Detecting and diagnosing prior and likelihood sensitivity with power-scaling

Noa Kallioinen Topi Paananen Paul-Christian Bürkner Aki Vehtari

Probabilistic Machine Learning group, Department of Computer Science, Aalto University

Finnish Center for Artificial Intelligence

Cluster of Excellence SimTech, University of Stuttgart

Overview

Background

$$p(\theta|y) \propto p(\theta)p(y|\theta)$$

- Influence and interplay of both *prior* and *likelihood*
- Both have influence on the posterior
- Both need to be specified by the modeller
- *priorsense*: uses power-scaling to detect and diagnose influence

Motivation and goals

Clear need for modern sensitivity analysis tools

Example: Only 16% of psychology studies using Bayesian methods included a prior sensitivity analysis[1]

[1] R. van de Schoot, S. D. Winter, O. Ryan, *et al.*, “A systematic review of Bayesian articles in psychology: The last 25 years.,” *Psychological Methods*, vol. 22, no. 2, pp. 217–239, Jun. 2017. DOI: 10.1037/met0000100.

Motivation and goals

Clear need for modern sensitivity analysis tools

Example: Only 16% of psychology studies using Bayesian methods included a prior sensitivity analysis[1]

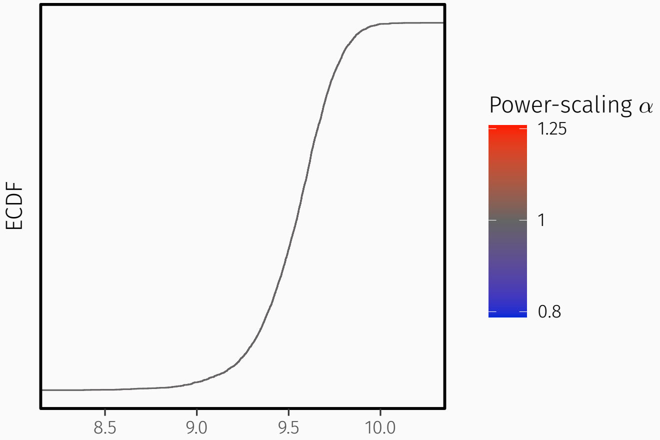
These tools should be:

- semi-automated
- computationally efficient
- trustworthy and self-diagnosing
- applicable to a wide range of models

[1] R. van de Schoot, S. D. Winter, O. Ryan, *et al.*, "A systematic review of Bayesian articles in psychology: The last 25 years.," *Psychological Methods*, vol. 22, no. 2, pp. 217–239, Jun. 2017. DOI: 10.1037/met0000100.

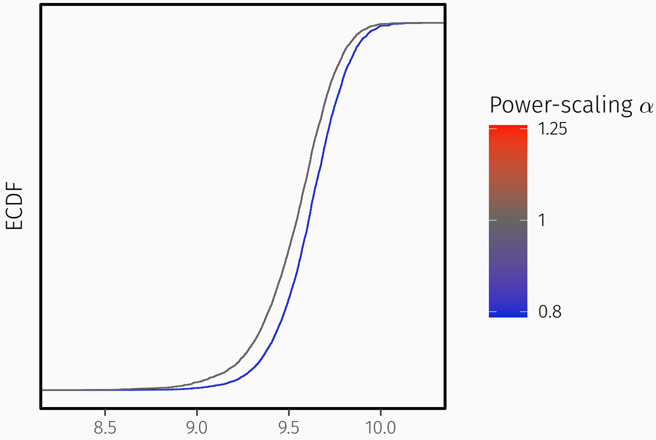
Project outcome

`priorsense`: A tool which *quantifies* and *visualises* how ‘weakening’ or ‘strengthening’ the prior or likelihood influences the posterior



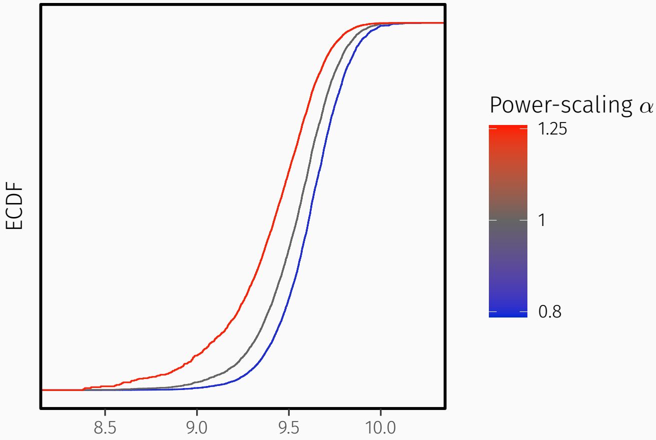
Project outcome

`priorsense`: A tool which *quantifies* and *visualises* how ‘weakening’ or ‘strengthening’ the prior or likelihood influences the posterior



Project outcome

`priorsense`: A tool which *quantifies* and *visualises* how ‘weakening’ or ‘strengthening’ the prior or likelihood influences the posterior



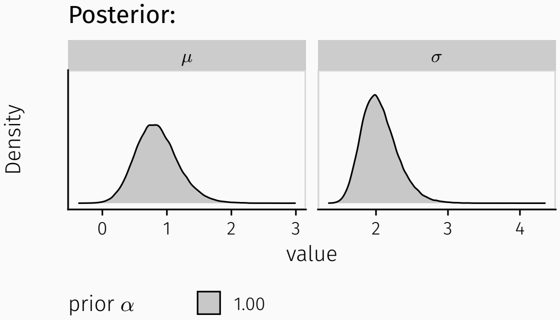
Approach

Envisioned workflow

Sensitivity analysis:

1. Fit a starting template model to data (using MCMC)
2. Estimate perturbed posteriors (modified prior / likelihood)
3. Determine change from base posterior to perturbed posteriors
4. If there is sensitivity, investigate cause further

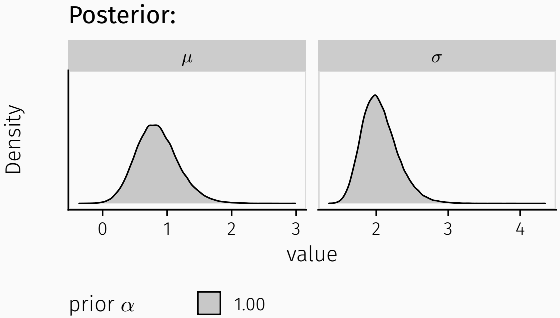
Which priors or likelihoods to compare?



Details

Which priors or likelihoods to compare?

We independently power-scale the prior and likelihood
$p(\theta|y) \propto \boldsymbol{p(\theta)^{\alpha}} p(y|\theta)$
$p(\theta|y) \propto p(\theta) \boldsymbol{p(y|\theta)^{\alpha}}$



Details

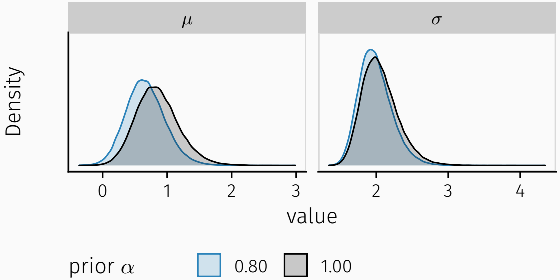
Which priors or likelihoods to compare?

We independently power-scale the prior and likelihood

$p(\theta|y) \propto \boldsymbol{p(\theta)^{\alpha}} p(y|\theta)$

$p(\theta|y) \propto p(\theta) \boldsymbol{p(y|\theta)^{\alpha}}$

Posterior:



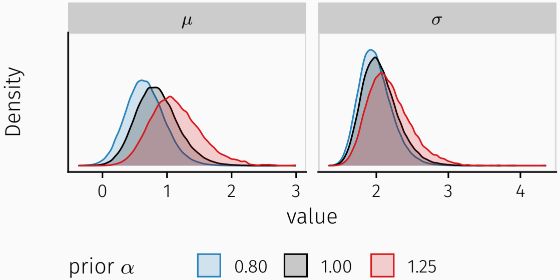
Which priors or likelihoods to compare?

We independently power-scale the prior and likelihood

$p(\theta|y) \propto \boldsymbol{p(\theta)^{\alpha}} p(y|\theta)$

$p(\theta|y) \propto p(\theta) \boldsymbol{p(y|\theta)^{\alpha}}$

Posterior:



Efficient estimation

We don’t want to refit the model each time we scale

Efficient estimation

We don’t want to refit the model each time we scale

- Estimate posterior properties with importance sampling

Efficient estimation

We don't want to refit the model each time we scale

- Estimate posterior properties with importance sampling
- Prior scaling: $w_{\alpha_{pr}}^{(s)} = \frac{p(\theta^{(s)})^{\alpha} p(y|\theta^{(s)})}{p(\theta^{(s)}) p(y|\theta^{(s)})} = p(\theta^{(s)})^{\alpha-1}$

Efficient estimation

We don't want to refit the model each time we scale

- Estimate posterior properties with importance sampling
- Prior scaling: $w_{\alpha_{pr}}^{(s)} = \frac{p(\theta^{(s)})^{\alpha} p(y|\theta^{(s)})}{p(\theta^{(s)}) p(y|\theta^{(s)})} = p(\theta^{(s)})^{\alpha-1}$
- Likelihood scaling: $w_{\alpha_{lik}}^{(s)} = \frac{p(\theta^{(s)}) p(y|\theta^{(s)})^{\alpha}}{p(\theta^{(s)}) p(y|\theta^{(s)})} = p(y|\theta^{(s)})^{\alpha-1}$

Efficient estimation

We don't want to refit the model each time we scale

- Estimate posterior properties with importance sampling
- Prior scaling: $w_{\alpha_{pr}}^{(s)} = \frac{p(\theta^{(s)})^{\alpha} p(y|\theta^{(s)})}{p(\theta^{(s)}) p(y|\theta^{(s)})} = p(\theta^{(s)})^{\alpha-1}$
- Likelihood scaling: $w_{\alpha_{lik}}^{(s)} = \frac{p(\theta^{(s)}) p(y|\theta^{(s)})^{\alpha}}{p(\theta^{(s)}) p(y|\theta^{(s)})} = p(y|\theta^{(s)})^{\alpha-1}$
- Use weights for expectations, ECDFs or resampling

Improving estimates

- Pareto-smoothed importance sampling (PSIS)[2] to stabilise and provide reliable diagnostic

[2] A. Vehtari, D. Simpson, A. Gelman, *et al.*, "Pareto Smoothed Importance Sampling," arXiv: 1507.02646 [stat]. (2021).
[3] T. Paananen, J. Piironen, P.-C. Bürkner, *et al.*, "Implicitly adaptive importance sampling," *Statistics and Computing*, vol. 31, p. 16, Feb. 9, 2021, ISSN: 1573-1375. DOI: 10.1007/s11222-020-09982-2.

Improving estimates

- Pareto-smoothed importance sampling (PSIS)[2] to stabilise and provide reliable diagnostic
- Importance weighted moment matching (IWMM)[3] to further improve (but more expensive)

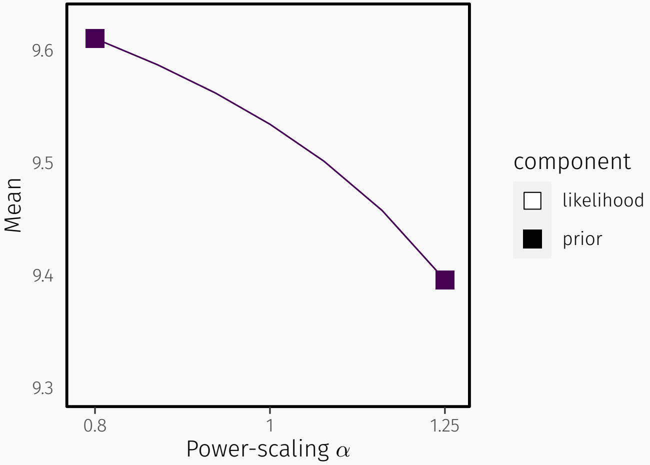
[2] A. Vehtari, D. Simpson, A. Gelman, *et al.*, "Pareto Smoothed Importance Sampling," arXiv: `1507.02646 [stat]`. (2021).

[3] T. Paananen, J. Piironen, P.-C. Bürkner, *et al.*, "Implicitly adaptive importance sampling," *Statistics and Computing*, vol. 31, p. 16, Feb. 9, 2021, ISSN: 1573-1375. DOI: `10.1007/s11222-020-09982-2`.

Sensitivity diagnostic

Change in quantities (e.g. mean) or divergences

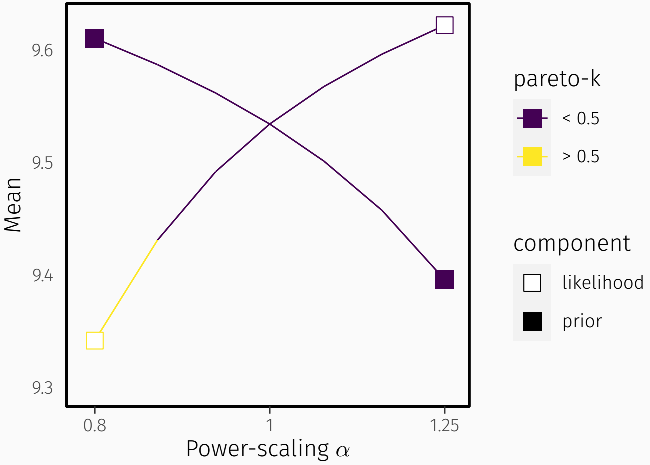
By default, we use the Cumulative Jensen-Shannon distance



Sensitivity diagnostic

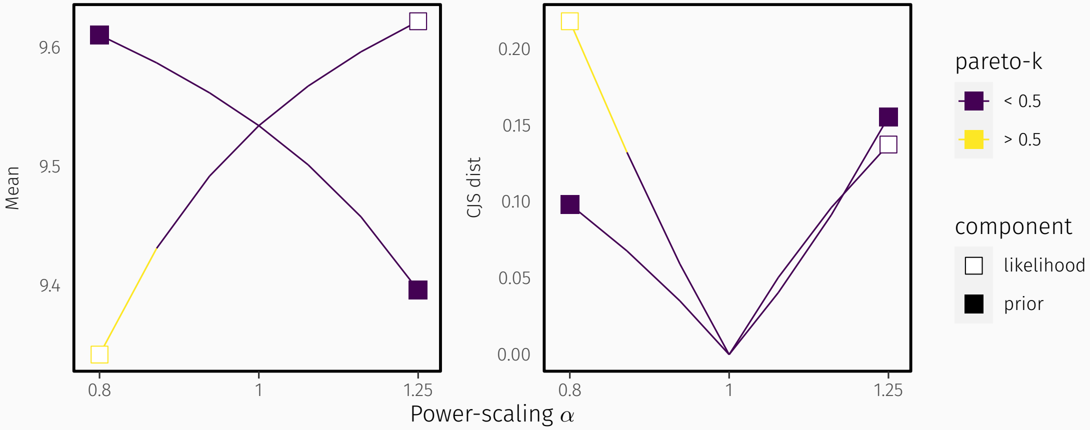
Change in quantities (e.g. mean) or divergences

By default, we use the Cumulative Jensen-Shannon distance



Change in quantities (e.g. mean) or divergences

By default, we use the Cumulative Jensen-Shannon distance



Diagnosing sensitivity

Prior sensitivity	Likelihood sensitivity	Possible cause

Diagnosing sensitivity

Prior sensitivity	Likelihood sensitivity	Possible cause
no	yes	diffuse/weak prior

Diagnosing sensitivity

Prior sensitivity	Likelihood sensitivity	Possible cause
no	yes	diffuse/weak prior
yes	no	weak likelihood

Diagnosing sensitivity

Prior sensitivity	Likelihood sensitivity	Possible cause
no	yes	diffuse/weak prior
yes	no	weak likelihood
yes	yes	prior-data conflict

Case studies

Case study: Body fat measurement

Approximate expensive bodyfat measurement from easier measurements[4]

Normal linear regression model

Original prior: $\beta^k \sim \mathsf{normal}(0, 1)$ without scaling covariates

[4] R. W. Johnson, "Fitting Percentage of Body Fat to Simple Body Measurements," *Journal of Statistics Education*, vol. 4, no. 1, p. 6, Mar. 1996, ISSN: 1069-1898. DOI: 10.1080/10691898.1996.11910505.

Case study: Body fat measurement

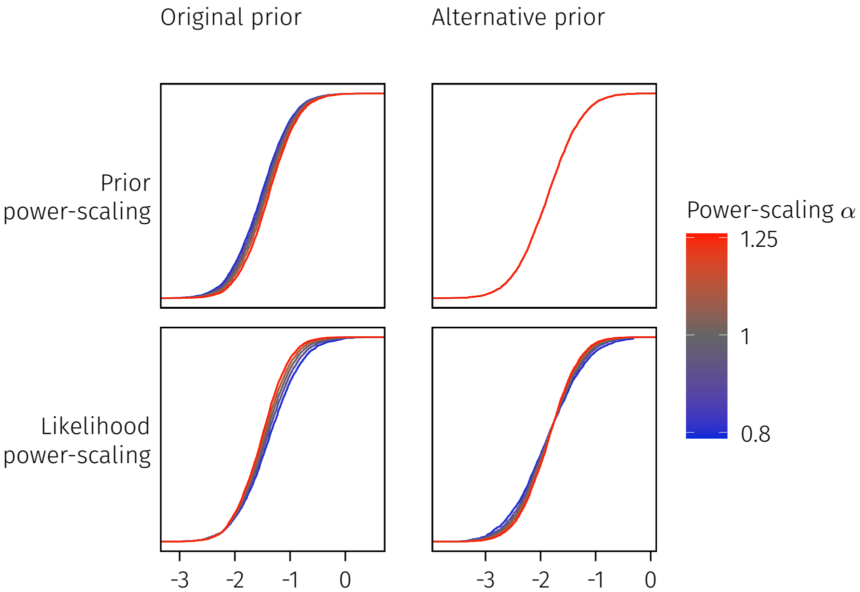
Parameter	Prior sensitivity	Likelihood sensitivity	Comment
β^{wrist}	**0.12**	0.09	prior-data conflict
β^{weight}	0.02	0.12	
β^{thigh}	0.01	0.08	
β^{neck}	0.01	0.11	
β^{knee}	0.01	0.1	
β^{hip}	0.01	0.11	
β^{height}	0.00	0.09	
β^{forearm}	0.02	0.12	
β^{chest}	0.01	0.08	
β^{biceps}	0.01	0.09	
β^{ankle}	0.02	0.1	
β^{age}	0.03	0.12	
β^{abdomen}	0.00	0.09	

Case study: Body fat measurement

Alternative data-scaled prior $\beta^k \sim \mathrm{normal}(0, 2.5 s_y / s_x)$

Case study: Body fat measurement

Alternative data-scaled prior $\beta^k \sim \text{normal}(0, 2.5 s_y / s_x)$



Case study: Motorcycle crash

Modelling head acceleration during a motorcycle crash[5].

$$\begin{aligned} y &\sim \mathrm{normal}(f(x), \exp(g(x))) \\ f &\sim \mathrm{GP} \\ g &\sim \mathrm{GP} \end{aligned}$$

Priors are specified on GP parameters (length-scale and magnitude)

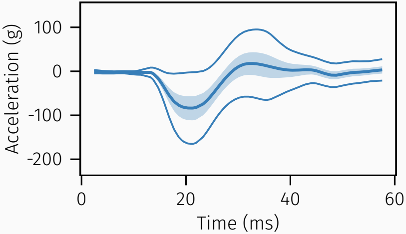
[5] B. W. Silverman, “Some aspects of the spline smoothing approach to non-parametric regression curve fitting,” *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 47, no. 1, pp. 1–21, Sep. 1985. DOI: `10.1111/j.2517-6161.1985.tb01327.x`.

Case study: Motorcycle crash

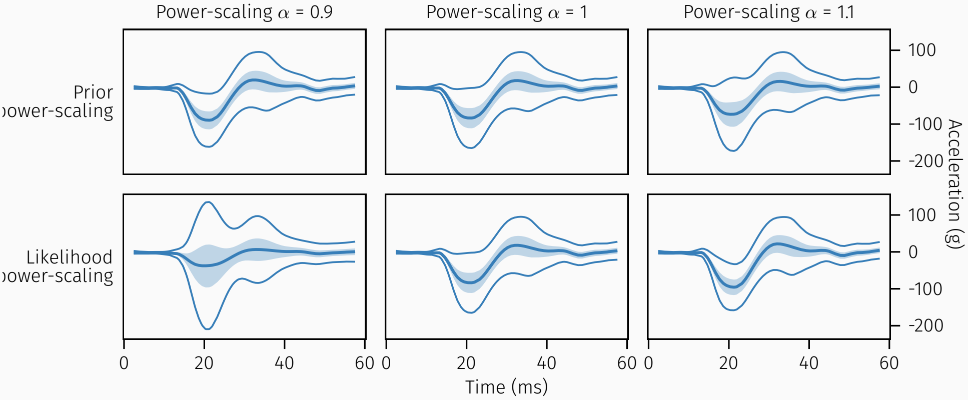
Modelling head acceleration during a motorcycle crash[5].

$$
\begin{aligned}
y &\sim \mathrm{normal}(f(x), \exp(g(x))) \\
f &\sim \mathrm{GP} \\
g &\sim \mathrm{GP}
\end{aligned}
$$

Priors are specified on GP parameters (length-scale and magnitude)

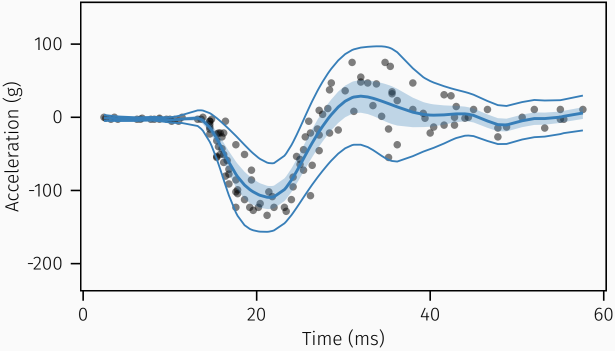


[5] B. W. Silverman, "Some aspects of the spline smoothing approach to non-parametric regression curve fitting," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 47, no. 1, pp. 1–21, Sep. 1985. DOI: `10.1111/j.2517-6161.1985.tb01327.x`.



Power-scaling α = 0.9
Power-scaling α = 1
Power-scaling α = 1.1
Prior
power-scaling
Likelihood
power-scaling
100
0
-100
-200
Acceleration (g)
0
20
40
60
Time (ms)

With a wider prior on the length-scale



Conclusions

Limitations and cautions

- Power-scaling is a single and very specific type of perturbation

Limitations and cautions

- Power-scaling is a single and very specific type of perturbation
- Sensitivity of correlated parameters can be difficult to interpret

Limitations and cautions

- Power-scaling is a single and very specific type of perturbation
- Sensitivity of correlated parameters can be difficult to interpret
- Sensitivity is not inherently problematic:
- Carefully consider how you address sensitivity
- Do not repeatedly adjust priors just to remove warnings

Summary

- Power-scaling is an intuitive perturbation of a distribution
- PSIS and IWMM enable efficient computation
- Can detect prior-data conflict or weak likelihood
- Fits into a modern Bayesian (MCMC-based) workflow with minimal adjustments required

Summary

- Power-scaling is an intuitive perturbation of a distribution
- PSIS and IWMM enable efficient computation
- Can detect prior-data conflict or weak likelihood
- Fits into a modern Bayesian (MCMC-based) workflow with minimal adjustments required

- Pre-print: `arxiv:2107.14054`
- `priorsense` R package: `https://github.com/n-kall/priorsense`
- Python version (ArviZ) in development