



12.8 Principal component analysis

Douglas Wilhelm Harder, LEL, M.Math.

dwharder@uwaterloo.ca

dwharder@gmail.com





Introduction

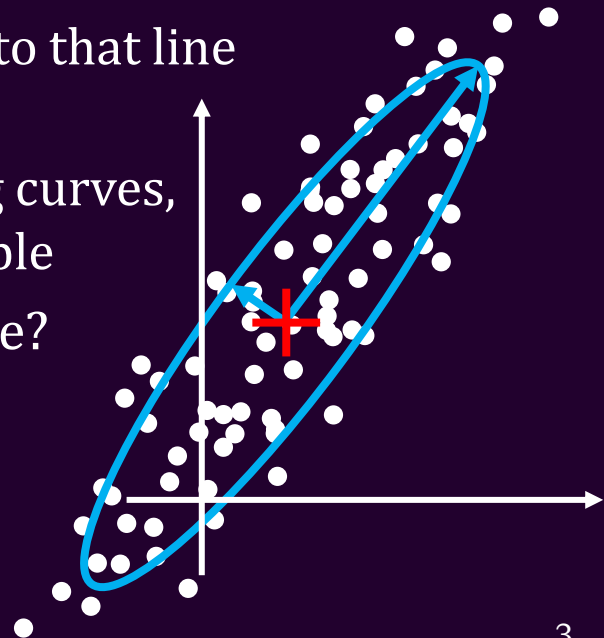
- In this topic, we will
 - Introduce the idea of finding correlations between data
 - Discuss how we can have factors and responses we are measuring through numerous trials
 - Describe how our goal is to find those orthogonal vectors that best describe the data
 - Claim that these can be found simply by finding the eigenvalues and eigenvectors of a positive semi-definite matrix
 - Look at an example





Data analysis

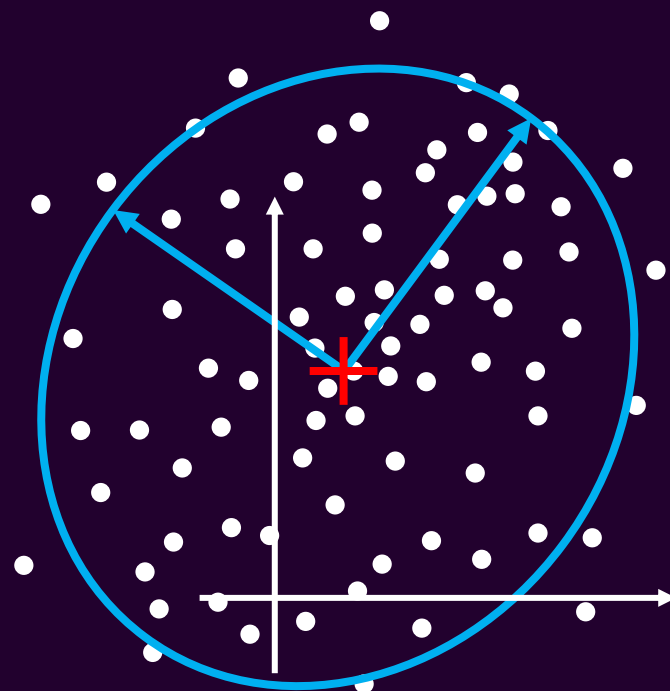
- You are reading two remote sensors
 - You plot one reading versus the other
 - There is obviously a relationship between the two
 - The data has a mean or average
 - Can we describe the relationship between these readings?
 - The total distance from each point to that line is minimized?
 - Recall the least-squares best-fitting curves, but without a dependent variable
 - Can we describe the data by an ellipse?





Data analysis

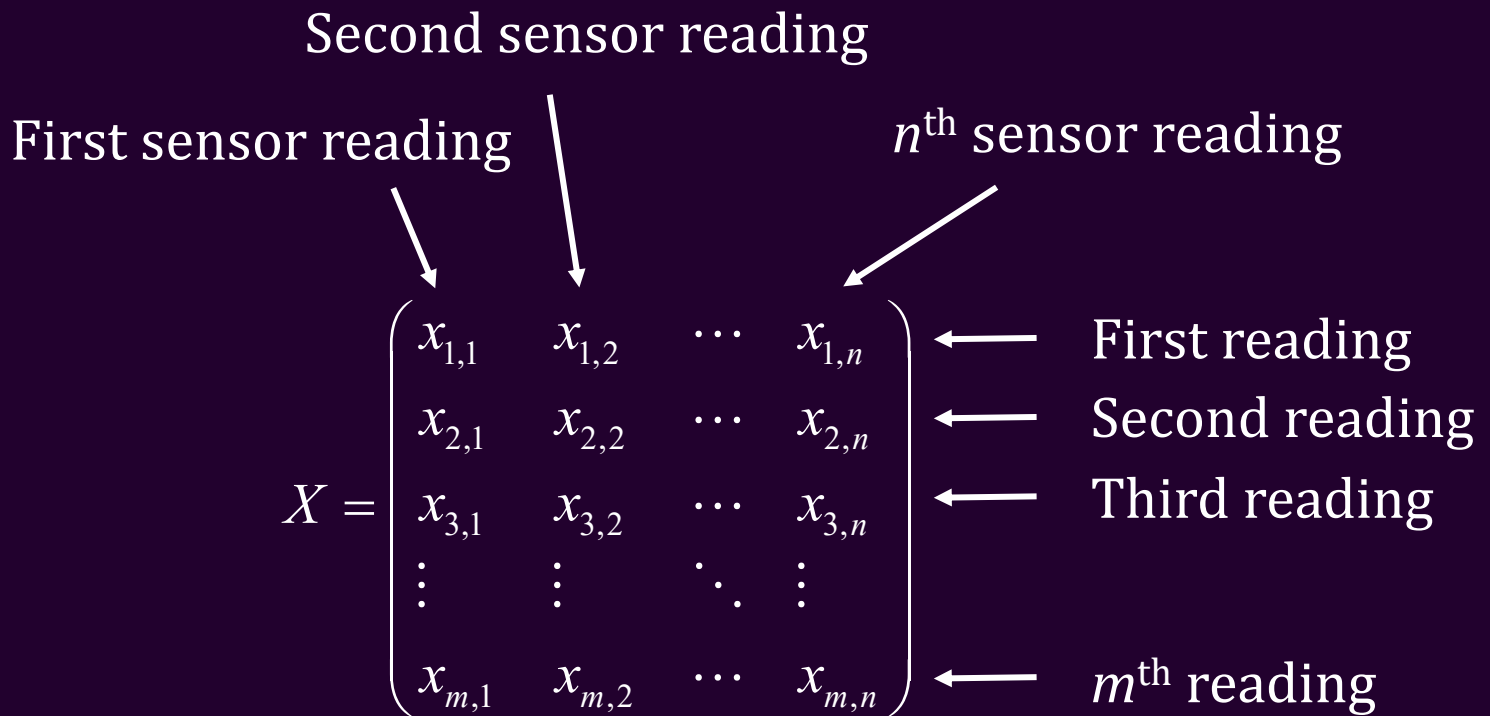
- Suppose we had this data
 - There seems to be a weaker relationship between the readings





Data analysis

- Suppose we have the following set-up:





Data analysis

- Suppose you are developing a polymer blend
 - See *Methodology for Optimization of Polymer Blends Composition* by Alessandra Martins Coelho *et al.*
 - You have created m samples, and with each sample, you measured n different characteristics:
 - Some may be measurements during the combination (factors):
 - The proportion of polymers and other compounds
 - The amount of solvent
 - The time taken
 - The temperature
 - Others may be properties of the resulting product (responses):
 - Modulus of elasticity
 - Tensile strength at break
 - Elongation at rupture
 - Question: are there relationships between these?





Data analysis

- Such analysis of factors and responses has been developing for centuries
 - In the late 1800s, William Gosset writing under the pseudonym “Student” derived the Student’s t -distribution for analyzing variation in factors and responses with few samples
 - He happened to be the Head Brewer and Head Experimental Brewer at Guinness
- The algorithm we are looking at is described as *principal component analysis* and was derived in 1901 by Karl Pearson





Principal component analysis

- Our goal is given data $\mathbf{x}_1, \dots, \mathbf{x}_m$ in n variables, to:
 - First subtract off the mean or average of these vectors
 - Find one vector that defines a line that minimizes the distance of all factors and responses to that line
 - This is called the first principal component
 - Say this vector is $\mathbf{w}_1$
 - Next, find the perpendicular component of each vector to that first vector, so replace $\mathbf{x}_k \leftarrow \text{perp}_{\mathbf{w}_1}(\mathbf{x}_k) = \mathbf{x}_k - \text{proj}_{\mathbf{w}_1}(\mathbf{x}_k)$
 - Having removed this, repeat the process to find the next principal component $\mathbf{w}_2$
 - This vector must be perpendicular to $\mathbf{w}_1$
 - Repeat n times to get an orthogonal basis representing the data





Principal component analysis

- This sounds difficult, but fortunately, it is quite straight-forward
 - Given the vectors $\mathbf{x}_1, \dots, \mathbf{x}_m$, each storing n measurements of factors or responses, find the average:

$$\bar{\mathbf{x}} = \frac{1}{m} \sum_{k=1}^m \mathbf{x}_k$$

- Define the matrix

$$X = (\mathbf{x}_1 - \bar{\mathbf{x}} \quad \mathbf{x}_2 - \bar{\mathbf{x}} \quad \cdots \quad \mathbf{x}_m - \bar{\mathbf{x}})^T$$





Covariance matrix

- This sounds difficult, but fortunately, it is quite straight-forward
 - The covariance matrix is the $n \times n$ matrix $C = \frac{1}{m} X^T X$
 - Specifically, the $(i, j)^{\text{th}}$ entry is

$$\begin{aligned} c_{i,j} &= \frac{1}{m} \sum_{k=1}^m \left((\mathbf{x}_k)_i - \bar{x}_i \right) \left((\mathbf{x}_k)_j - \bar{x}_j \right) \\ &= \frac{1}{m} \sum_{k=1}^m x_{k,i} x_{k,j} \end{aligned}$$

- Eigenvectors don't change under scalar multiplication, so we will deal with $C = X^T X$





Powers of a matrix

- The matrix $C = X^T X$ is positive semi-definite:
 - It is symmetric
 - All the eigenvalues are greater than or equal to zero
- Therefore, $C = U D U^T$ where:
 - U is an orthogonal matrix comprised of eigenvectors
 - D is a diagonal matrix containing the eigenvalues along the diagonal
 - If necessary, reorder the eigenvalues from largest to smallest, swapping the entries of D and columns of U as necessary
 - The unit eigenvectors define the principal components
 - $U \sqrt{\frac{1}{m} D}$ describes the variation of the principal components:
 - $\sqrt{\frac{1}{m} \lambda_k} \hat{\mathbf{u}}_k$ describes the variation the k^{th} principal component





Powers of a matrix

- We can examine how quickly the eigenvalues drop:
 - If the eigenvalues are all approximately the same, there is little correlation between the factors and/or responses
 - If, however, there is a dominant eigenvalue, or a few dominant eigenvalues, these can possibly be used to control the responses by adjusting the factors





Example

- Thus, suppose we collect this data:

Proportion of solvent	Temperature	Elastic modulus
0.26	-5.1	27.7
0.33	-9.3	38.5
0.84	-22.6	54.2
0.86	-11.8	26.6
0.7	-10.3	21
0.42	-7.4	12.3
0.53	-11.2	35.2
0.81	-20.2	52.8
0.26	-2	21.8
0.08	2.6	22.9





Example

- The mean of this data is:

$$\bar{\mathbf{x}} = \begin{pmatrix} 0.509 \\ -9.37 \\ 31.3 \end{pmatrix}$$

$$\begin{pmatrix} 0.26 & -5.1 & 27.7 \\ 0.33 & -9.3 & 38.5 \\ 0.84 & -22.6 & 54.2 \\ 0.86 & -11.8 & 26.6 \\ 0.7 & -10.3 & 21 \\ 0.42 & -7.4 & 12.3 \\ 0.53 & -11.2 & 35.2 \\ 0.81 & -20.2 & 52.8 \\ 0.26 & -2 & 21.8 \\ 0.08 & 2.6 & 22.9 \end{pmatrix}$$





Example

- We subtract the mean from each row:

$$\bar{\mathbf{x}} = \begin{pmatrix} 0.509 \\ -9.37 \\ 31.3 \end{pmatrix}$$

$$X = \begin{pmatrix} -0.249 & 4.63 & -3.6 \\ -0.179 & 0.43 & 7.2 \\ 0.331 & -12.87 & 22.9 \\ 0.351 & -2.07 & -4.7 \\ 0.191 & -0.57 & -10.3 \\ -0.089 & 2.33 & -19 \\ 0.021 & -1.47 & 3.9 \\ 0.301 & -10.47 & 21.5 \\ -0.249 & 7.73 & -9.5 \\ -0.429 & 12.33 & -8.4 \end{pmatrix}$$

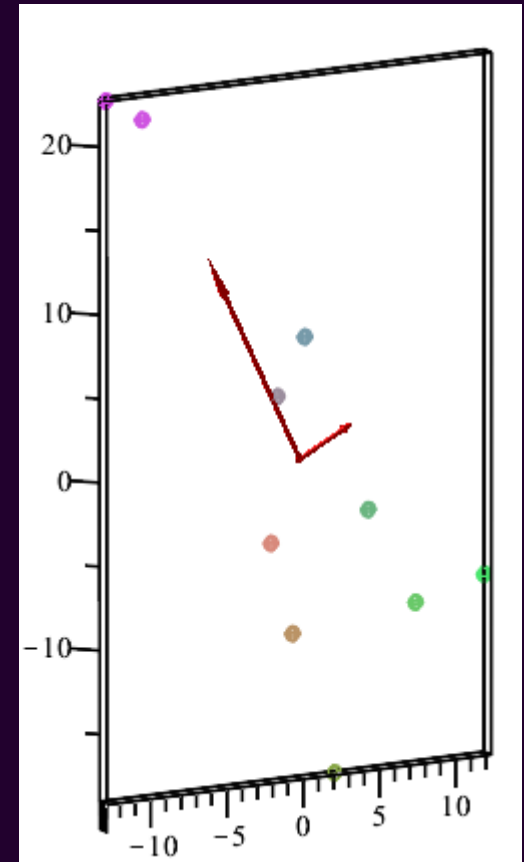




Example

- We can plot the data points and the eigenvectors multiplied as described:

$$14.40, \begin{pmatrix} 0.011 \\ -0.432 \\ 0.901 \end{pmatrix} \quad 4.05, \begin{pmatrix} -0.046 \\ 0.900 \\ 0.432 \end{pmatrix} \quad 0.097, \begin{pmatrix} -0.999 \\ -0.047 \\ -0.010 \end{pmatrix}$$





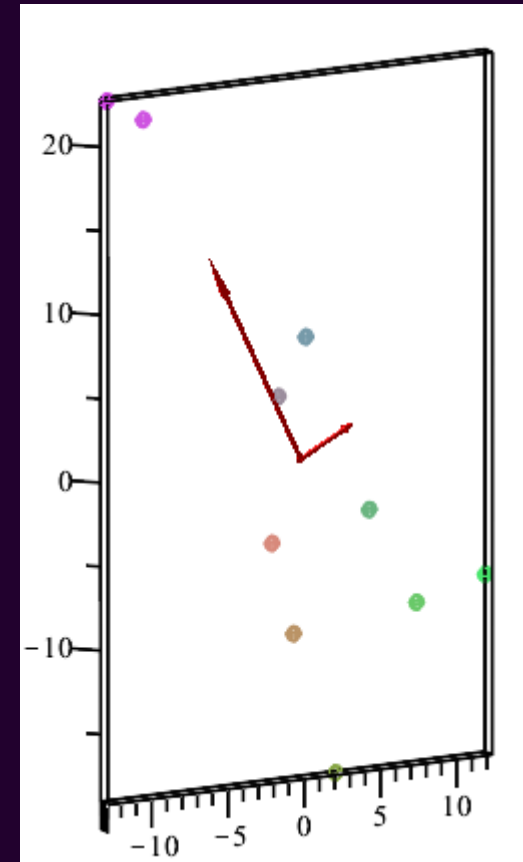
Example

- With the dominant eigenvector, we have:

$$\bar{\mathbf{x}} + \alpha \mathbf{w}_1 = \begin{pmatrix} 0.509 \\ -9.37 \\ 31.3 \end{pmatrix} + \alpha \begin{pmatrix} 0.011 \\ -0.432 \\ 0.901 \end{pmatrix}$$

- Thus, suppose our target elastic modulus is some value E , so $E = 31.3 + \alpha 0.901$
 - Thus, we can solve for α :
$$\alpha \leftarrow (E - 31.3)/0.901$$
 - We thus use this α to select the most appropriate proportion of solvent and temperature:

$$\begin{pmatrix} 0.509 + 0.011\alpha \\ -9.37 - 0.432\alpha \end{pmatrix}$$





Example

- For example, for a target elastic modulus of 20, we have
$$\alpha \leftarrow (20 - 31.3)/0.901 = -12.542$$
 - The most appropriate proportion of solvent and temperature are:

$$0.509 + 0.011\alpha = 0.371$$

$$-9.37 - 0.432\alpha = -3.952$$

- For a target elastic modulus of 40, we have
$$\alpha \leftarrow (40 - 31.3)/0.901 = 9.655$$
 - The most appropriate proportion of solvent and temperature are:

$$0.509 + 0.011\alpha = 0.615$$

$$-9.37 - 0.432\alpha = -13.541$$





Summary

- Following this topic, you now
 - Understand the idea behind principal component analysis
 - Know that, having subtracted off the average of the data, the matrix X can be used to find a scalar multiple of the covariance matrix
 - Are aware that the eigenvalues and eigenvectors can be used to determine if there is any strong correlation between the factors and responses





References

- [1] https://en.wikipedia.org/wiki/Principal_component_analysis





Acknowledgments

None so far.





Colophon

These slides were prepared using the Cambria typeface. Mathematical equations use Times New Roman, and source code is presented using Consolas. Mathematical equations are prepared in MathType by Design Science, Inc. Examples may be formulated and checked using Maple by Maplesoft, Inc.

The photographs of flowers and a monarch butter appearing on the title slide and accenting the top of each other slide were taken at the Royal Botanical Gardens in October of 2017 by Douglas Wilhelm Harder. Please see

<https://www.rbg.ca/>

for more information.





Disclaimer

These slides are provided for the NE 112 *Linear algebra for nanotechnology engineering* course taught at the University of Waterloo. The material in it reflects the authors' best judgment in light of the information available to them at the time of preparation. Any reliance on these course slides by any party for any other purpose are the responsibility of such parties. The authors accept no responsibility for damages, if any, suffered by any party as a result of decisions made or actions based on these course slides for any other purpose than that for which it was intended.

