# Bayesian inference and model inversion
# Variational + Markov chain Monte Carlo

Lionel Rigoux

Translational Neuro-Circuitry (TNC) Cologne

Translational Neuromodeling Unit (TNU) Zürich

Max-Planck-Institut
für Stoffwechselforschung

# Overview

Introduction to Bayesian inference (revisited)

Sampling methods (sampling)

Variational methods (approximation)

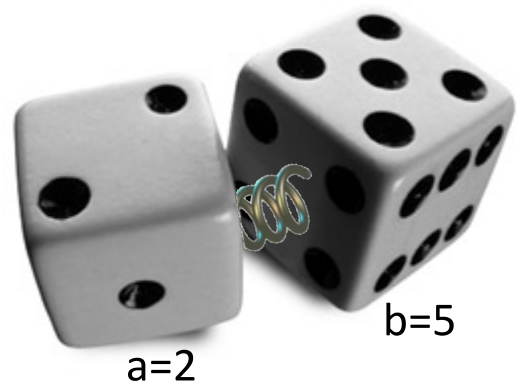
# Probability basics

Formalizes the degree of plausibility of events:

i. represented by real numbers
ii. should conform to intuition
iii. should be consistent


Normalization $$\sum_a p(a) = 1$$

Independence $$\mathbf{p}(\mathbf{a},\mathbf{b}) = \mathbf{p}(\mathbf{a})\mathbf{p}(\mathbf{b})$$




Conditioning $$\mathbf{p}(\mathbf{a},\mathbf{b}) = \mathbf{p}(\mathbf{a}|\mathbf{b})\mathbf{p}(\mathbf{b})$$

Marginalization $$\mathbf{p}(\mathbf{b}) = \sum_a p(a,b)$$

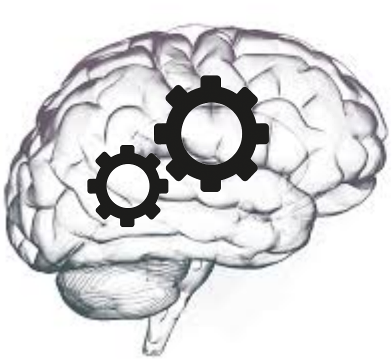
# Model inversion

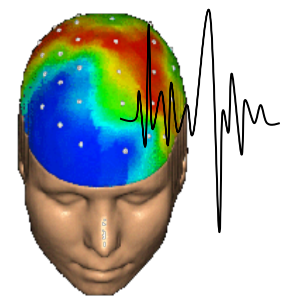
$$\mathbf{p}(\mathbf{y}|\boldsymbol{\theta},\mathbf{m})$$

forward problem

generative model

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





observation

$\boldsymbol{y}$

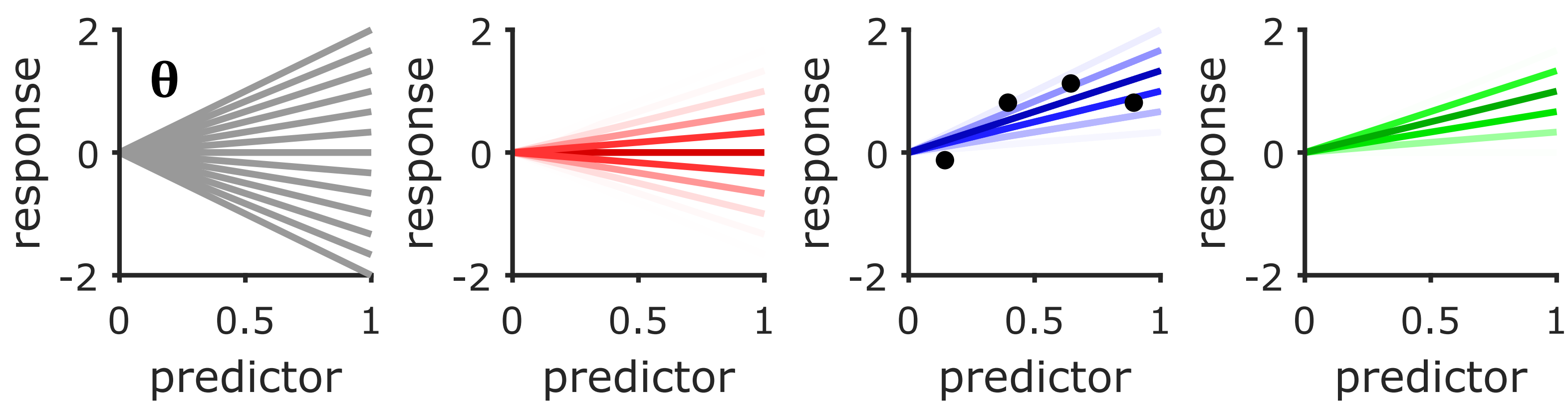
inverse problem

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

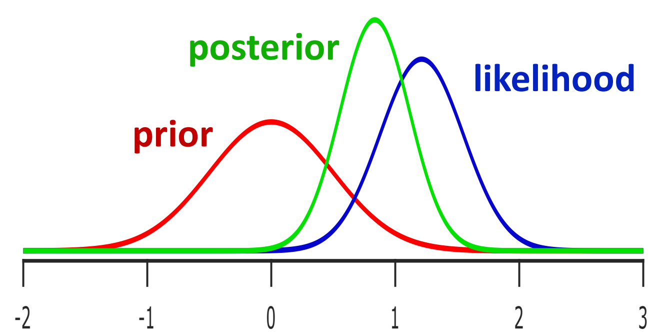
# Bayes rule

**Linear regression**

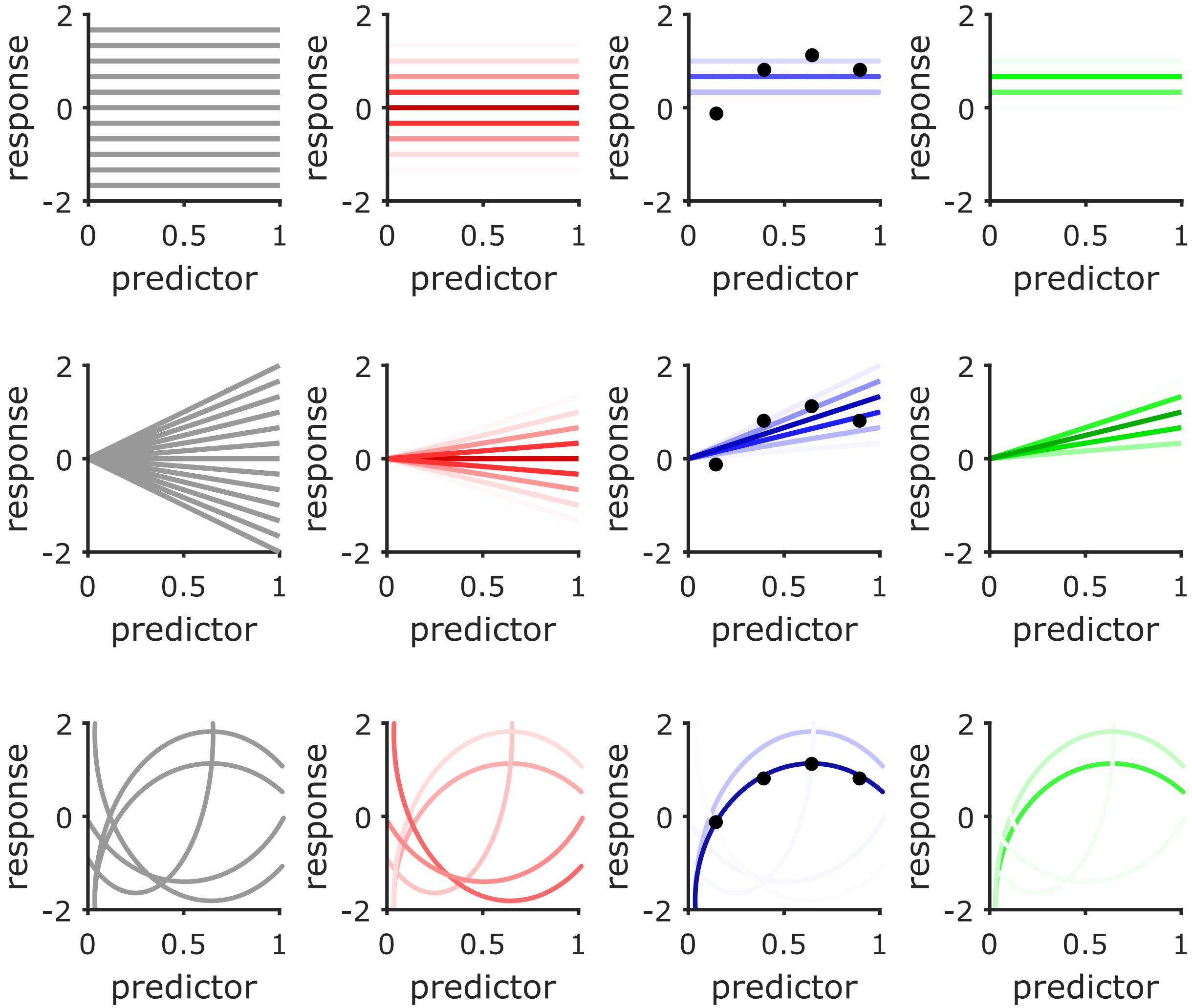
| *model* | *prior* | *likelihood* | *posterior* |
|---|---|---|---|
| $\mathbf{y} = \boldsymbol{\theta}\,\mathbf{x} + \boldsymbol{\varepsilon}$ <br> $\boldsymbol{\varepsilon} = \mathcal{N}\left(0, \sigma_y^2\right)$ | $\mathbf{p}(\boldsymbol{\theta}) = \mathcal{N}\left(0, \sigma_p^2\right)$ | $\mathbf{p}(\mathbf{y}\vert\boldsymbol{\theta}) = \mathcal{N}\left(\boldsymbol{\theta}\,\mathbf{x}, \boldsymbol{\sigma}_\mathbf{y}^2\right)$ | $\mathbf{p}(\boldsymbol{\theta}\vert\mathbf{y})$ |




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



# Model evidence

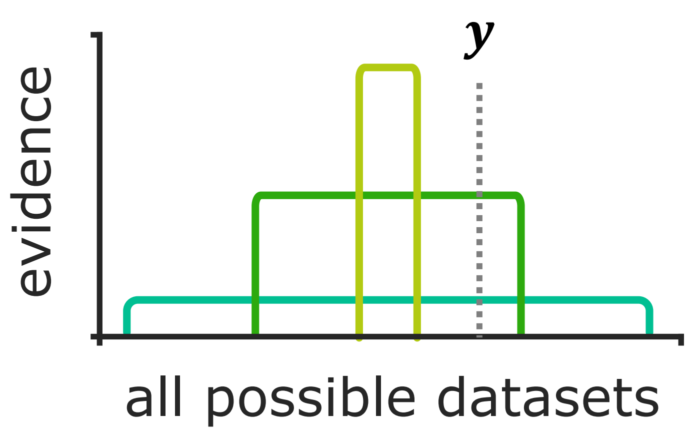


# Model evidence

Model evidence (marginal likelihood)

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

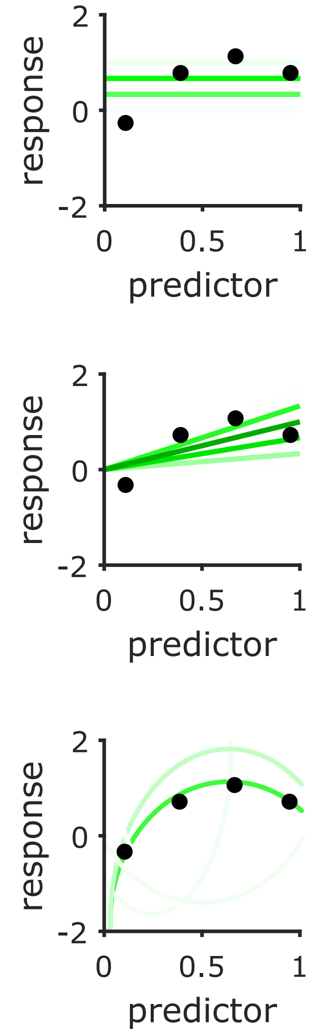
*"how likely are the data on average across plausible parametrization"*



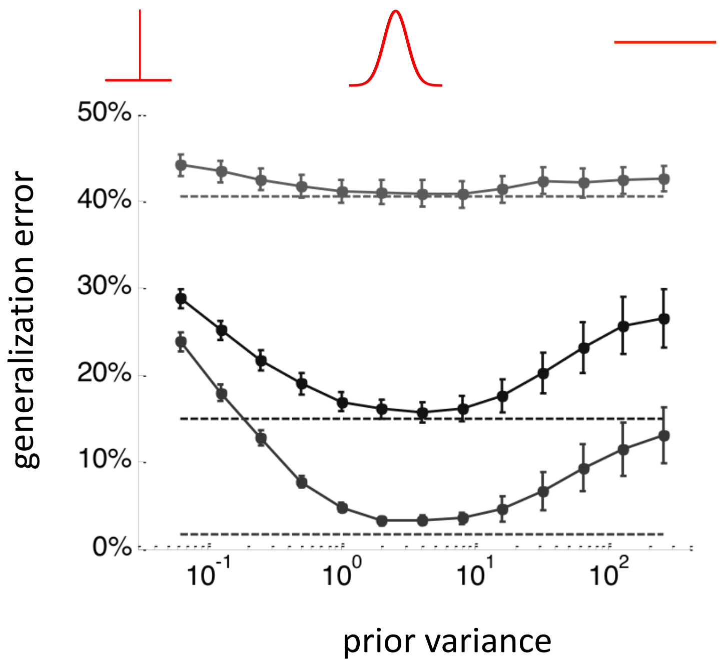

too simple
miss the data

just right

too complex
overfitting



# The bias-variance trade-off



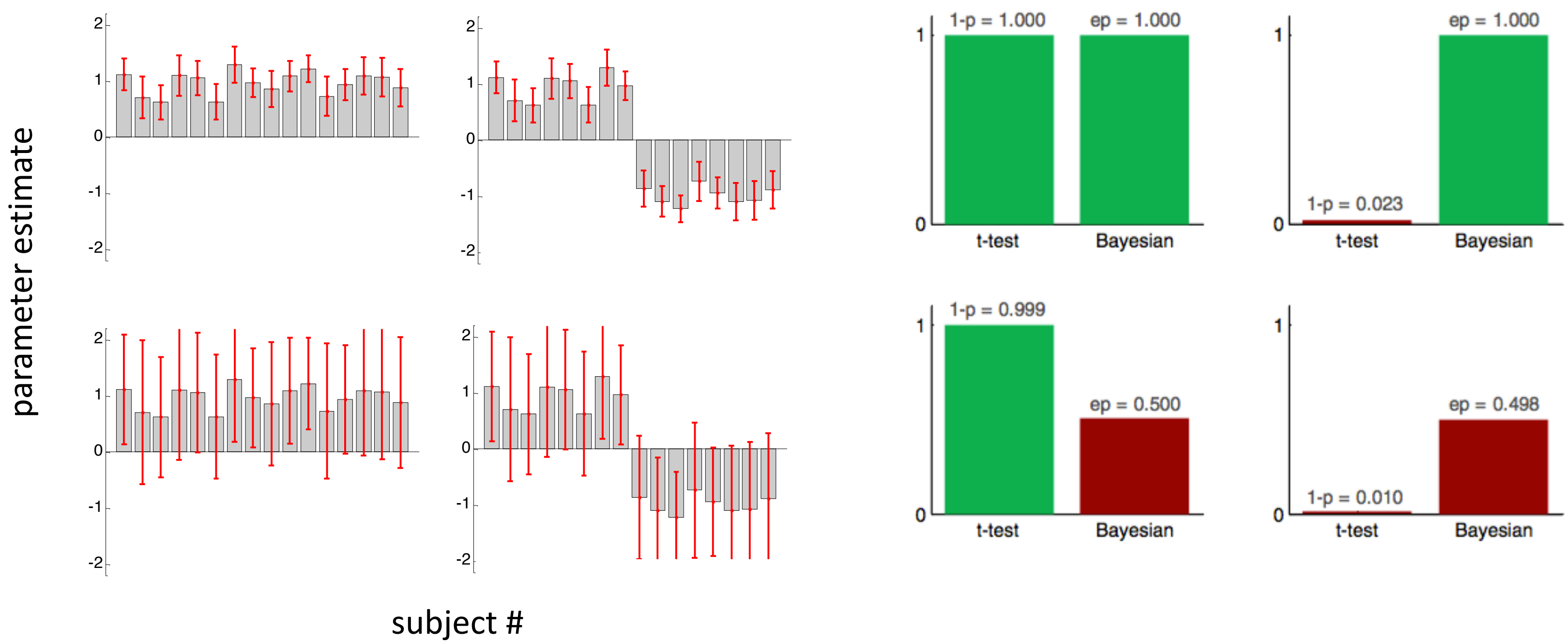

*Frequentist*

- always overfit (fit noise)
  > large variance
- estimate converge to the true value on average
  > unbiased

*Bayesian*

- regularized estimation
  > small variance
- estimate stays close to prior
  > biased

# Bayesian vs. frequentist hypothesis testing



# On the importance of priors

Priors allow to define:
- plausible values of computational parameters
- range of data patterns predicted by the model

Role of priors
- avoid overfitting (generalization error)
- anchor a complexity measure

Impact of priors
- on parameters: "shrinkage to the mean" effect (bias / regularization)
- on model evidence

# 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

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

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

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

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

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

# A little game

The un-normalized posterior:

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

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

# Markov Chain sampling

Markov Chain: stochastic process that evolve in time

- initial state $\theta_0$
- state evolve following a transition function $\mathrm{T}(\theta_{t+1}|\theta_t)$

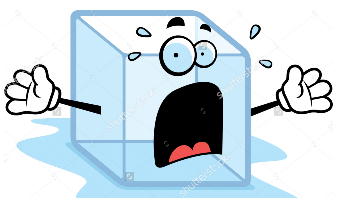
> In the long run the probability of visiting $\theta$ is called the *ergodic density*

# Metropolis Hastings algorithm

The Metropolis-Hastings algorithm

- start form $\theta_0$
- 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 $\mathbf{r} > \mathbf{X} \sim \mathbf{U}(0,1)$

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

else

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

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

# Metropolis Hastings algorithm: example

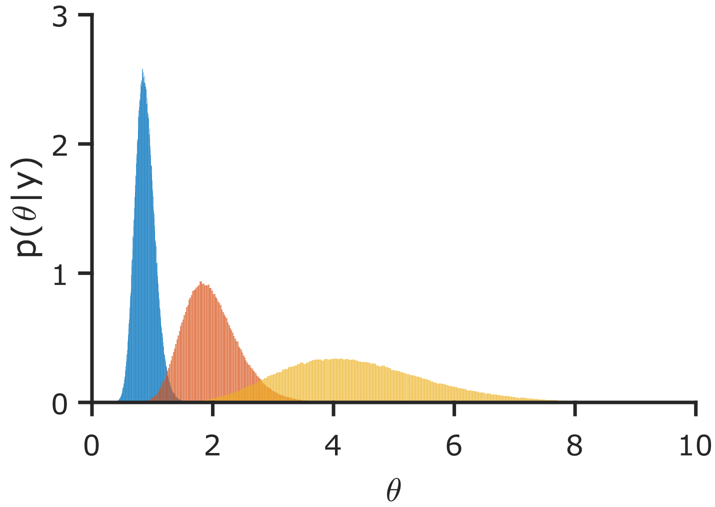
Logistic regression

$$\mathbf{p(y = 1|\theta) = s(\theta x)}$$

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

$$\mathbf{s = \frac{1}{1 + e^{-x}}}$$

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

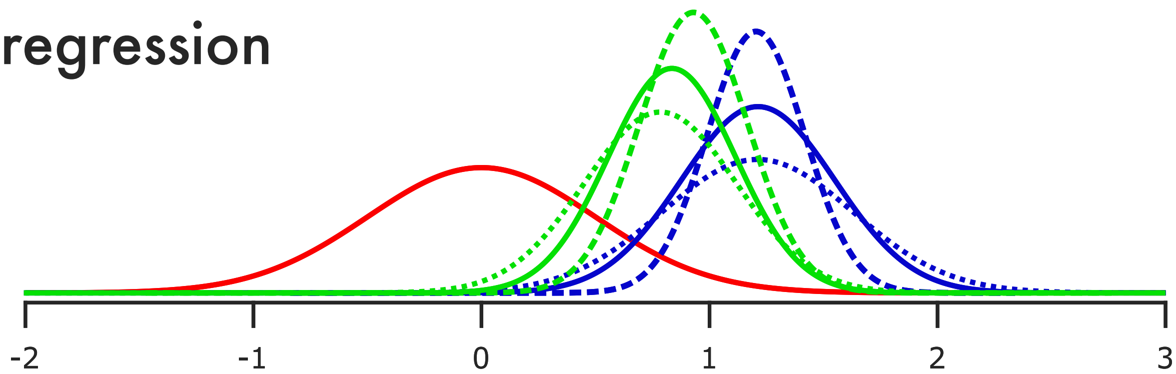


# Metropolis Hastings algorithm: multivariate case

## Example of linear regression

$$\mathbf{y} = \boldsymbol{\theta}\mathbf{x} + \boldsymbol{\varepsilon}$$

$$\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\lambda})$$




$$\mathbf{p}(\boldsymbol{\theta}, \boldsymbol{\lambda}) = \mathbf{p}(\boldsymbol{\theta})\mathbf{p}(\boldsymbol{\lambda}) = \mathcal{N}(\mathbf{0}, \boldsymbol{\sigma}_{\mathbf{0}})\mathbf{Ga}(\mathbf{a}, \mathbf{b})$$

## Blocked sampler

- start with $\theta_t = \theta_0$ and $\lambda_t = \lambda_0$
- repeat:
  - sample $\theta_{t+1}$ from $p(\theta|y, \lambda_t)$
  - sample $\lambda_{t+1}$ from $p(\lambda|y, \theta_{t+1})$

  Analytical expression (conjugacy)
  > no need for Markov chain!
- estimate $\mathbf{p}(\boldsymbol{\theta}, \boldsymbol{\lambda}|\mathbf{y})$ using LLN

# 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
- hard to tune and diagnose

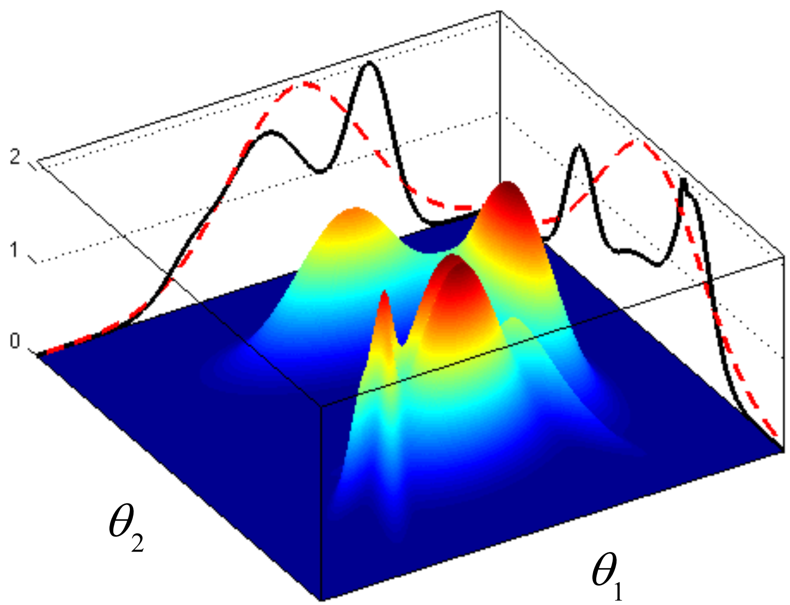
# Variational inference

*“variational inference is the thing you implement while you wait for your sampler to converge”*

David Blei

# Approximating the posterior

$$\mathbf{p}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 | \mathbf{y}, \mathbf{m}) = \frac{\mathbf{p}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 | \mathbf{m})\, \mathbf{p}(\mathbf{y} | \boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \mathbf{m})}{\mathbf{p}(\mathbf{y} | \mathbf{m})}$$




Mean field approximation —

$$\mathbf{p}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 | \mathbf{y}) \approx \mathbf{p}(\boldsymbol{\theta}_1 | \mathbf{y})\mathbf{p}(\boldsymbol{\theta}_2 | \mathbf{y})$$

Laplace approximation - - -

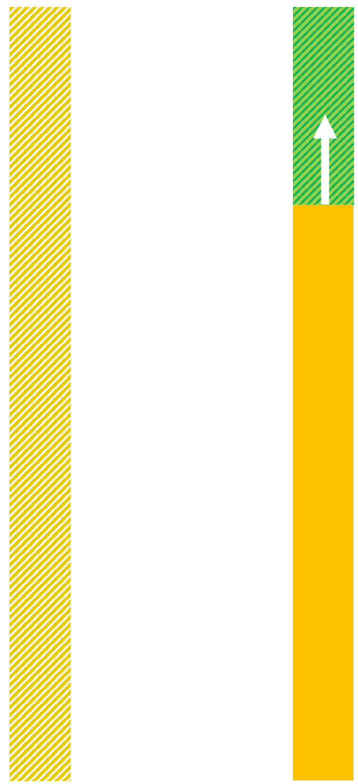
$$\mathbf{q}(\boldsymbol{\theta}_1 | \mathbf{y}) \approx \mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$$

finding $\mathbf{p}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 | \mathbf{y}, \mathbf{m})$ ⟶ finding $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2$

# Free Energy approximation

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

evidence



posterior approximation error

model evidence approximation

$-$

$$\underset{q(\theta)}{\operatorname{argmax}} F(q(\theta))$$

$+$

*#ELBO*

*#Jensen's Inequality*

*#Free energy*

## Free Energy approximation

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

1) Maximize the free energy using variational calculus

$$\mathbf{F} = \langle \log \mathbf{p(y|\theta_1, \theta_2, m)} + \log \mathbf{p(\theta_1, \theta_2)} \rangle_{\mathbf{q}} + \langle \log \mathbf{q(\theta_1, \theta_2)} \rangle_{\mathbf{q}}$$

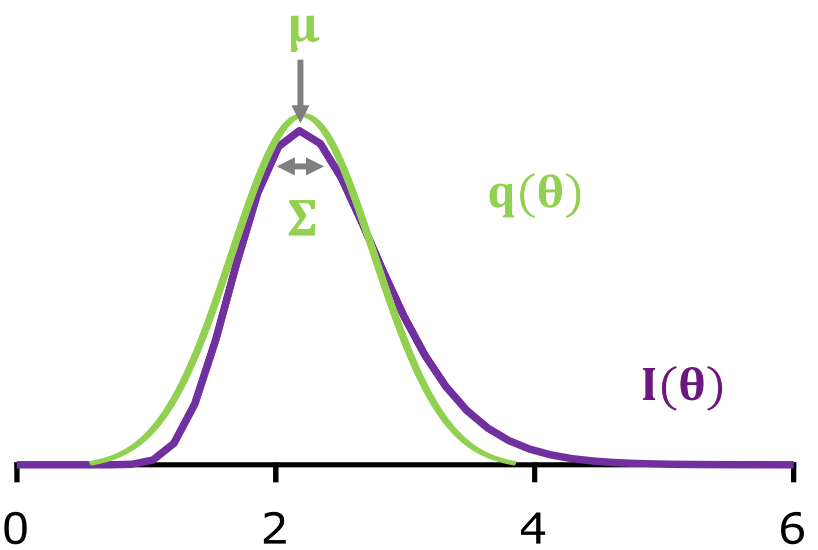
$$\frac{\partial \mathbf{F}}{\partial \mathbf{q(\theta_1)}} = \mathbf{0} \Rightarrow \mathbf{q(\theta_1)} \propto \exp\Big(\underbrace{\langle \log \mathbf{p(y|\theta_1, \theta_2, m)} + \log \mathbf{p(\theta_1, \theta_2)} \rangle_{\mathbf{q(\theta_2)}}}_{\text{variational free energy} \quad \mathbf{I(\theta_1)}}\Big)$$

2) Iterate over parameters until convergence

# Variational inference

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

*logistic regression*




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

*multivariate case*

Until convergence:

for all i:

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

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.

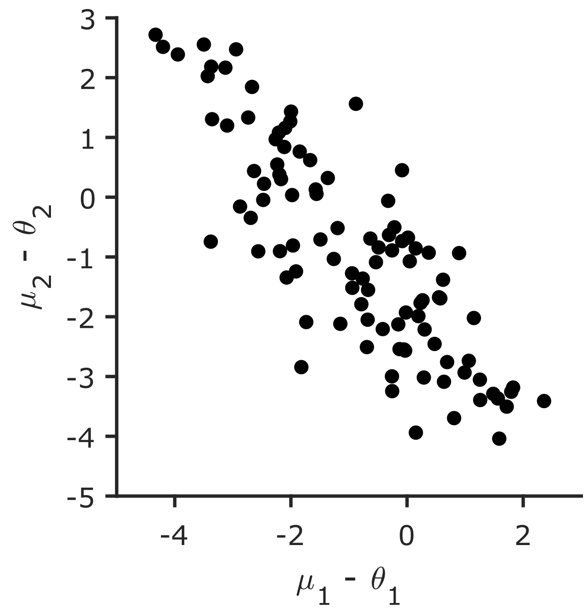
# Software

- 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 $\theta$
- invert your model (find $\mu$)
- compute estimation error $(\mu - \theta)$
- check bias/variance trade-off
- check for posterior / error correlation



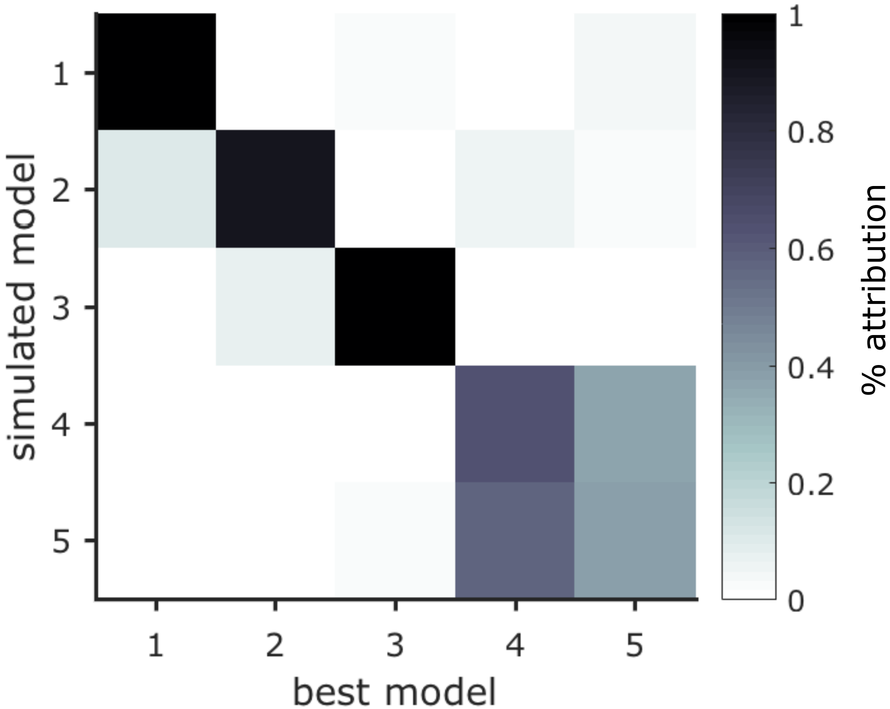
# Validating your model: hypothesis identifiability

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!