



Bayesian model inversion and selection

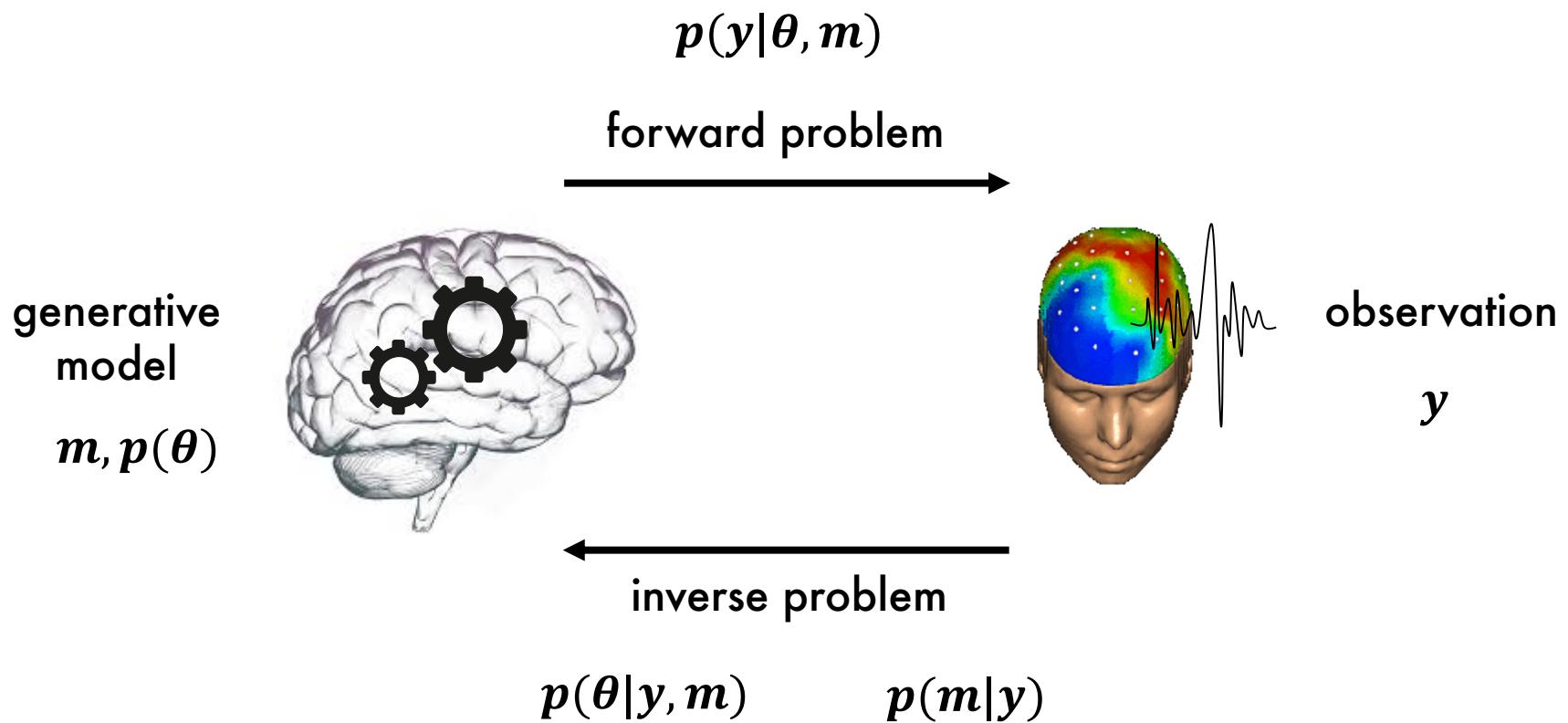
Lionel Rigoux

Translational Neuro-Circuitry (TNC) Cologne

Translational Neuromodeling Unit (TNU) Zürich



Overview



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

Marginal likelihood

Model evidence

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

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

$$p(y|m)$$

Compute the posterior and the evidence for a model

Monte-Carlo (sampling) methods

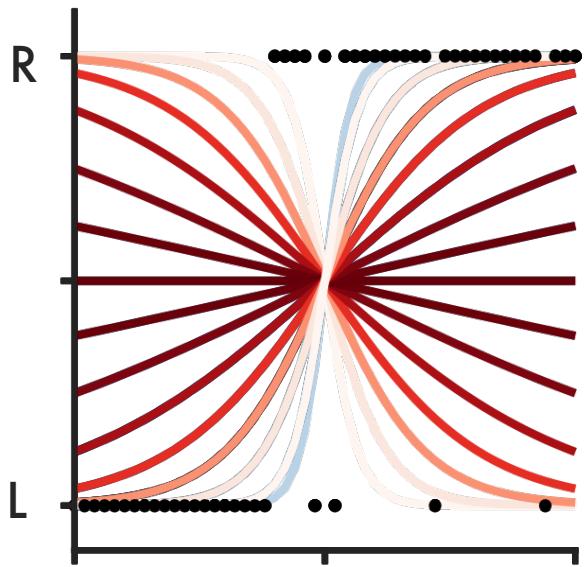
Variational methods

Good practices

online supplementary material

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

Example: logistic regression



Model prediction:

$$p(y=1|\theta, \beta) = \text{sig}(\theta u + \beta) = s$$

Likelihood:

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

Prior:

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

$$p(\beta) = \mathcal{N}(0, 0)$$

Sensitivity to orientation?



Bias?

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



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

$$E[(z - 3.5)^2] = \sum p(z)(z - 3.5)^2 = 2.9167$$



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

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

Law of
Large Numbers

Model evidence

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

Posterior moments

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

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

A little game

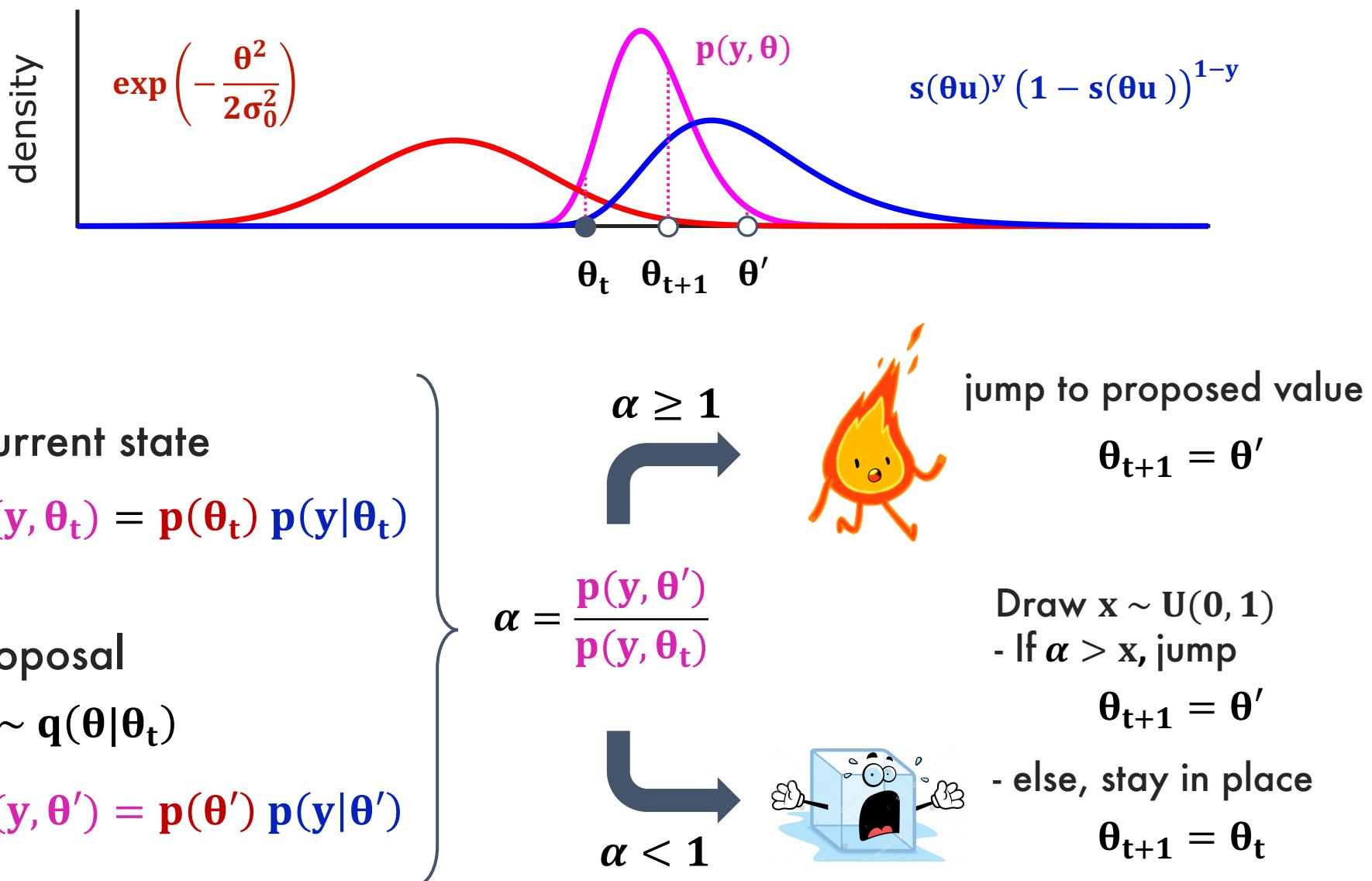
The joint as an un-normalized posterior:

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

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

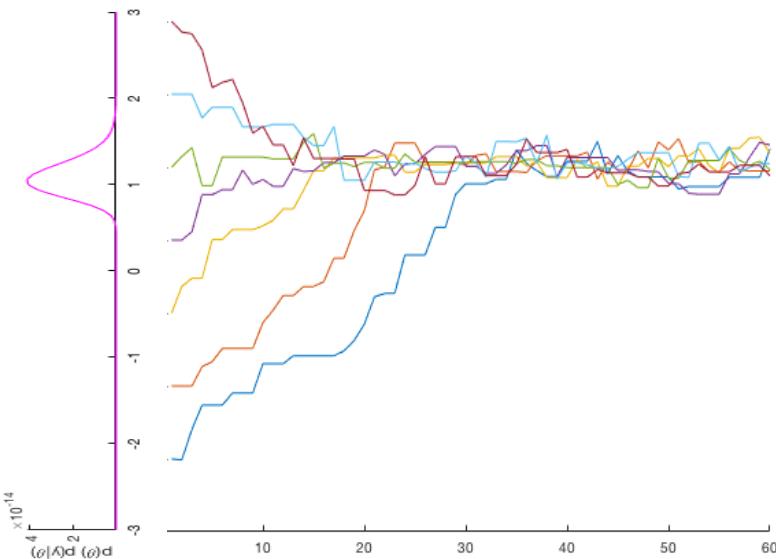


Metropolis-Hastings algorithm



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

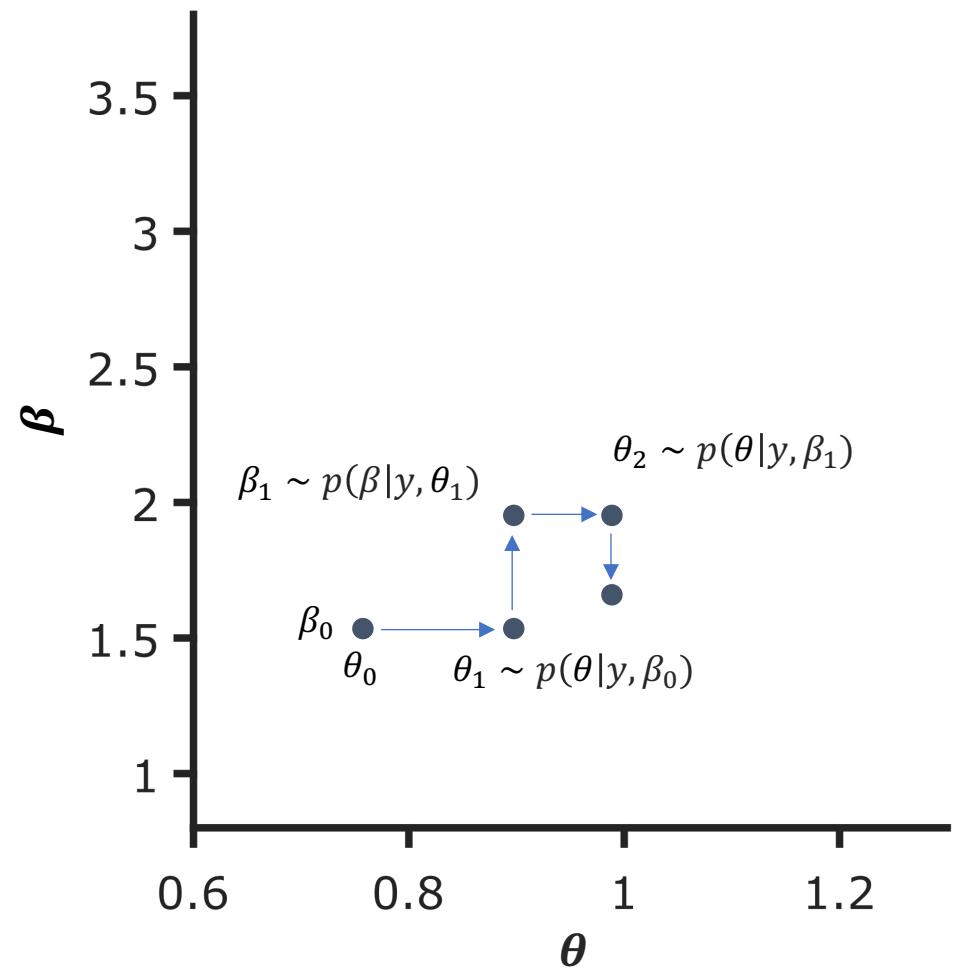
$$p(\theta|y, \beta) = \frac{p(y, \theta, \beta)}{p(y, \beta)}$$

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

Iterative sampling

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

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



Multivariate case

Using the law of large numbers:

- Posterior mean

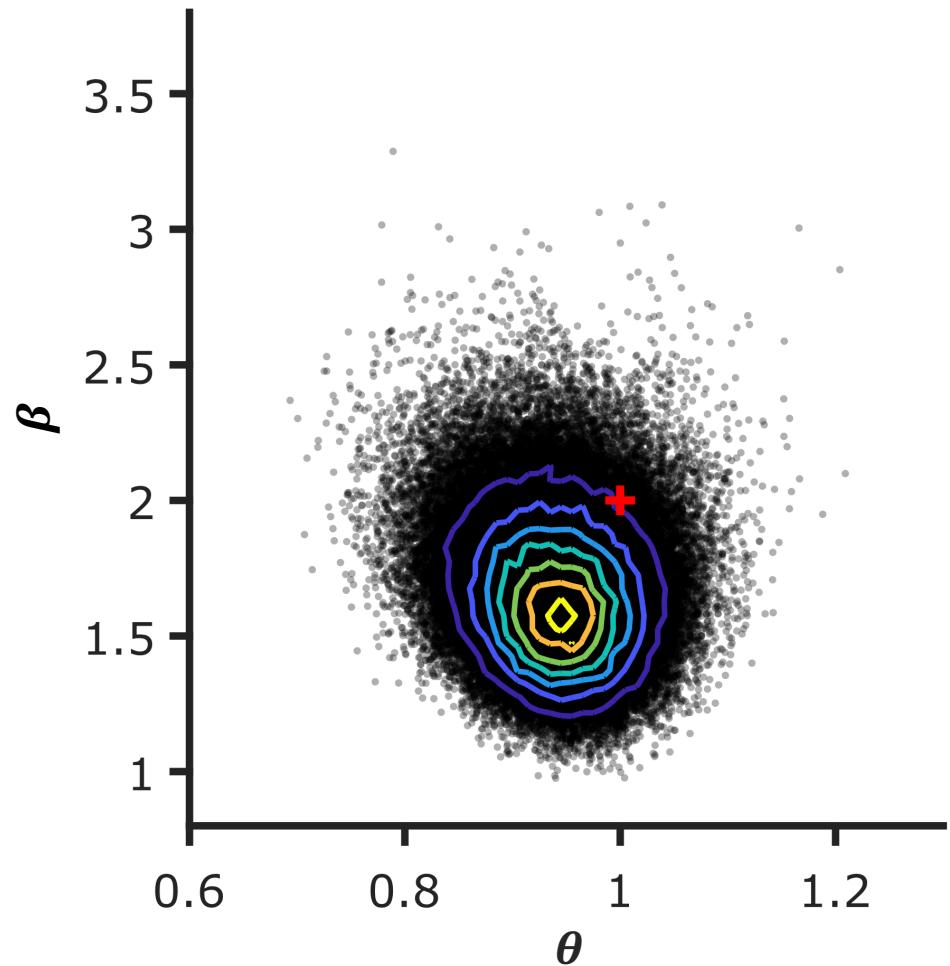
$$E[\theta|y] \approx \text{mean}(\theta_t)$$

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

- Posterior variance

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

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



Monte-Carlo inference

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

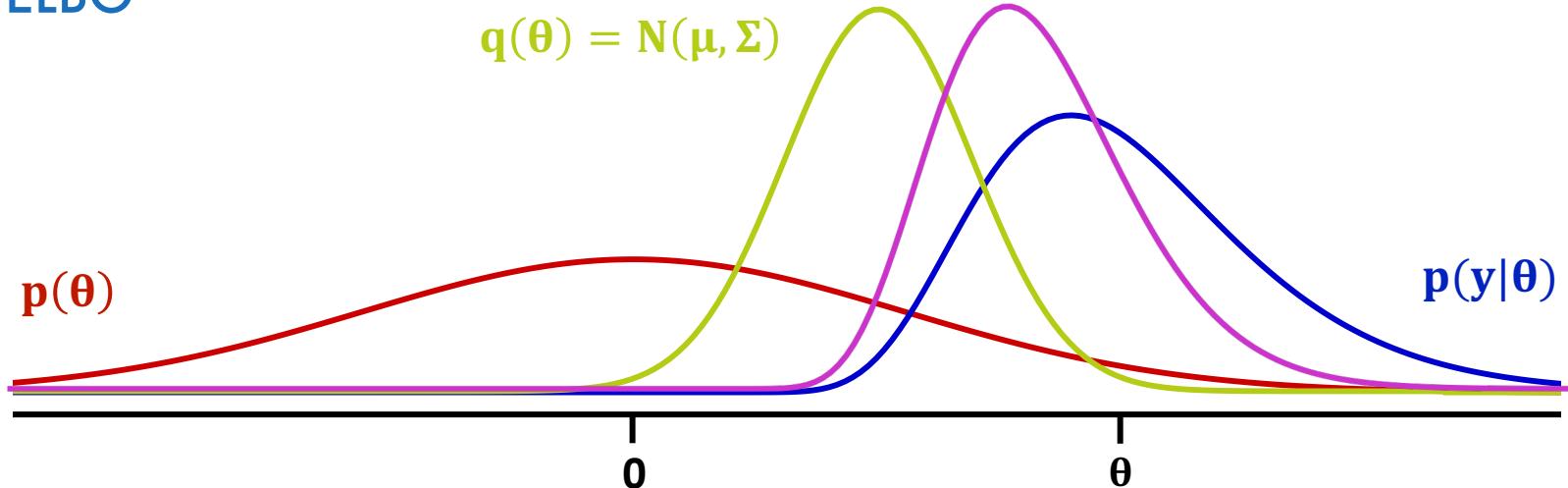
> The Law of Large Numbers guarantees that the sufficient statistics of the samples will converge to the true posterior moments.

Problems:

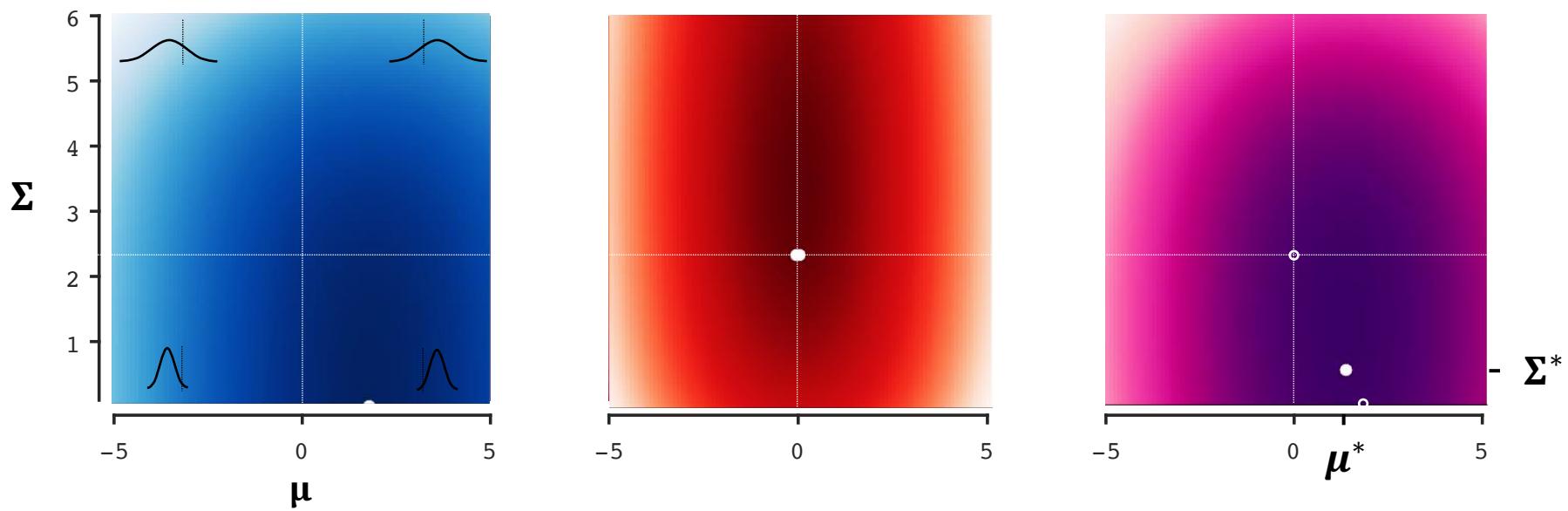
- computationally expensive
- does not scale well with the number of parameters
- no direct measure of model evidence
- hard to tune and diagnose

Variational Methods

The ELBO



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



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

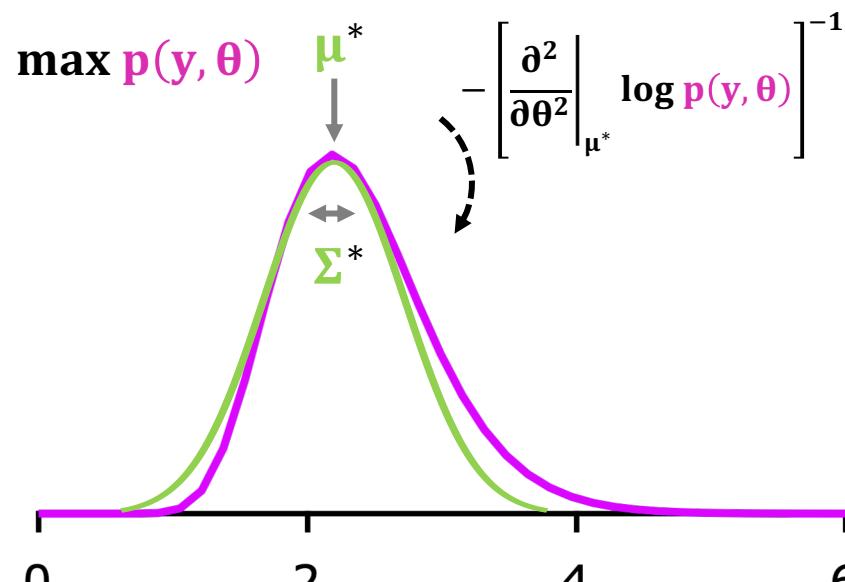
Variational Laplace

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

Stochastic gradient

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

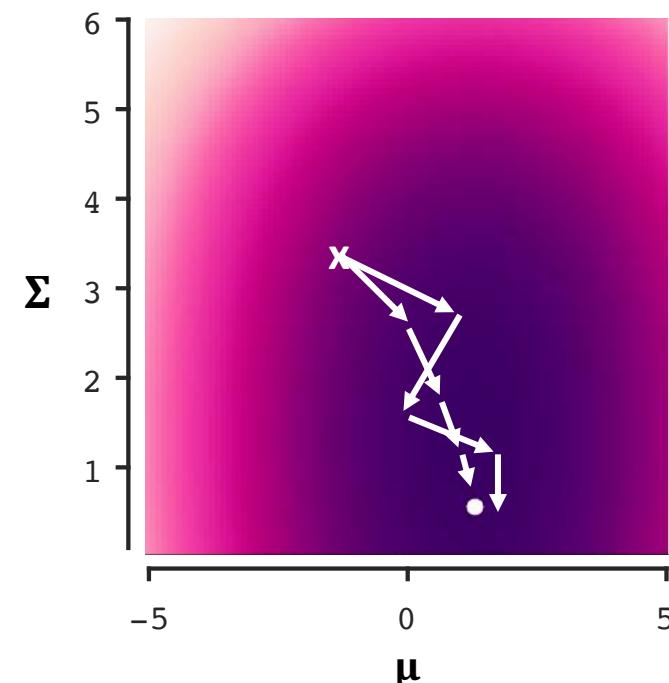
Find maximum: $\frac{d}{dq(\theta)} F_{\text{Laplace}} = 0$



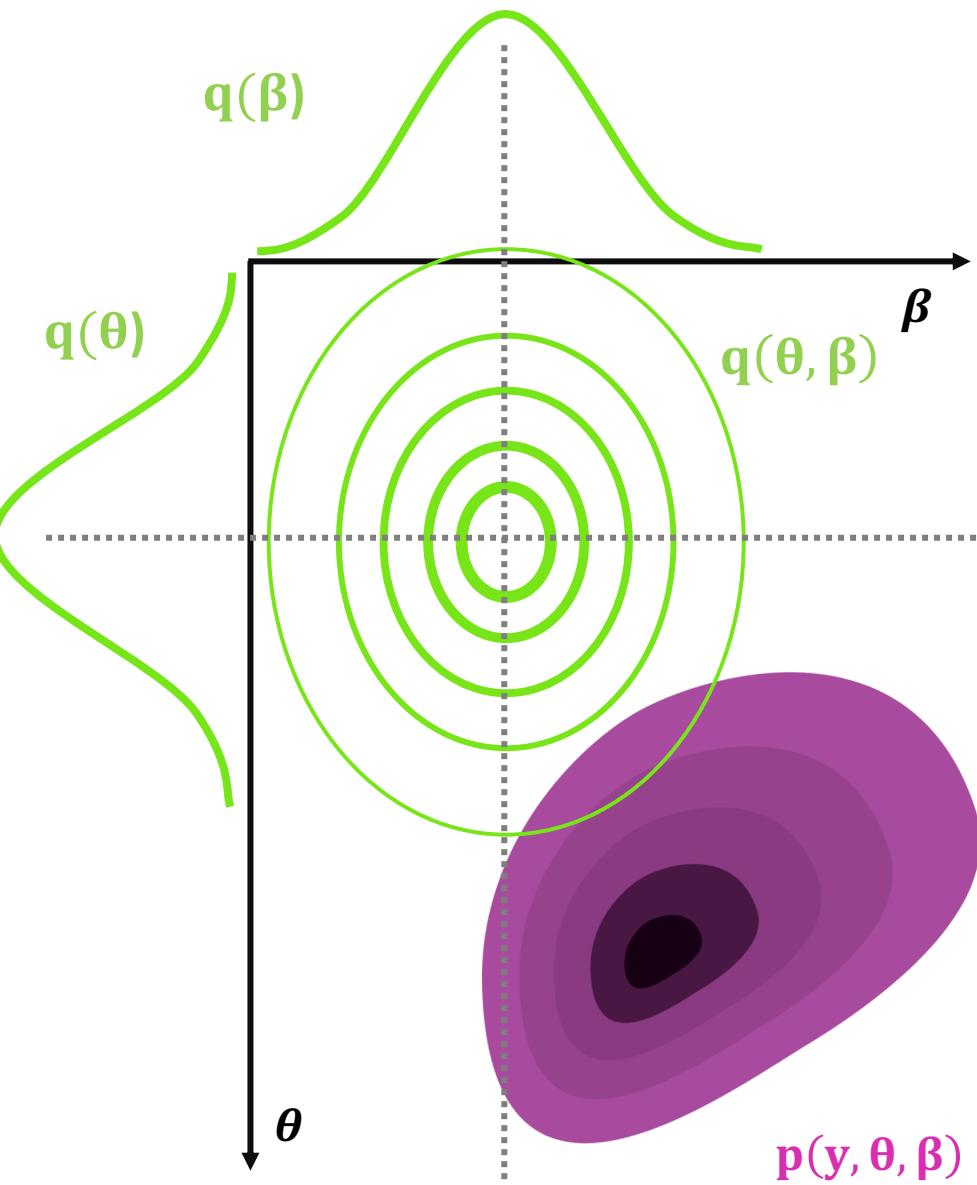
$$\log p(y) \approx \log p(y, \mu^*) + \frac{1}{2} [\log |\Sigma^*| + n_\theta \log(2\pi)]$$

Find maximum: gradient ascent

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



Multivariate posterior



Mean field approximation

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

$$q(\boldsymbol{\theta}, \boldsymbol{\beta}) \approx q(\boldsymbol{\theta})q(\boldsymbol{\beta})$$

Maximise Variational energy

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

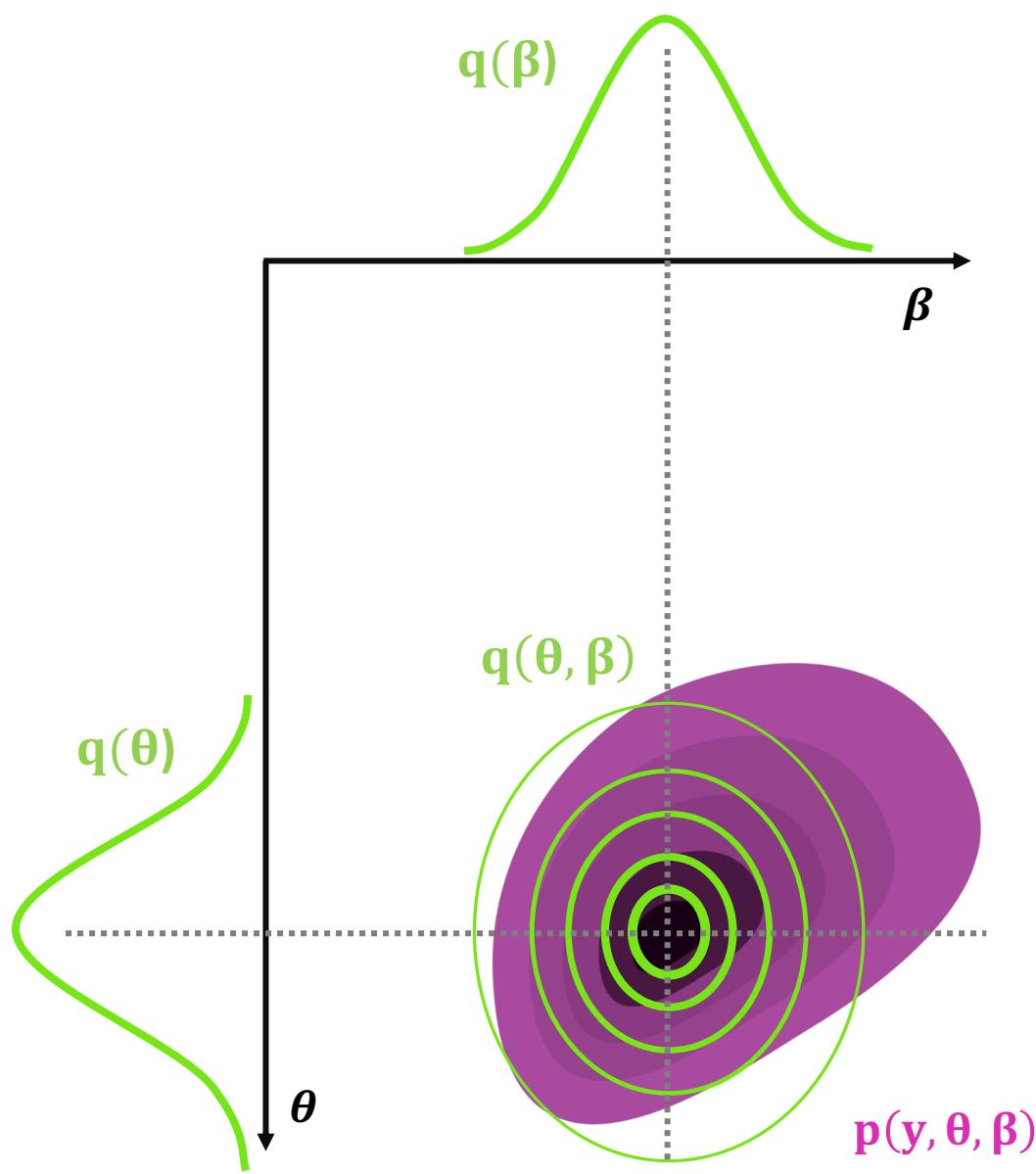
$$\approx \log p(\mathbf{y}, \boldsymbol{\theta}, \boldsymbol{\mu}_{\boldsymbol{\beta}}) + \dots$$

Iterative optimization

$$\boldsymbol{\mu}_i = \text{argmax } I(\boldsymbol{\theta}_i)$$

$$\boldsymbol{\Sigma}_i = - \left[\frac{\partial^2}{\partial \boldsymbol{\theta}_i^2} \Bigg|_{\boldsymbol{\mu}_i} I(\boldsymbol{\theta}_i) \right]^{-1}$$

Multivariate posterior



Mean field approximation

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

$$q(\boldsymbol{\theta}, \boldsymbol{\beta}) \approx q(\boldsymbol{\theta}) q(\boldsymbol{\beta})$$

Maximise Variational energy

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

$$\approx \log p(\mathbf{y}, \boldsymbol{\theta}, \mu_{\boldsymbol{\beta}}) + \dots$$

Iterative optimization

$$\mu_i = \text{argmax } I(\boldsymbol{\theta}_i)$$

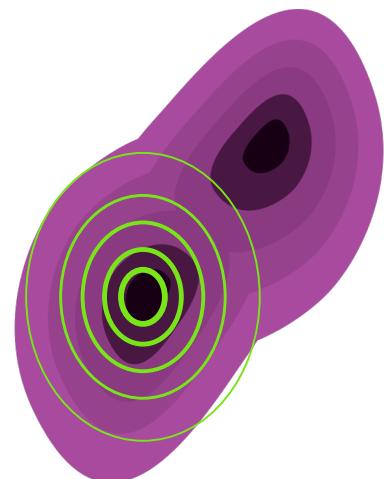
$$\Sigma_i = - \left[\frac{\partial^2}{\partial \theta_i^2} \Big|_{\mu_i} I(\boldsymbol{\theta}_i) \right]^{-1}$$

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.

Software

Variational

VBA-toolbox

TAPAS

SPM

Sampling

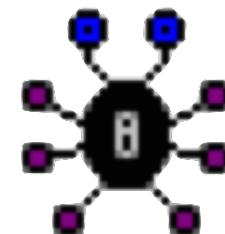
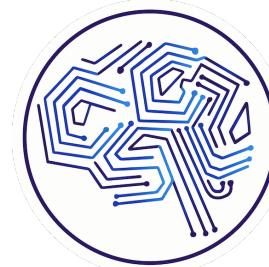
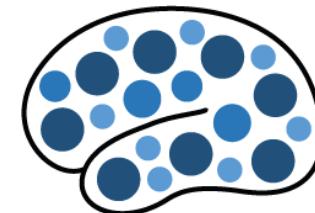
STAN

BUGS

JAGS

hBayesDM

hddm



JAGS

178 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

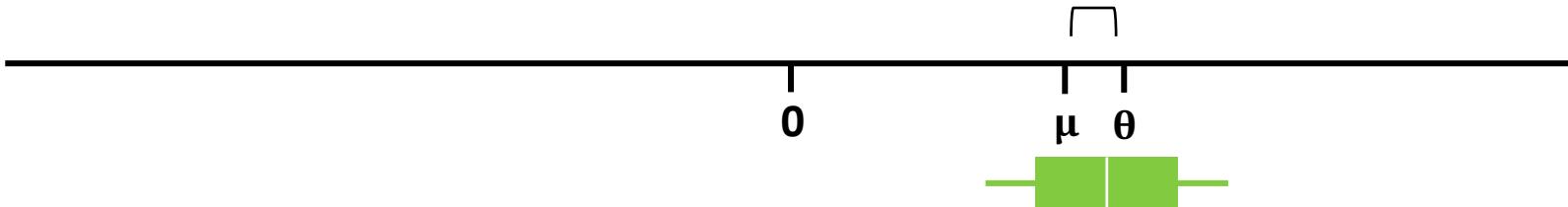
Design optimization, multisession, multimodal observations, ...



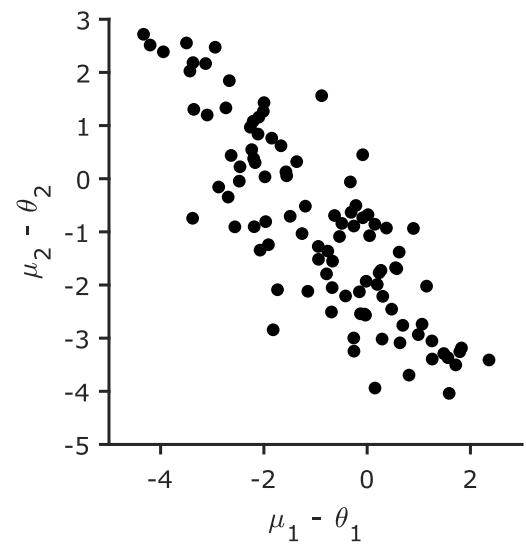
Need only the model description!

Validating your pipeline: parameter identifiability

estimation error
observation noise + numerical method



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



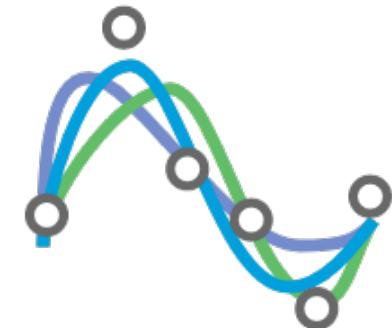
Thank you!

Online supplementary material

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

VBA-Toolbox

mbb-team.github.io/VBA-toolbox



Easy writing workflow

pandemics.gitlab.io



Variational variants

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

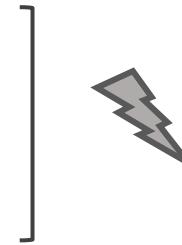
Jensen's
inequality



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

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

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



approximation error

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

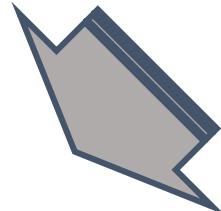


need to maximise wrt q

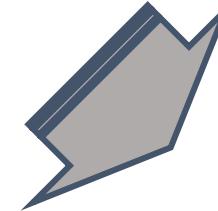


Assume parametric form

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



Analytical approximation



Sampling

Variational Laplace

Stochastic gradient