

DS-GA 3001.001

Special Topics in Data Science: Probabilistic Time Series Analysis

Project proposal instructions

Proposal due on 10/8 (by 6pm)

The project should be done in groups of 2 or 3 registered students (no exceptions unless you have explicitly cleared this with me first).

The **project** can take any of the following forms:

- In-depth **review** of some recent papers related to time series topics either covered in the lecture (e.g. generalizations of LDS, HMMs) or not (e.g. RNNs and other deep learning variants).
- high quality **software implementation** of algorithms related to the class that are currently not publicly available, to be released for public usage.
- application of a machine learning model to a previously unconsidered **dataset** and to a specific scientific question. For this: find some interesting (for you) data. Keep in mind the issue of stationarity. Make sure you have enough data to be able to fit a reasonable model (the more parameters, the more data you'll need).
- **extension to existing method**, or theoretical analysis of existing algorithm. This would likely have a scope outside the course, e.g. it could be the starting point for a longer research project.
- Multiple **models comparison** on existing dataset

If you are unsure if your idea fits into these types, talk to CS.

Proposal:

Write a ~1 page proposal that details the question you are planning to address, which dataset are you planning to use (if applicable), the family of algorithms used for the analysis, and how you plan to evaluate these methods. The goal is to check that you do have a plan, so add whatever details you have already worked out that may be relevant (within the space limits).

Important: explain also how the tasks will be allocated across team members (who does what).

Proposals should be uploaded on nyuclasses in **pdf**. We plan to provide some constructive feedback for each proposal by end of following week.

Some past projects:

1. **Model comparison:** HMMs vs RNNs for anomaly detection.
2. **Model comparison:** Automatic speech recognition - HMMs Acoustic models, Deep nets
3. **Implementation:** RNNs for multivariate time series with missing values (GLU-R)
4. **Object tracking:** implementation + model comparison
5. **Dataset:** Modeling protein sequences + Model comparison HMMs vs RNNs
6. **Dataset:** Forecasting Wikipedia web traffic
7. **Dataset:** Prediction of Beijing pollution levels
8. **New algorithm:** Gaussian Process Factor Analysis with Spectral Mixture kernels
9. **Review:** latent state space models for modeling neural data
10. **Voice synthesis:** speak like Pippa the pig

Some new ideas:

- **Finite sample** properties of linear models (*e.g. Kley et al*)
- **Probabilistic RNNs** — variational auto-encoders for time series
- **Volatility** models, e.g. ARCH/GARCH
- **Nonlinear** generalizations of a model considered in class
- Find a **new dataset** with dependencies and try to model it