Lab 11

Qiaoyu Liang

2023-04-02

Overview

In this lab you'll be fitting a second-order P-Splines regression model to foster care entries by state in the US, projecting out to 2030.

```
library(tidyverse)
library(here)
library(rstan)
library(tidybayes)
source(here("getsplines.R"))
```

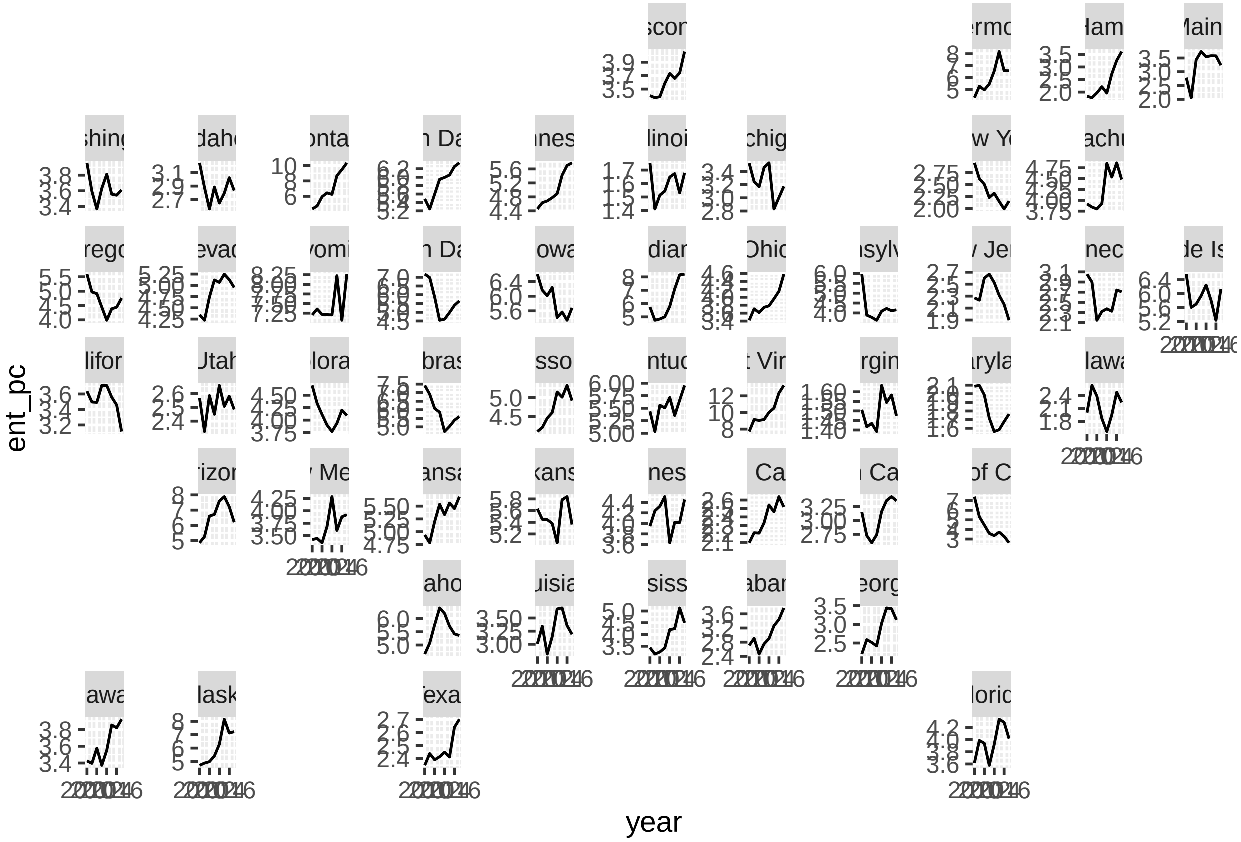
Here's the data

```
d <- read_csv(here("fc_entries.csv"))
```

Question 1

Make a plot highlighting trends over time by state. Might be a good opportunity to use `geofacet`. Describe what you see in a couple of sentences.

```
library(geofacet)

d |>
  ggplot(aes(year, ent_pc)) +
  geom_line()+
  facet_geo(~state, scales = "free_y")
```



The above plot demonstrates diverse patterns across various states. While we can see a generally increasing trend for some certain states, we can also see some states have a generally decreasing trend. Moreover, some states show no clear trend.

Question 2

Fit a hierarchical second-order P-Splines regression model to estimate the (logged) entries per capita over the period 2010-2017. The model you want to fit is

$$
\begin{aligned}
y_{st} &\sim N(\log \lambda_{st}, \sigma^2_{y,s}) \\
\log \lambda_{st} &= \alpha_k B_k(t) \\
\Delta^2 \alpha_k &\sim N(0, \sigma^2_{\alpha,s}) \\
\log \sigma_{\alpha,s} &\sim N(\mu_\sigma, \tau^2)
\end{aligned}
$$

Where $y_{s,t}$ is the logged entries per capita for state s in year t. Use cubic splines that have knots 2.5 years apart and are a constant shape at the boundaries. Put standard normal priors on standard deviations and hyperparameters.

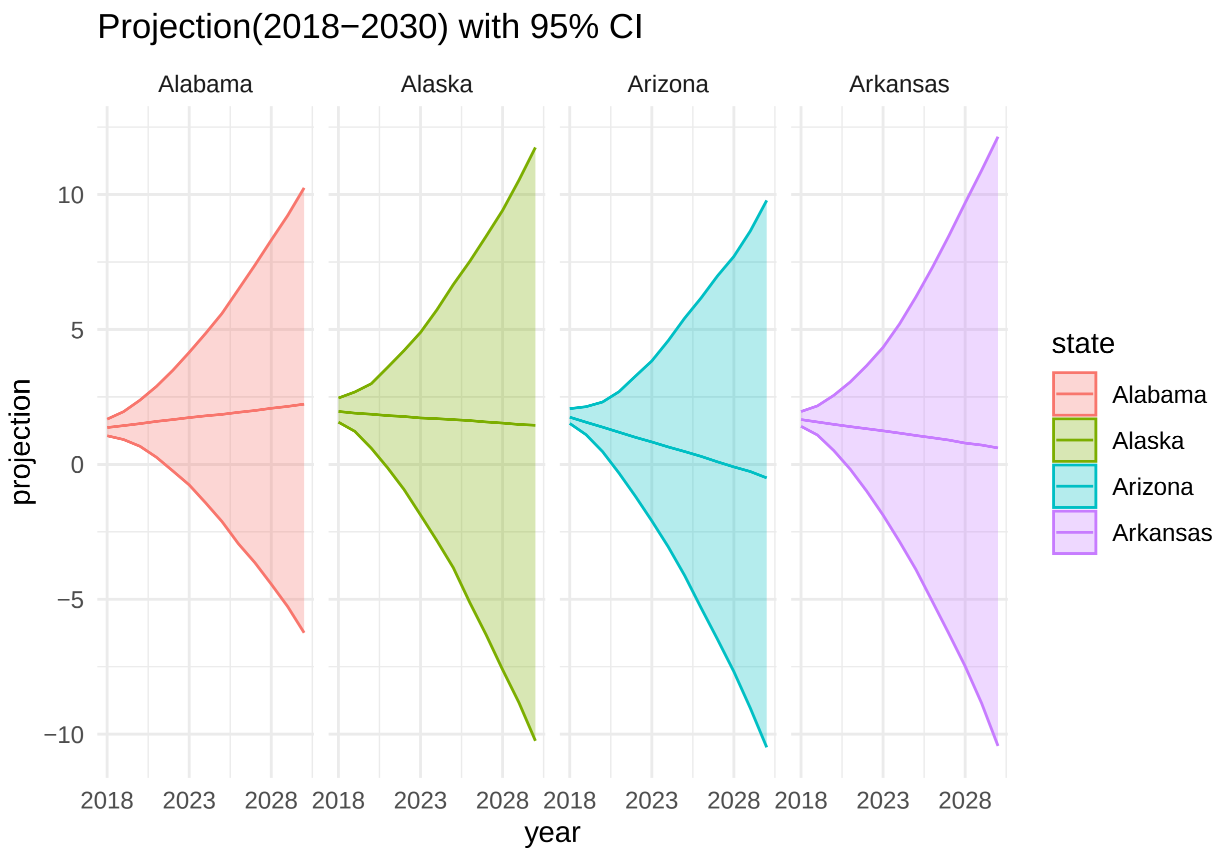
```r
years <- unique(d$year)
N <- length(years)
y <- log(d |>
  select(state, year, ent_pc) |>
  pivot_wider(names_from = "state", values_from = "ent_pc") |>
  select(-year) |>
  as.matrix())

res <- getsplines(years, 2.5)
B <- res$B.ik
K <- ncol(B)

stan_data <- list(N = N, y = y, K = K, S = length(unique(d$state)),
                  B = B)

mod <- stan(data = stan_data, file = "lab11.stan")
```

Question 3

Project forward entries per capita to 2030. Pick 4 states and plot the results (with 95% CIs). Note the code to do this in R is in the lecture slides.

```
proj_years <- 2018:2030
# Note: B.ik are splines for in-sample period
# has dimensions i (number of years) x k (number of knots)
# need splines for whole period
B.ik_full <- getsplines(c(years, proj_years), 2.5)$B.ik
K <- ncol(B) # number of knots in sample
K_full <- ncol(B.ik_full) # number of knots over entire period
proj_steps <- K_full - K # number of projection steps
# get your posterior samples
alphas <- rstan::extract(mod)[["alpha"]]
sigmas <- rstan::extract(mod)[["sigma_alpha"]] # sigma_alpha
sigma_ys <- rstan::extract(mod)[["sigma_y"]]
nsims <- nrow(alphas)
states = unique(d$state)
# first, project the alphas
alphas_proj <- array(NA, c(nsims, proj_steps, length(states)))
set.seed(1098)
# project the alphas
for(j in 1:length(states)){
first_next_alpha <- rnorm(n = nsims, mean = 2*alphas[,K,j] - alphas[,K-1,j],sd = sigmas[,j])
second_next_alpha <- rnorm(n = nsims, mean = 2*first_next_alpha - alphas[,K,j], sd = sigmas[,j])
alphas_proj[,1,j] <- first_next_alpha
alphas_proj[,2,j] <- second_next_alpha
# now project the rest
for(i in 3:proj_steps){ #!!! not over years but over knots
alphas_proj[,i,j] <- rnorm(n = nsims,
mean = 2*alphas_proj[,i-1,j] - alphas_proj[,i-2,j],
sd = sigmas[,j])
}
}
# now use these to get y's
y_proj <- array(NA, c(nsims, length(proj_years), length(states)))
for(i in 1:length(proj_years)){ # now over years
for(j in 1:length(states)){
all_alphas <- cbind(alphas[,,j], alphas_proj[,,j] )
this_lambda <- all_alphas %*% as.matrix(B.ik_full[length(years)+i, ])
y_proj[,i,j] <- rnorm(n = nsims, mean = this_lambda, sd = sigma_ys[,j])
}
}
# then proceed as normal to get median, quantiles etc
```

```
# Alabama, Alaska, Arizona, and Arkansas are chosen.
y_projs <- y_proj[,,c(1, 2, 3, 4)]
state_names <- unique(d$state)[c(1, 2, 3, 4)]
y_proj_median <- apply(y_projs, c(2, 3), median)
y_proj_upp <- apply(y_projs, c(2, 3), function(x) quantile(x, 0.975))
y_proj_low <- apply(y_projs, c(2, 3), function(x) quantile(x, 0.025))
df <- data.frame(
  year = rep(proj_years, length(c(1, 2, 3, 4))),
  med = as.vector(y_proj_median),
  upp = as.vector(y_proj_upp),
  low = as.vector(y_proj_low),
  state = rep(state_names, each = length(proj_years))
)
ggplot(df, aes(x = year, y = med, group = state, color = state)) +
  geom_line() +
  geom_ribbon(aes(ymin = low, ymax = upp, fill = state), alpha = 0.3) +
  labs(x = "year",
       y = "projection",
       title = "Projection(2018-2030) with 95% CI") +
  facet_wrap(~state, ncol = length(c(1, 2, 3, 4))) +
  scale_x_continuous(breaks = seq(min(proj_years), max(proj_years), by = 5)) +
  theme_minimal()
```



Question 4 (bonus)

P-Splines are quite useful in structural time series models, when you are using a model of the form

$$f(y_t) = \text{systematic part} + \text{time-specific deviations}$$

where the systematic part is model with a set of covariates for example, and P-splines are used to smooth data-driven deviations over time. Consider adding covariates to the model you ran above. What are some potential issues that may happen in estimation? Can you think of an additional constraint to add to the model that would overcome these issues?

Identification could be a potential issue in this setting. To overcome this issue, we can constrain splines so that splines sum to zero.