(variational) Bayesian inference

J. Daunizeau

ICM, Paris, France

TNU, Zurich, Switzerland

Overview of the talk

- ✓ Introduction to Bayesian inference
- ✓ The variational approach to approximate Bayesian inference
- ✓ VBA toolbox

Overview of the talk

- ✓ Introduction to Bayesian inference
- ✓ The variational approach to approximate Bayesian inference
- ✓ VBA toolbox

Probability theory: basics

Degree of plausibility desiderata:

- ❑ should be represented using real numbers (D1)
- ❑ should conform with intuition (D2)
- ❑ should be consistent (D3)

a=2

→ normalization: $$\sum_a P(a) = 1$$

b=5

a=2

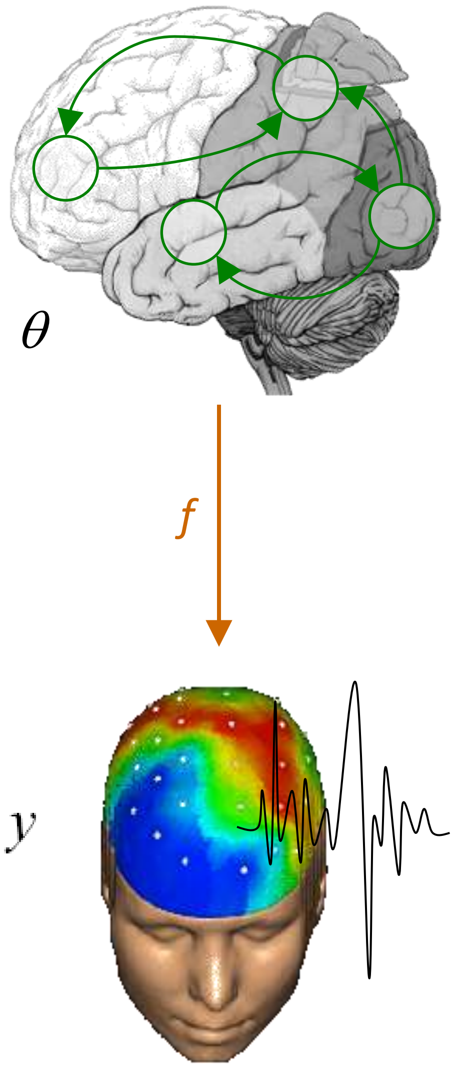
→ marginalization: $$P(b) = \sum_a P(a,b)$$

→ conditioning :

(*Bayes rule*)

$$P(a,b) = P(a|b)\,P(b) = P(b|a)\,P(a)$$

Deriving the likelihood function

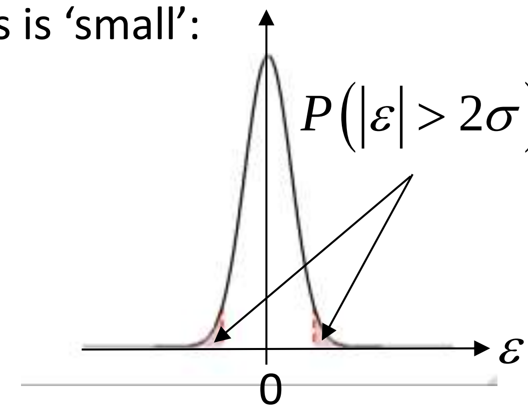


- Model of data with unknown parameters:

$$y = f(\theta) \qquad \text{e.g., GLM:} \qquad f(\theta) = X\theta$$

- But data is noisy: $y = f(\theta) + \varepsilon$

- Assume noise/residuals is 'small':



$$P(|\varepsilon| > 2\sigma) \approx 0.05$$

$$p(\varepsilon) \propto \exp\left(-\frac{1}{2\sigma^2}\varepsilon^2\right)$$

→ Distribution of data, *given fixed parameters*:

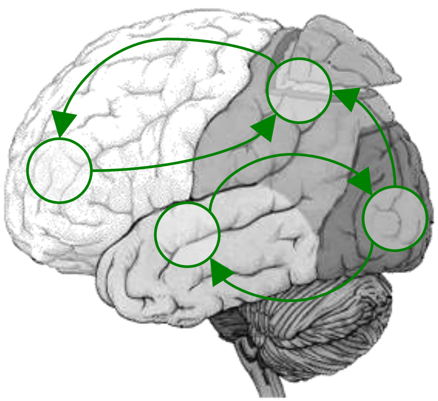
$$p(y|\theta) \propto \exp\left(-\frac{1}{2\sigma^2}\left(y - f(\theta)\right)^2\right)$$

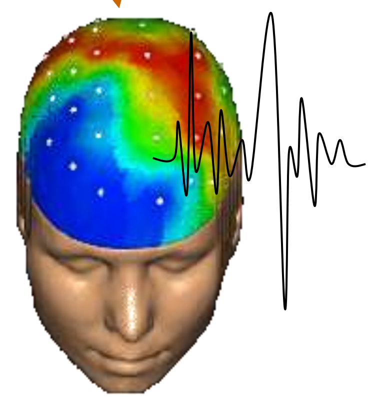
Probabilistic model inversion

forward problem

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

likelihood



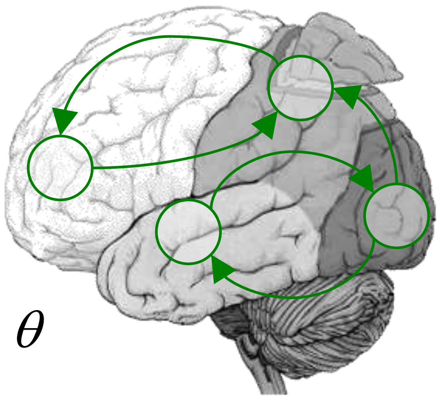


posterior distribution

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

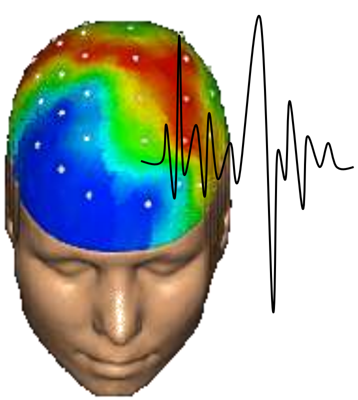
inverse problem

Posterior inference on model parameters



θ

generative model m

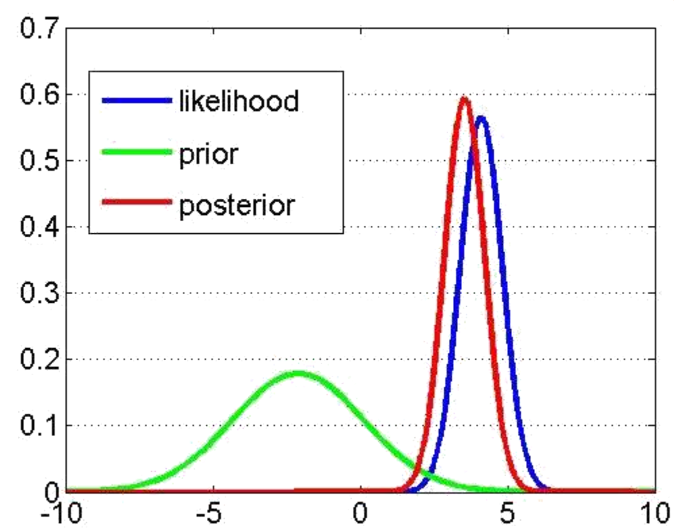


y

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

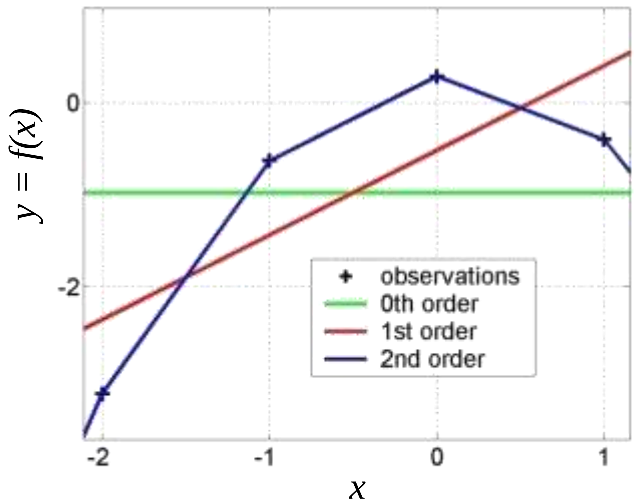
Prior: $p(\theta|m)$

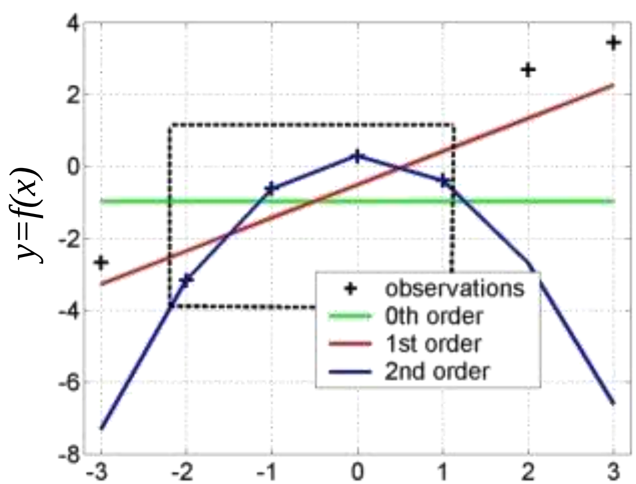
Bayes rule: $$p(\theta|y,m) = \frac{p(y|\theta,m)\,p(\theta|m)}{p(y|m)}$$



Bayesian model comparison

Principle of parsimony :
« plurality should not be assumed without necessity »

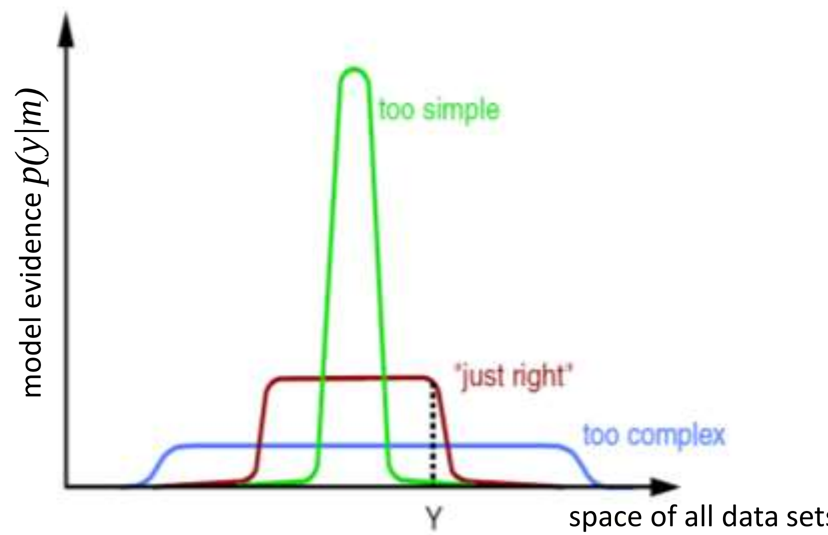




Model evidence:

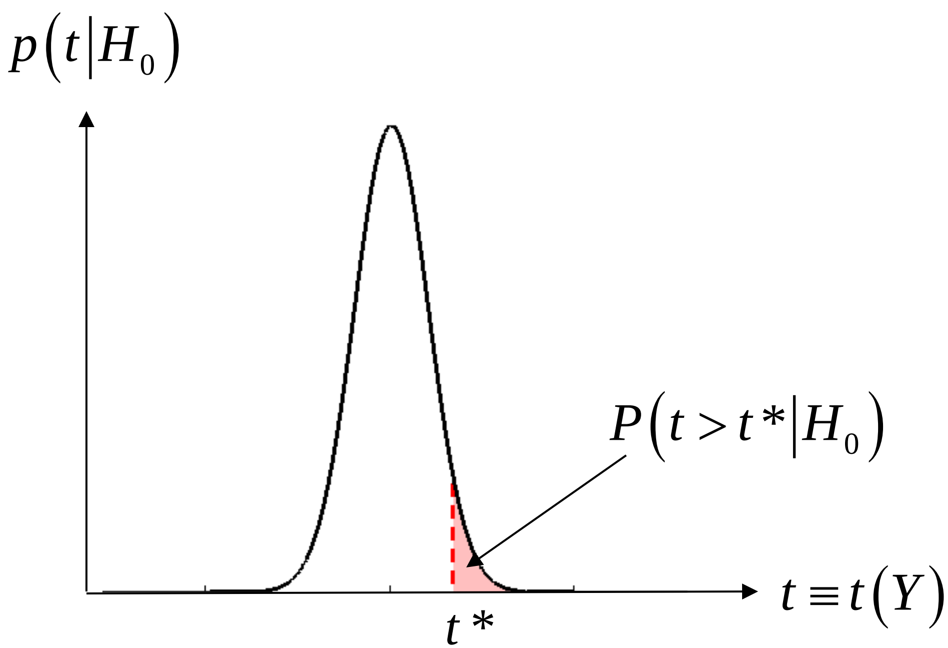
$$p(y|m) = \int p(y|\theta, m)\, p(\theta|m)\, d\theta$$

"Occam's razor" :



Bayesian versus frequentist hypothesis testing

- define the null, e.g.: $H_0 : \theta = 0$



- estimate parameters (obtain test stat.)

- apply decision rule, i.e.:

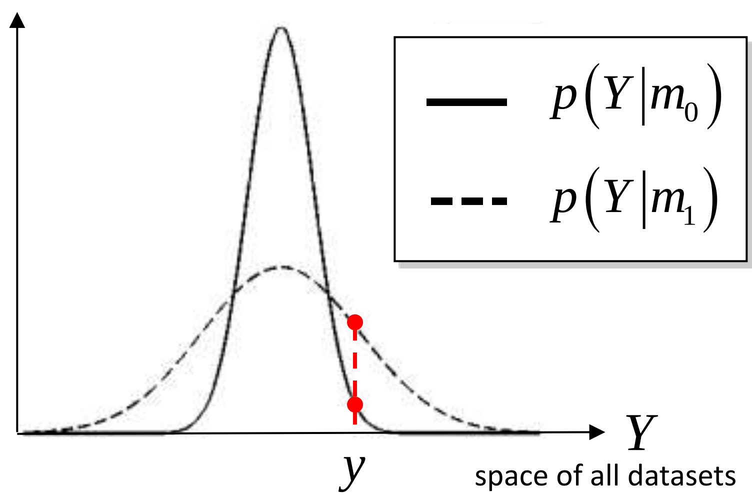
if $P(t > t^* | H_0) \leq \alpha$ then reject H0

classical (null) hypothesis testing

- define two alternative models, e.g.:

$$m_0 : p(\theta|m_0) = \begin{cases} 1 & \text{if } \theta = 0 \\ 0 & \text{otherwise} \end{cases}$$

$$m_1 : p(\theta|m_1) = N(0, \Sigma)$$



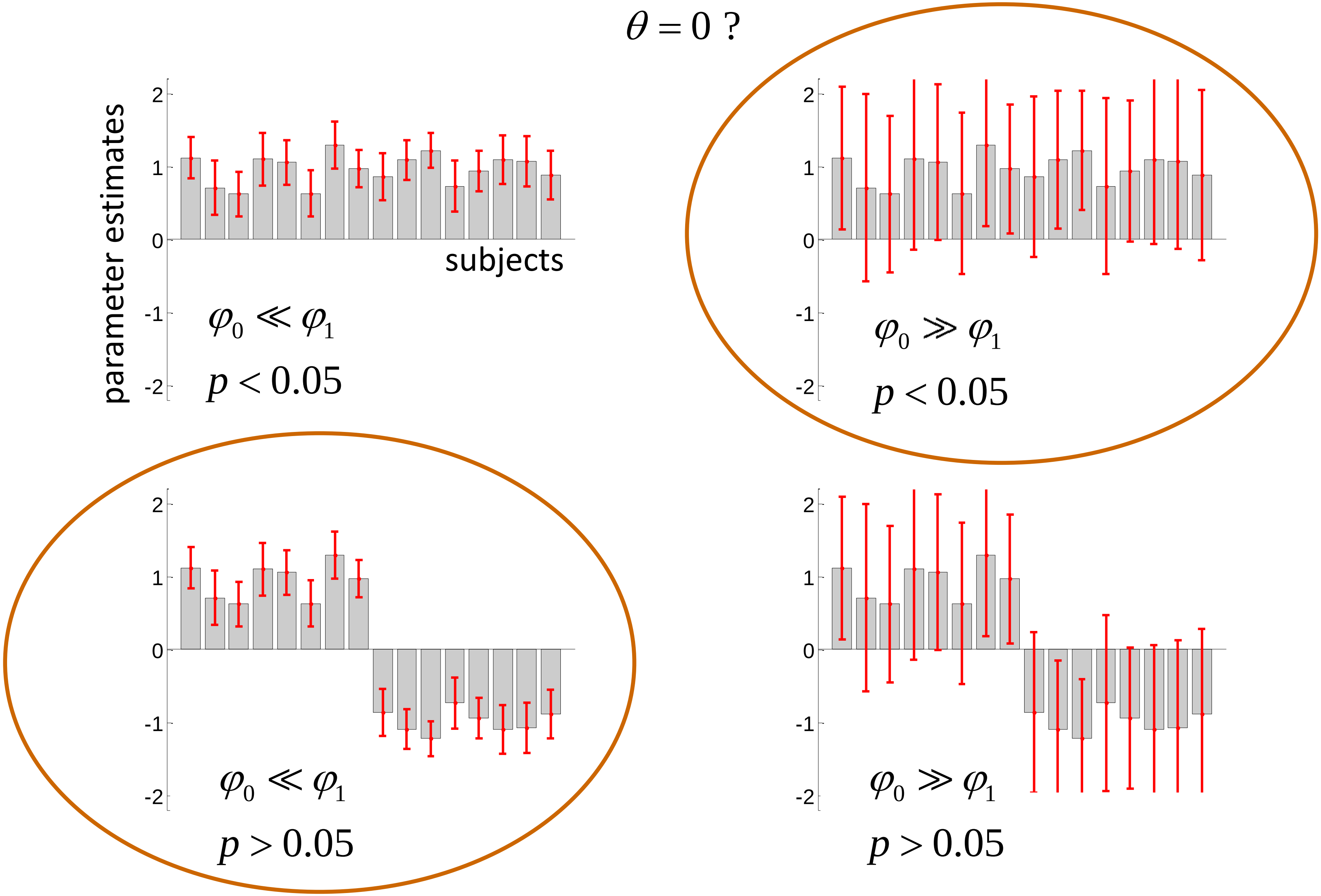
- apply decision rule, e.g.:

if $\dfrac{P(m_0|y)}{P(m_1|y)} \geq \alpha$ then accept m_0

Bayesian model comparison

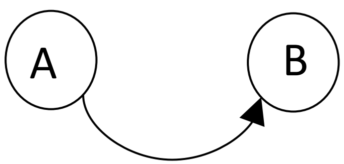
Group-level model selection

$\theta = 0$?

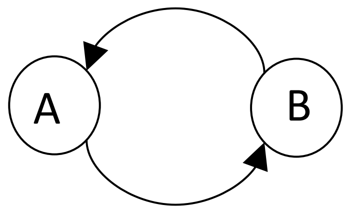


Family-level inference

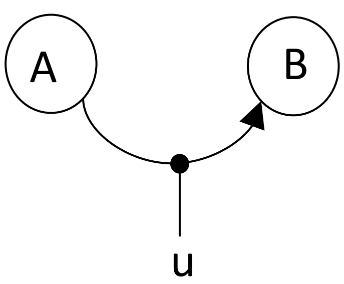
$P(m_1|y) = 0.04$



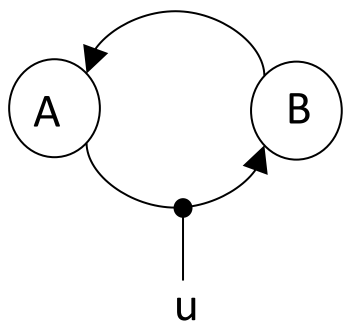
$P(m_2|y) = 0.25$



$P(m_2|y) = 0.01$



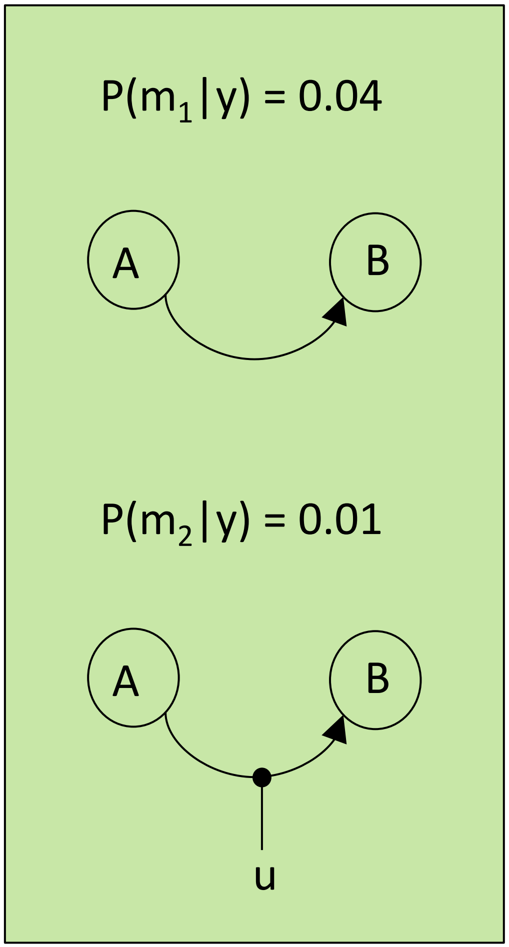
$P(m_2|y) = 0.7$



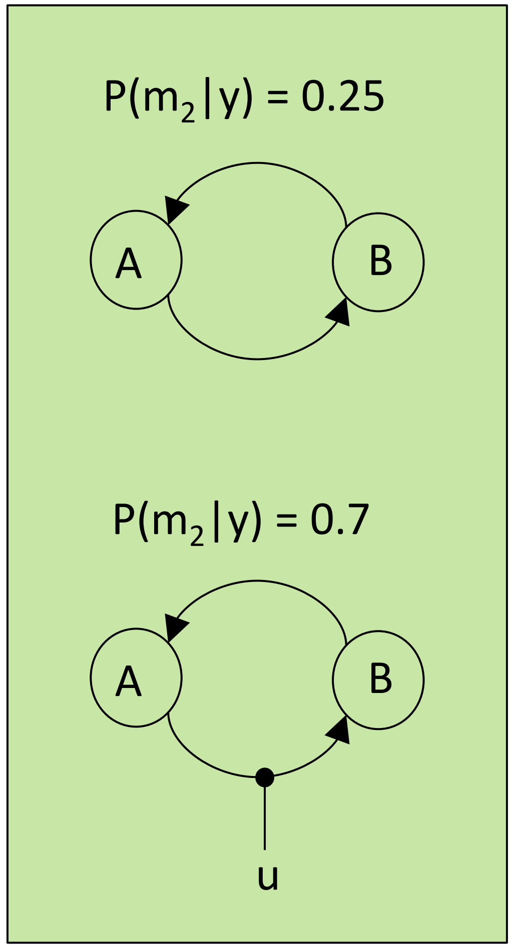
model selection error risk:

$$P\left(e=1|y\right)=1-\max_m P\left(m|y\right)$$
$$=0.3$$

Family-level inference



$P(f_1|y) = 0.05$



$P(f_2|y) = 0.95$

model selection error risk:

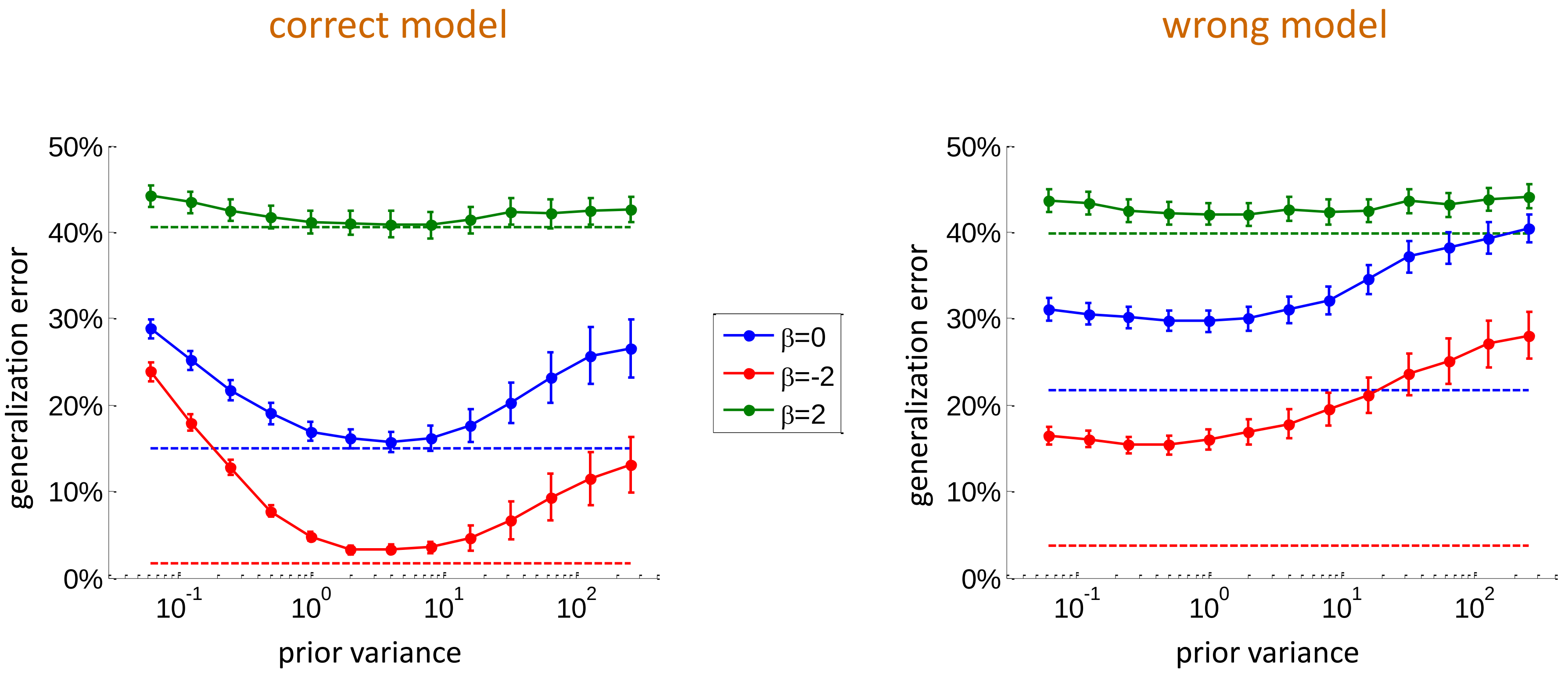
$$P\left(e=1\middle|y\right)=1-\max_m P\left(m\middle|y\right)$$
$$=0.3$$

family inference
(pool statistical evidence)

$$P\left(f\middle|y\right)=\sum_{m\in f}P\left(m\middle|y\right)$$

$$P\left(e=1\middle|y\right)=1-\max_f P\left(f\middle|y\right)$$
$$=0.05$$

Priors and the bias-variance trade-off



Type, role and impact of priors

- Types of priors:
  - ✓ Explicit priors on *model parameters* (e.g., Gaussian)
  - ✓ Implicit priors on *model functional form* (e.g., evolution & observation functions)
  - ✓ Choice of “interesting” *data features* (e.g., response magnitude vs response profile)

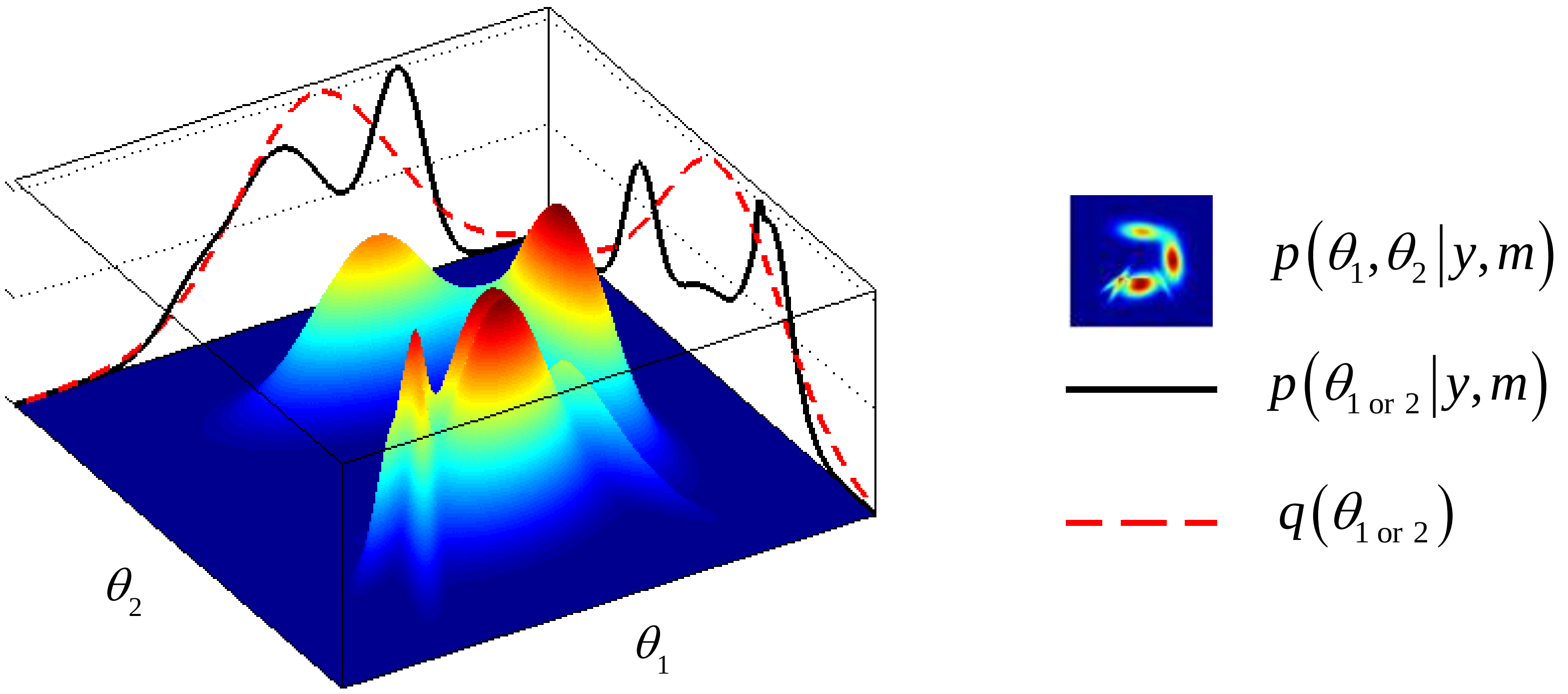
- Role of explicit priors (on model parameters):
  - ✓ Resolving the *ill-posedness* of the inverse problem
  - ✓ Avoiding *overfitting* (cf. generalization error)

- Impact of priors:
  - ✓ On parameter posterior distributions (cf. “shrinkage to the mean” effect)
  - ✓ On model evidence (cf. “Occam’s razor”)

Overview of the talk

- ✓ Introduction to Bayesian inference
- ✓ The variational approach to approximate Bayesian inference
- ✓ VBA toolbox

Why do we need approximations?



The Laplace approximation

$$t(\theta) = \ln p(y|\theta, m) + \ln p(\theta|m)$$

$$\approx t(\hat{\theta}) + \underbrace{(\theta - \hat{\theta})^{\mathrm{T}} \left.\frac{\partial t}{\partial \theta}\right|_{\hat{\theta}}}_{0} + \frac{1}{2}(\theta - \hat{\theta})^{\mathrm{T}} \underbrace{\left.\frac{\partial^2 t}{\partial \theta^2}\right|_{\hat{\theta}}}_{-H(\hat{\theta})} (\theta - \hat{\theta})$$

$$\ln p(y|m) = \ln \int \exp(t(\theta))\, d\theta$$

$$\approx \underbrace{t(\hat{\theta}) + \frac{p}{2}\ln 2\pi - \frac{1}{2}\ln\left|H(\hat{\theta})\right|}_{F_{\mathrm{Laplace}}}$$

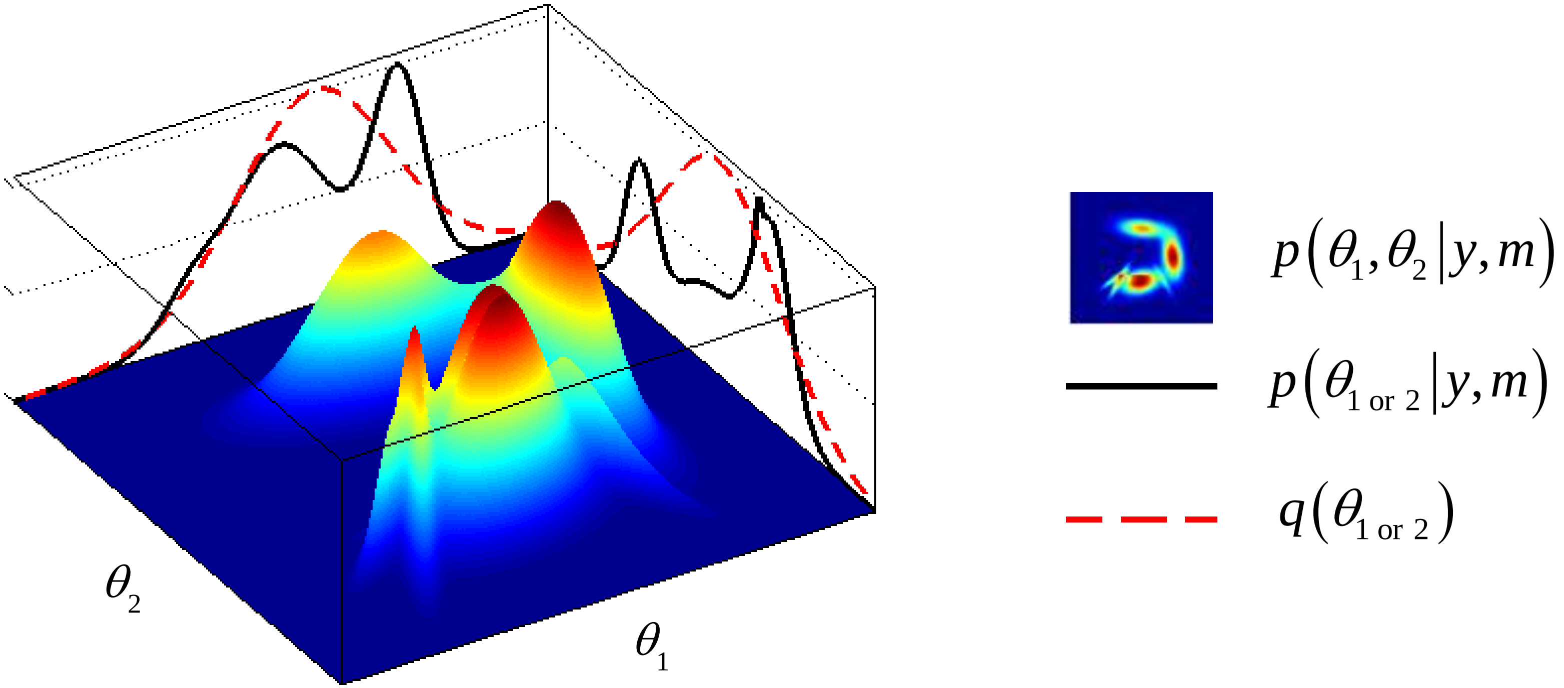
The Free energy lower bound

$$
\begin{aligned}
F &= \left\langle \ln p\left(y \middle| \theta, m\right) + \ln p\left(\theta \middle| m\right) \right\rangle_q + S(q) \\
&= \left\langle \ln p\left(y \middle| \theta, m\right) \right\rangle_q - KL\left(p(\theta \,|\, m) \;; q(\theta) \right) \\
&= \ln p(y \,|\, m) - KL\left(p(\theta \,|\, y, m) \;; q(\theta) \right)
\end{aligned}
$$

VB and the Free Energy

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

→ **VB** : maximize the free energy $F(q)$ w.r.t. the approximate posterior $q(\theta)$ under some (e.g., *mean field, Laplace*) simplifying constraint



The mean-field approximation

$$F = \left\langle \ln p\left(y \middle| \theta, m\right) + \ln p\left(\theta \middle| m\right) \right\rangle_q + S\left(q\right)$$

$$q\left(\theta\right) \approx q_1\left(\theta_1\right) q_2\left(\theta_2\right)$$

$$\frac{\delta F}{\delta q_2} = 0 \Rightarrow q_2\left(\theta_2\right) \propto \exp\left\langle \ln p\left(y \middle| \theta, m\right) + \ln p\left(\theta \middle| m\right) \right\rangle_{q_1}$$

The frequentist limit to the model evidence

$$F = \left\langle \ln p\left(y \middle| \theta, m\right) + \ln p\left(\theta \middle| m\right) \right\rangle_q + S\left(q\right)$$

$$\xrightarrow[\text{flat priors}]{p(\theta)\to 1} \left\langle \ln p\left(y \middle| \theta, m\right) \right\rangle_q + S\left(q\right)$$

$$\xrightarrow[\text{point mass approximation}]{q(\theta)\to \delta(\hat{\theta})} \underbrace{\ln p\left(y \middle| \hat{\theta}, m\right)}_{\text{frequentist log-likelihood}}$$

BIC and AIC

→ BIC: Laplace approximation at the asymptotic limit

$$\Sigma \xrightarrow{n\to\infty} \frac{1}{n} I_p$$

$$F_{\text{Laplace}} \xrightarrow{n\to\infty} \underbrace{\ln p\left(y \middle| \hat{\theta}, m\right) - \frac{p}{2} \ln n}_{\text{BIC}}$$

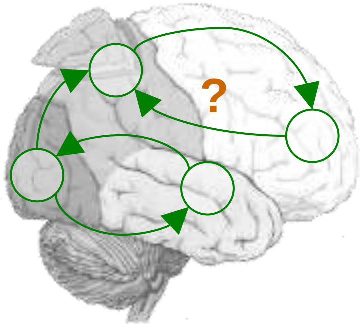
→ AIC: approximation to a frequentist KL-divergence risk!

$$AIC = \ln p\left(y \middle| \hat{\theta}, m\right) - p$$

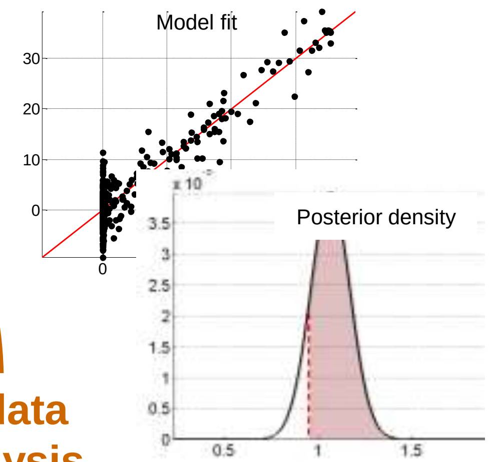
Overview of the talk

- ✓ Introduction: Bayesian inference
- ✓ The variational approach to approximate Bayesian inference
- ✓ VBA toolbox

The three core mathematical problems of VBA



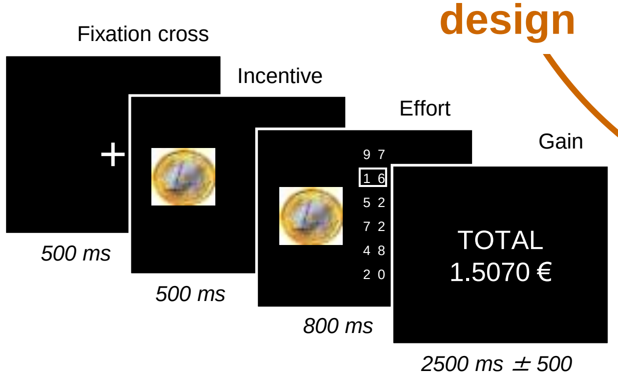
new set of alternative hypotheses



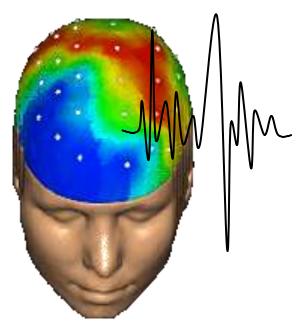
1. biological & behavioural predictions

3. data analysis (variational-Laplace approach)

2. experimental design

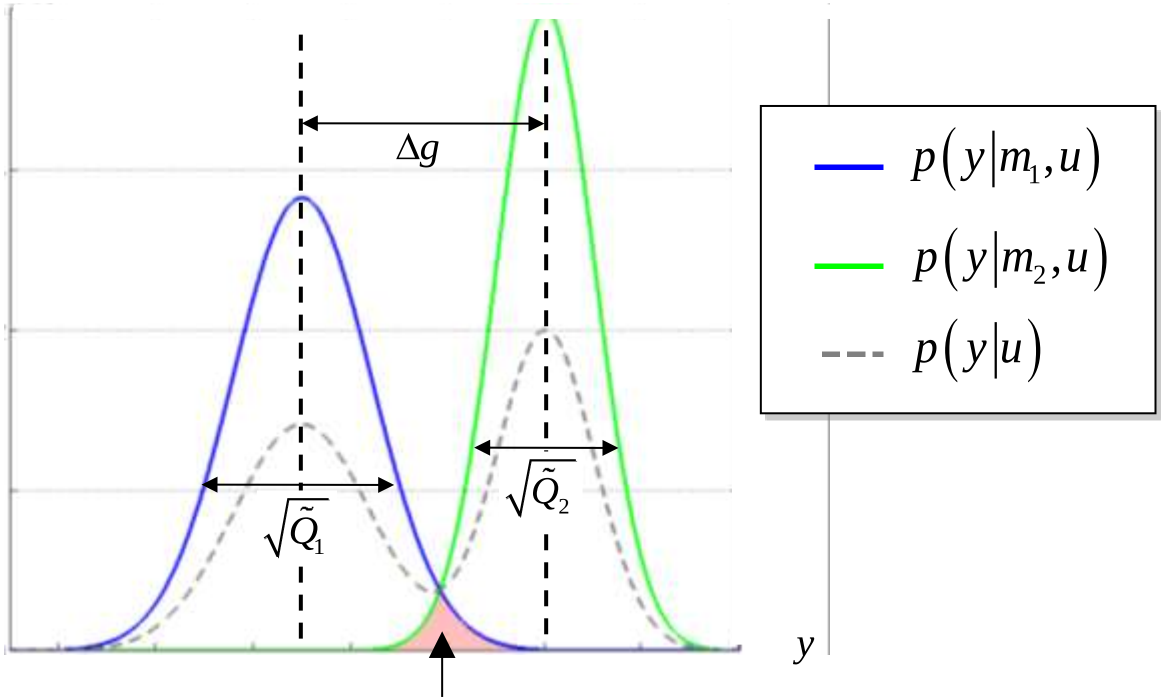


data acquisition



Selection error rate and the Laplace-Chernoff risk

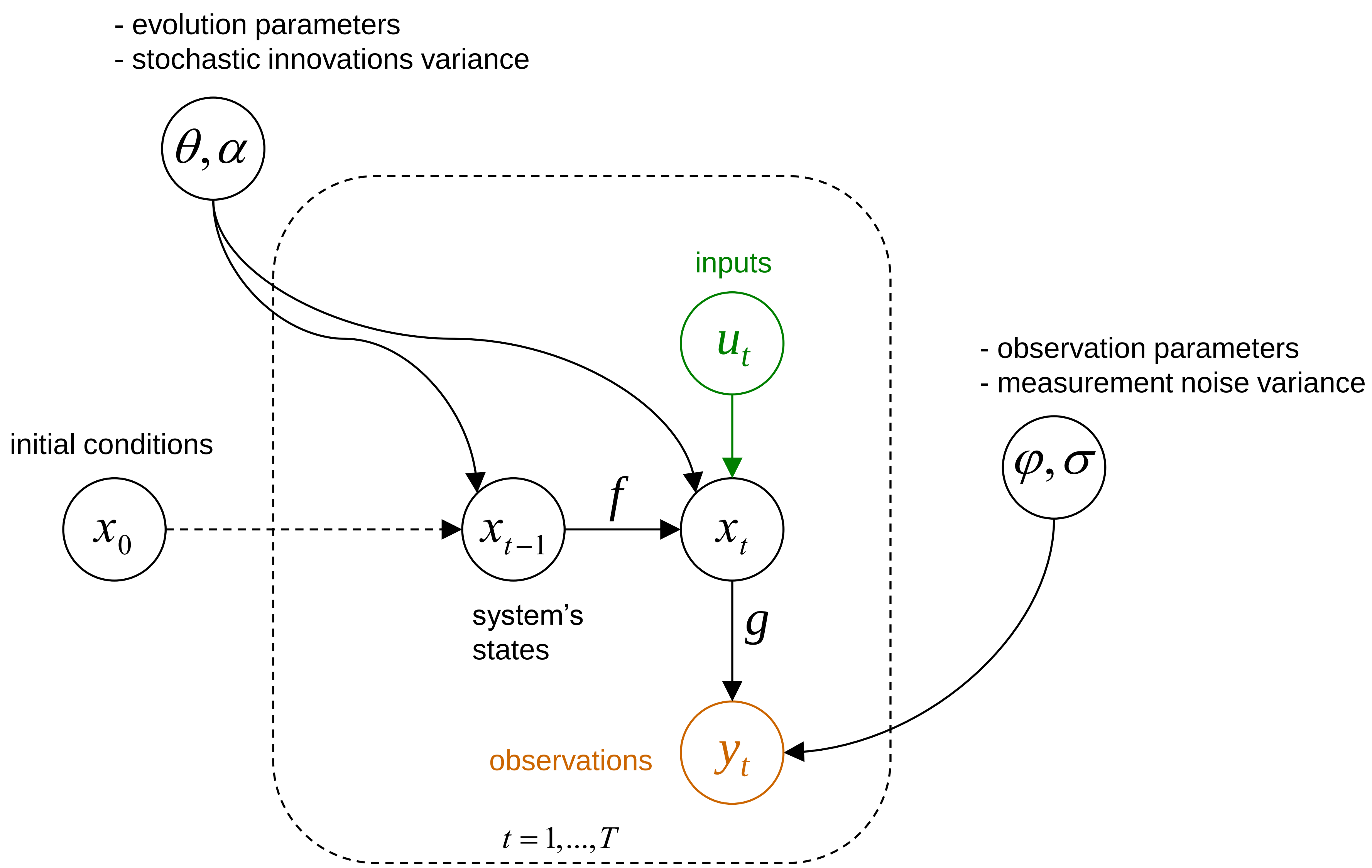
$$b_{LC}(u) = 1 - \frac{1}{2}\log\left(\frac{\Delta g(u)^2}{4\tilde{Q}(u)} + 1\right) \quad \text{if} \quad \tilde{Q}_1(u) \approx \tilde{Q}_2(u) \equiv \tilde{Q}(u)$$



$$p(\hat{e}=1|u) = 1 - \int_Y \max_m \left[p(m)\, p(y|m,u) \right] dy$$

Daunizeau, Friston et al., PLoS Comp. Biol. , 2011

VBA: model structure



Daunizeau, Friston et al., Physica D, 2009

VBA: how to?

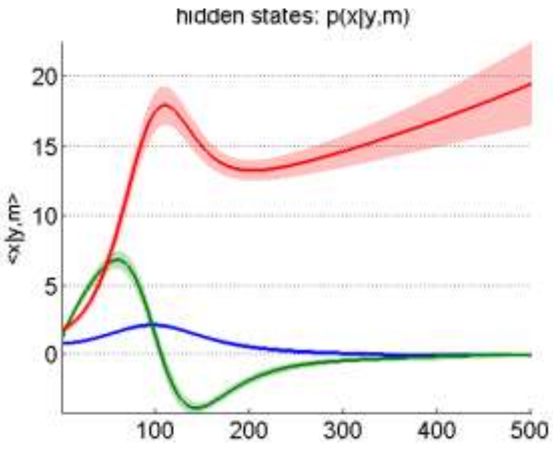
✓ You need to provide:

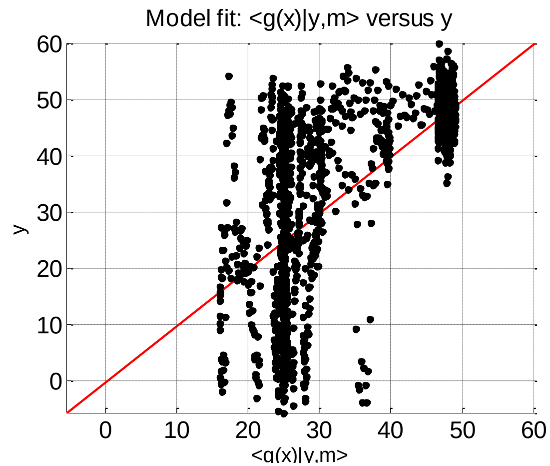
- The data
- The system's inputs (can be left blank)
- The observation and evolution functions (the latter can be left blank)
- The model dimensions

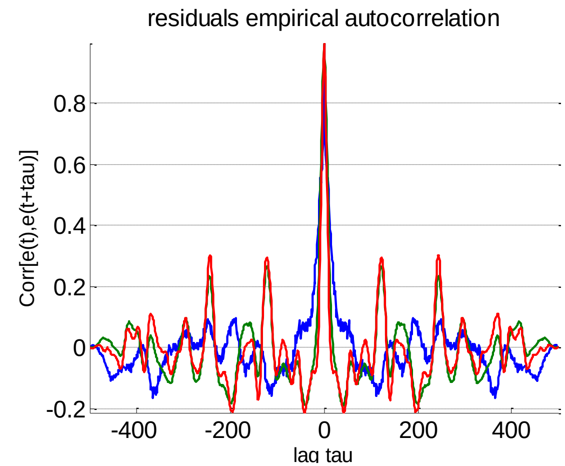
✓ You can specify (otherwise VBA uses defaults):

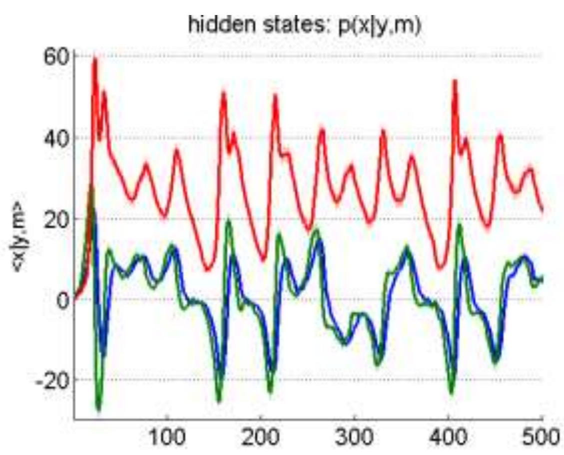
- The priors (mean and covariances)
- Inversion options (free-energy tolerance, display flag, etc…)

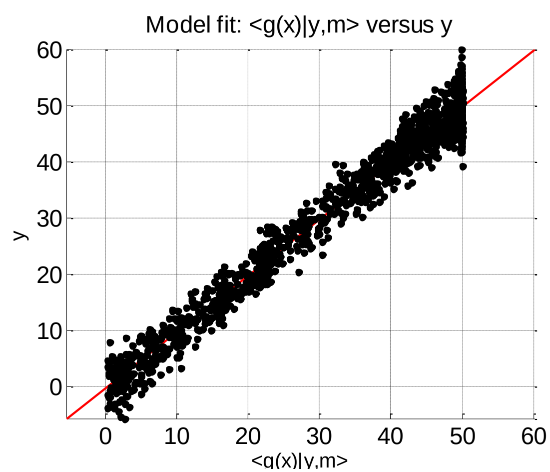
Model inversion diagnostics (I)

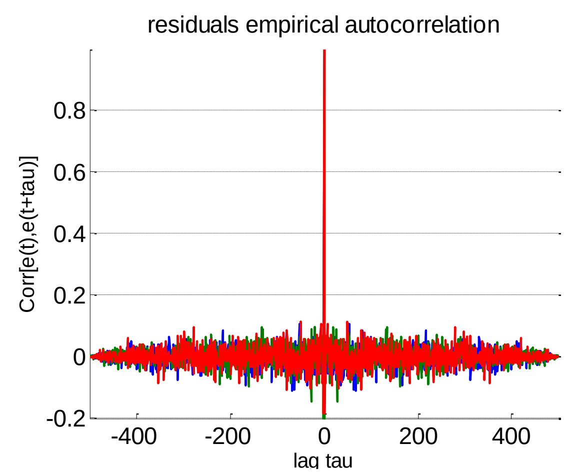




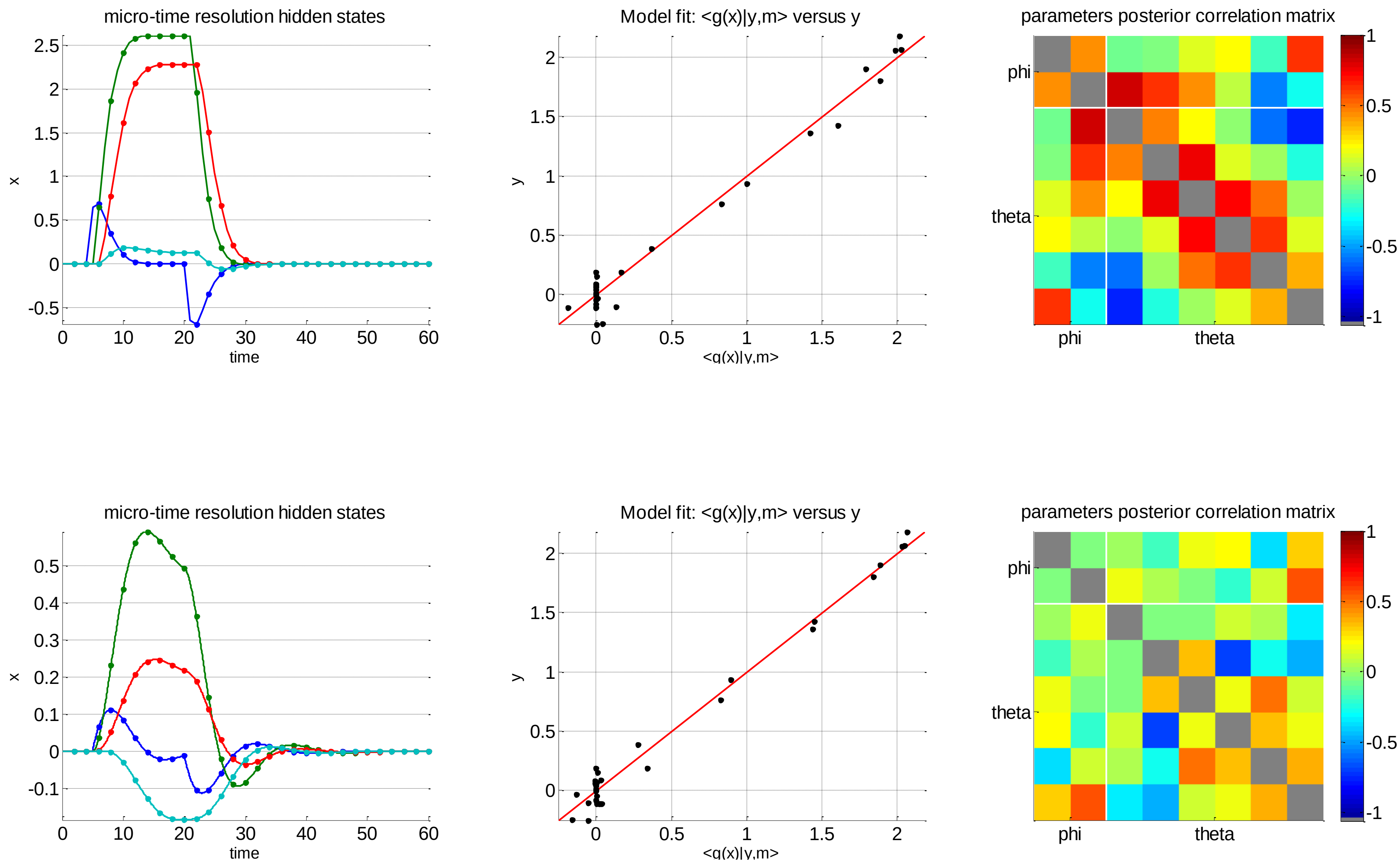




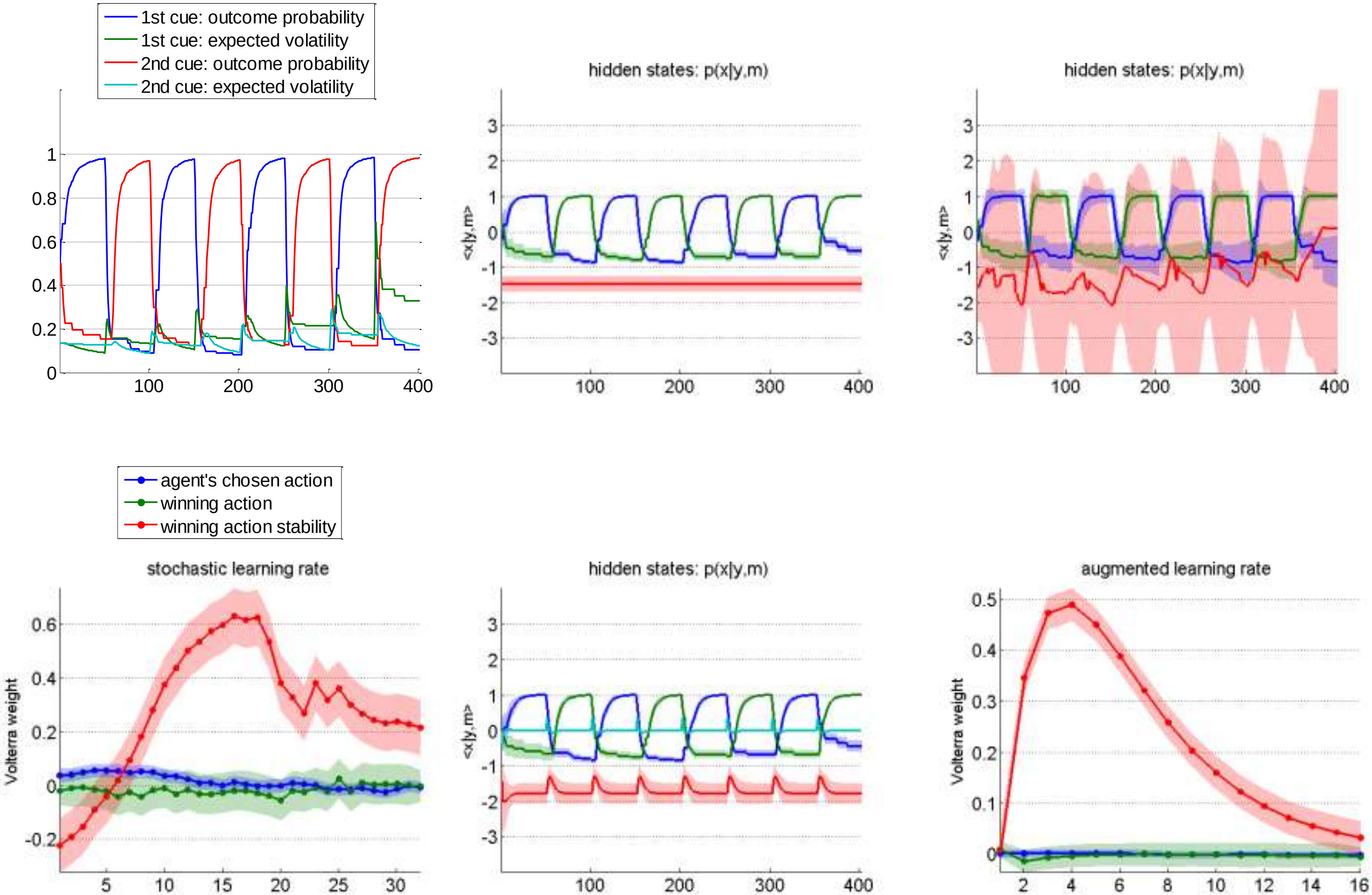




Model inversion diagnostics (II)



Model improvement



Statistical features

- ✓ Mediation analysis (Sobel test, Monte-Carlo sampling)
- ✓ Cross-validation (balanced accuracy, PRESS, etc…)
- ✓ Missing data handling
- ✓ Sparse estimation (approximate lasso)
- ✓ Autoregressive and/or state-dependent noise
- ✓ Arbitrary likelihood functions (e.g., reaction time distributions…)
- ✓ Mixed-effects modelling (empirical priors based upon group statistics)
- ✓ Clustering and Dirichlet processes
- ✓ On-line adaptive experimental designs
- ✓ …

VBA's (growing) library of models

- ✓ Learning models (Q-learning, hierarchical Bayesian learning, etc…)
- ✓ Decision models (delay/risk/effort discounting, etc…)
- ✓ Neural systems models (DCM, Hodgkin-Huxley, Fitz-Hugh-Nagumo, etc…)
- ✓ Dynamical systems models (double-well, Lorenz attractor, Henon map, etc…)
- ✓ …
- ✓ Recursive Theory of Mind
- ✓ Spiking models for calcium imaging
- ✓ Race models and other variants of diffusion-drift models
- ✓ …

Acknowledgements

- ✓ Karl J Friston, Klaas E Stephan
- ✓ Lionel Rigoux
- ✓ members of the MBB team