Bayesian model selection (BMS)

Klaas Enno Stephan

Translational Neuromodeling Unit
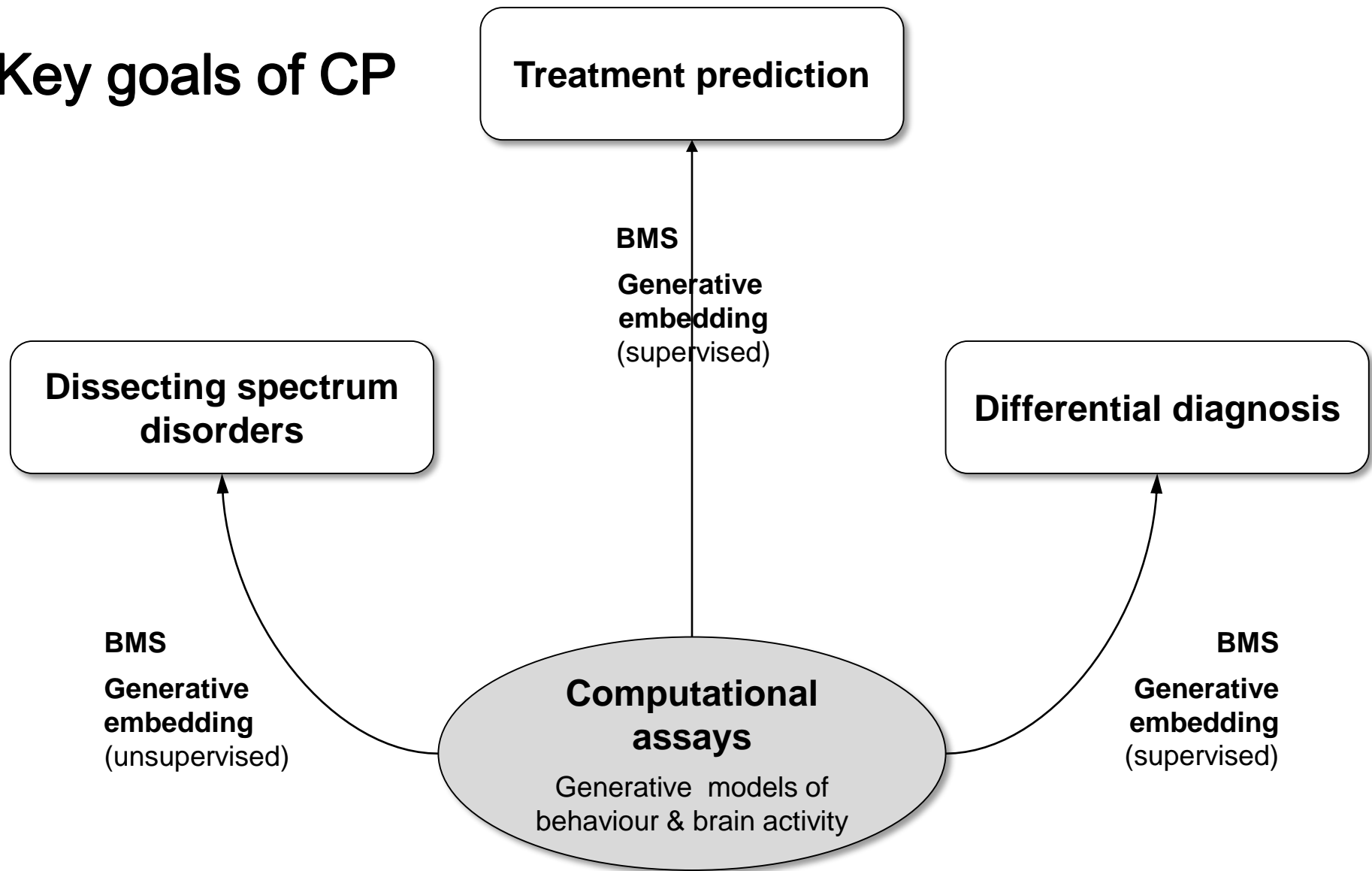
Universität Zürich UZH
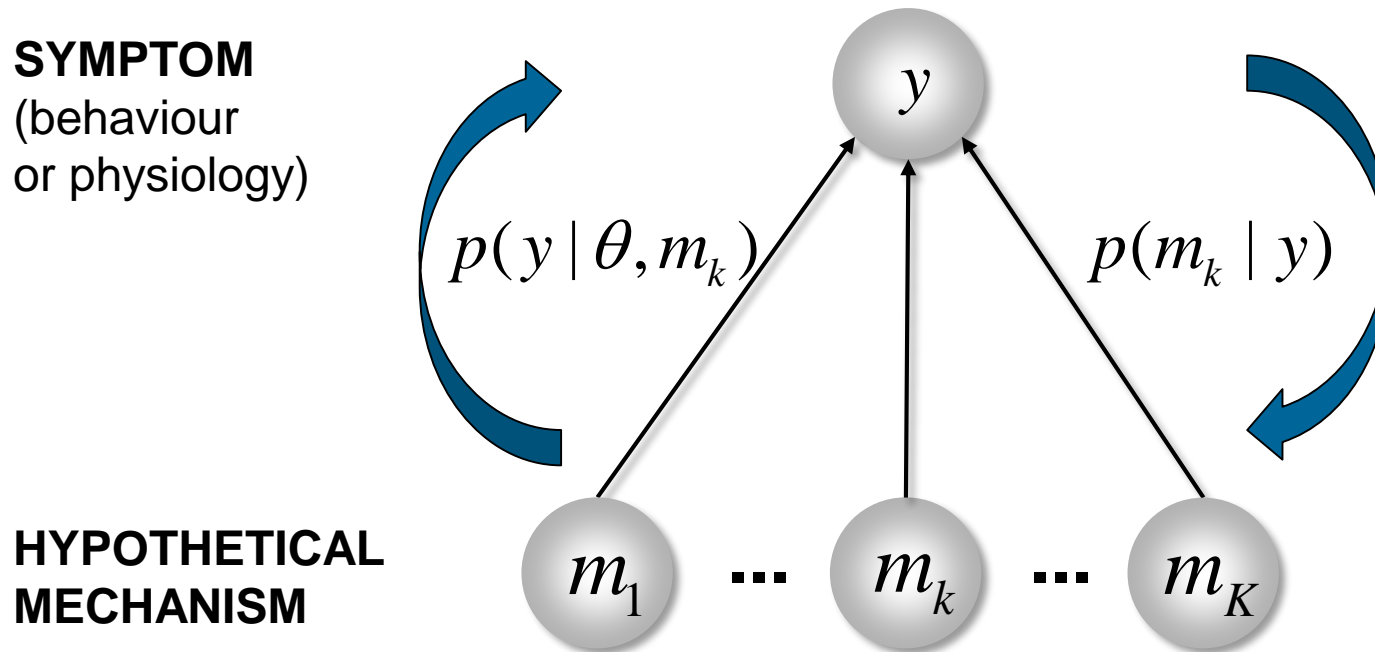
ETH

Eidgenössische Technische Hochschule Zürich
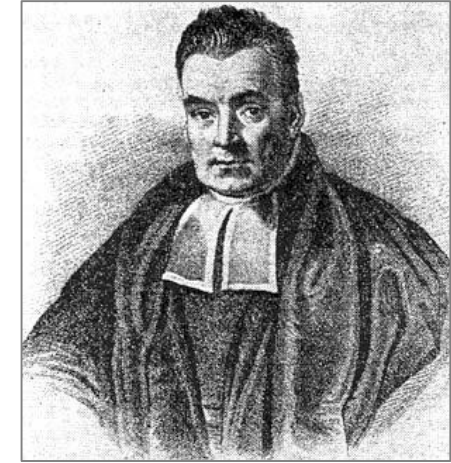Swiss Federal Institute of Technology Zurich

Key goals of CP

Treatment prediction

BMS

Generative embedding
(supervised)

Dissecting spectrum disorders

Differential diagnosis

BMS

Generative embedding
(unsupervised)

BMS

Generative embedding
(supervised)

Computational assays

Generative models of
behaviour & brain activity

BMS for differential diagnosis

SYMPTOM
(behaviour
or physiology)

**HYPOTHETICAL
MECHANISM**

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

Stephan et al. 2017, *NeuroImage*
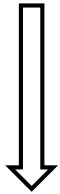
Reminder: Bayes' rule
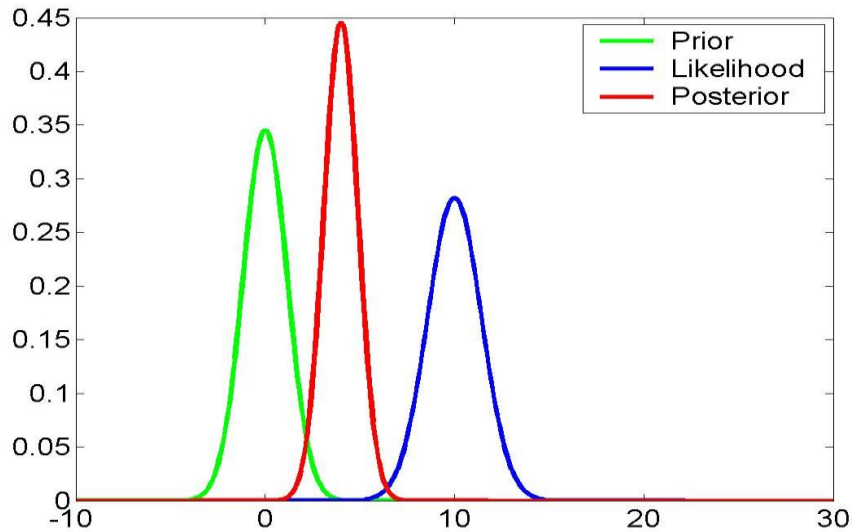
The Reverend Thomas Bayes
(1702-1761)

$$p(\boldsymbol{\theta} \mid \mathbf{y}, m) = \frac{p(\mathbf{y} \mid \boldsymbol{\theta}, m)\, p(\boldsymbol{\theta} \mid m)}{p(\mathbf{y} \mid m)}$$
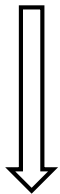
$$\text{posterior} = \frac{\text{likelihood} \times \text{prior}}{\text{evidence}}$$
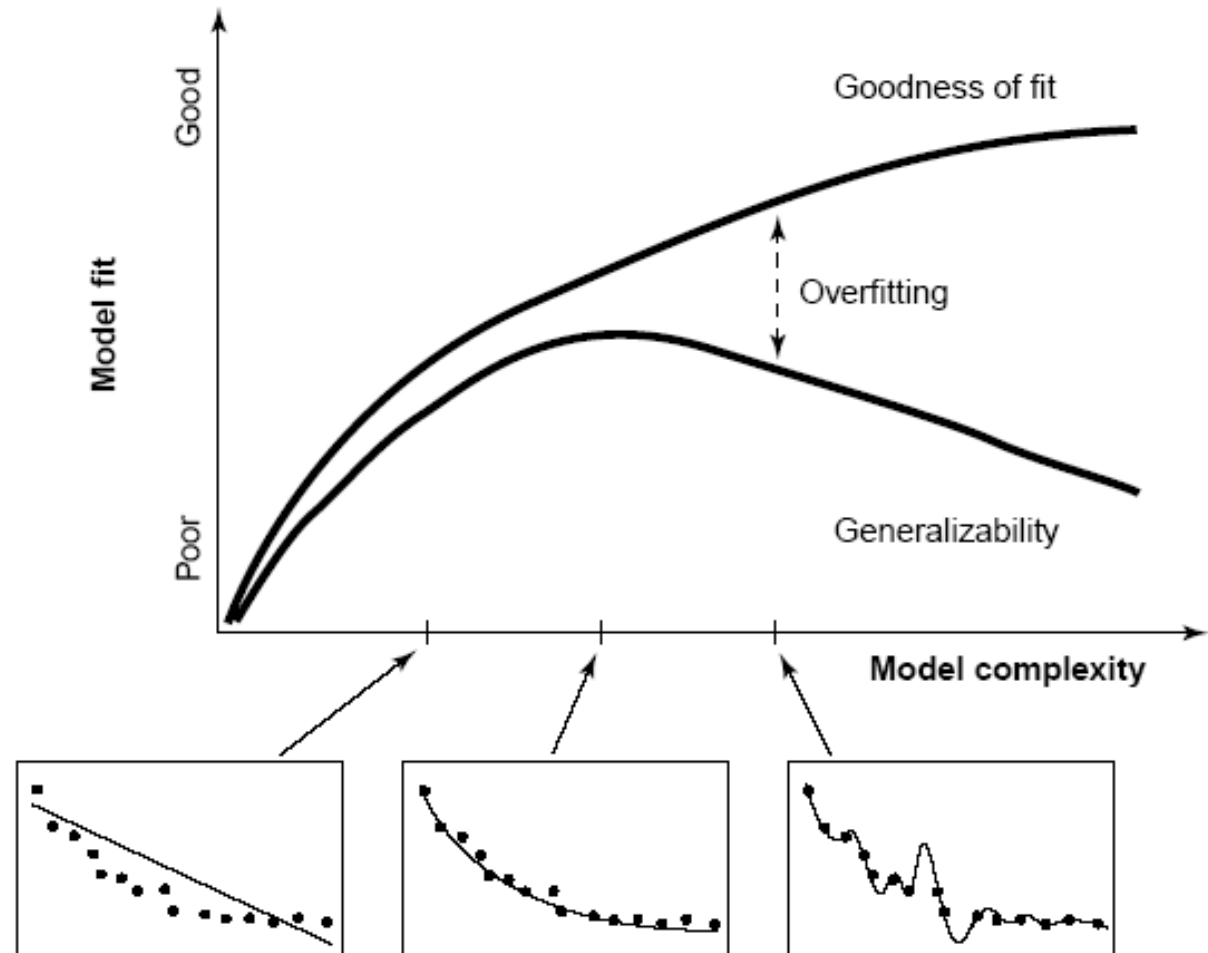
Model comparison and selection

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 major goal of generative modeling: comparing competing hypotheses about the mechanisms underlying observed data

  – a priori definition of hypothesis set (model space) is crucial

- BMS = selecting 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 and only allows one to compare the relative goodness of competing hypotheses within the pre-specified model space!

- Model validation requires external criteria (external to the measured data).

BMS: Defining the model space

- The model space M comprises all models m that are deemed plausible *a priori.* This is the researcher's hypothesis space.

- Defining M is quivalent to defining a prior over models, $p(m)$, such that a finite set of models has non-zero prior probability.

- Usually, all models are assigned equal prior probability.

all possible
models m
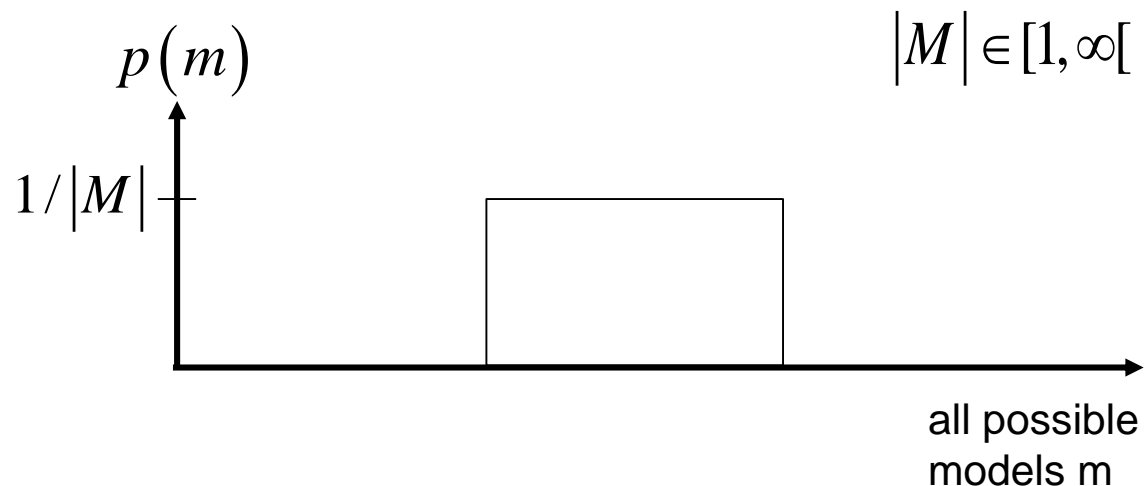
BMS: Applying Bayes' rule to models

- We wish to select the model that has the highest posterior probability given the data:

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

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

$$p(m_i \mid y) = \frac{p(y \mid m_i)}{\boxed{\sum_{j=1}^{|M|} p(y \mid m_j)}} \quad \text{identical for all models}$$

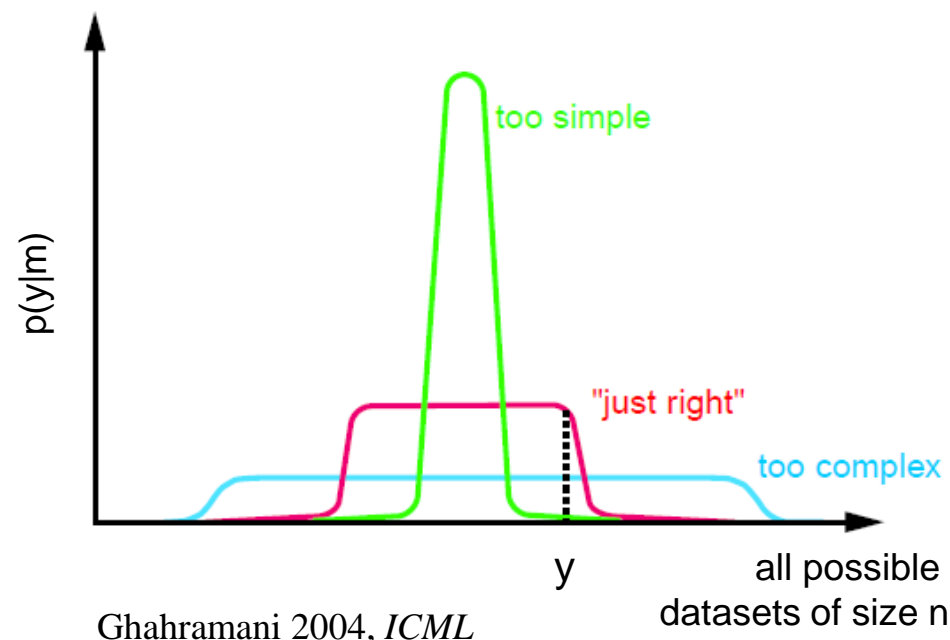- We can thus use the model evidence to compare and select models.

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 likelihood that randomly selected parameters from the prior would generate the observed data

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

Ghahramani 2004, *ICML*

Overly simple models are unlikely to generate the given data.

Overly complex models can generate many possible datasets and are thus unlikely to have generated the specific dataset that is given.

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?"

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

Ghahramani 2004, *ICML*

Overly simple models are unlikely to generate the given data.

Overly complex models can generate many possible datasets and are thus unlikely to have generated the specific dataset that is given.
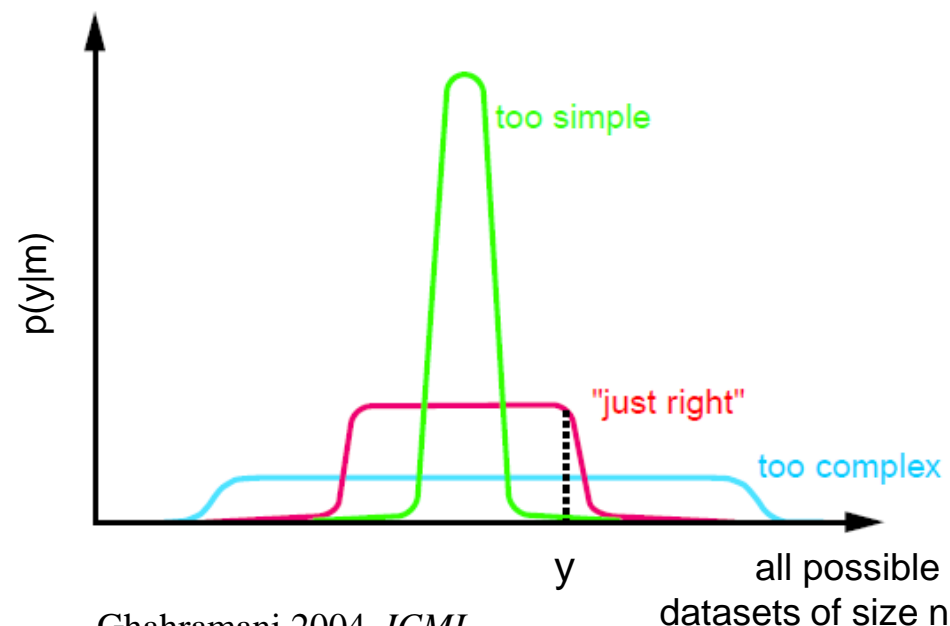
Model evidence: How can it be computed?

- **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

- Generally:
  - 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:

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

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

No. of parameters

Akaike Information Criterion: $\quad AIC = \log p(y\,|\,\theta, m) - p$

No. of data points

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

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.

true posterior $p(\theta|y)$

divergence KL$[q||p]$

best proxy $q(\theta)$

hypothesis class
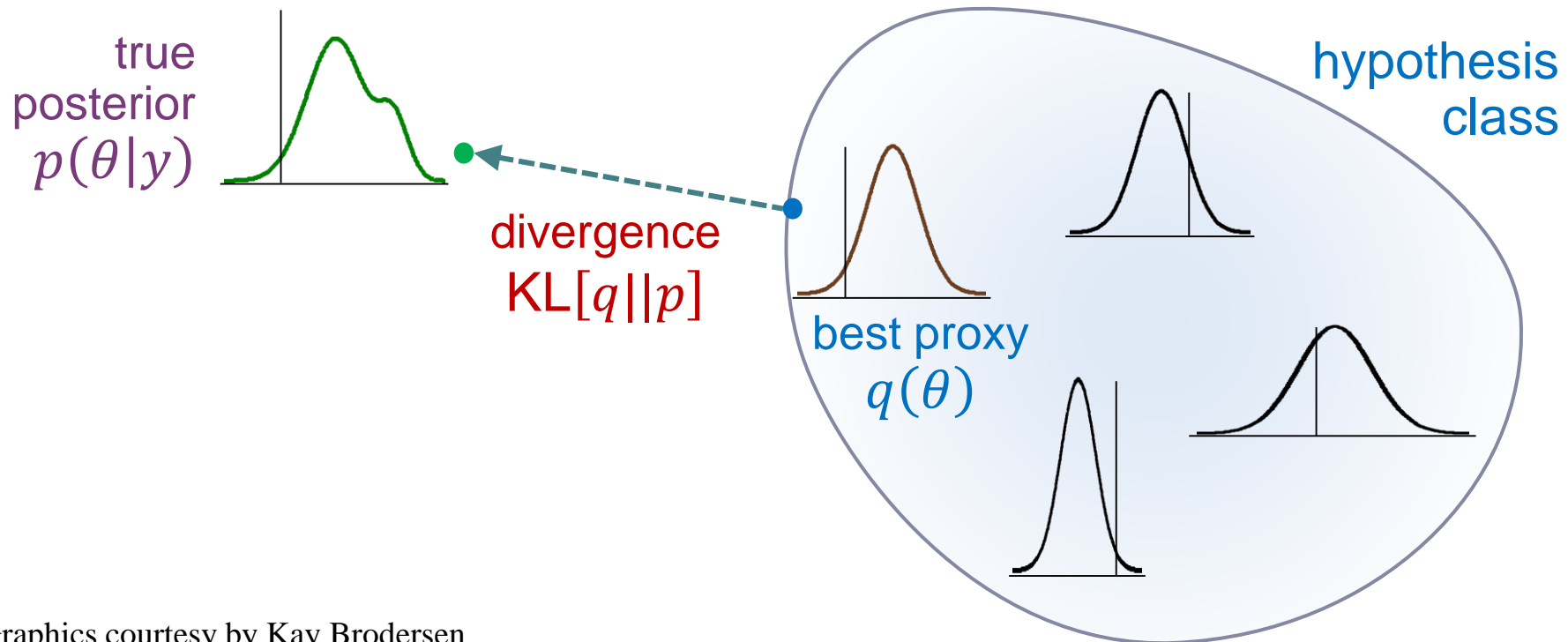
Recap: Variational Bayes

$$\ln p(y) = \underbrace{\text{KL}[q||p]}_{\substack{\textbf{divergence} \\ \geq 0 \\ \text{(unknown)}}} + \underbrace{F(q,y)}_{\substack{\textbf{neg. free} \\ \textbf{energy} \\ \text{(easy to evaluate} \\ \text{for a given } q\text{)}}}$$

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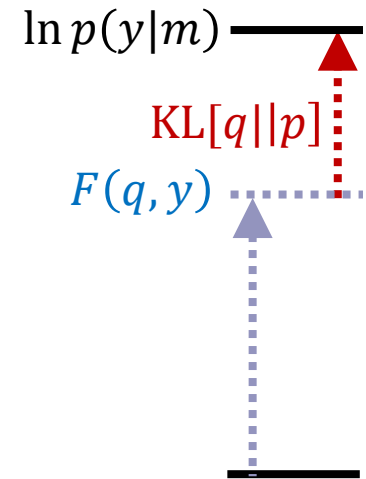
Maximizing $F(q, y)$ is equivalent to:

- minimizing $\text{KL}[q||p]$,

- 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
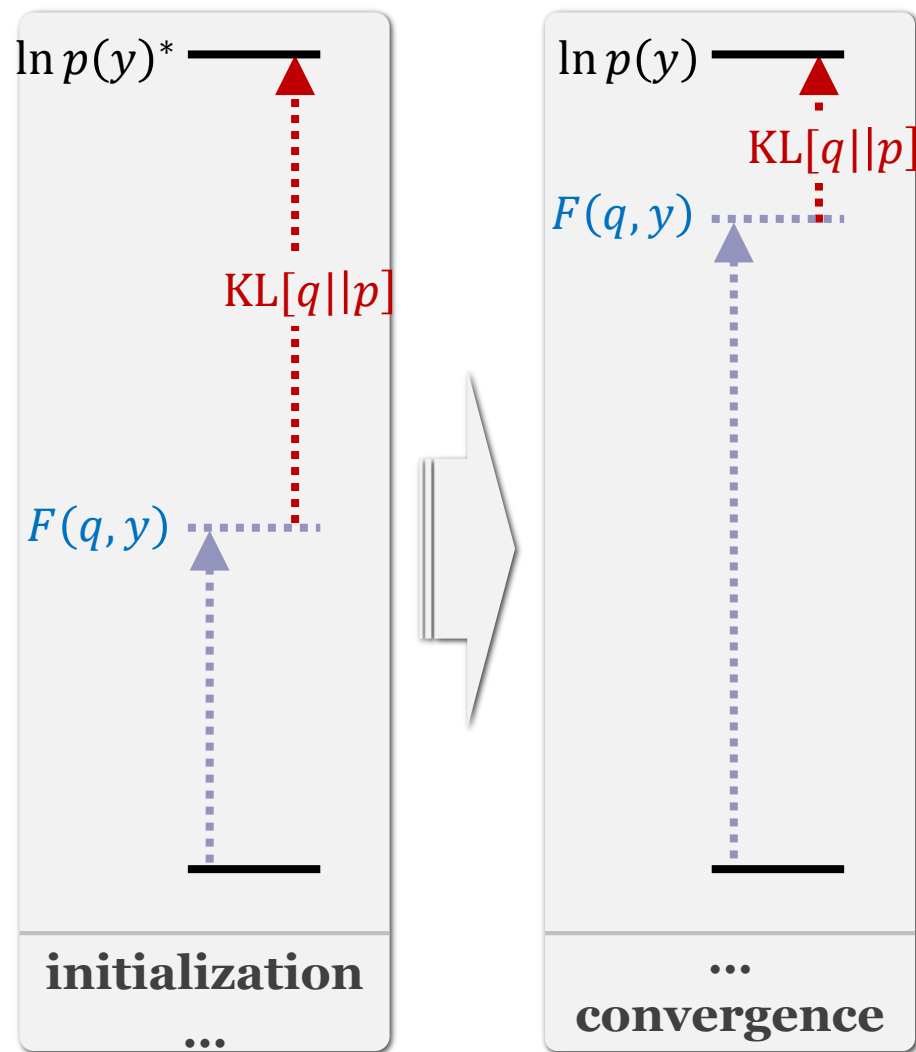
The negative free energy approximation F

F is a lower bound on the log model evidence:

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

$\ln p(y|m)$ ────

KL[q||p]

$F(q, y)$

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*

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):

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

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

- 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, $[0; \infty[$

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
$\geq$ 150	$\geq$ 99%	Very strong

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

Fixed effects BMS at group level

Group Bayes factor (GBF) for 1...*K* subjects:
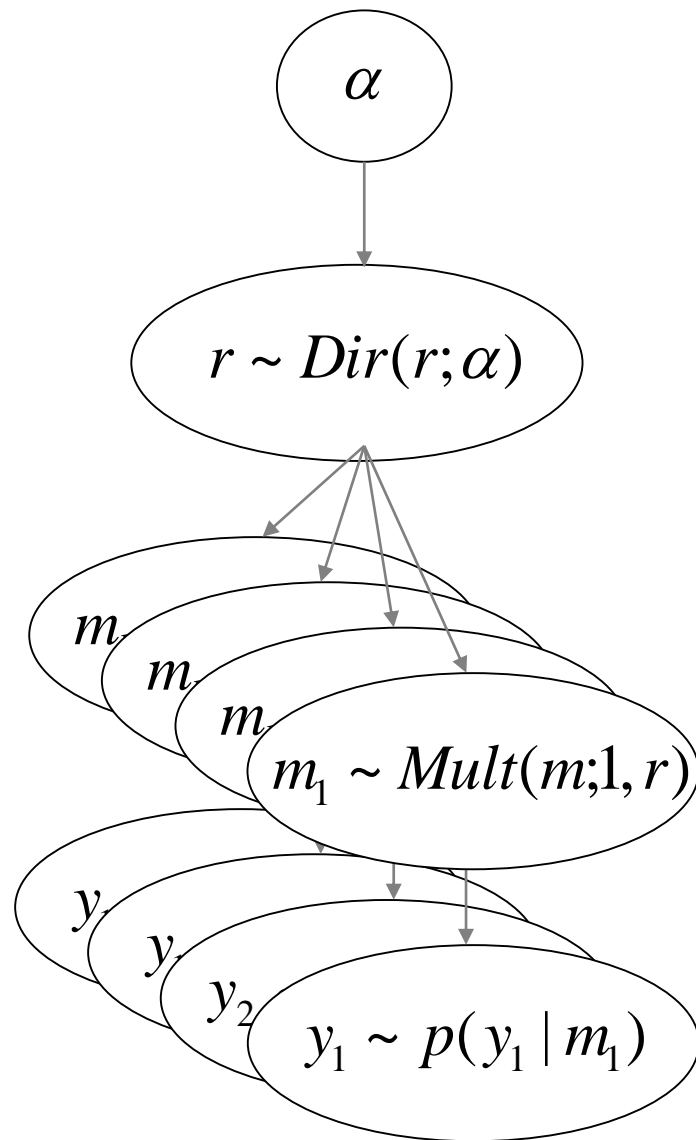
$$GBF_{ij} = \prod_k BF_{ij}^{(k)}$$

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

α

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

$r \sim Dir(r;\alpha)$

Dirichlet distribution of model probabilities r

$m_1 \sim Mult(m;1,r)$

Multinomial distribution of model labels m

Model inversion by Variational Bayes or MCMC

$y_1 \sim p(y_1 \mid m_1)$

Measured data y

Stephan et al. 2009, *NeuroImage*

Random effects BMS for heterogeneous groups

Dirichlet parameters α
= "occurrences" of models in the population
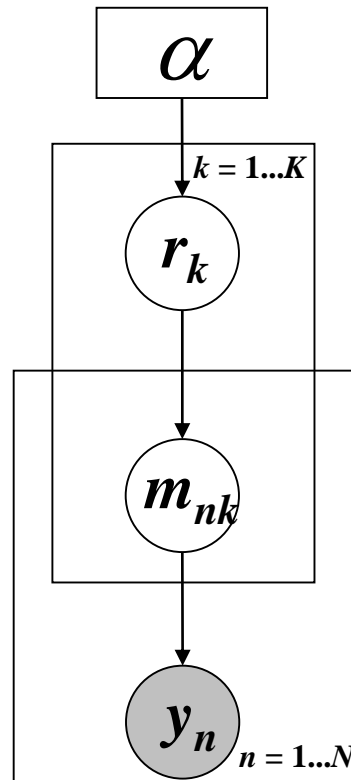
Dirichlet distribution of model probabilities r

Multinomial distribution of model labels m

Model inversion by Variational Bayes or MCMC

Measured data y

Stephan et al. 2009, *NeuroImage*

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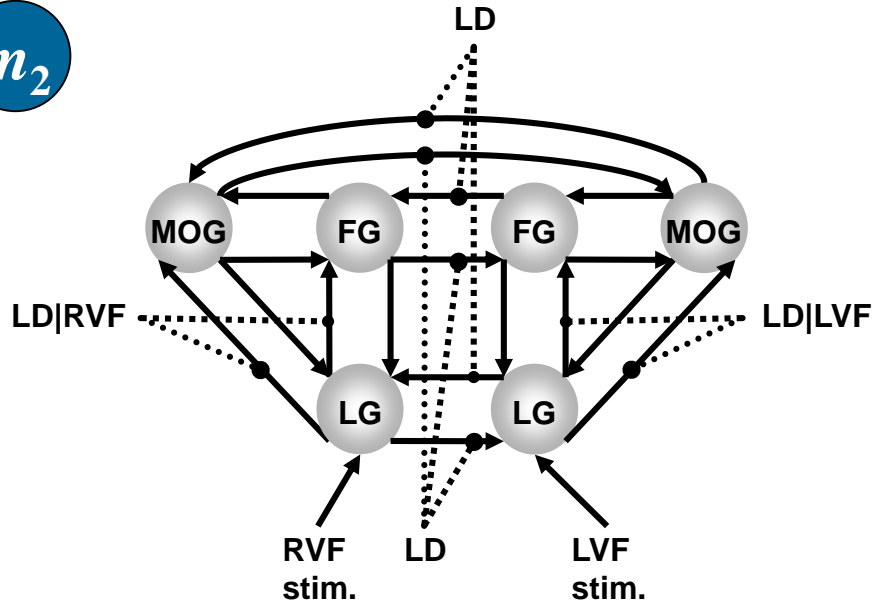
$$\exists k \in \{1\ldots K\}, \forall j \in \{1\ldots K \mid j \neq k\}:$$
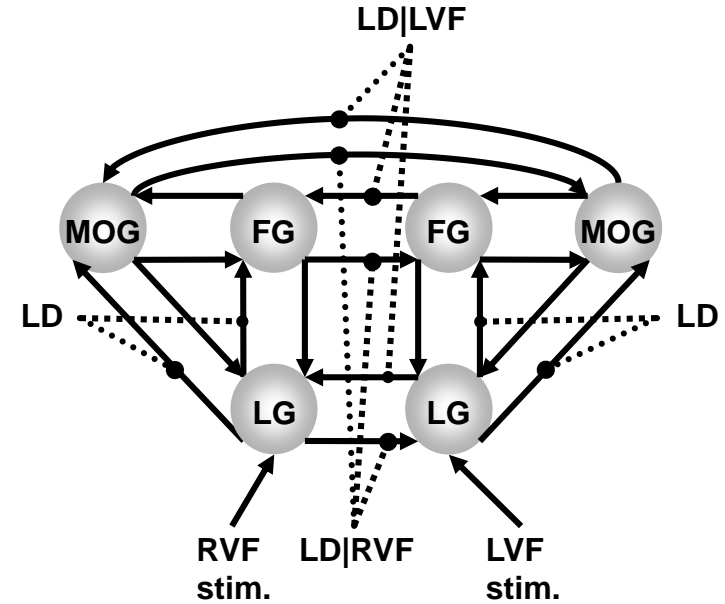$$\varphi_k = p(r_k > r_j \mid y; \alpha)$$

4. **protected exceedance probability (PXP)**: see below

Example: Hemispheric interactions during vision
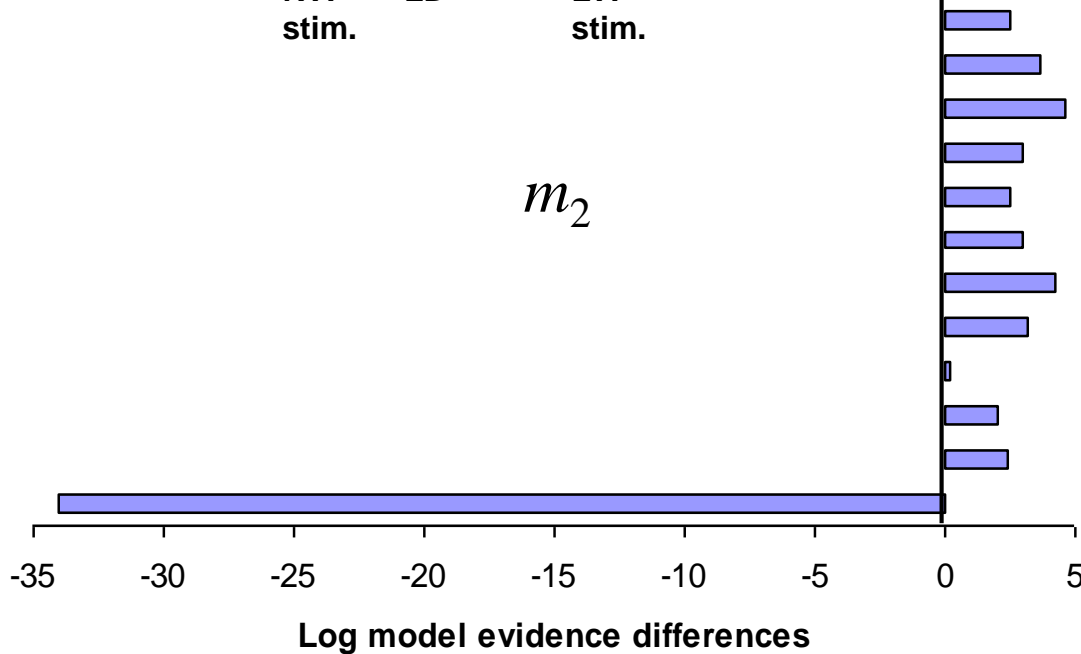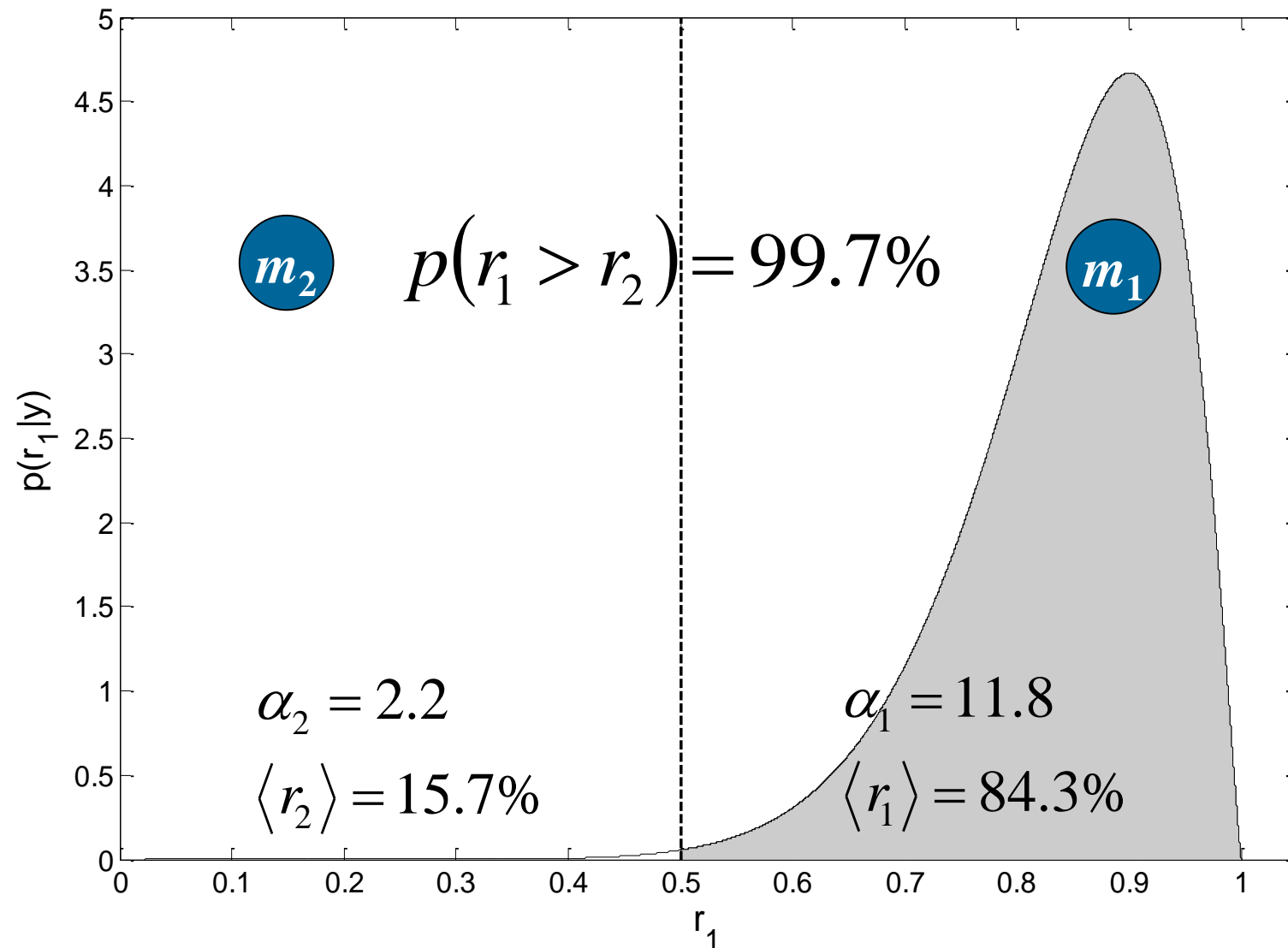

Data: Stephan et al. 2003, *Science*
Models: Stephan et al. 2007, *J. Neurosci.*

$$p(r_1 > r_2) = 99.7\%$$

m_2

m_1

$$\alpha_2 = 2.2$$

$$\langle r_2 \rangle = 15.7\%$$

$$\alpha_1 = 11.8$$

$$\langle r_1 \rangle = 84.3\%$$

Stephan et al. 2009, *NeuroImage*

Protected exceedance probability:
# Using BMA to protect against chance findings

- XPs 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: $r_k = 1/K$

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

$$PXP = P\left(r_k \geq r_{k'=k} \mid y\right)$$

$$= P\left(r_k \geq r_{k'=k} \mid y, H_1\right) P\left(H_1 \mid y\right) + P\left(r_k \geq r_{k'=k} \mid y, H_0\right) P\left(H_0 \mid y\right)$$

Rigoux et al. 2014, *NeuroImage*

Example: DA effects on strial connectivity

- task-free fMRI with 2x2 pharmacological design (sulpiride, bromocriptin, ON/OFF)

- stochastic DCM

- DA drug effects on the striatum vary widely across individuals and might average to zero

- BMS: identifies where DA drugs alter intrastriatal connectivity, even when the sign of this effect was not consistent across participants
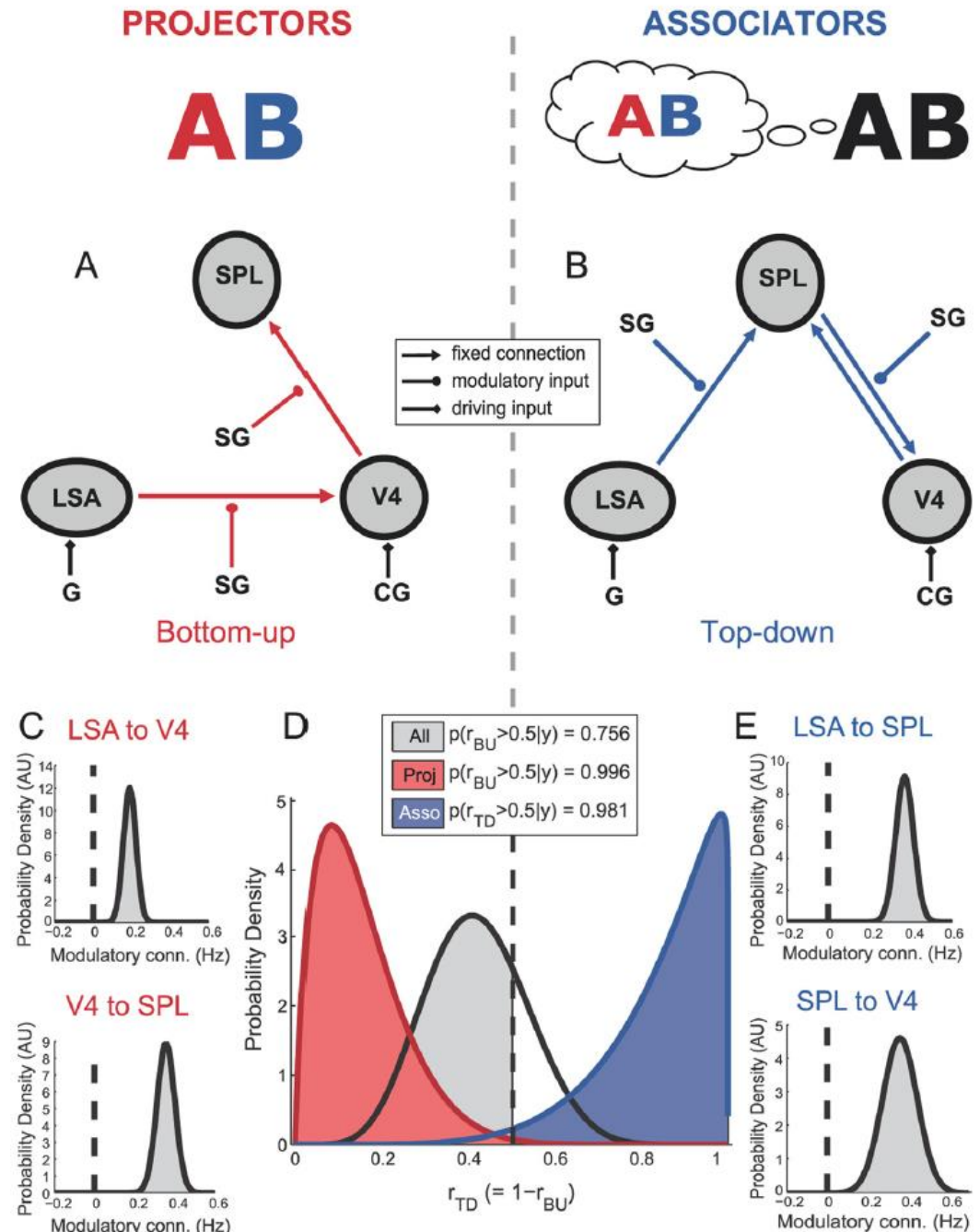
intra-striatal connectivity
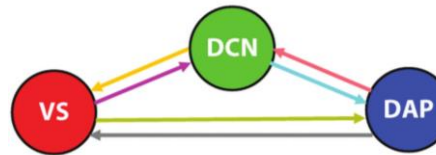
dopaminergic modulation

winning model

VS: ventral striatum
DCN: dorsal caudate nucleus
DAP: dorsal anterior putamen

Piray et al. 2017, *Cereb. Cortex*

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.*

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?

No.

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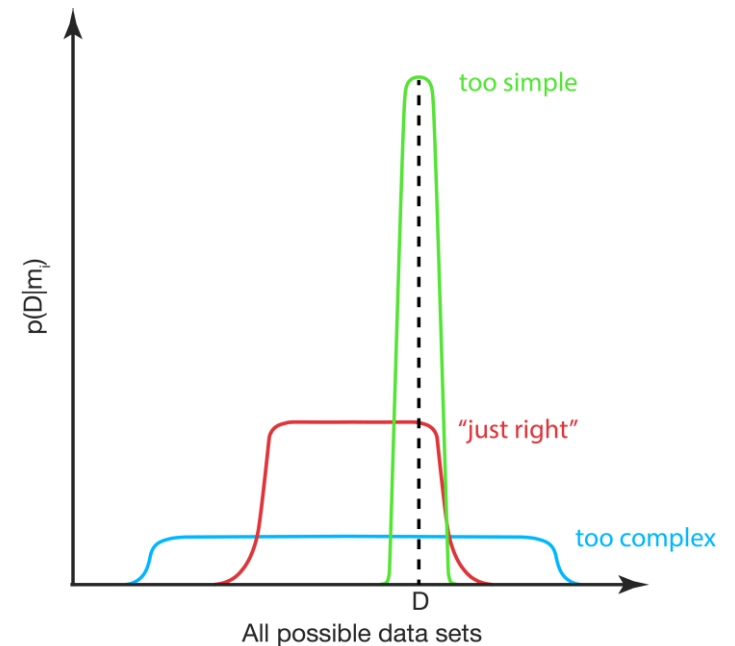
Overfitting at the level of models

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?

No.
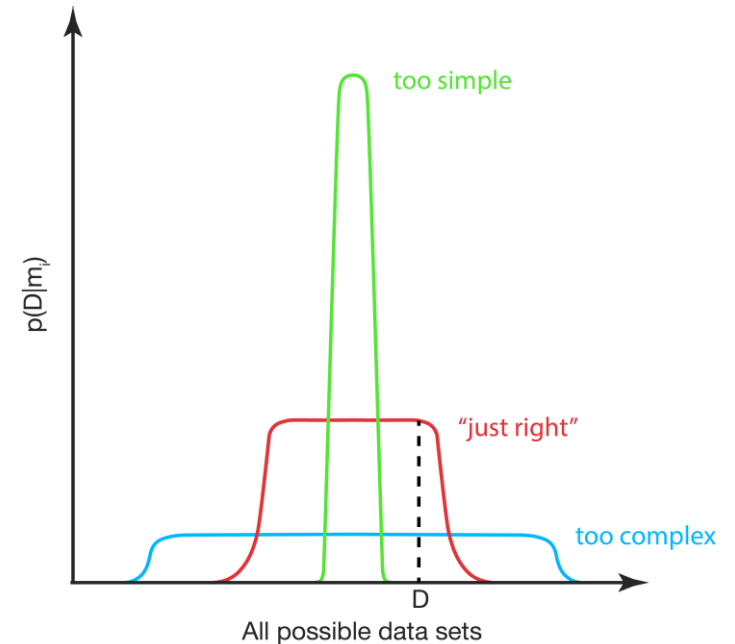
The more models are included in the model space, the risk of overfitting (at the level of models) increases, too.

Overfitting at the level of models

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

- Three solutions in the context of BMS:
  – regularisation: choosing priors over models p(m) such that model space is small
  – family-level BMS
  – Bayesian model averaging (BMA)

Family-level BMS

- rests on 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

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

- the uniform prior over models is replaced with a uniform prior over families:

  - For **FFX**, this means family prior $p(f) = \frac{1}{K}$ and model prior $p(m) = \frac{1}{KN_k}$.

  - **RFX**, fair priors over families is achieved by setting $\alpha_{prior}(m) = \frac{1}{N_k}$.

Example application of family-level BMS: Bipolar disorder

- Go/No-Go task to emotional faces with bipolar patients, at-risk individuals, healthy controls

- DCM used to explain interaction of motor inhibition and fear processing during this task.

- Family-level BMS showed differences in general model structure (bilinear vs. nonlinear models) between at-risk group vs. healthy controls and bipolar patients.

Breakspear et al. 2015, *Brain*
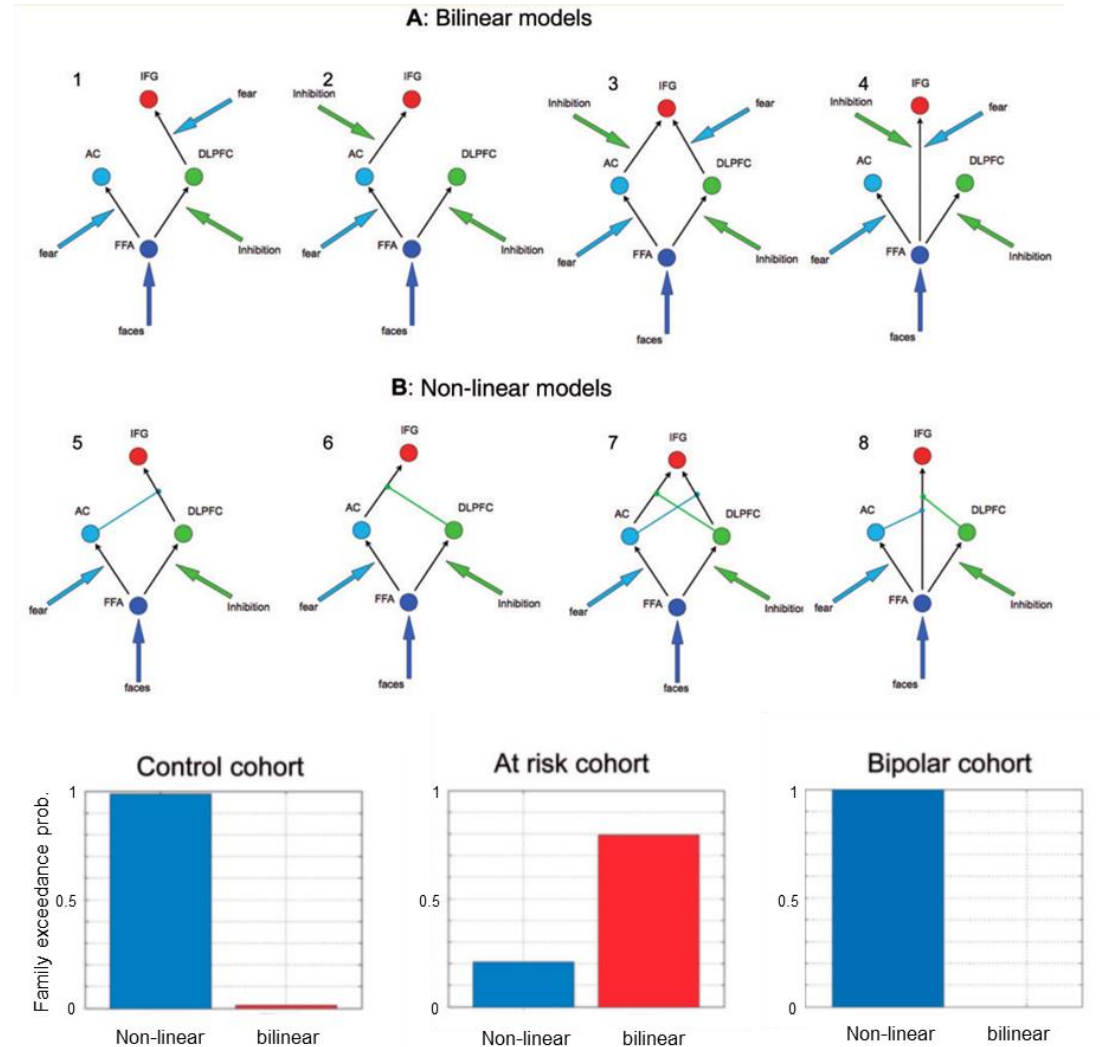
Bayesian Model Averaging (BMA)

- 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:

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

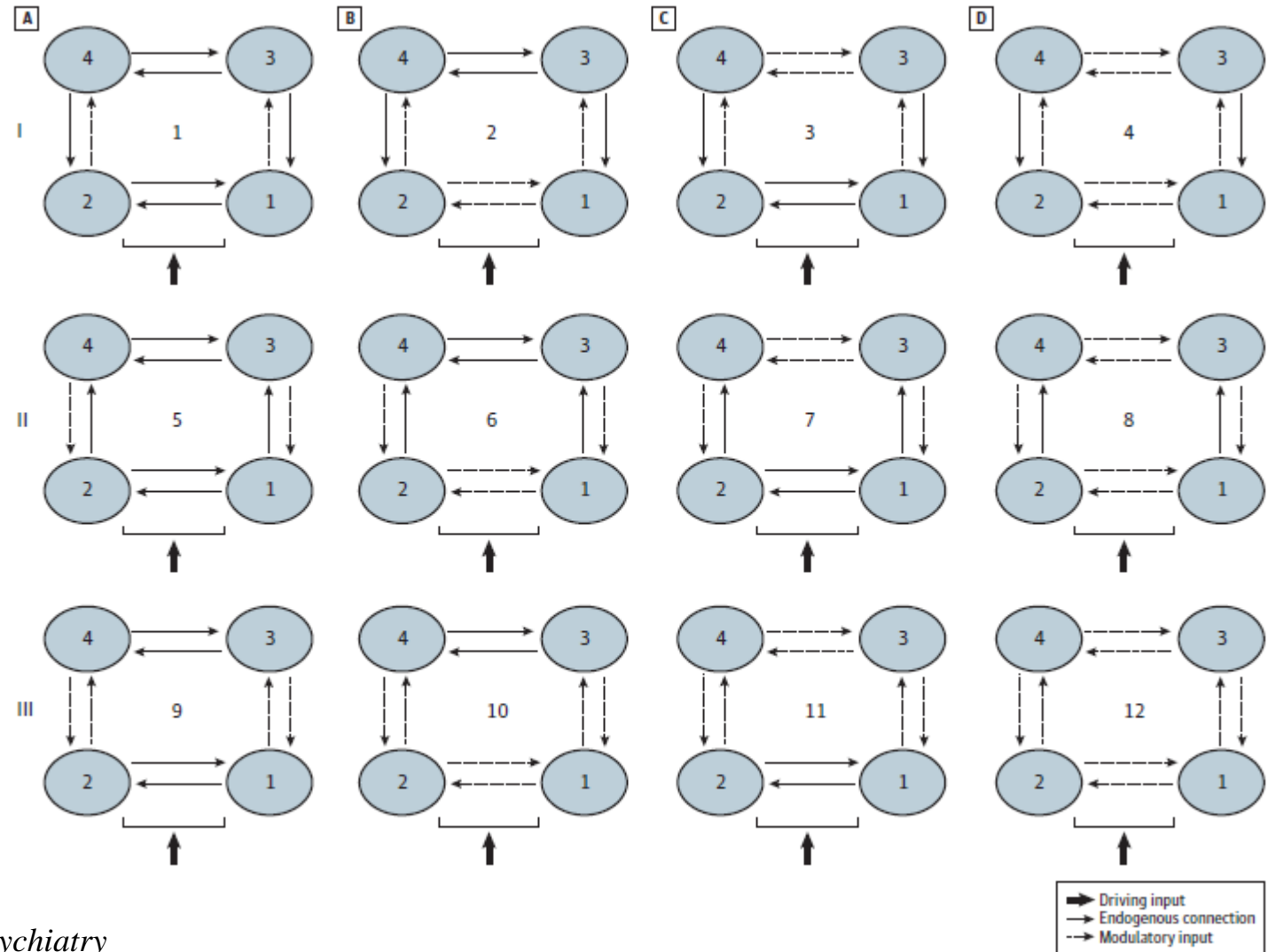$$= \sum_m p(\theta \mid y, m) p(m \mid y)$$

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

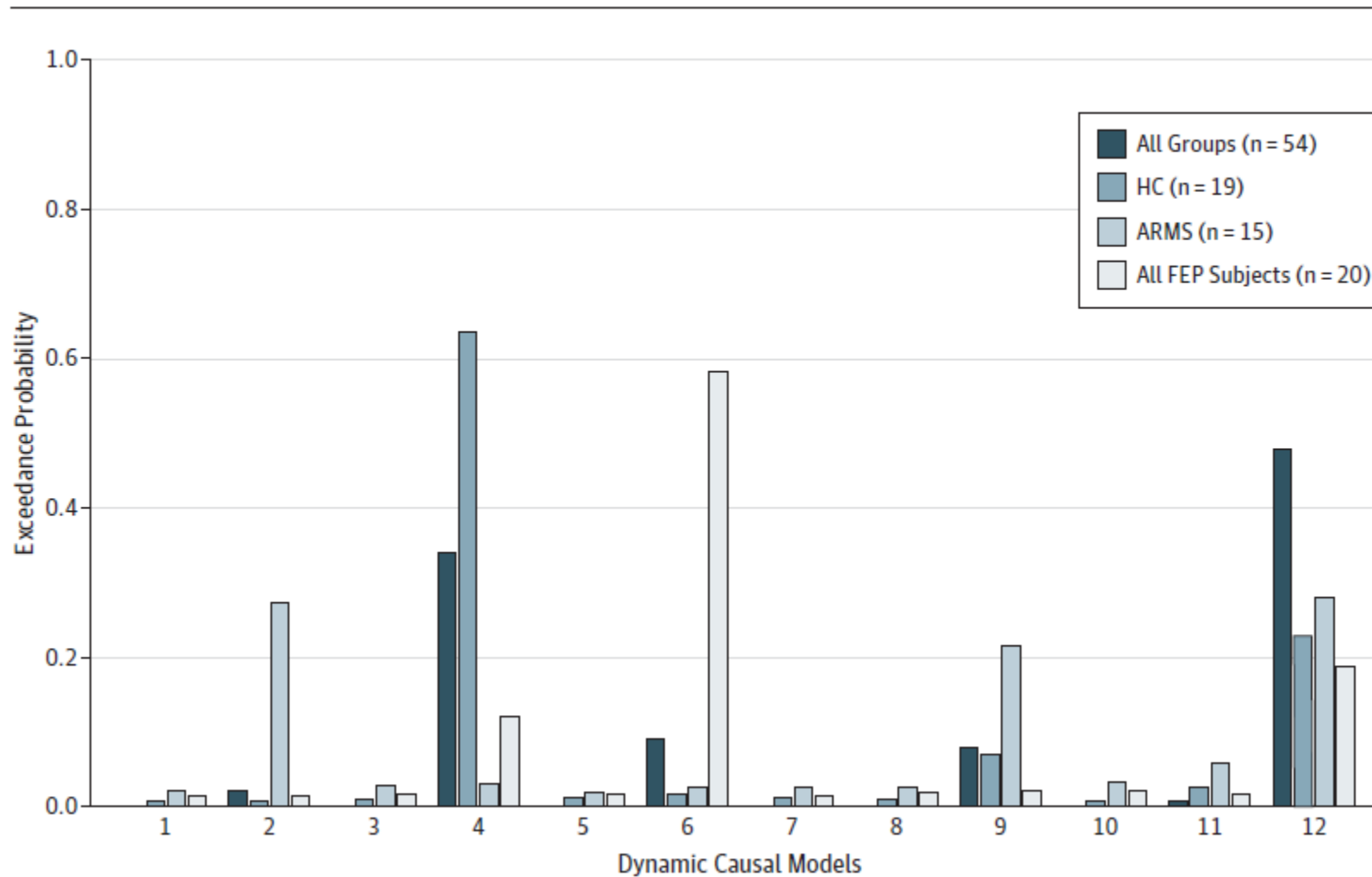Penny et al. 2010, *PLoS Comput. Biol.*

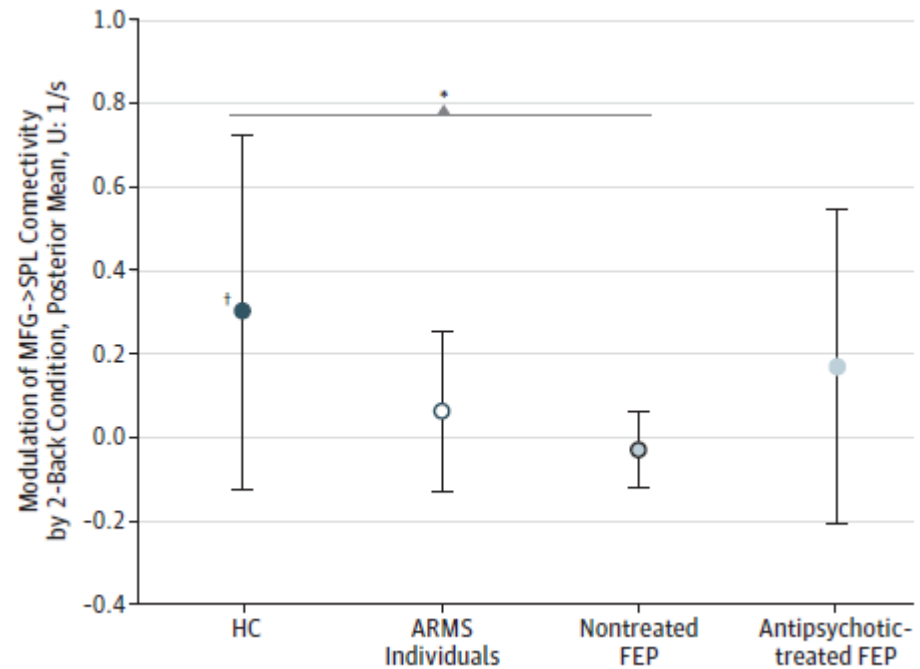# Prefrontal-parietal connectivity during working memory in schizophrenia

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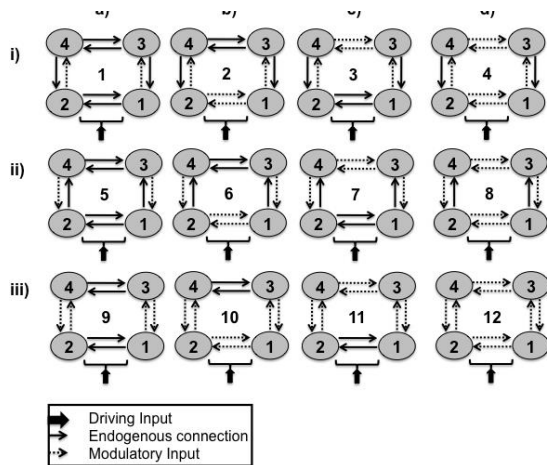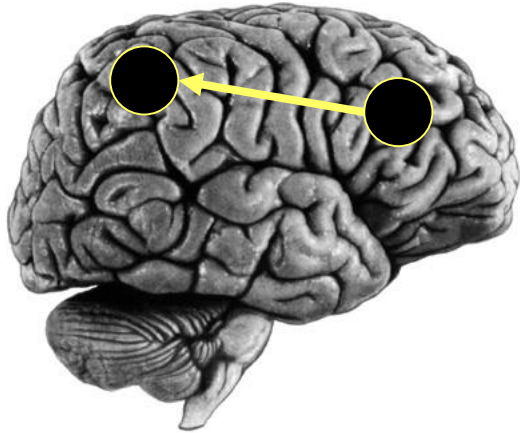
17 ARMS, 21 first-episode (13 non-treated), 20 controls

Schmidt et al. 2013, *JAMA Psychiatry*

definition of model space

inference on **model structure** or inference on **model parameters**?

inference on **individual models** or **model space partition**?

inference on **parameters of an optimal model** or **parameters of all models**?

optimal model structure assumed to be identical across subjects?

comparison of model families using FFX or RFX BMS

optimal model structure assumed to be identical across subjects?

BMA

yes

no

FFX BMS

RFX BMS

yes

no

FFX BMS

RFX BMS

FFX analysis of parameter estimates (e.g. BPA)

RFX analysis of parameter estimates (e.g. t-test, ANOVA)

Stephan et al. 2010, *NeuroImage*

Important topics I have not covered in this talk

- **Bayesian model reduction**
  - an efficient alternative to RFX-BMS (for nested models)
  - see: Friston KJ et al. (2016) NeuroImage 128:413-431

- estimating the model evidence with **sampling-based methods**
  - for a tutorial review, see
    https://www.biorxiv.org/content/10.1101/471417v1.abstract

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