# Fitting a model: VB & MCMC

Lionel Rigoux

Max Planck Institute for Metabolism Research

Translational Neuro-Circuitry Group

Max-Planck-Institut
für Stoffwechselforschung

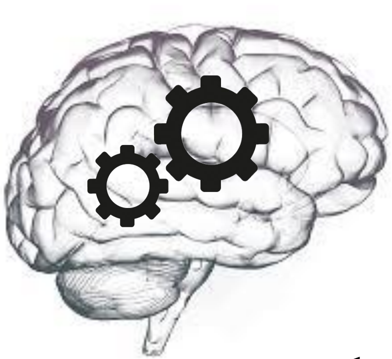
# Overview

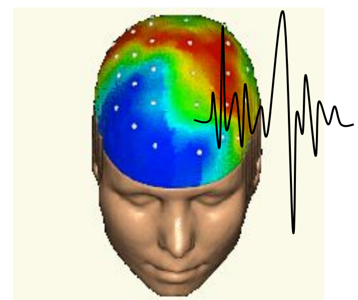
$$\mathbf{p(y|\theta, m)}$$

forward problem

generative model

$\mathbf{m, p(\theta)}$





observation

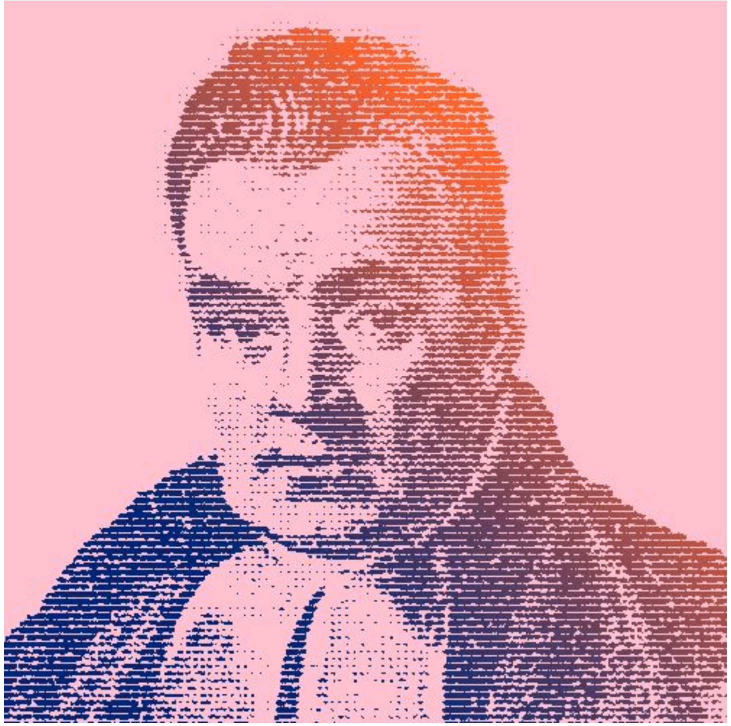
$\mathbf{y}$

inverse problem

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

# Bayes rule

Divine Benevolence
*or*
An Attempt to Prove That the Principal End of the Divine Providence and Government Is the Happiness of His Creatures

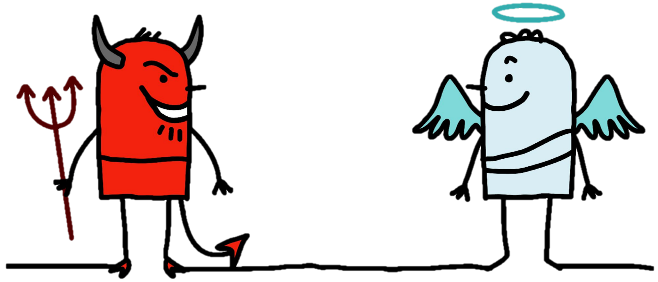


$$\mathbf{p(y|\theta, m)p(\theta|m)}$$

An Introduction to the Doctrine of Fluxions and a Defence of the Mathematicians Against the Objections of the Author of The Analyst

# Bayes rule

trust the
data!



do not jump
to conclusion!

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

# Bayes rule

Joint distribution

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

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

Expectation

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

Marginal likelihood

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

Model evidence

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

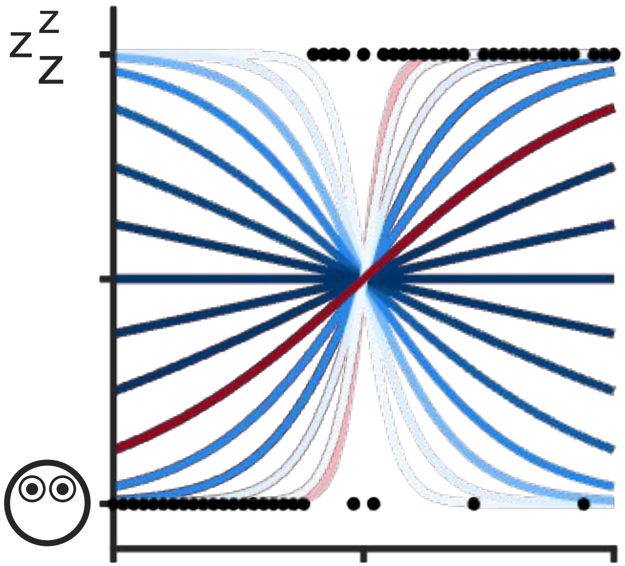
## Overview

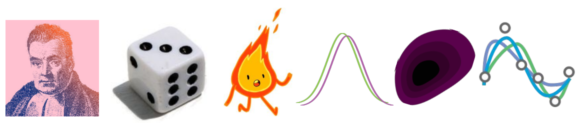
*How to compute the posterior belief and the model evidence in practice?*

Monte-Carlo (sampling) methods

Variational methods

# Example: logistic regression





Model prediction: psychometric (logistic) function

$$\mathbf{p(y_t = 1 | u_t, \theta, \beta) = sig\,(\theta u_t + \beta) = s_t}$$

Likelihood

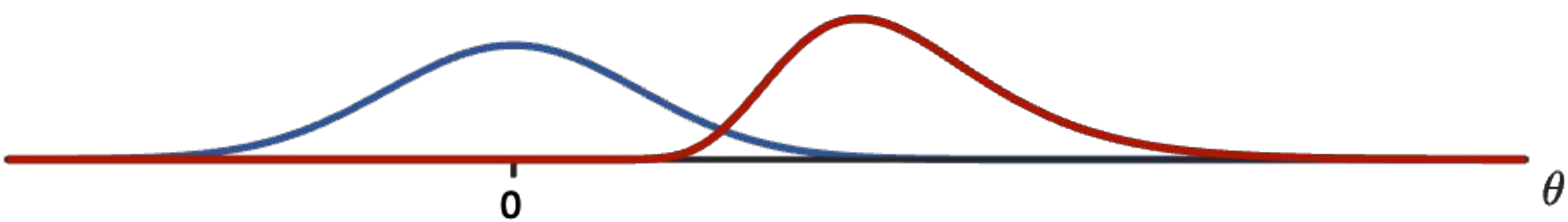
$$y_t \sim \mathcal{B}(s_t)$$

$$\mathbf{p(y|\theta, \beta)} = \prod \mathbf{s_t{}^y\ (1 - s_t)^{1-y}}$$

Prior

$$\boldsymbol{\theta \sim \mathcal{N}(\mu_\theta, \sigma_\theta^2)} \qquad \boldsymbol{\beta \sim \mathcal{N}(0, 0)}$$

$$\mathbf{p(\theta) = exp}\left(-\frac{1}{2}\frac{(\boldsymbol{\theta} - \boldsymbol{\mu_\theta})^2}{\boldsymbol{\sigma_\theta^2}}\right)\frac{1}{\sqrt{2\pi\sigma^2}}$$



# Example: logistic regression

Joint

$$\mathbf{p(y,\theta,\beta)} \propto \prod \mathbf{s_t}^{\mathbf{y}} \mathbf{(1-s_t)}^{\mathbf{1-y}} \mathbf{e}^{-\frac{(\theta-\mu_\theta)^2}{2\sigma_\theta^2}}$$

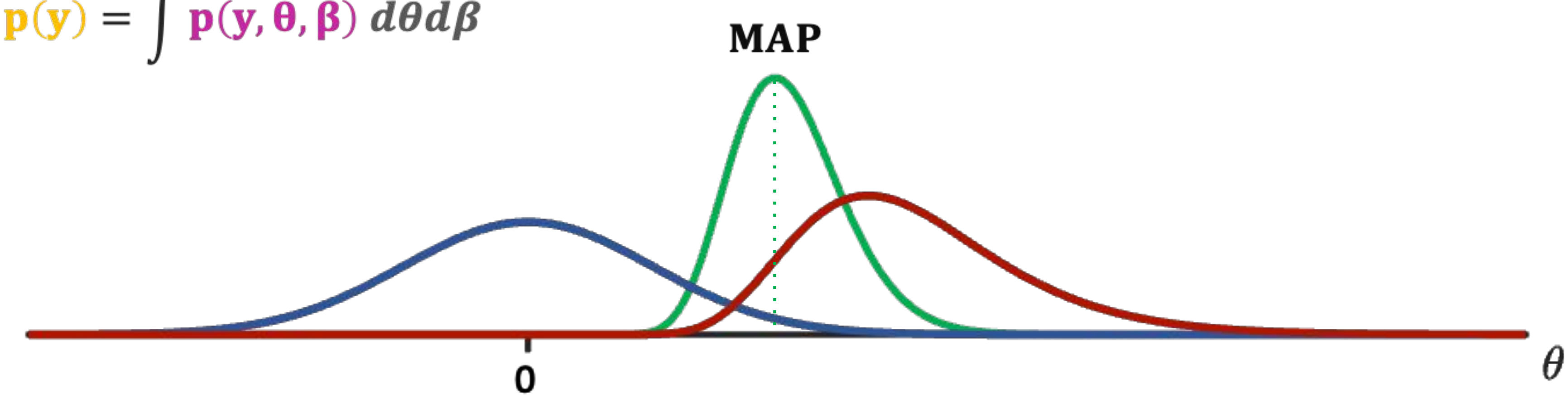
$$\log \mathbf{p(y,\theta,\beta)} = \sum \boldsymbol{y} \log \mathbf{s_t} + (\mathbf{1}-\boldsymbol{y}) \log(\mathbf{1}-\mathbf{s_t}) - \frac{(\boldsymbol{\theta}-\boldsymbol{\mu_\theta})^2}{\mathbf{2}\,\boldsymbol{\sigma_\theta^2}} + c$$

Posterior

$$\mathbf{p(\theta,\beta|y)} \propto \mathbf{p(y,\theta,\beta)} \qquad \mathbf{MAP} = \underset{\theta,\beta}{\operatorname{argmax}}\, \mathbf{p(y,\theta,\beta)}$$

Model evidence

$$\mathbf{p(y)} = \int \mathbf{p(y,\theta,\beta)}\, d\theta d\beta$$



# Sampling (Monte Carlo)

*"When one tries continuously, one ends up succeeding. Thus, the more one fails, the greater the chance that it will work."*

***Les Shadoks***

# Monte-Carlo methods

*How many turns to win?*

# Monte-Carlo methods

Expectation (theoretical mean)

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

Variance (theoretical distance to the mean)

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

$$\mathbf{E}[\mathbf{z}_1 + \cdots + \mathbf{z}_n \geq 63] = ? \qquad \mathbf{n} \approx \frac{63}{\mathbf{E}[\mathbf{z}]} = 18$$

## Law of Large Numbers

Expectation ≈ Empirical mean

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

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

# Monte-Carlo methods

## Model evidence

Arithmetic estimator

$$\mathbf{p(y)} = \mathbf{E}[\mathbf{p(y|\theta)}]_{\mathbf{p(\theta)}} \approx \frac{\mathbf{1}}{\mathbf{n}} \sum \mathbf{p}(\mathbf{y}|\boldsymbol{\theta}_i)$$

Samples from prior

$$\boldsymbol{\theta}_i \sim \mathbf{p(\theta)}$$

*Harmomic estimator, Gibb's estimator,*

*Annealed importance sampling, etc.*

## Posterior moments

Mean

$$\boldsymbol{\mu} = \mathbf{E}[\boldsymbol{\theta}]_{\mathbf{p(\theta|y)}} \approx \frac{\mathbf{1}}{\mathbf{n}} \sum \boldsymbol{\theta}_i$$

Variance

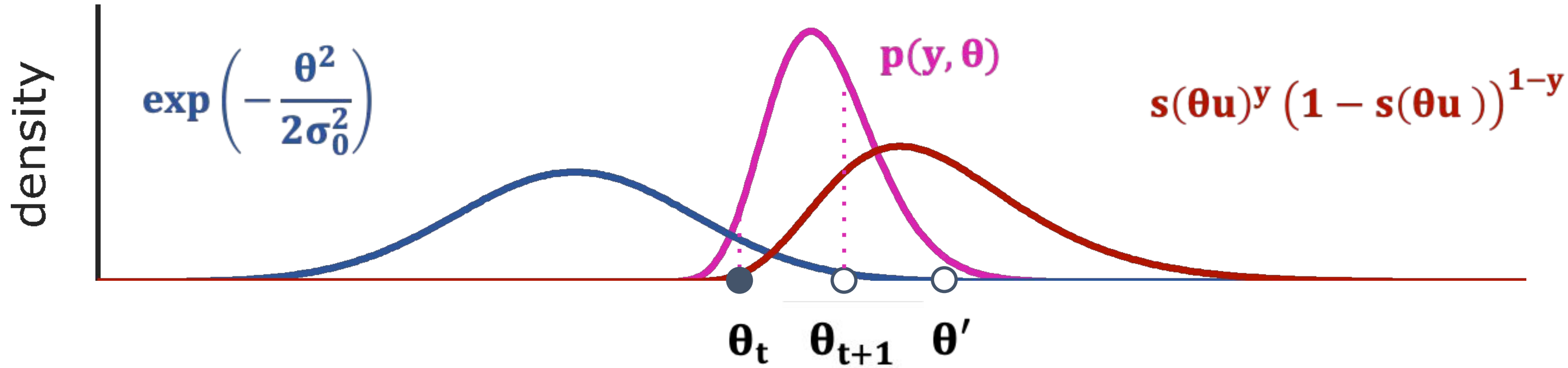
$$\boldsymbol{\Sigma} = \mathbf{E}\left[(\boldsymbol{\theta} - \boldsymbol{\mu})^2\right]_{\mathbf{p(\theta|y)}} \approx \frac{\mathbf{1}}{\mathbf{n}} \sum (\boldsymbol{\theta}_i - \hat{\boldsymbol{\mu}})^2$$

Samples *from posterior*

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

# Hot and cold game

# 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*

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

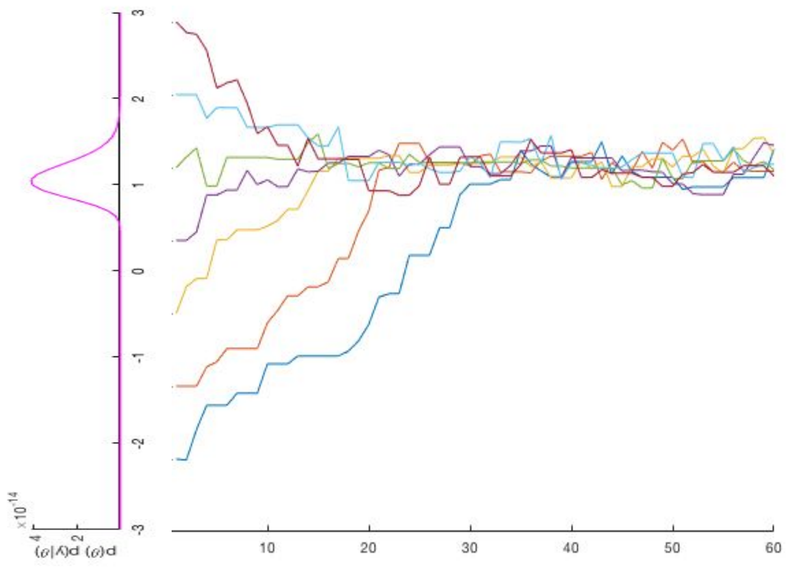
$\alpha < 1$

*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 (decimation)
- Step size (Goldilocks principle)

# Multivariate case

Write conditional posteriors

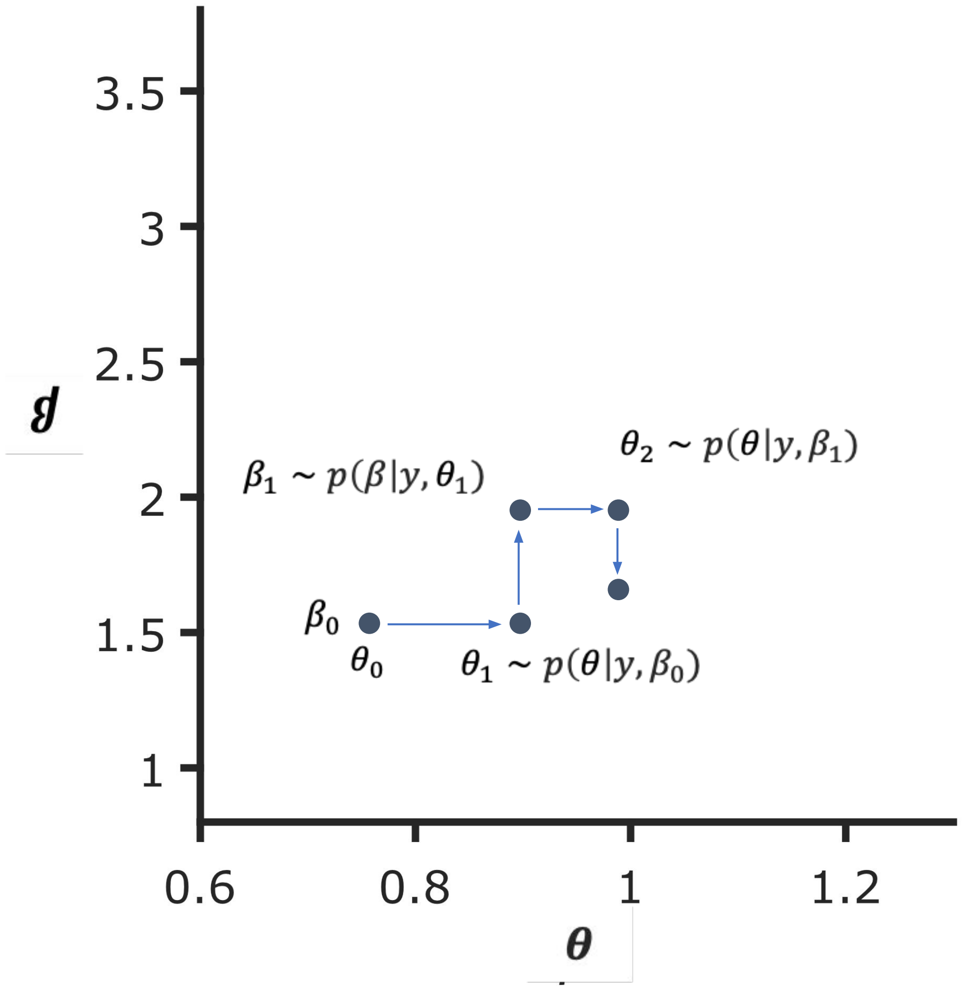
$$\mathbf{p}(\boldsymbol{\theta}|\mathbf{y},\boldsymbol{\beta}) = \frac{\mathbf{p}(\mathbf{y},\boldsymbol{\theta},\boldsymbol{\beta})}{\mathbf{p}(\mathbf{y},\boldsymbol{\beta})}$$

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

Iterative sampling

$$\boldsymbol{\theta}_t \sim \mathbf{p}(\boldsymbol{\theta}|\mathbf{y},\boldsymbol{\beta}_{t-1})$$

$$\boldsymbol{\beta}_t \sim \mathbf{p}(\boldsymbol{\beta}|\mathbf{y},\boldsymbol{\theta}_t)$$



# Multivariate case

Using the law of large numbers:

*Posterior mean*

$$\mathbf{E}[\boldsymbol{\theta}|\mathbf{y}] \approx \mathbf{mean}(\boldsymbol{\theta}_t)$$

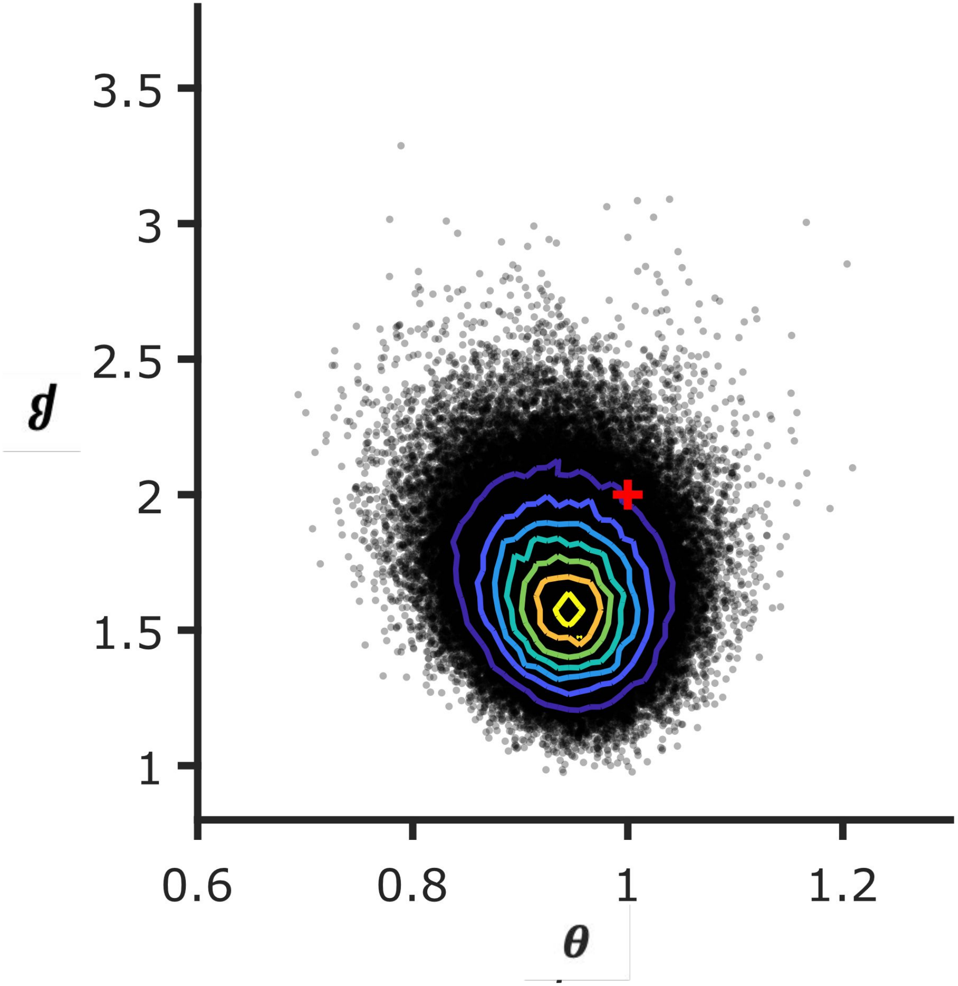
$$\mathbf{E}[\boldsymbol{\beta}|\mathbf{y}] \approx \mathbf{mean}(\boldsymbol{\beta}_t)$$

*Posterior variance*

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

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

*Covariance, etc.*



# Monte-Carlo inference

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

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

**Problems**

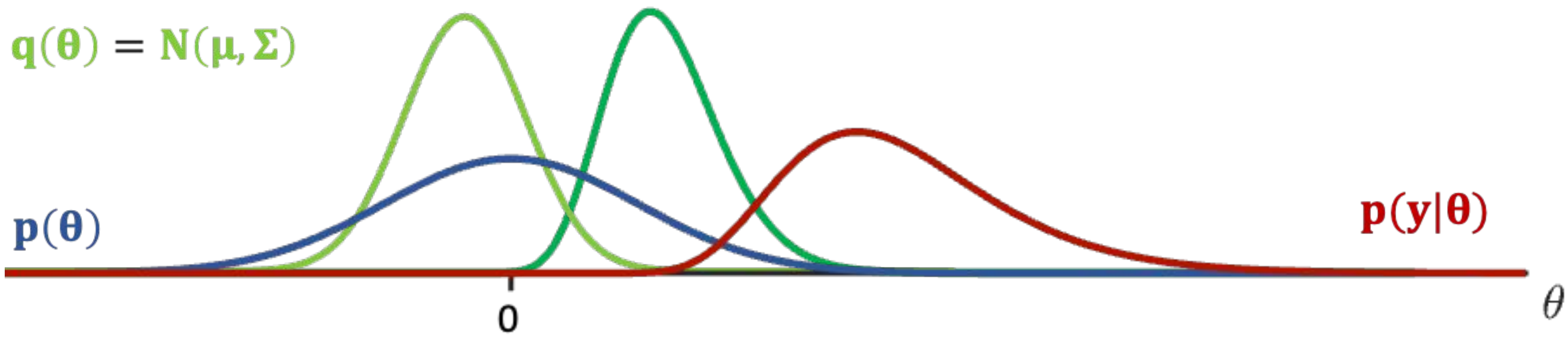
- computationally expensive
- does not scale well with the number of parameters
- hard to tune and diagnose
- no direct measure of model evidence

# Variational methods

*”[Variational inference is the thing you implement while you wait for your Monte-Carlo sampling to finish.]"*

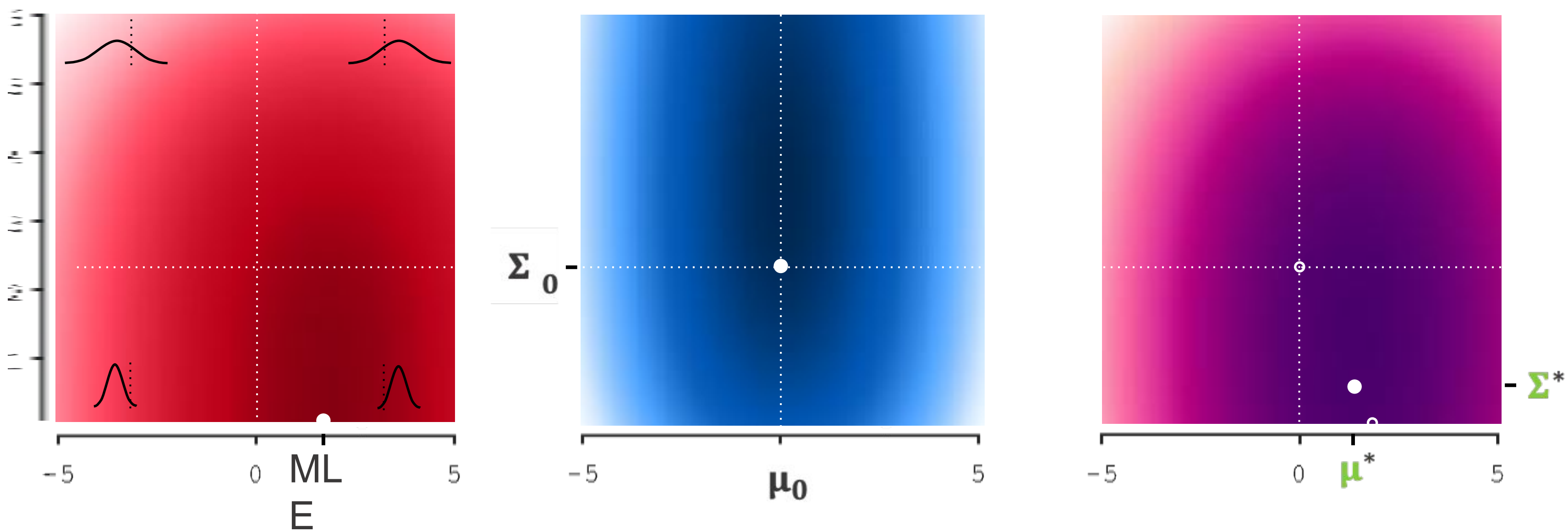
***David Blei***

# Approximate posterior




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

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



# Evidence LOwer Bound

candidate distribution $\mathbf{q}(\theta)$

Jensen's inequality

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

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

$$= \log \mathbf{E}\left[\frac{\mathbf{p}(\mathbf{y}, \theta)}{\mathbf{q}(\theta)}\right]_{\mathbf{q}(\theta)}$$

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

ELB
O

$< \mathbf{p}(\mathbf{y})$

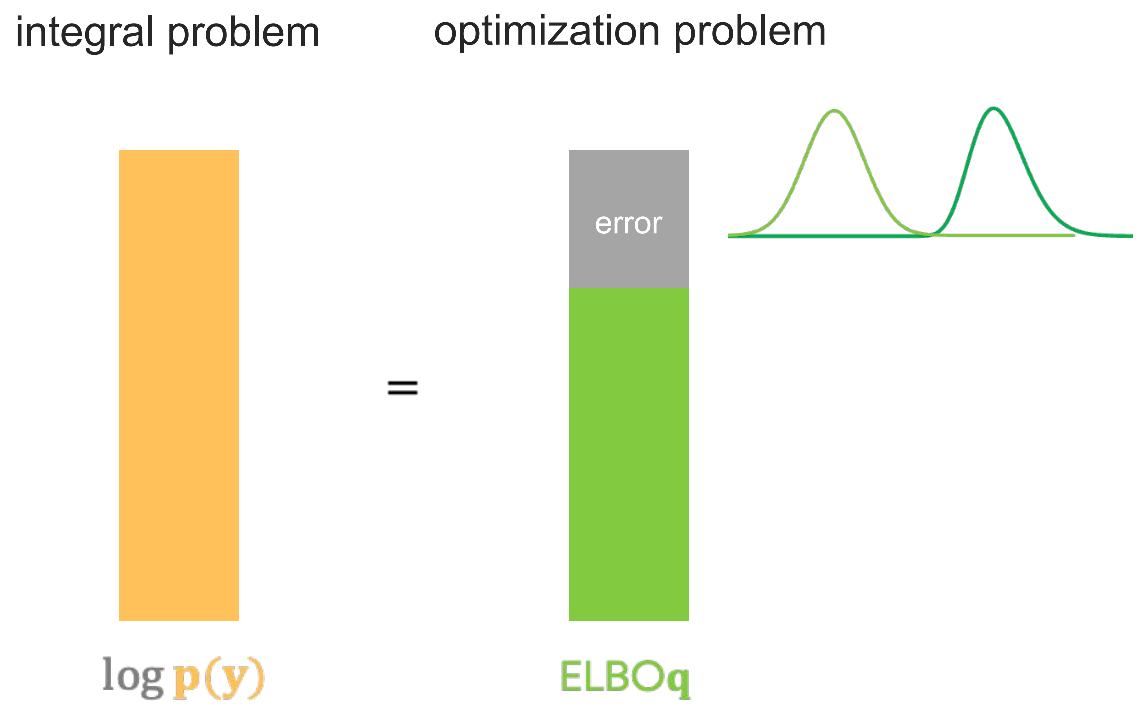
error

$> \mathbf{0}$

*"it's called the negative variational free energy"*

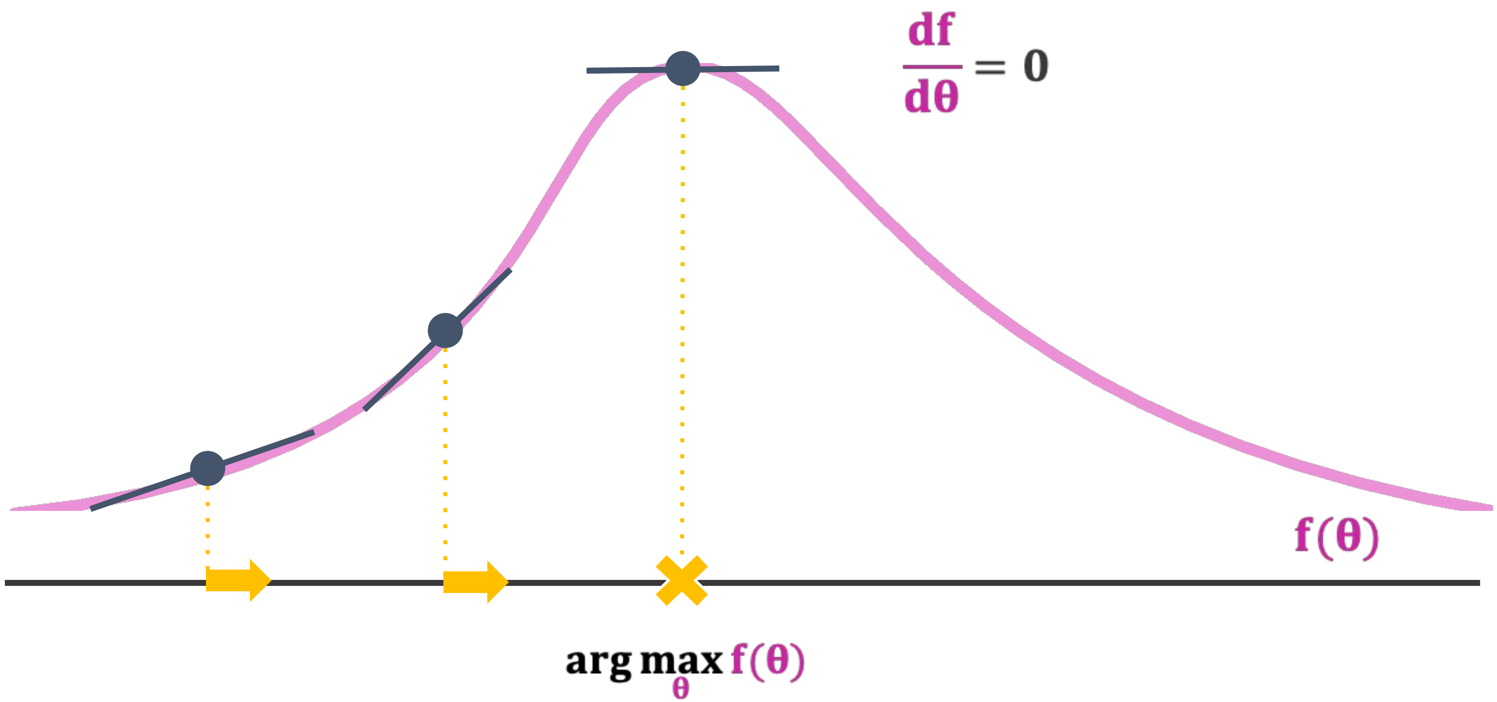
***Karl Friston***

# Evidence LOwer Bound



# Finding a maximum

analytical approach: *null derivative*
solve conditions for an extremum




numerical approach: *gradient ascent*
follow slope until it gets flat

# Maximizing the ELBO: numerical approach

## Objective

$$\log p(y) \approx \max \mathbf{E}\left[\log \frac{p(y,\theta)}{q(\theta)}\right]_{q(\theta)}$$

## Stochastic gradient

$$\nabla \mathrm{ELBO} = \mathbf{E}\left[\nabla \log q(\theta) \left(\log \frac{p(y,\theta)}{q(\theta)}\right)\right]_{q(\theta)}$$

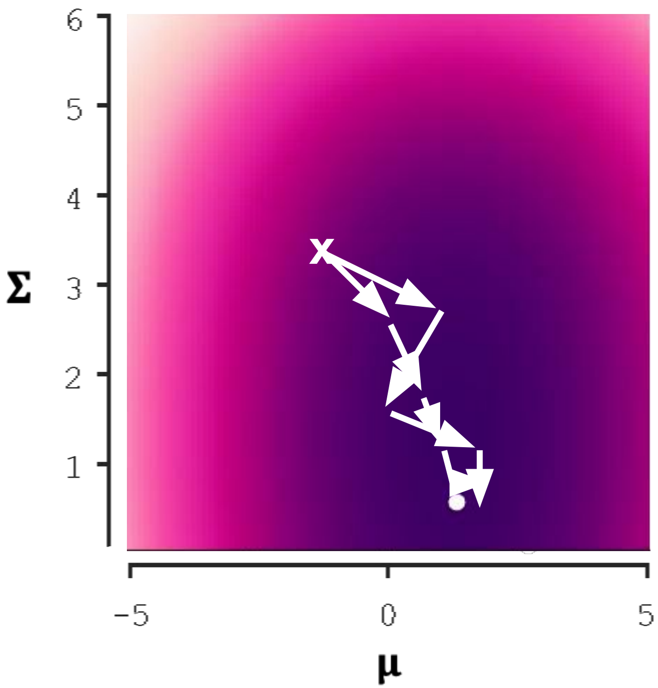
> $q$ derivable, samplable distribution

$$\theta_i^t \sim q^t(\theta) = N(\mu^t, \Sigma^t)$$

MC approximation of expetations

## Solution

Gradient ascent



# Maximizing the ELBO : analytical approach

## Objective

$$\log p(y) \approx \max \mathbf{E}\left[\log \frac{p(y,\theta)}{q(\theta)}\right]_{q(\theta)}$$

## Variational Laplace

> $q$ exponential family

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

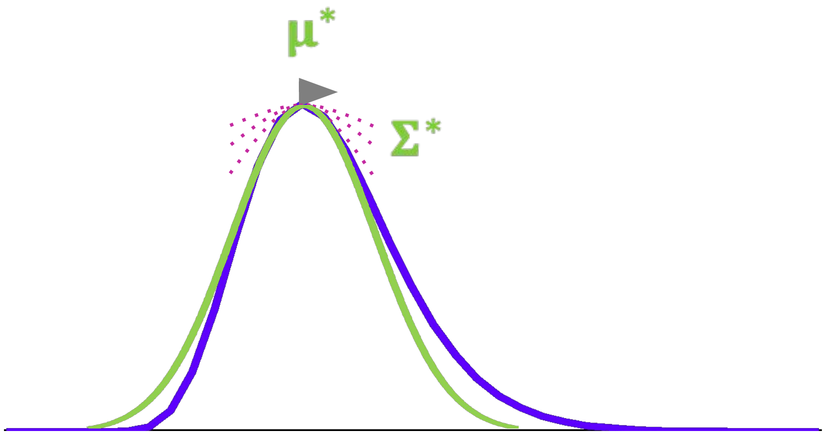
analytical approximation

$$\mathbf{ELBO} \approx \mathbf{ELBO}_{\mathrm{Laplace}}$$

find maximum

$$\frac{\mathbf{d}}{\mathbf{d}q(\theta)}\mathbf{ELBO}_{\mathrm{Laplace}} = 0$$

## Solution




$\mu^* = \mathrm{argmax}\; p(y, \theta)$ = **MAP**

$$\Sigma^* = -\left[\left.\frac{\partial^2}{\partial\theta^2}\right|_{\mu^*} \log p(y,\theta)\right]^{-1}$$

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

# Multivariate posterior

**Mean field approximation**

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

**Variational energy**

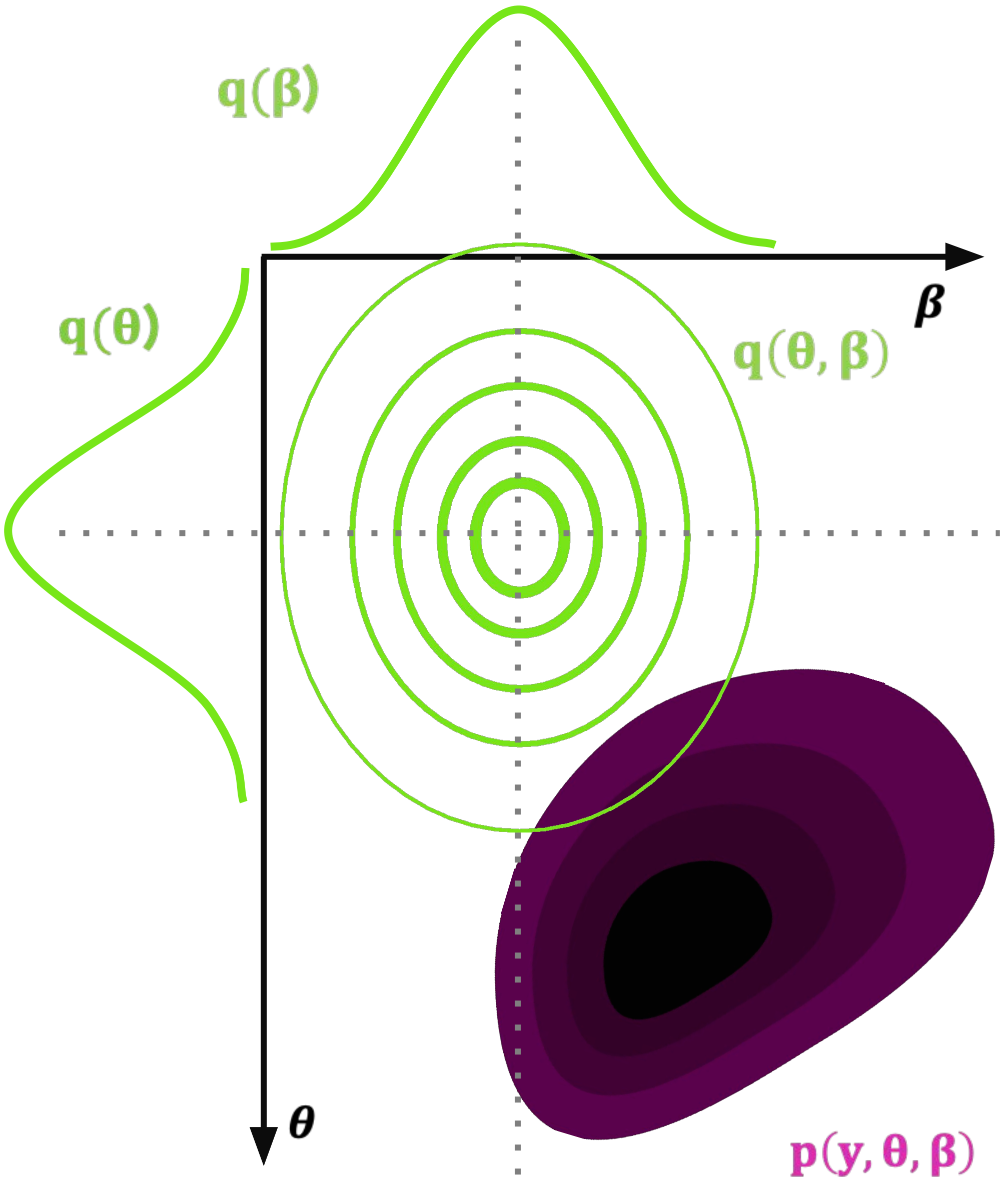
$$I(\theta) = E[\log p(y, \theta, \beta)]_{q(\beta)}$$

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

**Iterative optimization**

$$\mu_i = \operatorname{argmax} I(\theta_i)$$

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



# Multivariate posterior

**Mean field approximation**

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

**Variational energy**

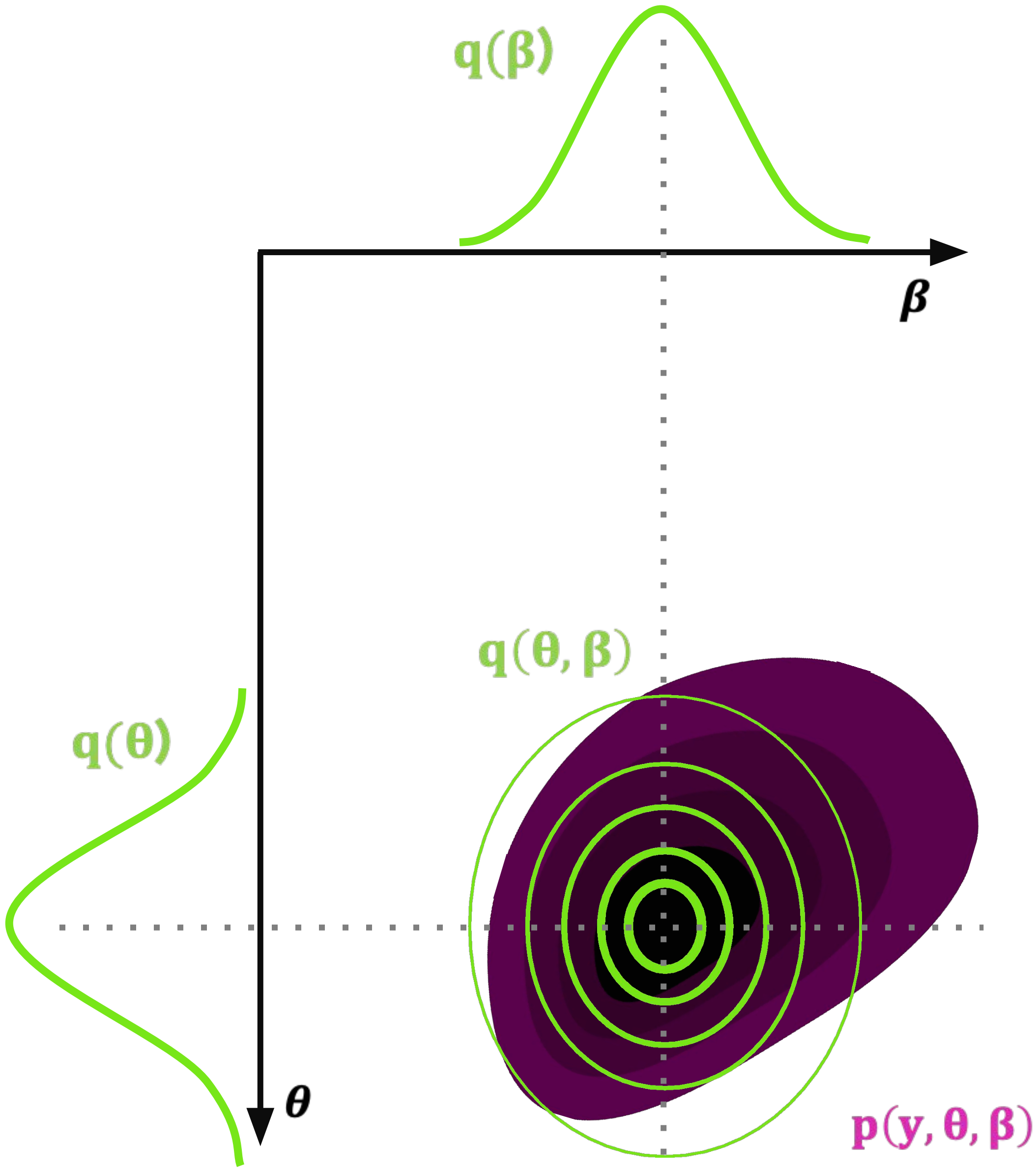
$$I(\theta) = E[\log p(y, \theta, \beta)]_{q(\beta)}$$

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

**Iterative optimization**

$$\mu_i = \operatorname{argmax} I(\theta_i)$$

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



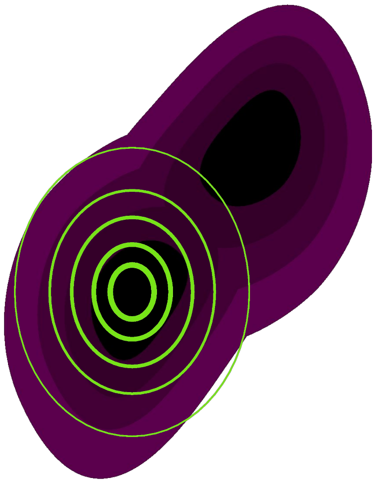
# Variational inference

Approximate posterior with a parametric distribution and find the parametrization which maximizes the ELBO.

This requires multiple approximations (Jensen, shape of the posterior, Laplace & mean-field) to be tractable.

**Problems**

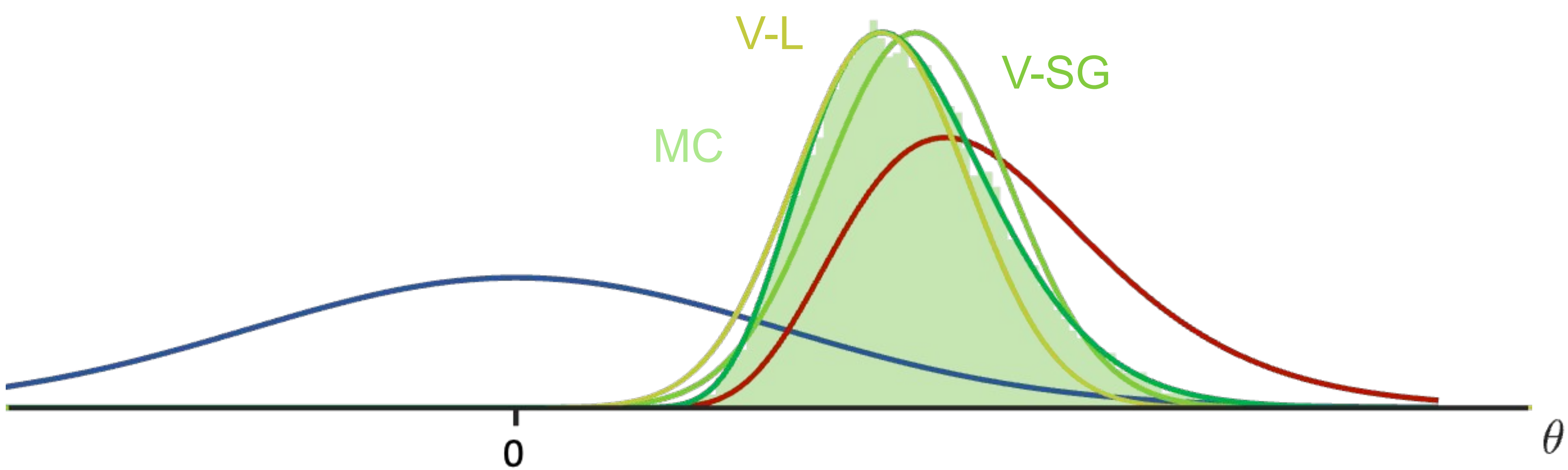
- does not converge to the true posterior
- can get stuck in local optimum



## 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. Good for **small models / scarce data**.
- ✔ **Variational methods** are fast & scalable computations of an approximation of the posterior. Good for **large models / large data**.



# Software

## Variational

- VBA-toolbox
- TAPAS
- SPM

## Sampling

- STAN
- BUGS
- JAGS
- hBayesDM
- hddm

# VBA Toolbox

377 published papers

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

Online wiki + Q&A

Simulation

Inversion (single subject, hierarchical)

Model selection (families, btw groups, btw conditions)

Visual diagnostics

Design optimization, multisession, multimodal observations, …

Need only the model description!

# Thank you!

**Online supplementary material**

*github.com/lionel-rigoux/tutorial-bayesian-inference*

- interactive app
- code of all algorithms
- selected references

**VBA-Toolbox**

*mbb-team.github.io/VBA-toolbox*

**Easy and reproducible writing workflow**

*pandemics.gitlab.io*