Bayesian model calibration using preposterior generalized cross-validation

Spencer Woody[1] Novin Ghaffari[1] Lauren Hund[2]

[1]The University of Texas at Austin

[2]Sandia National Laboratories

September 14, 2018

Summary

- Address the problem of *physical parameter identifiability* by avoiding discrepancy terms
- Account for model discrepancy via *power likelihood*,

$$\pi_w(\theta \mid y) \propto [f(y \mid \theta)]^w \pi(\theta)$$

- Use preposterior generalized cross-validation (GCV) to select w by credible interval quantile matching

Problem set-up

(Brown and Hund, 2018)

Consider a *dynamic materials experiment*

- Apply a boundary condition to a system
- Measure a functional output
- Calibrate model input parameters

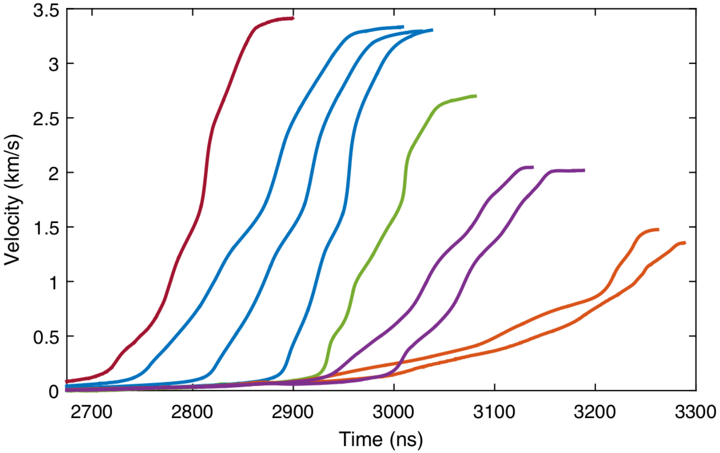
Here we study the compressibility properties of **tantalum** ($_{73}$Ta) by applying to it a dynamic magnetic field

The data and parameters

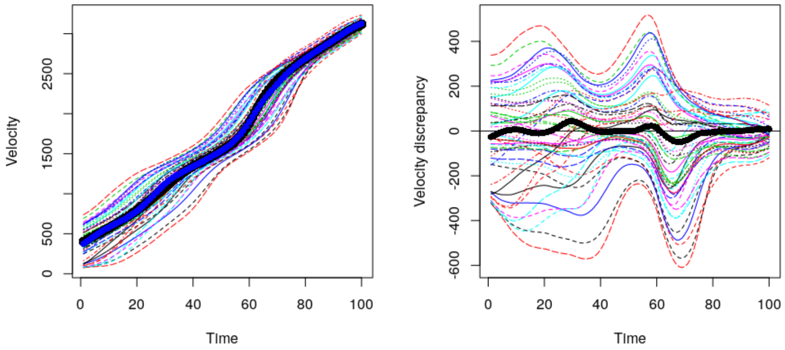
For experiments $j = 1, \ldots, 9$,

- $y(x_{ij})$: Observed time-velocity functional response resulting from impulse
- $\eta(x_{ij};\theta)$: Output of wave code simulator to model velocity
- θ: Calibration parameters for the simulator

The data and parameters



The data and parameters



The dark blue line is the observed $y(x)$

Other lines represent model output $\eta(x;\theta)$ for different θ

Applications of BMC

- Prediction (interpolation or extrapolation)
- Assess level of misspecification in computer models
- **Physical parameter estimation**

Inferential goal

Inverse problem: back out $\theta = (\alpha, \gamma)$ calibration parameters using observed data

- $\alpha = (B_0, B_0')$ are *parameters of interest*, bulk modulus B_0 and its pressure derivative B_0'
- $\gamma = (\rho_0, x_{\text{Al}}, x_{\text{Ta}}, BC_{\text{scale}})$ are *nuisance parameters*

given that $\eta(x_{ij}, \theta)$ **cannot perfectly describe** $y(x_{ij})$

Only B_0, B_0', and ρ_0 are common to all experiments j

GP discrepancy term

Kennedy and O'Hagan (2002) [KOH]

$$y(x_{ij}) = \eta(x_{ij};\theta) + \delta(x_{ij}) + \epsilon(x_{ij})$$

- GP discrepancy with squared-exponential kernel

$$\delta(x_j) \sim \mathcal{N}(0, \Sigma_j^{\delta})$$
$$\Sigma_j^{\delta}[i,i'] = \tau_1^2 \exp\left[-\frac{1}{2\tau_2^2}(x_{i'j} - x_{ij})^2\right]$$

- Observation error, for known $\Sigma_j^{\epsilon} = \text{diag}(\{\sigma_{ij}^2\}_{i=1}^n)$

$$\epsilon(x_j) \sim \mathcal{N}(0, \Sigma_j^{\epsilon})$$

- Specifying prior $\pi(\theta)$ leads to posterior $\pi(\theta|y_j) \propto f(y_j|\theta)\pi(\theta)$ using MVN likelihood

GP discrepancy term

- Requires $\mathcal{O}(n^3)$ computation for covariance matrix inversion
- Has potential for numerical instability
- *Difficult or impossible to jointly identify θ and $\delta(x)$*
  - Exception if we have strong prior information on $\delta(x)$ (Brynjarsdóttir and O'Hagan, 2014)

Scaling likelihood by effective sample size

Brown and Hund (2018)

Drop the discrepancy term, so now the model is

$$y(x_{ij}) \mid \theta, \phi_j \stackrel{\text{iid}}{\sim} \mathcal{N}(\eta(x_{ij};\theta), \phi_j)$$
$$\pi(\theta, \phi_j) \propto \pi(\theta) \cdot \phi_j^{-1},$$

and the sampling model is known to be misspecified.

Scaling by effective sample size

Issues:

- Model misspecification manifested through *autocorrelation of residuals*
- Because of functional nature of output, n may be chosen to be *arbitrarily large*

Solution: scale likelihood effective sample size (ESS), n_{ej} via

$$\pi(\theta|y_j) \propto \left[\prod_{i=1}^{n} f(y_{ij}|\theta, \phi_j)\right]^{n_{ej}/n} \pi(\theta),$$

calculated from autocorrelation of residuals τ_j by $n_{ej}/n = 1/\tau_j$.

On finding the ESS

1) Find MLE $\hat{\theta} = \arg\min_\theta \|\eta(x_j;\theta) - y(x_j)\|_2$
2) Calculate resulting empirical discrepancy $\hat{\delta}(x_j) = y(x_j) - \eta(x_j;\hat{\theta})$
3) Calculate autocorrelation $\hat{\tau}_j$ from $\hat{\delta}(x_j)$
4) $\hat{n}_{ej} = n/\hat{\tau}_j$

Uses concept of "information gain," similar to Holmes and Walker (2017) and Lyddon, Holmes, and Walker (2017).

This approach is used to retain correct variance in posterior

Generalized Bayesian posterior

Traditional Bayesian inference relies on the concept of well-specified models, which may be difficult, impossible, or inconvenient.

Much recent work has been conducted on power-likelihood methods, yielding the *generalized Bayesian posterior*, for $0 \leq w \leq 1$,

$$\pi_w(\theta \mid y) = \frac{[f(y \mid \theta)]^w \pi(\theta)}{\int [f(y \mid \theta)]^w \pi(\theta) \mathrm{d}\theta}$$

Previously, Brown and Hund (2018) used $w = n_{ej}/n$
How else can we find w?

Power likelihood

Bissiri, Holmes, and Walker (2016) use a loss function $l(\theta, x)$ to connect observations x with parameters θ

They argue via a decision theory approach that “a valid and coherent update” to $\pi(\theta)$ exists in the posterior of the form

$$\pi_w(\theta \mid y) \propto \exp(-wl(\theta, x))\pi(\theta).$$

Power likelihood

$$\pi_w(\theta \mid y) \propto [f(y \mid \theta)]^w \pi(\theta)$$

- Miller and Dunson (2018) determin w by assigning a prior to the KL-divergence between the "idealized data" $X_{1:n}$ under the misspecified model and the observed data $x_{1:n}$, $d_n(X_{1:n}, x_{1:n})$
- Grünwald and van Ommen (2017) select w to minimize posterior expected log-loss using a leave-one-out cross-validation method
- Syring and Martin (2018) use coverage based on bootstrap resampling to tune credible intervals to have nominal frequentist coverage rates

Power likelihood

Main idea of our method:

Let $C_{w,\alpha}(y)$ represent an equal-tailed $1-\alpha$-level posterior credible interval for θ coming from $\pi_w(\theta \mid y) \propto [f(y \mid \theta)]^w \pi(\theta)$.

Select the power to be $w^\star$, such that

$$\Pr(\theta \in C_{w^\star,\alpha}(y)) = 1-\alpha$$

Toy example

Prior:

$$\theta \sim \mathcal{N}(0,1)$$

Sampling model:

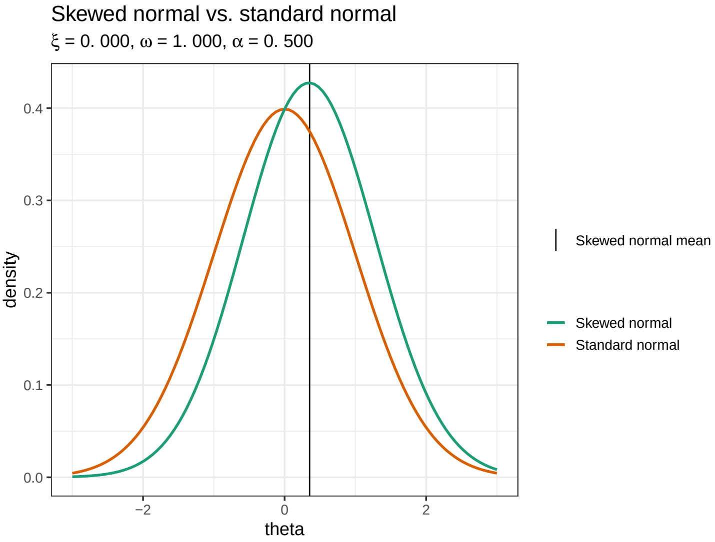
$$(y_i \mid \theta) \sim \mathcal{N}(\theta,1)$$

Reality:

$$(y_i \mid \theta) \sim \mathcal{SN}(\xi = \theta, \omega = 1, \alpha = 1/2),$$

i.e. the skewed-normal with location θ, scale 1, and shape 1/2 (mean $\approx$ 0.353, sd $\approx$ 0.933)

Toy example


Skewed normal vs. standard normal
ξ = 0. 000, ω = 1. 000, α = 0. 500
density
0.4
0.3
0.2
0.1
0.0
−2
0
2
theta
Skewed normal mean
Skewed normal
Standard normal

Toy example

Take a sample $y_i \sim \mathcal{N}(\theta, 1)$ of size $N = 25$

Generalized posterior:

$$\pi_w(\theta \mid y) \propto \prod_{i=1}^{n} [\mathcal{N}(y; \theta, 1)]^w \pi(\theta)$$

Question: how much do we need to discount the likelihood (i.e. what to set w) to achieve *nominal frequentist coverage* of posterior credible intervals?

Toy example

For $k = 1, \ldots, K$

- Generate $\theta_k \sim \mathcal{N}(0, 1)$ for
- For each k, generate $y_{ki} \sim \mathcal{SN}(\theta_k, 1, 1/2)$, $i = 1, \ldots, N$
- For each w on some grid, draw samples from $\pi_w(\theta \mid y)$, using misspecified normal likelihood
- Check whether a 90% credible interval covers θ

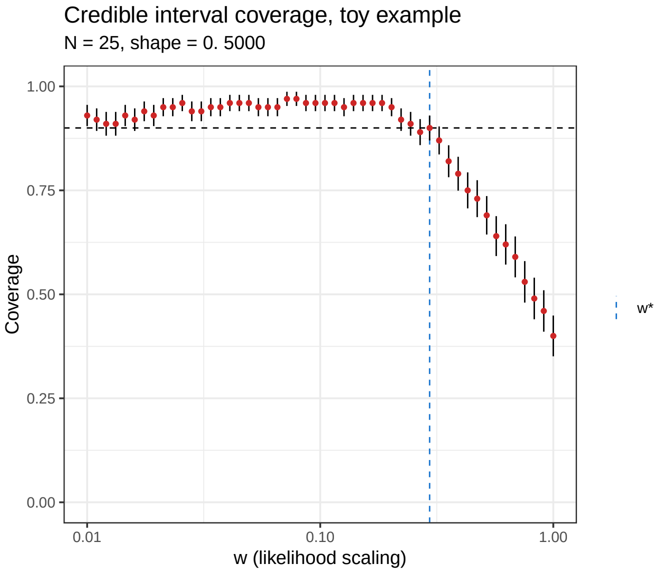
Get a Monte Carlo estimate of coverage probability for each w

Toy example

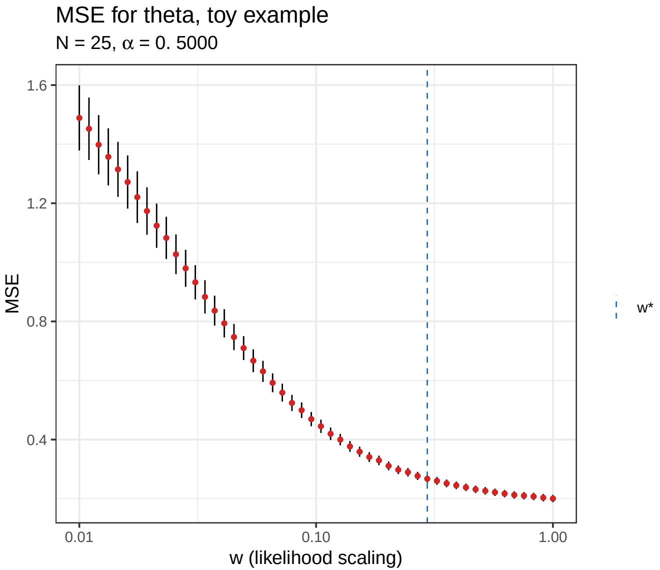
Consider these cases:

- $w = 0 \Rightarrow \pi_w(\theta \mid y) = \pi(\theta)$, (no update) $\Rightarrow$ trivially have nominal coverage rate
- Small positive $w \Rightarrow$ close to prior but favoring observed mean $\Rightarrow$ coverage rate too high
- $w = 1 \Rightarrow \pi_w(\theta \mid y)$ concentrates on (biased) empirical mean $\Rightarrow$ low coverage

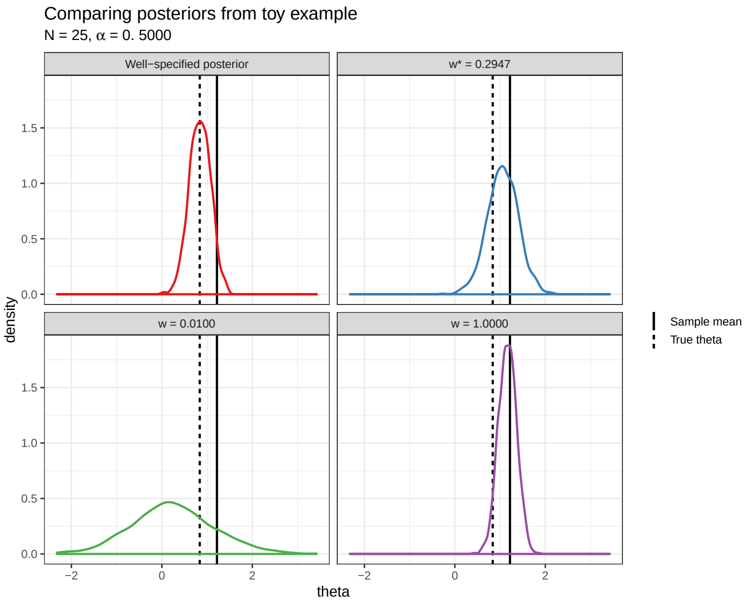
Toy example


Credible interval coverage, toy example
N = 25, shape = 0. 5000
1.00
0.75
0.50
0.25
0.00
Coverage
0.01
0.10
1.00
w (likelihood scaling)
w*

Toy example


MSE for theta, toy example
N = 25, α = 0. 5000
1.6
1.2
0.8
0.4
MSE
0.01
0.10
1.00
w (likelihood scaling)
w*

Toy example


Comparing posteriors from toy example
N = 25, α = 0. 5000
Well−specified posterior
w* = 0.2947
w = 0.0100
w = 1.0000
density
theta
0.0
0.5
1.0
1.5
−2
0
2
Sample mean
True theta

Generalized cross-validation

Key idea: choose w such that we believe the credible intervals to have nominal frequentist coverage.

Preposterior analysis

Preliminaries: empirical estimate of discrepancy

- Find MLE of θ,

$$\hat{\theta} = \arg\min_{\theta} \|y(x_j) - \eta(x_j; \theta)\|_2$$

- Find MLE hyperparameters $\hat{\tau}_1^2$ and $\hat{\tau}_2^2$ for resulting empirical discrepancy,

$$\hat{\delta}(x_j) = y(x_j) - \eta(x_j; \hat{\theta})$$

These estimates will generally be different for each experiment

Preposterior analysis

Pseudodata generation:

- Generate $\tilde{\theta} \sim \pi(\tilde{\theta})$, calculate *pseudotruth* $\eta(x;\tilde{\theta})$
- Generate *pseudodiscrepancy* [1]

$$\tilde{\delta}(x) \sim \mathcal{N}(0, \hat{\Sigma}_\delta),$$

  with $\hat{\Sigma}_\delta$ from estimated GP hyperparameters $\hat{\tau}_1^2$ and $\hat{\tau}_2^2$
- Calculate *pseudoexperimental* data,

$$\tilde{y}(x) = \eta(x;\tilde{\theta}) + \tilde{\delta}(x)$$

Similar to Arendt et al. (2016), who use priors for GP parameters instead of estimating them

[1] NB: Hats $\hat{\cdot}$ represent estimates, tildes $\tilde{\cdot}$ represent simulated data

Preposterior analysis

For one value of w,

(i) Generate pseudodata $\tilde{y}(x)$ for one $\tilde{\theta} \sim \pi(\tilde{\theta})$
(ii) Sample from GBP $\pi_w(\theta|\tilde{y})$
(iii) Check for coverage of $\tilde{\theta}$ in the credible interval $C_{w,\alpha}(\tilde{y})$ using draws from GBP
(iv) Repeat for many $\tilde{\theta} \sim \pi(\tilde{\theta})$

This yields Monte Carlo estimates of frequentist coverage probabilities

Repeat this procedure along a grid of w

Preposterior analysis

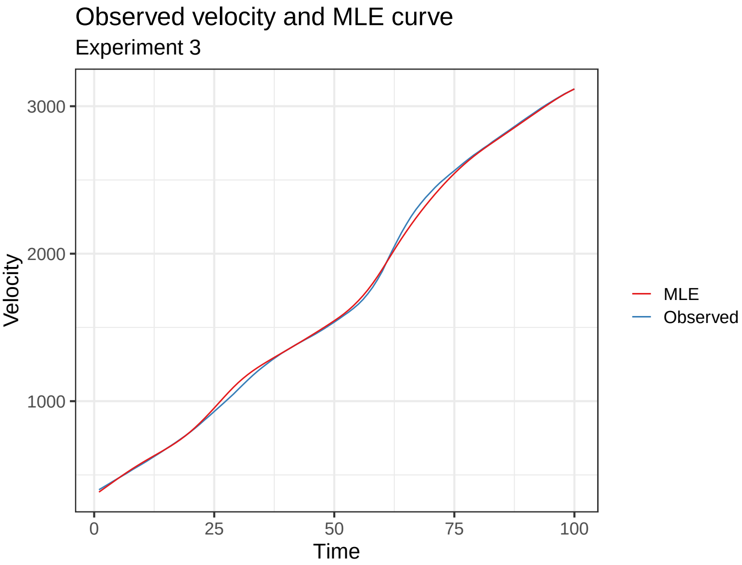
We can also consider other cross-validation metrics

- MSE of estimating θ
- MSE of estimating $\eta(x;\tilde{\theta})$
- Posterior predictive coverage of $\tilde{y}(x)$

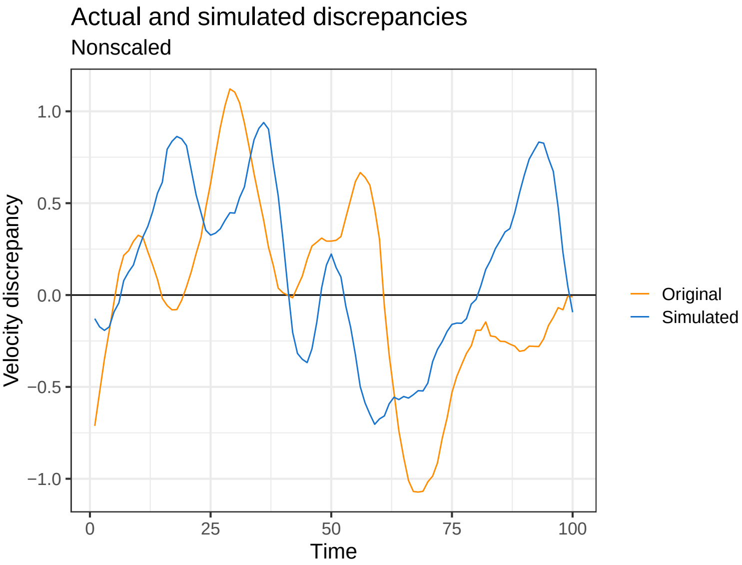
Preposterior analysis

Importantly, discrepancy function used in selecting w, but otherwise not included in the posterior for θ (avoid identifiability issue)

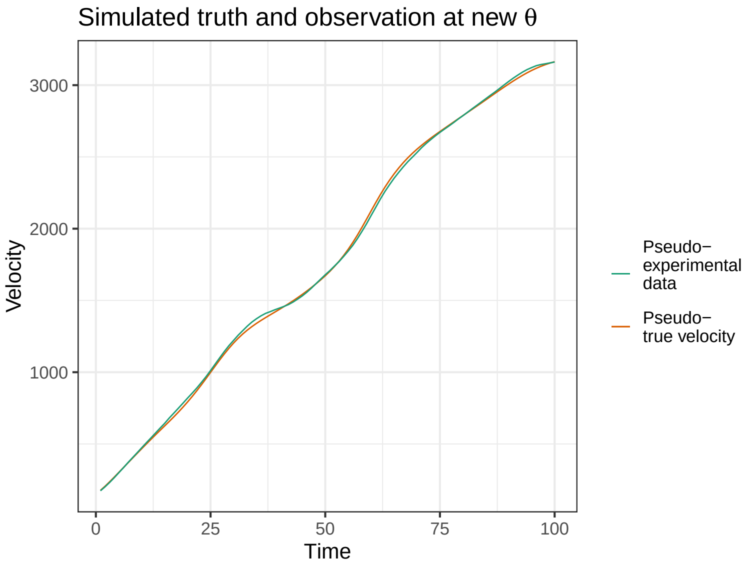
Example: one pseudo dataset



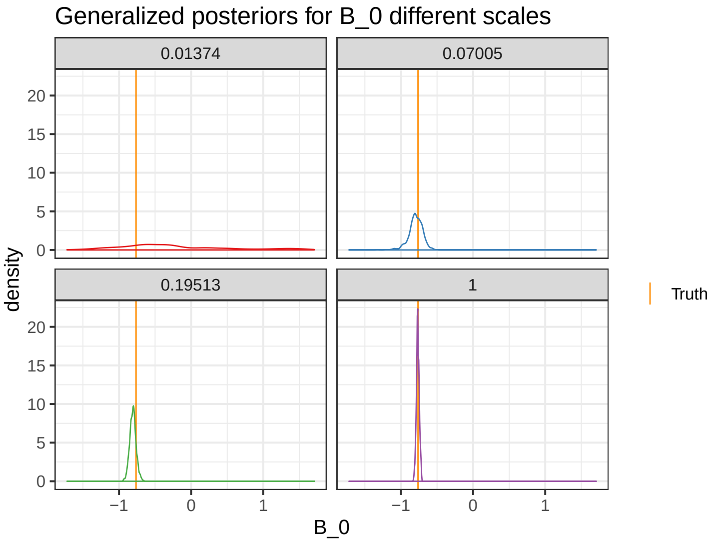
Example: one pseudo dataset



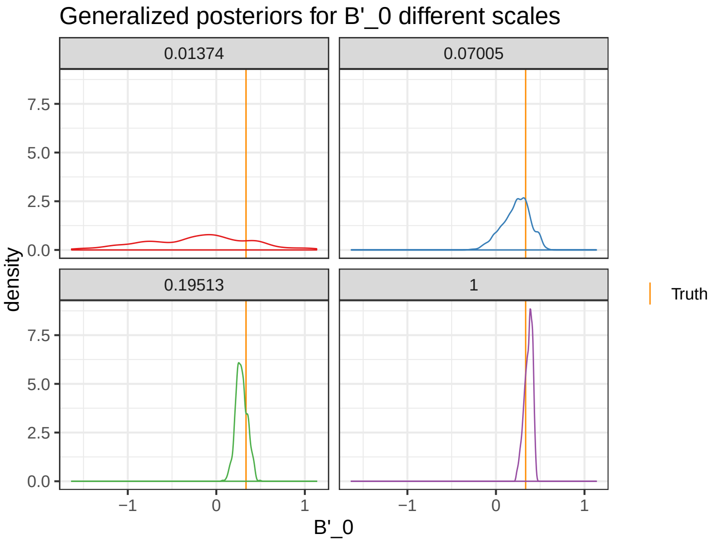
Example: one pseudo dataset



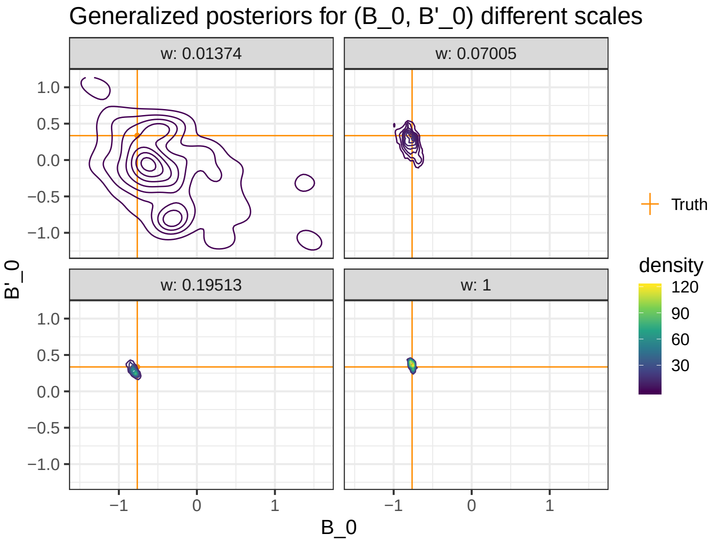
Example: one pseudo dataset



Example: one pseudo dataset



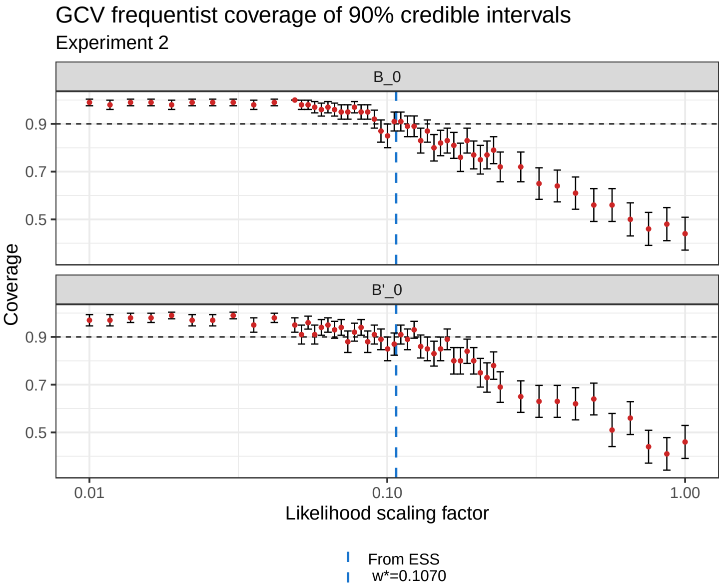
Example: one pseudo dataset



Results on one experiment

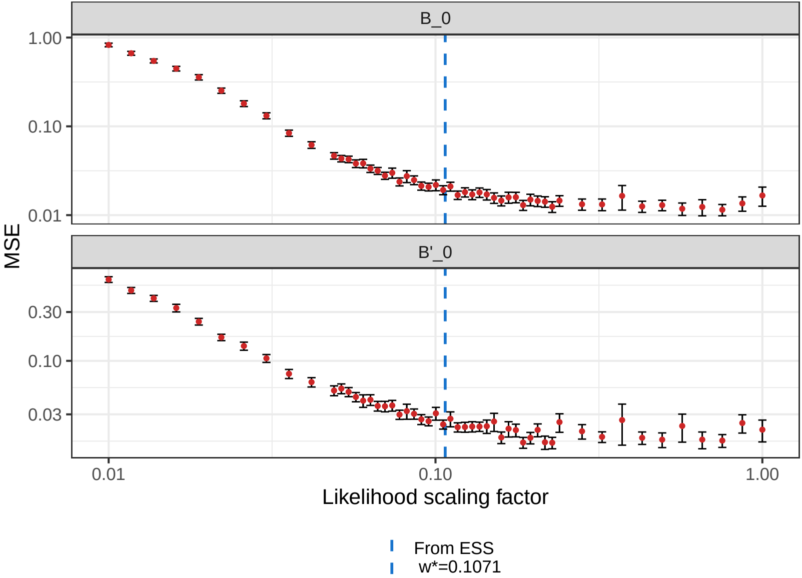
- 100 instantiations of $\tilde{\theta} \sim \pi(\tilde{\theta})$
- Evaluate coverage estimate on a grid of w on a log-scale
- Compare optimal $w^\star$ to scaling factor chosen by Brown and Hund (2018)

Results on one experiment



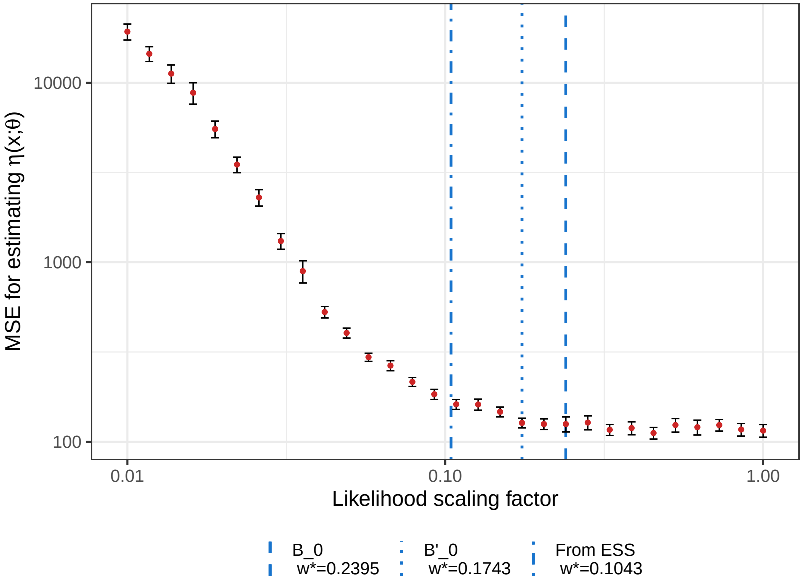
GCV for selecting likelihood scaling

Experiment 2



GCV on η(x;θ) error for selecting likelihood scaling

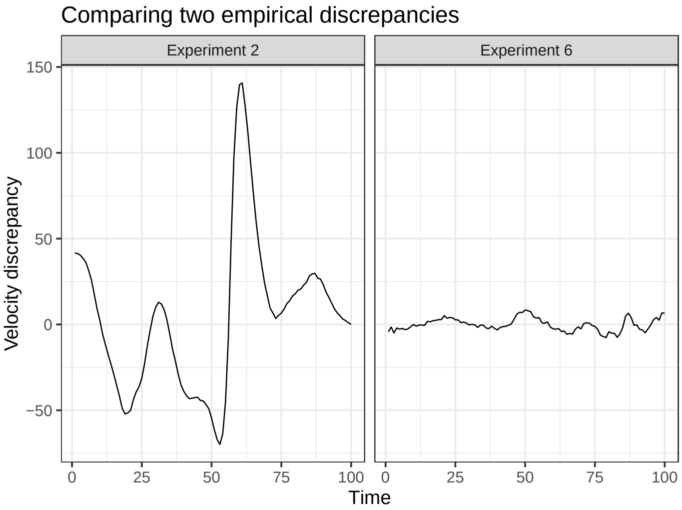
Experiment 3



Results

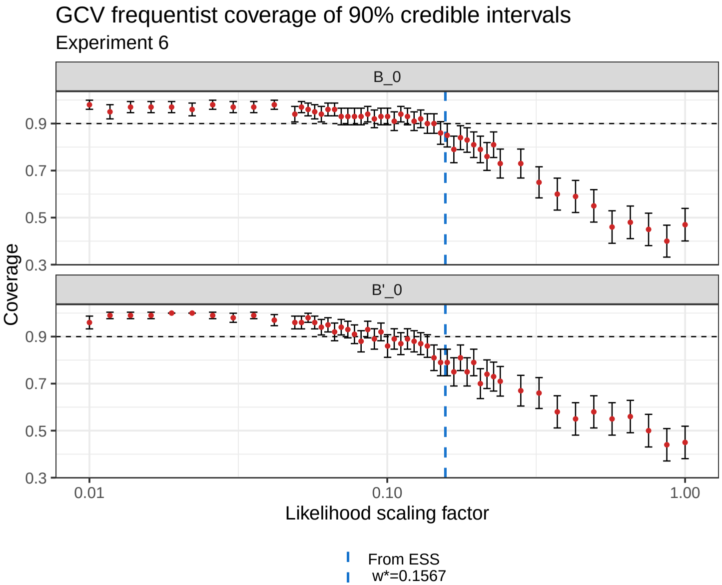
Is there general agreement with Brown and Hund (2018)?

Results



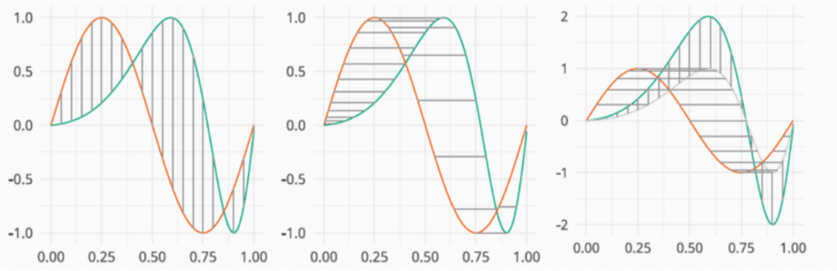
$ESS_2 = 10.70$, $ESS_6 = 15.67$

Results for another experiment



Other considerations

We can consider other ways of generating $\tilde{\delta}$, and generating $\tilde{\theta}$



Measuring phase variability using vertical distance.

Measuring phase variability using horizontal distance.

Need to measure both vertical and horizontal distance with both phase and amplitude.

Open questions

- Can we show that a solution $w^\star$ exists?
- Does $w^\star$ selected with this method scale with n?
- Show agreement with Brown and Hund (2018)?
- Can we extend this to predictive interval evaluation?
- What about extrapolation to other settings?

Conclusion

Email: `spencer.woody@utexas.edu`

References I

Paul D. Arendt, Daniel W. Apley, and Wei Chen. A preposterior analysis to predict identifiability in the experimental calibration of computer models. *IIE Transactions*, 48(1):75–88, 2016. doi: 10.1080/0740817X.2015.1064554. URL https://doi.org/10.1080/0740817X.2015.1064554.

P. G. Bissiri, C. C. Holmes, and S. G. Walker. A general framework for updating belief distributions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(5):1103–1130, 2016. doi: 10.1111/rssb.12158. URL https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/rssb.12158.

J. L. Brown and L. B. Hund. Estimating material properties under extreme conditions by using Bayesian model calibration with functional outputs. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 67(4): 1023–1045, 2018. doi: 10.1111/rssc.12273. URL https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/rssc.12273.

References II

Jenn Brynjarsdóttir and Anthony O'Hagan. Learning about physical parameters: The importance of model discrepancy. *Inverse Problems*, 30, 11 2014.

Peter Grünwald and Thijs van Ommen. Inconsistency of bayesian inference for misspecified linear models, and a proposal for repairing it. *Bayesian Anal.*, 12(4):1069–1103, 12 2017. doi: 10.1214/17-BA1085. URL https://doi.org/10.1214/17-BA1085.

Chris Holmes and Stephen Walker. Assigning a value to a power likelihood in a general bayesian model. *Biometrika*, 104, 01 2017.

Marc C. Kennedy and Anthony O'Hagan. Bayesian calibration of computer models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(3):425–464, 2002. doi: 10.1111/1467-9868.00294. URL https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/1467-9868.00294.

S. Lyddon, C. Holmes, and S. Walker. General Bayesian Updating and the Loss-Likelihood Bootstrap. *ArXiv e-prints*, September 2017.

References III

Jeffrey W. Miller and David B. Dunson. Robust Bayesian inference via coarsening. *Journal of the American Statistical Association*, 0(ja):1–31, 2018. doi: 10.1080/01621459.2018.1469995. URL https://doi.org/10.1080/01621459.2018.1469995.

N. Syring and R. Martin. Calibrating general posterior credible regions. *ArXiv e-prints*, September 2018.