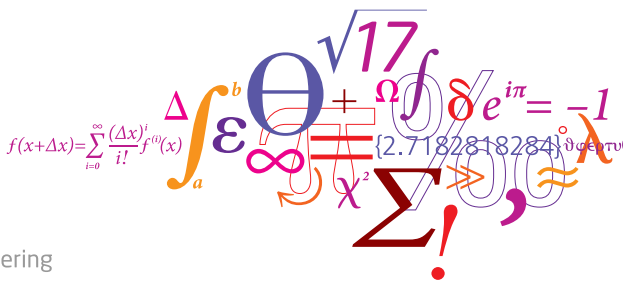
Project description

Filipe Rodrigues

Francisco Pereira



DTU Management Engineering
Department of Management Engineering

Outline

- Rules
- Important dates
- Common mistakes, misconceptions and advice
- Great projects from last year
- Take-away ideas

Rules

- Topic is **free** (creativity highly encouraged!)
- No constraints on the dataset
- If in doubt, talk with us!
- We also provide some suggestions for projects

Important dates

Important dates

- April 24 - Milestone 1
  - Research question
  - First draft of model(s) that you plan to try out (PGM + generative process)
  - Initial notebook (descriptive stats, data preparation)

Important dates

- April 24 - Milestone 1
  - Research question
  - First draft of model(s) that you plan to try out (PGM + generative process)
  - Initial notebook (descriptive stats, data preparation)
- May 15 - Final delivery (Part I)
  - Full self-explanatory notebook
- May 18 - Final delivery (Part II)
  - 2-page report (excluding figures and tables)

Common mistakes, misconceptions and advice

- No point putting priors on observed variables without any parents, unless doing imputation, otherwise these variables are always given

Common mistakes, misconceptions and advice

- No point putting priors on observed variables without any parents, unless doing imputation, otherwise these variables are always given
- Makes no sense to include observed variables that block the information path in the PGM
  - Consider for example: Intelligence $\rightarrow$ Course grade $\rightarrow$ Recommendation letter
  - If course grade is always observed, then it blocks the path between intelligence and recommendation letter. Therefore, "Intelligence $\rightarrow$ Course grade" and "Course grade $\rightarrow$ Recommendation letter" become two independent models.
  - There is no point estimating them jointly because there is no flow of information between the two (unless there is an alternative path, of course...)

Common mistakes, misconceptions and advice

- No point putting priors on observed variables without any parents, unless doing imputation, otherwise these variables are always given
- Makes no sense to include observed variables that block the information path in the PGM
  - Consider for example: Intelligence $\rightarrow$ Course grade $\rightarrow$ Recommendation letter
  - If course grade is always observed, then it blocks the path between intelligence and recommendation letter. Therefore, "Intelligence $\rightarrow$ Course grade" and "Course grade $\rightarrow$ Recommendation letter" become two independent models.
  - There is no point estimating them jointly because there is no flow of information between the two (unless there is an alternative path, of course...)
- Don't just think about what variables depend on other variables, and their distribution types. Think also of how you should model those dependencies (e.g. how to condition the parameters of a Beta distribution on another variable?)

Common mistakes, misconceptions and advice

- Be careful with the inclusion of latent variables that are discrete. Remember that they require special treatment in STAN/Pyro

Common mistakes, misconceptions and advice

- Be careful with the inclusion of latent variables that are discrete. Remember that they require special treatment in STAN/Pyro
- Choice of priors is often more of an art than a science. However, there are some recommendations and guidelines that you can try to follow, and people in statistics write papers about this a lot, but in the end it often boils down to a lot of trial and error. I recommend that you have a look at: https://github.com/stan-dev/stan/wiki/Prior-Choice-Recommendations

Common mistakes, misconceptions and advice

- Be careful with the inclusion of latent variables that are discrete. Remember that they require special treatment in STAN/Pyro
- Choice of priors is often more of an art than a science. However, there are some recommendations and guidelines that you can try to follow, and people in statistics write papers about this a lot, but in the end it often boils down to a lot of trial and error. I recommend that you have a look at: https://github.com/stan-dev/stan/wiki/Prior-Choice-Recommendations
- Try different types of inference algorithms: VI and MCMC, and make sure they are doing the right thing and you can trust the results

Common mistakes, misconceptions and advice

- Be careful with the inclusion of latent variables that are discrete. Remember that they require special treatment in STAN/Pyro
- Choice of priors is often more of an art than a science. However, there are some recommendations and guidelines that you can try to follow, and people in statistics write papers about this a lot, but in the end it often boils down to a lot of trial and error. I recommend that you have a look at: https://github.com/stan-dev/stan/wiki/Prior-Choice-Recommendations
- Try different types of inference algorithms: VI and MCMC, and make sure they are doing the right thing and you can trust the results
- It is a generally good idea to start with a simple model and incrementally make it more complex towards the idealised/conceived PGM. Plus, try to have some baseline models for comparison whenever possible

Common mistakes, misconceptions and advice

- Be careful with the inclusion of latent variables that are discrete. Remember that they require special treatment in STAN/Pyro
- Choice of priors is often more of an art than a science. However, there are some recommendations and guidelines that you can try to follow, and people in statistics write papers about this a lot, but in the end it often boils down to a lot of trial and error. I recommend that you have a look at: https://github.com/stan-dev/stan/wiki/Prior-Choice-Recommendations
- Try different types of inference algorithms: VI and MCMC, and make sure they are doing the right thing and you can trust the results
- It is a generally good idea to start with a simple model and incrementally make it more complex towards the idealised/conceived PGM. Plus, try to have some baseline models for comparison whenever possible
- Use ancestral sampling to generate artificial data, and run inference on the model using Pyro/STAN to see if it is able to recover the true values/parameters that were used to generate the data. This is a great way of guarantying that the model is correctly implemented, and that inference is working correctly

2019 proj.: Analysis group membership in Social Networks I

- Graph data of human interaction is called social networks. The interaction could for example being friends on Facebook or talking to each other.
- The project inferred groups in social networks based on Game Of Thrones. Take a look at the data here https://github.com/mathbeveridge/gameofthrones
- Tried different models. Results below is from a Mixed membership stochastic block model - MMSBM.

2019 proj.: Analysis group membership in Social Networks II

1. For $n = 1..N$
   1. draw $\boldsymbol{\pi}_n \sim \text{Dirichlet}(\boldsymbol{\alpha})$
2. For $k, k' = \{1, 2, \ldots, K\} \times \{1, 2, \ldots, K\}$ where $k \neq k'$
   1. draw $\lambda_{k,k'} \sim \text{Gamma}(a, b)$
3. For k in $1..K$
   1. draw $\lambda_{k,k} \sim \text{Gamma}(c, d)$
4. For all pairs $i, j \in \{1, 2, ...N\} \times \{1, 2, ...N\}$ where $i \neq j$
   1. $x_{i,j} \sim \text{Poisson}(\boldsymbol{\pi}_i^T \lambda \boldsymbol{\pi}_j)$;

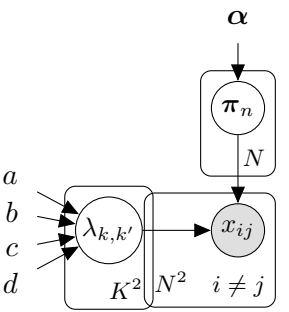


Figure: PGM for Mixed membership stochastic block model

2019 proj.: Analysis group membership in Social Networks III

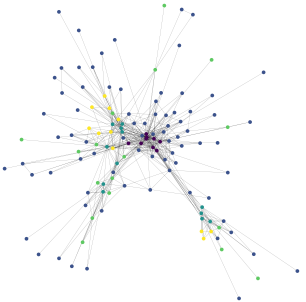


Figure: Each node is a GoT character from season 5. Colors indicate one of 5 inferred groups.

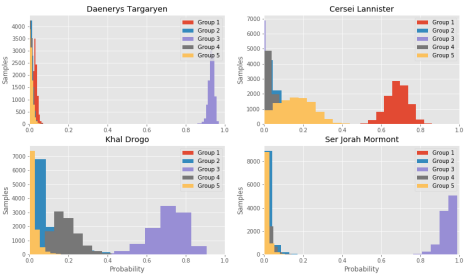


Figure: The histogram shows posterior probabilities of belonging to one of 5 groups for 4 main characters of GoT

2019 project: Bayesian semi-parametric soccer analysis

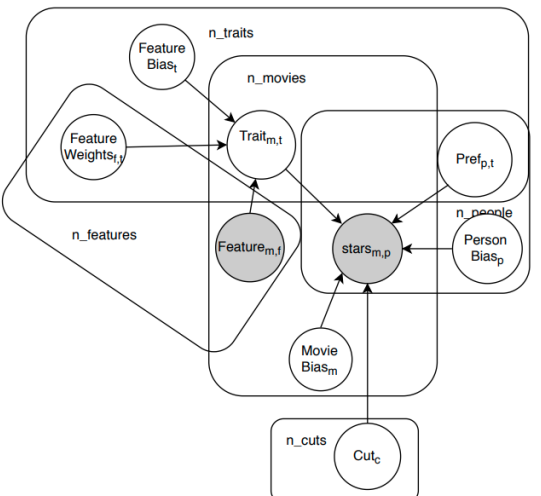
- Apply the time-to-event analysis methods to analyze the intensity with which goal scoring occurs in soccer matches
- Tried different time-to-event models from the class of Cox proportional hazards models and extend these to incorporate non-linear effects by use of the highly flexible Gaussian processes from the Bayesian non-parametrics toolbox
- The hazard in the i-th match at time interval j is

$$\lambda_{ij} = \lambda_j \exp(\beta_1 * skill\_gap + \beta_2(t) * x\_goal)$$

- Baseline of Cox model is a GP: $\log \lambda_0 \sim \mathcal{GP}(\mathbf{0}, K(t, t'))$
- Implemented everything in STAN and did a very good experimental evaluation

2019 project: Movie recommendations - MovieLens dataset

- Inspired by the example in Bishop's "Model-based Machine Learning" book, but with a good mix of creativity, lots of experimentation and insightful discussions

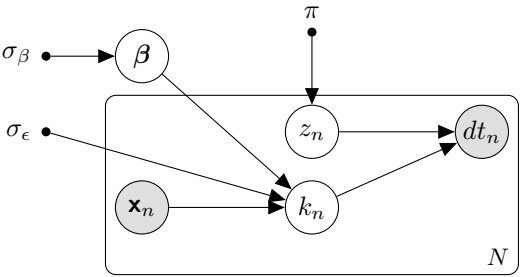


Take-away ideas

- Dwell time prediction in Copenhagen (Movia)
- Mixture of (interpretable) experts
- Predicting COVID-19 spread

Dwell time prediction in Copenhagen (Movia)

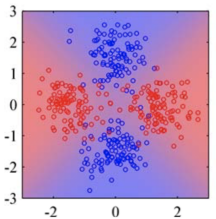
- Remember the dwell time example PGM?



- It was quite naive in many aspects
- The goal of this project is to make a more realistic model of dwell time
  - Combine with weather information (we already have it)
  - Allow the probability of stopping $p(z_n = 1)$ to depend on other features (previous buses, previous stops for that bus, weather, etc.)
  - Similar for the k_n variables: turn it into a time-series model!
  - Can use Gaussian processes instead of linear relationships
- Real data from Movia for a few bus stops in Copenhagen

Mixture of (interpretable) experts

- We often seek interpretable models (e.g. linear models, like linear regression and logistic regression)
- However, their linear assumptions are typically too naive :-(



- We can make progress by subdividing the input space into non-linear regions
  - E.g. fully-connected neural network that, given the input $\mathbf{x}$, assigns it to one of the sub-regions $z_n \sim \text{Bernoulli}(z_n|\text{NNet}(\mathbf{x}, \boldsymbol{\theta}))$
  - For each sub-region, build an interpretable model (e.g. logistic regression, discrete choice model, Poisson regression)
- Each linear model only has to "specialize" in modelling well the data in its sub-region!

Mixture of (interpretable) experts

- Goal: jointly infer non-linear sub-region assignment model that subdivides the input space, and a linear interpretable model for each sub-region
- A possible generative story:

1. Draw NNet parameters $\boldsymbol{\theta} \sim \mathcal{N}(\boldsymbol{\theta}|\mathbf{0}, \mathbf{I})$
2. For each sub-region $r \in \{1, \ldots, R\}$
    1. Draw coefficients of sub-region-specific linear model $\boldsymbol{\eta}_r \sim \mathcal{N}(\boldsymbol{\eta}_r|\mathbf{0}, \mathbf{I})$
3. For each data point $\mathbf{x}_n$
    1. Draw sub-region assignment $z_n \sim \text{Bernoulli}(z_n|\text{NNet}(\mathbf{x}, \boldsymbol{\theta}))$
    2. Draw class label from corresponding linear model $y_n \sim \text{Bernoulli}(y_n|\text{Sigmoid}(\mathbf{x}_n, \boldsymbol{\eta}_{z_n}))$

Predicting COVID-19

- Unfortunately, COVID-19 became an unavoidable topic in our world during this semester
- Predicting its spread is by no means a minor task, but it is a very important one and very prone to a PGM way of thinking:
  - It can be cast as a multivariate linear dynamical system (confirmed, deaths, recovered)
  - One can see clear dependencies (deaths and recovered clearly depend on confirmed)
  - There are many other factors that could be useful to predict (e.g. population density, WHO indicators of the country, which measures were taken)
- All data is available here:

  https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_covid_19_time_series/