Math 253: Mathematical Methods for Data Visualization

– Course introduction and overview (Spring 2020)

Dr. Guangliang Chen

Department of Math & Statistics
San José State University

What is this course about?

Context: Modern data sets often have hundreds, thousands, or even millions of features (or attributes). ⟵ large dimension

This course focuses on the statistical/machine learning task of dimension reduction, also called dimensionality reduction, which is the process of reducing the number of input variables of a data set under consideration, for the following benefits:

- It reduces the **running time** and **storage space**.
- Removal of **multi-collinearity** improves the interpretation of the parameters of the machine learning model.
- It can also clean up the data by reducing the **noise**.
- It becomes easier to **visualize the data** when reduced to very low dimensions such as 2D or 3D.

There are two different kinds of dimension reduction approaches:

- **Feature selection** approaches try to find a subset of the original features variables.

  Examples: *subset selection*, *stepwise selection*, *Ridge and Lasso regression*. ⟵ Already covered in Math 261A

- **Feature extraction** transforms the data in the high-dimensional space to a space of fewer dimensions. ⟵ Focus of this course

  Examples: *principal component analysis (PCA)*, *ISOmap*, and *linear discriminant analysis (LDA)*.

Dimension reduction methods to be covered in this course:

- **Linear projection methods**:
  - PCA (for unlabled data),
  - LDA (for labled data)
- **Nonlinear embedding methods**:
  - Multidimensional scaling (MDS), ISOmap
  - Locally linear embedding (LLE)
  - Laplacian eigenmaps

Use of dimension reduction

Dimension reduction can greatly help with the following statistical and machine learning tasks:

- **Regression** (Math 261A)
- **Classification** (Math 251)
- **Clustering** (Math 252)
- **Visualization** ⟵ this course

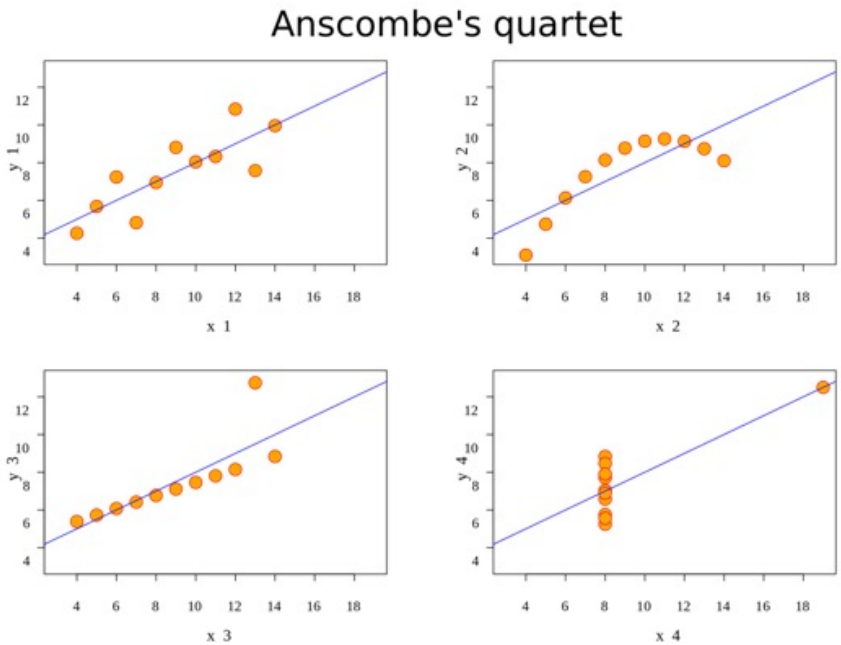
What is data visualization and why do we use it?

Data visualization is the graphic representation of data. According to Friedman (2008) "*the main goal of data visualization is to communicate information clearly and effectively through graphical means*."

Why it is important: A picture is worth a thousand words – especially when you are trying to understand trends, outliers, and patterns in data sets that include thousands or even millions of variables.

Data visualization can provide insight that descriptive statistics cannot, see example on next slide.

Example: Anscombe's Quartet (Francis Anscombe, 1973): The four datasets have almost identical variance, mean, correlation between X and Y coordinates, and linear regression lines.



However, the patterns are very different when plotted on a graph.

This illustration highlights why it's important to visualize data and not just rely on descriptive statistics.

Rational of this course

This course covers the following:

- Central topic: **dimension reduction**
- Supporting tools
  - **programming basics**, and **3D data plotting**
  - **Advanced linear algebra**
- Main application (motivation): **Data visualization**

It will also prepare you for Math 251 (classification).

Structure of this course

This course is divided into three parts:

Part 1. **Programming basics and 3D data plotting**

Part 2. **Advanced linear algebra**

Part 3. **Dimension reduction methods**

Overall, this course is 70% theory + 30% programming.
(Math 251 will be the opposite)

History of this course

This course is the further development of the following two courses:

- Fall 2016: **Math 298 Special Study**[1] (Student: Xiaohong Liu, Topic: Data visualization).
- Fall 2015: **Math 285 Selected Topics in High Dimensional Data Modeling**[2]

[1] http://www.sjsu.edu/faculty/guangliang.chen/298Report_Liu.pdf
[2] http://www.sjsu.edu/faculty/guangliang.chen/Math285F15.html

Relevance of the course

This course is used in the following ways:

- A prerequisite to **Math 251 Statistical and Machine Learning Classification**
- An elective for the regular **MS Statistics degree**
- A requirement by the **Machine Learning Specialization** (along with Math 251)
- A required course by a **new MS Data Science degree** (pending)

Prerequisites of the course

- Math 32 multivariable calculus* (method of Lagrange multipliers)
- Math 129A linear algebra* (strong linear algebra knowledge and skills are crucial)
- Math 163 probability theory (mathematical maturity)

*Requires a B or better grade

Textbook

"Foundations of Data Science", Avrim Blum, John Hopcroft, and Ravindran Kannan, by Cambridge University Press (expected March 2020).[3]

An unofficial version (January 2018) is publicly available from the authors' website.[4]

Meanwhile, lecture slides by the instructor and material from other sources (blogs, papers, etc.) will be posted on the course website.

[3] https://www.cambridge.org/us/academic/subjects/computer-science/pattern-recognition-and-machine-learning/foundations-data-science

[4] https://www.cs.cornell.edu/jeh/book.pdf

Computing

This course will use MATLAB as the main programming language due to its advantages in matrix computing and data plotting.

Unfortunately, MATLAB is commercial software and SJSU does not have a student license. A few options are:

- Download a **free 30-day trial version** of the newest MATLAB with all toolboxes at `https://www.mathworks.com`
- Use the computer lab in MacQuarrie Hall 221 (limited access)
- Buy **MATLAB student license** (platform+statistics and machine learning toolbox, $59.99/year).

New License for MATLAB Student R2019b

To purchase product for an existing license, select it in My Account first.

Add-on Products USD 10.00

Offer valid only for new license purchase.
Explore Areas of Study

Info

- MATLAB selected.

Sort By Category | Sort By Name

Add to Cart

MATLAB Product Family	Price	Add
MATLAB and Simulink Student Suite Includes MATLAB, Simulink, Control System Toolbox, Curve Fitting Toolbox, DSP System Toolbox, Image Processing Toolbox, Instrument Control Toolbox, Optimization Toolbox, Parallel Computing Toolbox, Signal Processing Toolbox, Statistics and Machine Learning Toolbox, Symbolic Math Toolbox	USD 99.00	☐
MATLAB	USD 49.00	☑
Parallel Computing		
Parallel Computing Toolbox	~~USD 29.00~~ USD 10.00	☐
Math and Optimization		
Statistics and Machine Learning Toolbox	~~USD 29.00~~ USD 10.00	☑
Curve Fitting Toolbox	~~USD 29.00~~ USD 10.00	☐

A few words about **Python**:

- I also plan to integrate Python into the course as much as I can.
- For those of you who are proficient in Python, you are encouraged to use Python to do your homework and project.
- If you don't know the Python language yet, you are recommended to take Math 167PS concurrently with this course.
- I will try to provide sample Python codes or refer you to online resources.

Data sets to be used in this course

We will use the following data for learning and practice:

- **MNIST Handwritten Digits**[5]: 70,000 digital images of size 28x28 of handwritten digits 0...9 collected from about 250 people
- **Fashion-MNIST**[6]: Same size and format with MNIST, but the images contain clothes instead

[5] `http://yann.lecun.com/exdb/mnist/`
[6] `https://github.com/zalandoresearch/fashion-mnist`

- **USPS Zip Code Data**[7]: 9,300 size 16x16 grayscale images of handwritten digits scanned from envelops
- **20 Newsgroups Data**[8]: about 19,000 text documents that are divided into 20 groups (according to their topics)

Smaller data sets such as the **Wine Quality Data Set**[9] from the *UCI Machine Learning Repository*[10] will also be used for teaching demonstration and homework assignments.

[7] `http://statweb.stanford.edu/~tibs/ElemStatLearn/data.html`
[8] `http://qwone.com/~jason/20Newsgroups/`
[9] `https://archive.ics.uci.edu/ml/datasets/wine+quality`
[10] `http://archive.ics.uci.edu/ml/`

Requirements of this course

- **Homework** (20%): Assigned roughly weekly
- **Midterm 1** (25%): Tuesday, March 10, in class
- **Midterm 2** (33%): Tuesday, April 21, in class
- **Final project** (22%):
  - Project proposal (2%): due Friday, April 10, 5pm
  - Project presentation (5%): Wednesday, May 13, 9:45am-12pm
  - Project report (15%): Wednesday, May 13, 9:45am-12pm

Homework policy

The homework assignments will typically contain both theory and programming questions.

- You must submit homework on time in order to receive full credit (late homework will receive a 20% penalty for each extra day).
- You may collaborate on homework but must write everything on your own.
- For theory questions, your answers must have necessary supporting steps.
- For programming questions, results must be presented in a concise and meaningful manner, e.g., by using figures or tables, with support codes.
- Your lowest homework score will be dropped.

What is allowed when doing the homework

Collaboration is encouraged on homework, but only for the understanding/ learning parts. That is, you may (without needing to acknowledge):

- Discuss homework questions together;
- Come up with a solution together;
- Help each other with certain step or line of code;
- Compare answers with each other.

However, you must write your code and/or steps individually on your own.

What is considered cheating or plagiarism in homework

Some examples (where cheating is involved) are

- Copy other people's work partly or in full
- Use other people's products (such as plots and code) as your own submission (even with acknowledgment)
- Give your work to other people for copying or studying
- One person does all the coding and shares it with others

- Copy solution or code found online (even with acknowledgment).[11]

Cheating and plagiarism in any form may lead to a failing grade for the course, and additionally will be reported to the Office of Student Conduct per SJSU policy.[12]

[11] However, you can study it and after you fully understand it, rewrite the steps or code independently by yourself.

[12] `http://info.sjsu.edu/static/catalog/integrity.html`

Midterms

The course has two midterms, both of which are closed-book but cheat sheets of certain size will be allowed.

The second midterm is cumulative, thus equivalent to an early final.

They will cover the **theory** component of the course (in contrast, the programming component of the course is covered by homework and the final project).

The final project

The course will end with a final project to be selected between each individual student and the instructor (proposal due April 10).

The students will need to give a short **oral presentation** in class to report their findings and meanwhile write **a report of 5+ pages**.

Both the presentation and report will be graded based on correctness, clarity, depth, completeness, and originality.

More details will be given later in class.

Some final reminders

This course is

- new (subject to changes as needed)
- very challenging (theory, or programming, or both)
- demanding

However, the course is very important and useful, as it lays the theory and programming foundations for you to learn machine learning.

Assignments

1. Take the MATLAB Onramp tutorial[13] (if you haven't)
2. Complete the background survey[14] (if you haven't)
3. Install software (both MATLAB and Python) on your computer.
4. Read the descriptions of the data sets mentioned in this presentation, and download them to your machine
5. Study my notes on linear algebra review (see course webpage)

[13] `https://www.mathworks.com/learn/tutorials/matlab-onramp.html`
[14] `https://forms.gle/efY1omFtXACcTkca8`

Waitlist policy

To request an add code, you need to

- **write down your name and sign on the attendance sheet**, and
- **complete the background survey**[15] as soon as possible.

I will rely on the survey result to determine your eligibility and give out add codes (if seats are available).

[15] `https://forms.gle/efY1omFtXACcTkca8`

Questions?