Multivariate Statistical Methods for Big Data Analysis and Process Improvement

Lecture 1: Introduction

Dr. Brandon Corbett

Course notes for ChE 765/SEP 767, McMaster University

Cancer Thought Experiment

Bad news: you have just been diagnosed with rare "McMaster Grad School" cancer. You are offered a choice of doctors to design your treatment plan (note all actual treatment will be done by medical doctors/nurses). Who will you choose and why?

Dr. Garcia, MD

- Bachelor of health science from McMaster
- Top of class at Harvard Med school
- 25 years of experience
- Has treated hundreds of patients, of which a few have had similar cases to yours

Dr. Smith, PhD

- Bachelor of engineering from McMaster
- No minimal medical training
- Leading expert on statistical analysis of historical data
- Has access to a database containing medical records of 1,000,000 cancer patients
  - 500 cases of McMaster Grad School cancer

Instructor

Dr. Brandon Corbett - Adjunct Professor in Chemical Engineering

- PhD from McMaster in 2016
  - Data based modeling of batch processes
  - Focus on developing models for advanced control systems
- Postdoctoral fellowship
  - Worked with Dr. John F. MacGregor: founding father of multivariate stats in engineering
- Senior research scientist with Sartorius Corporate Research
  - Full-time (”day job”)
  - Research in AI applications in pharmaceutical manufacturing
  - Optimizing manufacturing for vaccines and other biotherapeutics
- 15th time teaching this course at McMaster, about 40th total
- Full disclosure: my time is limited

Introductions

- What is your name?
- What is your program/department?
- What was the last dataset you worked with?

Course overview: objectives and what we will cover

Course title

Course Title

"Multivariate Statistical Methods for Big Data Analysis and Process Improvement"

- What is "big data"?
  - Why do you care about "big data"?
- What are "multivariate statistical methods"?
- What is meant by "process improvement"?

Definition of Big Data?

"Big Data" is an evolving term... (a buzz word)

My take on the "5 Vs" of big data in engineering:

1. **Volume:**
2. **Velocity:**
3. **Variety:**
4. **Veracity:**
5. **Variability:**

Web of science stats on “Big Data”

Papers including key words “Big Data”: 47,188

4,341 this year alone

Papers with “Big Data” in title: 8,645

Need to limit scope for this course!

Back to course title

“Multivariate Statistical Methods for Big Data Analysis and Process Improvement”

- What does “multivariate statistical methods” mean?

Definition of Multivariate

From Wikipedia: **Multivariate statistics** is a subdivision of statistics encompassing the simultaneous observation and analysis of more than one outcome variable*. The application of multivariate statistics is multivariate analysis.

My addition: * and/or (many) more than one input variable

Aside: don’t be afraid of the word “statistics”

Further limitation on scope

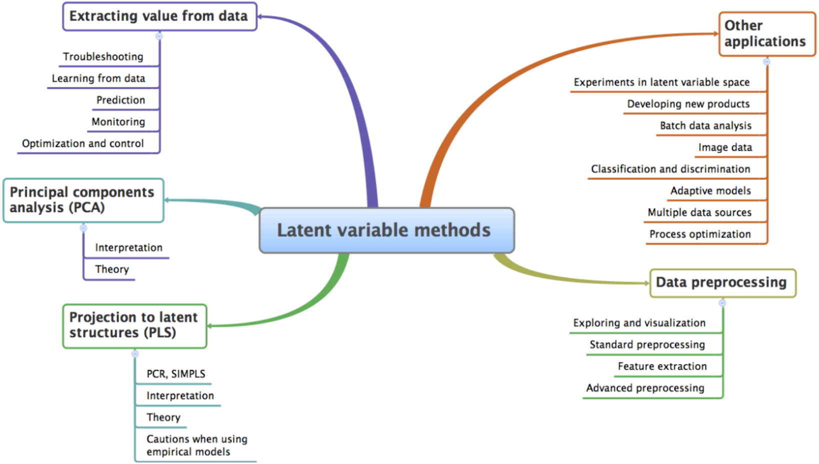
“Multivariate Statistical Methods for Big Data Analysis and Process Improvement”

- What does “Process Improvement” involve?
  - Particularly in an engineering context
  - What steps are involved in improving a process?

Process Improvement

1. Define the process that you are improving
2. Define what constitutes improvement (almost always $$)
   - Desired product quality
   - Improved uptime
   - Faster grade transitions
   - Fewer wasted resources
3. Figure out what is happening right now
   - Note: Data-based answers require data!
4. **Figure out what to change**

What we will cover



Latent variable methods

Extracting value from data
- Troubleshooting
- Learning from data
- Prediction
- Monitoring
- Optimization and control

Principal components analysis (PCA)
- Interpretation
- Theory

Projection to latent structures (PLS)
- PCR, SIMPLS
- Interpretation
- Theory
- Cautions when using empirical models

Other applications
- Experiments in latent variable space
- Developing new products
- Batch data analysis
- Image data
- Classification and discrimination
- Adaptive models
- Multiple data sources
- Process optimization

Data preprocessing
- Exploring and visualization
- Standard preprocessing
- Feature extraction
- Advanced preprocessing

“Latent Variable”

What do you think “Latent Variables” are?

Motivation for “latent variable” approach

In engineering, we don’t always measure “causal” variables.

- Good engineering practice involves taking many measurements
- Good engineering practice involves measuring the cheapest variables
- Reality dictates that we won’t always carry out designed experiments
- In order to correctly analyze our data we need to understand underlying (latent) factors
- This is especially true if we are making changes!

Administrative Details

Email Policy

- Given our limited time and the number of students, emails will be answered within one week of their reception
- ALWAYS include ChE 765 or SEP 767 in the title or your email may not be answered
- ONLY send emails from your official McMaster email address, emails from other addresses will not be answered
- I reserve the right to alter this policy as needed to meet new circumstances

Credits

- Dr. John MacGregor
    - Designed and taught this course
    - Was my post-doc supervisor
- Kevin Dunn
    - Instructed the course when I took it in 2011
    - Slides and curriculum largely based on his (more on this in a moment)
- School of Engineering Practice
    - Reinstating the course
    - Financial support

Contact Information

Dr. Brandon Corbett

corbeb@mcmaster.ca

Availability:

- I will be available after class
- Otherwise, schedule meetings by email

Course material (Avenue)

Please check Avenue regularly!

- Lecture slides
  - To be posted immediately before class
- Assignments (more details to come)

Grading

Project 70%

Paper Summary 15%

Assignments 5%

Midterm 10%

Grading allocation is subject to change

Course letter grades will be given using the standard method

Project (70%)

- Completed individually
- 70% of your mark
- Scope should be appropriate (thesis chapter/research paper)
- You will work with instructors 1:1
- Evaluation will be a 30 minute presentation and oral defense
- More details to come later in course

Paper summary (15%)

- To be completed in groups of 4
- You will be assigned a journal article related to the material in the upcoming class
- Prepare a 5 minute presentation
- Presentations will take place at the start of lecture
- Rules:
  - Designated presenter (only one group member may present)
  - Maximum of ONE slide
  - Maximum of five bullet points
  - Maximum of one figure

Assignments (5%)

- About 4 assignments + prereq quiz
- Approximately 1-2 typed pages per assignment
- Assignments must be submitted on Avenue prior to class
    - Failure to submit before class will result in a zero
    - No late assignments will be accepted

Midterm (10%)

- In class, closed notes, closed book
- I will give you a practice midterm
- Should be easy if you are following the course work (assignments etc.)
- Multiple choice questions
- Passing the midterm is mandatory to pass the course

Overall expectations

This is a 700 level graduate course

- Independent thought and understanding of concepts is required
- This is not a course on how to use a software tool
- You will learn various was of solving data problems
- There is no single correct method
    - There are plenty of incorrect methods

Academic Integrity:

THIS IS YOUR FIRST AND ONLY WARNING

This behaviour can result in serious consequences:

- The grade of zero on an assignment
- Loss of credit with a notation on the transcript
- **Suspension or expulsion from the university**

DO NOT copy from ANY source for this course

Course pre-requisites

- Basic ability in Matlab or Python is helpful (ie import data, plots etc.)
- Basic knowledge of linear algebra (we will do a small amount of review)
- Basic knowledge of univariate statistics (ie average and standard deviation)

Chemical engineering knowledge is NOT required but will be helpful

In the past, the course has deliberately been "code free"

- Numerical demos in Excel
- Modeling examples in ProMV (more in a minute)

- Do you have experience writing simple Python scripts?
- Would you like to learn to use Python?
- Would you feel comfortable seeing Python demos?

Discussion...

Vote: "Should I show Python examples in class time"

Aspen ProMV is the official software package provided for the course

- No coding experience needed
- Available on Horizon VM through SEP
- EVERYONE SHOULD TRY TO ACCESS ProMV THIS WEEK

See avenue for course schedule

Homework

- Pre-req quiz, worth 1%, due at the beginning of next lecture
- Think of 1 to 2 “concepts” which are unmeasurable but that clearly are a latent variable
  - Health
  - Quality of life