Final exam - Advanced Statistical Methods

L. Egidi, F. Pauli, M. Trevisani

Spring 2024

Contents

Exam's rules

- Prepare your final report/presentation by using RMarkdown. Any template (pdf, html, word, slides, power point, etc.) is admissible.
- As you'll see below, there are four sections, one for each course's specific part. You need to answer all the questions.
- Send by email the professor your project presentation 2/3 days before your exam takes place. You'll discuss your homework at the exam's day.

Multilevel modeling

Consider the dataset `nyc_arrests.txt` related to the police stops (you find the dataset in the Exam section in the Moodle page): for further details, see G& H book, Chapter 15, and lecture notes about hierarchical models, slides 123–128.

- Fit the model (24) in the notes from a Bayesian and a frequentist point of view, using `rstanarm`/`rstan` and `lme4` packages.
- Provide the estimates (also from a graphical perspective) and comment the results. Hint: use the `bayesplot` package when you can.
- Fit directly a negative binomial model using `rstanarm`/`rstan` and compare it with the previous estimates.

- Compare models in terms of predictive information criteria.
- Divide the dataset in training/test and make some predictions.
- Extend the model: write a modeling extension (fit is not required) where a further continuous covariate `income_ethnicity`, expressing an average income for the ethnicity e in New-York City, is available. (Hint: there is not a unique way to incorporate it). Finally, discuss the eventual sign of the estimated coefficient from a "socio-political" perspective.

Causal inference

Consider the dataset `electric_wide.txt` about the Electric Company example (you find the dataset in the Exam section in the Moodle page): for further details see the ROS book, Chapter 19.2, and lecture notes from slide 19.

- Fit the models (6), (7) and (8) from lecture notes from a Bayesian and a frequentist point of view, using `rstanarm`/`rstan` and `lme4` packages.
- Provide the estimates (also from a graphical perspective) and comment the results.
- Write in formulas and fit a multilevel/hierarchical model on these data and compare the results with those previously obtained.

Semiparametric regression

Consider the data in https://www.kaggle.com/datasets/sameersmahajan/seattle-house-sales-prices.

```
d=read.csv("house_sales.csv",header=TRUE)
str(d)
```

Price is your response variable y.

- Fit a semiparametric regression model

$$y_i = f(x_i) + \epsilon_i, \quad \epsilon_i \sim \ IID \ \ \mathcal{N}(0, \sigma^2),$$

where x_i is `sqft_living`.

- Discuss the model fit, explain why the model is not adequate and propose how to improve it.
- Devise a model to assess whether there is any seasonality in the price.

Mixed membership models

Choose only *one* among the two choices below:

1. Consider a subset of the data of end-of-year addresses of the presidents of the Italian Republic (in the course Team files) from 1992 (beginning of Scalfaro office) to 2021 (end of Mattarella first office). Select nouns, adjectives and verbs with a frequency of at least 5 occurrences as textual units of analysis. Fit a topic model by opportunely choosing the model structure – whether and how incorporating metadata (President and speech year) for detecting any significant influence on topic structure, and number of topics. Comment on the results and provide an effective visualization to help understanding.

2. Consider a sample of biodiversity data of `Rlda` R package (`abundance` or `birds` described in the package vignette and the related JSS and KBS articles). Fit a mixed membership model suited for the specific type of response variable. Comment on the advantages of considering a mixed membership assumption in place of a *full* membership assumption of a finite mixture model in this application.