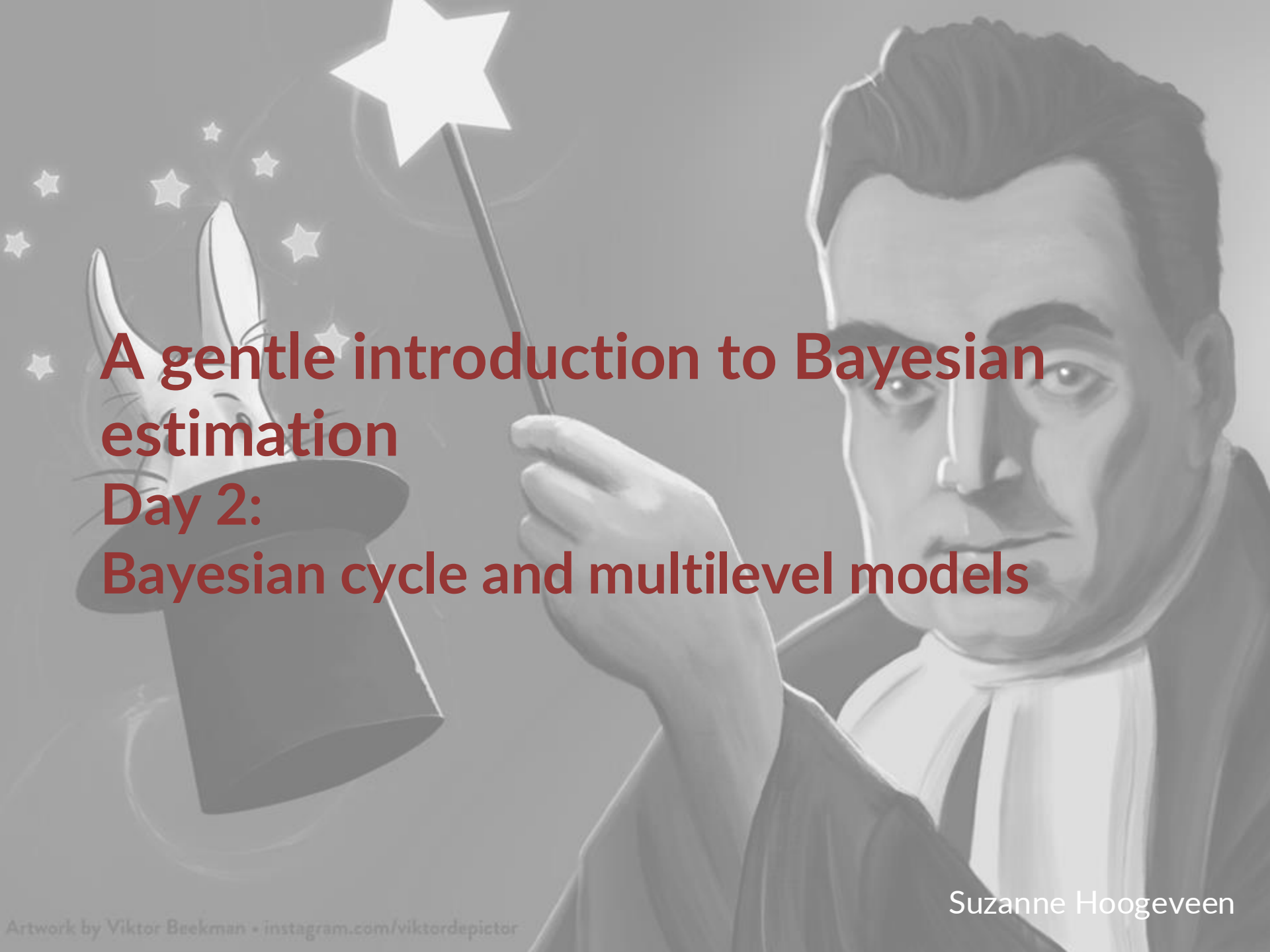


A gentle introduction to Bayesian estimation
Day 2:
Bayesian cycle and multilevel models

Suzanne Hoogeveen

Bayes: when to worry

- Aka: how to understand what you're actually doing

Based on slides by Rens van de Schoot.

Dear dr. X,
We would kindly invite you to review this paper about [*interesting topic Y*]

Dear dr. X,
We would kindly invite you to review this paper about [*interesting topic Y*]

Because of the small sample size (n=20) we used Bayesian estimation. Hox et al. (2012) showed that a multilevel model with only 20 clusters could be estimated with Bayesian statistics whereas maximum likelihood estimation could not.

Hox, J., van de Schoot. R., & Matthijsse, S. (2012). How few countries will do? Comparative survey analysis from a Bayesian perspective. Survey Research Methods, 6, 87-93.

Dear dr. X,
We would kindly invite you to review this paper about [*interesting topic Y*]

Because of the small sample size (n=20) we used Bayesian estimation. Hox et al. (2012) showed that a multilevel model with only 20 clusters could be estimated with Bayesian statistics whereas maximum likelihood estimation could not.

Since we are no experts in Bayesian estimation we relied on the default settings.
The results are completely in line with our hypothesis: there is a significant difference between the two groups. All is fine, please accept our paper for publication.

Hox, J., van de Schoot. R., & Matthijsse, S. (2012). How few countries will do? Comparative survey analysis from a Bayesian perspective. Survey Research Methods, 6, 87-93.

Making decisions implementing Bayes

- Naively applying Bayesian methods can be dangerous for three main reasons:

Making decisions implementing Bayes

- Naively applying Bayesian methods can be dangerous for three main reasons:
  - First, the exact influence of the priors is often not well understood and priors *might* have a huge impact on the study results;

Making decisions implementing Bayes

- Naively applying Bayesian methods can be dangerous for three main reasons:
  - First, the exact influence of the priors is often not well understood and priors *might* have a huge impact on the study results;
  - Second, akin to many elements of frequentist statistics, some Bayesian features can be easily misinterpreted;

Making decisions implementing Bayes

- Naively applying Bayesian methods can be dangerous for three main reasons:
  - First, the exact influence of the priors is often not well understood and priors *might* have a huge impact on the study results;
  - Second, akin to many elements of frequentist statistics, some Bayesian features can be easily misinterpreted;
  - Third, reporting on Bayesian statistics follows its own rules since there are elements included in the Bayesian framework that are fundamentally different from frequentist settings.

WAMBS checklist

- When to **Worry** and how to **Avoid** the **Misuse** of **Bayesian Statistics**
  - 10 main points that should be thoroughly checked when applying Bayesian statistics

Depaoli & van de Schoot (2017). https://doi.org/10.1037/met0000065
van de Schoot et al. (2020)
van de Schoot, et al. (2021). https://doi.org/10.1038/s43586-020-00001-2

Disclaimer

Bayesian inference is sometimes seen as:

1. A panacea (no more ‘the model failed to converge!’)
2. A minefield (subjective priors! divergent transitions! days of estimation time!)

Disclaimer

Bayesian inference is sometimes seen as:

1. A panacea (no more ‘the model failed to converge!’)
2. A minefield (subjective priors! divergent transitions! days of estimation time!)

Unfortunately, it won’t solve all your statistical problems …

Disclaimer

Bayesian inference is sometimes seen as:

1. A panacea (no more 'the model failed to converge!')
2. A minefield (subjective priors! divergent transitions! days of estimation time!)

Unfortunately, it won't solve all your statistical problems …

… but it can be a powerful and flexible tool that forces you to critically evaluate your data, your modeling assumptions and your results

WAMBS checklist

- 10 main points that should be checked when applying Bayesian analysis:

a) Issues to check before running the analysis (*prior knowledge*)

WAMBS checklist

- 10 main points that should be checked when applying Bayesian analysis:

a) Issues to check before running the analysis (*prior knowledge*)
b) Issues to check after running the analysis, but before interpreting the results (*sampling diagnostics*)

WAMBS checklist

- 10 main points that should be checked when applying Bayesian analysis:

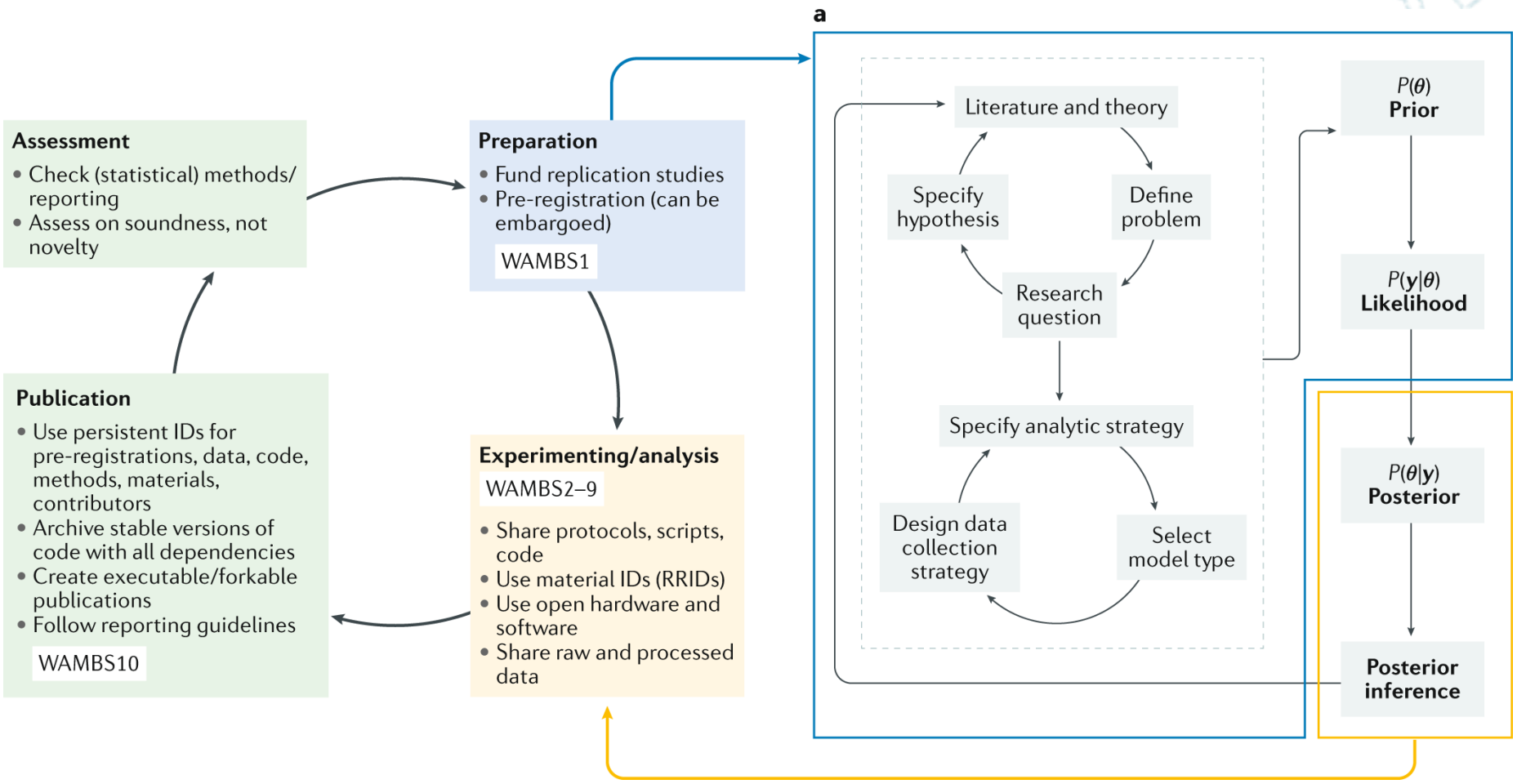
a) Issues to check before running the analysis (*prior knowledge*)
b) Issues to check after running the analysis, but before interpreting the results (*sampling diagnostics*)
c) Assessing the robustness of the results (*sensitivity analyses*)

WAMBS checklist

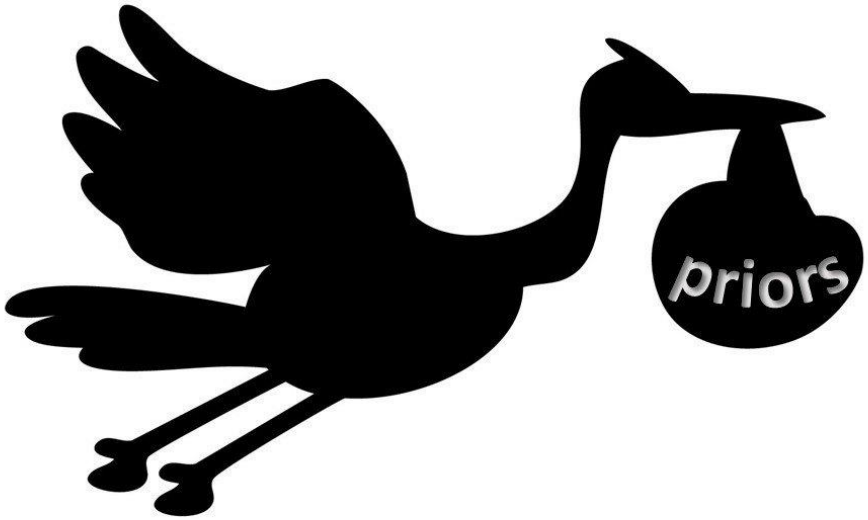
- 10 main points that should be checked when applying Bayesian analysis:

a) Issues to check before running the analysis (*prior knowledge*)
b) Issues to check after running the analysis, but before interpreting the results (*sampling diagnostics*)
c) Assessing the robustness of the results (*sensitivity analyses*)
d) Reporting the results (*transparency and reproducibility*)

(Bayesian) research cycle



Stage 1: before the analysis



Where do your priors come from?

Priors

When specifying priors, it is important to recognize that prior distributions fall into three main classes related to the amount of (un)certainty they contribute to the model about a given parameter:

1. non-informative priors (diffuse, flat)
2. weakly-informative priors
3. informative priors

Priors

When specifying priors, it is important to recognize that prior distributions fall into three main classes related to the amount of (un)certainty they contribute to the model about a given parameter:

1. non-informative priors (diffuse, flat)
2. weakly-informative priors
3. informative priors

All have pros and cons, e.g.:

- Diffuse priors work fine for estimation, but not for testing with Bayes factors
- Informative priors convey domain knowledge but can have strong impact on the posterior
- Diffuse and weakly-informative priors can lead to implausible prior predictions (especially on transformed scales).

1. Understanding priors

- The prior can only be fully understood in the context of the likelihood
- *That is to say, it's all relative*

Gelman, A., Simpson, D., & Betancourt, M. (2017). https://doi.org/10.3390/e19100555

1. Understanding priors

- The prior can only be fully understood in the context of the likelihood
- *That is to say, it's all relative*

- A normal(0,1) prior can be highly informative in a linear regression model that models response time data in seconds
- The same prior can be weakly informative as the intercept in a binomial (logit) model

Gelman, A., Simpson, D., & Betancourt, M. (2017). https://doi.org/10.3390/e19100555

1. Understanding priors

- The prior can only be fully understood in the context of the likelihood
- *That is to say, it's all relative*

- A normal(0,1) prior can be highly informative in a linear regression model that models response time data in seconds
- The same prior can be weakly informative as the intercept in a binomial (logit) model

- Also be aware of the combination of priors on the fixed and random parts of the model: a diffuse prior on the fixed part + a diffuse prior on the between-group variability can lead to strange prior predictions

Gelman, A., Simpson, D., & Betancourt, M. (2017). https://doi.org/10.3390/e19100555

Guidelines

Determine what strategy is most appropriate in the given project:

- Could prior information be found in the literature (e.g., empirical studies, reviews, meta-analyses)?
- Are there experts on the subject that can be consulted?
- What general knowledge is available about the model parameters?

Gather this information strategically and keep a log of the decisions (for transparency and your future self :))

Guidelines

- Visualize priors (prior distribution + prior predictions)
- Provide information on:
  1. justification for specific prior setting (also when using defaults!)
  2. exact specification of all priors
- Conduct sensitivity analyses to assess the impact on the posterior estimates (see point 7 and 8)
- If differences arise (which is not problematic in itself), explain and interpret them

Running example

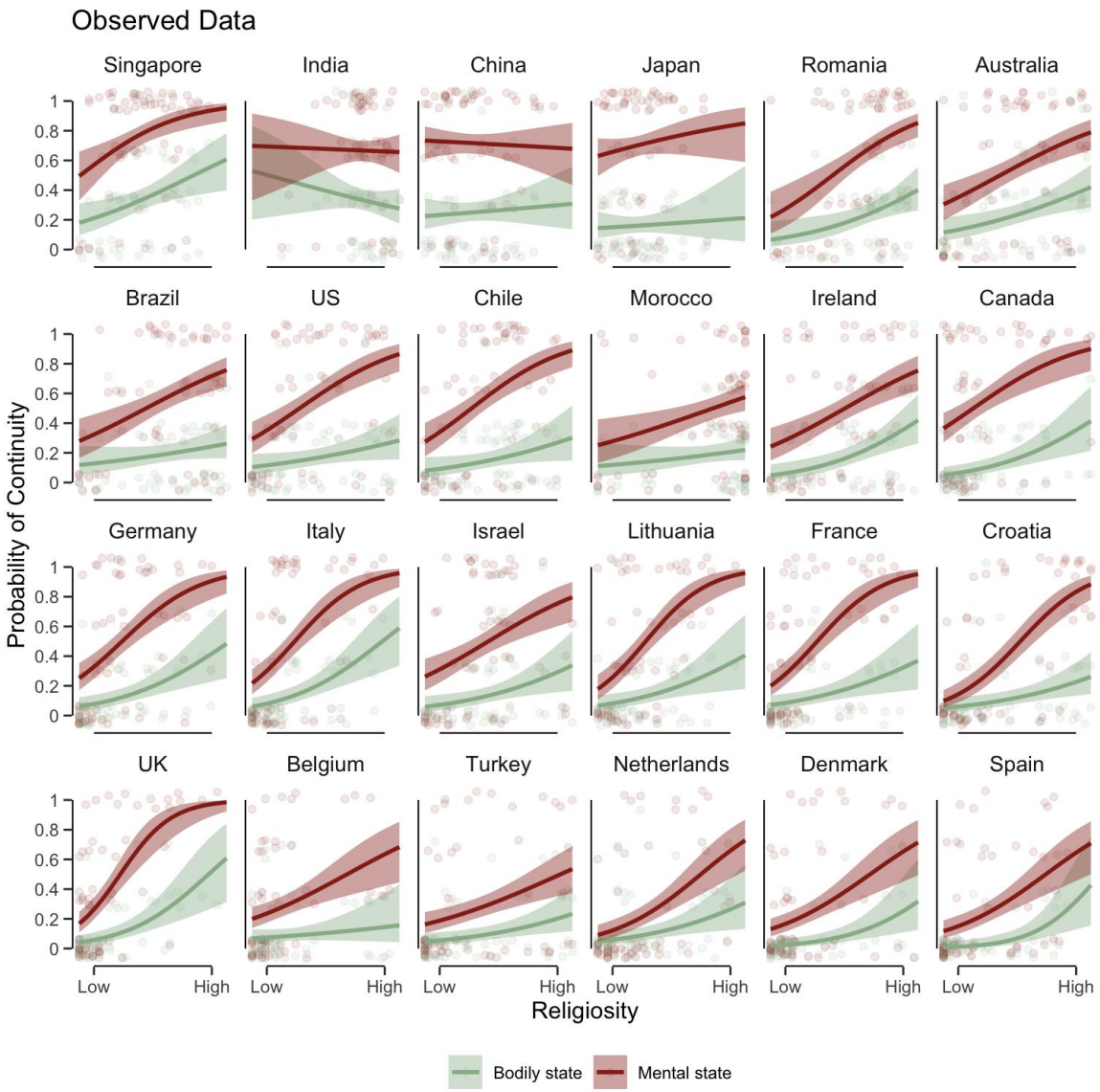
- Cross-cultural study (10000 subjects from 24 countries) on afterlife beliefs and mind-body dualism:
- Hypothesis: high-level mental states (e.g., love) are more likely to be judged as continuing after biological death than bodily states (e.g., hunger)
- Subjects read a story about a grandmother who dies and are asked to indicate to what extent she is still has certain states
- Each subject provided 6 'yes' or 'no' responses about continuation across 2 conditions (*mental*: love, knowledge, desire; *bodily*: hunger, working brains, hearing)

Running example

- Cross-cultural study (10000 subjects from 24 countries) on afterlife beliefs and mind-body dualism:
- Hypothesis: high-level mental states (e.g., love) are more likely to be judged as continuing after biological death than bodily states (e.g., hunger)
- Subjects read a story about a grandmother who dies and are asked to indicate to what extent she is still has certain states
- Each subject provided 6 'yes' or 'no' responses about continuation across 2 conditions (*mental*: love, knowledge, desire; *bodily*: hunger, working brains, hearing)
- Here: subset of 60 subjects per country (*N*=1440)
- We used a multilevel aggregated binomial model:

```
brm(formula = response | trials(3) ~ 1 + state_cond + relig + (1 +
state_cond + relig | country), family = binomial, …)
```

Running example



1. Understanding priors

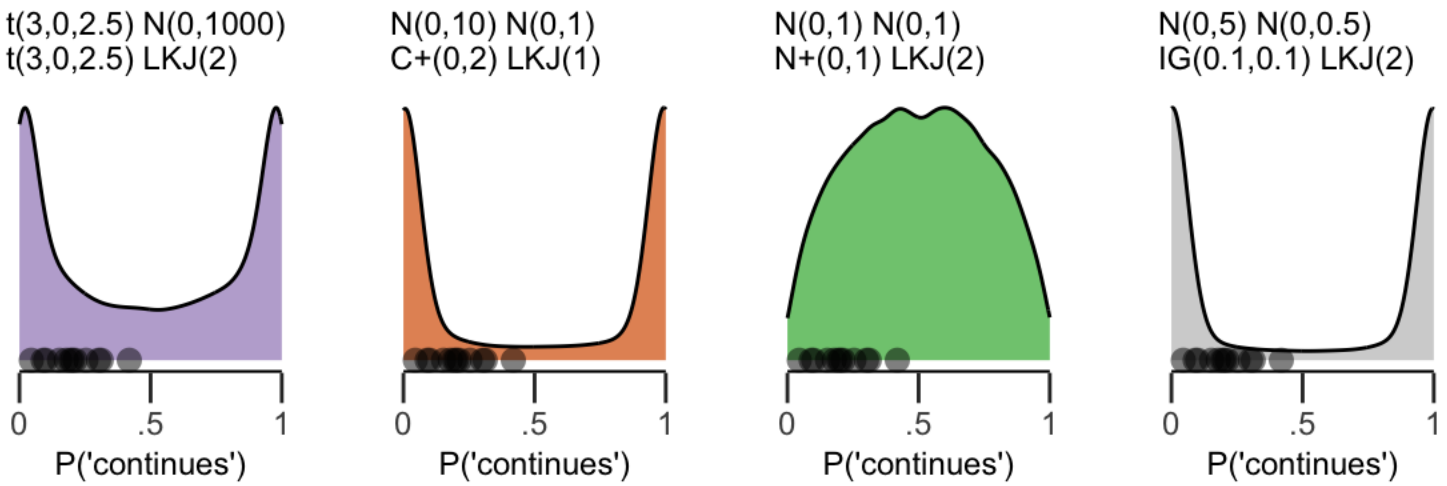
- Priors are set on the logit-transfomed scale
- We have priors on:
  1. intercept (i.e., overall probability of saying ‘continues’)
  2. experimental effect (i.e., difference in probability of saying ‘continues’ between mental and physical states)
  3. between-country variation (in intercepts and effects)
  4. correlation structure of random effects

1. Understanding priors

- What do we know?
  - Previous studies: mean state effect (difference mental states continuation vs bodily states continuation): ~16%
  - Previous studies: standard deviation across sites/countries: ~15%

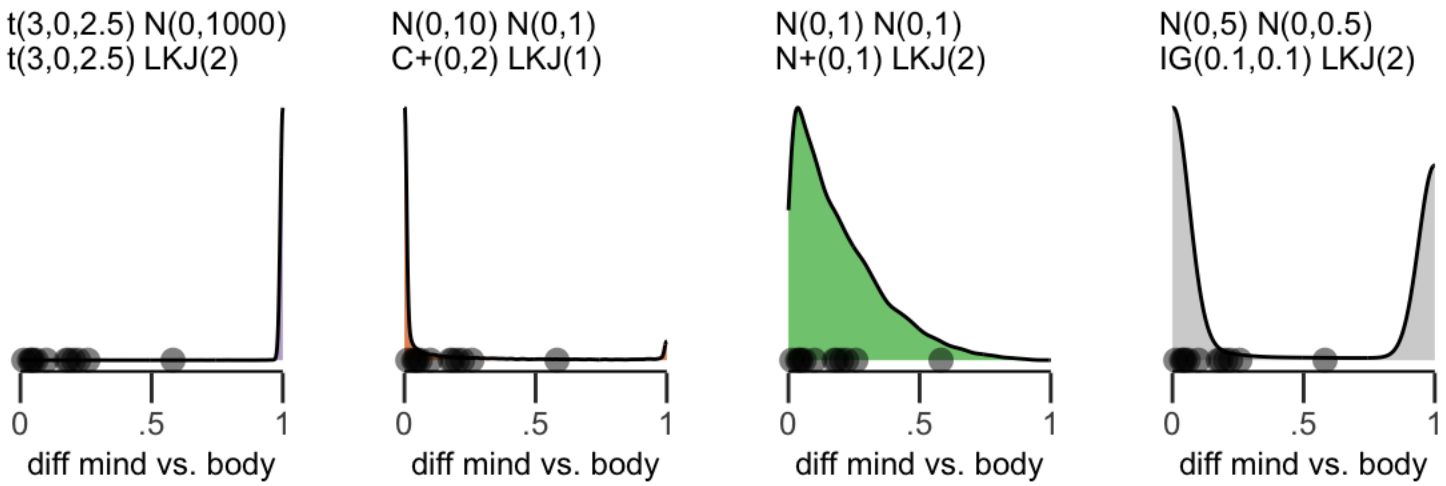
- What do we want?
  - We inspect prior predictions for different prior settings. We aim for distributions that are weakly informative and make sensible predictions (on the response scale)

Prior predictions



- What do we conclude?
  - Predictions from both the *brms* default settings (purple) and our preregistered prior settings (orange) are unrealistic; both predict that all responses with be either complete cessation or continuity.
  - Selection in green: allowing all rates, with slightly less mass at the extremes.

Prior predictions



- What do we conclude?
  - The *brms* default priors on the effects are much too wide, predicting an unlikely difference of 100% between conditions. The preregistered priors predict a modest effect, but due to the wide prior on the variation between countries, this results in a very strong prediction of no effect.
  - Selection: modest effect with most mass between 0 and 50% difference

Stage 2: after analysis, before interpretation

- Are the analysis outputs sufficiently reliable to be interpreted?

    2. Assessing convergence
    3. Assessing convergence with more samples
    4. Assessing posterior distributions
    5. Assessing effective sample size
    6. Assessing posterior predictive checks

2. Assessing convergence

Determining whether the MCMC chains have converged can be based on:

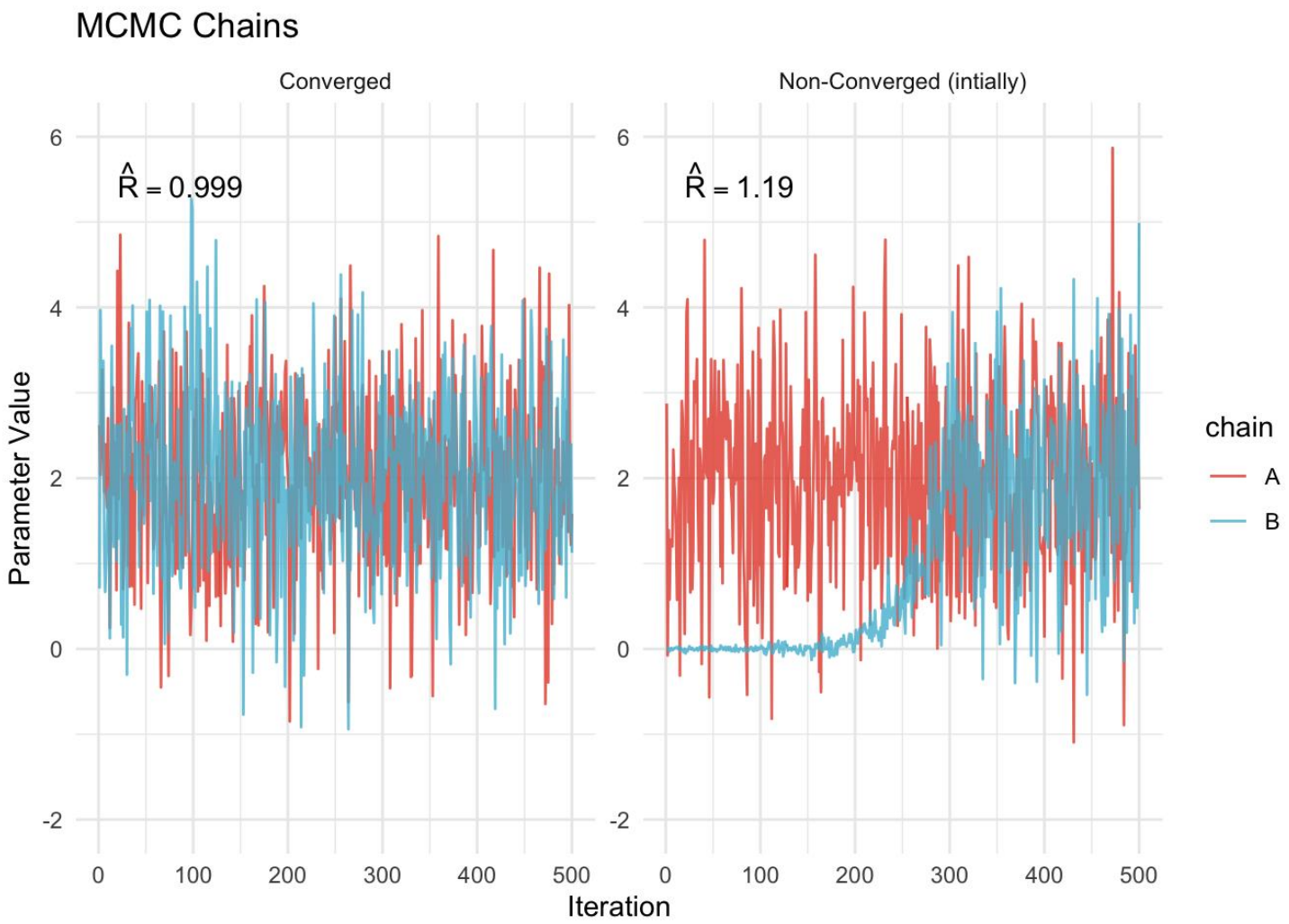
a) Visual inspection of 'traceplots'

b) $\widehat{\mathrm{R}}$ ('rhat') diagnostic (aka potential scale reduction factor, PSRF or Gelman and Rubin diagnostic)

c) Geweke diagnostic

Gelman, A., & Rubin, D. B. (1992). https://doi.org/10.1214/ss/1177011136
Geweke, J. (1992).

2. Assessing convergence



2. Assessing convergence

- $\hat{R}$ is based the variance between chains relative to the variance within chains:

Let:

- m: number of MCMC chains
- n: number of iterations per chain
- θ_{ij}: the j-th draw from the i-th chain
- $\bar{\theta}_i$: mean of chain i
- $\bar{\theta}$: overall mean across all chains

Then the components of the R-hat statistic are:

1. Between-chain variance

$$B = \frac{n}{m-1} \sum_{i=1}^{m} \left(\bar{\theta}_i - \bar{\theta} \right)^2$$

2. Within-chain variance

$$W = \frac{1}{m} \sum_{i=1}^{m} \left(\frac{1}{n-1} \sum_{j=1}^{n} \left(\theta_{ij} - \bar{\theta}_i \right)^2 \right)$$

3. Estimated marginal posterior variance

$$\hat{V} = \frac{n-1}{n} W + \frac{1}{n} B$$

4. Potential scale reduction factor (PSRF, $\hat{R}$)

$$\hat{R} = \sqrt{\frac{\hat{V}}{W}}$$

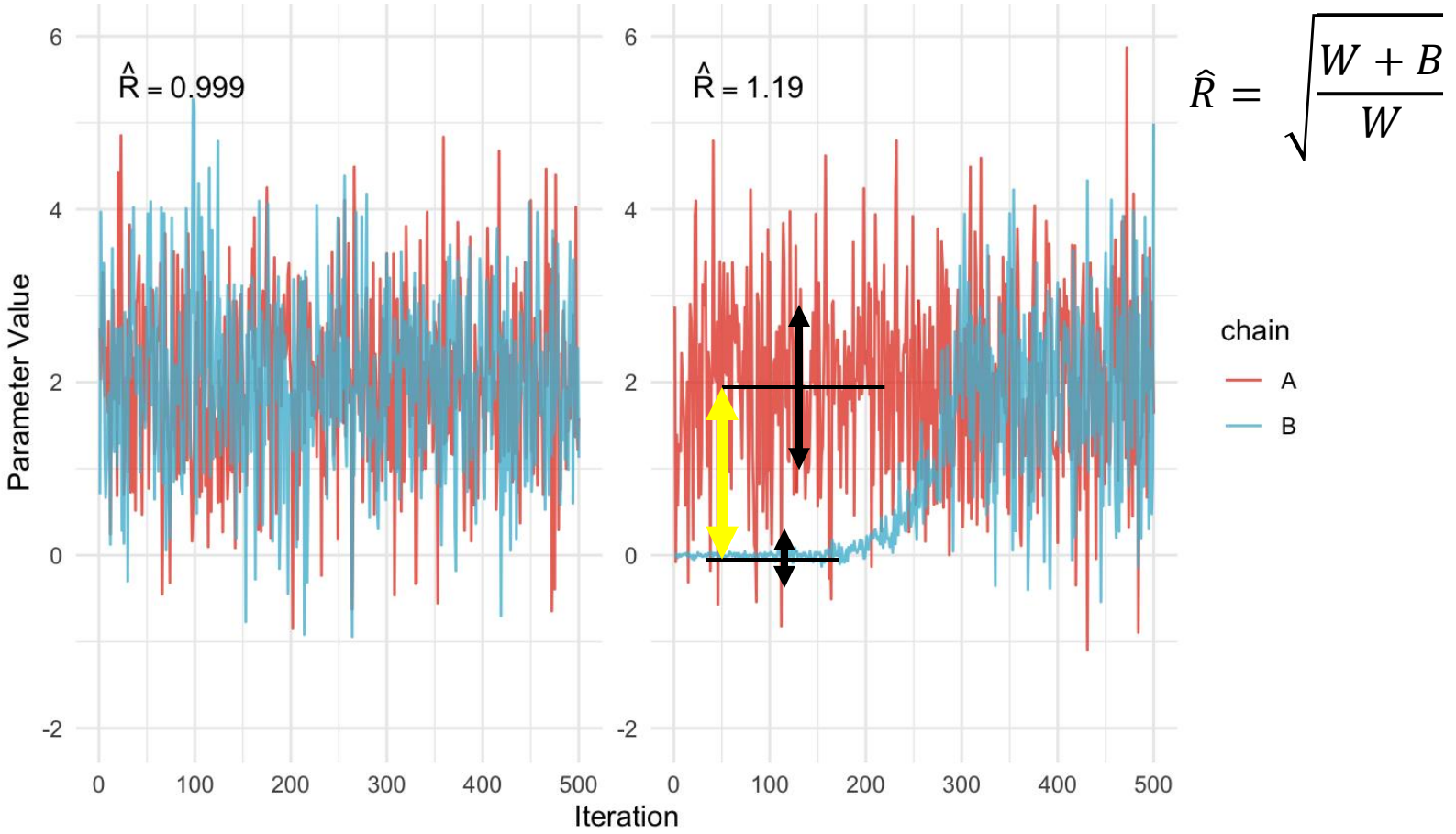
$\hat{R}$ = 1: ideal

$\hat{R}$ > 1.01: 'worth inspecting'

$\hat{R}$ > 1.1: 'problematic'

2. Assessing convergence

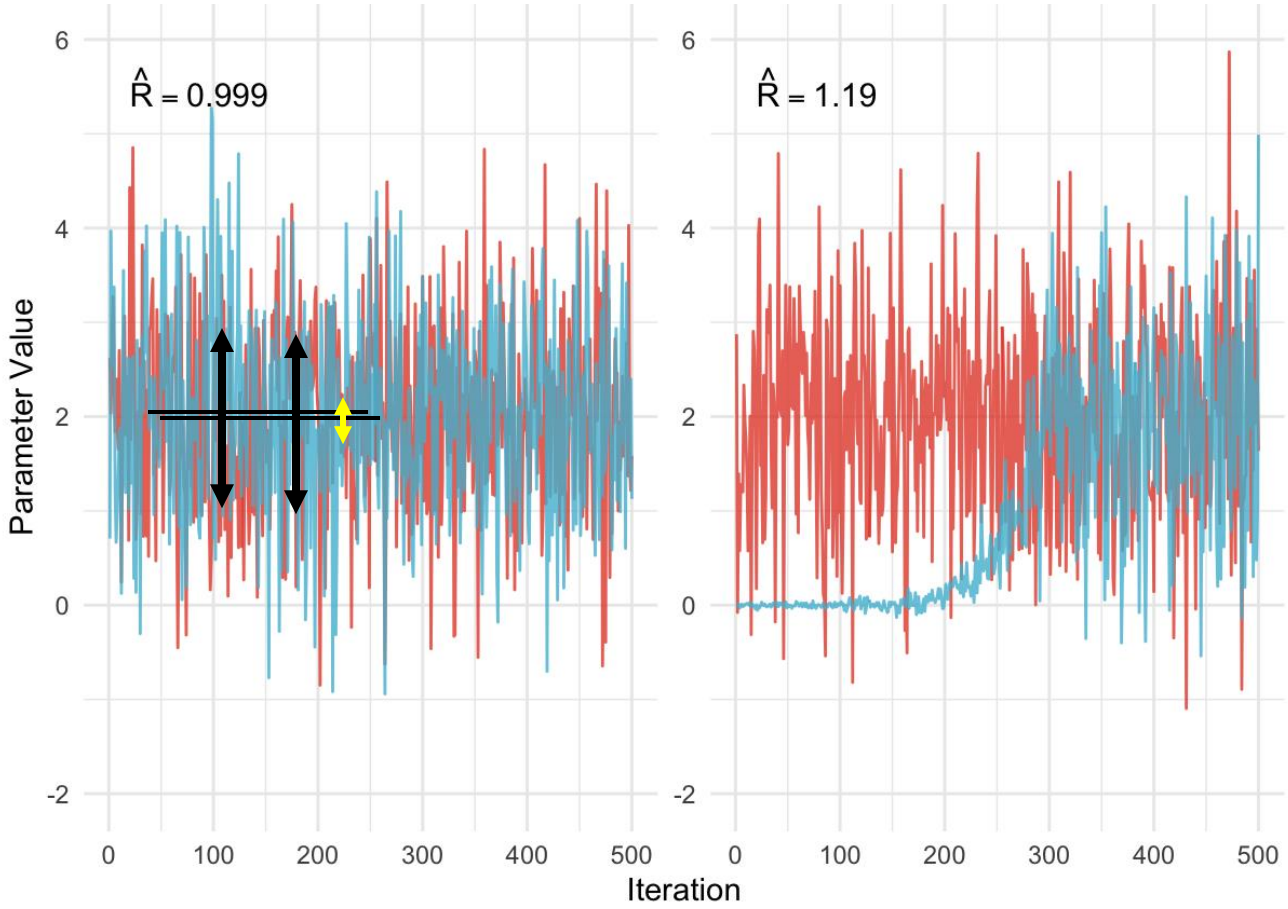
- $\hat{R}$ is based the variance between chains relative to the variance within chains:



$$\hat{R} = \sqrt{\frac{W + B}{W}}$$

2. Assessing convergence

- $\hat{R}$ is based the variance between chains relative to the variance within chains:



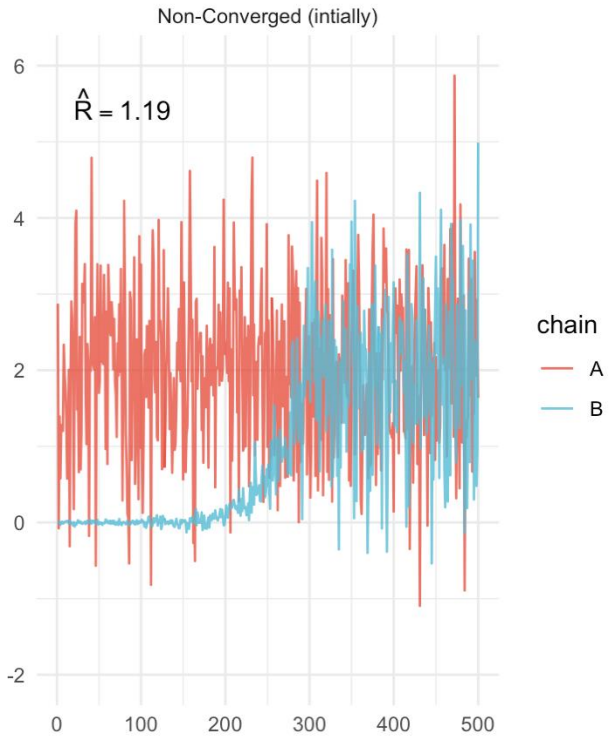
$$\hat{R} = \sqrt{\frac{W + B}{W}}$$

2. Assessing convergence

- The Geweke diagnostic gives a test statistic (z-score) for a difference in means of the first (10%) and last (50%) portion of a chain.
- If the chains have converged, the Geweke diagnostic should not be larger than +-1.96 (i.e., 95% confidence interval).

2. Assessing convergence

- The Geweke diagnostic gives a test statistic (z-score) for a difference in means of the first (10%) and last (50%) portion of a chain.
- If the chains have converged, the Geweke diagnostic should not be larger than +-1.96 (i.e., 95% confidence interval).

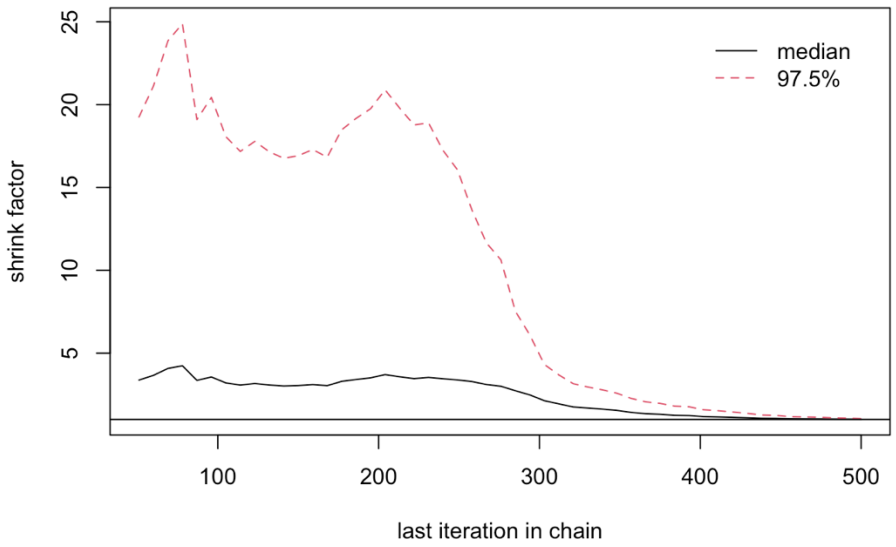


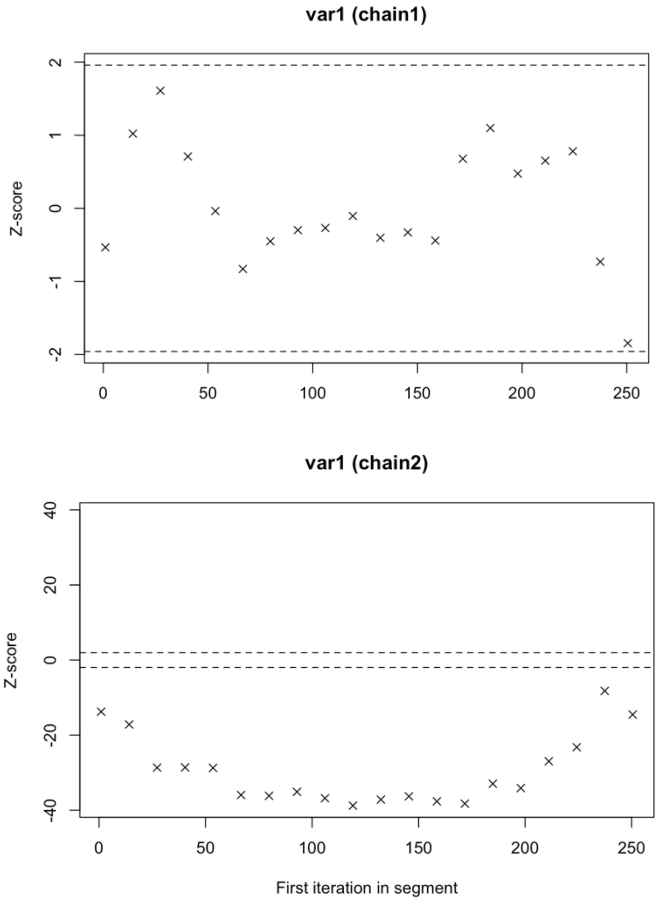

```
geweke.diag(mcmc_list)
```

```
[[1]]

Fraction in 1st window = 0.1
Fraction in 2nd window = 0.5

   var1
-0.5344


[[2]]

Fraction in 1st window = 0.1
Fraction in 2nd window = 0.5

  var1
-13.76
```

2. Assessing convergence

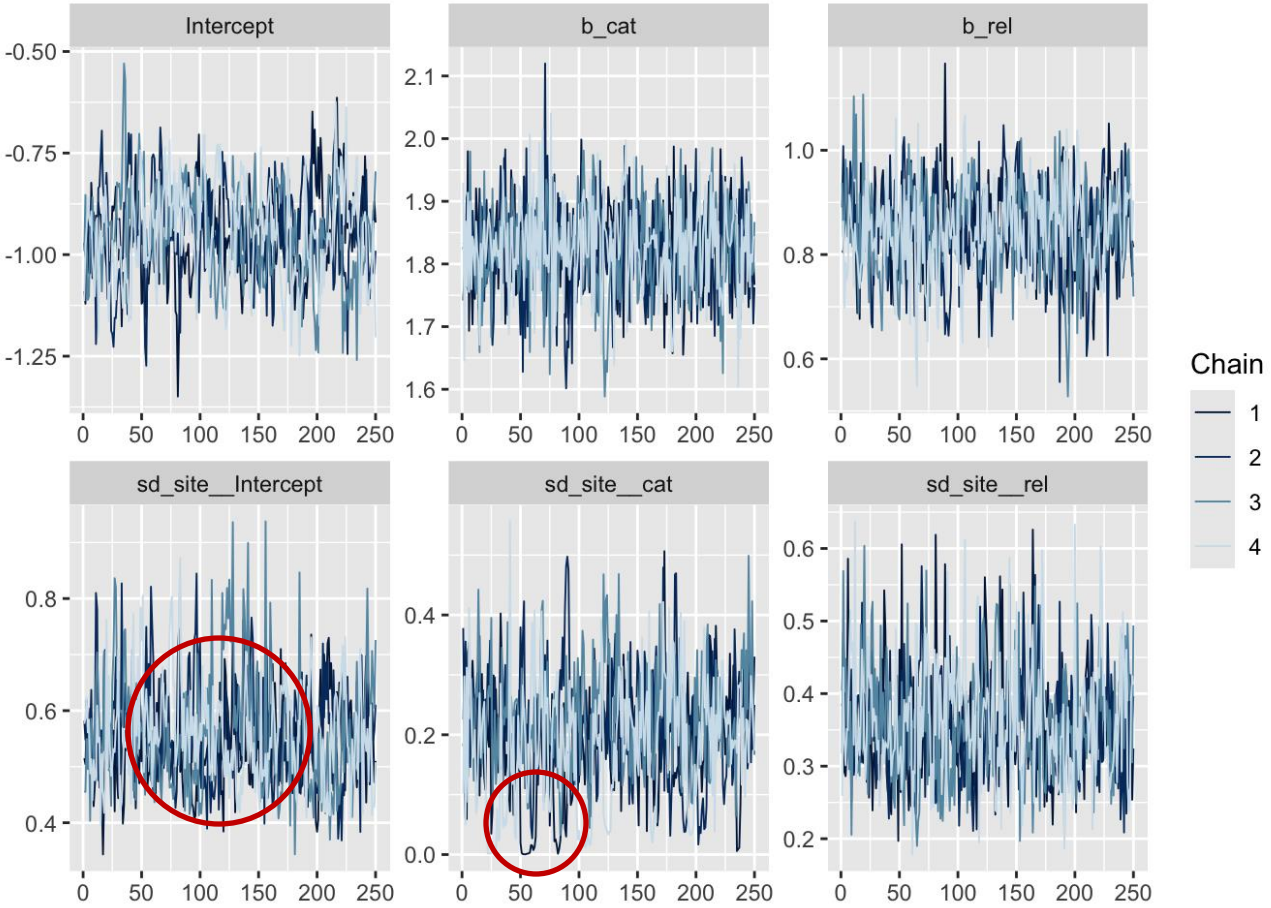
- $\hat{\mathbb{R}}$ is given per parameter, across chains (between-chain convergence)
- Geweke diagnostic is given per parameter, per chain (within-chain convergence/stability)





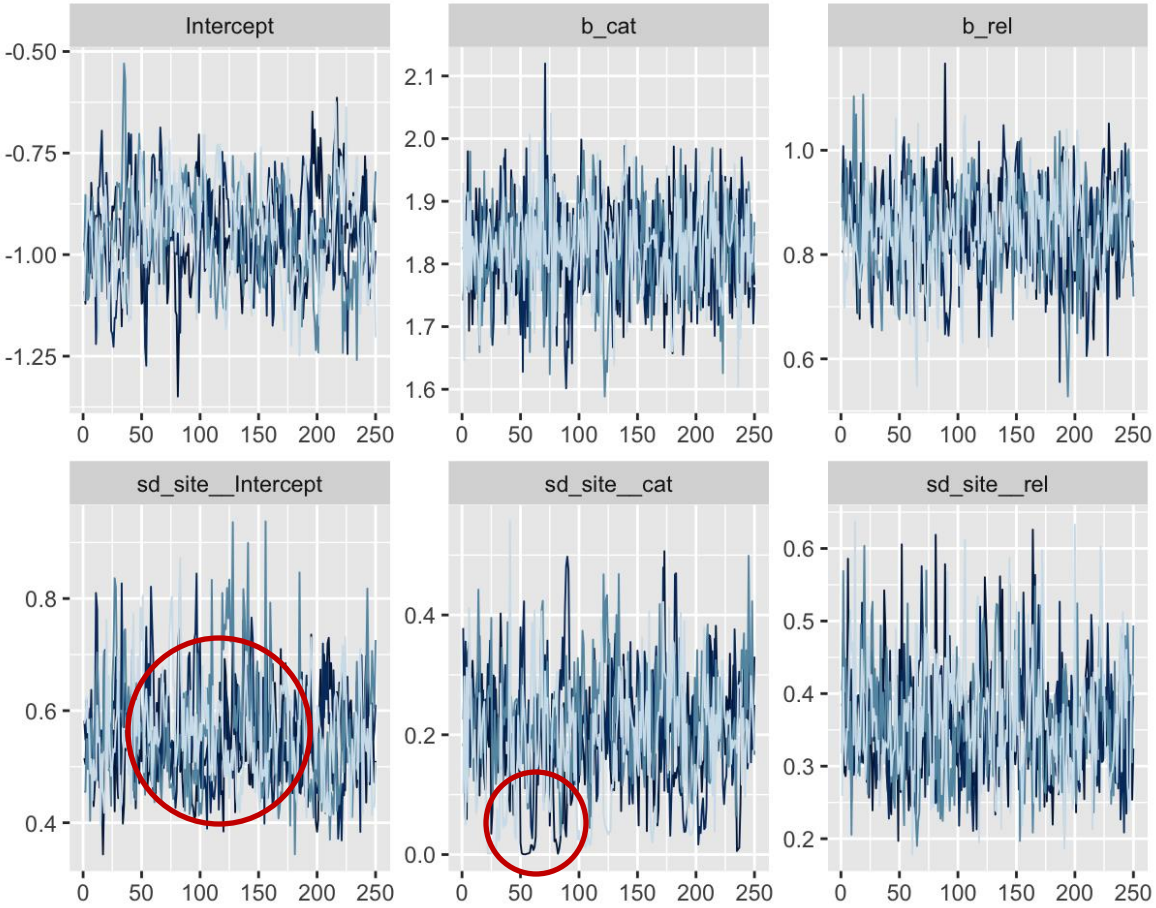
2. Assessing convergence

$N_{iter} = 500, N_{warmup} = 250$



2. Assessing convergence

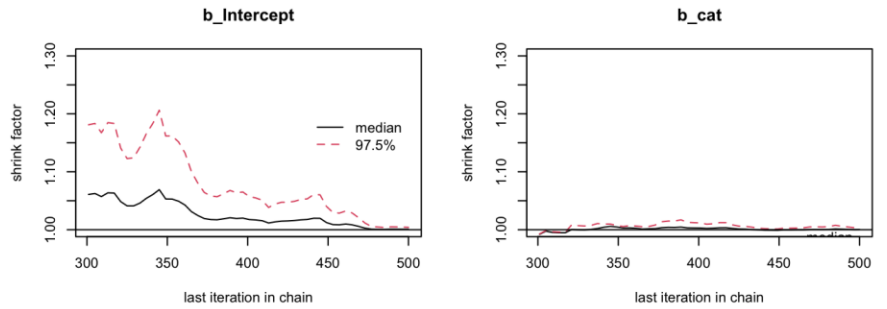
$N_{iter} = 500, N_{warmup} = 250$

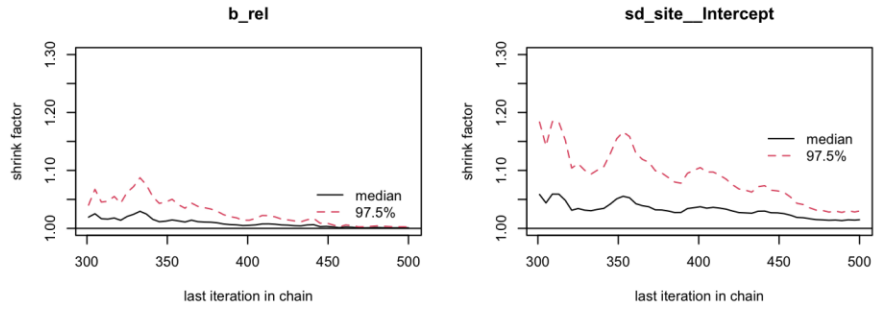


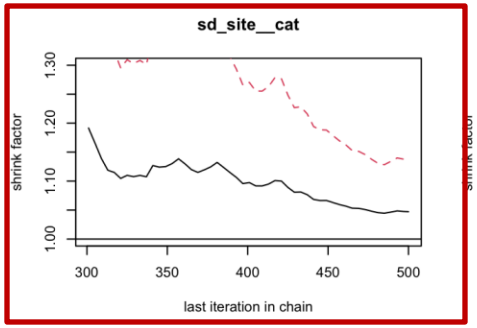

```
Potential scale reduction factors:

                    Point est. Upper C.I.
b_Intercept               1.00       1.00
b_cat                     1.00       1.00
b_rel                     1.00       1.00
sd_site__Intercept        1.01       1.03
sd_site__cat              1.05       1.14
sd_site__rel              1.00       1.01

Multivariate psrf

1.04
```

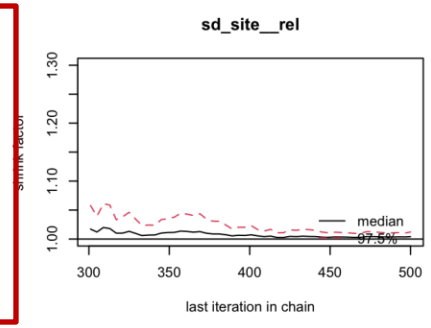
2. Assessing convergence

$\hat{R}$

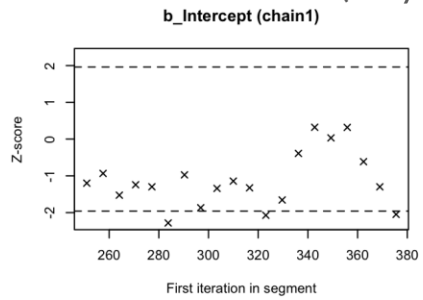


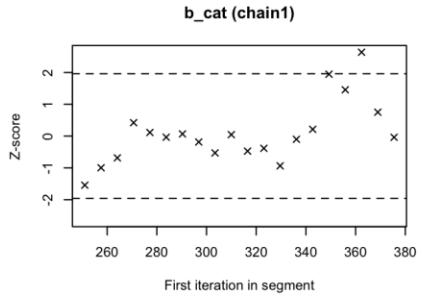


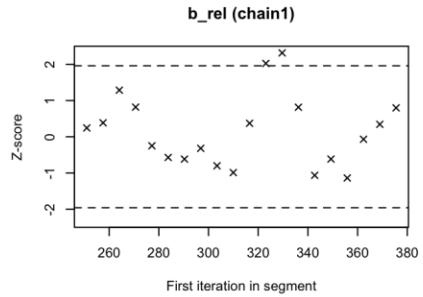


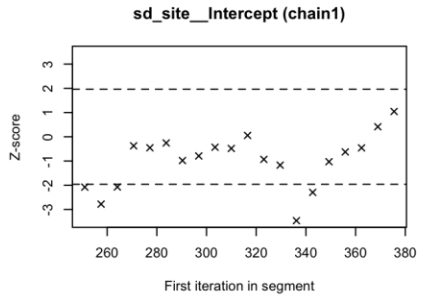


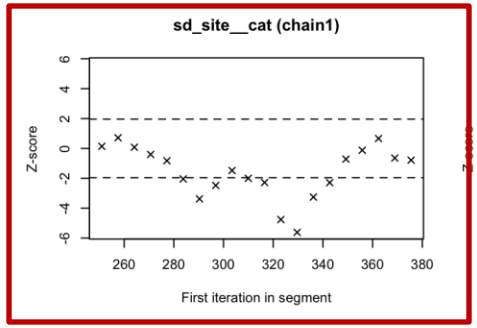
Geweke
(only chain 1)

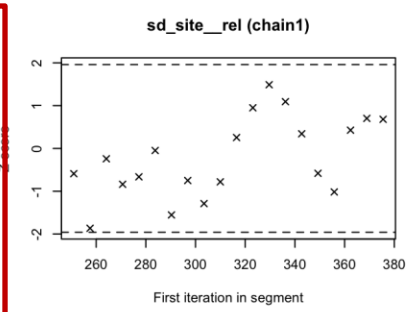












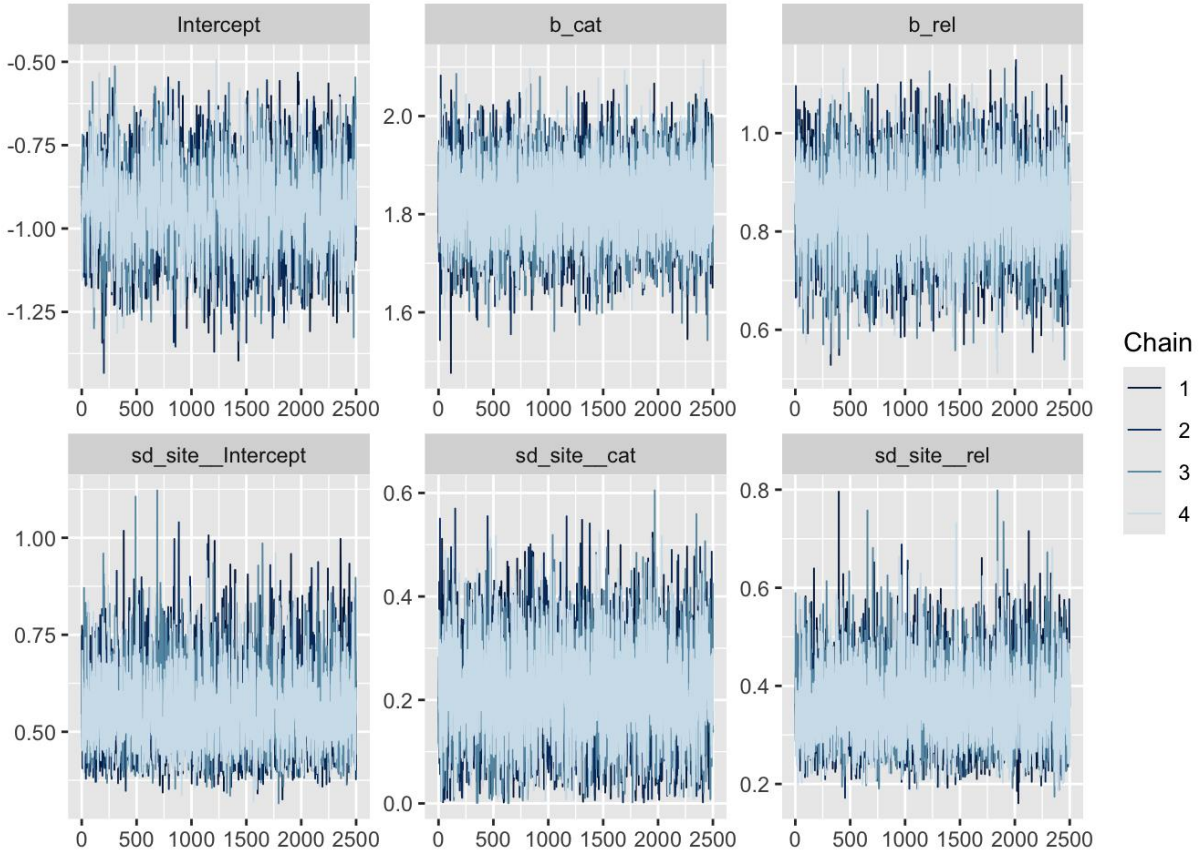
2. Assessing convergence

- Conclusion: not extremely bad, but also not great. Especially the between-country variability in the state category effect (*sd_cat*) is somewhat problematic.

- We aren't satisfied yet, and want to increase the number of iterations.

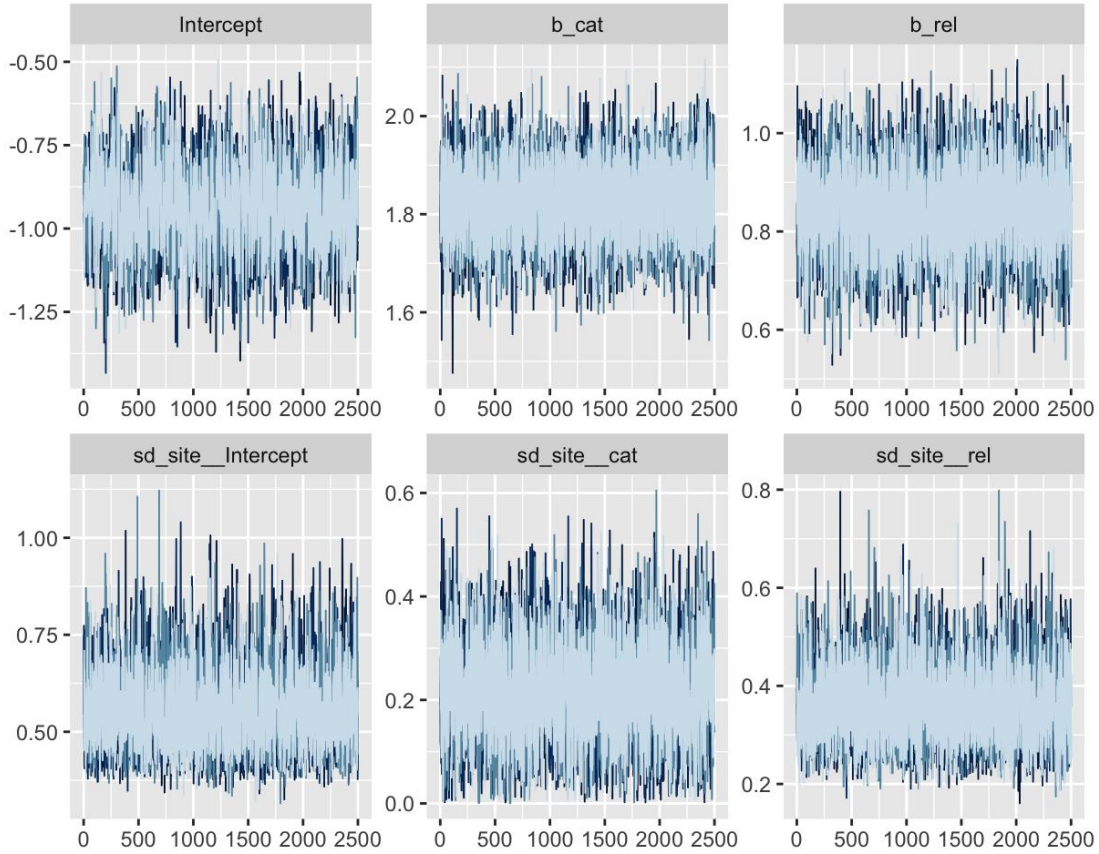
3. Assessing convergence with more iterations

$N_{iter} = 5000, N_{warmup} = 2500$



3. Assessing convergence with more iterations

N_{iter} = 5000, N_{warmup} = 2500




```
Potential scale reduction factors:

                    Point est. Upper C.I.
b_Intercept                  1       1.00
b_cat                        1       1.00
b_rel                        1       1.00
sd_site__Intercept           1       1.01
sd_site__cat                 1       1.00
sd_site__rel                 1       1.00

Multivariate psrf

1

[[3]]

Fraction in 1st window = 0.1
Fraction in 2nd window = 0.5

       b_Intercept              b_cat              b_rel sd_site__Intercept
            0.2467            -0.2113            -1.5816             0.2810
      sd_site__cat       sd_site__rel
           -1.2516            -0.5551


[[4]]

Fraction in 1st window = 0.1
Fraction in 2nd window = 0.5

       b_Intercept              b_cat              b_rel sd_site__Intercept
           0.86144           -0.07803            0.30301            0.65305
      sd_site__cat       sd_site__rel
           0.68209           -0.25264
```

3. Assessing convergence with more iterations

- We can also calculate *relative bias:* the difference in the estimated parameters between the initial model (M1) and the second model with more iterations (M2)

$$\text{bias} = 100 * \frac{\text{model with double iterations} - \text{initial converged model}}{\text{initial converged model}}$$

3. Assessing convergence with more iterations

- We can also calculate *relative bias:* the difference in the estimated parameters between the initial model (M1) and the second model with more iterations (M2)

$$\text{bias} = 100 * \frac{\text{model with double iterations} - \text{initial converged model}}{\text{initial converged model}}$$

- Bias should be small; rule of thumb:
  - if relative deviation is < |5|%, do not worry;
  - if relative deviation > |5|%, rerun with 4x nr of iterations.

3. Assessing convergence with more iterations

- We can also calculate *relative bias:* the difference in the estimated parameters between the initial model (M1) and the second model with more iterations (M2)

$$\text{bias} = 100 * \frac{\text{model with double iterations} - \text{initial converged model}}{\text{initial converged model}}$$

- Bias should be small; rule of thumb:
  - if relative deviation is < |5|%, do not worry;
  - if relative deviation > |5|%, rerun with 4x nr of iterations.

- Note that relative bias can only be interpreted in the context of the model parameters and substantive knowledge
  - E.g., with a regression coefficient of 0.0001, a 10% deviation might not be relevant. With an intercept of 50, a 10% devation might be quite meaningful

3. Assessing convergence with more iterations

Summary of convergence diagnostics and bias

Convergence diagnostics for the model with few iterations (250+250) and the model with more iterations (2500+2500)

Parameter	Estimate	Rhat	Geweke	Bias
continuity	-0.961 | -0.948	1.001 | 1.001	-0.555 | -0.258	-1.34%
state	1.826 | 1.827	1 | 1	-0.975 | -0.367	0.09%
religiosity	0.847 | 0.845	1.001 | 1	0.604 | -0.324	-0.17%
sd(continuity)	0.556 | 0.559	1.015 | 1.002	-0.26 | -0.209	0.51%
sd(state)	0.201 | 0.211	1.047 | 1	0.606 | -0.044	4.62%
sd(religiosity)	0.358 | 0.36	1.004 | 1.002	0.55 | -0.112	0.68%

4. Assessing posterior distributions

- The precision, or smoothness, of the histogram should be checked visually for each model parameter; we do not want gaps or other abnormalities.

- Note that visual inspection relates strongly to the *effective sample size* that the *brms* output also gives (see point 5)

4. Assessing posterior distributions

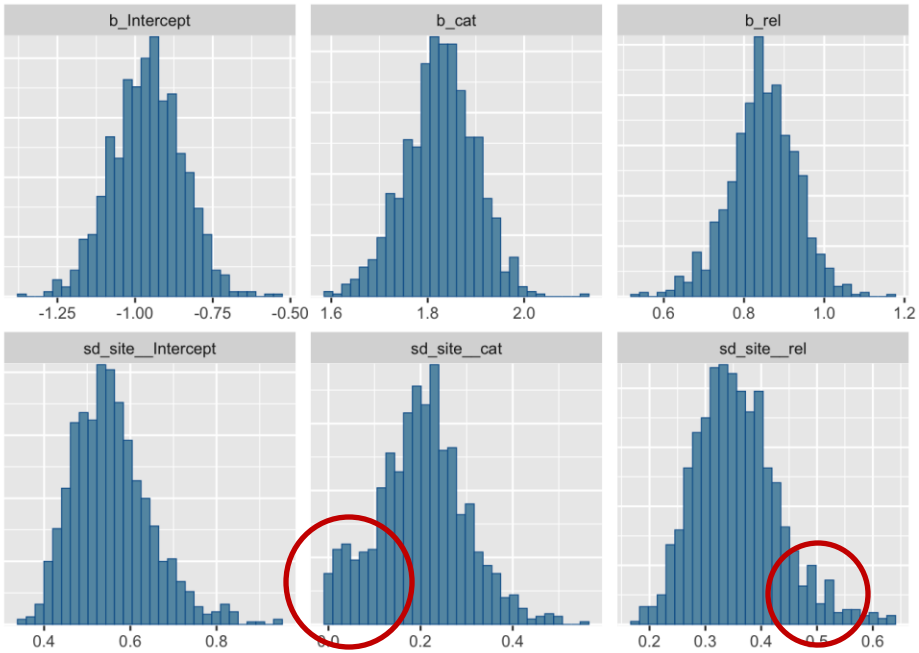
- The precision, or smoothness, of the histogram should be checked visually for each model parameter; we do not want gaps or other abnormalities.

- Note that visual inspection relates strongly to the *effective sample size* that the *brms* output also gives (see point 5)

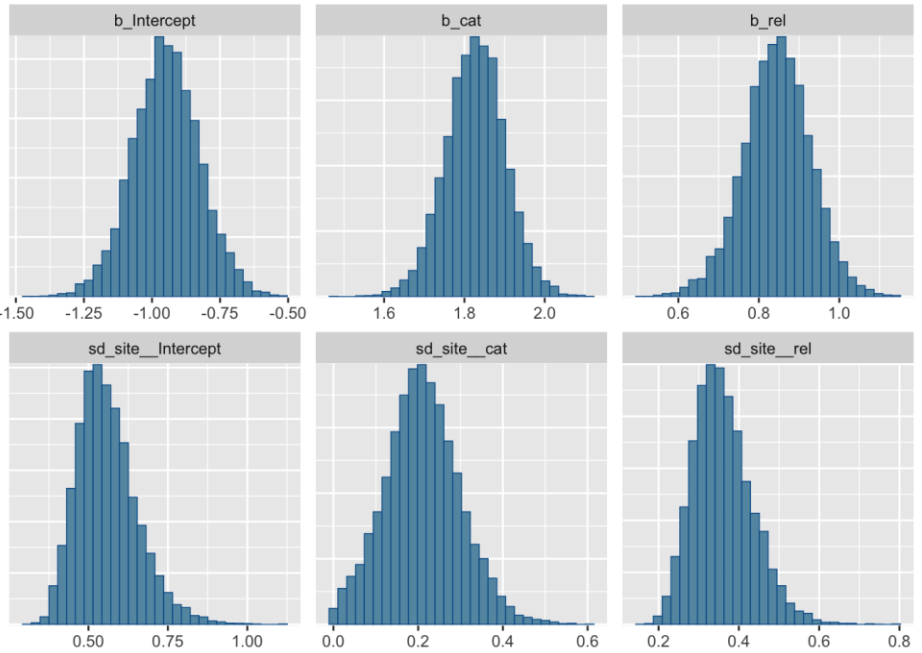
- We're not interpreting the posterior estimates yet, just evaluating the shape of the distribution!

4. Assessing posterior distributions

N_{iter} = 500, N_{warmup} = 250



N_{iter} = 5000, N_{warmup} = 2500



5. Assessing ESS and autocorrelation

- The effective sample size (ESS) is a measure of the number of independent samples in a Markov Chain Monte Carlo (MCMC) chain and reflects the efficiency of the algorithm.

5. Assessing ESS and autocorrelation

- The effective sample size (ESS) is a measure of the number of independent samples in a Markov Chain Monte Carlo (MCMC) chain and reflects the efficiency of the algorithm.

- It accounts for the *autocorrelation* in the chain, which can reduce the effective number of samples.
  - A higher ESS indicates that the MCMC chain is more efficient and provides more reliable estimates.

5. Assessing ESS and autocorrelation

- The effective sample size (ESS) is a measure of the number of independent samples in a Markov Chain Monte Carlo (MCMC) chain and reflects the efficiency of the algorithm.

- It accounts for the *autocorrelation* in the chain, which can reduce the effective number of samples.
  - A higher ESS indicates that the MCMC chain is more efficient and provides more reliable estimates.

- *brms* / Stan gives a warning if the number of ESS is too small.

5. Assessing ESS and autocorrelation

- MCMC iterations are typically dependent on each other
  - E.g., if iteration *t* of a Markov chain produces an estimate of .34 for a regression coefficient, then iteration *t*+1 will produce an estimate correlated with the previous one.

5. Assessing ESS and autocorrelation

- MCMC iterations are typically dependent on each other
  - E.g., if iteration *t* of a Markov chain produces an estimate of .34 for a regression coefficient, then iteration *t*+1 will produce an estimate correlated with the previous one.

- Amount of autocorrelation depends on sampling algorithm (Stan vs JAGS), model complexity, parameter type

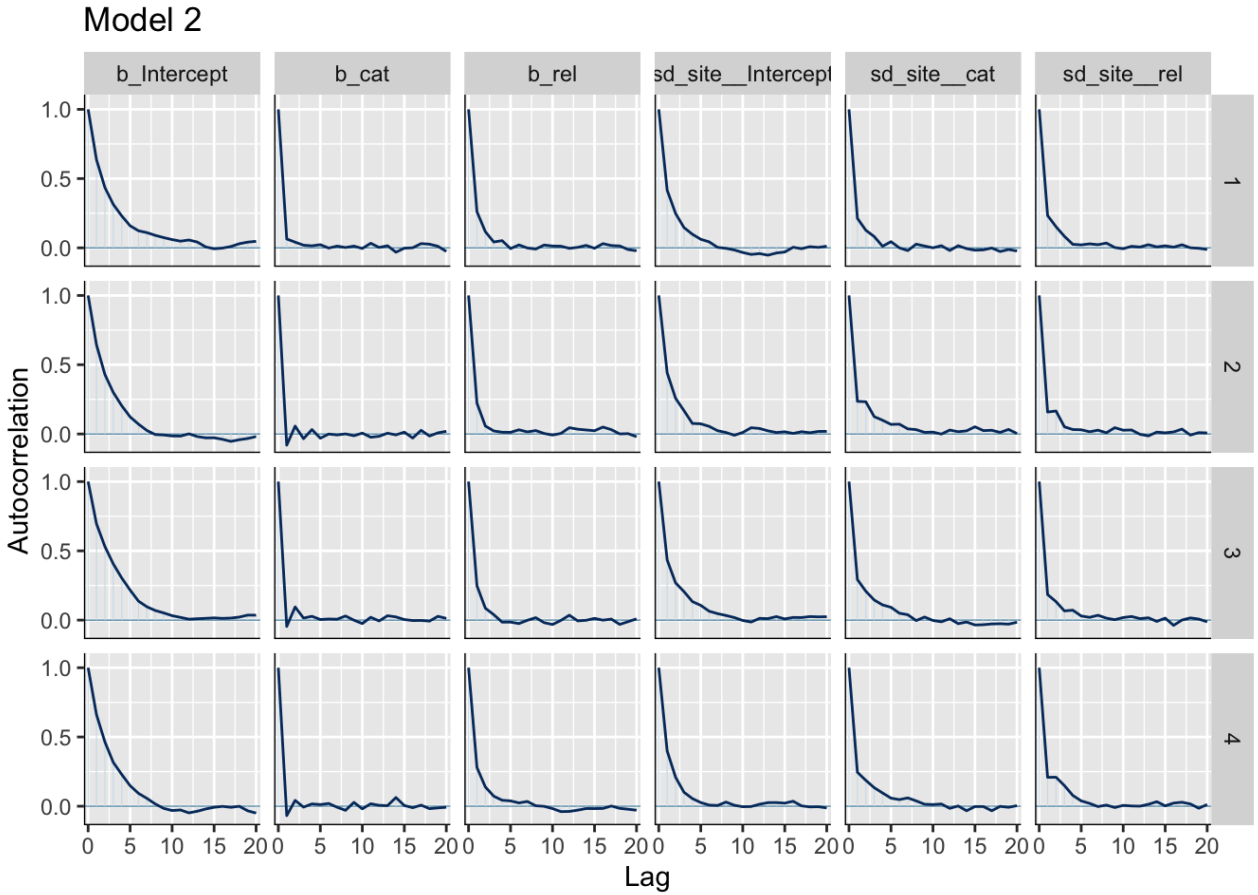
5. Assessing ESS and autocorrelation

- MCMC iterations are typically dependent on each other
  - E.g., if iteration *t* of a Markov chain produces an estimate of .34 for a regression coefficient, then iteration *t*+1 will produce an estimate correlated with the previous one.

- Amount of autocorrelation depends on sampling algorithm (Stan vs JAGS), model complexity, parameter type

- Historically, *thinning* has been used to reduce autocorrelation in MCMC chains, but this is no longer recommended. Instead, it is better to increase the number of iterations and warmup samples to ensure that the chains are well-mixed and that the effective sample size is sufficient.

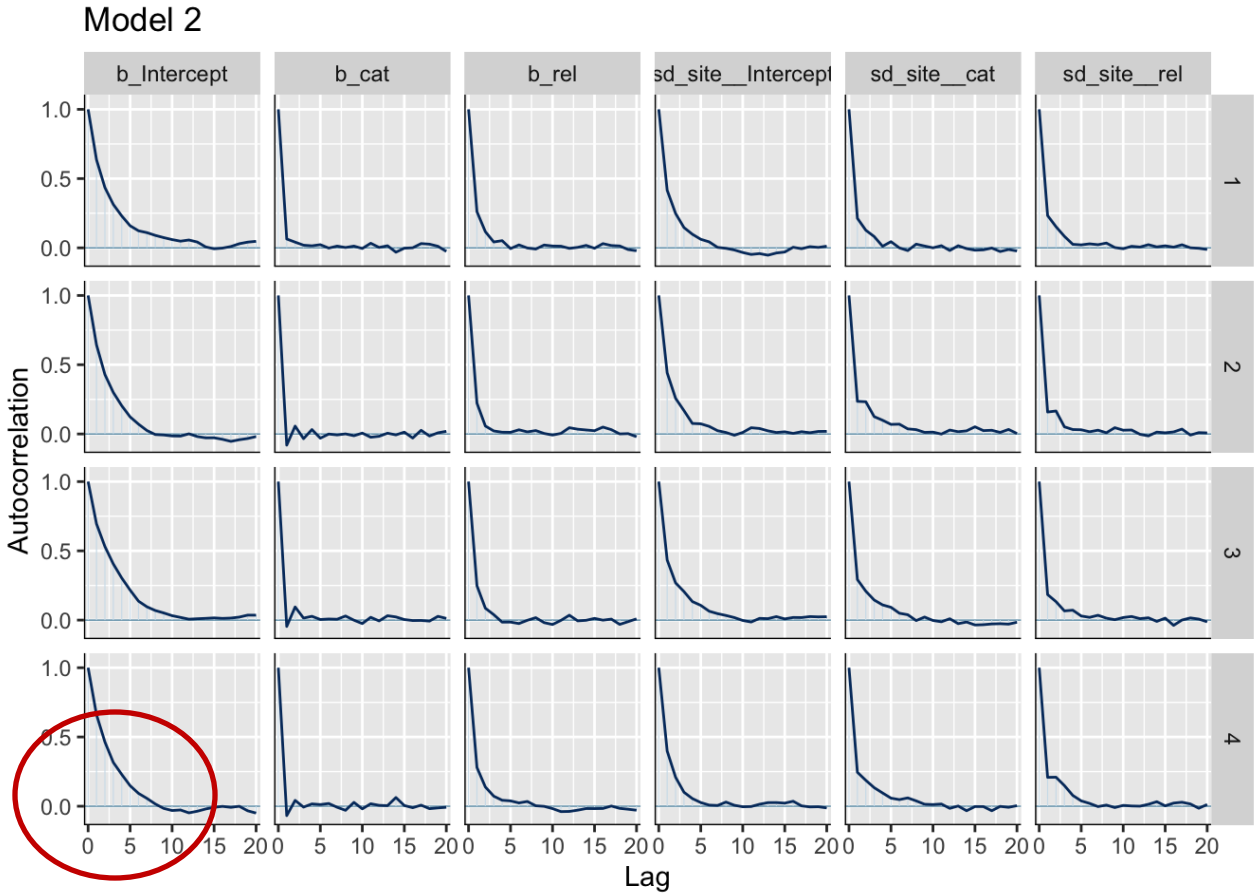
5. Assessing ESS and autocorrelation



Ratio ESS / $N_{samples}$

```
b_Intercept      b_cat      b_rel sd_site__Intercept
  0.1902673  0.7749418  0.5226856          0.3179952
sd_site__cat sd_site__rel
  0.3568674  0.4257796
```

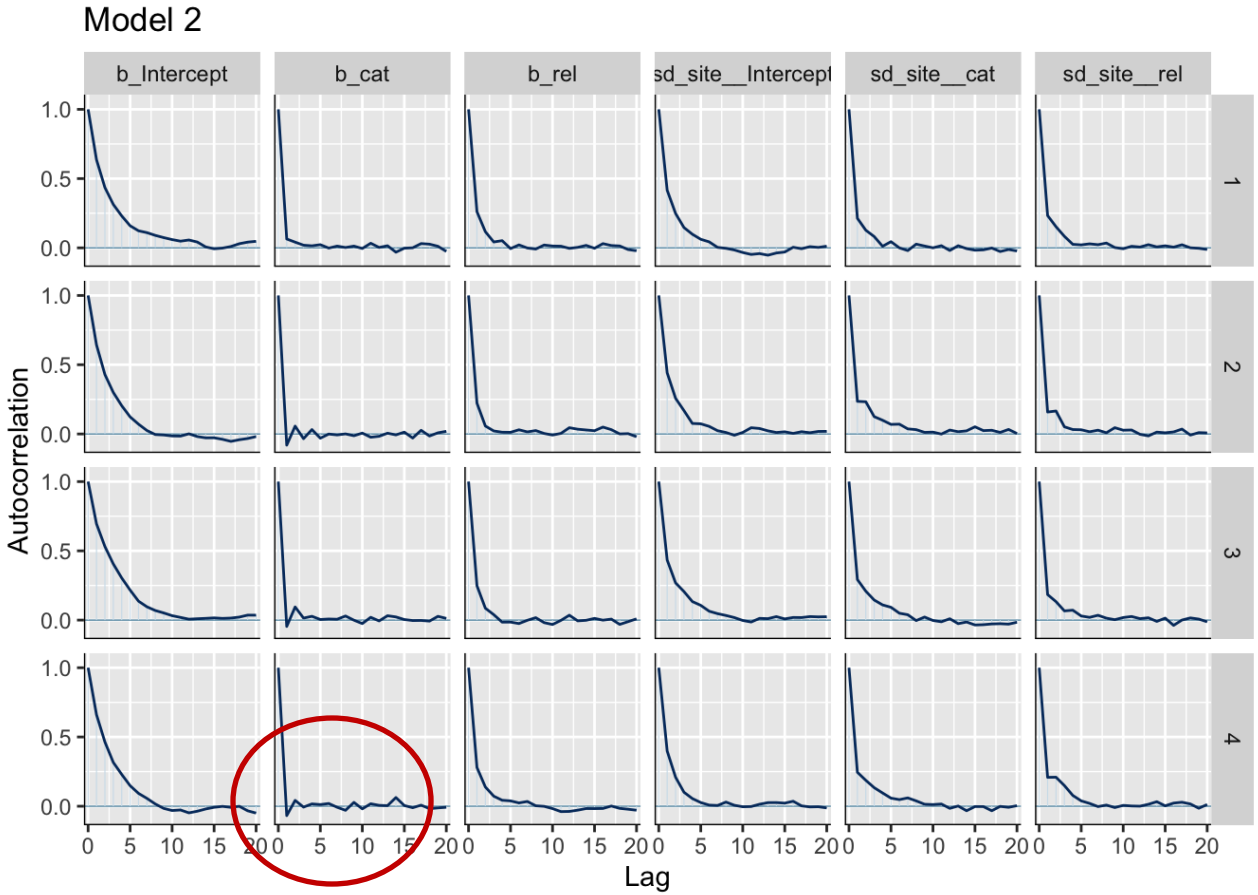
5. Assessing ESS and autocorrelation



Ratio ESS / $N_{samples}$

```
b_Intercept       b_cat       b_rel sd_site__Intercept
   0.1902673   0.7749418   0.5226856         0.3179952
sd_site__cat sd_site__rel
   0.3568674   0.4257796
```

5. Assessing ESS and autocorrelation



Ratio ESS / $N_{samples}$

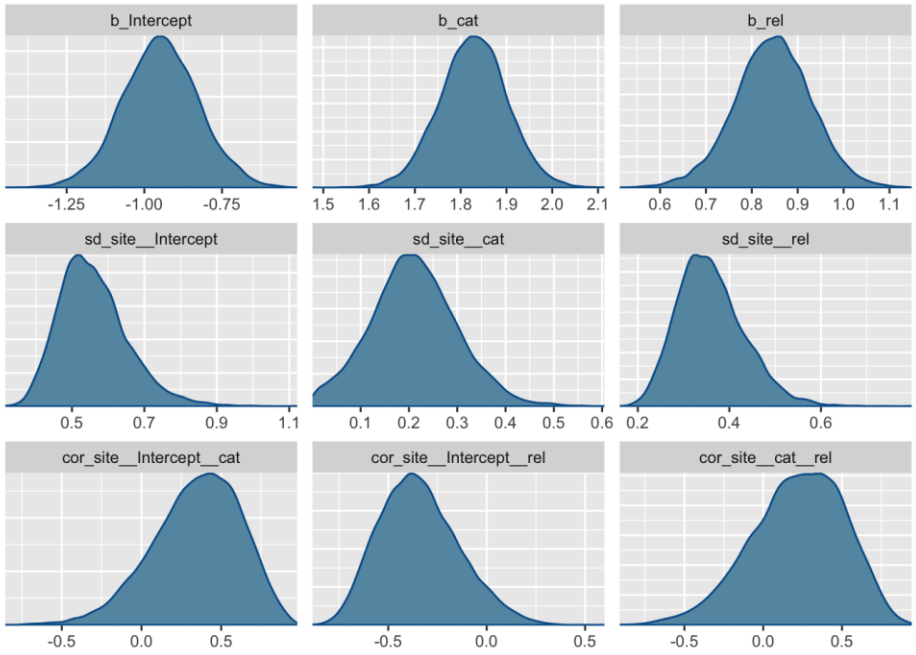
```
b_Intercept        b_cat        b_rel sd_site__Intercept
  0.1902673    0.7749418    0.5226856          0.3179952
sd_site__cat sd_site__rel
  0.3568674    0.4257796
```

6. Posterior predictive checks

- For the *marginal posterior distributions* of the parameters we want to make sure that they:
  1. Are smooth
  2. Make substantive sense (e.g., positive effect of mental vs bodily states on continuity)
  3. Are not too wide; the posterior SD and CI should not be larger than the scale of the original parameter

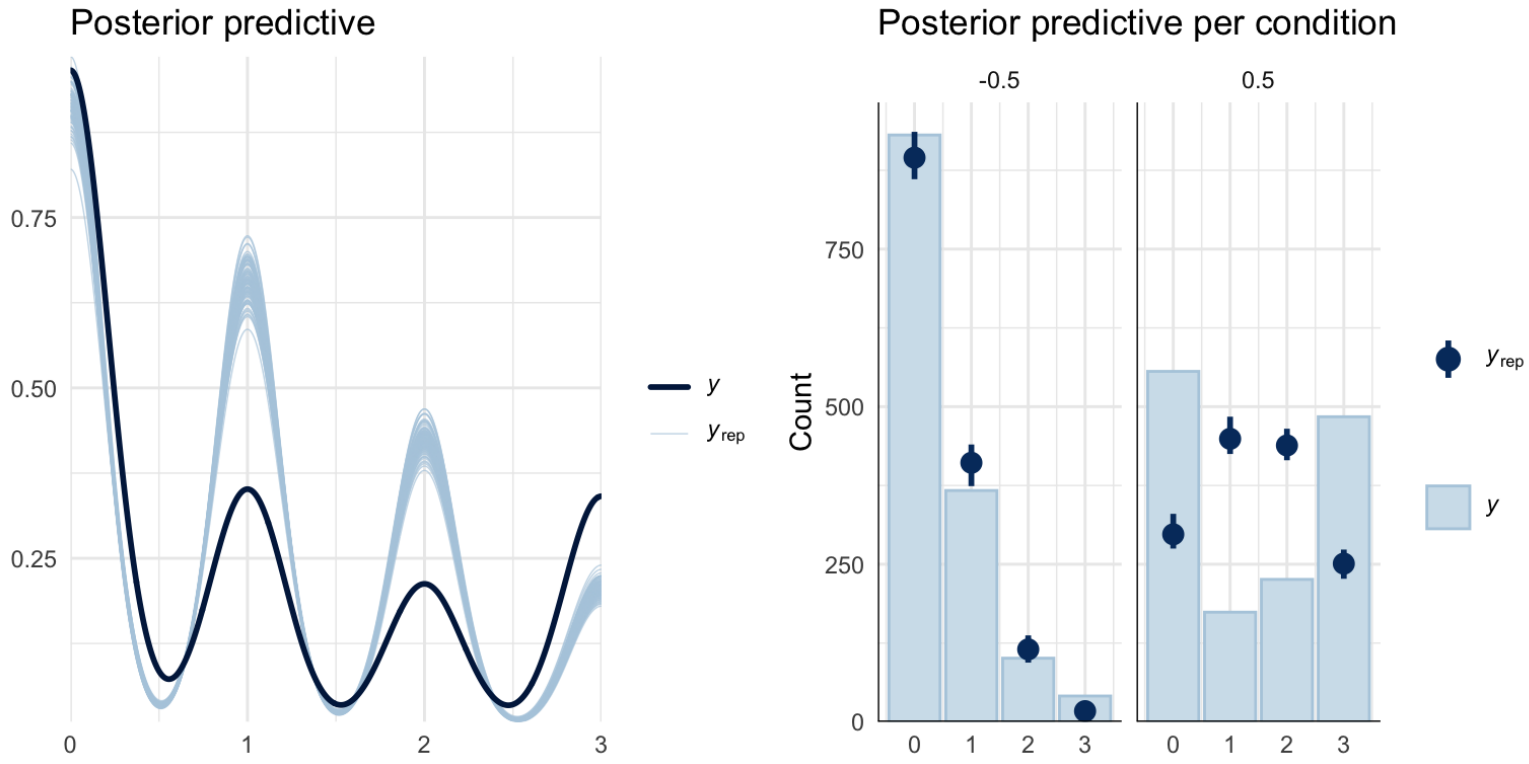
6. Posterior predictive checks

- For the *marginal posterior distributions* of the parameters we want to make sure that they:
  1. Are smooth
  2. Make substantive sense (e.g., positive effect of mental vs bodily states on continuity)
  3. Are not too wide; the posterior SD and CI should not be larger than the scale of the original parameter



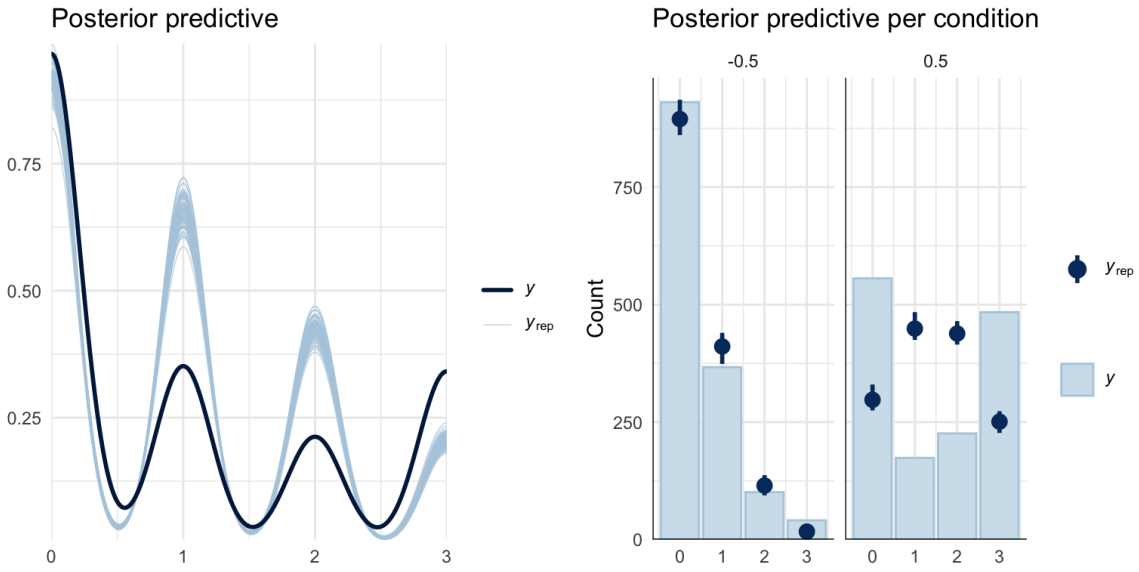
6. Posterior predictive checks

- In addition, we want to look at the posterior predictions (on the response scale)



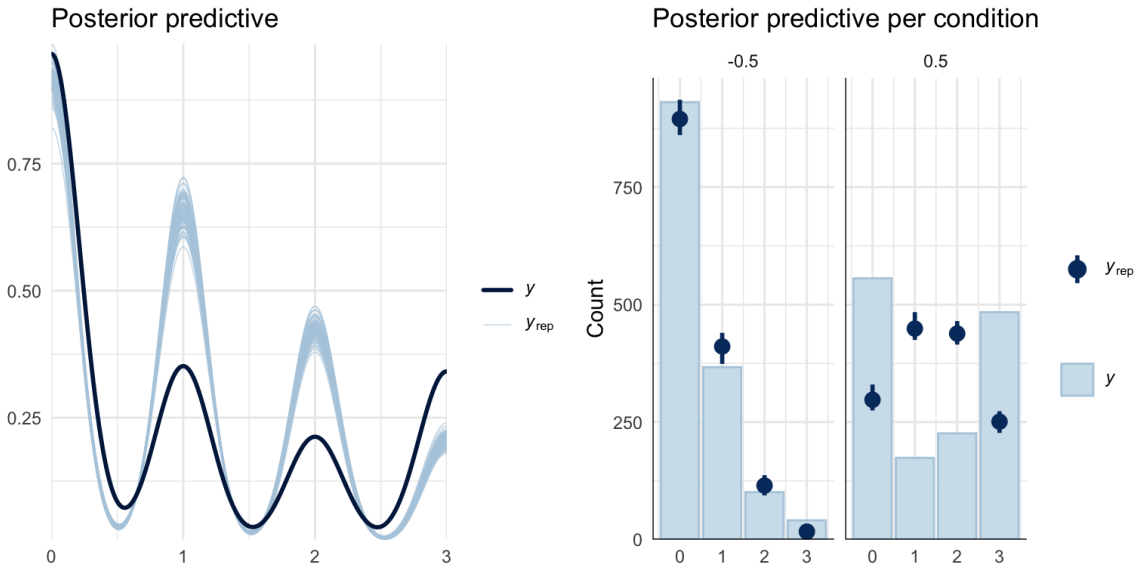
6. Posterior predictive checks

- Posterior predictions do not look great; most likely the aggregated binomial model is not the best fit for these data.



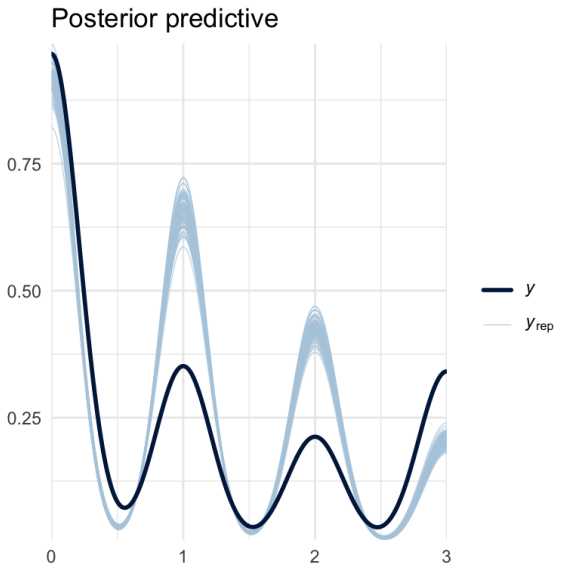
6. Posterior predictive checks

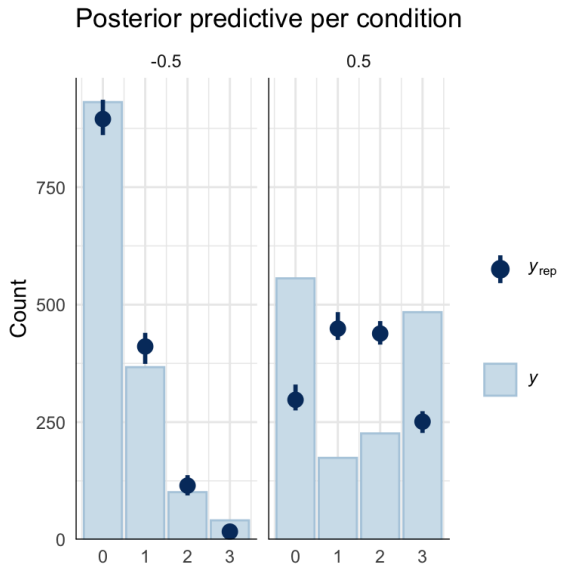
- Posterior predictions do not look great; most likely the aggregated binomial model is not the best fit for these data.
- Note however, that you sometimes need to make a trade-off between model complexity and model fit; simpler models are typically easier to estimate and interpret, but might show more misfit to the exact data pattern.



6. Posterior predictive checks

- Posterior predictions do not look great; most likely the aggregated binomial model is not the best fit for these data.
- Note however, that you sometimes need to make a trade-off between model complexity and model fit; simpler models are typically easier to estimate and interpret, but might show more misfit to the exact data pattern.
- As a researcher you need to decide what fit is good enough, and ideally conduct sensitivity analyses on other likelihood functions





Stage 3: Robustness

- Are the model estimates robust against alternative plausible specifications?

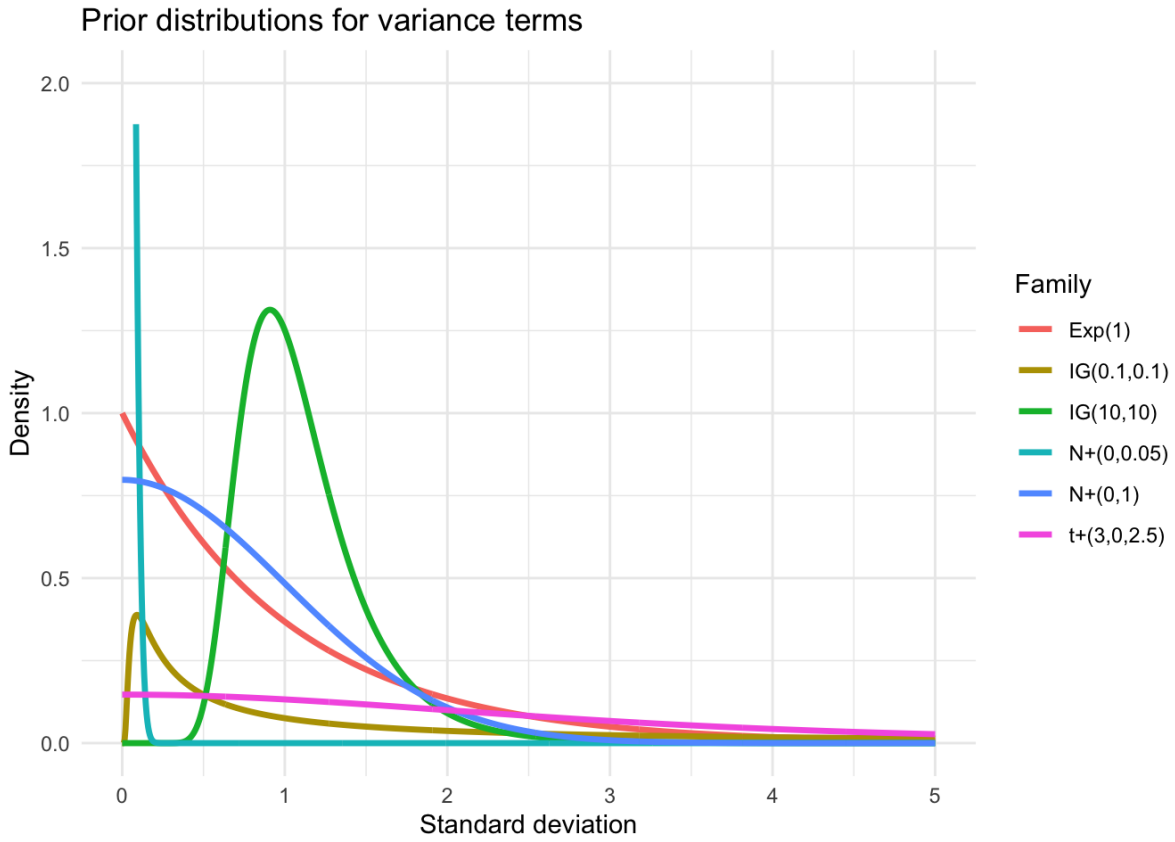
7. Prior sensitivity checks: variance terms
8. Prior sensitivity checks: informativeness of effect parameters
9. Sensitivity checks on model specification (likelihood)

7. Prior sensitivity: variance

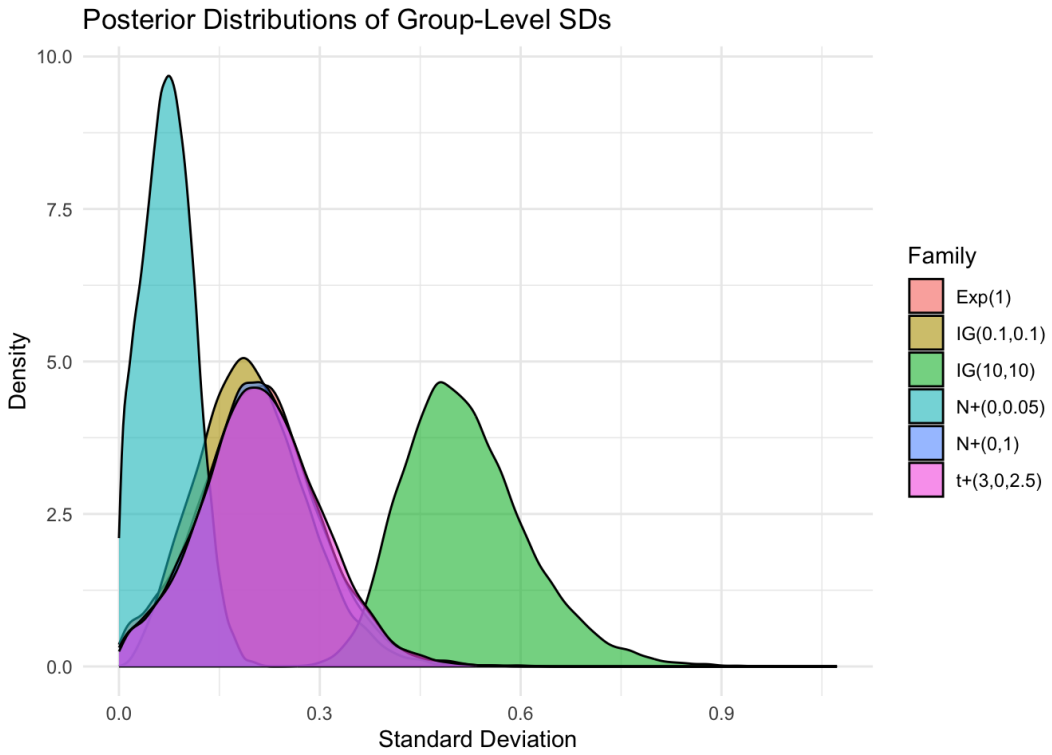
- In step 1 (understanding priors) we considered different families of priors on the between-country variance terms. For illustration, we add a few highly informative ones as well

7. Prior sensitivity: variance

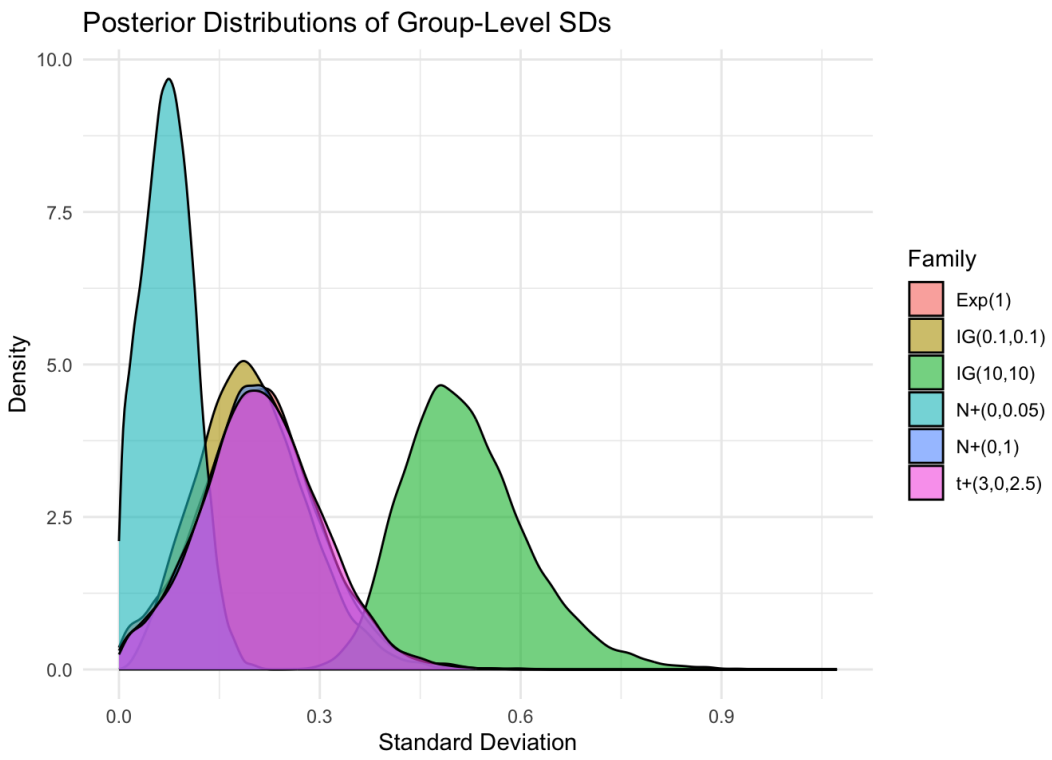
- In step 1 (understanding priors) we considered different families of priors on the between-country variance terms. For illustration, we add a few highly informative ones as well

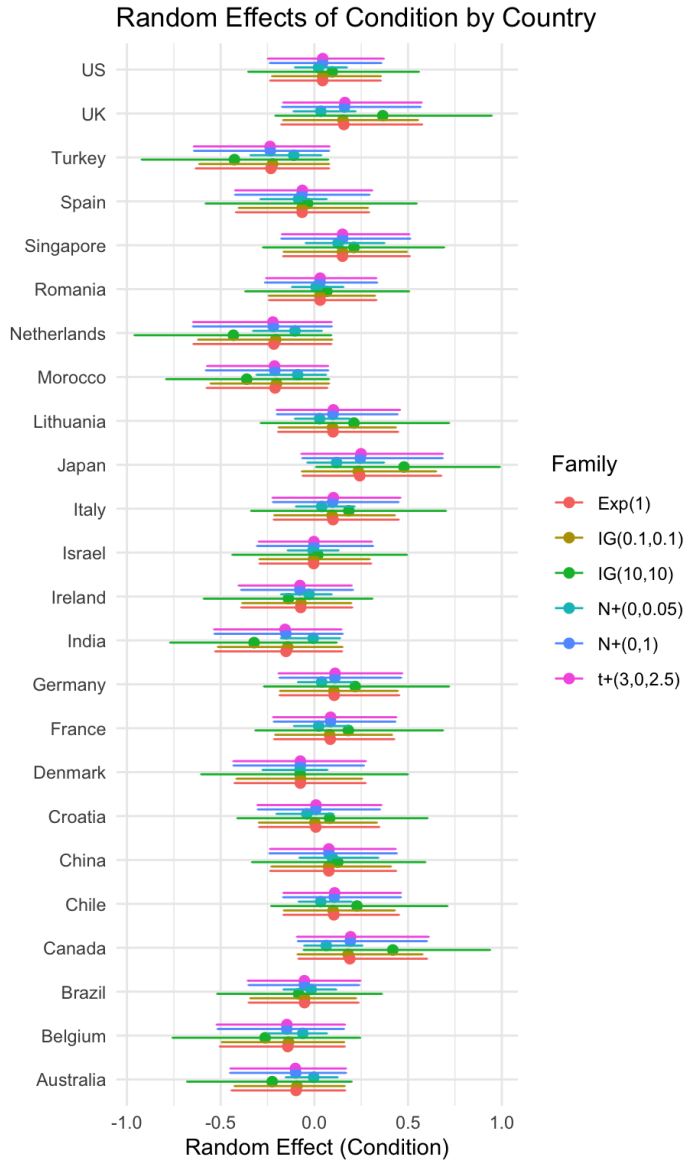


7. Prior sensitivity: variance



7. Prior sensitivity: variance





8. More prior sensitivity: effects

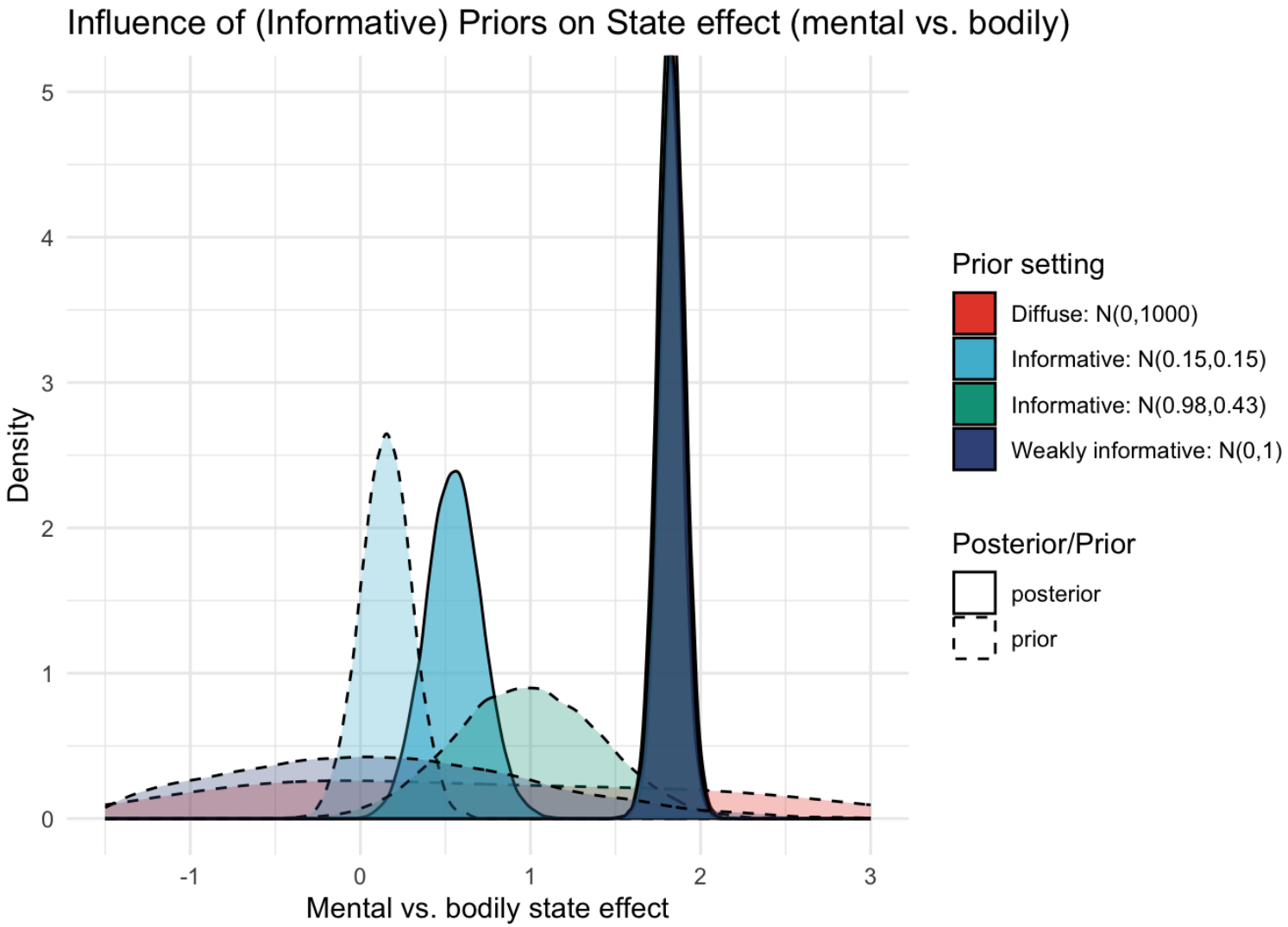
- We can also vary the priors on the effects
- This is not so much about different families, but different levels of informativeness (i.e., certainty), e.g.:

8. More prior sensitivity: effects

- We can also vary the priors on the effects

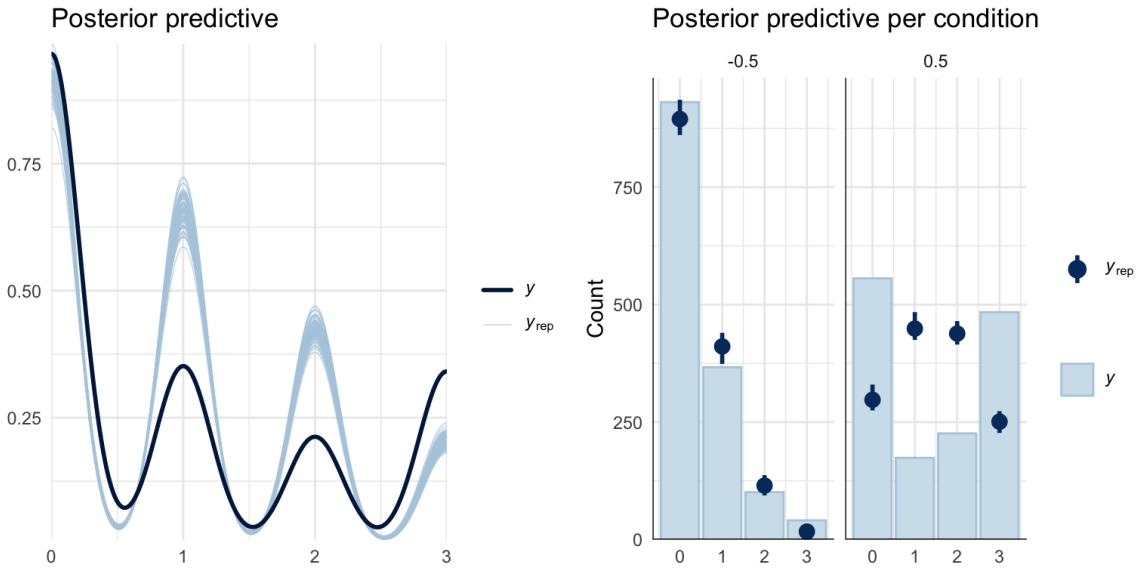
- This is not so much about different families, but different levels of informativeness (i.e., certainty), e.g.:
  - Diffuse: N(0, 1000)
    - *brms* default (approx.), non-sensible predictions
  - Weakly-informative: N(0, 1)
    - Reasonable, but not restrictive
  - Informative: N(0.98, 0.43)
    - Derived from the literature
  - Informative: N(0.15, 0.15)
    - Mistakenly derived from literature (15%, 15% ≠ N(0.15, 0.15))

8. More prior sensitivity: effects



9. Sensitivity checks on the model

- We saw that posterior predictive check indicated somewhat acceptable but not great model fit
  - Fails to capture zero-inflation and/or extremity responses (0/3 and 3/3)



9. Sensitivity checks on the model

- We saw that posterior predictive check indicated somewhat acceptable but not great model fit
  - Fails to capture zero-inflation and/or extremity responses (0/3 and 3/3)

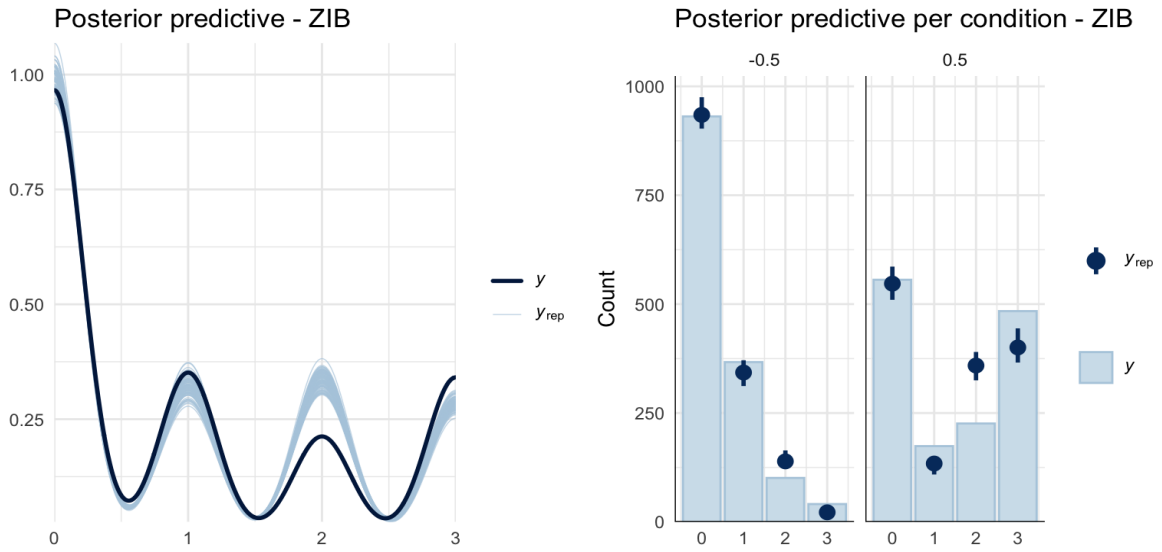
- Alternative models:
  - Zero-inflated binomial

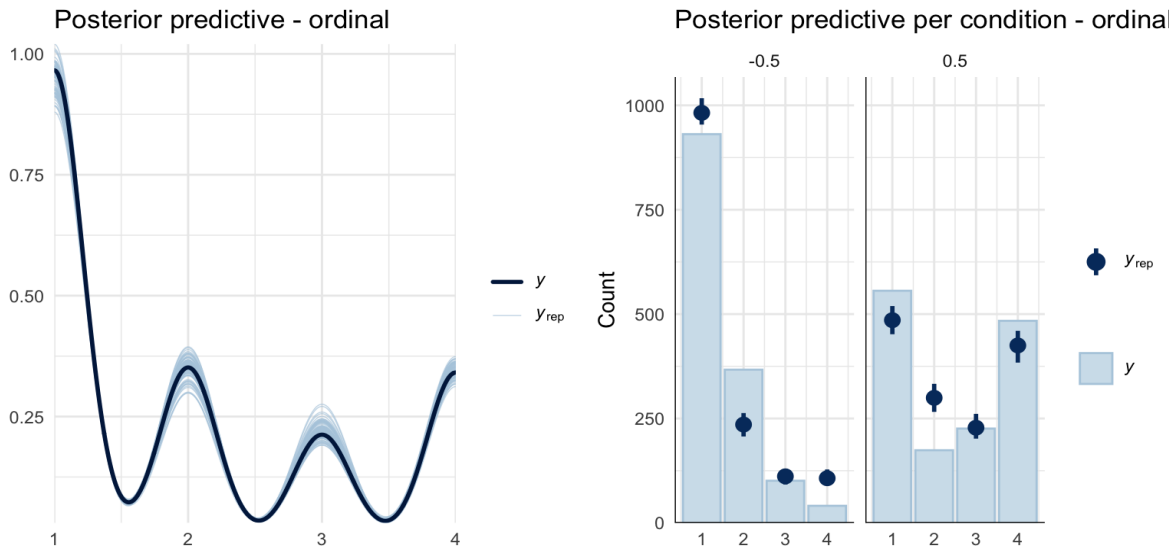
```
brm(formula = response | trials(3) ~ 1 + state_cond + relig +
(1 + state_cond + relig | country),
zi ~ 1 + state_cond + rel,
family = zero_inflated_binomial("logit"), ...)
```

  - Ordinal model

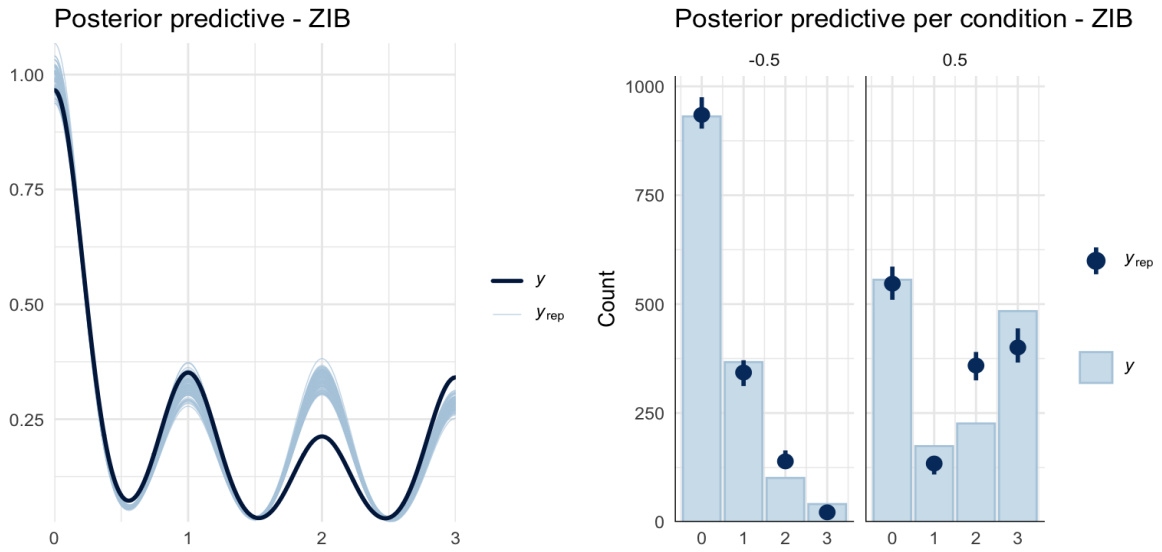
```
brm(formula = response_cat ~ 1 + state_cond + relig + (1 +
state_cond + relig | country),
family = cumulative("logit"), ...)
```

9. Sensitivity checks on the model

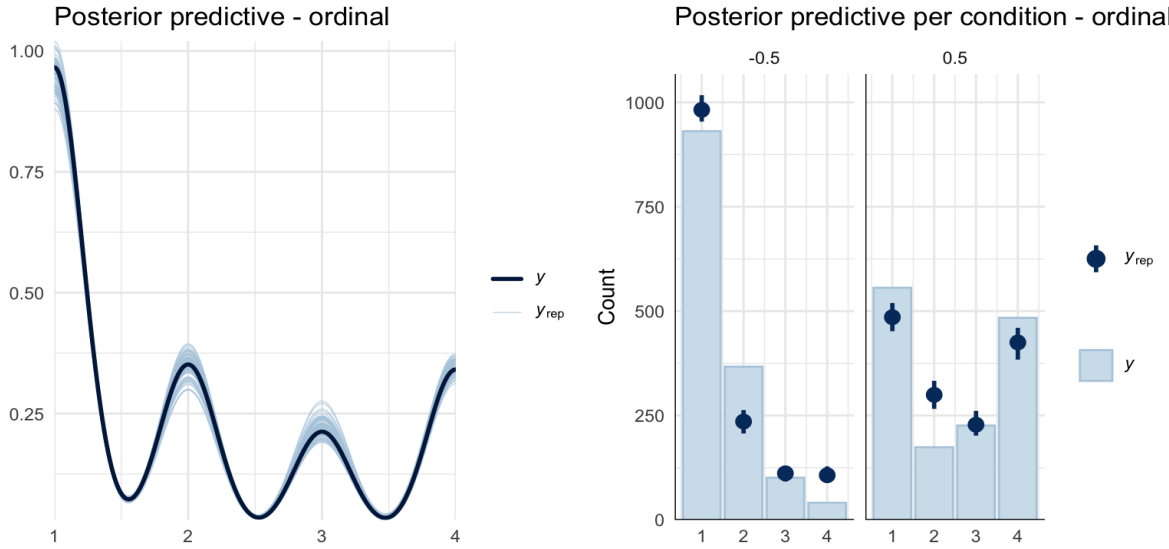




9. Sensitivity checks on the model



Slightly off on 2/3, good fit for pattern of conditions



Perfect fit overall responses, slightly off on differences between conditions

9. Sensitivity checks on the model

- Overall, very similar pattern across models (and raw data): clear effect of mental vs bodily states on continuity judgments

Stage 4: Reporting findings

- For comprehensibility:
  - Interpret results in the Bayesian framework, with reference to the full posterior distribution and uncertainty

- For transparency and reproducibility:
  - Report all details on packages, model specification, sampling method, priors etc.

10. Reporting findings

- In the Bayesian framework, we have not only point estimates, but full posterior distributions of all parameters, capturing the *uncertainty* in the true value

- We can summarize with posterior mean or median + credible interval

10. Reporting findings

- In the Bayesian framework, we have not only point estimates, but full posterior distributions of all parameters, capturing the *uncertainty* in the true value

- We can summarize with posterior mean or median + credible interval

- This CI is different from a frequentist confidence interval:
  - 95% confidence interval: across many repetitions of the study (under the same circumstances), 95% of the confidence intervals will contain the true value
  - 95% credible interval: based on the current data, there is a 95% probability that the true value is within the interval.

10. Reporting findings

Aspects to report:

1) Estimates + credible intervals (+ figure!)
2) Software and packages used (with version)
3) Discussion of sampling settings (number of chains, number of interaction, warmup, seeds etc.)
4) Discussion of sampling diagnostics (e.g., $\hat{R}$, effective sample size)
5) Discussion of priors (justify choices, and report sensitivity analyses)
6) Perhaps: model fit or model comparison metrics (e.g., loo, Bayes factors)

10. Reporting findings

1) Estimates:

- The estimate for the fixed intercept (overall continuity) is -0.95 95% CI [-1.19; -0.7], this translates into 0.279 [0.112, 0.557] on the probability scale, meaning that on average, people judge 28% [11%, 56%] of states to continue after physical death. Estimates range from 14% in Spain to 52% in Singapore.

10. Reporting findings

1) Estimates:

- The estimate for the fixed intercept (overall continuity) is -0.95 95% CI [-1.19; -0.7], this translates into 0.279 [0.112, 0.557] on the probability scale, meaning that on average, people judge 28% [11%, 56%] of states to continue after physical death. Estimates range from 14% in Spain to 52% in Singapore.
- The estimate for the fixed effect of state category is 1.83 [1.68; 1.97], which means 35.4% [16.2%, 49.5%] percentage points higher continuity for the mental state category compared to the bodily state category. Estimates range from 21.5% in the Netherlands to 46.9% in Japan.

10. Reporting findings

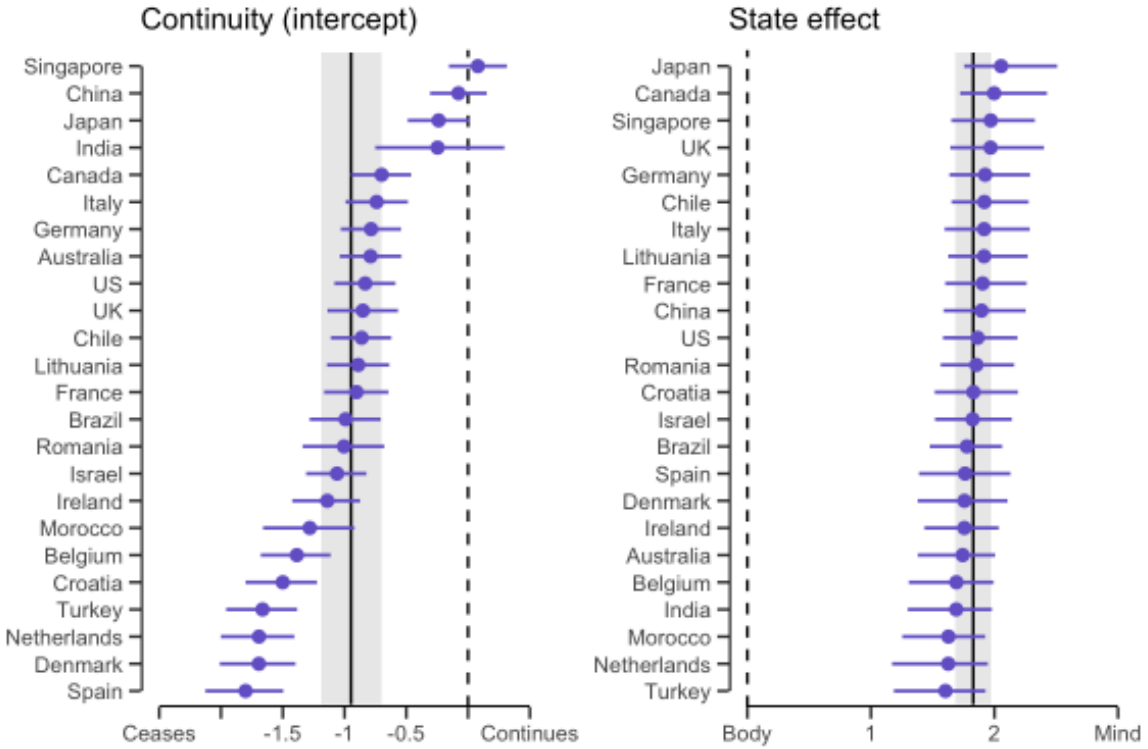


Figure 4: Estimated country-level effects (posterior medians) of the mental vs bodily state effect (left) and overall continuity (right) in increasing order. Each dot represents a country. The errorbars give the 95% credible interval for each country. The vertical lines denote the posterior median of the overall mean (fixed effect) of the respective effect with the 95% credible interval in the shaded bands. The dashed lines indicates zero.

10. Reporting findings

2) software + 3) sampling settings

- We used the *brms* package (Bürkner, 2017) to fit Bayesian multilevel models in R, which relies on the Stan language (Carpenter et al., 2017) with weakly informative priors. The model was run with 4 chains, each with 5000 iterations and a warmup of 2500 iterations (total post-warmup N=10000). We used the *cmdstanr* backend for efficient sampling.

10. Reporting findings

2) software + 3) sampling settings

- We used the *brms* package (Bürkner, 2017) to fit Bayesian multilevel models in R, which relies on the Stan language (Carpenter et al., 2017) with weakly informative priors. The model was run with 4 chains, each with 5000 iterations and a warmup of 2500 iterations (total post-warmup N=10000). We used the *cmdstanr* backend for efficient sampling.

4) sampling diagnostics

- The model diagnostics indicated good convergence (largest $\hat{R}$ = 1.0017 for the random intercept in Germany) and sufficient effective sample sizes for all parameters (median $\widehat{N}_{eff}$ = 4258). The smallest $\widehat{N}_{eff}$ = 1903 for the overall intercept, indicating that there is some autocorrelation in the chains.

10. Reporting findings

5) Discussion of priors and sensitivity analyses

- We used an aggregated binominal model with a ‘logit’ link function and weakly informative priors derived from comparable previous studies on the transformed intercept (N(0,1)), on the state effect (N(0,1)), on the standard deviation of the random effects across countries (N+(0,1)), and an LKJ(2) prior on the correlation between random effects.

10. Reporting findings

5) Discussion of priors and sensitivity analyses

- We used an aggregated binominal model with a 'logit' link function and weakly informative priors derived from comparable previous studies on the transformed intercept (N(0,1)), on the state effect (N(0,1)), on the standard deviation of the random effects across countries (N+(0,1)), and an LKJ(2) prior on the correlation between random effects.
- Sensitivity checks on the priors and model specification indicated that the conclusions are robust to different reasonable prior settings (e.g., inverse gamma, exponential or diffuse Student t priors on the between-country variance; diffuse N(0,1000) prior on the regression coefficients) and model specifications (e.g., a zero-inflated binomial model and an ordinal model). Details of these sensitivity checks are provided in the appendix.

(Bayesian) research cycle