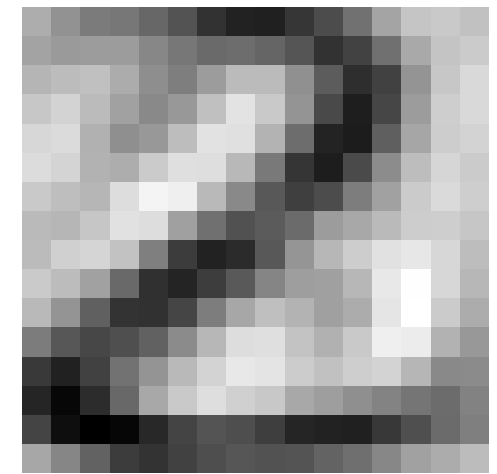
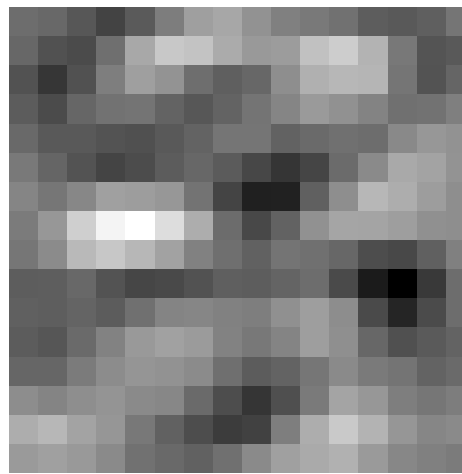
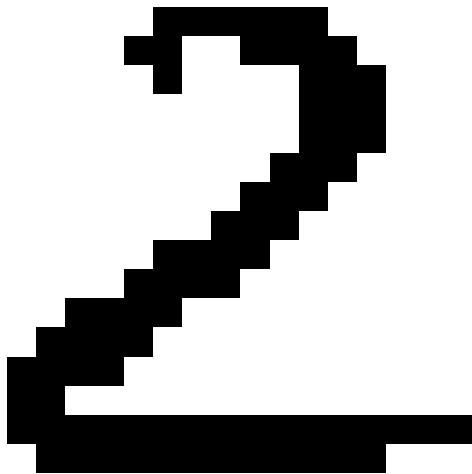


# Advanced Machine Learning

## *Principal Component Analysis (PCA)*

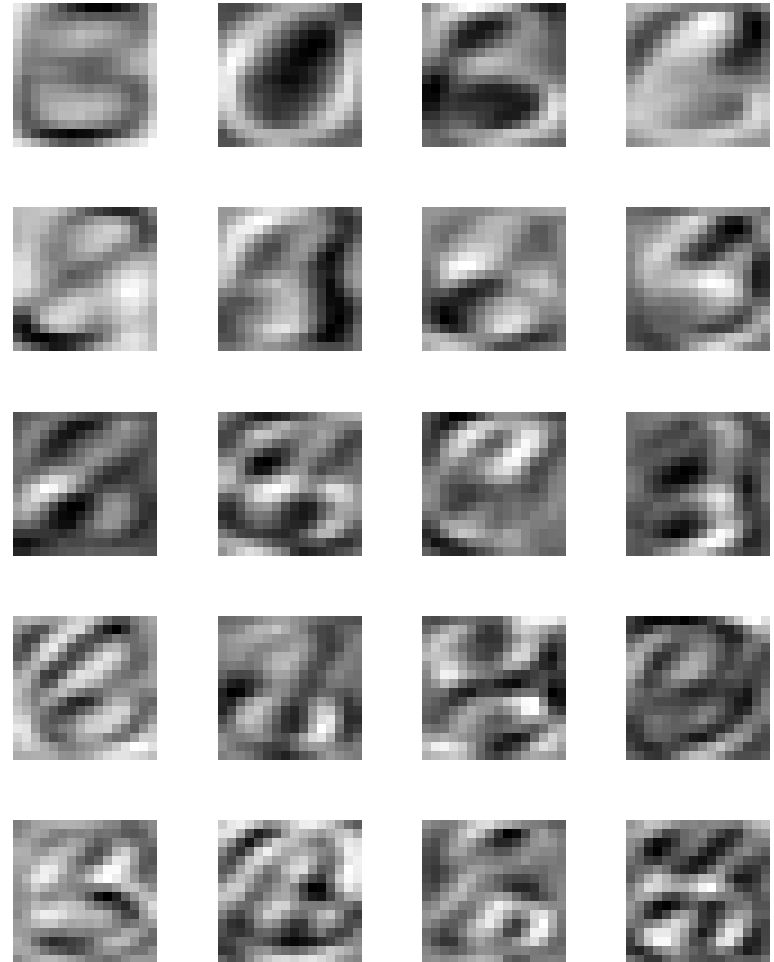
1.6 -1.1 -1.6 2.1 -0.52 2.8 0.72 0.7 -0.68 -0.41 -1.4 -1.5 -0.54 -0.62 1.3 -1.4 -0.27 0.74 0.77 -1



*Covariance matrices, dimensionality reduction, PCA, Duality*

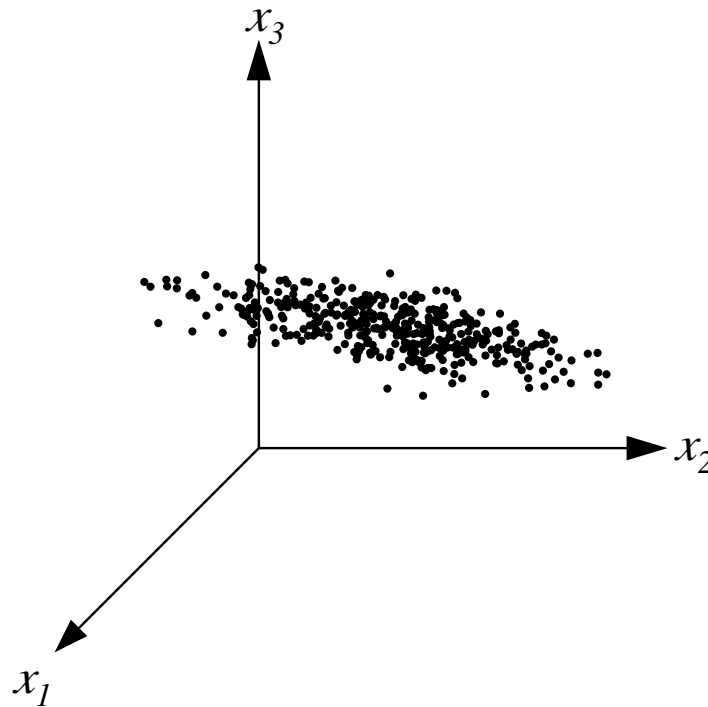
# Outline

1. **Covariance Matrices**
2. Principal Component Analysis
3. Duality



# Spread of Data

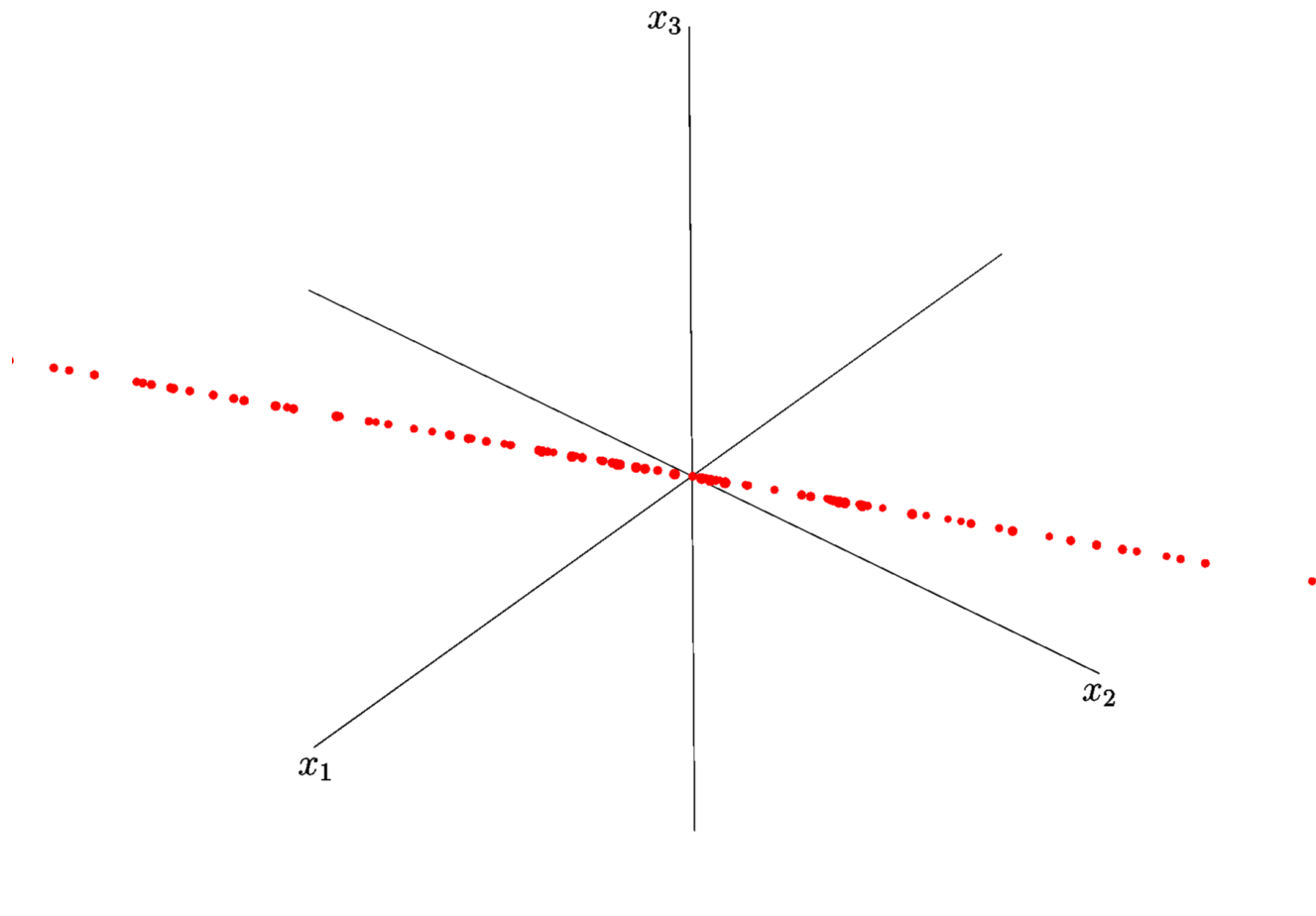
- Often data varies significantly in only some directions



- Reduce dimensions by projecting onto low dimensional subspace with maximum variation

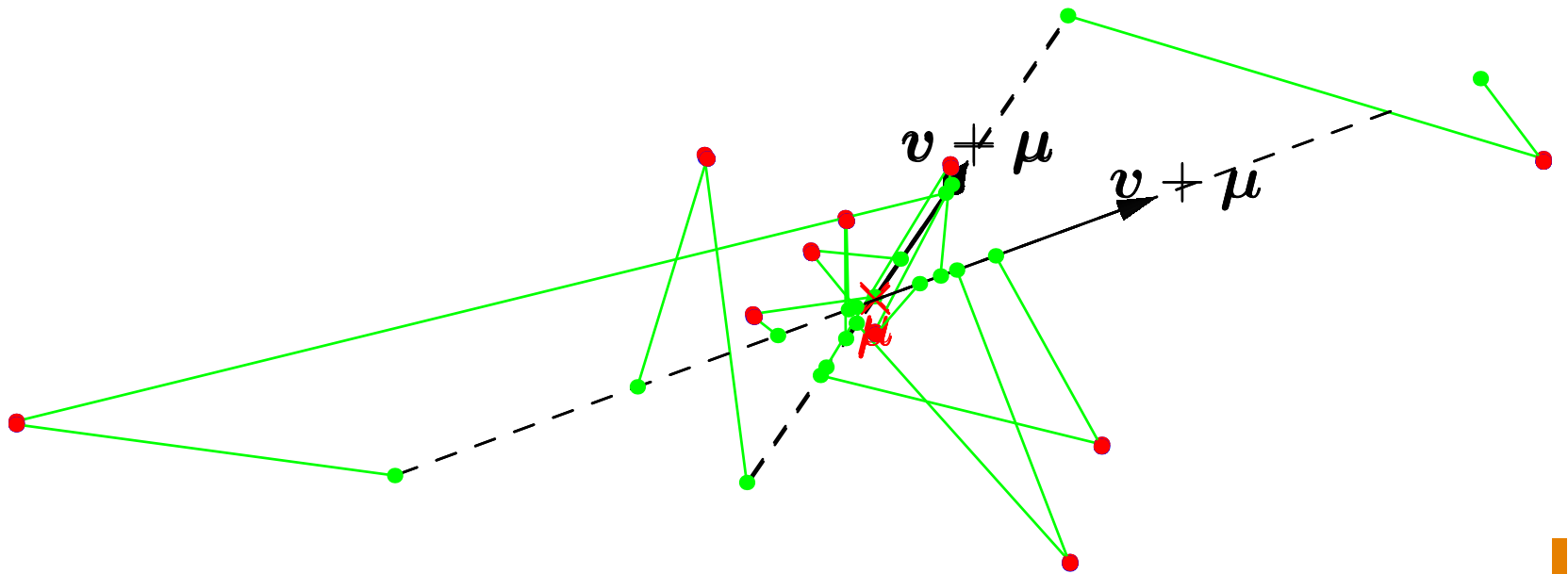
# Looking is not Enough

Can't spot low dimensional data by looking at numbers



# Dimensionality Reduction

- Often helpful to consider only directions where data varies significantly
- Want to find directions along which data has its greatest variation



# Direction of Maximum Variation

- Look for the vector  $\mathbf{v}$  with  $\|\mathbf{v}\|^2 = 1$  to maximise

$$\sigma^2 = \frac{1}{m-1} \sum_{i=1}^m (\mathbf{v}^\top (\mathbf{x}_i - \boldsymbol{\mu}))^2$$

- This is a constrained optimisation problem
- Solve by maximising Lagrangian

$$\mathcal{L} = \frac{1}{m-1} \sum_{k=1}^m (\mathbf{v}^\top (\mathbf{x}_k - \boldsymbol{\mu}))^2 - \lambda (\|\mathbf{v}\|^2 - 1)$$

- $\lambda$  is a Lagrange multiplier

# Direction of Maximum Variation

- Expanding the Lagrangian

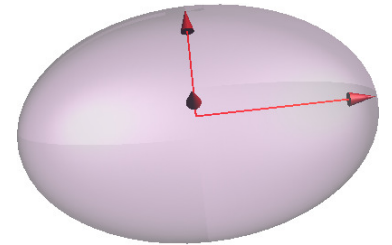
$$\begin{aligned}\mathcal{L} &= \frac{1}{m-1} \sum_{k=1}^m \left( \mathbf{v}^\top (\mathbf{x}_k - \boldsymbol{\mu}) \right)^2 - \lambda (\|\mathbf{v}\|^2 - 1) \blacksquare \\ &= \frac{1}{m-1} \sum_{k=1}^m \left( \mathbf{v}^\top (\mathbf{x}_k - \boldsymbol{\mu}) (\mathbf{x}_k - \boldsymbol{\mu})^\top \mathbf{v} \right) - \lambda (\|\mathbf{v}\|^2 - 1) \blacksquare \\ &= \mathbf{v}^\top \left( \frac{1}{m-1} \sum_{k=1}^m (\mathbf{x}_k - \boldsymbol{\mu}) (\mathbf{x}_k - \boldsymbol{\mu})^\top \right) \mathbf{v} - \lambda (\|\mathbf{v}\|^2 - 1) \blacksquare \\ &= \mathbf{v}^\top \mathbf{C} \mathbf{v} - \lambda (\mathbf{v}^\top \mathbf{v} - 1) \blacksquare\end{aligned}$$

- Extrema of the Lagrangian

$$\nabla \mathcal{L} = 2(\mathbf{C} \mathbf{v} - \lambda \mathbf{v}) = \mathbf{0} \blacksquare \quad \Rightarrow \quad \mathbf{C} \mathbf{v} = \lambda \mathbf{v} \blacksquare$$

# Direction of Maximum Variation

- The eigenvectors are directions that are extrema of the variance



- The variance in direction  $\mathbf{v}$  is equal to

$$\begin{aligned}\sigma^2 &= \frac{1}{m-1} \sum_{i=1}^m \left( \mathbf{v}^\top (\mathbf{x}_i - \boldsymbol{\mu}) \right)^2 \\ &= \mathbf{v}^\top \mathbf{C} \mathbf{v} = \lambda \mathbf{v}^\top \mathbf{v} = \lambda\end{aligned}$$

- The variance is maximised by the eigenvector with the maximum eigenvalue



# Covariance Matrix

- The **covariance matrix** is defined as

$$\mathbf{C} = \frac{1}{m-1} \sum_{k=1}^m (\mathbf{x}_k - \boldsymbol{\mu})(\mathbf{x}_k - \boldsymbol{\mu})^\top$$

- The components  $C_{ij}$  measure how the  $i^{th}$  and  $j^{th}$  components co-vary

$$C_{ij} = \frac{1}{m-1} \sum_{k=1}^m (x_{ik} - \mu_i)(x_{jk} - \mu_j)$$

- C.f. covariance of random variables

$$\text{Cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]$$

# Outer Product

- Remember that the outer-product of two vectors is defined as

$$\mathbf{x} \mathbf{y}^T = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} (y_1 \ y_2 \ \cdots \ y_n) = \begin{pmatrix} x_1 y_1 & x_1 y_2 & \cdots & x_1 y_n \\ x_2 y_1 & x_2 y_2 & \cdots & x_2 y_n \\ \vdots & \vdots & \ddots & \vdots \\ x_n y_1 & x_n y_2 & \cdots & x_n y_n \end{pmatrix}$$

- C.f. Inner product

$$\mathbf{x}^T \mathbf{y} = (x_1 \ x_2 \ \cdots \ x_n) \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = x_1 y_1 + x_2 y_2 + \cdots + x_n y_n$$

# Matrix Form

- The covariance matrix is

$$\mathbf{C} = \frac{1}{m-1} \sum_{k=1}^m (\mathbf{x}_k - \boldsymbol{\mu})(\mathbf{x}_k - \boldsymbol{\mu})^T$$

- Define the matrix

$$\mathbf{X} = \frac{1}{\sqrt{m-1}} (\mathbf{x}_1 - \boldsymbol{\mu}, \mathbf{x}_2 - \boldsymbol{\mu}, \dots, \mathbf{x}_m - \boldsymbol{\mu})$$

- We can write the covariance matrix as

$$\mathbf{C} = \mathbf{X}\mathbf{X}^T$$

# Properties of Covariance Matrix

- The **quadratic form** of a vector and matrix is defined as

$$\mathbf{v}^T \mathbf{M} \mathbf{v}$$

- The quadratic form of a covariance matrix is non-negative for any vector

$$\mathbf{v}^T \mathbf{C} \mathbf{v} = \mathbf{v}^T \mathbf{X} \mathbf{X}^T \mathbf{v} = \mathbf{u}^T \mathbf{u} = \|\mathbf{u}\|^2 \geq 0$$

where  $\mathbf{u} = \mathbf{X}^T \mathbf{v}$

- Matrices with non-negative quadratic forms are known as **positive semi-definite**

# Eigenvalue Decomposition

- The eigenvectors of  $\mathbf{C}$  with the largest eigenvalues are known as the **principal components**
- The eigenvalues are all greater than or equal to zero
- Recall an eigenvector  $\mathbf{v}$  satisfies the equation

$$\mathbf{C}\mathbf{v} = \lambda\mathbf{v}$$

- Multiplying both sides by  $\mathbf{v}^T$

$$\mathbf{v}^T\mathbf{C}\mathbf{v} = \lambda\mathbf{v}^T\mathbf{v} = \lambda\|\mathbf{v}\|^2$$

but  $\mathbf{v}^T\mathbf{C}\mathbf{v} \geq 0$  and  $\|\mathbf{v}\|^2 > 0$  so

$$\lambda = \frac{\mathbf{v}^T\mathbf{C}\mathbf{v}}{\|\mathbf{v}\|^2} \geq 0$$

# Surface Defined by Matrix

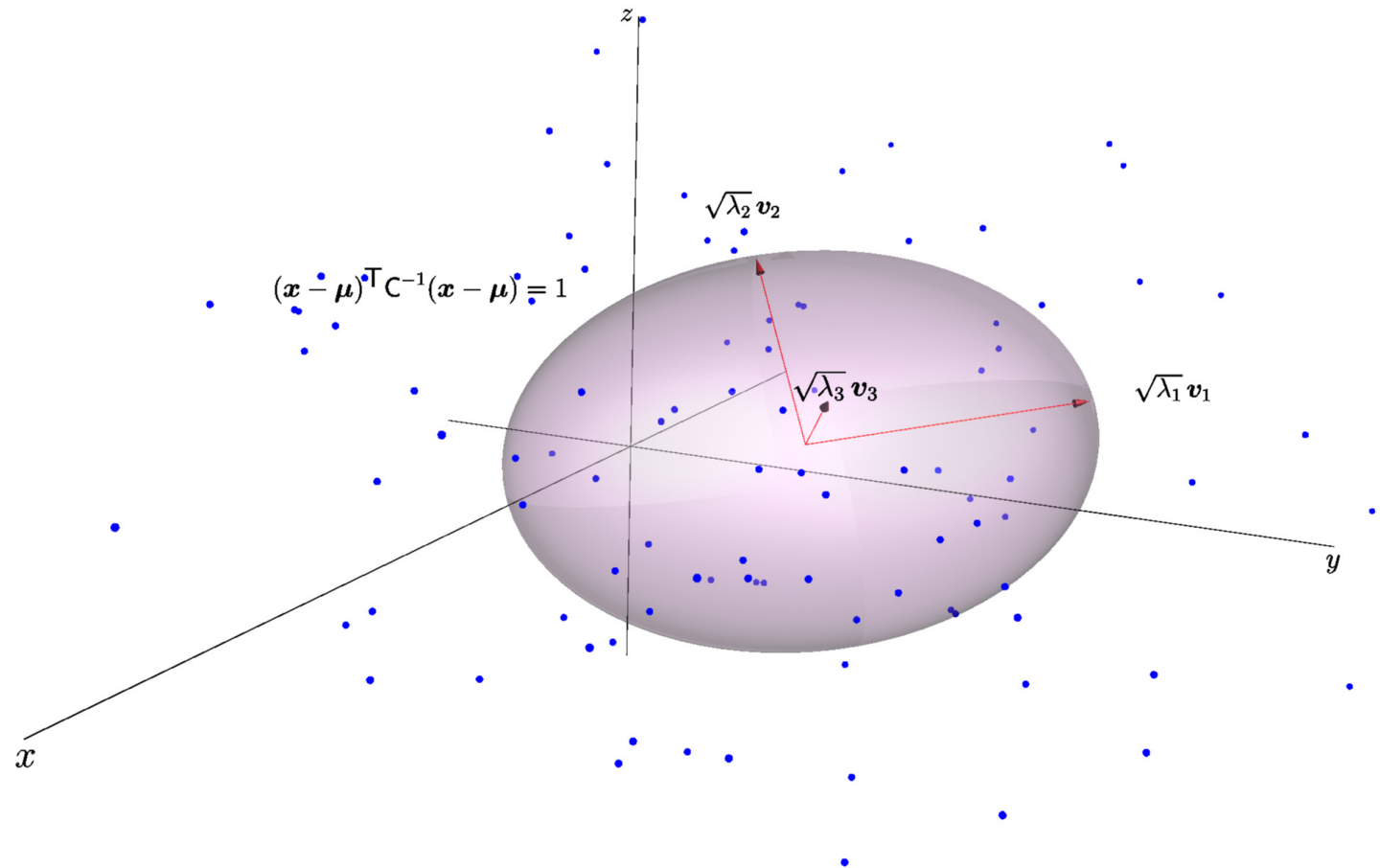
- The set of vectors  $x$  such that

$$x^T C^{-1} x = 1$$

defines a surface■

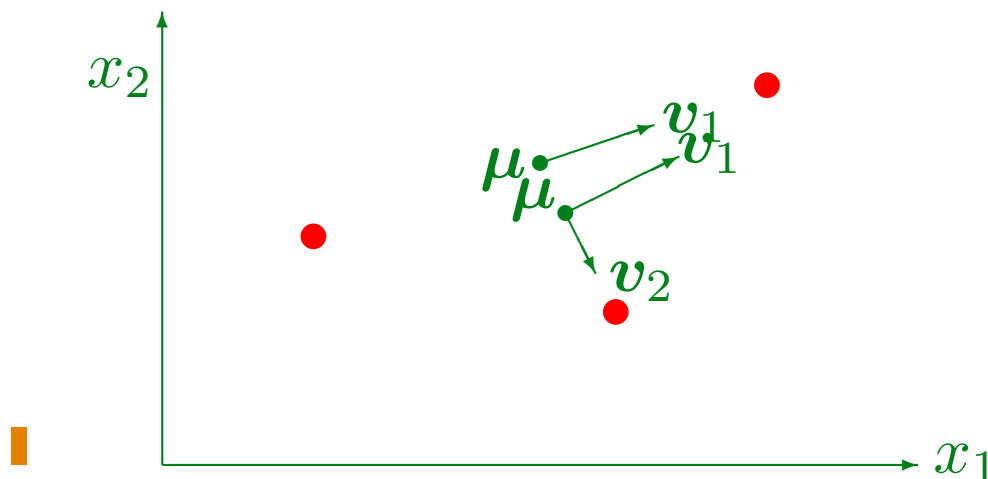
- The surface is an ellipsoid,  $\mathcal{E}$ ■
- The eigenvectors point in the direction of the principal axes of the ellipsoid■
- The radii of the principal axes are equal to the square root of the eigenvalues■

# Ellipsoid and Eigen Space



# Spanning Input Space

- A covariance matrix will have a zero eigenvalue only if there is no variation in the direction of the corresponding eigenvector
- A covariance matrix will have zero eigenvalues if the number of patterns are less than or equal to the number of dimensions
- A covariance matrix formed from  $p + 1$  patterns that are linearly independent (i.e. you cannot form any one out of  $p$  of the other patterns) will have no zero eigenvalues



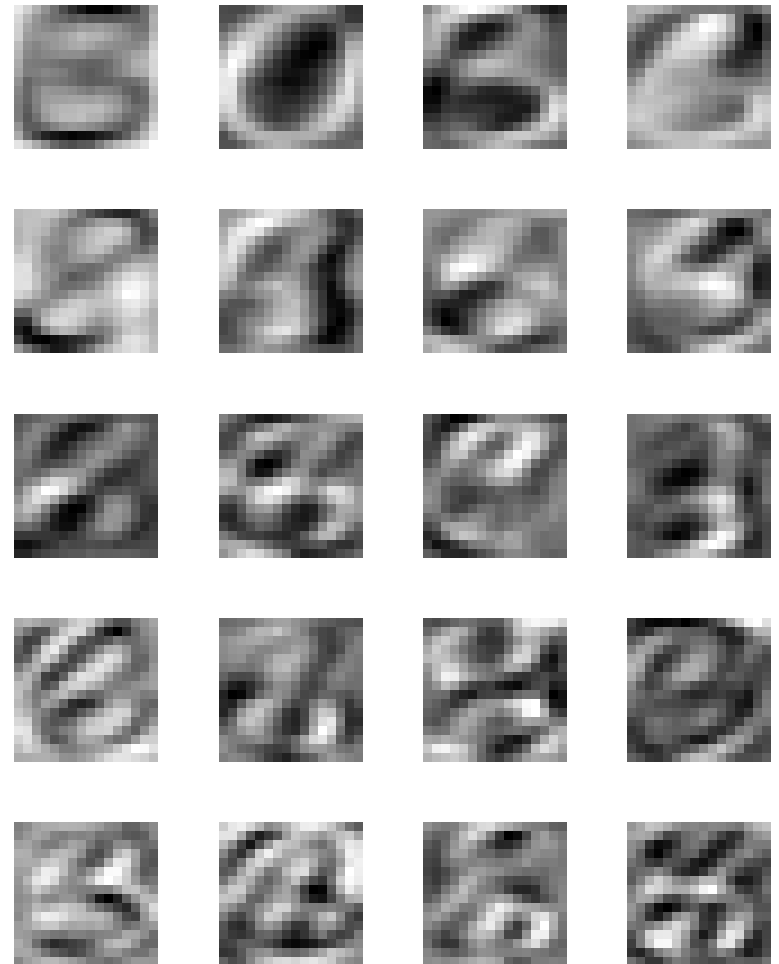


# Positive Definite

- Matrices with no zero eigenvalues are called **full rank** matrices (as opposed to rank deficient)■
- Full rank matrices are invertible, rank deficient matrices are singular and non-invertible■
- Full rank covariance matrices have positive eigenvalues only and are said to be **positive definite**■
- We would expect that when  $m > p$  the covariance matrix will be positive definite■ unless there are some symmetries that linearly constrain the patterns■

# Outline

1. Covariance Matrices
2. **Principal Component Analysis**
3. Duality



# Principal Component Analysis

- PCA occurs as follows■
  - ★ Construct the covariance matrix■
  - ★ Find the eigenvalues and eigenvectors■
  - ★ Keep the eigenvectors with the largest eigenvalues (principal components)■
  - ★ Project the inputs into the space spanned by the principal components ■
- We then use the projected inputs as inputs to our learning machine■

# Projection Matrix

- To project the inputs construct the projection matrix

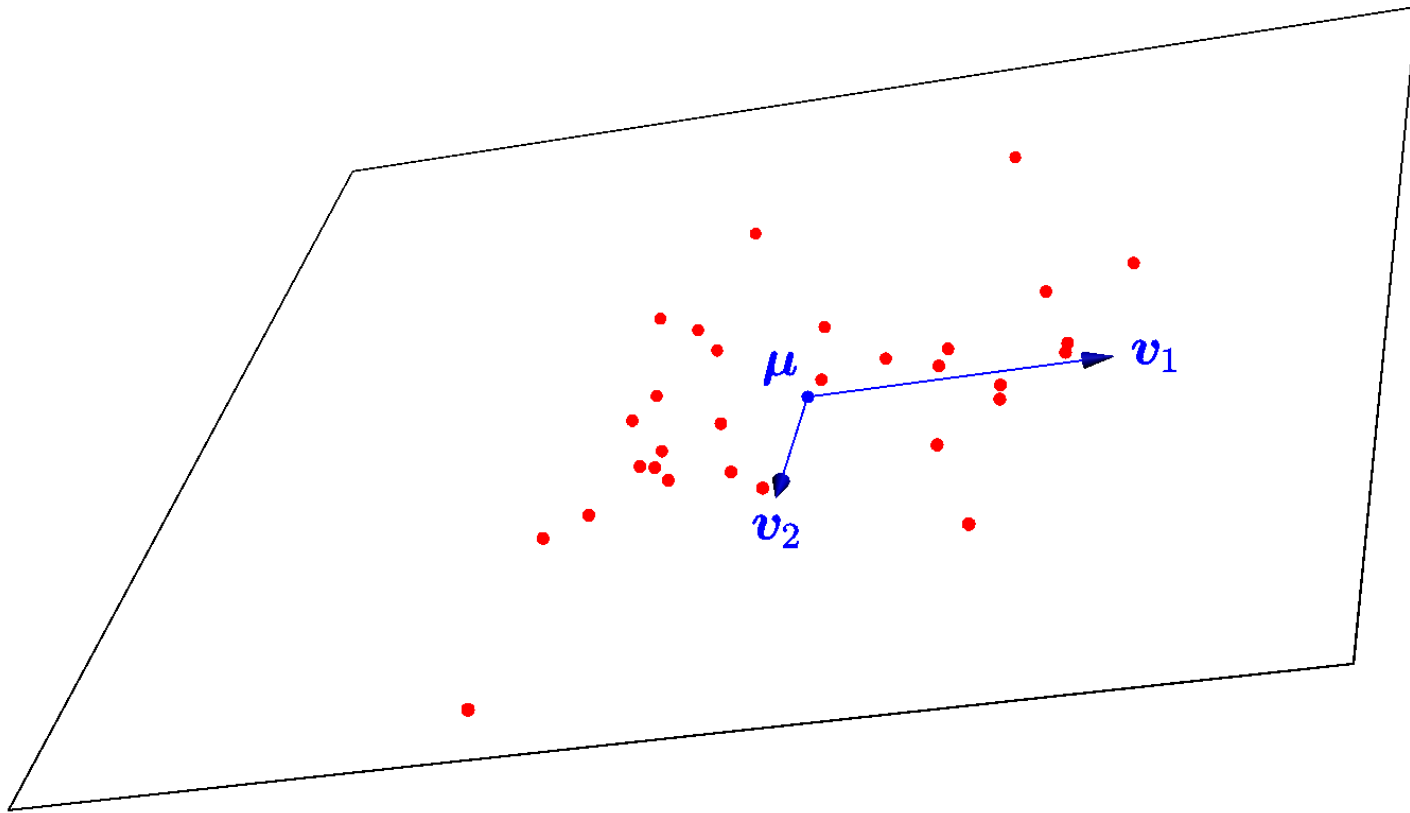
$$\mathbf{P} = \begin{pmatrix} \mathbf{v}_1^T \\ \mathbf{v}_2^T \\ \vdots \\ \mathbf{v}_k^T \end{pmatrix}$$

- $k < p$  is the number of principal components we keep
- Given a  $p$ -dimensional input pattern  $\mathbf{x}$  we can construct a  $k$ -dimensional representation  $\mathbf{z}$

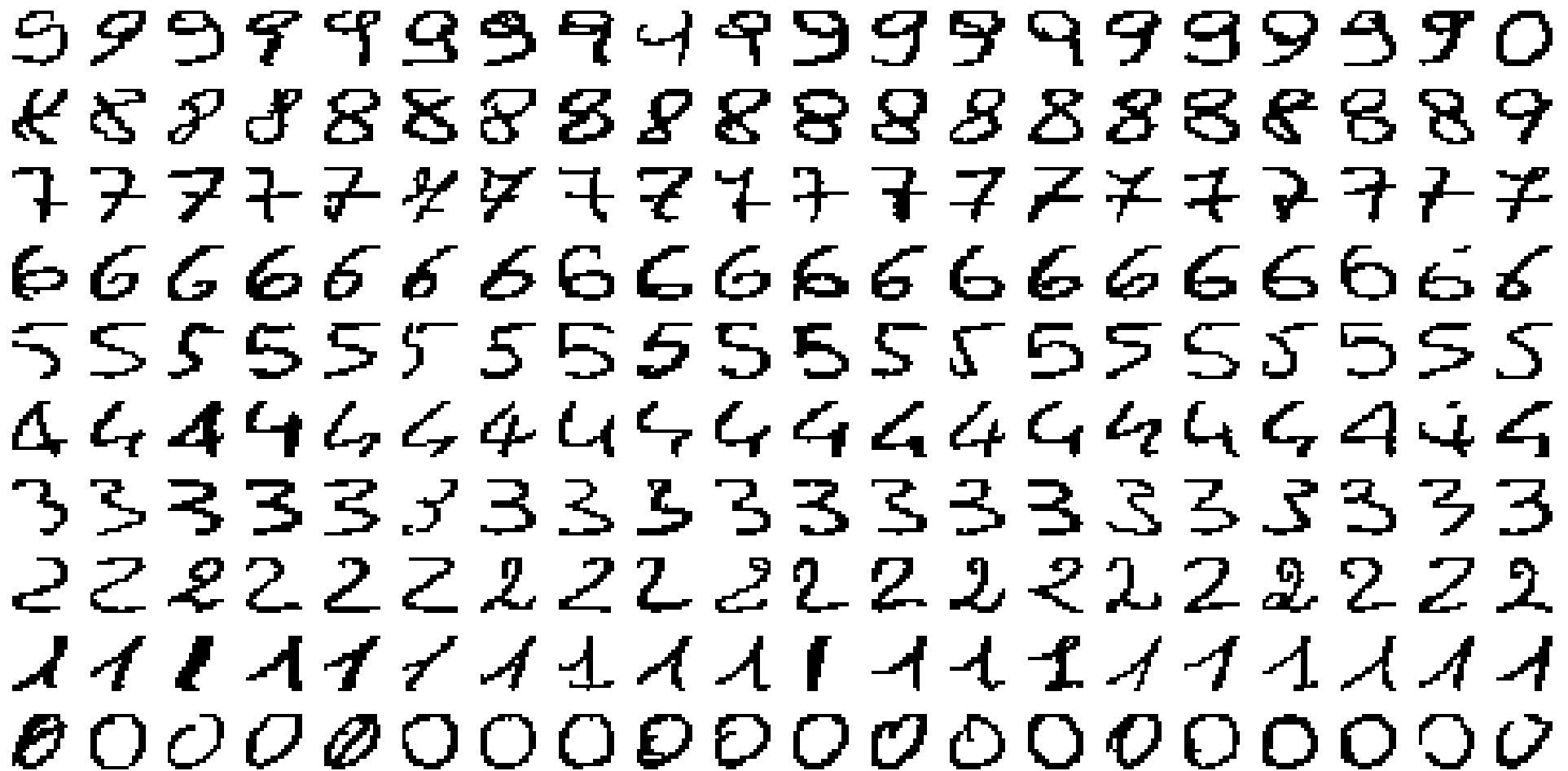
$$\mathbf{z} = \mathbf{P}(\mathbf{x} - \boldsymbol{\mu})$$

- Use  $\mathbf{z}$  as our new inputs

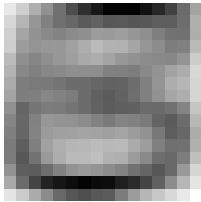
# Subspace Projection



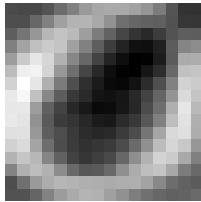
# Hand Written Digits



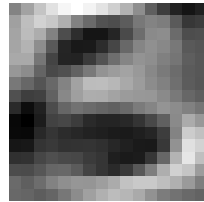
# Eigenvectors



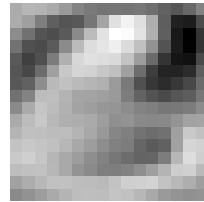
$\mu$



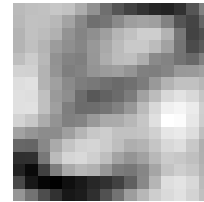
$v_1$



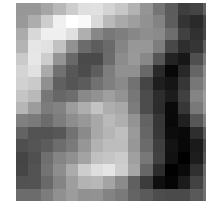
$v_2$



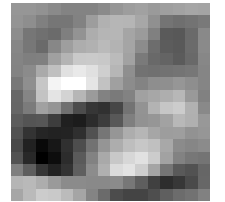
$v_3$



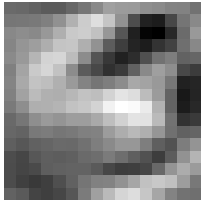
$v_4$



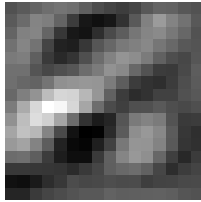
$v_5$



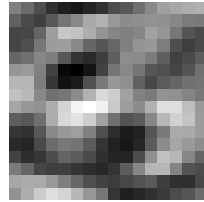
$v_6$



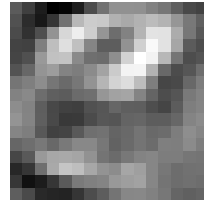
$v_7$



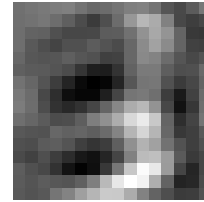
$v_8$



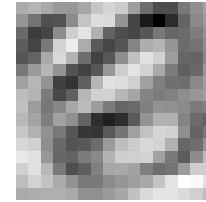
$v_9$



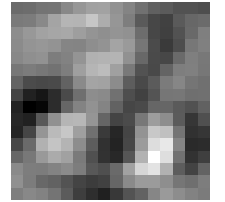
$v_{10}$



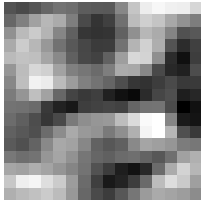
$v_{11}$



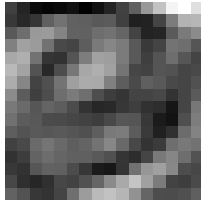
$v_{12}$



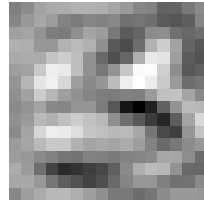
$v_{13}$



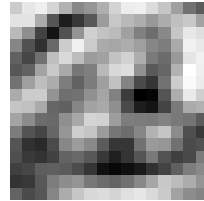
$v_{14}$



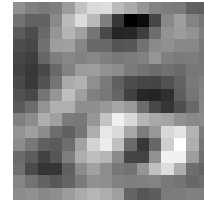
$v_{15}$



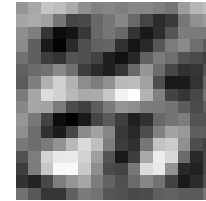
$v_{16}$



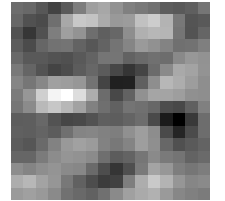
$v_{17}$



$v_{18}$



$v_{19}$



$v_{20}$

# Reconstruction

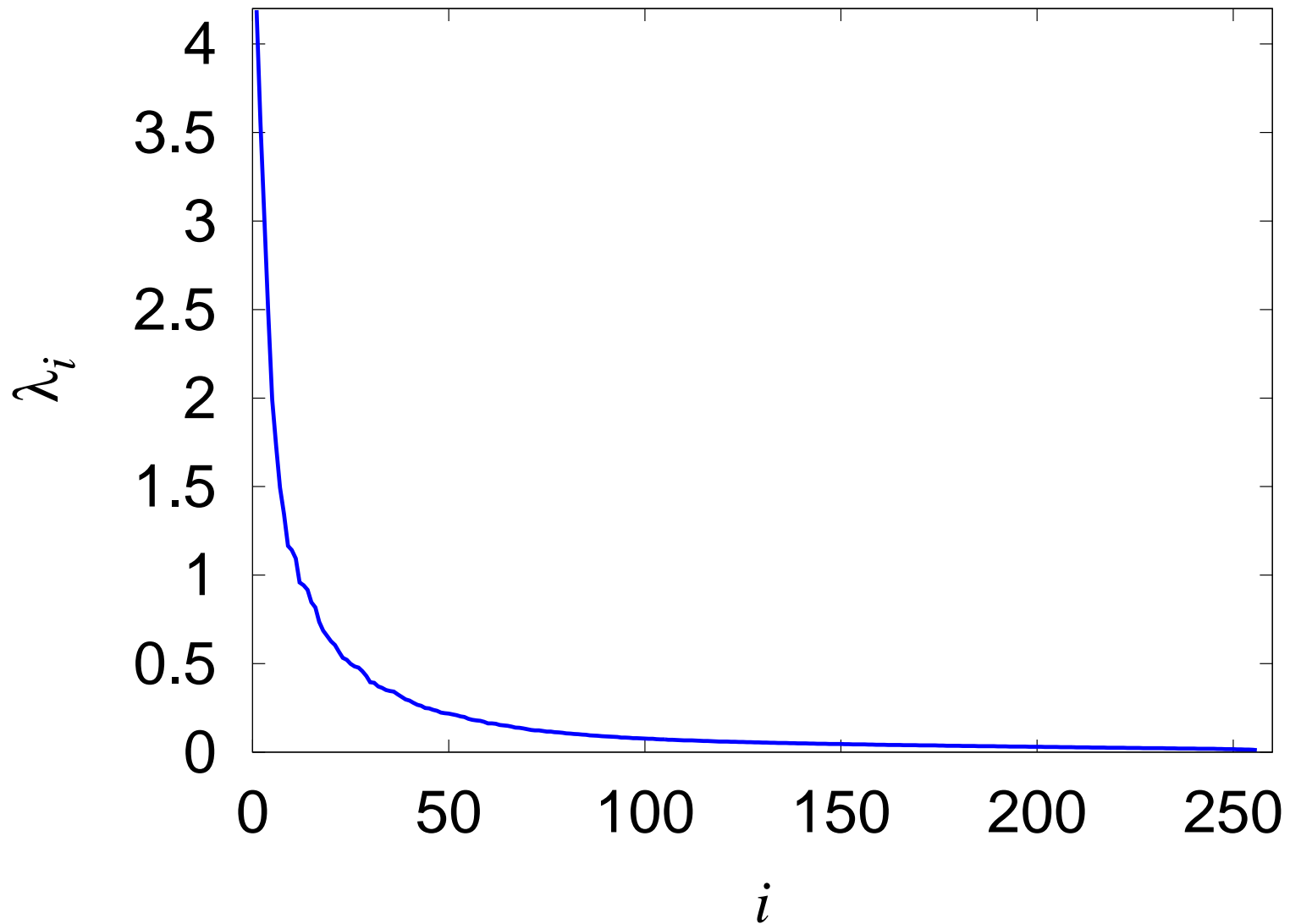
- Projecting into a subspace of eigenvectors can be seen as approximating the inputs by

$$\hat{\mathbf{x}}_i = \boldsymbol{\mu} + \sum_{j=1}^k z_j^i \mathbf{v}_j, \quad z_j^i = \mathbf{v}_j^\top (\mathbf{x}_i - \boldsymbol{\mu}), \quad \|\mathbf{v}_j\| = 1$$

- Principle component analysis projects the data into a subspace of size  $m$  with the minimal approximation error  $\mathbb{E}[\|\hat{\mathbf{x}}_i - \mathbf{x}_i\|^2]$  ■
- The loss of “energy” (or squared error) is equal to the sum of the eigenvalues in the directions that are ignored ■

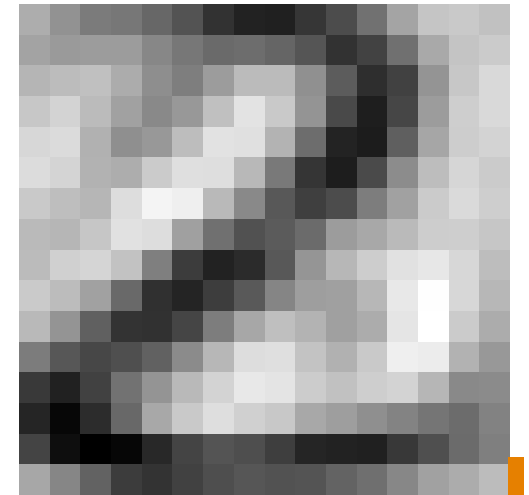
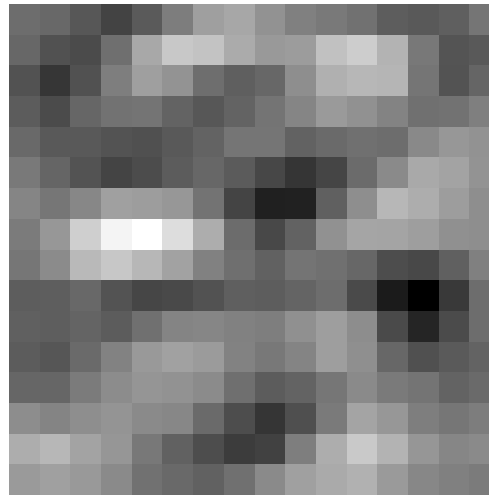
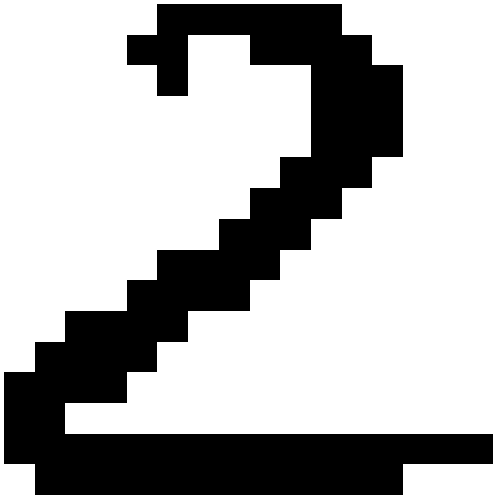


# Eigenvalues for Digits



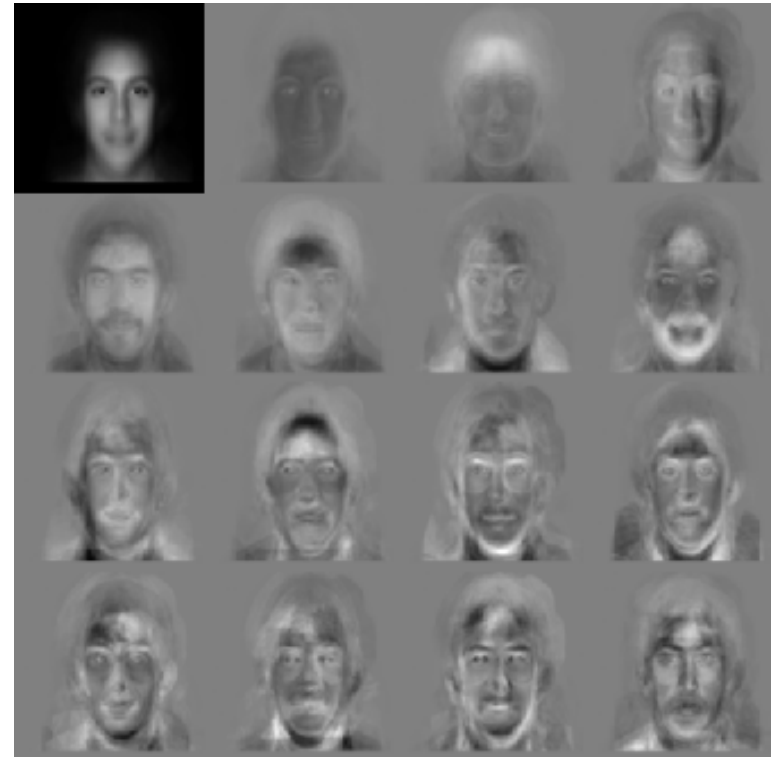
# Reconstruction from Eigenvectors

1.6 -1.1 -1.6 2.1 -0.52 2.8 0.72 0.7 -0.68 -0.41 -1.4 -1.5 -0.54 -0.62 1.3 -1.4 -0.27 0.74 0.77 -1



# Outline

1. Covariance Matrices
2. Principal Component Analysis
3. **Duality**



# PCA for Images

- An image often contains around  $p = 256 \times 256 = 64k$  pixels■
- In standard PCA we would create an  $p \times p$  matrix with over  $4 \times 10^9$  elements■
- This is intractable■
- $m$  images span at most a  $m - 1$  dimensional subspace■
- Usually this subspace will be much smaller than the space of all images  $m \ll p$ ■

# Dual Matrix

- The covariance  $\mathbf{C} = \mathbf{X}\mathbf{X}^\top$  is a  $p \times p$  matrix
- Consider the  $m \times m$  matrix  $\mathbf{D} = \mathbf{X}^\top \mathbf{X}$
- Suppose  $\mathbf{v}$  is an eigenvector of  $\mathbf{D}$

$$\mathbf{D}\mathbf{v} = \lambda\mathbf{v}$$

$$\mathbf{X}^\top \mathbf{X}\mathbf{v} = \lambda\mathbf{v}$$

$$\mathbf{X}\mathbf{X}^\top \mathbf{X}\mathbf{v} = \lambda\mathbf{X}\mathbf{v}$$

$$\mathbf{C}\mathbf{X}\mathbf{v} = \lambda\mathbf{X}\mathbf{v} \Rightarrow \mathbf{C}\mathbf{u} = \lambda\mathbf{u}$$

- $\mathbf{u} = \mathbf{X}\mathbf{v}$  (and  $\mathbf{v} \propto \mathbf{X}^\top \mathbf{u}$ )

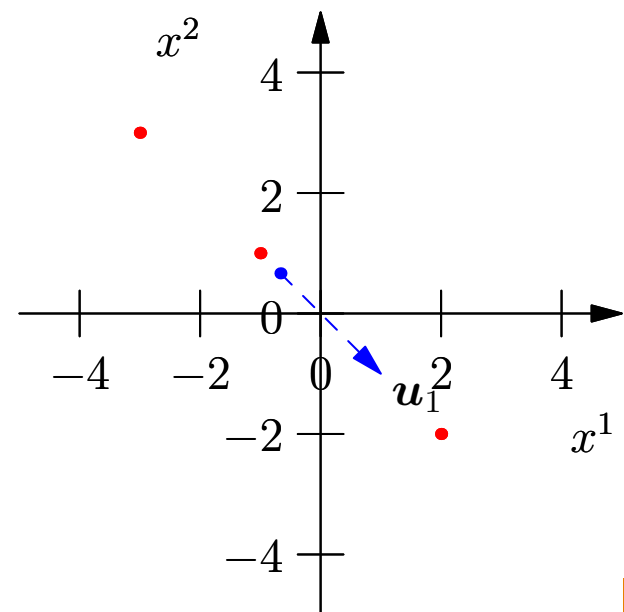
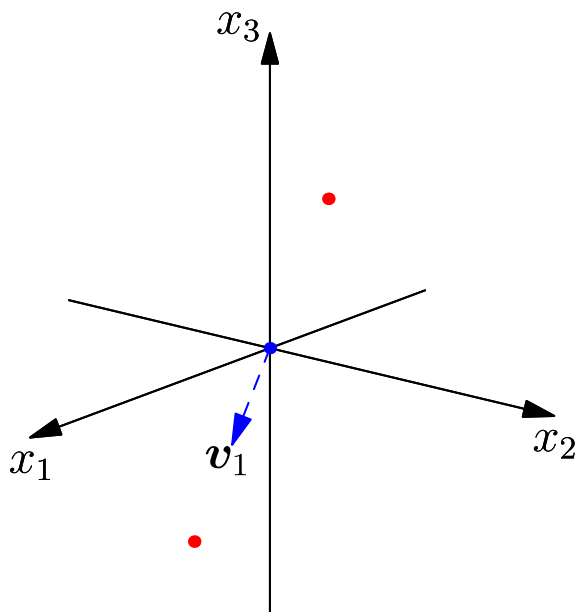
# Dual Matrix

- Matrices  $\mathbf{C} = \mathbf{X}\mathbf{X}^\top$  and  $\mathbf{D} = \mathbf{X}^\top\mathbf{X}$  have the same eigenvalues■
- Can use the dual  $m \times m$  matrix  $\mathbf{D}$  to find eigenvalues and eigenvectors of  $\mathbf{C}$ ■
- Note that  $\mathbf{D} = \mathbf{X}^\top\mathbf{X}$  has components  $D_{kl} \propto (\mathbf{x}_k - \boldsymbol{\mu})^\top(\mathbf{x}_l - \boldsymbol{\mu})$ ■
- Takes  $O(p \times m \times m)$  time to construct  $\mathbf{D}$ ■
- We work in a “dual space” which is the space spanned by the examples■

# What Does a Subspace Look Like?

- Consider  $\mathbf{y}^1 = \begin{pmatrix} 2 \\ 4 \\ 4 \end{pmatrix}$ ,  $\mathbf{y}^2 = \begin{pmatrix} 8 \\ 6 \\ 2 \end{pmatrix}$  with mean  $\boldsymbol{\mu} = \begin{pmatrix} 5 \\ 5 \\ 3 \end{pmatrix}$
- Subtracting the mean  $\mathbf{x}^i = \mathbf{y}^i - \boldsymbol{\mu}$  we can construct matrix

$$\mathbf{X} = \begin{pmatrix} x_1^1 & x_1^2 \\ x_2^1 & x_2^2 \\ x_3^1 & x_3^2 \end{pmatrix} = \begin{pmatrix} -3 & 3 \\ -1 & 1 \\ 2 & -2 \end{pmatrix}$$



# Summary

- PCA allows us to reduce the dimensionality of the inputs■
- We project the inputs into a sub-space where the data varies the most■
- We can work in either the original space ( $\mathbf{X}\mathbf{X}^T$ ) or the dual space ( $\mathbf{X}^T\mathbf{X}$ )■
- When we have many more features than examples (i.e.  $p \gg m$ ) then it is more efficient working in the dual space■
- We will see examples of dual spaces again when we look at SVMs■