# Bayesian model inversion and selection


Lionel Rigoux

Translational Neuro-Circuitry (TNC) Cologne

Translational Neuromodeling Unit (TNU) Zürich



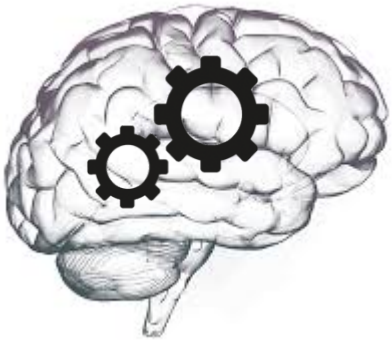
# Overview

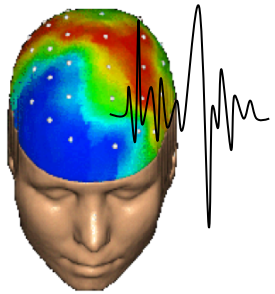
$p(y|\theta, m)$

forward problem

generative model + parameters

$m, p(\theta)$





observation

$y$

inverse problem

$p(\theta|y, m)$ $p(m|y)$

# Bayes rule

joint distribution

$$p(y, \theta|m)$$

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

Expectation

$$E[p(y|\theta, m)]_{p(\theta|m)}$$

Marginal likelihood

$$\int p(y, \theta|m)\, d\theta$$

Model evidence

$$p(y|m)$$

# Overview

Compute the posterior and the evidence for a model

- Monte-Carlo (sampling) methods
- Variational methods

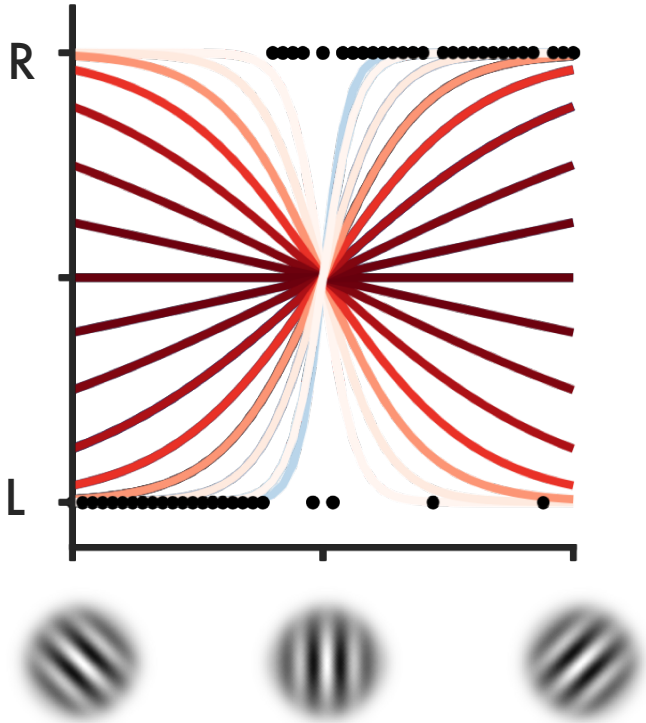
Compare multiple models and select the best

Good practices

# online supplementary material

https://github.com/lionel-rigoux/tutorial-bayesian-inference

# Example: logistic regression




*Sensitivity to orientation?*

*Bias?*

*How many parameters?*

Model prediction:

$$p(y = 1|\theta) = sig(\theta u + \beta) = s$$

Likelihood:

$$p(y|\theta, \beta) = \prod_y s^y \, (1 - s)^{1-y}$$

Prior:

$$p(\theta) = \mathcal{N}(0, \sigma_0^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{\theta^2}{2\sigma_0^2}\right)$$

Joint:

$$p(y, \theta, \beta) \propto \exp\left(-\frac{\theta^2}{2\sigma_0^2}\right) \prod_y s^y \, (1 - s)^{1-y}$$

# Sampling (Monte Carlo)

# Monte-Carlo methods

$$\mathbf{E}[\mathbf{z}] = \sum p(z)z = \sum_{z=1}^{6} \frac{1}{6} z = 3.5$$

$$\mathbf{E}\left[(\mathbf{z} - 3.5)^2\right] = \sum p(z)(z - 3.5)^2 = 2.9167$$


$$\mathbf{E}[\mathbf{z}] \approx \frac{1}{n} \sum_{i=1}^{n} z_i \qquad z_i \sim p(z)$$

$$\mathbf{E}[\mathbf{f}(\mathbf{z})] \approx \frac{1}{n} \sum_{i=1}^{n} f(z_i)$$

Law of
Large Numbers

## Model evidence

$$p(y) = E[p(y|\theta)]_{p(\theta)} \approx \frac{1}{n}\sum_{i=1}^{n} p(y|\theta_i) \qquad \theta_i \sim p(\theta)$$

## Posterior moments

$$\mu = E[\theta]_{p(\theta|y)} \approx \frac{1}{n}\sum_{i=1}^{n} \theta_i \qquad \theta_i \sim p(\theta|y)$$

$$\Sigma = E\left[(\theta - \mu)^2\right]_{p(\theta|y)} \approx \frac{1}{n}\sum_{i=1}^{n} (\theta_i - \hat{\mu})^2$$

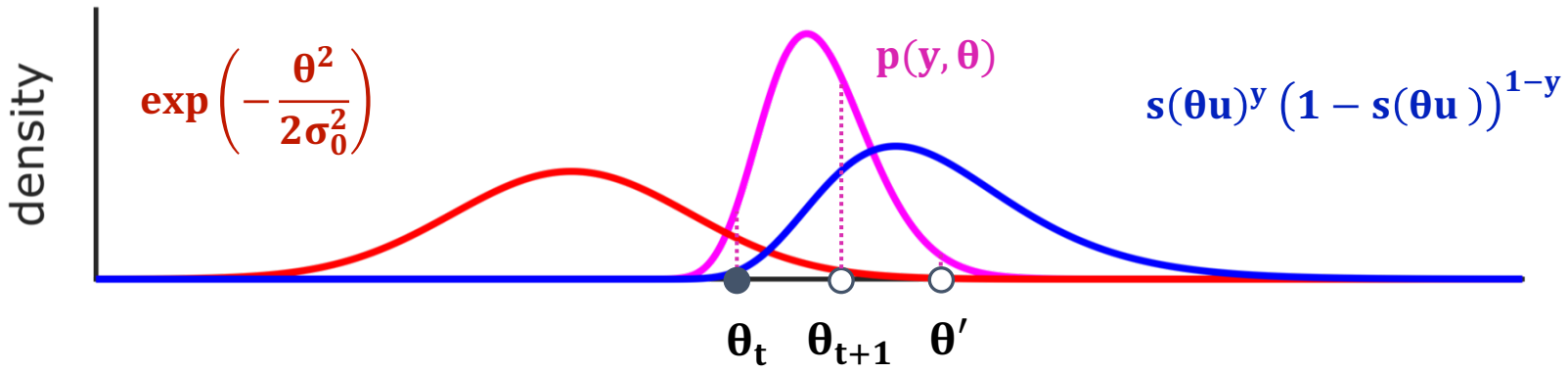
# A little game

The joint as an un-normalized posterior:

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

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

# Metropolis-Hastings algorithm




Current state

$p(y,\theta_t) = p(\theta_t)\,p(y|\theta_t)$

Proposal

$\theta' \sim q(\theta|\theta_t)$

$p(y,\theta') = p(\theta')\,p(y|\theta')$

$$\alpha = \frac{p(y,\theta')}{p(y,\theta_t)}$$

$\alpha \geq 1$

jump to proposed value

$\theta_{t+1} = \theta'$

$\alpha < 1$

Draw $x \sim U(0,1)$

- If $\alpha > x$, jump

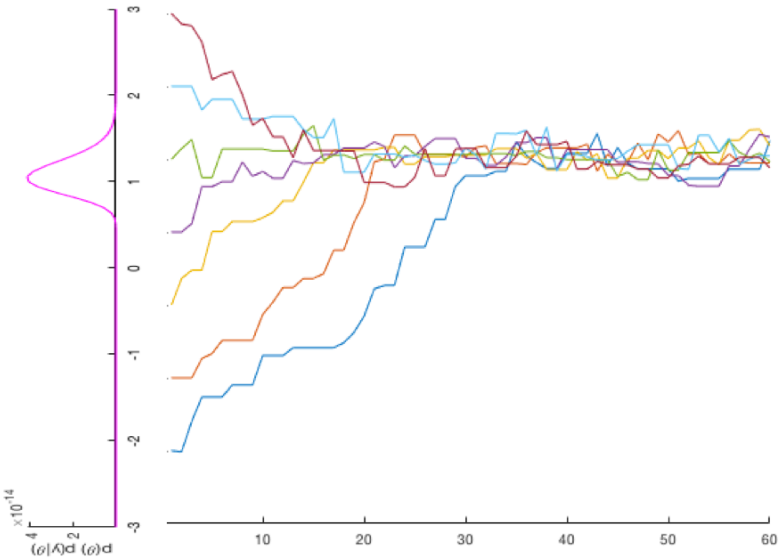
$\theta_{t+1} = \theta'$

- else, stay in place

$\theta_{t+1} = \theta_t$

# Did I sample right?

All sampling methods requires some "post-processing" and an extensive diagnostic to ensure the samples are representative.




1) Run multiple chains
2) Check:

- Convergence (eg. Geweke)
- Mixing (eg. Gelman-Rubin)
- Autocorrelation
- Step size (Goldilocks principle)

# Multivariate case

write conditional posteriors

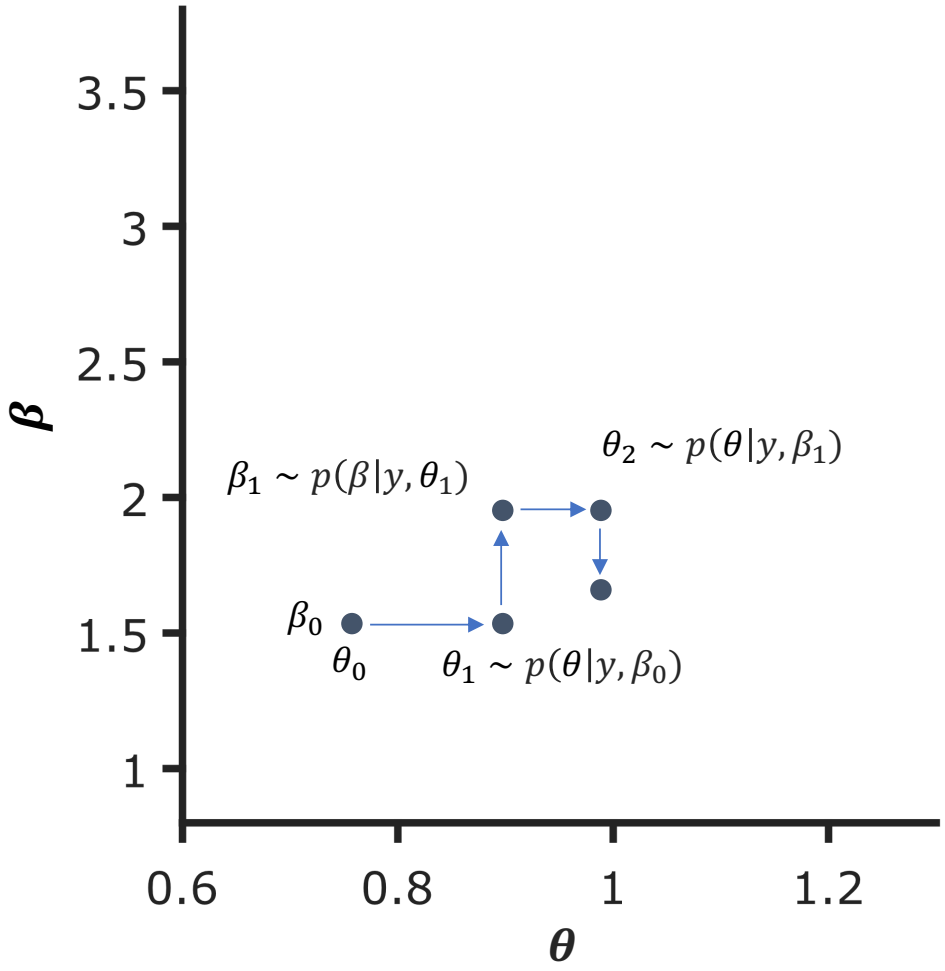
$$\mathbf{p(\theta|y,\beta) = \frac{p(y,\theta,\beta)}{p(y,\beta)}}$$

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

Iterative sampling

$$\mathbf{\theta_t \sim p(\theta|y,\beta_{t-1})}$$

$$\mathbf{\beta_t \sim p(\beta|y,\theta_t)}$$



# Multivariate case

Using the law of large numbers:

- Posterior mean

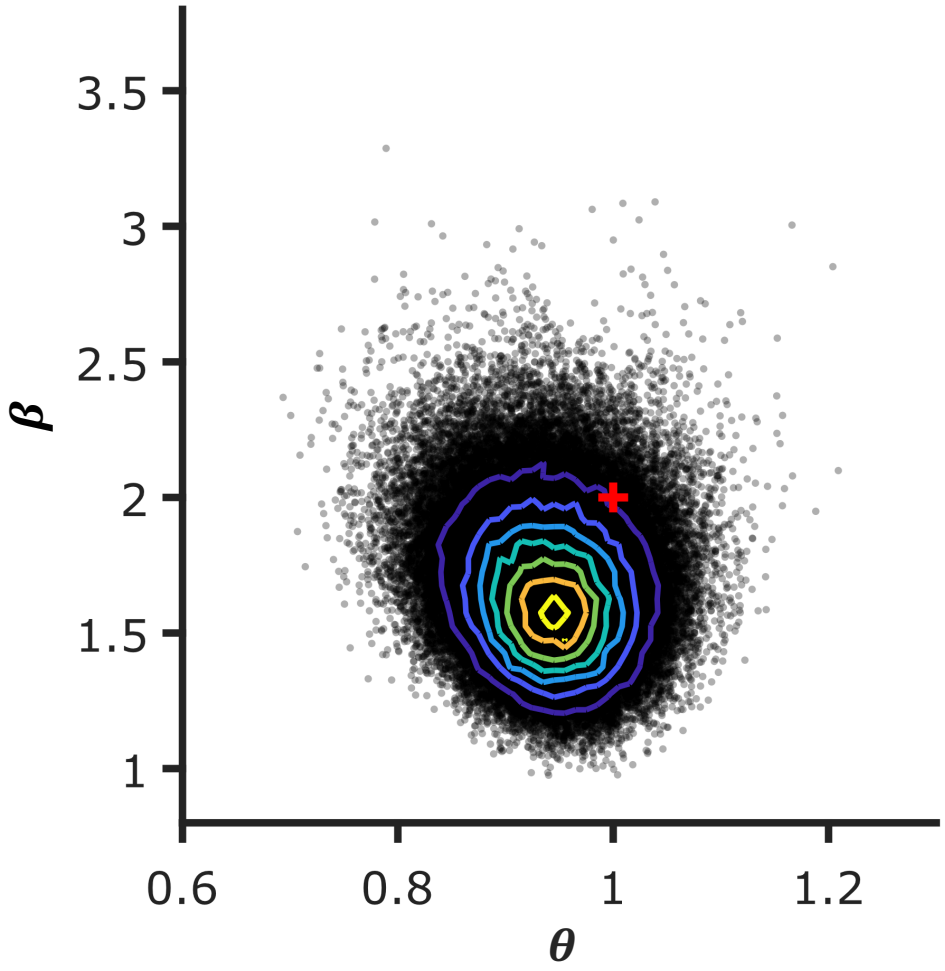
$$E[\theta|y] \approx mean(\theta_t)$$

$$E[\beta|y] \approx mean(\beta_t)$$

- Posterior variance

$$E\left[(\theta - \bar{\theta})^2 \middle| y\right] \approx var(\theta_t)$$

$$E\left[(\beta - \bar{\beta})^2 \middle| y\right] \approx var(\beta_t)$$



# Monte-Carlo inference

Monte-Carlo methods rely on sampling to estimate the posterior and the model evidence.

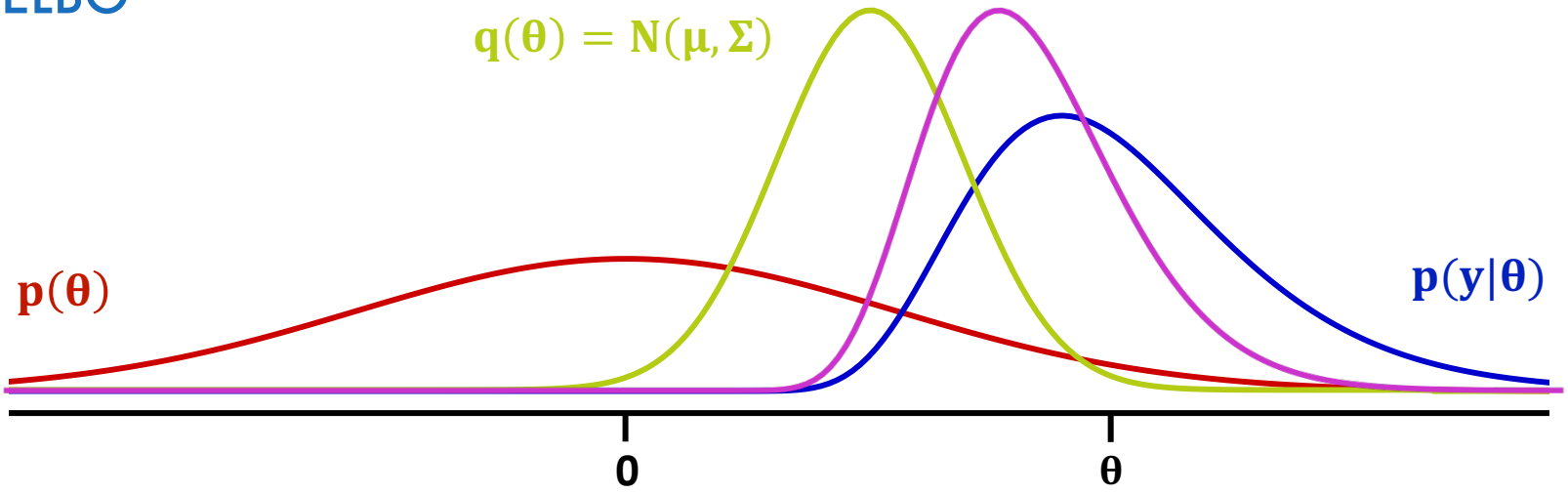
> The Law of Large Numbers guarantee that the sufficient statistics of the samples will converge to the true posterior moments.

Problems:

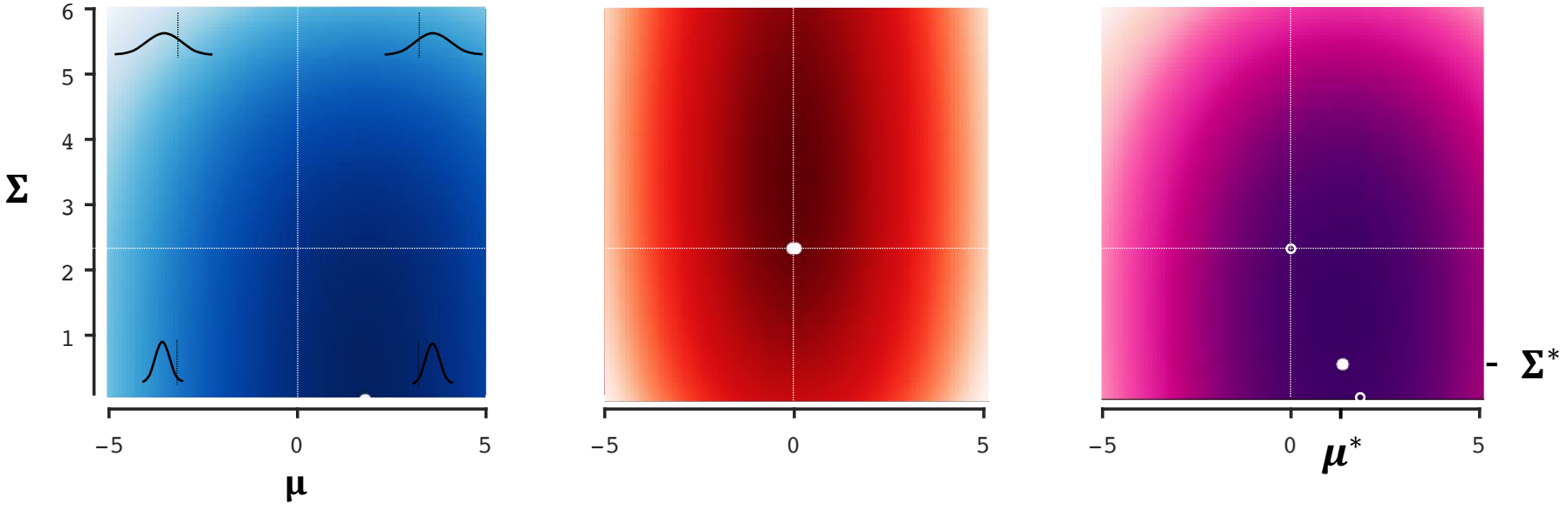
- computationally expensive
- does not scale well with the number of parameters
- no direct measure of model evidence
- hard to tune and diagnose

# Variational Methods

# The ELBO




$$E[\log p(y|\theta)]_q + E\left[\log \frac{p(\theta)}{q(\theta)}\right]_q = E\left[\log \frac{p(y,\theta)}{q(\theta)}\right]_q$$



## Variational Laplace

Maximize using $q(\theta) = N(\mu, \Sigma)$

$$F = \int \log p(y, \theta)\, q(\theta)\, d\theta + H(q(\theta))$$

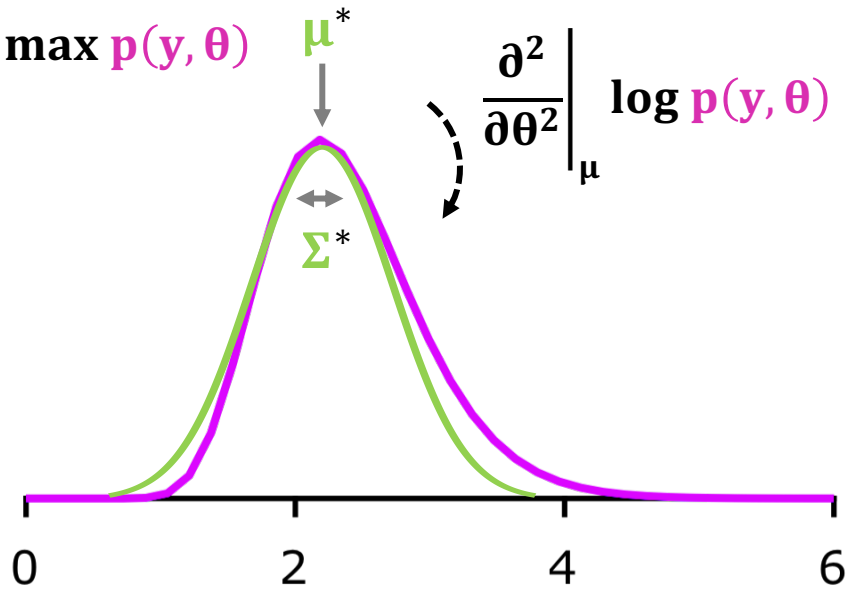
## Stochastic gradient

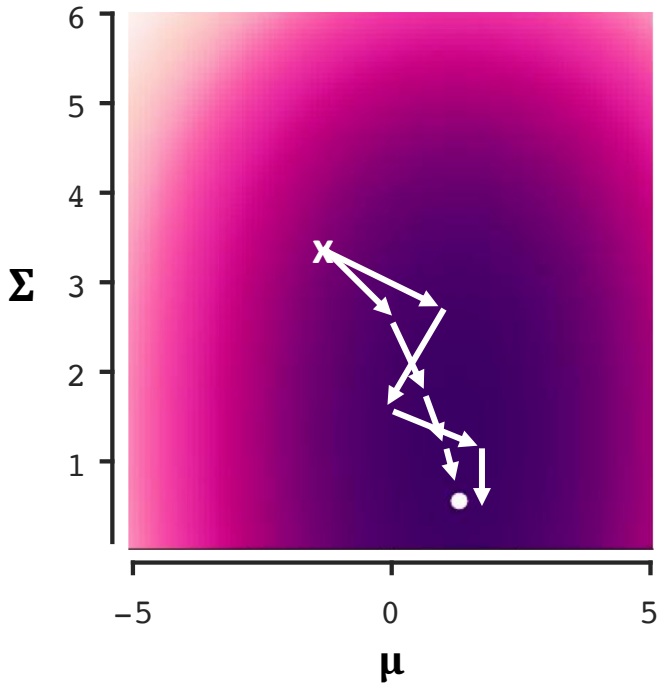
*Analytical approximation:* $F \approx F_{Laplace}$

*Find maximum:* $\frac{d}{dq(\theta)} F_{Laplace} = 0$

$$\nabla F = E_q[\nabla \log q(\theta)\, (\log \frac{p(y, \theta)}{q(\theta)})]$$

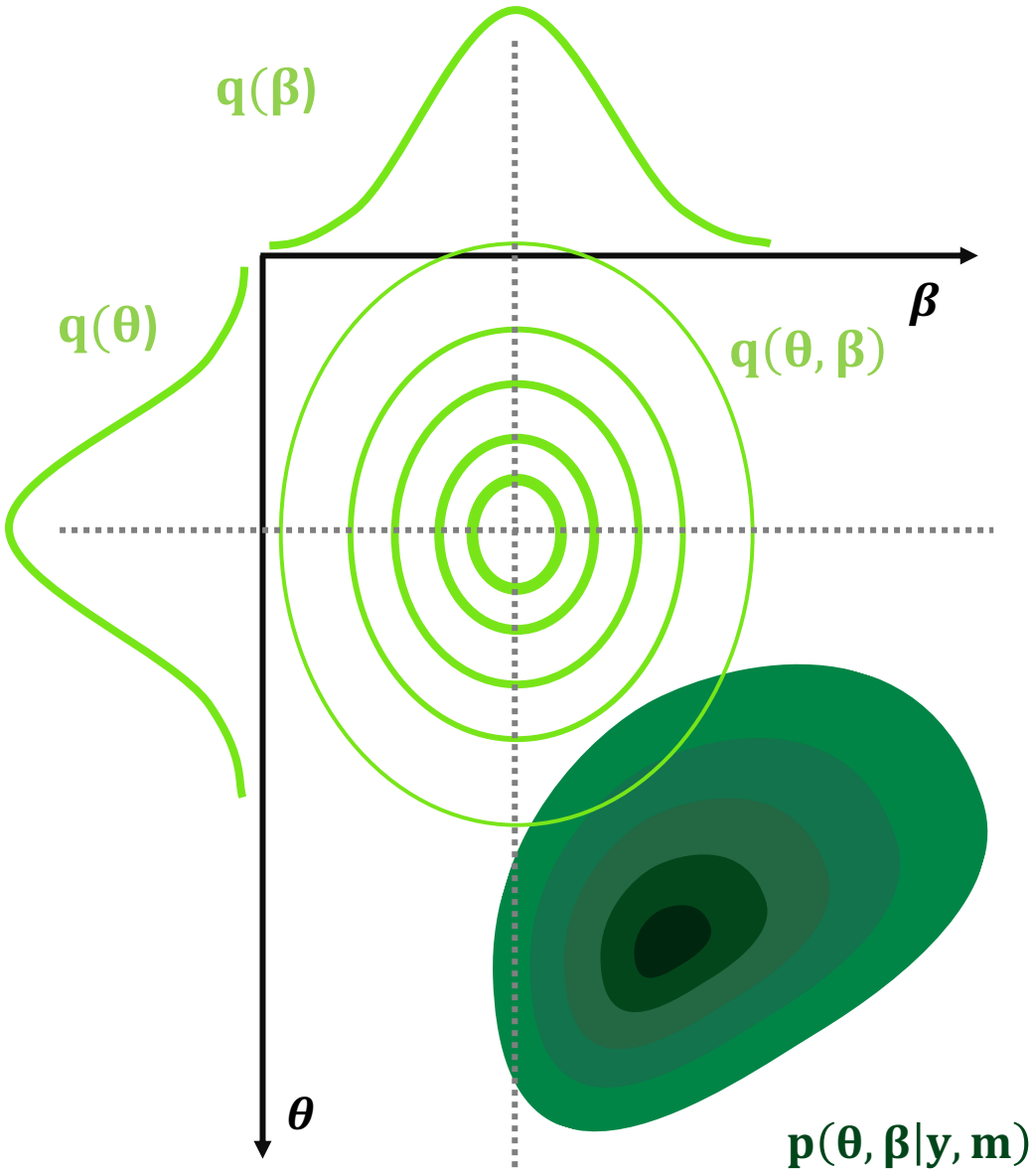
*Find maximum:* **gradient ascent**




$$\log p(y) \approx \log p(y, \mu^*) + \frac{1}{2}[\log |\Sigma^*| + n_\theta \log(2\pi)]$$

# Multivariate posterior




## Mean field approximation

$$p(\theta, \beta|y) \approx p(\theta|y)p(\beta|y)$$

$$q(\theta, \beta) \approx q(\theta)q(\beta)$$

## Maximise Variational energy

$$I(\theta) = E[\log p(y, \theta, \beta)]_{q(\beta)}$$

$$\approx \log p(y, \theta_1, \mu_\beta) + \ldots$$

## Iterative optimization

$$\mu_i = \text{argmax}\, I(\theta_i)$$

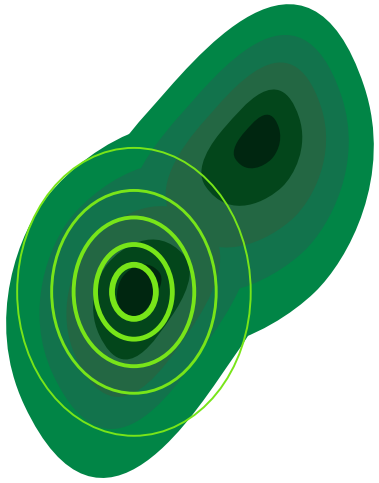
$$\Sigma_i = -\left[\frac{\partial^2}{\partial \theta_i^2}\bigg|_{\mu_i} I(\theta_i)\right]^{-1}$$

# Variational inference

Summarize the posterior to its sufficient statistics (mean, variance) and optimize those values wrt the ELBO.
This requires multiple approximations (Jensen/Free-energy, Gaussian posterior, Laplace, mean-field) to be tractable.

Problems:

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



# Take home message

Model evidence (normalization factor of the posterior) is in general intractable and calls for numerical methods.

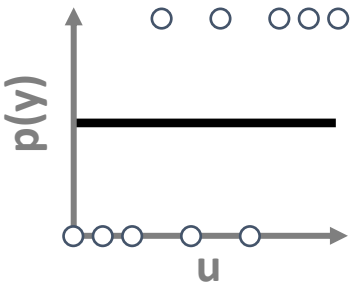
Sampling methods give a computationally expensive estimation of the true posterior.

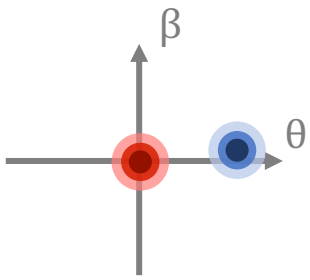
Variational methods are  fast & scalable computations of an approximation of the posterior.

# Model selection

$$E[p(y|\theta, m)]_{p(\theta|m)}$$

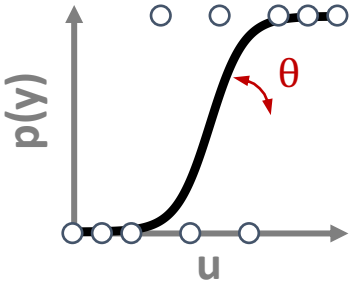
**m1**

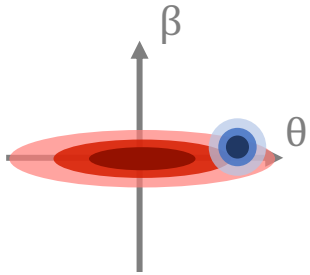



**low**
*too simple*

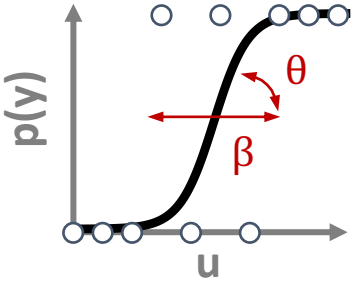
**m2**

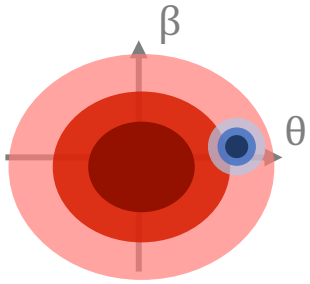



**high**
*just right*

**m3**



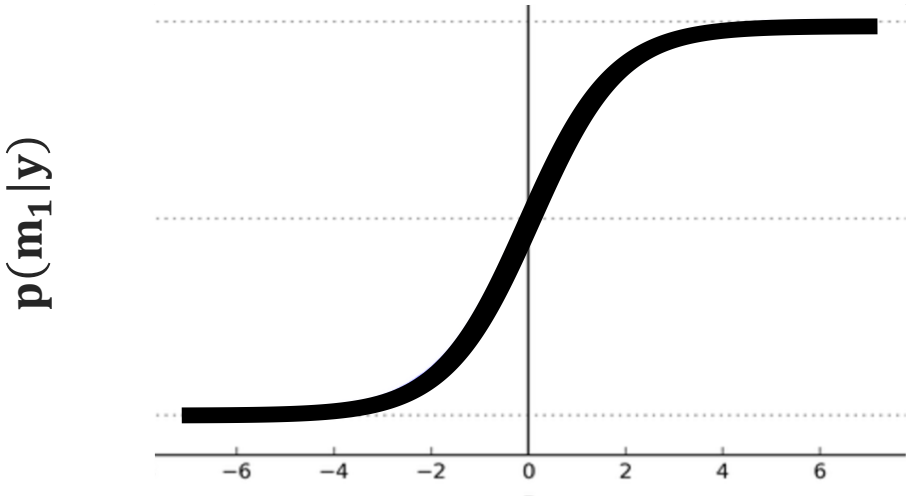




**low**
*too complex*

# Model Selection: single subject

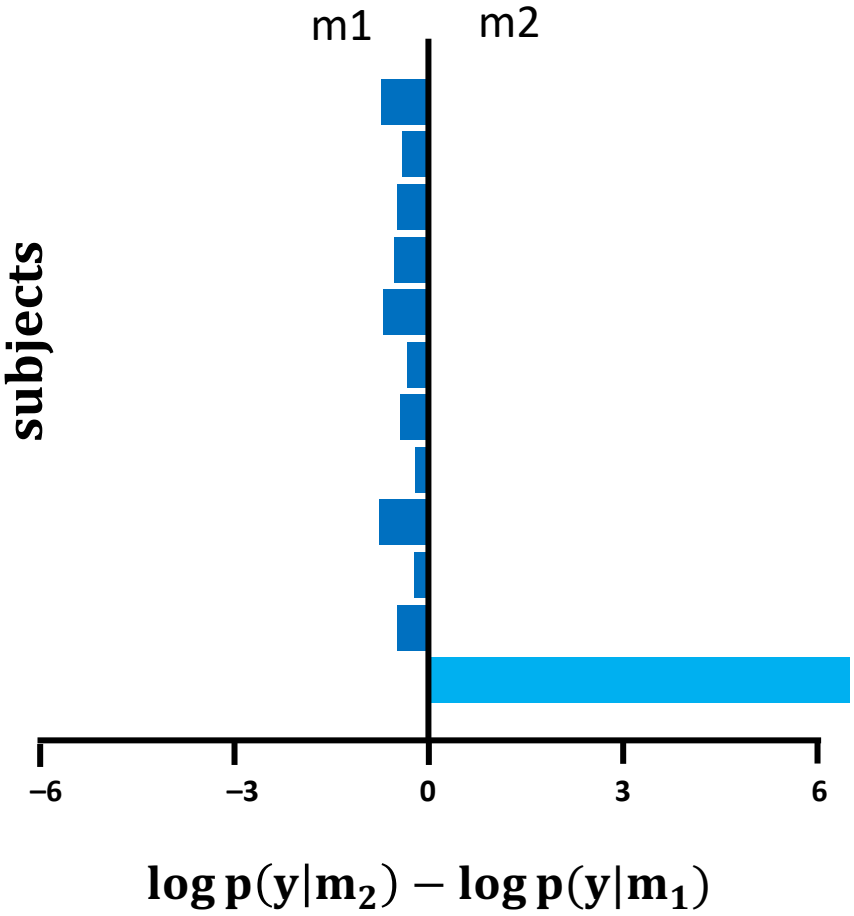
uniform prior

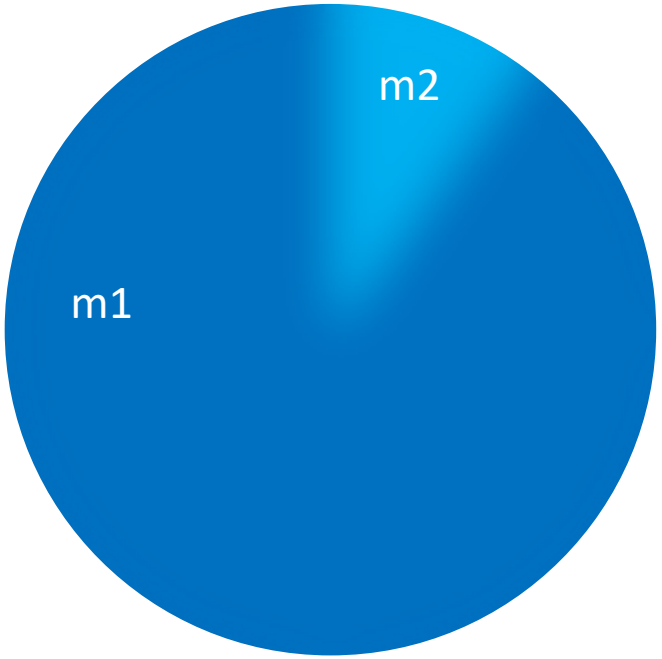
$$p(m_i|y) \quad \Rightarrow \quad p(m_i|y) = \frac{e^{\log p(y|m_i)}}{\sum_j e^{\log p(y|m_j)}}$$




$$\log p(y|m_1) - \log p(y|m_2) = \log B_{1,2}$$

# Model Selection: group analysis



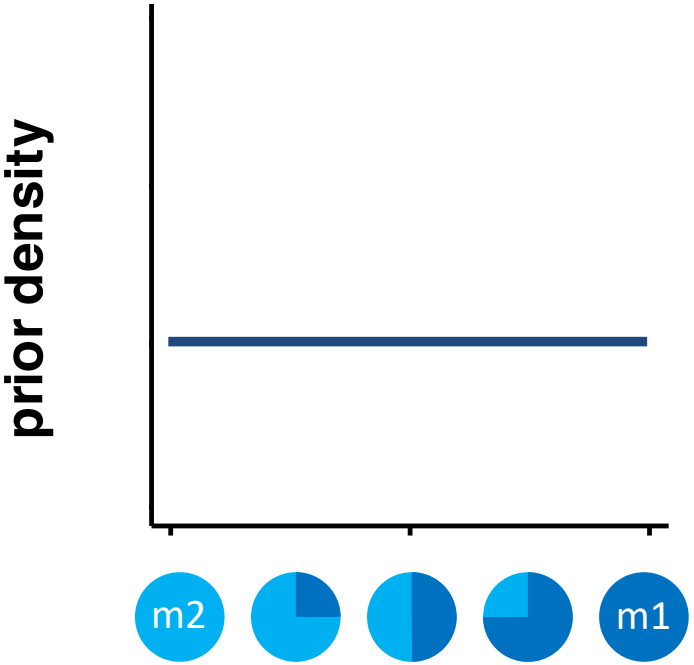

*If I pick a new subject at random, which model is the most likely to explain their data?*

# Model Selection: group analysis

## Hierarchical model

$$\mathbf{p(f) = Dirichlet(\alpha_0)}$$

$$\mathbf{p(y, f) = \prod_s p(y_s|m_s)p(m_s|f)p(f)}$$

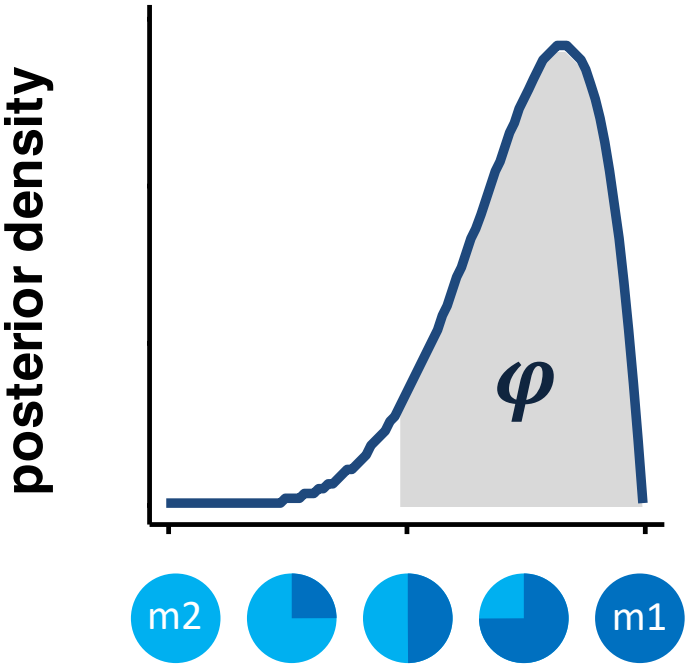



## Inference over frequencies

$$\mathbf{Ef = E[f_{m1}|y]}$$

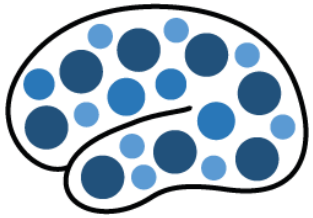
$$\boldsymbol{xp} = \mathbf{p}(\mathbf{f}_{m1} > \boldsymbol{f}_{m2})$$

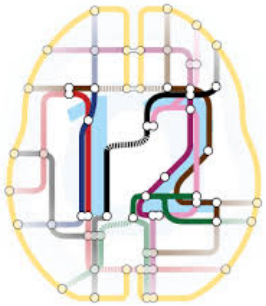
" Ef = 0.78, pxp = 0.98 "

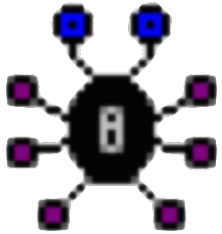


# Software

- Variational
  - VBA-toolbox
  - TAPAS
  - SPM
- Sampling
  - STAN
  - BUGS
  - JAGS
  - hBayesDM
  - hddm







JAGS

# VBA Toolbox

~ 120 published papers

85 demos (tutorial, Q-learning, HGF, DCMs, etc)

Online wiki + Q&A

Need only the model description!

Simulation

Inversion (single subject, hierarchical)

Model selection (families, btw groups, btw conditions)

Visual diagnostics

Design optimization, multisession, multimodal observations, …

# Thank you!

***Online supplementary material***

https://github.com/lionel-rigoux/tutorial-bayesian-inference

***VBA-Toolbox***

https://mbb-team.github.io/VBA-toolbox

***Easy writing workflow***

https://lionel-rigoux.github.io/pandemic/

# Variational variants

$$\log p(y) = \log \int p(y,\theta)\, d\theta$$

$$= \log \int \frac{p(y,\theta)}{q(\theta)} q(\theta) d\theta$$

*Jensen's inequality*

$$\geq \int \log \frac{p(y,\theta)}{q(\theta)} q(\theta) d\theta$$

$$= \int \log p(y,\theta)\, q(\theta)\, d\theta + S(q(\theta))$$

*approximation error*

$$KL[q(\theta)||p(\theta|y)]$$

*need to maximise wrt q*

*Assume parametric form*

$$q(\theta) = N(\mu, \Sigma)$$

*Analytical approximation*

**Variational Laplace**

*Sampling*

**Stochastic gradient**