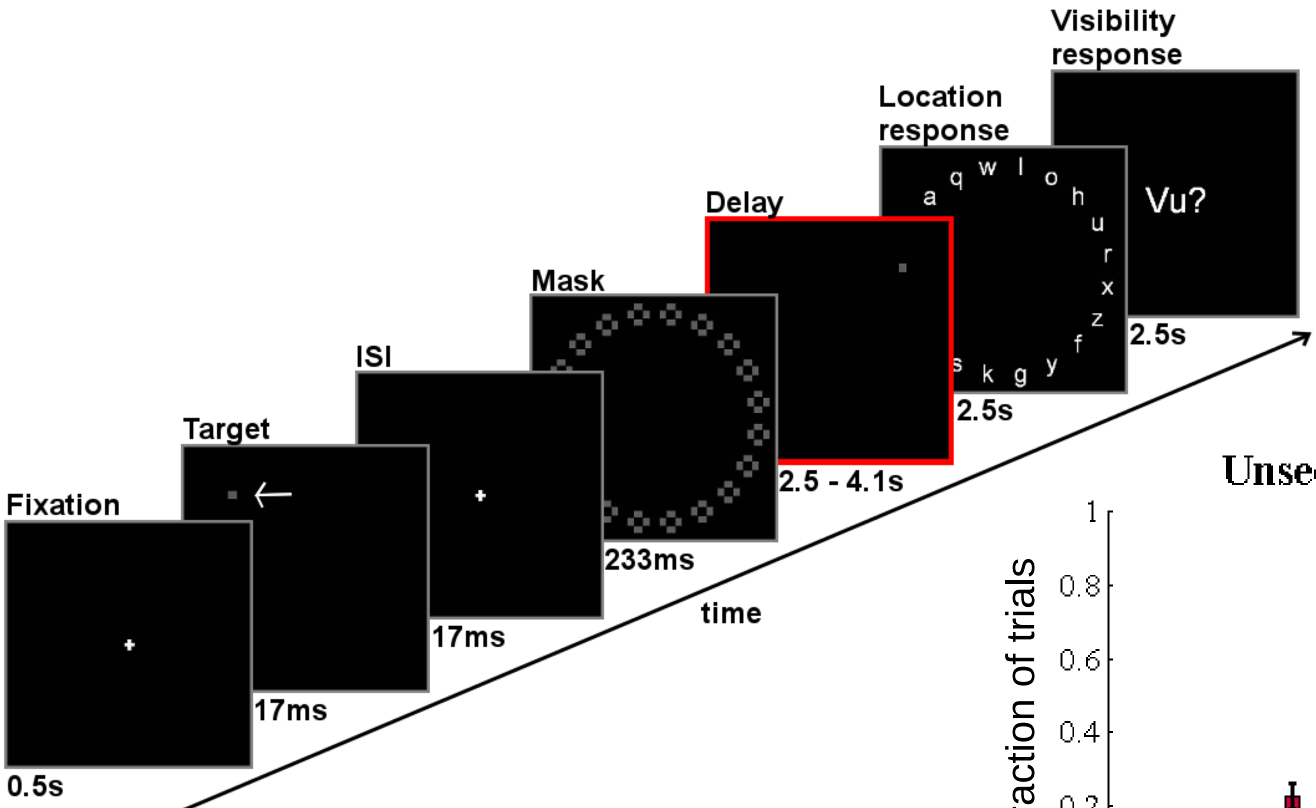
# Model comparison with Bayesian statistics

Florent Meyniel
CEA-Saclay
ENP – May 21st 2015

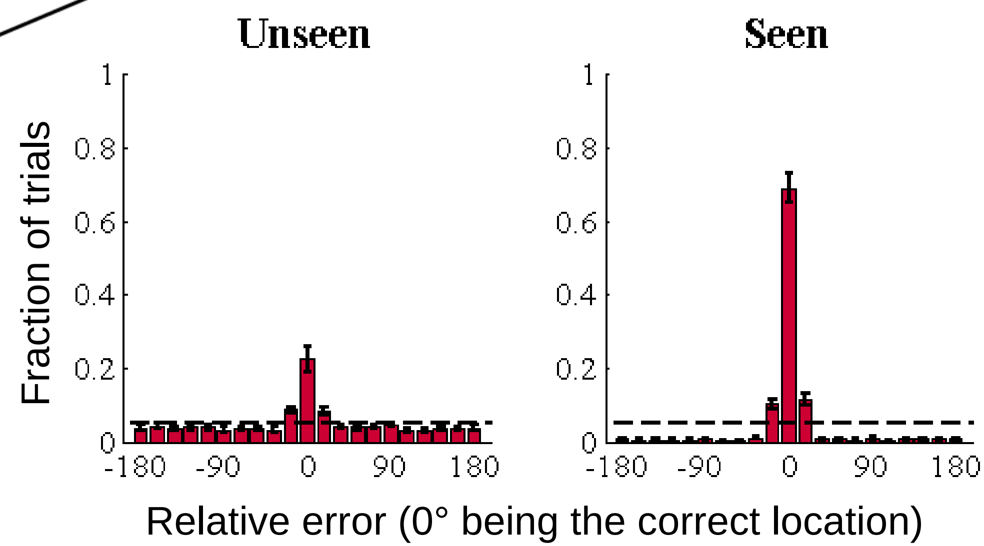
# Starting with an example

Question: Is there any information processed when the stimulus is not consciously perceived?

Response with necked eye: yes!
How to quantify the evidence?




(data & paradigm: Darinka Trübutschek et al.)




**Three hypotheses (=models) about the data:**
**M1:** the responses are random (= uniform distribution)
**M2:** the responses are informed, more or less concentrated around the correct answer (Gaussian distribution with *unknown* variance)
**M3:** an *unknown* fraction of random responses, the others are informed (with *unknown* variance)

*What is the fraction of random guesses in the 'Unseen' trials?*

*Is it more likely that there is something (M3) in 'Unseen' trials rather than nothing (M1)?*

# Topics addressed:

- The notion of conditional probabilities and Bayes' rule
- Characterization of a model (e.g. guessing unknown parameters) with Bayesian statistics
- Quantifying the evidence supporting a model (or a hypothesis) with Bayesian statistics
- Simpler is better: Bayesian statistics automatically penalize complexity
- Bayesian model comparison and hypothesis testing

# Conditional probability and Bayes' rule: going back and forth between observations and assumptions

We often (if not always) estimate plausibility given some prior information and / or assumptions.

In probability theory, this corresponds to **conditional probabilities**.
It can be linked to the 'If …, then …' reasoning.

If he is a trader, then he is likely to be rich:

p(rich | trader) = high *correlation*

If it rains, then the ground is likely to be wet:

p(wet | rain) = high *physical causation*

If it is a square then it is a rectangle:

p(rectangle | square)=1 *nested properties*

Conditional probabilities characterize an epistemic dependence, not a causal link. The symmetry of this dependence is known as Bayes' rule:

$$p(A|B)=\frac{p(A,B)}{P(B)}$$

$$=p(B|A)\frac{p(A)}{p(B)}$$

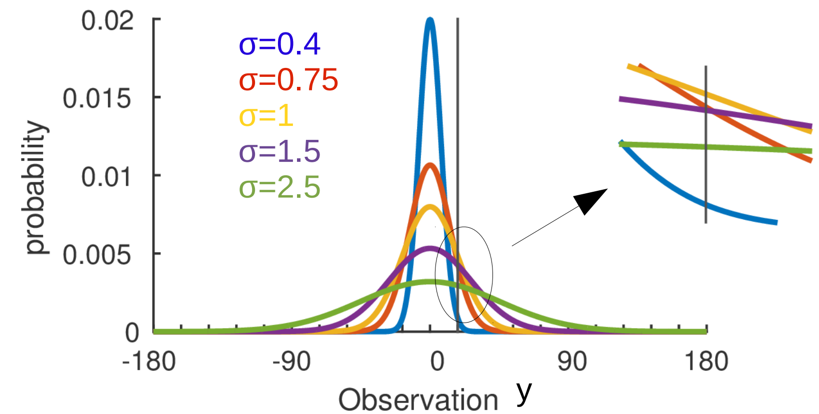
Bayes' rule affords **inference** about assumptions given actual data:
p(assumption | observations) ~ p(observations | assumption) * p(assumption)

# Model parameters and data: going back and forth with Bayes' rule

Likelihood of **observations**:

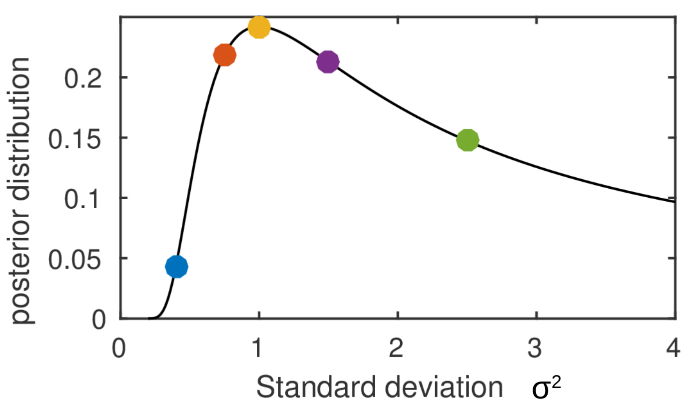
$$p(y|Gaussian, \mu=0, \sigma^2) = N(\mu=0, \sigma^2)$$

The **model**
A simple example: Model2
A Gaussian process
with mean μ=0, std = $\sigma^2$




Go the other way around with Bayes' rule

$$\underbrace{p(\sigma^2|Gaussian, \mu=0, y)}_{\text{Posterior knowledge about } \sigma^2} = \underbrace{p(y|Gaussian, \mu=0, \sigma^2)}_{\text{Likelihood of the data assuming } \sigma^2} \underbrace{p(\sigma^2|Gaussian)}_{\text{Prior knowledge about } \sigma^2 \text{ (may be constant)}} \underbrace{\frac{1}{p(y)}}_{\text{constant}}$$




Observed data: y=36°

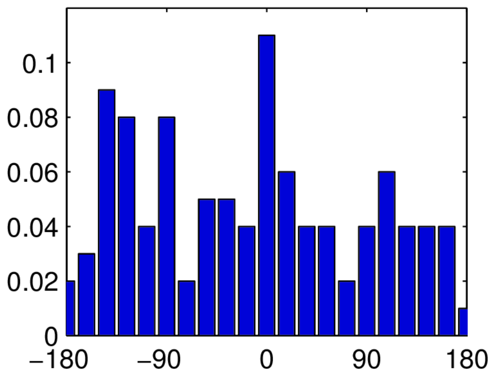
(For a collection of data: $y_1$=36°, $y_2$=0°, ..., use the product:

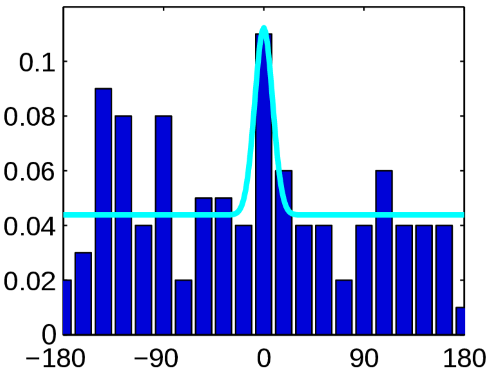
$$p(y_1, y_2, \ldots, \sigma^2|\mu=0) \propto p(y_1|\mu=0, \sigma^2) \ldots p(y_2|\mu, \sigma^2)\, p(\sigma^2)$$

# Bayesian inference of the unknown parameters given the observed data

**Characteristics** | **Simulated data (100 trials)** | **Best Fit of M3 (fits proportion + σ)** | **Maximum A Posterior value**

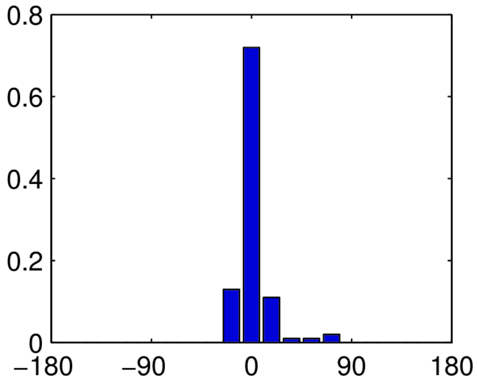
20% informed responses (σ=0.5) + 80% random guesses

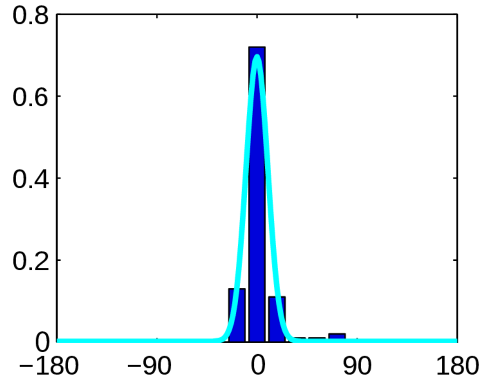



Proportion=0.08
σ=0.45

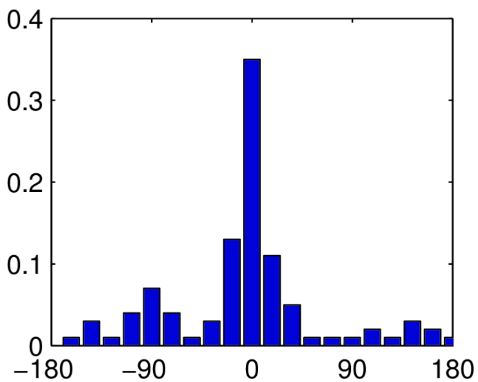
95% informed responses (σ=0.5) + 5% random guesses

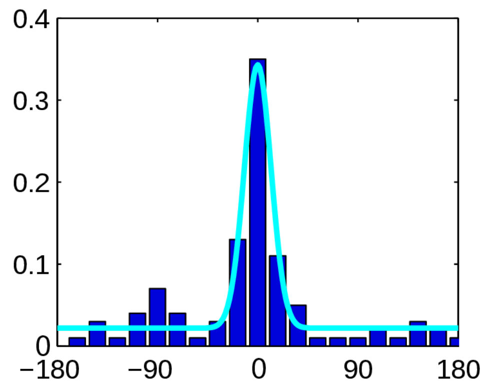



Proportion=0.96
σ=0.55

50% informed responses (σ=0.5) + 50% random guesses




Proportion=0.54
σ=0.67

# Model and data: going back and forth with Bayes' rule (again)

The **posterior probability of the model** quantifies
the plausibility of this model given some data: p(M1 | y)

It allows direct comparison between models:
e.g. 'Given our data, model #1 is 10 times more probable than model #2'

Following Bayes rule:
p(M1 | y) ~ p(y | M1)p(M1)

The dependence between the posterior and the data depends on p(y | M1), known as **model evidence**, a.k.a. marginal likelihood

In the absence of informative prior about models: p(M1) = p(M2) = constant and the ratio of posterior model probabilities is determined by the ratio of model evidence.

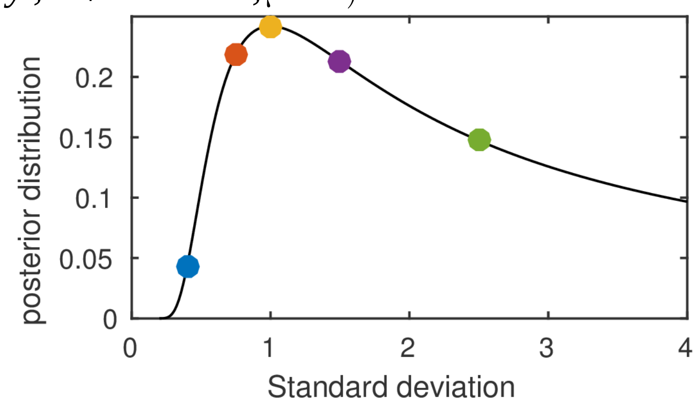
# p(data | model) quantifies the model evidence
## irrespective of any unknown parameters

Back to the example of M2, a Gaussian process with 0 mean and unknown variance.
What we want: $p(y|Gaussian, \mu=0)$
What we know: $p(y,\sigma^2|Gaussian,\mu=0) = p(y|Gaussian,\mu=0,\sigma^2)\,p(\sigma^2|Gaussian)$

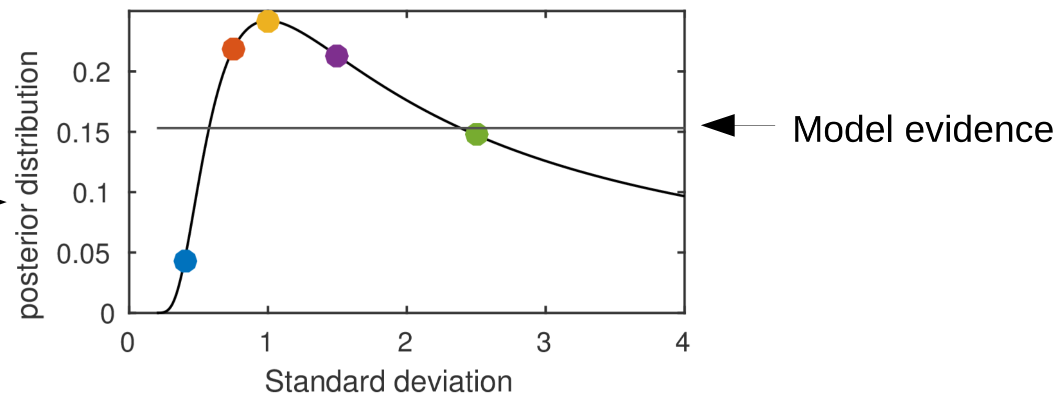
The trick: get rid of the parameter σ by averaging over all possible values (marginalization)

$$
\begin{aligned}
p(y|Gaussian,\mu=0) &= \int p(y,\sigma^2|Gaussian,\mu=0)\,d\sigma \\
&= \int p(y|Gaussian,\mu=0,\sigma^2)\,p(\sigma^2|Gaussian)\,d\sigma
\end{aligned}
$$

$p(y,\sigma^2|Gaussian,\mu=0)$

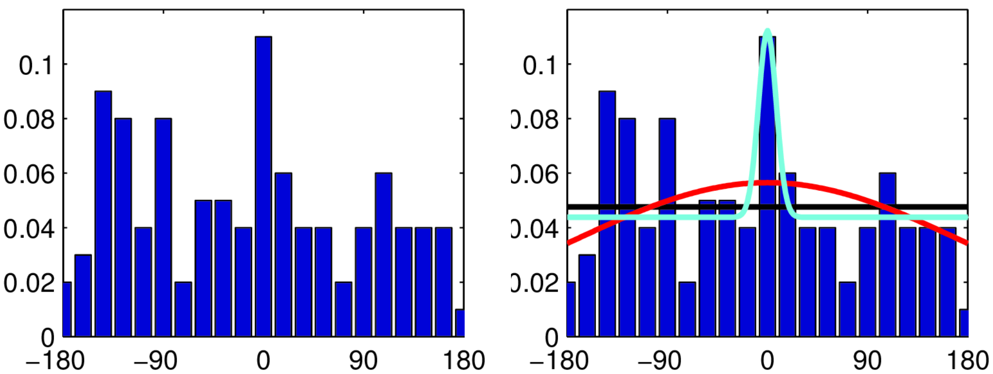
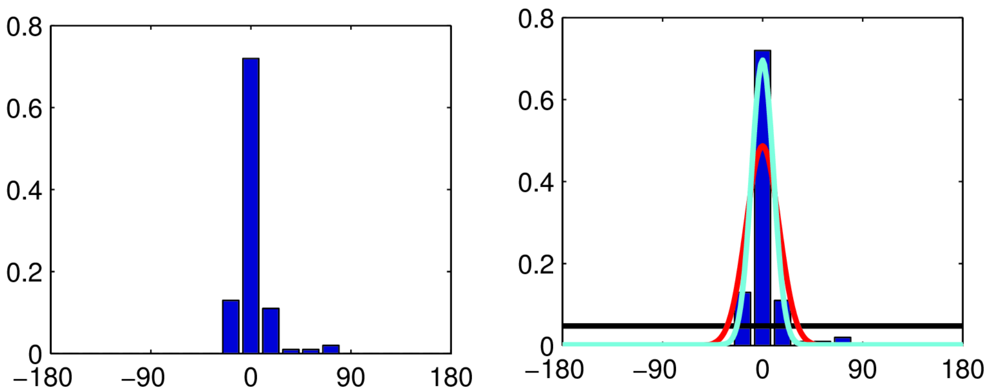
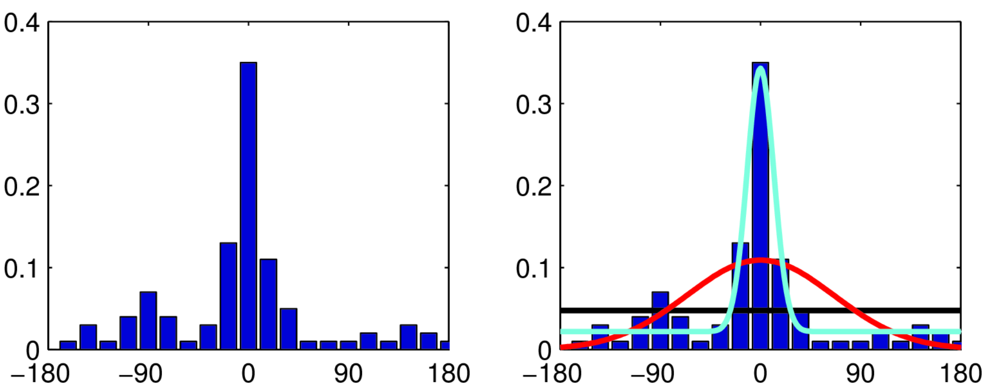



Averaging, uniform prior




**Same logic across our 3 models:**
**M1:** the responses are random (= uniform distribution) — No unknown parameter
**M2:** the responses are informed, more or less concentrated around the correct answer (Gaussian distribution with *unknown* variance) — 1 unknown parameter
**M3:** an *unknown* fraction of responses are random, and the other informed (with *unknown* variance) — 2 unknown parameters

# The model evidence allows direct comparison between models

| Characteristics | Simulated data | Best Fit M1, M2, M3 | Model evidence (log scale) | |
|---|---|---|---|---|
| 20% informed responses (σ=0.5) + 80% random guesses |  | | M1 -304<br>M3 -338<br>M2 -1737 | M1 is $e^{34} \sim 10^{14}$ times more likely than M3 |
| 95% informed responses (σ=0.5) + 5% random guesses |  | | M2 -212<br>M3 -235<br>M1 -304 | M2 is $e^{23} \sim 10^{9}$ times more likely than M3 |
| 50% informed responses (σ=0.5) + 50% random guesses |  | | M3 -286<br>M1 -304<br>M2 -871 | M3 is $e^{18} \sim 10^{7}$ times more likely than M1 |

# Model evidence and related concepts for model comparison

$$\frac{p(M1|y)}{p(M2|y)} = \frac{p(y|M1)}{p(y|M2)} \frac{p(M1)}{p(M2)}$$

Known as:
Odds
Posterior odds

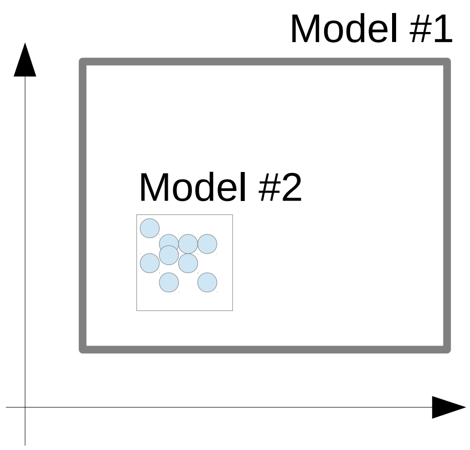
Known as:
Bayes factor
Evidence ratio
Marginal likelihood ratio

Known as:
Prior odds

The model evidence p(y|M) can be difficult to compute exactly. Approximations include:
- The Bayesian Information Criterion
- The Akaike Information Criterion
- The Watanabe-Akaike information Criterion

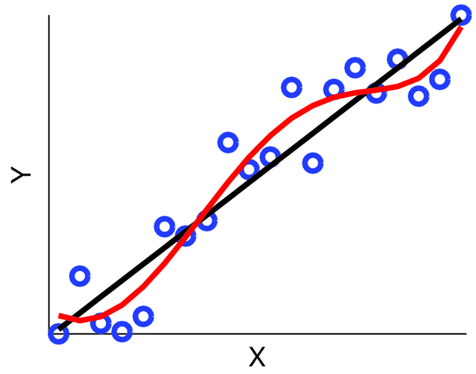
# Simpler is better

The 'simpler is better' preference is also called the **principle of parsimony**, or Ockham's razor




A preference for specific explanations

There is a conflict between the principle of parsimony and the selection of models based on the **maximization of likelihood** (= minimization of errors)



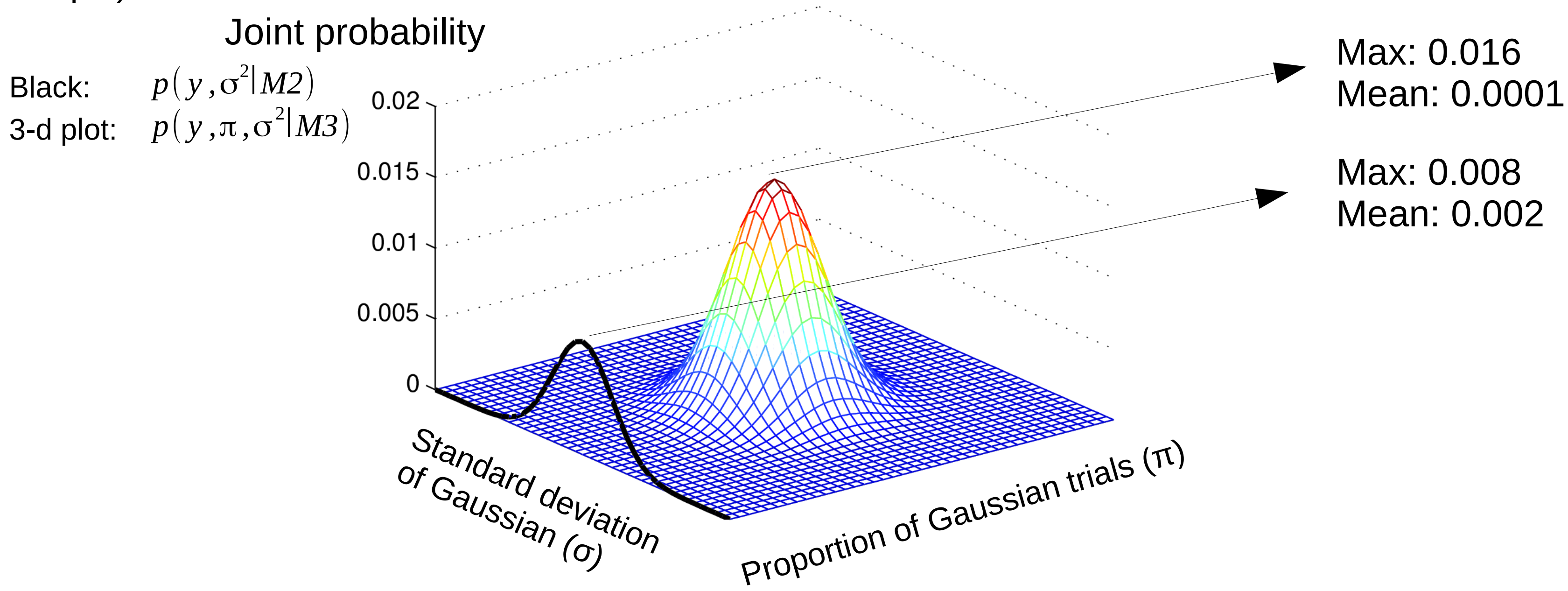

Black fit: $Y = \beta_0 + \beta_1 X + \text{error}$

Red fit: $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \beta_4 X^4 + \beta_5 X^5 + \text{error}$

More free parameters (almost always) ensure a better fit.
→ the criterion of likelihood maximization should be corrected to **penalize complexity**

# Automatic penalization of complexity with the Bayesian approach

(fictitious example)



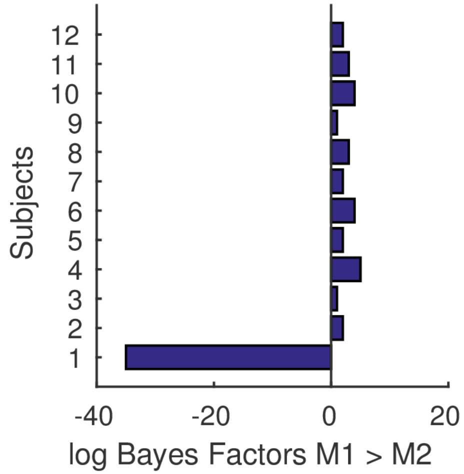

$$p(y|M2)=\int p(y|M2,\sigma^2)\,p(\sigma^2|M2)\,d\sigma$$
$$p(y|M3)=\int\int p(y|M3,\sigma^2,\pi)\,p(\sigma^2,\pi|M3)\,d\sigma\,d\pi$$

Here:
The maximum of the distribution (= maximal accuracy of fit) is larger for the more complex model
The mean of the distribution (= model evidence) is larger for the simpler model.

**Integration over the parameter space penalizes complexity: the model evidence gets 'diluted' in larger parameter space**

# Bayesian inference with subject and group levels: hierarchical models

- Solution 1: a single hierarchical model, with the subject level nested in the group level. Since there is only one model, it provides a group-level Bayes factor
- Solution 2: proceed with 2 steps
  - Fit the data at the subject level and collect model evidence for each subject and each model
  - Perform a group-level analysis.
    - Product of subject-level model evidence = fixed-effect analysis (but may be driven by a single subject)
    - Use a random-effect approach to compute the exceedance probability for each model (= probability that this more is more likely than any other in the general population). See Stephan, NeuroImage 2009.



# Binary hypothesis testing as a particular case of Bayesian model comparison

- The classical t-test
  - H0 (null-model): the mean is exactly 0
  - H1 (alternative model): the mean is different from 0 (and unknown)
- Larger t-values provide evidence to reject the null-model
- The logic seems similar to Bayesian Model Comparison. → See Valentin Wyart's presentation for a worked-out example of 'Bayesian' t-test.

# The advantage of Bayesian over classical hypothesis tests

What the classical **p-value** really is:
*The probability of obtaining a test statistic at least as extreme as the one that was observed, assuming that the null hypothesis is true and the data were generated according to a known sampling plan* (Wagenmakers 2015)

So... smaller p-values indicate stronger evidence that there is an effect?
→ no, they indicate more evidence against the null hypothesis.
So... larger p-value indicates there is no effect?
→ no, they indicate the data are not extreme under the null hypothesis.
Well... p-values quantify some statistical evidence??
→ no. The evidence against the null is over-estimated and the bias increases with the sample size (Wagenmakers 2007)

By contrast, Bayesian statistics:
→ are easier to interpret: 'given my data, it is 100 times more likely that there is an effect rather than no effect'
→ can quantify symmetrically the absence of effect
→ are less biased by sample size
→ can take into account prior knowledge
→ can quantify the plausibility of hypotheses tailored to specific designs.

# Practical recommendations

- For simple use, e.g. t-test, regression... an online tool to compute bayesian statistics: http://pcl.missouri.edu/bayesfactor
- Fit of linear models, existing codes include the Matlab function spm_PEB.m from the SPM toolbox: http://www.fil.ion.ucl.ac.uk/spm/ (this function will estimate the fit of your linear model, and the model evidence for model comparison; also allows hierarchical models)
- More sophisticated models
  - You can make your own codes. Several toolboxes facilitate tricky Bayesian computations, such as Markov Chain Monte Carlo sampling: WinBUG (in R); PyMC (Python); Stan (C++, interface with R, Python, Matlab...); Church (a programming language for probabilistic generative models https://probmods.org)
  - A Matlab toolbox for stochastic models: https://code.google.com/p/mbb-vb-toolbox/

# Selected references

- A graphical illustration of Bayes' rule
  - Puga & Altman, 2015, Nature Method, *Bayes' Theorem*
- A general and very good texbook for basic and advanced Bayesian data analysis:
  - Gelman, Carlin, Stern, Dunson, Vehtari, Rubin, 2014 (Third Edition) *Bayesian Data Analysis*
- Troubles with classical t-tests, and a Bayesian solution
  - Wagenmakers, 2007, Psychonomic Bulletin & Review, *A practical solution to the pervasive problems of p values*
- A variational Bayes approximation of model evidence + group-level analysis
  - Stephan, Penny, Daunizeau, Moran, Friston, 2009, NeuroImage, *Bayesian model selection for group studies*
  - Penny, 2012, NeuroImage, *Comparing Dynamic Causal Models using AIC, BIC and Free Energy*
- Bayesian t-test (companion paper of http://pcl.missouri.edu/bayesfactor)
  - Rouder, Speckman, Sun & Morey, 2009, Psychonomic Bulletin & Review, *Bayesian t-tests for accepting and rejecting the null hypothesis*
- Joshua Tenenbaum & Noah Goodman on-line textbook for probabilistic models (adapted to cognitive science):
  - https://probmods.org