ISYE 7405 - Fall 2020 DUE: 12/10/2020

Final Project Guidelines

Prof. Shihao Yang *Email: shihao.yang@isye.gatech.edu*
TA: Tianyi Liu *Email: tianyiliu@gatech.edu*

Overview

During this project you will produce a paper presenting an application of multivariate statistical methods to one (or multiple) datasets of your interest.

Grading

This group project will consist 30% of your final grade, with the following decomposition: Topic definition 4%, proposal 6%, final presentation 4%, paper 6%, and technical merit 10%.

The grading rubric will be based on the writing, analysis, and context of your paper. For example, how organized, clearly written and comprehensible is the report? Were the analyses chosen and carried out properly? Were your conclusions about the data set sensible and clearly justified by numerical and graphical evidence?

Your team

You can form the group of one or two people. That is, you can either complete the project yourself, or team up with one another classmate. All students in the same group will receive the same grade.

Requirement for final paper

The paper should be organized like a scientific publication, with the following sections:

- Title, abstract, and significant text.
- Description of research question / issues (either scientific or statistical question)
- Description of data
- Presentation of statistical analysis of data
  - Methods: what analyses were done and why. If there is any challenge in analysis, describe your approach to tackle the problem.
  - Results: No verbatim of computer output, but a small number of tailored tables and graphics may be appropriate.

  - Conclusion: Convey your findings to broader audience.

- Discussion on limitation

The paper should be at least 2 pages but should not exceed 6 pages, including figures and tables. Please follow PNAS research article template. If you are using latex (preferred but not required), you can directly use the overleaf template: https://www.overleaf.com/latex/templates/template-for-preparing-your-research-report-submission-to-pnas-using-overleaf/fzcbzjvpvnxn

Statistical analysis on your dataset

By end of the semester, we will have seen (include but not limit to):

- hypothesis testing:
  - Hotelling T^2 test (one-sample or two-sample)
- dimension reduction
  - singular value decomposition
  - principal component analysis
  - factor analysis
- classification
  - logistic regression
  - linear discriminant analysis
  - support vector machine
- cluster analysis
  - EM algorithm
  - K-means
  - Hierarchical clustering

Choose one or two from the methods we covered in class.

Alternatively, some of you may want to tackle a project involving a multivariate method that we will not cover in this course, or to compare the performance of several statistical methods on datasets. These are all attractive options.

Potential topics and data sources

COVID-19 is a natural topic with immediate relevance to all of us. Data sources include

- CDC website https://www.cdc.gov/coronavirus/2019-ncov/cases-updates/index.html
- Covid-tracking project https://covidtracking.com/
- JHU https://coronavirus.jhu.edu/
- Apple mobility https://covid19.apple.com/mobility or Google mobility https://www.google.com/covid19/mobility/
- County-level data in Georgia https://dph.georgia.gov/covid-19-daily-status-report
- US Census https://catalog.data.gov/dataset/conditions-contributing-to-deaths-involving-coronavirus-disease-2019-covid-19-by-age-group
- New York Times https://github.com/nytimes/covid-19-data
- Social Economical status https://github.com/OpportunityInsights/EconomicTracker
- European data https://www.ecdc.europa.eu/en/covid-19/data

In order to have multivariate data, you can focus on state-, county-, or city-level data, rather than national level data as a whole.

You can also choose other topics from your field of study. If you choose to use datasets available to public, UCI Machine Learning Repository (http://archive.ics.uci.edu/ml/) contains a number of preprocessed multivariate datasets. The following list contains many topics and data sets you may be interested:

- America's Best Colleges - U.S. News & World Reports
- American FactFinder
- The Baseball Archive
- The Bureau of Justice Statistics
- The Bureau of Labor Statistics
- The Bureau of Transportation Statistics
- The Census Bureau
- Data and Story Library (DASL)
- Data Sets, UCLA Statistics Department

- DIG Stats
- Economic Research Service, US Department of Agriculture
- Energy Information Administration
- Eurostat
- Exploring Data
- FedStats
- The Gallop Organization
- International Fuel Prices
- Journal of Statistics Education Data Archive
- Kentucky Derby Race Statistics
- Hong Kong Jockey Club Race Statistics
- National Center for Education Statistics
- National Center for Health Statistics
- National Climatic Data Center
- National Geophysical Data Center
- National Oceanic and Atmospheric Administration
- Sports Data Resources
- Statistics Canada
- StatLib—Datasets Archive
- UK Government Statistical Service
- United Nations: Cyber SchoolBus Resources

Project deliverables

Project proposal

You need to submit the project proposal by Oct 27, 2020. The project proposal will need to have the following elements:

- Team membership.

- Title of the paper.
- Abstract / Significant text that defines the topic, or the subtopic within COVID-19 if you choose COVID-19 as topic
- The exact datasets (with hyperlink) that you are going to analyze
- Methods that you plan to use

The format should follow the title page (first page) of the PNAS paper template. The purpose of this proposal is to make sure you have a clear plan for the paper, and also to minimize overlap between different teams. The instructor or TA will provide feedback and you might need to revise the proposal to incorporate those comments, then the proposal will be deemed finalized. After the proposal is finalized:

- Team membership cannot change.
- Title cannot change.
- Abstract / Significant text can be revised, but you cannot change the topic/subtopic.
- The proposed datasets can be expanded but not dropped. That is, you can add more data later but you must analyze the existing dataset you proposed in the proposal.
- Methods that you plan to use can be freely revise without constraints.

Presentation

The last two lectures (11/19 and 11/24) is reserved for each team to present your project with some preliminary results. The presentation will be conducted online, and each team will have 8 minutes plus 2 minutes of Q&A. There is no format requirement for the presentation. During presentation, you do not need to have your final paper ready.

Final paper

The final paper will due on Dec 10, 2020. Late submission will be heavily penalized. No paper submission will be accepted strictly after Dec 13, 2020.