Bayesian model selection & averaging

Klaas Enno Stephan

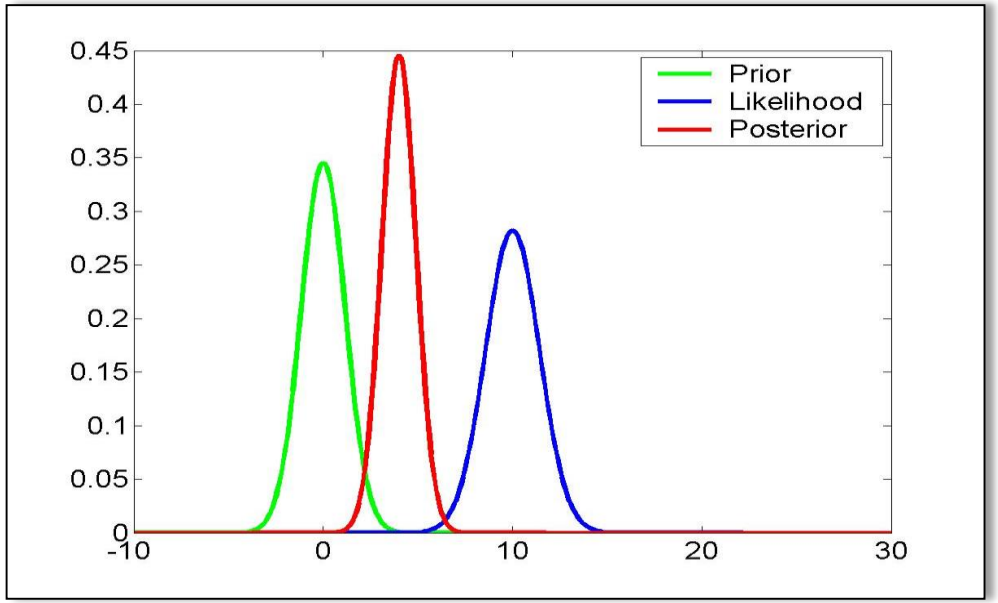
Translational Neuromodeling Unit

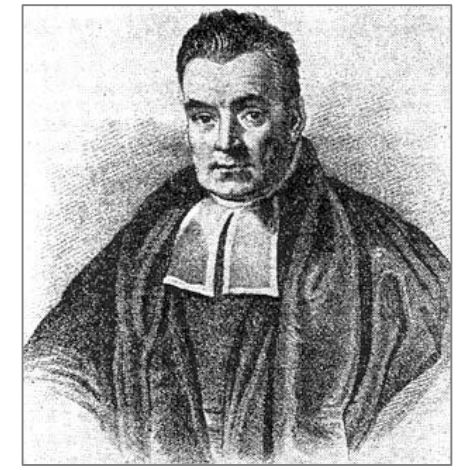
Universität Zürich[UZH]

ETH

Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Bayes' theorem



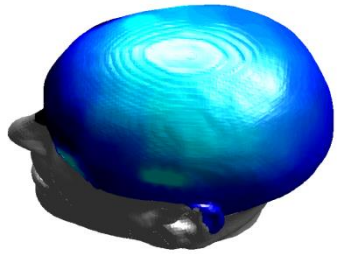


The Reverend Thomas Bayes
(1702-1761)

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

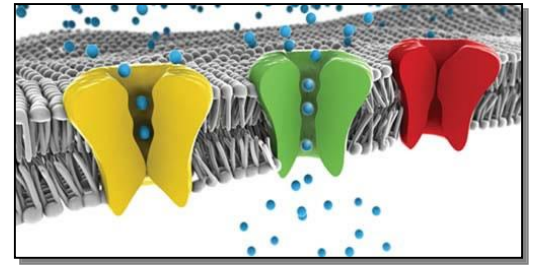
posterior = likelihood • prior / evidence

Generative model



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

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

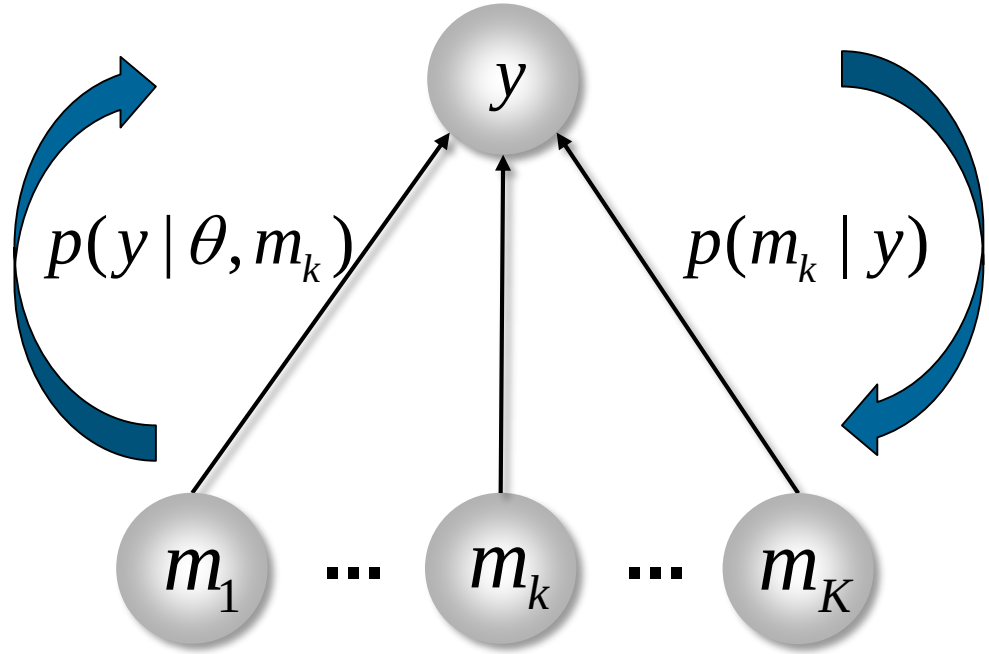


1. enforces mechanistic thinking: how could the data have been caused?
2. generate synthetic data (observations) by sampling from the prior – can model explain certain phenomena at all?
3. inference about parameters → p(θ|y)
4. inference about model structure: formal approach to disambiguating mechanisms → p(y|m) or p(m|y)

Long-term goal: Differential diagnosis based on generative models of disease symptoms

SYMPTOM
(behaviour
or physiology)

HYPOTHETICAL MECHANISM

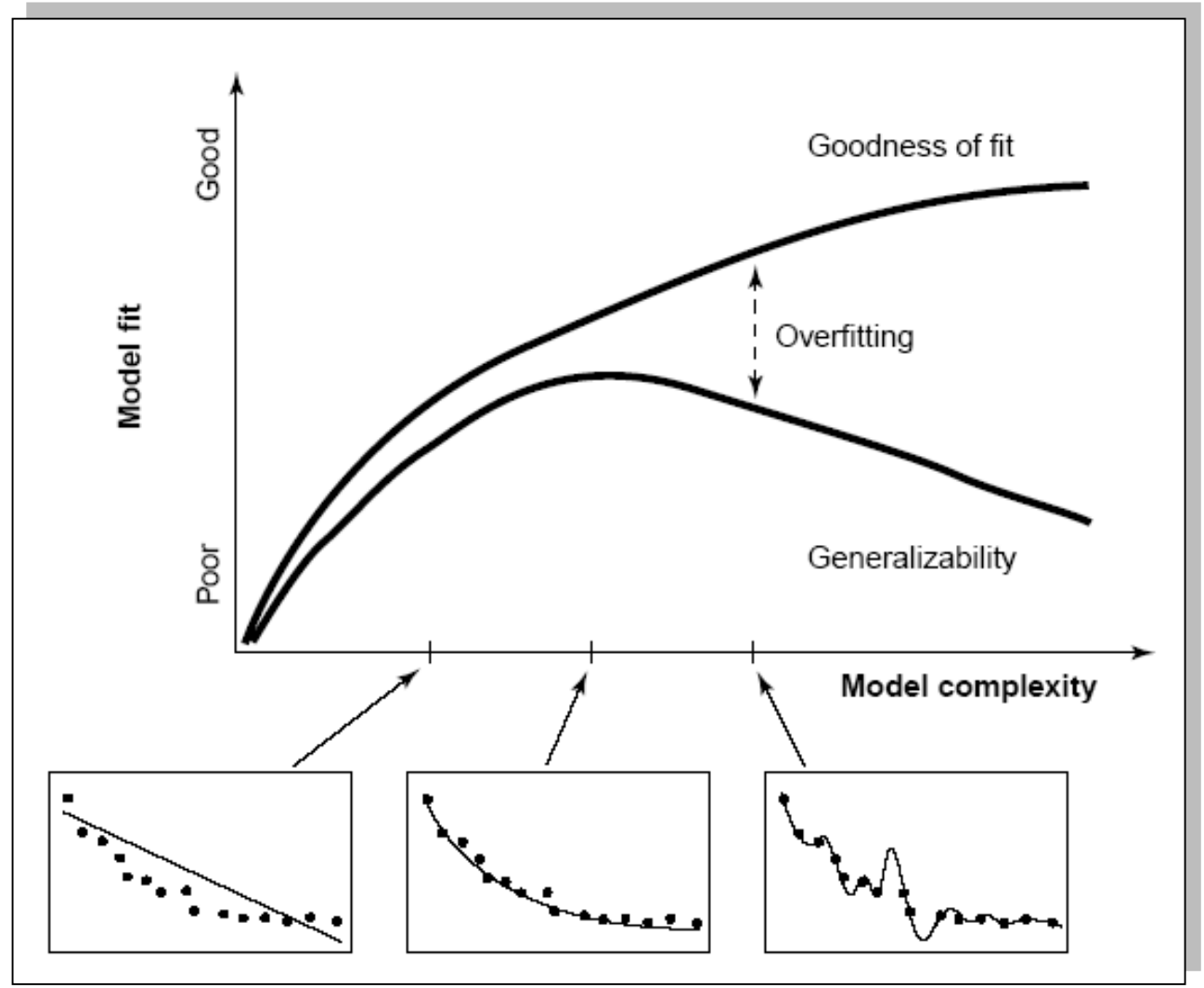


Model comparison and selection

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?

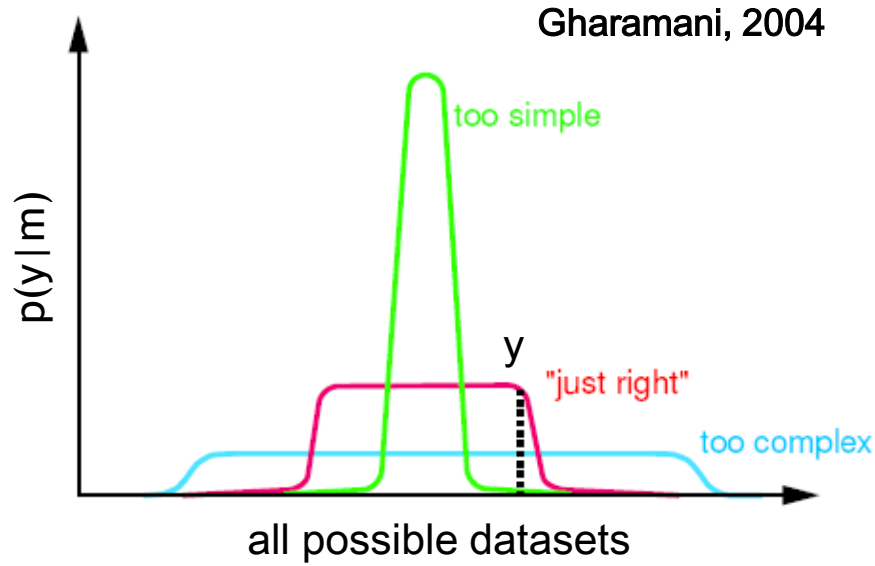


Pitt & Miyung (2002) *TICS*

Bayesian model selection (BMS)

Model evidence (marginal likelihood):

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



→ 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:

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

Approximations to the model evidence in DCM

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

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

No. of parameters

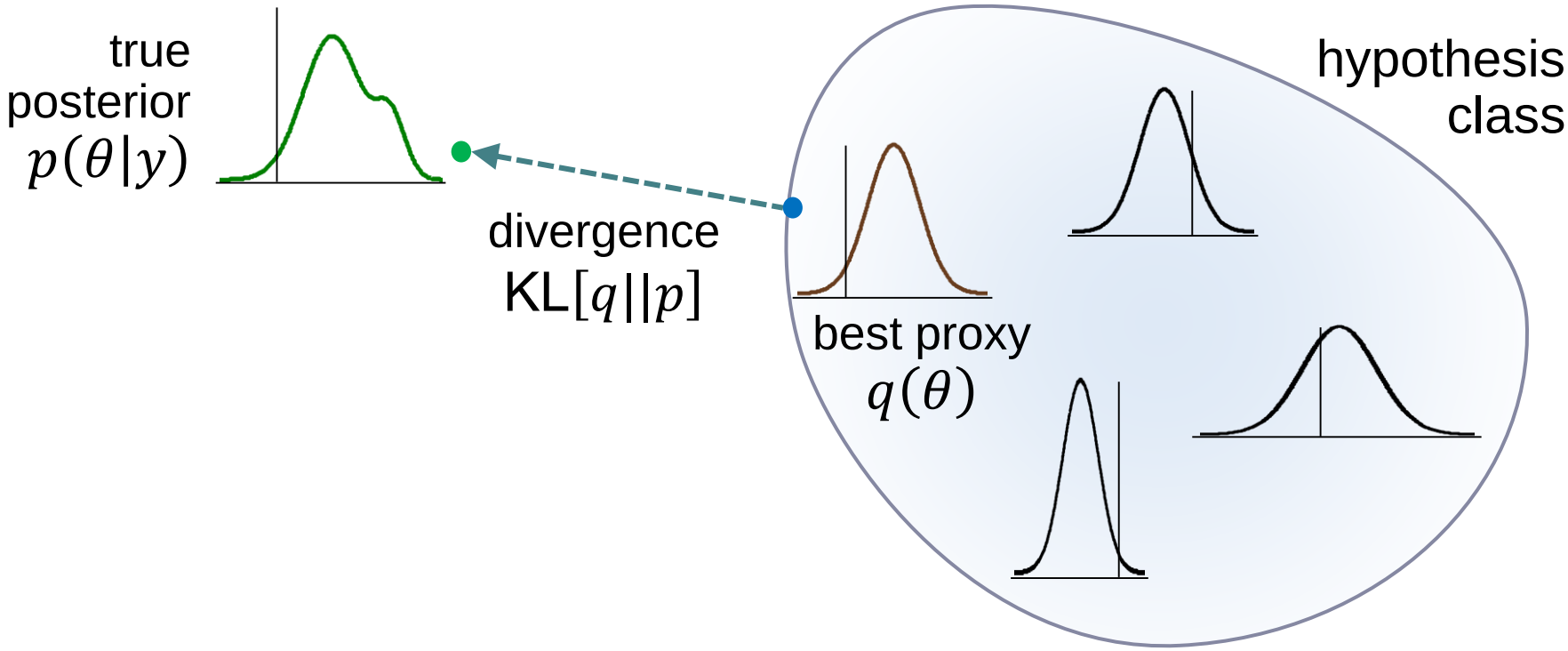
No. of data points

Penny et al. 2004a, *NeuroImage*

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.



The (negative) free energy approximation F

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

divergence ≥ 0 (unknown)

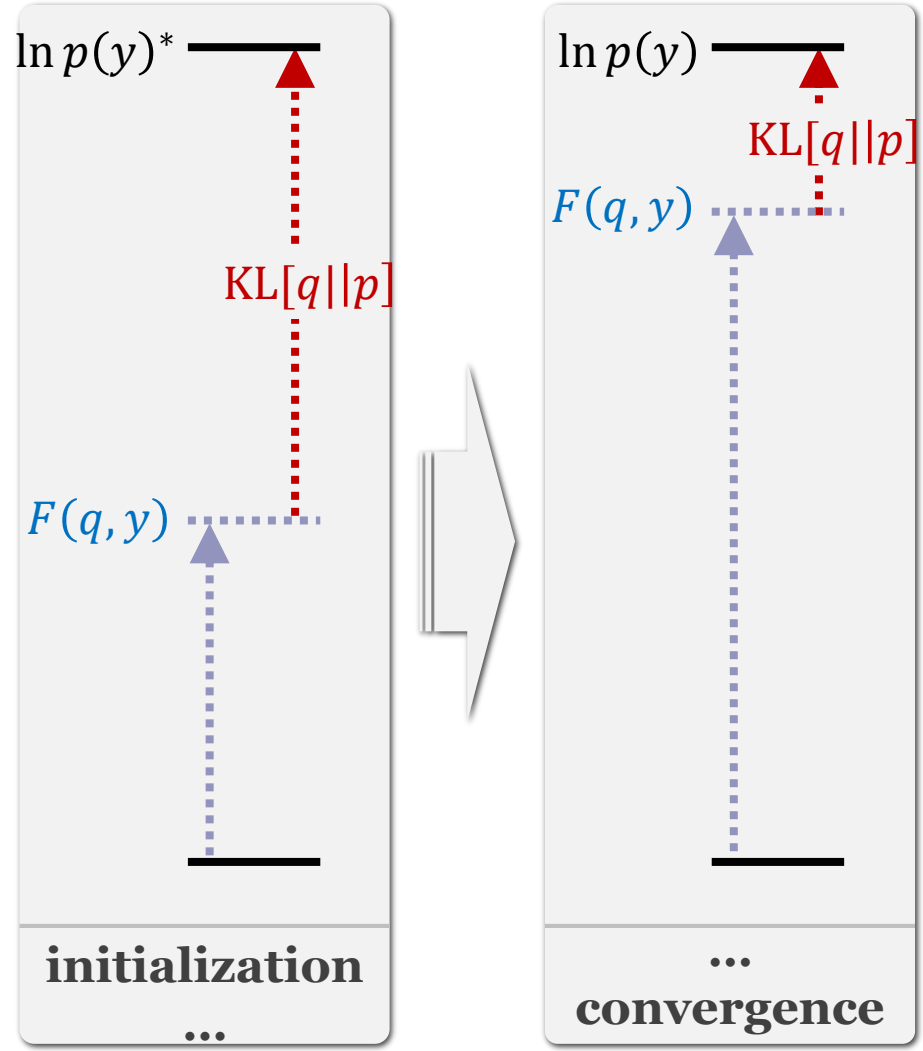
neg. free energy (easy to evaluate for a given q)

$F(q,y)$ 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

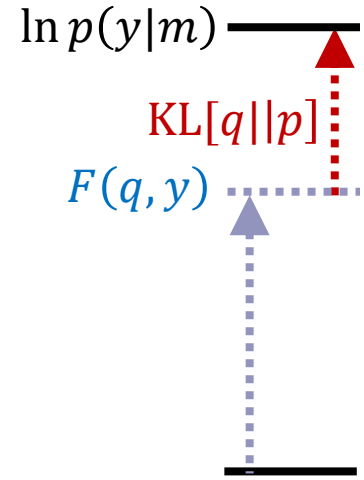
When $F(q,y)$ is maximized, $q(\theta)$ is our best estimate of the posterior.



The (negative) free energy approximation F

F is a lower bound on the log model evidence, where the bound is determined by the KL divergence between an approximate posterior q and the true posterior::

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



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

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

The complexity term in *F*

- In contrast to AIC & BIC, the complexity term of the negative free energy *F* accounts for parameter interdependencies.

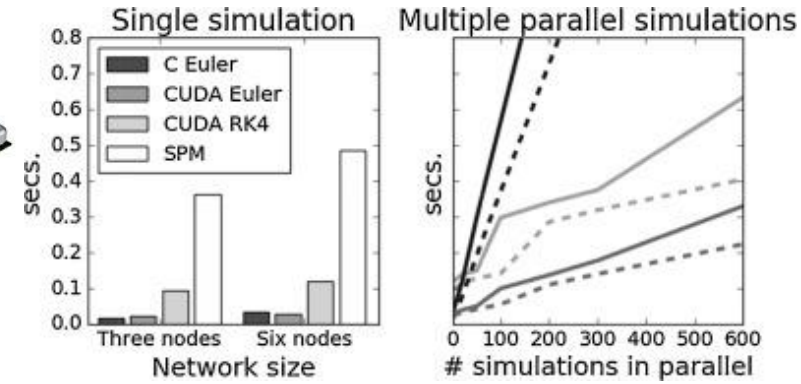
$$KL\left[q(\theta), p(\theta \mid m)\right]$$

$$= \frac{1}{2}\ln\left|C_\theta\right| - \frac{1}{2}\ln\left|C_{\theta|y}\right| + \frac{1}{2}\left(\mu_{\theta|y} - \mu_\theta\right)^T C_\theta^{-1}\left(\mu_{\theta|y} - \mu_\theta\right)$$

- determinant = measure of “volume” (space spanned by the eigenvectors of the matrix)

- 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

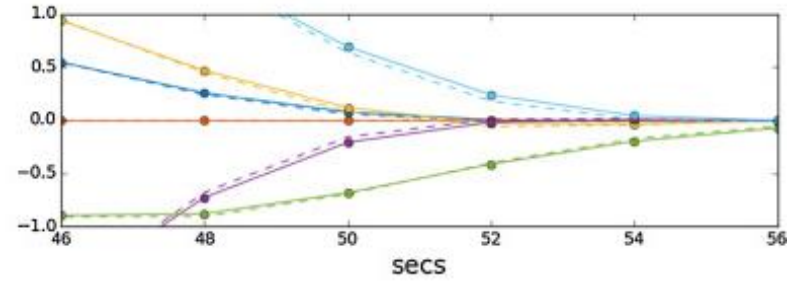
mpdcm: Computing the evidence by sampling



$$\left.\begin{aligned}\dot{x} &= f(x, u_1, \theta_1)\\ \dot{x} &= f(x, u_2, \theta_2)\\ &\vdots\\ \dot{x} &= f(x, u_1, \theta_1)\end{aligned}\right\}$$

mpdcm_integrate(dcms)

$$\left\{\begin{aligned} y_1\\ y_2\\ \vdots\\ y_3\end{aligned}\right.$$



www.translationalneuromodeling.org/tapas

Aponte et al. 2015., *J. Neurosci Meth.*

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:

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

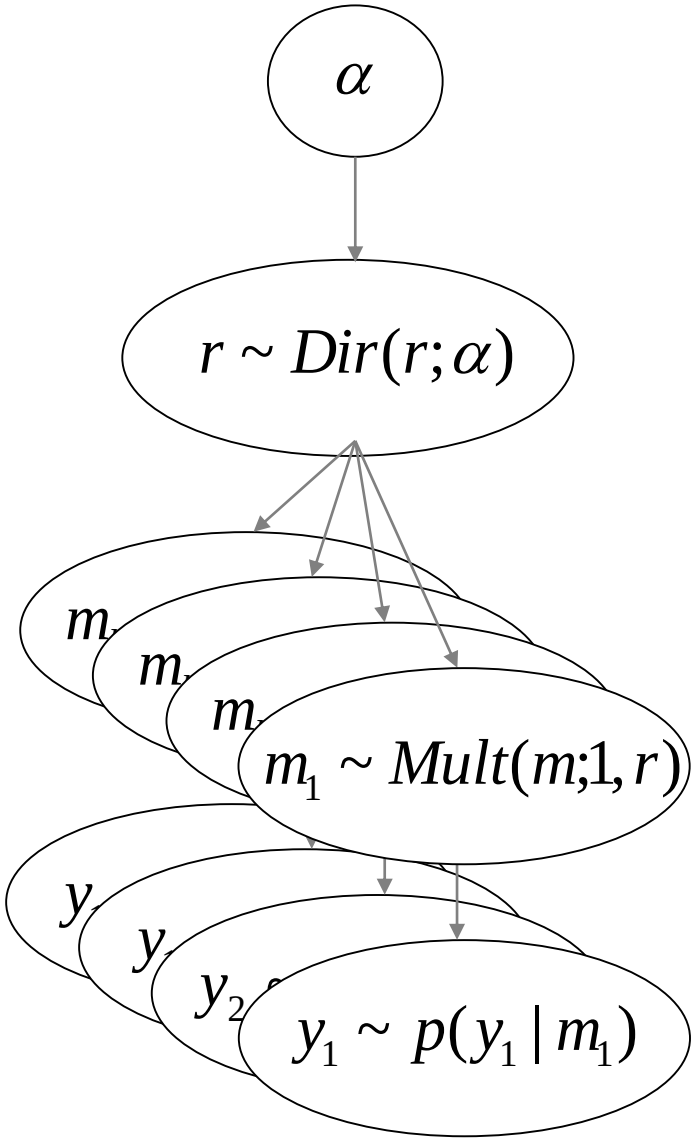
Average Bayes factor (ABF):

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

Problems:

- blind with regard to group heterogeneity
- sensitive to outliers

Random effects BMS for heterogeneous groups



Dirichlet parameters α
= "occurrences" of models in the population

Dirichlet distribution of model probabilities r

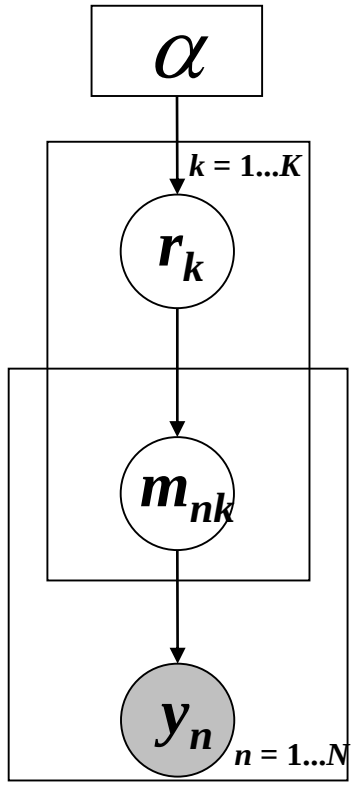
Multinomial distribution of model labels m

Measured data y

Model inversion by Variational Bayes (VB) or MCMC

Stephan et al. 2009a, *NeuroImage*
Penny et al. 2010, *PLoS Comp. Biol.*

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

Model inversion by Variational Bayes (VB) or MCMC

Measured data y

Stephan et al. 2009a, *NeuroImage*
Penny et al. 2010, *PLoS Comp. Biol.*

Reminder: VB in a nutshell (mean-field approximation)

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

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

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

❷ Mean field approx.

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

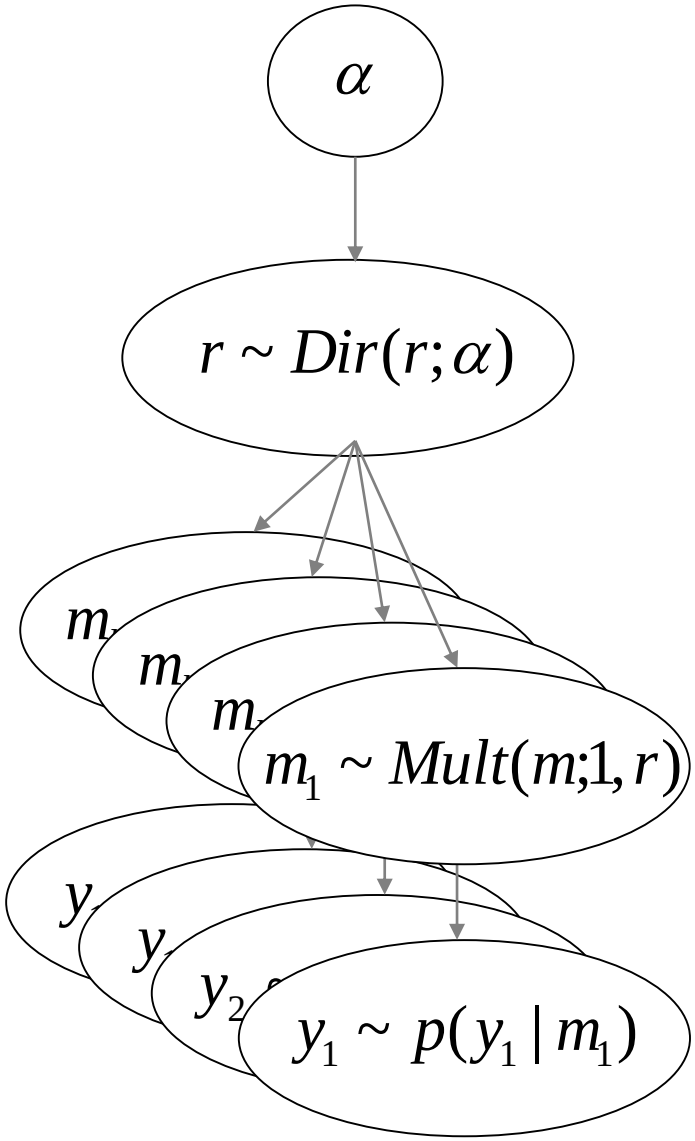
❸ Maximise neg. free energy wrt. q = minimise divergence, by maximising variational energies

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

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

❹ Iterative updating of sufficient statistics of approx. posteriors (e.g., by gradient ascent).

VB solution for random effects BMS



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

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

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

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

Stephan et al. 2009, *NeuroImage*

❶ Write down joint probability and take the log

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

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

❷ Mean field approx.

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

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

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

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

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

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

❹ Iterative updating of sufficient statistics of approx. posteriors

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

Until convergence

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

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

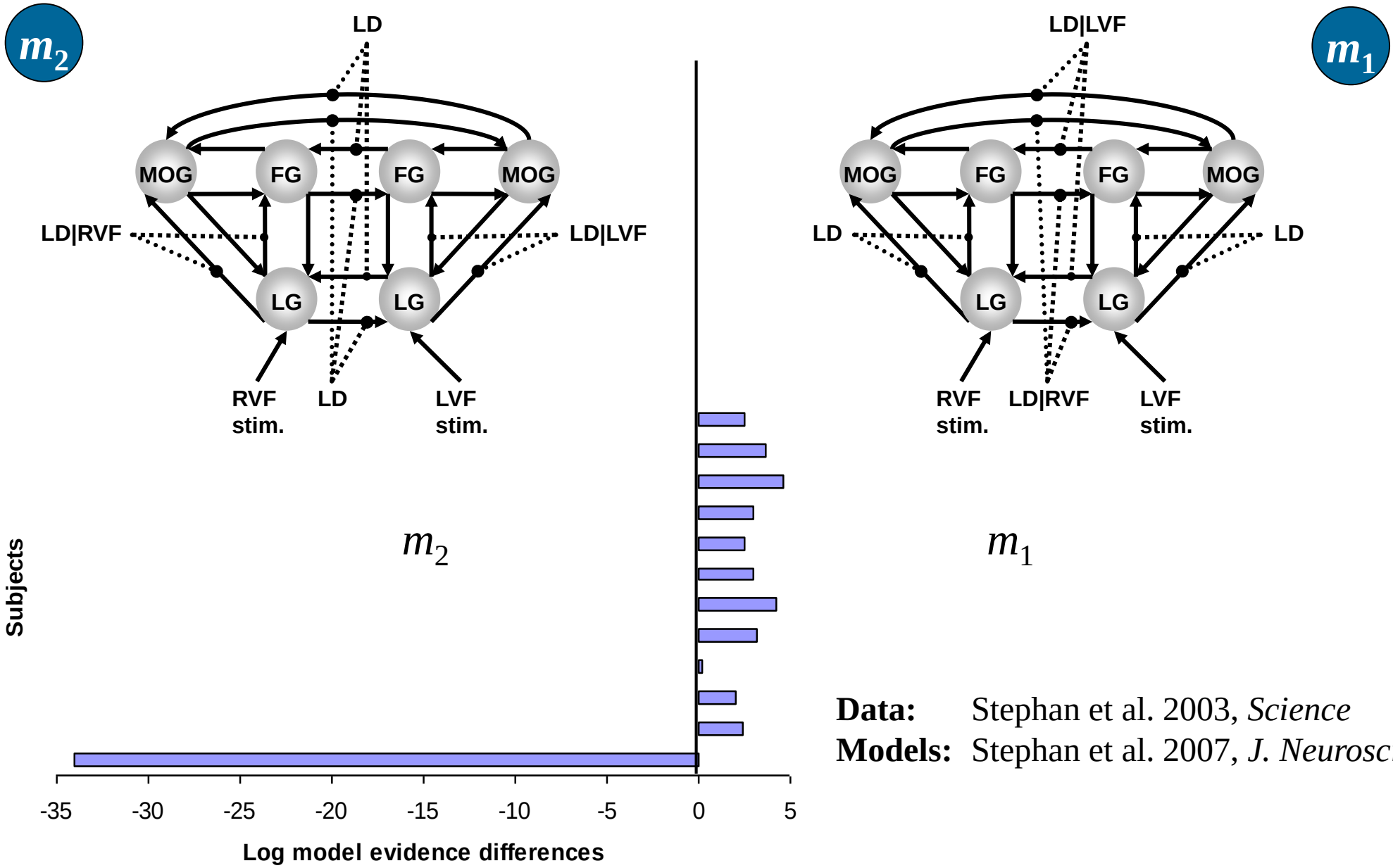
$$\beta_k = \sum_n g_{nk}$$

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

end

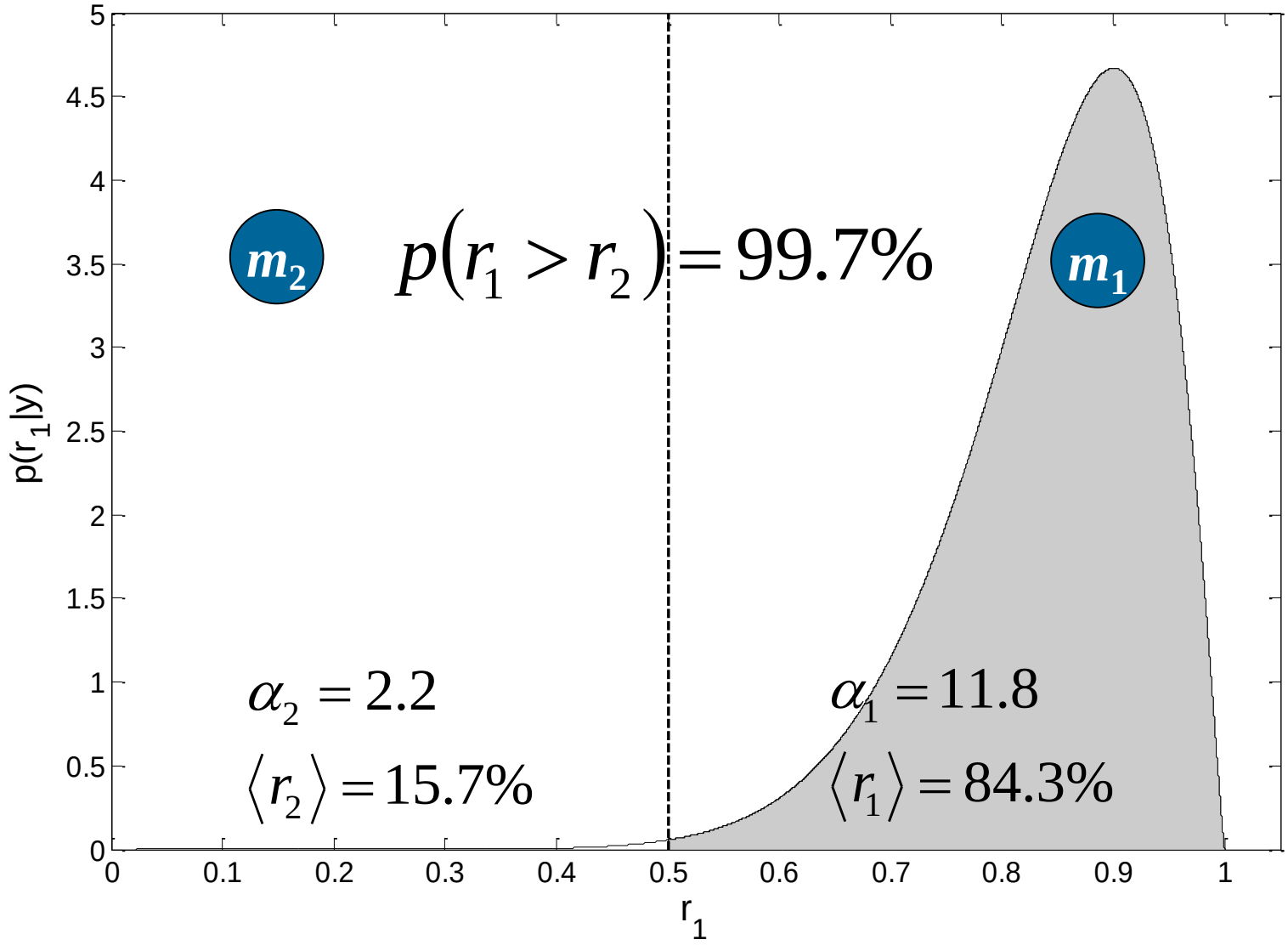
$g_{nk} = q(m_{nk} = 1)$ our (normalized) posterior belief that model *k* generated the data from subject *n*

$\beta_k = \sum_n g_{nk}$ expected number of subjects whose data we believe were generated by model *k*



Data: Stephan et al. 2003, *Science*

Models: Stephan et al. 2007, *J. Neurosci.*



Stephan et al. 2009a, *NeuroImage*

How can we report the results of random effects BMS?

1. **Dirichlet parameter estimates**

$$\alpha$$

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

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

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

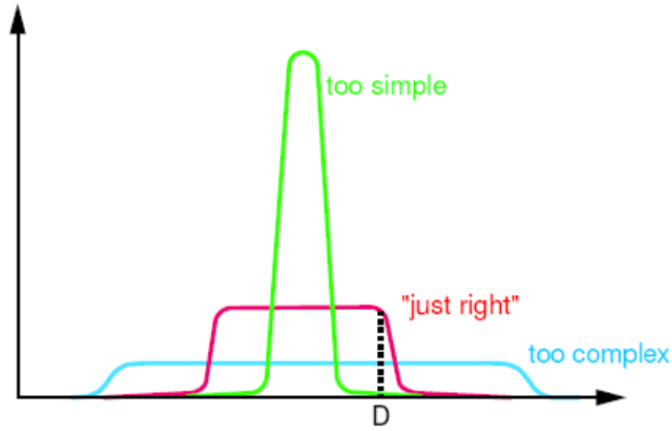
$$\exists k \in \{1...K\}, \forall j \in \{1...K \mid j \neq k\}:$$

$$\varphi_k = p(r_k > r_j \mid y; \alpha)$$

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

Overfitting at the level of models

- ↑ #models ⇒ ↑ risk of overfitting
- solutions:
  - regularisation: definition of model space = choosing priors p(m)
  - family-level BMS
  - Bayesian model averaging (BMA)



posterior model probability:

$$p(m \mid y)$$

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

BMA:

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

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

Model space partitioning or: Comparing model families

- partitioning model space into K subsets or families:

$$M = \{ f_1, \ldots, f_K \}$$

- pooling information over all models in these subsets allows one to compute the probability of a model family, given the data

$$p(f_k)$$

- effectively removes uncertainty about any aspect of model structure, other than the attribute of interest (which defines the partition)

Stephan et al. 2009, *NeuroImage*
Penny et al. 2010, *PLoS Comput. Biol.*

Family-level inference: fixed effects

- We wish to have a uniform prior at the family level:

$$p(f_k) = \frac{1}{K}$$

- This is related to the model level via the sum of the priors on models:

$$p(f_k) = \sum_{m \in f_k} p(m)$$

- Hence the uniform prior at the family level is:

$$\forall m \in f_k : p(m) = \frac{1}{K|f_k|}$$

- The probability of each family is then obtained by summing the posterior probabilities of the models it includes:

$$p(f_k \mid y_{1..N}) = \sum_{m \in f_k} p(m \mid y_{1..N})$$

Penny et al. 2010, *PLoS Comput. Biol.*

Family-level inference: random effects

- The frequency of a family in the population is given by:

$$s_k = \sum_{m \in f_k} r_m$$

- In RFX-BMS, this follows a Dirichlet distribution, with a uniform prior on the parameters α (see above).

$$p(s) = Dir(\alpha)$$

- A uniform prior over family probabilities can be obtained by setting:

$$\forall m \in f_k : \alpha_{prior}(m) = \frac{1}{|f_k|}$$

Stephan et al. 2009, *NeuroImage*
Penny et al. 2010, *PLoS Comput. Biol.*

Family-level inference: random effects – a special case

- When the families are of equal size, one can simply sum the posterior model probabilities within families by exploiting the agglomerative property of the Dirichlet distribution:

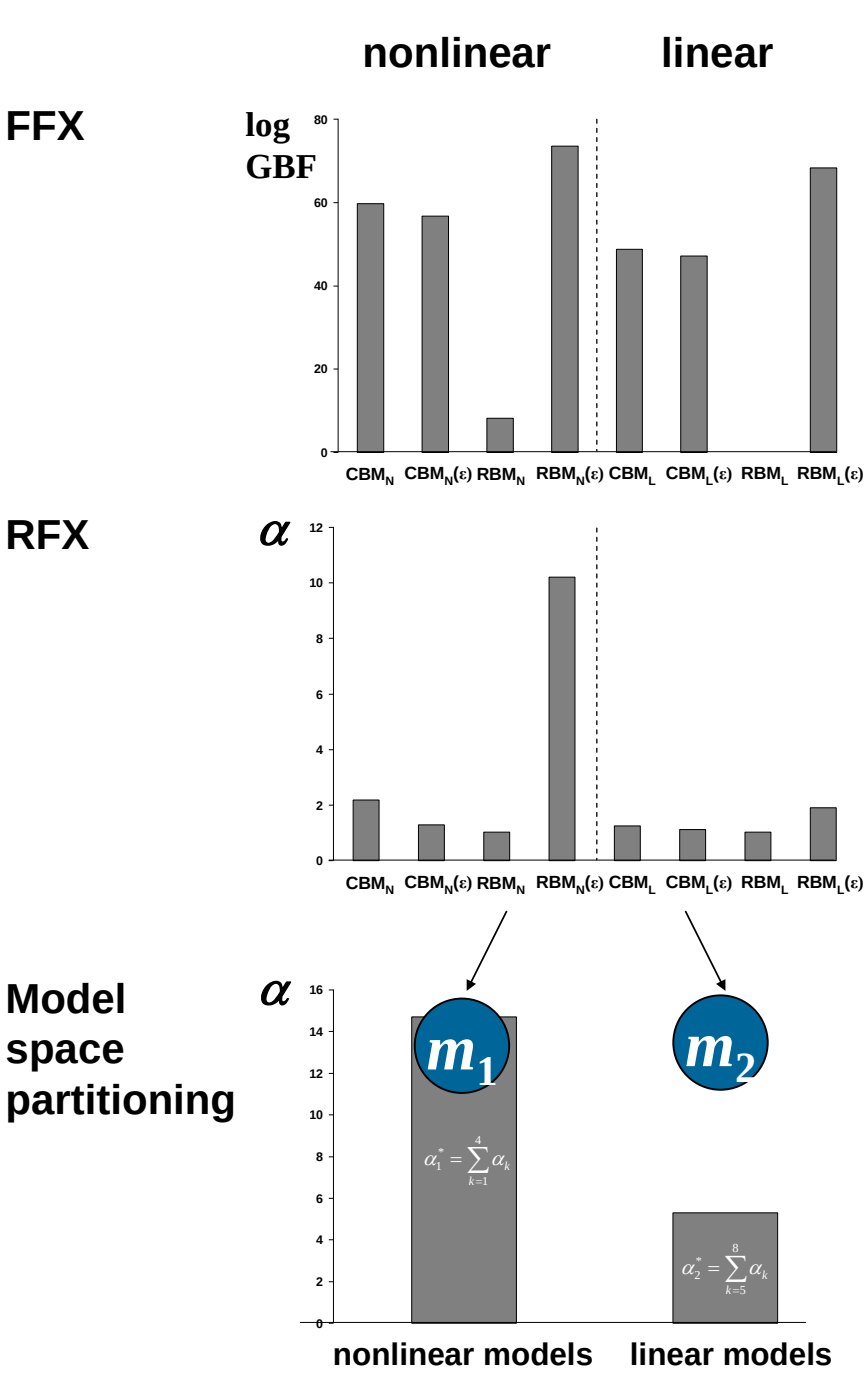
$$\left(r_1, r_2, ..., r_K\right) \sim Dir\left(\alpha_1, \alpha_2, ..., \alpha_K\right)$$

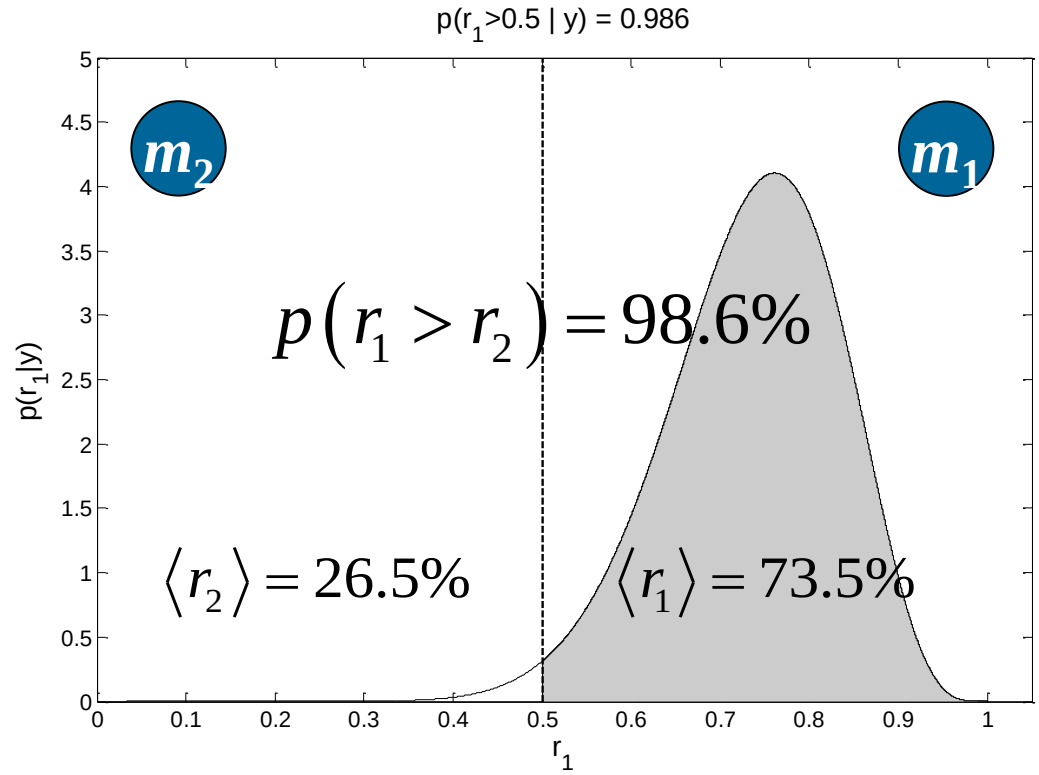
$$\Rightarrow r_1^* = \sum_{k \in N_1} r_k, r_2^* = \sum_{k \in N_2} r_k, ..., r_J^* = \sum_{k \in N_J} r_k$$

$$\sim Dir\left(\alpha_1^* = \sum_{k \in N_1} \alpha_k, \alpha_2^* = \sum_{k \in N_2} \alpha_k, ..., \alpha_J^* = \sum_{k \in N_J} \alpha_k\right)$$

Stephan et al. 2009, *NeuroImage*

Model space partitioning: comparing model families

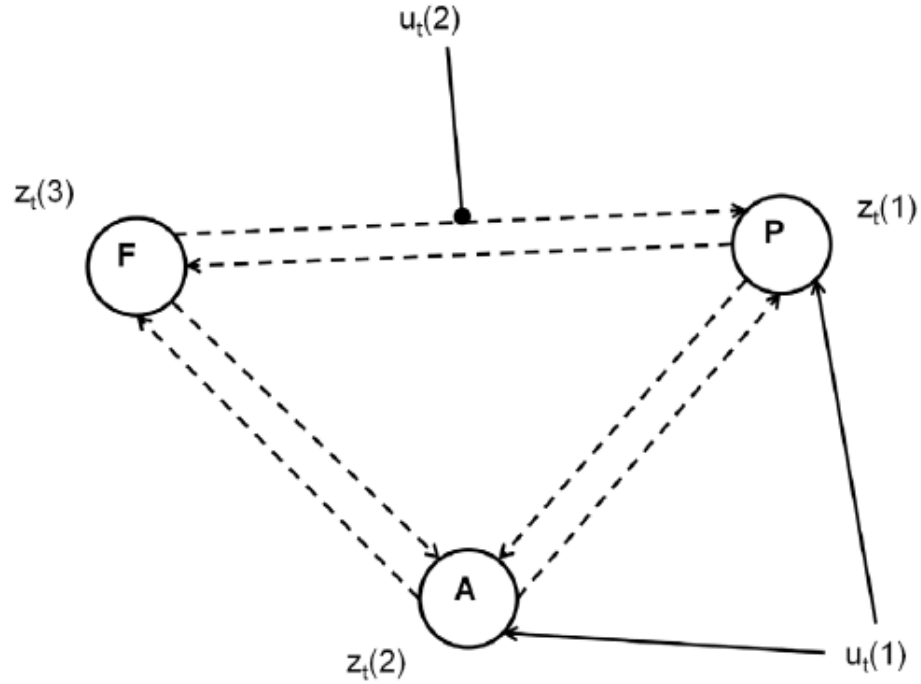


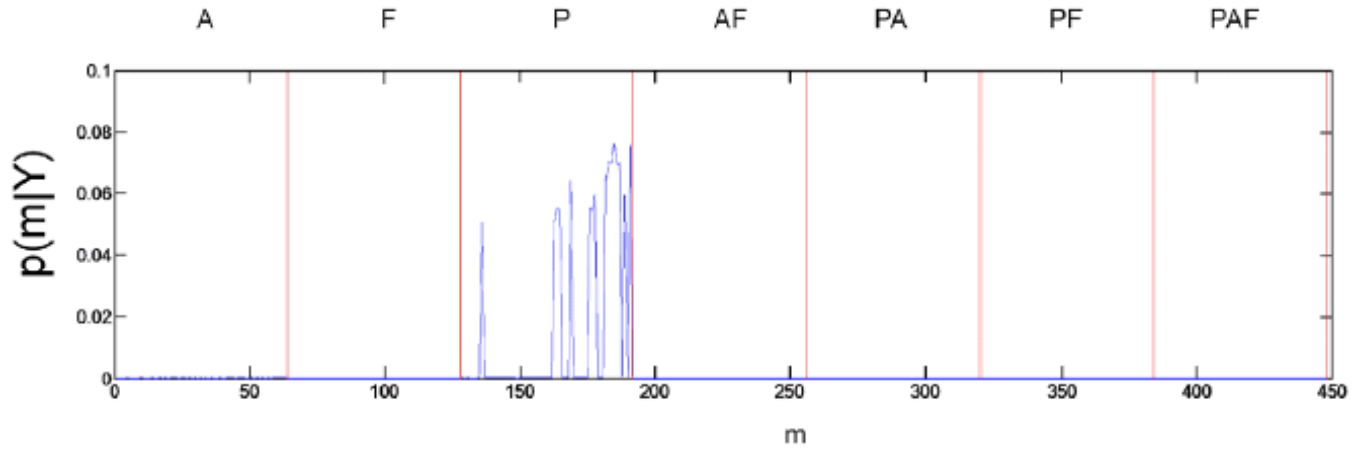


Stephan et al. 2009, *NeuroImage*

Comparing model families – a second example

- data from Leff et al. 2008, J. Neurosci
- one driving input, one modulatory input
- $2^6 = 64$ possible modulations
- $2^3 - 1$ input patterns
- $7 \times 64 = 448$ models
- integrate out uncertainty about modulatory patterns and ask where auditory input enters





Penny et al. 2010, *PLoS Comput. Biol.*

Bayesian Model Averaging (BMA)

- abandons dependence of parameter inference on a single model and takes into account model uncertainty
- uses the entire model space considered (or an optimal family of models)
- averages parameter estimates, weighted by posterior model probabilities
- represents a particularly useful alternative
  - when none of the models (or model subspaces) considered clearly outperforms all others
  - when comparing groups for which the optimal model differs

single-subject BMA:

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

group-level BMA:

$$p(\theta_n \mid y_{1..N})$$
$$= \sum_m p(\theta_n \mid y_n, m) p(m \mid y_{1..N})$$

NB: $p(m|y_{1..N})$ can be obtained by either FFX or RFX BMS

Penny et al. 2010, *PLoS Comput. Biol.*

Protected exceedance probability: Using BMA to protect against chance findings

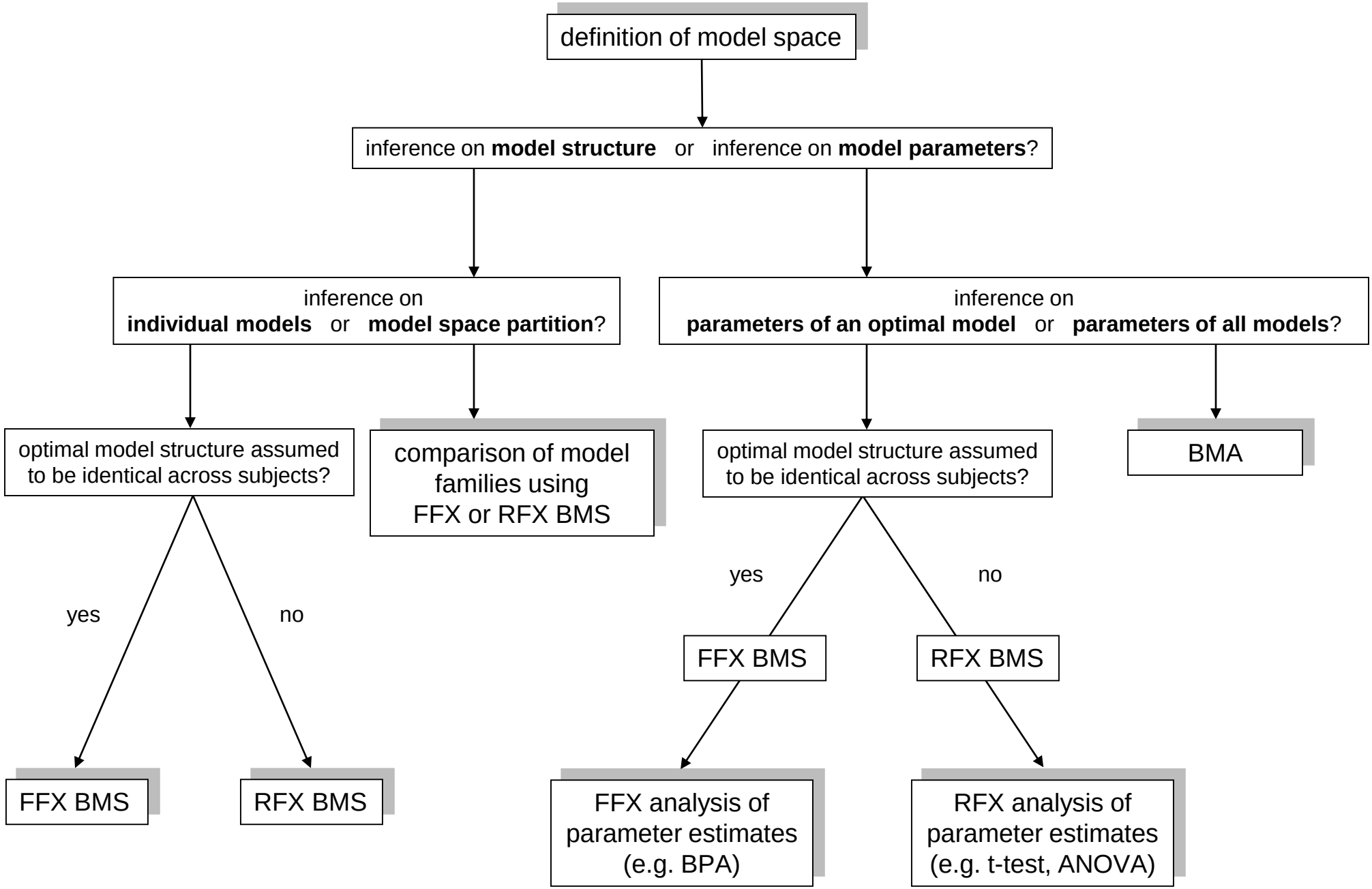
- EPs express our confidence that the posterior probabilities of models are different – under the hypothesis H_1 that models differ in probability: $r_k \neq 1/K$
- does not account for possibility "null hypothesis" H_0: $r_k = 1/K$
- **Bayesian omnibus risk (BOR)** of wrongly accepting H_1 over H_0:

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

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

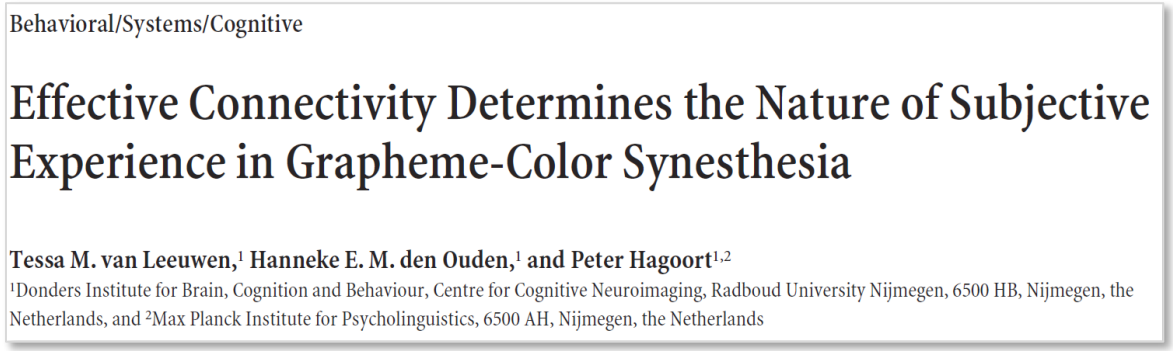
$$\begin{aligned}\tilde{\varphi}_k &= P(r_k \geq r_{k' \neq k}|y) \\ &= P(r_k \geq r_{k' \neq k}|y, H_1)P(H_1|y) + P(r_k \geq r_{k' \neq k}|y, H_0)P(H_0|y) \\ &= \varphi_k(1 - P_0) + \frac{1}{K}P_0\end{aligned}$$

Rigoux et al. 2014, *NeuroImage*



Stephan et al. 2010, *NeuroImage*

Some examples of empirical BMS/BMA applications

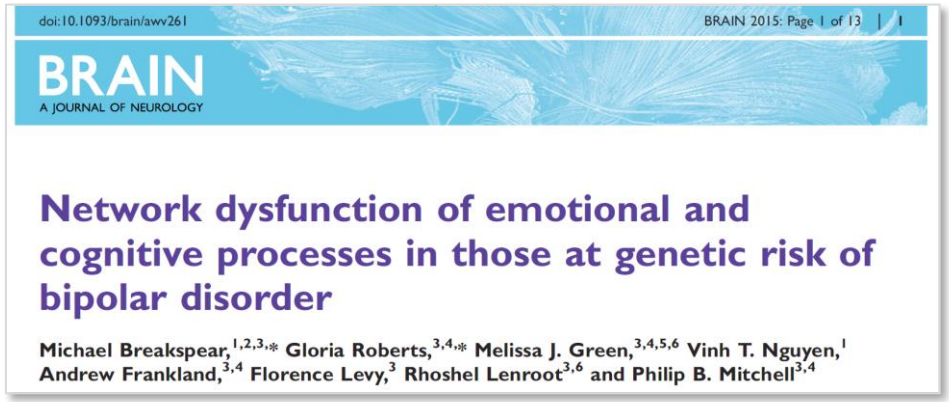


Behavioral/Systems/Cognitive

Effective Connectivity Determines the Nature of Subjective Experience in Grapheme-Color Synesthesia

Tessa M. van Leeuwen,[1] Hanneke E. M. den Ouden,[1] and Peter Hagoort[1,2]
[1]Donders Institute for Brain, Cognition and Behaviour, Centre for Cognitive Neuroimaging, Radboud University Nijmegen, 6500 HB, Nijmegen, the Netherlands, and [2]Max Planck Institute for Psycholinguistics, 6500 AH, Nijmegen, the Netherlands

van Leeuwen et al. 2011,
J. Neurosci.



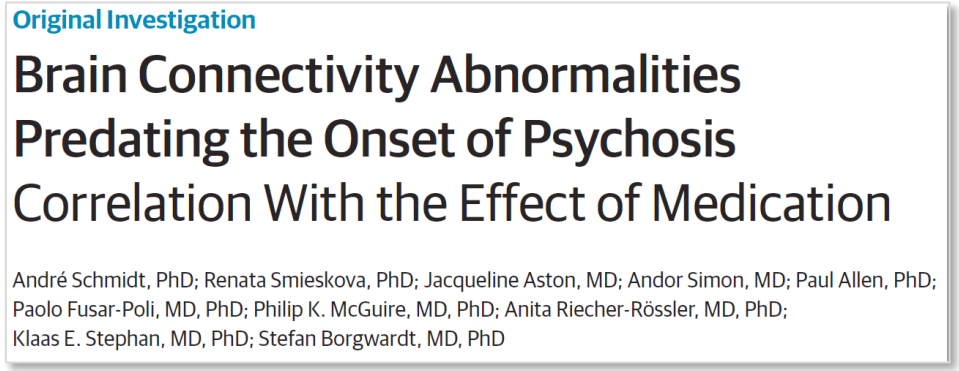
doi:10.1093/brain/aww261 BRAIN 2015: Page 1 of 13 | 1

BRAIN
A JOURNAL OF NEUROLOGY

Network dysfunction of emotional and cognitive processes in those at genetic risk of bipolar disorder

Michael Breakspear,[1,2,3,*] Gloria Roberts,[3,4,*] Melissa J. Green,[3,4,5,6] Vinh T. Nguyen,[1] Andrew Frankland,[3,4] Florence Levy,[3] Rhoshel Lenroot[3,6] and Philip B. Mitchell[3,4]

Breakspear et al. 2015,
Brain



Original Investigation

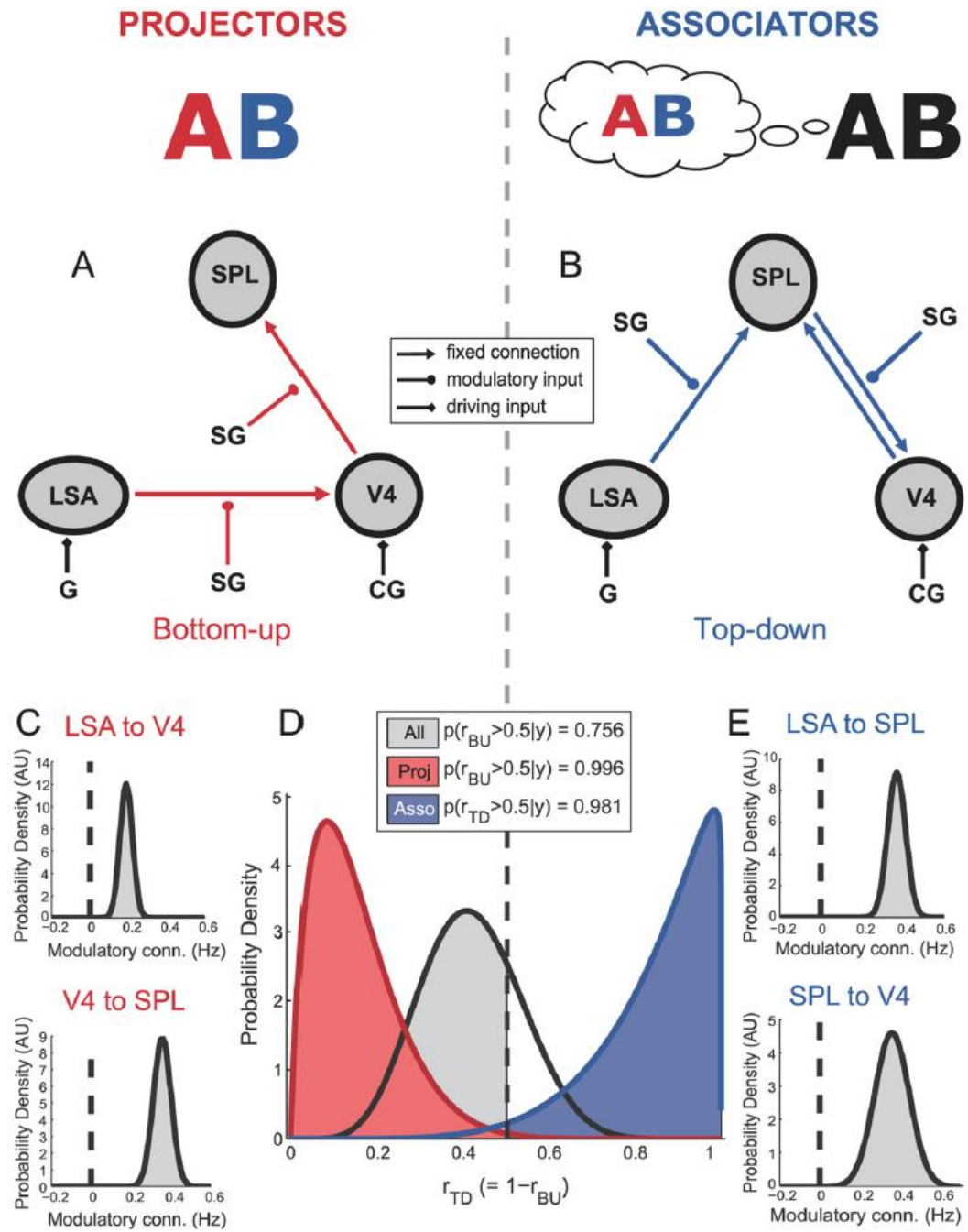
Brain Connectivity Abnormalities Predating the Onset of Psychosis
Correlation With the Effect of Medication

André Schmidt, PhD; Renata Smieskova, PhD; Jacqueline Aston, MD; Andor Simon, MD; Paul Allen, PhD; Paolo Fusar-Poli, MD, PhD; Philip K. McGuire, MD, PhD; Anita Riecher-Rössler, MD, PhD; Klaas E. Stephan, MD, PhD; Stefan Borgwardt, MD, PhD

Schmidt et al. 2013,
JAMA Psychiatry

Application: Synaesthesia

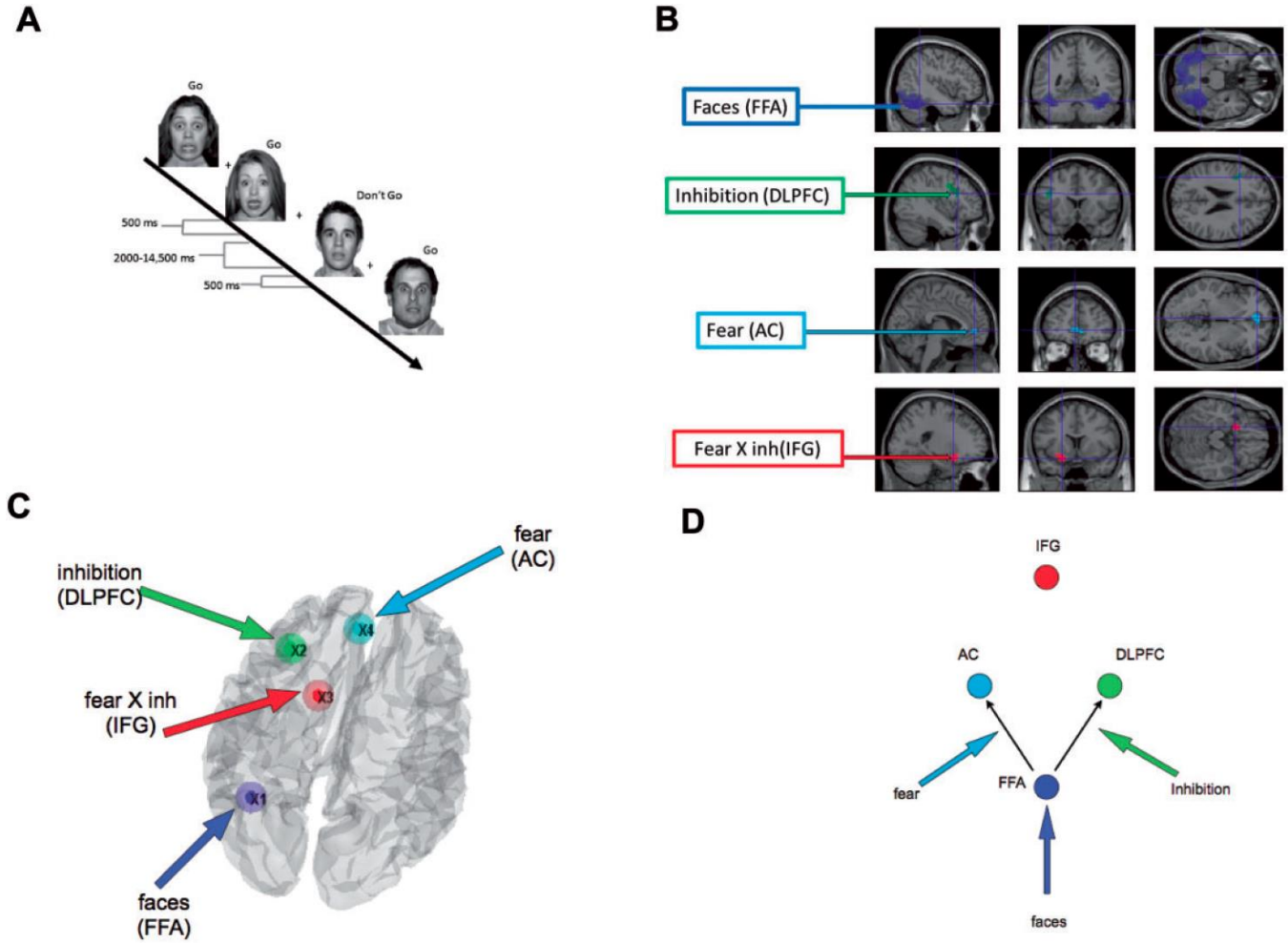
- “projectors” experience color externally colocalized with a presented grapheme
- “associators” report an internally evoked association
- across all subjects: no evidence for either model
- but BMS results map precisely onto projectors (bottom-up mechanisms) and associators (top-down)



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

Go/No-Go task to emotional faces (bipolar patients, at-risk individuals, controls)

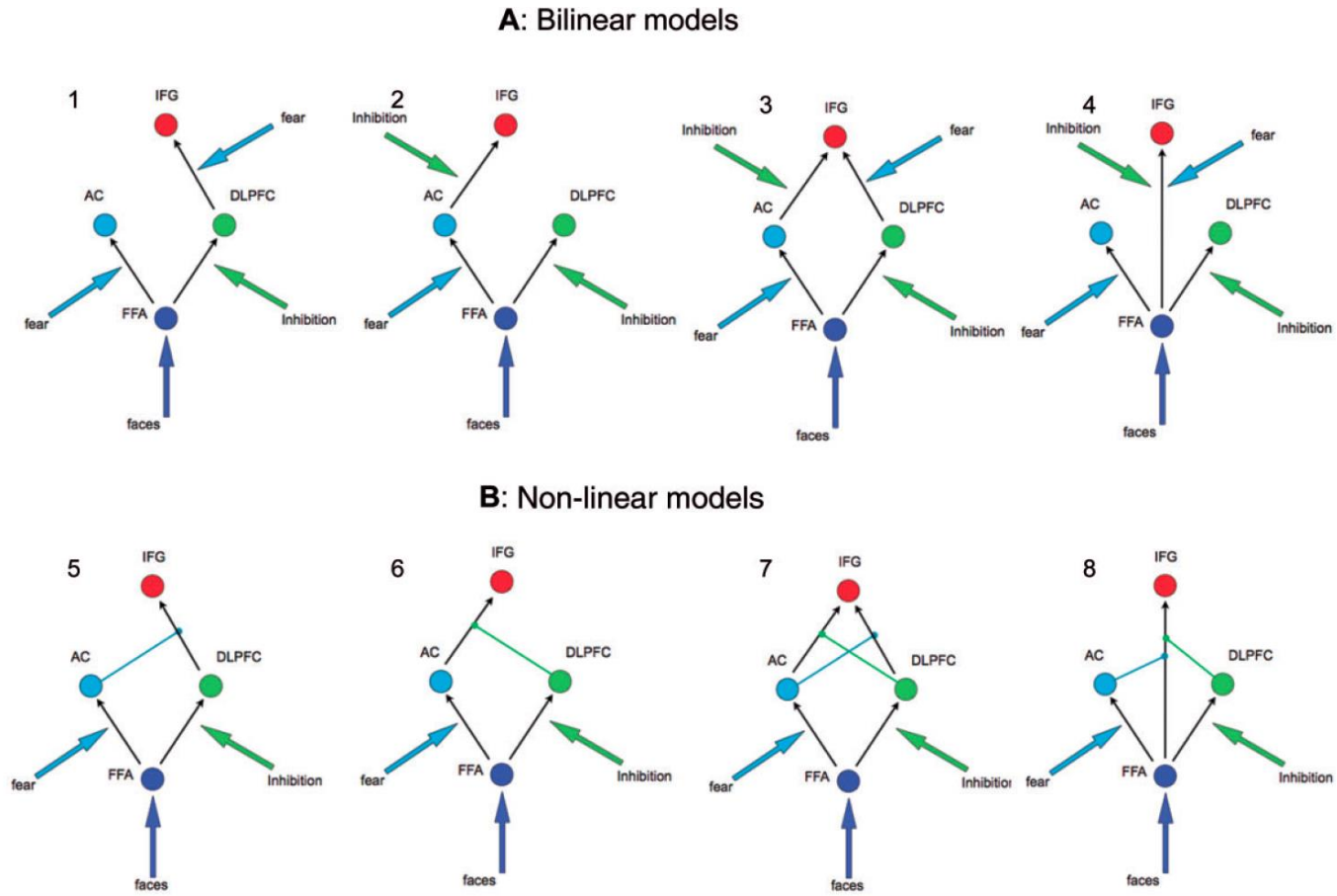
- interaction of motor inhibition and fear perception
- hypoactivation of left IFG in the at-risk group during fearful distractor trials
- What is the most likely circuit mechanism explaining the fear x inhibition interaction in IFG?



Breakspear et al. 2015, *Brain*

Model space

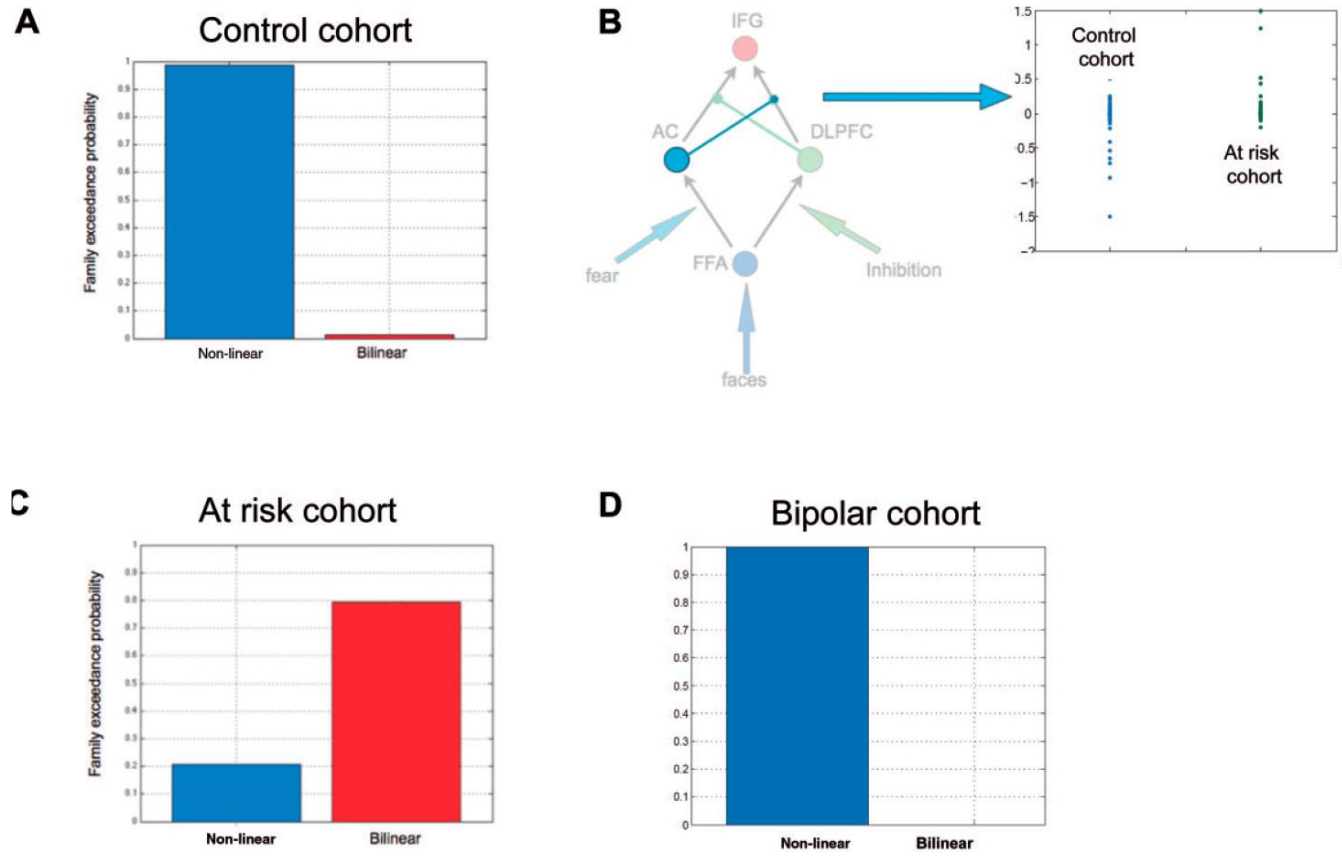
- models of serial (1-3), parallel (4) and hierarchical (5-8) processes



Breakspear et al. 2015, *Brain*

Family-level BMS

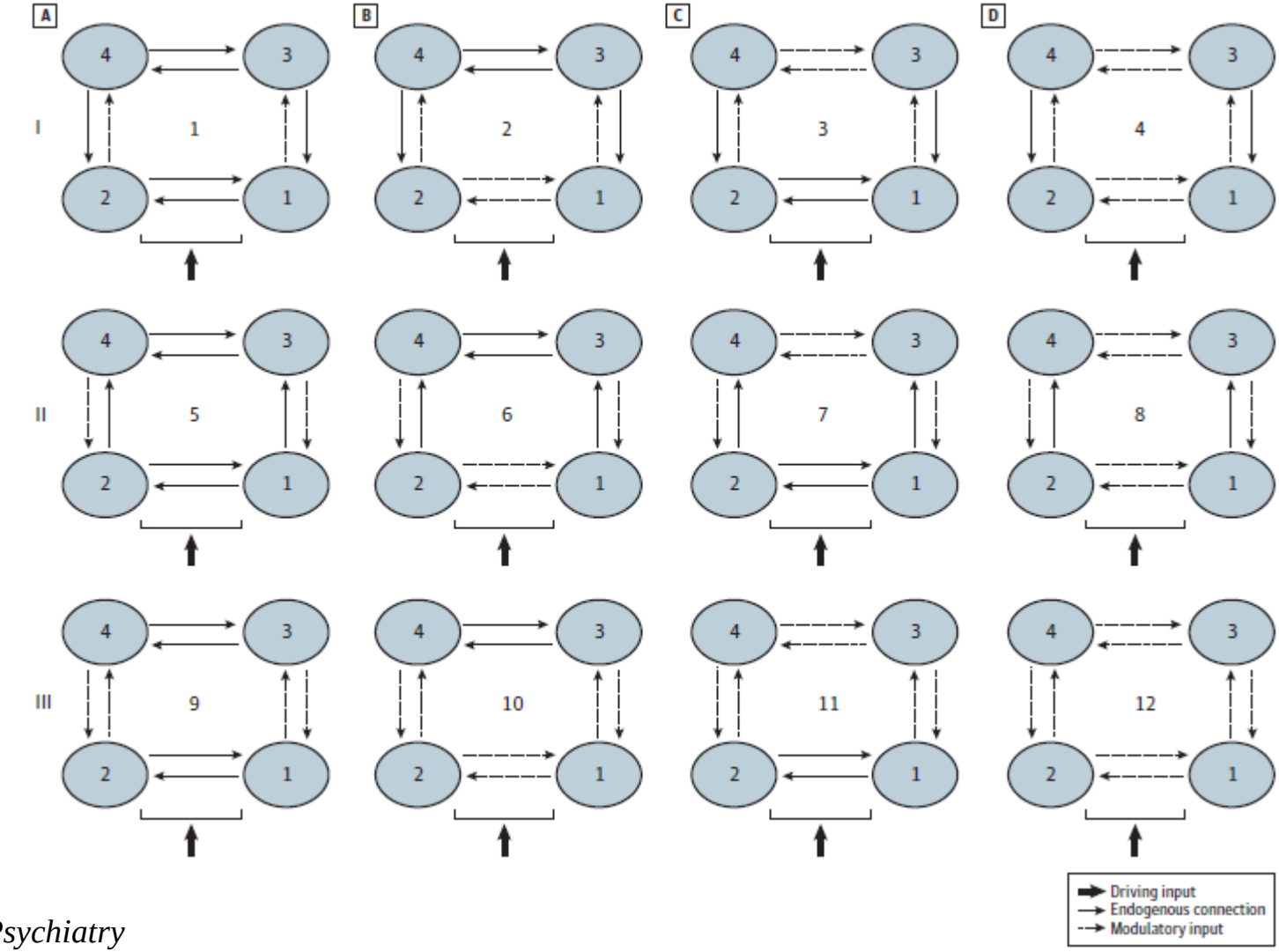
- family-level comparison: nonlinear models more likely than bilinear ones in both healthy controls and bipolar patients
- at-risk group: bilinear models more likely
- significant group difference in ACC modulation of DLPFC→IFG interaction



Breakspear et al. 2015, *Brain*

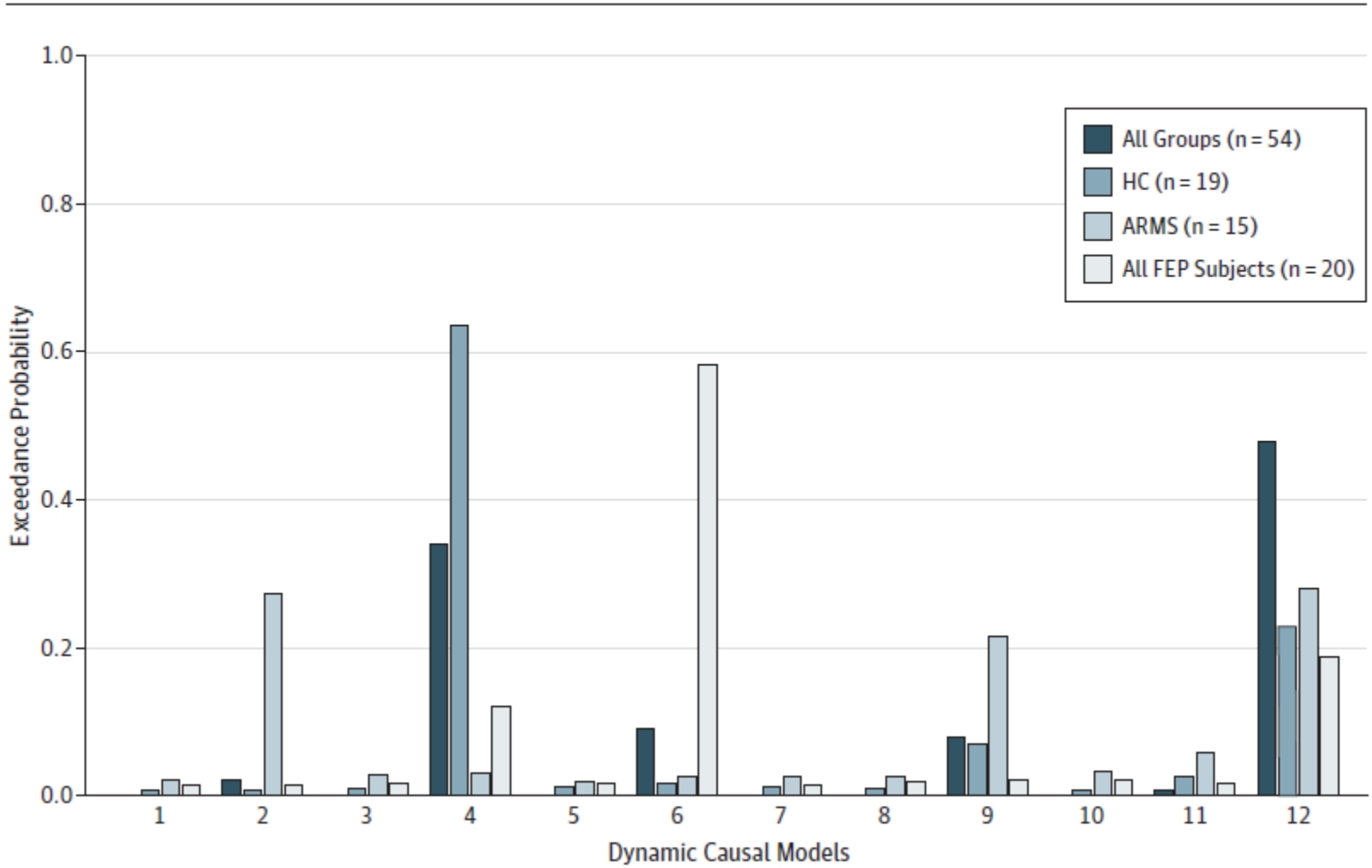
Prefrontal-parietal connectivity during working memory in schizophrenia

- 17 at-risk mental state (ARMS) individuals
- 21 first-episode patients (13 non-treated)
- 20 controls



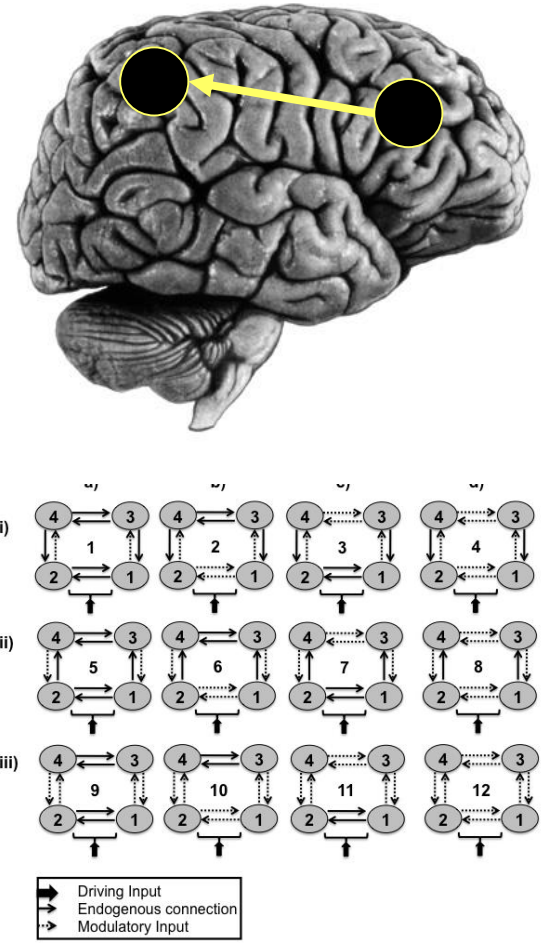
Schmidt et al. 2013, *JAMA Psychiatry*

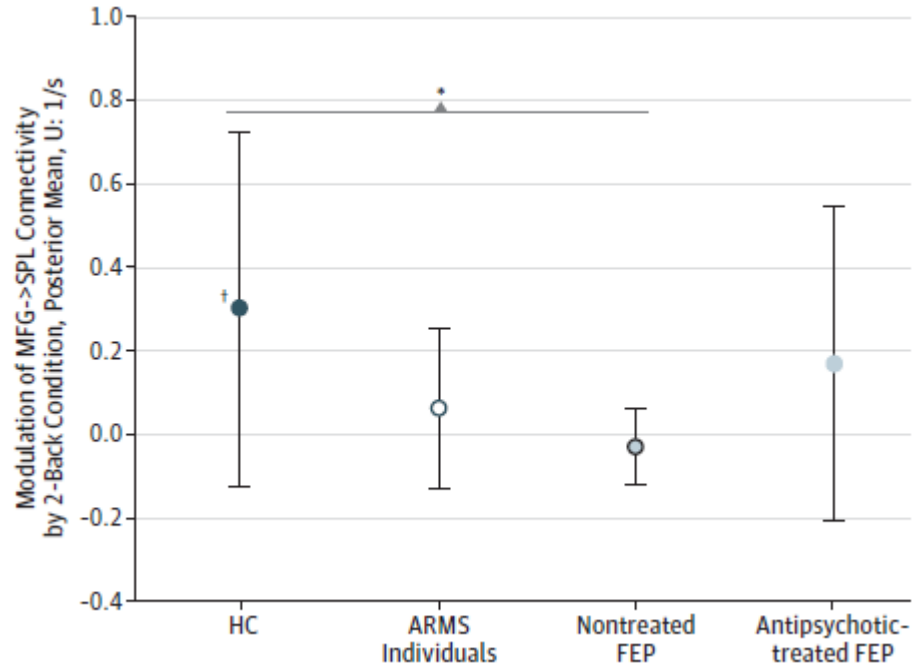
BMS results for all groups



Schmidt et al. 2013, *JAMA Psychiatry*

BMA results: PFC → PPC connectivity





17 ARMS, 21 first-episode (13 non-treated), 20 controls

Schmidt et al. 2013, *JAMA Psychiatry*

Further reading on BMS

- Penny WD, Stephan KE, Mechelli A, Friston KJ (2004a) 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.

Thank you