Model inversion: Variational Bayes and Markov Chain Monte Carlo

Lionel Rigoux

Translational Neuro-Circuitry (TNC) Cologne

Translational Neuromodeling Unit (TNU) Zürich

Max-Planck-Institut für Stoffwechselforschung

Overview

Bayesian inference

Sampling methods

Variational methods

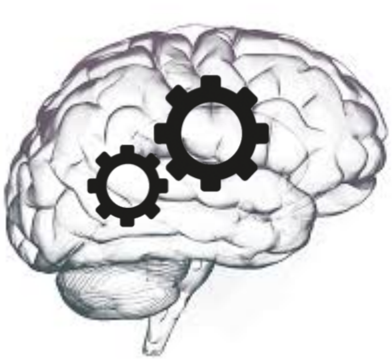
Model inversion

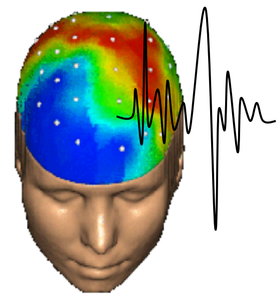
$\mathbf{p}(\mathbf{y}|\boldsymbol{\theta}, \mathbf{m})$

forward problem

generative model

$\boldsymbol{m}, \boldsymbol{\theta}$





observation

$\boldsymbol{y}$

inverse problem

$\mathbf{p}(\boldsymbol{\theta}|\mathbf{y})$ $\mathbf{p}(\boldsymbol{m}|\mathbf{y})$

Bayes rule

Linear regression

model

$$\mathbf{y} = \boldsymbol{\theta}\,\mathbf{x} + \boldsymbol{\varepsilon}$$
$$\boldsymbol{\varepsilon} = \mathcal{N}\left(0, \sigma_y^2\right)$$

prior

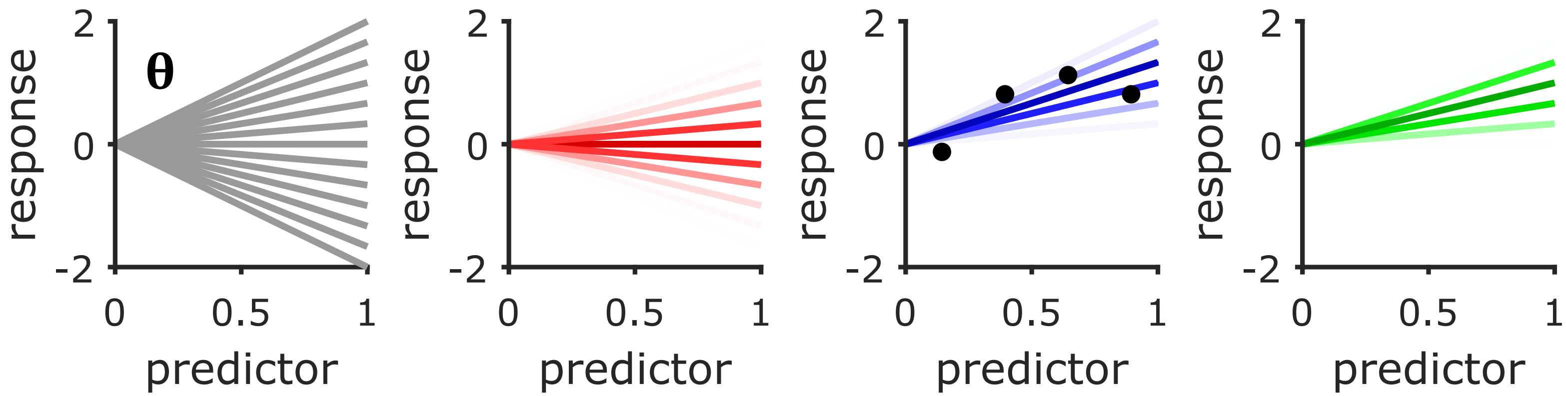
$$\mathbf{p}(\boldsymbol{\theta}) = \mathcal{N}\left(0, \sigma_p^2\right)$$

likelihood

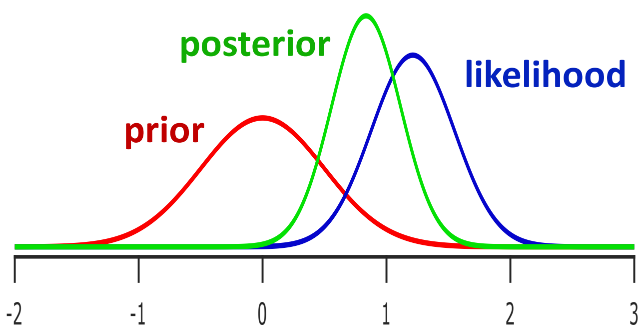
$$\mathbf{p}(\mathbf{y}|\boldsymbol{\theta}) = \mathcal{N}\left(\boldsymbol{\theta}\,\mathbf{x}, \sigma_y^2\right)$$

posterior

$$\mathbf{p}(\boldsymbol{\theta}|\mathbf{y})$$



$$\mathbf{p}(\boldsymbol{\theta}|\mathbf{y},\mathbf{m}) = \frac{\mathbf{p}(\boldsymbol{\theta}|\mathbf{m})\,\mathbf{p}(\mathbf{y}|\boldsymbol{\theta},\mathbf{m})}{\mathbf{p}(\mathbf{y}|\mathbf{m})}$$



Inference in practice

How to compute the posterior?

- Write $\mathbf{p(\theta|m)}$ $\mathbf{p(y|\theta, m)}$

1. recognize it looks like a know distribution (conjugacy)

$$\mathbf{p(\theta|m)} = \mathcal{N}(\mu_0, \sigma_0^2)$$
$$\mathbf{p(y|\theta, m)} = \mathcal{N}(\theta, \sigma_y^2)$$

$$\mathbf{p(\theta|y, m)} = \mathcal{N}(\mu', \sigma^{2\prime})$$
$$\mathbf{p(y|m)} = \text{[analytical solution]}$$

$$\sigma^{2\prime} = \left(\frac{1}{\sigma_0^2} + \frac{n}{\sigma_y^2}\right)^{-1} \qquad \mu' = \left(\sigma^{2\prime}\right)^{-1}\left(\frac{\mu_0}{\sigma_0^2} + \frac{\sum_n y_n}{\sigma_y^2}\right)$$

2. use variational Bayes or Monte Carlo methods

Monte-Carlo methods

If you can not calculate it,
simulate many random trials and see what happens...

Law of large numbers

- simulate many independent draws of a random variable
- the average of the results will converge to the true expected value (probability mean)

$$\mathbf{E[\theta] = \int p(\theta|y,m)\theta \approx \frac{1}{n}\sum_{n} \theta_n}$$

A little game

The un-normalized posterior:

$$p(\theta|y,m) \propto p(\theta|m)\, p(y|\theta,m) = \tilde{p}(\theta|y,m)$$

- is not a probability
- gives the relative plausibility of parameter values

Markov Chain sampling

Markov Chain: stochastic process that evolve in time

- initial state θ_0
- state evolve following a transition function $\mathrm{T}(\theta_{t+1}|\theta_t)$

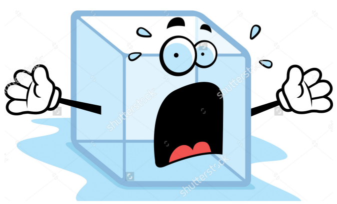
> In the long run the probability of visiting θ is called the *ergodic density*

Metropolis Hastings algorithm

The Metropolis-Hastings algorithm

- start form θ_0
- propose a new value according to $\mathrm{T}'(\theta'|\theta_t)$
- look for guidance

$$\mathbf{r} = \frac{\tilde{\mathbf{p}}(\boldsymbol{\theta}'|\mathbf{y},\mathbf{m})}{\tilde{\mathbf{p}}(\boldsymbol{\theta}_t|\mathbf{y},\mathbf{m})}$$



<1

>1

$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}'$

jump to proposed value

if $\mathbf{r} > \mathbf{X} \sim \mathbf{U}(\mathbf{0},\mathbf{1})$

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}'$$

else

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t$$

$$\boxed{\text{ergodic density} = \mathbf{p}(\boldsymbol{\theta}|\mathbf{y},\mathbf{m})}$$

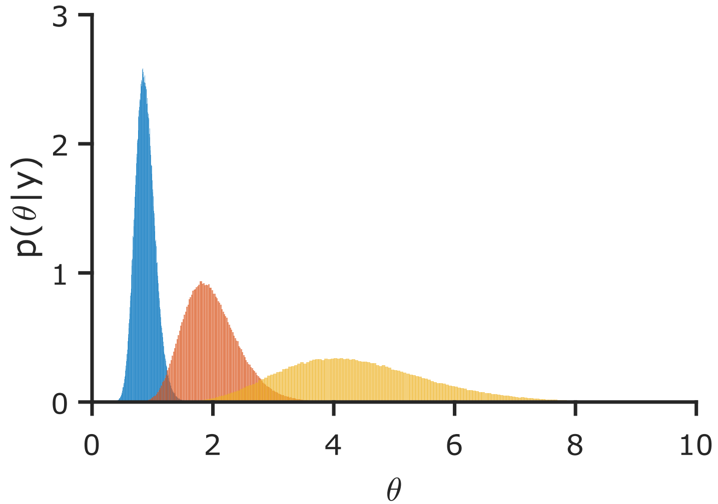
Metropolis Hastings algorithm: example

Logistic regression

$$\mathbf{p(y = 1|\theta) = s(\theta x)} \qquad \mathbf{s = \frac{1}{1 + e^{-x}}}$$

$$\mathbf{p(\theta) = \mathcal{N}(0, \sigma_0^2)}$$

$$\mathbf{\tilde{p}(\theta|y) = \exp\left(-\frac{\theta^2}{2\sigma_0^2}\right) \times \prod_y s(\theta x)^y \left(1 - s(\theta x)\right)^{1-y}}$$



Inference using sampling

Multivariate case:

- write conditional posteriors $\mathbf{p}(\boldsymbol{\theta}^1|\mathbf{y}, \boldsymbol{\theta}^2, \boldsymbol{\theta}^3, \dots, \boldsymbol{\theta}^n)$
- sample in turn $\boldsymbol{\theta}^1$, $\boldsymbol{\theta}^2$, etc.

Further tricks

- Gibbs sampling (conditional has a known form)
- Collapsing (marginalization)
- Blocking (multivariate sampling)
- Adaptive step, Langevin, …

Monte-Carlo inference

Sample in turn from all the conditional (Gibbs) or the un-normalized conditional (Markov chain/Metropolis-Hastings) posterior.

> Sufficient statistics converge to the true value.

Problems:

- computationally expensive
- does not scale well
- no direct measure of model evidence (Chib & Jeliazkov estimator)
- hard to tune and diagnose (jump size, burn in, sample autocorrelation, convergence, ...)

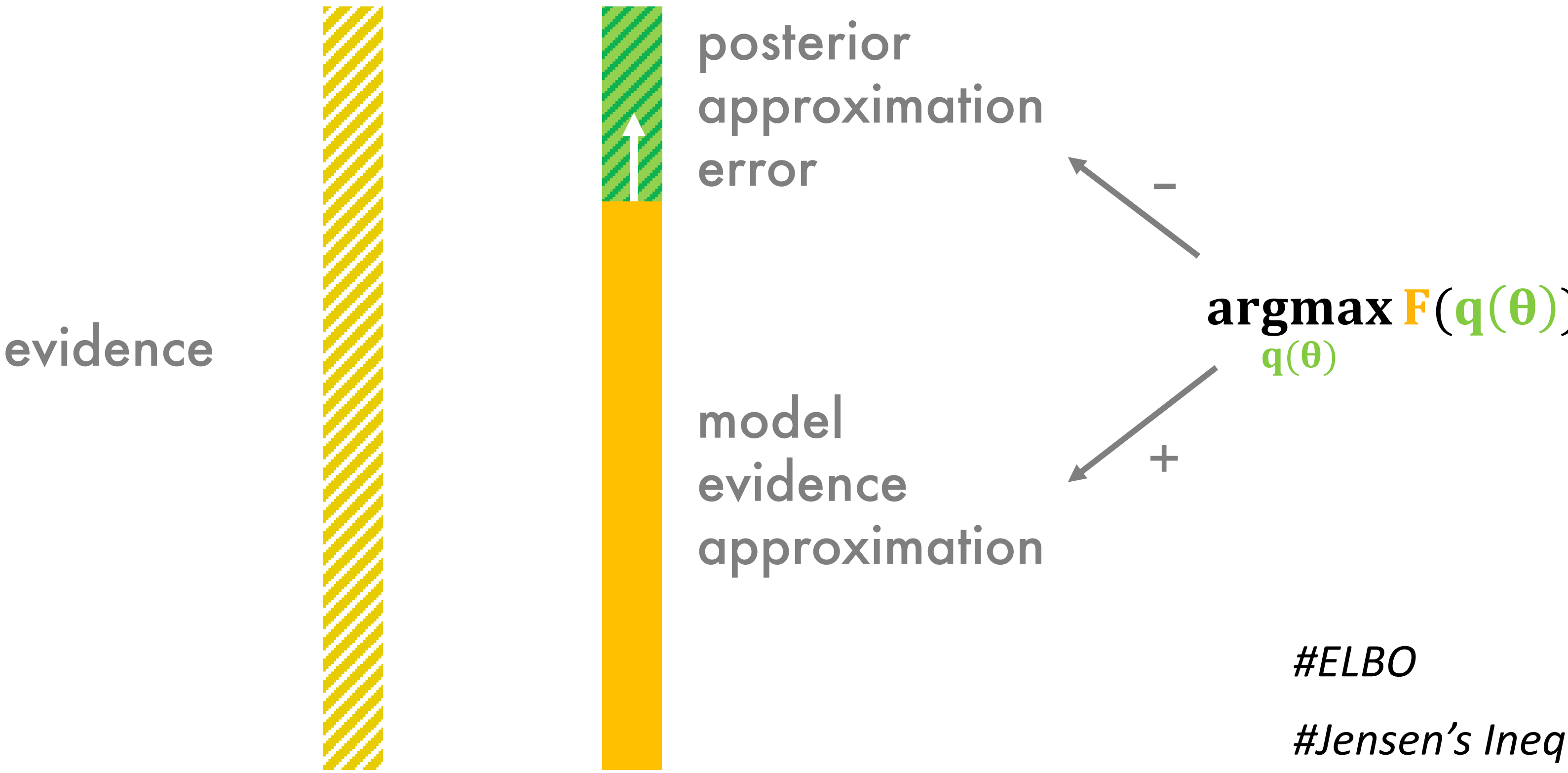
Variational inference

"variational inference is the thing you implement while you wait for your sampler to converge"

David Blei

Free Energy approximation

$$\log p(y|m) = F(q(\theta), y) + KL[q(\theta)||p(\theta|y, m)]$$

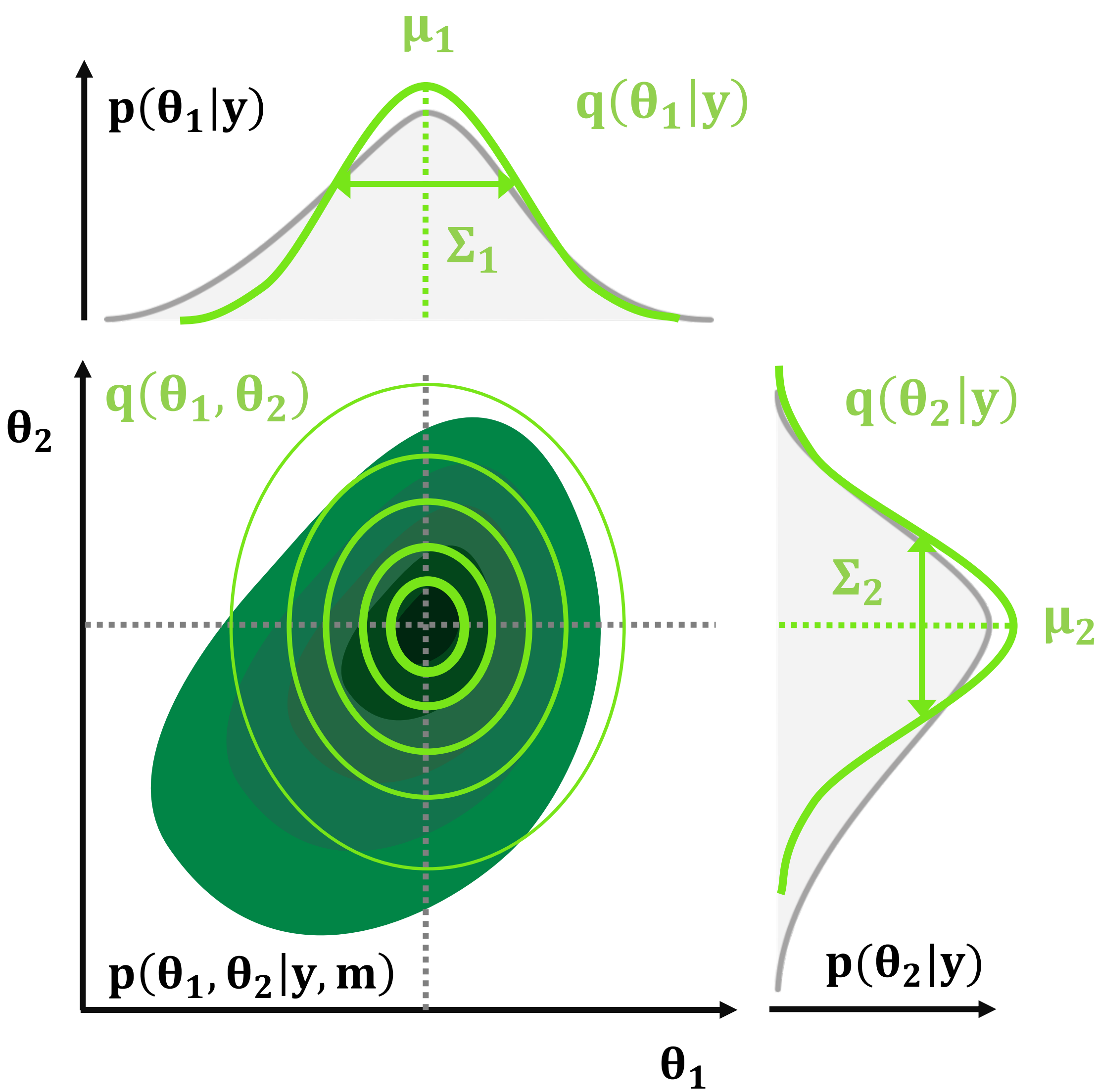


#ELBO

#Jensen's Inequality

#Free energy

Posterior approximation



Mean field approximation

$$p(\theta_1, \theta_2|y) \approx p(\theta_1|y)p(\theta_2|y)$$

Laplace approximation

$$q(\theta_i|y) \approx \mathcal{N}(\mu_i, \Sigma_i)$$

finding $p(\theta_1, \theta_2|y, m)$

$\downarrow$

finding $\mu_1, \mu_2, \Sigma_1, \Sigma_2$

Free Energy maximization

Approximating the model evidence = maximizing the ELBO wrt $\mathbf{q}(\boldsymbol{\theta})$

1) Maximize the free energy using variational calculus

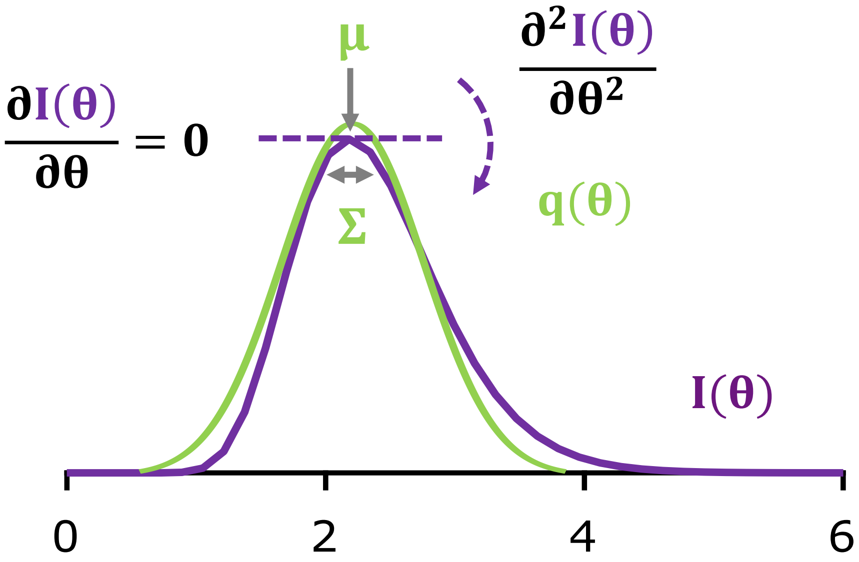
$$\mathbf{F} = \langle \log \mathbf{p}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) + \log \mathbf{p}(\mathbf{y}|\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \mathbf{m}) \rangle_{\mathbf{q}} + \langle \log \mathbf{q}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) \rangle_{\mathbf{q}}$$

$$\frac{\partial \mathbf{F}}{\partial \mathbf{q}(\boldsymbol{\theta}_1)} = \mathbf{0} \quad \Longrightarrow \mathbf{q}(\boldsymbol{\theta}_1) \propto \exp\Big(\underbrace{\langle \log \mathbf{p}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) + \log \mathbf{p}(\mathbf{y}|\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \mathbf{m}) \rangle_{\mathbf{q}(\boldsymbol{\theta}_2)}}_{\text{variational free energy } \mathbf{I}(\boldsymbol{\theta}_1)} \Big)$$

Posterior approximation

Find $\mathcal{N}(\mu, \Sigma)$ that best approximates $I(\theta)$

logistic regression



$$I(\theta) = \exp\left(-\frac{\theta^2}{2\sigma_0^2}\right) \times \prod_y s(\theta x)^y \left(1 - s(\theta x)\right)^{1-y}$$

multivariate case

Until convergence:

for all i:

- $\mu_i = \max_{\theta_i}\left(I(\theta_i)\right)$
- $\Sigma_i = -\left[\left.\frac{\partial^2}{\partial\theta_i^2}\right|_{\mu_i} I(\theta_i)\right]^{-1}$

end

end

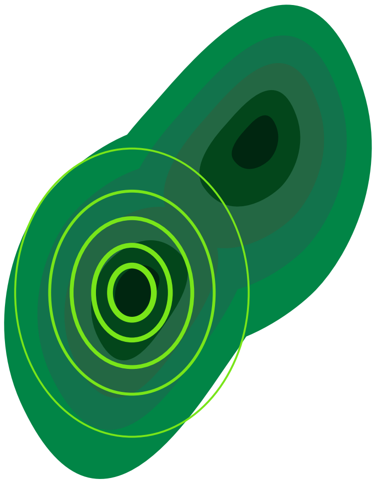
Variational inference

Summarize the posterior to its sufficient statistics (mean, variance) and optimize those values wrt the evidence lower bound.

This requires multiple approximations (free-energy, mean-field, Laplace) to be tractable.

Problems:

- does not converge to the true posterior
- can get stuck in local optimum



Bayesian inference methods: summary

Model evidence (normalization factor of the posterior) is in general intractable.

Sampling methods give a computationally expensive estimation of the true posterior.

Variational methods are fast and scalable but potentially inaccurate.

Software

Variational

- VBA-toolbox
- TAPAS
- SPM

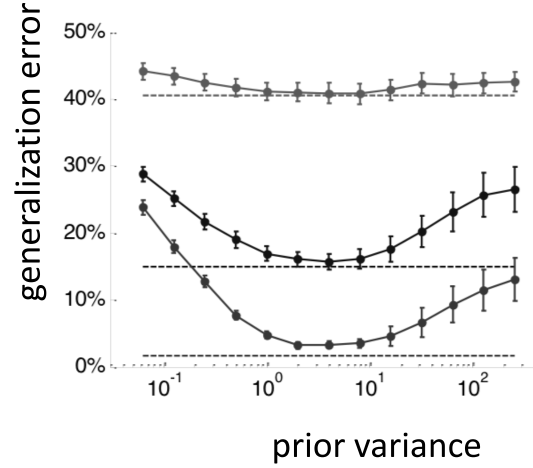
Sampling

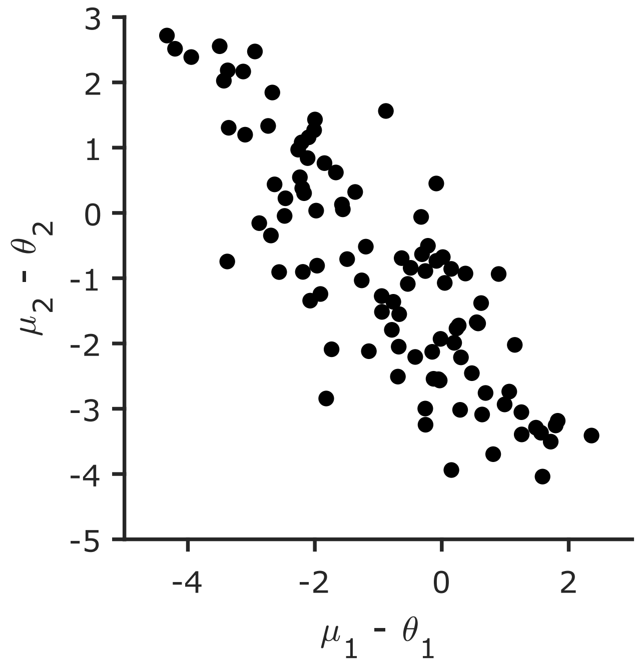
- STAN
- BUGS
- JAGS
- hBayesDM
- hddm

Validating your model: parameters

Checking if your parameters are identifiable:

- simulate data using your design with realistic θ
- invert your model (find μ)
- compute estimation error $(\mu - \theta)$
    - check effect of prior mean
    - check effect of prior variance
    - check for posterior / error correlation

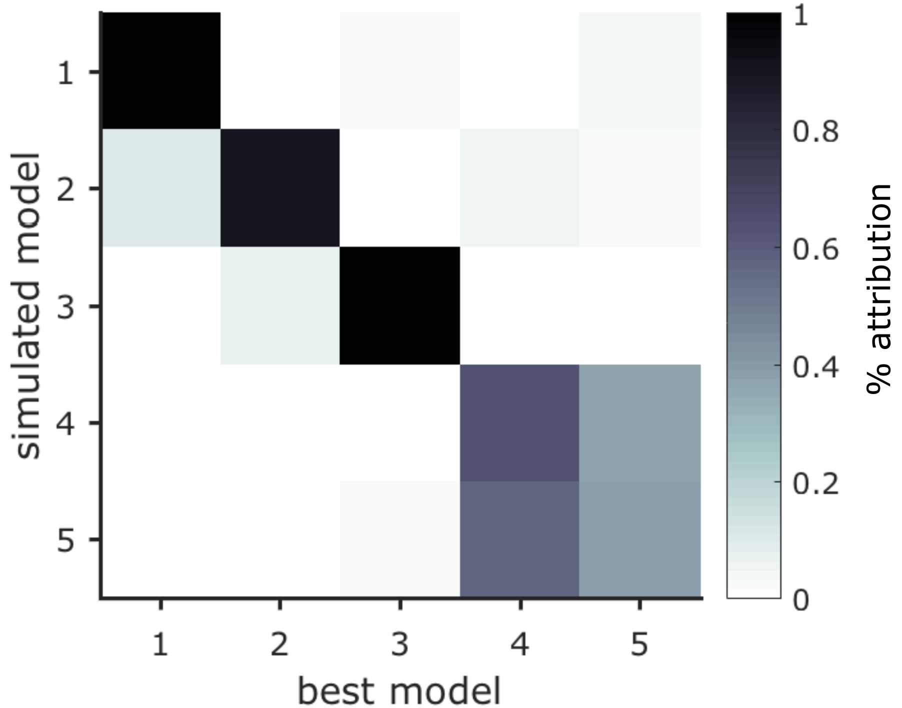




Validating your model: hypothesis identifiability

Checking if your models are identifiable:

- simulate all models
- compute evidence of each hypothesis for each dataset (BMS)
- count misattributions and build confusion matrix



Thank you!