Machine Learning

Lecture. 12.

Mark Girolami

girolami@dcs.gla.ac.uk

Department of Computing Science
University of Glasgow

PCA

UNIVERSITY of GLASGOW

- Principal Component Analysis

PCA

UNIVERSITY
of
GLASGOW

- Principal Component Analysis
- Feature Extraction

PCA

- Principal Component Analysis
- Feature Extraction
- Dimensionality Reduction

PCA

- Principal Component Analysis
- Feature Extraction
- Dimensionality Reduction
- Data Compression

PCA

- Principal Component Analysis
- Feature Extraction
- Dimensionality Reduction
- Data Compression
- Data Visualisation

Image Representation

Image Representation

UNIVERSITY *of* GLASGOW

- 64×64 grey-scale images with $64^2 = 4096$ pixels

Image Representation

UNIVERSITY *of* GLASGOW

- 64×64 grey-scale images with $64^2 = 4096$ pixels
- Pixel values in the range $0 - 2^8(256)$

Image Representation

UNIVERSITY *of* GLASGOW

- 64×64 grey-scale images with $64^2 = 4096$ pixels
- Pixel values in the range $0 - 2^8 (256)$
- Each image represented as a $M = 4096 \times 1$ dimensional vector $\mathbf{x}$

Image Representation

UNIVERSITY of GLASGOW

- 64×64 grey-scale images with $64^2 = 4096$ pixels
- Pixel values in the range $0 - 2^8(256)$
- Each image represented as a $M = 4096 \times 1$ dimensional vector $\mathbf{x}$
- Total number of possible images produced from representation is 256^{4096}

Image Representation

UNIVERSITY of GLASGOW

- 64×64 grey-scale images with $64^2 = 4096$ pixels
- Pixel values in the range $0 - 2^8(256)$
- Each image represented as a $M = 4096 \times 1$ dimensional vector $\mathbf{x}$
- Total number of possible images produced from representation is 256^{4096}
- Staggeringly large number i.e. $256^{4096} = 2^{8\times4096}$. Number of atoms in the entire universe 2^{784}

Image Representation

UNIVERSITY of GLASGOW

- 64×64 grey-scale images with $64^2 = 4096$ pixels
- Pixel values in the range $0 - 2^8(256)$
- Each image represented as a $M = 4096 \times 1$ dimensional vector $\mathbf{x}$
- Total number of possible images produced from representation is 256^{4096}
- Staggeringly large number i.e. $256^{4096} = 2^{8\times 4096}$. Number of atoms in the entire universe 2^{784}
- Pixel representation very powerful - overfitting almost certain

Classification

UNIVERSITY
of
GLASGOW

- Discriminate between faces with and without spectacles

Classification

- Discriminate between faces with and without spectacles
- Devise classifier based on image pixel values

Classification

UNIVERSITY *of* GLASGOW

- Discriminate between faces with and without spectacles
- Devise classifier based on image pixel values
-

$$\log \frac{P(C=1|\mathbf{x})}{P(C=0|\mathbf{x})} = \mathbf{w}^{\mathsf{T}}\mathbf{x}$$

  $C = 1 \equiv$ Spectacles & $C = 0 \equiv$ No Spectacles

Classification

- Discriminate between faces with and without spectacles
- Devise classifier based on image pixel values
- $$\log \frac{P(C=1|\mathbf{x})}{P(C=0|\mathbf{x})} = \mathbf{w}^{\mathsf{T}}\mathbf{x}$$

  $C = 1 \equiv$ Spectacles & $C = 0 \equiv$ No Spectacles
- As $\mathbf{w}$ has dimension $M = 4096 \times 1$ with only 400 available samples overfitting is highly likely

Classification

- Discriminate between faces with and without spectacles
- Devise classifier based on image pixel values
- $$\log \frac{P(C=1|\mathbf{x})}{P(C=0|\mathbf{x})} = \mathbf{w}^{\mathsf{T}}\mathbf{x}$$

  $C = 1 \equiv$ Spectacles & $C = 0 \equiv$ No Spectacles
- As $\mathbf{w}$ has dimension $M = 4096 \times 1$ with only 400 available samples overfitting is highly likely
- Alleviate problem by extracting a smaller number of informative features

Image Representation

UNIVERSITY
of
GLASGOW

- Consider the variability observed in the images.

Image Representation

- Consider the variability observed in the images.
- Differences in pose (head on, facing diagonally, looking up, down)

Image Representation

UNIVERSITY
of
GLASGOW

- Consider the variability observed in the images.
- Differences in pose (head on, facing diagonally, looking up, down)
- Facial expression (grinning, smiling, scowling, open mouthed etc)

Image Representation

UNIVERSITY
of
GLASGOW

- Consider the variability observed in the images.
- Differences in pose (head on, facing diagonally, looking up, down)
- Facial expression (grinning, smiling, scowling, open mouthed etc)
- Wearing of glasses, presence of beard, male, female

Image Representation

UNIVERSITY
of
GLASGOW

- Consider the variability observed in the images.
- Differences in pose (head on, facing diagonally, looking up, down)
- Facial expression (grinning, smiling, scowling, open mouthed etc)
- Wearing of glasses, presence of beard, male, female
- Shape of face, lighting, ...

Image Representation

UNIVERSITY *of* GLASGOW

- Consider the variability observed in the images.
- Differences in pose (head on, facing diagonally, looking up, down)
- Facial expression (grinning, smiling, scowling, open mouthed etc)
- Wearing of glasses, presence of beard, male, female
- Shape of face, lighting, ...
- Variability in images due to a small number (smaller than 256^{4096}) of degrees of freedom

Linear Subspace

- In other words the data lies in a lower-dimensional feature space which accounts for all of the information or variability in the images

Linear Subspace

UNIVERSITY
of
GLASGOW

- In other words the data lies in a lower-dimensional feature space which accounts for all of the information or variability in the images
- By extracting these features from original representation it may be possible to overcome potential generalisation problems

Linear Subspace

UNIVERSITY
of
GLASGOW

- In other words the data lies in a lower-dimensional feature space which accounts for all of the information or variability in the images
- By extracting these features from original representation it may be possible to overcome potential generalisation problems
- Images of faces may be described by a subspace of the 4096 dimensional pixel space

Linear Subspace

UNIVERSITY
of
GLASGOW

- Assume M dimensional data actually lies within a P dimensional subspace where $P << M$

Linear Subspace

UNIVERSITY of GLASGOW

- Assume M dimensional data actually lies within a P dimensional subspace where $P << M$
- Further assume that the subspace is linear

Linear Subspace

UNIVERSITY of GLASGOW

- Assume M dimensional data actually lies within a P dimensional subspace where $P << M$
- Further assume that the subspace is linear
- Orthonormal basis vectors (coordinates) span subspace i.e. $\{\boldsymbol{\beta}_1 \cdots \boldsymbol{\beta}_P\}$ where each $\boldsymbol{\beta}_p \in \mathbb{R}^D$

Linear Subspace

UNIVERSITY *of* GLASGOW

- Assume M dimensional data actually lies within a P dimensional subspace where $P << M$
- Further assume that the subspace is linear
- Orthonormal basis vectors (coordinates) span subspace i.e. $\{\boldsymbol{\beta}_1 \cdots \boldsymbol{\beta}_P\}$ where each $\boldsymbol{\beta}_p \in \mathbb{R}^D$
- Data point $\mathbf{x}$ approximated by linear combination of basis vectors

$$\mathbf{x}_n \approx \sum_{p=1}^{P} u_{np}\boldsymbol{\beta}_p = \mathbf{B}\mathbf{u}_n$$

where $D \times P$ dimensional matrix $\mathbf{B} = [\boldsymbol{\beta}_1 \cdots \boldsymbol{\beta}_P]$ and $\mathbf{u}_n$ is a $P \times 1$ dimensional vector.

PCA Derivation

UNIVERSITY *of* GLASGOW

- Consider limiting case of $P = 1$ i.e. data $\mathbf{X}$ is modeled as residing around a 1-dimensional linear subspace $\boldsymbol{\beta}_1$

PCA Derivation

- Consider limiting case of $P = 1$ i.e. data $\mathbf{X}$ is modeled as residing around a 1-dimensional linear subspace $\boldsymbol{\beta}_1$
- Squared reconstruction error incurred by approximation $\mathbf{x}_n = u_{1n}\boldsymbol{\beta}_1$ defined as

$$\mathcal{E} = \frac{1}{N}\sum_{n=1}^{N}(\mathbf{x}_n - u_{1n}\boldsymbol{\beta}_1)^2$$

PCA Derivation

UNIVERSITY of GLASGOW

- Consider limiting case of $P = 1$ i.e. data $\mathbf{X}$ is modeled as residing around a 1-dimensional linear subspace $\boldsymbol{\beta}_1$
- Squared reconstruction error incurred by approximation $\mathbf{x}_n = u_{1n}\boldsymbol{\beta}_1$ defined as

$$\mathcal{E} = \frac{1}{N}\sum_{n=1}^{N}(\mathbf{x}_n - u_{1n}\boldsymbol{\beta}_1)^2$$

- Taking derivatives with respect to each u_{1n} and setting to zero gives

$$\frac{\partial \mathcal{E}}{\partial u_{1n}} = -\frac{2}{N}(\boldsymbol{\beta}_1^{\mathsf{T}}\mathbf{x}_n - u_{1n}) = 0 \Rightarrow u_{1n} = \boldsymbol{\beta}_1^{\mathsf{T}}\mathbf{x}_n$$

PCA Derivation

- Plugging this value back into the expression for $\mathcal{E}$ yields

$$\begin{aligned}
\mathcal{E} &= \frac{1}{N}\sum_{n=1}^{N}(\mathbf{x}_n - u_{1n}\boldsymbol{\beta}_1)^2 \\
&= \frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_n^{\mathsf{T}}\mathbf{x}_n - 2u_{1n}\boldsymbol{\beta}_1^{\mathsf{T}}\mathbf{x}_n + u_{1n}^2\boldsymbol{\beta}_1^{\mathsf{T}}\boldsymbol{\beta}_1 \\
&= \frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_n^{\mathsf{T}}\mathbf{x}_n - 2u_{1n}^2 + u_{1n}^2\boldsymbol{\beta}_1^{\mathsf{T}}\boldsymbol{\beta}_1 \\
&= \frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_n^{\mathsf{T}}\mathbf{x}_n - u_{1n}^2 = \frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_n^{\mathsf{T}}\mathbf{x}_n - \boldsymbol{\beta}_1^{\mathsf{T}}\mathbf{x}_n\mathbf{x}_n^{\mathsf{T}}\boldsymbol{\beta}_1
\end{aligned}$$

PCA Derivation

UNIVERSITY *of* GLASGOW

- So to minimise our reconstruction error we require to maximise

$$\frac{1}{N}\sum_{n=1}^{N}\boldsymbol{\beta}_1^{\mathsf{T}}\mathbf{x}_n\mathbf{x}_n^{\mathsf{T}}\boldsymbol{\beta}_1 = \frac{1}{N}\boldsymbol{\beta}_1^{\mathsf{T}}\mathbf{X}^{\mathsf{T}}\mathbf{X}\boldsymbol{\beta}_1 = \boldsymbol{\beta}_1^{\mathsf{T}}\widehat{\mathbf{C}}\boldsymbol{\beta}_1$$

PCA Derivation

UNIVERSITY of GLASGOW

- So to minimise our reconstruction error we require to maximise

$$\frac{1}{N}\sum_{n=1}^{N}\boldsymbol{\beta}_1^{\mathsf{T}}\mathbf{x}_n\mathbf{x}_n^{\mathsf{T}}\boldsymbol{\beta}_1 = \frac{1}{N}\boldsymbol{\beta}_1^{\mathsf{T}}\mathbf{X}^{\mathsf{T}}\mathbf{X}\boldsymbol{\beta}_1 = \boldsymbol{\beta}_1^{\mathsf{T}}\widehat{\mathbf{C}}\boldsymbol{\beta}_1$$

- Subject to $\boldsymbol{\beta}_1^{\mathsf{T}}\boldsymbol{\beta}_1 = 1$ where the sample covariance matrix is denoted as $\widehat{\mathbf{C}}$ (remember that each $\mathbf{X}$ is zero mean).

Variance Maximisation

UNIVERSITY *of* GLASGOW

- Note that minimisation of reconstruction error by maximisation of

$$\frac{1}{N}\sum_{n=1}^{N}\boldsymbol{\beta}_1^{\mathsf{T}}\mathbf{x}_n\mathbf{x}_n^{\mathsf{T}}\boldsymbol{\beta}_1 = \frac{1}{N}\sum_{n=1}^{N}u_{1n}^2$$

provides projections which have maximum variance so are maximally informative.

Variance Maximisation

UNIVERSITY
of
GLASGOW

- Note that minimisation of reconstruction error by maximisation of

$$\frac{1}{N}\sum_{n=1}^{N}\boldsymbol{\beta}_1^{\mathsf{T}}\mathbf{x}_n\mathbf{x}_n^{\mathsf{T}}\boldsymbol{\beta}_1 = \frac{1}{N}\sum_{n=1}^{N}u_{1n}^2$$

  provides projections which have maximum variance so are maximally informative.

- Minimisation of reconstruction error requires to find projection which maximises variance of projection - retain as much information as possible.

Variance Maximisation

UNIVERSITY
of
GLASGOW

- Remember that we are restricting each basis-vector to have unit norm in which case we require to create the Langrangian (Refer to the Week 5 notes)

$$\boldsymbol{\beta}_1^{\mathsf{T}} \widehat{\mathbf{C}} \boldsymbol{\beta}_1 - \lambda_1 \boldsymbol{\beta}_1^{\mathsf{T}} \boldsymbol{\beta}_1$$

and maximise with respect to $\boldsymbol{\beta}_1$.

Variance Maximisation

UNIVERSITY *of* GLASGOW

- Remember that we are restricting each basis-vector to have unit norm in which case we require to create the Langrangian (Refer to the Week 5 notes)

$$\boldsymbol{\beta}_1^{\mathsf{T}}\widehat{\mathbf{C}}\boldsymbol{\beta}_1 - \lambda_1\boldsymbol{\beta}_1^{\mathsf{T}}\boldsymbol{\beta}_1$$

  and maximise with respect to $\boldsymbol{\beta}_1$.

- The corresponding vector of partial derivatives gives

$$\frac{\partial}{\partial\boldsymbol{\beta}_1} = \widehat{\mathbf{C}}\boldsymbol{\beta}_1 - \lambda_1\boldsymbol{\beta}_1$$

Variance Maximisation

UNIVERSITY *of* GLASGOW

- Remember that we are restricting each basis-vector to have unit norm in which case we require to create the Langrangian (Refer to the Week 5 notes)

$$\boldsymbol{\beta}_1^{\mathsf{T}} \widehat{\mathbf{C}} \boldsymbol{\beta}_1 - \lambda_1 \boldsymbol{\beta}_1^{\mathsf{T}} \boldsymbol{\beta}_1$$

  and maximise with respect to $\boldsymbol{\beta}_1$.

- The corresponding vector of partial derivatives gives

$$\frac{\partial}{\partial \boldsymbol{\beta}_1} = \widehat{\mathbf{C}} \boldsymbol{\beta}_1 - \lambda_1 \boldsymbol{\beta}_1$$

- Setting to zero obtain an eigenvalue problem

$$\widehat{\mathbf{C}} \boldsymbol{\beta}_1 = \lambda_1 \boldsymbol{\beta}_1$$

Variance Maximisation

UNIVERSITY of GLASGOW

- As the variance of the projection is defined $\boldsymbol{\beta}_1^{\mathsf{T}} \widehat{\mathbf{C}} \boldsymbol{\beta}_1$ then for $\boldsymbol{\beta}_1^{\mathsf{T}} \boldsymbol{\beta}_1 = 1$ it should be clear that the variance of the projection is equal to λ_1 the associated eigenvalue.

Variance Maximisation

UNIVERSITY *of* GLASGOW

- As the variance of the projection is defined $\boldsymbol{\beta}_1^\mathsf{T}\widehat{\mathbf{C}}\boldsymbol{\beta}_1$ then for $\boldsymbol{\beta}_1^\mathsf{T}\boldsymbol{\beta}_1 = 1$ it should be clear that the variance of the projection is equal to λ_1 the associated eigenvalue.
- We have now found the direction $\boldsymbol{\beta}_1$ which maximises the variance of the projection $\boldsymbol{\beta}_1^\mathsf{T}\mathbf{x}$ and correspondingly minimises the reconstruction error

$$\mathcal{E} = \frac{1}{N}\sum_{n=1}^{N}(\mathbf{x}_n - u_{1n}\boldsymbol{\beta}_1)^2$$

This is referred to as the First Principal Direction and the projections of the data in this direction are the Principal Components in this direction.

Finding Additional Directions

UNIVERSITY of GLASGOW

- Now we want to find another direction vector $\boldsymbol{\beta}_2$ which will satisfy $\boldsymbol{\beta}_1^{\mathsf{T}}\boldsymbol{\beta}_2 = 0$ and $\boldsymbol{\beta}_2^{\mathsf{T}}\boldsymbol{\beta}_2 = 1$

Finding Additional Directions

UNIVERSITY *of* GLASGOW

- Now we want to find another direction vector $\boldsymbol{\beta}_2$ which will satisfy $\boldsymbol{\beta}_1^{\mathsf{T}}\boldsymbol{\beta}_2 = 0$ and $\boldsymbol{\beta}_2^{\mathsf{T}}\boldsymbol{\beta}_2 = 1$
- The approximations of points in data space will now take the form of

$$\mathbf{x}_n \approx \sum_{p=1}^{P=2} u_{np}\boldsymbol{\beta}_p$$

Finding Additional Directions

UNIVERSITY *of* GLASGOW

- Now we want to find another direction vector $\boldsymbol{\beta}_2$ which will satisfy $\boldsymbol{\beta}_1^\mathsf{T}\boldsymbol{\beta}_2 = 0$ and $\boldsymbol{\beta}_2^\mathsf{T}\boldsymbol{\beta}_2 = 1$
- The approximations of points in data space will now take the form of

$$\mathbf{x}_n \approx \sum_{p=1}^{P=2} u_{np}\boldsymbol{\beta}_p$$

- Reconstruction error is

$$\mathcal{E} = \frac{1}{N}\sum_{n=1}^{N}(\mathbf{x}_n - u_{1n}\boldsymbol{\beta}_1 - u_{2n}\boldsymbol{\beta}_2)^2$$

it is straightforward to see that $u_{2n} = \boldsymbol{\beta}_2^\mathsf{T}\mathbf{x}_n$

Finding Additional Directions

UNIVERSITY *of* GLASGOW

- The reconstruction error can obtained as the following where the orthonormal characteristics of both directions has been exploited

$$\frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_n^{\mathsf{T}}\mathbf{x}_n - \boldsymbol{\beta}_1^{\mathsf{T}}\mathbf{x}_n\mathbf{x}_n^{\mathsf{T}}\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2^{\mathsf{T}}\mathbf{x}_n\mathbf{x}_n^{\mathsf{T}}\boldsymbol{\beta}_2$$

Finding Additional Directions

UNIVERSITY *of* GLASGOW

- The reconstruction error can obtained as the following where the orthonormal characteristics of both directions has been exploited

$$\frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_n^{\mathsf{T}}\mathbf{x}_n - \boldsymbol{\beta}_1^{\mathsf{T}}\mathbf{x}_n\mathbf{x}_n^{\mathsf{T}}\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2^{\mathsf{T}}\mathbf{x}_n\mathbf{x}_n^{\mathsf{T}}\boldsymbol{\beta}_2$$

- It is clear that given $\boldsymbol{\beta}_1$ then we require to obtain a solution for

$$\widehat{\mathbf{C}}\boldsymbol{\beta}_2 = \lambda_2\boldsymbol{\beta}_2$$

subject to the orthonormal constraints imposed.

Projection and Deflation

UNIVERSITY of GLASGOW

- If $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ are orthonormal then

$$\begin{aligned}\mathbf{x} &= u_1\boldsymbol{\beta}_1 + u_2\boldsymbol{\beta}_2 \\ &= (\mathbf{x}^{\mathsf{T}}\boldsymbol{\beta}_1)\boldsymbol{\beta}_1 + (\mathbf{x}^{\mathsf{T}}\boldsymbol{\beta}_2)\boldsymbol{\beta}_2\end{aligned}$$

Projection and Deflation

- If $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ are orthonormal then

$$\begin{aligned} \mathbf{x} &= u_1\boldsymbol{\beta}_1 + u_2\boldsymbol{\beta}_2 \\ &= (\mathbf{x}^{\mathsf{T}}\boldsymbol{\beta}_1)\boldsymbol{\beta}_1 + (\mathbf{x}^{\mathsf{T}}\boldsymbol{\beta}_2)\boldsymbol{\beta}_2 \end{aligned}$$

- Where $(\mathbf{x}^{\mathsf{T}}\boldsymbol{\beta}_2)\boldsymbol{\beta}_2$ is the projection orthogonal to $(\mathbf{x}^{\mathsf{T}}\boldsymbol{\beta}_1)\boldsymbol{\beta}_1$ so projection orthogonal to first principal direction is

$$(\mathbf{x}^{\mathsf{T}}\boldsymbol{\beta}_2)\boldsymbol{\beta}_2 = (\mathbf{I} - \boldsymbol{\beta}_1\boldsymbol{\beta}_1^{\mathsf{T}})\mathbf{x}$$

Projection and Deflation

UNIVERSITY of GLASGOW

- If $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ are orthonormal then

$$\begin{aligned} \mathbf{x} &= u_1\boldsymbol{\beta}_1 + u_2\boldsymbol{\beta}_2 \\ &= (\mathbf{x}^\mathsf{T}\boldsymbol{\beta}_1)\boldsymbol{\beta}_1 + (\mathbf{x}^\mathsf{T}\boldsymbol{\beta}_2)\boldsymbol{\beta}_2 \end{aligned}$$

- Where $(\mathbf{x}^\mathsf{T}\boldsymbol{\beta}_2)\boldsymbol{\beta}_2$ is the projection orthogonal to $(\mathbf{x}^\mathsf{T}\boldsymbol{\beta}_1)\boldsymbol{\beta}_1$ so projection orthogonal to first principal direction is

$$(\mathbf{x}^\mathsf{T}\boldsymbol{\beta}_2)\boldsymbol{\beta}_2 = (\mathbf{I} - \boldsymbol{\beta}_1\boldsymbol{\beta}_1^\mathsf{T})\mathbf{x}$$

- Applying this to all of the data gives

$$\mathbf{X}(\mathbf{I} - \boldsymbol{\beta}_1\boldsymbol{\beta}_1^\mathsf{T})^\mathsf{T} = \mathbf{X}(\mathbf{I} - \boldsymbol{\beta}_1\boldsymbol{\beta}_1^\mathsf{T})$$

Projection and Deflation

UNIVERSITY *of* GLASGOW

- We can think of this operation as removing from the D-dimensional data the component that lies in the direction of the first principal direction. In other words we are deflating the matrix $\mathbf{X}$ and thus reducing its rank from D to $D-1$ i.e. removing one direction component, the principal direction.

Projection and Deflation

UNIVERSITY of GLASGOW

- We can think of this operation as removing from the D-dimensional data the component that lies in the direction of the first principal direction. In other words we are deflating the matrix $\mathbf{X}$ and thus reducing its rank from D to $D-1$ i.e. removing one direction component, the principal direction.
- Consider then the covariance of this deflated data matrix $\widetilde{\mathbf{X}} = \mathbf{X}(\mathbf{I} - \boldsymbol{\beta}_1\boldsymbol{\beta}_1^\mathsf{T})$ i.e. $\frac{1}{N}\widetilde{\mathbf{X}}^\mathsf{T}\widetilde{\mathbf{X}}$

$$
\begin{aligned}
&= \frac{1}{N}(\mathbf{I} - \boldsymbol{\beta}_1\boldsymbol{\beta}_1^\mathsf{T})\widetilde{\mathbf{X}}^\mathsf{T}\widetilde{\mathbf{X}}(\mathbf{I} - \boldsymbol{\beta}_1\boldsymbol{\beta}_1^\mathsf{T}) \\
&= \frac{1}{N}\Big(\mathbf{X}^\mathsf{T}\mathbf{X} - \boldsymbol{\beta}_1\boldsymbol{\beta}_1^\mathsf{T}\mathbf{X}^\mathsf{T}\mathbf{X} - \mathbf{X}^\mathsf{T}\mathbf{X}\boldsymbol{\beta}_1\boldsymbol{\beta}_1^\mathsf{T} + \boldsymbol{\beta}_1\boldsymbol{\beta}_1^\mathsf{T}\mathbf{X}^\mathsf{T}\mathbf{X}\boldsymbol{\beta}_1\boldsymbol{\beta}
\end{aligned}
$$

Projection and Deflation

UNIVERSITY of GLASGOW

- Taking this expression term by term we see that the right hand term can be written as

$$\boldsymbol{\beta}_1 \left(\boldsymbol{\beta}_1^{\mathsf{T}} \mathbf{X}^{\mathsf{T}} \mathbf{X} \boldsymbol{\beta}_1 \right) \boldsymbol{\beta}_1^{\mathsf{T}} = \boldsymbol{\beta}_1 \left(N \lambda_1 \right) \boldsymbol{\beta}_1^{\mathsf{T}} = N \lambda_1 \boldsymbol{\beta}_1 \boldsymbol{\beta}_1^{\mathsf{T}}$$

Projection and Deflation

- Taking this expression term by term we see that the right hand term can be written as

$$\boldsymbol{\beta}_1 \left(\boldsymbol{\beta}_1^\mathsf{T} \mathbf{X}^\mathsf{T} \mathbf{X} \boldsymbol{\beta}_1 \right) \boldsymbol{\beta}_1^\mathsf{T} = \boldsymbol{\beta}_1 \left(N\lambda_1 \right) \boldsymbol{\beta}_1^\mathsf{T} = N\lambda_1 \boldsymbol{\beta}_1 \boldsymbol{\beta}_1^\mathsf{T}$$

- For

$$\boldsymbol{\beta}_1 \boldsymbol{\beta}_1^\mathsf{T} \mathbf{X}^\mathsf{T} \mathbf{X} = \boldsymbol{\beta}_1 (N\lambda_1 \boldsymbol{\beta}_1^\mathsf{T}) = N\lambda_1 \boldsymbol{\beta}_1 \boldsymbol{\beta}_1^\mathsf{T}$$

Projection and Deflation

UNIVERSITY of GLASGOW

- Taking this expression term by term we see that the right hand term can be written as

$$\boldsymbol{\beta}_1 \left(\boldsymbol{\beta}_1^{\mathsf{T}} \mathbf{X}^{\mathsf{T}} \mathbf{X} \boldsymbol{\beta}_1 \right) \boldsymbol{\beta}_1^{\mathsf{T}} = \boldsymbol{\beta}_1 \left(N\lambda_1 \right) \boldsymbol{\beta}_1^{\mathsf{T}} = N\lambda_1 \boldsymbol{\beta}_1 \boldsymbol{\beta}_1^{\mathsf{T}}$$

- For

$$\boldsymbol{\beta}_1 \boldsymbol{\beta}_1^{\mathsf{T}} \mathbf{X}^{\mathsf{T}} \mathbf{X} = \boldsymbol{\beta}_1 (N\lambda_1 \boldsymbol{\beta}_1^{\mathsf{T}}) = N\lambda_1 \boldsymbol{\beta}_1 \boldsymbol{\beta}_1^{\mathsf{T}}$$

- and

$$\mathbf{X}^{\mathsf{T}} \mathbf{X} \boldsymbol{\beta}_1 \boldsymbol{\beta}_1^{\mathsf{T}} = N\lambda_1 \boldsymbol{\beta}_1 \boldsymbol{\beta}_1^{\mathsf{T}}$$

Projection and Deflation

- Plugging these into the expression for the covariance we obtain

$$
\begin{aligned}
\widetilde{\mathbf{C}} &= \frac{1}{N}\widetilde{\mathbf{X}}^{\mathsf{T}}\widetilde{\mathbf{X}} \\
&= \frac{1}{N}\mathbf{X}^{\mathsf{T}}\mathbf{X} - \lambda_1\boldsymbol{\beta}_1\boldsymbol{\beta}_1^{\mathsf{T}} = \widehat{\mathbf{C}} - \lambda_1\boldsymbol{\beta}_1\boldsymbol{\beta}_1^{\mathsf{T}}
\end{aligned}
$$

Projection and Deflation

UNIVERSITY of GLASGOW

- Plugging these into the expression for the covariance we obtain

$$
\begin{aligned}
\widetilde{\mathbf{C}} &= \frac{1}{N}\widetilde{\mathbf{X}}^{\mathsf{T}}\widetilde{\mathbf{X}} \\
&= \frac{1}{N}\mathbf{X}^{\mathsf{T}}\mathbf{X} - \lambda_1 \boldsymbol{\beta}_1 \boldsymbol{\beta}_1^{\mathsf{T}} = \widehat{\mathbf{C}} - \lambda_1 \boldsymbol{\beta}_1 \boldsymbol{\beta}_1^{\mathsf{T}}
\end{aligned}
$$

- Find the principal direction of deflated covariance matrix $\widetilde{\mathbf{C}}$ by solving

$$\widetilde{\mathbf{C}}\boldsymbol{\beta}_2 = \lambda_2 \boldsymbol{\beta}_2$$

Projection and Deflation

UNIVERSITY *of* GLASGOW

- Plugging these into the expression for the covariance we obtain

$$
\begin{aligned}
\widetilde{\mathbf{C}} &= \frac{1}{N}\widetilde{\mathbf{X}}^{\mathsf{T}}\widetilde{\mathbf{X}} \\
&= \frac{1}{N}\mathbf{X}^{\mathsf{T}}\mathbf{X} - \lambda_1 \boldsymbol{\beta}_1 \boldsymbol{\beta}_1^{\mathsf{T}} = \widehat{\mathbf{C}} - \lambda_1 \boldsymbol{\beta}_1 \boldsymbol{\beta}_1^{\mathsf{T}}
\end{aligned}
$$

- Find the principal direction of deflated covariance matrix $\widetilde{\mathbf{C}}$ by solving

$$\widetilde{\mathbf{C}}\boldsymbol{\beta}_2 = \lambda_2 \boldsymbol{\beta}_2$$

- Then $\boldsymbol{\beta}_2^{\mathsf{T}}\boldsymbol{\beta}_2 = 1$ and as $\widetilde{\mathbf{X}}$ resides in $D-1$ dimensional subspace orthogonal to the first principal direction $\boldsymbol{\beta}_1$ then $\boldsymbol{\beta}_1^{\mathsf{T}}\boldsymbol{\beta}_2 = 0$ must hold.

Projection and Deflation

UNIVERSITY *of* GLASGOW

- We will see further on that continuing this joint matrix deflation and solving of the associated eigenvalue problems will provide a set of eigenvectors $\{\boldsymbol{\beta}_1 \cdots \boldsymbol{\beta}_D\}$ and associated eigenvalues $\{\lambda_1 \cdots \lambda_D\}$ which provide an orthonormal basis for the data which when truncated at $P << D$ will provide the minimum reconstruction error, in the least squares sense, of the data.

Reconstruction Error

UNIVERSITY *of* GLASGOW

- The overall data reconstruction error can be written as

$$
\begin{aligned}
\mathcal{E} &= \frac{1}{N}\sum_{n=1}^{N}\left(\mathbf{x}_n - \sum_{p=1}^{P} u_{pn}\boldsymbol{\beta}_p\right)^2 \\
&= \frac{1}{N}\sum_{n=1}^{N}\left(\mathbf{x}_n^{\mathsf{T}}\mathbf{x}_n - \sum_{p=1}^{P}\boldsymbol{\beta}_p^{\mathsf{T}}\mathbf{x}_n\mathbf{x}_n^{\mathsf{T}}\boldsymbol{\beta}_p\right) \\
&= \frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_n^{\mathsf{T}}\mathbf{x}_n - \sum_{p=1}^{P}\lambda_p
\end{aligned}
$$

Reconstruction Error

UNIVERSITY of GLASGOW

- Now if there is no truncation and $P = D$ then $\mathcal{E}$ is clearly zero in which case

$$
\begin{aligned}
0 &= \frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_n^{\mathsf{T}}\mathbf{x}_n - \sum_{p=1}^{P}\lambda_p - \sum_{p'=P+1}^{D}\lambda_{p'} \\
&= \mathcal{E} - \sum_{p'=P+1}^{D}\lambda_{p'} \\
\Rightarrow &\;\; \mathcal{E} = \sum_{p'=P+1}^{D}\lambda_{p'}
\end{aligned}
$$

Reconstruction Error

UNIVERSITY
of
GLASGOW

- The reconstruction error is composed of the sum of the eigenvalues associated with the principal components discarded in the truncation

Reconstruction Error

UNIVERSITY *of* GLASGOW

- The reconstruction error is composed of the sum of the eigenvalues associated with the principal components discarded in the truncation
- As the first principal component provides the largest reduction in error and the second principal component (PC) is obtained from the deflated covariance matrix $\widehat{\mathbf{C}} - \lambda_1 \boldsymbol{\beta}_1 \boldsymbol{\beta}_1^{\mathsf{T}}$ then the reduction in error obtained by the second PC will be smaller than that obtained from the first as such $\lambda_1 \geq \lambda_2 \geq \lambda_3 \cdots \geq \lambda_D$.

Reconstruction Error

UNIVERSITY of GLASGOW

- The reconstruction error is composed of the sum of the eigenvalues associated with the principal components discarded in the truncation
- As the first principal component provides the largest reduction in error and the second principal component (PC) is obtained from the deflated covariance matrix $\widehat{\mathrm{C}} - \lambda_1 \boldsymbol{\beta}_1 \boldsymbol{\beta}_1^\mathsf{T}$ then the reduction in error obtained by the second PC will be smaller than that obtained from the first as such $\lambda_1 \geq \lambda_2 \geq \lambda_3 \cdots \geq \lambda_D$.
- This means that by studying the distribution of the eigenvalues we can potentially identify the intrinsic dimension of the data by assessing which dimensions incur the main contributions to the overall reconstruction error.

Reconstruction Error

- If we define the $D \times D$ matrix $\mathbf{B}$ whose columns are $\boldsymbol{\beta}_p$ and the $D \times D$ diagonal matrix $\mathbf{D}$ whose elements are each λ_p then the covariance matrix can be represented in terms of the associated eigenvalues and eigenvectors as

$$\widehat{\mathbf{C}} = \mathbf{B}\mathbf{D}\mathbf{B}^{\mathsf{T}}$$

Illustrative Example

UNIVERSITY
of
GLASGOW

- Consider 200 samples of 2-dimensional data denoted by the matrix $\mathbf{X}$.

Illustrative Example

UNIVERSITY
of
GLASGOW

- Consider 200 samples of 2-dimensional data denoted by the matrix $\mathbf{X}$.
- The data is drawn from two 2-D isotropic Gaussian distributions centered at [-2, -2] and [+2, +2]

Illustrative Example

UNIVERSITY *of* GLASGOW

- Consider 200 samples of 2-dimensional data denoted by the matrix $\mathbf{X}$.
- The data is drawn from two 2-D isotropic Gaussian distributions centered at [-2, -2] and [+2, +2]
- Generate a random 10×2 matrix $\mathbf{A}$ and apply the transformation $\widetilde{\mathbf{Y}} = \mathbf{X}\mathbf{A}$ such that the data has now been projected from the original 2-D space into a 10-D representation.

Illustrative Example

UNIVERSITY *of* GLASGOW

- Consider 200 samples of 2-dimensional data denoted by the matrix $\mathbf{X}$.
- The data is drawn from two 2-D isotropic Gaussian distributions centered at [-2, -2] and [+2, +2]
- Generate a random 10×2 matrix $\mathbf{A}$ and apply the transformation $\widetilde{\mathbf{Y}} = \mathbf{X}\mathbf{A}$ such that the data has now been projected from the original 2-D space into a 10-D representation.
- Finally we set $\mathbf{Y} = \widetilde{\mathbf{Y}} + \boldsymbol{\epsilon}$ where $\boldsymbol{\epsilon}$ is isotropic noise with variance 2

Illustrative Example

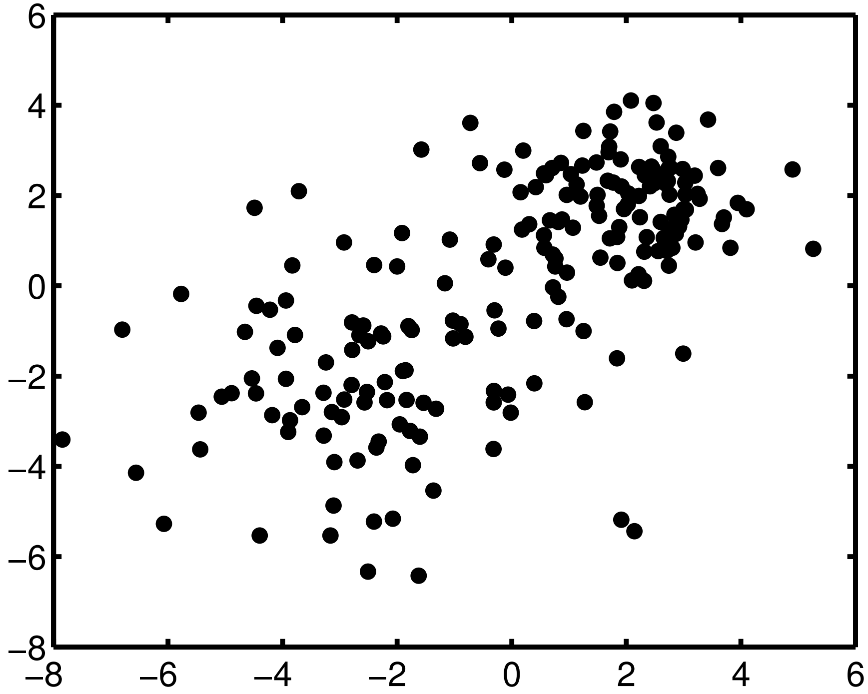


Figure 1: A scatter diagram of the 2-D data.

Illustrative Example

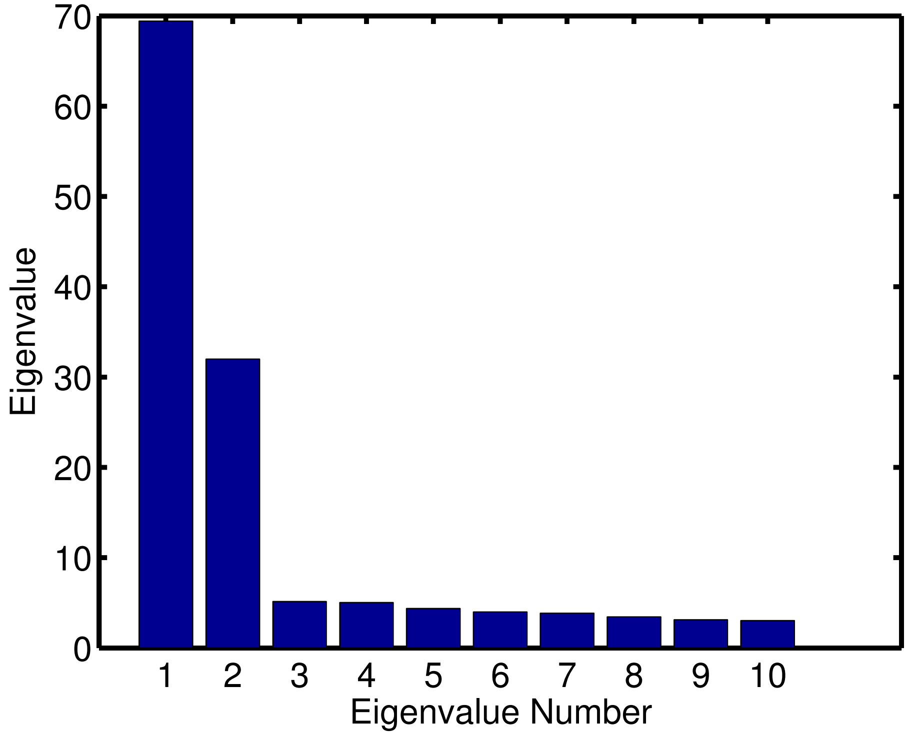
UNIVERSITY *of* GLASGOW

- Given this 10-D data let us perform PCA on the data and study how the errors are distributed throughout the ten dimensions by plotting the 10 eigenvalues $\lambda_1 \cdots \lambda_{10}$.

Illustrative Example

UNIVERSITY of GLASGOW

- Given this 10-D data let us perform PCA on the data and study how the errors are distributed throughout the ten dimensions by plotting the 10 eigenvalues $\lambda_1 \cdots \lambda_{10}$.



Illustrative Example

UNIVERSITY *of* GLASGOW

- The face data matrix $\mathbf{X}$ has dimension 400×4096 and so the covariance matrix will have dimension 4096×4096 which is huge relative to the number of examples available.

Illustrative Example

UNIVERSITY of GLASGOW

$$
\begin{aligned}
\widehat{\mathbf{C}} &= \mathbf{B}\mathbf{D}\mathbf{B}^{\mathsf{T}} \\
&\Rightarrow \frac{1}{N}\mathbf{X}^{\mathsf{T}}\mathbf{X} = \mathbf{B}\mathbf{D}\mathbf{B}^{\mathsf{T}} \\
&\Rightarrow \frac{1}{N}\mathbf{X}^{\mathsf{T}}\mathbf{X}\mathbf{B} = \mathbf{B}\mathbf{D} \\
&\Rightarrow \frac{1}{N}\mathbf{X}\mathbf{X}^{\mathsf{T}}\mathbf{X}\mathbf{B} = \mathbf{X}\mathbf{B}\mathbf{D} \\
&\Rightarrow \frac{1}{N}\mathbf{X}\mathbf{X}^{\mathsf{T}}\mathbf{U} = \mathbf{U}\mathbf{D}
\end{aligned}
$$

where we have defined $\mathbf{U} = \mathbf{X}\mathbf{B}$. Now as there are only N non-zero eigenvalues then we can see that

$$\frac{1}{N}\mathbf{X}\mathbf{X}^{\mathsf{T}}\mathbf{U} = \mathbf{U}\mathbf{D}$$

Face Images

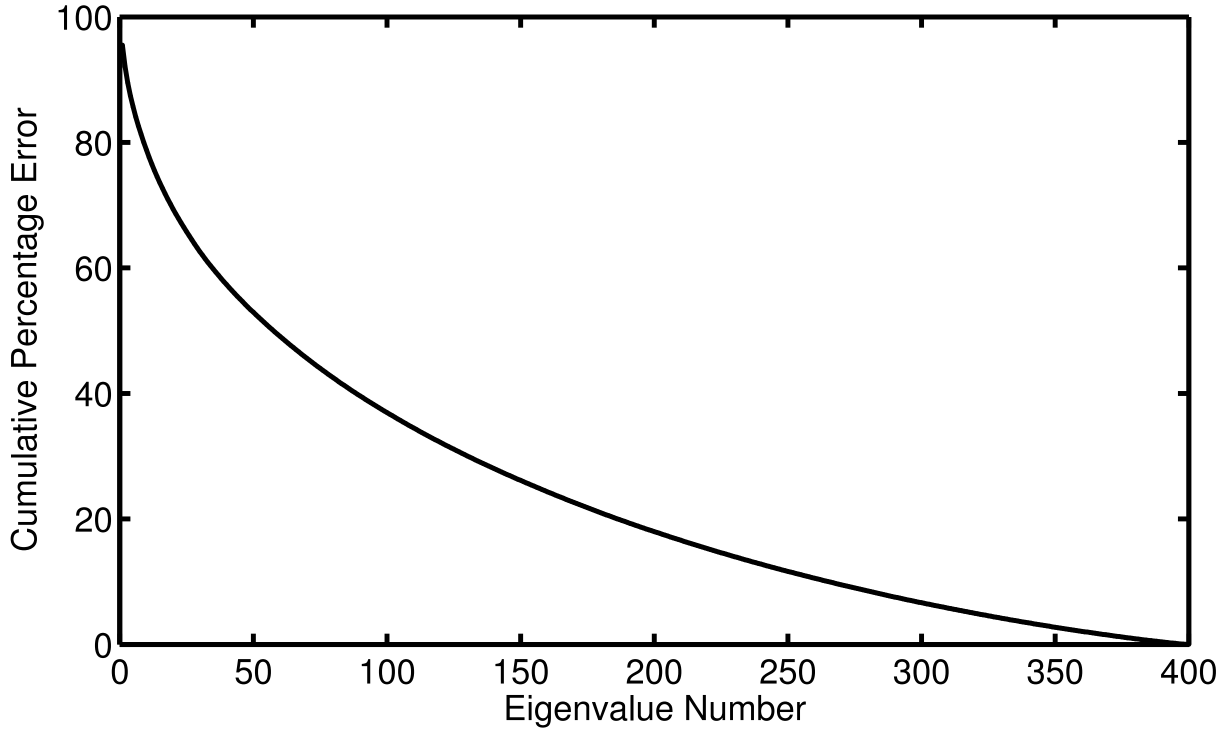


Figure 2: The percentage reconstruction error as principal components are included within the image representation.

Face Images

UNIVERSITY of GLASGOW

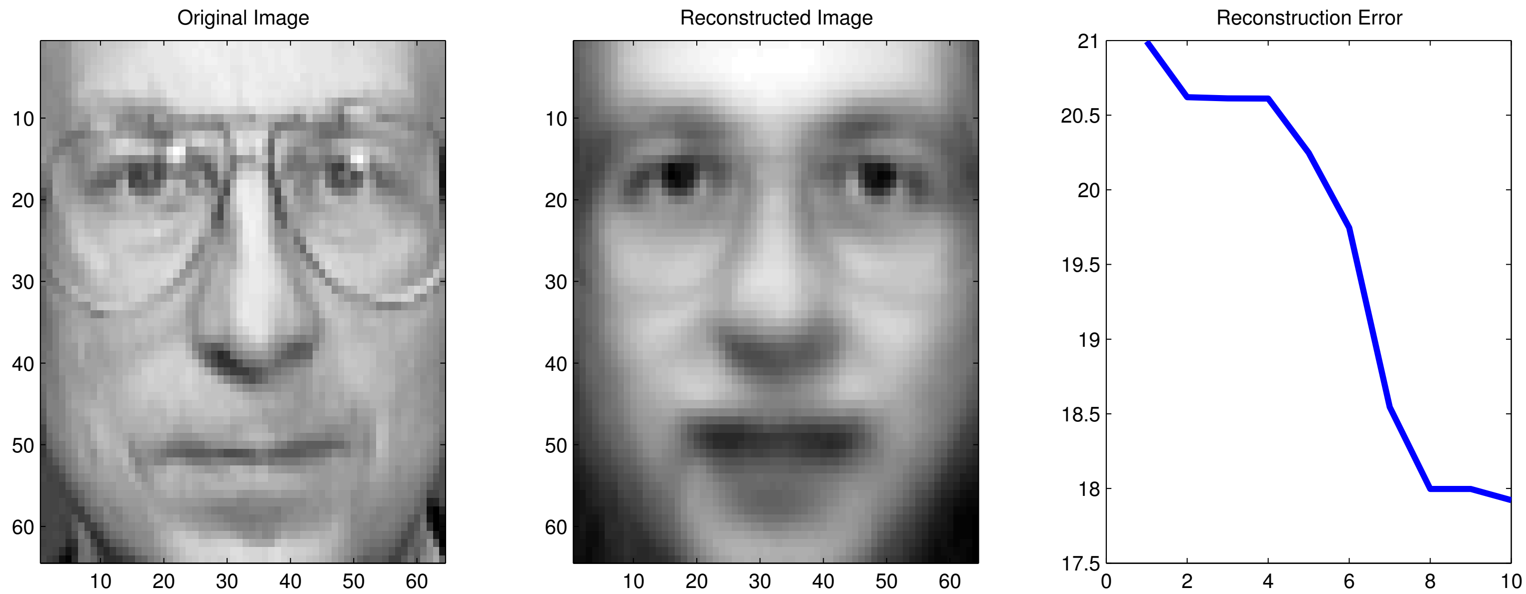


Figure 3: The original image (left) and the reconstructed image (middle) after ten principal components have been employed. The right hand plot shows how the error has decreased for this particular face over the ten PC's employed.

Face Images

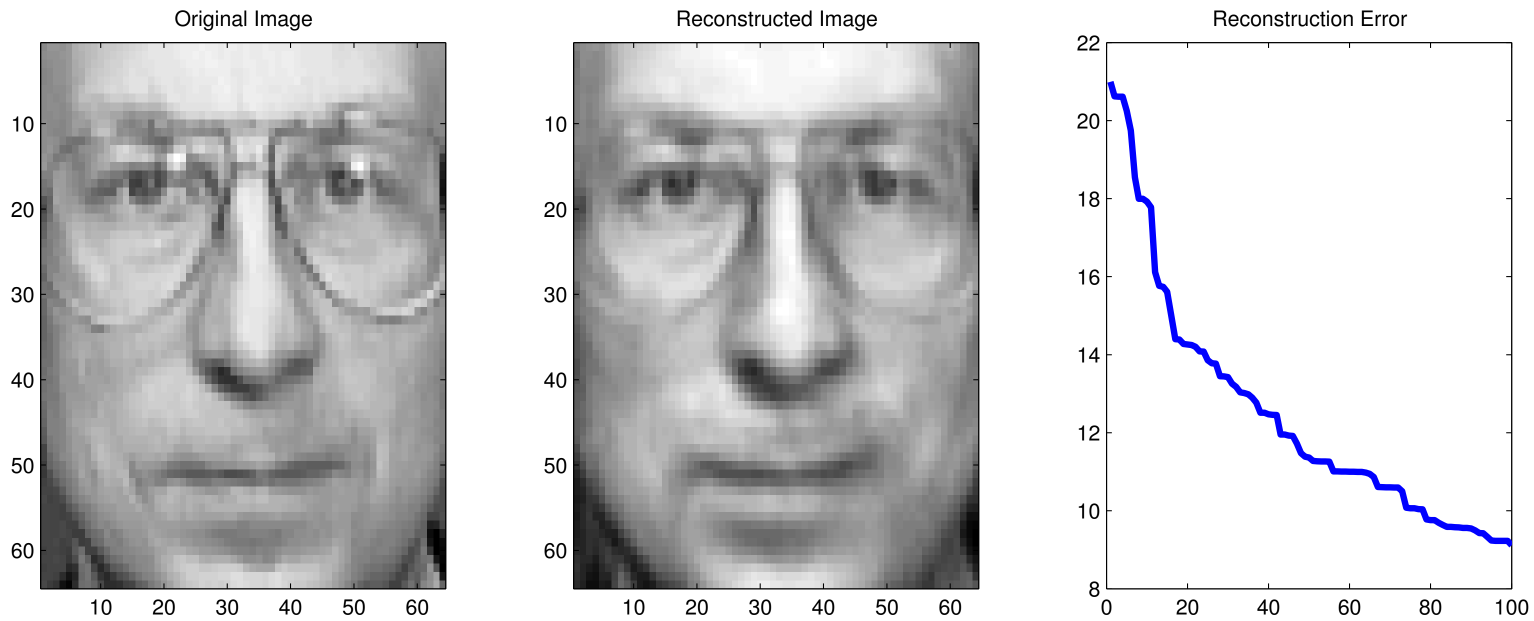


Figure 4: The original image (left) and the reconstructed image (middle) after one hundred principal components have been employed. The right hand plot shows how the error has decreased for this particular face over the one hundred PC's employed.

Generalisation

UNIVERSITY
of
GLASGOW

- Recall that the variance of predictions made by linear regression models on data points $\mathbf{x}_*$ can be given as

$$\sigma^2 \mathbf{x}_*^{\mathsf{T}} (\mathbf{X}^{\mathsf{T}} \mathbf{X})^{-1} \mathbf{x}_*$$

Generalisation

UNIVERSITY *of* GLASGOW

- Recall that the variance of predictions made by linear regression models on data points $\mathbf{x}_*$ can be given as

$$\sigma^2 \mathbf{x}_*^\mathsf{T} (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1} \mathbf{x}_*$$

- and as

$$\mathbf{X}^\mathsf{T}\mathbf{X} = N\mathbf{B}\mathbf{D}\mathbf{B}^\mathsf{T}$$

Generalisation

- Recall that the variance of predictions made by linear regression models on data points $\mathbf{x}_*$ can be given as

$$\sigma^2 \mathbf{x}_*^\mathsf{T} (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1} \mathbf{x}_*$$

- and as

$$\mathbf{X}^\mathsf{T}\mathbf{X} = N\mathbf{B}\mathbf{D}\mathbf{B}^\mathsf{T}$$

- then given that $\mathbf{B}$ is an orthonormal matrix such that $\mathbf{B}^\mathsf{T}\mathbf{B} = \mathbf{I}$ then $\mathbf{B}^{-1} = \mathbf{B}^\mathsf{T}$ we can write

Generalisation

UNIVERSITY *of* GLASGOW

- Recall that the variance of predictions made by linear regression models on data points $\mathbf{x}_*$ can be given as

$$\sigma^2 \mathbf{x}_*^\top (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{x}_*$$

- and as

$$\mathbf{X}^\top \mathbf{X} = N\mathbf{B}\mathbf{D}\mathbf{B}^\top$$

- then given that $\mathbf{B}$ is an orthonormal matrix such that $\mathbf{B}^\top \mathbf{B} = \mathbf{I}$ then $\mathbf{B}^{-1} = \mathbf{B}^\top$ we can write
-

$$(\mathbf{X}^\top \mathbf{X})^{-1} = \frac{1}{N}\mathbf{B}\mathbf{D}^{-1}\mathbf{B}^\top = \frac{1}{N}\sum_{p=1}^{D} \frac{1}{\lambda_p} \boldsymbol{\beta}_p \boldsymbol{\beta}_p^\top$$

Visualistion

UNIVERSITY *of* GLASGOW

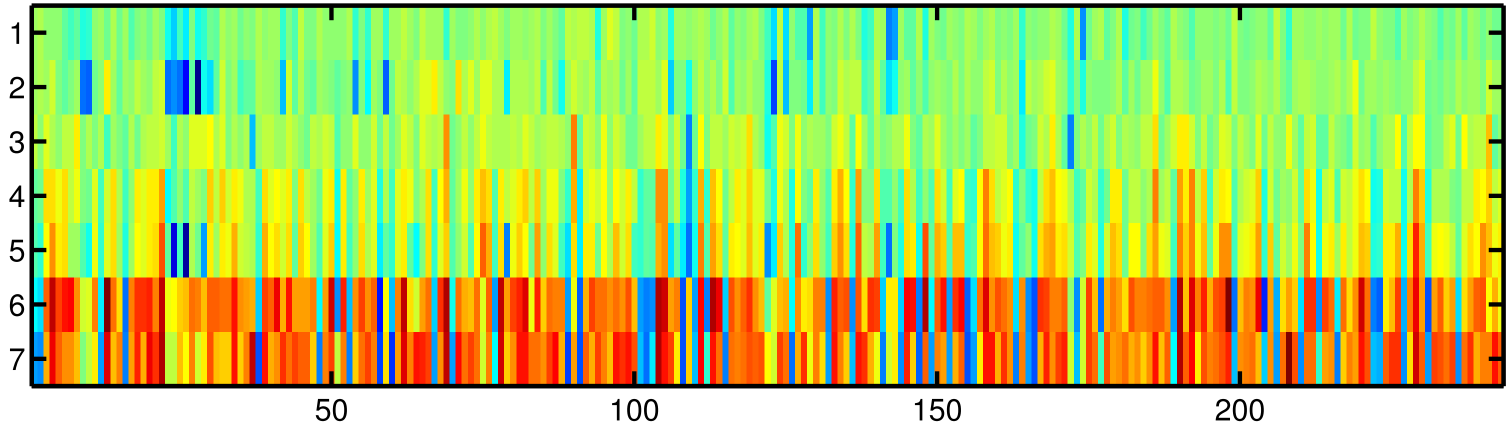


Figure 5: The differential gene expression levels of 243 genes measured at seven time points.

Visualistion

UNIVERSITY of GLASGOW

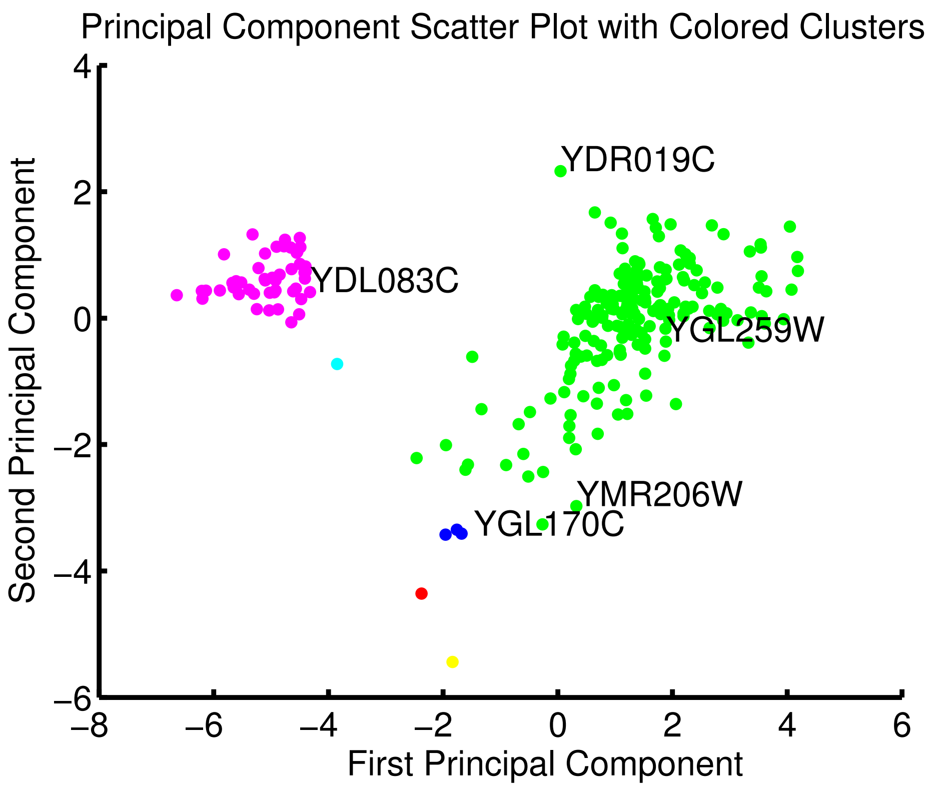


Figure 6: The projection of the differential gene expression levels of 243 genes onto the first two principal directions.

Computing a PCA

UNIVERSITY *of* GLASGOW

- The following iteration will converge to the principal eigenvector of the covariance matrix $\mathbf{C}$.

Computing a PCA

UNIVERSITY
of
GLASGOW

- The following iteration will converge to the principal eigenvector of the covariance matrix $\mathbf{C}$.

$$\begin{aligned} \mathbf{x}_t &= \mathbf{C}\mathbf{y}_{t-1} \\ \mathbf{y}_t &= \frac{\mathbf{x}_t}{\sqrt{\mathbf{x}_t^\top \mathbf{x}_t}} \end{aligned}$$

as $t \to \infty$ then $\mathbf{y}_t \to \boldsymbol{\beta}_1$ and $\sqrt{\mathbf{x}_t^\top \mathbf{x}_t} \to \lambda_1$. Covariance matrix is deflated as detailed previously $\mathbf{C} \leftarrow \mathbf{C} - \lambda_1 \boldsymbol{\beta}_1 \boldsymbol{\beta}_1^\top$ and the above iteration is applied to the deflated matrix to obtain the second eigenvector and associated eigenvalue. This is repeated until all the eigenvector/value pairs are obtained.