Hierarchical linear regression

ISTA 410 / INFO 510: Bayesian Modeling and Inference

U. of Arizona School of Information

October 19, 2020

Outline

Last time:

- Model comparison using WAIC / approximate LOO-CV

Today (and forthcoming):

- Hierarchical and generalized linear models
- Mixture models
- Modal approximations, EM and related algorithms

Logistics

Second half of the semester

Rough outline of the rest of the semester:

- Hierarchical regression, GLMs, mixtures (2-3 weeks)
- Graphical models; Bayes and Markov networks (2 weeks)
- Time series models; HMM and Kalman filters (2 weeks)
- Gaussian processes (1 week)
- ???

Midterm and “participation”

- Midterm: an assignment spanning the various topics we’ve covered
    - Not longer than a regular homework, but less restricted to a particular topic
- Solution presentations
    - Pick a problem from a list (or choose your own)
    - Prepare a solution and give a short presentation of it at the beginning of class

Hierarchical linear models

Recap: linear regression as a Bayesian model

Remember the basic framework we had for a linear model in the Bayesian setting:

$$
\begin{aligned}
y_i &\sim \mathrm{Normal}(\mu_i, \sigma) \\
\mu_i &= \alpha + \beta \cdot \mathrm{x_i} \\
\sigma &\sim \mathrm{HalfCauchy}(\phi) \\
\beta_i &\sim \mathrm{Normal}(0, \sigma_\beta) \\
\alpha &\sim \mathrm{Normal}(0, \sigma_\alpha)
\end{aligned}
$$

(different prior choices possible, of course!)

Recap: hierarchical models

Recall the idea of a hierarchical model:

- Observations are grouped into clusters
- Model parameters for each group come from a prior distribution dependent on population-level *hyperparameters*
- Allows for "partial pooling"; clusters don't all have the same model parameters, but some information is shared across clusters
- Effect: shrinkage toward population parameters, especially for clusters with few observations

Non-Bayesian terminology

There are a few pieces of terminology that are common in the frequentist statistical literature that correspond to these Bayesian concepts:

- *fixed-effects*: the pooled model; same coefficients across all groups
- *random-effects*: an unpooled model; varying model coefficients across groups
- *mixed-effects*: a hierarchical model; effects are varying, but not completely decoupled (also, can have some effects pooled, some not)

Example

Example from BDA sec. 15.2: US presidential election forecasting

- US presidential elections carried out on a state-by-state basis
- Idea: forecast the vote shares in each state based on state, regional, national variables
- Example variables:
  - Vote share from previous year
  - Economic growth
  - Demographic data

Example

Preliminary model: fixed effects (fully pooled)

- Fit an ordinary linear regression (with some normalizing priors) to all of the predictors, pooling all years 1948 - 1988
- Hold out 1992 for testing
- Evaluate model with a posterior predictive check
- Upgrade to a hierarchical model
- Compare the models by their forecasts and by WAIC/LOO

Model structure

Model structure for the fixed-effects model:

$$
\begin{aligned}
y_i &\sim \mathrm{Normal}(\mu_i, \sigma) \\
\mu_i &= \alpha + \beta \cdot \mathrm{x_i} \\
\beta_i &\sim \mathrm{Normal}(0, 1) \\
\sigma &\sim \mathrm{HalfCauchy}(5) \\
\alpha &\sim \mathrm{Normal}(0, 0.2)
\end{aligned}
$$

y = Democratic vote share by state

x_i = various national/state predictors, region code

A posterior predictive check

One of the clear problems: the model cannot account for correlations between states in a single year

- In a year where candidate A performs better than expected in State X, we expect them to also do better than expected in State Y
- When fitting the model, our fixed-effects model treats all years as equivalent
- To detect this, BDA examines the average national residuals
  - For each sampled β, compute the residual $(y_i - \mu_i)$
  - Average residuals across states ("nationwide realized residual"); root-sum-square
  - Compute the same for draws y^{rep} from posterior predictive

A posterior predictive check

The fixed-effects model “fails” the posterior predictive check: the residuals from replicated values are considerably smaller than those from observed values, suggesting that there is additional variability in the observations that is not captured by the model.

A hierarchical model

In the book, the model is expanded by adding 11 national yearly predictors and 44 regional ones:

$$\begin{aligned}
y_{i,t} &\sim \text{Normal}(\mu_i, \sigma) \\
\mu_{i,t} &= \alpha + \beta \cdot \mathsf{x_i} + \delta_t + \gamma_{r(i),t} \\
\delta_t &\sim \text{Normal}(0, \tau_\delta) \\
\gamma_{r(i),t} &\sim \text{Normal}(0, \tau_{\gamma_r}) \\
\beta_i &\sim \text{Normal}(0, 1) \\
\sigma &\sim \text{HalfCauchy}(5) \\
\tau_. &\sim \text{HalfCauchy}(5) \\
\alpha &\sim \text{Normal}(0, 0.2)
\end{aligned}$$

I'm going to do something simpler: δ_t only

Our hierarchical model

$$
\begin{aligned}
y_{i,t} &\sim \text{Normal}(\mu_i, \sigma) \\
\mu_{i,t} &= \alpha + \beta \cdot \mathsf{x_i} + \delta_t \\
\delta_t &\sim \text{Normal}(0, \tau_\delta) \\
\beta_i &\sim \text{Normal}(0, 1) \\
\tau_\delta &\sim \text{HalfCauchy}(5)\sigma & \sim \text{HalfCauchy}(5) \\
\alpha &\sim \text{Normal}(0, 0.2)
\end{aligned}
$$

Our hierarchical model

Some preliminary notes:

- Notice: β not directly affected by the hierarchical structure (still the same β estimated for all years)
- Mean of δ not estimated, only variance
- Partial pooling on δ_t, not on β

Notes on the hierarchical model

Some observations:

- Principal effect on forecasting: more uncertainty
- Similar state-by-state point predictions – why? No estimate for δ_{1992}
- Real world models: use polling data, etc. to estimate values for the extra variables, incorporate these into the forecast
- Still, some improvement in certain forecast quantities; e.g., mode of electoral vote share matches true values

Summary

Today:

- Hierarchical linear models, partial pooling

Next:

- More on hierarchical regression
- GLMs
- Mixture models