# Visual Analytics
# Lecture week 7: PCA

Ian T. Nabney

University of BRISTOL

# Learning Outcomes

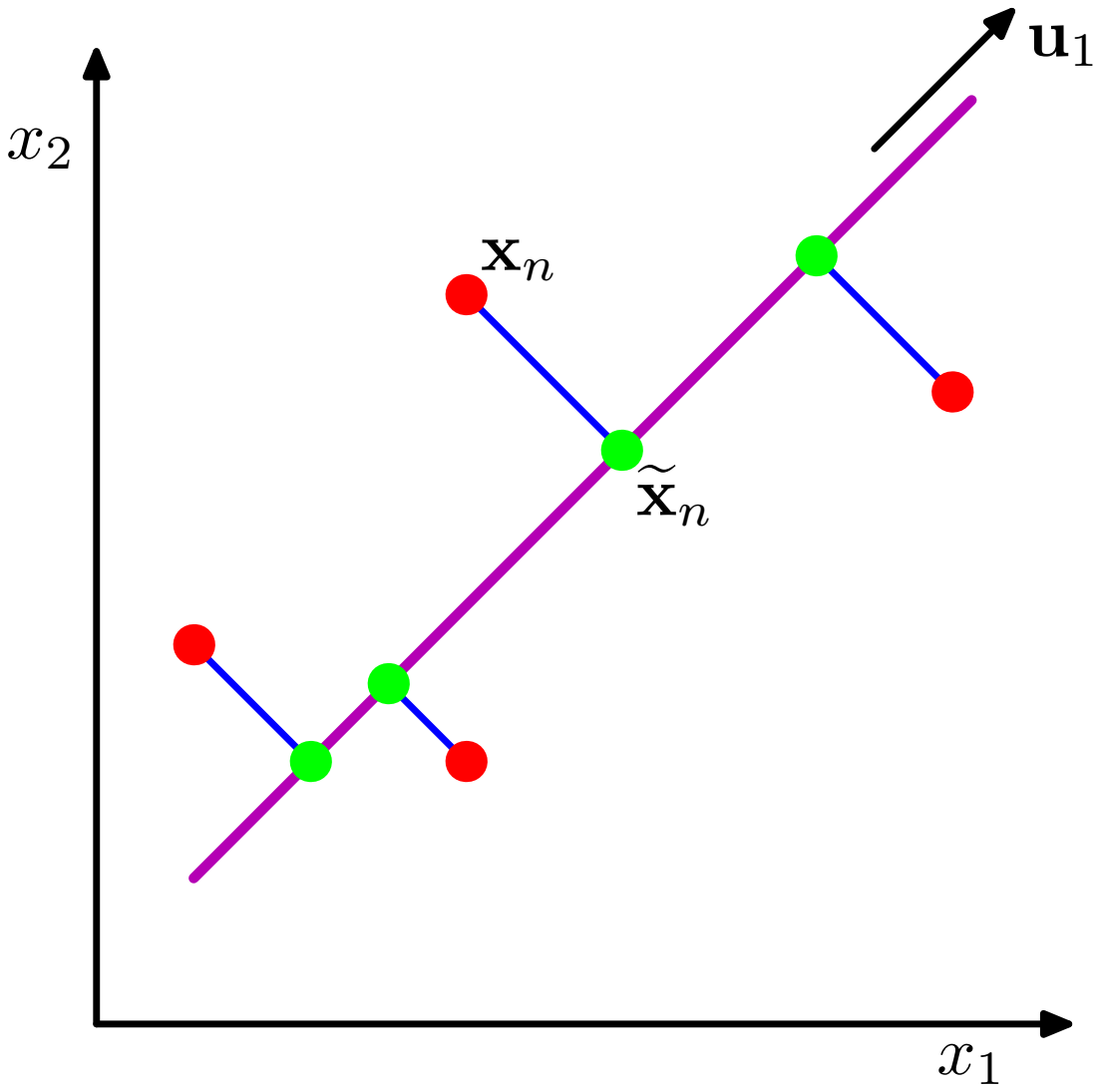
- Reading: Section 12.1 and Section 4.1.4 of Bishop
- Understand the simplest form of dimensionality reduction: Principal Component Analysis
- Able to explain PCA in terms of covariance matrix structure
- Able to use Fisher discriminant to project data with class labels

University of BRISTOL

# PCA

- PCA is widely used for applications such as dimensionality reduction, lossy data compression, feature extraction, and data visualisation.
- It is also known as the Karhunen-Loève transformation.
- There are two commonly used definitions of PCA that give rise to the same algorithm:
  - PCA can be defined as the orthogonal projection of the data onto a lower dimensional linear subspace, known as the principal subspace, such that the variance of the projected data is maximised.
  - It can also be defined as the linear projection that minimises the average projection cost, defined as the mean squared distance between the data points and their projections.

# Data Projection




- Magenta line is the principal subspace
- Orthogonal projection of data points (red dots) maximises the variance of the projected points (green dots)
- Alternatively it minimises the sum-of-squares of the projection errors, indicated by the blue lines

University of BRISTOL

# Maximum variance formulation

- Consider a data set of observations $\{\mathbf{x}_n\}$ where $n = 1, \ldots, N$, and $\mathbf{x}_n$ is a Euclidean variable with dimensionality $D$.
- Our goal is to project the data onto a space having dimensionality $M < D$ while maximizing the variance of the projected data.
- For the moment, we shall assume that the value of $M$ is given. Later in this unit, we shall consider techniques to determine an appropriate value of $M$ from the data.

University of BRISTOL

# First principal component

- Define direction of one-dimensional space using a $D$-dimensional unit vector $\mathbf{u}_1$, so that $\mathbf{u}_1^{\mathrm{T}}\mathbf{u}_1 = 1$.
- Each data point $\mathbf{x}_n$ is then projected onto a scalar value $\mathbf{u}_1^{\mathrm{T}}\mathbf{x}_n$.
- The mean of the projected data is $\mathbf{u}_1^{\mathrm{T}}\bar{\mathbf{x}}$ where $\bar{\mathbf{x}}$ is the sample set mean given by

$$\bar{\mathbf{x}} = \frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_n \tag{1}$$

- The variance of the projected data is given by

$$\frac{1}{N}\sum_{n=1}^{N}\left\{\mathbf{u}_1^{\mathrm{T}}\mathbf{x}_n - \mathbf{u}_1^{\mathrm{T}}\bar{\mathbf{x}}\right\}^2 = \mathbf{u}_1^{\mathrm{T}}\mathbf{S}\mathbf{u}_1 \tag{2}$$

where $\mathbf{S}$ is the data covariance matrix defined by

$$\mathbf{S} = \frac{1}{N}\sum_{n=1}^{N}(\mathbf{x}_n - \bar{\mathbf{x}})(\mathbf{x}_n - \bar{\mathbf{x}})^{\mathrm{T}}. \tag{3}$$

University of BRISTOL

- We now maximize the projected variance $\mathbf{u}_1^{\mathrm{T}}\mathbf{S}\mathbf{u}_1$ with respect to $\mathbf{u}_1$. This is a constrained maximization from the normalization condition $\mathbf{u}_1^{\mathrm{T}}\mathbf{u}_1 = 1$.
- To enforce this constraint, we introduce a Lagrange multiplier $\lambda_1$, and then make an unconstrained maximization of

$$\mathbf{u}_1^{\mathrm{T}}\mathbf{S}\mathbf{u}_1 + \lambda_1\left(1 - \mathbf{u}_1^{\mathrm{T}}\mathbf{u}_1\right). \tag{4}$$

- By setting the derivative with respect to $\mathbf{u}_1$ equal to zero, we see that this quantity has a stationary point when

$$\mathbf{S}\mathbf{u}_1 = \lambda_1\mathbf{u}_1 \tag{5}$$

  which shows that $\mathbf{u}_1$ is an eigenvector of $\mathbf{S}$.
- Left-multiply by $\mathbf{u}_1^{\mathrm{T}}$ and make use of $\mathbf{u}_1^{\mathrm{T}}\mathbf{u}_1 = 1$, we see that the variance is given by

$$\mathbf{u}_1^{\mathrm{T}}\mathbf{S}\mathbf{u}_1 = \lambda_1 \tag{6}$$

  and so the variance will be a maximum when we set $\mathbf{u}_1$ equal to the eigenvector having the largest eigenvalue $\lambda_1$.
- This eigenvector is known as the first principal component.

University of BRISTOL

# Other principal components

- Define additional principal components in an incremental fashion by choosing each new direction to maximise the projected variance amongst all possible directions orthogonal to those already considered.
- In the case of an $M$-dimensional projection space, the optimal linear projection for which the variance of the projected data is maximized is defined by the $M$ eigenvectors $\mathbf{u}_1, \ldots, \mathbf{u}_M$ of the data covariance matrix $\mathbf{S}$ corresponding to the $M$ largest eigenvalues $\lambda_1, \ldots, \lambda_M$.
- For visualisation, we retain the top two eigenvectors, but more can be chosen for general dimensionality reduction.
- The computational cost of computing the full eigenvector decomposition for a matrix of size $D \times D$ is $O(D^3)$.
- If $M \ll D$ then it may be more efficient to use other techniques, such as the power method which scales as $O(MD^2)$.

University of BRISTOL

# Projection error minimisation

- We introduce a complete orthonormal set of $D$-dimensional basis vectors $\{\mathbf{u}_i\}$ where $i = 1, \ldots, D$ that satisfy

$$\mathbf{u}_i^{\mathrm{T}} \mathbf{u}_j = \delta_{ij}. \tag{7}$$

- We can write any point $\mathbf{x}_n$ as

$$\mathbf{x}_n = \sum_{i=1}^{D} \left( \mathbf{x}_n^{\mathrm{T}} \mathbf{u}_i \right) \mathbf{u}_i. \tag{8}$$

- Our goal is to approximate this data point using a representation involving a restricted number $M < D$ of variables corresponding to a projection onto a lower-dimensional subspace. The $M$-dimensional linear subspace can be represented, without loss of generality, by the first $M$ of the basis vectors, and so we approximate each data point $\mathbf{x}_n$ by

$$\widetilde{\mathbf{x}}_n = \sum_{i=1}^{M} z_{ni} \mathbf{u}_i + \sum_{i=M+1}^{D} b_i \mathbf{u}_i \tag{9}$$

where the $\{z_{ni}\}$ depend on the particular data point, whereas the $\{b_i\}$ are constants that are the same for all data points.

- We choose the $\{\mathbf{u}_i\}$, the $\{z_{ni}\}$, and the $\{b_i\}$ so as to minimize the distortion introduced by the reduction in dimensionality.
- As our distortion measure, we shall use the squared distance between the original data point $\mathbf{x}_n$ and its approximation $\widetilde{\mathbf{x}}_n$, averaged over the data set, so that our goal is to minimize

$$J = \frac{1}{N}\sum_{n=1}^{N} \|\mathbf{x}_n - \widetilde{\mathbf{x}}_n\|^2. \tag{10}$$

- Substituting for $\widetilde{\mathbf{x}}_n$, setting the derivative with respect to $z_{nj}$ to zero, and making use of the orthonormality conditions, we obtain

$$z_{nj} = \mathbf{x}_n^{\mathrm{T}}\mathbf{u}_j \tag{11}$$

  where $j = 1, \ldots, M$.
- Similarly, setting the derivative of $J$ with respect to $b_i$ to zero, and again making use of the orthonormality relations, gives

$$b_j = \overline{\mathbf{x}}^{\mathrm{T}}\mathbf{u}_j \tag{12}$$

  where $j = M+1, \ldots, D$.

University of BRISTOL

- We obtain

$$\mathbf{x}_n - \widetilde{\mathbf{x}}_n = \sum_{i=M+1}^{D} \left\{ (\mathbf{x}_n - \bar{\mathbf{x}})^{\mathrm{T}} \mathbf{u}_i \right\} \mathbf{u}_i \tag{13}$$

  from which we see that the displacement vector from $\mathbf{x}_n$ to $\widetilde{\mathbf{x}}_n$ lies in the space orthogonal to the principal subspace, because it is a linear combination of $\{\mathbf{u}_i\}$ for $i = M+1, \ldots, D$.

- We therefore obtain an expression for the distortion measure $J$ as a function purely of the $\{\mathbf{u}_i\}$ in the form

$$J = \frac{1}{N} \sum_{n=1}^{N} \sum_{i=M+1}^{D} \left( \mathbf{x}_n^{\mathrm{T}} \mathbf{u}_i - \bar{\mathbf{x}}^{\mathrm{T}} \mathbf{u}_i \right)^2 = \sum_{i=M+1}^{D} \mathbf{u}_i^{\mathrm{T}} \mathbf{S} \mathbf{u}_i. \tag{14}$$

- There remains the task of minimizing $J$ with respect to the $\{\mathbf{u}_i\}$, which is a constrained minimization due to the orthonormality conditions.
- This is obtained by choosing the $\{\mathbf{u}_i\}$ to be eigenvectors of the covariance matrix given by

$$\mathbf{S}\mathbf{u}_i = \lambda_i \mathbf{u}_i \tag{15}$$

  where $i = 1, \ldots, D$, and as usual the eigenvectors $\{\mathbf{u}_i\}$ are chosen to be orthonormal.

# Distortion measure

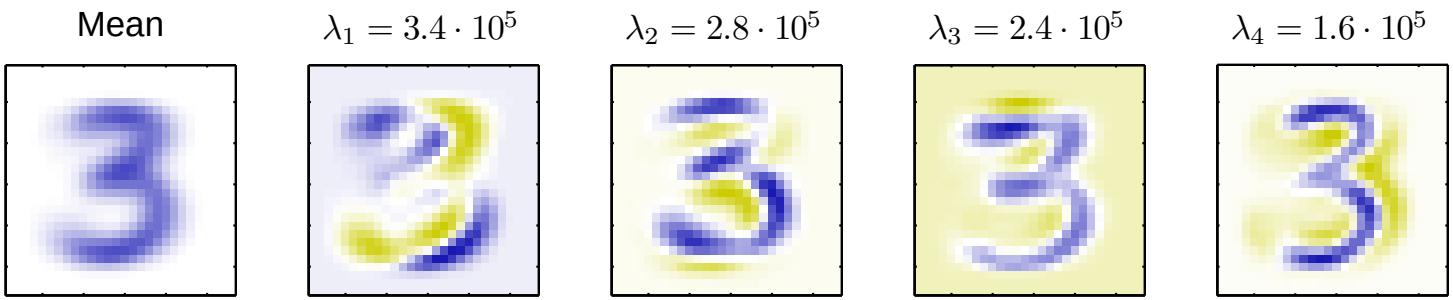
- The corresponding value of the distortion measure is

$$J = \sum_{i=M+1}^{D} \lambda_i \tag{16}$$

  which is simply the sum of the eigenvalues of those eigenvectors that are orthogonal to the principal subspace.
- We therefore obtain the minimum value of $J$ by selecting these eigenvectors to be those having the $D - M$ smallest eigenvalues, and hence the eigenvectors defining the principal subspace are those corresponding to the $M$ largest eigenvalues.
- Although we have considered $M < D$, the PCA analysis still holds if $M = D$, in which case there is no dimensionality reduction but simply a rotation of the coordinate axes to align with principal components.

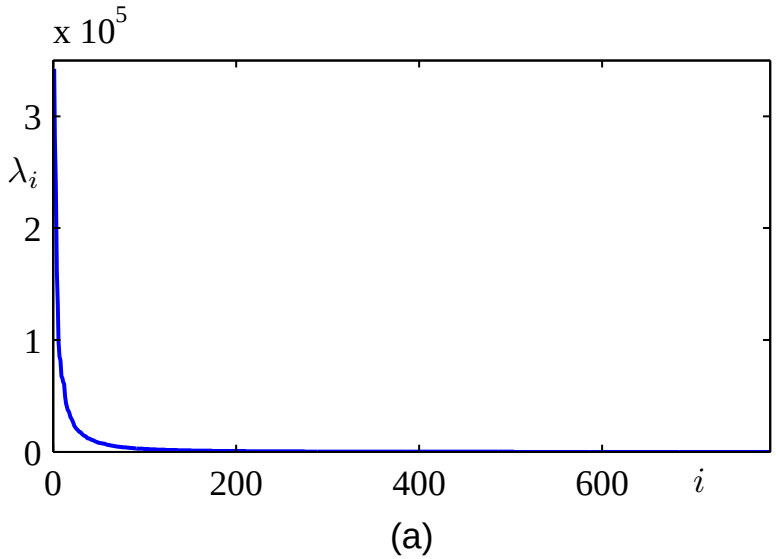
# Applying PCA to digit data

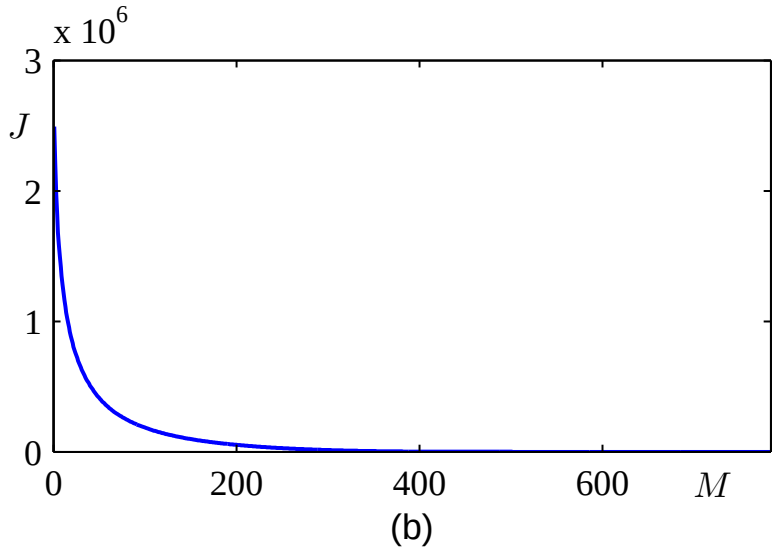
- Because each eigenvector of the covariance matrix is a vector in the original $D$-dimensional space, we can represent the eigenvectors as images of the same size as the data points.




- The mean vector $\bar{\mathbf{x}}$ along with the first four PCA eigenvectors $\mathbf{u}_1, \ldots, \mathbf{u}_4$ for the off-line digits data set, together with the corresponding eigenvalues.
- It is essential to subtract the mean vector before computing the principal components.

# Eigen analysis




- **a** Eigenvalue spectrum for the off-line digits data set.
- **b** Sum of the discarded eigenvalues, which represents the sum-of-squares distortion $J$ introduced by projecting the data onto a principal component subspace of dimensionality $M$.

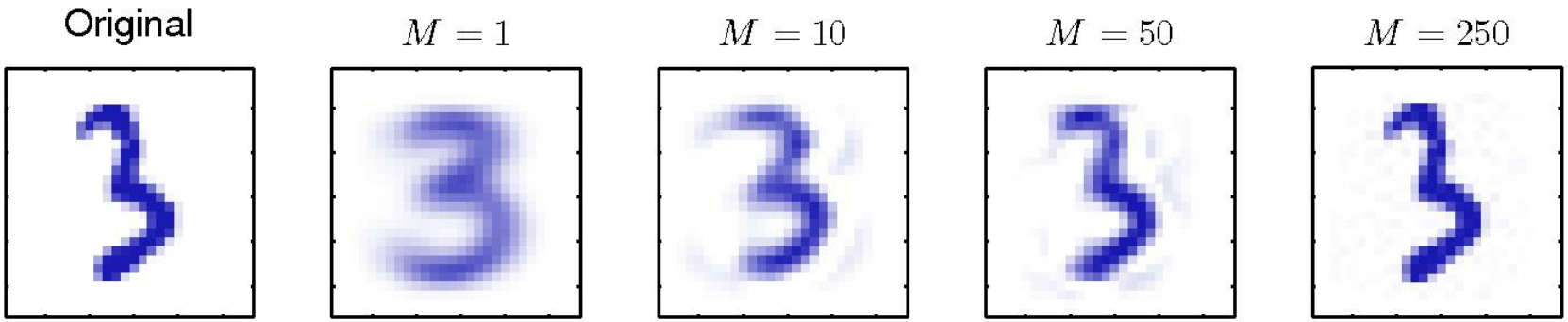
# PCA: reconstruction

1. We can write the PCA approximation to a data vector $\mathbf{x}_n$ in the form

$$\widetilde{\mathbf{x}}_n = \sum_{i=1}^{M}(\mathbf{x}_n^{\mathrm{T}}\mathbf{u}_i)\mathbf{u}_i + \sum_{i=M+1}^{D}(\overline{\mathbf{x}}^{\mathrm{T}}\mathbf{u}_i)\mathbf{u}_i \tag{17}$$

$$= \overline{\mathbf{x}} + \sum_{i=1}^{M}\left(\mathbf{x}_n^{\mathrm{T}}\mathbf{u}_i - \overline{\mathbf{x}}^{\mathrm{T}}\mathbf{u}_i\right)\mathbf{u}_i. \tag{18}$$

2. This represents a compression of the data set, because for each data point we have replaced the $D$-dimensional vector $\mathbf{x}_n$ with an $M$-dimensional vector having components $\left(\mathbf{x}_n^{\mathrm{T}}\mathbf{u}_i - \overline{\mathbf{x}}^{\mathrm{T}}\mathbf{u}_i\right)$.




PCA reconstructions: as $M$ increases the reconstruction becomes more accurate and would become perfect when $M = D = 28 \times 28 = 784$.

# PCA for data pre-processing

- The goal is not dimensionality reduction but rather the transformation of a data set in order to standardize certain of its properties. This can be important in allowing subsequent pattern recognition algorithms to be applied successfully to the data set.
- Typically, it is done when the original variables are measured in various different units or have significantly different variability.
- Standardising the dataset means rescaling it so that each variable has zero mean and unit variance: the covariance matrix is rescaled to the correlation matrix.
- With PCA we can rescale so that different variables are decorrelated. write the eigenvector equation (15) in the form

$$\mathbf{SU} = \mathbf{UL} \tag{19}$$

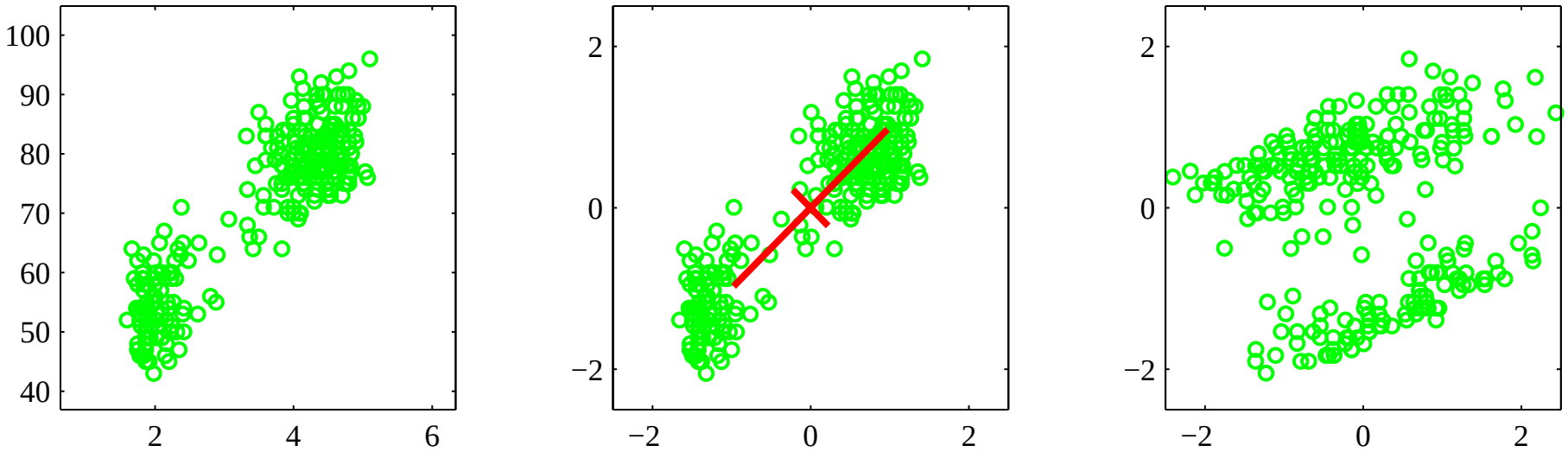
where $\mathbf{L}$ is a $D \times D$ diagonal matrix with elements $\lambda_i$. Then we transform each data point $\mathbf{x}_n$ by

$$\mathbf{y}_n = \mathbf{L}^{-1/2}\mathbf{U}^{\mathrm{T}}(\mathbf{x}_n - \bar{\mathbf{x}}) \tag{20}$$

where $\bar{\mathbf{x}}$ is the sample mean.

University of BRISTOL

# Case study: sphering the Old Faithful dataset

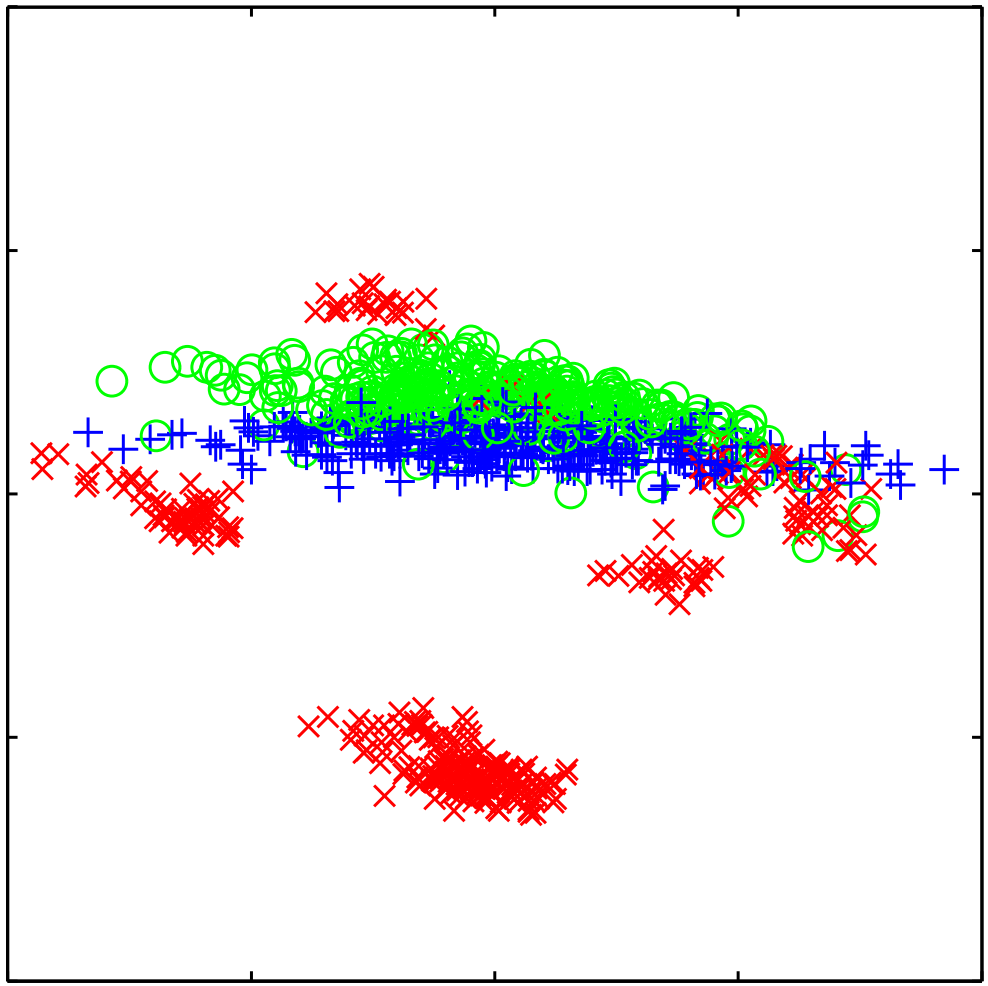



The effects of linear pre-processing applied to the Old Faithful dataset[1]. The plot on the left shows the original data. The centre plot shows the result of standardizing the individual variables to zero mean and unit variance. Also shown are the principal axes of this normalized data set, plotted over the range $\pm\lambda_i^{1/2}$. The plot on the right shows the result of sphering of the data to give it zero mean and unit covariance.

[1] https://gist.github.com/curran/4b59d1046d9e66f2787780ad51a1cd87

# PCA for data visualisation: oil flow dataset

- Each data point is projected onto a two-dimensional ($M = 2$) principal subspace.
- The red, blue, and green points correspond to the 'laminar', 'homogeneous', and 'annular' flow configurations respectively.



University of BRISTOL

# Semi-supervised learning

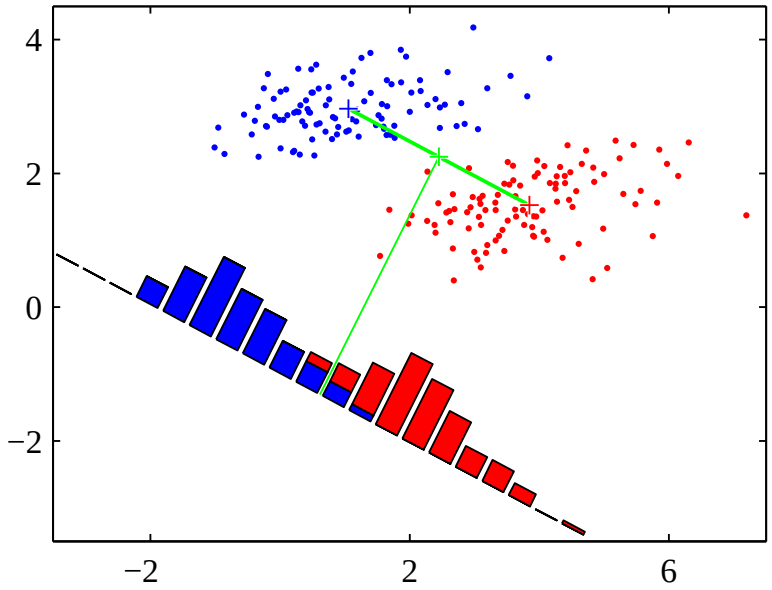
- PCA is an unsupervised learning algorithm.
- What can we do if we have some additional information: class labels? This is called semi-supervised learning, because we use the labels to train a model but not in the recall phase.
- The Fisher linear discriminant takes account of class information in choosing directions to project to: it maximises class separation in the projected space.

# Fisher's discriminant: two classes

- Consider a two-class problem in which there are $N_1$ points of class $\mathcal{C}_1$ and $N_2$ points of class $\mathcal{C}_2$.
- We project data to a one-dimensional space using $y = \mathbf{w}^{\mathrm{T}}\mathbf{x}$.
- The simplest measure of the separation of the classes, when projected onto $\mathbf{w}$, is the separation of the projected class means. This suggests that we might choose $\mathbf{w}$ so as to maximize

$$m_2 - m_1 = \mathbf{w}^{\mathrm{T}}(\mathbf{m}_2 - \mathbf{m}_1),$$

where $\mathbf{w}$ is constrained to unit length.




Two classes well separated in original space that overlap significantly when projected onto line joining their means. This is because class covariances are strongly non-diagonal.

- Compute the within-class covariance matrix

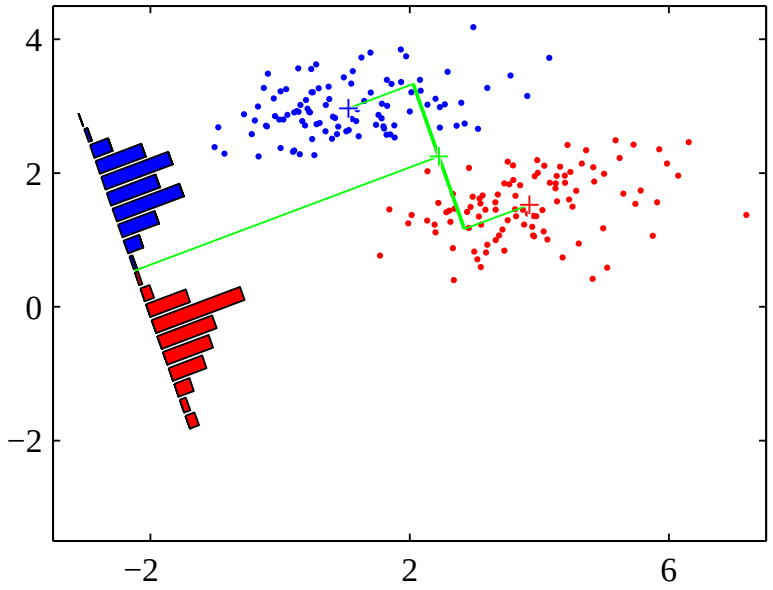
$$\mathbf{S}_{\mathrm{W}} = \sum_{n \in \mathcal{C}_1} (\mathbf{x}_n - \mathbf{m}_1)(\mathbf{x}_n - \mathbf{m}_1)^{\mathrm{T}} + \sum_{n \in \mathcal{C}_2} (\mathbf{x}_n - \mathbf{m}_2)(\mathbf{x}_n - \mathbf{m}_2)^{\mathrm{T}}. \tag{21}$$

and the between-class covariance matrix

$$\mathbf{S}_{\mathrm{B}} = (\mathbf{m}_2 - \mathbf{m}_1)(\mathbf{m}_2 - \mathbf{m}_1)^{\mathrm{T}} \tag{22}$$

and maximise

$$J(\mathbf{w}) = \frac{\mathbf{w}^{\mathrm{T}} \mathbf{S}_{\mathrm{B}} \mathbf{w}}{\mathbf{w}^{\mathrm{T}} \mathbf{S}_{\mathrm{W}} \mathbf{w}} \tag{23}$$




Corresponding projection based on the Fisher linear discriminant, showing the greatly improved class separation.

# Fisher's discriminant for multiple classes

- $K > 2$ classes, linear 'features' $y_k = \mathbf{w}_k^{\mathrm{T}}\mathbf{x}$, where $k = 1, \ldots, D'$.
- Generalization of the within-class covariance matrix

$$\mathbf{S}_{\mathrm{W}} = \sum_{k=1}^{K} \mathbf{S}_k \tag{24}$$

where (note not normalised)

$$\mathbf{S}_k = \sum_{n \in \mathcal{C}_k} (\mathbf{x}_n - \mathbf{m}_k)(\mathbf{x}_n - \mathbf{m}_k)^{\mathrm{T}}$$

- Total scatter matrix

$$\mathbf{S}_{\mathrm{T}} = \sum_{n=1}^{N} (\mathbf{x}_n - \mathbf{m})(\mathbf{x}_n - \mathbf{m})^{\mathrm{T}}$$

and between-class scatter matrix

$$\mathbf{S}_{\mathrm{B}} = \sum_{k=1}^{K} N_k (\mathbf{m}_k - \mathbf{m})(\mathbf{m}_k - \mathbf{m})^{\mathrm{T}}$$

$$\mathbf{S}_{\mathrm{T}} = \mathbf{S}_{\mathrm{W}} + \mathbf{S}_{\mathrm{B}}.$$

University of BRISTOL

- One choice of optimisation criterion is

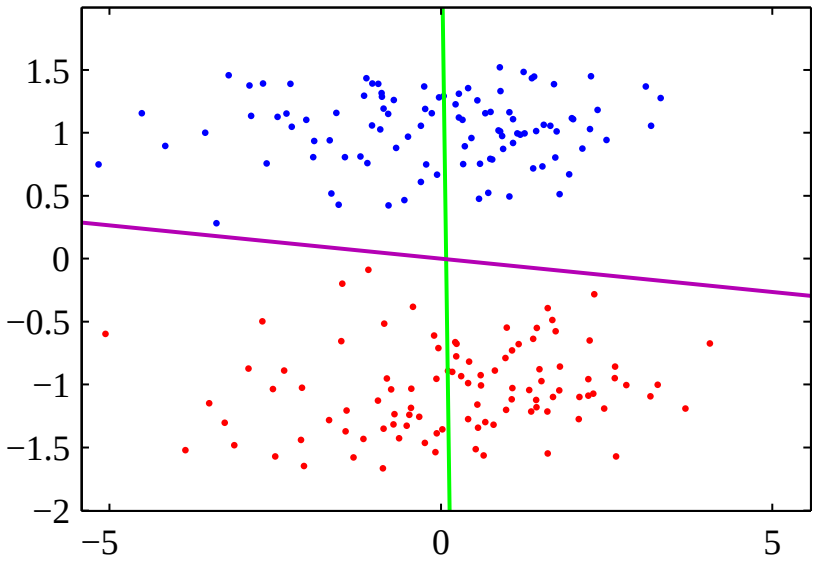
$$J(\mathbf{w}) = \mathrm{Tr}\left\{(\mathbf{W}\mathbf{S}_{\mathrm{W}}\mathbf{W}^{\mathrm{T}})^{-1}(\mathbf{W}\mathbf{S}_{\mathrm{B}}\mathbf{W}^{\mathrm{T}})\right\}.$$

  which can be optimised by computing the eigenvectors[2] of $\mathbf{S}_{\mathrm{W}}^{-1}\mathbf{S}_{\mathrm{B}}$ that correspond to the $D'$ largest eigenvalues.
- $\mathbf{S}_{\mathrm{B}}$ is composed of the sum of $K$ matrices, each of which is an outer product of two vectors and therefore of rank 1. In addition, only $(K-1)$ of these matrices are independent as a result of the constraint.
- Thus, $\mathbf{S}_{\mathrm{B}}$ has rank at most equal to $(K-1)$ and so there are at most $(K-1)$ nonzero eigenvalues.
- So we are therefore unable to find more than $(K-1)$ linear 'features' by this means.

[2] Actually better to compute generalised eigenvalues of $\mathbf{S}_{\mathrm{W}}$ and $\mathbf{S}_{\mathrm{B}}$

# Comparing PCA and linear discriminant

- Data in two dimensions, belonging to two classes shown in red and blue, is projected onto a single dimension.
- PCA chooses the direction of maximum variance, shown by the magenta curve, which leads to strong class overlap.
- Fisher linear discriminant takes account of the class labels and leads to a projection onto the green curve giving much better class separation.




University of BRISTOL

# Learning Outcomes

- Understand the simplest form of dimensionality reduction: Principal Component Analysis
- Able to explain PCA in terms of covariance matrix structure
- Able to use Fisher discriminant to project data with class labels