Bayesian inference and Bayesian model selection

Klaas Enno Stephan

Translational Neuromodeling Unit

Universität Zürich[UZH]

Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Lecture as part of "Methods & Models for fMRI data analysis",
University of Zurich & ETH Zurich, 28 November 2017

With slides from and many thanks to:

Kay Brodersen,

Will Penny,

Sudhir Shankar Raman

Why should I know about Bayesian inference?

Because Bayesian principles are fundamental for

- **statistical inference** in general

- **system identification**

- **translational neuromodeling** ("computational assays")
  - computational psychiatry
  - computational neurology

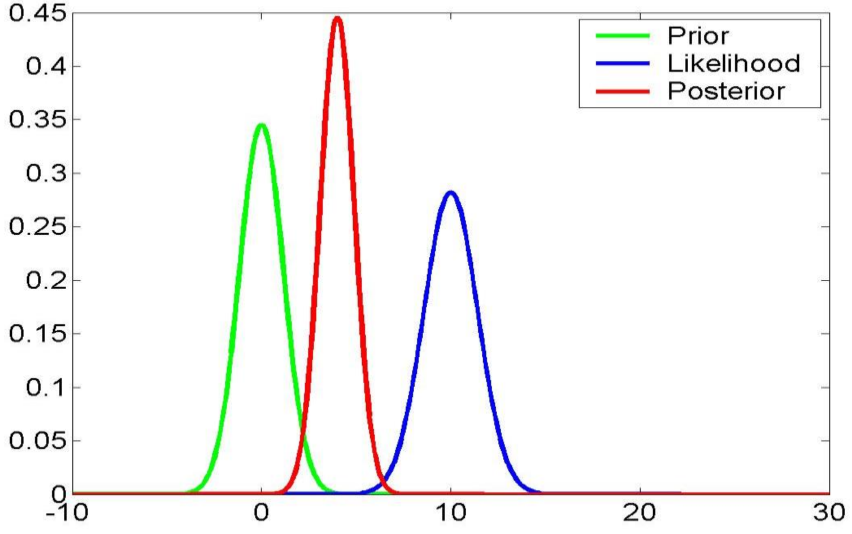
- contemporary **theories of brain function** (the "Bayesian brain")
  - predictive coding
  - free energy principle
  - active inference

Why should I know about Bayesian inference?

Because Bayesian principles are fundamental for

- **statistical inference** in general
- **system identification**
- **translational neuromodeling** ("computational assays")
  - computational psychiatry
  - computational neurology
- contemporary **theories of brain function** (the "Bayesian brain")
  - predictive coding
  - free energy principle
  - active inference

Bayes' theorem



The Reverend Thomas Bayes
(1702–1761)

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

posterior = likelihood · prior / evidence

“Bayes‘ Theorem describes, how an ideally rational person processes information."

Wikipedia

Bayes' Theorem

Given data y and parameters θ, the joint probability is:

$$p(y,\theta) = p(\theta \mid y)\,p(y) = p(y \mid \theta)\,p(\theta)$$

Eliminating $p(y,\theta)$ gives Bayes' rule:

posterior

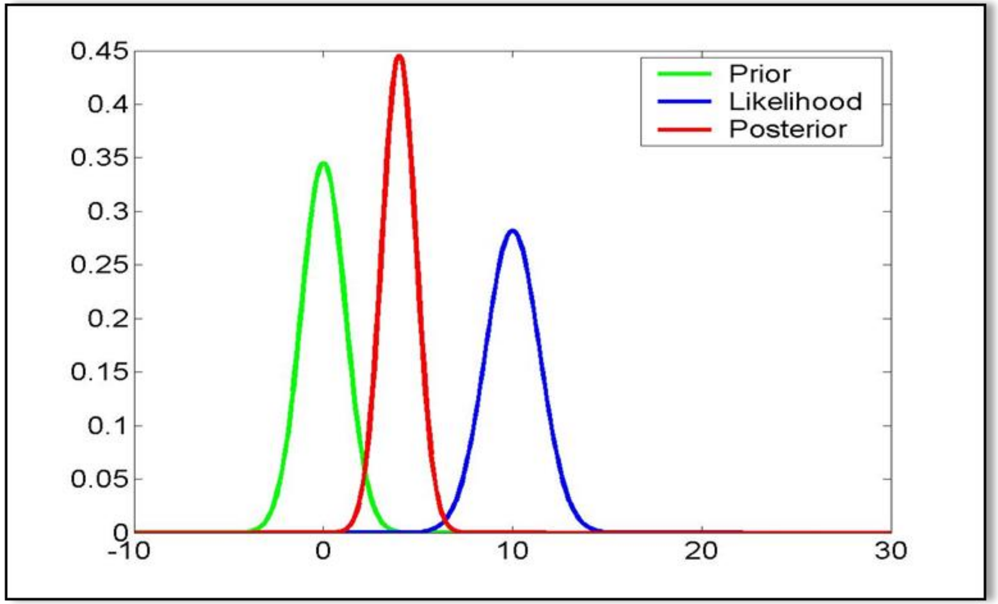
likelihood

prior

$$P(\theta \mid y) = \frac{p(y \mid \theta)\,p(\theta)}{p(y)}$$

evidence

Bayesian inference: an animation



Generative models

- specify a joint probability distribution over all variables (observations and parameters)

- require a likelihood function and a prior:

$$p(y,\theta \mid m) = p(y \mid \theta, m)\, p(\theta \mid m) \propto p(\theta \mid y, m)$$

- can be used to randomly generate synthetic data (observations) by sampling from the prior
  - we can check in advance whether the model can explain certain phenomena at all

- model comparison based on the model evidence

$$p(y \mid m) = \int p(y \mid \theta, m)\, p(\theta \mid m)\, d\theta$$

Principles of Bayesian inference

⇨ Formulation of a **generative model**

Model → likelihood function $p(y|\theta)$

Model → prior distribution $p(\theta)$

⇨ Observation of **data**

Measurement → data y

⇨ **Model inversion** – updating one's beliefs

$$p(\theta \mid y) \propto p(y \mid \theta)\, p(\theta)$$

→ maximum a posteriori (MAP) estimates

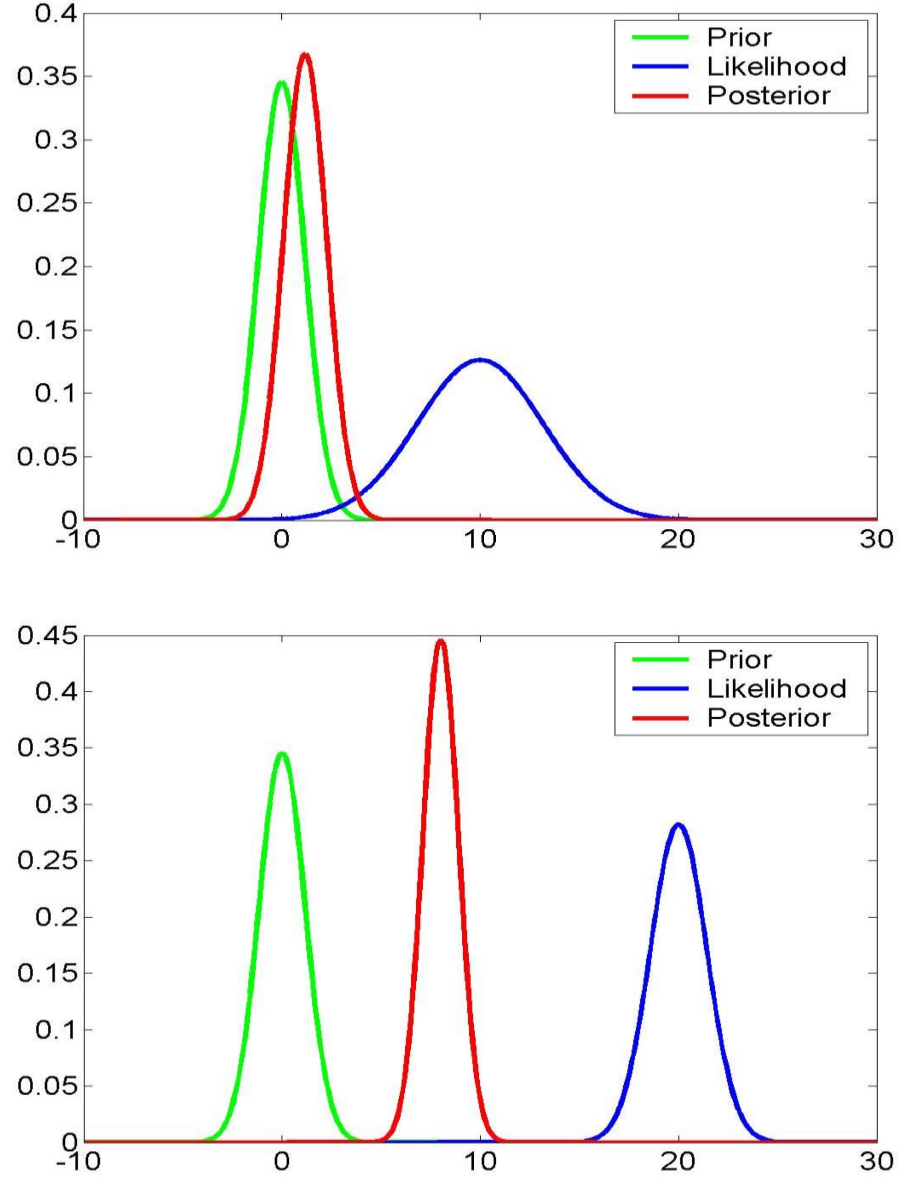
→ model evidence

Priors

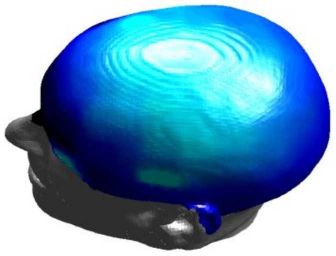
Priors can be of different sorts, e.g.

- empirical (previous data)
- empirical (estimated from current data using a hierarchical model → "empirical Bayes")
- uninformed
- principled (e.g., positivity constraints)
- shrinkage

Example of a shrinkage prior

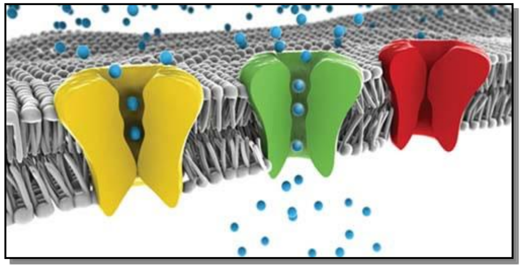


Generative models



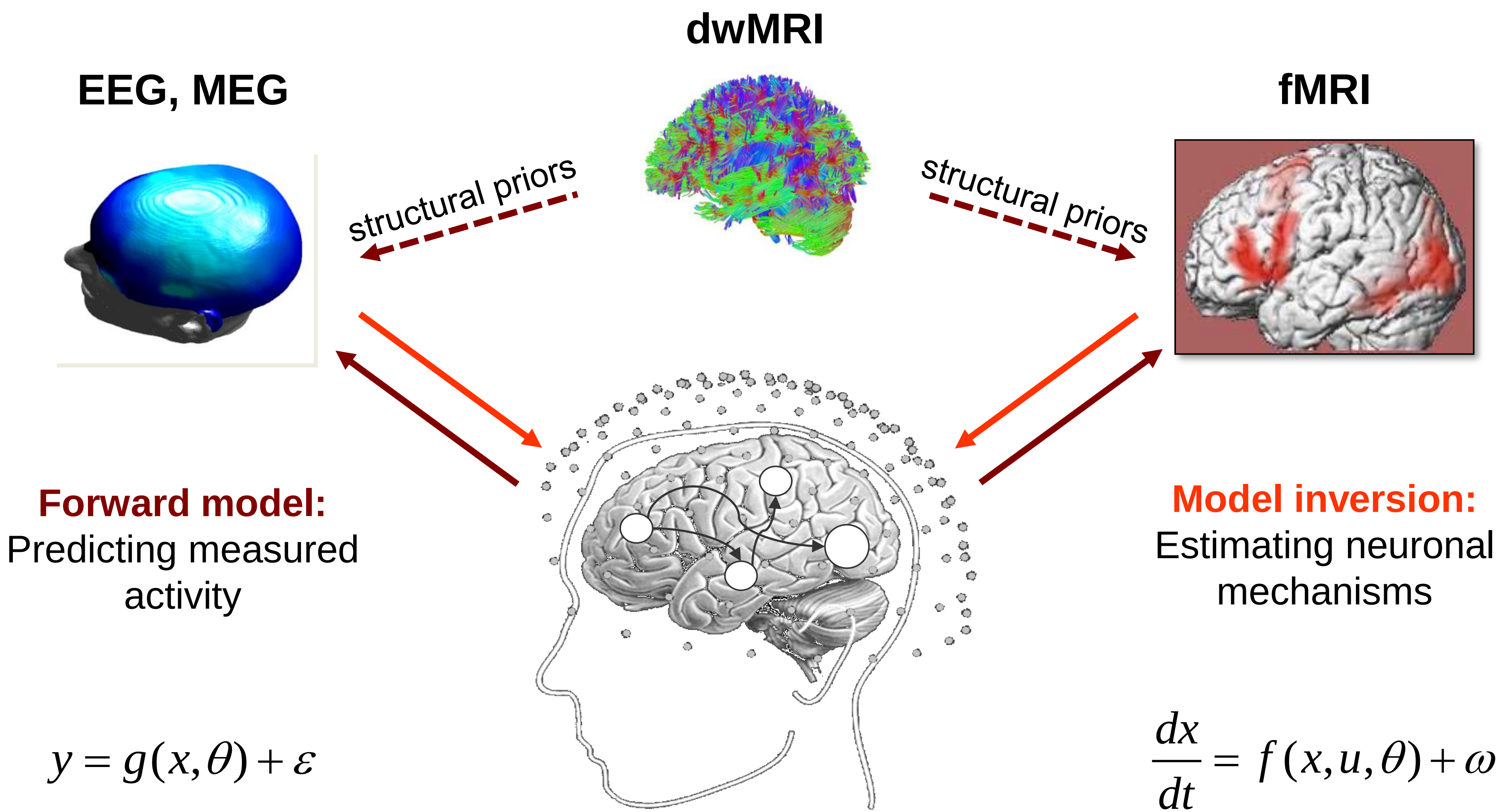
$$p(y|\theta,m) \cdot p(\theta|m)$$

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



1. enforce mechanistic thinking: how could the data have been caused?
2. generate synthetic data (observations) by sampling from the prior – can model explain certain phenomena at all?
3. inference about parameters → $p(\theta|y)$
4. inference about model structure → $p(y|m)$ or $p(m|y)$
5. model evidence: index of model quality

A generative modelling framework for fMRI & EEG: Dynamic causal modeling (DCM)



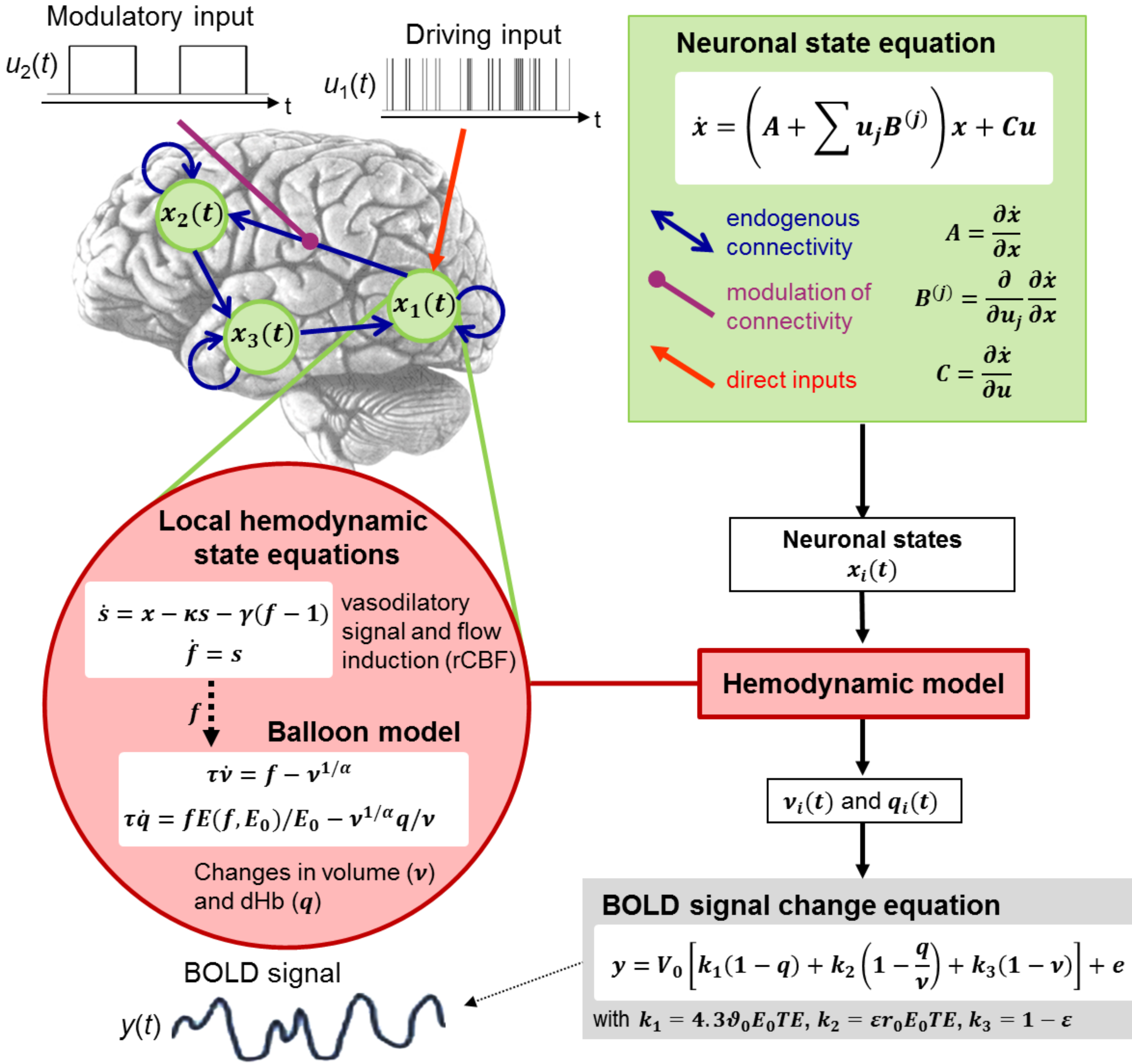
$$y = g(x, \theta) + \varepsilon$$

$$\frac{dx}{dt} = f(x, u, \theta) + \omega$$

Friston et al. 2003, *NeuroImage*

Stephan et al. 2009, *NeuroImage*

DCM for fMRI



Stephan et al. 2015, *Neuron*

Bayesian system identification

$$u(t)$$

Neural dynamics

$$dx/dt = f(x,u,\theta)$$

Observer function

$$y = g(x,\theta) + \varepsilon$$

$$p(y \mid \theta, m) = N(g(\theta), \Sigma(\theta))$$
$$p(\theta, m) = N(\mu_\theta, \Sigma_\theta)$$

Inference on model structure

$$p(y \mid m) = \int p(y \mid \theta, m)\, p(\theta)\, d\theta$$

Inference on parameters

$$p(\theta \mid y, m) = \frac{p(y \mid \theta, m)\, p(\theta, m)}{p(y \mid m)}$$

Design experimental inputs

Define likelihood model

Specify priors

Invert model

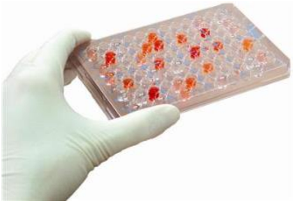
Make inferences

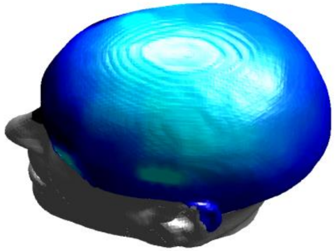
Why should I know about Bayesian inference?

Because Bayesian principles are fundamental for

- **statistical inference** in general
- **system identification**
- **translational neuromodeling** ("computational assays")
  - computational psychiatry
  - computational neurology
- contemporary **theories of brain function** (the "Bayesian brain")
  - predictive coding
  - free energy principle
  - active inference

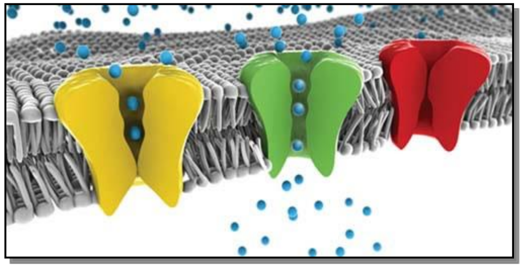
Generative models as "computational assays"

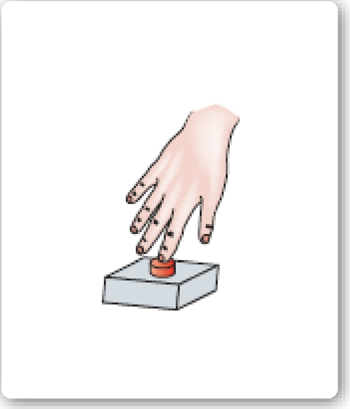




$$p(y \mid \theta, m) \cdot p(\theta \mid m)$$

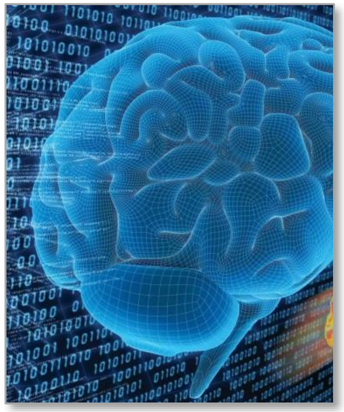
$$p(\theta \mid y, m)$$





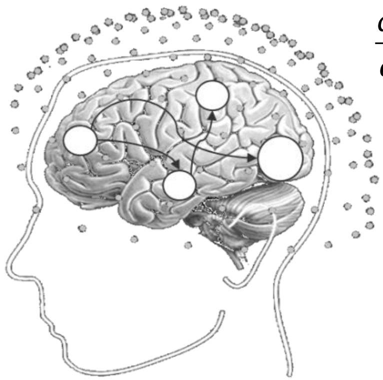
$$p(y \mid \theta, m) \cdot p(\theta \mid m)$$

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

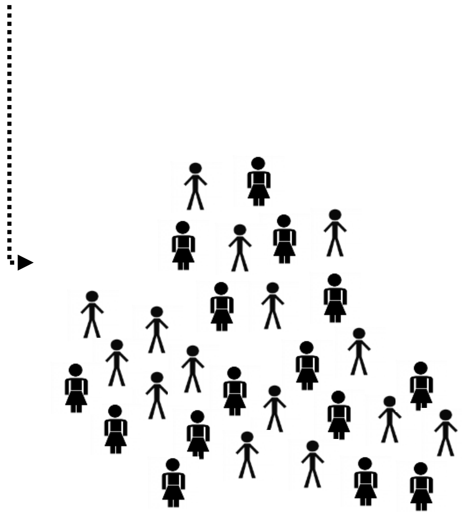


Translational Neuromodeling

❶ Computational assays:
Models of disease mechanisms

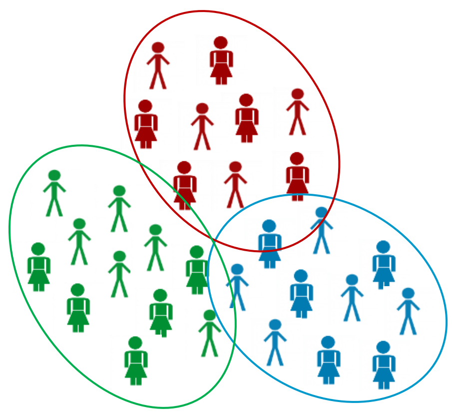


$$\frac{dx}{dt} = f(x,u,\theta) + \omega$$



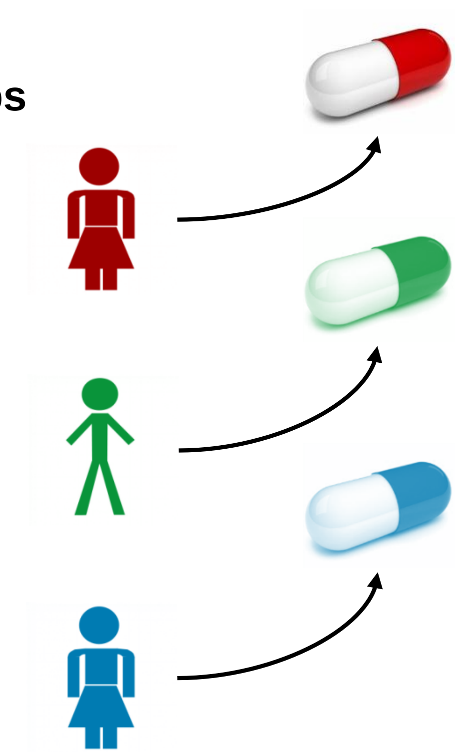
❷ Application to brain activity and behaviour of individual patients

❸ Detecting physiological subgroups (based on inferred mechanisms)



- disease mechanism A
- disease mechanism B
- disease mechanism C

❹ Individual treatment prediction

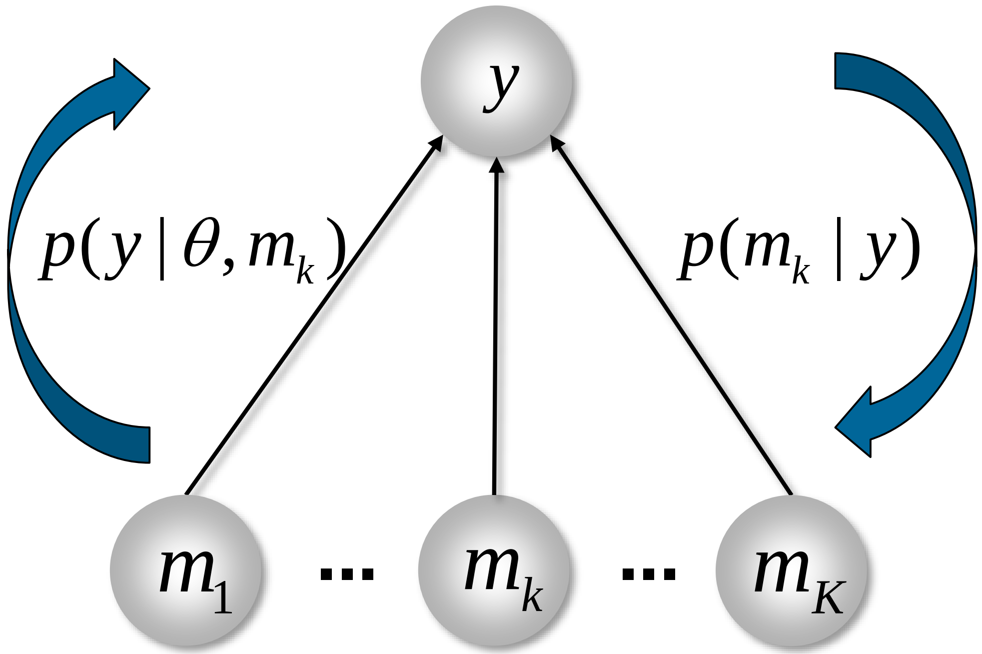


Stephan et al. 2015, *Neuron*

Differential diagnosis based on generative models of disease symptoms

SYMPTOM
(behaviour
or physiology)

HYPOTHETICAL
MECHANISM



$$p(m_k \mid \mathrm{y}) = \frac{p(y \mid m_k)\, p(m_k)}{\sum_k p(y \mid m_k)\, p(m_k)}$$

Stephan et al. 2016, *NeuroImage*

Why should I know about Bayesian inference?

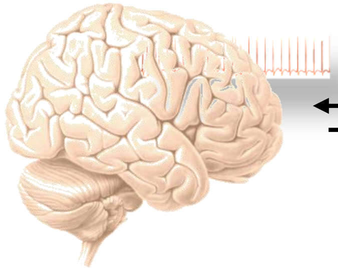
Because Bayesian principles are fundamental for

- **statistical inference** in general
- **system identification**
- **translational neuromodeling** ("computational assays")
  - computational psychiatry
  - computational neurology
- contemporary **theories of brain function** (the "Bayesian brain")
  - predictive coding
  - free energy principle
  - active inference

Perception = inversion of a hierarchical generative model

neuronal states

environm. states
others' mental states
bodily states



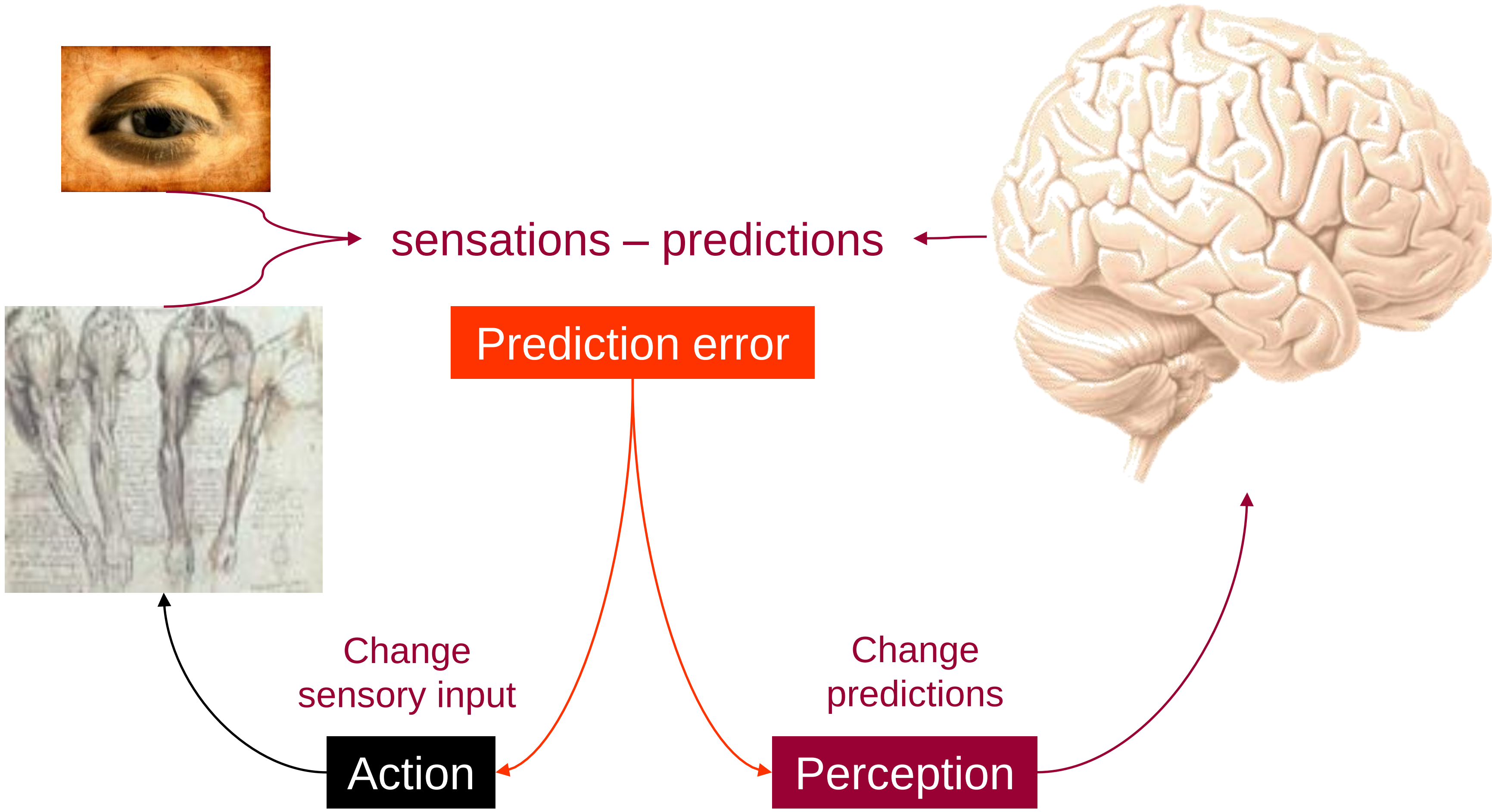
forward model

$$p(y \mid x, m) \cdot p(x \mid m)$$

$$p(x \mid y, m)$$

perception

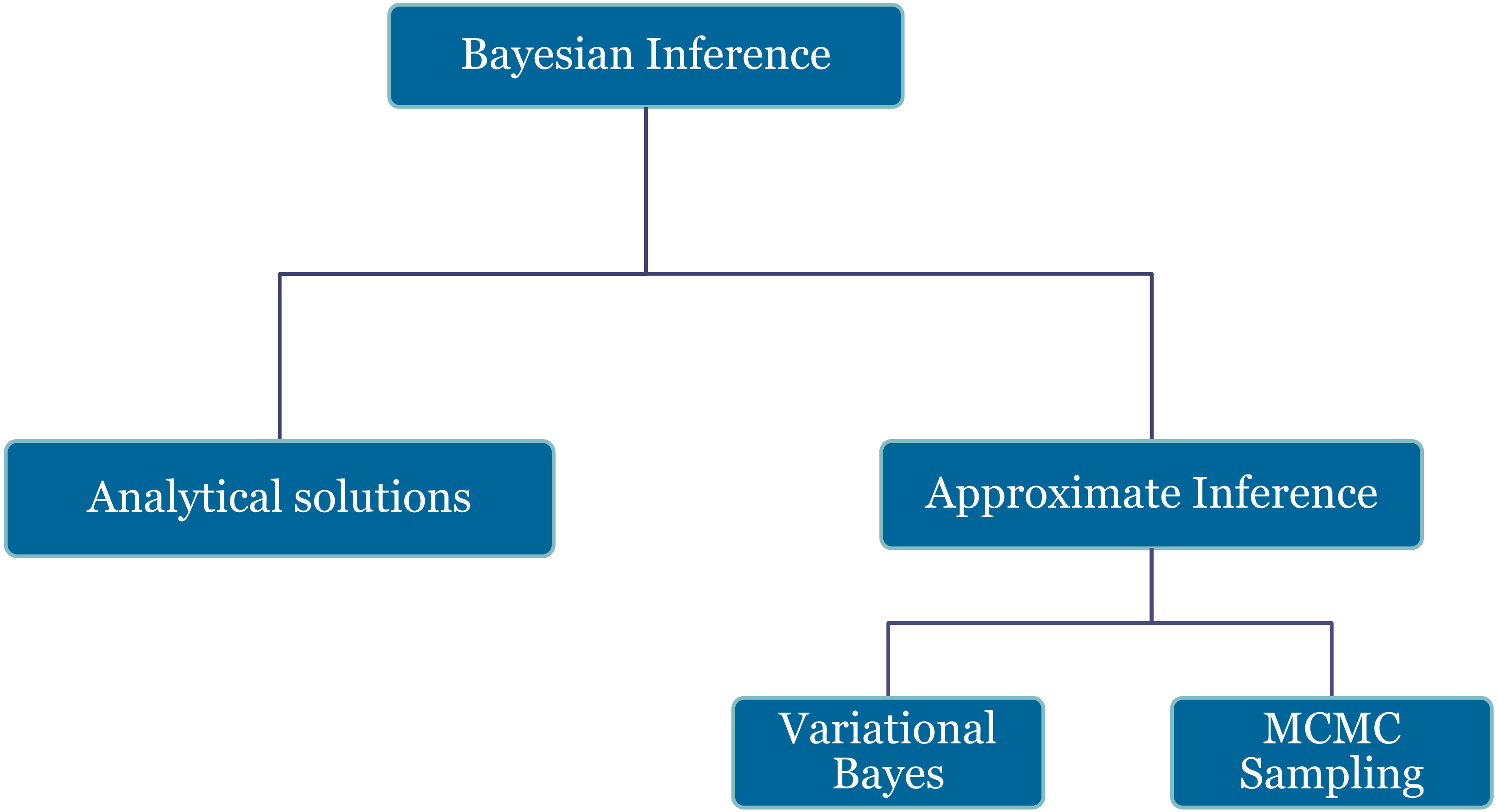
Example: free-energy principle and active inference



Maximizing the evidence (of the brain's generative model)
= minimizing the surprise about the data (sensory inputs).

Friston et al. 2006,
J Physiol Paris

How is the posterior computed = how is a generative model inverted?



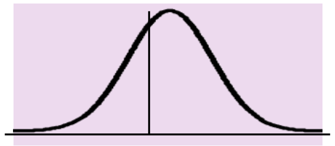
How is the posterior computed = how is a generative model inverted?

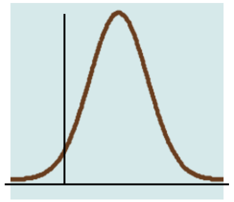
- **compute the posterior analytically**
  - requires conjugate priors
  - even then often difficult to derive an analytical solution
- **variational Bayes (VB)**
  - often hard work to derive, but fast to compute
  - cave: local minima, potentially inaccurate approximations
- **sampling methods (MCMC)**
  - guaranteed to be accurate in theory (for infinite computation time)
  - but may require very long run time in practice
  - convergence difficult to prove

Conjugate priors

If the posterior p(θ|x) is in the same family as the prior p(θ), the prior and posterior are called "conjugate distributions", and the prior is called a "conjugate prior" for the likelihood function.

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





same form

⇨ analytical expression for posterior

⇨ examples (likelihood-prior):

- Normal-Normal
- Normal-inverse Gamma
- Binomial-Beta
- Multinomial-Dirichlet

Posterior mean & variance of univariate Gaussians

$$y = \theta + \varepsilon$$

Likelihood & Prior

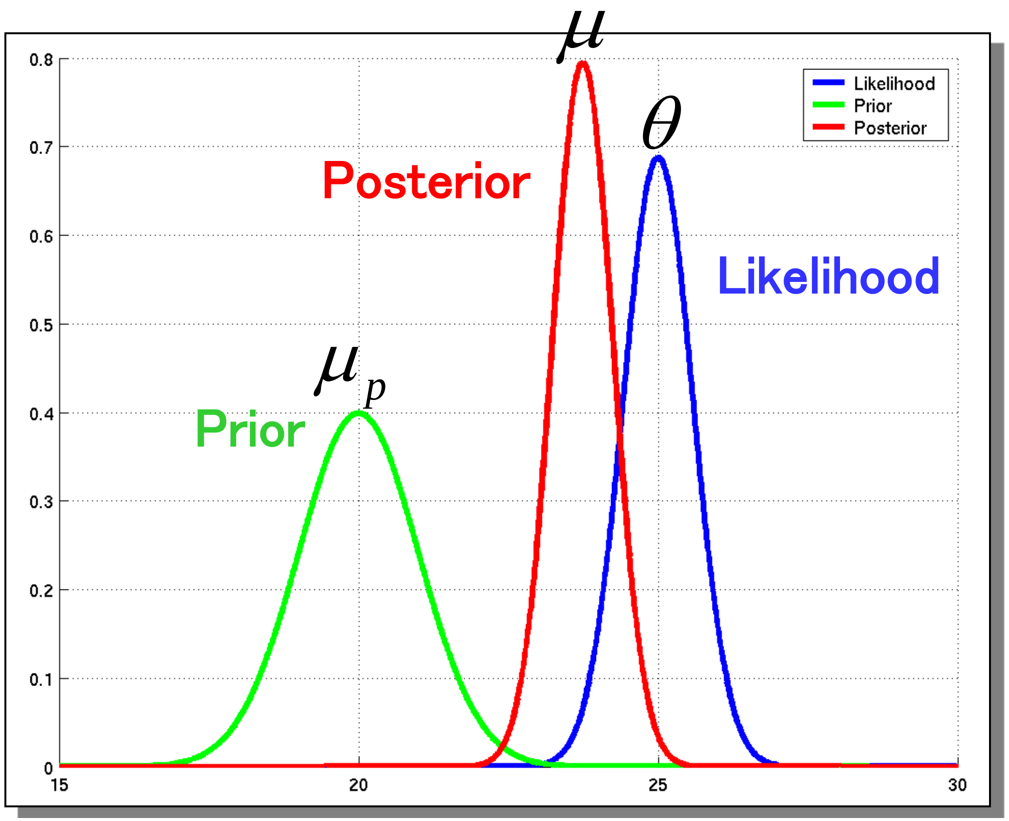
$$p(y \mid \theta) = N(\theta, \sigma_e^2)$$

$$p(\theta) = N(\mu_p, \sigma_p^2)$$

Posterior: $p(\theta \mid y) = N(\mu, \sigma^2)$

$$\frac{1}{\sigma^2} = \frac{1}{\sigma_e^2} + \frac{1}{\sigma_p^2}$$

$$\mu = \sigma^2 \left(\frac{1}{\sigma_e^2}\theta + \frac{1}{\sigma_p^2}\mu_p \right)$$



Posterior mean = variance-weighted combination of prior mean and data mean

Same thing – but expressed as precision weighting

Likelihood & prior

$$p(y\,|\,\theta) = N(\theta, \lambda_e^{-1})$$

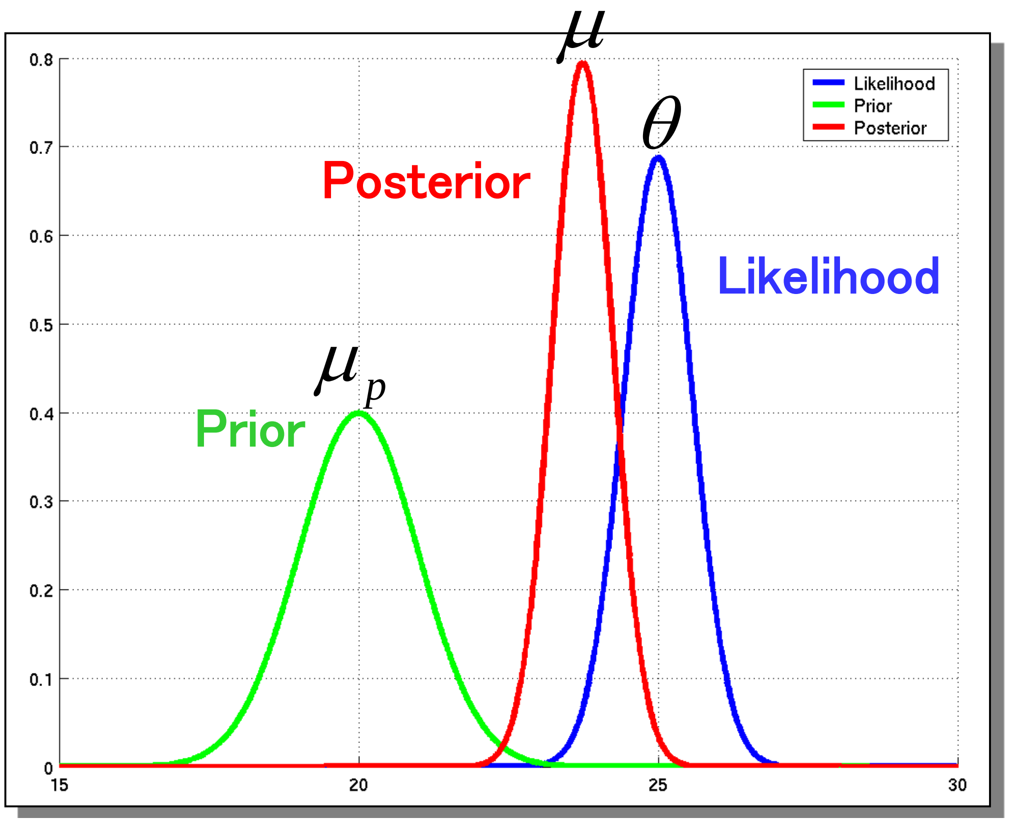
$$p(\theta) = N(\mu_p, \lambda_p^{-1})$$

$$y = \theta + \varepsilon$$

Posterior: $p(\theta\,|\,y) = N(\mu, \lambda^{-1})$

$$\lambda = \lambda_e + \lambda_p$$

$$\mu = \frac{\lambda_e}{\lambda}\theta + \frac{\lambda_p}{\lambda}\mu_p$$

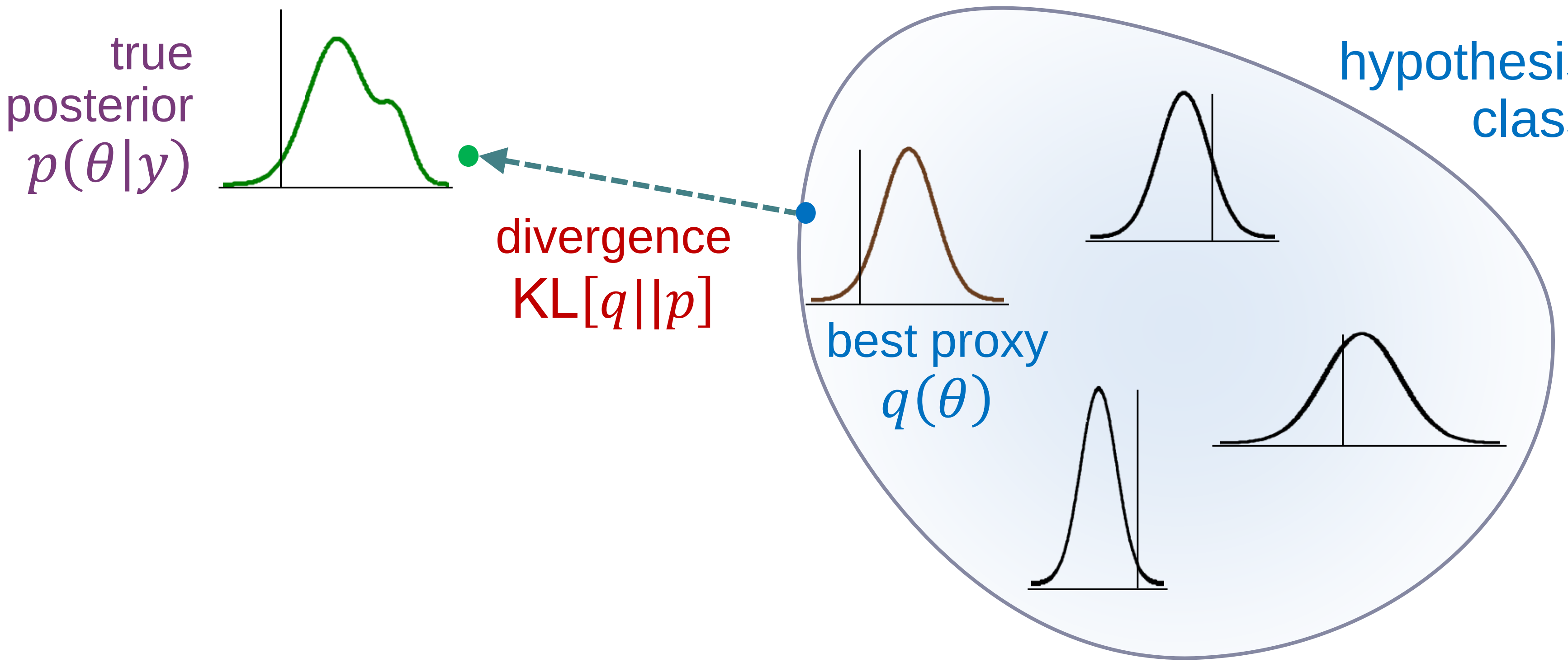


Relative precision weighting

Variational Bayes (VB)

Idea: find an approximate density $q(\theta)$ that is maximally similar to the true posterior $p(\theta|y)$.

This is often done by assuming a particular form for q (fixed form VB) and then optimizing its sufficient statistics.



Kullback–Leibler（KL）divergence

- asymmetric measure of the difference between two probability distributions P and Q

- Interpretations of $D_{KL}(P\|Q)$:
  - "Bayesian surprise" when Q=prior, P=posterior: measure of the information gained when one updates one's prior beliefs to the posterior P
  - a measure of the information lost when Q is used to approximate P

- non-negative: ≥0 (zero when P=Q)

$$D_{\mathrm{KL}}(P\|Q) = \sum_i P(i) \ln \frac{P(i)}{Q(i)}.$$

$$D_{\mathrm{KL}}(P\|Q) = \int_{-\infty}^{\infty} p(x) \ln \frac{p(x)}{q(x)} \, \mathrm{d}x,$$

Variational calculus

Standard calculus
Newton, Leibniz, and others

- functions $f: x \mapsto f(x)$
- derivatives $\frac{\mathrm{d}f}{\mathrm{d}x}$

Example: maximize the likelihood expression $p(y|\theta)$ w.r.t. θ

Variational calculus
Euler, Lagrange, and others

- functionals $F: f \mapsto F(f)$
- derivatives $\frac{\mathrm{d}F}{\mathrm{d}f}$

Example: maximize the entropy $H[p]$ w.r.t. a probability distribution $p(x)$

Leonhard Euler
(1707 – 1783)
Swiss mathematician, ‘Elementa Calculi Variationum’

Variational Bayes

$$\ln p(y) = \underbrace{\mathrm{KL}[q||p]}_{\text{divergence}} + \underbrace{F(q,y)}_{\text{neg. free energy}}$$

divergence ≥ 0 (unknown)

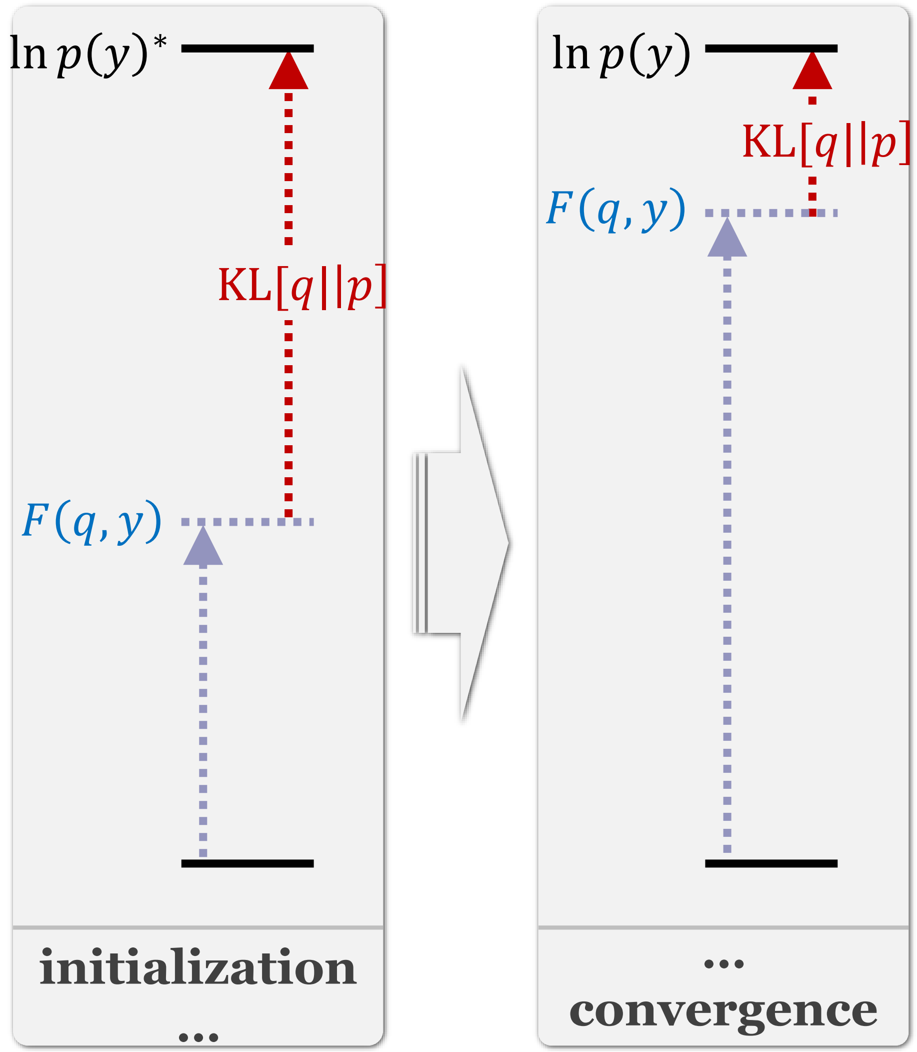
neg. free energy (easy to evaluate for a given q)

$F(q)$ is a functional wrt. the approximate posterior $q(\theta)$.

Maximizing $F(q,y)$ is equivalent to:

- minimizing $\mathrm{KL}[q||p]$
- tightening $F(q,y)$ as a lower bound to the log model evidence

When $F(q,y)$ is maximized, $q(\theta)$ is our best estimate of the posterior.



Derivation of the (negative) free energy approximation

- See whiteboard!
- (or Appendix to Stephan et al. 2007, NeuroImage 38: 387-401)

Mean field assumption

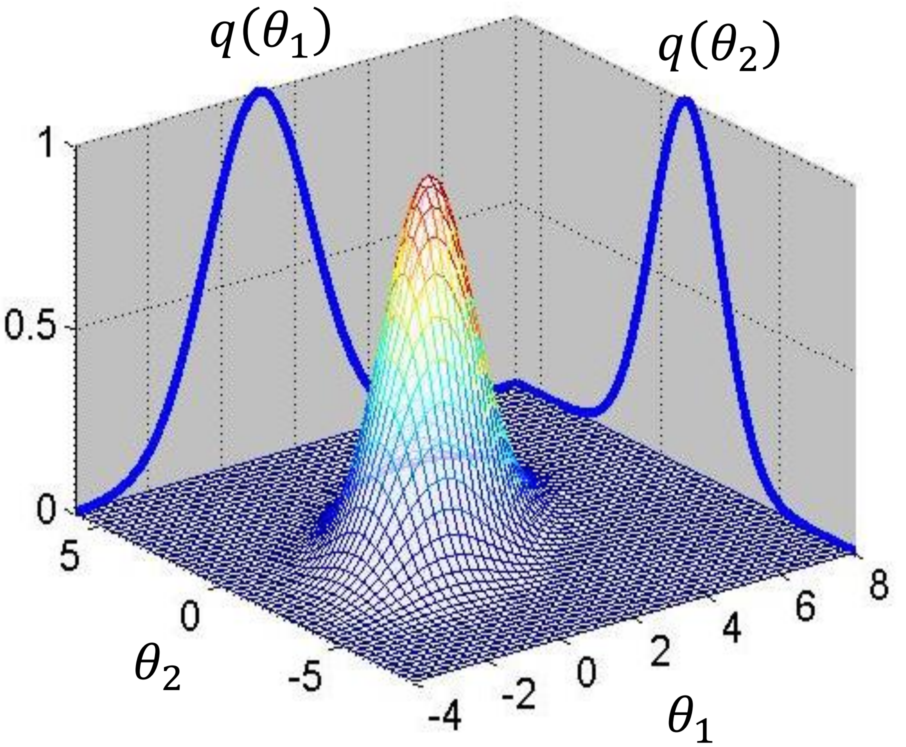
Factorize the approximate posterior $q(\theta)$ into independent partitions:

$$q(\theta) = \prod_i q_i(\theta_i)$$

where $q_i(\theta_i)$ is the approximate posterior for the i^{th} subset of parameters.

For example, split parameters and hyperparameters:

$$p(\theta, \lambda \mid y) \approx q(\theta, \lambda) = q(\theta) q(\lambda)$$



Jean Daunizeau, www.fil.ion.ucl.ac.uk/~jdaunize/presentations/Bayes2.pdf

VB in a nutshell (under mean-field approximation)

❶ Neg. free-energy approx. to model evidence.

$$\ln p(y|m) = F + KL\left[q(\theta,\lambda), p(\theta,\lambda|y)\right]$$

$$F = \left\langle \ln p(y|\theta,\lambda)\right\rangle_q - KL\left[q(\theta,\lambda), p(\theta,\lambda|m)\right]$$

❷ Mean field approx.

$$p(\theta,\lambda|y) \approx q(\theta,\lambda) = q(\theta)q(\lambda)$$

❸ Maximise neg. free energy wrt. q = minimise divergence, by maximising variational energies

$$q(\theta) \propto \exp(I_\theta) = \exp\left[\left\langle \ln p(y,\theta,\lambda)\right\rangle_{q(\lambda)}\right]$$

$$q(\lambda) \propto \exp(I_\lambda) = \exp\left[\left\langle \ln p(y,\theta,\lambda)\right\rangle_{q(\theta)}\right]$$

❹ Iterative updating of sufficient statistics of approx. posteriors by gradient ascent.

VB (under mean-field assumption) in more detail

$$
\begin{aligned}
F(q,y) &= \int q(\theta)\ln\frac{p(y,\theta)}{q(\theta)}\,\mathrm{d}\theta \\
&= \int \prod_i q_i \times \left(\ln p(y,\theta) - \sum_i \ln q_i\right)\mathrm{d}\theta \\
&= \int q_j \prod_{\setminus j} q_i \left(\ln p(y,\theta) - \ln q_j\right)\mathrm{d}\theta - \int q_j \prod_{\setminus j} q_i \sum_{\setminus j} \ln q_i \,\mathrm{d}\theta \\
&= \int q_j \Bigg(\underbrace{\int \prod_{\setminus j} q_i \ln p(y,\theta)\,\mathrm{d}\theta_{\setminus j}}_{\langle \ln p(y,\theta)\rangle_{q_{\setminus j}}} - \ln q_j \Bigg)\mathrm{d}\theta_j - \int q_j \int \prod_{\setminus j} q_i \ln \prod_{\setminus j} q_i \,\mathrm{d}\theta_{\setminus j}\,\mathrm{d}\theta_j \\
&= \int q_j \ln \frac{\exp\left(\langle \ln p(y,\theta)\rangle_{q_{\setminus j}}\right)}{q_j}\,\mathrm{d}\theta_j \quad + \quad c \\
&= -\mathrm{KL}\left[q_j \,||\, \exp\left(\langle \ln p(y,\theta)\rangle_{q_{\setminus j}}\right)\right] \quad + \quad c
\end{aligned}
$$

mean-field assumption:
$q(\theta) = \prod_i q_i(\theta_i)$

VB (under mean-field assumption) in more detail

In summary:

$$F(q,y) = -\mathrm{KL}\left[q_j || \exp\left(\langle \ln p(y,\theta) \rangle_{q_{\setminus j}}\right)\right] + c$$

Suppose the densities $q_{\setminus j} \equiv q\left(\theta_{\setminus j}\right)$ are kept fixed. Then the approximate posterior $q\left(\theta_j\right)$ that maximizes $F(q,y)$ is given by:

$$\begin{aligned} q_j^* &= \arg\max_{q_j} F(q,y) \\ &= \frac{1}{Z} \exp\left(\langle \ln p(y,\theta) \rangle_{q_{\setminus j}}\right) \end{aligned}$$

Therefore:

$$\ln q_j^* = \underbrace{\langle \ln p(y,\theta) \rangle_{q_{\setminus j}}}_{=:I(\theta_j)} - \ln Z$$

This implies a straightforward algorithm for variational inference:

❶ Initialize all approximate posteriors $q(\theta_i)$, e.g., by setting them to their priors.

❷ Cycle over the parameters, revising each given the current estimates of the others.

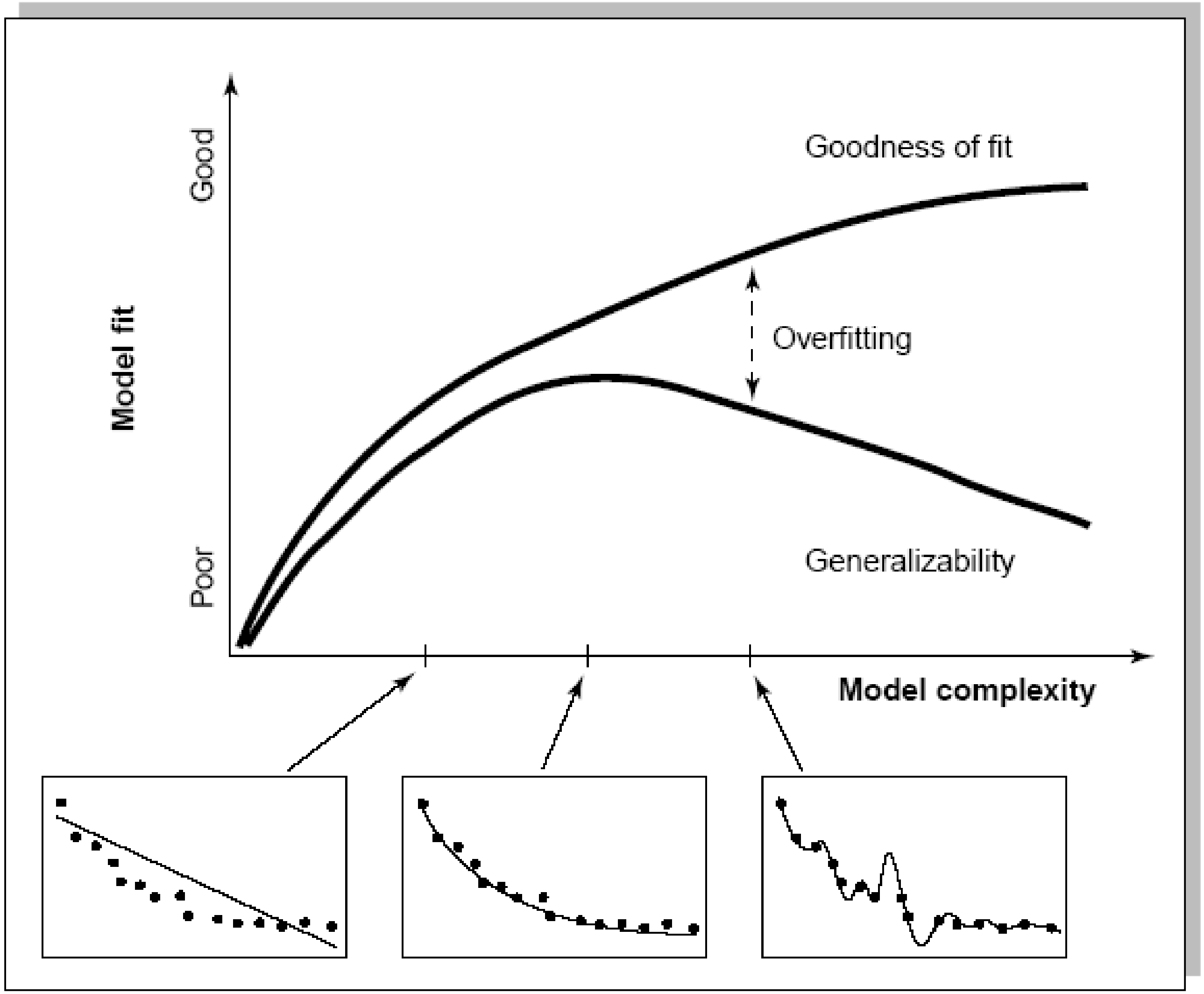
❸ Loop until convergence.

Model comparison and selection

Given competing hypotheses on structure & functional mechanisms of a system, which model is the best?

Which model represents the best balance between model fit and model complexity?

For which model m does $p(y|m)$ become maximal?

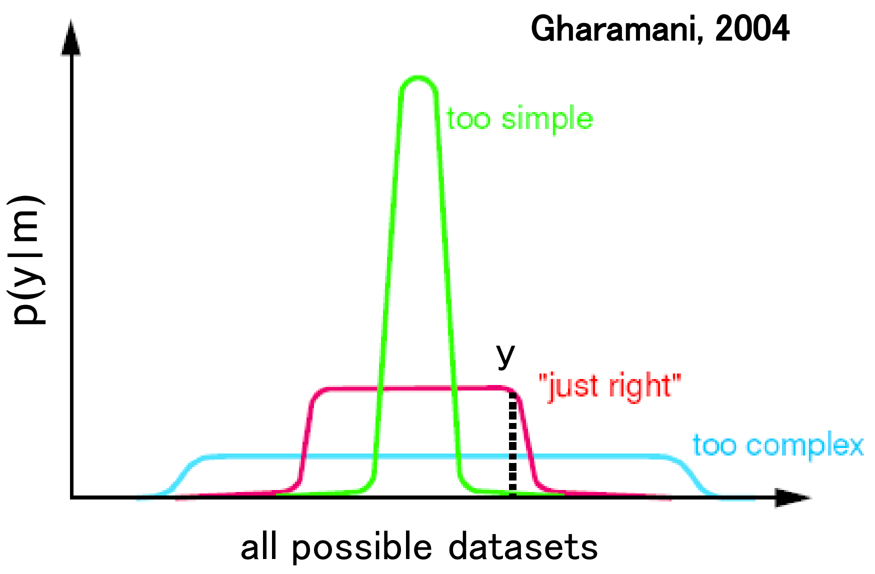


Pitt & Miyung (2002) *TICS*

Bayesian model selection (BMS)

Model evidence (marginal likelihood):

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



→ accounts for both accuracy and complexity of the model

→ "If I randomly sampled from my prior and plugged the resulting value into the likelihood function, how close would the predicted data be – on average – to my observed data?"

Various approximations, e.g.:

– negative free energy, AIC, BIC

McKay 1992, *Neural Comput.*
Penny et al. 2004a, *NeuroImage*

Model space (hypothesis set) M

Model space M is defined by prior on models.

Usual choice: flat prior over a small set of models.

$$p(m) = \begin{cases} 1/|M| \text{ if } m \in M \\ 0 \text{ if } m \notin M \end{cases}$$

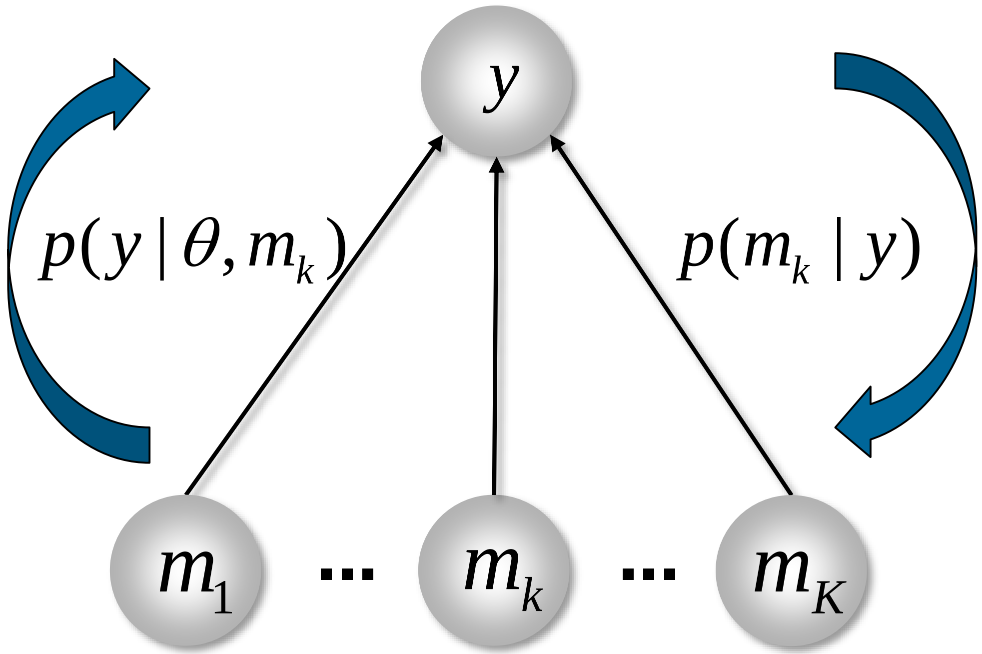
In this case, the posterior probability of model i is:

$$p(m_i \mid y) = \frac{p(y \mid m_i)\, p(m_i)}{\sum_{j=1}^{|M|} p(y \mid m_j)\, p(m_j)} = \frac{p(y \mid m_i)}{\sum_{j=1}^{|M|} p(y \mid m_j)}$$

Differential diagnosis based on generative models of disease symptoms

SYMPTOM
(behaviour
or physiology)

HYPOTHETICAL
MECHANISM



$$p(m_k \mid \mathrm{y}) = \frac{p(y \mid m_k)\, p(m_k)}{\sum_k p(y \mid m_k)\, p(m_k)}$$

Stephan et al. 2016, *NeuroImage*

Approximations to the model evidence

Logarithm is a monotonic function → Maximizing log model evidence = Maximizing model evidence

Log model evidence = balance between fit and complexity

$$\log p(y \mid m) = accuracy(m) - complexity(m)$$
$$= \log p(y \mid \theta, m) - complexity(m)$$

Akaike Information Criterion:

$$AIC = \log p(y \mid \theta, m) - p$$

(p: No. of parameters)

Bayesian Information Criterion:

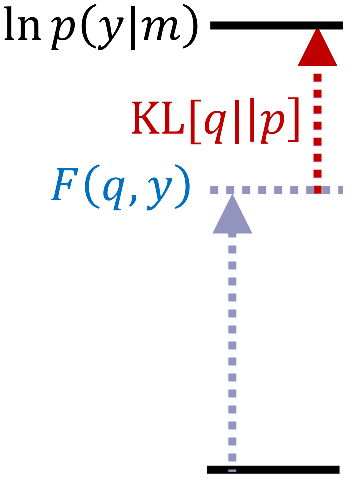
$$BIC = \log p(y \mid \theta, m) - \frac{p}{2} \log N$$

(N: No. of data points)

The (negative) free energy approximation F

F is a lower bound on the log model evidence:

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



Like AIC/BIC, F is an accuracy/complexity tradeoff:

$$F = \underbrace{\langle \log p(y|\theta,m)\rangle}_{accuracy} - \underbrace{KL\left[q(\theta), p(\theta|m)\right]}_{complexity}$$

The (negative) free energy approximation

- Log evidence is thus expected log likelihood (wrt. q) plus 2 KL's:

$$\log p(y|m)$$
$$= \langle \log p(y|\theta, m) \rangle - KL\left[q(\theta), p(\theta|m)\right] + KL\left[q(\theta), p(\theta|y,m)\right]$$

$$F = \log p(y|m) - KL\left[q(\theta), p(\theta|y,m)\right]$$
$$= \underbrace{\langle \log p(y|\theta, m) \rangle}_{accuracy} - \underbrace{KL\left[q(\theta), p(\theta|m)\right]}_{complexity}$$

The complexity term in *F*

- In contrast to AIC & BIC, the complexity term of the negative free energy *F* accounts for parameter interdependencies. Under Gaussian assumptions about the posterior (Laplace approximation):

$$KL\left[q(\theta), p(\theta \mid m)\right]$$
$$= \frac{1}{2}\ln\left|C_\theta\right| - \frac{1}{2}\ln\left|C_{\theta|y}\right| + \frac{1}{2}\left(\mu_{\theta|y} - \mu_\theta\right)^T C_\theta^{-1}\left(\mu_{\theta|y} - \mu_\theta\right)$$

- The complexity term of *F* is higher
  - the more independent the prior parameters (↑ effective DFs)
  - the more dependent the posterior parameters
  - the more the posterior mean deviates from the prior mean

Bayes factors

To compare two models, we could just compare their log evidences.

But: the log evidence is just some number – not very intuitive!

A more intuitive interpretation of model comparisons is made possible by Bayes factors:

positive value, [0; ∞[

$$B_{12} = \frac{p(y \mid m_1)}{p(y \mid m_2)}$$

Kass & Raftery classification:

B_{12}	$p(m_1 \mid y)$	Evidence
1 to 3	50-75%	weak
3 to 20	75-95%	positive
20 to 150	95-99%	strong
≥ 150	≥ 99%	Very strong

Kass & Raftery 1995, *J. Am. Stat. Assoc.*

Fixed effects BMS at group level

Group Bayes factor (GBF) for 1...*K* subjects:

$$GBF_{ij} = \prod_k BF_{ij}^{(k)}$$

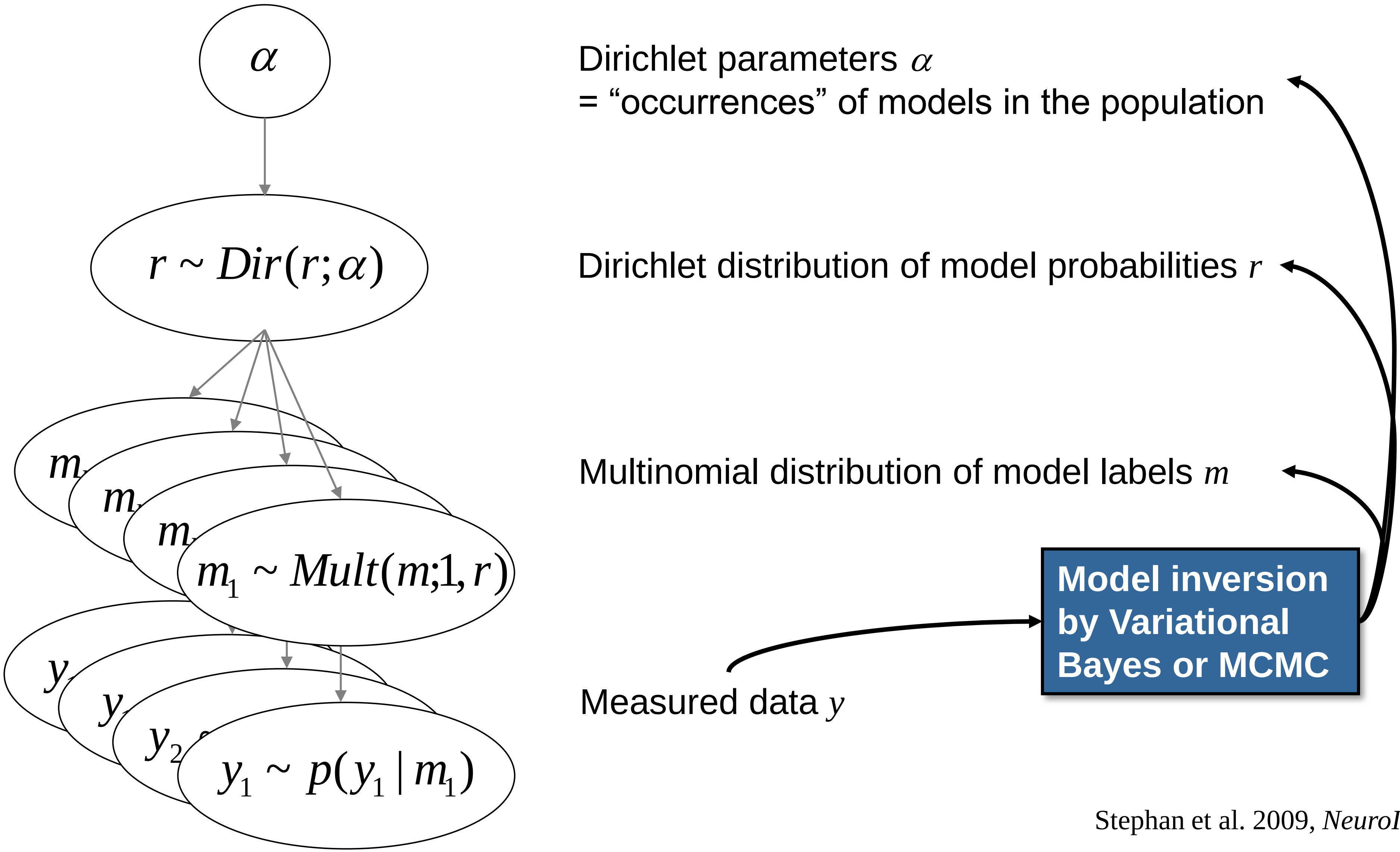
Average Bayes factor (ABF):

$$ABF_{ij} = \sqrt[K]{\prod_k BF_{ij}^{(k)}}$$

Problems:

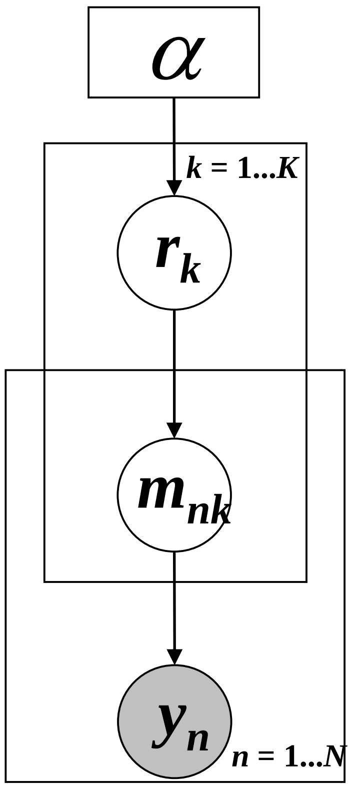
- blind with regard to group heterogeneity
- sensitive to outliers

Random effects BMS for heterogeneous groups



Stephan et al. 2009, *NeuroImage*

Random effects BMS for heterogeneous groups



Dirichlet parameters α
= “occurrences” of models in the population

Dirichlet distribution of model probabilities r

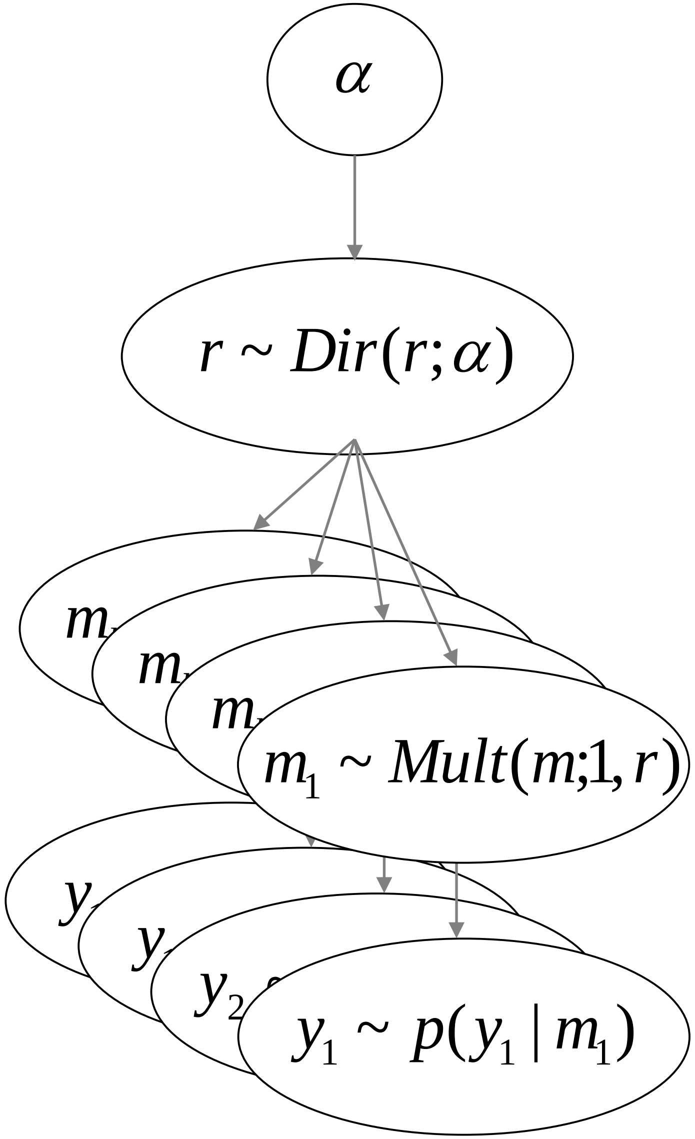
Multinomial distribution of model labels m

Measured data y

Model inversion by Variational Bayes (VB) or MCMC

Stephan et al. 2009, *NeuroImage*

Random effects BMS



$$p(r|\alpha) = Dir(r,\alpha) = \frac{1}{Z(\alpha)} \prod_k r_k^{\alpha_k - 1}$$

$$Z(\alpha) = \prod_k \Gamma(\alpha_k) \Big/ \Gamma\left(\sum_k \alpha_k\right)$$

$$p(m_n | r) = \prod_k r_k^{m_{nk}}$$

$$p(y_n | m_{nk}) = \int p(y|\vartheta) p(\vartheta | m_{nk}) d\vartheta$$

Stephan et al. 2009, *NeuroImage*

❶ Write down joint probability and take the log

$$p\left(y,r,m\right)=p\left(y\,|\,m\right)p\left(m\,|\,r\right)p(r\,|\,\alpha_0)$$

$$=p(r\,|\,\alpha_0)\left[\prod_n p\left(y_n\,|\,m_n\right)p\left(m_n\,|\,r\right)\right]$$

$$=\frac{1}{Z\left(\alpha_0\right)}\left[\prod_k r_k^{\alpha_{0k}-1}\right]\left[\prod_n p\left(y_n\,|\,m_n\right)\prod_k r_k^{m_{nk}}\right]$$

$$=\frac{1}{Z\left(\alpha_0\right)}\prod_n\left[\prod_k\left[p\left(y_n|m_{nk}\right)r_k\right]^{m_{nk}}\,r_k^{\alpha_{0k}-1}\right]$$

$$\ln p\left(y,r,m\right)=-\ln Z\left(\alpha_0\right)+\sum_n\sum_k\left(\left(\alpha_{0k}-1\right)\ln r_k+m_{nk}\left(\log p\left(y_n\,|\,m_{nk}\right)+\ln r_k\right)\right)$$

❷ Mean field approx.

$$q(r,m) = q(r)q(m)$$

❸ Maximise neg. free energy wrt. q = minimise divergence, by maximising variational energies

$$q(r) \propto \exp\left(I(r)\right)$$

$$q(m) \propto \exp\left(I(m)\right)$$

$$I(r) = \left\langle \log p(y,r,m) \right\rangle_{q(m)}$$

$$I(m) = \left\langle \log p(y,r,m) \right\rangle_{q(r)}$$

❹ Iterative updating of sufficient statistics of approx. posteriors

$\alpha = \alpha_0$

$\alpha_0 = [1,\ldots,1]$

Until convergence

$$u_{nk} = \exp\left(\ln p(y_n \mid m_{nk}) + \Psi(\alpha_k) - \Psi\left(\sum_k \alpha_k\right)\right)$$

$$g_{nk} = \frac{u_{nk}}{\sum_k u_{nk}}$$

$$\beta_k = \sum_n g_{nk}$$

$$\alpha = \alpha_0 + \beta$$

end

$$g_{nk} = q(m_{nk} = 1)$$

our (normalized) posterior belief that model k generated the data from subject n

$$\beta_k = \sum_n g_{nk}$$

expected number of subjects whose data we believe were generated by model k

Four equivalent options for reporting model ranking by random effects BMS

1. **Dirichlet parameter estimates**

$$\alpha$$

2. **expected posterior probability** of obtaining the k-th model for any randomly selected subject

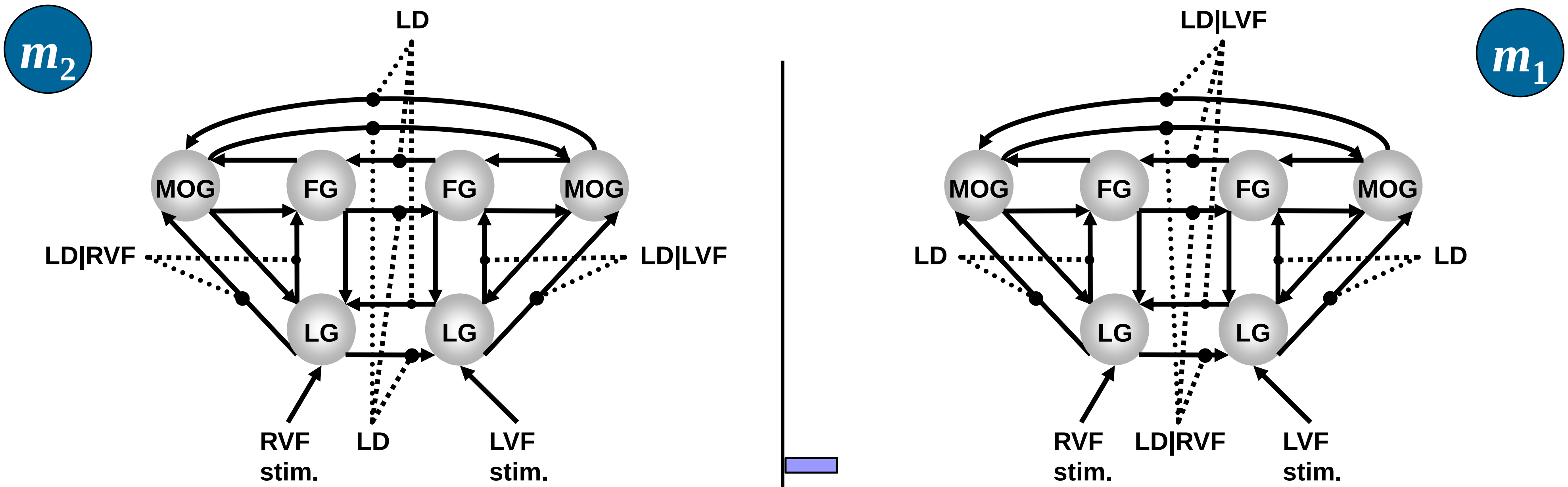
$$\langle r_k \rangle_q = \alpha_k / (\alpha_1 + \ldots + \alpha_K)$$

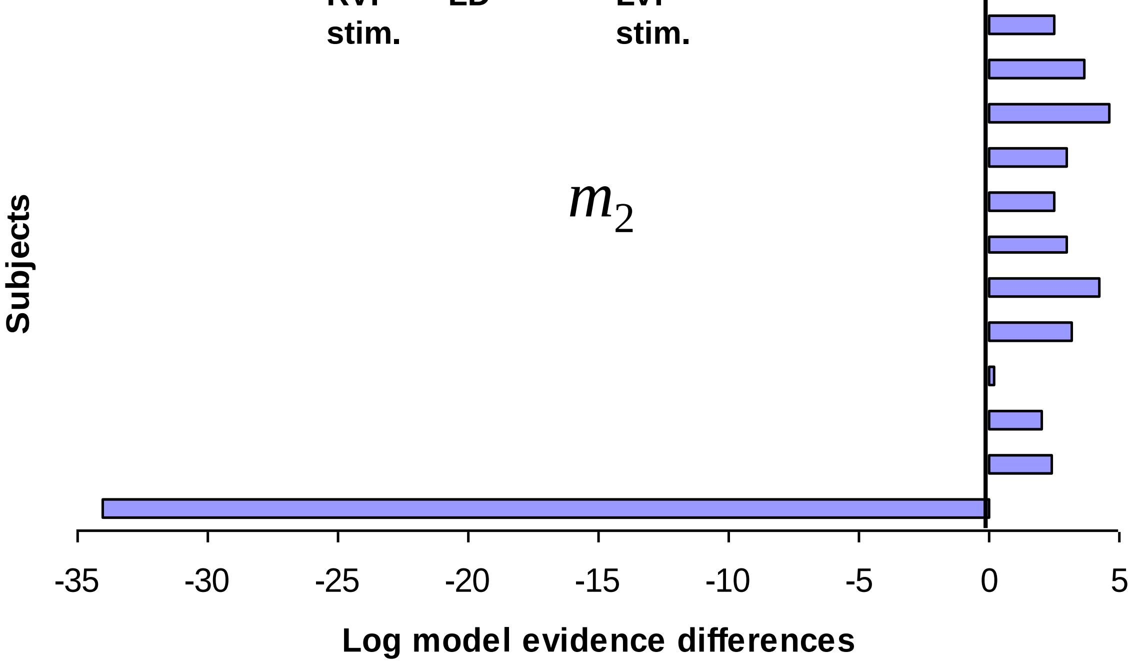
3. **exceedance probability** that a particular model *k* is more likely than any other model (of the *K* models tested), given the group data

$$\exists k \in \{1 \ldots K\}, \forall j \in \{1 \ldots K \mid j \neq k\}:$$
$$\varphi_k = p(r_k > r_j \mid y; \alpha)$$

4. **protected exceedance probability**: see below

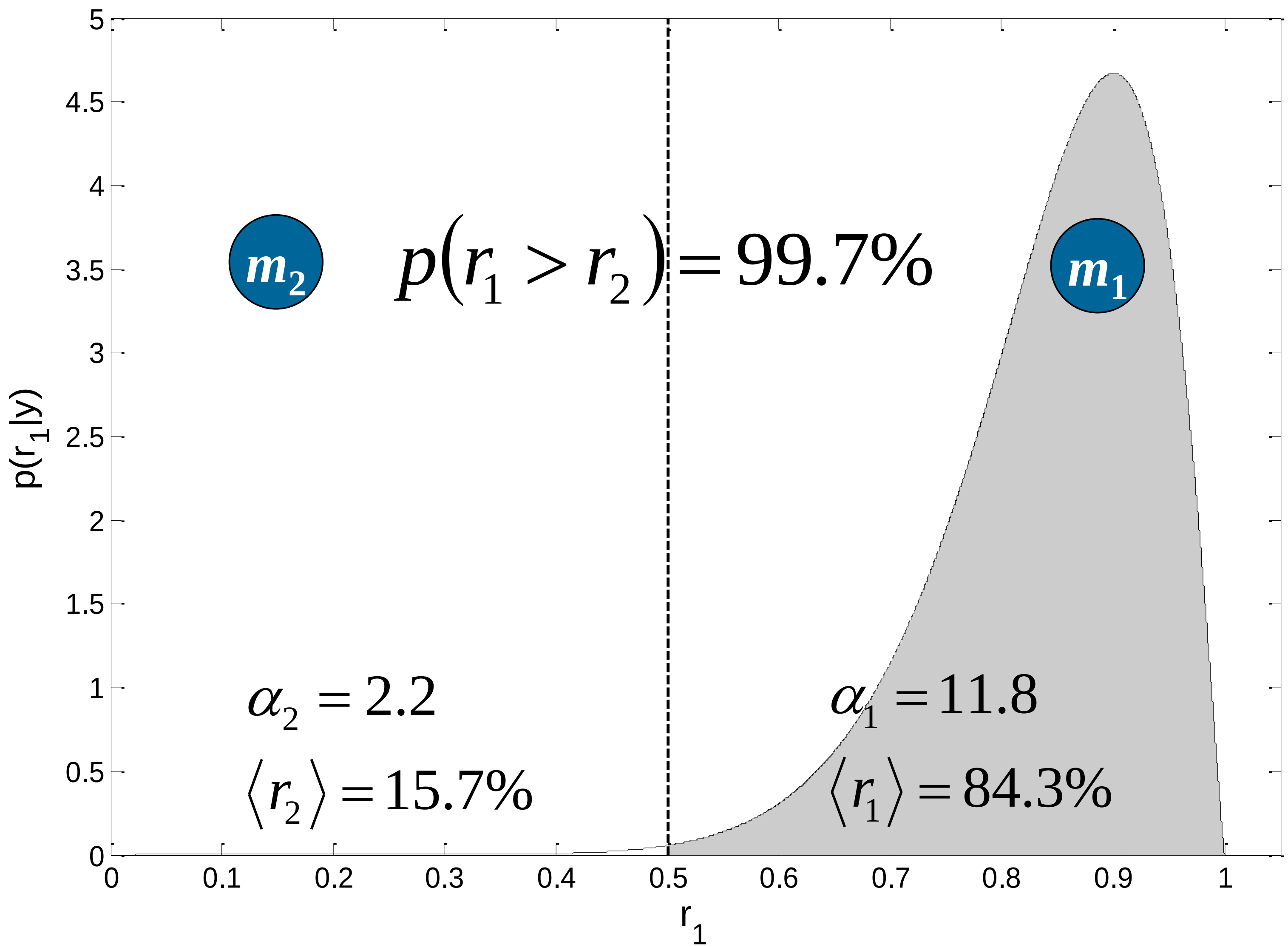
Example: Hemispheric interactions during vision





m_1

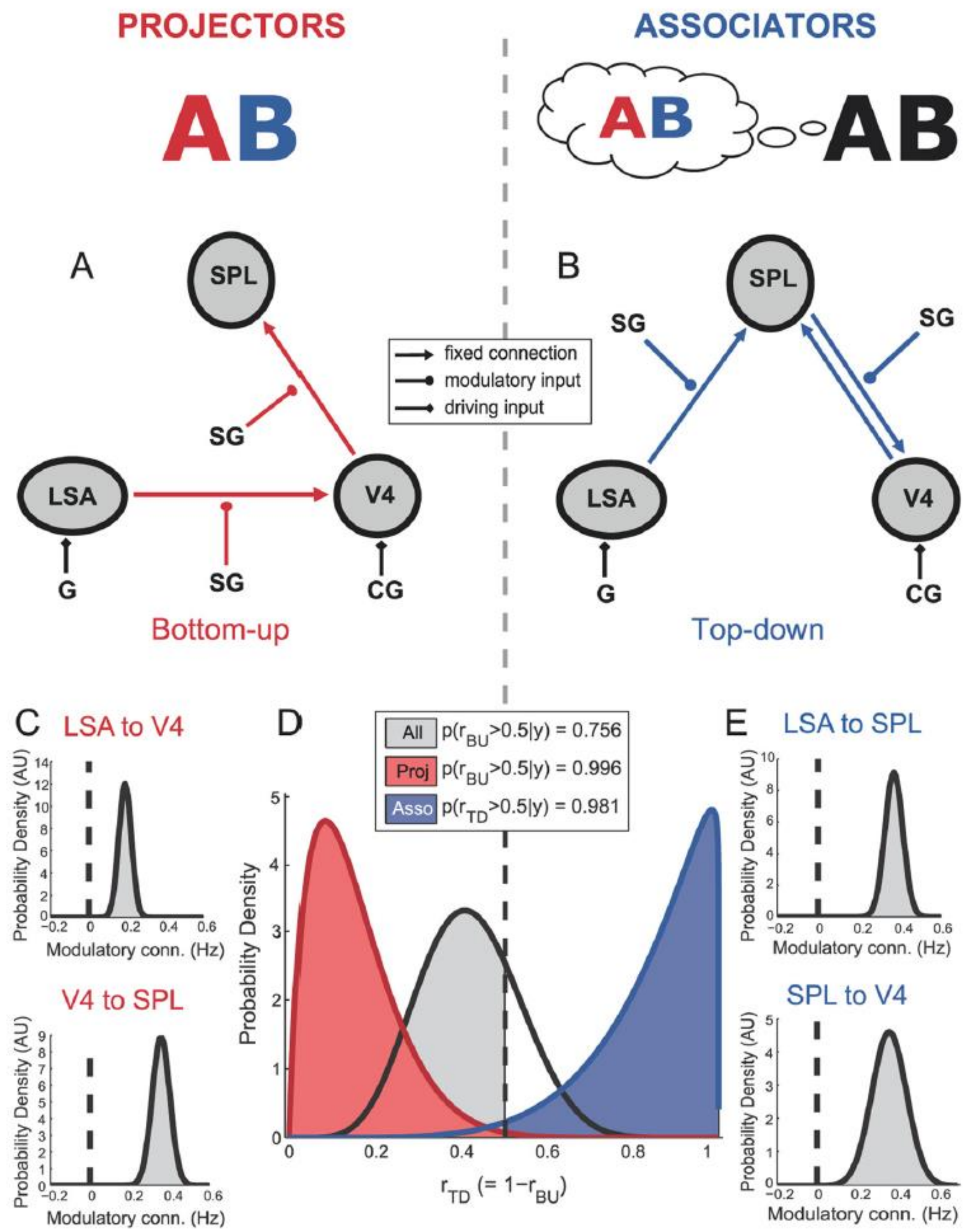
Data: Stephan et al. 2003, *Science*
Models: Stephan et al. 2007, *J. Neurosci.*



Stephan et al. 2009a, *NeuroImage*

Example: Synaesthesia

- "projectors" experience color externally colocalized with a presented grapheme
- "associators" report an internally evoked association
- across all subjects: no evidence for either model
- but BMS results map precisely onto projectors (bottom-up mechanisms) and associators (top-down)



van Leeuwen et al. 2011, *J. Neurosci.*

Protected exceedance probability: Using BMA to protect against chance findings

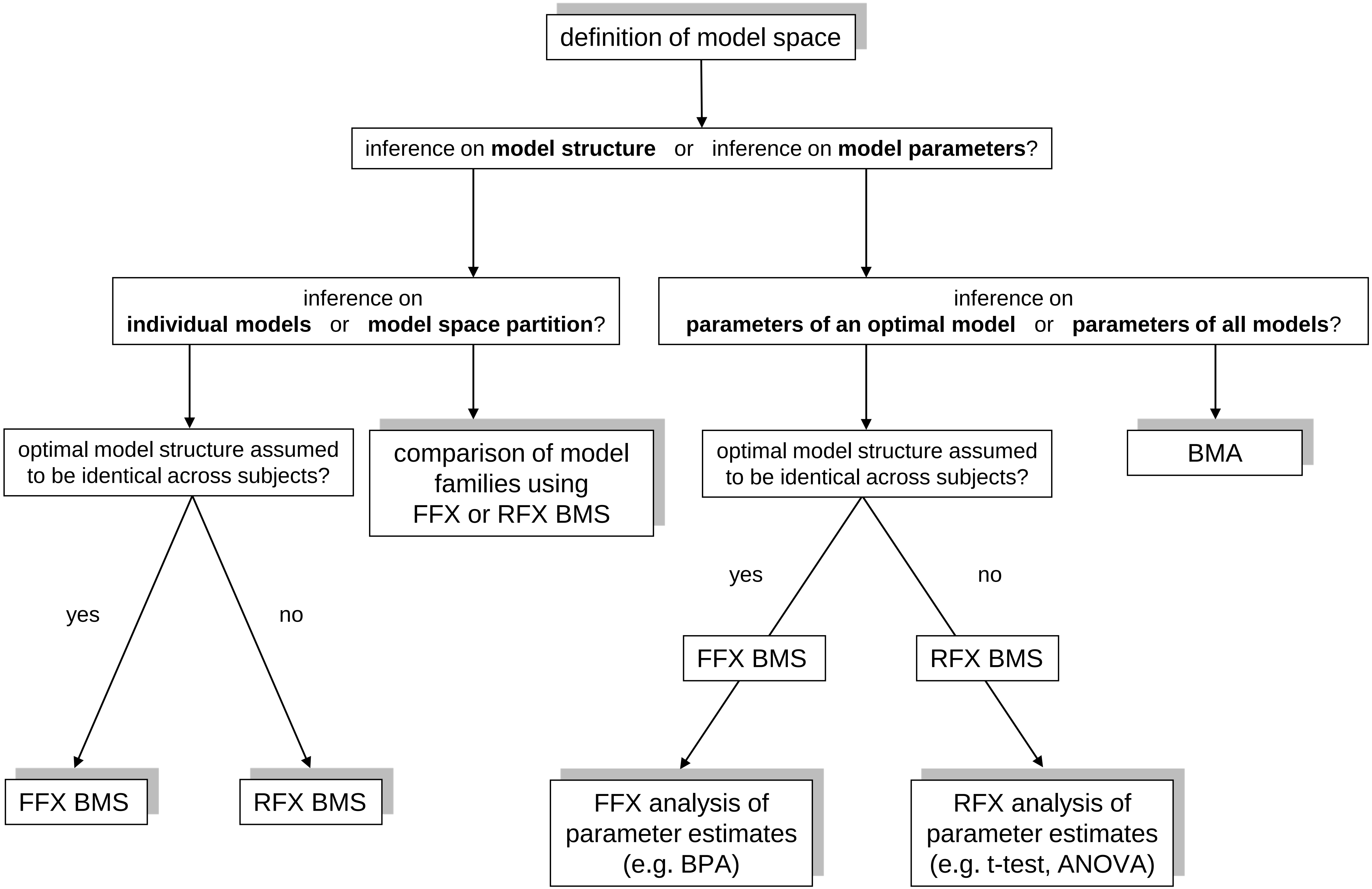
- EPs express our confidence that the posterior probabilities of models are different – under the hypothesis H_1 that models differ in probability: $r_k \neq 1/K$
- does not account for possibility "null hypothesis" H_0: $r_k = 1/K$
- **Bayesian omnibus risk (BOR)** of wrongly accepting H_1 over H_0:

$$P_o = \frac{1}{1 + \frac{p(m|H_1)}{p(m|H_0)}}.$$

- **protected EP**: Bayesian model averaging over H_0 and H_1:

$$\begin{aligned}\widetilde{\varphi}_k &= P(r_k \geq r_{k' \neq k} | y) \\ &= P(r_k \geq r_{k' \neq k} | y, H_1) P(H_1 | y) + P(r_k \geq r_{k' \neq k} | y, H_0) P(H_0 | y) \\ &= \varphi_k (1 - P_0) + \frac{1}{K} P_0\end{aligned}$$

Rigoux et al. 2014, *NeuroImage*



Stephan et al. 2010, *NeuroImage*

Further reading

- Penny WD, Stephan KE, Mechelli A, Friston KJ (2004) Comparing dynamic causal models. NeuroImage 22:1157-1172.
- Penny WD, Stephan KE, Daunizeau J, Joao M, Friston K, Schofield T, Leff AP (2010) Comparing Families of Dynamic Causal Models. PLoS Computational Biology 6: e1000709.
- Penny WD (2012) Comparing dynamic causal models using AIC, BIC and free energy. Neuroimage 59: 319-330.
- Rigoux L, Stephan KE, Friston KJ, Daunizeau J (2014) Bayesian model selection for group studies – revisited. NeuroImage 84: 971-985.
- Stephan KE, Weiskopf N, Drysdale PM, Robinson PA, Friston KJ (2007) Comparing hemodynamic models with DCM. NeuroImage 38:387-401.
- Stephan KE, Penny WD, Daunizeau J, Moran RJ, Friston KJ (2009) Bayesian model selection for group studies. NeuroImage 46:1004-1017.
- Stephan KE, Penny WD, Moran RJ, den Ouden HEM, Daunizeau J, Friston KJ (2010) Ten simple rules for Dynamic Causal Modelling. NeuroImage 49: 3099-3109.
- Stephan KE, Iglesias S, Heinzle J, Diaconescu AO (2015) Translational Perspectives for Computational Neuroimaging. Neuron 87: 716-732.
- Stephan KE, Schlagenhauf F, Huys QJM, Raman S, Aponte EA, Brodersen KH, Rigoux L, Moran RJ, Daunizeau J, Dolan RJ, Friston KJ, Heinz A (2017) Computational Neuroimaging Strategies for Single Patient Predictions. NeuroImage 145: 180-199.

Thank you