Bayesian model selection (BMS)

Klaas Enno Stephan

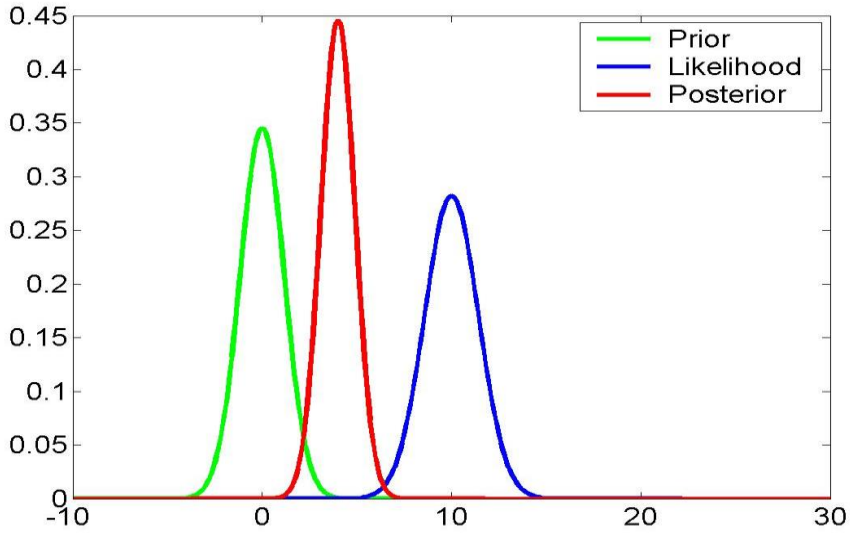
Translational Neuromodeling Unit

Universität Zürich[UZH]

ETH

Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Reminder: Bayes' rule



The Reverend Thomas Bayes
(1702-1761)

Likelihood
(data)

Prior
(prediction)

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

Posterior
(inference)

Evidence
(normalisation term)

θ: parameters
y: data

"... the theorem expresses how a degree of belief, expressed as a probability, should rationally change to account for the availability of related evidence."

Wikipedia

Model comparison and selection

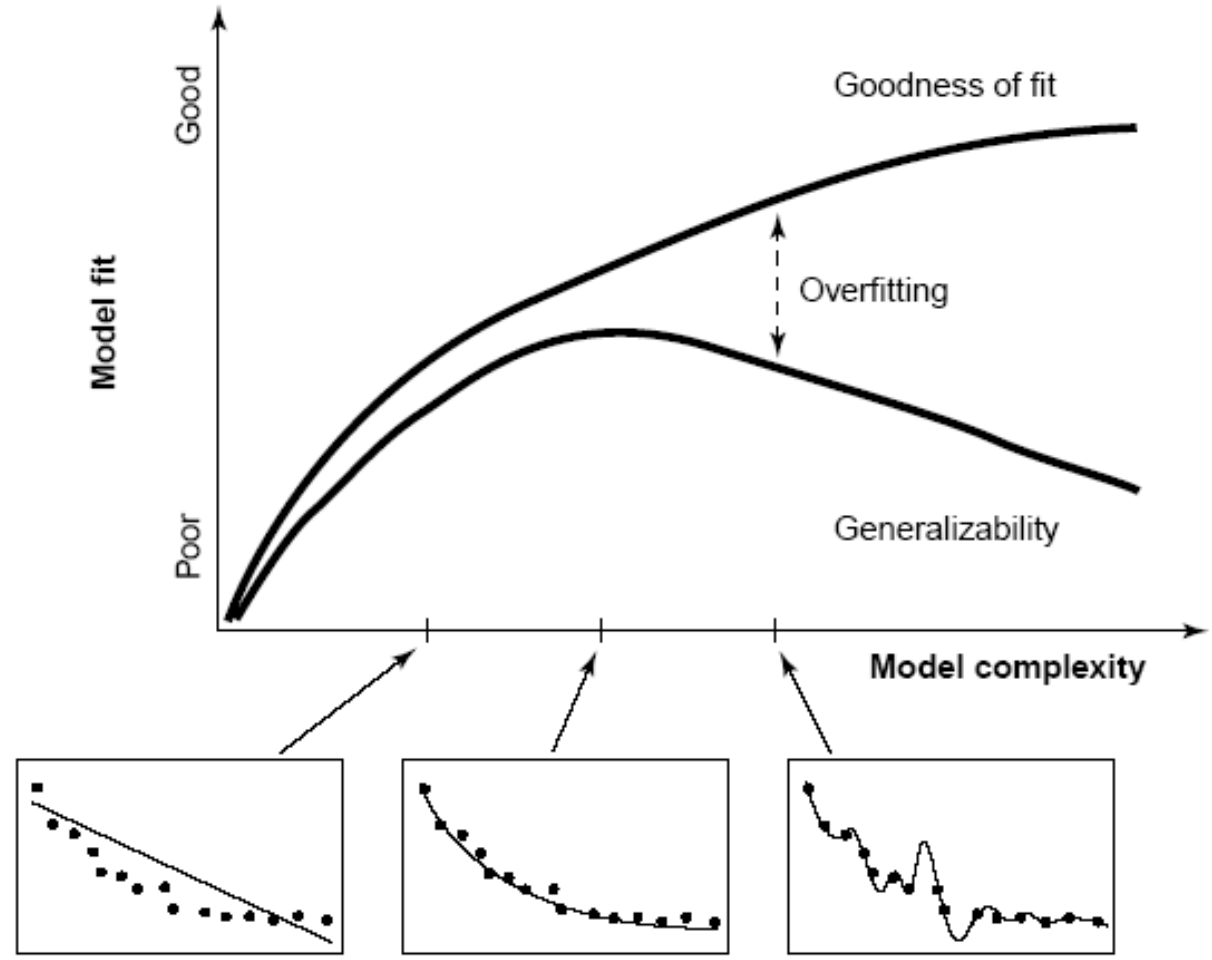
Given competing hypotheses on structure & functional mechanisms of a system, which model is the best?

⇩

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)

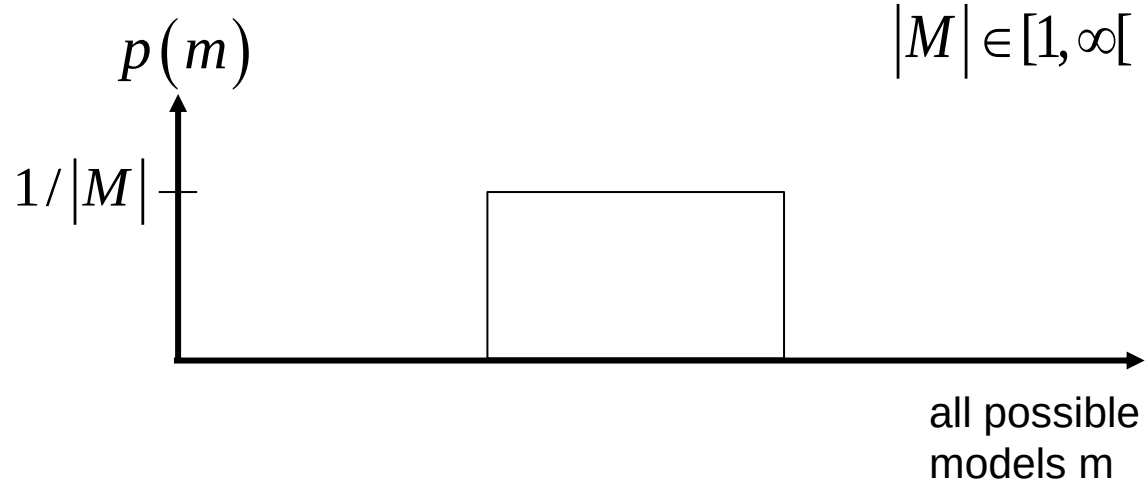
- A priori definition of hypothesis set (model space) is crucial
  - BMS selects the model within the pre-specified model space that has the highest probability of having generated the data

- Model selection is not equal to model validation!
  - compares the relative goodness of competing hypotheses within the pre-specified model space but does not "validate" any of the models

- Model validation requires external criteria (external to the measured data)
  - e.g. predictive validity: can model predict new (unseen) data?

BMS: Defining the model space

- Model space M: all models m that are deemed plausible *a priori.*
  This is the researcher's hypothesis space.
- Defining M is equivalent to defining a prior over models, $p(m)$, such that a finite set of models has non-zero prior probability.
- Usually, all models in M are assigned equal prior probability:



BMS: Applying Bayes' rule to models

- Goal: select the model that has the highest posterior probability given the data:

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

- Under uniform priors over models, all prior model probabilities are identical:

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

identical for all models i

- We can thus use the model evidence to compare and select models.

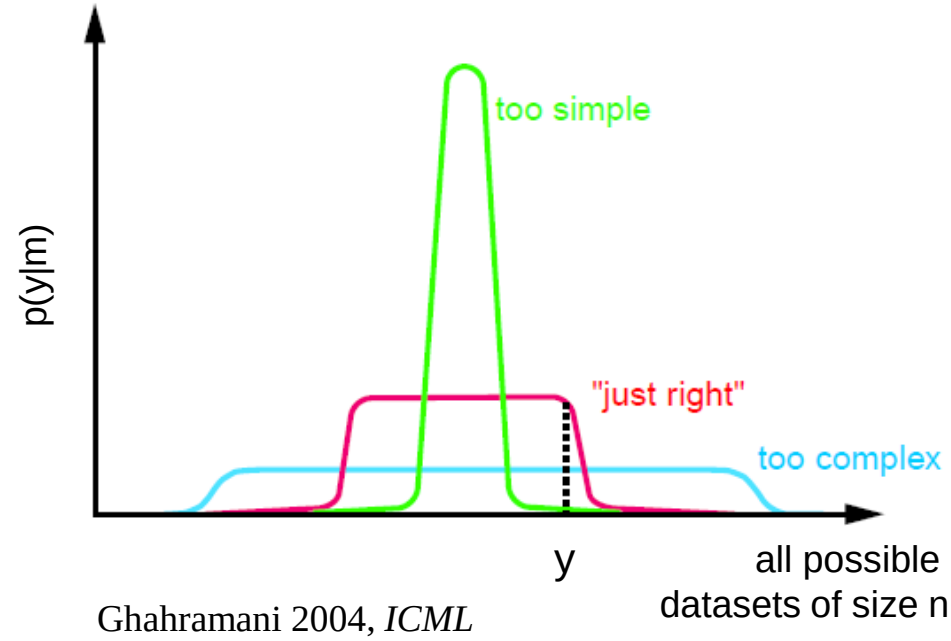
Model evidence: How can it be interpreted?

Model evidence = marginal likelihood:

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

⟹ expected probability that randomly selected parameters from the prior would generate the observed data

⟹ measure of model goodness: log evidence accounts for both accuracy and complexity of the model



Ghahramani 2004, *ICML*

Overly simple models can only generate few datasets. Overly complex models can generate many possible datasets, each with low probability.

Both are thus unlikely to have generated a given dataset.

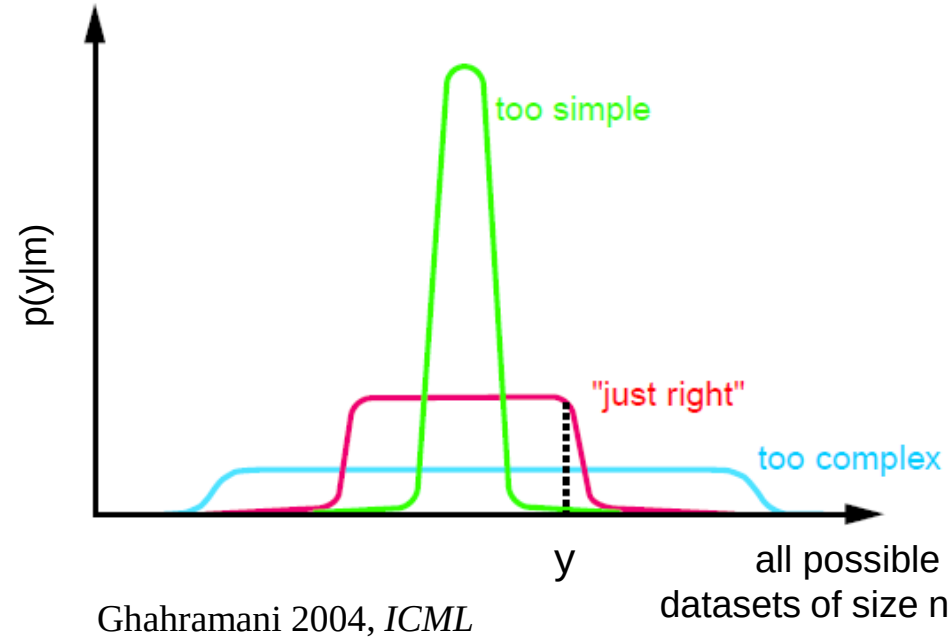
Model evidence: How can it be interpreted?

Model evidence = marginal likelihood:

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

⇒ "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?"

⇒ measure of model goodness: log evidence accounts for both accuracy and complexity of the model



Ghahramani 2004, *ICML*

Overly simple models can only generate few datasets. Overly complex models can generate many possible datasets, each with low probability.

Both are thus unlikely to have generated a given dataset.

Model evidence: How can it be computed/approximated?

- **Analytically:**
  - only possible in rare cases (e.g. linear Gaussian models)
- **By approximation:**
  - Akaike Information Criterion (AIC)
  - Bayesian Information Criterion (BIC)
  - Variational Bayes (VB)
    - negative free energy as lower bound approximation to the log evidence
  - Sampling-based methods
    - prior arithmetic mean estimator
    - posterior harmonic mean estimator
    - thermodynamic integration (TI)

AIC/BIC approximations to the model evidence

- For numerical reasons, we typically approximate the log evidence (the logarithm is a strictly monotonic function).
- The log evidence (and its approximations) can be decomposed into accuracy and complexity terms:

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

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

Two expressions of AIC/BIC in the literature

- definition of AIC/BIC shown on the previous slide:
  AIC/BIC serve as (asymptotic) approximations of the log model evidence
  - for details, see Raftery (1995) Sociological Methodology 25: 111-163, or Penny et al. (2004) NeuroImage 22: 1157-1172
- often the negative of these expressions is used
  - essentially an approximation to surprise
  - not in line with original BIC paper by Schwarz (1978) Annals of Statistics, 6: 461-464
- in principle, either version is fine for model selection
  - but when using AIC/BIC for random effects BMS (see below), AIC/BIC expressions that approximate log evidence must be used

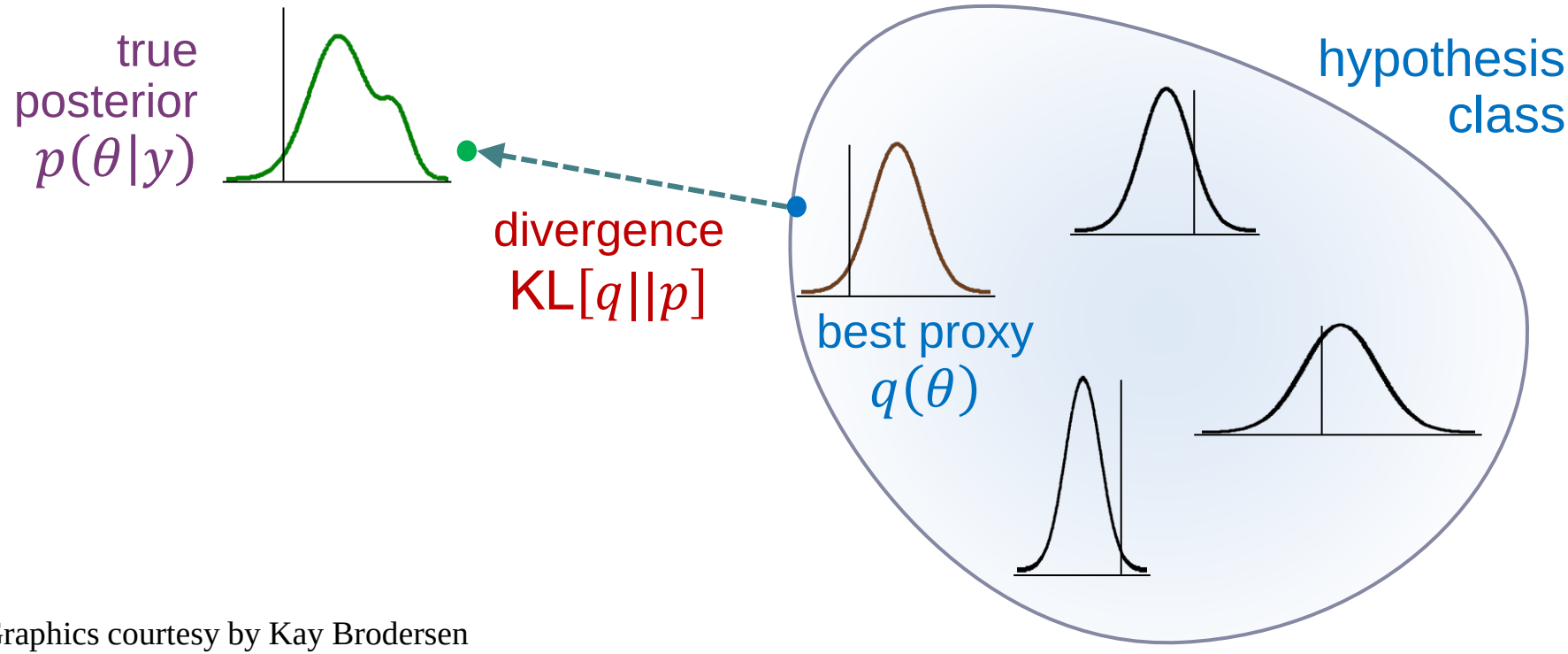
Pros and cons of AIC/BIC

- **easy and fast to compute**
  - only the MLE of the parameters and counting (parameters and data points) is needed.
- but: a **limited concept of model complexity**
  - all parameters contribute equally to complexity, regardless of how they can affect the data
  - e.g. do not take into account dependencies amongst parameters
- a richer concept of model complexity is provided by the **negative free energy** (aka **"evidence lower bound", ELBO**)

Recap: Variational Bayes (VB)

Basic idea of VB: 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 optimising its sufficient statistics.



Graphics courtesy by Kay Brodersen

Recap: Variational Bayes

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

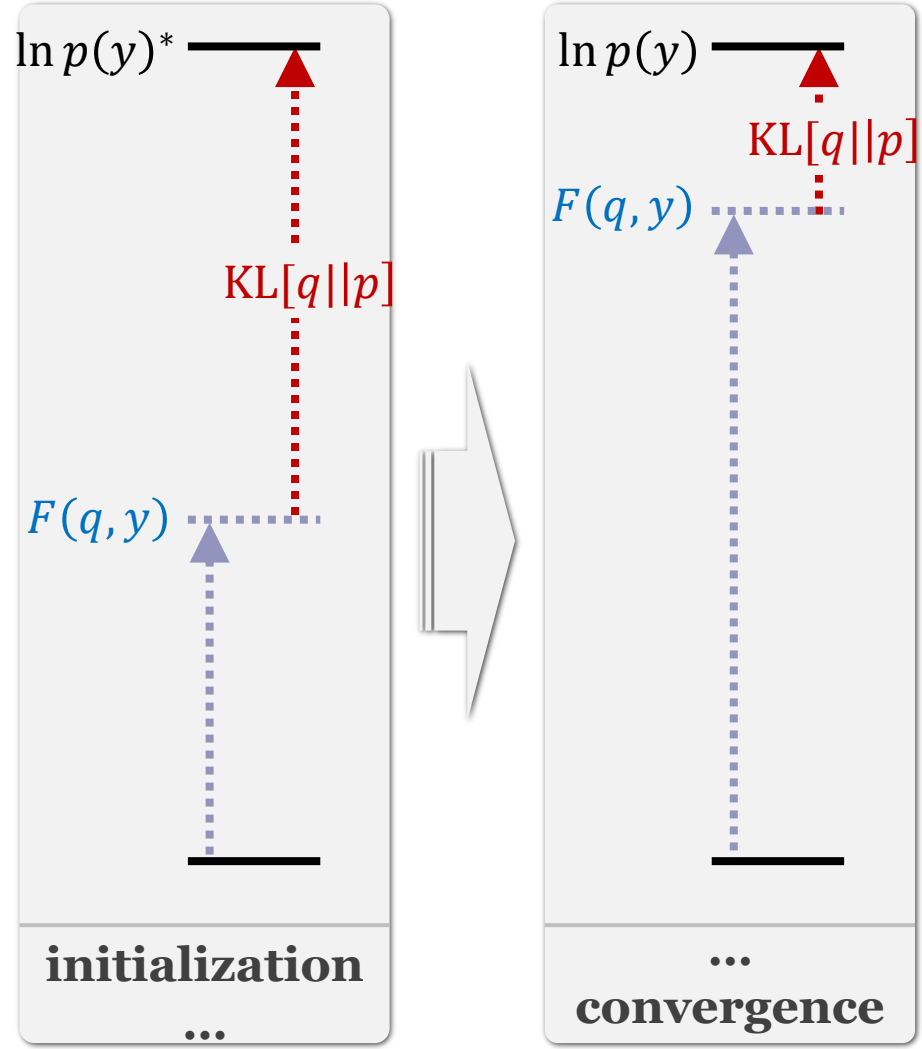
divergence ≥ 0 (unknown)

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

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

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

- minimizing $\mathrm{KL}[q||p]$,
- making $q(\theta)$ the best possible estimate of the posterior,
- tightening $F(q,y)$ as a lower bound to the log model evidence.

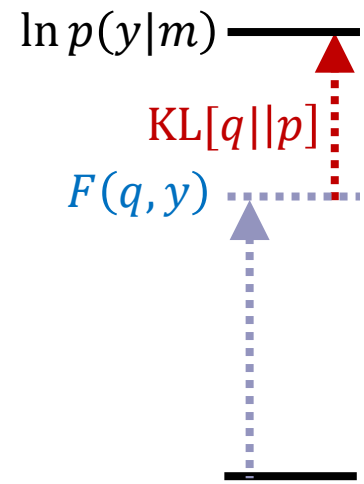


Graphics courtesy by Kay Brodersen

The negative free energy approximation F

F is a lower bound on the log model evidence:

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



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 derivations can be found in many papers; e.g. see appendix to Stephan et al. 2007, *NeuroImage* (not the first!)

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:

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

positive value, [0; ∞[

Kass & Raftery classification:

B_{12}	$p(m_1\|y)$	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:

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

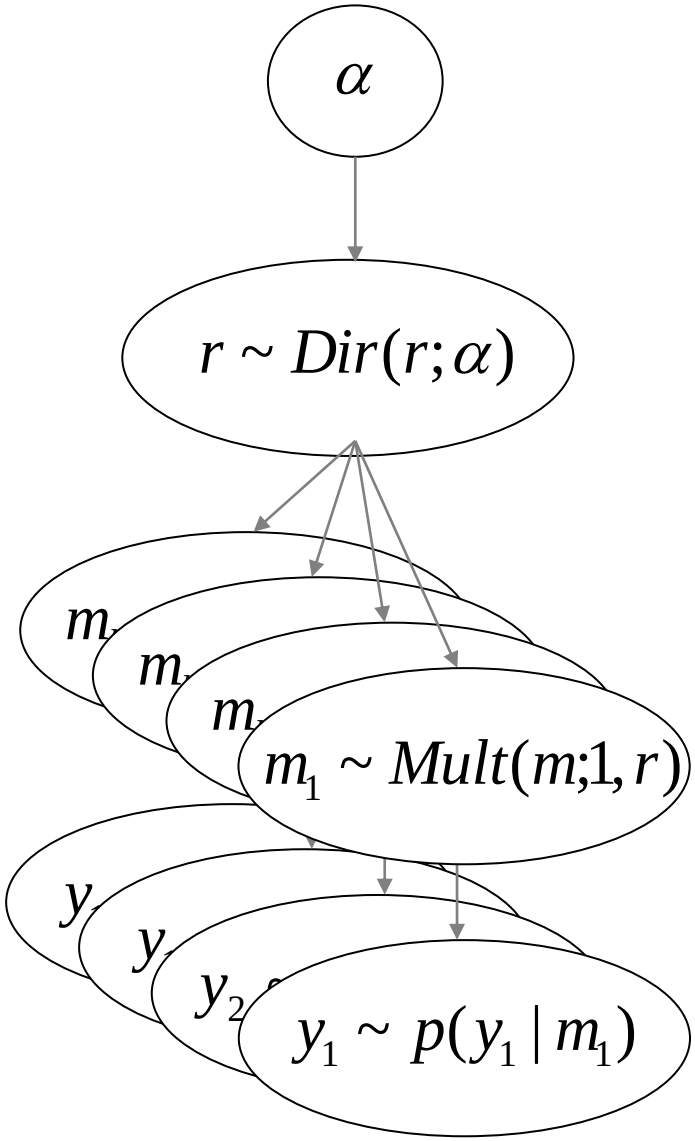
Average Bayes factor (ABF):

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

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

Model inversion by Variational Bayes or MCMC

Measured data y

Stephan et al. 2009, *NeuroImage*

Four options for reporting model ranking by random effects BMS

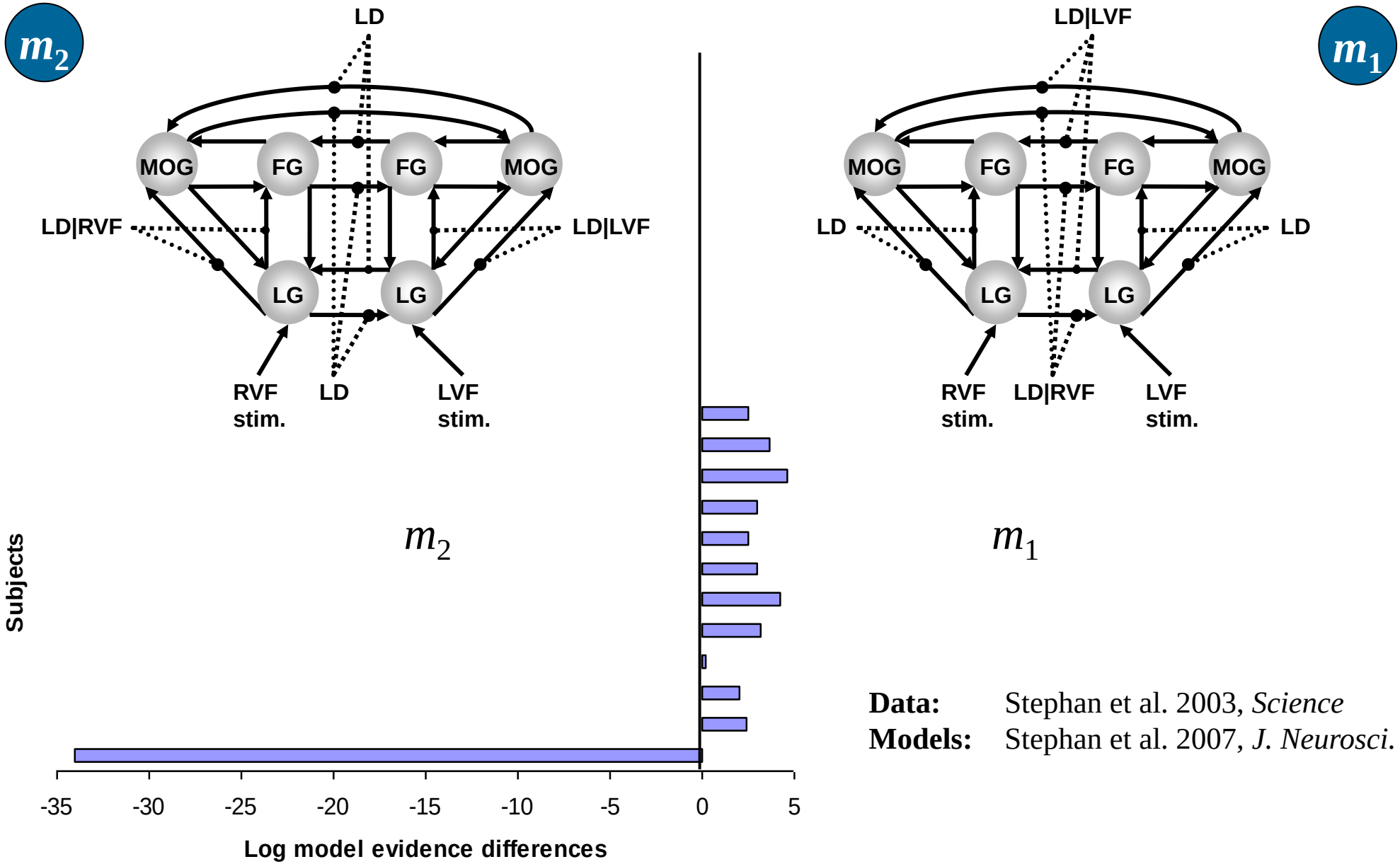
1. **Dirichlet parameter estimates** α

2. **expected posterior probability** of obtaining the k-th model for any randomly selected subject $\left\langle r_k \right\rangle_q = \alpha_k / (\alpha_1 + \ldots + \alpha_K)$

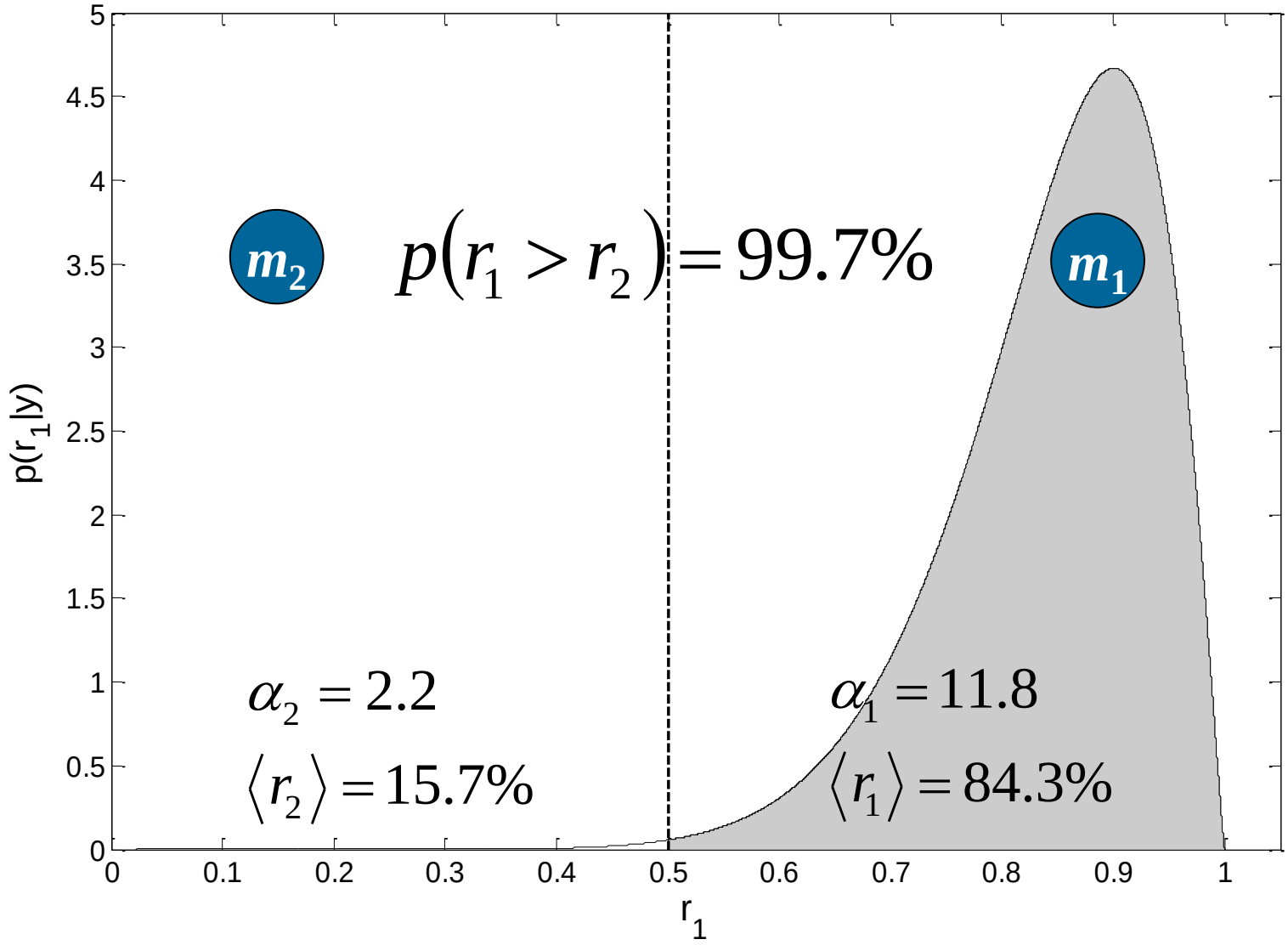
3. **exceedance probability (XP)** that a particular model *k* is more likely than any other model (of the *K* models tested), given the group data $\exists k \in \{1...K\}, \forall j \in \{1...K \mid j \neq k\}:$ $\varphi_k = p(r_k > r_j \mid y; \alpha)$

4. **protected exceedance probability (PXP)** see Rigoux et al. (2014) and below

Example: Hemispheric interactions during vision



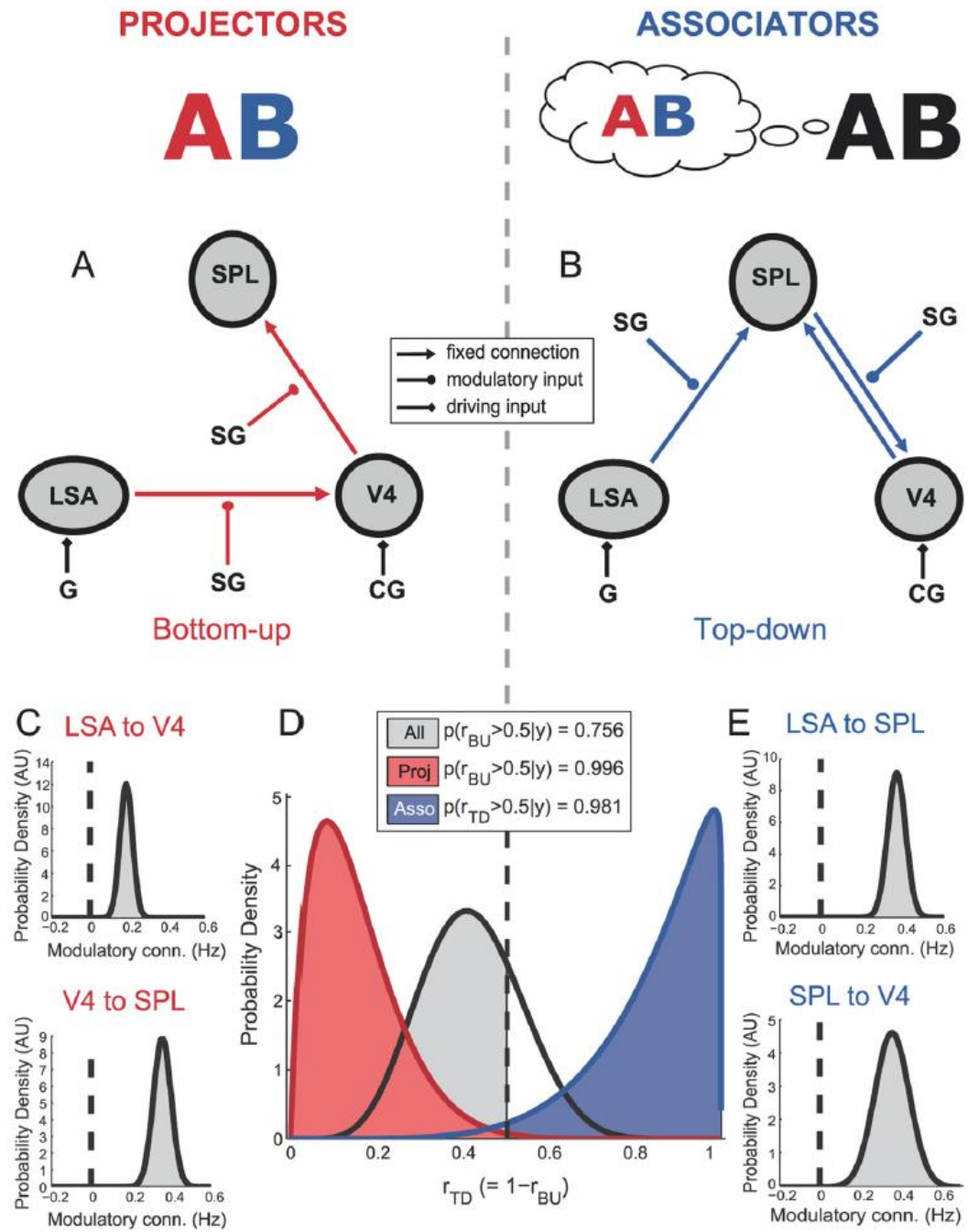
Data: Stephan et al. 2003, *Science*
Models: Stephan et al. 2007, *J. Neurosci.*



Stephan et al. 2009, *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 the possibility of the "null hypothesis" H_0: $r_k = 1/K$
- **Bayesian omnibus risk (BOR)** of model frequencies being equal (H_0:)

$$P_0 = \frac{1}{1 + \frac{p(m \mid H_1)}{p(m \mid H_0)}}$$

- **protected XP**: Bayesian model averaging (BMA, see below) over H_1 and H_0:

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

Rigoux et al. 2014, *NeuroImage*

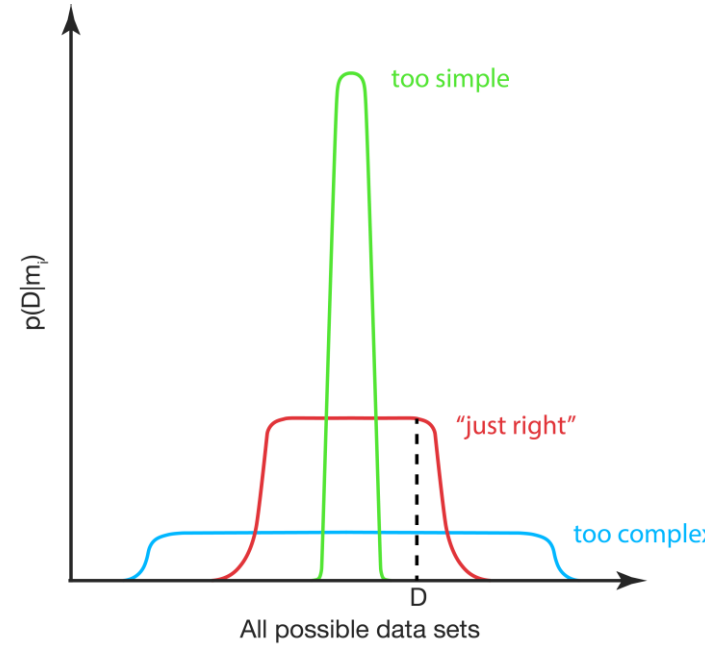
Overfitting at the level of models

Common question:

There is an infinite number of possible models for a given dataset. Should we not strive for a "brute force" search and compare as many models as possible?

Not necessarily.

The more models are included in the model space, the risk of overfitting (at the level of models) increases, too.



Animation courtesy of Stefan Frässle

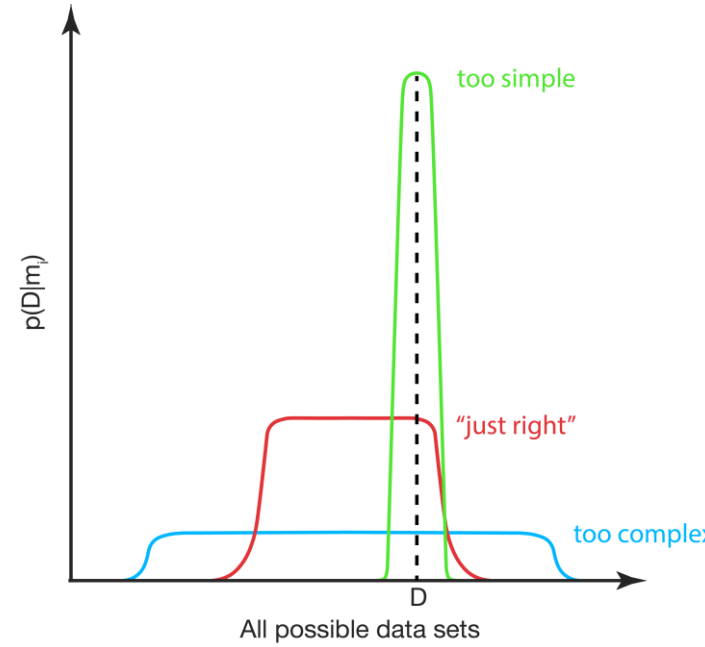
Overfitting at the level of models

Common question:

There is an infinite number of possible models for a given dataset. Should we not strive for a "brute force" search and compare as many models as possible?

Not necessarily.

The more models are included in the model space, the risk of overfitting (at the level of models) increases, too.



Animation courtesy of Stefan Frässle

Overfitting at the level of models

In brief: $\uparrow$ #models $\Rightarrow$ $\uparrow$ risk of overfitting

Three solutions in the context of BMS:

❶ Regularisation: choosing priors over models p(m) such that model space is small

❷ Family-level BMS

❸ Bayesian model averaging (BMA)

Overfitting at the level of models

In brief: ↑ #models ⇒ ↑ risk of overfitting

Three solutions in the context of BMS:

❶ Regularisation: choosing priors over models p(m) such that model space is small

❷ Family-level BMS (not discussed in today's talk for lack of time)

❸ Bayesian model averaging (BMA)

❸ 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 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:

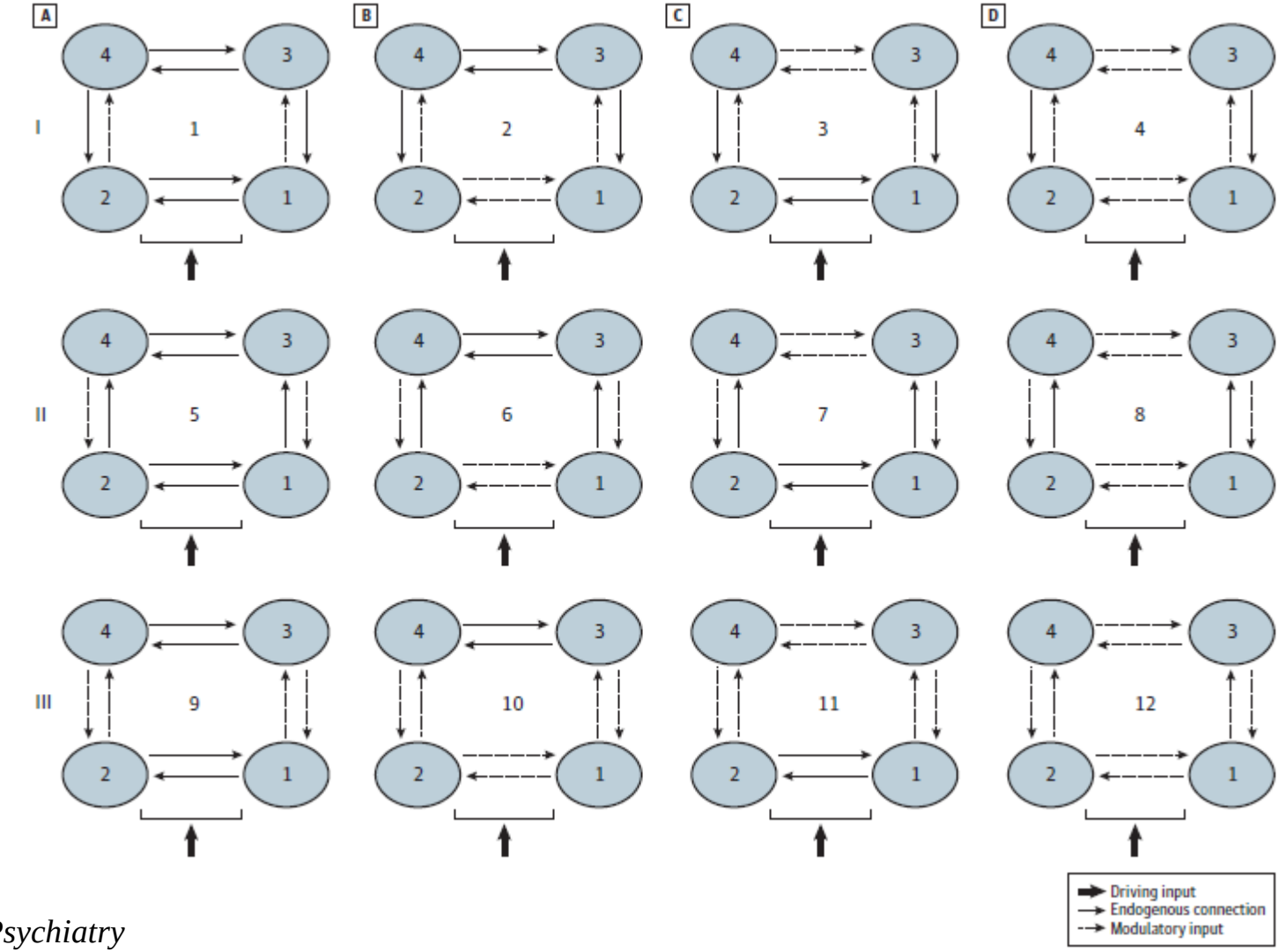
$$p(\theta_n \mid y_{1..N})$$
$$= \sum_m p(\theta_n \mid y_n, m) p(m \mid y_{1..N})$$

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

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

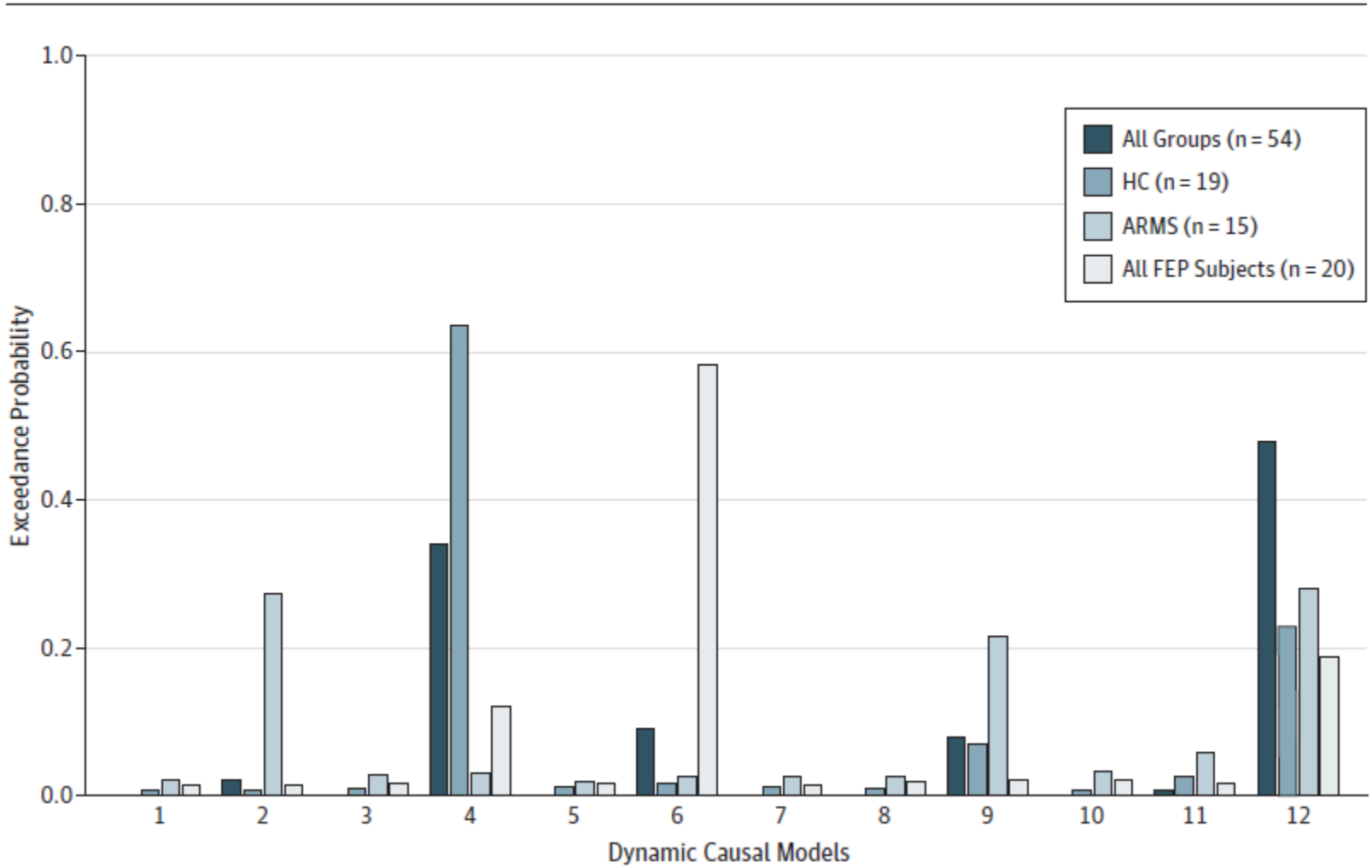
Prefrontal-parietal connectivity during working memory in schizophrenia

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



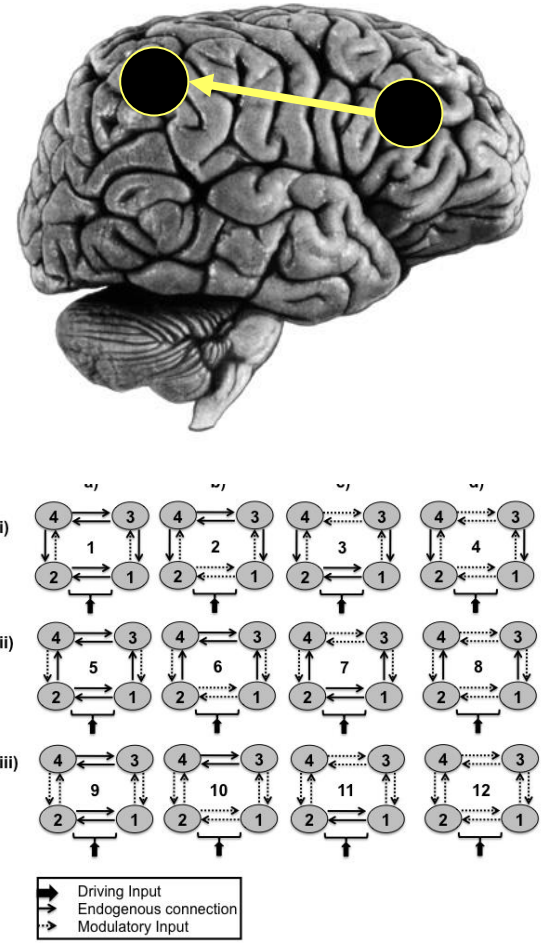
Schmidt et al. 2013, *JAMA Psychiatry*

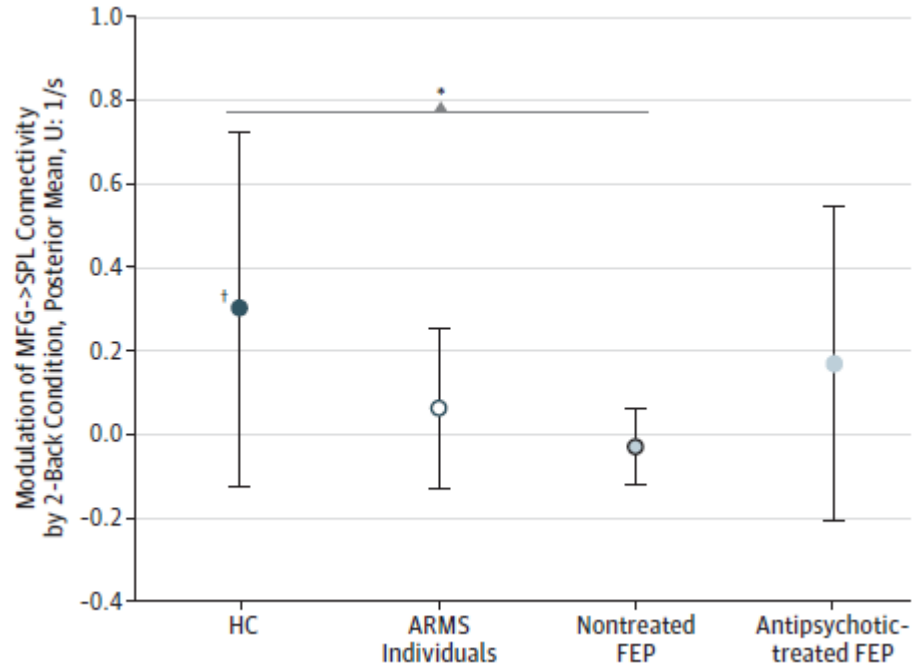
BMS results for all groups



Schmidt et al. 2013, *JAMA Psychiatry*

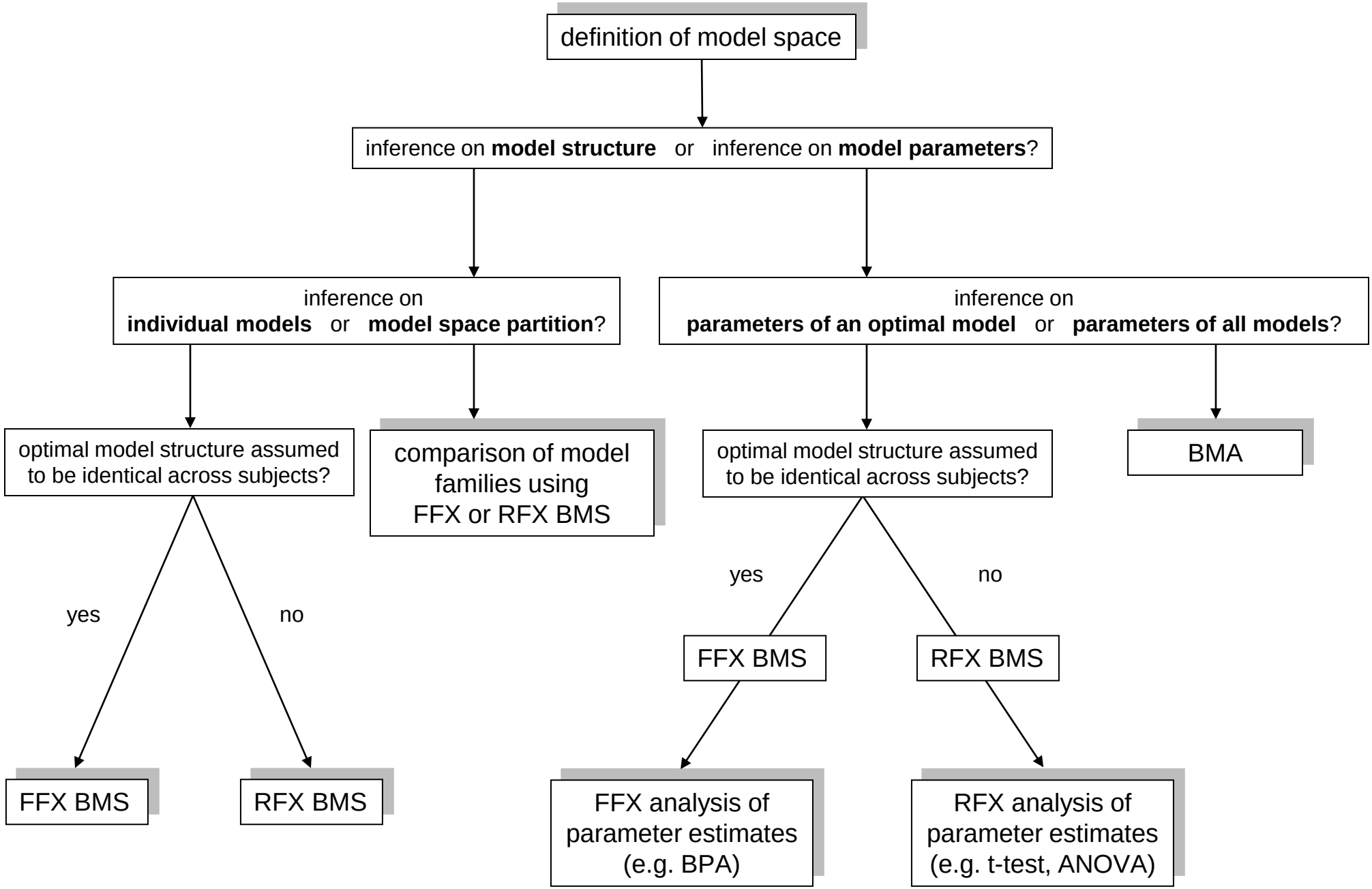
BMA results: PFC → PPC connectivity





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

Schmidt et al. 2013, *JAMA Psychiatry*



Stephan et al. 2010, *NeuroImage*

Important topics I have not covered in this talk

- **Bayesian model reduction**
  - an efficient alternative to RFX-BMS (applicable to nested models only)
  - see:
    Friston KJ et al. (2016) NeuroImage 128:413-431
    https://www.sciencedirect.com/science/article/pii/S105381191501037X

- estimating the model evidence with **sampling-based methods**, such as **thermodynamic integration**
  - for a tutorial review, see:
    Aponte, Yao et al. (2022) Cognitive Neurodynamics
    https://link.springer.com/article/10.1007/s11571-021-09696-9

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, 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