Homework 05

Spencer Pease

5/11/2020

Questions

Q1

Q1.a

For a student i in school j, our Bayesian random effects one-way analysis of variance model is written as:

$$
\begin{aligned}
y_i &= \alpha_{j[i]} + \epsilon_i, \\
\epsilon_i &\overset{iid}{\sim} N(0, \sigma_y^2), \\
\alpha_j &\overset{iid}{\sim} N(\mu_\alpha, \sigma_\alpha^2)
\end{aligned}
$$

where the standard deviation of error in estimating individual student performance (σ_y), the mean performance across all schools (μ_α), and the standard deviation in performance across all schools (σ_α) are the unknown parameters to be estimated.

Q1.b

For this scenario, we set the prior distributions to:

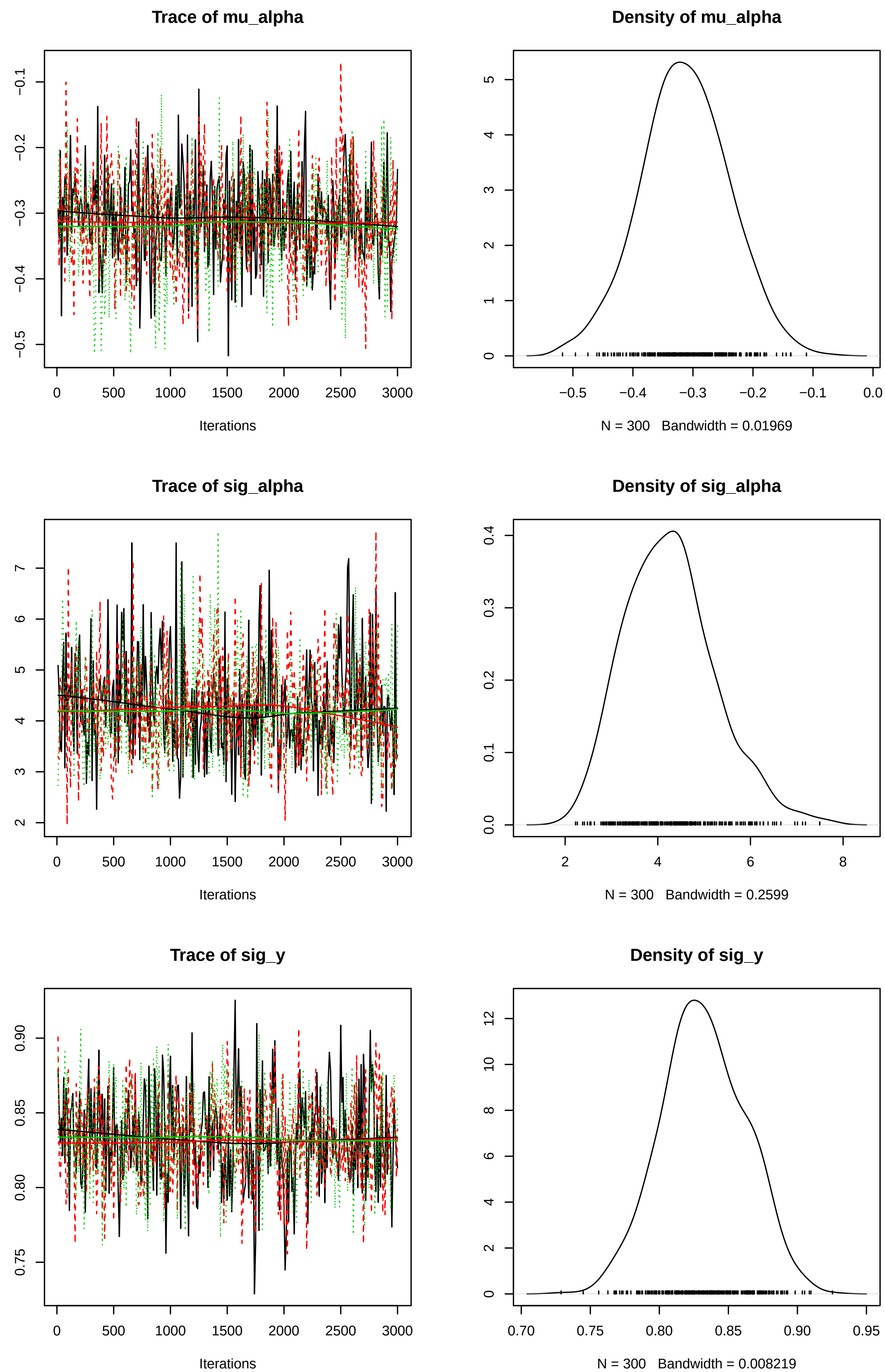
$$
\begin{aligned}
\mu_\alpha &\sim N(0, 0.0001) \\
\sigma_\alpha &\sim \text{Gamma}(1, 0.0001) \\
\sigma_y &\sim \text{Gamma}(1, 0.0001)
\end{aligned}
$$

Q1.c

```
## Compiling model graph
##    Resolving undeclared variables
##    Allocating nodes
## Graph information:
##    Observed stochastic nodes: 1672
##    Unobserved stochastic nodes: 98711
##    Total graph size: 100388
##
## Initializing model
```

This model was fit using JAGS.

Q1.d


Trace of mu_alpha
Iterations
Density of mu_alpha
N = 300   Bandwidth = 0.01969
Trace of sig_alpha
Iterations
Density of sig_alpha
N = 300   Bandwidth = 0.2599
Trace of sig_y
Iterations
Density of sig_y
N = 300   Bandwidth = 0.008219

Running the model for 3000 iterations is enough to achieve convergence.

Q1.e

Table 1: Summary of posterior distribution of prior parameters

Parameter	2.5%	25%	50%	75%	97.5%
mu	-0.460	-0.361	-0.314	-0.264	-0.171
sigma alpha	2.602	3.544	4.221	4.825	6.388
sigma y	0.772	0.812	0.831	0.853	0.889

We can get the summary of our posterior distributions, and look at the density plots of the distributions in question *(1.d)*.

Q2

Q2.a

Q2.b

Q2.c

Q3

Q3.a

Table 2: Total fertility rates, Netherlands, 1950-2020

Period Start	TFR
1950	3.052
1955	3.097
1960	3.166
1965	2.795
1970	2.100
1975	1.598
1980	1.515
1985	1.555
1990	1.592
1995	1.599
2000	1.740
2005	1.746
2010	1.732
2015	1.660

The start of Phase III of the fertility model is defined by two consecutive five-year increases of TFR while staying below a TFR of 2. Looking at TFR data for the Netherlands, we see that Phase III starts with the period beginning in **1985**.

Q3.b

We now fit an order 1 autoregressive model to the subset of Netherlands TFR data in Phase III, and extract some model parameters below. *Note that the AR(1) model was fit using the "mle" method.*

Table 3: Netherlands Phase III AR(1) model parameters

Mean	AR param.	Error var.
1.644	0.608	0.003

Q3.c

A post-transition (Phase III) model of TFR change, as proposed in *Lee RD, Tuljapurkar S (1994), Stochastic population forecasts for the United States: beyond high, medium, and low*, is defined as:

$$f_{c,t+1} \sim N(\mu + \rho(f_{c,t} - \mu), s^2)$$

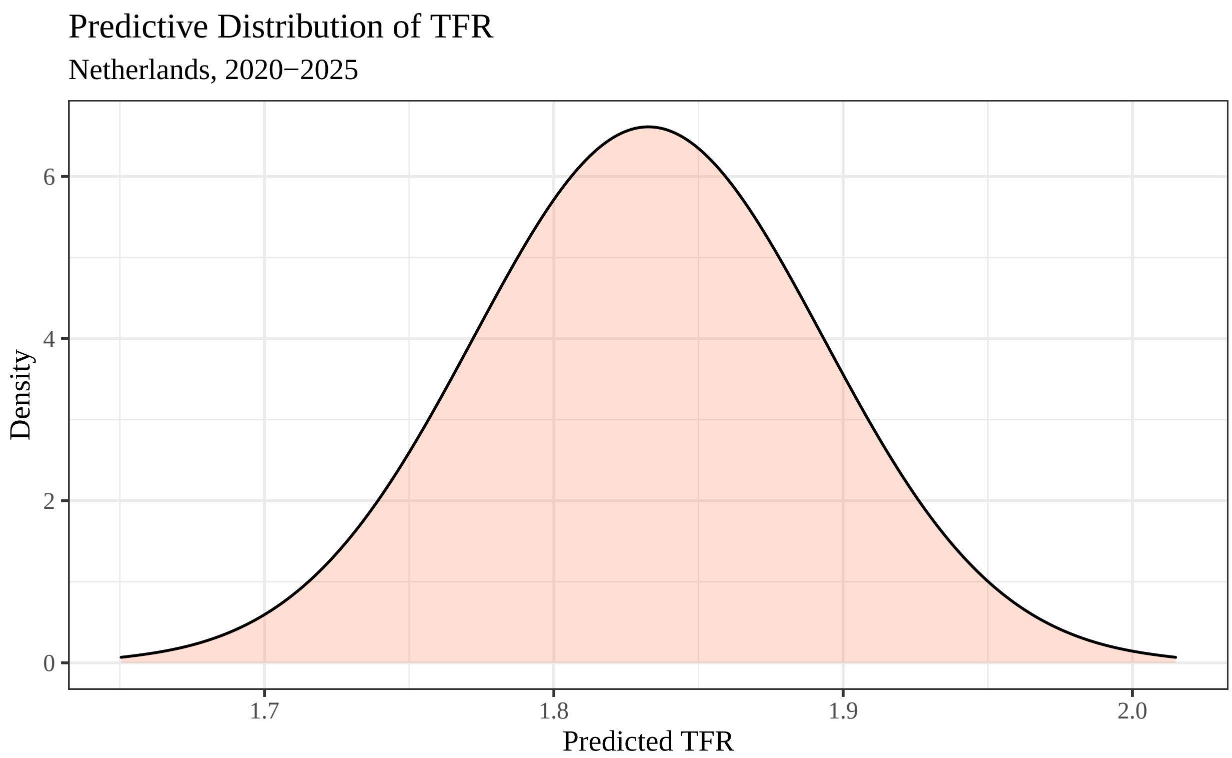
where $f_{c,t}$ is the TFR of country c in the five-year period starting at t, μ is the approximate replacement-level fertility 2.1, ρ is the autoregressive parameter, and s is the standard deviation of the random errors.

Plugging in these values, we analytically find the distribution of Netherlands TFR for 2020-2025 to be:

$$f_{c,2020} \sim N(1.833, 0.059^2)$$

Table 4: Predictive distribution summary of Netherlands TFR, 2020-2025

Mean	Median	2.5% PI	97.5% PI
1.833	1.833	1.717	1.948

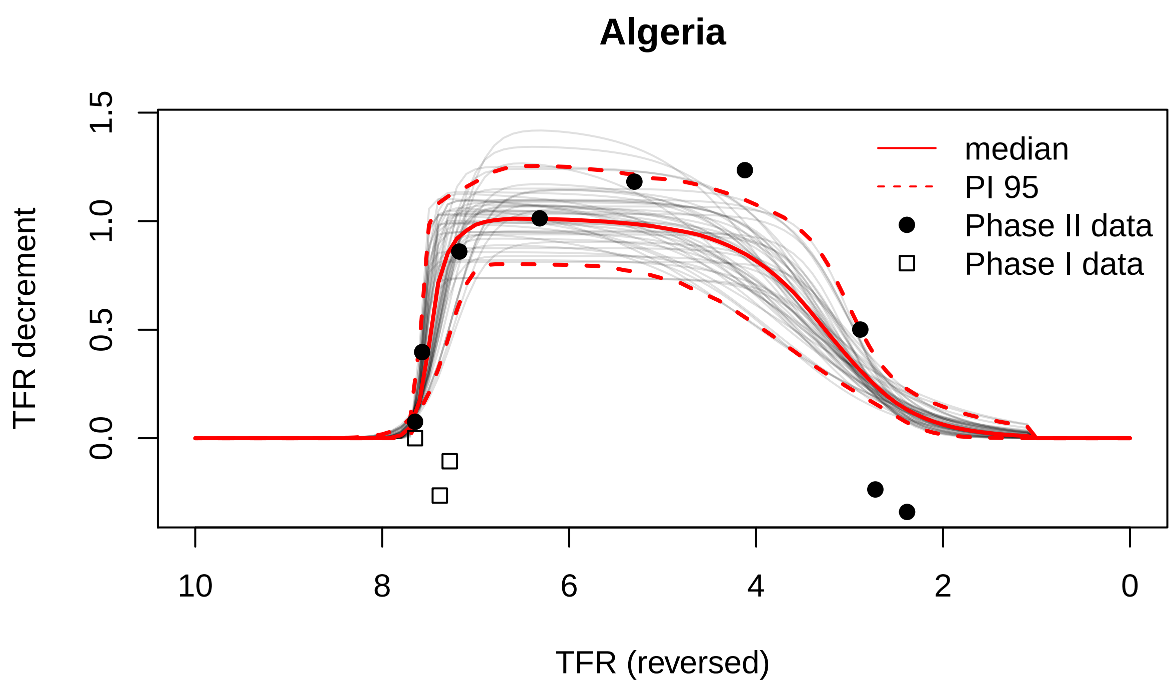


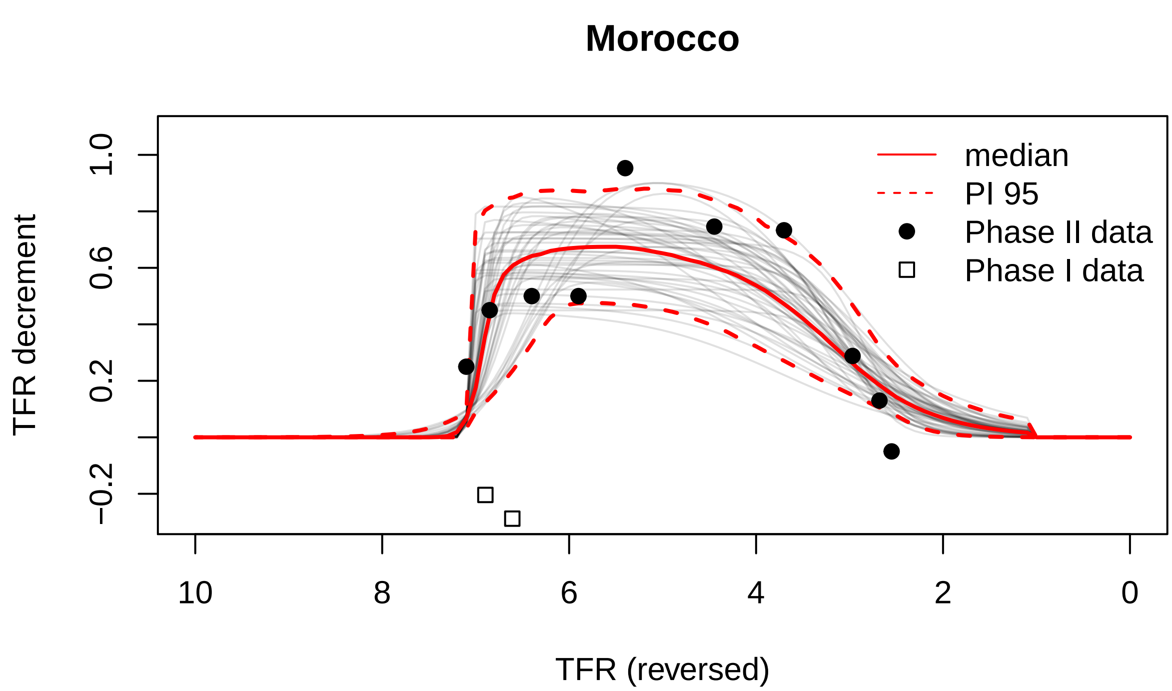
Q4

Q4.a

The fully converged simulation is loaded using the `README` file contained with the data.

Q4.b





From the above graphs comparing the Phase II double logistic models for Algeria and Morocco, we see that the TFR decrements maintains higher values across TFR in Algeria for both the median and 95% PI, suggesting that fertility is declining faster in Algeria than Morocco. We can also observe this trend by looking directly at the TFR for both countries over time:

Table 5: TFR over time

Period Start	Algeria	Morocco
1950	7.278	6.608
1955	7.384	6.896
1960	7.648	7.100
1965	7.648	6.850
1970	7.572	6.400
1975	7.175	5.900
1980	6.315	5.400
1985	5.302	4.430
1990	4.120	3.700
1995	2.885	2.965
2000	2.384	2.670
2005	2.724	2.530
2010	2.960	2.600
2015	3.050	2.420

Q4.c

By getting the TFR trajectories for Algeria and Morocco, we can find the posterior predictive probability for many conditions.

First, we can determine the probability of Algeria having a higher TFR than Morocco in each five-year period from 2020 through 2095 by finding the mean number of times Algeria has a higher TFR than Morocco across all simulations:

Period Start	Pr(DZA > MAR)
2020	0.704
2025	0.646
2030	0.637
2035	0.622
2040	0.604
2045	0.624
2050	0.619
2055	0.619
2060	0.635
2065	0.620
2070	0.612
2075	0.633
2080	0.609
2085	0.617
2090	0.616
2095	0.618

We can also find the probability that the TFR of Algeria will be higher than that of Morocco in all five-year periods from 2020 through 2095 by finding the mean number of simulations where Algeria has a higher TFR than Morocco for all periods. This value is calculated to be **0.207**.

Appendix

```r
# Prep work ------------------------------------------------------------

# Load libraries
library(dplyr)
library(tidyr)
library(ggplot2)
library(R2jags)
library(bayesTFR)

# Data
data("egsingle", package = "mlmRev")
data("tfr", package = "wpp2019")
tfr_sim_dir <- "./data/sim01192018"

tfr_all <- tfr %>%
  select(-country_code, -last.observed) %>%
  pivot_longer(
    -name,
    values_to = "tfr",
    names_pattern = "^(.*)-",
    names_to = "year",
    names_ptypes = list(year = numeric())
  )

# Control randomness
set.seed(9876)


# Question 1 -----------------------------------------------------------

edu_data <- egsingle %>%
  filter(year == .5) %>%
  select(childid, schoolid, math) %>%
  pivot_wider(names_from = "schoolid", values_from = "math") %>%
  tibble::column_to_rownames("childid") %>%
  as.matrix()


# Question 1c ----------------------------------------------------------

n_students <- dim(edu_data)[1]
n_schools <- dim(edu_data)[2]

edu_model <- jags.model(
  "./edu_jags_model.txt",
  data = list(Y = edu_data, n_students = n_students, n_schools = n_schools),
  n.chains = 3
)

edu_sample_priors <- coda.samples(
  model = edu_model,
```

```r
  variable.names = c("mu_alpha", "sig_y", "sig_alpha"),
  n.iter = 3000,
  thin = 10
)

edu_sample_schools <- coda.samples(
  model = edu_model,
  variable.names = "alpha_j",
  n.iter = 3000,
  thin = 10
)


# Question 1d ------------------------------------------------------------

plot(edu_sample_priors)


# Question 1e ------------------------------------------------------------

edu_quantiles_tbl <- summary(edu_sample_priors)[["quantiles"]] %>%
  as_tibble() %>%
  mutate(Parameter = c("mu", "sigma alpha", "sigma y")) %>%
  select(Parameter, everything())


knitr::kable(
  edu_quantiles_tbl,
  booktabs = TRUE,
  digits = 3,
  caption = "Summary of posterior distribution of prior parameters"
)


# Question 2 -------------------------------------------------------------

hnd_tfr <- tfr_all %>% filter(name == "Honduras") %>% select(-name)


# Question 2a ------------------------------------------------------------

# run.tfr.mcmc() or nls()?


# Question 3 -------------------------------------------------------------

nld_tfr <- tfr_all %>% filter(name == "Netherlands") %>% select(-name)


# Question 3a ------------------------------------------------------------

nld_phase3_year <- nld_tfr %>%
  arrange(year) %>%
```

```r
  filter(tfr < 2) %>%
  mutate(
    year_diff = lead(year) - year,
    period_5 = year_diff == 5 & lag(year_diff) == 5,
    two_increases = tfr > lag(tfr, 1) & tfr < lead(tfr, 1)
  ) %>%
  filter(period_5 & two_increases) %>%
  slice(1) %>%
  pull(year)

knitr::kable(
  nld_tfr,
  booktabs = TRUE,
  digits = 3,
  col.names = c("Period Start", "TFR"),
  caption = "Total fertility rates, Netherlands, 1950-2020"
)

# Question 3b -------------------------------------------------------------

nld_model <- nld_tfr %>%
  filter(year >= nld_phase3_year) %>%
  pull(tfr) %>%
  ar(aic = FALSE, order.max = 1, method = "mle")

nld_model_results <- tibble(
  Mean = nld_model$x.mean,
  `AR param.` = nld_model$ar,
  `Error var.` = nld_model$var.pred
)

knitr::kable(
  nld_model_results,
  booktabs = TRUE,
  digits = 3,
  caption = "Netherlands Phase III AR(1) model parameters"
)

# Question 3c -------------------------------------------------------------

nld_ar_rho <- nld_model$ar
rep_tfr <- 2.1
nld_tfr_2015 <- nld_tfr %>% filter(year == 2015) %>% pull(tfr)

nld_pred_mean <- rep_tfr + nld_ar_rho * (nld_tfr_2015 - rep_tfr)
nld_pred_sd <- sqrt(nld_model$var.pred)

nld_pred_dist <- qnorm(seq(.001, .999, .001), mean = nld_pred_mean, sd = nld_pred_sd)

nld_pred_tbl <- tibble(
  Mean = nld_pred_mean,
  Median = median(nld_pred_dist),
  `2.5% PI` = Mean - 1.96 * nld_pred_sd,
```

```r
  `97.5% PI` = Mean + 1.96 * nld_pred_sd
)

knitr::kable(
  nld_pred_tbl,
  booktabs = TRUE,
  digits = 3,
  caption = "Predictive distribution summary of Netherlands TFR, 2020-2025"
)

ggplot(tibble(nld_pred_dist), aes(x = nld_pred_dist)) +
  geom_density(fill = "coral", alpha = .25) +
  theme_bw() +
  theme(text = element_text(family = "serif")) +
  labs(
    title = "Predictive Distribution of TFR",
    subtitle = "Netherlands, 2020-2025",
    x = "Predicted TFR",
    y = "Density"
  )


# Question 4 --------------------------------------------------------------

tfr_dza_mar <- tfr_all %>%
  filter(name %in% c("Algeria", "Morocco")) %>%
  pivot_wider(names_from = name, values_from = "tfr") %>%
  rename(`Period Start` = year)


# Question 4.a ------------------------------------------------------------

tfr_phase2_mcmc  <- get.tfr.mcmc(tfr_sim_dir)
tfr_phase3_mcmc <- get.tfr3.mcmc(tfr_sim_dir)
tfr_pred  <- get.tfr.prediction(tfr_sim_dir)


# Question 4.b ------------------------------------------------------------

DLcurve.plot(
  tfr_phase2_mcmc,
  country = "Algeria",
  nr.curves = 50,
  pi = 95
)

DLcurve.plot(
  tfr_phase2_mcmc,
  country = "Morocco",
  nr.curves = 50,
  pi = 95
)
```

```r
knitr::kable(
  tfr_dza_mar,
  booktabs = TRUE,
  digits = 3,
  caption = "TFR over time"
)

# Question 4c -----------------------------------------------------------

tfr_traj_dza <- get.tfr.trajectories(tfr_pred, country = "Algeria")[-(1:2), ]
tfr_traj_mar <- get.tfr.trajectories(tfr_pred, country = "Morocco")[-(1:2), ]

prob_tfr_dza_higher <- rowMeans(tfr_traj_dza > tfr_traj_mar)

prob_tfr_dza_higher_tbl <- tibble(
  period_start = as.integer(names(prob_tfr_dza_higher)) - 3,
  prob_dza_higher = prob_tfr_dza_higher
)

prob_tfr_dza_higher_all <-
  sum(apply(tfr_traj_dza > tfr_traj_mar, 2, all)) / ncol(tfr_traj_dza)

knitr::kable(
  prob_tfr_dza_higher_tbl,
  booktabs = "TRUE",
  digits = 3,
  col.names = c("Period Start", "Pr(DZA > MAR)")
)
```