

SVD, Multivariate Gaussians, Principal Component Analysis (PCA) and Dimensionality Reduction

Michael S. Murillo

Computational Mathematics, Science and Engineering
Michigan State University

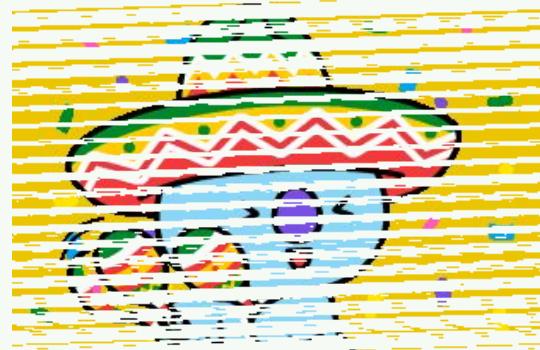


MidTerm Projects

Congratulations on your project

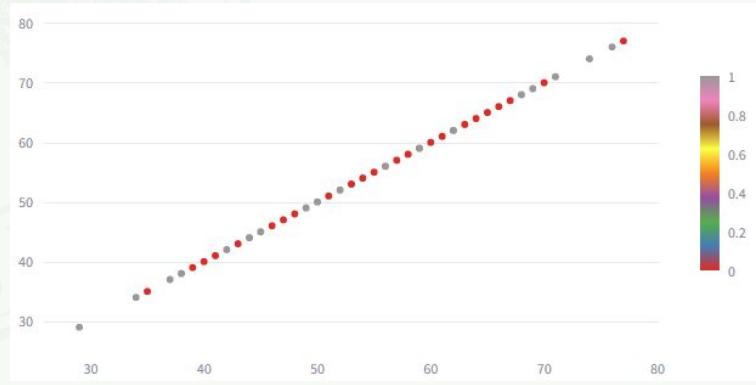
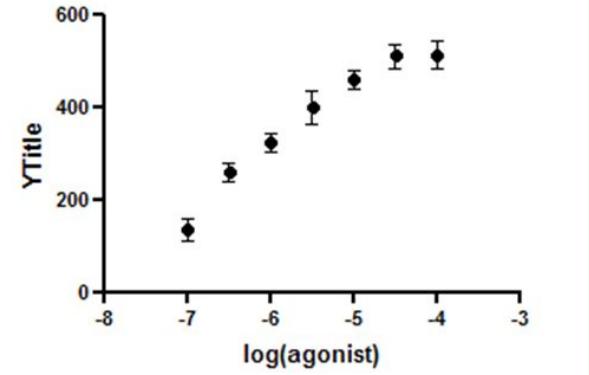
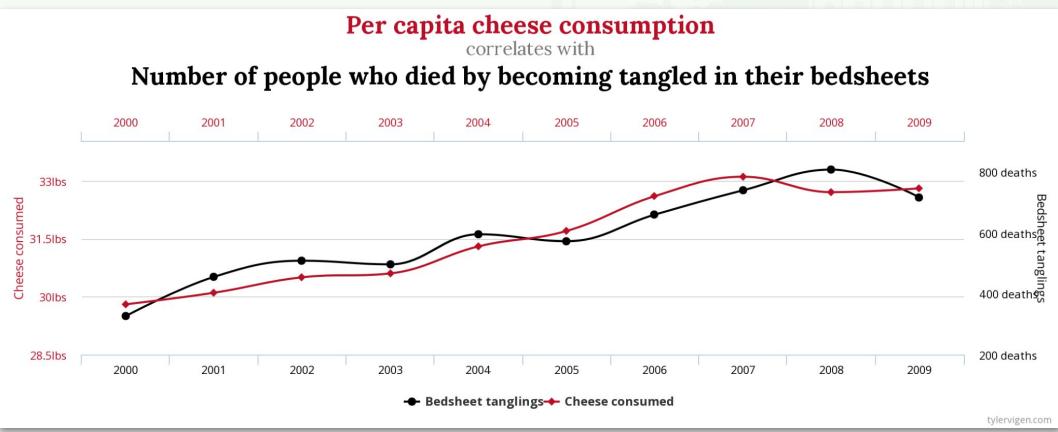
Comments:

- ❖ Project is not assignment, it is a showcase
 - Programming ability
 - Creativity
 - Problem Solving



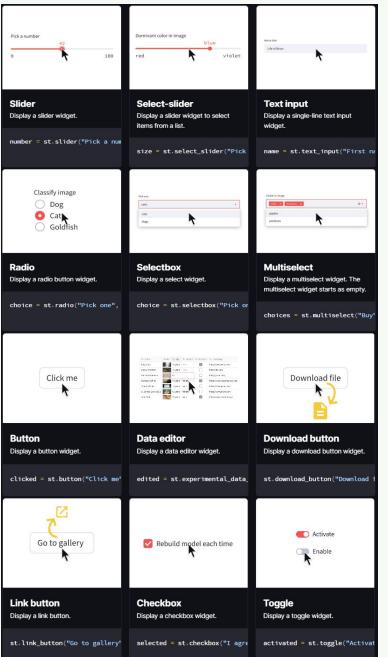
MidTerm Projects

Visualization is the proof for your project claims, but alone it doesn't carry anything



MidTerm Projects

Webapp Design



□ Section 1

□ Section 2

□ Section 3

□ Tab 1

□ Tab 2

□ Tab 3



The federal funds rate, the interest rate at which banks lend to each other that determines borrowing costs for consumers, has gone from probably zero in spring of 2020 to 5.25% to 5.5%. Source: Board of Governors of the Federal Reserve System. [Read more](#)

Covid-up disrupted supply chains and led to soaring inflation. The Federal Reserve embarked on an aggressive campaign beginning about a year and a half ago to combat that, significantly raising interest rates.

Debt is now much more expensive. The average 30-year fixed mortgage rate jumped 7% in August, reaching its highest level in 21 years.

A rule of thumb definition for a recession is two consecutive quarters of declining GDP, which did occur last year. Another key indicator is an inverted yield curve, where short-term rates are higher than long-term ones, which has been the case since 2022.

Yet, the economy has remained more resilient than expected. Gross Domestic Product, or GDP, grew faster than anticipated in the second quarter at 6.6%, the fourth consecutive quarter of growth. Third-quarter GDP data won't be until later this month, but forecasters surveyed by the Federal Reserve Bank of Philadelphia predict higher output for the next three quarters than they did previously.

Looking at the labor market, the unemployment rate is historically low at 3.8%. While hiring seemed to be decelerating in August, Friday's report from the Bureau of Labor Statistics show the economy added 355,000 jobs in September, far more than the 175,000 economists expected.



The New-York Times.

SATURDAY BOOK REVIEW SUPPLEMENT.



- The story must make sense
- The option should affect unnecessary materials

Calendar



Lecture and ICA
on PCA, SD



Project
Presentation

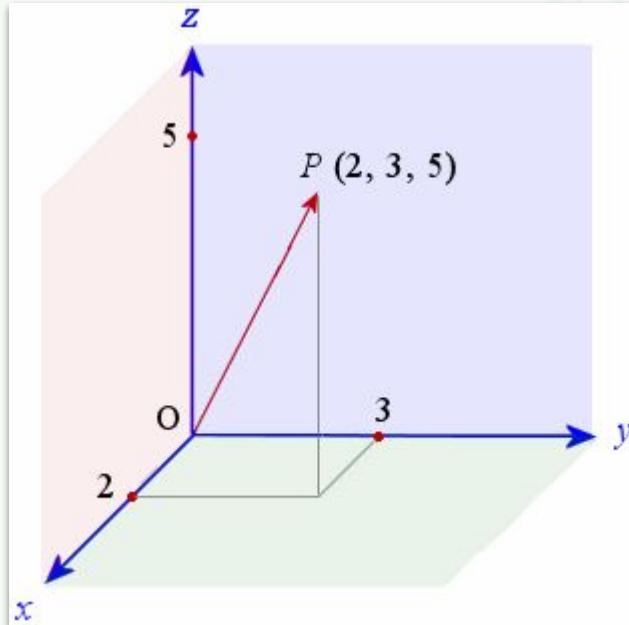
Project
Submission

Class Ends!



Vectors

In engineering and physical sciences, vectors are used to describe quantities that have both a magnitude/size and a direction.



Examples of vectors are:

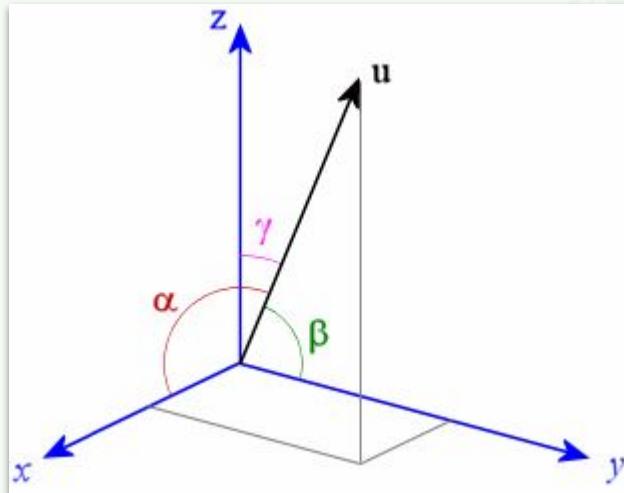
- velocity (which is speed and a direction)
- distance
- force

Vectors have components, usually in our usual 3D physical space.

Dot Products from Vector Analysis

Important for DS, we are interested in the “projection” of the vector on the axes.

This is equivalent to, say, “*what is the x velocity?*”



$$\vec{a} = (a_1 \vec{i} + a_2 \vec{j} + a_3 \vec{k})$$

$$\vec{b} = (b_1 \vec{i} + b_2 \vec{j} + b_3 \vec{k})$$

$$\vec{a} \cdot \vec{b} = (\vec{a}_1 \vec{i} + \vec{a}_2 \vec{j} + \vec{a}_3 \vec{k}) \cdot (\vec{b}_1 \vec{i} + \vec{b}_2 \vec{j} + \vec{b}_3 \vec{k})$$

$$= a_1 b_1 + a_2 b_2 + a_3 b_3$$

A 3D Cartesian coordinate system with axes labeled x, y, and z. A vector a is shown originating from the origin. It is decomposed into components along the axes, represented by unit vectors \vec{i} , \vec{j} , and \vec{k} . The components are labeled a_x , a_y , and a_z respectively.

$$\mathbf{a} \cdot \mathbf{b} = |\mathbf{a}| |\mathbf{b}| \cos \theta$$
$$\cos \theta = \frac{\mathbf{a} \cdot \mathbf{b}}{|\mathbf{a}| |\mathbf{b}|}$$

Dot Products are Inner Products

$$\vec{A}^T = \begin{bmatrix} A_1 & A_2 & A_3 \end{bmatrix} \quad \vec{B} = \begin{bmatrix} B_1 \\ B_2 \\ B_3 \end{bmatrix}$$

$$\begin{bmatrix} A_1 & A_2 & A_3 \end{bmatrix} \begin{bmatrix} B_1 \\ B_2 \\ B_3 \end{bmatrix} = A_1 B_1 + A_2 B_2 + A_3 B_3 = \vec{A} \cdot \vec{B}$$

$$\begin{aligned}\vec{a} &= (a_1\vec{i} + a_2\vec{j} + a_3\vec{k}) \\ \vec{b} &= (b_1\vec{i} + b_2\vec{j} + b_3\vec{k}) \\ \vec{a} \cdot \vec{b} &= (a_1\vec{i} + a_2\vec{j} + a_3\vec{k}) \cdot (b_1\vec{i} + b_2\vec{j} + b_3\vec{k}) \\ &= a_1b_1 + a_2b_2 + a_3b_3\end{aligned}$$

Matrix multiplication can be interpreted as a dot product (or, projection).

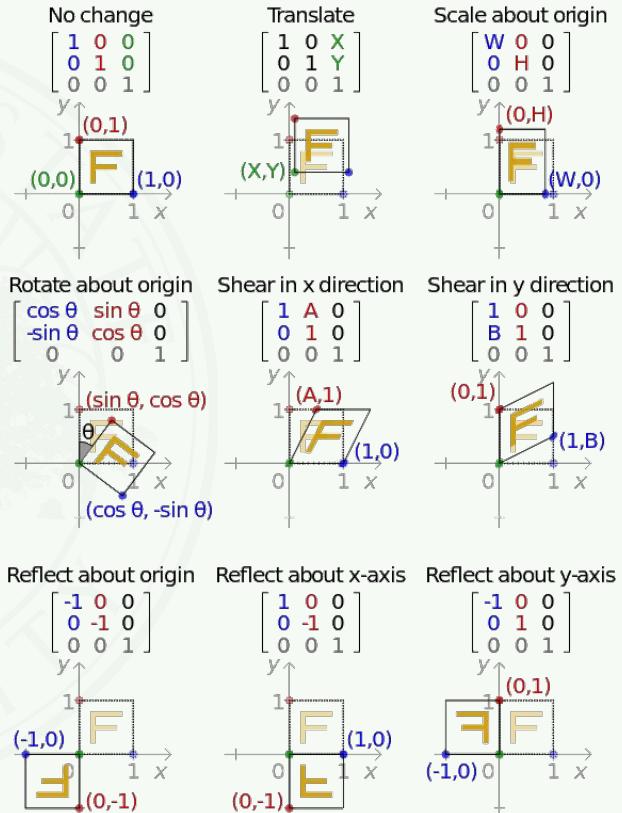
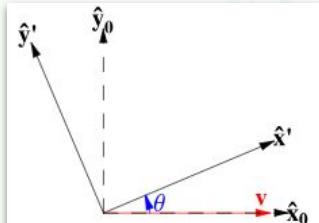


Matrices as Transformations

Matrix multiplication of a vector can be seen as a transformation of that vector to a new vector.

Rotation is just one type of transformation:

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{bmatrix} \cdot \begin{bmatrix} x \\ y \end{bmatrix}$$



Diagonalization of a Matrix

Consider a square matrix A . We can perform a “similarity transformation” to connect A to another matrix D .

We will assume that D is a diagonal matrix.

$$A = PDP^{-1}$$

- d_i are the “eigenvalues”
- p_i are the “eigenvectors”

$$AP = PD$$

$$\begin{bmatrix} a_{11} & a_{12} & \dots \\ a_{21} & a_{22} & \dots \\ \vdots & \vdots & \dots \end{bmatrix} \begin{bmatrix} | & | & \dots \\ p_1 & p_2 & \dots \\ | & | & \dots \end{bmatrix} = \begin{bmatrix} | & | & \dots \\ p_1 & p_2 & \dots \\ | & | & \dots \end{bmatrix} \begin{bmatrix} d_1 & 0 & \dots \\ 0 & d_2 & \dots \\ \vdots & \vdots & \dots \end{bmatrix}$$

$$Ap_i = d_i p_i \quad \text{stretch a vector with no rotation}$$



Diagonalization As a Change of Basis

1. A transforms \mathbf{x} into \mathbf{y}

$$A\mathbf{x} = \mathbf{y}$$

2. solve for x , given A and y

$$A = PDP^{-1}$$

$$PDP^{-1}\mathbf{x} = \mathbf{y}$$

$$D\boxed{P^{-1}\mathbf{x}} = \boxed{P^{-1}\mathbf{y}}$$

original data projected onto eigenvalues

components act independently in new basis

$$D\mathbf{x}' = \mathbf{y}'$$

$$\mathbf{x}' = D^{-1}\mathbf{y}'$$

$$\mathbf{x} = PD^{-1}P^{-1}\mathbf{y}$$



Statistical Properties of the Data Matrix X

We have explored statistics a bit earlier in the course.

Let's connect the statistical properties of the data to newer ideas, such as the SVD.



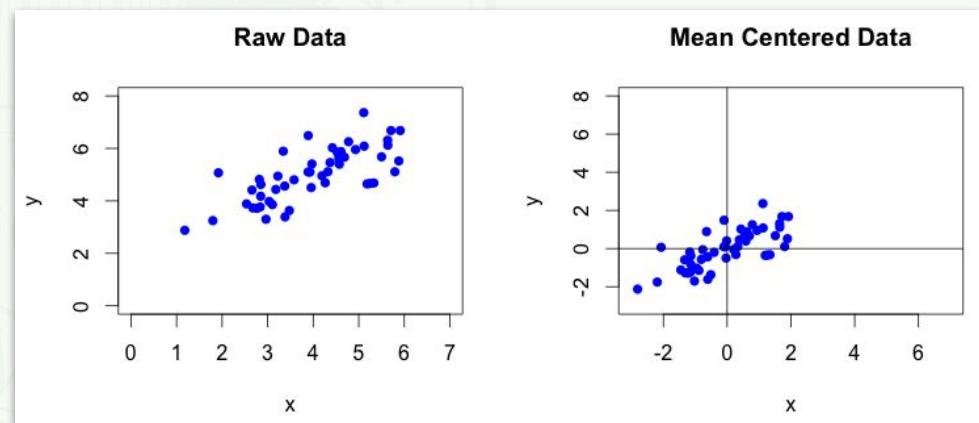
Remove Mean From Data Matrix

$$X = \begin{bmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \\ x_{31} & x_{32} \end{bmatrix}$$

↑ ↓
features samples

$$\mu_1 \quad \mu_2$$

$$X' = \begin{bmatrix} x_{11} - \mu_1 & x_{12} - \mu_2 \\ x_{21} - \mu_1 & x_{22} - \mu_2 \\ x_{31} - \mu_1 & x_{32} - \mu_2 \end{bmatrix}$$



Covariance Reminder

$$\begin{aligned}\text{cov}(X, Y) &= \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] \\ &= \mathbb{E}[XY - X\mathbb{E}[Y] - \mathbb{E}[X]Y + \mathbb{E}[X]\mathbb{E}[Y]] \\ &= \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] - \mathbb{E}[X]\mathbb{E}[Y] + \mathbb{E}[X]\mathbb{E}[Y] \\ &= \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y],\end{aligned}$$



Variance and Covariance

Here, X is mean centered.

Recall that the covariance is defined in terms of centered data.

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

$$\begin{bmatrix} \text{Var}(x_1) & \dots & \text{Cov}(x_n, x_1) \\ \vdots & \ddots & \vdots \\ \text{Cov}(x_n, x_1) & \dots & \text{Var}(x_n) \end{bmatrix}$$

Let's use the smallest non-trivial data matrix.

$$X = \begin{bmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \\ x_{31} & x_{32} \end{bmatrix}$$

$$\begin{aligned} X^T X &= \begin{bmatrix} x_{11} & x_{21} & x_{31} \\ x_{12} & x_{22} & x_{32} \end{bmatrix} \begin{bmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \\ x_{31} & x_{32} \end{bmatrix}, \\ &= \begin{bmatrix} x_{11}^2 + x_{21}^2 + x_{31}^2 & x_{11}x_{12} + x_{21}x_{22} + x_{31}x_{32} \\ x_{12}x_{11} + x_{22}x_{21} + x_{32}x_{31} & x_{12}^2 + x_{22}^2 + x_{32}^2 \end{bmatrix} \end{aligned}$$



Note on Normalization and Choice of Transpose

You don't always see this definition.

You might also see:

- no factor of $n-1$,
- a factor of n instead,
- transpose might be reversed.

Note that these C are square,
but have different meanings.

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

$$C_1 = X^T X$$

$$C_2 = X X^T$$

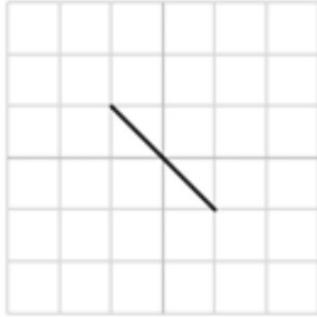


Geometric Interpretation I

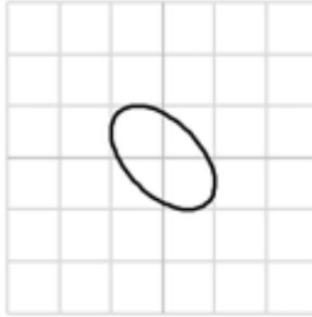
Correlation matrix

Correlation matrix has 1 in diagonals and values between -1 and 1 inclusive in off-diagonals. Ellipse size remains the same (always touches square of side 2 units). Whether the correlation is positive or negative can be observed by the orientation of the ellipse. The amount of correlation can be interpreted by how thin the ellipse is.

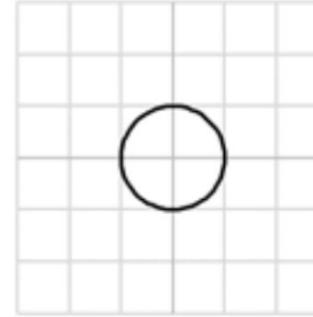
$$\begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}$$



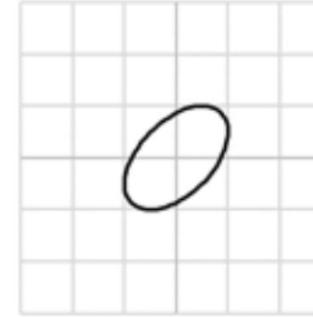
$$\begin{bmatrix} 1 & -0.5 \\ -0.5 & 1 \end{bmatrix}$$



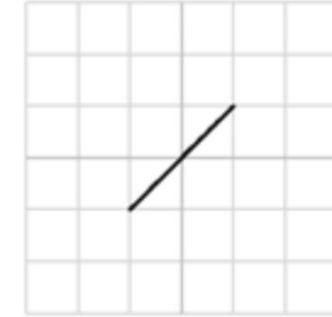
$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$



$$\begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$$



$$\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$$

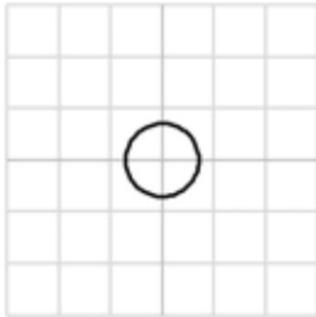


Geometric Interpretation II

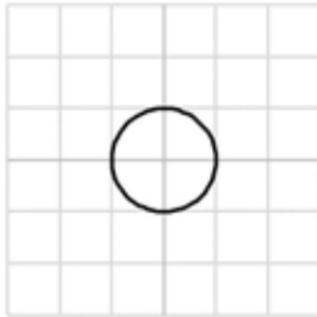
Diagonal matrix

Zeros in off-diagonals means zero correlation. Ellipse axes are parallel to coordinate axes (no rotation).

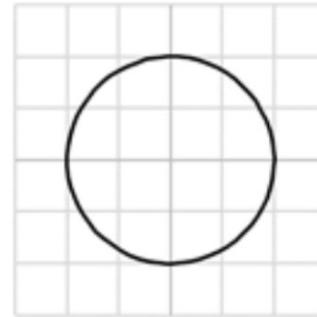
$$\begin{bmatrix} 0.5 & 0 \\ 0 & 0.5 \end{bmatrix}$$



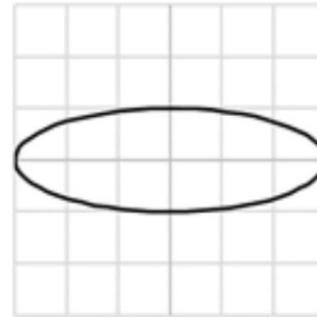
$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$



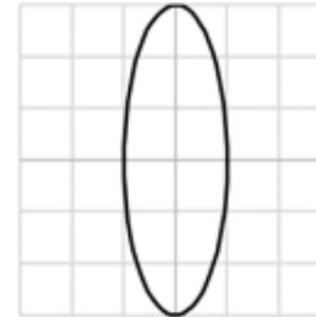
$$\begin{bmatrix} 4 & 0 \\ 0 & 4 \end{bmatrix}$$



$$\begin{bmatrix} 9 & 0 \\ 0 & 1 \end{bmatrix}$$



$$\begin{bmatrix} 1 & 0 \\ 0 & 9 \end{bmatrix}$$

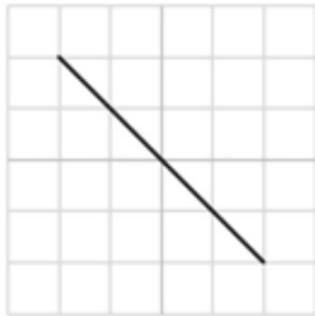


Geometric Interpretation III

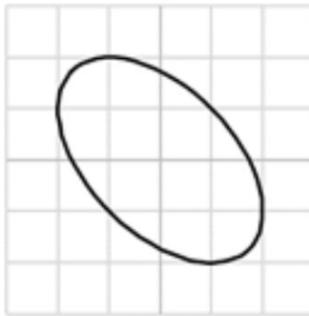
Same values in diagonals

Ellipse is rotated 45 degrees if correlation (off-diagonal) is positive regardless of its magnitude. Similarly -45 degrees if correlation is negative.

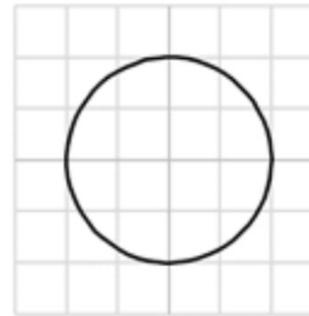
$$\begin{bmatrix} 4 & -4 \\ -4 & 4 \end{bmatrix}$$



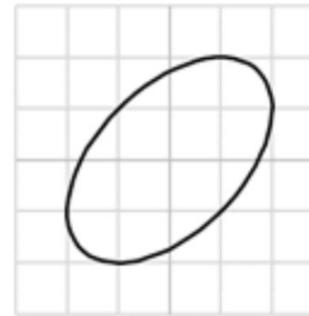
$$\begin{bmatrix} 4 & -2 \\ -2 & 4 \end{bmatrix}$$



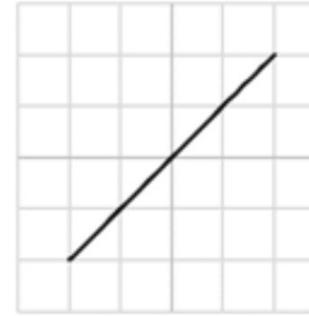
$$\begin{bmatrix} 4 & 0 \\ 0 & 4 \end{bmatrix}$$



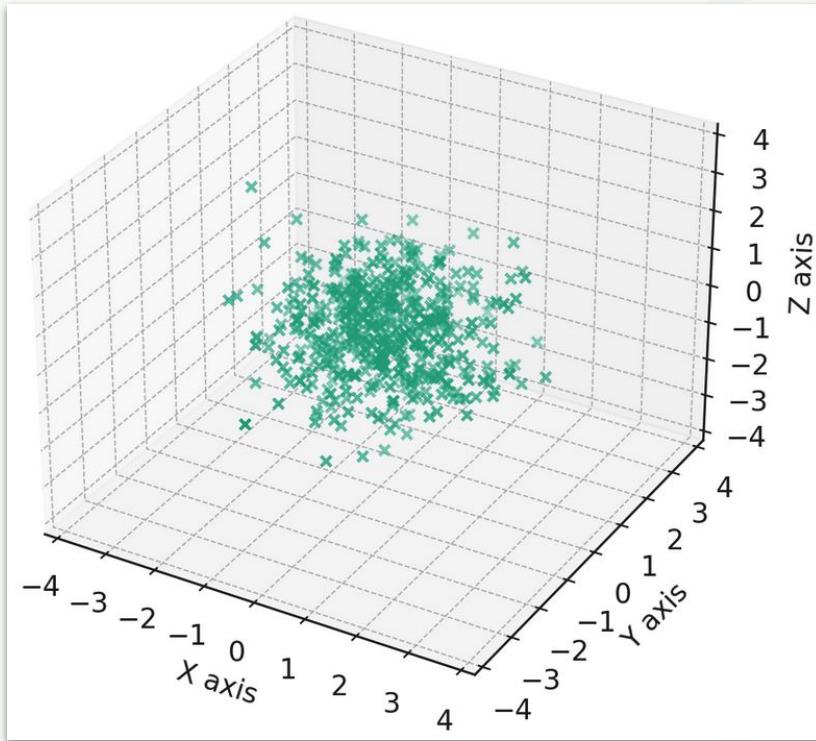
$$\begin{bmatrix} 4 & 2 \\ 2 & 4 \end{bmatrix}$$



$$\begin{bmatrix} 4 & 4 \\ 4 & 4 \end{bmatrix}$$



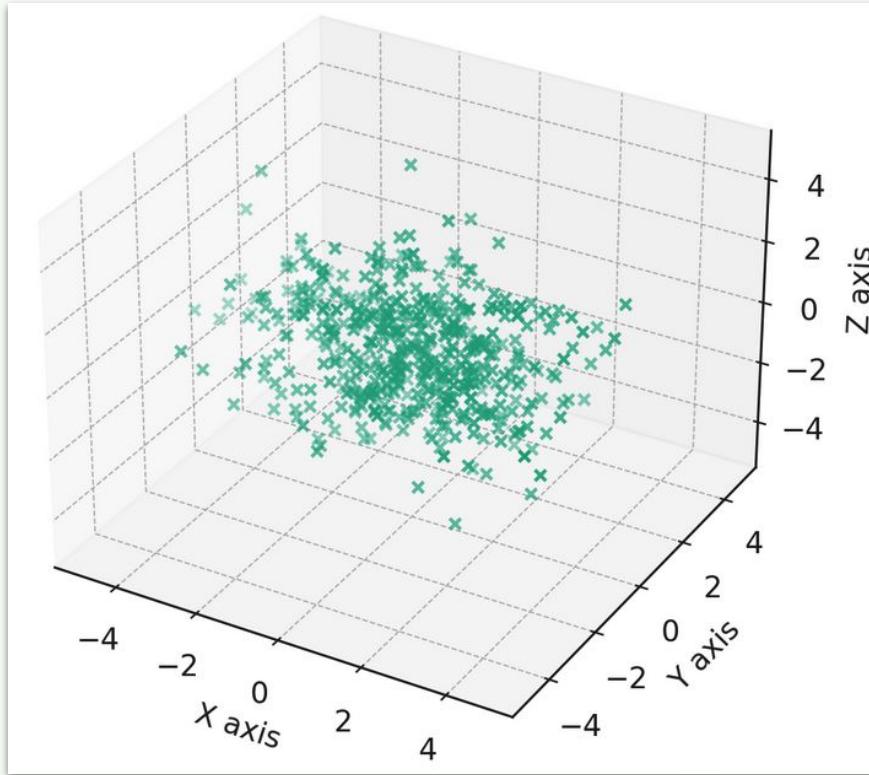
Let's examine some data sets!



covariance matrix

$$\begin{pmatrix} 0.996 & -0.039 & 0.028 \\ -0.039 & 0.998 & -0.098 \\ 0.028 & -0.098 & 0.982 \end{pmatrix}$$

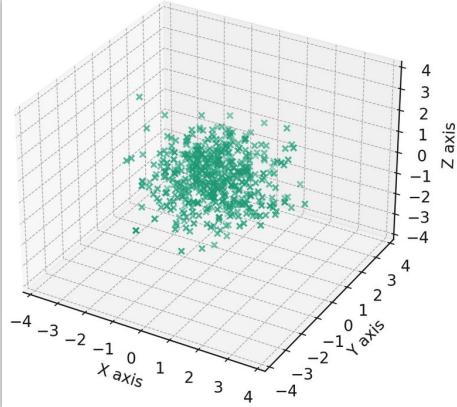
Let's examine some data sets!



covariance matrix

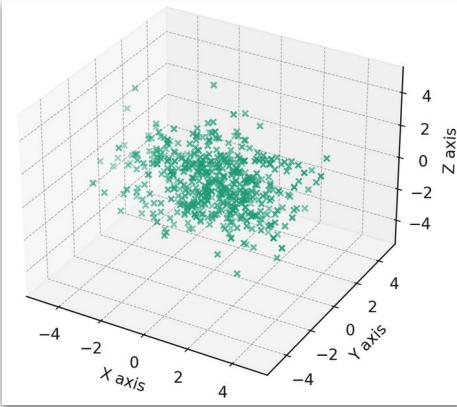
$$\begin{pmatrix} 3.204 & 0.113 & -0.046 \\ 0.113 & 0.911 & 0.092 \\ -0.046 & 0.092 & 1.976 \end{pmatrix}$$

Let's examine some data sets!



$$\begin{pmatrix} 0.996 & -0.039 & 0.028 \\ -0.039 & 0.998 & -0.098 \\ 0.028 & -0.098 & 0.982 \end{pmatrix}$$

Note that these covariance matrices are already (approximately) diagonal.



$$\begin{pmatrix} 3.204 & 0.113 & -0.046 \\ 0.113 & 0.911 & 0.092 \\ -0.046 & 0.092 & 1.976 \end{pmatrix}$$

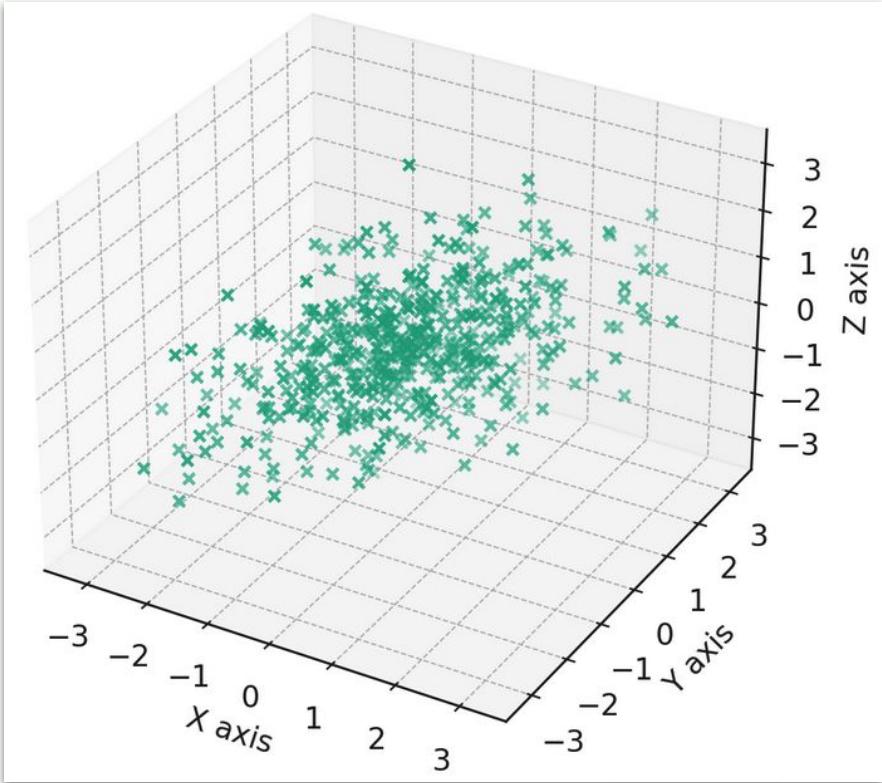
$$f(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^k \det(\Sigma)}} \exp \left(-\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right)$$

↗

covariance matrix



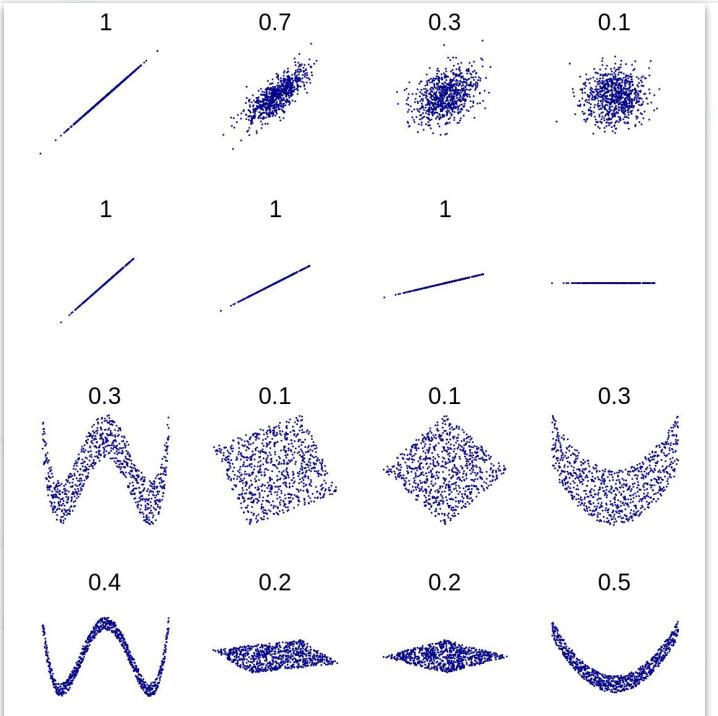
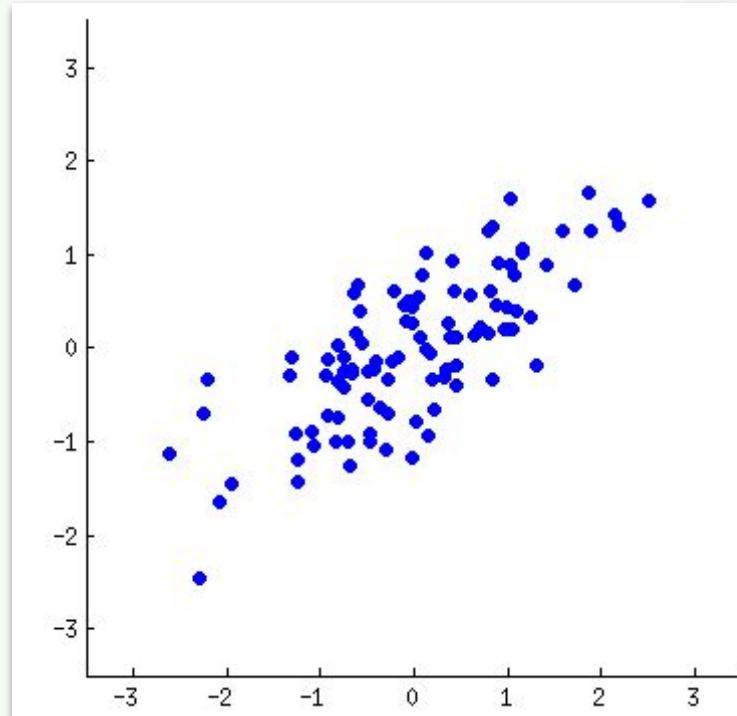
Non-Diagonal Case



$$\begin{pmatrix} 1.052 & 0.821 & 0.262 \\ 0.821 & 0.958 & 0.461 \\ 0.262 & 0.461 & 0.979 \end{pmatrix}$$



Diagonalizing the Covariance Matrix: Why?



Diagonalizing the Covariance Matrix: Math

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

$$= PDP^{-1}$$

$$P^{-1}\mathcal{C}P = D$$

$$P^{-1}X^T X P = D$$

$$P^T X^T X P = (XP)^T XP$$



SVD of Covariance Matrix

$$X^T X = (U \Sigma V^T)^T U \Sigma V^T,$$

$$= V \Sigma^T U^T U \Sigma V^T,$$

$$= V \Sigma^T \Sigma V^T,$$

$$= V \Sigma^2 V^T,$$

$$= V \Lambda V^T,$$

Note the importance of the matrix V in the covariance.

When we write a matrix in this form, we say we have “diagonalized” the matrix.

The diagonal matrix Λ contains the eigenvalues, which are the squares of the singular values.



Expand Correlation in Singular Values (SVs)

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

$$= V \Lambda V^T$$

$$= \sum_i \lambda_i v_i v_i^T$$

$$\lambda_i = \sigma_i^2$$

Each vector in V “explains” some of the variance in the dataset.

The first column of V contributes the most because it is associated with the largest SV in the sum.



New Coordinate System: Project Onto V

How can we understand this result?

$$X^T X = V \Sigma^2 V^T,$$

$$V^T X^T X V = \Sigma^2,$$

$$(X V)^T X V = \Sigma^2,$$

$$(Y)^T Y = \Sigma^2$$



PCA: Columns of V

The columns of V are called the “*principal components*” (directions) of X .

The first principal component of X is:

$$Xv_1$$

The variance of the first principal component of X is:

$$\text{Var}(Xv_1) = \frac{\sigma_1^2}{n - 1}$$



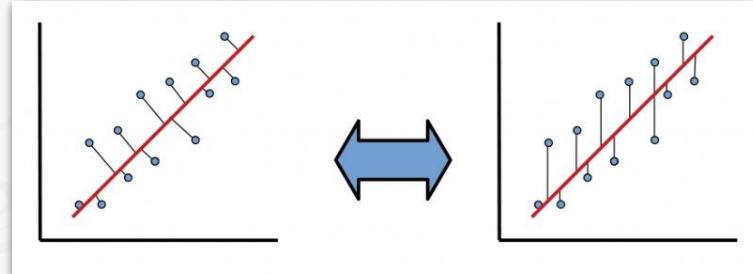
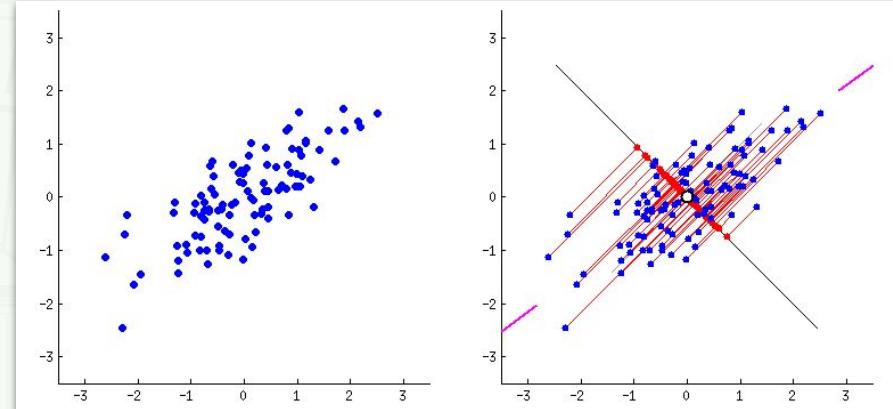
Principal Component Analysis (PCA)

Given a dataset, the principal components are vectors v_i that best describe the variance in the data.

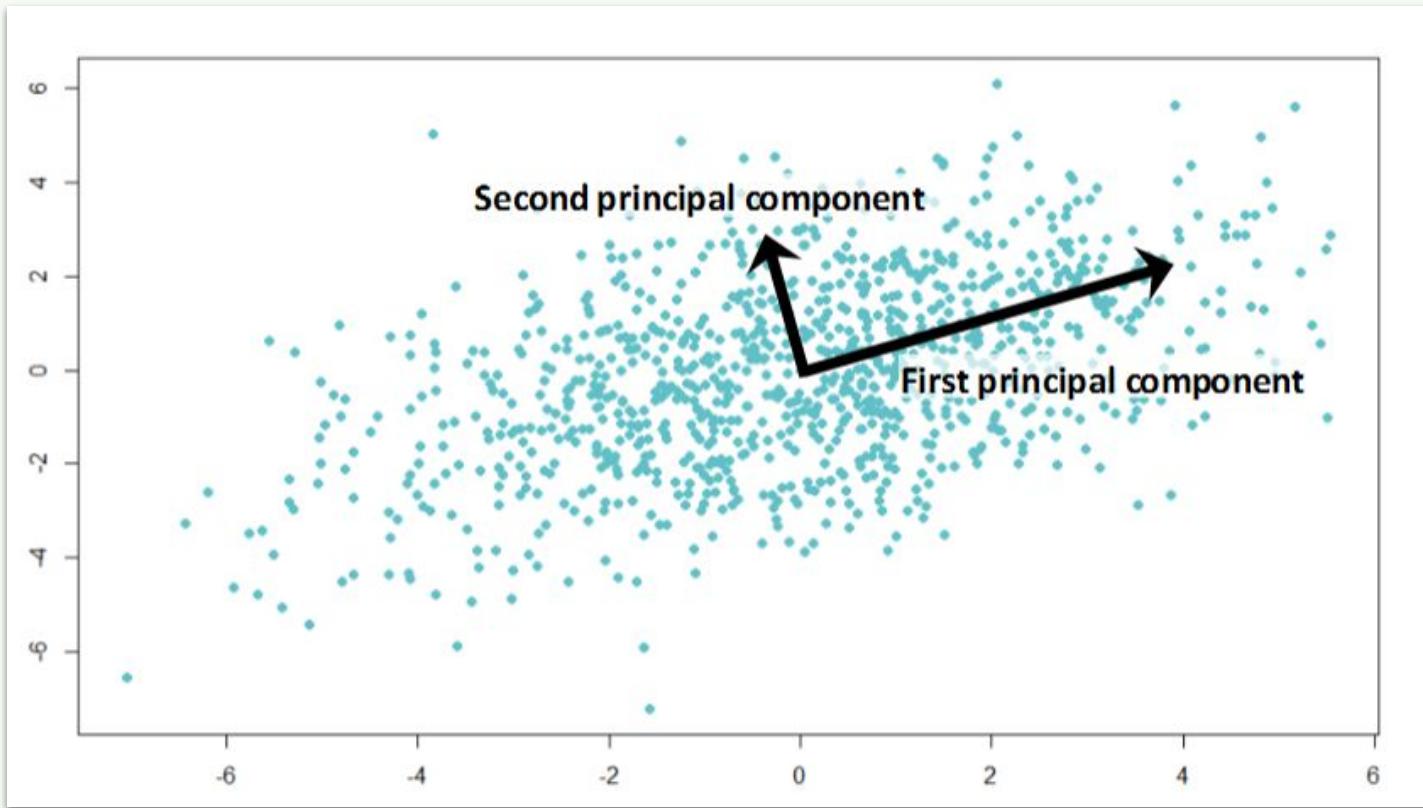
The vectors are orthogonal to each other.

The “best fit” is defined by minimizing the *perpendicular* distance of the data to the line(s).

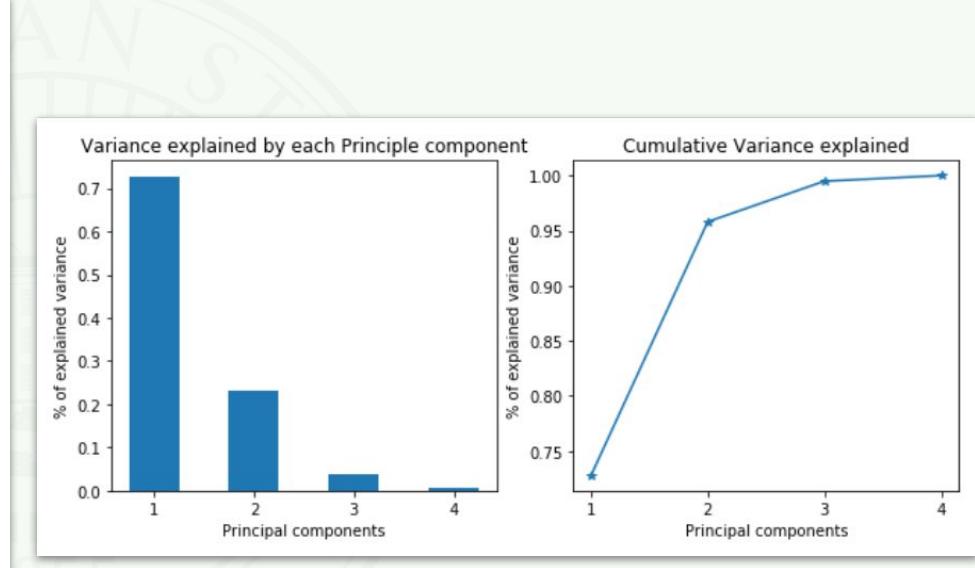
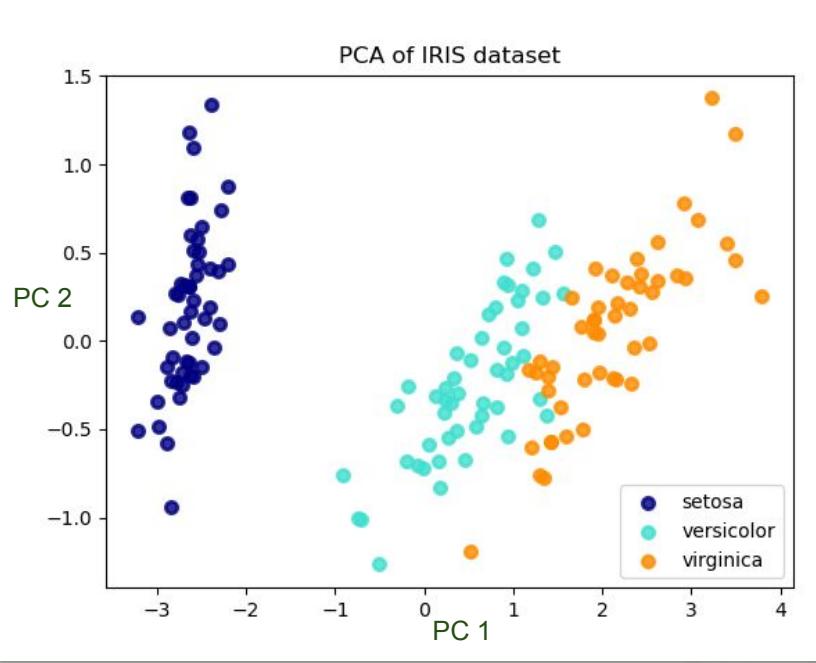
Don't confuse the best fit line in PCA with a line used for modeling.



PCA Vectors

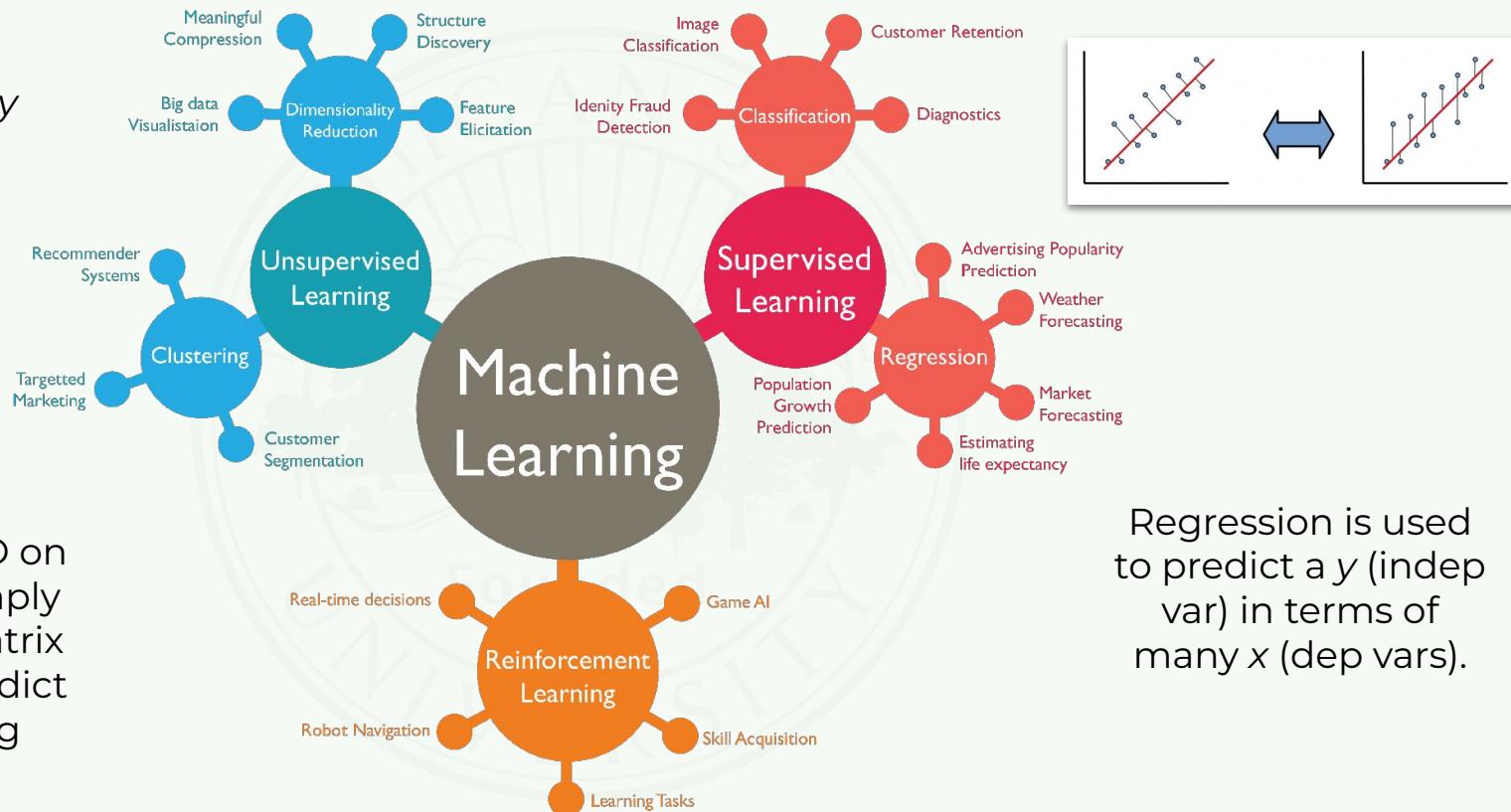


Iris Example



Organization of Machine Learning

There is no known y in unsupervised learning.

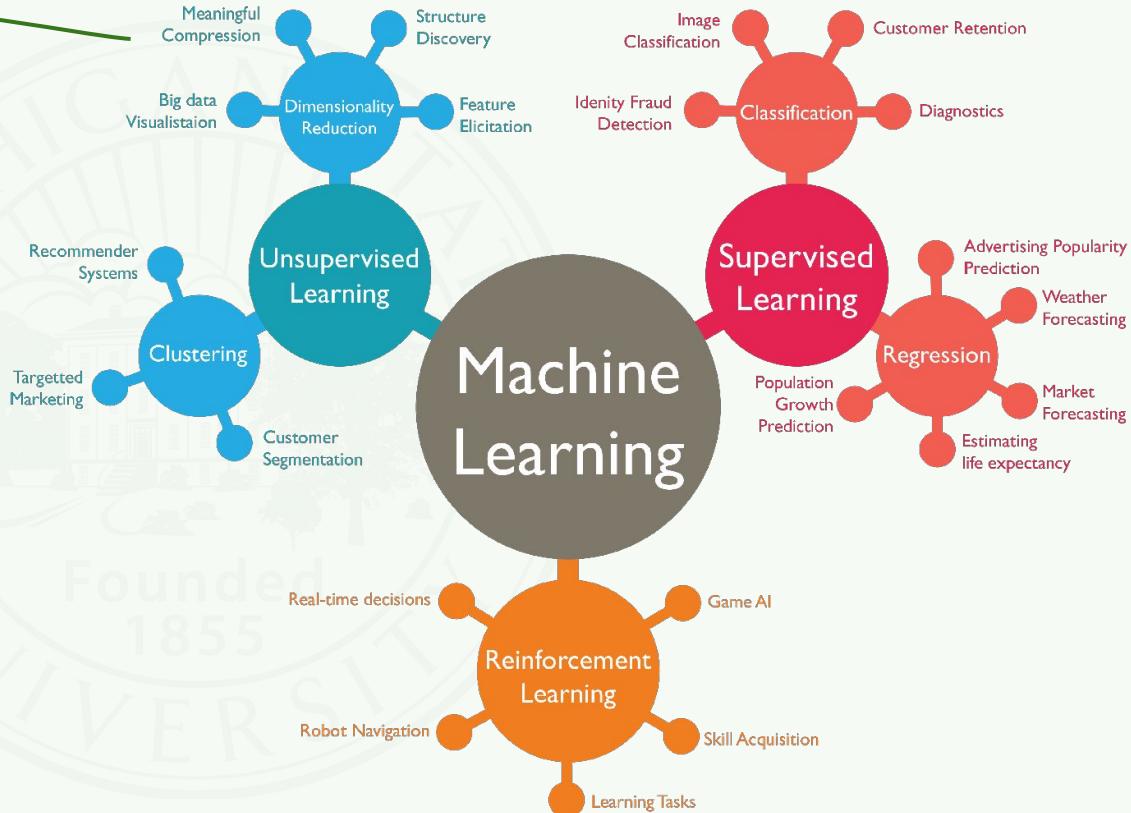
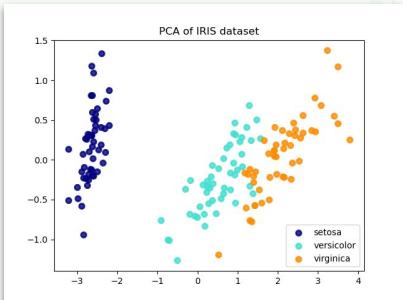


When we did SVD on an image, we simply used the data matrix X with no y to predict or use in finding weights.

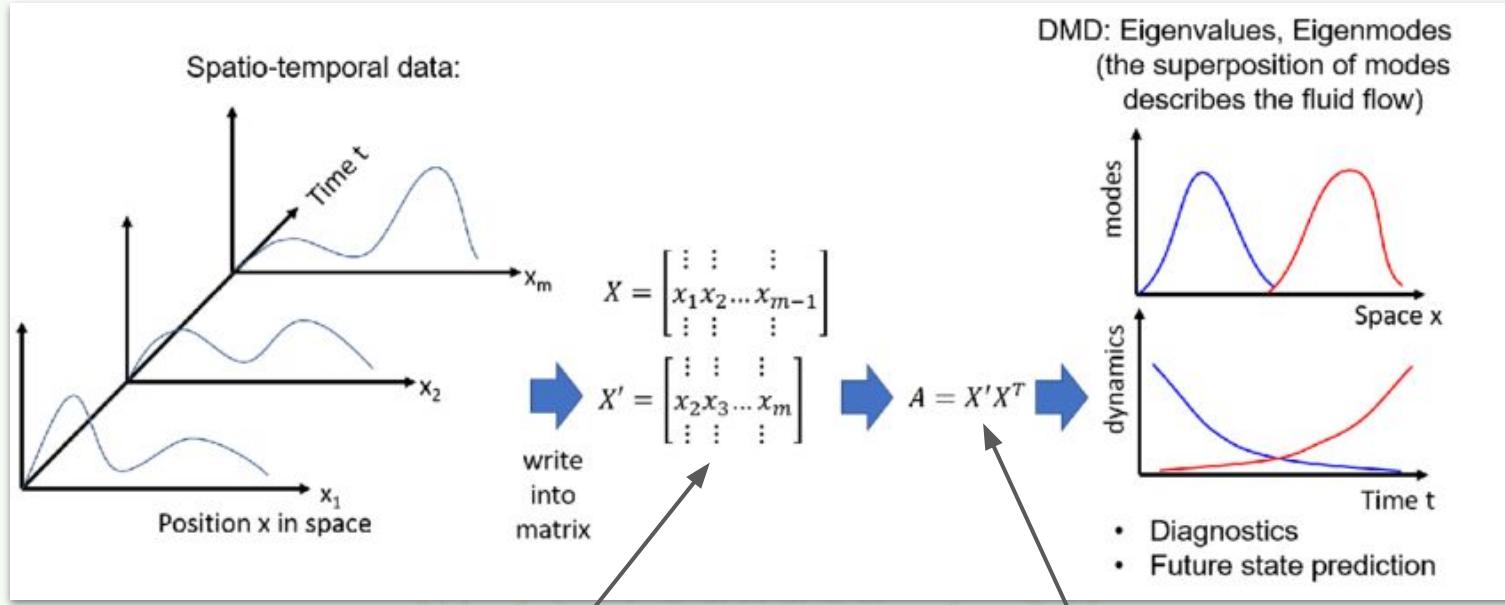
Regression is used to predict a y (indep var) in terms of many x (dep vars).

Dimensionality Reduction

dimensionality reduction:
transform data to a
lower-dimensional space
that preserves important
features of the data



Note on Spatiotemporal Data



In this case, the values in each column are not different data samples, but values at specific spatial points. X^TX and XX^T have important physical meanings: averages over space and time.

For time dependent data we can form generalizations of the correlation matrix that includes changes over time. That correlation matrix contains interesting information about the dynamics.



SVD/PCA Are Common, But Not Always Best

SVD vs. CUR

$$\text{SVD: } A = U \Sigma V^T$$

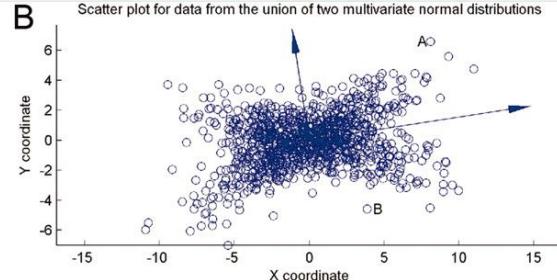
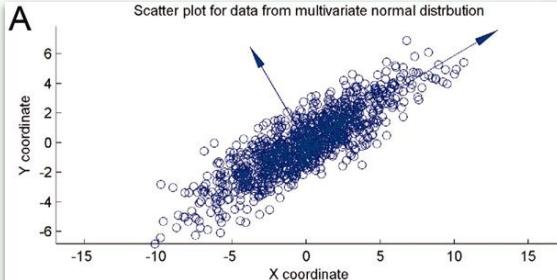
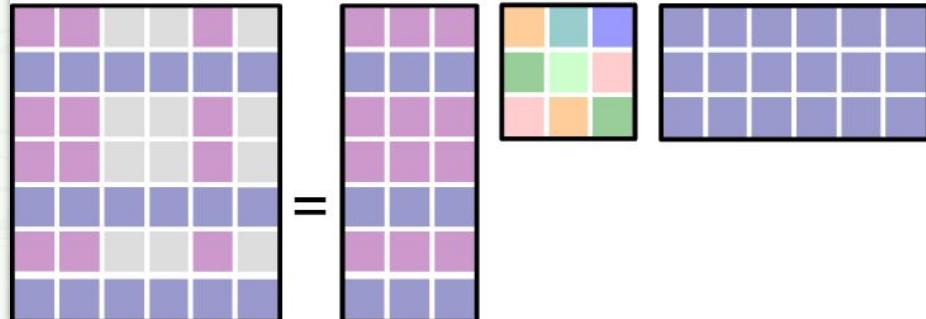
Huge but sparse Big and dense
sparse and small

$$\text{CUR: } A = C U R$$

Huge but sparse Big but sparse
dense but small

J. Leskovec, A. Rajaraman, J. Ullman: Mining of Massive Datasets, <http://www.mmds.org>

A $m \times n$ **C** $m \times k$ **U** $k \times k$ **R** $k \times n$



PCA is based on covariance, which may not best capture the structure in the data.