Applied Bayesian Modeling module 11:
Model checking - Part 2 (out-of-sample)

Leontine Alkema, lalkema@umass.edu
Fall 2022

Model checking

- We can now fit a whole range of Bayesian models (using Stan).
- Important question: how well does a model fit the data?
- To discuss:
  - Part 1: in-sample validation
    General diagnostic plots, e.g. using residuals, and posterior predictive checks (based on simulating data from the fitted model)
  - Part 2: Measures of predictive accuracy based on out-of-sample validation

Model checks based on cross-validation

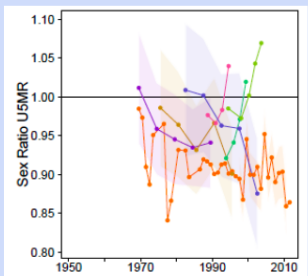
- Issue with in-sample checks like posterior predictive checks: we use the data twice
- Out-of-sample model validation: check how well a model predicts "new observations" that were not used for model fitting
- Set-up:
  - Fit the model to a training set and validate predictive accuracy of the model for the left-out test data
  - Summarize model performance, e.g., in terms of measures for accuracy of
    - point predictions
    - probabilistic predictions (evaluating the distribution as opposed to a point prediction)
- Good: actual predictions for new data, usually quite re-assurring if you find that your model predicts left-out data well.

Out-of-sample validation: examples from Alkema lab

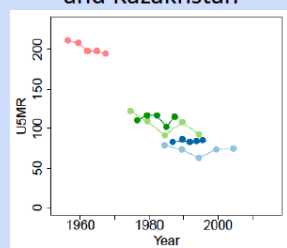
- We develop Bayesian models to estimate and project population health and indicators such as child mortality, maternal mortality, abortion, ..
- Typically, the main challenge is that data are limited for country-periods of interest and/or subject to data quality issues

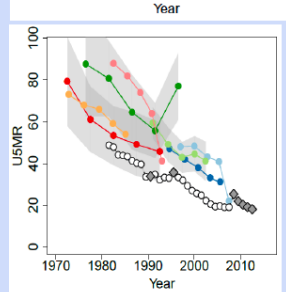
Why we need models to answer these questions...

Sex ratio of child mortality in India

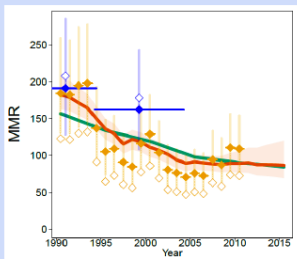


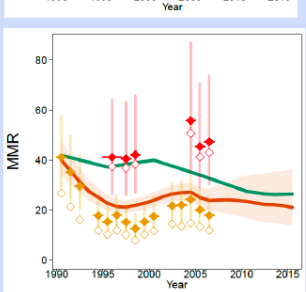
Child mortality in Papua New Guinea and Kazakhstan

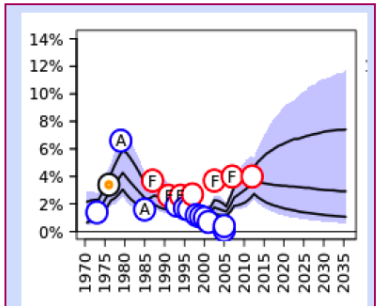




Maternal mortality in Ecuador and Thailand







Use of traditional contraceptive methods in Indonesia

Out-of-sample validation: examples from Alkema lab (ctd)

- Typical validation exercises: leave out data
  - at random (minimum check),
  - after a given year (to check accuracy of projections),
  - for one country at-the-time (to check how well the hierarchical models work to predict outcomes in countries with limited/no data),
  - grouped by survey (if there are several observations per survey)
- Then construct predictive distribution for left-out data points and summarize accuracy of point predictions and distribution

Measures of accuracy for point predictions

- Measures given by some summary of errors e_i,
  e.g. $e_i = y_i - \hat{y}_i$ or $e_i = (y_i - \hat{y}_i)/\hat{y}_i$ (relative errors).
- Examples
  - Mean error: mean(e)
  - Median error: median(e_i)
  - Mean squared error (MSE): mean(e_i^2)
  - Median absolute error: median$(|e_i|)$

Example: maternal mortality ratio (MMR) estimation

- We left out (exercise I) 20% of all MMR observations at random, and (II) all observations after 2007 (to check forecasts).
- Error = observed MMR - posterior median MMR, and Relative error = error/posterior median MMR.

TABLE 4

Validation results based on left-out observations, for developed and developing countries. *The outcome measures are: median error (ME), absolute error (MAE), relative error (MRE) and absolute relative error (MARE) for the MMR (per 100,000 live births), as well as the % of left-out observations below and above the 80% prediction interval (PI) based on the training set. Results for exercise II refer to the most recent left-out observation in each country.*

	# of left-out observations	error in MMR		relative error (%)		outside 80% PI	
		ME	MAE	MRE	MARE	% Below	% Above
Exercise I: appr. 20% of observations were excluded at random							
Developed countries	187	0.4	1.9	3.5	23.7	9.1	5.9
Developing countries	248	1.6	8.3	2.3	16.8	4.4	6.9
Exercise II: all observations in and after 2007 were excluded							
Developed countries	43	0.2	1.5	2.5	30.0	11.6	4.7
Developing countries	80	6.5	17.1	15.2	31.0	7.5	11.2

Alkema et al (2017); "A Bayesian approach to the global estimation of maternal mortality"

Accuracy for probabilistic predictions (evaluating the distribution as opposed to a point prediction)

- Simple but informative summary of prediction accuracy is the coverage of prediction intervals (PIs) for left-out y_is.
- A $(1-\alpha)100\%$ PI for y_i is based on the respective quantiles of $p(\tilde{y}_i|\boldsymbol{y}_{train})$.
- Coverage of PIs refers to the % of left-out observations that fall inside their respective PI (should be close to $\alpha \cdot 100\%$, with $\alpha/2 \cdot 100\%$ above/below).
- See MMR example for 80% PIs:

	# of left-out observations	error in MMR ME	MAE	relative error (%) MRE	MARE	outside 80% PI % Below	% Above
Exercise I: appr. 20% of observations were excluded at random							
Developed countries	187	0.4	1.9	3.5	23.7	9.1	5.9
Developing countries	248	1.6	8.3	2.3	16.8	4.4	6.9
Exercise II: all observations in and after 2007 were excluded							
Developed countries	43	0.2	1.5	2.5	30.0	11.6	4.7
Developing countries	80	6.5	17.1	15.2	31.0	7.5	11.2

Accuracy for probabilistic predictions (ctd)

- Our general aim for probabilistic predictions: sharpness subject to calibration, where
  - Calibration refers to the correct coverage of prediction intervals
  - Sharpness refers to the width of prediction intervals: narrower prediction intervals (sharper predictions) convey more information and are thus preferred.

  Some score rules exist that combine both (to compare models).

Summary model checks based on (actual) cross-validation

- Set-up: Fit the model to a training set and validate predictive accuracy of the model for the left-out test data
- Good: actual predictions for new data, usually quite reassuring if you find that your model predicts left-out
- Bad: computationally intensive, not necessarily clear what data to leave out
- Re computationally intensive...
  Wouldn't it be nice if someone would have figured out some solution to that?

Approximate leave-one-out (LOO) cross-validation

- LOO: use $\boldsymbol{y}_{-i}$, all data except for observation i, to predict y_i.
- Ideally, we would get $p(\tilde{y}_i|\boldsymbol{y}_{-i})$ by fitting the model to $\boldsymbol{y}_{-i}$ but this is usually computationally expensive.
- Common approach: try to estimate $p(\tilde{y}_i|\boldsymbol{y}_{-i})$
- PSIS-LOO refers to using Pareto-smoothed importance sampling (PSIS) to estimate $p(\tilde{y}_i|\boldsymbol{y}_{-i})$.
  - Implemented in R package loo;
    For lots of info on approach and functions, see
    https://mc-stan.org/loo/index.html
  - Technical details outside class material (very accessible 2017 paper, and recently extended https://arxiv.org/abs/1507.02646)
  - Important bonus output: a check to see if the result is reliable or whether you should obtain $p(\tilde{y}_i|\boldsymbol{y}_{-i})$ by fitting the model to $\boldsymbol{y}_{-i}$
- Next: approach, interpretation of main PSIS-LOO results ($\hat{k}$ and ELPDs), further usage, using radon data

Summary of PSIS-LOO approach

- What we want:

$$p(\tilde{y}_i|\boldsymbol{y}_{-i}) \quad = \quad \int p(\tilde{y}_i|\boldsymbol{\theta})p(\boldsymbol{\theta}|\boldsymbol{y}_{-i})d\boldsymbol{\theta} \approx 1/S \sum_{s=1}^{S} p(\tilde{y}_i|\boldsymbol{\theta}^{(s)}), \quad (1)$$

  where $\boldsymbol{\theta}^{(s)} \sim p(\boldsymbol{\theta}|\boldsymbol{y}_{-i})$, so samples from a model fit excluding y_i.
- We do NOT have a model fit excluding y_i but we DO have $\theta^{(s)} \sim p(\boldsymbol{\theta}|\boldsymbol{y})$, samples from a model fit including y_i.
- Approach: calculate weighs $w_i^{(s)}$ associated with each posterior sample $\boldsymbol{\theta}^{(s)} \sim p(\boldsymbol{\theta}|\boldsymbol{y})$ such that

$$p(\tilde{y}_i|\boldsymbol{y}_{-i}) \quad \approx \quad \frac{\sum_{s=1}^{S} p(\tilde{y}_i|\boldsymbol{\theta}^{(s)})w_i^{(s)}}{\sum_{s=1}^{S} w_i^{(s)}} \quad (2)$$

  where $\boldsymbol{\theta}^{(s)} \sim p(\boldsymbol{\theta}|\boldsymbol{y})$.
- In PSIS-LOO approach, a pareto distribution is fitted to obtain the weights.

PSIS-LOO for the radon data

- To calculate the weights, we need to save posterior samples $\log(p(y_i|\boldsymbol{\theta}^{(s)}))$
- In `Stan`, you can add this outcome to the generated quantities block (see example in Rmd)

```
vector[N] log_lik; // pointwise log-likelihood for LOO
for (i in 1:N)
  log_lik[i] = normal_lpdf(y[i] | mu[i], sigma_y);
```

- When using `brm`, you can call the `loo` directly with the model-fit-object as argument

```
library(loo)
loo_fit <- loo(fit)
```

- Observation-specific outcomes are in `loo_fit$pointwise`.

What are "k estimates"?

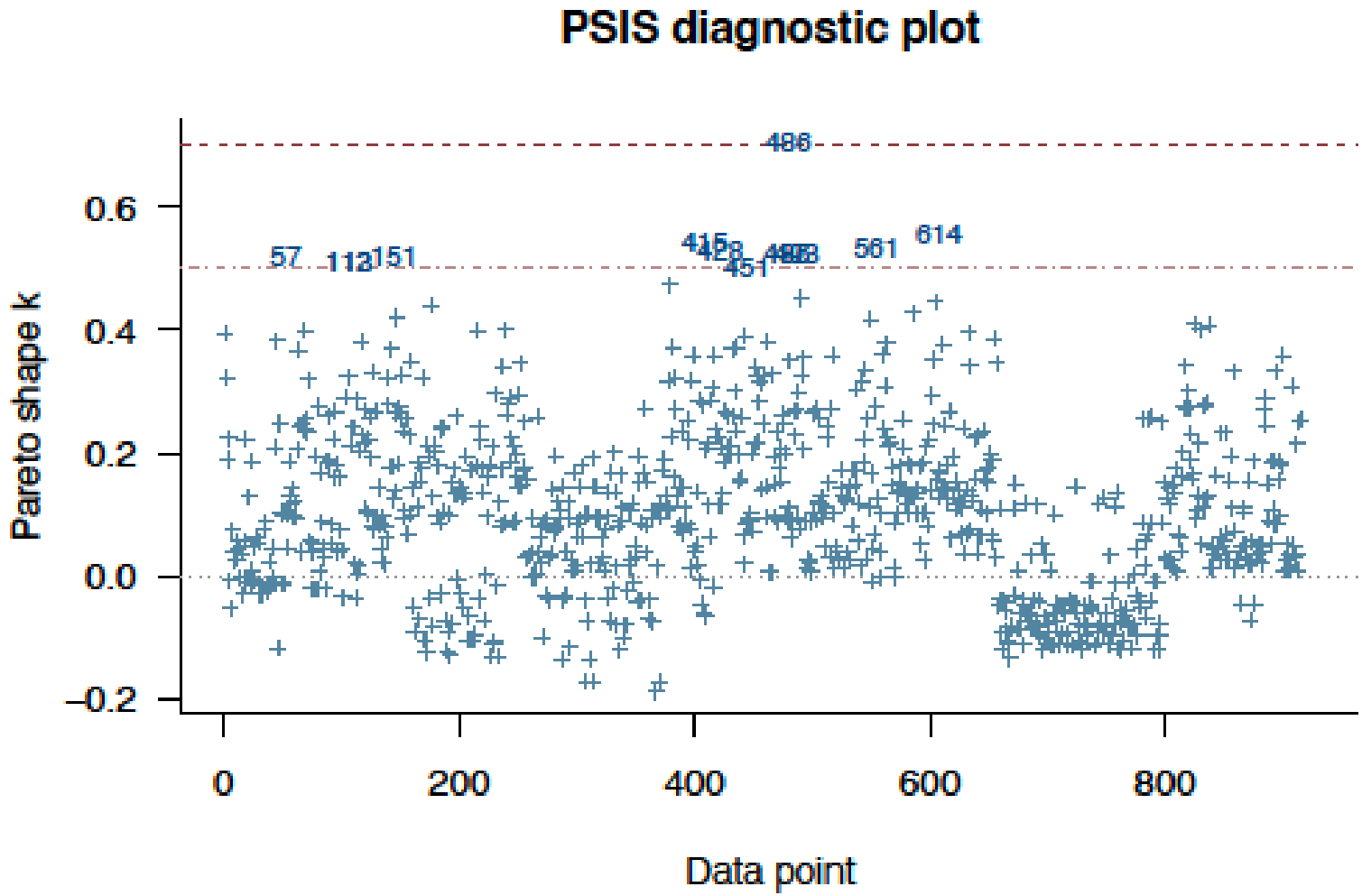
- When you call `loo`, you get a message regarding Pareto k estimates, i.e.

```
> loo_fit
All Pareto k estimates are good (k < 0.5)
See help('pareto-k-diagnostic') for details.
```

- Output for PSIS-LOO calculations includes a $\hat{k}$ for each observation, which is the estimated shape parameter of the pareto distribution fitted to the right tail of importance weights
$r_i^{(s)} = 1/p(y_i|\boldsymbol{\theta}^{(s)}) \propto p(\boldsymbol{\theta}^{(s)}|\boldsymbol{y}_{-i})/p(\boldsymbol{\theta}^{(s)}|\boldsymbol{y})$
$\Rightarrow$ larger $\hat{k}$ refers to more outlying weights, i.e. $p(\boldsymbol{\theta}^{(s)}|\boldsymbol{y}_{-i})$ being very different from $p(\boldsymbol{\theta}^{(s)}|\boldsymbol{y})$.
- If $\hat{k}$ is larger than 0.5 (or 0.7), then
  1. it is flagged as possibly influential
  2. PSIS-LOO approximation of $p(y_i|\boldsymbol{y}_{-i})$ may not be very good, it is recommended to get $p(y_i|\boldsymbol{y}_{-i})$ directly by fitting the model without observation i.

$\hat{k}$: Identifying influential points

- Example: $\hat{k}$ diagnostics for a radon model; point 486 is highlighted as possibly being influential



PSIS-LOO log-predictive densities

- For each observation i, we can obtain through the PSIS approach an estimate (and SE) for the expected log pointwise predictive density (ELPD)

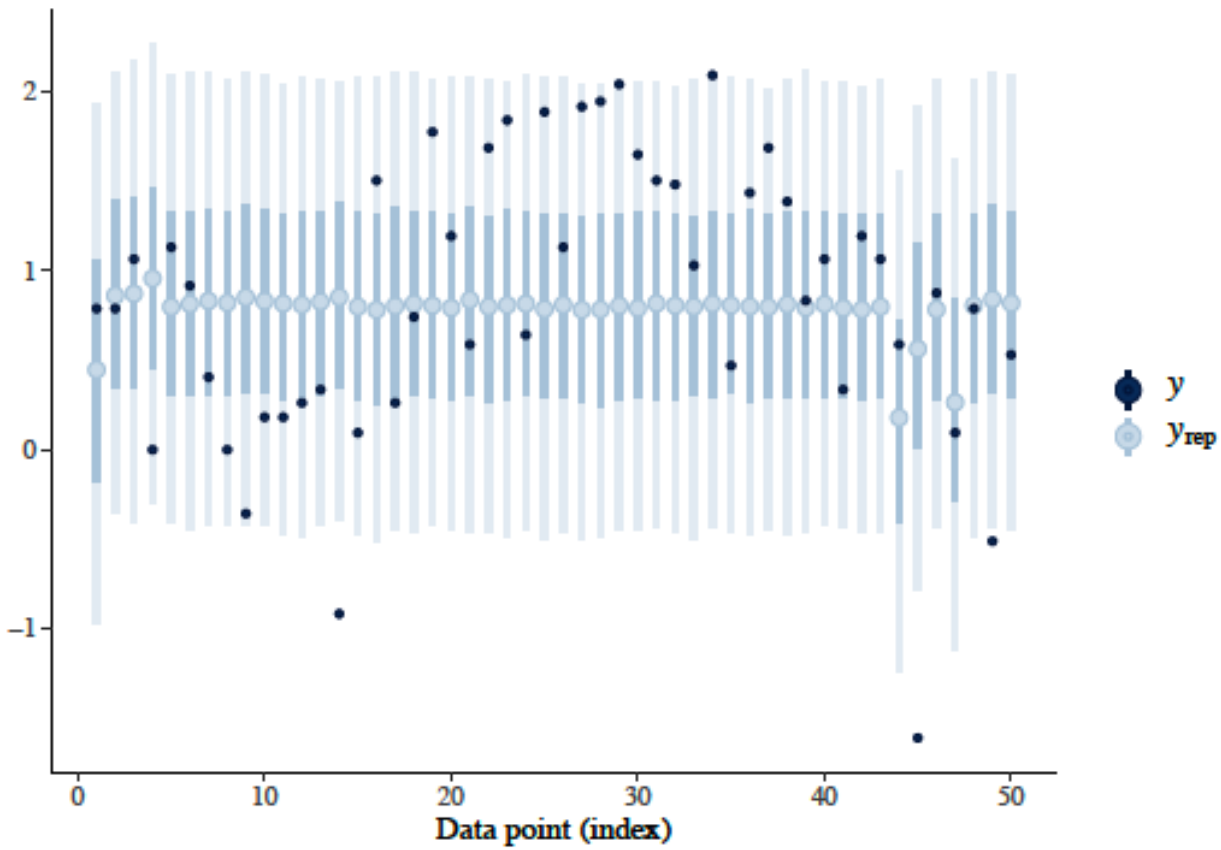
$$ELPD_i = \log(p(y_i|\boldsymbol{y}_{-i}))$$

- These pointwise (observation-specific) values can be compared across models to check which models are better/worse at predicting specific observations.
  - If ($ELPD_i$ model 1) $>$ ($ELPD_i$ model 2) for some observation i, which model do we prefer for predicting that observation?
    Model 1, bigger is better.
- We will use these in HW5 to compare models for the switching problem.

PSIS-LOO predictive checks

- Using the weights, we can generate other outcomes of interest that are relevant for model checking, similar to outcomes discussed for in-sample model checking.
- See https://mc-stan.org/bayesplot/reference/PPC-loo.html

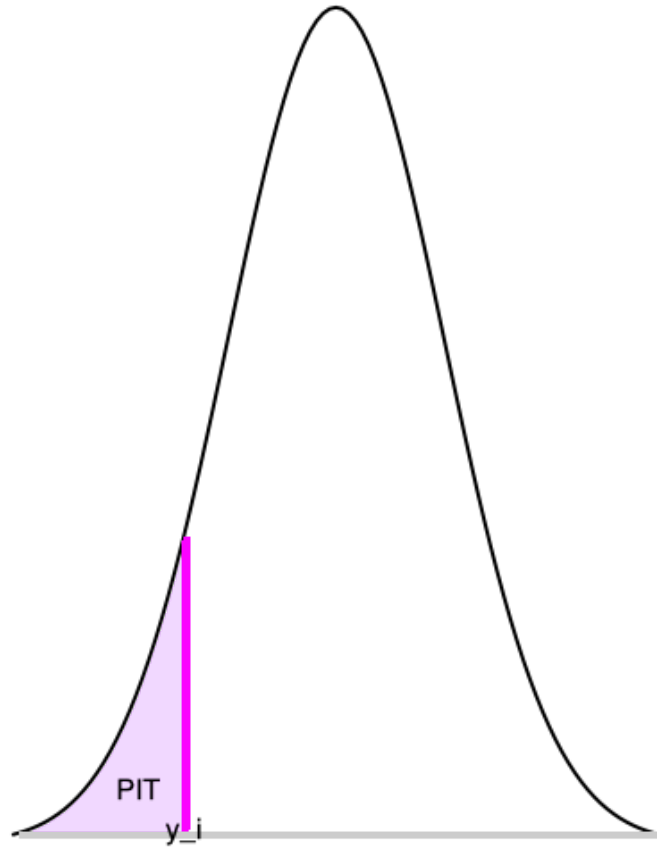
Example: prediction intervals for the first 50 observations



Probability integral transform (PIT) outcomes

- Introduction to PITs:
  - Let random variable $X \sim f_X(x)$ with $F_X(x) = P(X \leq x)$,
  - Define the probability integral transform (PIT) of X as random variable $Y = F_x(X)$,
  - Then $Y \sim U(0, 1)$
- In our notation/application for LOO:
  - If $y_i \sim p(\tilde{y}_i|\boldsymbol{y}_{-i})$,
  - then $\text{PIT}_i = P(\tilde{y}_i \leq y_i|\boldsymbol{y}_{-i}) \sim U(0, 1)$.

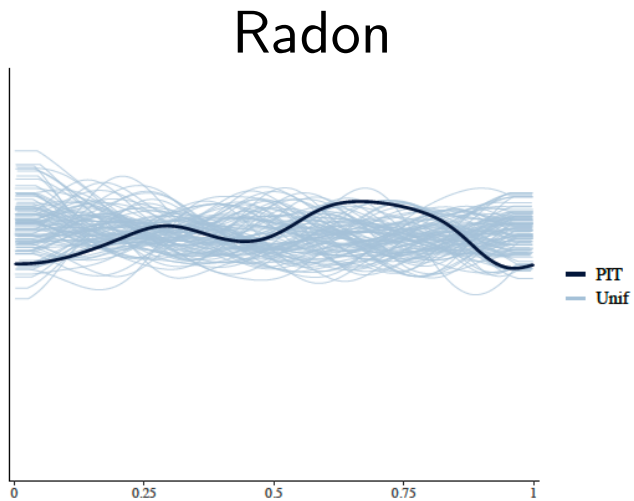
  The PIT value quantifies where y_i falls in its predictive distribution.



Note that in practice, when using LOO, LOO-PIT is expected to be close to uniform, as opposed to exactly uniform(which only holds true when the sample size goes to infinity).

PSIS-LOO PIT densities

- Recap: If $y_i \sim p(\tilde{y}_i|\boldsymbol{y}_{-i})$, then LOO-PIT$_i = P(\tilde{y}_i \leq y_i|\boldsymbol{y}_{-i}) \sim U(0,1)$, and the PIT value shows where y_i falls in its LOO predictive distribution
- Plot: comparison of PITs to U(0,1), thin lines illustrate 100 densities based on n draws from U(0,1)
- PIT frown shape (as compared to uniform) indicates that the predictive distributions are too broad (uncertain) compared to the data



Summary approximate LOO cross-validation

- PSIS-LOO approach is a computationally convenient approach to do approximate cross-validation, and provides diagnostic measures ($\hat{k}$) to check if approximation is ok
- The PSIS-LOO approach can be used for
  - model checking for one model (influential points, goodness of fit)
  - comparing two models, or a small set of models
- Additional notes:
  - The default loo-function can only be used if we can write $p(\boldsymbol{y}|\boldsymbol{\theta}) = \prod_{i=1}^{n} p(y_i|\boldsymbol{\theta})$ but there is a recent extension for non-factorized models (https://link.springer.com/article/10.1007/s00180-020-01045-4)
  - When working with time series data, see https://www.tandfonline.com/doi/full/10.1080/00949655.2020.1783262 for leave-future-out cross-validation (LFO-CV).

Model checking: summary of part 1 and 2

- We discussed
  - General diagnostic plots, e.g. of residuals
  - Posterior predictive checks based on simulating data from the fitted model
  - Measures of predictive accuracy based on out-of-sample validation, including approximate leave-one-out validation
- What "goodness of fit" outcomes you need to check depends on your outcome of interest.
- Next: Bayesian workflow, model comparison