Business Analytics & Machine Learning

High-Dimensional Problems

Prof. Dr. Martin Bichler

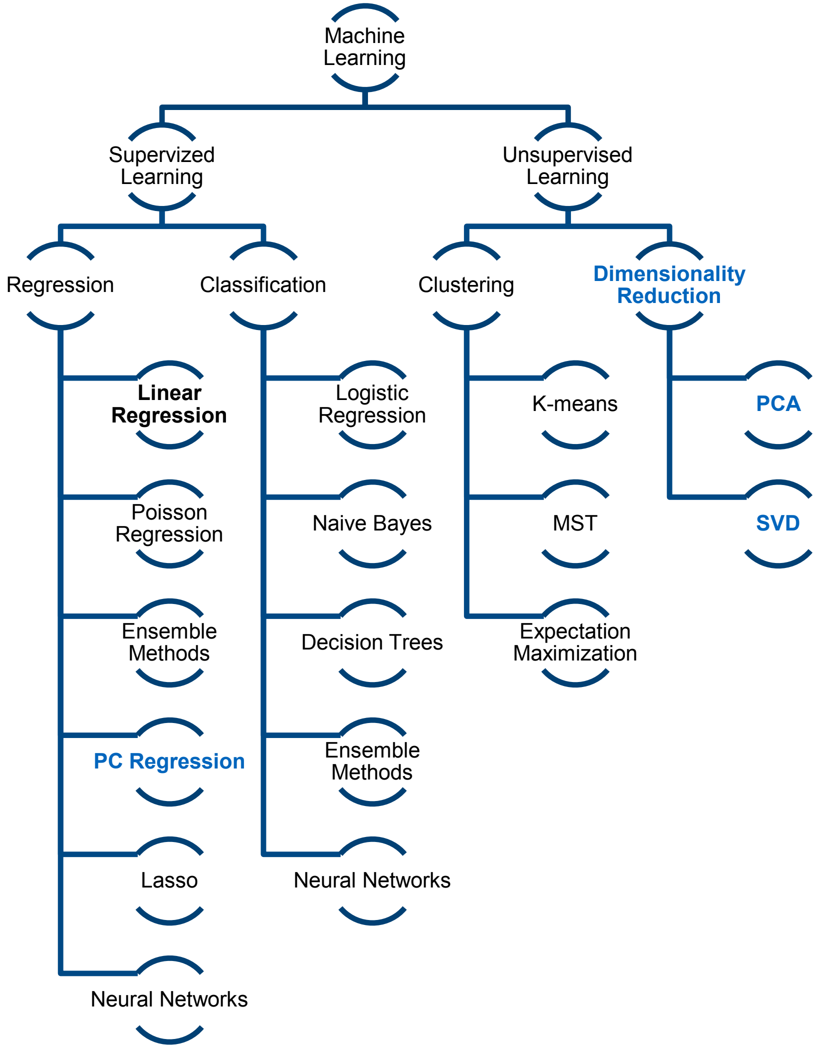
Department of Computer Science

School of Computation, Information, and Technology

Technical University of Munich

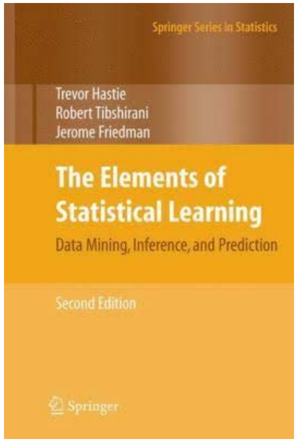
Course Content

- Introduction
- Regression Analysis
- Regression Diagnostics
- Logistic and Poisson Regression
- Naive Bayes and Bayesian Networks
- Decision Tree Classifiers
- Data Preparation and Causal Inference
- Model Selection and Learning Theory
- Ensemble Methods and Clustering
- **Dimensionality Reduction**
- Convex Optimization
- Neural Networks
- Reinforcement Learning

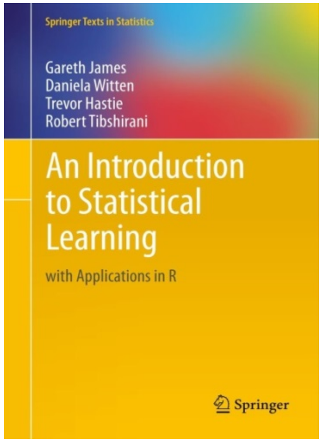


Recommended Literature

- **The Elements of Statistical Learning**
  - Trevor Hastie, Robert Tibshirani, Jerome Friedman
  - https://web.stanford.edu/~hastie/ElemStatLearn/
  - Section: 3.4 – 3.6


Springer Series in Statistics
Trevor Hastie
Robert Tibshirani
Jerome Friedman
The Elements of Statistical Learning
Data Mining, Inference, and Prediction
Second Edition
Springer

- **An Introduction to Statistical Learning: With Applications in R**
  - Gareth James, Trevor Hastie, Robert Tibshirani
  - http://www-bcf.usc.edu/~gareth/ISL/
  - Section: 6, 10.2



- https://www.cs.princeton.edu/picasso/mats/PCA-Tutorial-Intuition_jp.pdf

Autline for Today

- **Overview**
- Linear algebra revisited
- Principal Component Analysis
- Singular Value Decomposition
- PC regression

Principal Component Analysis (PCA)

Principal component analysis (PCA) converts a set of possibly correlated variables into a (possibly smaller) set of values of linearly uncorrelated variables called principal components.

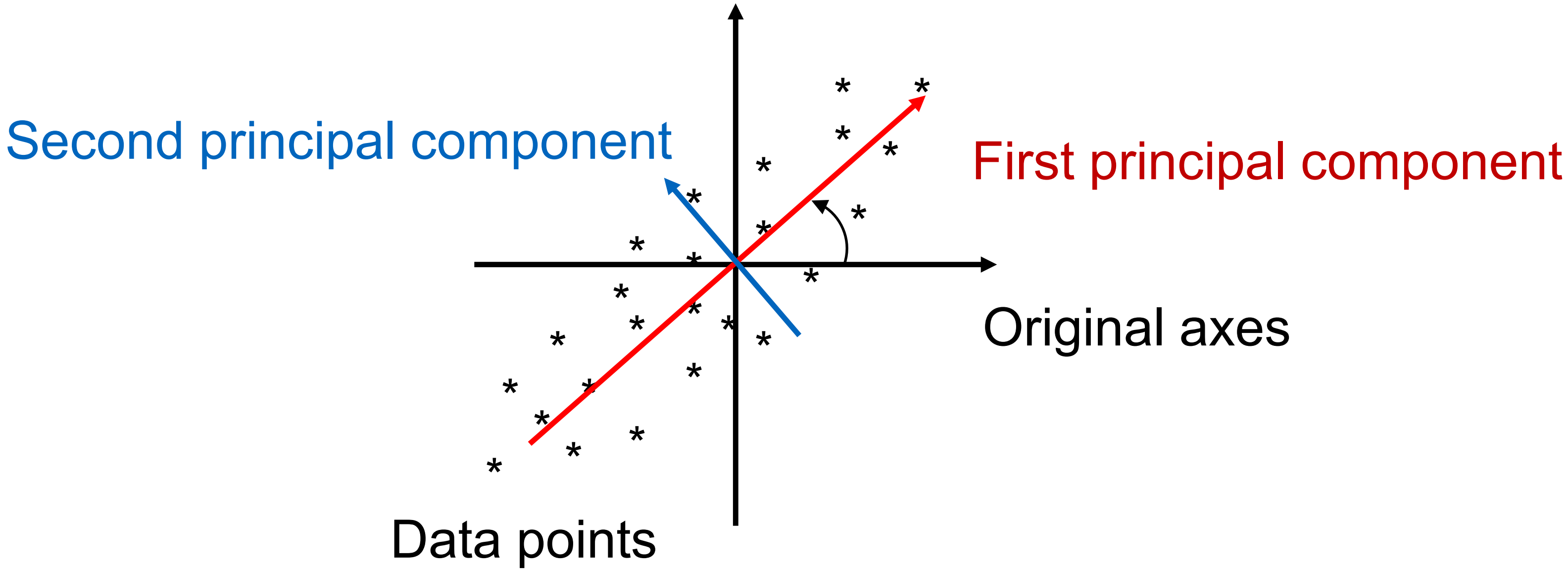
The first principal component has the largest possible variance, and each succeeding component in turn has the highest variance possible under the constraint that it is orthogonal to (i.e., uncorrelated with) the preceeding components.

The principal components are orthogonal (they are the Eigenvectors of the symmetric covariance matrix).

PCA is sometimes referred to as the Karhunen–Loève transform (KLT) and related to singular value decomposition (SVD).

Principal Component Analysis

Suppose we have a population measured on p random variables $X_1, \ldots, X_p$.
Note that these random variables represent the p axes of the Cartesian coordinate system in which the population resides.
Our goal is to develop a new set of $k \leq p$ axes (linear combinations of the original p axes) in the directions of greatest variability. This is accomplished by rotating the axes.



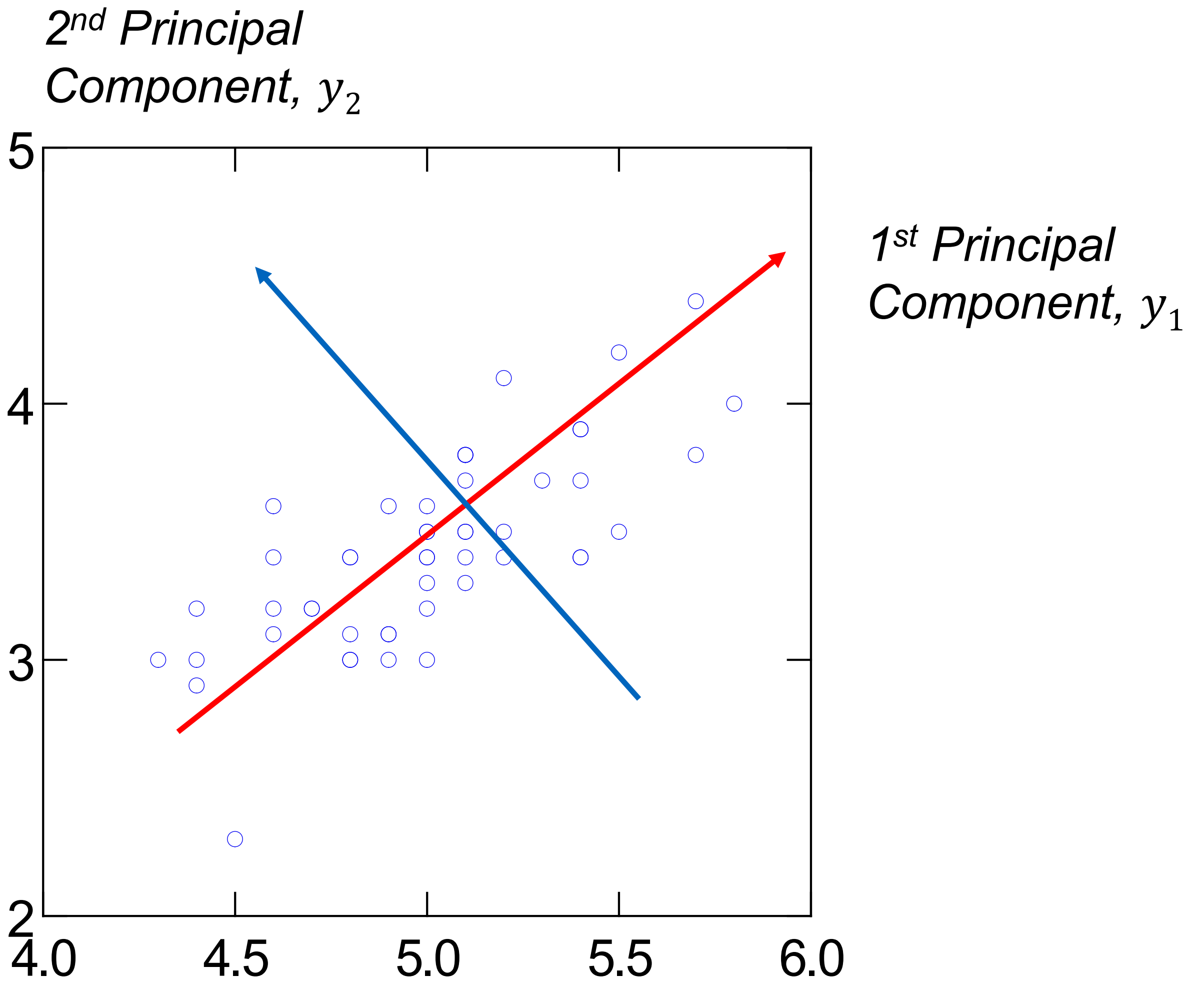
Second principal component

First principal component

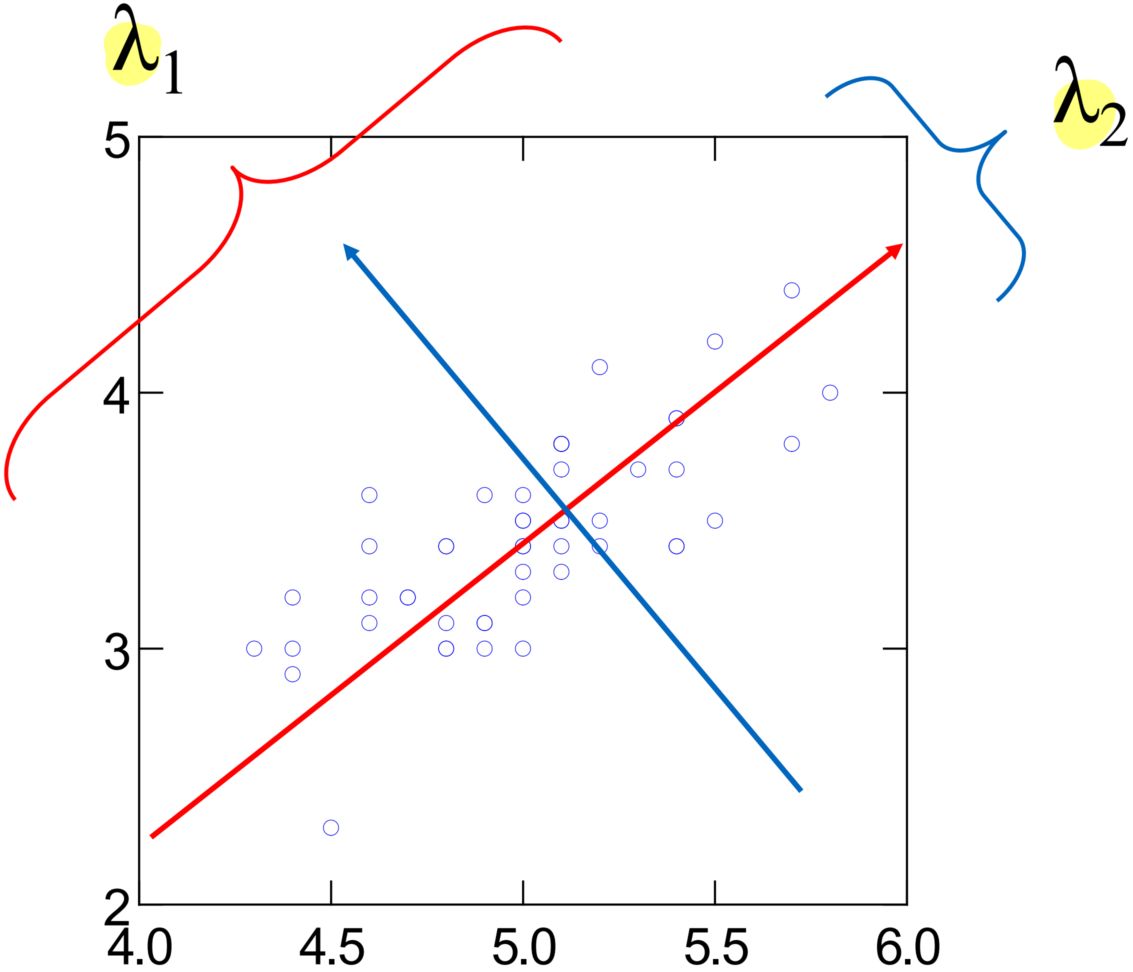
Original axes

Data points

Principal Components are Eigenvectors

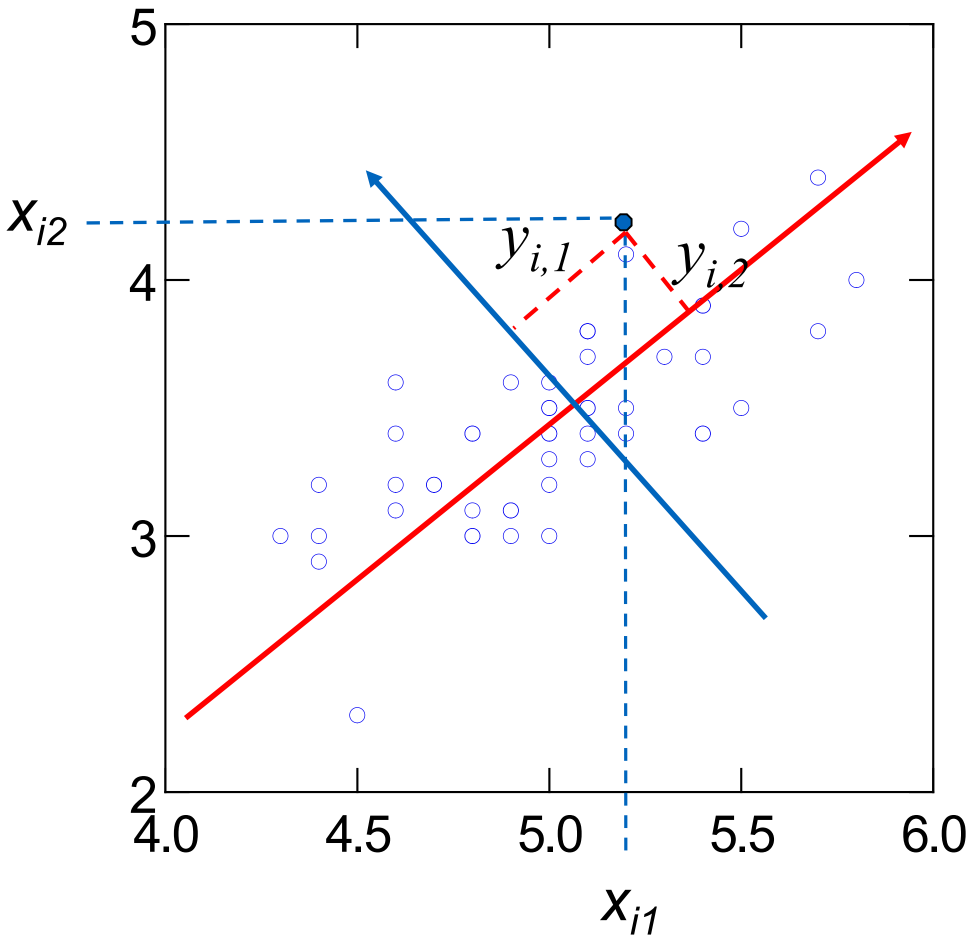


PCA Eigenvalues



Eigenvalues λ_1 explain the proportion of the total variance explained by PC1.

PCA Scores



The PCA score for any of the x is just it's coefficient in each of the y.

PCA

From p original variables: $x_1, x_2, \ldots, x_p$:

Produce k (or less) new variables: $y_1, y_2, \ldots, y_k$ as linear combinations of the original variables x_i.

$$\left.\begin{aligned}
y_1 &= a_{11}x_1 + a_{12}x_2 + \ldots + a_{1p}x_p \\
y_2 &= a_{21}x_1 + a_{22}x_2 + \ldots + a_{2p}x_p \\
&\vdots \\
y_k &= ak_1x_1 + ak_2x_2 + \ldots + a_{kp}x_p
\end{aligned}\right\} \quad \text{y_k's are the Principal Components}$$

such that:

y_k's are uncorrelated (orthogonal)

y_1 explains as much as possible of original variance in data

y_2 explains as much as possible of remaining variance

etc.

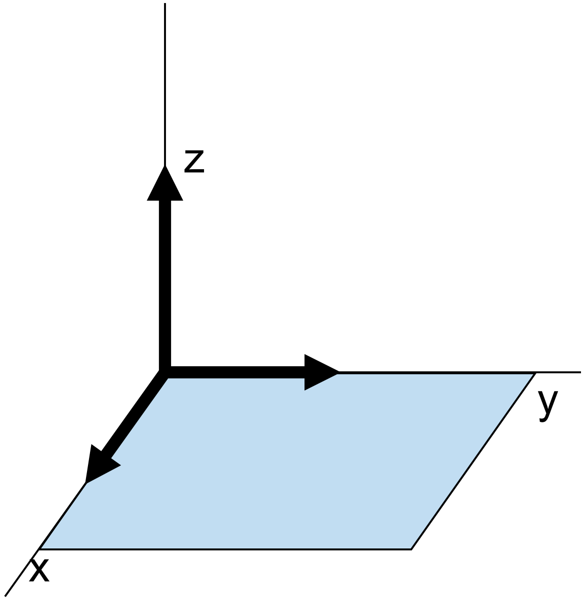
Outline for today

- Overview
- **Linear algebra revisited**
- Principal Component Analysis
- Singular Value Decomposition
- PCA regression and regularization

Vector Spaces

Formally, a **vector space** is a set of vectors which is closed under addition and multiplication by real numbers.
A **subspace** is a subset of a vector space which is a vector space itself, e.g., the plane $z = 0$ (i.e., $\mathbb{R}^2$) is a subspace of $\mathbb{R}^3$.
Subspaces must include the origin (zero vector).

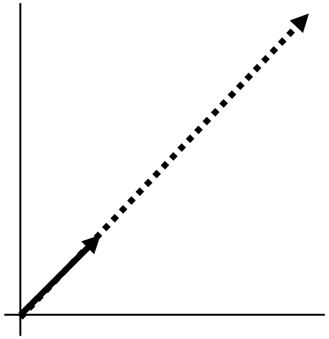
The notion of planes in $\mathbb{R}^3$ may be extended to **hyperplanes** in $\mathbb{R}^n$ (of dimension $n - 1$).



Matrices as Linear Transformations

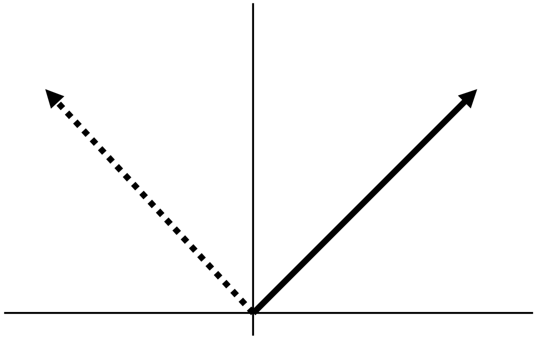
λI

$$\begin{pmatrix} \underline{5} & 0 \\ 0 & \underline{5} \end{pmatrix}\begin{pmatrix} 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 5 \\ 5 \end{pmatrix}$$



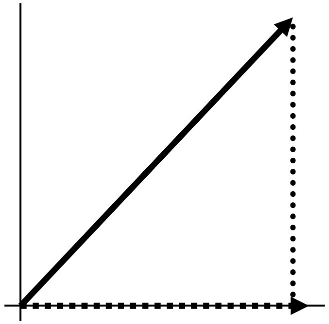
(stretching)

$$\begin{pmatrix} 0 & -\underline{1} \\ \underline{1} & 0 \end{pmatrix}\begin{pmatrix} 1 \\ 1 \end{pmatrix} = \begin{pmatrix} -1 \\ 1 \end{pmatrix}$$



(rotation)

$$\begin{pmatrix} \underline{1} & 0 \\ 0 & 0 \end{pmatrix}\begin{pmatrix} 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$$



(projection)

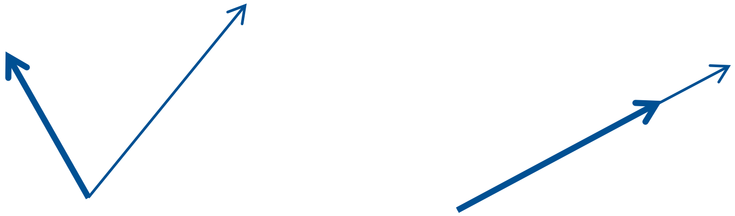
Linear Independence

Vectors $v_1, \dots, v_k$ are linearly independent if $c_1 v_1 + \dots + c_k v_k = 0$ implies $c_1 = \dots = c_k = 0$

$$\begin{pmatrix} 1 & 0 \\ 2 & 3 \\ 1 & 3 \end{pmatrix} \begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

This means, the null space is the set of solutions for u and v.

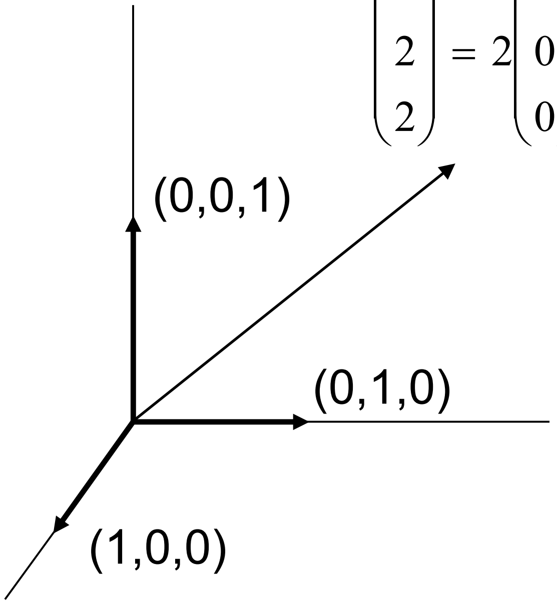
If the null space contained only $(u, v) = (0,0)$, then the columns are linearly independent.



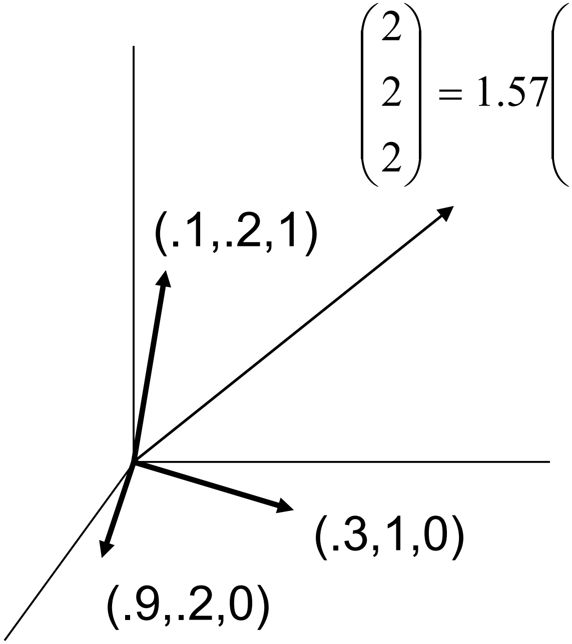
Linear Independence and Basis

If all vectors in a vector space may be expressed as linear combinations of $v_1, \dots, v_k$, then $v_1, \dots, v_k$ **span** the space.

$$\begin{pmatrix}2\\2\\2\end{pmatrix} = 2\begin{pmatrix}1\\0\\0\end{pmatrix} + 2\begin{pmatrix}0\\1\\0\end{pmatrix} + 2\begin{pmatrix}0\\0\\1\end{pmatrix}$$



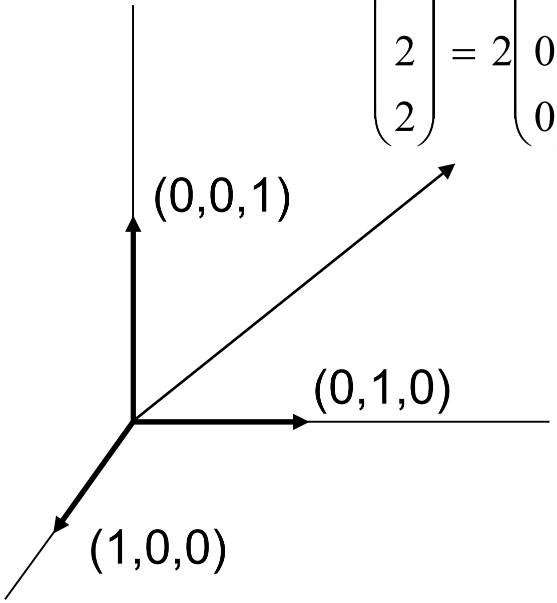
$$\begin{pmatrix}2\\2\\2\end{pmatrix} = 1.57\begin{pmatrix}.9\\.2\\0\end{pmatrix} + 1.29\begin{pmatrix}.3\\1\\0\end{pmatrix} + 2\begin{pmatrix}.1\\.2\\1\end{pmatrix}$$



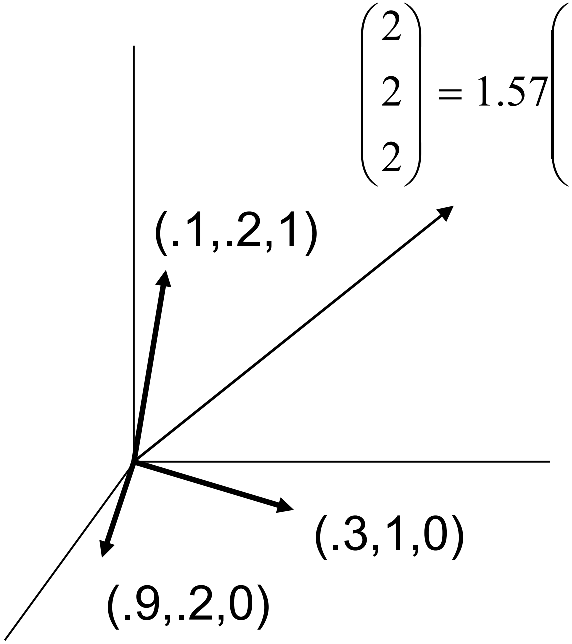
Linear Independence and Basis

A **basis** is a set of linearly independent vectors which span the space.
With three dimensions we need tree basis vectors of dimension 3.
A basis is a minimal set of spanning vectors (not necessarily orthogonal).

$$\begin{pmatrix}2\\2\\2\end{pmatrix} = 2\begin{pmatrix}1\\0\\0\end{pmatrix} + 2\begin{pmatrix}0\\1\\0\end{pmatrix} + 2\begin{pmatrix}0\\0\\1\end{pmatrix}$$



$$\begin{pmatrix}2\\2\\2\end{pmatrix} = 1.57\begin{pmatrix}.9\\.2\\0\end{pmatrix} + 1.29\begin{pmatrix}.3\\1\\0\end{pmatrix} + 2\begin{pmatrix}.1\\.2\\1\end{pmatrix}$$



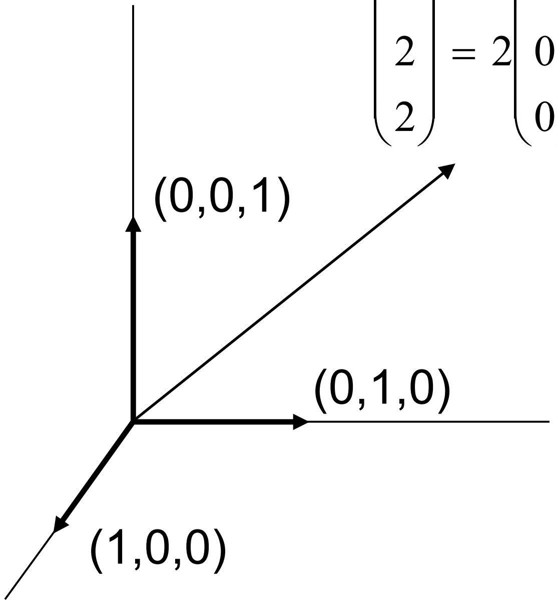
Linear Independence and Basis

Two vectors are orthogonal if their dot product is 0.

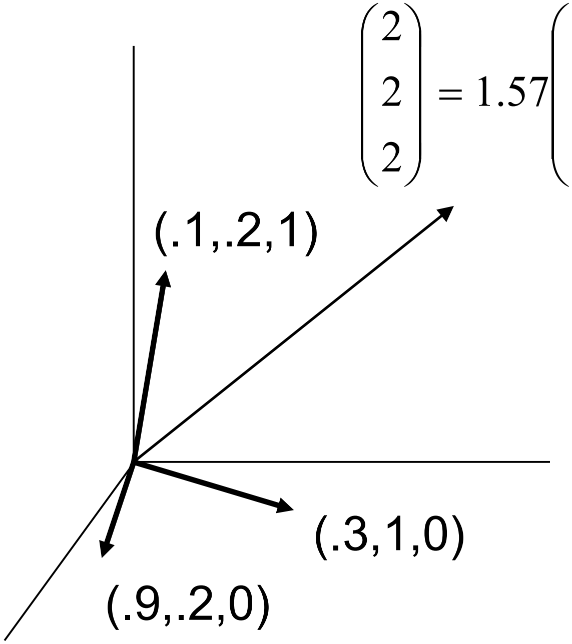
An orthogonal basis consists of orthogonal vectors.

An orthonormal basis consists of orthogonal vectors of unit length.

$$\begin{pmatrix}2\\2\\2\end{pmatrix} = 2\begin{pmatrix}1\\0\\0\end{pmatrix} + 2\begin{pmatrix}0\\1\\0\end{pmatrix} + 2\begin{pmatrix}0\\0\\1\end{pmatrix}$$



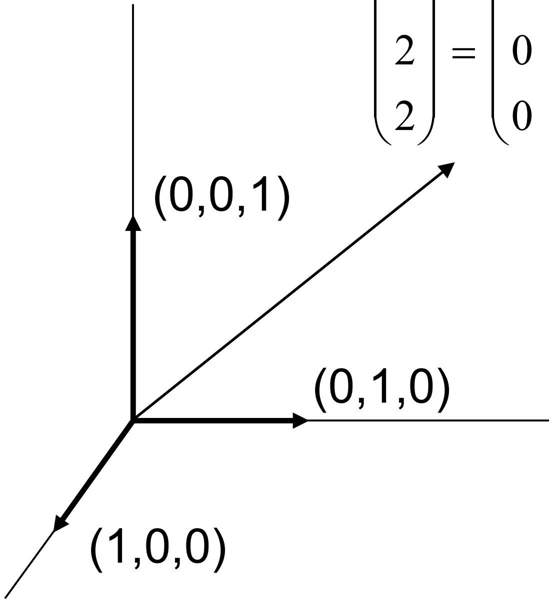
$$\begin{pmatrix}2\\2\\2\end{pmatrix} = 1.57\begin{pmatrix}.9\\.2\\0\end{pmatrix} + 1.29\begin{pmatrix}.3\\1\\0\end{pmatrix} + 2\begin{pmatrix}.1\\.2\\1\end{pmatrix}$$



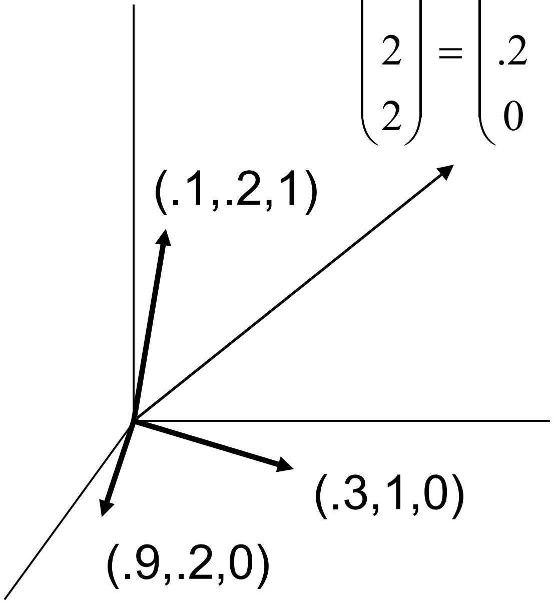
Basis Transformations

We may write $v = (2,2,2)$ in terms of an alternate basis:

$$\begin{pmatrix}2\\2\\2\end{pmatrix} = \overset{B}{\begin{pmatrix}1 & 0 & 0\\0 & 1 & 0\\0 & 0 & 1\end{pmatrix}}\begin{pmatrix}2\\2\\2\end{pmatrix}$$



$$\begin{pmatrix}2\\2\\2\end{pmatrix} = \overset{B'}{\begin{pmatrix}.9 & .3 & .1\\.2 & 1 & .2\\0 & 0 & 1\end{pmatrix}}\begin{pmatrix}1.57\\1.29\\2\end{pmatrix}$$



Components of $(1.57,1.29,2)$ are projections of v onto the new basis vectors B'. Transform back to the original basis B via the inverse of the transformation matrix T: $B'T = B$ and $B' = BT^{-1}$ (note that in the example $B = I$ such that $B' = T^{-1}$).

Eigenvalues and Eigenvectors

Eigenvectors of a linear transformation (i.e., a matrix) $\boldsymbol{A}$ are not rotated (but will be scaled by the corresponding Eigenvalue) when $\boldsymbol{A}$ is applied.

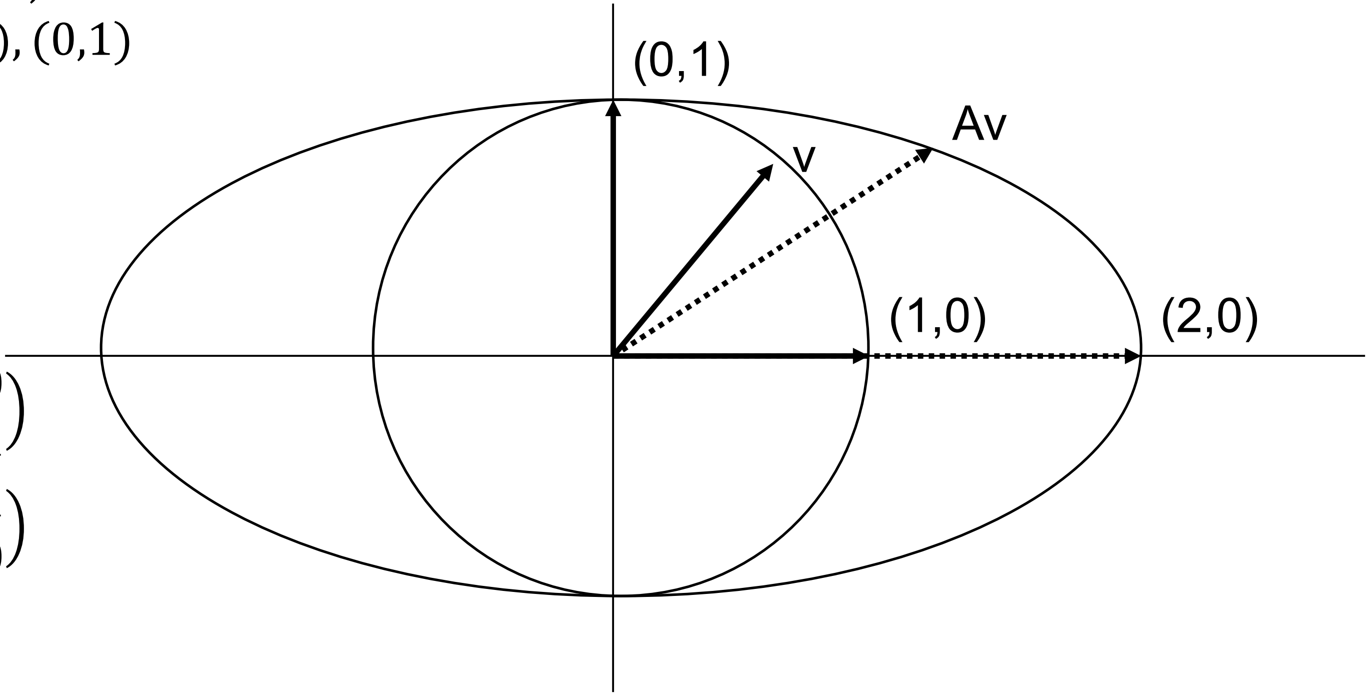
An Eigenvector stays on its span after the transformation! Other vectors do not.

Eigenvalues $\lambda = 2, 1$ with
Eigenvectors $(1{,}0), (0{,}1)$

$$A = \begin{pmatrix} 2 & 0 \\ 0 & 1 \end{pmatrix}$$

$$\begin{pmatrix} 2 & 0 \\ 0 & 1 \end{pmatrix}\begin{pmatrix} 0 \\ 1 \end{pmatrix} = 1\begin{pmatrix} 0 \\ 1 \end{pmatrix}$$

$$\begin{pmatrix} 2 & 0 \\ 0 & 1 \end{pmatrix}\begin{pmatrix} 1 \\ 0 \end{pmatrix} = 2\begin{pmatrix} 1 \\ 0 \end{pmatrix}$$



Eigenvalues and Eigenvectors

Let $\boldsymbol{A}$ be an $p \times p$ matrix with Eigenvalue λ and corresponding Eigenvector v. Thus $\mathbf{A}v = \lambda v$. This equation may be written

$$\mathbf{A}v - \lambda v = 0$$

given

$$\boxed{(\mathbf{A} - \lambda \mathbf{I}_p)v = 0}$$

- Solving the equation $|\mathbf{A} - \lambda\mathbf{I}_p| = 0$ for λ leads to all the Eigenvalues of **A**.
  - A determinant can be viewed as the volume scaling factor of the linear transformation described by the matrix!
  - If the determinant is 0, then the space described by the transformation via $(\mathbf{A} - \lambda \mathrm{I}_p)$ is 0, i.e., we get a scalar λ that scales the vector.
- On expending the determinant $|\mathbf{A} - \lambda\mathbf{I}_p|$, we get a polynomial in λ.
- This polynomial is called the **characteristic polynomial** of **A**.
- The equation $|\mathbf{A} - \lambda\mathbf{I}_p| = 0$ is called the **characteristic equation** of **A**.

Eigenvalues and Eigenvectors

$$\mathbf{A} = \begin{bmatrix} -4 & -6 \\ 3 & 5 \end{bmatrix}$$

- For the equation $|A - \lambda I_p| = 0$ let us first derive the characteristic polynomial of $\boldsymbol{A}$.
- We get: $$\mathbf{A} - \lambda\mathbf{I}_2 = \begin{bmatrix} -4 & -6 \\ 3 & 5 \end{bmatrix} - \lambda\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} -4-\lambda & -6 \\ 3 & 5-\lambda \end{bmatrix}$$
  $$|\mathbf{A} - \lambda\mathbf{I}_2| = (-4-\lambda)(5-\lambda) + 18 = \lambda^2 - \lambda - 2$$
- We now solve the characteristic equation of **A**.

  $$\lambda^2 - \lambda - 2 = 0 \Rightarrow (\lambda - 2)(\lambda + 1) = 0 \Rightarrow \underline{\lambda = 2} \text{ or } \underline{-1}$$
- The Eigenvalues of $\boldsymbol{A}$ are 2 and –1.
- The corresponding Eigenvectors are found by using these values of λ in the equation $\underline{(\mathbf{A} - \lambda\mathbf{I}_2)v = 0}$.

Reminder: Eigenvectors for λ = 2

We solve the equation $(\mathbf{A} - 2\mathbf{I}_2)v = 0$ for v. The matrix $(\mathbf{A} - 2\mathbf{I}_2)$ is obtained by subtracting 2 from the diagonal elements of **A.**

- We get:

$$\begin{bmatrix} -6 & -6 \\ 3 & 3 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = 0$$

- This leads to the system of equations:

$$\begin{cases} -6v_1 - 6v_2 = 0 \\ 3v_1 + 3v_2 = 0 \end{cases}$$

giving $v_1 = -v_2$. The solutions to this system of equations are $\underline{v_1 = -r,\ v_2 = r}$, where *r* is a scalar. Thus, the Eigenvectors of **A** corresponding to λ = 2 are nonzero vectors of the form

$$r \begin{bmatrix} -1 \\ 1 \end{bmatrix}$$

Eigenvectors for λ = –1

We solve the equation $(\mathbf{A} + 1\mathbf{I}_2)v = 0$ for v. The matrix $(\mathbf{A} + 1\mathbf{I}_2)$ is obtained by adding 1 to the diagonal elements of **A**.

We get:

$$\begin{bmatrix} -3 & -6 \\ 3 & 6 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = 0$$

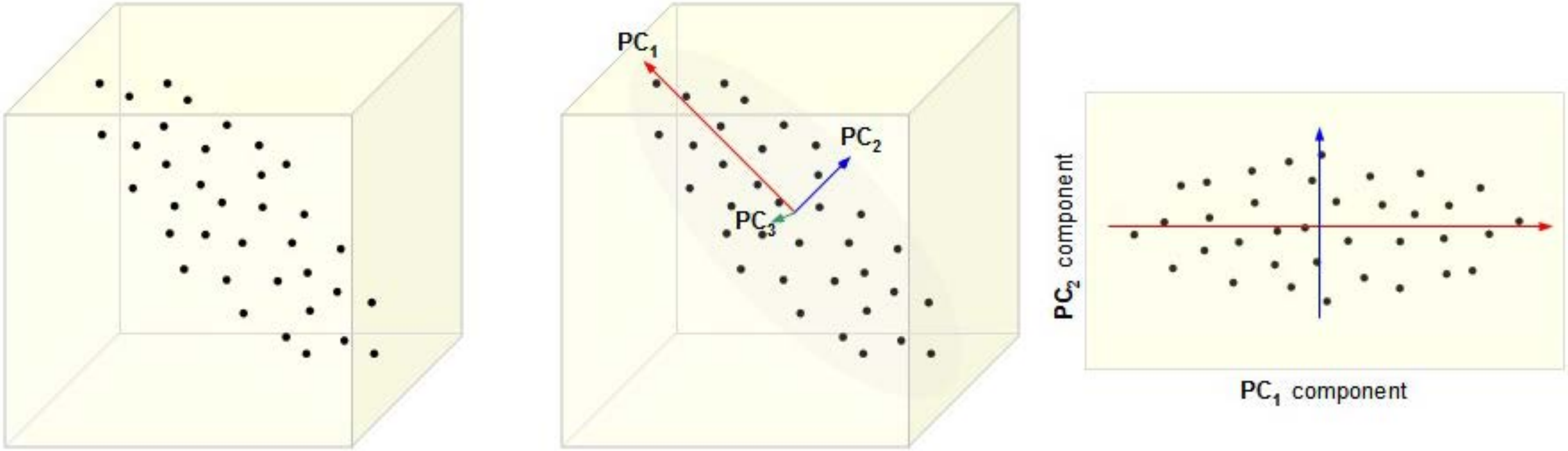
This leads to the system of equations

$$\begin{cases} -3v_1 - 6v_2 = 0 \\ 3v_1 + 6v_2 = 0 \end{cases}$$

Thus $v_1 = -2v_2$. The solutions to this system of equations are $\underline{v_1 = -2s \text{ and } v_2 = s}$, where s is a scalar. Thus, the Eigenvectors of **A** corresponding to λ = –1 are nonzero vectors of the form

$$s\begin{bmatrix} -2 \\ 1 \end{bmatrix} ✓$$

Principal Components in 3 Dimensions



Principal components are the **Eigenvectors** of the covariance or correlation matrix. They show the direction of maximum variance.

The **Eigenvalues** λ explain the amount of variance along that axis, and the proportion of the overall variance explained by the PC.

So **PCA** simply takes points expressed in the standard **basis** and transforms them into points expressed in an eigenvector **basis**.

Can you explain the Eigenvectors and Eigenvalues of a matrix?

Outline for today

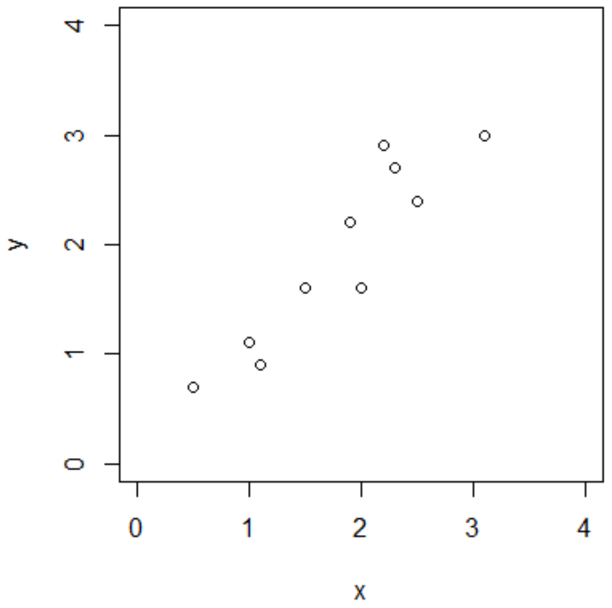
- Overview
- Linear algebra revisited
- **Principal Component Analysis**
- Singular Value Decomposition
- PCA regression and regularization

PCA Example –STEP 1

Center data by subtracting the mean as we are rotating axes around the origin later on.

Data (A):	
x	y
2.5	2.4
0.5	0.7
2.2	2.9
1.9	2.2
3.1	3.0
2.3	2.7
2	1.6
1	1.1
1.5	1.6
1.1	0.9

Zero mean Data (X)	
x	y
.69	.49
-1.31	-1.21
.39	.99
.09	.29
1.29	1.09
.49	.79
.19	-.31
-.81	-.81
-.31	-.31
-.71	-1.01



PCA Example –STEP 2

Calculate the covariance matrix, which summarizes the relationship between variables

$$cov = \Sigma = \begin{pmatrix} .616555556 & .615444444 \\ .615444444 & .716555556 \end{pmatrix}$$

Since the non-diagonal elements in this covariance matrix are positive, we should expect that both the x and y variable increase together.

Correlation and covariance matrices are all positive semi-definite matrices. Thus its eigenvalues are always positive or null.

Variance-Covariance vs. Correlation Matrix

$$\Sigma = \begin{bmatrix} \mathrm{E}[(X_1-\mu_1)(X_1-\mu_1)] & \mathrm{E}[(X_1-\mu_1)(X_2-\mu_2)] & \cdots & \mathrm{E}[(X_1-\mu_1)(X_n-\mu_n)] \\ \mathrm{E}[(X_2-\mu_2)(X_1-\mu_1)] & \mathrm{E}[(X_2-\mu_2)(X_2-\mu_2)] & \cdots & \mathrm{E}[(X_2-\mu_2)(X_n-\mu_n)] \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{E}[(X_n-\mu_n)(X_1-\mu_1)] & \mathrm{E}[(X_n-\mu_n)(X_2-\mu_2)] & \cdots & \mathrm{E}[(X_n-\mu_n)(X_n-\mu_n)] \end{bmatrix}.$$

The **covariance matrix** centers each variable on the mean, but the scale matters. Should be used only, when the variables are measured in comparable units and the differences in variance are important for interpretation.
A covariance matrix implicitly involves centering of the data already.

If variables are measured in different units use the **correlation matrix**:

$$\boxed{\rho = corr(X,Y) = \frac{cov(X,Y)}{\sigma_X \sigma_Y}}$$

Use correlation if differences in variance across the variables are not meaningful.

PCA Example –STEP 3

Calculate the Eigenvectors and Eigenvalues of the covariance matrix ∑:

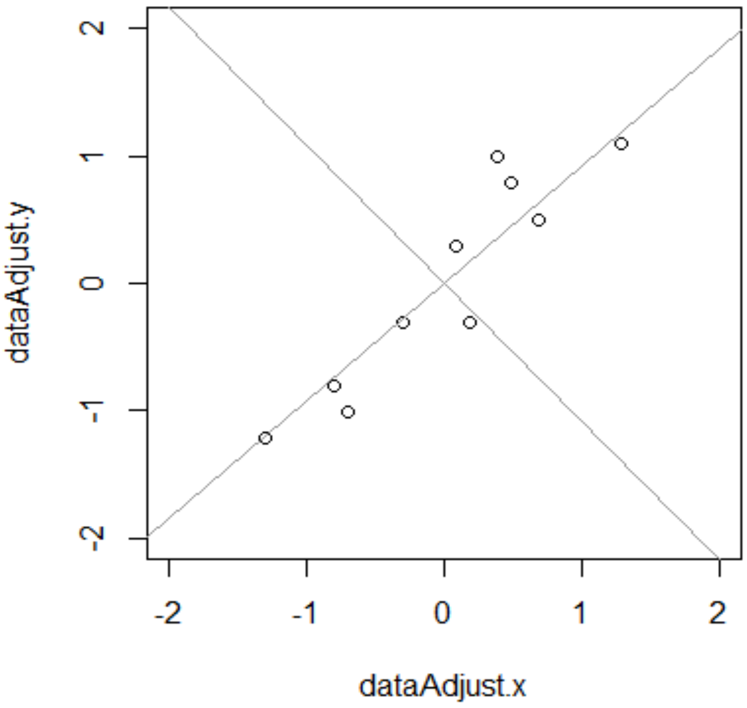
$$Eigenvalues = \begin{pmatrix} .0490833989 \\ 1.28402771 \end{pmatrix}$$

$$Eigenvectors = \begin{pmatrix} -.735178656 & .677873399 \\ .677873399 & .735178656 \end{pmatrix}$$

Once Eigenvectors are found from the covariance matrix, the next step is to order them by Eigenvalue, highest to lowest. This gives you the components in order of significance: Φ

$$\Phi = \begin{pmatrix} 0.677873399 & -.735178656 \\ 0.735178656 & .677873399 \end{pmatrix}$$

PCA Example –STEP 3



Eigenvectors are plotted as diagonal dotted lines on the plot.

Note they are perpendicular to each other.

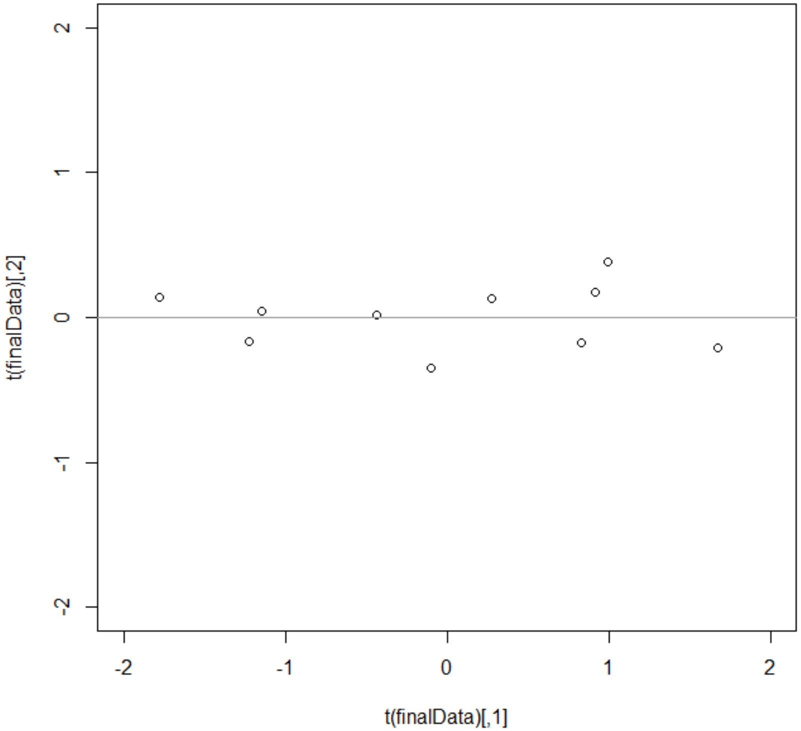
The second Eigenvector gives us the other, less important, pattern in the data, that all the points follow the main line, but are off to the side of the main line by some amount.

Taking both Eigenvectors and Rotate

Final data: $\mathbf{Z} = \mathbf{X} \cdot \mathbf{\Phi}$

If you take all Eigenvectors in $\mathbf{\Phi}$, you will get the original data rotated so that the Eigenvectors are the axes! Rotation is equivalent to a basis transformation by an orthonormal basis.

Z1 (PC1)	Z2 (PC2)
0.82797019	-0.17511531
-1.77758033	0.14285723
0.99219749	0.38437499
0.27421042	0.13041721
1.67580142	-0.20949846
0.91294910	0.17528244
-0.09910944	-0.34982470
-1.14457216	0.04641726
-0.43804614	0.01776463
-1.22382056	-0.16267529



PCA Example –STEP 4

Now, if you like, you can decide to ignore the components of lesser significance.

You do lose some information, but if the Eigenvalues are small, you don't lose much:

- p dimensions in your data
- calculate p Eigenvectors and Eigenvalues
- choose only the first $k < p$ Eigenvectors
- final data set has only k dimensions

↖ dimensions reduced without losing a significant information.

PCA Example –STEP 4

Feature vector = $(v_1\ v_2\ v_3 \ldots v_{\ k})$

We can either form a feature vector with both of the Eigenvectors:

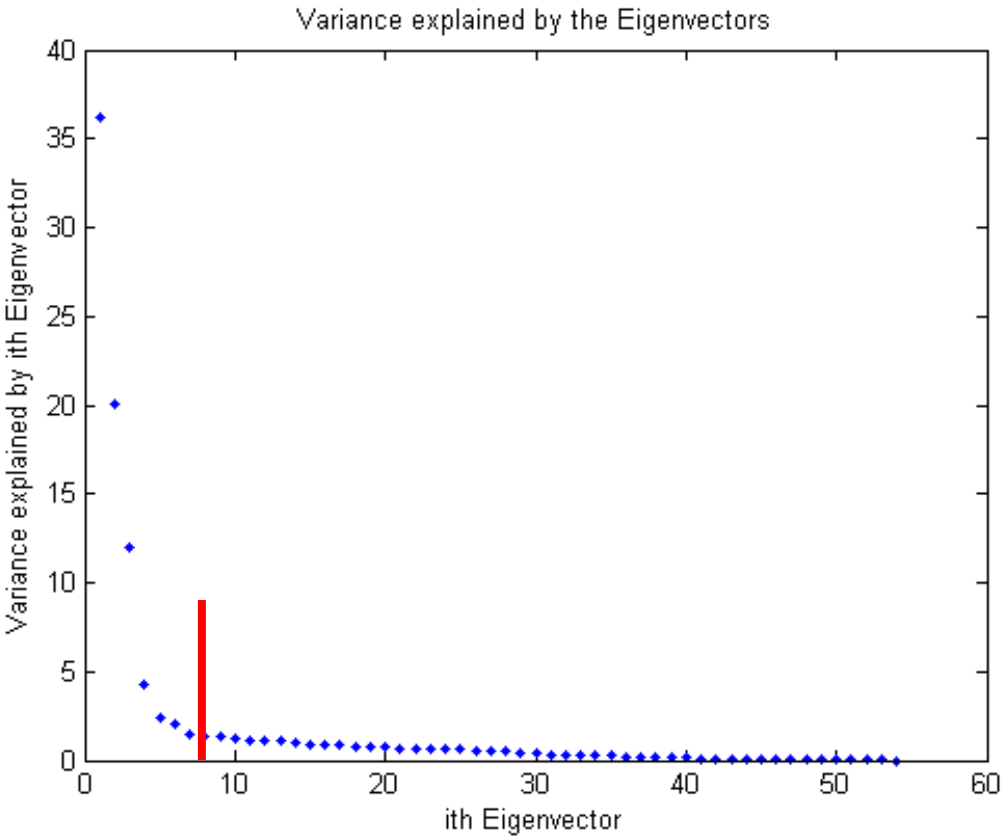
$$\begin{pmatrix} 0.677873399 & -.735178656 \\ 0.735178656 & .677873399 \end{pmatrix}$$

Or we can choose to leave out the smaller, less significant component and only have a single column:

$$\begin{pmatrix} 0.677873399 \\ 0.735178656 \end{pmatrix}$$

How many Principal Components?

- Check the distribution of Eigenvalues in a scree plot (below).
- Take enough Eigenvalues to cover 80-90% of the total variance.



Taking only 1 Eigenvector: Data Reduction

Taking only one Eigenvector: $\mathbf{Z} = \mathbf{X} \cdot \mathbf{\Phi}$

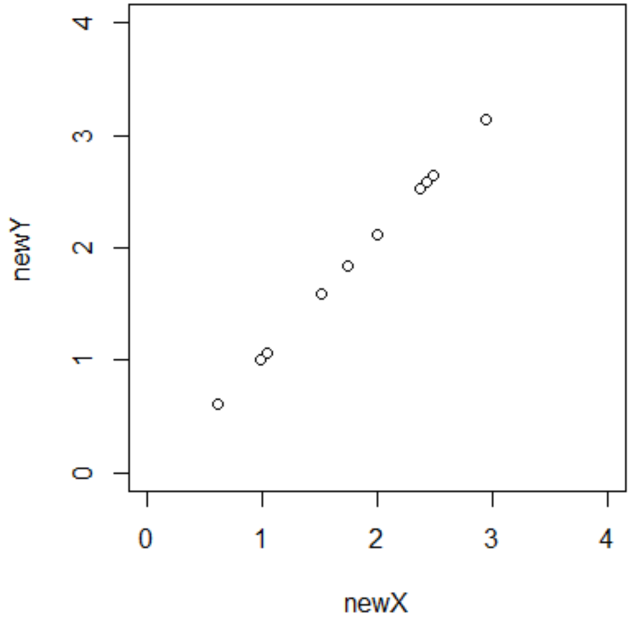
$$\Phi = \left(\boxed{\begin{matrix}0.677873399\\0.735178656\end{matrix}} \quad \begin{matrix}-.735178656\\.677873399\end{matrix}\right)$$

$$\begin{pmatrix} 0.69 & 0.49 \\ -1.31 & -1.21 \\ 0.39 & 0.99 \\ 0.09 & 0.29 \\ 1.29 & 1.09 \\ 0.49 & 0.79 \\ 0.19 & -0.31 \\ -0.81 & -0.81 \\ -0.31 & -0.31 \\ -0.71 & -1.01 \end{pmatrix} \cdot \begin{pmatrix} 0.677873399 \\ 0.735178656 \end{pmatrix} = \begin{pmatrix} 0.827970186 \\ -1.77758033 \\ 0.992197494 \\ 0.274210416 \\ 1.67580142 \\ 0.912949103 \\ -0.0991094375 \\ -1.14457216 \\ -0.438046137 \\ -1.22382056 \end{pmatrix}$$

Reconstruction of Original Data with one EV

- Restore original data: $\widehat{A} \approx \mathbf{Z}\boldsymbol{\Phi}^{\mathrm{T}}$ + original mean
  - $\mathbf{A} \approx$ PCA Scores · Eigenvectors + original mean

$\mathbf{Z}$	$\boldsymbol{\Phi}$
0.827970186	0.6778734
-1.77758033	0.7351787
0.992197494	
0.274210416	
1.67580142	
0.912949103	
-0.0991094375	
-1.14457216	
-0.438046137	
-1.22382056	



- If we reduce the dimensionality, then, when reconstructing the data, we lose those dimensions we chose to discard.
- Note that if all p Eigenvectors are used, then $\boldsymbol{\Phi}\boldsymbol{\Phi}^{\mathrm{T}} = I$.

Summary of Steps 1-3

1.

$\mathbf{A} =$

Data (A):

x	y
2.5	2.4
0.5	0.7
2.2	2.9
1.9	2.2
3.1	3.0
2.3	2.7
2	1.6
1	1.1
1.5	1.6
1.1	0.9

2.

$\mathbf{X} =$

Zero mean Data (X)

x	y
.69	.49
-1.31	-1.21
.39	.99
.09	.29
1.29	1.09
.49	.79
.19	-.31
-.81	-.81
-.31	-.31
-.71	-1.01

3.

$$\sum = \begin{pmatrix} 0.616555556 & 0.615444444 \\ 0.615444444 & 0.716555556 \end{pmatrix}$$

4.

$$Eigenvalues = \begin{pmatrix} .0490833989 \\ 1.28402771 \end{pmatrix}$$

5.

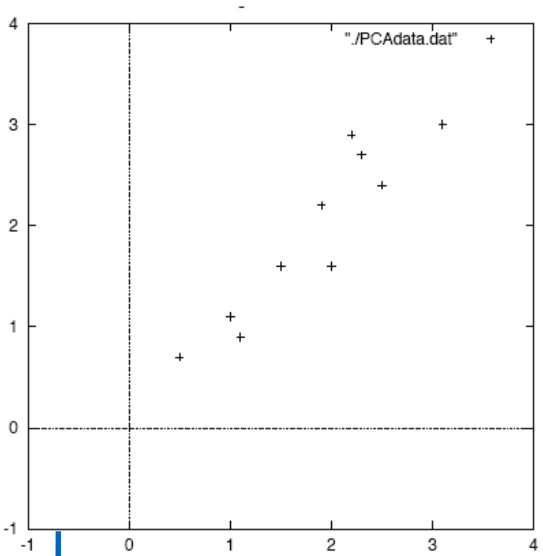
$$Eigenvectors = \begin{pmatrix} -.735178656 & 0.677873399 \\ 0.677873399 & 0.735178656 \end{pmatrix}$$

Then order Eigenvectors after decreasing Eigenvalues in matrix

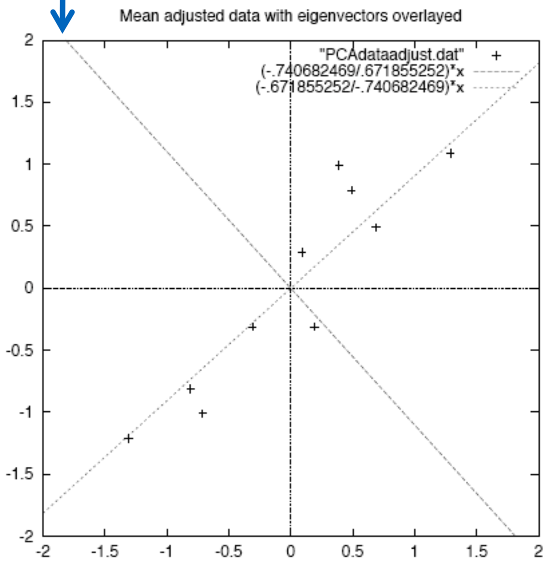
$\mathbf{\Phi}$, the principal components: $\mathbf{\Phi} = \begin{pmatrix} 0.677873399 & -.735178656 \\ 0.735178656 & 0.677873399 \end{pmatrix}$

6.

$\mathbf{A} =$



$\mathbf{X} =$



PCA and Aggregation of Attributes

Matrix $\mathbf{\Phi}$ also allows aggregating similar attributes.

Each element of the Eigenvectors represents the contribution of a given variable to a component.

- In the example below the attributes volume, length, width, and depth all have high impact on the first component. They could as well be considered representations of a latent variable “size”, while speed1 and speed2 could be aggregated to “speed”.
- Similarly, determine **latent variables** such as „social status“ in customer data.

Principal Components – Matrix $\mathbf{\Phi}$

	1	2	3	4	5	6	7	8
Volume	**0.948**	−0.094	−0.129	0.228	0.040	0.036	0.136	0.055
Length	**0.906**	0.302	−0.064	−0.209	0.128	−0.144	−0.007	−0.050
Width	**0.977**	−0.128	−0.031	0.032	0.103	−0.017	−0.014	0.129
Depth	**0.934**	−0.276	−0.061	0.014	0.074	0.129	0.154	−0.038
Speed1	0.552	**0.779**	−0.196	−0.133	−0.099	0.143	−0.038	0.018
Speed2	−0.520	**0.798**	−0.157	0.222	0.109	−0.038	0.071	0.004
Radius	0.398	0.311	**0.862**	0.038	0.008	0.022	−0.002	−0.005

Applications to Image Compression

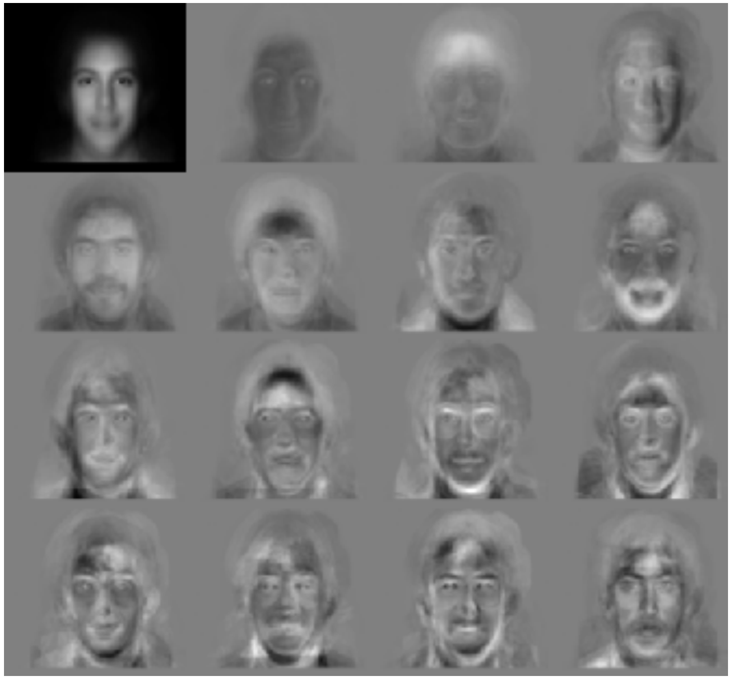
Images:
Here the rows of a matrix are grey scale values of a pixel.

Compression:
Each image can be approximated by projection onto the first few principal components.

Recognition:
Measure difference to existing pictures on the new axes derived from the PCA.

PCA is also known as Karhunen-Loève transformation.

First principal component (i.e., Eigenvector with largest Eigenvalue)



Other principal components

Assumptions of PCA

doesn't work not numeric

PCA assumes relationships among variables are **LINEAR**.

- cloud of points in p-dimensional space has linear dimensions that can be effectively summarized by the principal axes.
- If the structure in the data is **NONLINEAR**, the principal axes will not be an efficient and informative summary of the data.

PCA uses the Euclidean distance among points assuming **continuous variables**. With discrete variables special techniques are in order (e.g., correspondence analysis).

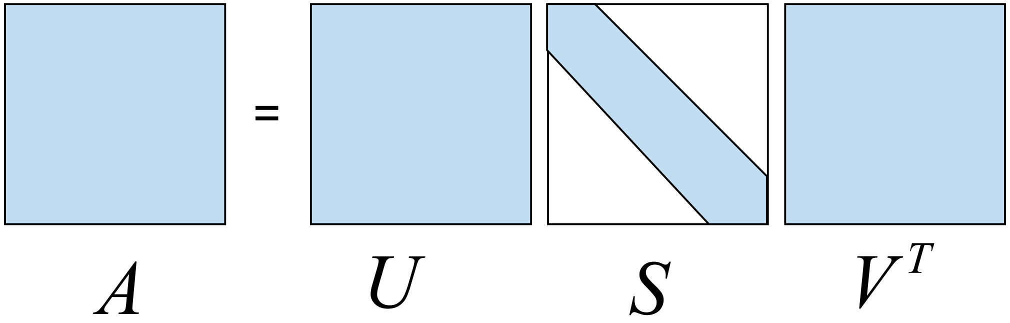
Outline for today

- Overview
- Linear algebra revisited
- Principal Component Analysis
- **Singular Value Decomposition (outlook)**
- PCA regression

Singular Value Decomposition (SVD)

For any matrix $\boldsymbol{A} \in R^{n \times p}$, there exist orthogonal matrices $\boldsymbol{U}, \boldsymbol{V}$ and a diagonal matrix $\boldsymbol{S}$, such that all the diagonal values s_i of $\boldsymbol{S}$ are non-negative.

$$A = USV^T$$



SVD and PCA

The diagonal values of $\boldsymbol{S}$ are called the **singular values**. It is accustomed to sort them by size.

The columns of $\boldsymbol{U}$ are called the **left singular vectors**.

The columns of $\boldsymbol{V}$ are called the **right singular vectors**.

$$A = USV^T$$

Singular values of the SVD decomposition of the matrix $\boldsymbol{A}$ **are the square roots of the Eigenvalues** of the matrix $(\boldsymbol{AA^t})$ or $(\boldsymbol{A^tA})$.

The columns of $\boldsymbol{V}$ are the **principal axes**, while the columns of $\boldsymbol{US}$ are principal component scores of the centered matrix $\boldsymbol{X}$ in PCA.

Singular values are related to the eigenvalues of the covariance matrix via $\boxed{\lambda_i = {s_i^2}/{_{n-1}}.}$

SVD (on Uncentered Data)

Data $\boldsymbol{A}$ =

10	20	10
2	5	2
8	17	7
9	20	10
12	22	11

1st row of $\boldsymbol{A} = [10\ 20\ 10]$

Since $\boldsymbol{A} = U\, S\, V^t$,

then $A(1,) = U(1,) \cdot \boldsymbol{S} \cdot V^t$

$= [24.3\ \ 0.21\ \ -0.228] \cdot V^t$

→ $A(1,) = 24.3\, v_1 + 0.21\, v_2 + -0.228\, v_3$

where v_1, v_2, v_3 are rows of V^t

$[24.3, 0.21, -0.228] \cdot [0.41, 0.73, 0.55]^T = \mathbf{10}$

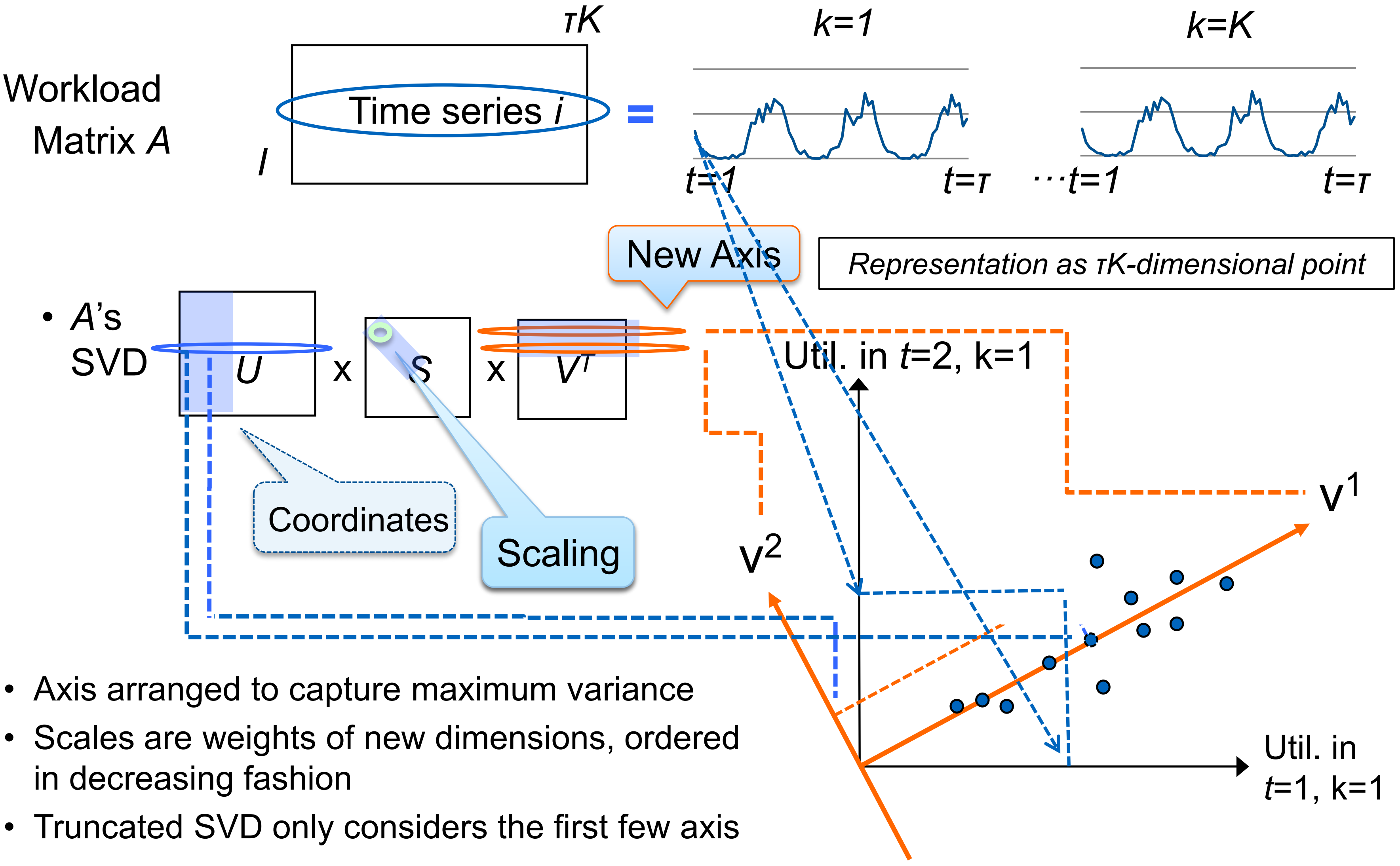
$A = U\, S\, V^t$

where U = $\begin{matrix} 0.50 & 0.14 & -0.19 \\ 0.12 & -0.35 & 0.07 \\ 0.41 & -0.54 & 0.66 \\ 0.49 & -0.35 & -0.67 \\ 0.56 & 0.66 & 0.27 \end{matrix}$

where S = $\begin{matrix} 48.6 & 0 & 0 \\ 0 & 1.5 & 0 \\ 0 & 0 & 1.2 \end{matrix}$

and V^t = $\begin{matrix} 0.41 & 0.82 & 0.40 \\ 0.73 & -0.56 & 0.41 \\ 0.55 & 0.12 & -0.82 \end{matrix}$

Example: Time Series of I Customers



- Axis arranged to capture maximum variance
- Scales are weights of new dimensions, ordered in decreasing fashion
- Truncated SVD only considers the first few axis

Computing SVD

$$A = USV^T$$

- $\boldsymbol{U}$ and $\boldsymbol{V}$ are orthogonal matrices, such that $\boldsymbol{V^T} = \boldsymbol{V^{-1}}$ *and* $\boldsymbol{U^T} = \boldsymbol{U^{-1}}$
- Finding $\boldsymbol{U}, \boldsymbol{S}, \boldsymbol{V}$ requires finding Eigenvectors of $\boldsymbol{A^t A}$.
  - det($\boldsymbol{A^t A} - \lambda \boldsymbol{I}$) gives us the Eigenvalues λ of $\boldsymbol{A^t A}$.
- The corresponding Eigenvectors are found by using these values of λ in the equation $(\mathbf{A} - \lambda \boldsymbol{I})v = 0$, providing $\boldsymbol{V}$.
- $\boldsymbol{S}$ has the square roots of the Eigenvalues λ in the diagonals.
- Knowing that $\boldsymbol{AV} = \boldsymbol{US}$, one can derive $\boldsymbol{U}$.

Today, there are more advanced numerical techniques to compute the SVD more effectively.

Principal Components: Summary

PCA takes a data matrix of n objects by p variables, which may be correlated, and summarizes it by uncorrelated axes (principal components or principal axes) that are linear combinations of the original p variables.

The first k components display as much as possible of the variation in the data.

The first PC is the direction of maximum variance from origin. Subsequent PCs are orthogonal to first PC and describe maximum residual variance.

PCA can be quite useful to combat multicollinearity in the regression analysis.

How do Eigenvectors play a role in Principal Component Analysis?

Outline for today

- Overview
- Linear algebra revisited
- Principal Component Analysis
- Singular Value Decomposition
- **PCA regression**

Reminder: Multiple Linear Regression

Equation: $\mathrm{Y} = \boldsymbol{X}\beta + \varepsilon$

- Y: $n \times 1$ vector of observed values
- $\boldsymbol{X}$: $n \times p$ matrix of independent values
- β: $p \times 1$ vector of regression parameters
- ε: $n \times 1$ vector of residuals

OLS estimator: $\hat{\beta} = (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}y$

Multicollinearity in Linear Regression Models

$\min(\|\boldsymbol{X}\beta - y\|^2)$

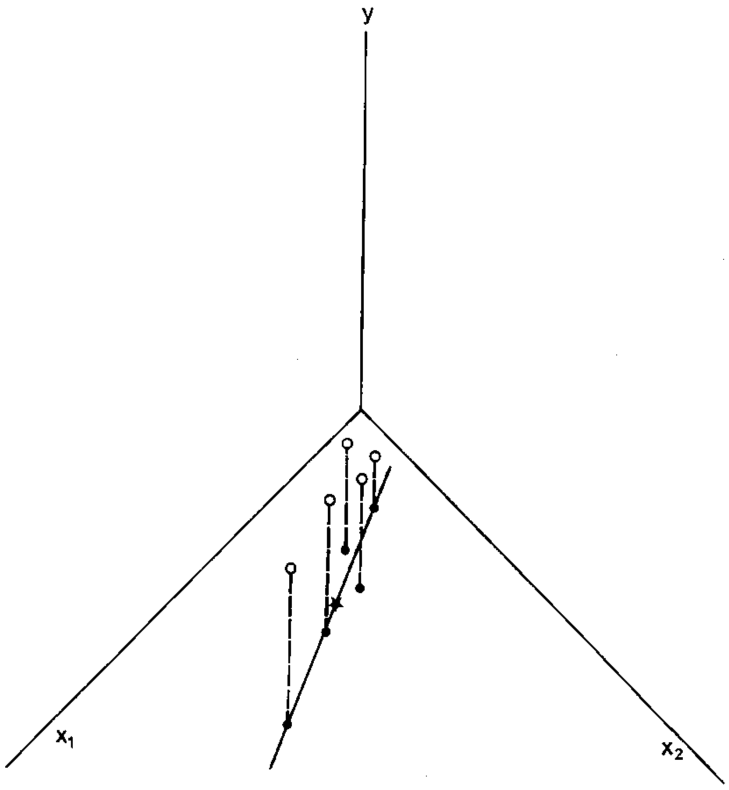
MLR solution, requires $n \geq p$
(variable selection) and nearly orthogonal $\boldsymbol{X}$.

If $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$ is not full rank:

- No unique solution to normal equations.

If the columns of $\boldsymbol{X}$ are highly correlated:

- Leads to unstable equation/plane
  in the direction with little variability.



Considerations in High Dimensions

While p can be extremely large, the number of observations n is often limited due to cost, sample availability, etc.

Data sets containing more features than observations are often referred to a **high-dimensional**.

When the number of features p is as large as, or larger than, the number of observations n, OLS should not be performed.

- It is too **flexible** and hence overfits the data.

Forward stepwise selection, ridge regression, lasso, and PCR are particularly useful for performing regression in the high-dimensional setting.

Solutions to Multicollinearity

Subset selection

- best subset, backward, forward, stepwise selection of features
- (already discussed in the context of the linear regression)

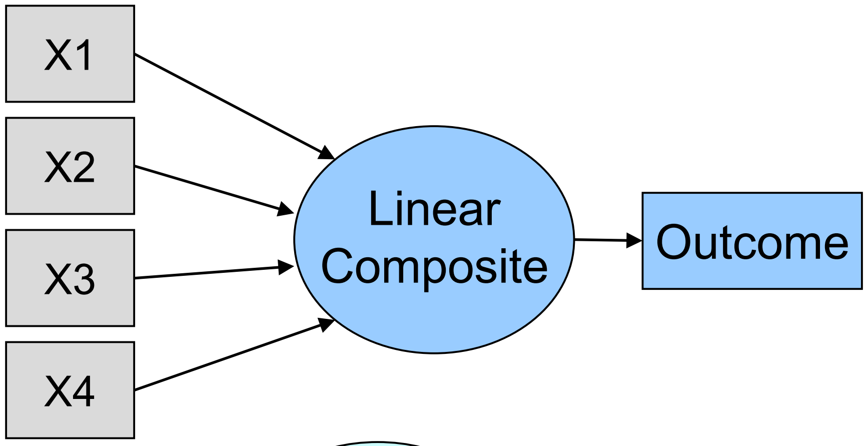
Using derived input

- **Principal component regression**
- Partial least squares

Coefficient shrinkage (regularization) => next class

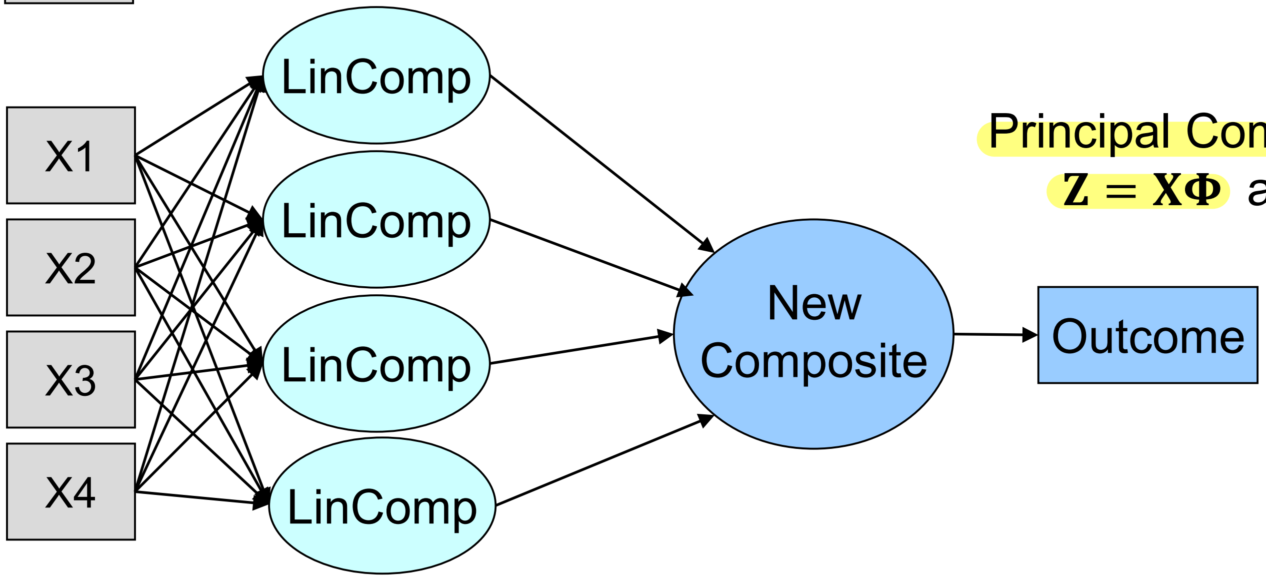
- Ridge regression
- Lasso (least absolute shrinkage and selection operator)

Multiple Regression vs. Principal Component Regression



Multiple Linear Regression

$$y = \mathbf{X}\beta + \varepsilon$$



Principal Component Regression

$\mathbf{Z} = \mathbf{X\Phi}$ and $y = \mathbf{Z}\gamma + \varepsilon$

learn regression on the PC features

Back to our Example

$$\begin{pmatrix} 0.69 & 0.49 \\ -1.31 & -1.21 \\ 0.39 & 0.99 \\ 0.09 & 0.29 \\ 1.29 & 1.09 \\ 0.49 & 0.79 \\ 0.19 & -0.31 \\ -0.81 & -0.81 \\ -0.31 & -0.31 \\ -0.71 & -1.01 \end{pmatrix} \cdot \begin{pmatrix} 0.677873399 \\ 0.735178656 \end{pmatrix} = \begin{pmatrix} 0.827970186 \\ -1.77758033 \\ 0.992197494 \\ 0.274210416 \\ 1.67580142 \\ 0.912949103 \\ -0.0991094375 \\ -1.14457216 \\ -0.438046137 \\ -1.22382056 \end{pmatrix} \qquad \mathbf{X} \cdot \mathbf{\Phi} = \mathbf{Z}$$

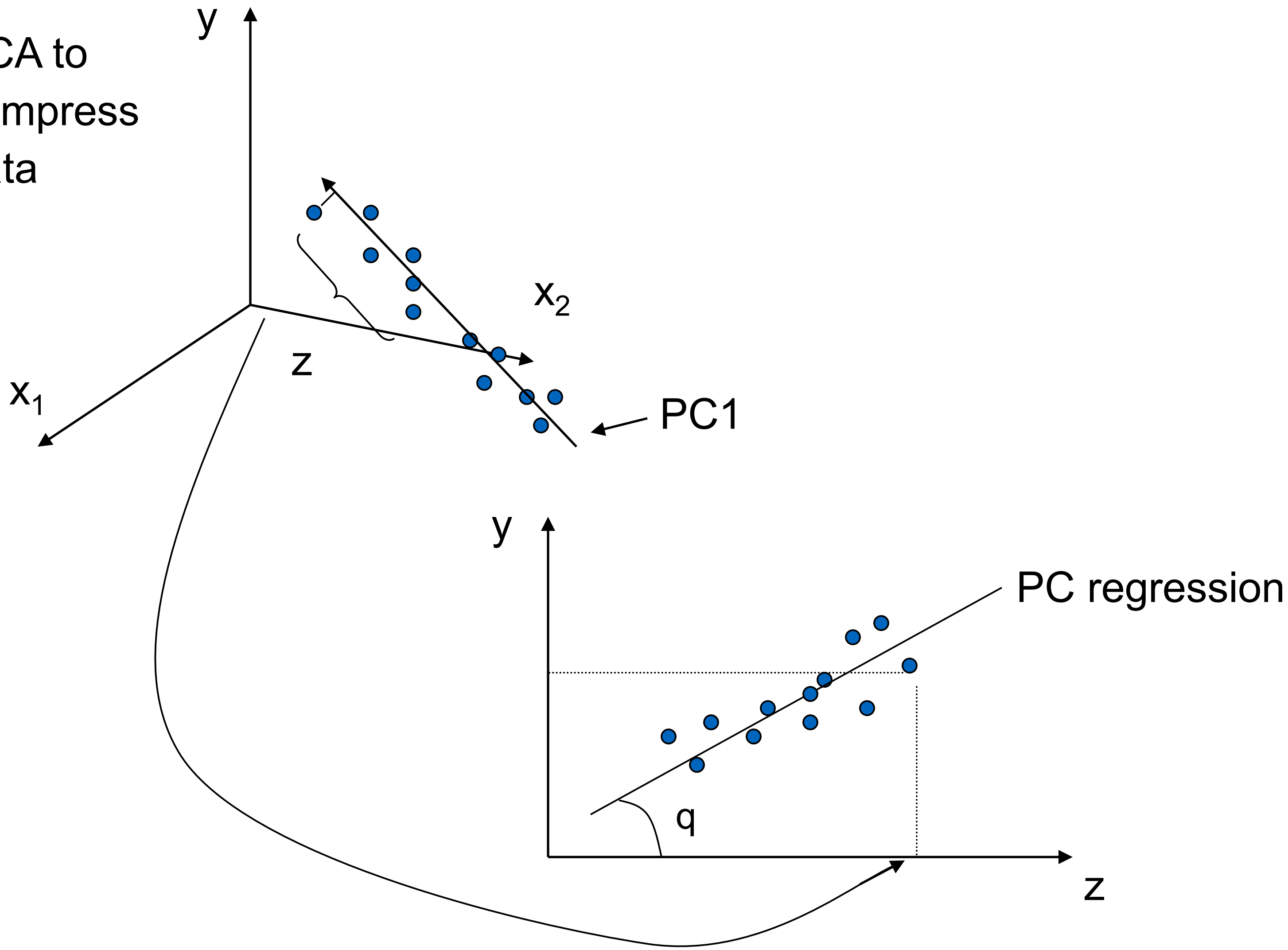
PC regression: $y = \mathbf{Z}\gamma + \varepsilon$

Now, we need to estimate γ via the OLS estimator $\hat{\gamma} = (\mathbf{Z}^t\mathbf{Z})^{-1}\mathbf{Z}^t y$

- The independent variables are now the principal components in $\mathbf{Z}$.
- If only a subset of the principal components in $\mathbf{Z}$ is used for the regression, the coefficients in γ remain the same as the PCs are orthogonal.
- Remember, the PCs are linear combinations of all original variables.

Principal Component Regression

PCA to compress data



Principal Components Regression

PCR will tend to do well in cases when the first few principal components are sufficient to capture most of the variation in the predictors as well as the relationship with the response.

Note that even though PCR provides a simple way to perform regression using $k < p$ predictors, it *is not* a feature selection method.

In PCR, the number of principal components is typically chosen by cross-validation.

Regularization (Next Class)

If the linear model is correct for a given problem, then the OLS prediction is unbiased, and has the lowest variance among all linear unbiased estimators.

But there can be (and often exist) biased estimators with smaller MSE.

Generally, by **regularizing** the estimator in some way, its variance will be reduced; if the corresponding increase in bias is small, this will be worthwhile.

Examples of regularization:

- subset selection (forward, backward, all subsets)
- ridge regression
- the lasso

Next class