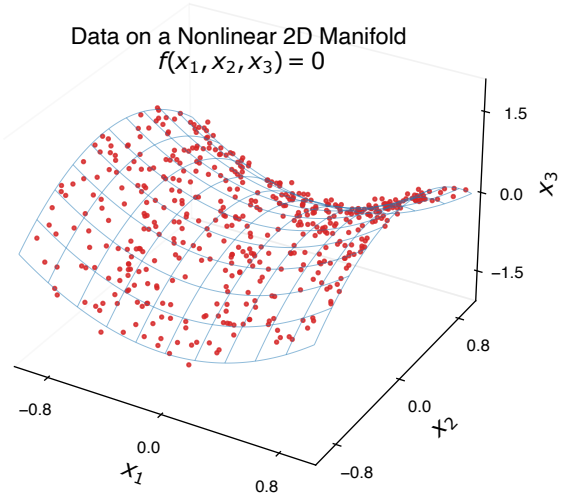
Representations of such data only requires k variables, and not n. E.g. *You only need longitude and latitude to represent a location on the earth's surface, not the full 3D coordinates.*



Data often lies in a low dimensional manifold

Identifying the low dimensional structure of the data is useful for multiple reasons:

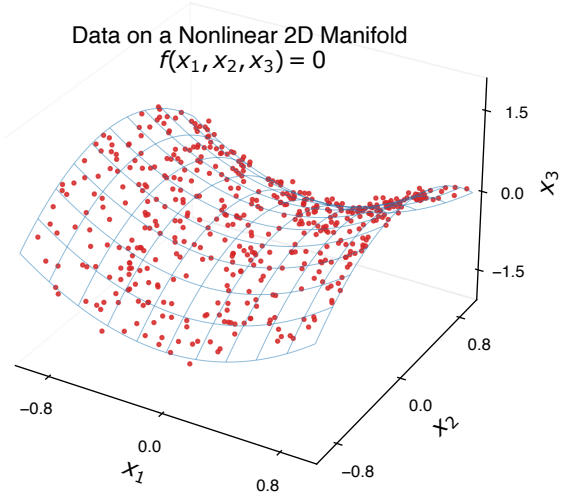
1. Data compression
2. Visualization
3. Noise reduction
4. Modelling the generative process



Dimensionality reduction

Dimensionality reduction methods help identify the low dimensional manifold in which the data lies.

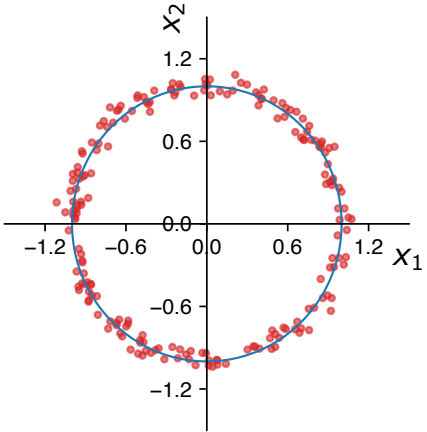
The oldest of these methods is the principal component analysis (PCA), which help identiy low dimensional subspace on which the data lies.



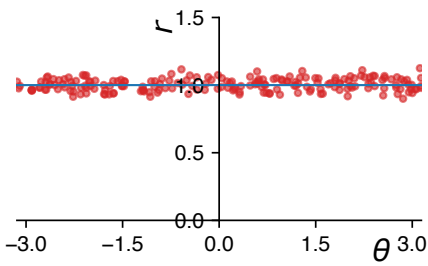
Principal Component Analysis (PCA)

Change of variables can often show the low dimensional structure of the data.

Original Data $\mathbf{D} \in \mathbb{R}^{N \times 2}$



Transformed Data $\hat{\mathbf{D}} \in \mathbb{R}^{N \times 2}$



Nonlinear transformation

$$\theta = \text{atan2}\,(x_2, x_1)$$

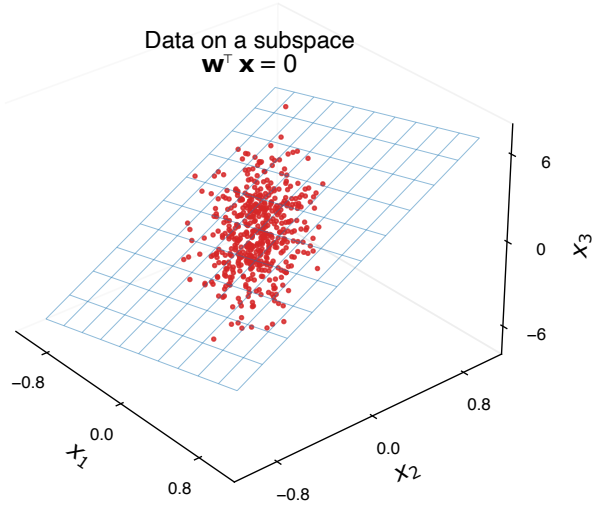
$$r = \sqrt{x_1^2 + x_2^2}$$

Principal Component Analysis (PCA)

If the data lies in an low dimensional subspace (hyperplane), then a linear transformation could be used to identify

PCA helps identify the low dimensional subspace on which the data lies.

The output of PCA is an orthonormal basis for $\mathbb{R}^n$, called the principal components. The principal components are arragned in the order of decreasing variance along their respective directions.



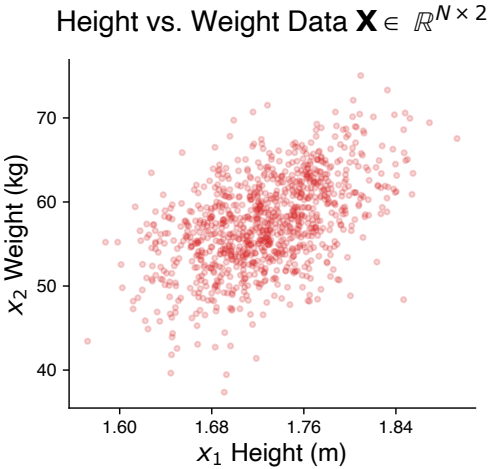
Principal Component Analysis (PCA)

Consider the height and weight data of $m = 1000$ individuals. The scatter plot of the data is shown on the right.

This data is stored in an array $\mathbf{X} \in \mathbb{R}^{m \times 2}$, where the rows are the measurements from each individual, and the first column is the height and the second column is the weight.

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}_1 & \mathbf{x}_2 \end{bmatrix} = \begin{bmatrix} \tilde{\mathbf{x}}_1^\top \\ \tilde{\mathbf{x}}_2^\top \\ \vdots \\ \tilde{\mathbf{x}}_m^\top \end{bmatrix} = \begin{bmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \\ \vdots & \vdots \\ x_{m1} & x_{m2} \end{bmatrix}$$

The spread of points appears to be large in particular direction.



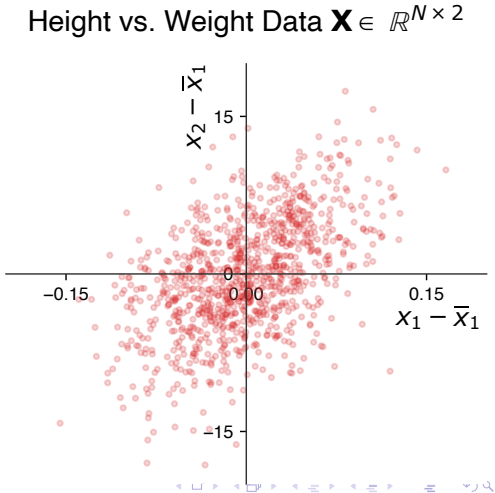
Principal Component Analysis (PCA)

We pose the following problem: what is the direction along which the spread of points from $\mathbf{X}$ is the largest?
We first **remove the mean** from the data points, i.e. move it to the origin.

$$\overline{\mathbf{X}} = \begin{bmatrix} \mathbf{x}_1 - \overline{x}_1\mathbf{1} & \mathbf{x}_2 - \overline{x}_2\mathbf{1} \end{bmatrix}, \ \overline{x}_i = \frac{1}{m}\sum_{j=1}^{m} x_{ji}$$

From this point forward, we assume $\mathbf{X}$ is centered, i.e. has zero mean.
We define the spead of points from $\mathbf{X}$ along a direction $\mathbf{w}_1 \in \mathbb{R}^2$ as the sum of squared norms of the orthogonal projections of the points onto the subspace spanned by $\mathbf{w}_1$.

$$V(\mathbf{w}_1) = \sum_{i=1}^{m} \|\mathbf{w}_1\mathbf{w}_1^\top \tilde{\mathbf{x}}_i\|_2^2$$



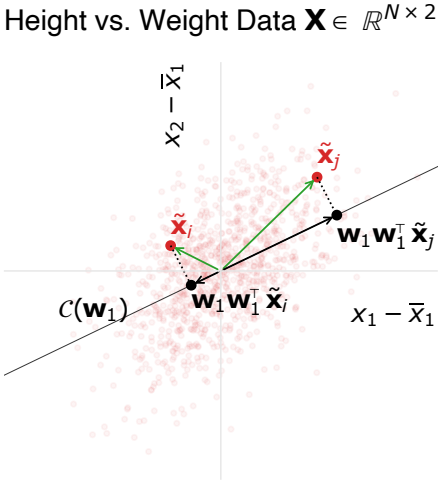
Principal Component Analysis (PCA)

The variance is now a function of $\mathbf{w}_1$, and the problem is to find the direction $\mathbf{w}_1$ that maximizes $V(\mathbf{w}_1)$, i.e.

$$\begin{aligned} \arg\max_{\mathbf{w}_1} \quad & \sum_{i=1}^{m} \|\mathbf{w}_1^\top \tilde{\mathbf{x}}_i\|_2^2 \\ \text{subject to} \quad & \|\mathbf{w}_1\|_2 = 1 \end{aligned}$$

Note that this is equivalent to the following minimization problem:

$$\begin{aligned} \arg\min_{\mathbf{w}_1} \quad & \sum_{i=1}^{m} \|\tilde{\mathbf{x}}_i - \mathbf{w}_1 \mathbf{w}_1^\top \tilde{\mathbf{x}}_i\|_2^2 \\ \text{subject to} \quad & \|\mathbf{w}_1\|_2 = 1 \end{aligned}$$

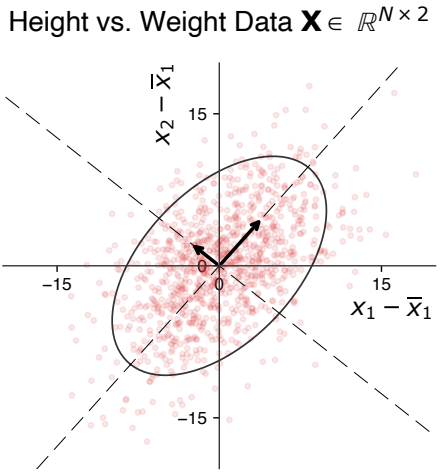


Principal Component Analysis (PCA)

Let the solution to the previous optimization problem be $\mathbf{p}_1$. This direction is referred to as the *first principal component*.

Once we've identied $\mathbf{p}_1$, we can then search for the next direction $\mathbf{w}_2$ that is orthogonal to $\mathbf{w}_1$ and maximizes the variance of the data its direction.

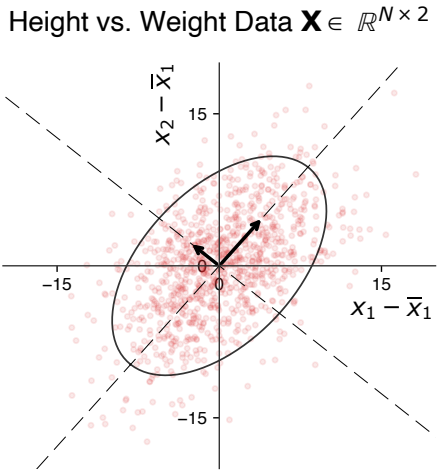
In $\mathbb{R}^2$, once we find $\mathbf{p}_1$, we know $\mathbf{p}_2$ (the *second principal component*) as well. Why?



Principal Component Analysis (PCA)

The two vectors $\mathbf{p}_1$ and $\mathbf{p}_2$ form an orthonormal basis for $\mathbb{R}^2$, and are called the *principal components* of the data $\mathbf{X}$.

Every point can be represented as a linear combination of the principal components.



Principal Component Analysis (PCA)

- We can now extend this to $\mathbb{R}^n$. $\mathbf{X} \in \mathbb{R}^{m \times n}$ and $\tilde{\mathbf{x}}_i \in \mathbb{R}^n$.
  We can search for the principal components $\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_n \in \mathbf{R}^n$, one after the other, by maximizing the variance of the data along a single direction, which is orthogonal to the previously identified directions.

- The j^{th} principal component is the solution to the following optimization problem:

$$\begin{aligned} \arg\max_{\mathbf{w}_j} \quad & \sum_{i=1}^{m} \|\mathbf{w}_j^\top \tilde{\mathbf{x}}_i\|_2^2 \\ \text{subject to} \quad & \|\mathbf{w}_j\|_2 = 1 \\ & \mathbf{w}_j^\top \mathbf{w}_k = 0, \ 1 \leq k < j \end{aligned}$$

- This is the iterative method for obtaining the principal components.

Principal Component Analysis (PCA)

- We can also pose this as a combined problem as the following, where, $\mathbf{W} = \begin{bmatrix} \mathbf{w}_1 & \mathbf{w}_1 & \cdots & \mathbf{w}_n \end{bmatrix}$

$$\begin{aligned} \arg\max_{\mathbf{W}} \quad & \sum_{j=1}^{n} \sum_{i=1}^{m} \|\mathbf{w}_j^\top \tilde{\mathbf{x}}_i\|_2^2 \\ \text{subject to} \quad & \mathbf{W}^\top \mathbf{W} = \mathbf{I} \\ & \mathbf{W}^\top \mathbf{X}^\top \mathbf{X} \mathbf{W} \text{ is diagonal.} \end{aligned}$$

This optimization problem will result in the same solution as the iterative method discussed earlier..

Principal Component Analysis (PCA)

- The principal components are the eigenvectors of the covariance matrix $\mathbf{X}^\top\mathbf{X}$.
- The principal components form an orthonormal basis in which the data points are decorrelated.
- The total variance in the data is the sum of the variances along each of the principal components.

$$\mathbf{X}^\top\mathbf{X} = \mathbf{P}\mathbf{D}\mathbf{P}^\top \implies trace\left(\mathbf{X}^\top\mathbf{X}\right) = trace\left(\mathbf{D}\right) = \sum_{i=1}^{n} d_{ii}$$

$d_{ii} \geq 0$, why?
d_{ii} is the variance of the data along the i^{th} principal component.
We will arrange the eigenvectors in $\mathbf{P}$ such that, $d_{11} \geq d_{22} \geq \cdots \geq d_{nn} \geq 0$.

Principal Component Analysis (PCA)

- The PCA allows us to uncover a linear structure in the data.

- Let's assume that the data that we measure or observe is generated by the following process,

$$\tilde{\mathbf{x}} = \mathbf{A}\tilde{\mathbf{z}} + \overline{\mathbf{x}}^{\top}, \quad \tilde{\mathbf{x}}, \tilde{\mathbf{z}}, \overline{\mathbf{x}} \in \mathbb{R}^n$$

  where,
  - $\tilde{\mathbf{z}}$ is a data point from the latent space, such that $\frac{1}{m}\mathbf{Z}^{\top}\mathbf{Z} = \mathbf{I}$, $\mathbf{Z} \in \mathbb{R}^{m \times n}$.
  - $\mathbf{A} \in \mathbb{R}^{n \times n}$ is matrix that transforms the latent space to the observation space.
  - $\overline{\mathbf{x}} \in \mathbb{R}^n$ is the mean of the data.

Where is SVD in all this?

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}_1 & \mathbf{x}_2 & \cdots & \mathbf{x}_n \end{bmatrix} = \begin{bmatrix} \tilde{\mathbf{x}}_1^\top \\ \tilde{\mathbf{x}}_2^\top \\ \vdots \\ \tilde{\mathbf{x}}_m^\top \end{bmatrix} \longrightarrow \mathbf{X}^\top \mathbf{X} = \mathbf{P}\mathbf{D}\mathbf{P}^\top$$

We could have obtained $\mathbf{P}$ and $\mathbf{D}$ using the SVD of $\mathbf{X}$!

$$\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top \implies \mathbf{X}^\top \mathbf{X} = \mathbf{V}\mathbf{\Sigma}^\top \mathbf{U}^\top \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top = \mathbf{V}\mathbf{\Sigma}^2\mathbf{V}^\top$$

$$\mathbf{P} = \mathbf{V} \quad \text{and} \quad \mathbf{D} = \mathbf{\Sigma}$$

Dimensionsality Reduction with PCA

- PCA allows us to reduce the dimensionality of the data by projecting the data onto a lower dimensional subspace.

$$\mathbf{X}^\top \mathbf{X} = \mathbf{P}\mathbf{D}\mathbf{P}^\top = \sum_{i=1}^{n} d_{ii}\mathbf{p}_i\mathbf{p}_i^\top$$

Let's assume that none of the eigenvalues are zero, i.e. $d_{ii} > 0,\ \forall i$.

- We can obtain the latent space representation of the data through the following transformation,

$$\mathbf{Z}^\top = \mathbf{D}^{-\frac{1}{2}}\mathbf{P}^\top\mathbf{X}^\top \implies \mathbf{X}^\top = \mathbf{P}\mathbf{D}^{\frac{1}{2}}\mathbf{Z}^\top$$

- If some of the d_{ii}s are small, the we can get an approximation of the observed data $\mathbf{X}$ by using only the first $k < n$ principal components,

$$\hat{\mathbf{X}}^\top = \mathbf{P}_{n\times k}\mathbf{D}^{\frac{1}{2}}_{k\times k}\mathbf{Z}^\top_{k\times m}$$

For each data point $\tilde{\mathbf{x}}_i$ we only keep the first k elements of $\tilde{\mathbf{z}}_i$.

Dimensionsality Reduction with PCA

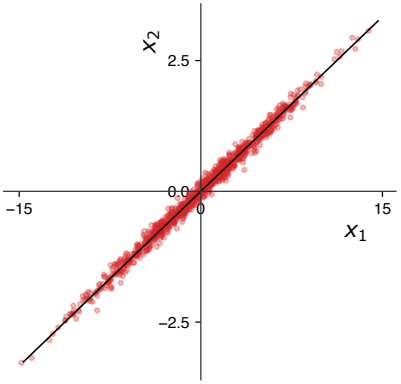
We get a good approximation of $\mathbf{X}$ ($2N$ numbers), by retaining only the projections onto the first principal component, which requires only N numbers.

$$\hat{\mathbf{X}} = \mathbf{p}_1 \mathbf{p}_1^\top \mathbf{X}$$

How much information did we loose with $\hat{\mathbf{X}}$? Or how much did we incur?

Variance lost = Approx. Error = d_{22}

Data lies on a low dimensional subspace

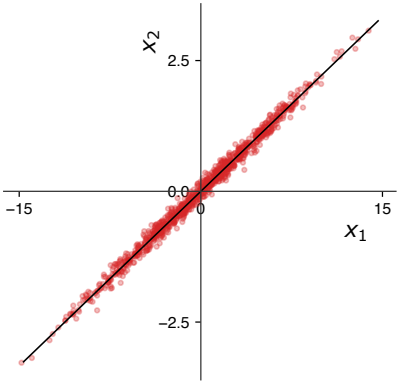


Dimensionsality Reduction with PCA

In general, if we approximate with the first k principal components,

$$\text{Variance lost} = \text{Approx. Error} = \sum_{i=k+1}^{n} d_{ii}$$

Data lies on a low dimensional subspace



Generative view of PCA

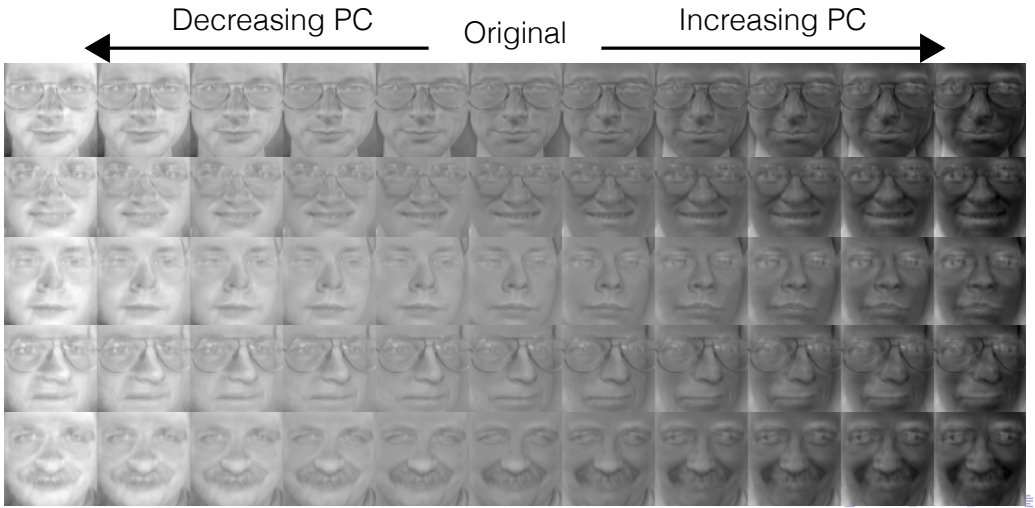
Generating new data points

Original Faces

Generated Faces

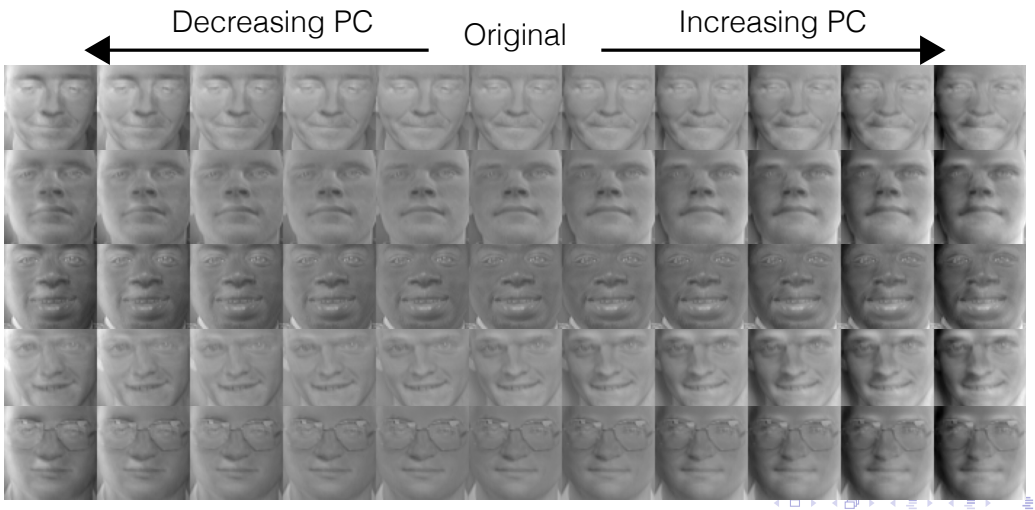
Meaning of the PCs

Changing PC 1



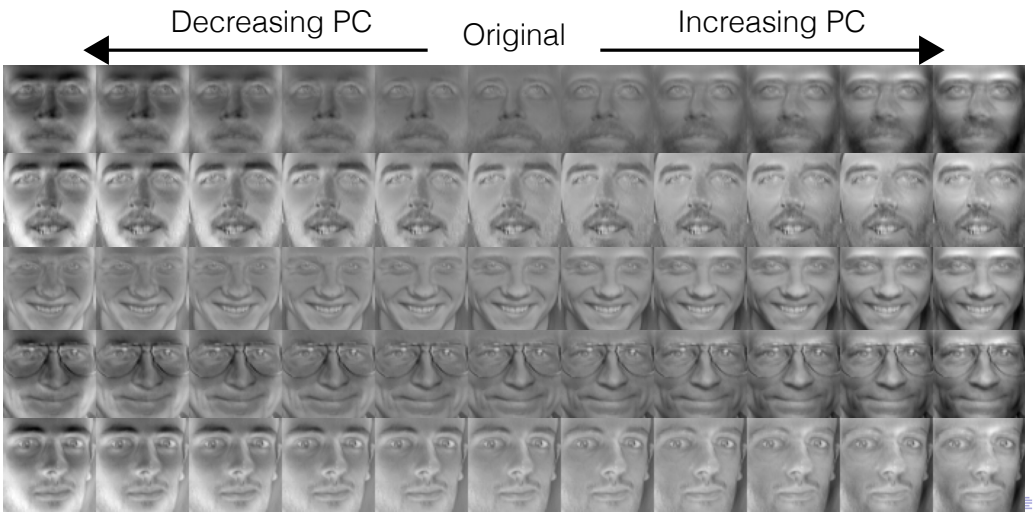
Meaning of the PCs

Changing PC 2



Meaning of the PCs

Changing PC 3



Meaning of the PCs

Changing PC 4

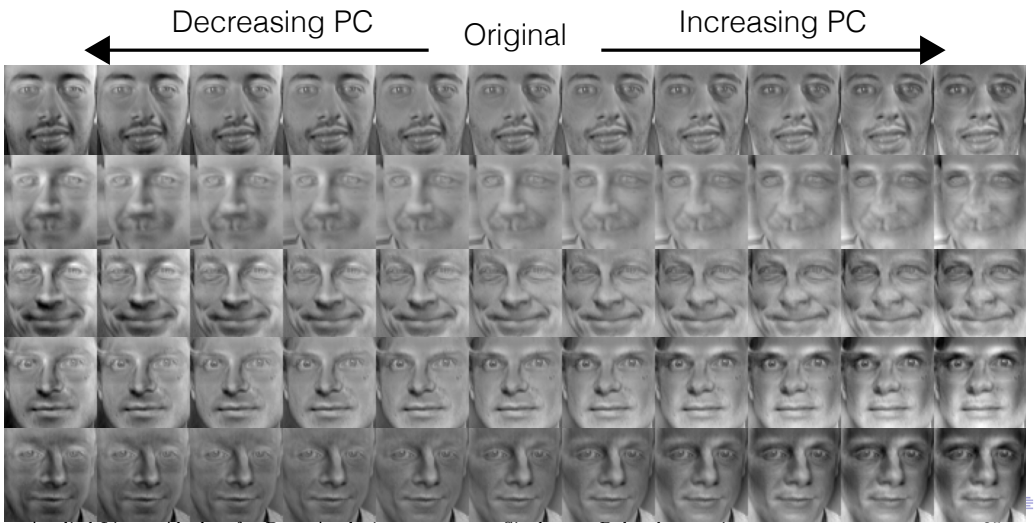


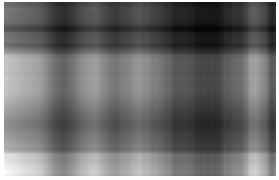
Image Compression with SVD

Image Size: 1208×1880 numbers

Image Compression with SVD

Rank k Reconstruction $(n + m + 1) \times k$ numbers

Rank-1 Reconst.



Rank-5 Reconst.

Rank-10 Reconst.

Rank-20 Reconst.

Rank-40 Reconst.

Rank-80 Reconst.

Rank-100 Reconst.

Rank-150 Reconst.