Machine Learning

Session 5 Dimensionality Reduction

- Introduction to Dimensionality reduction
- Properties of the projections
- Principal Component Analysis and Dimensionality Reduction
- Examples

Chap 12 of C. Bishop book

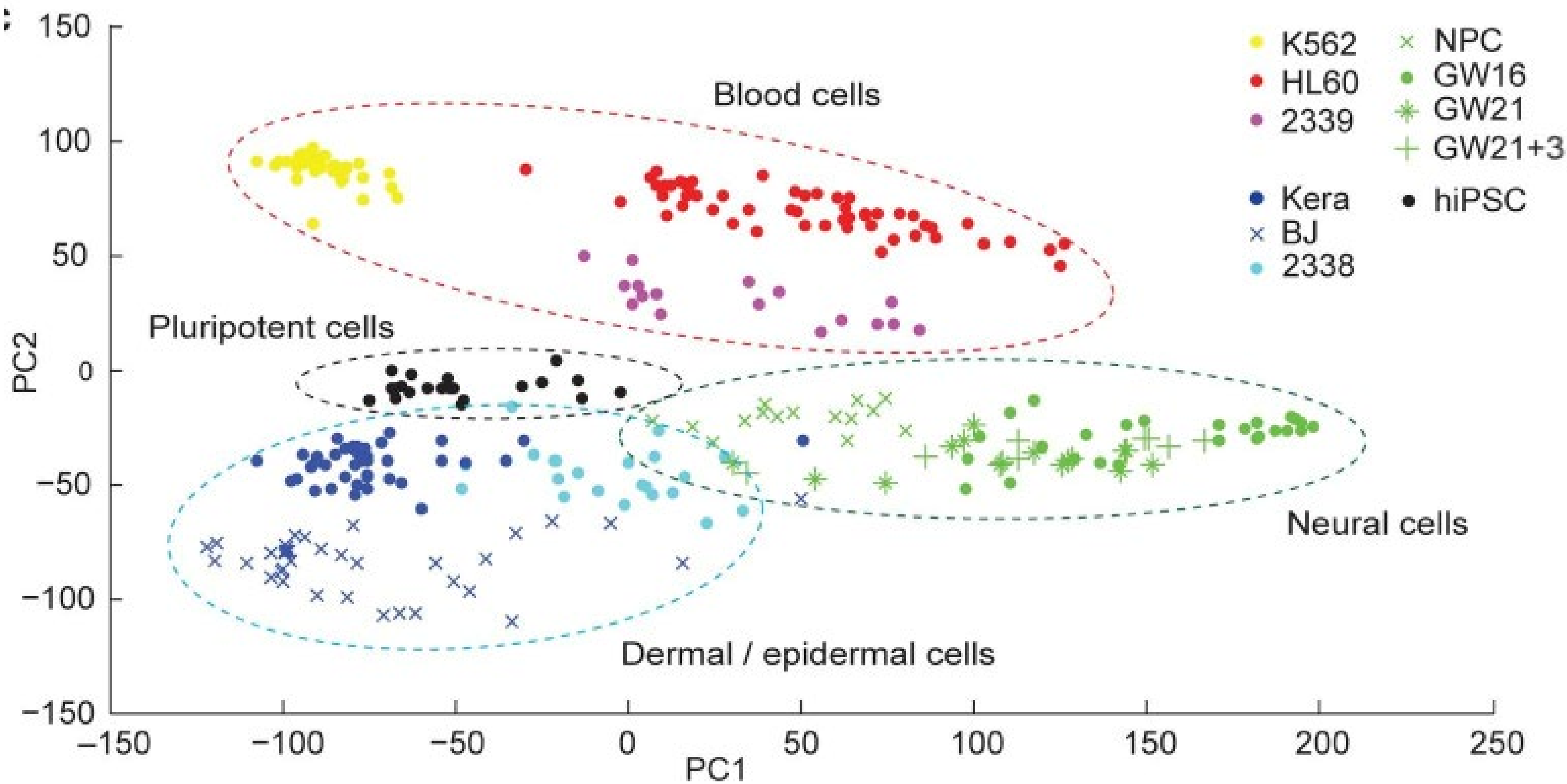
https://www.youtube.com/watch?v=HMOI_lkzW08

https://www.youtube.com/watch?v=_UVHneBUBW0

https://towardsdatascience.com/understanding-pca-fae3e243731d

Introduction to Dimensionality Reduction

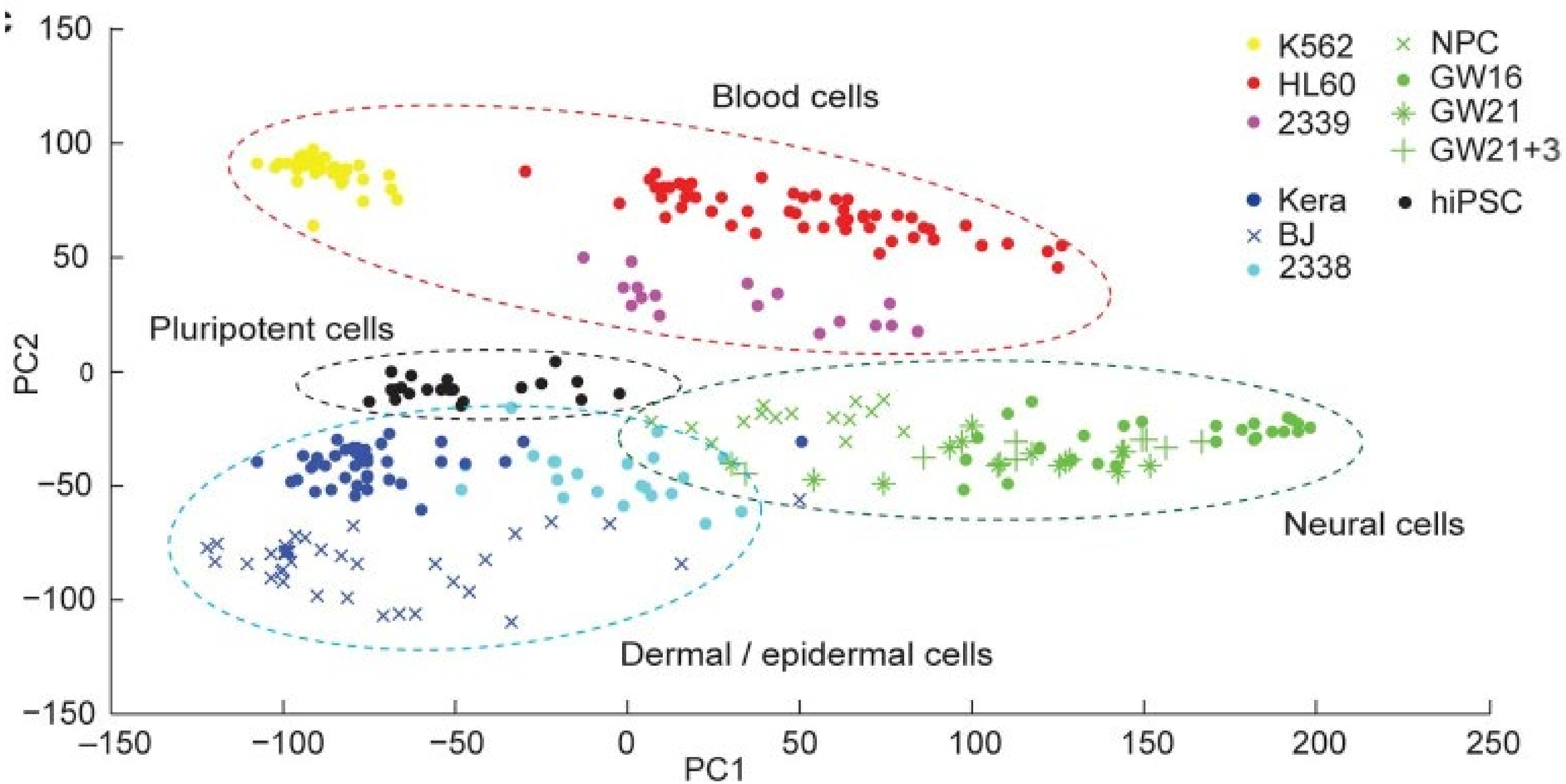
Motivation



Pollen et al. [Nat Biotechnol. 2014 Oct; 32(10): 1053–1058.]

Introduction to Dimensionality Reduction

Motivation



Pollen et al. [Nat Biotechnol. 2014 Oct; 32(10): 1053–1058.]

Take a cell and extract the gens of them. Do that for different cell types

Two axis (‘Principal Component 1 and 2’) object of this lecture

Each dot represents a single cell.

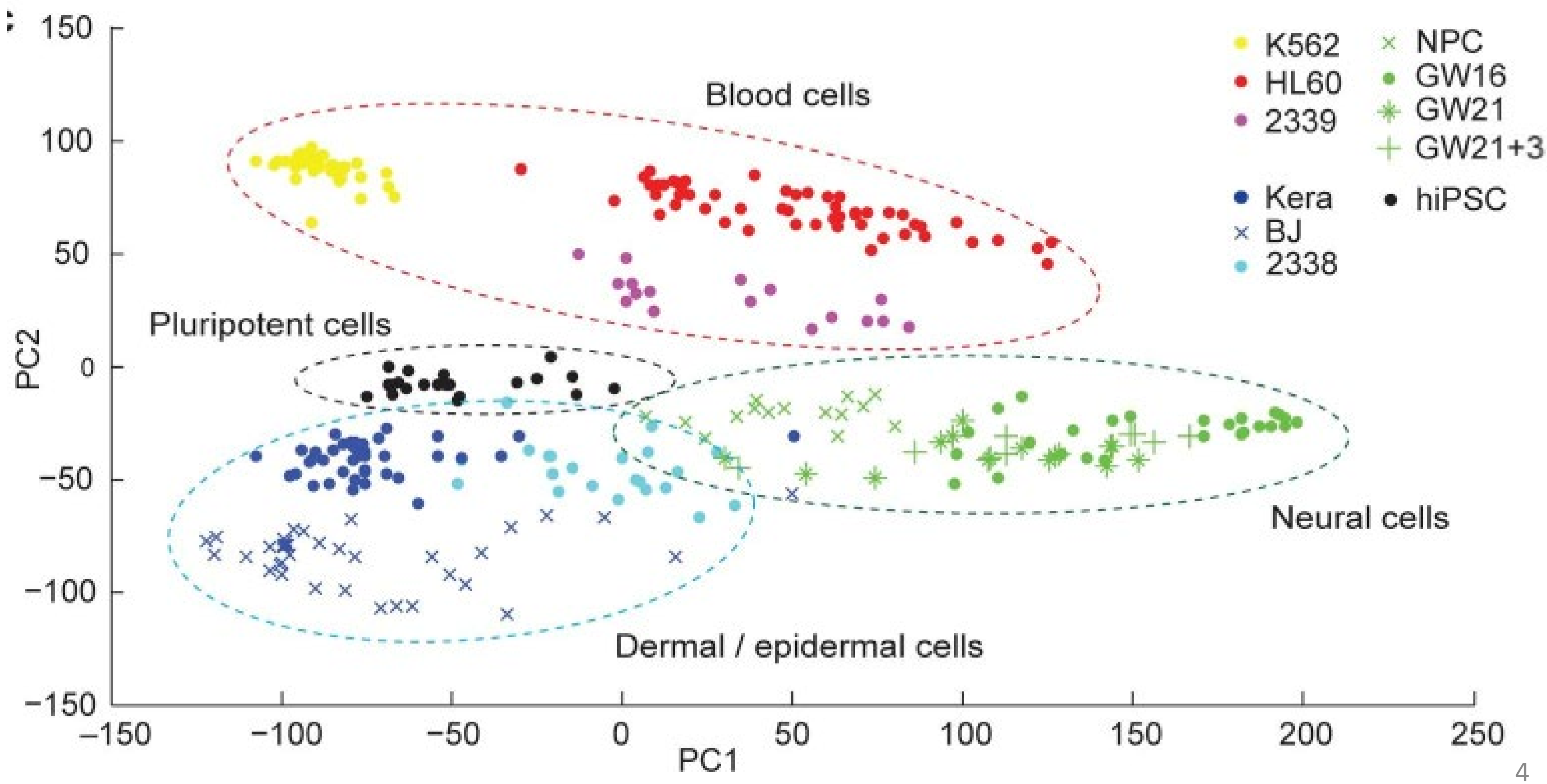
There are about 10,000 transcribed genes in each cell: a cell represented as a points $\mathbf{x}^{(n)} \in \mathbb{R}^{10000}$

Cells with similar transcriptions should cluster.

Introduction to Dimensionality Reduction

The question of this lecture:

1. How the 10,000 genes of each cell get compressed to a single point in the 2D plot?
2. **Principal Component Analysis (PCA)** is a method for compressing a lot of data into something that captures the essence of the original data.
   It tries to find this by focusing on the things that are different between the cells



Introduction to Dimensionality Reduction

Visualizing the data:

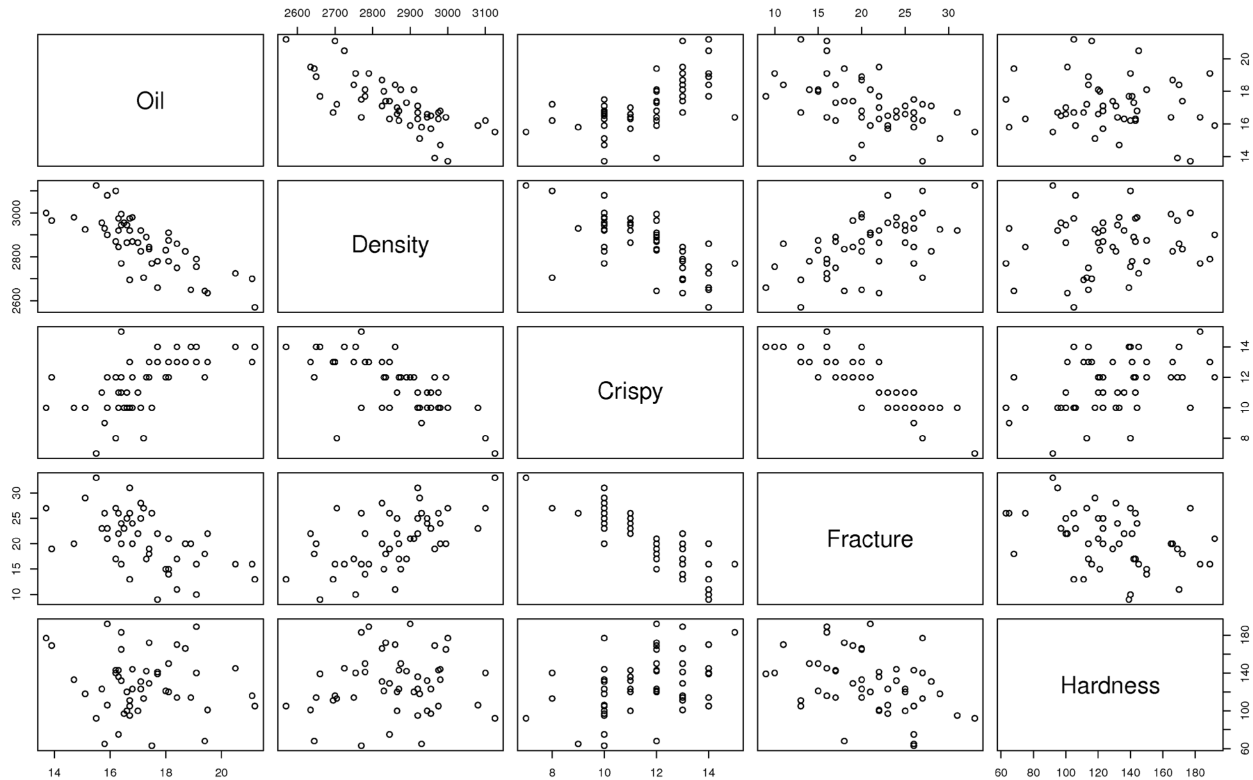
Imagine a $N \times D$ design matrix M, with N observations, each one with D variables
Example: $N = 50$ samples of, each one with $D = 5$ variables.

	x_1	x_2	x_3	x_4	x_5
$\mathbf{x}^{(1)^\top}$					
..					
$\mathbf{x}^{(N)^\top}$					

Introduction to Dimensionality Reduction

Visualizing the data:

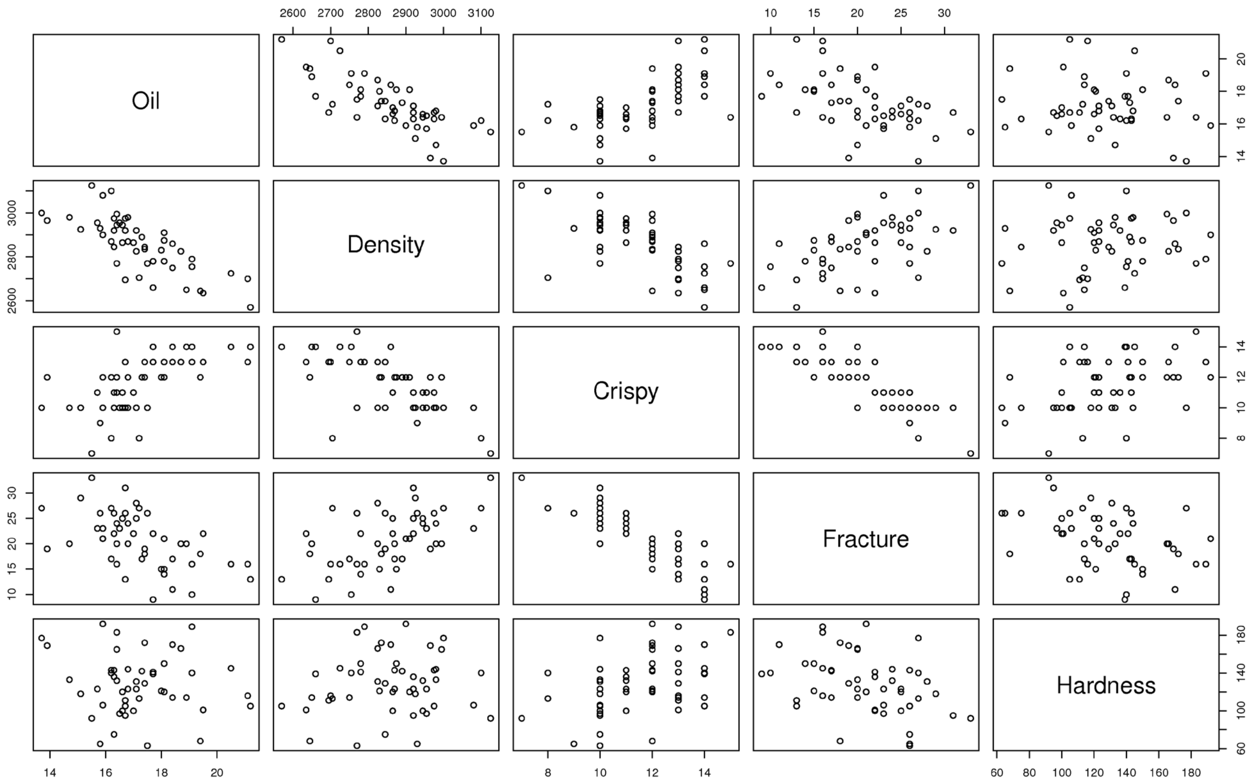
For $D = 5$, we need to represent the $\frac{D(D-1)}{2}$ scatterplots:



Introduction to Dimensionality Reduction

Visualizing the data:

For $D = 5$, we need to represent the $\frac{D(D-1)}{2}$ scatterplots:



Can we do better?

PCA uses correlations between variables (redundancies) to represent the data

Introduction to Dimensionality Reduction

In general

- For many datasets, data points lie close to a manifold of **much lower dimensionality** compared to that of the original data space

- Training continuous latent variable models is often called **dimensionality reduction**, since there are typically fewer latent dimensions.

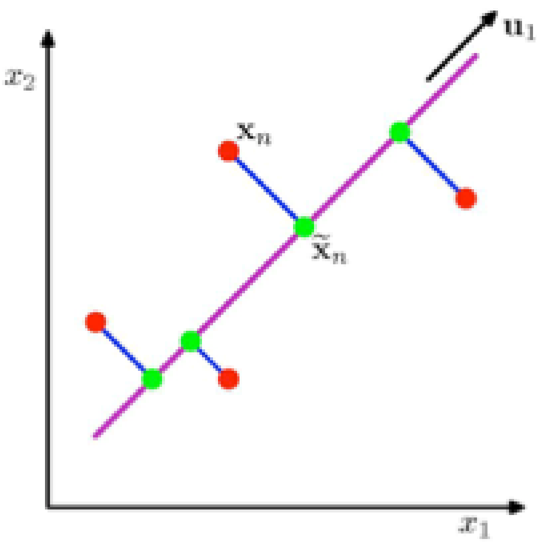
- Often there are some unknown underlying causes of the data.

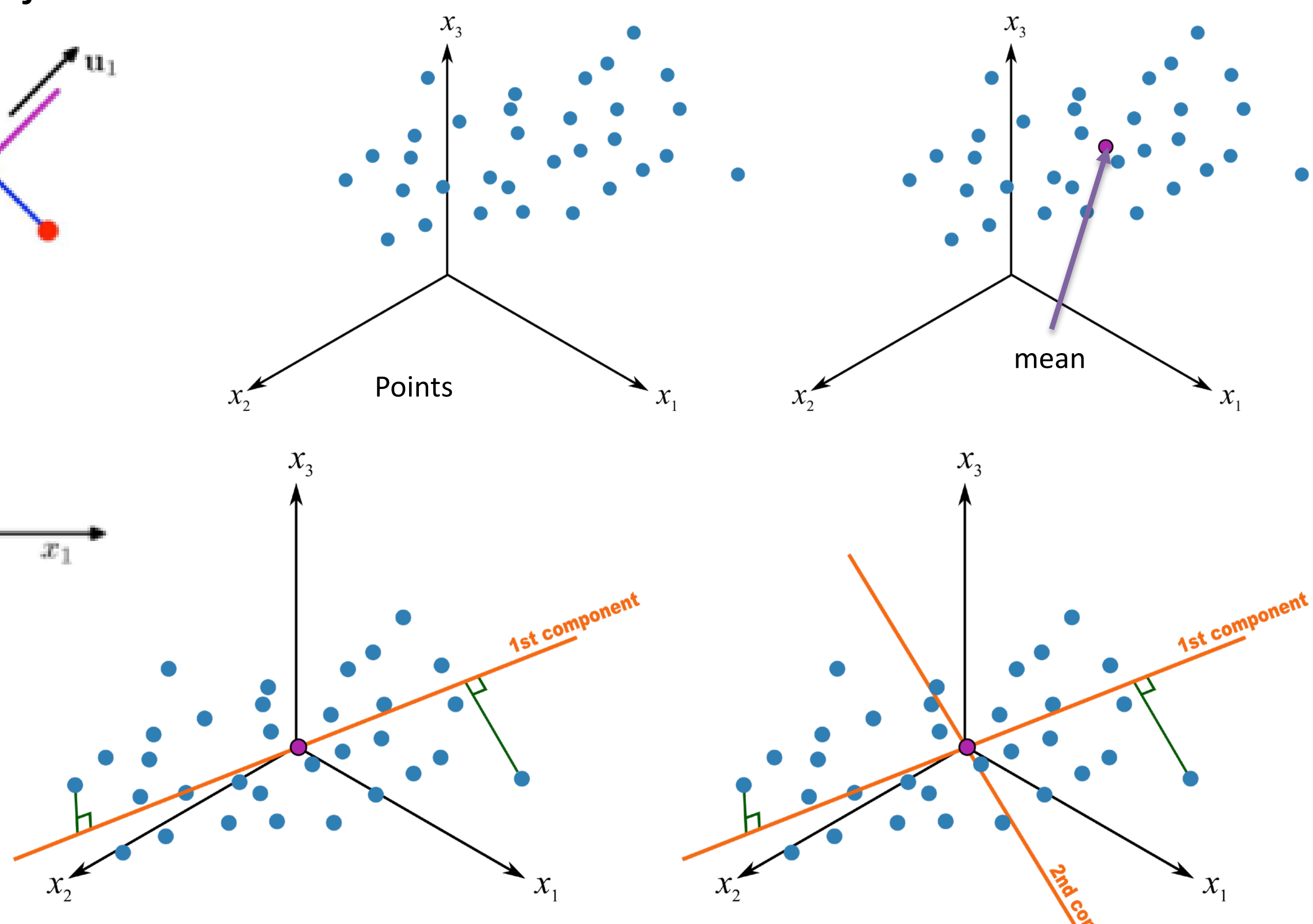
- PCA can be understood as a ***Continuous Latent Variable*** model, in contrast to the mixture of Gaussians, which has *discrete* latent variables (cluster memberships)

Introduction to Dimensionality Reduction

Main Idea of PCA:

We project each point **orthogonally** to a line (or subspace), such that the variance of the projected data is maximized.





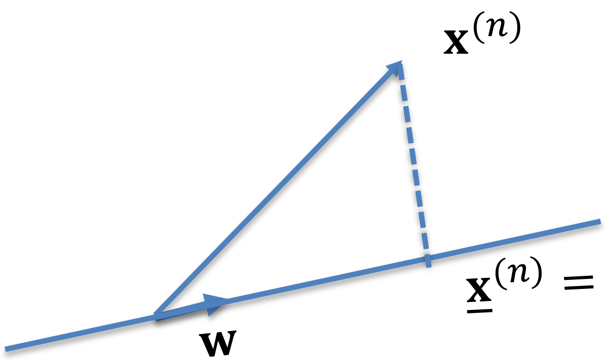
Principal directions

Properties of Projections

Main Idea of PCA:

Given data $\mathcal{D} = \{\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(N)}\}$. We project each point **orthogonally** to a line (or subspace), such that the variance of the projected data is maximized.

Preliminary (1) Orthogonal projection of a point $\mathbf{x}^{(n)}$ to a vector $\mathbf{w}$



$$\underline{\mathbf{x}}^{(n)} = \left(\mathbf{w}^\top \mathbf{x}^{(n)}\right)\mathbf{w} = r^{(n)}\mathbf{w}$$

Remember:

$$\mathbf{x}^\top \mathbf{w} = \|\mathbf{x}\|\|\mathbf{w}\|\cos\alpha$$

- The result is a number
- Order does not matter
- If $\|\mathbf{w}\|=1$, then $\mathbf{x}^\top \mathbf{w} = \|\mathbf{x}\|\cos\alpha$ is the projection of $\mathbf{x}$ onto $\mathbf{w}$

Remarks (always, **always** take $\mathbf{w}$ with norm 1)

$r^{(n)} = \mathbf{w}^\top \mathbf{x}^{(n)}$ is a number: **value of the projection** onto $\mathbf{w}$

$\underline{\mathbf{x}}^{(n)}$ is a vector: the **projected point**

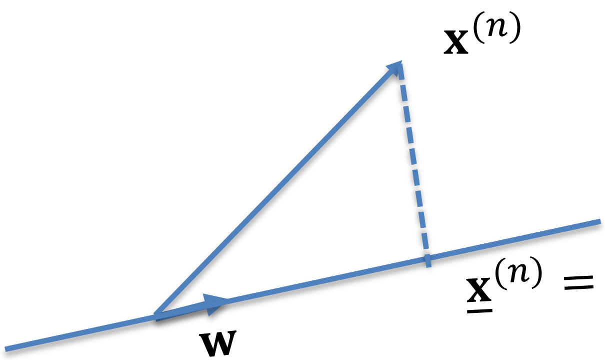
The difference between $\mathbf{x}^{(n)}$ and $\underline{\mathbf{x}}^{(n)}$ is the **loss**

Properties of Projections

Main Idea of PCA:

Given data $\mathcal{D} = \{\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(N)}\}$. We project each point **orthogonally** to a line (or subspace), such that the variance of the projected data is maximized.

Preliminary (1) Orthogonal projection of a point $\mathbf{x}^{(n)}$ to a vector $\mathbf{w}$



$$\underline{\mathbf{x}}^{(n)} = \left(\mathbf{w}^\top \mathbf{x}^{(n)}\right)\mathbf{w} = r^{(n)}\mathbf{w}$$

Remember:

$$\mathbf{x}^\top \mathbf{w} = \|\mathbf{x}\|\|\mathbf{w}\|\cos\alpha$$

- The result is a number
- Order does not matter
- If $\|\mathbf{w}\|$=1, then $\mathbf{x}^\top \mathbf{w} = \|\mathbf{x}\|\cos\alpha$ is the projection of $\mathbf{x}$ onto $\mathbf{w}$

Example:

$\mathbf{x}^{(n)} = \begin{pmatrix} 2 \\ 4 \end{pmatrix}$ on the direction $\begin{pmatrix} 1 \\ 1 \end{pmatrix}$?

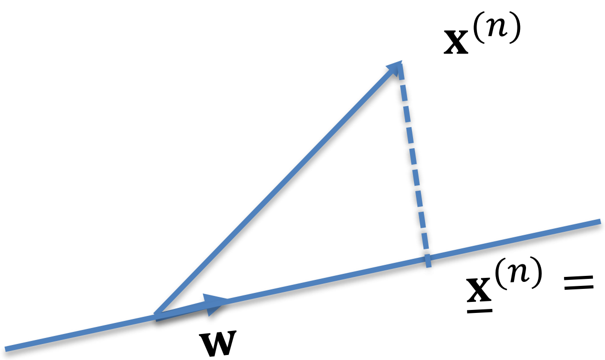
- Projection **value?**
- Vector projection?
- Loss?

Properties of Projections

Main Idea of PCA:

Given data $\mathcal{D} = \{\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(N)}\}$. We project each point **orthogonally** to a line (or subspace), such that the variance of the projected data is maximized.

Preliminary (1) Orthogonal projection of a point $\mathbf{x}^{(n)}$ to a vector $\mathbf{w}$



$$\underline{\mathbf{x}}^{(n)} = \left(\mathbf{w}^\top \mathbf{x}^{(n)}\right)\mathbf{w} = r^{(n)}\mathbf{w}$$

Remember:

$$\mathbf{x}^\top \mathbf{w} = \|\mathbf{x}\|\|\mathbf{w}\|\cos\alpha$$

- The result is a number
- Order does not matter
- If $\|\mathbf{w}\|$=1, then $\mathbf{x}^\top \mathbf{w} = \|\mathbf{x}\|\cos\alpha$ is the projection of $\mathbf{x}$ onto $\mathbf{w}$

Example:

$\mathbf{x}^{(n)} = \binom{2}{4}$ on the direction $\binom{1}{1}$. First, normalize to have unit norm, $\mathbf{w} = \begin{pmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{pmatrix}$ then

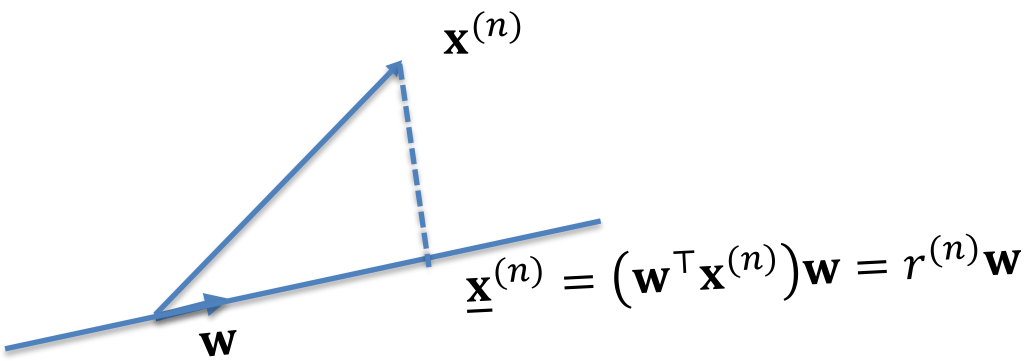
- Projection **value**: $r^{(n)} = \mathbf{w}^\top \mathbf{x}^{(n)} = \left(1/\sqrt{2}, 1/\sqrt{2}\right)\binom{2}{4} = 3\sqrt{2}$ means $3\sqrt{2}$ times the length of $\mathbf{w}$

Properties of Projections

Main Idea of PCA:

Given data $\mathcal{D} = \{\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(N)}\}$. We project each point **orthogonally** to a line (or subspace), such that the variance of the projected data is maximized.

Preliminary (1) Orthogonal projection of a point $\mathbf{x}^{(n)}$ to a vector $\mathbf{w}$



Remember:

$$\mathbf{x}^\top \mathbf{w} = \|\mathbf{x}\| \|\mathbf{w}\| \cos\alpha$$

- The result is a number
- Order does not matter
- If $\|\mathbf{w}\|$=1, then $\mathbf{x}^\top \mathbf{w} = \|\mathbf{x}\| \cos\alpha$ is the projection of $\mathbf{x}$ onto $\mathbf{w}$

Example:

$\mathbf{x}^{(n)} = \begin{pmatrix} 2 \\ 4 \end{pmatrix}$ on the direction $\begin{pmatrix} 1 \\ 1 \end{pmatrix}$. First, normalize $\mathbf{w}$ to have unit norm, $\mathbf{w} = \begin{pmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{pmatrix}$ then

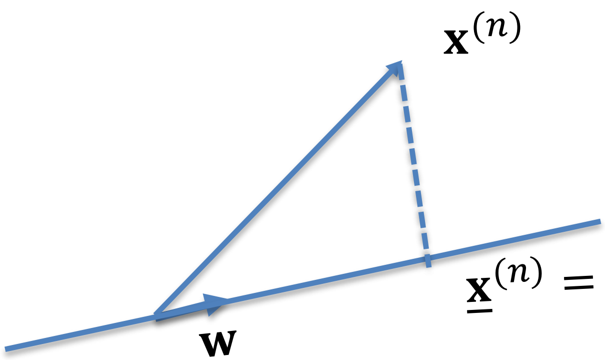
- Projection **value**: $r^{(n)} = \mathbf{w}^\top \mathbf{x}^{(n)} = \left(1/\sqrt{2}, 1/\sqrt{2}\right) \begin{pmatrix} 2 \\ 4 \end{pmatrix} = 3\sqrt{2}$ means $3\sqrt{2}$ times the length of $\mathbf{w}$
- Vector projection: $\underline{\mathbf{x}}^{(n)} = r^{(n)}\mathbf{w} = 3\sqrt{2} \begin{pmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{pmatrix} = \begin{pmatrix} 3 \\ 3 \end{pmatrix}$

Properties of Projections

Main Idea of PCA:

Given data $\mathcal{D} = \{\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(N)}\}$. We project each point **orthogonally** to a line (or subspace), such that the variance of the projected data is maximized.

Preliminary (1) Orthogonal projection of a point $\mathbf{x}^{(n)}$ to a vector $\mathbf{w}$



$$\underline{\mathbf{x}}^{(n)} = \left(\mathbf{w}^\top \mathbf{x}^{(n)}\right)\mathbf{w} = r^{(n)}\mathbf{w}$$

Remember:

$$\mathbf{x}^\top \mathbf{w} = \|\mathbf{x}\|\|\mathbf{w}\|\cos\alpha$$

- The result is a number
- Order does not matter
- If $\|\mathbf{w}\|$=1, then $\mathbf{x}^\top \mathbf{w} = \|\mathbf{x}\|\cos\alpha$ is the projection of $\mathbf{x}$ onto $\mathbf{w}$

Example:

$\mathbf{x}^{(n)} = \binom{2}{4}$ on the direction $\binom{1}{1}$. First, normalize $\mathbf{w}$ to have unit norm, $\mathbf{w} = \begin{pmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{pmatrix}$ then

- Projection **value**: $r^{(n)} = \mathbf{w}^\top \mathbf{x}^{(n)} = \left(1/\sqrt{2}, 1/\sqrt{2}\right)\binom{2}{4} = 3\sqrt{2}$ means $3\sqrt{2}$ times the length of $\mathbf{w}$
- Vector projection: $\underline{\mathbf{x}}^{(n)} = r^{(n)}\mathbf{w} = 3\sqrt{2}\begin{pmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{pmatrix} = \binom{3}{3}$
- Loss: $\mathbf{x}^{(n)} - \underline{\mathbf{x}}^{(n)} = \binom{-1}{1}$

Introduction to Dimensionality Reduction

Main Idea of PCA:

Given data $\mathcal{D} = \left\{\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(N)}\right\}$. We project each point **orthogonally** to a line (or subspace), such that the variance of the projected data is maximized.

Preliminary (2)

The mean of the projections of the points in $\mathcal{D}$ onto **w**, is the projected mean

$$
\begin{aligned}
\mathrm{E}\left[r^{(n)}\right] &= \mathrm{E}\left[\mathbf{w}^\top \mathbf{x}^{(n)}\right] \\
&= \mathbf{w}^\top \mathrm{E}\left[\mathbf{x}^{(n)}\right] \\
&= \mathbf{w}^\top \boldsymbol{\mu}
\end{aligned}
$$

Introduction to Dimensionality Reduction

Main Idea of PCA:

Given data $\mathcal{D} = \{\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(N)}\}$. We project each point **orthogonally** to a line (or subspace), such that the variance of the projected data is maximized.

Preliminary (2)

The mean of the projections of the points in $\mathcal{D}$ onto **w**, is the projected mean

$$\begin{aligned} \mathrm{E}\left[r^{(n)}\right] &= \mathrm{E}\left[\mathbf{w}^\top \mathbf{x}^{(n)}\right] \\ &= \mathbf{w}^\top \mathrm{E}\left[\mathbf{x}^{(n)}\right] \\ &= \mathbf{w}^\top \boldsymbol{\mu} \end{aligned}$$

Exemple: $\mathbf{x}^{(1)} = \begin{pmatrix} 2 \\ 4 \end{pmatrix}, \mathbf{x}^{(2)} = \begin{pmatrix} 1 \\ 3 \end{pmatrix}, \mathbf{x}^{(3)} = \begin{pmatrix} 1 \\ -3 \end{pmatrix}$ amb $\mathbf{w} = \begin{pmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{pmatrix}$

Introduction to Dimensionality Reduction

Main Idea of PCA:

Given data $\mathcal{D} = \{\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(N)}\}$. We project each point **orthogonally** to a line (or subspace), such that the variance of the projected data is maximized.

Preliminary (2)

The mean of the projections of the points in $\mathcal{D}$ onto $\mathbf{w}$, is the projected mean

$$\begin{aligned} \mathrm{E}\left[r^{(n)}\right] &= \mathrm{E}\left[\mathbf{w}^\top \mathbf{x}^{(n)}\right] \\ &= \mathbf{w}^\top \mathrm{E}\left[\mathbf{x}^{(n)}\right] \\ &= \mathbf{w}^\top \boldsymbol{\mu} \end{aligned}$$

Exemple: $\mathbf{x}^{(1)} = \begin{pmatrix} 2 \\ 4 \end{pmatrix}, \mathbf{x}^{(2)} = \begin{pmatrix} 1 \\ 3 \end{pmatrix}, \mathbf{x}^{(3)} = \begin{pmatrix} 1 \\ -3 \end{pmatrix}$ amb $\mathbf{w} = \begin{pmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{pmatrix}$

$\boldsymbol{\mu} = \begin{pmatrix} 4 \\ 4 \end{pmatrix}$, $\qquad \mathrm{E}\left[r^{(n)}\right] = \frac{4}{3}\sqrt{2}$

Introduction to Dimensionality Reduction

Main Idea of PCA:

Given data $\mathcal{D} = \{\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(N)}\}$. We project each point **orthogonally** to a line (or subspace), such that the variance of the projected data is maximized.

Preliminary (3)

The variance of the projections of the points in $\mathcal{D}$ onto $\mathbf{w}$ is the projected covariance matrix of the data

Mean of the projected data

$$
\begin{aligned}
\mathrm{Var}\left[r^{(n)}\right] = \mathrm{Var}\left[\mathbf{w}^\top \mathbf{x}^{(n)}\right] &= \mathrm{E}\left[\left(\mathbf{w}^\top \mathbf{x}^{(n)} - \mathbf{w}^\top \boldsymbol{\mu}\right)^2\right] \\
&= \mathrm{E}\left[\left(\mathbf{w}^\top \mathbf{x}^{(n)} - \mathbf{w}^\top \boldsymbol{\mu}\right)\left(\mathbf{w}^\top \mathbf{x}^{(n)} - \mathbf{w}^\top \boldsymbol{\mu}\right)^\top\right] \\
&= \mathrm{E}\left[\mathbf{w}^\top \left(\mathbf{x}^{(n)} - \boldsymbol{\mu}\right)\left(\mathbf{x}^{(n)} - \boldsymbol{\mu}\right)^\top \mathbf{w}\right] \\
&= \mathbf{w}^\top \mathrm{E}\left[\left(\mathbf{x}^{(n)} - \boldsymbol{\mu}\right)\left(\mathbf{x}^{(n)} - \boldsymbol{\mu}\right)^\top\right] \mathbf{w} \\
&= \mathbf{w}^\top \boldsymbol{\Sigma} \mathbf{w}
\end{aligned}
$$

Conclusion: The projected data has mean $\mathbf{w}^\top \boldsymbol{\mu}$ and variance $\mathbf{w}^\top \boldsymbol{\Sigma} \mathbf{w}$

Principal Component Analysis (PCA)

- **Objective**: Find a low-dimensional space such that the variance of the projected data is maximized (Hotelling 1933)

1. Find a vector $\mathbf{w}_1$ that maximizes $\mathrm{Var}[r]$ subject to $\|\mathbf{w}_1\| = 1$

Lagrangian $\mathcal{L}(\mathbf{w}_1, \lambda_1) = \mathbf{w}_1^\top \boldsymbol{\Sigma} \mathbf{w}_1 - \lambda_1(\mathbf{w}_1^\top \mathbf{w}_1 - 1)$

Principal Component Analysis (PCA)

- **Objective**: Find a low-dimensional space such that the variance of the projected data is maximized (Hotelling 1933)

1. Find a vector $\mathbf{w}_1$ that maximizes $\mathrm{Var}[r]$ subject to $\|\mathbf{w}_1\| = 1$

Lagrangian $\mathcal{L}(\mathbf{w}_1, \lambda_1) = \mathbf{w}_1^\top \mathbf{\Sigma} \mathbf{w}_1 - \lambda_1(\mathbf{w}_1^\top \mathbf{w}_1 - 1)$

Remember: $\frac{\partial \mathbf{x}^\top \mathbf{A} \mathbf{x}}{\partial \mathbf{x}} = 2\mathbf{A}\mathbf{x}$

$\frac{\partial \mathbf{x}^\top \mathbf{x}}{\partial \mathbf{x}} = 2\mathbf{x}$

- Take derivative of $\mathcal{L}$ w.r.t. $\mathbf{w}_1$:

$$\mathbf{\Sigma}\mathbf{w}_1 = \lambda_1 \mathbf{w}_1$$

that is, $\mathbf{w}_1$ ***is an eigenvector*** of $\mathbf{\Sigma}$

Principal Component Analysis (PCA)

- **Objective**: Find a low-dimensional space such that the variance of the projected data is maximized (Hotelling 1933)

1. Find a vector $\mathbf{w}_1$ that maximizes $\mathrm{Var}[r]$ subject to $\|\mathbf{w}_1\| = 1$

Lagrangian $\mathcal{L}(\mathbf{w}_1, \lambda_1) = \mathbf{w}_1^\top \boldsymbol{\Sigma} \mathbf{w}_1 - \lambda_1(\mathbf{w}_1^\top \mathbf{w}_1 - 1)$

Remember: $\frac{\partial \mathbf{x}^\top \mathbf{A}\mathbf{x}}{\partial \mathbf{x}} = 2\mathbf{A}\mathbf{x}$

$\frac{\partial \mathbf{x}^\top \mathbf{x}}{\partial \mathbf{x}} = 2\mathbf{x}$

- Take derivative of $\mathcal{L}$ w.r.t. $\mathbf{w}_1$:

$$\boldsymbol{\Sigma}\mathbf{w}_1 = \lambda_1 \mathbf{w}_1$$

that is, $\mathbf{w}_1$ ***is an eigenvector*** of $\boldsymbol{\Sigma}$

- Left-multiplying with $\mathbf{w}_1^\top$:

$$\mathbf{w}_1^\top \boldsymbol{\Sigma} \mathbf{w}_1 = \lambda_1 \mathbf{w}_1^\top \mathbf{w}_1 = \lambda_1$$

The **variance in the $\mathbf{w}_1$ direction is the largest eigenvalue** of $\boldsymbol{\Sigma}$

Principal Component Analysis (PCA)

- **Objective**: Find a low-dimensional space such that the variance of the projected data is maximized (Hotelling 1933)

1. Find a vector $\mathbf{w}_1$ that maximizes $\mathrm{Var}[r]$ subject to $\|\mathbf{w}_1\| = 1$

Lagrangian $\mathcal{L}(\mathbf{w}_1, \lambda_1) = \mathbf{w}_1^\top \mathbf{\Sigma} \mathbf{w}_1 - \lambda_1(\mathbf{w}_1^\top \mathbf{w}_1 - 1)$

Remember: $\frac{\partial \mathbf{x}^\top \mathbf{A} \mathbf{x}}{\partial \mathbf{x}} = \mathbf{2Ax}$ $\frac{\partial \mathbf{x}^\top \mathbf{x}}{\partial \mathbf{x}} = \mathbf{2x}$

- Take derivative of $\mathcal{L}$ w.r.t. $\mathbf{w}_1$:

$$\mathbf{\Sigma} \mathbf{w}_1 = \lambda_1 \mathbf{w}_1$$

that is, $\mathbf{w}_1$ ***is an eigenvector*** of $\mathbf{\Sigma}$

- Left-multiplying with $\mathbf{w}_1^\top$:

$$\mathbf{w}_1^\top \mathbf{\Sigma} \mathbf{w}_1 = \lambda_1 \mathbf{w}_1^\top \mathbf{w}_1 = \lambda_1$$

The **variance in the $\mathbf{w}_1$ direction is the largest eigenvalue** of $\mathbf{\Sigma}$

The 1st Principal Component is the eigenvector with the largest eigenvalue

Principal Component Analysis (PCA)

- **Objective**: Find a low-dimensional space such that the variance of the projected data is maximized (Hotelling 1933)

2. **Second Principal Component**: Find $\mathbf{w}_2$ that maximizes $\mathrm{Var}[r]$, subject to $\|\mathbf{w}_2\| = 1$ and $\mathbf{w}_2$ be orthogonal to $\mathbf{w}_1$

$$\mathcal{L}(\mathbf{w}_1, \lambda_1, \eta) = \mathbf{w}_2^\top \mathbf{\Sigma} \mathbf{w}_2 - \lambda_2(\mathbf{w}_2^\top \mathbf{w}_2 - 1) + \eta(\mathbf{w}_2^\top \mathbf{w}_1 - 1)$$

Principal Component Analysis (PCA)

- **Objective**: Find a low-dimensional space such that the variance of the projected data is maximized (Hotelling 1933)

2. **Second Principal Component**: Find $\mathbf{w}_2$ that maximizes $\mathrm{Var}[r]$, subject to $||\mathbf{w}_2|| = 1$ and $\mathbf{w}_2$ be orthogonal to $\mathbf{w}_1$

$$\mathcal{L}(\mathbf{w}_1, \lambda_1, \eta) = \mathbf{w}_2^\top \mathbf{\Sigma} \mathbf{w}_2 - \lambda_2(\mathbf{w}_2^\top \mathbf{w}_2 - 1) + \eta(\mathbf{w}_2^\top \mathbf{w}_1 - 1)$$

- Deriving w.r.t. $\mathbf{w}_2$ and left-multiplying the equation with $\mathbf{w}_1^\top$ we get $\mathbf{\Sigma}\mathbf{w}_2 = \lambda_2 \mathbf{w}_2$, that is, $\mathbf{w}_2$ is another eigenvector of $\mathbf{\Sigma}$
- As before, the **variance in the $\mathbf{w}_2$ direction is the second largest eigenvalue** of $\mathbf{\Sigma}$

Principal Component Analysis (PCA)

- **Objective**: Find a low-dimensional space such that the variance of the projected data is maximized (Hotelling 1933)

2. **Second Principal Component**: Find $\mathbf{w}_2$ that maximizes $\mathrm{Var}[r]$, subject to $||\mathbf{w}_2|| = 1$ and $\mathbf{w}_2$ be orthogonal to $\mathbf{w}_1$

$$\mathcal{L}(\mathbf{w}_1, \lambda_1, \eta) = \mathbf{w}_2^\top \mathbf{\Sigma} \mathbf{w}_2 - \lambda_2(\mathbf{w}_2^\top \mathbf{w}_2 - 1) + \eta(\mathbf{w}_2^\top \mathbf{w}_1 - 1)$$

   - Deriving w.r.t. $\mathbf{w}_2$ and left-multiplying the equation with $\mathbf{w}_1^\top$ we get $\mathbf{\Sigma}\mathbf{w}_2 = \lambda_2 \mathbf{w}_2$, that is, $\mathbf{w}_2$ is another eigenvector of $\mathbf{\Sigma}$
   - As before, the **variance in the $\mathbf{w}_2$ direction is the second largest eigenvalue** of $\mathbf{\Sigma}$

3. One can proceed incrementally to obtain the rest of Principal Components

Principal Component Analysis (PCA)

- The **Principal Components** are the eigenvectors of the Sample Covariance. The eigenvalue is the variance in the eigenvector direction

- The SVD of a square matrix (spectral decomposition): $\mathbf{\Sigma} = \mathbf{U}\mathbf{D}\mathbf{U}^\top$

The eigenvector $\mathbf{u}_i$ **(principal component $\boldsymbol{i}$)** is the column $\boldsymbol{i}$ of the $\mathbf{U}$ matrix.
Its eigenvalue λ_i (**variance** in the is the non-zero value of the column i in the diagonal matrix $\mathbf{D}$

- The vectors $\mathbf{u}_i$ are ordered in decreasing variance: $\lambda_1 \geq \lambda_2 \geq \lambda_3$ ...
- $\mathbf{u}_i^\top \mathbf{u}_j = \delta_{ij}$: the Principal Components are orthonormal $\mathbf{U}\mathbf{U}^\top = \mathbf{U}^\top\mathbf{U} = \mathbf{I}$
- The Covariance $\mathbf{\Sigma} = \lambda_1 \mathbf{u_1} \mathbf{u_1}^\top + \cdots + \lambda_D \mathbf{u}_D \mathbf{u}_D^\top$

Principal Component Analysis (PCA)

- The **Principal Components** are the eigenvectors of the Sample Covariance. The eigenvalue is the variance in the eigenvector direction

- The SVD of a square matrix (spectral decomposition): $\mathbf{\Sigma} = \mathbf{U}\mathbf{D}\mathbf{U}^\top$

- PCA involves evaluating the mean $\boldsymbol{\mu}$ and the covariance matrix $\mathbf{\Sigma}$ of the data set and then finding the eigenvectors of $\mathbf{\Sigma}$ corresponding to the largest eigenvalues

Principal Component Analysis (PCA)

- The **Principal Components** are the eigenvectors of the Sample Covariance. The eigenvalue is the variance in the eigenvector direction.
  2. If we take the M Principal Components: $\mathbf{W} = [\mathbf{u}_1, \dots, \mathbf{u}_M]$, then we can project any data point $\mathbf{x}^{(n)}$ into the subspace centred on $\boldsymbol{\mu}$ with the basis vectors $\langle \mathbf{u}_1, \dots, \mathbf{u}_M \rangle$

  – The **projected point** of $\mathbf{x}^{(n)}$ in such a subspace (orthogonal) is the vector

$$\mathbf{W}^{\top}\left(\mathbf{x}^{(n)} - \boldsymbol{\mu}\right)$$

  Note: the result is in $\langle \mathbf{u}_1, \dots, \mathbf{u}_M \rangle$ coordinate system (previous $r^{(n)}$in Mdimensions)

Principal Component Analysis (PCA)

- The **Principal Components** are the eigenvectors of the Sample Covariance. The eigenvalue is the variance in the eigenvector direction.
  2. If we take the M Principal Components: $\mathbf{W} = [\mathbf{u}_1, \dots, \mathbf{u}_M]$, then we can project any data point $\mathbf{x}^{(n)}$ into the subspace centred on $\boldsymbol{\mu}$ with the basis vectors $\langle \mathbf{u}_1, \dots, \mathbf{u}_M \rangle$

  – The **projected point** of $\mathbf{x}^{(n)}$ in such a subspace (orthogonal) is the vector

  $$\mathbf{W}^\top\left(\mathbf{x}^{(n)} - \boldsymbol{\mu}\right)$$

  Note: the result is in $\langle \mathbf{u}_1, \dots, \mathbf{u}_M \rangle$ coordinate system (previous $r^{(n)}$in Mdimensions)

  – The **reconstruction error of the point** $\mathbf{x}^{(n)}$ is $\left\|\mathbf{x}^{(n)} - \underline{\mathbf{x}}^{(n)}\right\|^2$

Principal Component Analysis (PCA)

- The **Principal Components** are the eigenvectors of the Sample Covariance. The eigenvalue is the variance in the eigenvector direction.
  2. If we take the M Principal Components: $\mathbf{W} = [\mathbf{u}_1, \dots, \mathbf{u}_M]$, then we can project any data point $\mathbf{x}^{(n)}$ into the subspace centred on $\boldsymbol{\mu}$ with the basis vectors $\langle \mathbf{u}_1, \dots, \mathbf{u}_M \rangle$

  – The **projected point** of $\mathbf{x}^{(n)}$ in such a subspace (orthogonal) is the vector

  $$\mathbf{W}^\top(\mathbf{x}^{(n)} - \boldsymbol{\mu})$$

  Note: the result is in $\langle \mathbf{u}_1, \dots, \mathbf{u}_M \rangle$ coordinate system (previous $r^{(n)}$in Mdimensions)

  – The **reconstruction error of the point** $\mathbf{x}^{(n)}$ is $\left\|\mathbf{x}^{(n)} - \underline{\mathbf{x}}^{(n)}\right\|^2$

  – The **reconstruction error of all the dataset** is the addition of this reconstruction from all the points

  $$\sum_n \left\|\mathbf{x}^{(n)} - \underline{\mathbf{x}}^{(n)}\right\|^2$$

Principal Component Analysis (PCA)

3. We are interested in the sum of the variances of each component of $\mathbf{x}^{(n)}$

$$\mathrm{Var}\left[\mathbf{x}^{(n)}\right] = \sum_{i=1}^{D} \mathbf{\Sigma}_{ii} = \mathrm{trace}(\mathbf{\Sigma}) = \mathrm{tr}(\mathbf{U}\mathbf{D}\mathbf{U}^{\top}) = \mathrm{tr}(\mathbf{U}^{\top}\mathbf{U}\mathbf{D}) = \mathrm{tr}(\mathbf{D}) = \sum_{i=1}^{D} \lambda_i$$

Principal Component Analysis (PCA)

3. We are interested in the sum of the variances of each component of $\mathbf{x}^{(n)}$

$$\mathrm{Var}\left[\mathbf{x}^{(n)}\right] = \sum_{i=1}^{D} \boldsymbol{\Sigma}_{ii} = \mathrm{trace}(\boldsymbol{\Sigma}) = \mathrm{tr}(\mathbf{U}\mathbf{D}\mathbf{U}^{\top}) = \mathrm{tr}(\mathbf{U}^{\top}\mathbf{U}\mathbf{D}) = \mathrm{tr}(\mathbf{D}) = \sum_{i=1}^{D} \lambda_i$$

- In the **reduced space**, the **variance explained by the** M principal Components will be $\mathrm{Var}_M = \sum_{i=1}^{M} \lambda_i$

Principal Component Analysis (PCA)

3. We are interested in the sum of the variances of each component of $\mathbf{x}^{(n)}$

$$\mathrm{Var}\left[\mathbf{x}^{(n)}\right] = \sum_{i=1}^{D} \mathbf{\Sigma}_{ii} = \mathrm{trace}(\mathbf{\Sigma}) = \mathrm{tr}(\mathbf{U}\mathbf{D}\mathbf{U}^\top) = \mathrm{tr}(\mathbf{U}^\top\mathbf{U}\mathbf{D}) = \mathrm{tr}(\mathbf{D}) = \sum_{i=1}^{D} \lambda_i$$

- In the **reduced space**, the **variance explained by the** M principal Components will be $\mathrm{Var}_M = \sum_{i=1}^{M} \lambda_i$
- The Covariance matrix in the principal components basis is diagonal. So the new r.v. are uncorrelated.

Principal Component Analysis (PCA)

3. We are interested in the sum of the variances of each component of $\mathbf{x}^{(n)}$

$$\mathrm{Var}\left[\mathbf{x}^{(n)}\right] = \sum_{i=1}^{D} \boldsymbol{\Sigma}_{ii} = \mathrm{trace}(\boldsymbol{\Sigma}) = \mathrm{tr}(\mathbf{U}\mathbf{D}\mathbf{U}^{\top}) = \mathrm{tr}(\mathbf{U}^{\top}\mathbf{U}\mathbf{D}) = \mathrm{tr}(\mathbf{D}) = \sum_{i=1}^{D} \lambda_i$$

- In the **reduced space**, the **variance explained by the** M principal Components will be $\mathrm{Var}_M = \sum_{i=1}^{M} \lambda_i$
- The Covariance matrix in the principal components basis is diagonal.
  So the new r.v. are uncorrelated.
- The **proportion of the variance** explained by the M Principal Components is

  $\mathrm{PofV} = \frac{\lambda_1 + \ldots + \lambda_M}{\lambda_1 + \ldots + \lambda_M + \cdots + \lambda_D}$, typically choose M such that PofV > **90%**

Principal Component Analysis (PCA)

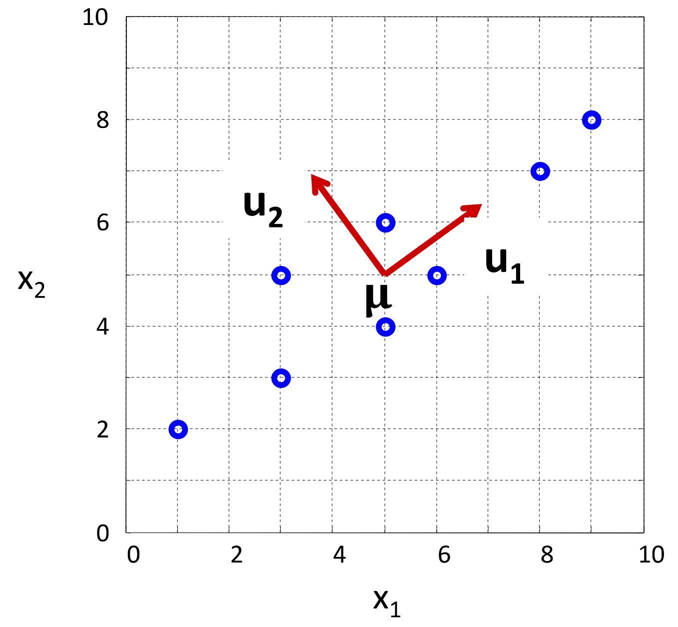
3. We are interested in the sum of the variances of each component of $\mathbf{x}^{(n)}$

$$\mathrm{Var}\left[\mathbf{x}^{(n)}\right] = \sum_{i=1}^{D} \mathbf{\Sigma}_{ii} = \mathrm{trace}(\mathbf{\Sigma}) = \mathrm{tr}(\mathbf{U}\mathbf{D}\mathbf{U}^{\top}) = \mathrm{tr}(\mathbf{U}^{\top}\mathbf{U}\mathbf{D}) = \mathrm{tr}(\mathbf{D}) = \sum_{i=1}^{D} \lambda_i$$

- In the **reduced space**, the **variance explained by the** M principal Components will be $\mathrm{Var}_M = \sum_{i=1}^{M} \lambda_i$
- The Covariance matrix in the principal components basis is diagonal.
  So the new r.v. are uncorrelated.
- The **proportion of the variance** explained by the M Principal Components is

  $\mathrm{PofV} = \frac{\lambda_1 + \ldots + \lambda_M}{\lambda_1 + \ldots + \lambda_M + \cdots + \lambda_D}$, typically choose M such that PofV > **90%**
- The **scree graph** plots value of the variance against eigenvector number. Also PofV against M

Example: What PCA does



Problem statement: we have the dataset:

X= {(1,2), (3,3), (3,5), (5,4), (5,6), (6,5), (8,7), (9,8)}

-Let's first plot the data and get an idea

- Project the point $(5, 6)^T$ on the two principal components

Solution

Let $\mathbf{X}=\begin{pmatrix} 1 & 3 & 3 & 5 & 5 & 6 & 8 & 9 \\ 2 & 3 & 5 & 4 & 6 & 5 & 7 & 8 \end{pmatrix}$; $\boldsymbol{\mu} = \begin{pmatrix} 5 \\ 5 \end{pmatrix}$

The sample covariance is

$$\boldsymbol{\Sigma_x} = \frac{\mathbf{1}}{\mathbf{8}}\mathbf{XX^T}\text{-}\boldsymbol{\mu\mu}^{\mathbf{T}} = \begin{pmatrix} 6.25 & 4.25 \\ 4.25 & 3.5 \end{pmatrix}$$

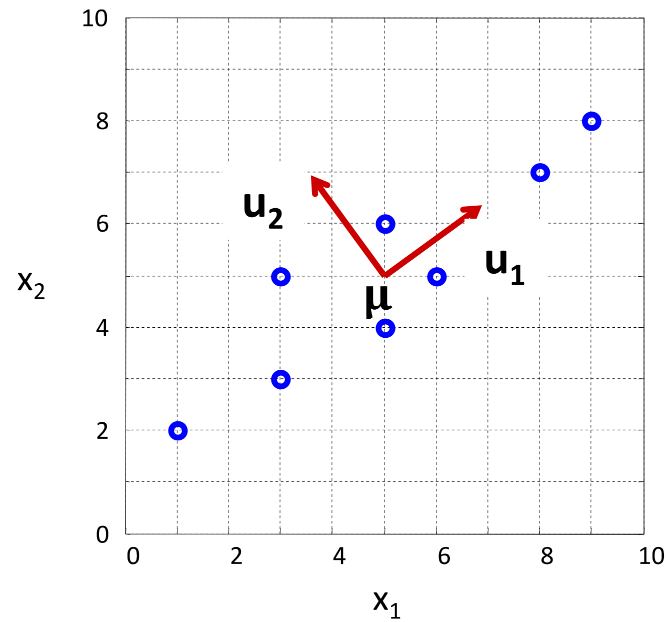
The eigenvectors and eigenvalues are the zeros of the characteristic equation:

$$\Sigma_x \mathrm{u} = \lambda \mathrm{u} \;\Rightarrow |\Sigma_x \mathrm{u} - \lambda \mathrm{I}| = 0 \Rightarrow \begin{vmatrix} 6.25-\lambda & 4.25 \\ 4.25 & 3.5-\lambda \end{vmatrix} = 0 \;\Rightarrow \lambda_1 = 9.34;\; \lambda_2 = 0.41$$

The eigenvectors are the solution of the system:

$$\begin{pmatrix} 6.25 & 4.25 \\ 4.25 & 3.5 \end{pmatrix}\begin{pmatrix} u_{11} \\ u_{12} \end{pmatrix} = 9.34\begin{pmatrix} u_{11} \\ u_{12} \end{pmatrix} \Rightarrow \boldsymbol{u_1} = \begin{pmatrix} u_{11} \\ u_{12} \end{pmatrix} \mathbf{=} \begin{pmatrix} 0.81 \\ 0.59 \end{pmatrix} \text{ with } |\mathbf{u_1}|=1 \;\text{ Same: } \boldsymbol{u_2} = \begin{pmatrix} u_{21} \\ u_{22} \end{pmatrix} \mathbf{=} \begin{pmatrix} -0.59 \\ 0.81 \end{pmatrix}$$

Example: What PCA does



Spectral decomposition $\boldsymbol{\Sigma}_{\boldsymbol{x}}=\begin{pmatrix}6.25 & 4.25\\ 4.25 & 3.5\end{pmatrix}=\mathbf{UDU}^{\mathrm{T}}=\begin{pmatrix}0.81 & -0.59\\ 0.59 & 0.81\end{pmatrix}\begin{pmatrix}9.34 & 0\\ 0 & 0.41\end{pmatrix}\begin{pmatrix}0.81 & -0.59\\ 0.59 & 0.81\end{pmatrix}^{\mathrm{T}}$

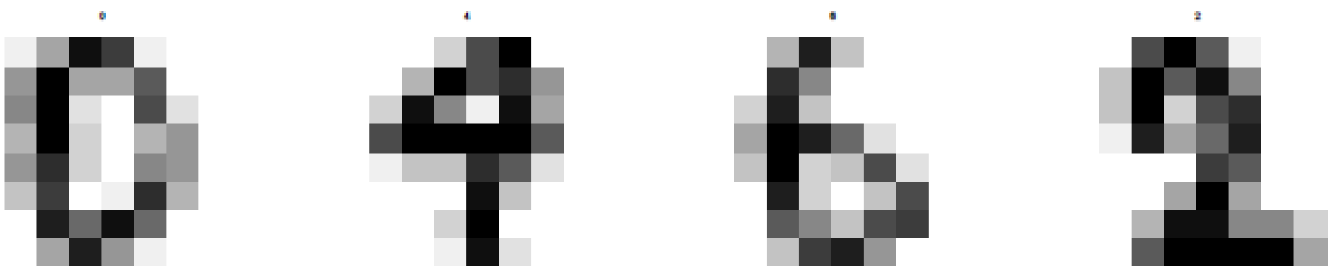
Projection of $(5,6)^{\mathrm{T}}$ on the first Principal Component:

$$\mathbf{u}_1^{\mathrm{T}}\left(\begin{pmatrix}5\\6\end{pmatrix}-\begin{pmatrix}5\\5\end{pmatrix}\right)=0.59 \quad \text{And into the second component} \quad \mathbf{u}_2^{\mathrm{T}}\left(\begin{pmatrix}5\\6\end{pmatrix}-\begin{pmatrix}5\\5\end{pmatrix}\right)=0.81$$

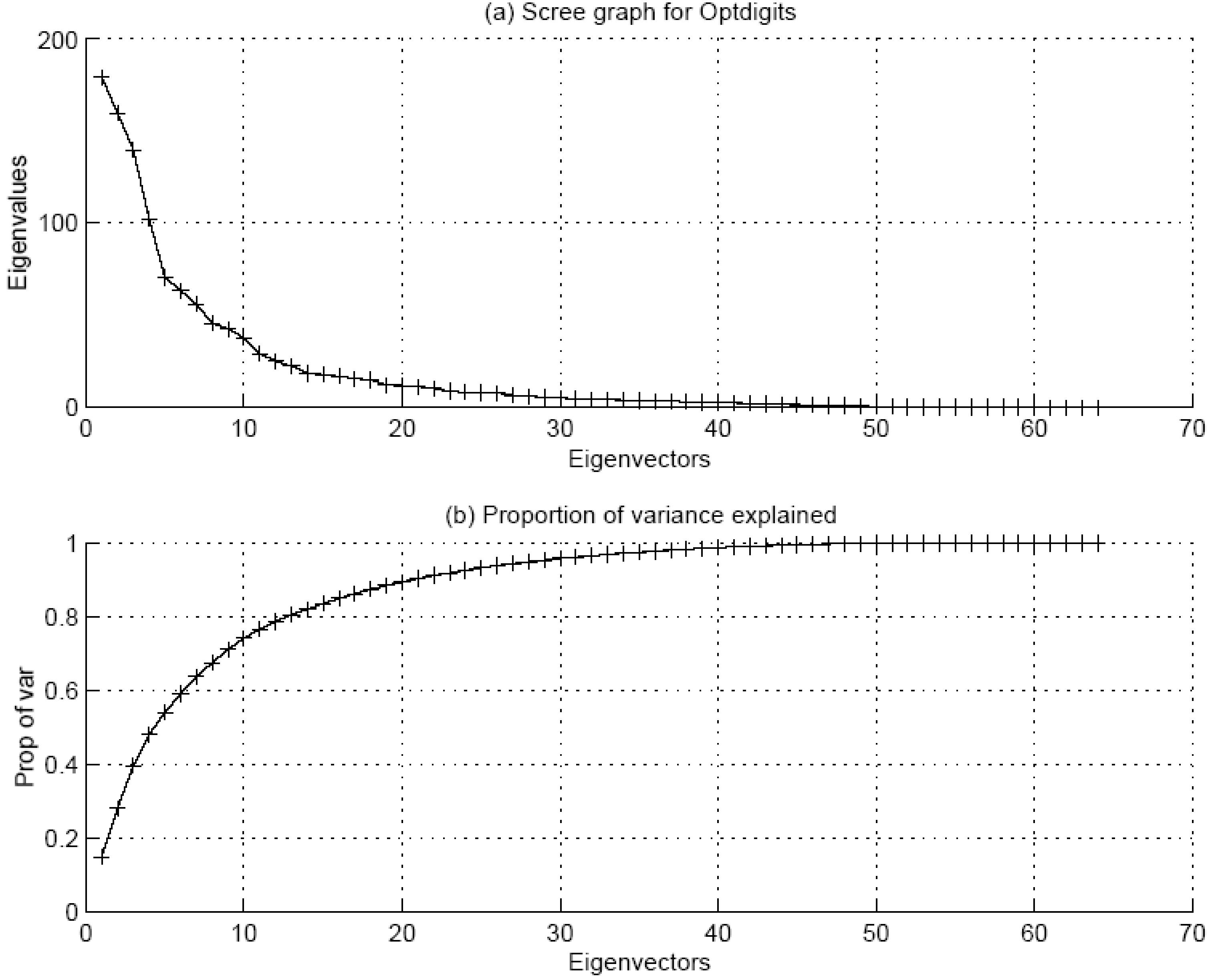
The new coordinates of $(5,\ \ 6)^{\mathrm{T}}$ in the basis $\langle\boldsymbol{\mu},\mathbf{u}_1,\mathbf{u}_2\rangle$ are $(0.59,\ \ 0.81)^{\mathrm{T}}$

Example: OPTIDIGITS dataset

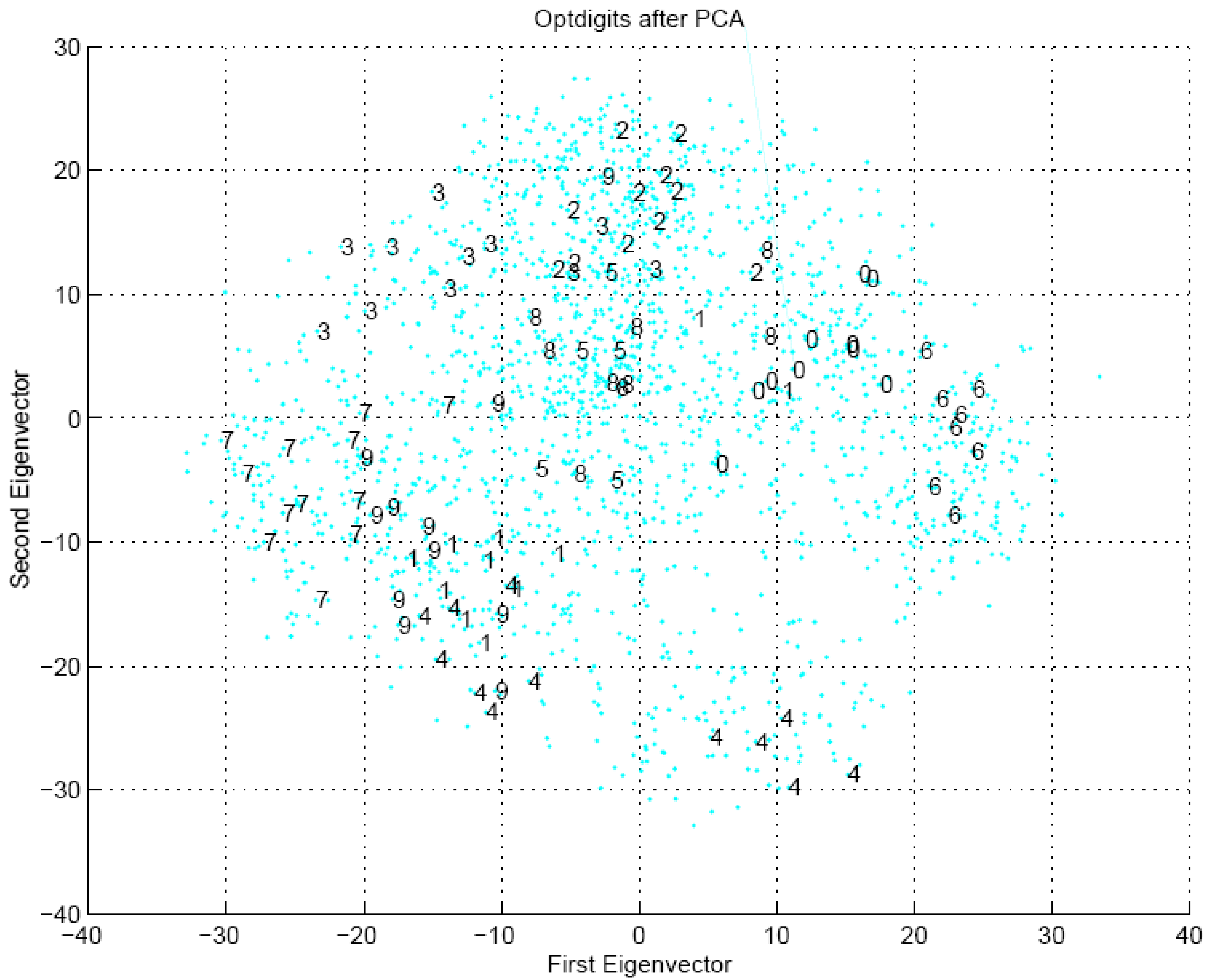
- OPTDIGITS data set contains 5620 instances of digitized handwritten digits in range 0–9.
- Each digit is a $\mathbb{R}^{64}$ vector: $8 \times 8 = 64$ pixels, 16 grayscales.



Example: OPTIDIGITS Scree Graph



Example: OPTIDIGITS visualization: Projection on the 2 first eigens (from R^{64} to R^2)



PCA application to FACES

- Run PCA on 2429 19x19 grayscale images (CBCL database)

- Data compression: We can get good reconstructions with only 3 components.

- Pre-processing: We can apply a standard classifier to latent representation – PCA with 3 components obtains 79% accuracy on face/non-face discrimination in test data, vs. 77% for a mixture of Gaussians with 84 components.

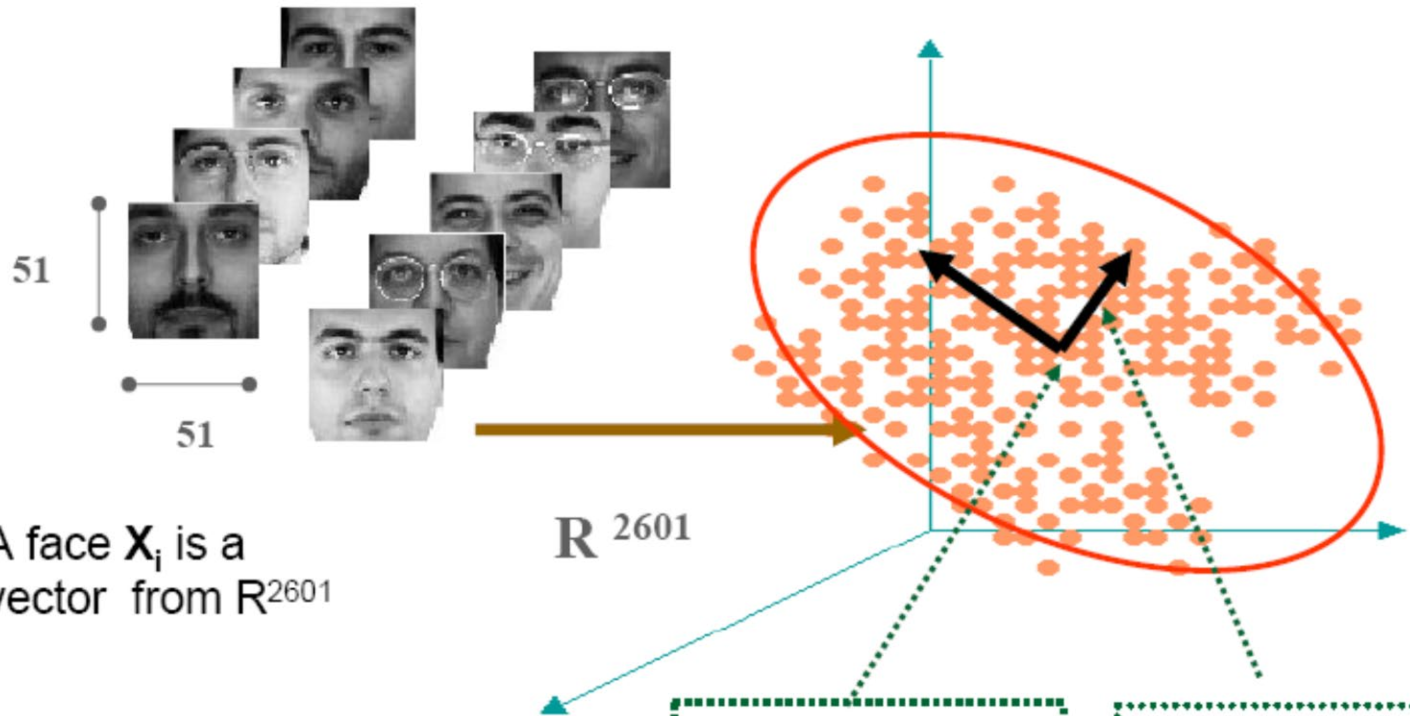
- Data visualization: by projecting the data onto the first two principal components.

PCA application to FACES

Learned Basis

- Run PCA on 2429 19x19 grayscale images (CBCL database)

PCA application to FACES



A face $\mathbf{X}_i$ is a vector from R^{2601}

$$\overline{X} = \frac{1}{n} \sum_{i=1..n} \vec{X}_i$$

Basis vectors are the eigenvalues of the Covariance Matrix

PCA application to FACES