University of Western Ontario
Winter 2021

SS9878/CS9878: Analysis of High Dimensional Noisy Data

1 Instructor Information

Grace Y. Yi, Ph.D.
Professor, Canada Research Chair in Data Science (Tier 1)
Department Statistical and Actuarial Sciences
Department of Computer Science
University of Western Ontario
http://fisher.stats.uwo.ca/faculty/yyi/
https://covid-19-canada.uwo.ca

Email: gyi5@uwo.ca

Office Hour: Wednesdays 8:30-9:30pm

Note: The instructor has to handle a large volume of emails on a daily basis. When it is necessary to contact her by email, students need to use their Western (@uwo.ca) email addresses (so the emails won't be mis-treated as spams by the university email system), together with the subject title "**SS9878/CS9878**" (so the instructor can promptly notice your emails).

2 Course Information

Lecture Time: Wednesdays and Thursdays, 7:00- 8:30pm EST
Lecture Format: Synchronous online teaching via Zoom
Access Information: Meeting ID: 976 7478 8175; Passcode: 197003

Facility Requirements and Key Notes

- A computer with the reliable internet access to Zoom (as well as workable video and audio functions) is required for each student who takes or audits this course. All students are required to turn on the video during the lecture times.
- All the lectures, presentations, and office hours will be delivered via Zoom with the same access code shown above.
- The due date of submitting the course work is **April 10, 2021. No extensions will be granted**. It is your responsibility to carefully plan your time.

Course Audience

This is a graduate topic course cross-listed by the Department of Statistical and Actuarial Sciences (DSAS) and the Department of Computer Science (DCS). It is open to graduate students in DSAS, DCS, and Data Analytic Master program. This course focuses the discussion on the theory and methods. Hands-on experience on implementations of various methods is not the target, though some implementation software packages are to be discussed.

Prerequisite

Having basic statistics knowledge such as likelihood, conditional expectations, and regression would be important to well appreciate this course.

3 Course Outline

Thanks to the advancement of modern technology in acquiring data, data with diverse features and big volume are becoming more accessible than ever. While the abundant volume of data presents great opportunities for researchers to extract useful information for new knowledge gain and sensible decision making, big data present great challenges. A very important, but often overlooked challenge is the quality and provenance of the data. Big data are not automatically useful; big data are often raw and involve considerable noise.

Typically, the challenges presented by measurement error and missing observations are particularly intriguing. Measurement error and missing data arise ubiquitously from various fields including health sciences, epidemiological studies, survey research, environmental studies, economics, and so on. They have been a long standing concern in data analysis and have attracted extensive research interest over the past few decades.

It has been well documented that ignoring measurement error or missing data in statistical analyses may lead to erroneous or even misleading results. The effects of measurement error or missing data are, however, complex, and are affected by various factors. Further, measurement error or missing data present considerable challenges in analyzing high dimensional data. The primary aim of this course is to lead students through these important areas that have attracted extensive interest, and the ultimate objective is to broaden graduate students' view at an advanced level and to equip them with critical thinking skills.

There will not be a single textbook for this course, but the lecture notes will be organized in a coherent and self-contained manner. Some course material is based on the monograph

"G. Y. Yi (2017). *Statistical Analysis with Measurement Error or Misclassification: Strategy, Method and Application.* Springer Science+Business Media LLC, New York."

4 Sketch of Topics

PART 1: A BRIEF REVIEW AND PREPARATION

- Convex Optimization
- Matrix Operation
- Conditional Expectation
- Likelihood Method
- Estimating Function

PART 2: HIGH DIMENSIONAL DATA ANALYSIS

- Ridge Regression
- Least Absolute Shrinkage and Selection Operator (LASSO)
- Penalize Likelihood

PART 3: MEASUREMENT ERROR MODELS

- Examples Arising from Distinct Contexts
- Overview of Measurement Error Problems
- Methods of Addressing Measurement Error Effects
  - Regression Calibration
  - Simulation-Extrapolation (SIMEX)
  - Estimation Equation

PART 4: MISSING DATA PROBLEMS

- Introduction and Examples
- Missing Data Mechanisms
- Analysis Methods
  - Imputation methods
  - Likelihood-based methods
  - Inverse Probability Weighted GEE

5 Evaluation Scheme

$$15\%\ \textit{participation} + 35\%\ \textit{presentation} + 50\%\ \textit{course work}$$

Participation

This includes the student's attendance to classes, the involvement and participation with the class discussions, and the efforts paid to the course.

Presentation

Each student is going to give a presentation on a paper (or a few papers) on a topic concerning high dimensional data analysis, measurement error, or missing data. Students can choose papers on their own or ask the instructor to help them identify papers to present. The presentation length is about 10 to 30 minutes, to be determined after the registration number of the course is finalized. Each presentation will be evaluated by both the instructor and audience. Your presentation slides should be sent to the instructor (at least) a day prior to your presentation.

Time: The last few classes

Course Work

This course is intended to engage students with active thinking and explorations. Meanwhile, it is understood that the students in this courses have different backgrounds and come from different programs. To accommodate these features, the course work is designated differently in order to give the students more opportunities to showcase their talent/potential. You may choose ***one*** of the following tasks to be your contributions to the course work:

- Problems Designed by the Instructor:

  The number of questions is similar to the next bullet point.

- Problems Chosen on Your Own:

  A student may complete X full problems, or a mix of Y full problems and Z sub-problems, from a reference book (listed below) at you own selection, where $Z = (X - Y) \times 3$ (it is understood that $Y < X$ here).

  - For Ph.D. students in DSAS: $X = 7$
  - For Master's students in DSAS: $X = 5$
  - Other students: $X = 4$

  Note: Rather than requiring *formal* assignments to be submitted on scheduled times, you are given the freedom to choose problems to work with. If you are unable to tackle the problems completely, you are strongly encouraged to make efforts to solve problems as much as you can.

- Course Summary:

  You may write an essay about this course. The essay should be in the format of a scientific paper with a length about 20 - 30 pages. The contents of the essay should include two components: (1) a complete summary of the course topics, and (2) your own thoughts on extensions of some relevant topis.

- Extension of An Existing Topic:

  You may choose a specific topic of you interest and read a relevant research paper (or a couple of research papers if you wish). Extend the development of that research paper by using the knowledge you have learned in this course. Your extensions are expected to be well described and comprehensive with technical details. A paper-format report of length 10 - 30 pages is expected.

- Your Own Topic:

  You may identify a new problem on your own and write a report about it. The topic should be pertinent to high dimensional data analysis, measurement error, or missing data. The report should be in the format of a scientific paper with a length about 10 - 30 pages.

- Your Own Problems and Solutions:

  You may create a set of new problems on high dimensional data analysis, measurement error, or missing data you think of or modify from existing sources, together with the solutions of those problems. The number of problems can be as many as you want, but is expected no less than seven.

- Software Package:

  You may choose a paper on high dimensional data analysis, measurement error, or missing data, and develop a software package for the public to use. The developed package should be reliable and will be posted at a public platform such as CRAN or GitHub. Check with the instructor before you start.

- Real Application:

  You may find an available data set and implement a method (or some methods) you have learned from this course to analyze the data. A complete report of the analysis should be prepared in the scientific paper format.

Note on Course Work

- The course work should be prepared in a self-contained manner with each notation clearly defined. It is expected to be laid out in a research manuscript format, including a title, an abstract, and references, along with the main text. The layout of the contents should be logic and flow smoothly.

- The course work should be prepared neatly in Latex. A .pdf file together with a .tex file is expected to be submitted.

- The course work should be completed and submitted on an individual basis. However, if you think discussing with your peers can help you output more valuable outcomes, you may do so. In this case, please clearly write a statement to point out: (1) how your work is benefitted from the discussion, (2) who is involved with the discussion, and (3) what part(s) are identical to your peer(s)' work.

- **Due Date: April 10, 2021**

6 Reference Books

- Yi, G. Y. (2017). *Statistical Analysis with Measurement Error or Misclassification: Strategy, Method and Application.* Springer Science+Business Media LLC, New York.
- Carroll, R.J., Ruppert, D., Stefanski, L.A., and Crainiceany, C.M. (2006). *Measurement Error in Nonlinear Models*, 2nd ed., Chapman & Hall.
- Little, R. J. A. and Rubin, D. B. (2002). *Statistical Analysis with Missing Data*, 2nd Ed., New York: Wiley.
- James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An Introduction to Statistical learning, with Application in R.* Springer Science+Business Media LLC, New York.

7 Disclaimer

- The lecture materials are only intended for your own use. Some of them might be the on-going research of the instructor and her co-authors that has not been published yet. Please do NOT distribute the lecture notes without the instructor's permission.
- Citation: In case you need to cite some of the lecture material for your future work, you may refer it as

  "Yi, G.Y. (2021). Lecture Notes of "SS9878/CS9878 - Analysis of High Dimensional Noisy Data", University of Western Ontario."

8 General Information

Accommodation and Accessibility

If you are unable to meet a course requirement due to illness or other serious circumstances, you must provide valid medical or supporting documentation to the Academic Counselling Office of your home faculty as soon as possible. If you are a Science student, the Academic Counselling Office of the Faculty of Science is located in WSC 140, and can be contacted at *scibmsac@uwo.ca*. For further information, please consult the university's medical illness policy at
http://www.uwo.ca/univsec/pdf/academic_policies/appeals/accommodation_medical.pdf

Academic Policies

The website for Registrarial Services is

http://www.registrar.uwo.ca

In accordance with policy,

http://www.uwo.ca/its/identity/activatenonstudent.html, the centrally administered e-mail account provided to students will be considered the individual's official university e-mail address. It is the responsibility of the account holder to ensure that e-mail received from the University at his/her official university address is attended to in a timely manner.

Scholastic offences are taken seriously and students are directed to read the appropriate policy, specifically, the definition of what constitutes a Scholastic Offence, at this website: *http://www.uwo.ca/univsec/pdf/academic_policies/appeals/scholastic_discipline_undergrad.pdf.*

Support Services

Please contact the course instructor if you require lecture or printed material in an alternate format or if any other arrangements can make this course more accessible to you. You may also wish to contact Services for Students with Disabilities (SSD) at 661-2111 ext. 82147 if you have questions regarding accommodation.

The policy on Accommodation for Students with Disabilities can be found at
www.uwo.ca/univsec/pdf/academic_policies/appeals/accommodation_disabilities.pdf

The policy on Accommodation for Religious Holidays can be found at
http://www.uwo.ca/univsec/pdf/academic_policies/appeals/accommodation_religious.pdf