Welcome to Advanced Data Analysis (PHYS 605)

Prof. Claudia Gomes da Rocha

✉ claudia.gomesdarocha@ucalgary.ca

Department of Physics and Astronomy
Faculty of Science, University of Calgary

UNIVERSITY OF
CALGARY

Important material to study off-class:

Sequences and Time Series (some things in Chapter 18 of Dr. Jackel's notes)

- Understand the concept of **stationarity** in series
- Understand the use of auto-correlation calculation in predictive models
- Lag plots
- Partial auto-correlation functions
- Study the most common ways of obtaining auto-correlated sequences from random noise data:
  - ➢ Distributed Lag
  - ➢ Moving average
  - ➢ Autoregressive

This content is also very well detailed here:
https://www.itl.nist.gov/div898/handbook/pmc/section4/pmc444.htm

And on the book available at the UCalgary library: Time Series Analysis: Forecasting and Control by George E. P. Box, et al. (5th or 4th Edition), Wiley

Go over the really good tutorials of Pandas about "Chart Visualization" and Matlab "Plot Gallery":
https://pandas.pydata.org/docs/dev/user_guide/visualization.html
https://www.mathworks.com/products/matlab/plot-gallery.html

Previously in the course, we learned about optimization schemes in which we **fit known functions** into data to determine some fundamental characteristics of the processes they represent or to identify some underlying signal within the noise.

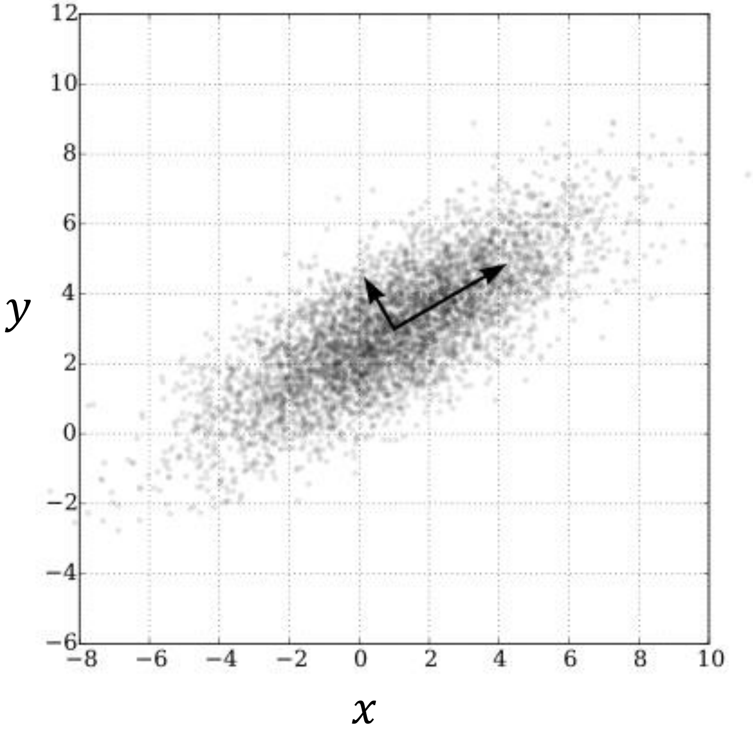
However, the signal within the data may reflect a linear combination of a set of processes that **do not follow some closed-form function or law.** For example, the distribution of rainfall within a region may represent a linear combination of the temperature, humidity, atmosphere pressure and other physical variables associated with the locality. Each of these variables changes in time in a highly complex fashion and a full analytical description with a well-defined mathematical expression may be not feasible.

How can we analyze this rainfall multivariate data and decompose it into a set of meaningful functions?

Information about the relationship between **two random variables** x and y is straightforward given by the covariance σ_{xy}. But what about a set of **3 or more** random variables (multivariate dataset)? It turns out that some very interesting multivariate relationships can be identified from the set of covariance pairs σ_{ij} (i and j can be any random variable from the multivariate set) and basic matrix mathematics.

Principal Component Analysis (PCA)

PCA is a technique that converts a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called **principal components**. PCA is often used to simplify data, reduce noise, recognize patterns, and reduce the number of features that will represent the original dataset in a compressed way; PCA is a statistical approach that can be used to analyze **high-dimensional data** and capture the most important information from it. This is done by transforming the original data into a **lower-dimensional** space while collating highly correlated variables together.



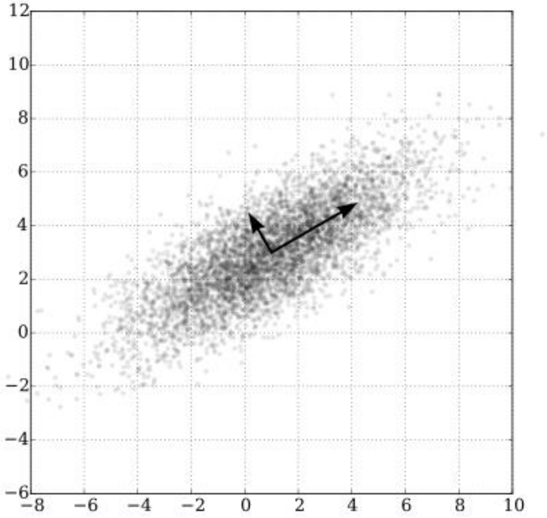
PCA of a multivariate Gaussian distribution centered at (1,3) with a standard deviation of 3 in roughly the (0.866, 0.5) direction and of 1 in the orthogonal direction. The vectors shown are the eigenvectors of the covariance matrix scaled by the square root of the corresponding eigenvalue and shifted so their tails are at the mean.

https://en.wikipedia.org/wiki/Principal_component_analysis

Principal Component Analysis (PCA)

Intuition: PCA can be thought of as fitting a *p*-dimensional ellipsoid to the data, where each axis of the ellipsoid represents a **principal component**. If some axis of the ellipsoid is small, then the variance along that axis is also small.

To find the axes of the ellipsoid, **we must first center the values of each variable in the dataset on 0 by subtracting the mean of the variable's observed values from each of those values.** These transformed values are used instead of the original observed values for each of the variables. Then, we compute the **covariance matrix of the data and calculate the eigenvalues and corresponding eigenvectors of this covariance matrix.** Then we must **normalize each of the orthogonal eigenvectors to turn them into unit vectors.** Once this is done, each of the mutually orthogonal **unit eigenvectors can be interpreted as an axis of the ellipsoid fitted to the data.** This choice of basis will transform the covariance matrix into a diagonalized form, in which the diagonal elements represent the variance of each axis. **The proportion of the variance that each eigenvector represents can be calculated by dividing the eigenvalue corresponding to that eigenvector by the sum of all eigenvalues.**



https://en.wikipedia.org/wiki/Principal_component_analysis

Principal Component Analysis (PCA)

PCA is essentially an eigen-decomposition of the covariance matrix of a dataset.

PCA is a statistical procedure that uses an orthogonal transformation that converts a set of correlated variables to a set of uncorrelated variables. PCA is a widely used tool in exploratory data analysis and in machine learning for predictive models.

PCA is considered an unsupervised learning algorithm technique used to examine the interrelations among a set of variables. In unsupervised learning, the algorithm explores the inherent structure of the data without using explicit labels or target values (and no training!).

A main goal of PCA is to reduce the dimensionality of a dataset while preserving the most important patterns or relationships between the variables without any prior knowledge of the target variables.

- Covariance decomposition (more transparent to understand, but not as general)
- Singular Value Decomposition (SVD) (maybe not as intuitive, but more general)

$$X = U\Sigma V^*$$

meaning, the singular value decomposition of an $m \times n$ complex matrix X is the factorization above where U is an $m \times m$ complex unitary matrix, Σ is an $m \times n$ rectangular diagonal matrix with non-negative real numbers on the diagonal, V is an $n \times n$ complex unitary matrix, and V^* is the conjugate transpose of V. Such decomposition always exists for any complex matrix.

Principal Component Analysis (PCA)

Step 1 - Data normalization (or standardization)

Step 2 - Covariance matrix calculation

Step 3 - Eigenvectors and eigenvalues (from the covariance matrix)
Geometrically, an eigenvector represents a direction such as “vertical” or “90 degrees”. An eigenvalue, on the other hand, is a number representing the amount of variance present in the data for a given direction. Each eigenvector has its corresponding eigenvalue.

Step 4 - Selection of principal components
There are as many pairs of eigenvectors and eigenvalues as the number of variables in the data. In a dataset, for instance, containing monthly expenses, age, and consumer rate, there will be three pairs. Not all the pairs are relevant. So, the eigenvector with the highest eigenvalue corresponds to the first principal component. The second principal component is the eigenvector with the second highest eigenvalue, and so on.

Step 5 - Data transformation in new dimensional space
This step involves re-orienting the original data onto a new subspace defined by the principal components. This reorientation is done by multiplying the original data by the previously computed eigenvectors.

Principal Component Analysis (PCA)

Some math behind … (on the board)