Week-1 | Summary

Karthik Thiagarajan

1. Misc 1
    1.1. Notation 1
    1.2. Data-matrix 2
    1.3. Data-point 2
2. Centering the dataset 3
3. Covariance matrix 4
4. Optimization problem 5
5. Principal components 6
6. Projections 8
7. Reconstruction error revisited 9
8. Variance captured 10
9. Compression 11

1. Misc

1.1. Notation

Scalars:

$$x_1, x_2, y_1, y_2, z_2, z_2, a, b, \alpha, \beta$$

Column vector:

$$\mathbf{x} \in \mathbb{R}^d$$

$$\mathbf{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_d \end{bmatrix}$$

Row vector:

$$\mathbf{x} \in \mathbb{R}^d$$

$$\mathbf{x}^T = \begin{bmatrix} x_1 & \cdots & x_d \end{bmatrix}$$

Matrix:

$$\mathbf{X} \in \mathbb{R}^{d \times n}$$

1.2. Data-matrix

$$\mathbf{X} \in \mathbb{R}^{d \times n}$$

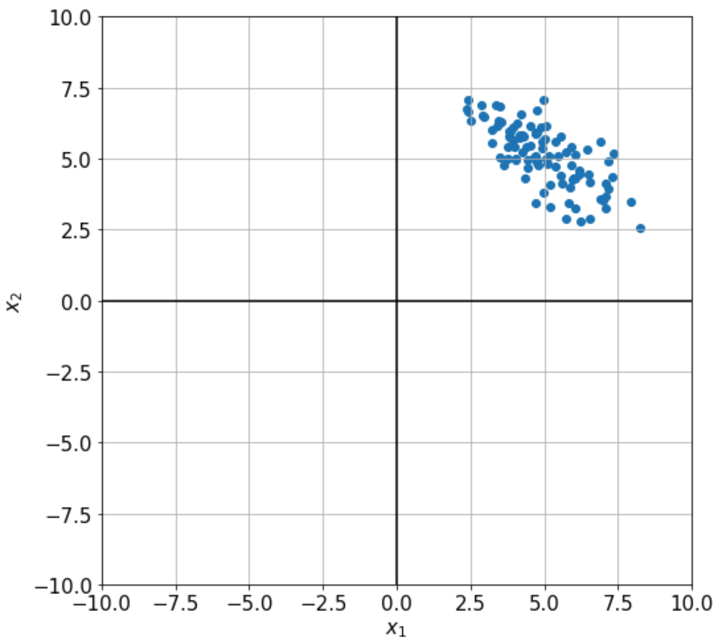
- $d \rightarrow$ number of features
- $n \rightarrow$ number of data-points

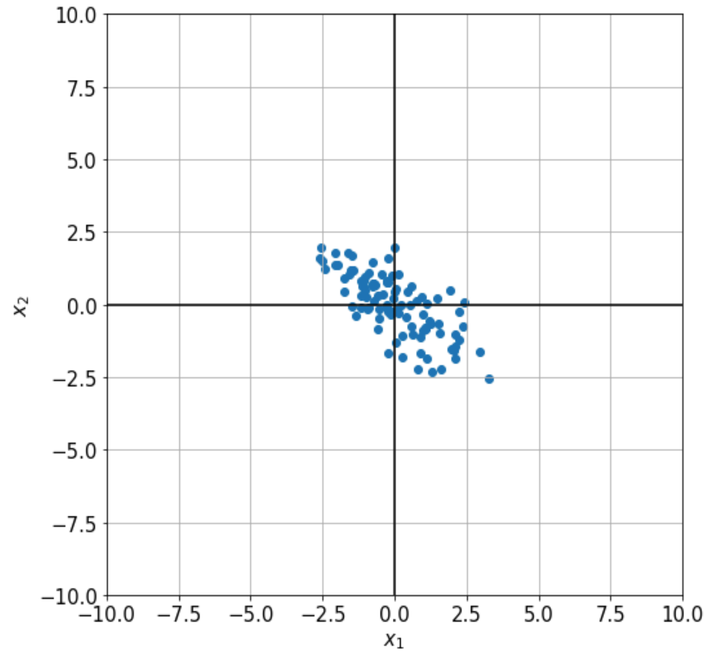
$$\mathbf{X} = \begin{bmatrix} | & & | \\ \mathbf{x}_1 & \cdots & \mathbf{x}_n \\ | & & | \end{bmatrix}$$

1.3. Data-point

$$\mathbf{x}_i \in \mathbb{R}^d$$

2. Centering the dataset





$$\overline{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i$$

If the dataset is already centered, $\overline{\mathbf{x}} = \mathbf{0}$. If $\overline{x} \neq \mathbf{0}$, do the following:

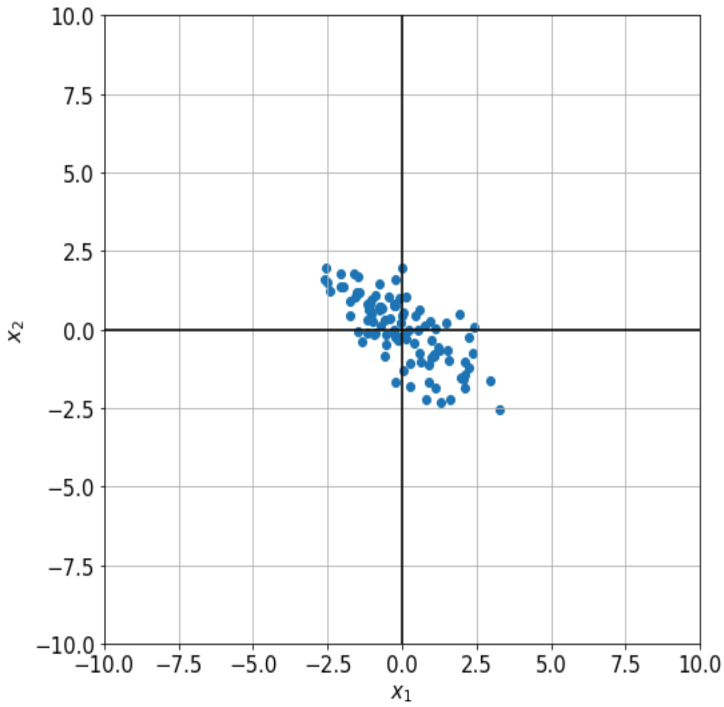
$$\mathbf{x}_i' = \mathbf{x}_i - \overline{\mathbf{x}}$$

$$\mathbf{X}_c = \begin{bmatrix} | & & | \\ \mathbf{x}_1' & \cdots & \mathbf{x}_n' \\ | & & | \end{bmatrix}$$

$\mathbf{X}_c$ is the centered data-matrix.

Remark: From now we will work only with the centered data-matrix and will be calling it $\mathbf{X}$ (the subscript c will be dropped)

3. Covariance matrix



$$\mathbf{C} = \begin{bmatrix} 1 & -0.9 \\ -0.9 & 1 \end{bmatrix}$$

<u>Shape</u>

$$\mathbf{C} \in \mathbb{R}^{d \times d}$$

<u>Outer-product form</u>

$$\mathbf{C} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^T$$

<u>Matrix-form</u>

$$\mathbf{C} = \frac{1}{n} \mathbf{X}\mathbf{X}^T$$

<u>Scalar form</u>

$$C_{pq} = \frac{1}{n} \sum x_{ip} x_{iq}$$

C_{pq} captures the covariance between the p^{th} feature and the q^{th} feature.

As a special case:

$$C_{pp} = \frac{1}{n} \sum x_{ip}^2$$

C_{pp} captures the variance of the p^{th} feature.

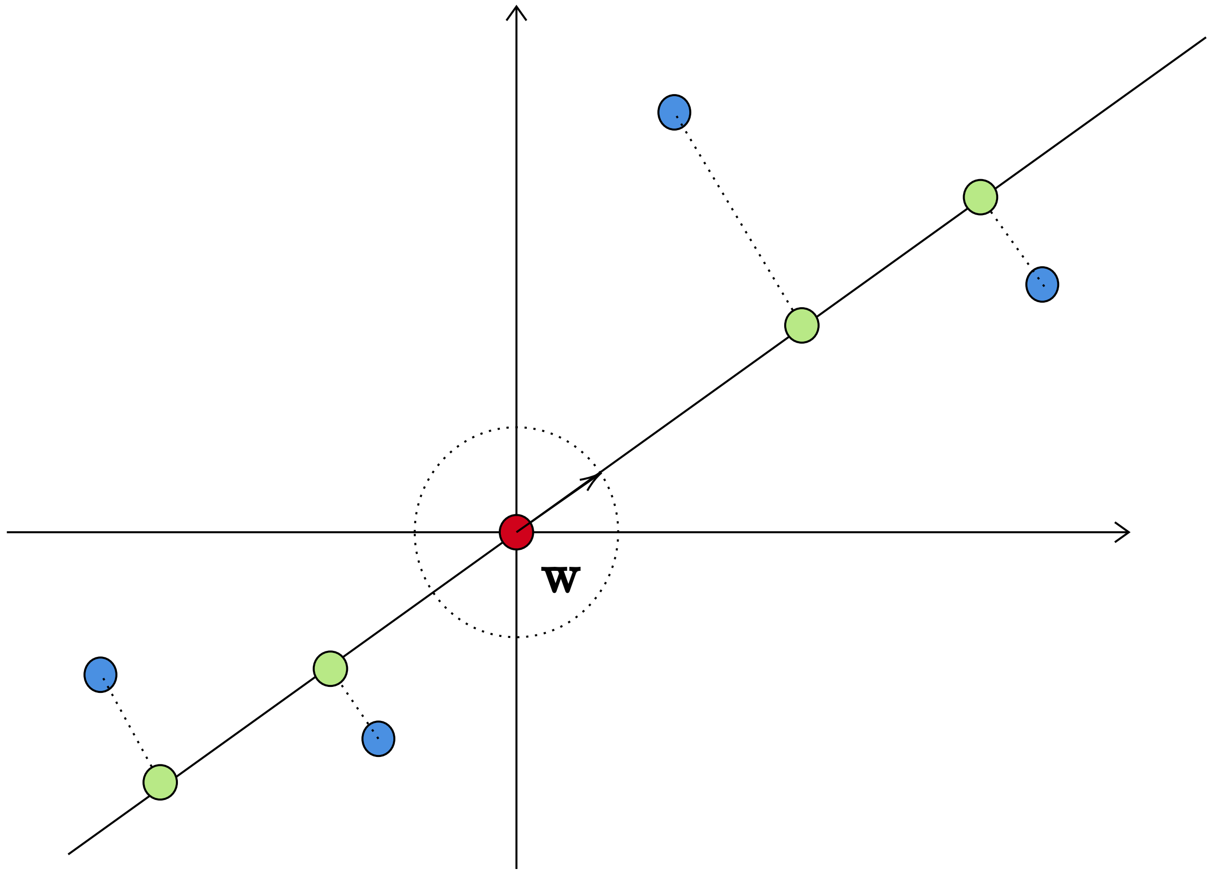
Properties

- $\mathbf{C}^T = \mathbf{C}$
- All eigenvalues of $\mathbf{C}$ are non-negative.
  - $\lambda_1 \geqslant \cdots \geqslant \lambda_d \geqslant 0$
- There is an orthonormal basis for $\mathbb{R}^d$ made up of eigenvectors of $\mathbf{C}$
  - $\{\mathbf{w}_1, \cdots, \mathbf{w}_d\}$
  - This comes from the spectral theorem.

Note: If $\mathbf{C}$ is a square matrix, then $(\lambda, \mathbf{w})$ is said to be an eigenvalue-eigenvector pair if $\mathbf{Cw} = \lambda \mathbf{w}$. Note that $\mathbf{w} \neq \mathbf{0}$ for it to be an eigenvector.

Remark: $\mathbf{w}_i$ will always represent a unit-norm vector in the rest of the document.

4. Optimization problem



Minimizing the reconstruction error

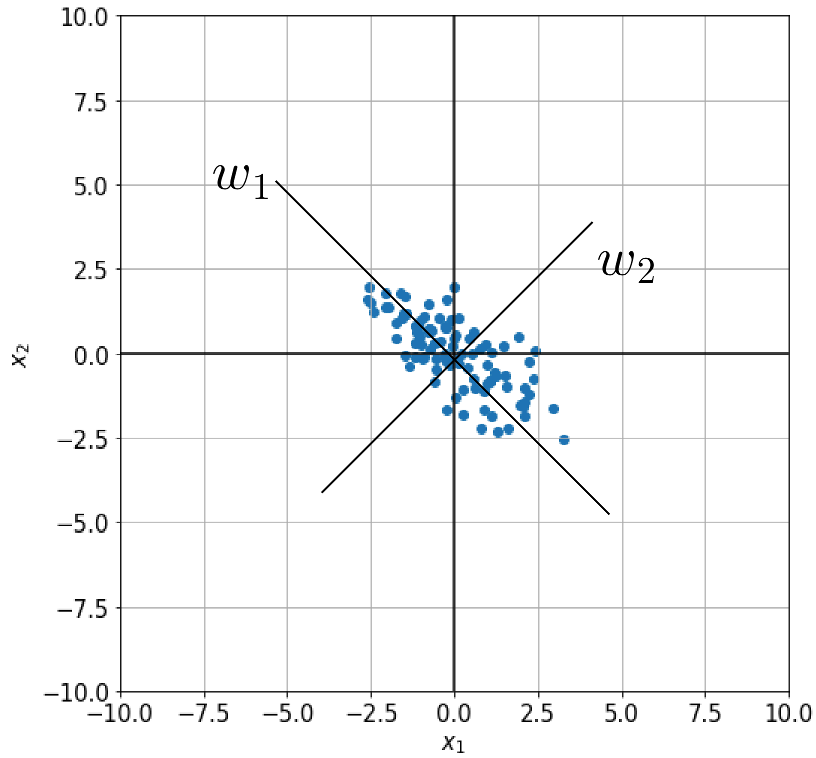
$$\min_{\mathbf{w}} \quad \frac{1}{n}\sum_{i=1}^{n} ||\mathbf{x}_i - \left(\mathbf{x}_i^T\mathbf{w}\right)\mathbf{w}||^2$$

Maximizing the variance

$$\max_{\mathbf{w}} \ \mathbf{w}^T\mathbf{C}\mathbf{w}$$

Both forms are equivalent to each other.

5. Principal components



Let $(\lambda_1, \mathbf{w}_1), \cdots, (\lambda_d, \mathbf{w}_d)$ be the eigen-pairs of $\mathbf{C}$, where $\lambda_1 \geqslant \cdots \geqslant \lambda_d$ and $\{\mathbf{w}_1, \cdots, \mathbf{w}_d\}$ is an orthonormal basis for $\mathbb{R}^d$. $\mathbf{w}_i$ is termed the i^{th} principal component of $\mathbf{C}$. To be more precise:

$$\mathbf{C}\mathbf{w}_i = \lambda_i \mathbf{w}_i$$

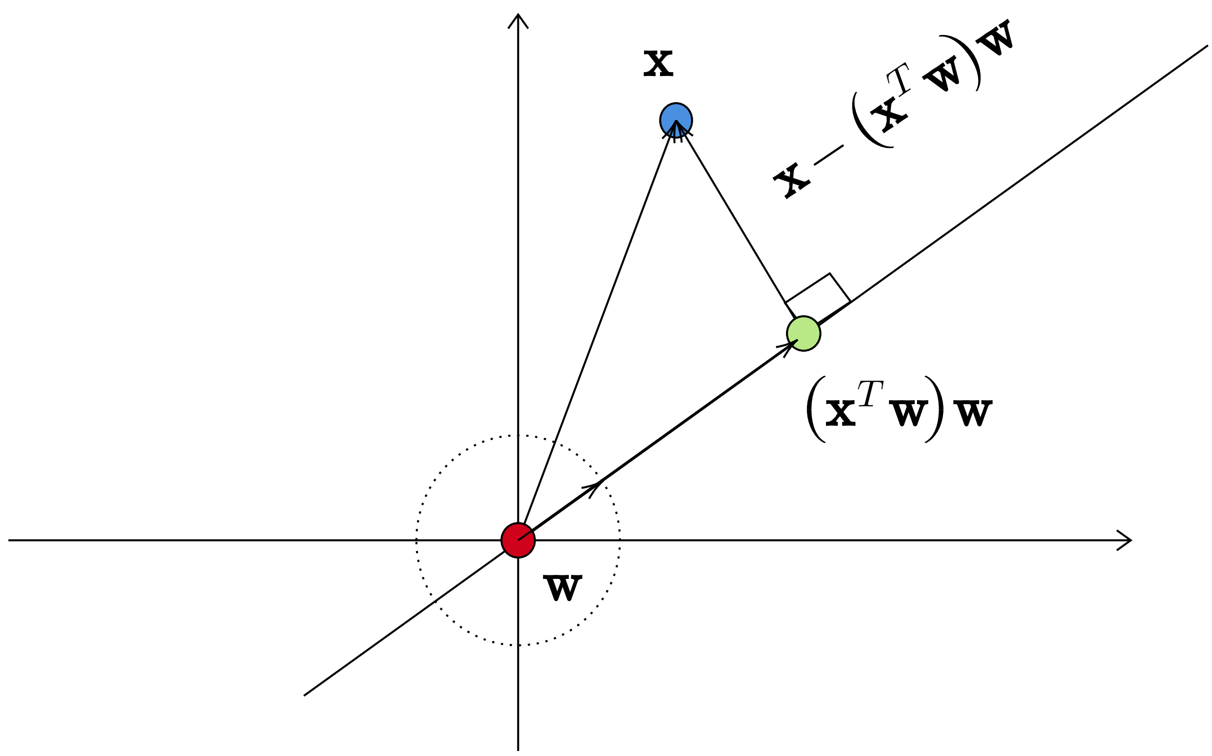
$$\mathbf{w}_i^T \mathbf{w}_j = \begin{cases} 1, & i = j \\ 0, & i \neq j \end{cases}$$

$$\lambda_1 = \max_{\mathbf{w}} \quad \mathbf{w}^T \mathbf{C} \mathbf{w}$$

$$\mathbf{w}_1 = \arg\max_{\mathbf{w}} \quad \mathbf{w}^T \mathbf{C} \mathbf{w}$$

$$\mathbf{w}_1^T \mathbf{C} \mathbf{w}_1 = \lambda_1$$

6. Projections



(Vector) Projection of $\mathbf{x}_i$ onto the j^{th} PC

$$\left(\mathbf{x}_i^T\mathbf{w}_j\right)\mathbf{w}_j$$

Scalar projection of $\mathbf{x}_i$ onto the j^{th} PC (or) coordinate of the data-point along this direction:

$$\mathbf{x}_i^T\mathbf{w}_j$$

The projection of a data-point $\mathbf{x}_i$ onto the top k principal components:

$$\mathbf{x}_i' = \left(\mathbf{x}_i^T\mathbf{w}_1\right)\mathbf{w}_1 + \cdots + \left(\mathbf{x}_i^T\mathbf{w}_k\right)\mathbf{w}_k$$

To represent the reconstruction and scalar projections in matrix form:

$$\mathbf{W} \in \mathbb{R}^{d\times k}$$

$$\mathbf{W} = \begin{bmatrix} | & & | \\ \mathbf{w}_1 & \cdots & \mathbf{w}_k \\ | & & | \end{bmatrix}$$

<u>Scalar projections</u>

$$\mathbf{X}' \in \mathbb{R}^{k \times n}$$

$$\mathbf{X}' = \begin{bmatrix} \mathbf{x}_1^T \mathbf{w}_1 & & \mathbf{x}_n^T \mathbf{w}_k \\ | & \cdots & | \\ \mathbf{x}_1^T \mathbf{w}_k & & \mathbf{x}_n^T \mathbf{w}_k \end{bmatrix}$$

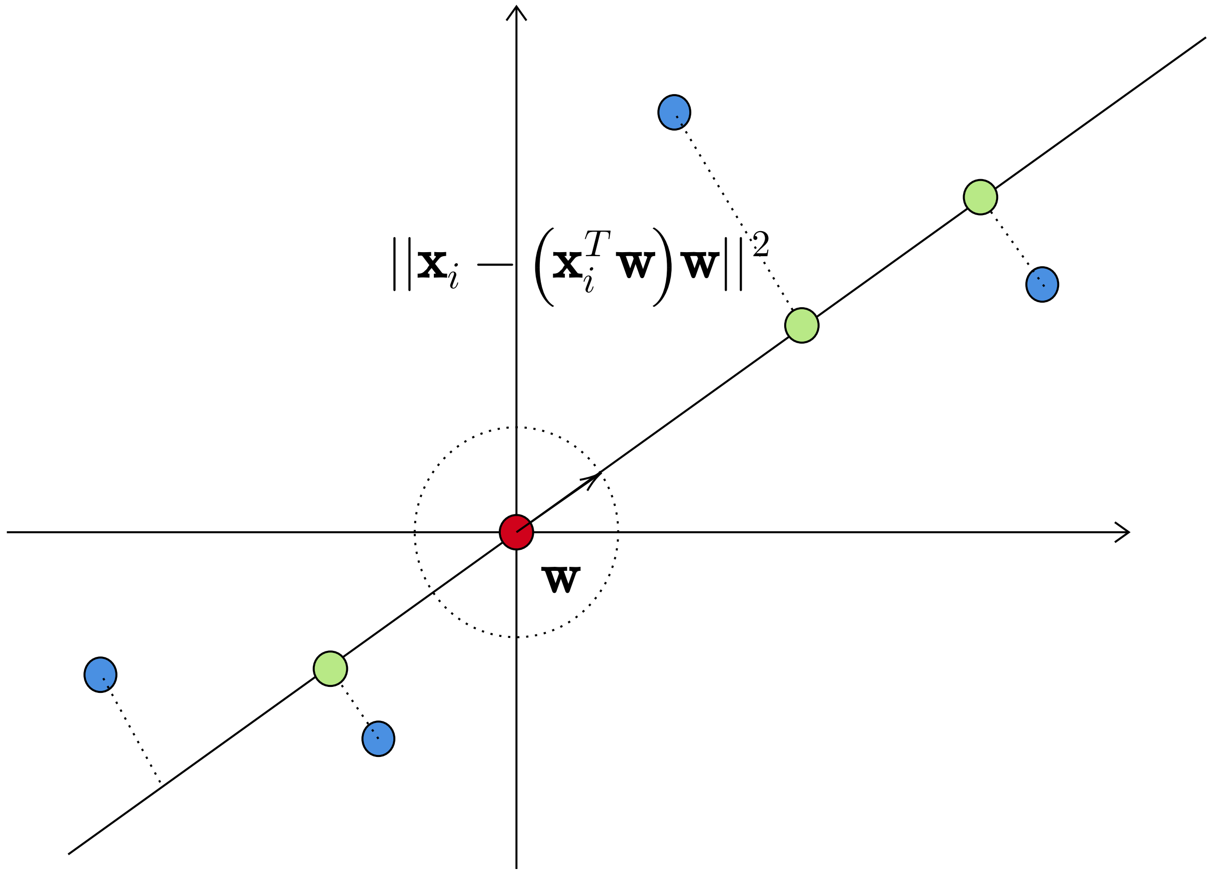
$$\mathbf{X}' = \mathbf{W}^T \mathbf{X}$$

<u>Reconstruction</u>

$$\mathbf{X}' \in \mathbb{R}^{d \times n}$$

$$\mathbf{X}' = \mathbf{W}\mathbf{W}^T \mathbf{X}$$

7. Reconstruction error revisited (for k directions)



$$\frac{1}{n}\sum_{i=1}^{n} ||\mathbf{x}_i - \mathbf{x}_i'||^2$$

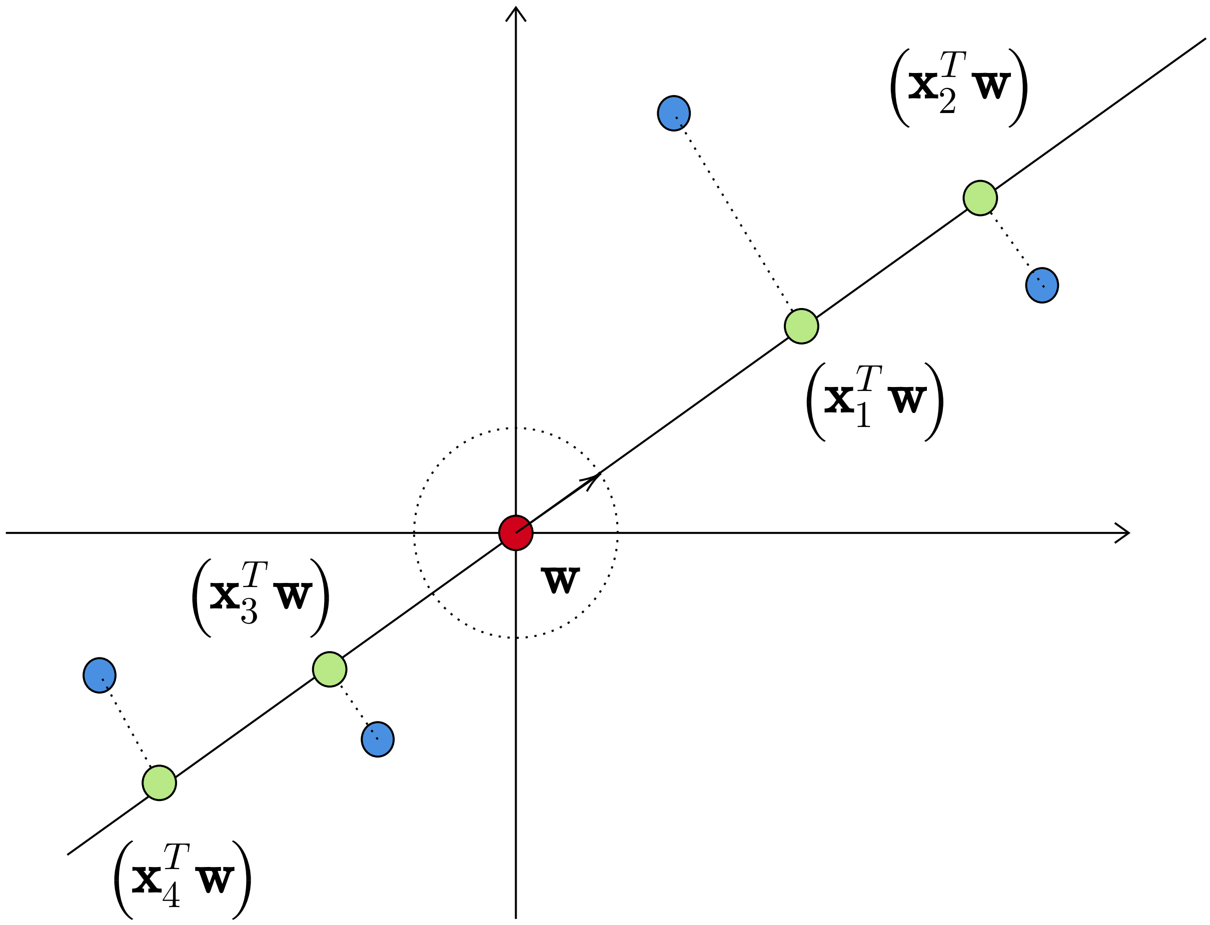
$$\frac{1}{n}\sum_{i=1}^{n} \left\|\mathbf{x}_i - \sum_{j=1}^{k} \left(\mathbf{x}_i^T \mathbf{w}_j\right)\mathbf{w}_j\right\|^2$$

8. Variance captured

Total variance:

$$\lambda_1 + \cdots + \lambda_d$$

Variance along a given direction $\mathbf{w}$ (unit vector):



$$\frac{1}{n}\sum_{i=1}^{n}\left(\mathbf{x}_i^T\mathbf{w}\right)^2$$

$$\mathbf{w}^T\mathbf{C}\mathbf{w}$$

Proportion of variance captured by top k PCs:

$$\frac{\lambda_1 + \cdots + \lambda_k}{\lambda_1 + \cdots + \lambda_d}$$

Heuristic to choose the value of k: smallest value that captures 95% of the variance in the dataset.

9. Compression

Reconstruction

$$\frac{nk+dk}{dn} = \frac{k(d+n)}{dn}$$

Retaining only scalar projections

$$\frac{kn}{dn} = \frac{k}{d}$$