DRAWING INFERENCES FROM MODELS; MODEL COMPARISON

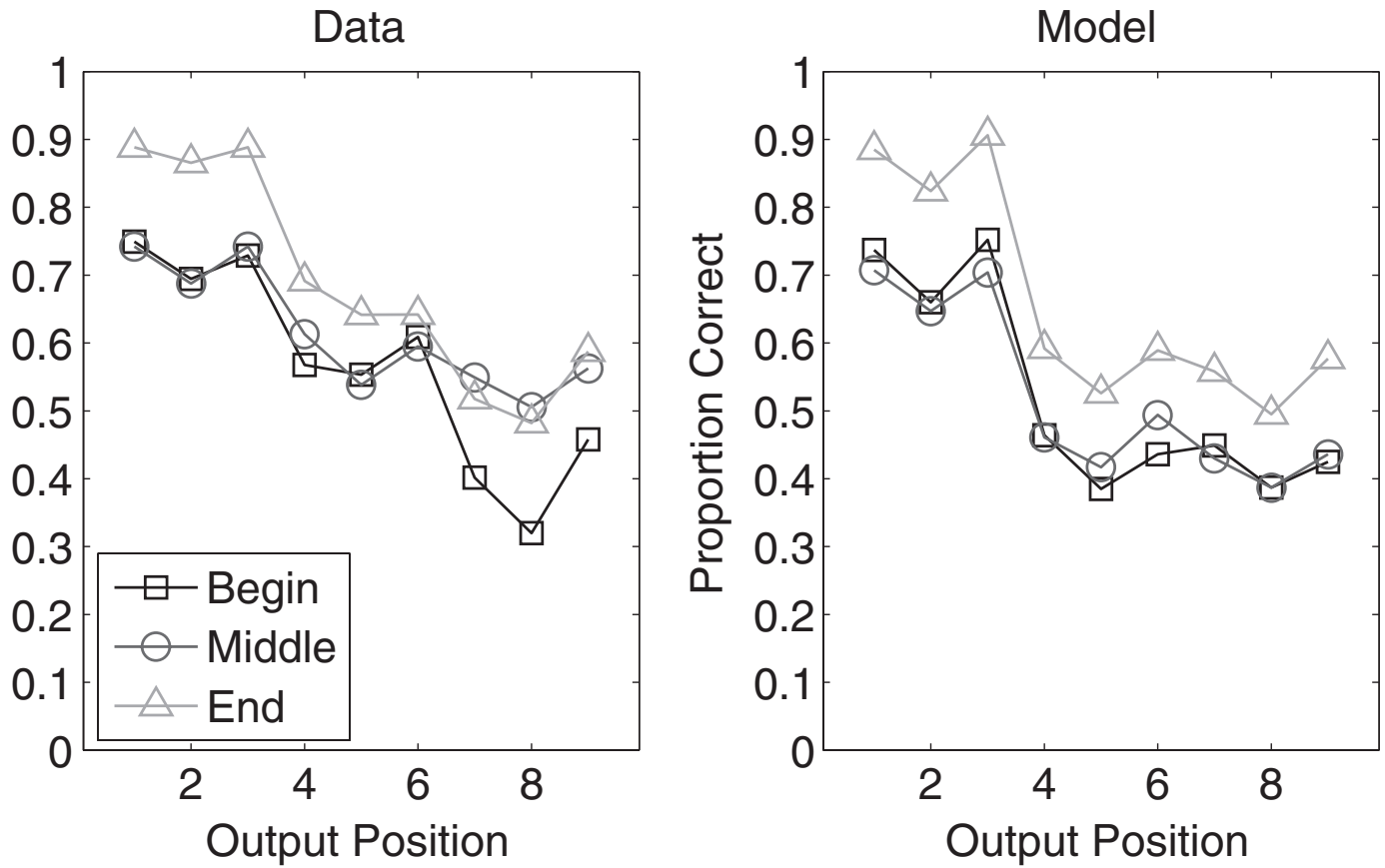
Chris and Cas and Simon

Outline

- Drawing inferences from models
- Comparing models on fit
- Model complexity
- Akaike's Information criterion
- Bayesian Information Criterion
- Priors

DRAWING CONCLUSIONS FROM MODELS

Farrell (2012), Psych Rev



Many publications report only a single model

What does this tell us?

It works (but not much else); *sufficiency*

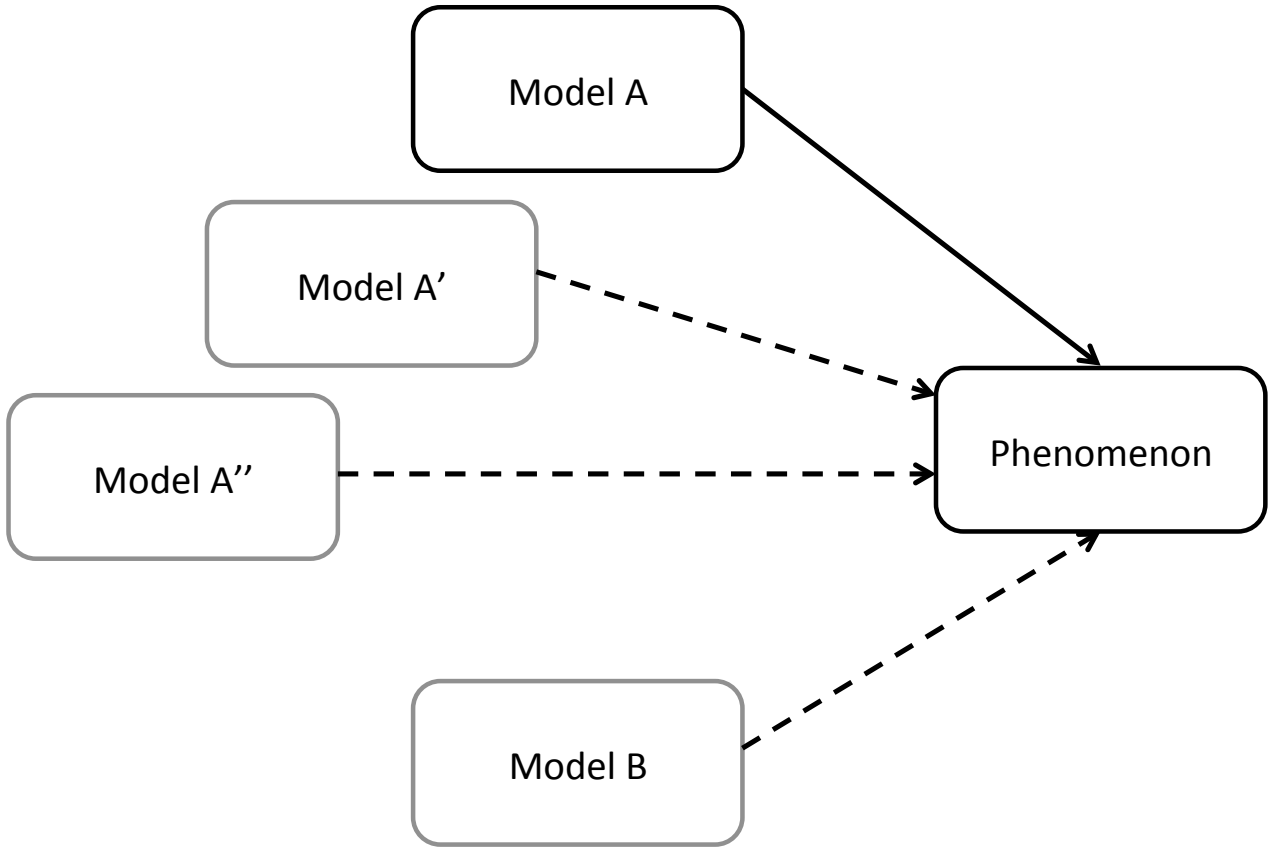
Parameter estimation

- Parameters can be informative
  - Does the accumulation rate in an accumulator model differ between conditions?
  - Do people pay more attention to some dimensions in categorization?
  - Is the value function in prospect theory concave or convex (or linear)?

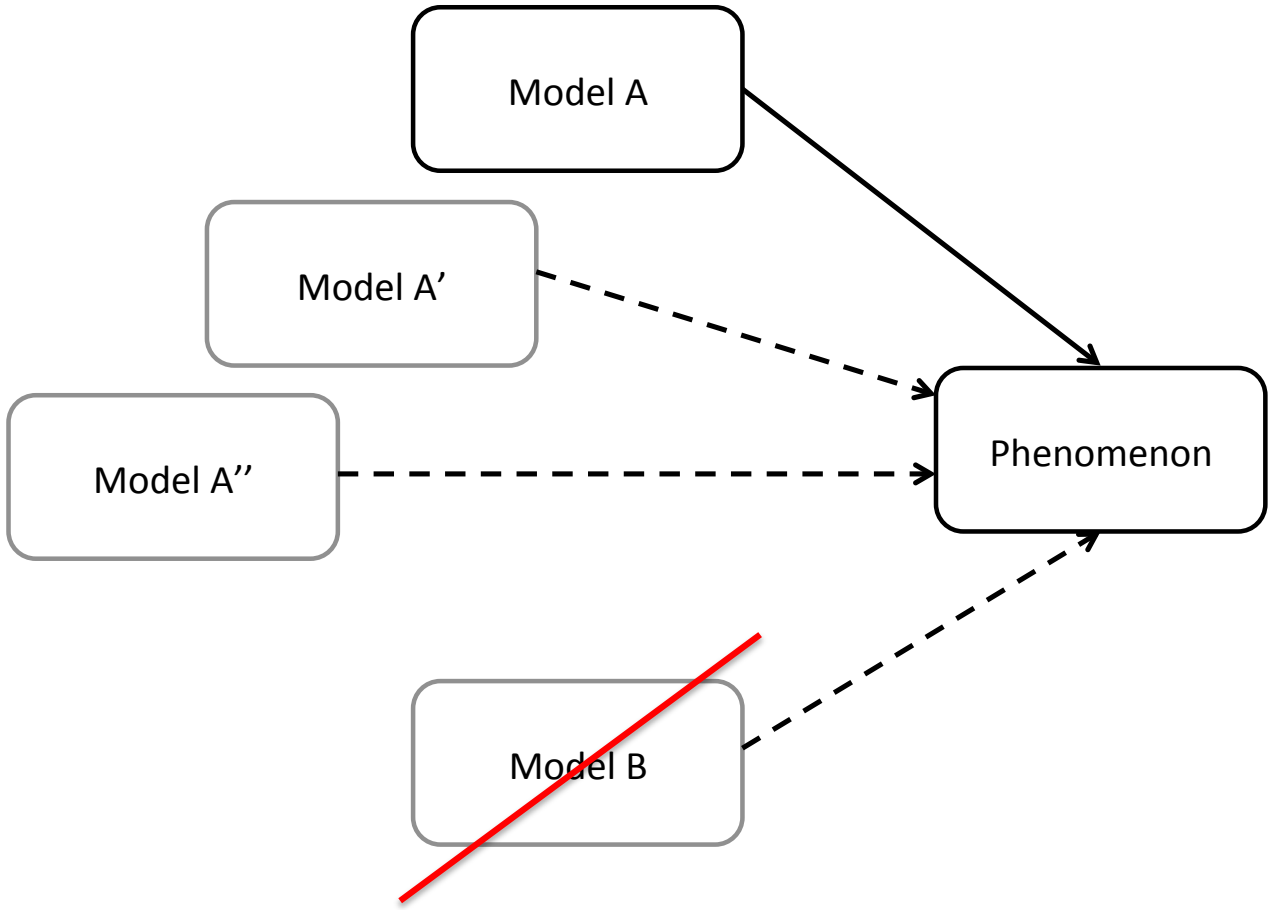
Try and break the model

- Credit assignment
  - Does a mechanism usefully contribute to a model?
  - Does a parameter do what we think it is doing?
- Klaus exploring TBRS*: setting threshold to 0 doesn't actually change the predictions

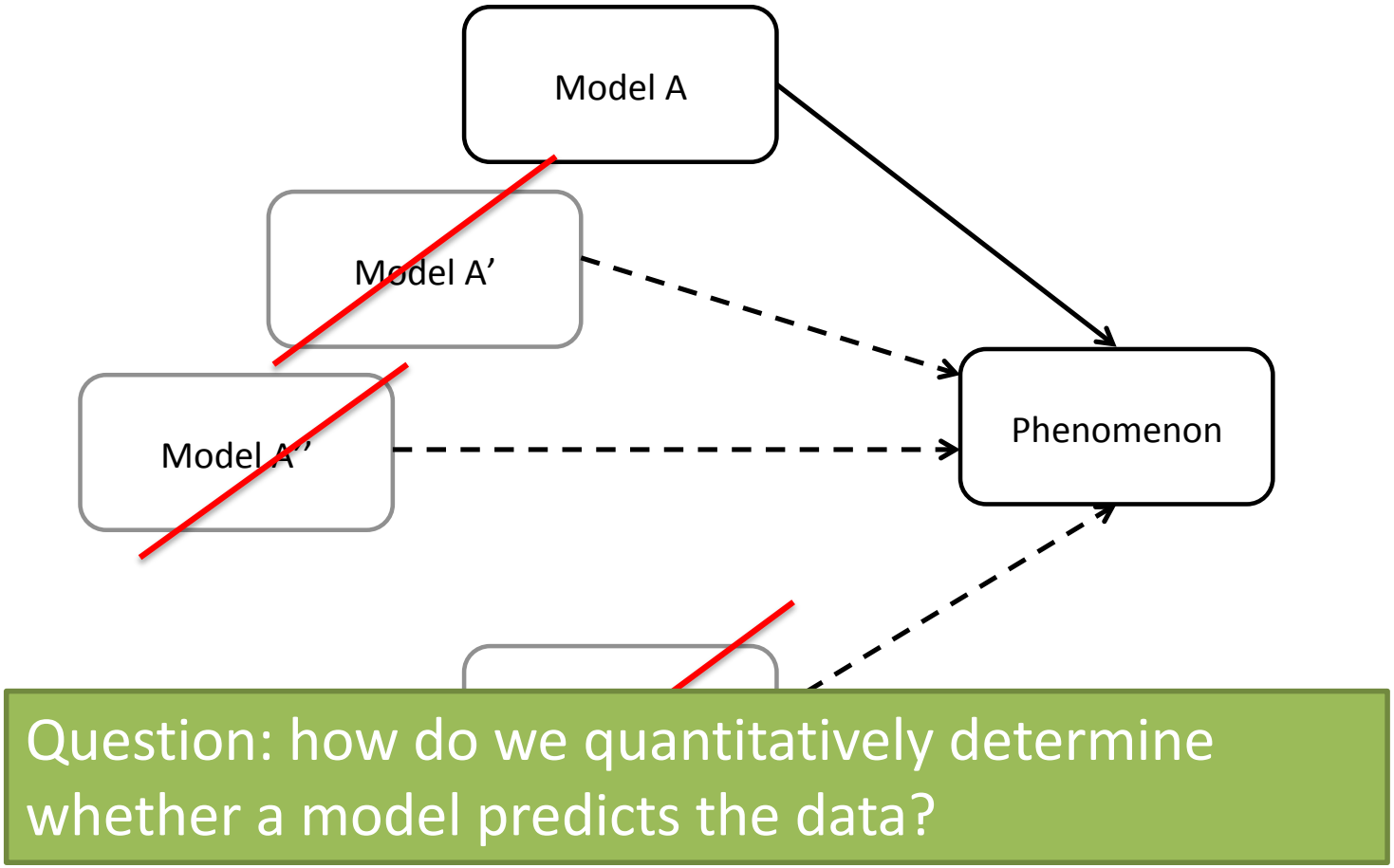
Model comparison



Model comparison



Model comparison



Question: how do we quantitatively determine whether a model predicts the data?

Comparing models based on fit (Cas)

- Let's to back to psychophysics

Polynomial Law of Sensation

SUE DOE NIHM *Chang Ri Law University*

ABSTRACT: *A new theory proposes that sensation grows as a polynomial function of physical intensity. The theory reproduces all of the published data perfectly without error. The degree of the polynomial is independent of whether category ratings or magnitude estimations are used as the dependent variable; it is independent of stimulus range, number of categories, value of the standard, first stimulus, modulus, stimulus spacing, and all other contextual features of the experiment except the number of stimuli. Because the polynomial law always provides a superior fit to the data, it should supersede the logarithmic and power laws of sensation.*

are never fit exactly by the function. There are always systematic deviations. Second, the exponent depends on whether ratings or magnitude estimations are the dependent variable, on the range of the stimuli, on the value of the standard (if any), and on a variety of other experimental conditions. It also depends on the statistical estimation procedures used to fit it. The same data can lead to different exponents, and different data can lead to the same exponent.

Consequently, during my stay in the United

Nimh paper turns out to be very important:

American Psychologist
Volume 32, Issue 9, September 1977, Page 782
The Polynomial Law
Douglas G. Detterman[a] and Stephen K. Reed[b]
[a]Case Western Reserve University, Cleveland, OH, US
[b]Case Western Reserve University, Cleveland, OH, US

We were very impressed with Professor Sue Doe Nihm's (November 1976) polynomial law of sensation, which states that the degree of the polynomial is always one less than the number of stimuli. However, a distinguished visitor to our university, Professor Hoff Witt of the Frohliche Hochschule, has found that the law applies not only to psychophysical data but to psychological data in general. In recognition of Professor Witt's generalization of Nihm's law, we hope other psychologists will join us in referring to their joint contribution as the Nihm-Witt law of just enough numbers. The important implication of this law is, of course, that psychology's promise has been fulfilled. We now have a single law descriptive of all psychological data. The work of Professor Nihm and Professor Witt, as well as our own work, has convinced us that no single psychological law will ever be more powerful.

Exercise

- Implement the polynomial psychophysics law
- Look at NihmTemplate.R

Is the Nihm polynomial model a good model?

- If not, why not?
- Isn’t a model powerful (=good) if it can account for lots of different types of data?

MODEL COMPLEXITY

What does the Sue Doe Nihm example tell us?

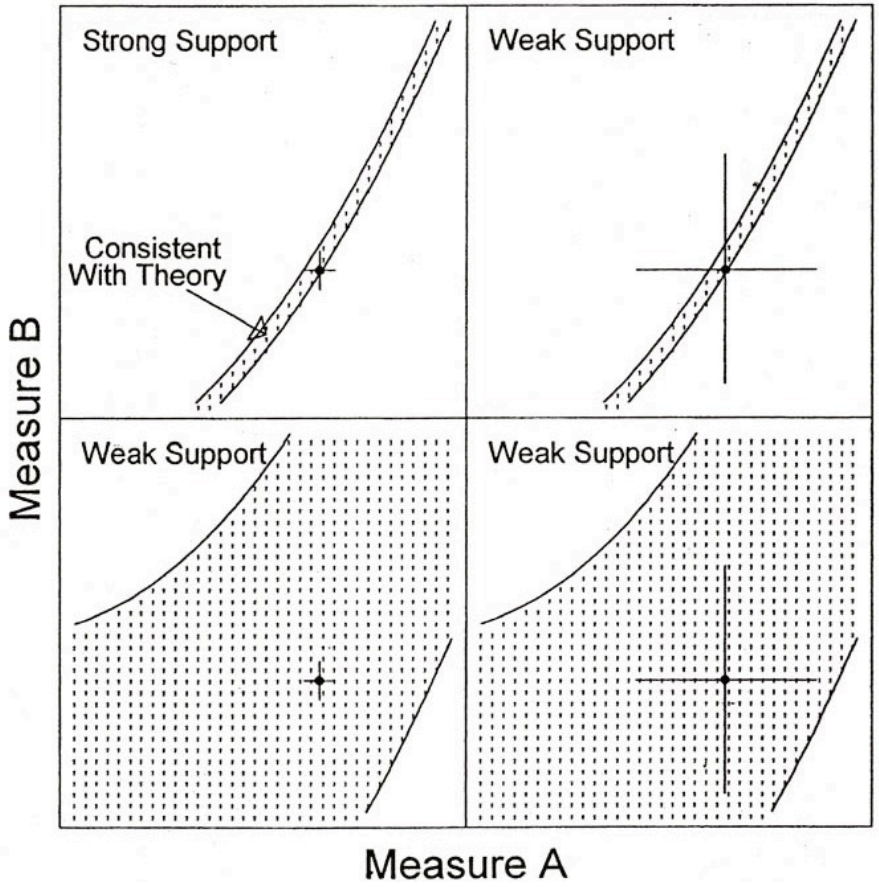
- A powerful theory is not necessarily a good thing
- Powerful = flexible: can fit any data
  - Unfalsifiable
- Tension
  - Parameter estimation: striving for a good fit
  - We don't want **too** good a fit
    - Fit signal and not noise
  - Account for flexibility in models

Isn’t all this a problem for mathematical modelling?

- All this flexibility is hidden in verbal theorizing and reasoning from mental models
  - Submission “Theory X predicts a particular pattern of data. In fact, we observed the reverse pattern”
  - Reviewer 3 (author of Theory X): “Well, actually, theory would be consistent with reverse pattern if we additionally assume...”
- Verbal theories predict qualitative patterns
- We can *quantify* flexibility in models, just like we quantify the fit, using mathematical/computational models

A reminder

Roberts & Pashler (2000)



Where are we going with this?

- Knowing a model accounts for some data tells us something
- Knowing a model accounts for data better (or worse) than other models tells us more
- We need to take into account the flexibility/complexity of models when comparing them—pure fit isn't everything

MODEL COMPLEXITY (MODEL FLEXIBILITY)

What factors determine complexity? (Myung & Pitt, 1997)

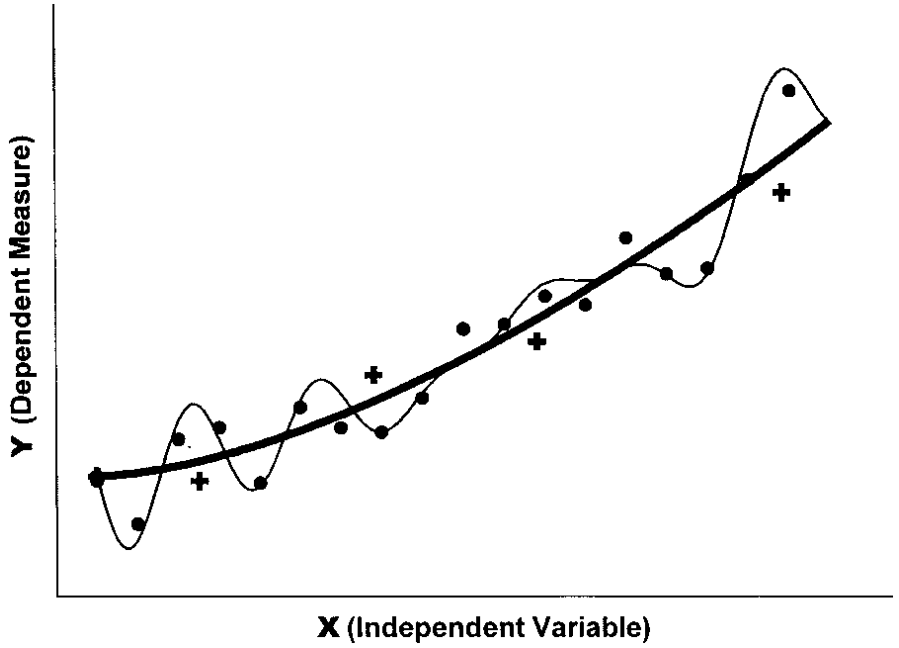
1. Number of parameters
2. Expansion of parameter space
    1. Bounding parameters gives less complex model
3. Functional form of the model
    1. How “wriggly” is the model?
    2. Do the predictions change much as the parameters change?
    3. (Average) curvature of the likelihood surface

Why is complexity so bad?

- Complex models fit noise as well as “true” underlying process
- Complex models will fit data well, but will give poor predictions for other data to which they haven’t been fit

Fitting systematic process vs noise

- Underlying cognitive process is systematic
- Measurement error is noise

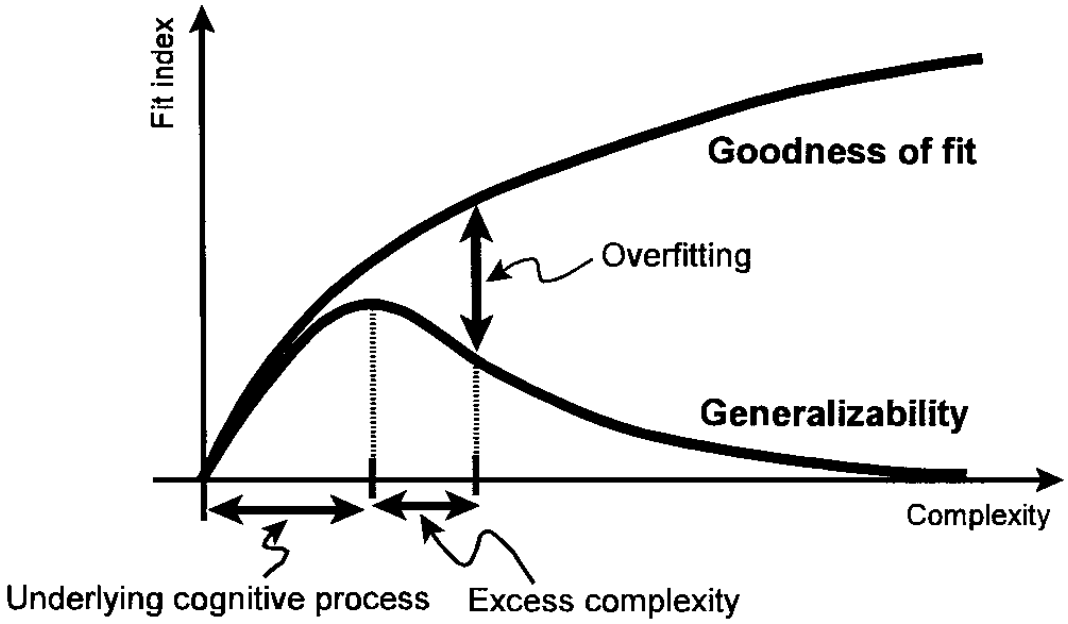


From Pitt, Myung & Zhang (2002)

Exercise

- Data coming from a polynomial
- Your job: find the best model
- Polynomial, or any other model
- Test data set (from known generating process)

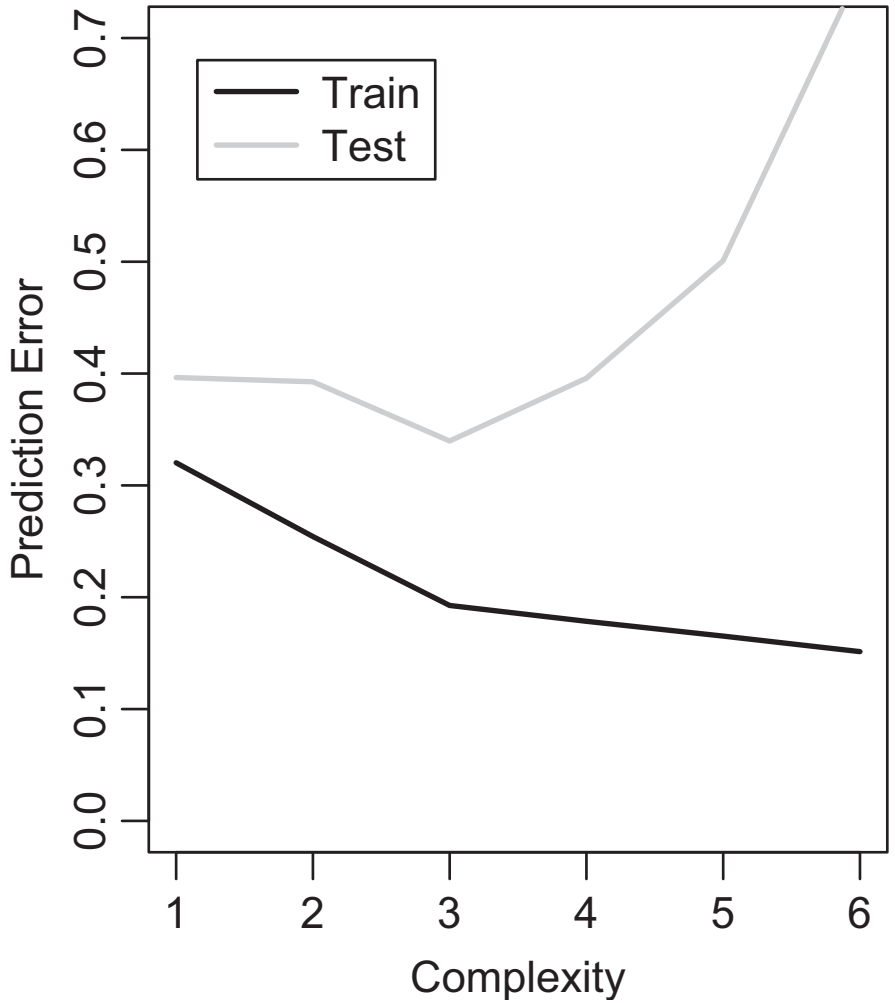
Fitting vs overfitting



Generalizability: How well does the model account for a different data set that wasn't fit?

From Pitt, Myung & Zhang (2002)

Cross-validation: predicting data that haven't been fit



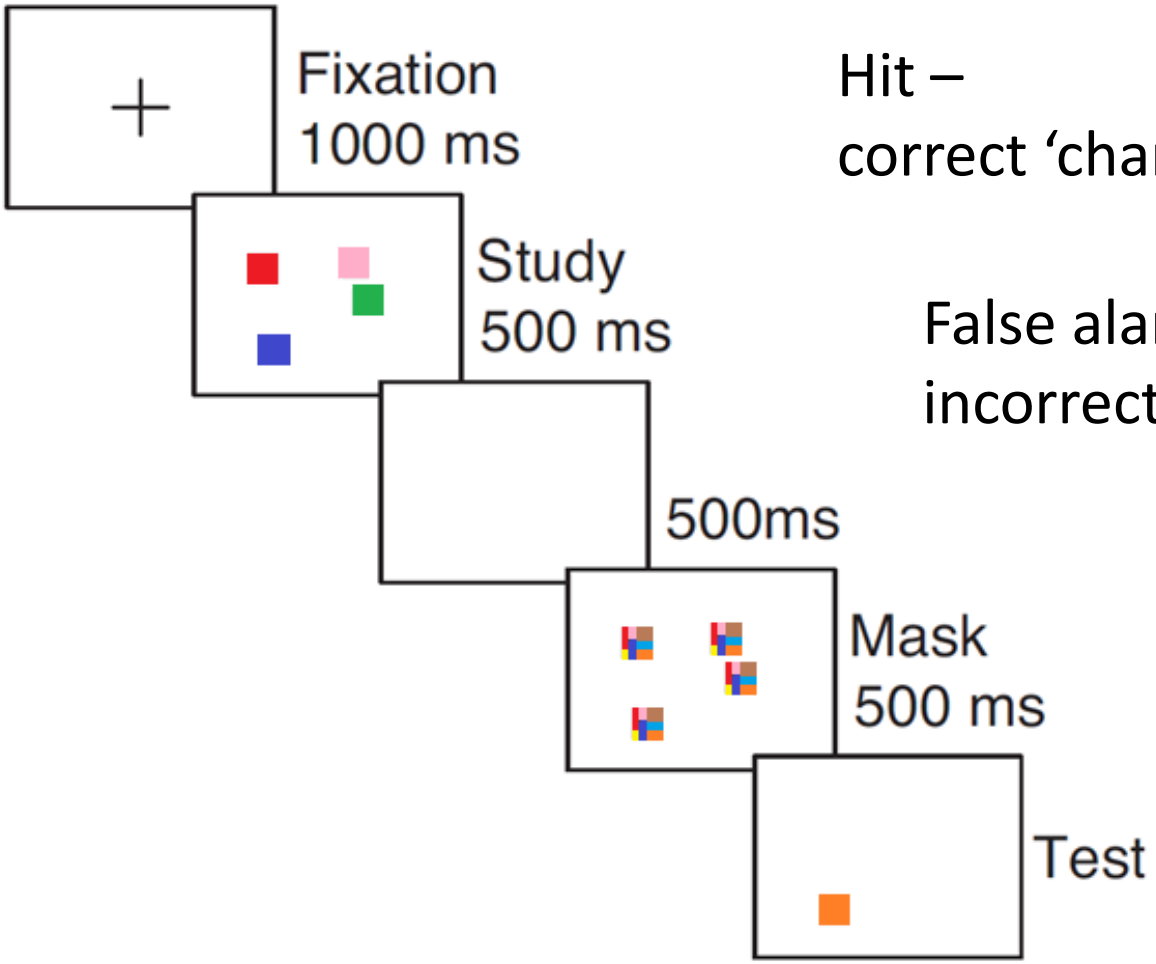
Model comparison needs to take complexity into account

- Likelihood ratio test
- **Information Criteria**
  - Akaike's Information Criterion
  - Bayesian Information Criterion
- Cross-validation
- Minimum description length
  - Normalised maximum likelihood
- Landscaping
- **Bayes Factors**

- Complexity means that we can’t just use log-likelihood (deviance) to compare models
- We need to account for differences between the models in complexity
  - Number of parameters
  - Functional form

INFORMATION CRITERIA (CD)

Change Detection



Hit –
correct ‘change’ response

False alarm –
incorrect ‘change’ response

Experiment

- Data from change detection experiment
- Set size 4
  - k = 35 change responses out of 40 change trials
  - k = 5 change responses out of of 40 same trials
- Set size 8
  - k = 25 change responses out of 40 change trials
  - k = 15 change responses out of 40 same trials

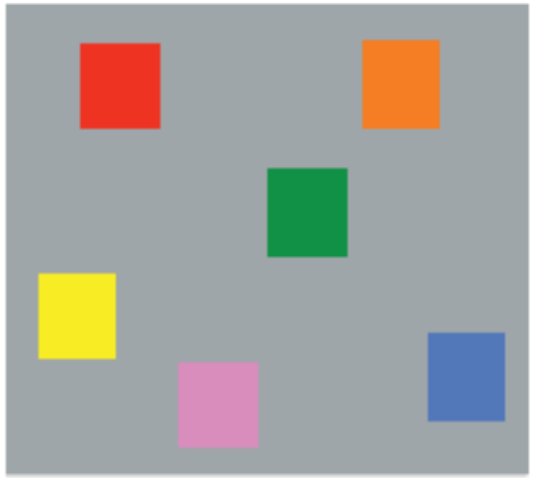
Data-generating Process

- Model change/same responses using Binomial distribution
- Success/failure on N trials with a fixed probability of success

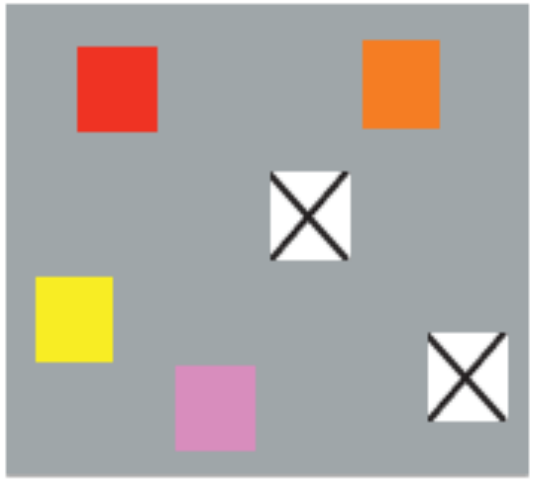
$$\binom{N}{k} p^k (1-p)^{N-k}$$

- Probability generated by cognitive model

You see this

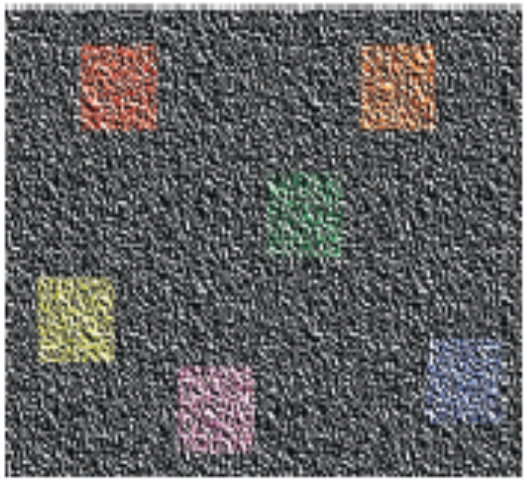


a. Do you remember this?



Slots

b. Or this?



Resources

slots

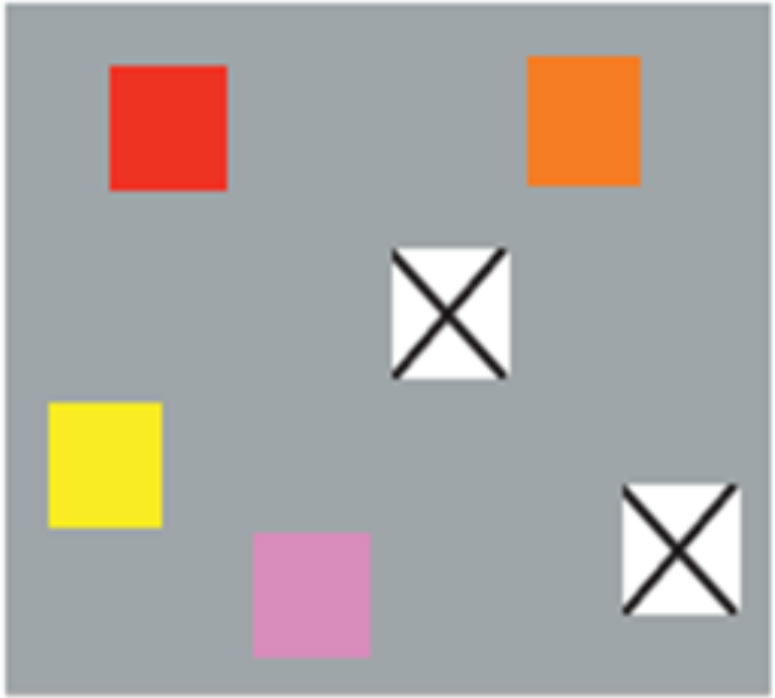
$$d = {}^{k}/_{N}$$

Hit

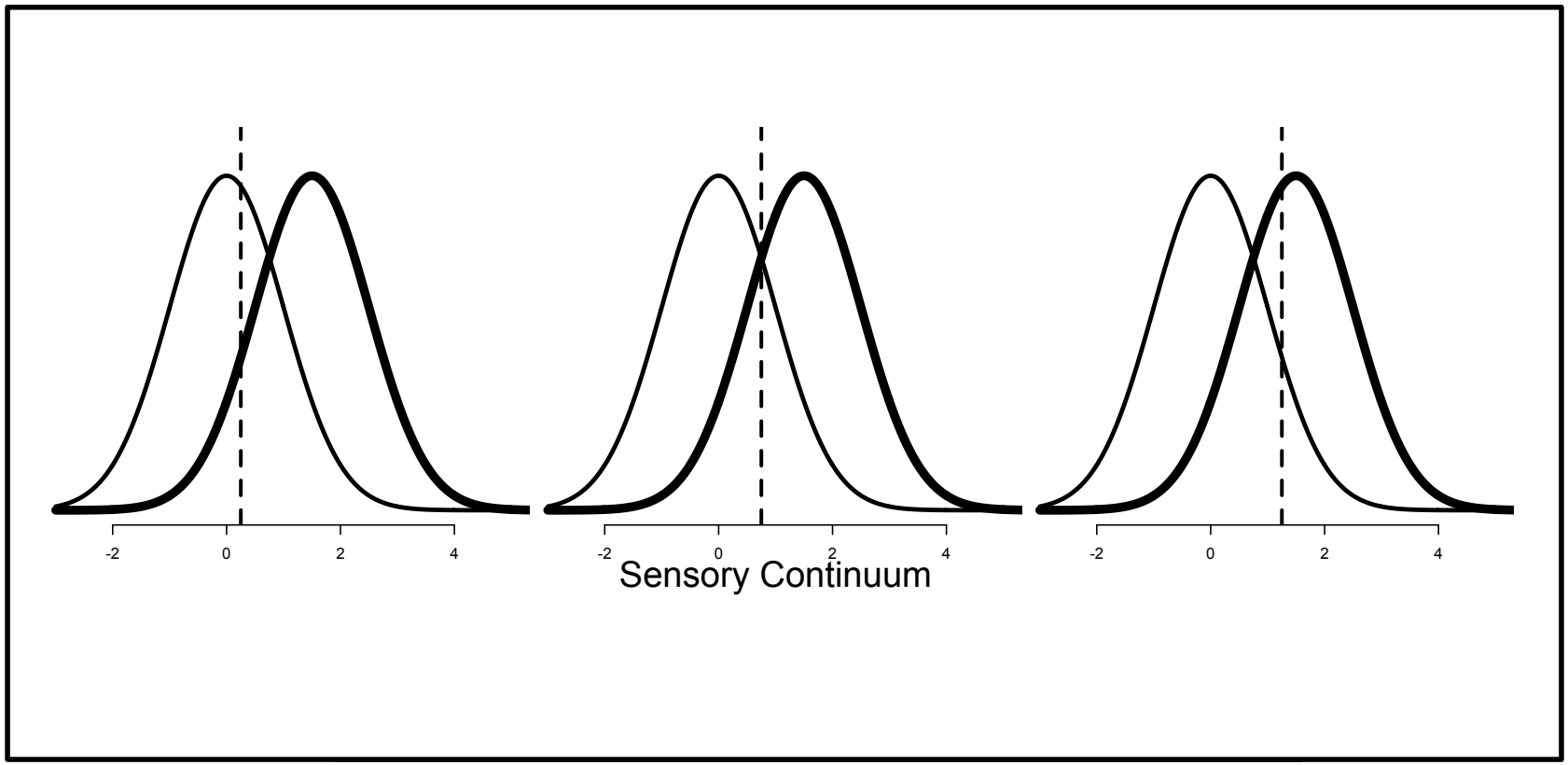
$$d + (1 - d)g$$

False Alarm

$$(1 - d)g$$



resource



PREDICTIONS

slots

$$p(h_i) = d_i + (1 - d_i)g$$

$$p(f_i) = (1 - d_i)g$$

resource

$$p(h_i) = \Phi\left(\frac{d'_i}{2} - \frac{\log\beta}{d'_i}\right)$$

$$p(f_i) = \Phi\left(-\frac{d'_i}{2} - \frac{\log\beta}{d'_i}\right)$$

Exercise

- Get maximum-likelihood parameter estimates for slots model

AKAIKE’S INFORMATION CRITERION

Likelihoods and information (SF)

- Deviance (-2 ln L) is a measure of the fit of the model to data
- Also a measure of the information lost between the data and the model
- Kullback-Leibler measure of information

Discrete

$$KL = \sum_{i=1}^{I} p_i \log \frac{p_i}{\pi_i}$$

Continuous

$$KL = \int R(x) \log \frac{R(x)}{p(x|\theta)} dx$$

$$KL = \int R(x) \log R(x) dx - \int R(x) \log p(x|\theta) dx$$

Information lost when going from Reality to the model

Information in Reality

Information shared between model and Reality

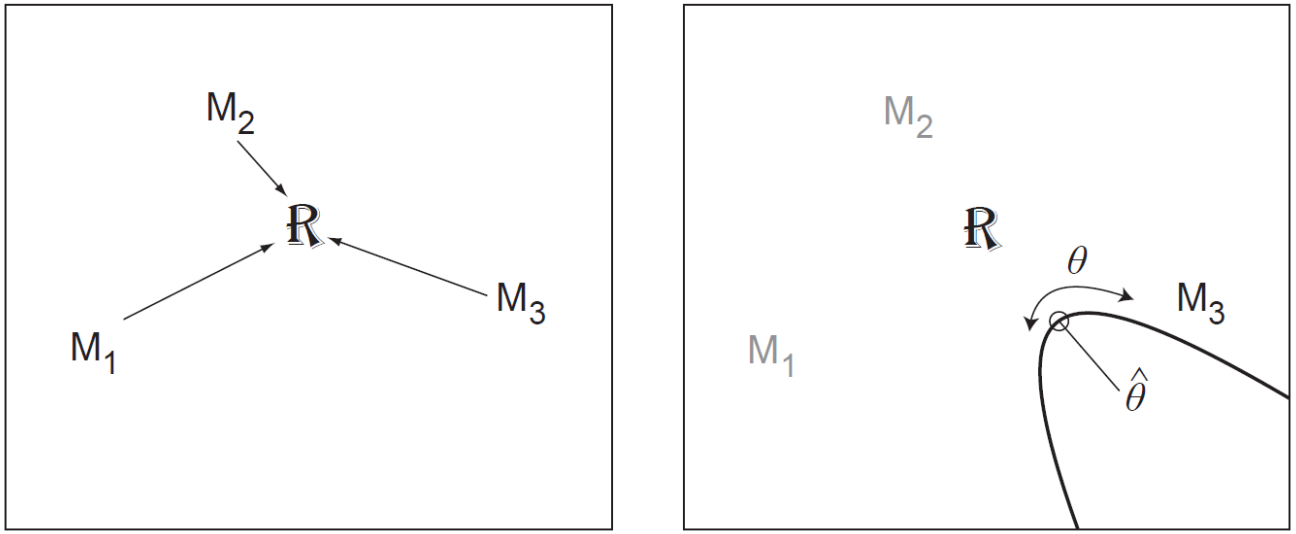
Kullback-Leibler divergence

$$KL = \int R(x) \log R(x) dx - \int R(x) \log p(x|\theta) dx$$

- R(x) is Reality: a constant
- Only thing that can change while modelling data is log *p* (*x*|*theta*)
  - Log likelihood!
- Minimize Kullback-Leibler divergence by maximizing log-likelihood

Akaike (e.g., 1974)

- Log-likelihood is a biased measure of KL divergence
- Same set of data used to estimate parameters and calculate log-likelihood
  - Fit noise as well as systematic effects (Reality)



Information space

Akaike's Information Criterion (AIC) (CD)

- Correct log-likelihood (as measure of KL divergence) for bias
- What does it look like?
- AIC = $-2 \ln L + 2K$
  - Where K is the number of parameters
- AIC = fit + parsimony
  - $2K$ punishes more complicated models (those with more parameters)

Exercise: Compare Models

- Slot and resource models both fit the data
- Use AIC to decide which gives most parsimonious solution

Exercise

- We have fit Model A and Model B to five participants individuals
  - Deviance (model A) = c(100,200,300,400,500)
  - Deviance (model B) = c(50,100,300,400,600)
- Model A has 2 free parameters per participant
- Model B has 3 free parameters per participant
- Calculate AIC for each individual, for each model
- Advanced: calculate total AIC for all participants

AICc: Corrected AIC

- AIC turns out to be a bit rubbish for small samples (ie where N/K is small)
- AICc corrects AIC for small samples

$$AICc = AIC + \frac{2K(K+1)}{n-K-1}$$

AICc behaves like AIC for large samples

Bayesian Information Criterion

- BIC = -2 *ln* L + K *ln* N

  K: number of parameters
  N: number of data points

- Looks like AIC: fit + penalty for complexity

Exercise

- Calculate BIC for slot and resource models
  - What is N?
- Which is the best model?

What is N?

$$\binom{N}{k} p^k (1-p)^{N-k}$$

A few notes about AIC/BIC

- On a log scale: difference of, e.g., 2 just as big for 2 vs 4 as for 1002 vs 1004
- Similar in form, difference in penalty term
  - AIC: $2K$
  - BIC: $\ln(N)K$
  - BIC more punishing (prefers simplicity) for $\log(N)>2$ (around $N=8$)
  - Difference in theoretical interpretation, but used interchangeably in literature to compare models

AIC vs BIC: which to use?

- Report both (?)
- BIC more useful than AIC for nested models
- Other comparisons (see Kuha, 2004; Wagenmakers & Farrell, 2004)
- What interpretation do you want to take?

1 min contemplation

- AIC (AICc) and BIC correct for model complexity
- How do they fall short in doing so? What aspects of complexity do they miss?

1 min contemplation

- AIC (AICc) and BIC correct for model complexity
- How do they fall short in doing so? What aspects of complexity do they miss?
- Extension of parameter space
- **Complexity arising from functional form**

(Brief mention): Minimum description length

$$-\ln L + 0.5k \ln\left(\frac{n}{2\pi}\right) + \ln \int \sqrt{\det(I(\theta))}\, d\theta$$

- Rissanen (1996); see Pitt & Myung (2002) for a quick intro
- Corrects for complexity in functional form of the model
- Complexity = curvature of likelihood surface
  - More peaked/curvy = more complex

Bayesian model comparison with Bayes Factors

Baby Bayes

Hypothesis Testing

- Two hypotheses, H_0 and H_1
- Observe some data D
- Under which hypothesis is the data more likely?
- Want to know which is larger
  - $p(D|H_0)$ or $p(D|H_1)$
- A Bayes factor is the ratio

$$BF_{1.0} = \frac{p(D|H_1)}{p(D|H_0)}$$

Bayes Factor

- Relative probability of the data under two hypotheses
  - $BF_{10} = 2$ means the data are twice as likely under H_1 than H_0.
  - $BF_{10} = 0.5$ says the data are twice as likely under H_0 than H_1.
  - $BF_{10} = 1$ tells us the data are equally likely under both hypotheses.

Interpretation of Bayes Factors

- $1 < BF < 3$ "Barely worth mentioning"
- $3 < BF < 10$ "Substantial"
- $BF > 10$ "Strong"

ESP

- Have to predict whether a coin toss will yield a heads or tails
- Do this for 20 trials
- Crazy people predict that people can do this with probability > 0.5.

Bayesian Hypothesis Test

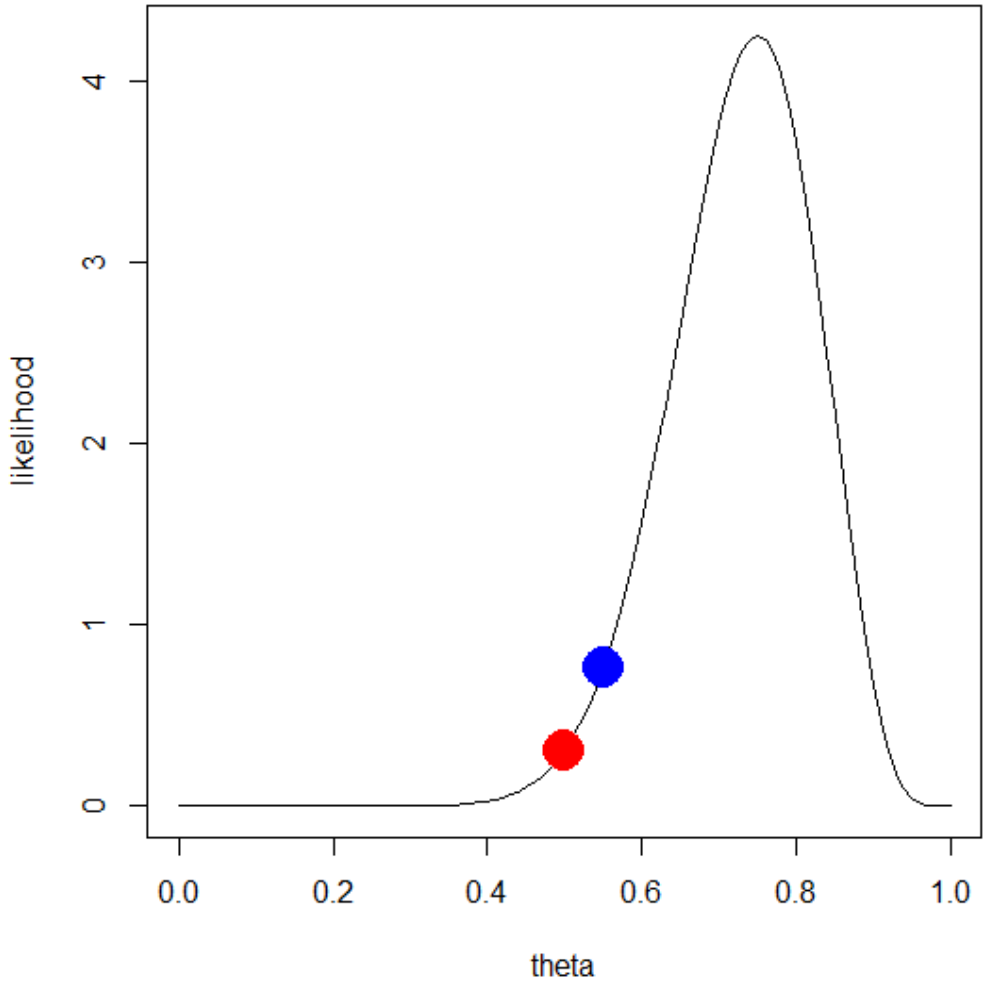
- Back to our example with θ
- Clear null hypothesis of $\theta = 0.5$
  - H_0: $\theta = 0.5$
- What is the alternative?
  - H_A: $\theta = 0.55$

Likelihood – 15 correct

BF=2.5

H_0: $\theta = 0.5$

H_A: $\theta = 0.55$

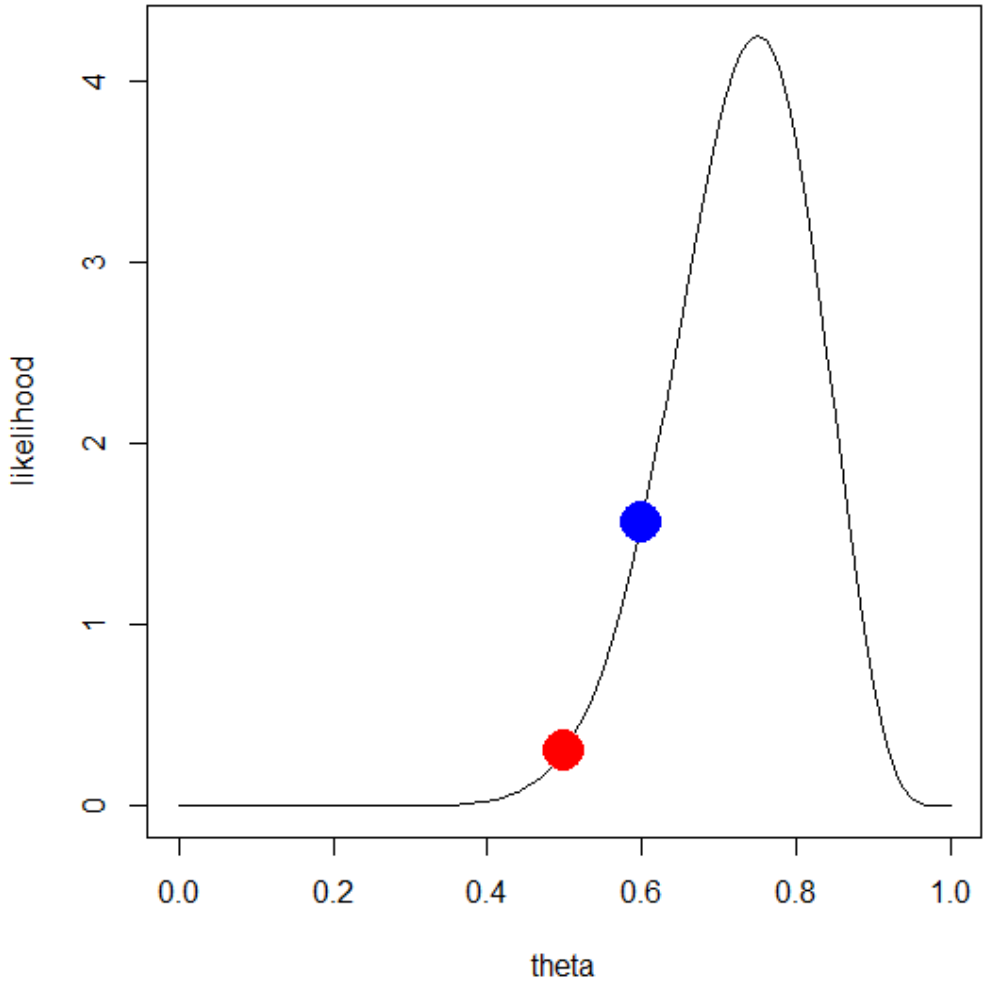


Likelihood – 15 correct

H_0: $\theta = 0.5$

H_A: $\theta = 0.6$

BF=5

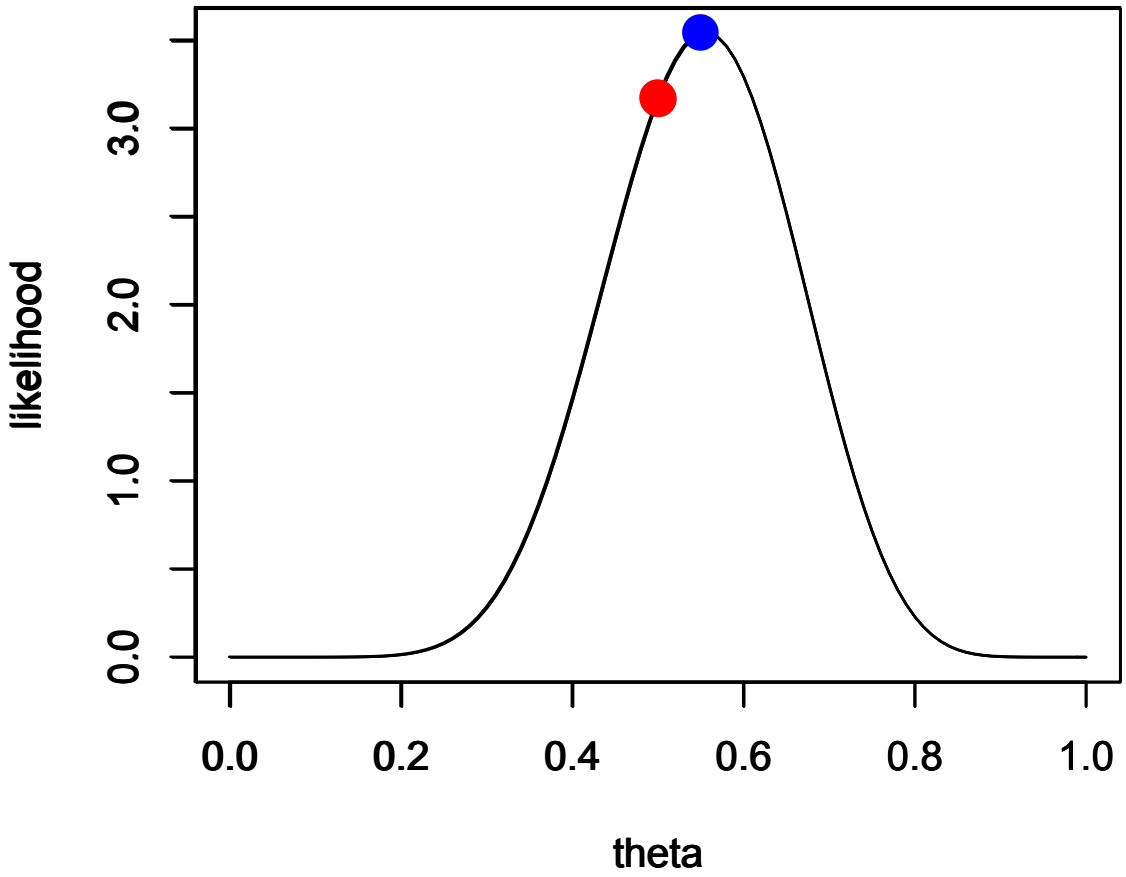


Likelihood – 11 correct

H_0: $\theta = 0.5$

H_A: $\theta = 0.55$

BF=1.11

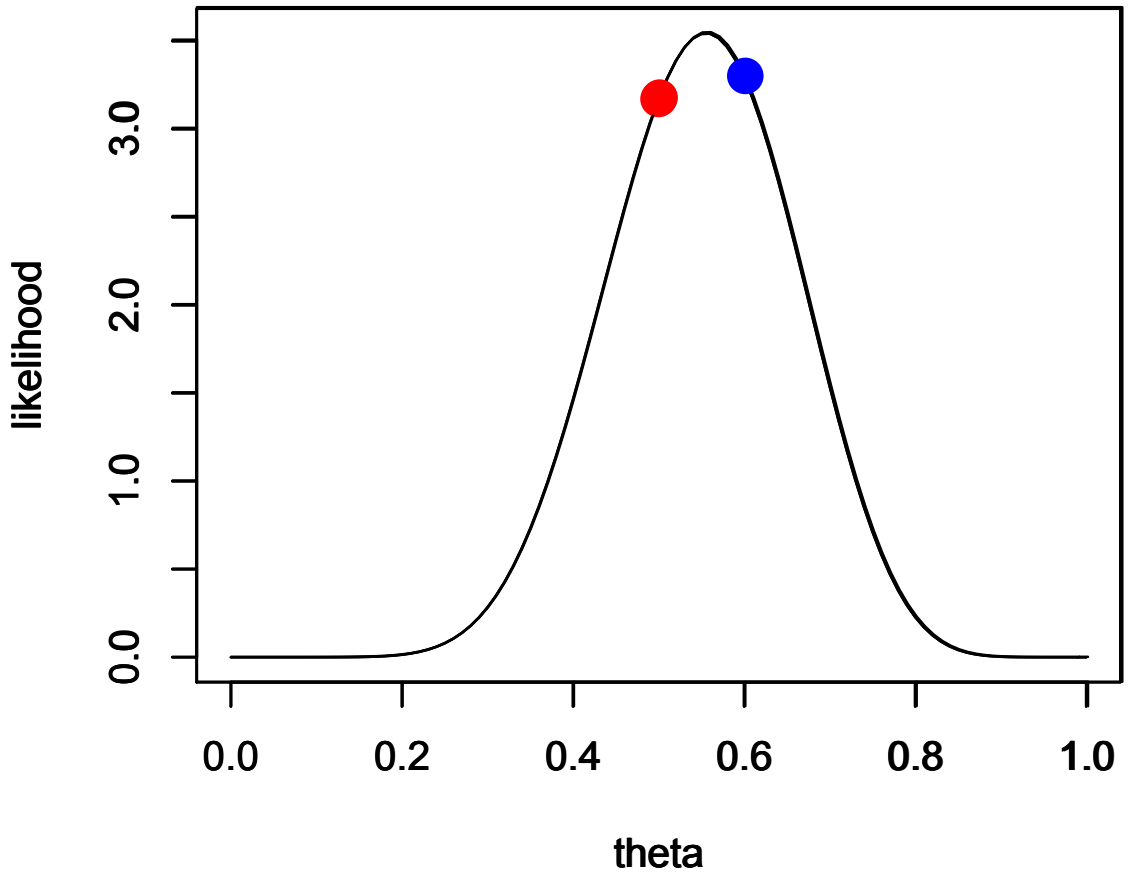


Likelihood – 11 correct

BF=1.04

H_0: $\theta = 0.5$

H_A: $\theta = 0.6$



Alternative Hypothesis

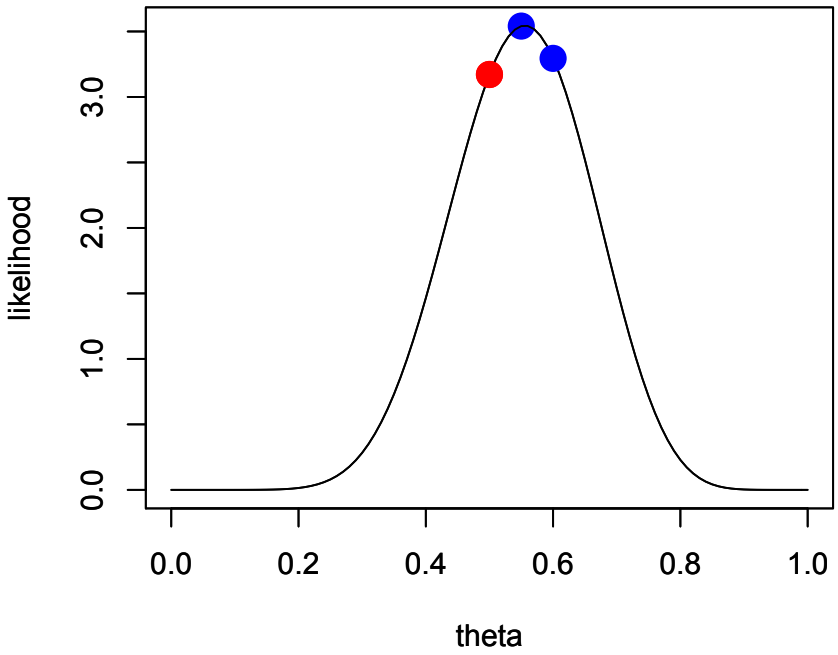
- What if we didn't know which to expect?

  H_A: $\theta = 0.55$

  H_A: $\theta = 0.6$

- Could say that each of the hypotheses could be true, with equal probability

BF = 1.08



Alternative Hypothesis

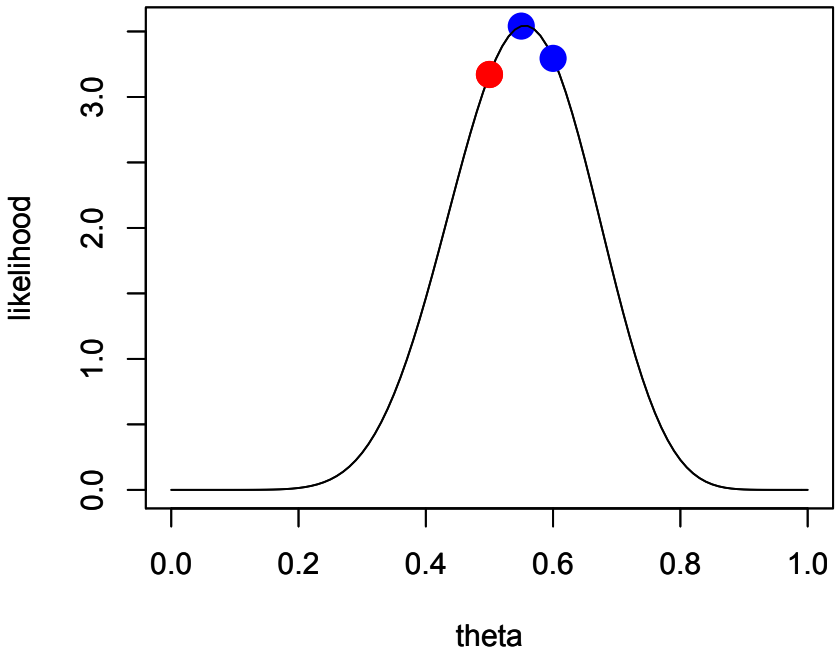
- What if we didn't know which to expect?

  H_A: $\theta = 0.55$

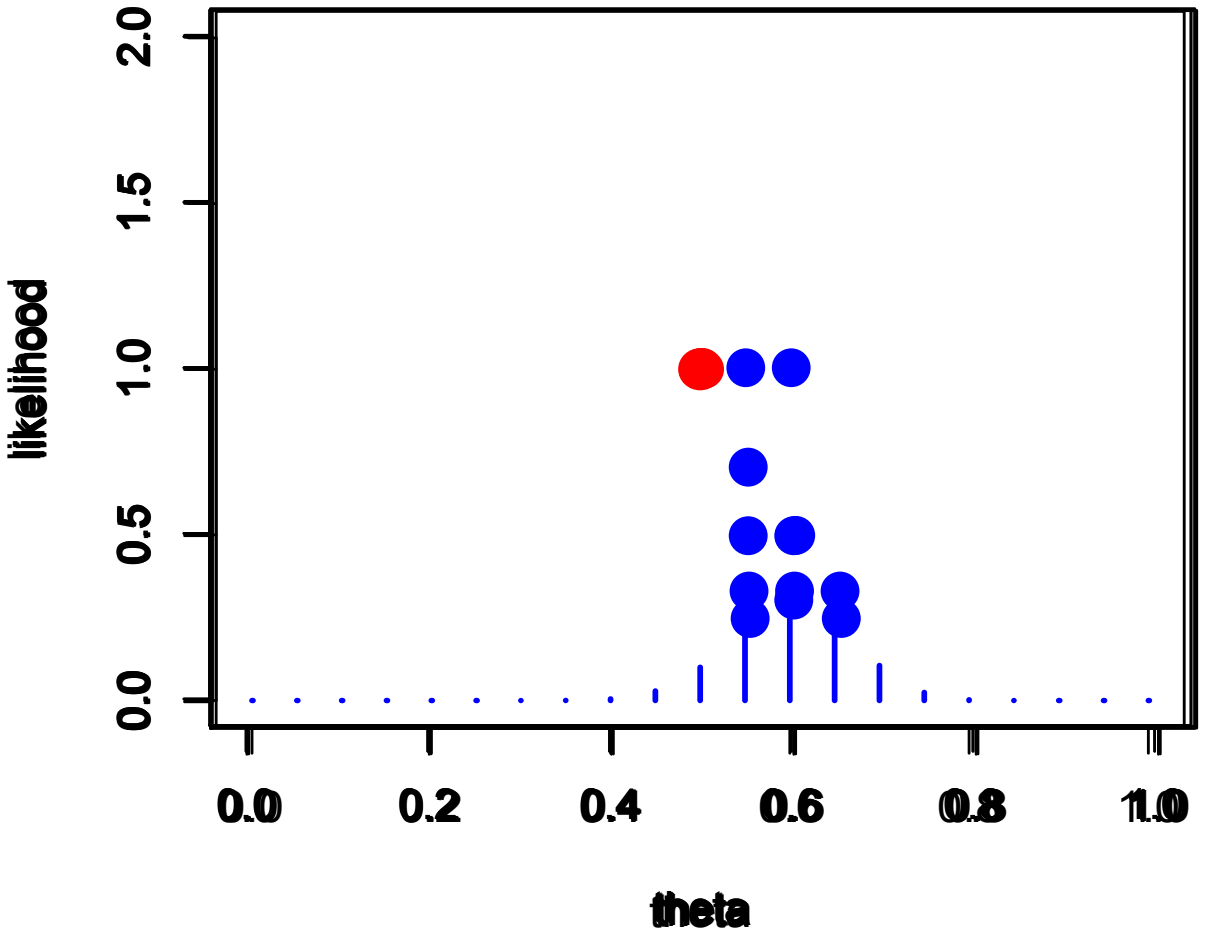
  H_A: $\theta = 0.6$

- Could say that each of the hypotheses could be true, but $\theta = 0.55$ is 70% likely, and $\theta = 0.6$ is 30% likely.

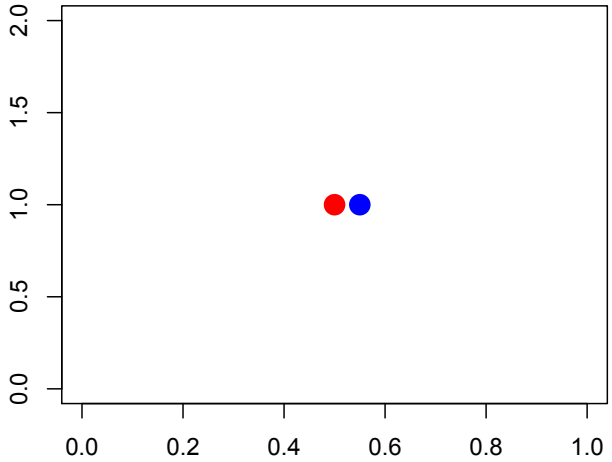
BF = 1.09

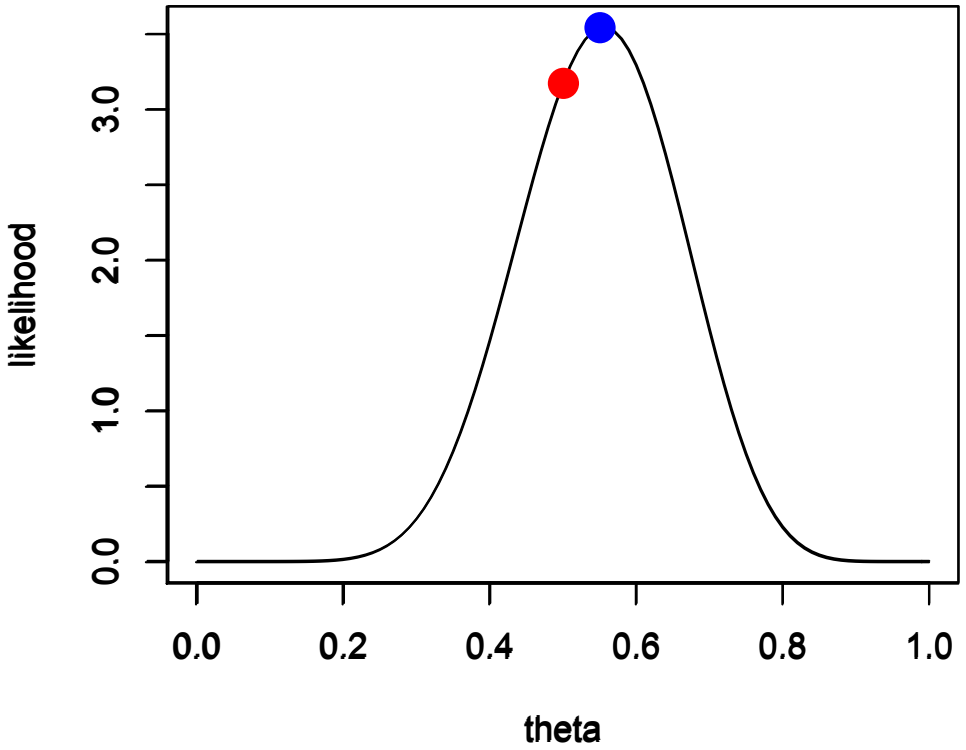


Alternative Hypothesis



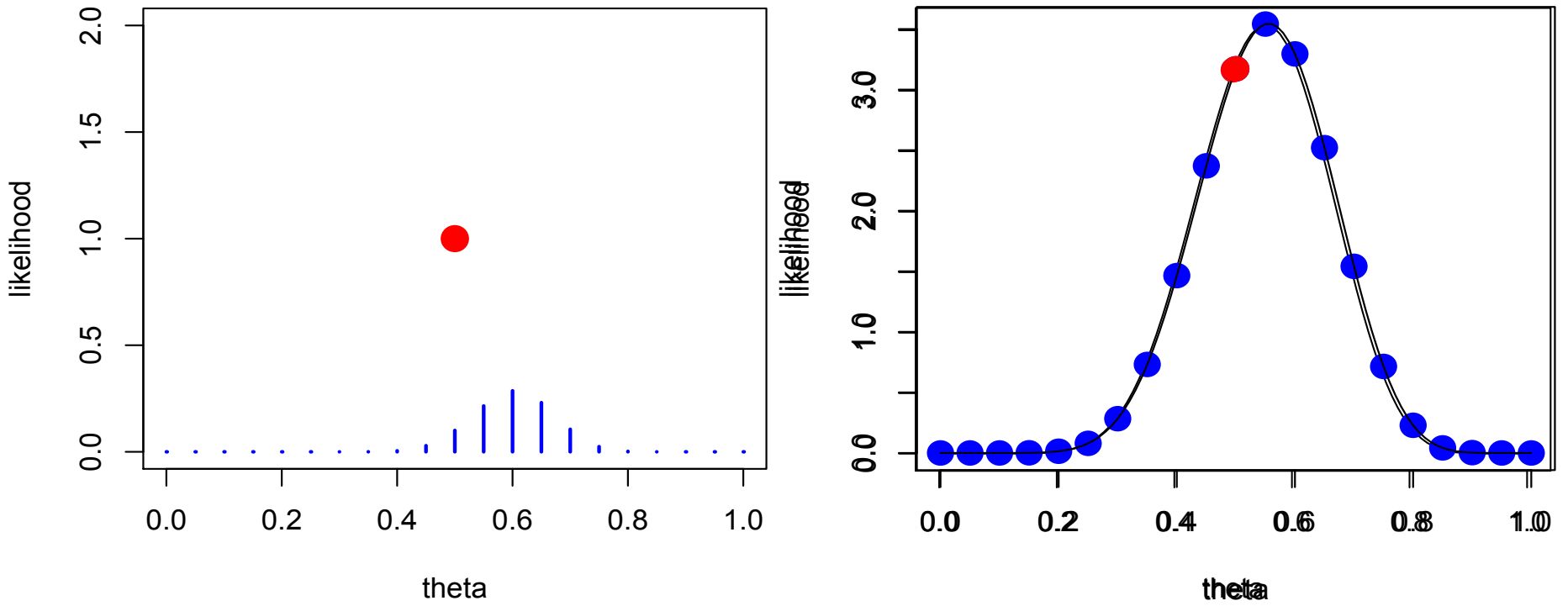
Bayesian Hypothesis Testing





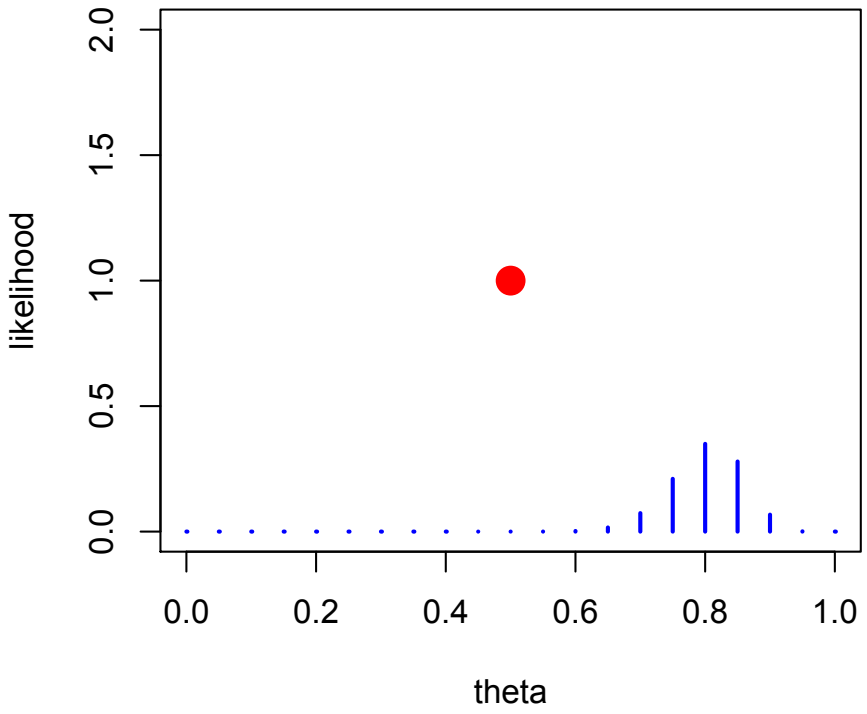
$p(D|H_0) = 3.36$ $p(D|H_A) = 3.72$ $BF = 3.72/3.36$

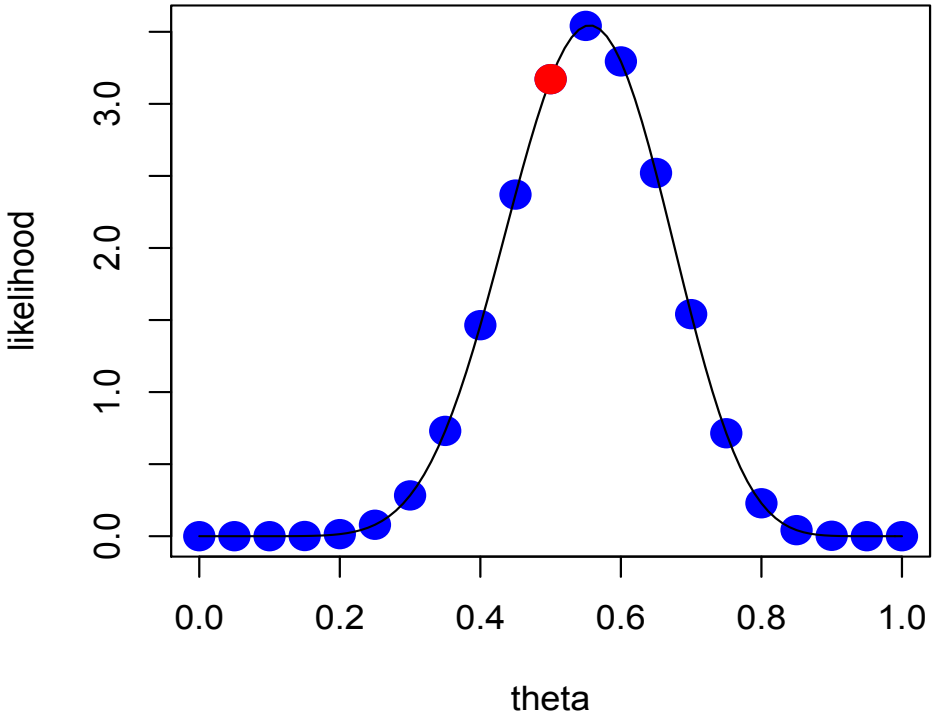
Bayesian Hypothesis Testing



$p(D|H_0) = 3.36$ $p(D|H_A) = 2.9$ $BF = 2.9/3.36$

Bayesian Hypothesis Testing





$p(D|H_0) = 3.36$ $p(D|H_A) = 0.33$ $BF = 0.33/3.36$

Grown-up Bayes (SF)

$$\underbrace{P(\boldsymbol{\theta}|y)}_{posterior} = \underbrace{(P(y|\boldsymbol{\theta})}_{likelihood} \times \underbrace{P(\boldsymbol{\theta}))}_{prior} / \underbrace{P(y)}_{evidence} .$$

$$\underbrace{P(\boldsymbol{\theta}|y, M)}_{posterior} = \underbrace{(P(y|\boldsymbol{\theta}, M)}_{likelihood} \times \underbrace{P(\boldsymbol{\theta}|M))}_{prior} / \underbrace{P(y|M)}_{evidence}.$$

Evidence is the probability of the data given the model
We can use this to compare models: under which model are the data more probable?

Marginal likelihood (evidence)

$$p(y|M) = \int p(y|\boldsymbol{\theta}, M)p(\boldsymbol{\theta}|M)d\boldsymbol{\theta}$$

This is the *average* likelihood (in contrast to maximum likelihood)

Weighted average---weighted by prior

We can't "cheat" by maximising, look across entire parameter space

Bayes Factor (e.g., Kass &Raftery, 1995)

$$BF_{ij} = \frac{p(y|M_i)}{p(y|M_j)} = \frac{\int p(y|\boldsymbol{\theta}, M_i)p(\boldsymbol{\theta}|M_i)d\boldsymbol{\theta}}{\int p(y|\boldsymbol{\theta}, M_j)p(\boldsymbol{\theta}|M_j)d\boldsymbol{\theta}}.$$

Relative evidence for one model over the other

How much more likely are the data under one model rather than the other

This is not the posterior probability of the models

Bayes Factor (e.g., Kass &Raftery, 1995)

$$BF_{ij} = \frac{p(y|M_i)}{p(y|M_j)} = \frac{\int p(y|\boldsymbol{\theta}, M_i)p(\boldsymbol{\theta}|M_i)d\boldsymbol{\theta}}{\int p(y|\boldsymbol{\theta}, M_j)p(\boldsymbol{\theta}|M_j)d\boldsymbol{\theta}}.$$

Yuck!

Relative evidence for one model over the other

How much more likely are the data under one model rather than the other

This is not the posterior probability of the models

Calculating marginal likelihoods

(methods discussed in book)

- Approximation
  - Laplace approximation
  - Bayesian Information Criterion (BIC)
- **Numerical integration**
- Sampling methods (e.g., from JAGS)
  - Importance sampling
  - Savage-Dickey ratio
  - Trans-dimensional MCMC

NUMERICAL METHODS

Back to marginal likelihoods

$$p(y|M) = \int p(y|\boldsymbol{\theta}, M)p(\boldsymbol{\theta}|M)d\boldsymbol{\theta}$$

This is the *average* likelihood (in contrast to maximum likelihood)

Weighted average---weighted by prior

Let's do this in R

$$p(y) = \iint p(y|k,g)p(k)p(g)\,dk\,dg$$

- Question: what are we assuming about the priors here?

We need priors on k and g (CD)

$$k \sim \text{Uniform}(0,8)$$

$$g \sim \text{Uniform}(0,1)$$

Weighted average

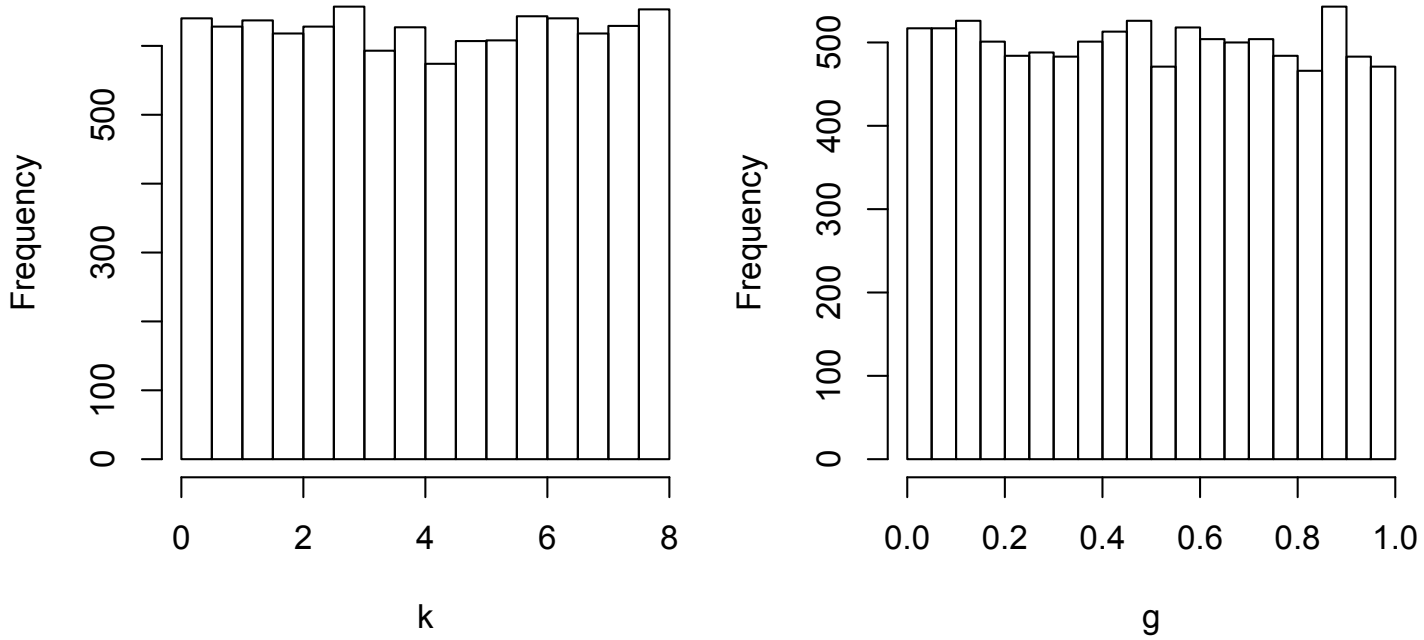
- Average of 5 and 10

$$\frac{5+10}{2}$$

- Weighted average - 70% of 5, 30% of 10
- Dumb method

$$\frac{5+5+5+5+5+5+5+7+7+7}{10}$$

Weighted average by sampling



For pairs of k and g, calculate the likelihood of the observed data

Then, average.

Exercise: Marginal Likehood

- Estimate marginal likelihood for slot model given data
- Sample parameters from prior distributions
  - Assume independence, sample independently
- Plug each sample from the prior into likelihood function to get $p(d|theta)$
- Average the likelihoods
- Only do the slot model
- Advanced: do for resource model

Exercise: Bayes Factor

- From the final folder, get the code for generating the marginal likelihood of the resource model
- Exercise: estimate marginal likelihood, and then a Bayes Factor

Why is it dumb?

- Very inefficient
- In this example, roughly half of the samples will have zero likelihood

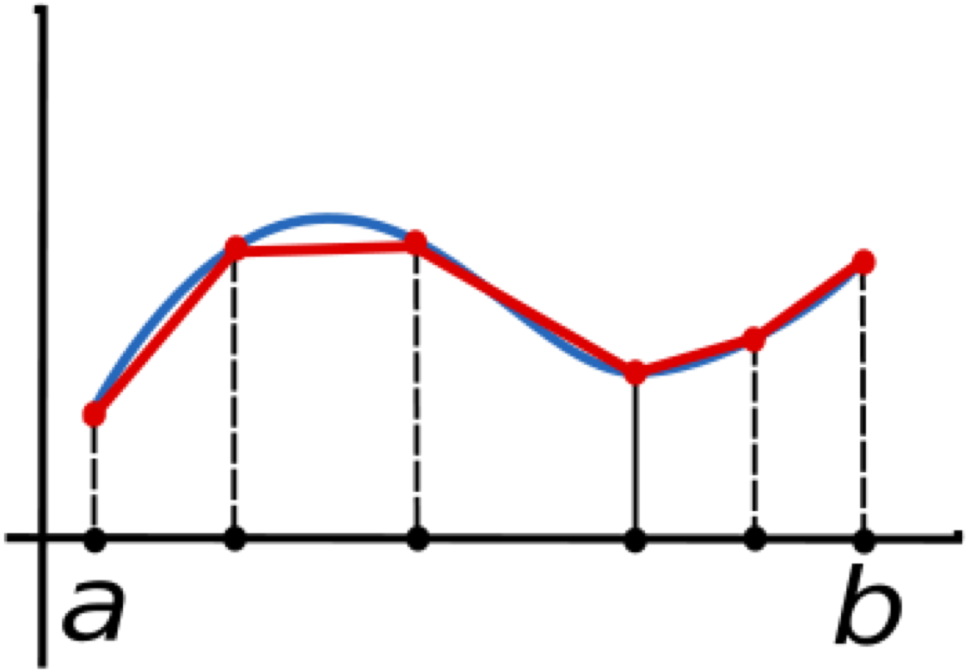
Why is it dumb?

- Very inefficient
- In this example, roughly half of the samples will have zero likelihood
- Can improve this with bett
  - Use posterior distribution
    - Importance sampling
    - Bridge sampling

NUMERICAL INTEGRATION TO CALCULATE BAYES FACTORS (SF)

Numerical integration

- For simple problems (not too many parameters; < 9 say Kass & Raftery, 1995), we can numerically integrate



(from Wikipedia; CC)

Cubature: Adaptive Integration

- Walkthrough

THE IMPORTANCE OF PRIORS (CD)

Exercise

- Change the priors
- Resource model: order constrained

Critical warning

- In parameter estimation, influence of priors diminishes with more data
- This is not the case for model comparison (Bayes Factors)
- Selection of priors is critically important

Setting Priors

- Priors are part of the model specification
- So how do you set them?
- (I think) Best to think about priors in terms of the predictions of models about data

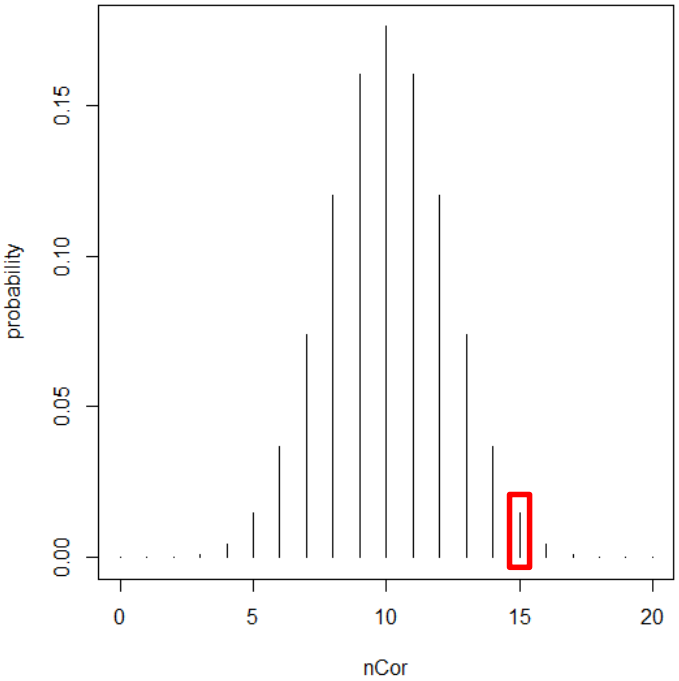
Back to Baby Bayes

- Bayes factors have an inbuilt penalty for complexity

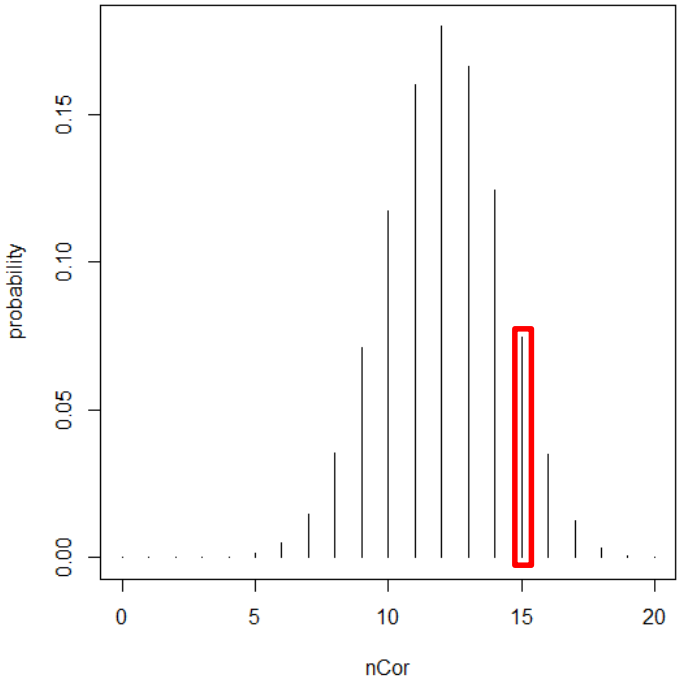
Prediction

- Bayes factors evaluate the predictions of models

$H_0: \theta = 0.5$



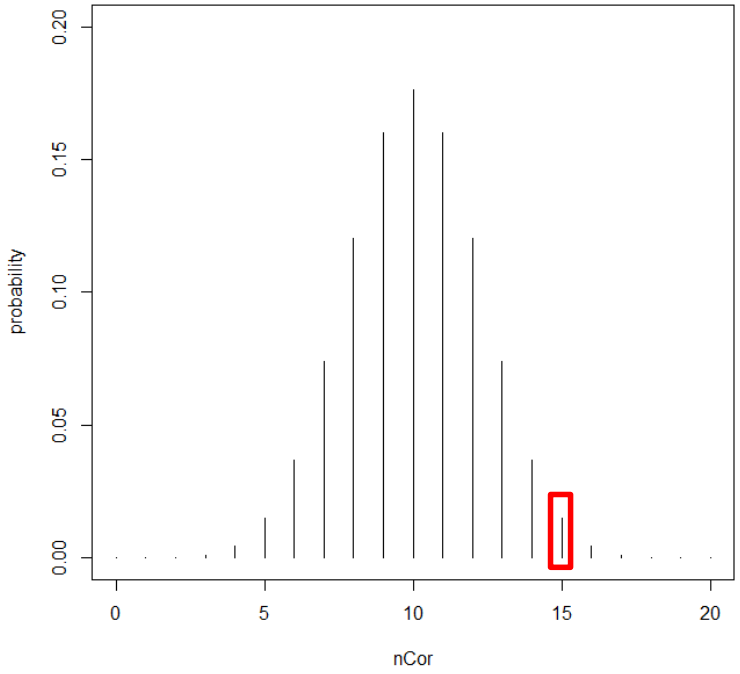
$H_1: \theta = 0.6$

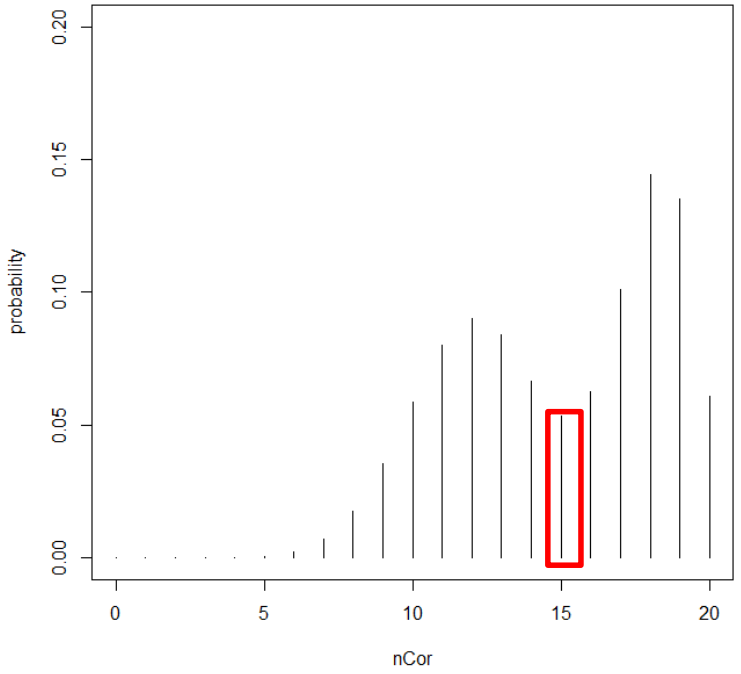


Prediction

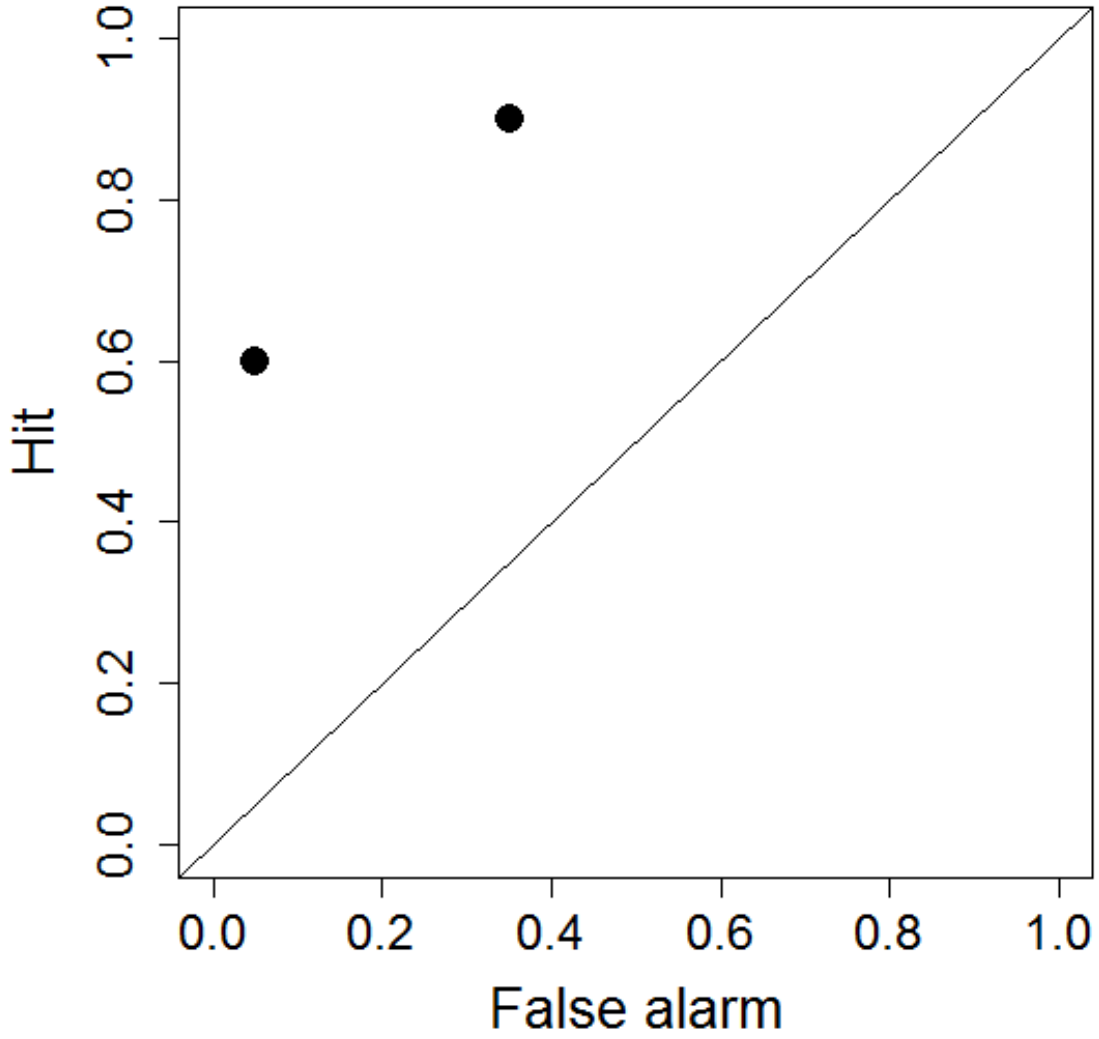
- Bayes factors evaluate the predictions of models

$H_0: \theta = 0.5$ $H_1: \theta = 0.6$ or 0.9

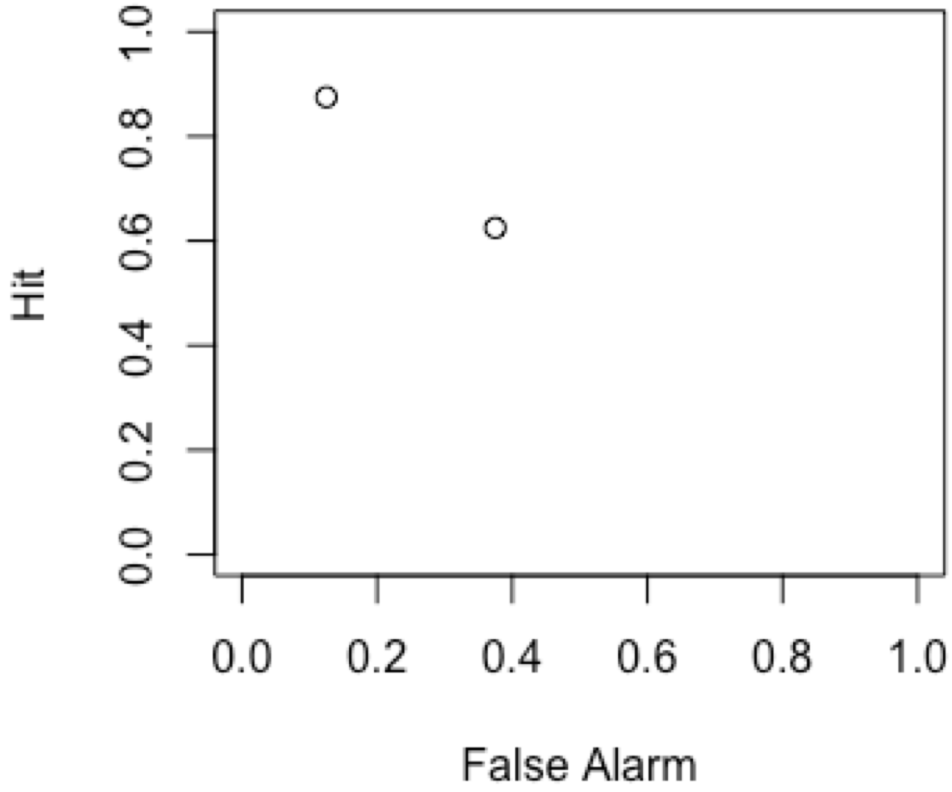




ROC Plots



Fake Experiment Data



Predictions from cognitive model?

- What does a slot model predict?

$$d = {}^{k}/_{N}$$

- Depends on parameter values

$$\underline{\text{Hit}}$$

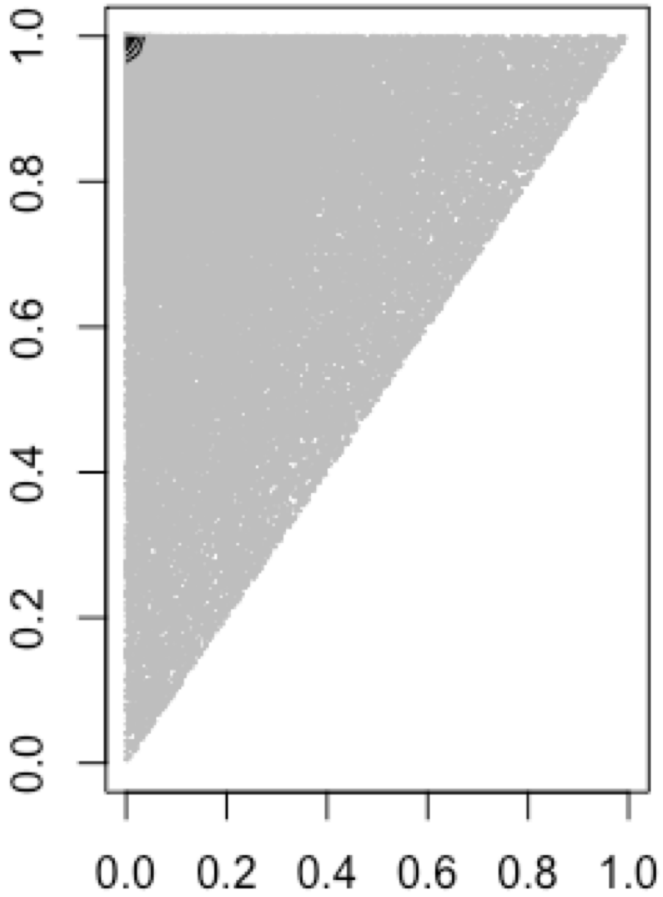
$$d + (1 - d)g$$

$$\underline{\text{False Alarm}}$$

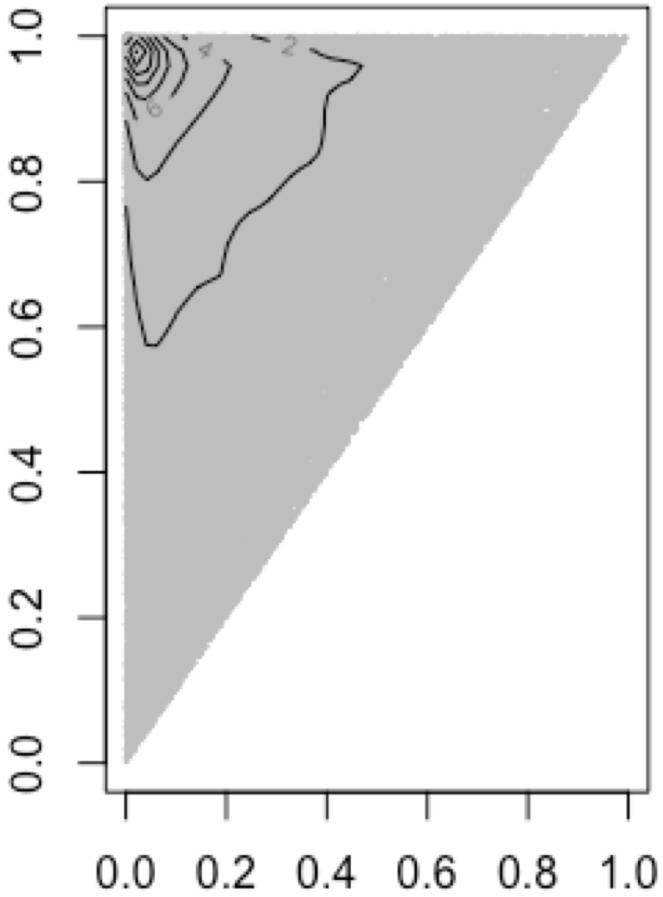
$$(1 - d)g$$

With Uniform Prior

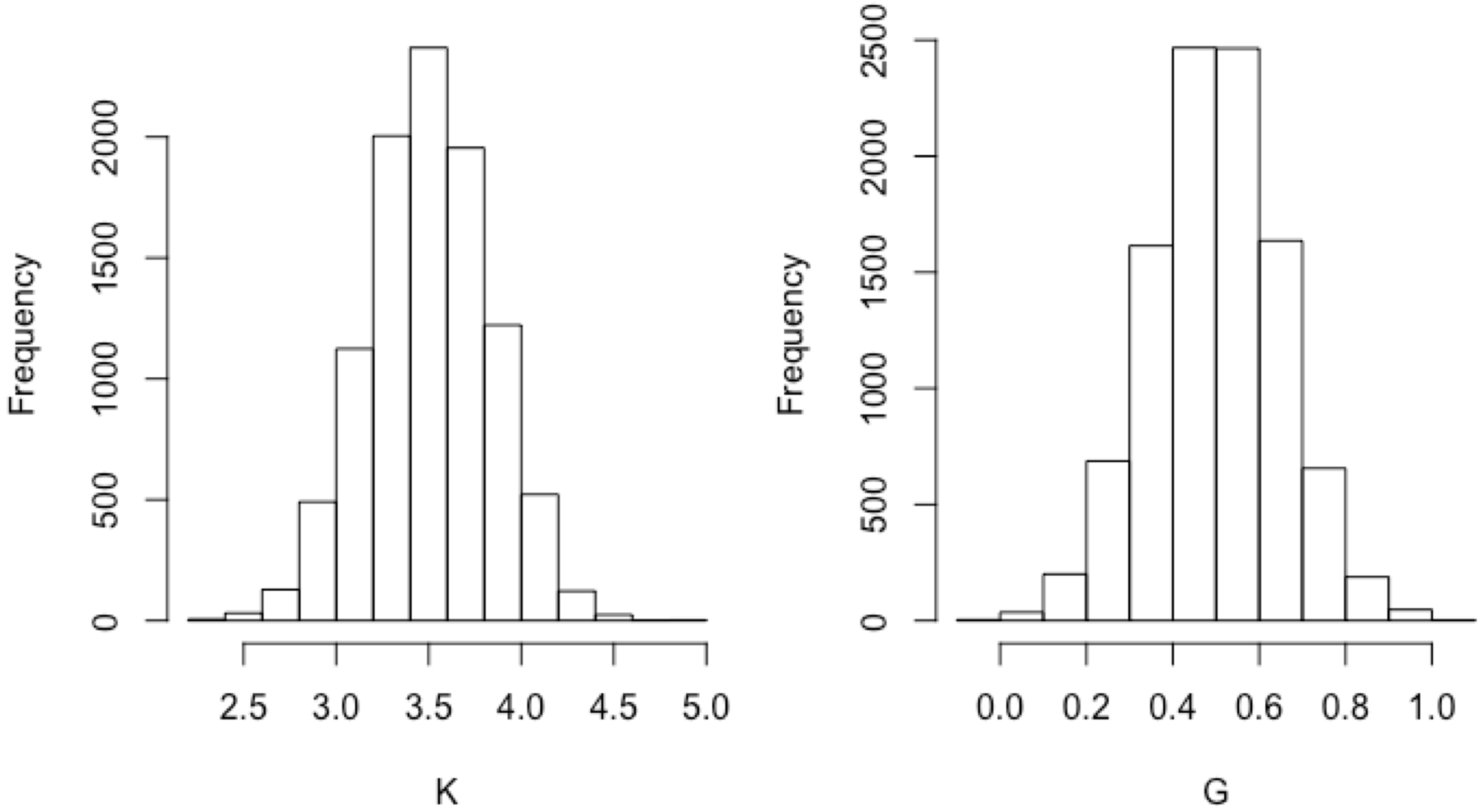
Set Size 4



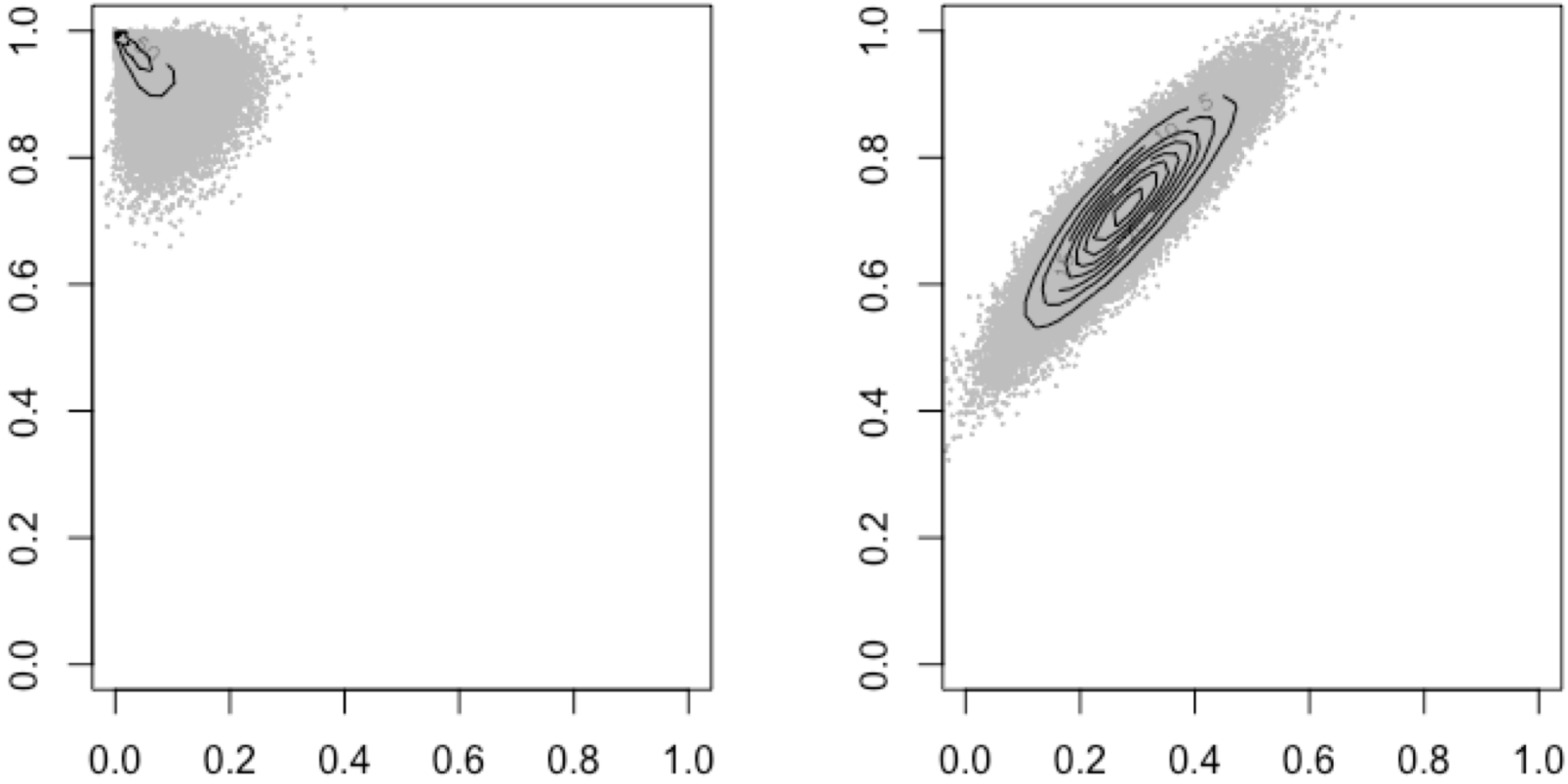
Set Size 8



Better Priors

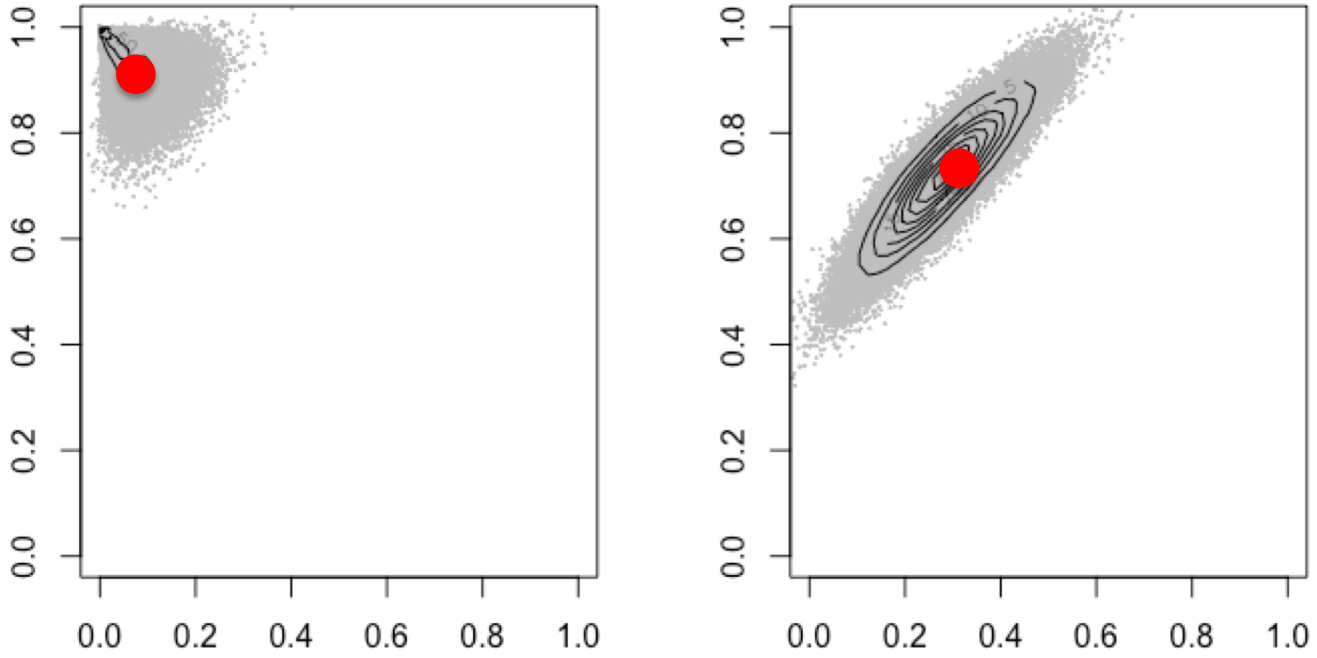


Predictions with Better Priors



Why do better predictions matter?

- Marginal likelihood comes from these predictions



Model Predictions

- Prior + Likelihood determines prediction
- Model that makes better prediction preferred
- So of course priors matter

- Where do these priors come from?
  - Theoretical knowledge about parameters
  - Inspect prior predictives

BAYESIAN INFORMATION CRITERION (SF)

BF interpretation of Bayesian Information Criterion

- Schwartz (1978)
- Estimates $p(y|M)$
  - We can convert back to marginal likelihood by exp(-0.5*BIC)
- If we have time: calculate Bayes Factor using the BICs for slot and resource models
  - Why might they different to our other Bayes Factors?

BIC assumptions

- BIC approximation to marginal likelihood makes some extreme asymptotic (N approaches infinity) assumptions
  - The ML estimate of theta is equal to the mode of the posterior
  - Normality of the likelihood
- Assumes a *unit information* prior: the prior is obtained from the amount of information we expect from a single data point (worked out from the data!)

SUMMARY

Summary

- The whole point of modelling is to reach a conclusion about psychological theory
- Modelling can be informative in several ways:
  - Showing a model works (sufficiency)
  - Model parameter values as a measure of psychological processes
  - Showing a model gives a better account of the data than other models when we correct for number of free parameters/complexity
  - Model comparison more useful in modelling (vs statistics) because our models are often functionally quite different