



Lecture 3: Advanced Bayesian topics



Data Science and Information Theory, ED127 course (2025)

Florent Leclercq

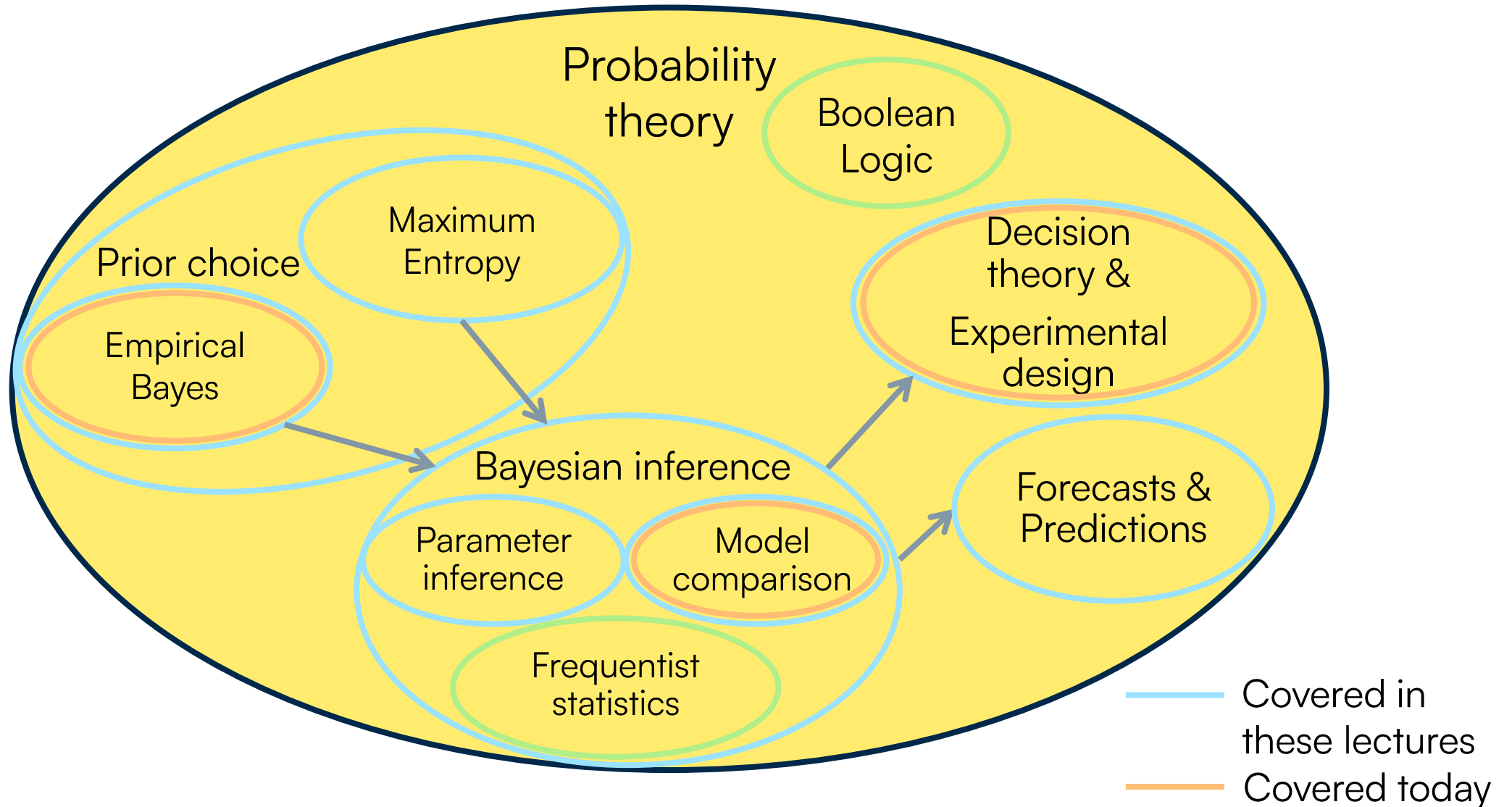
www.florent-leclercq.eu

Institut d'Astrophysique de Paris
CNRS & Sorbonne Université

2 APRIL 2025

Gates of the Arctic National Park, Alaska

Jaynes's “probability theory”: an extension of ordinary Boolean logic



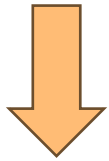
03

BAYESIAN MODEL COMPARISON

Three levels of inference

LEVEL 1:

I have selected a model \mathcal{M} and a prior $p(\theta|\mathcal{M})$



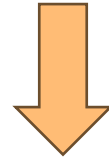
PARAMETER INFERENCE

What are the values of θ preferred by the data, assuming that the model \mathcal{M} is true?

$$p(\theta|d, \mathcal{M}) = \frac{p(d|\theta, \mathcal{M})p(\theta|\mathcal{M})}{p(d|\mathcal{M})}$$

LEVEL 2:

Actually, there are several possible models $\mathcal{M}_1, \mathcal{M}_2 \dots$



MODEL COMPARISON

What is the relative plausibility of the different models $\mathcal{M}_1, \mathcal{M}_2$ given the data?

$$\mathcal{B}_{12} = \frac{p(d|\mathcal{M}_1)}{p(d|\mathcal{M}_2)}$$

LEVEL 3:

None of the models is clearly the best



MODEL AVERAGING

What is the inference on the parameters accounting for model uncertainty?

$$p(\theta|d) = \sum_i p(\theta|d, \mathcal{M}_i)p(\mathcal{M}_i|d)$$



MODEL COMPARISON

Bayesian inference

- Define:
 - data d
 - model θ
 - model parameters \mathcal{M}
- Specify likelihood and prior
- Infer posterior and evidence

INPUTS		OUTPUTS	
$p(d \theta, \mathcal{M})$	$\times p(\theta \mathcal{M})$	$= p(\theta d, \mathcal{M})$	$\times p(d \mathcal{M})$
Likelihood	Prior	Posterior	Evidence

Model selection

- Now apply Bayes' theorem to models \mathcal{M}_i rather than parameters:

$$p(\mathcal{M}_i|d) = \frac{p(d|\mathcal{M}_i)p(\mathcal{M}_i)}{p(d)}$$

- The “meta-evidence” (normalisation) can be written as a sum over models:

$$p(d) = \sum_j p(d|\mathcal{M}_j)p(\mathcal{M}_j)$$

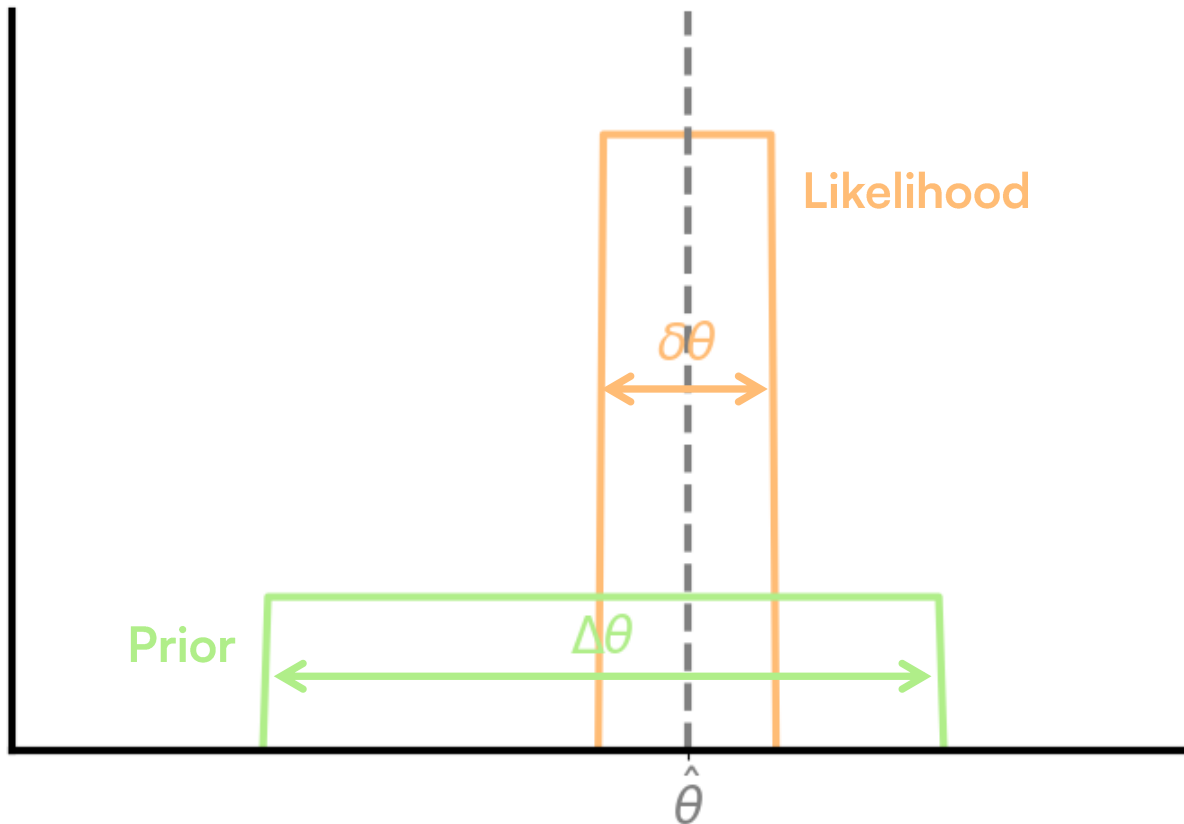
- The evidence for each model can be written as an integral:

$$p(d|\mathcal{M}_j) = \int p(d|\theta, \mathcal{M}_j)p(\theta|\mathcal{M}_j) d\theta$$

- So we can compare the posterior probabilities for two models:

$$\text{Posterior odds: } \frac{p(\mathcal{M}_1|d)}{p(\mathcal{M}_2|d)} = \frac{p(d|\mathcal{M}_1)}{p(d|\mathcal{M}_2)} \frac{p(\mathcal{M}_1)}{p(\mathcal{M}_2)} \quad \text{Bayes factor: } \mathcal{B}_{12} = \frac{p(d|\mathcal{M}_1)}{p(d|\mathcal{M}_2)} \quad \text{Prior odds: } \frac{p(\mathcal{M}_1)}{p(\mathcal{M}_2)}$$

Evidence for a toy model

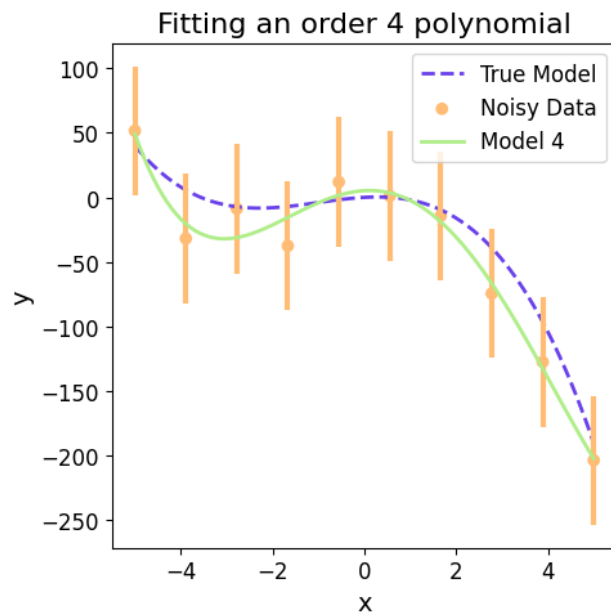
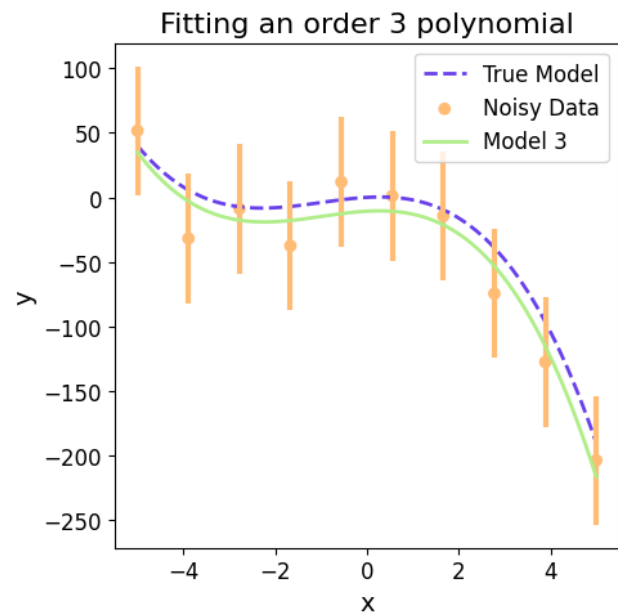
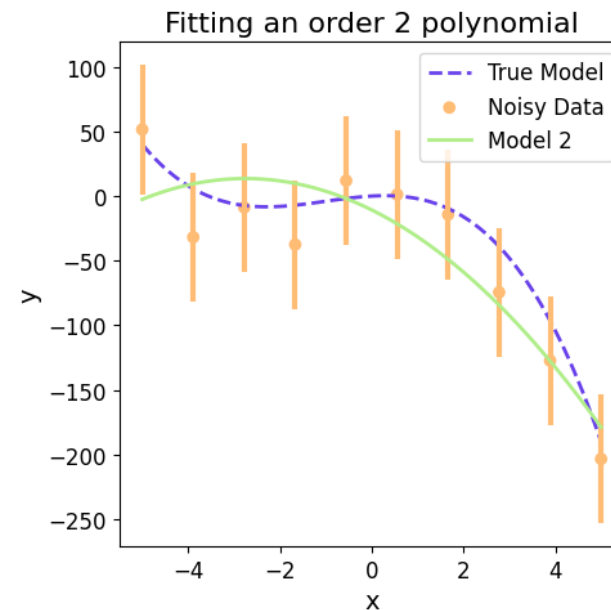
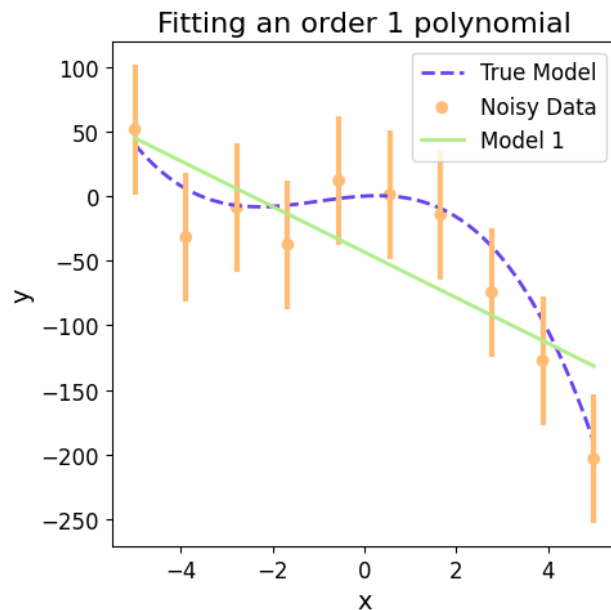
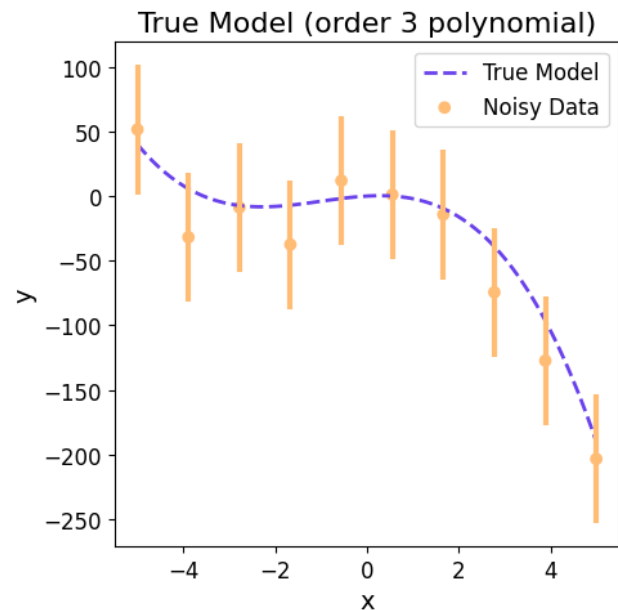


$$\begin{aligned} p(d|\mathcal{M}) &= \int p(d|\theta, \mathcal{M})p(\theta|\mathcal{M}) d\theta \\ &= p(d|\hat{\theta}, \mathcal{M})p(\hat{\theta}|\mathcal{M}) \delta\theta \\ &= L(\hat{\theta}) \frac{\delta\theta}{\Delta\theta} \end{aligned}$$

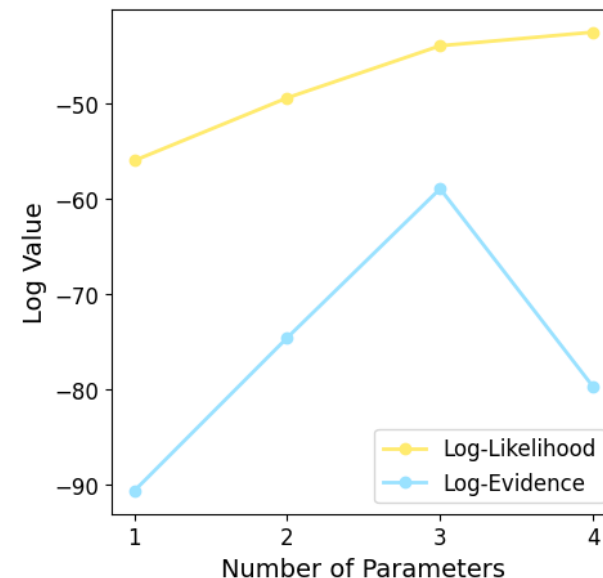
Goodness-of-fit
factor (rewards
accurate predictions)

Occam's factor
(disfavors large
priors)

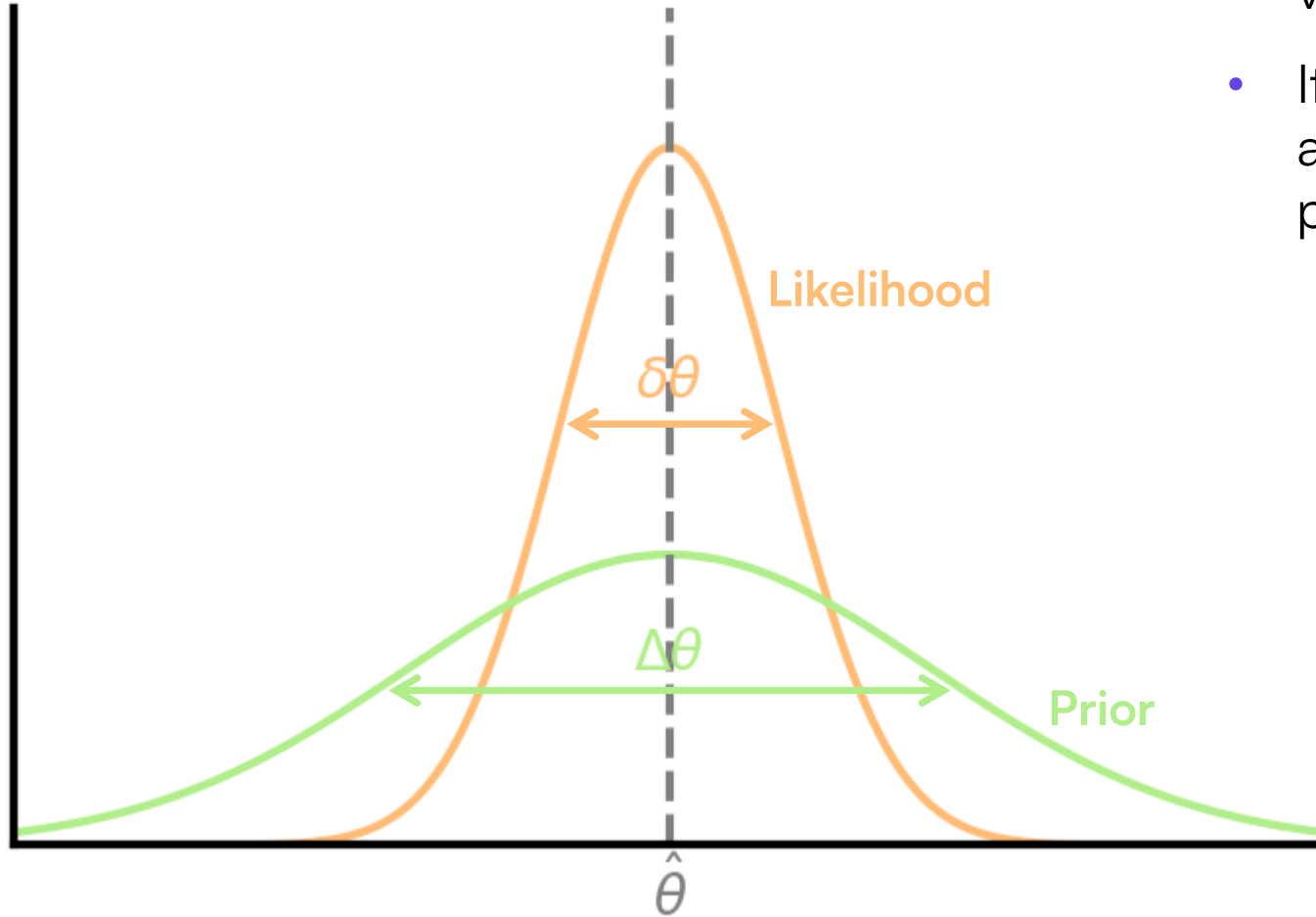
Evidence for polynomial fits



Likelihood and Evidence vs Number of Parameters



An automatic Occam's razor



- The Bayes' factor balances the quality of fit versus the model complexity.
- It rewards highly predictive models (if they are accurate), penalising “wasted” parameter space.

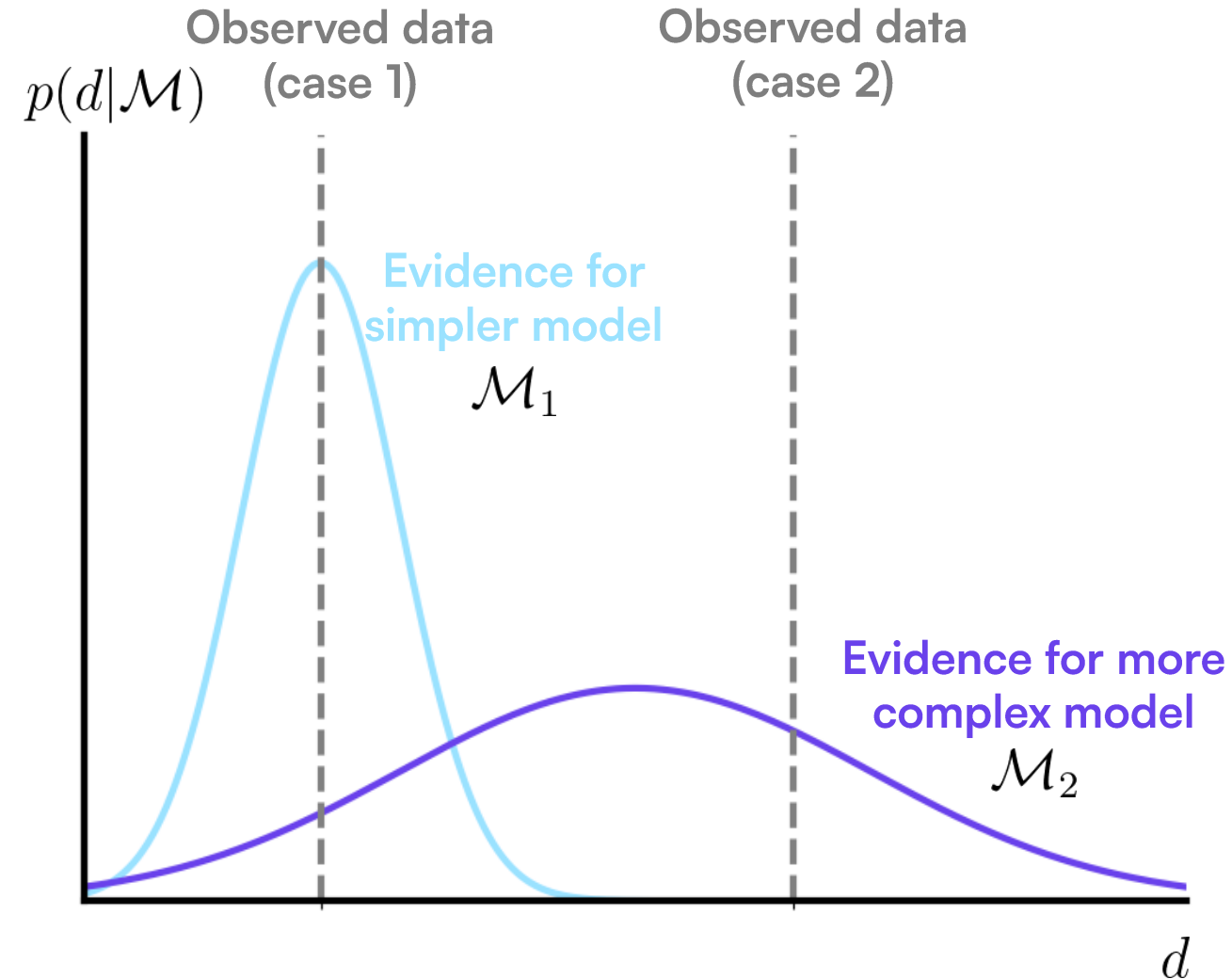
$$\begin{aligned} p(d|\mathcal{M}) &= \int L(\theta)p(\theta|\mathcal{M}) d\theta \\ &\approx L(\hat{\theta}) p(\hat{\theta}|\mathcal{M}) \delta\theta \\ &\approx \frac{\delta\theta}{\Delta\theta} L(\hat{\theta}) \end{aligned}$$

Occam's factor

With four parameters I can fit an elephant, and with five I can make him wiggle his trunk.

John von Neumann


The evidence as predictive probability



- The evidence can be understood as the predictive probability of the data d , under the model \mathcal{M} :
 - In case 1: the simpler model \mathcal{M}_1 is preferred, as it made a sharp prediction that has been verified
 - In case 2: the more complex model \mathcal{M}_2 is preferred, as its additional complexity is required by the data

Decisiveness and the Bayes factor

- Bayes factor: $\mathcal{B}_{12} = \frac{p(d|\mathcal{M}_1)}{p(d|\mathcal{M}_2)}$
- We can always write: $p(d|\mathcal{M}_1) = b p(d|\mathcal{M}_2)$

- Then: $\mathcal{B}_{12} = \frac{b}{1}$ and $\mathcal{B}_{21} = \frac{1}{b}$


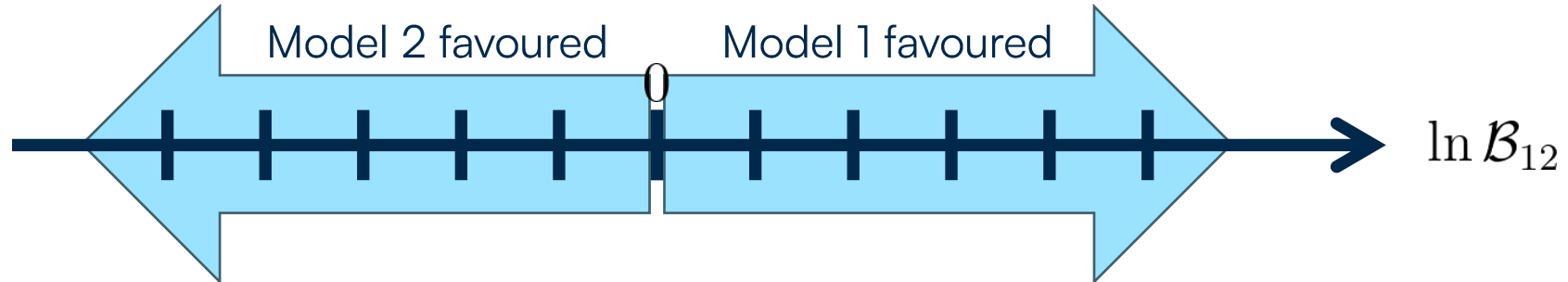
Grows linearly Asymptotes to zero

- Therefore, we take the logarithm to define a measure of decisiveness:

$$\left. \begin{array}{l} \ln \mathcal{B}_{12} = \ln b \\ \ln \mathcal{B}_{21} = -\ln b \end{array} \right\} \rightarrow \text{Now } \mathcal{B}_{12} \text{ and } \mathcal{B}_{21} \text{ are treated on an equal footing}$$

Jeffreys' scale: scale for the strength of evidence

- Decisiveness can be represented graphically:

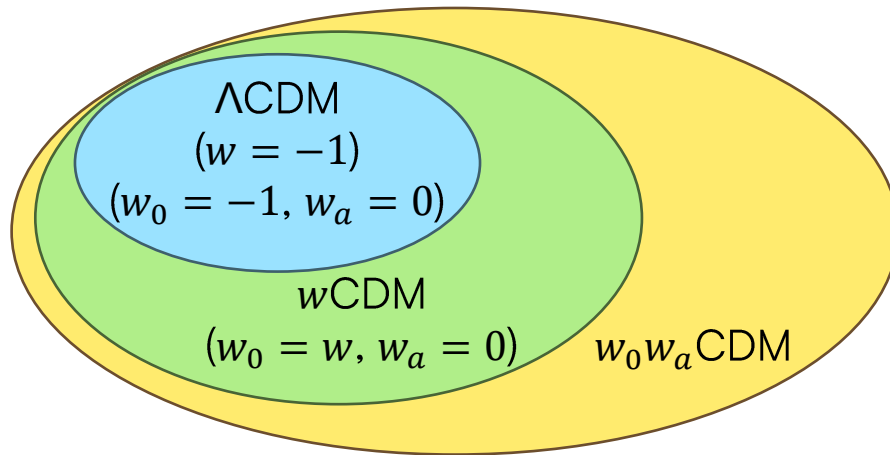


- A (slightly modified) [Jeffreys' scale](#) to measure the strength of evidence:

$ \ln B_{12} $	Relative odds	Favoured models' probability	Interpretation
< 1.0	$< 3:1$	< 0.750	Inconclusive
< 2.5	$< 12:1$	< 0.923	Weak
< 5.0	$< 150:1$	< 0.993	Moderate
> 5.0	$> 150:1$	> 0.993	Strong

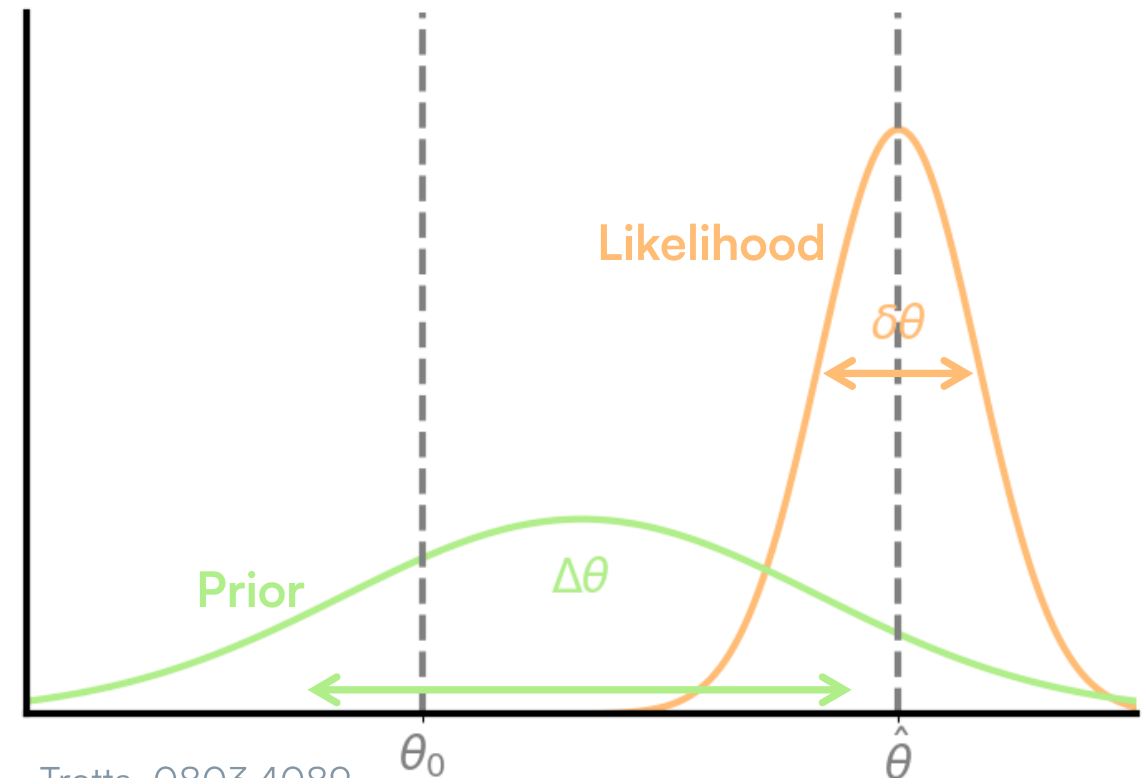
A particular case: nested models

- A frequent case is when \mathcal{M}_1 is a complex model, with prior $p(\theta|\mathcal{M}_1)$, which reduces to a simpler model \mathcal{M}_0 for a certain value of the parameter, e.g. $\theta = \theta_0$. \mathcal{M}_1 and \mathcal{M}_0 are called nested models.
- Example in cosmology:



- Is the extra complexity of \mathcal{M}_1 warranted by the data?

- Define $\lambda \equiv \frac{\hat{\theta} - \theta_0}{\delta\theta}$
- Then for “informative” data:
$$\ln \mathcal{B}_{01} = \ln \frac{\Delta\theta}{\delta\theta} - \frac{\lambda^2}{2}$$



Trotta, 0803.4089

A particular case: nested models

$$\ln \mathcal{B}_{01} = \ln \frac{\Delta\theta}{\delta\theta} - \frac{\lambda^2}{2}$$

Wasted parameter space
(favours simpler model)

Mismatch between prediction
and observation (favours more
complex model)

Likelihood

$$\lambda \equiv \frac{\hat{\theta} - \theta_0}{\delta\theta}$$

$$I_{10} \equiv \log_{10} \frac{\Delta\theta}{\delta\theta}$$

Prior

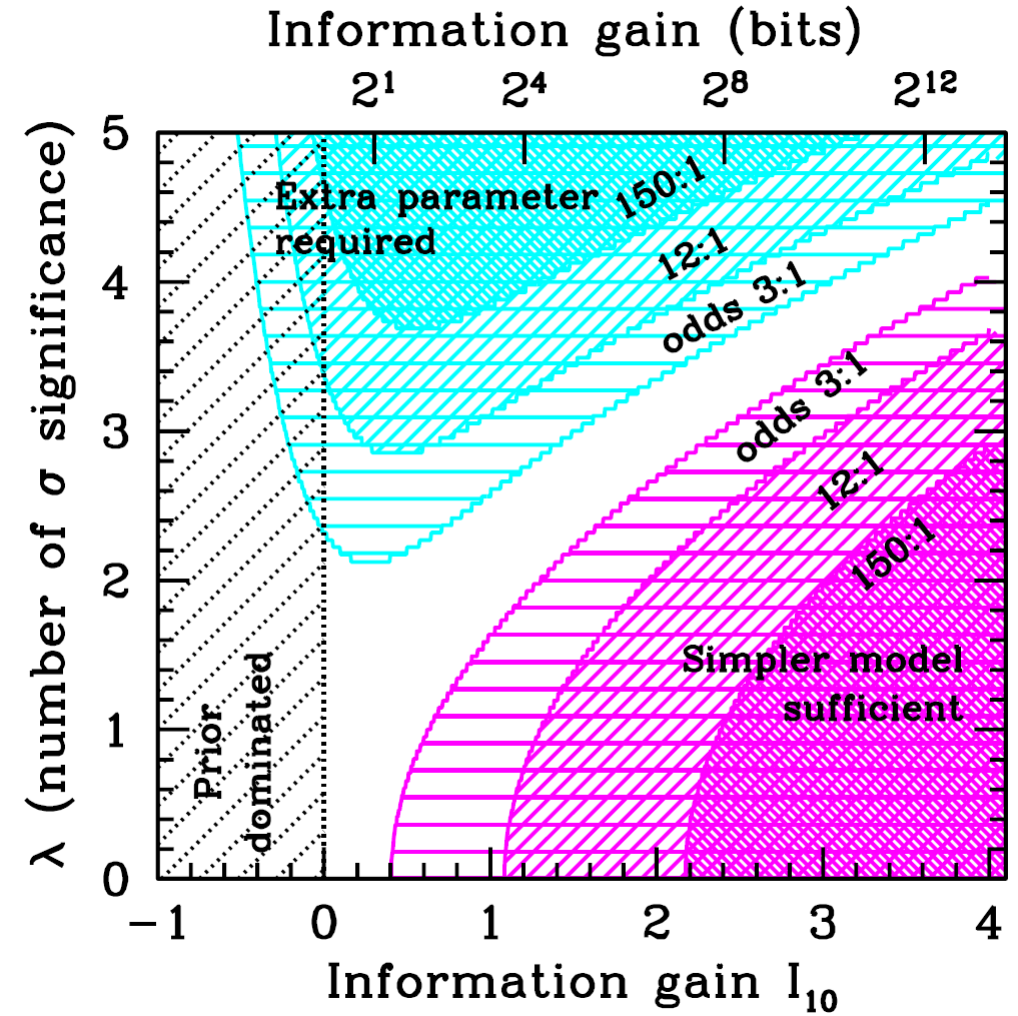
$\Delta\theta$

$\hat{\theta}$

θ_0

wider prior (fixed data)

wider likelihood (fixed prior and significance)



Model comparison example: looking for a signal

- Problem considered: we want to know if there is evidence for a constant signal in a set of noisy data.
- We perform model comparison between two models:
 - \mathcal{M}_0 : the data mean has a known value $\mu = \mu^*$ (e.g. there is no signal, $\mu^* = 0$)
 - \mathcal{M}_1 : the data mean has unknown value $\mu \neq \mu^*$ (e.g. there is a signal, $\mu^* \neq 0$)
- Assume that data points are drawn from a Gaussian with known variance and unknown mean. What is the likelihood and summary statistics for the problem?
- Assume that the prior on the unknown mean is Gaussian. What is the analytic form for the evidence?
- How does the evidence depend on the hyperparameters of the problem?

Model comparison example: looking for a signal

- Likelihood: for one data point:

$$d_i \curvearrowright \mathcal{G}(\mu, \tau^{-1}) \quad \text{with} \quad \tau \equiv 1/\sigma^2$$

- For the full data set:

$$\begin{aligned} p(\{d_i\} | \mu, \tau) &= \prod_{i=1}^N p(d_i | \mu, \tau) \\ &\propto \exp \left[-\frac{1}{2} (N\tau) (\bar{d} - \mu)^2 \right] \\ &= p(\bar{d} | \mu) \end{aligned}$$

$$\text{with } \bar{d} \curvearrowright \mathcal{G}(\mu, (N\tau)^{-1}) \quad \bar{d} \equiv \frac{1}{N} \sum_{i=1}^N d_i$$

- The empirical mean \bar{d} acts as a sufficient summary statistics of the full data. For this problem we can just work with \bar{d} .

- Prior: a Gaussian will be a conjugate prior.

We assume:

$$\mu \curvearrowright \mathcal{G}(\mu_0, p_0^{-1})$$

- This prior is characterised by two hyperparameters (μ_0, p_0)

Model comparison example: looking for a signal

- Evidence: we integrate the product of the likelihood and the prior.
- For \mathcal{M}_1 with a Gaussian prior on μ , we have

$$\begin{aligned}
 p(\bar{d}|\mathcal{M}_1) &= \int p(\bar{d}|\mu)p(\mu|\mu_0, p_0) d\mu \\
 &\propto \int \exp\left[-\frac{1}{2}(N\tau)(\bar{d} - \mu)^2\right] \exp\left[-\frac{1}{2}p_0(\mu - \mu_0)^2\right] d\mu \\
 &\propto \exp\left[-\frac{1}{2}[(N\tau)^{-1} + p_0^{-1}](\bar{d} - \mu_0)^2\right] \\
 &\quad \text{(by completing the square and integrating)} \\
 &\propto \mathcal{G}(\mu_0, (N\tau)^{-1} + p_0^{-1})
 \end{aligned}$$

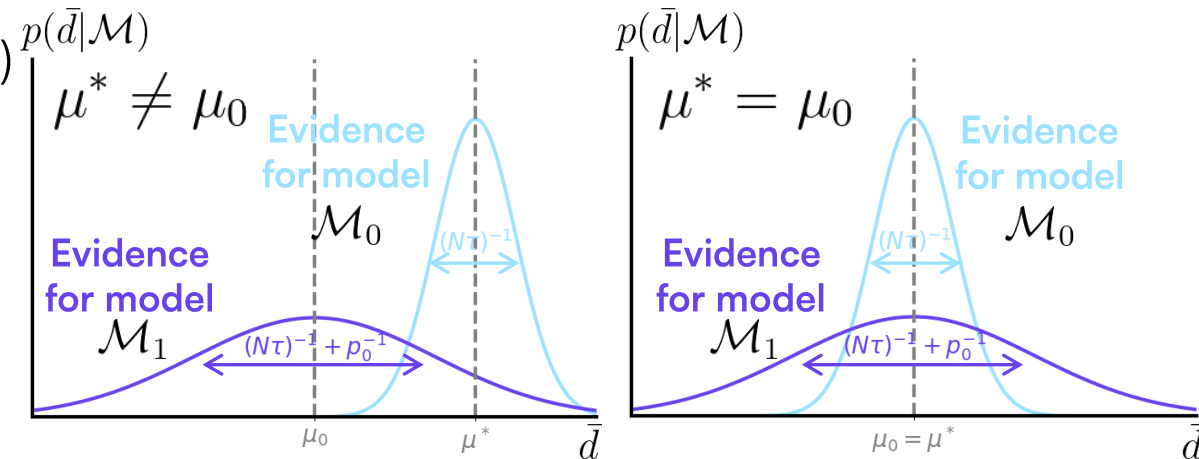
- For \mathcal{M}_0 , the prior is a Dirac delta distribution, $p(\mu) = \delta_D(\mu - \mu^*)$ giving an evidence:

$$p(\bar{d}|\mathcal{M}_0) \propto \mathcal{G}(\mu^*, (N\tau)^{-1})$$

- Bayes factor (including the constants):

$$\mathcal{B}_{01} = \sqrt{\frac{N\tau + p_0}{p_0}} \frac{\exp\left[-\frac{1}{2}(N\tau)(\bar{d} - \mu^*)^2\right]}{\exp\left[-\frac{1}{2}[(N\tau)^{-1} + p_0^{-1}](\bar{d} - \mu_0)^2\right]}$$

- The Bayes factor depends on the hyperparameters (μ_0, p_0) and μ^* . Different priors will change the conclusion of model comparison.
 - e.g. for $\mu^* = \mu_0$, the choice of p_0 will set the thresholds of decisiveness.





MODEL COMPARISON IN PRACTICE

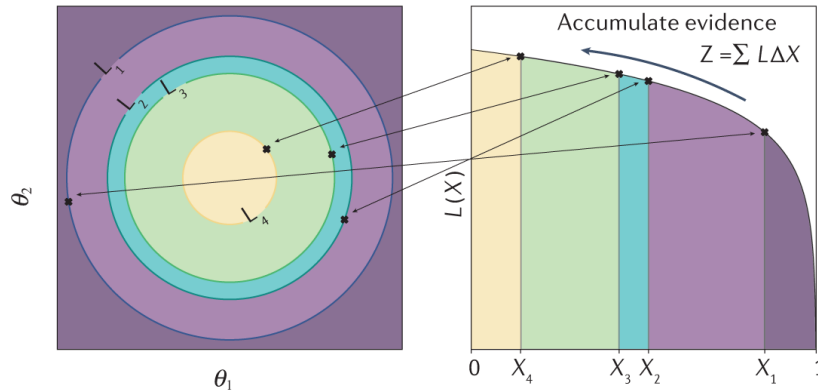
Model comparison in practice

- Full integration of the evidence (“thermodynamic integration”):
 - Nested sampling
 - MCEvidence
- Laplace approximation
- Special cases:
 - e.g. nested models \Rightarrow Savage-Dickey density ratio
- Approximate methods:
 - e.g. information criteria: AIC, BIC, DIC

Full thermodynamic integration

Nested sampling

- Original idea proposed by John Skilling in 2004: convert a D-dimensional integral into a 1D integral that can be done easily.

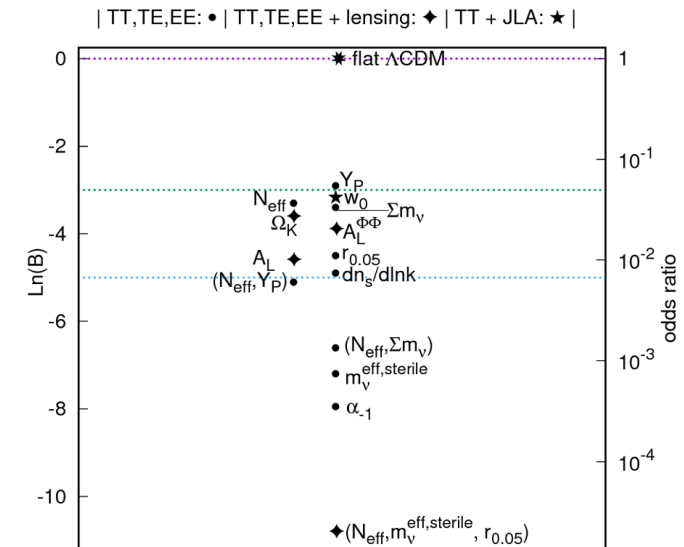


- As a by-product, it also produces posterior samples: parameter inference and model evidence are obtained simultaneously (\Rightarrow alternative to MCMC).
- Several implementations and enhancements: [MultiNest](#), [PolyChord](#).

[Skilling \(2004\)](#); [Mukherjee et al., astro-ph/0508461](#); [Feroz et al., 0809.3437](#); [Graff et al., 1110.2997](#); [Handley et al., 1502.01856](#)

MCEvidence

- After MCMC sampling, using k th nearest-neighbour distances in parameter space and the Mahalanobis distance metric.
- Implementation: [MCEvidence](#).
- Application to Planck (2015) MCMC chains: no evidence for extensions to the standard cosmological model



[Heavens et al., 1704.03472](#); [Heavens et al., 1704.03467](#)

Laplace approximation

- Fit a multivariate Gaussian to the likelihood close to its peak:

$$p(d|\theta, \mathcal{M}) \approx L_{\max} \exp \left[-\frac{1}{2}(\theta - \theta_{\max})^\top \mathbf{L}(\theta - \theta_{\max}) \right]$$

- Assume the prior is Gaussian with zero mean and precision matrix \mathbf{P} :

$$p(\theta|\mathcal{M}) = |2\pi\mathbf{P}^{-1}|^{-1/2} \exp \left[-\frac{1}{2}\theta^\top \mathbf{P}\theta \right]$$

- Then the evidence is:

$$p(d|\mathcal{M}) = L_{\max} \frac{|\mathbf{F}|^{-1/2}}{|\mathbf{P}|^{-1/2}} \exp \left[-\frac{1}{2}(\theta_{\max}^\top \mathbf{L} \theta_{\max} - \bar{\theta}^\top \mathbf{F} \bar{\theta}) \right] \quad \text{with} \quad \begin{cases} \mathbf{F} \equiv \mathbf{L} + \mathbf{P} \\ \bar{\theta} \equiv \mathbf{F}^{-1} \mathbf{L} \theta_{\max} \end{cases}$$

Best-fit factor

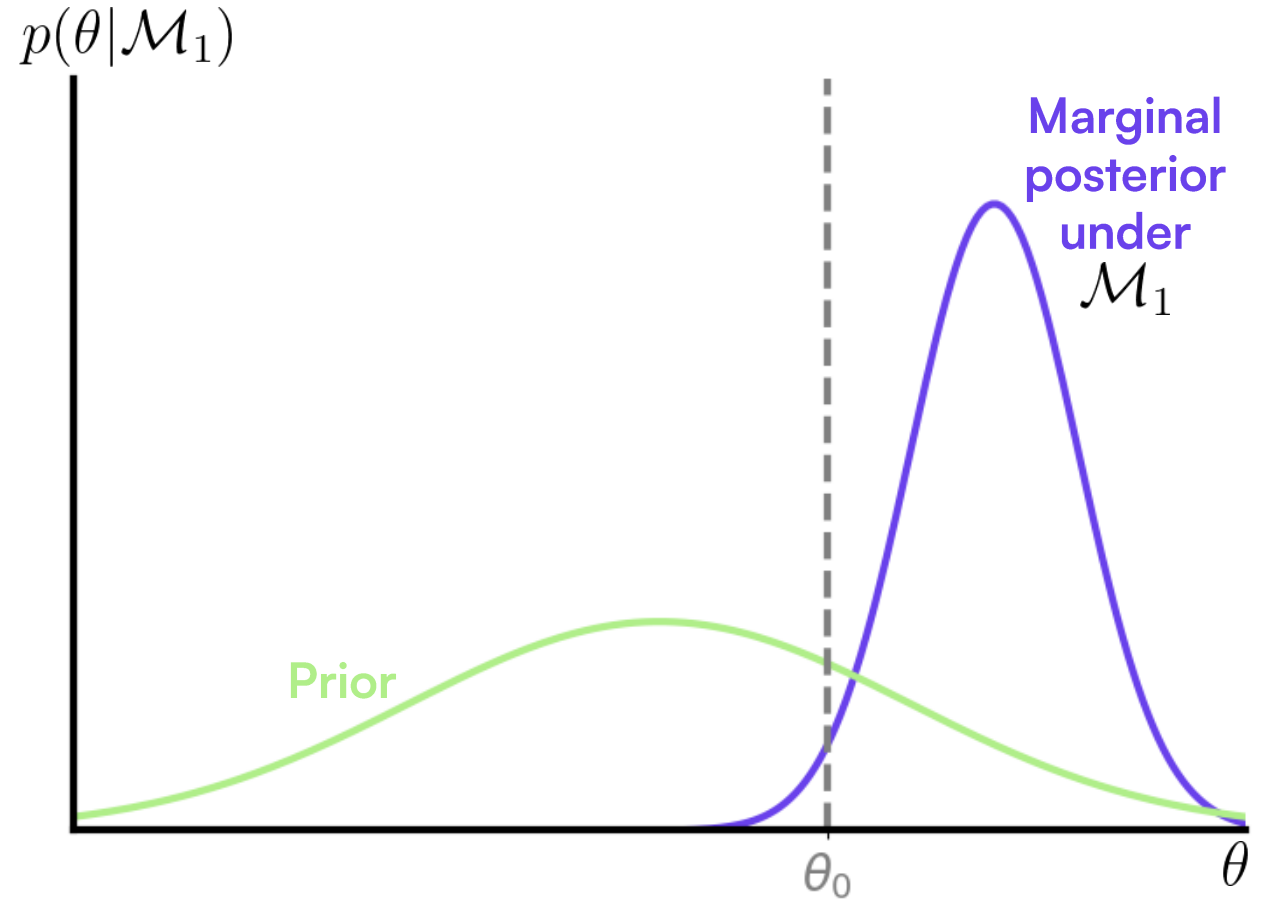
Occam's factor

Suppresses the likelihood of models for which parameter values that maximise the likelihood θ_{\max} differ from the posterior expectation value $\bar{\theta}$

Savage-Dickey density ratio (SDDR)

- This approach is applicable for nested models and provides an analytical solution for the Bayes factor.
- Assumptions:
 - Nested models: \mathcal{M}_1 with parameters (θ, ψ) reduces to \mathcal{M}_0 for $\theta = \theta_0$
 - Separable prior:
 $p(\theta, \psi | \mathcal{M}_1) = p(\theta | \mathcal{M}_1) p(\psi | \mathcal{M}_0)$
- Result (SDDR):

$$\mathcal{B}_{01} = \frac{p(\theta_0 | d, \mathcal{M}_1)}{p(\theta_0 | \mathcal{M}_1)}$$
- Interpretation: The Bayes factor is the ratio of the normalised marginal posterior in \mathcal{M}_1 over its prior, evaluated at the value of the parameter for which \mathcal{M}_1 reduces to \mathcal{M}_0 .



- The SDDR involves the (low-dimensional) posterior and prior of the extra parameter.
- It is calculable e.g. from MCMC samples drawn from the posterior under \mathcal{M}_1 .

Savage (1962); Dickey (1972); Verdinelli & Wasserman (1995)

Derivation of the SDDR

- We compute the evidence of model \mathcal{M}_0 in terms of \mathcal{M}_1 using the rules of probability theory:

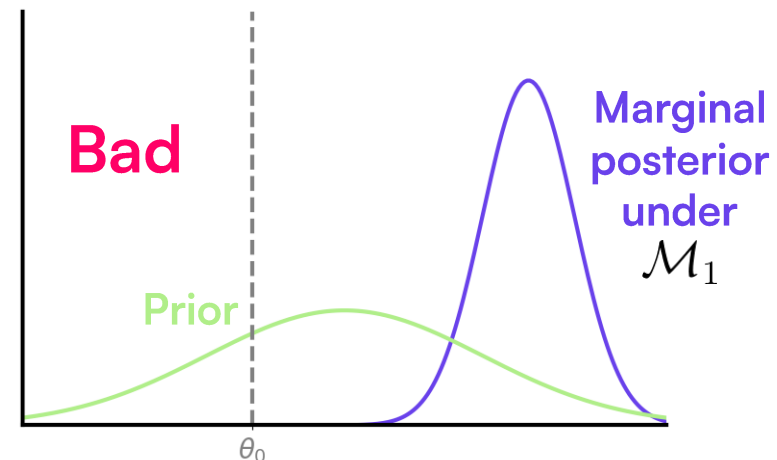
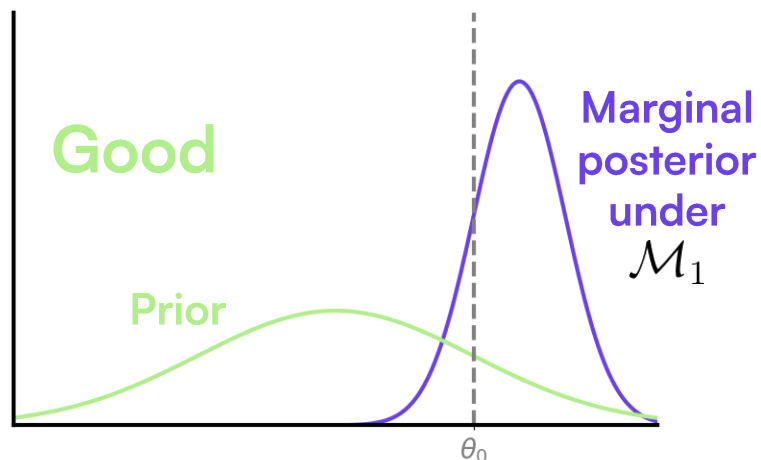
$$\begin{aligned} p(d|\mathcal{M}_0) &= \int p(d|\psi, \theta_0, \mathcal{M}_0) p(\psi|\mathcal{M}_0) d\psi \\ &= \int \frac{p(\psi, \theta_0|d, \mathcal{M}_1) p(d|\mathcal{M}_1)}{p(\theta_0, \psi|\mathcal{M}_1)} p(\psi|\mathcal{M}_0) d\psi && \text{(Bayes' theorem)} \\ &= \int \frac{p(\psi, \theta_0|d, \mathcal{M}_1) p(d|\mathcal{M}_1)}{p(\theta_0|\mathcal{M}_1) p(\psi|\mathcal{M}_0)} p(\psi|\mathcal{M}_0) d\psi && \text{(separable prior hypothesis)} \\ &= \frac{p(d|\mathcal{M}_1)}{p(\theta_0|\mathcal{M}_1)} \int p(\psi, \theta_0|d, \mathcal{M}_1) d\psi \\ &= \frac{p(d|\mathcal{M}_1)}{p(\theta_0|\mathcal{M}_1)} \int p(\psi|\theta_0, d, \mathcal{M}_1) p(\theta_0|d, \mathcal{M}_1) d\psi && \text{(product rule)} \\ &= \frac{p(d|\mathcal{M}_1)}{p(\theta_0|\mathcal{M}_1)} p(\theta_0|d, \mathcal{M}_1) \int p(\psi|\theta_0, d, \mathcal{M}_1) d\psi && \text{(normalisation to unity)} \\ &= p(d|\mathcal{M}_1) \frac{p(\theta_0|d, \mathcal{M}_1)}{p(\theta_0|\mathcal{M}_1)} \end{aligned}$$

- Therefore: $\mathcal{B}_{01} = \frac{p(d|\mathcal{M}_0)}{p(d|\mathcal{M}_1)} = \frac{p(\theta_0|d, \mathcal{M}_1)}{p(\theta_0|\mathcal{M}_1)}$

Savage (1962); [Dickey \(1972\)](#); [Verdinelli & Wasserman \(1995\)](#)

Comments on the Savage-Dickey density ratio

- For nested models and separable priors, the values of the common parameters ψ do not matter for the value of the Bayes factor.
 - Therefore, no need to spend time/resources to average the likelihoods over the common parameters!
 - Prior sensitivity analysis is simplified: only the prior on the additional parameter needs to be considered.
- The role of the prior on the additional parameter is clarified: the prior on is θ :
 - a Dirac delta distribution in \mathcal{M}_0 : $p(\theta|\mathcal{M}_0) = \delta_D(\theta - \theta_0)$
 - a wider distribution in \mathcal{M}_1 : $p(\theta|\mathcal{M}_1)$ (dilution of the predictive power of \mathcal{M}_0)
- The wider the prior, the stronger Occam's razor effect.
- The SDDR does not assume Gaussianity, but it does require sufficiently detailed sampling of the marginal posterior under \mathcal{M}_1 to evaluate reliably its value at $\theta = \theta_0$.



Information criteria

- In some cases, we need a simpler way to roughly rank models. Several [information criteria](#) exist to [approximate](#) Bayesian model comparison.

- Parameters:

- N : number of model parameters
- k : number of data points
- $-2 \ln L_{\max}$: best-fit χ^2

- Akaike information criterion:

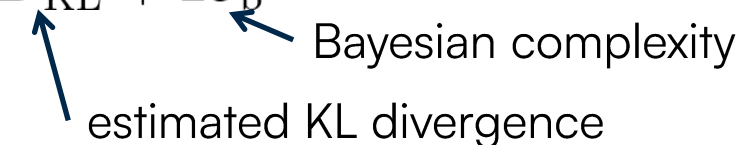
$$\text{AIC} \equiv -2 \ln L_{\max} + 2k$$

- Bayesian information criterion:

$$\text{BIC} \equiv -2 \ln L_{\max} + k \ln N$$

- Deviance information criterion:

$$\text{DIC} = -2\widehat{D}_{\text{KL}} + 2\mathcal{C}_b$$



[Trotta, 0803.4089](#)

- The best model is the one which minimises the AIC/BIC/DIC.
- The AIC and BIC penalise models differently as a function of the number of data points (stronger penalty with the BIC for $N > 7$).
- The BIC approximates the full Bayesian evidence with a Gaussian prior equivalent to $1/N$ -th of the data in the large N limit.
- The Bayesian evidence does not penalise models with parameters that are unconstrained by the data. Unmeasured parameters (posterior = prior) do not contribute to the evidence integral.
 - The DIC considers whether parameters are measured or not (via the Bayesian complexity).
- When possible, calculation of the Bayesian evidence is always preferable.
 - Note: none of these information criteria are Bayesian (not even the BIC). In Bayesian statistics, finding that the data are extremely implausible within a model does not invalidate the model in the absence of an explicit alternative model with better performance.

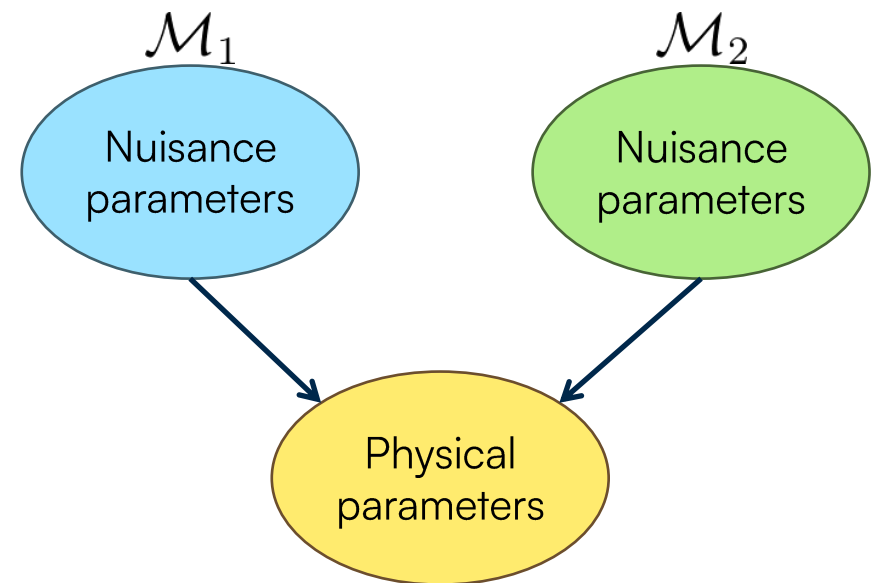


MODEL AVERAGING

Model averaging

- Imagine that two or more models explain the same effect (predict the same parameters). None is “better” than the others, as probed by the Bayesian evidence.
- Examples:
 - Weak lensing: different intrinsic alignment models:
 - Linear Alignment (LA)
 - Tidal Alignment Tidal Torque (TATT)
 - Empirical models based on simulations...
 - Structure formation:
 - Press-Schechter mass function
 - Sheth-Tormen mass function
 - Tinker *et al.* mass function
 - Jenkins *et al.* mass function...
- Model averaging:
 - includes model uncertainty into final parameter uncertainty
 - can be thought of as “third-level” Bayesian inference

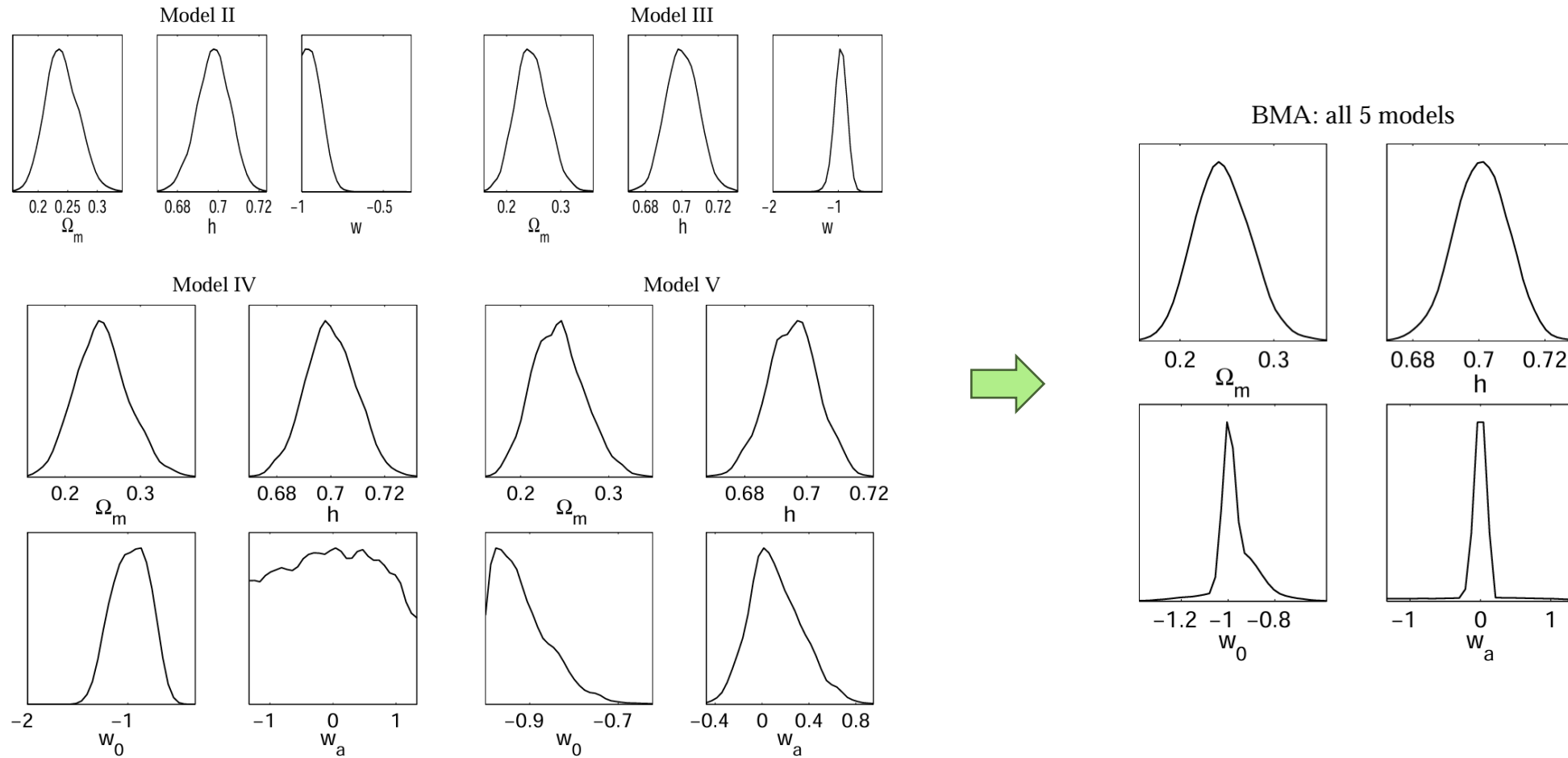
- Typical scenario:



$$p(\theta|d) = \sum_i p(\theta|d, \mathcal{M}_i) p(\mathcal{M}_i|d)$$

Model averaging: example

- Application to dark energy:



Bayesian model comparison: summary

- Bayesian model comparison extends Bayesian inference to the space of models, using evidence ratios.
- The Bayesian evidence balances the goodness of fit against the model complexity (number of parameters, prior volume).
- In practice,
 - Various approximations exist (SDDR for nested models, Laplace approximation, information criteria).
 - Algorithms exist that give parameter constraints and evidences.
- Setting priors for model comparison is important (often easier with nested models). Model comparison is prior-dependent.

03

BAYESIAN DECISION THEORY AND EXPERIMENTAL DESIGN



BAYESIAN DECISION THEORY

Bayesian decision theory

- Bayesian decision theory is a framework for **optimal decision-making**, given a set of possible actions and a state of uncertain knowledge, represented by a pdf $p(\theta|I)$ (usually the posterior from a Bayesian inference prior to decision-making).
- Notations:
 - $\{\theta\}$ = set of parameters (observed variables)
 - $\{a\}$ = set of possible actions
- **Expected utility hypothesis**: Given a set of gain functions $G(a|\theta)$, the optimal decision rule consists of performing the action that maximises the expected utility $U(a|I)$, defined by

$$U(a|I) \equiv \langle G(a|\theta) \rangle_{p(\theta|I)} = \int G(a|\theta) p(\theta|I) d\theta$$

- Thus, one should perform the action $a^* = \operatorname{argmax}_a U(a|I)$.

Example: Bayesian alerts

Exercise: Bayesian alerts

- We are looking for an event E . We have access to $p(E|I)$ and $p(\bar{E}|I) = 1 - p(E|I)$.
- There are two possible actions:
 - a_1 = raise the alert
 - a_2 = do nothing

- The utility functions are:

$$\begin{aligned} U(a_1|I) &= \underbrace{G(a_1|E)}_{\substack{\text{correct detection} \\ \text{(a "hit")}}} p(E|I) + \underbrace{G(a_1|\bar{E})}_{\substack{\text{false positive} \\ \text{(a "false alarm")}}} [1 - p(E|I)] \\ U(a_2|I) &= \underbrace{G(a_2|E)}_{\substack{\text{false negative} \\ \text{(a "miss")}}} p(E|I) + \underbrace{G(a_2|\bar{E})}_{\substack{\text{correct rejection}}} [1 - p(E|I)] \end{aligned}$$

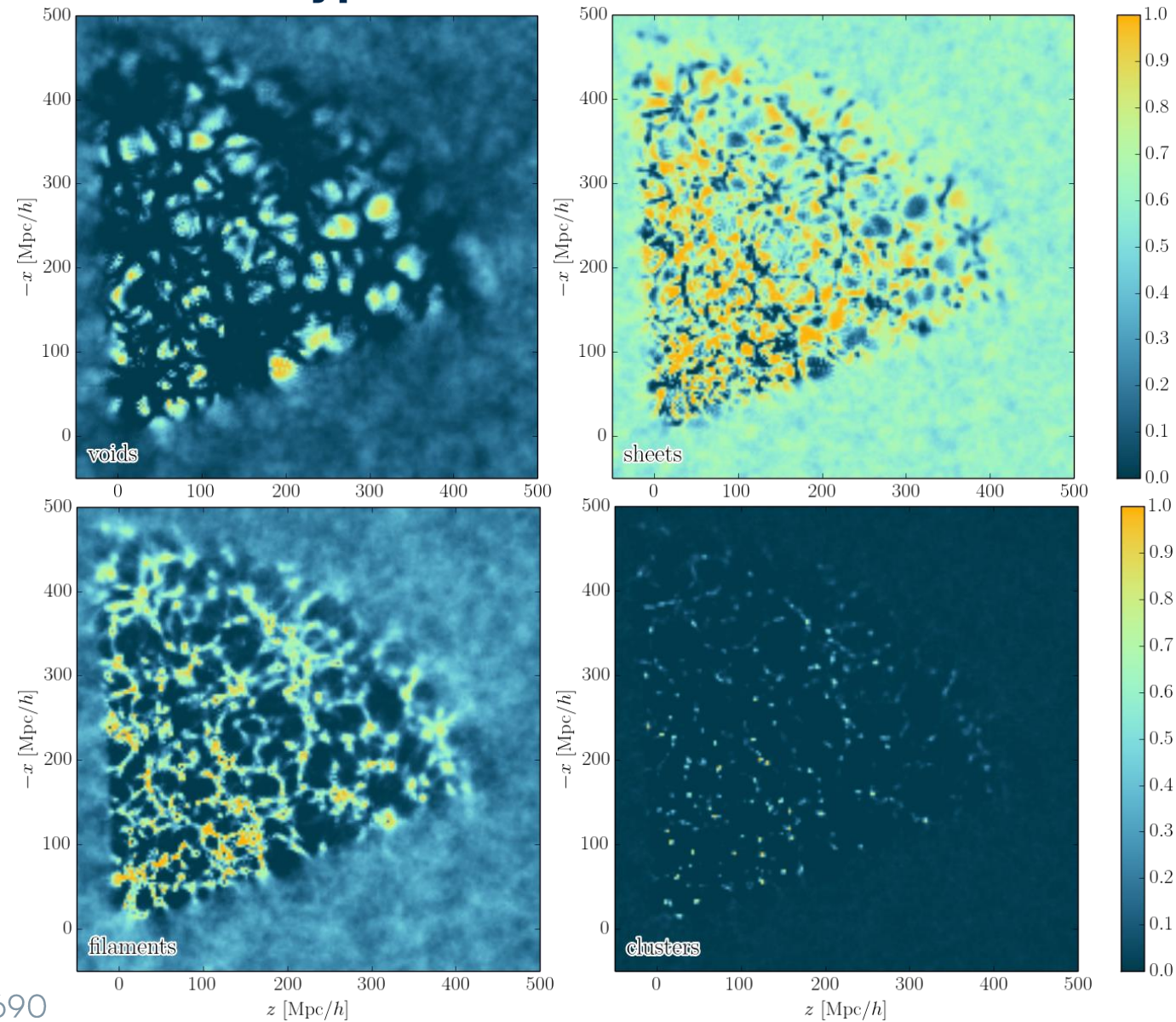
- A typical choice of gain functions:
$$\begin{aligned} G(a_1|E) &= G - C & G(a_1|\bar{E}) &= -C \\ G(a_2|E) &= 0 & G(a_2|\bar{E}) &= 0 \end{aligned}$$

the expected gain for a detection the cost of raising an alert

- Therefore, we have
$$U(a_1|I) = p(E|I)(G - C) + [1 - p(E|I)](-C)$$
$$U(a_2|I) = 0$$

➡ One should raise the alert if and only if $p(E|I) \geq \frac{C}{G}$

Classification of cosmic web-types



A decision rule for structure classification

- Space of “input features”:

$$\{T_0 = \text{void}, T_1 = \text{sheet}, T_2 = \text{filament}, T_3 = \text{cluster}\}$$

- Space of “actions”:

$$\{a_0 = \text{“decide void”}, a_1 = \text{“decide sheet”}, a_2 = \text{“decide filament”}, a_3 = \text{“decide cluster”}, a_{-1} = \text{“do not decide”}\}$$

- It is thus a problem of [Bayesian decision theory](#): one should take the action that maximises the utility

$$U(a_j(\vec{x}_k)|d) = \sum_{i=0}^3 G(a_j|T_i) \mathcal{P}(T_i(\vec{x}_k)|d)$$

- How to write down the gain functions?

Gambling with the Universe

- One proposal:

$$G(a_j | T_i) = \begin{cases} \frac{1}{\mathcal{P}(T_i)} - \alpha & \text{if } j \in \llbracket 0, 3 \rrbracket \text{ and } i = j & \text{"Winning"} \\ -\alpha & \text{if } j \in \llbracket 0, 3 \rrbracket \text{ and } i \neq j & \text{"Losing"} \\ 0 & \text{if } j = -1. & \text{"Not playing"} \end{cases}$$

- Without data, the expected utility is

$$U(a_j) = 1 - \alpha \quad \text{if } j \neq -1 \quad \text{"Playing the game"}$$

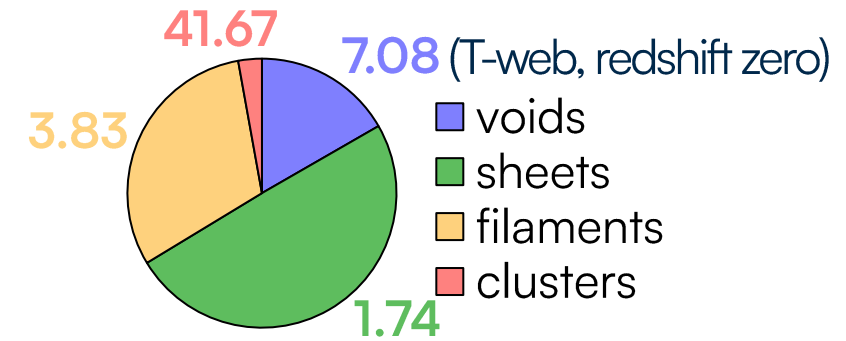
$$U(a_{-1}) = 0 \quad \text{"Not playing the game"}$$

- With $\alpha = 1$, it's a fair game \Rightarrow always play

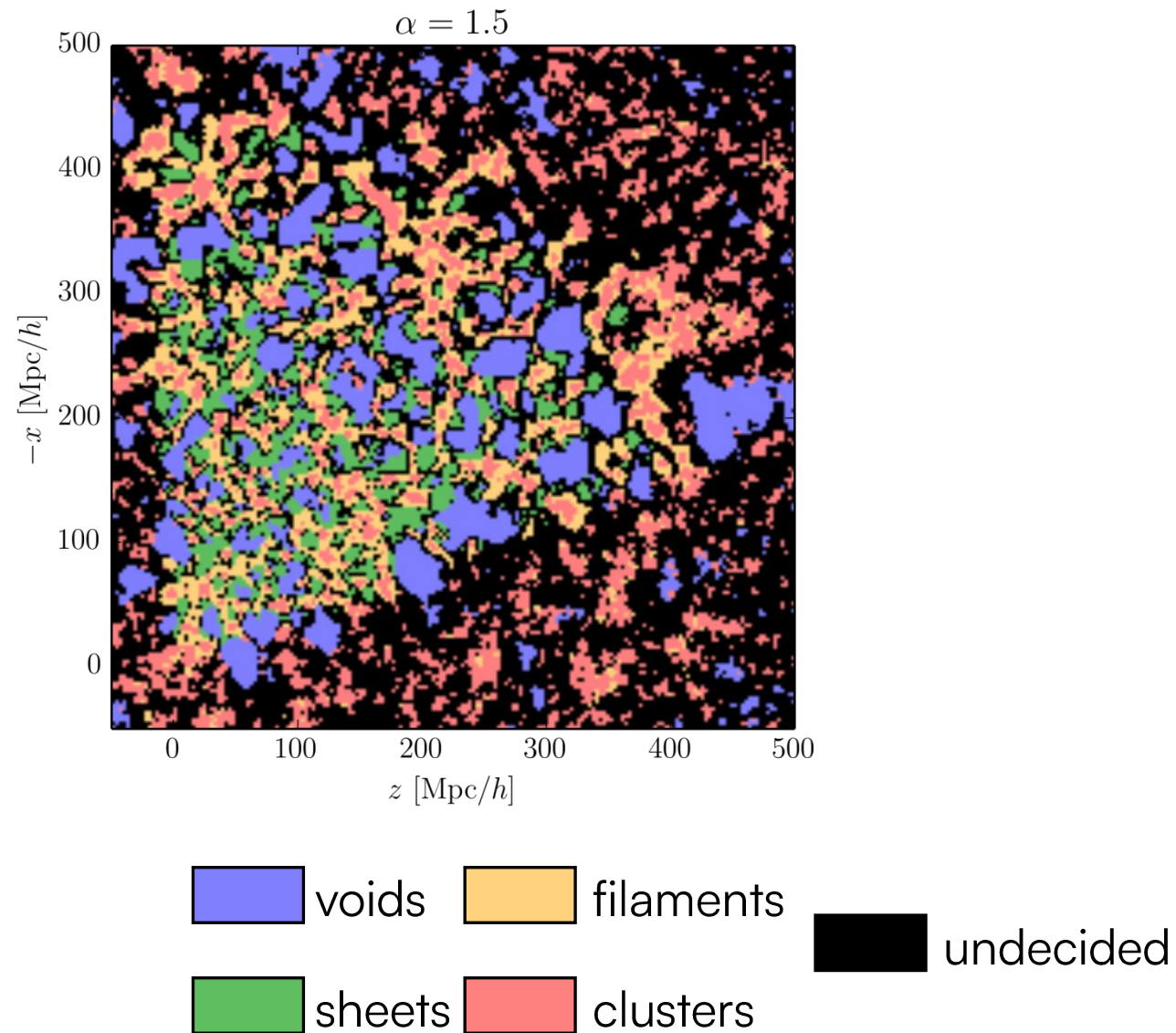
\Rightarrow "speculative map" of the LSS

- Values $\alpha > 1$ represent an aversion for risk

\Rightarrow increasingly "conservative maps" of the LSS



Playing the game...





BAYESIAN EXPERIMENTAL DESIGN

Experiment utility and optimisation

- Bayesian experimental design is an optimisation problem where we seek to optimise the expected utility of a future experiment.
- The optimisation problem is fully specified by the joint utility function $U(\xi, d, \theta|o)$ with
 - ξ : experimental design (parameter characterising the design of the new experiment)
 - d : new data to be acquired
 - θ : parameters of the problem, to be measured
 - o : result of the current experiment (all probabilities are conditional on o here)
- We can evaluate the expected utility:

$$\begin{aligned} U(\xi|o) &= \langle U(\xi, d, \theta|o) \rangle_{p(d, \theta|\xi, o)} \\ &= \iint U(\xi, d, \theta|o) p(d, \theta|\xi, o) \, dd \, d\theta \\ &= \iint U(\xi, d, \theta|o) \underbrace{p(d|\theta, \xi, o)}_{\text{predictive distribution of new experiment}} \underbrace{p(\theta|o)}_{\text{posterior of current experiment}} \, dd \, d\theta \end{aligned}$$

Experiment utility and optimisation

- Particular cases:

1. If the utility does not explicitly depend on the true values of the parameters to be measured (only on the quality of the future data): $U(\xi, d, \theta|o) = U(\xi, d|o)$

$$\text{Then } U(\xi|o) = \int U(\xi, d|o) \underbrace{p(d|\xi, \theta)}_{\text{predictive distribution of new experiment}} dd$$

predictive distribution of
new experiment

2. If the future data explicitly contribute to the “scientific return” $e = (\xi, d)$ (not only the experimental design), then one should not marginalise over d : $U(\xi, d, \theta|o) = U(e, \theta|o)$

$$\text{Then } U(e|o) = \int U(e, \theta|o) \underbrace{p(\theta|o)}_{\text{posterior of current experiment}} d\theta$$

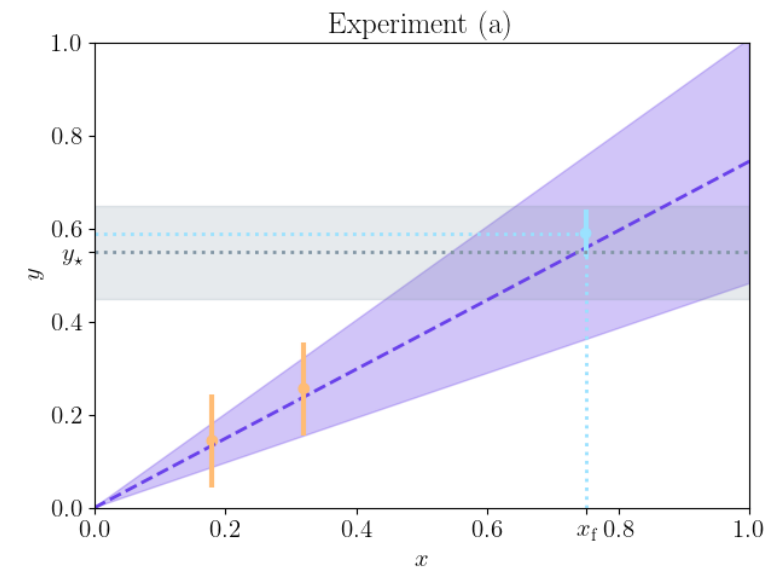
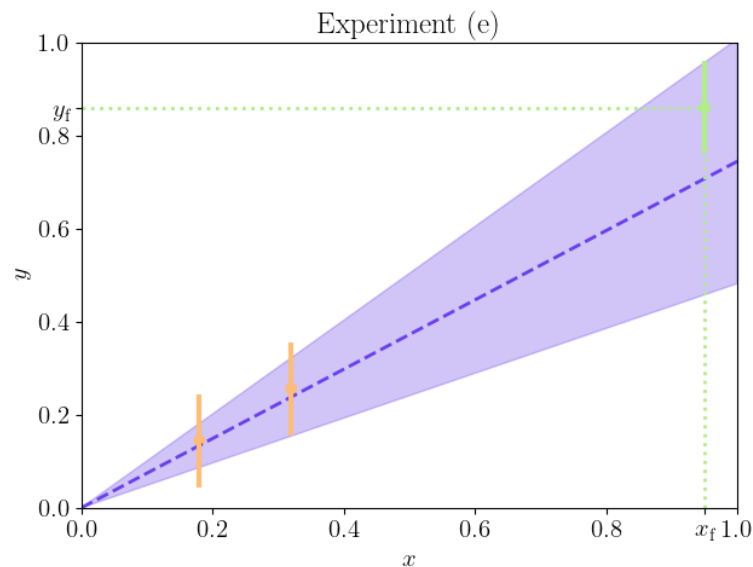
posterior of current
experiment

- More on Bayesian experimental design after we have studied information-theoretic measures of entropy and information.

Bayesian experimental design: example

Exercise: Bayesian linear model – Bayesian experimental design

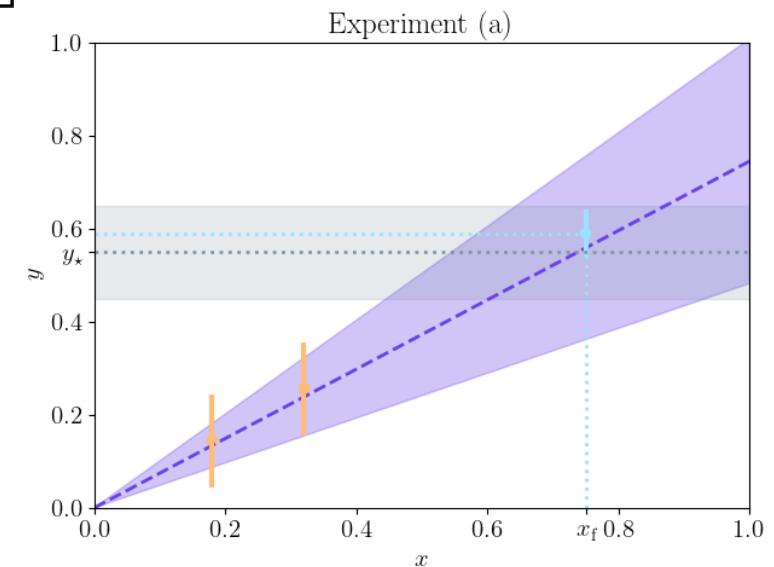
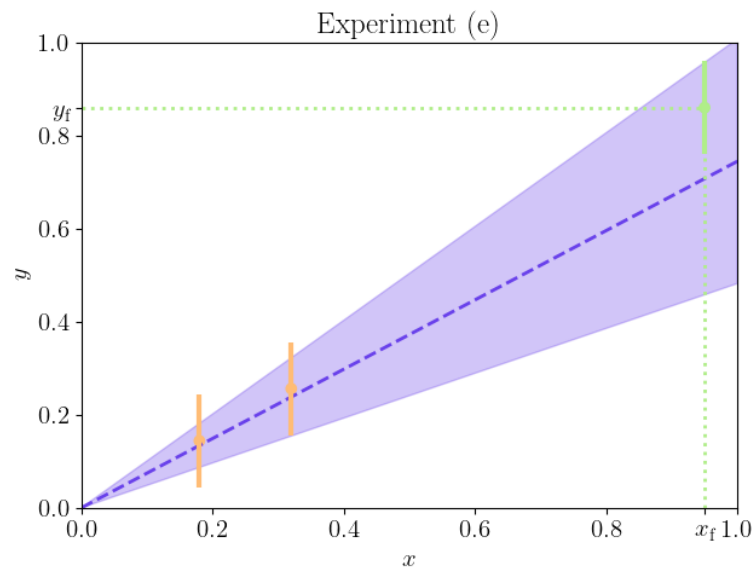
- Model: $y = mx$. We want to measure the slope m of this relationship.
- We have measured two points y_0 and y_1 with error σ at two locations x_0 and x_1 .
- We now have the choice between two (equally expensive) experiments:
 - Instrument (e): As accurate as today's instrument, will measure y_f at a much larger value x_f (so as to increase the lever arm in the measurement of the slope)
 - Instrument (a): Much more accurate instrument, but built so as to have a “sweet spot” at a certain value of y , called y_* , and much less accurate elsewhere.



Trotta et al., in Bayesian Methods in Cosmology (2010), chap. 5

Bayesian experimental design: example

- Which instrument should we go for? The answer should probably depend on how good our current knowledge of m is. Is the current uncertainty on m small enough to target accurately enough $x = x_*$ so that we get to the sweet spot $y_* = mx_*$?
- We can use for the utility of the inverse variance of the future posterior on m and assume for the noise levels of instrument a the toy model:
$$\tau_a^2 = \tau_*^2 \exp \left[\frac{(y - y_*)^2}{2\Delta^2} \right]$$
 where Δ is the width of the sweet spot.



Bayesian experimental design: example

- If we take a prior for m centred on zero with unit variance, the posterior of the current experiment, $p(m|o)$, is Gaussian with:

$$\text{mean: } \bar{m} \equiv \frac{x_0 y_0 + x_1 y_1}{\sigma^2 + x_0^2 + x_1^2} \quad \text{inverse variance: } F \equiv 1 + \frac{x_0^2 + x_1^2}{\sigma^2}$$

- After adding an independent data point at x_f with variance $\tau^2(x_f)$, the posterior of the next experiment is Gaussian with inverse variance: $F + \frac{x_f^2}{\tau^2(x_f)}$ which we choose as utility function.

- For experiment (e), the utility does not depend on the parameter, i.e.

$$U(e|o) = \int U(e, m|o) p(m|o) dm \quad \text{and the noise is constant, i.e. } \tau_e(x_f) = \tau_e$$

Therefore, $U(e|o) = F + \frac{x_f^2}{\tau_e^2}$. Maximising the utility is equivalent to maximising x_f (i.e. using the maximum lever arm possible).

Bayesian experimental design: example

- For experiment (a), we use the noise model

$$\tau_a^2(x_f) = \tau_\star^2 \exp \left[\frac{(y - y_\star)^2}{2\Delta^2} \right] = \tau_\star^2 \exp \left[\frac{(mx_f - y_\star)^2}{2\Delta^2} \right]$$

where Δ is the width of the sweet spot, using $y = mx_f$.

- The utility is $U(a, m|o) = F + \frac{x_f^2}{\tau_\star^2} \exp \left[-\frac{1}{2} \frac{(mx_f - y_\star)^2}{\Delta^2} \right]$

The expected utility is $U(a|o) = \int \left\{ F + \frac{x_f^2}{\tau_\star^2} \exp \left[-\frac{1}{2} \frac{(mx_f - y_\star)^2}{\Delta^2} \right] \right\} p(m|o) dm$

with $p(m|o) \propto \exp \left[-\frac{1}{2} F(m - \bar{m})^2 \right] \equiv \exp \left[-\frac{1}{2} \frac{(m - \bar{m})^2}{\Sigma^2} \right]$ and $\Sigma \equiv \frac{1}{\sqrt{F}}$

- The result is

$$U(a|o) = F + \frac{x_f^2}{\tau_\star^2} \exp \left[-\frac{1}{2} \frac{(\bar{m}x_f - y_\star)^2}{\Delta^2 + \Delta_y^2} \right] \frac{\Delta}{\sqrt{\Delta^2 + \Delta_y^2}} \quad \text{where } \Delta_y \equiv \Sigma x_f \text{ is the uncertainty at } x_f.$$

Bayesian experimental design: example

- In the case where $\Delta \gg \Delta_y$, maximising $U(a|o)$ is equivalent to maximising

$$\frac{x_f^2}{\tau_a^2(x_f)} = \frac{x_f^2}{\tau_\star^2} \exp \left[-\frac{1}{2} \frac{(\bar{m}x_f - y_\star)^2}{\Delta^2} \right]$$

- The solution is

$$x_f = \frac{y_\star}{\bar{m}} \frac{1 + \sqrt{1 + 8\Delta^2/y_\star^2}}{2}$$

- This is different from minimising the noise $\tau_a(x_f)$ which would have given

$$x_f = \frac{y_\star}{\bar{m}} \cdot \text{The term } \frac{1 + \sqrt{1 + 8\Delta^2/y_\star^2}}{2} \geq 1$$

increases the lever arm while staying in the “sweet spot” of experiment (a).

- In the case where $\Delta \ll \Delta_y$, maximising $U(a|o)$ is equivalent to maximising

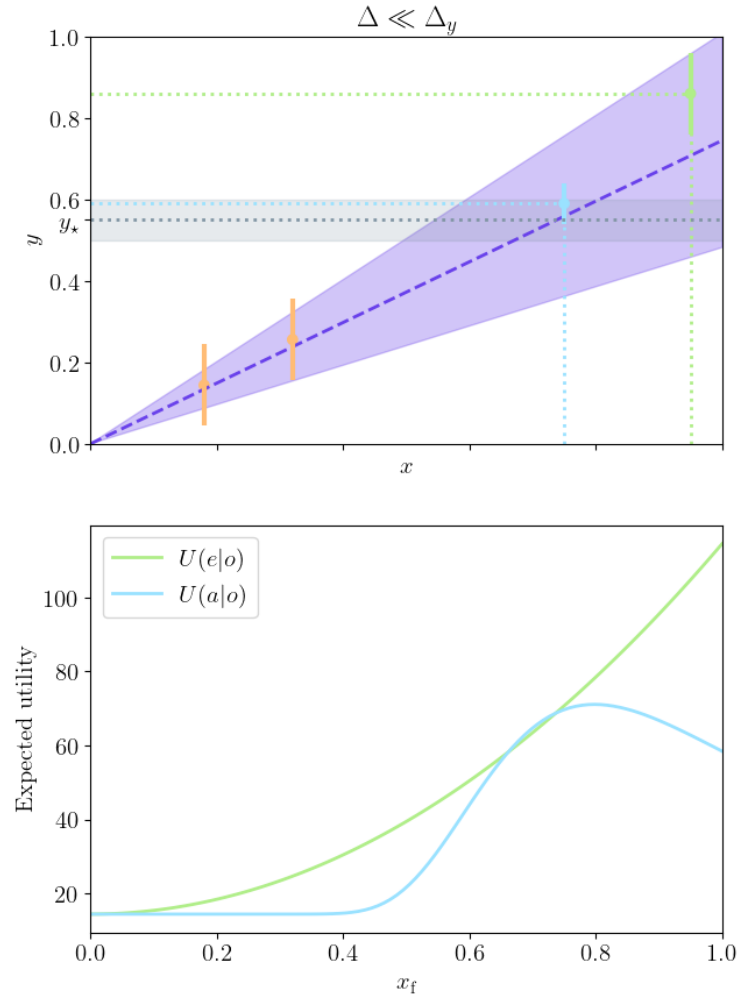
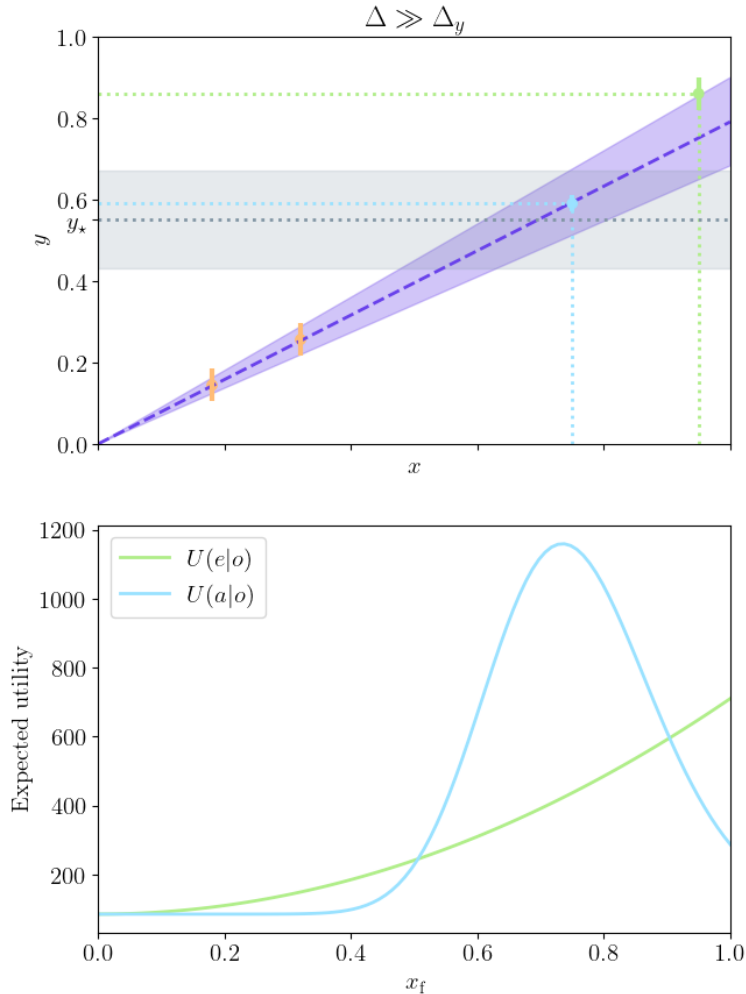
$$\frac{x_f^2}{\tau_\star^2} \exp \left[-\frac{1}{2} \frac{(\bar{m}x_f - y_\star)^2}{\Sigma^2 x_f^2} \right] \frac{\Delta}{\Sigma x_f}$$

- The solution is

$$x_f = \frac{y_\star}{\bar{m}} \frac{1 + \sqrt{1 - 4\Sigma^2/\bar{m}^2}}{2\Sigma^2/\bar{m}^2}$$

- x_f is real if $\bar{m} \geq 2\Sigma$, i.e. the slope is measured from current data with an accuracy better than 2Σ .
- If this is not the case, $U(a|o)$ is a monotonically increasing function of x_f , so it is maximised by maximising x_f , even if it means carrying out a very poor measurement ($\tau_a \rightarrow +\infty$ as $x_f \rightarrow +\infty$).

Bayesian experimental design: example



Interpretation: Designating an experiment that exploits a “sweet spot” is only feasible if our current uncertainty on the parameter to be measured is small enough compared to the “sweet spot” window of opportunity.

[Trotta et al., in Bayesian Methods in Cosmology \(2010\), chap. 5](#)

03

BAYESIAN NETWORKS AND HIERARCHICAL MODELS



BAYESIAN HIERARCHY: AN EXAMPLE

Bayesian hierarchy from latent variables

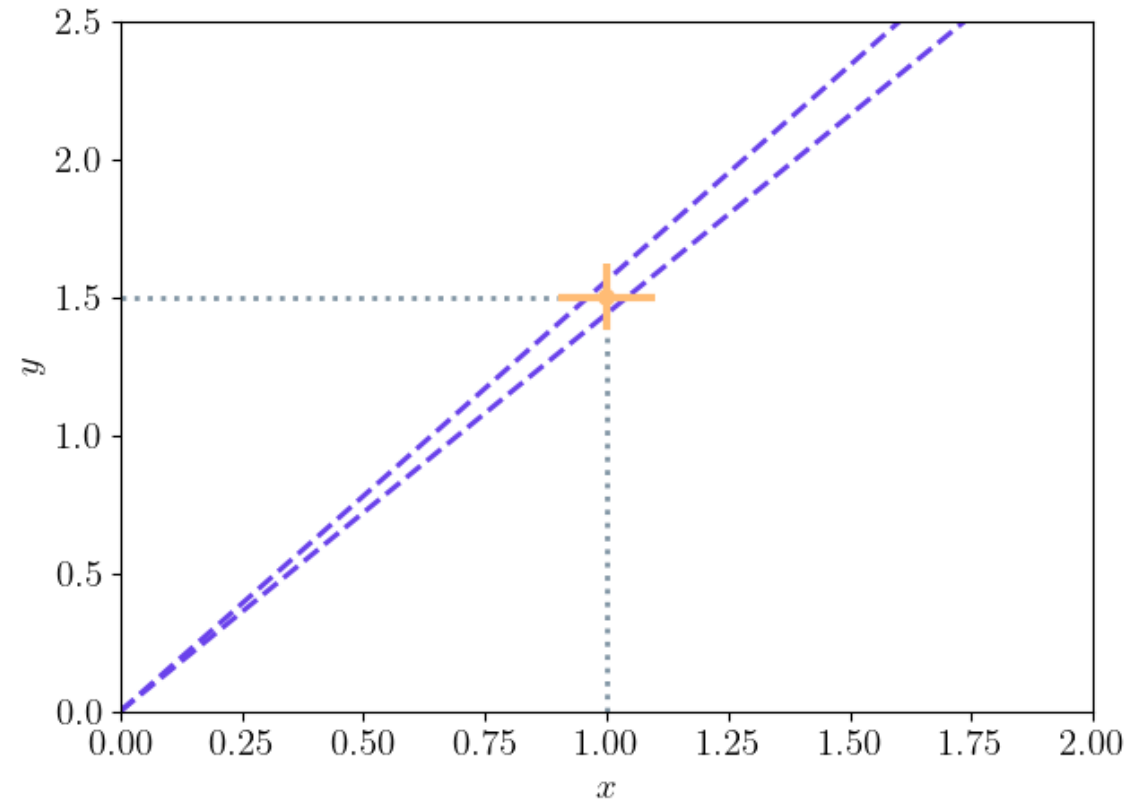
Exercise: Bayesian linear model – Bayesian hierarchical model

- Model: $y = mx$
- We measure X, Y , but they both have measurement errors. What is the posterior for the slope m ?
- Applying the first rule (“write down what you want to know”): we want to know $p(m|X, Y)$
- There are two unknown (“latent”) variables in the problem: the true values x, y
- Full joint pdf of the problem:

$$p(m, x, y, X, Y)$$

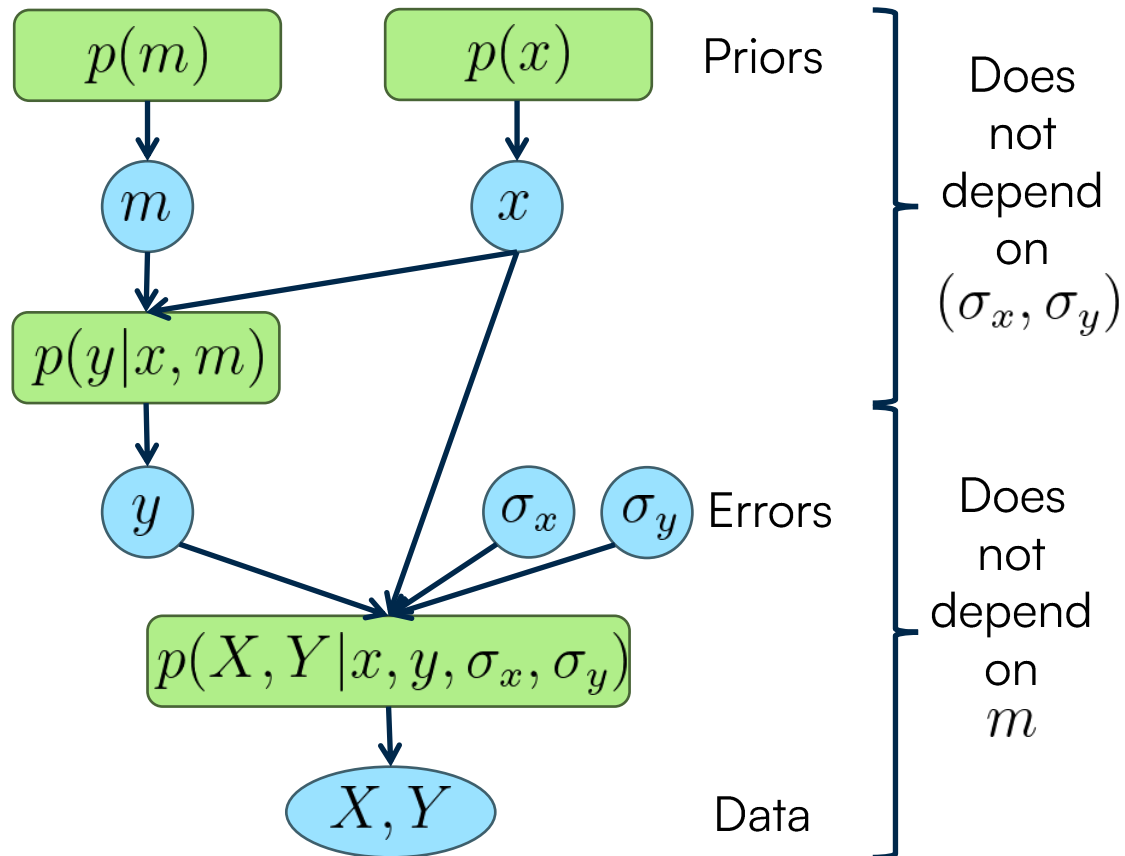
- Joint pdf of the target and observed variables:

$$p(m, X, Y) = \int p(m, x, y, X, Y) \, dx \, dy$$



Building the statistical model

- We construct a forward (generative) model of the data graphically:



- Apply Bayes' theorem:

$$p(m|X, Y) \propto p(X, Y|m)p(m)$$

- Introduce the latent variables and marginalise:

$$p(m|X, Y) \propto \iint p(X, Y, x, y|m)p(m) dx dy$$

- Expand first probability with the product rule:

$$p(m|X, Y) \propto \iint p(X, Y|x, y, m)p(x, y|m)p(m) dx dy$$

- Expand second probability with the product rule:

$$p(m|X, Y) \propto \iint p(X, Y|x, y, m)p(y|x, m)p(x|m)p(m) dx dy$$

- Simplify conditional dependencies:

$$p(X, Y|x, y, m) = p(X, Y|x, y)$$

$$p(x|m) = p(x)$$

- Apply physical relation: $p(y|x, m) = \delta_D(y - mx)$

- Integrate to get the final result:

$$p(m|X, Y) \propto \int p(X, Y|x, mx)p(x)p(m) dx$$

Inferring the slope

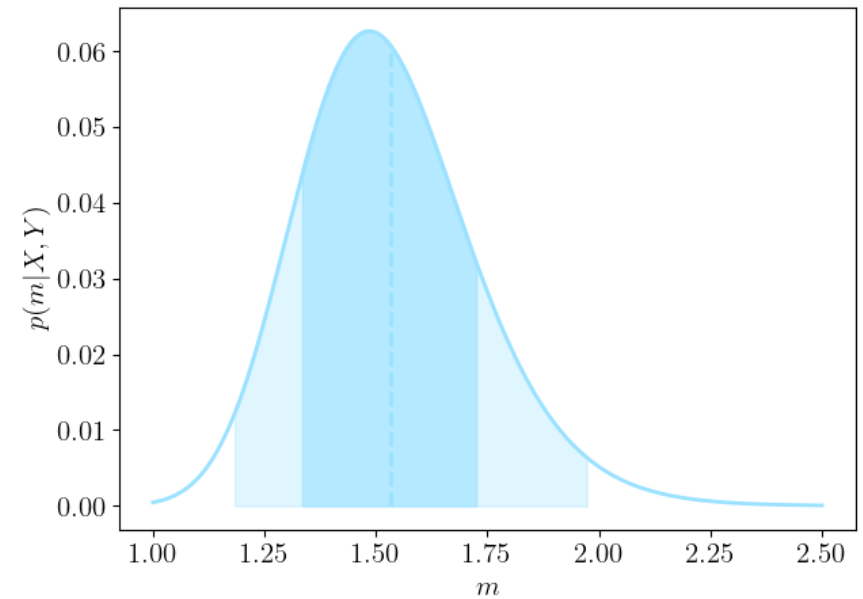
$$p(m|X, Y) \propto \int p(X, Y|x, mx)p(x)p(m) dx$$

- If the error distribution is Gaussian with zero mean, and if we take uniform priors on x and m :

$$p(m|X, Y) \propto \int_{-\infty}^{+\infty} e^{-\frac{1}{2} \frac{(X-x)^2}{\sigma_x^2}} e^{-\frac{1}{2} \frac{(Y-mx)^2}{\sigma_y^2}} dx$$

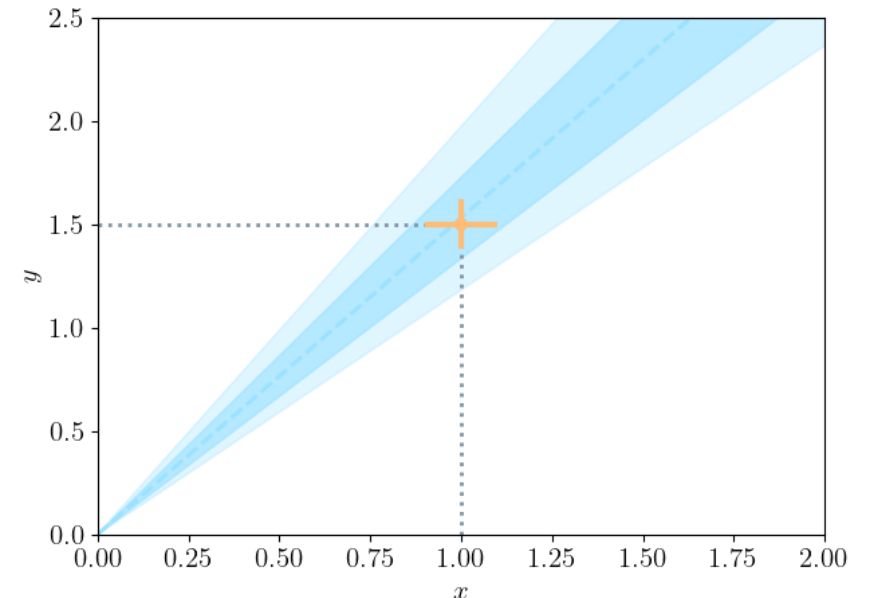
- Completing the square and integrating gives the marginal posterior for m :

$$p(m|X, Y) \propto \frac{\sigma_x \sigma_y}{\sqrt{\sigma_y^2 + m^2 \sigma_x^2}} \exp \left[-\frac{1}{2} \frac{(Y - mX)^2}{\sigma_y^2 + m^2 \sigma_x^2} \right]$$



$$X = 1.0, Y = 1.5$$

$$\sigma_x = 0.10, \sigma_y = 0.12$$



Inferring the full model and sampling

- The joint posterior for (x, m) is:

$$p(x, m|X, Y) \propto p(X, Y|x, m)p(x)p(m) \\ \propto e^{-\frac{1}{2}\frac{(X-x)^2}{\sigma_x^2}} e^{-\frac{1}{2}\frac{(Y-mx)^2}{\sigma_y^2}}$$

- At fixed x :

$$p(m|X, Y, x) \propto \exp \left[-\frac{1}{2} \frac{x^2(m - \frac{Y}{x})^2}{\sigma_y^2} \right]$$

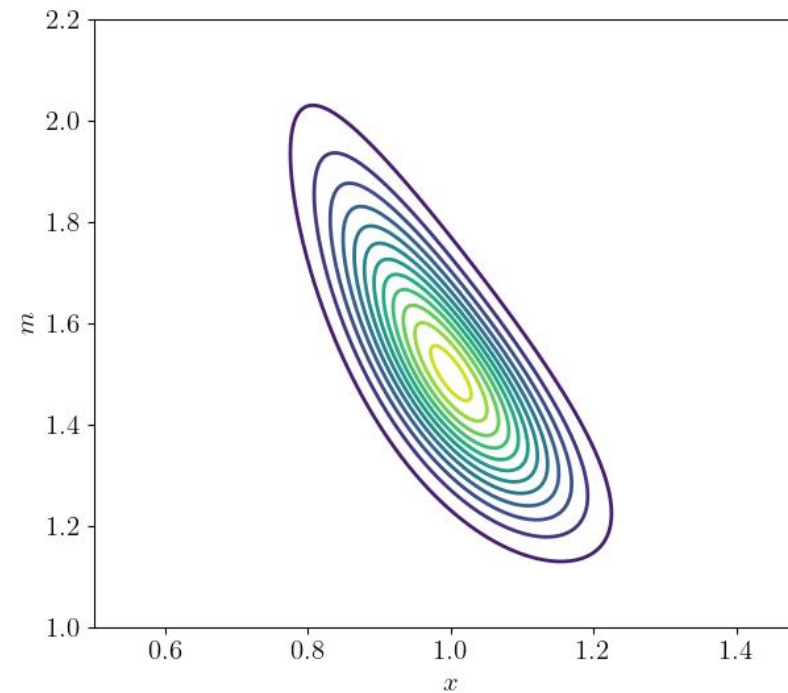
i.e. $p(m|X, Y, x) = \mathcal{G} \left(\frac{Y}{x}, \frac{\sigma_y^2}{x^2} \right)$

- At fixed m (combining the exponents and completing the square):

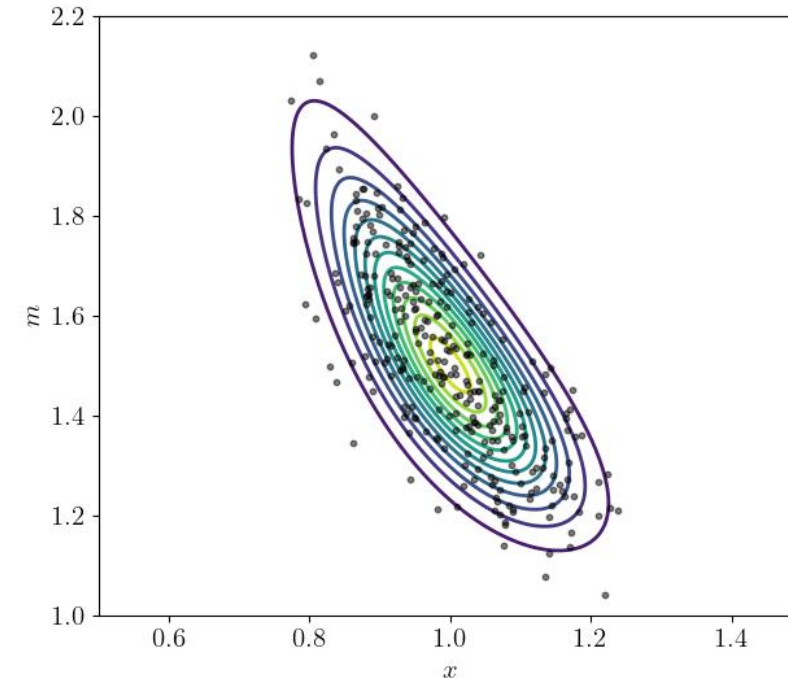
$$p(x|X, Y, m) = \mathcal{G} \left(\frac{\sigma_y^2 X + m\sigma_x^2 Y}{\sigma_y^2 + m^2\sigma_x^2}, \frac{\sigma_y^2\sigma_x^2}{\sigma_y^2 + m^2\sigma_x^2} \right)$$

- We can therefore use Gibbs sampling to draw samples from the joint posterior:

- $m \leftarrow p(m|X, Y, x)$
- $x \leftarrow p(x|X, Y, m)$



$$X = 1.0, Y = 1.5 \\ \sigma_x = 0.10, \sigma_y = 0.12$$



Bayesian hierarchical models and generalised linear regression

- At the heart of the method lies the fundamental problem of (generalised) [linear regression](#), in the presence of measurement errors on both the dependent and the independent variable and intrinsic scatter in the relationship.
- This is a general problem in any field dealing with objects with an intrinsic variability.
- The key parameter is the noise to population variance ratio, $r \equiv \frac{\sigma_x \sigma_y}{R_x}$. [March et al., 1102.3237](#)
 - For small r , the Bayesian marginal posterior on m is identical to the frequentist profile likelihood.
 - For large r , the Bayesian marginal posterior is broader but less biased than the profile likelihood.

- Model to be fitted:

$$y = mx + b$$

- Statistical model:

$$x_i \sim p(x|R_x) = \mathcal{G}(\mu_x, R_x) \quad \text{Population distribution}$$

$$y_i|x_i \sim \mathcal{G}(mx_i + b, R_y) \quad \text{Intrinsic variability}$$

$$X_i, Y_i|x_i, y_i \sim \mathcal{G}([x_i, y_i], C) \quad \text{Measurement error}$$

$$\text{usually } C = \begin{pmatrix} \sigma_x^2 & 0 \\ 0 & \sigma_y^2 \end{pmatrix}$$



BAYESIAN HIERARCHICAL MODELS

Bayesian hierarchical models for adapting the prior

- Simple Bayesian inference:

$$p(\theta|d) \propto p(d|\theta) \overset{\text{prior}}{p(\theta)}$$

- Inference with an adaptive prior depending on a latent variable:

$$p(\theta|d) \propto p(d|\theta) \overset{\text{prior}}{p(\theta|\eta)} \overset{\text{hyperprior}}{p(\eta)}$$

- ... or a full hierarchy of hyperpriors.

Examples:

- Cosmic microwave background:

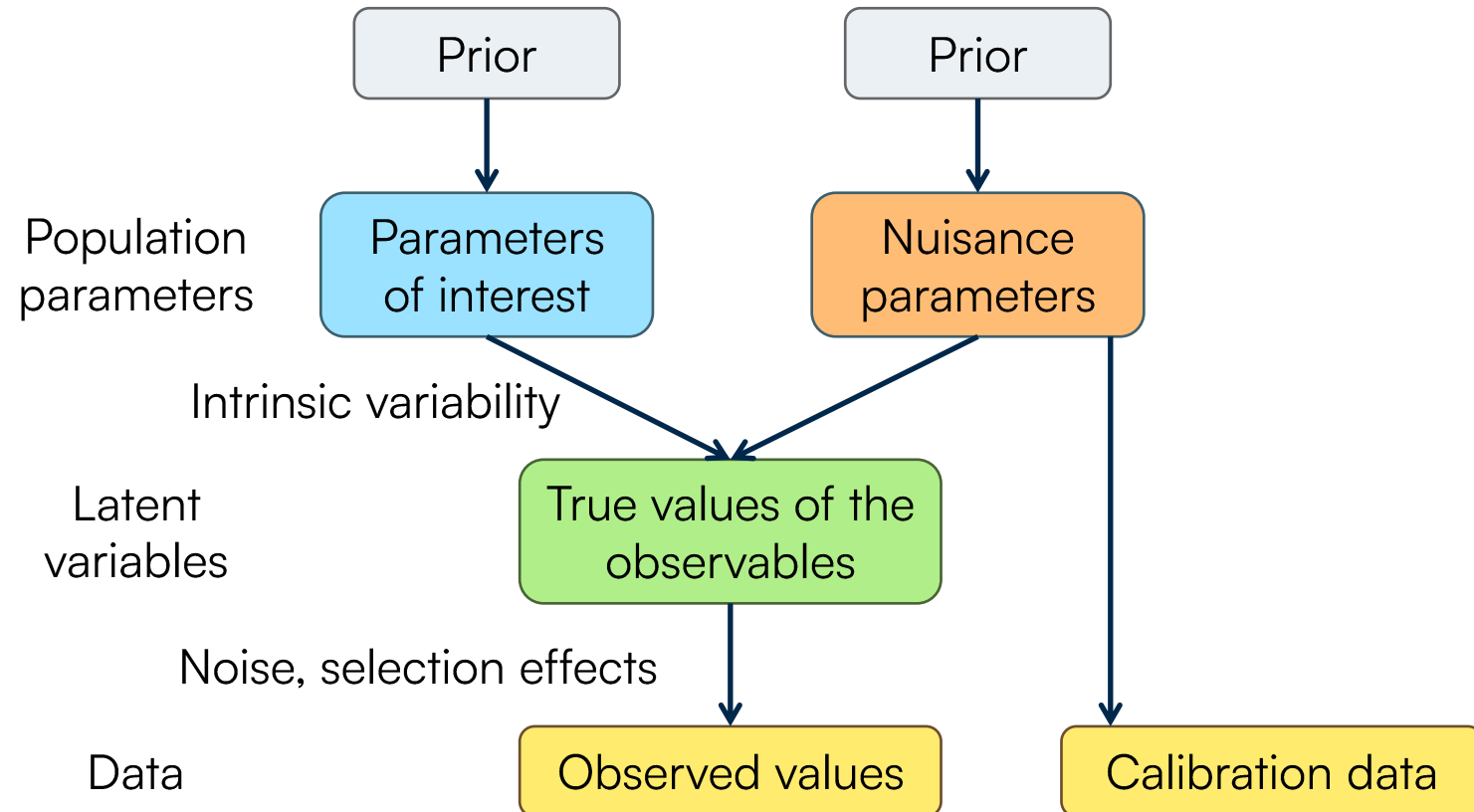
$$p(\{\Omega\}, \{C_\ell\}, s|d) \propto p(d|s) p(s|\{C_\ell\}) p(\{C_\ell\}|\{\Omega\}) p(\{\Omega\})$$

- Large-scale structure:

$$p(\{\Omega\}, \phi, g|d) \propto p(d|g) p(g|\phi) p(\phi|\{\Omega\}) p(\{\Omega\})$$

Many sources of variability

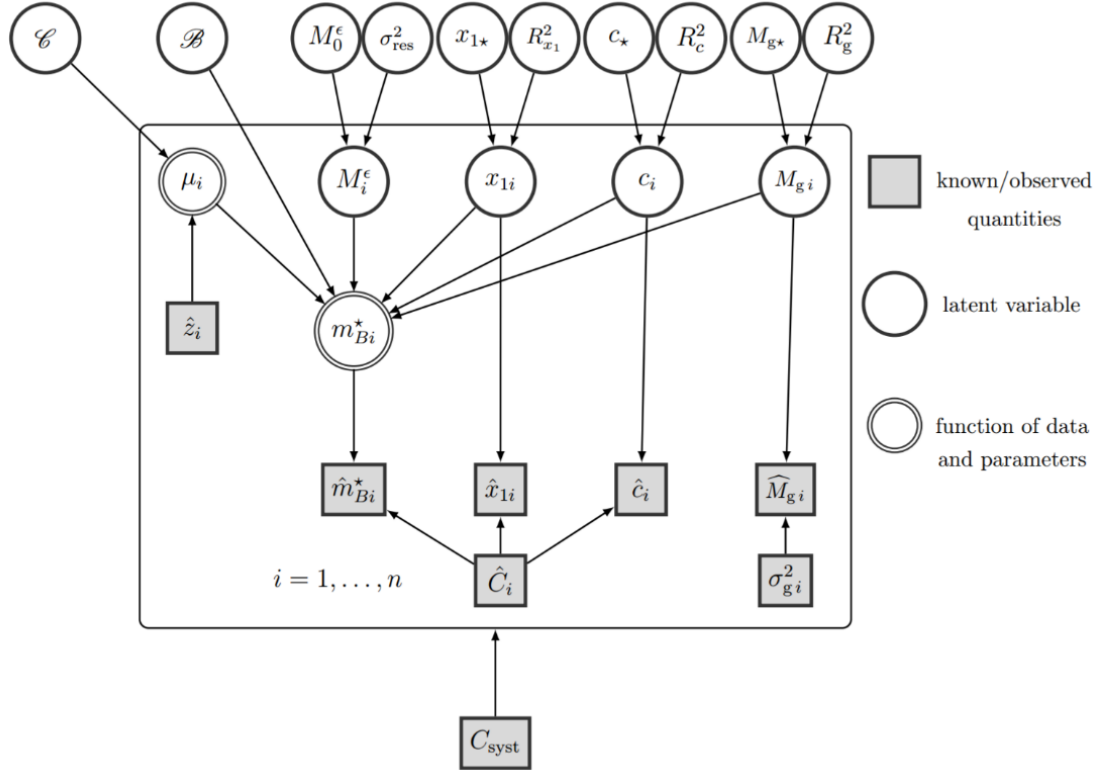
- You pick a lightbulb, and measure its brightness. What is it?
- There are many reasons why the value might vary:
 - It's picked from a box of bulbs of different brightnesses
 - The manufacturing process is imprecise
 - Measurement error
- *Any or all* of these may apply (and you may not know which).



Bayesian hierarchical models for complex problems

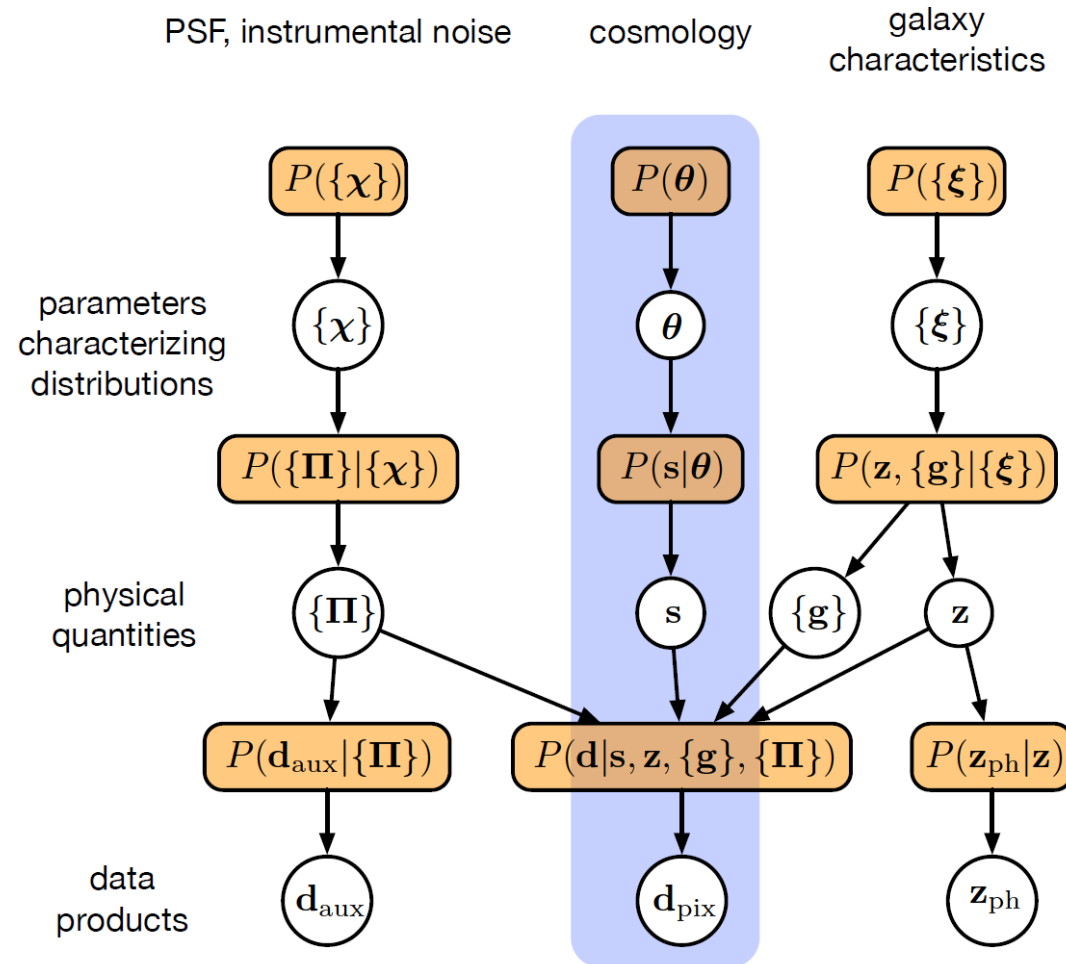
- How can we make sure all the errors are propagated correctly to the posterior?
- We split the inference problem into steps, where the full model is made up of **a series of sub-models**. The aim is to build a complete model of the data. It is a principled way to include systematic errors, selection effects, etc. (everything, really).
- The Bayesian Hierarchical Model (BHM) links the sub-models together, correctly **propagating uncertainties** in each sub-model from one level to the next.
- It also exposes what you need to know or assume. At each step you will (ideally) know the **conditional distributions**.
- All of the steps give rise to “**latent variables**”: parameters in sub-models, usually not of interest.
 - A particular sort of “nuisance parameter”
 - They still need to be accounted for (and marginalised over)
 - e.g., contribution of systematic error to a measurement
 - e.g., galactic dust flux in a noisy CMB pixel
 - These might very well be “signal” for a different purpose.
- Therefore, BHMs may have **very many parameters**.
- When you are using **sampling** for inference, **marginalisation** is “trivial”: just ignore those variables in the output.
 - Realistically, of course, there is usually some information there!

BHM example: supernova cosmology (BAHAMAS)



Parameter	Notation and Prior Distribution
Cosmological parameters	
Matter density parameter	$\Omega_m \sim \text{UNIFORM}(0, 2)$
Cosmological constant density parameter	$\Omega_\Lambda \sim \text{UNIFORM}(0, 2)$
Dark energy EOS	$w \sim \text{UNIFORM}(-2, 0)$
Hubble parameter	$H_0/\text{km/s/Mpc} = 67.3$
Covariates	
Coefficient of stretch covariate	$\alpha \sim \text{UNIFORM}(0, 1)$
Coefficient of color covariate	β (or β_0) $\sim \text{UNIFORM}(0, 4)$
Coefficient of interaction of color correction and z	$\beta_1 \sim \text{UNIFORM}(-4, 4)$
Jump in coefficient of color covariate	$\Delta\beta \sim \text{UNIFORM}(-1.5, 1.5)$
Redshift of jump in color covariate	$z_t \sim \text{UNIFORM}(0.2, 1)$
Coefficient of host galaxy mass covariate	$\gamma \sim \text{UNIFORM}(-4, 4)$
Population-level distributions	
Mean of absolute magnitude	$M_0^\epsilon \sim \mathcal{N}(-19.3, 2^2)$
Residual scatter after corrections	$\sigma_{\text{res}}^2 \sim \text{INV GAMMA}(0.003, 0.003)$
Mean of absolute magnitude, low galaxy mass	$M_0^{\text{lo}} \sim \mathcal{N}(-19.3, 2^2)$
SD of absolute magnitude, low galaxy mass	$\sigma_{\text{res}}^{\text{lo}^2} \sim \text{INV GAMMA}(0.003, 0.003)$
Mean of absolute magnitude, high galaxy mass	$M_0^{\text{hi}} \sim \mathcal{N}(-19.3, 2^2)$
SD of absolute magnitude, high galaxy mass	$\sigma_{\text{res}}^{\text{hi}^2} \sim \text{INV GAMMA}(0.003, 0.003)$
Mean of stretch	$x_{1*} \sim \mathcal{N}(0, 10^2)$
SD of stretch	$R_{x_1} \sim \text{LOG UNIFORM}(-5, 2)$
Mean of color	$c_* \sim \mathcal{N}(0, 1^2)$
SD of color	$R_c \sim \text{LOG UNIFORM}(-5, 2)$
Mean of host galaxy mass	$M_{g*} \sim \mathcal{N}(10, 100^2)$
SD of host galaxy mass	$R_g \sim \text{LOG UNIFORM}(-5, 2)$

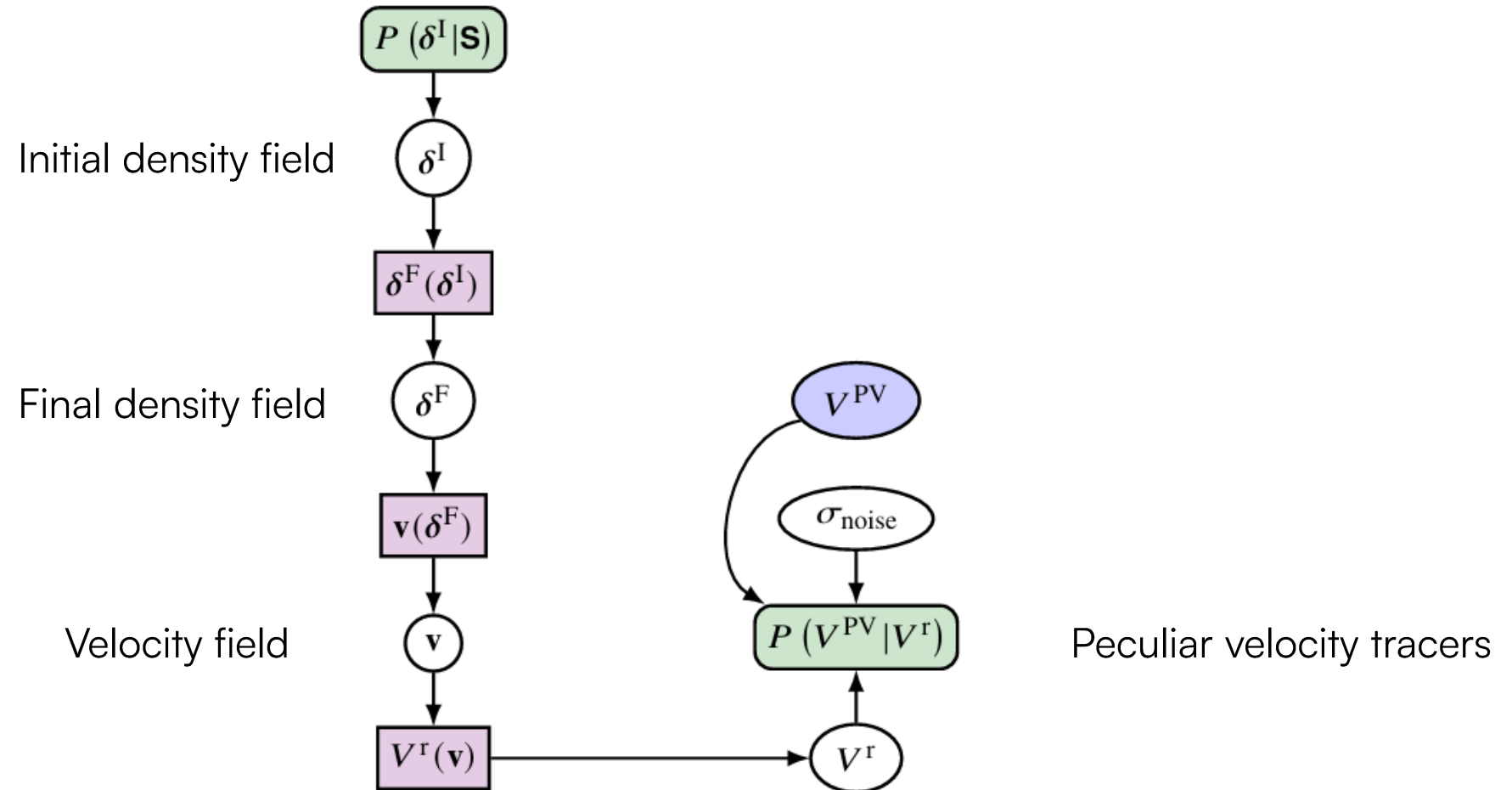
BHM example: weak lensing



Can include:

- Mask
- Intrinsic alignments
- Baryon feedback
- Shape measurement
- Photometric redshifts

BHM example: large-scale structure inference from peculiar velocity tracers

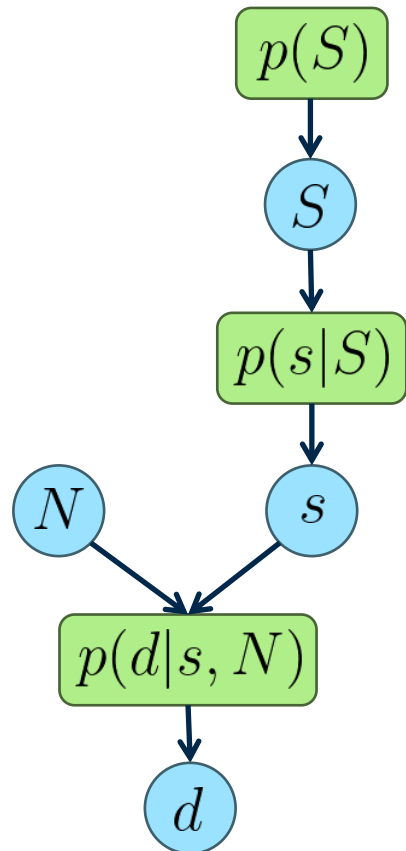


Back to Wiener filtering

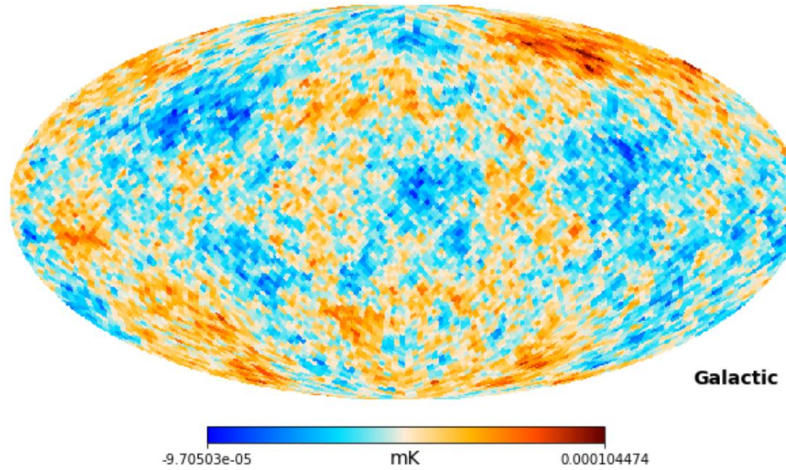
$$\mu_{s|d} = \mu_s + S(S + N)^{-1}(d - \mu_d)$$

$$C_{s|d} = S - S(S + N)^{-1}S$$

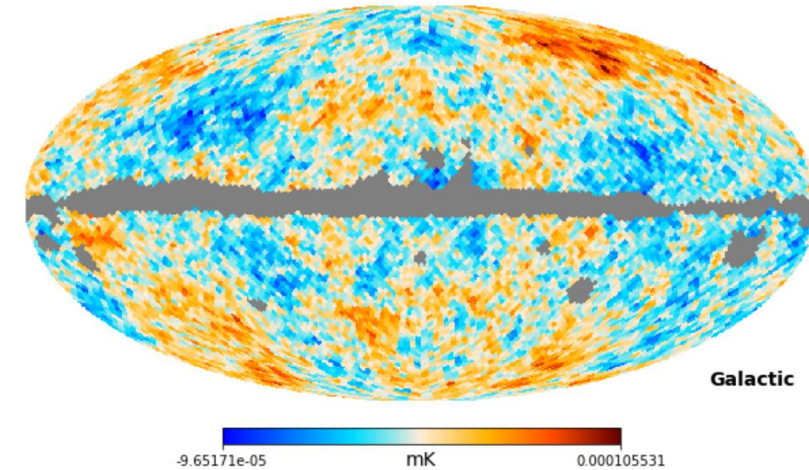
- As a BHM:



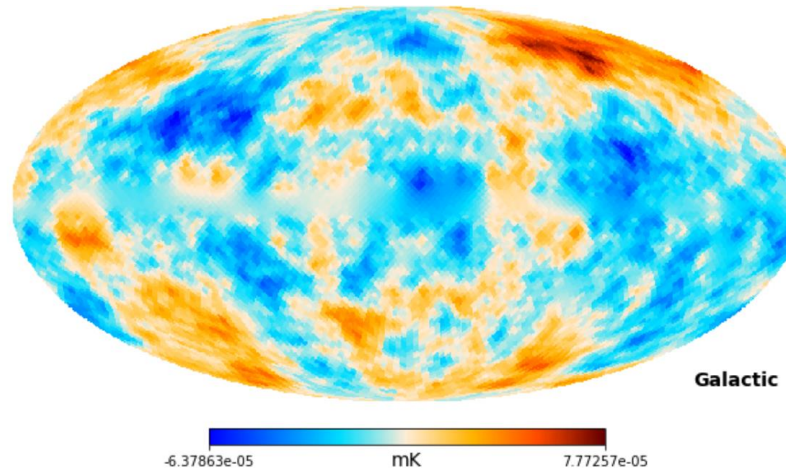
True signal



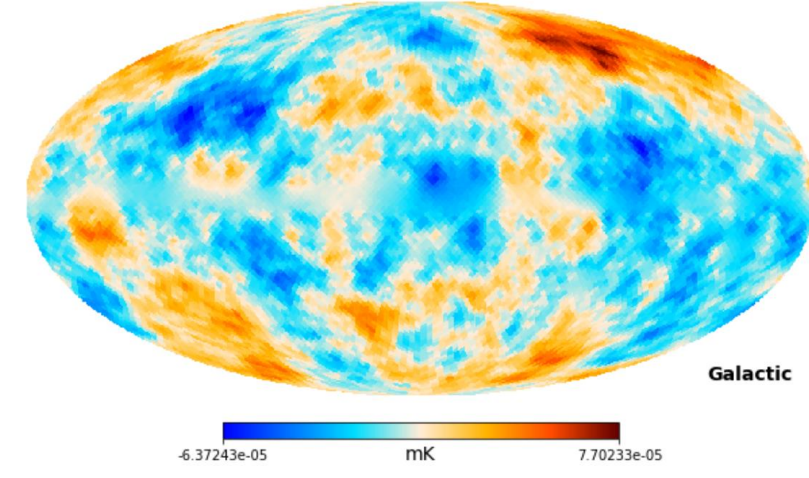
Simulated data d



Wiener filtered data (posterior mean)



One simulated signal

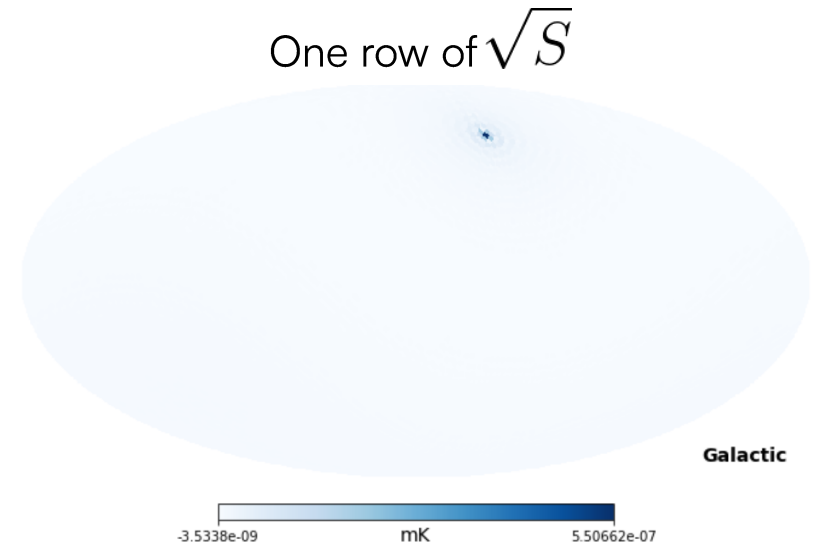
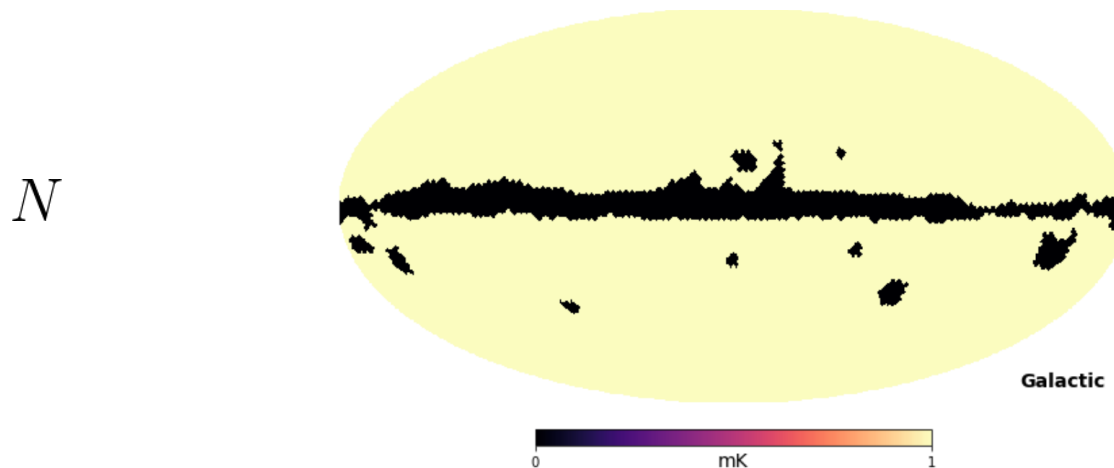
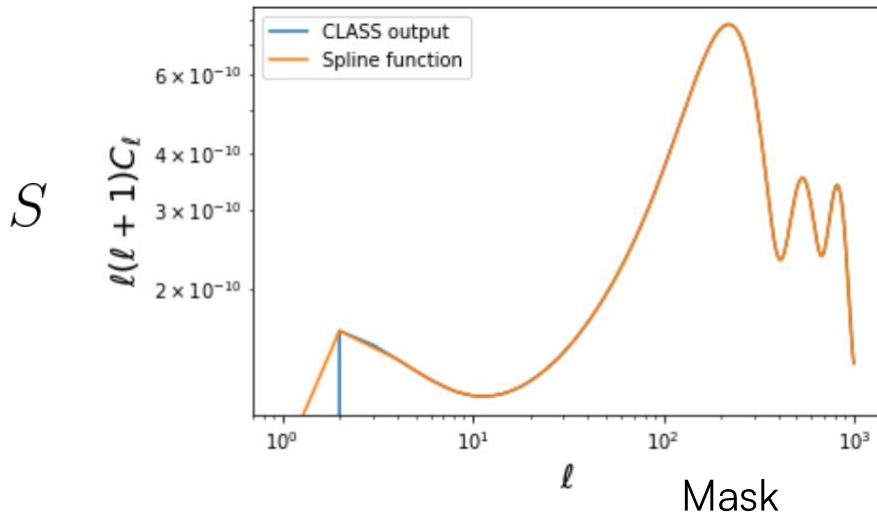


Back to Wiener filtering

$$\mu_{s|d} = \mu_s + S(S + N)^{-1}(d - \mu_d)$$

$$C_{s|d} = S - S(S + N)^{-1}S$$

- Problem: computing/representing $(S + N)^{-1}$ is difficult because S is sparse in harmonic/ Fourier space and N is sparse in configuration/real space.

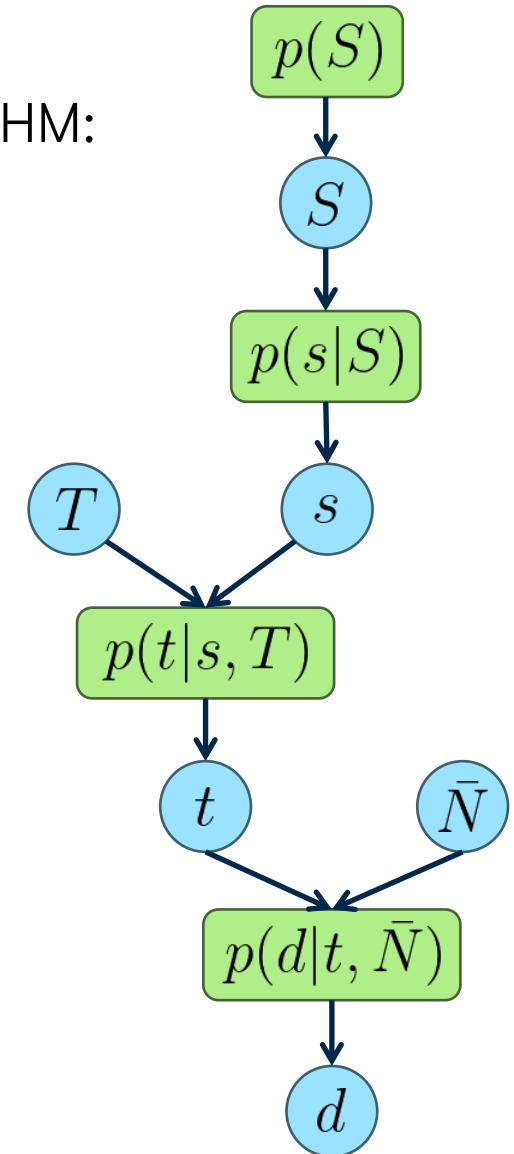


Messenger field and multivariate Wiener filtering

$$\mu_{s|d} = S(S + N)^{-1}d \quad (\text{assuming } \mu_s = \mu_d = 0)$$

$$C_{s|d} = S - S(S + N)^{-1}S$$

- As a BHM:



- Messenger field algorithm:

- Introduce an auxiliary Gaussian random field t with covariance matrix $T \equiv \tau I$.
- T (isotropic noise covariance matrix) is diagonal in any basis (harmonic/Fourier and configuration/real).
- Introduce $\bar{N} \equiv N - T$ (residual noise covariance matrix).

- Sampling:

- Goal: obtain samples of $p(s, t|d)$ via Gibbs sampling. We need the conditionals $p(s|d, t)$ and $p(t|s, d)$.
- $p(s|d, t) = p(s|t)$ is Gaussian with
 mean: $\mu_{s|t} = (S^{-1} + T^{-1})^{-1}T^{-1}t$ (assuming $\mu_s = \mu_t = 0$)
 covariance: $C_{s|t} = (S^{-1} + T^{-1})^{-1}$
- $p(t|s, d) \propto p(t|s)p(d|t)$ is Gaussian with
 mean: $\mu_{t|s,d} = (T^{-1} + \bar{N}^{-1})^{-1}T^{-1}s + (T^{-1} + \bar{N}^{-1})^{-1}\bar{N}^{-1}d$
 covariance: $C_{t|s,d} = (T^{-1} + \bar{N}^{-1})^{-1}$

Bayesian hierarchical models: summary

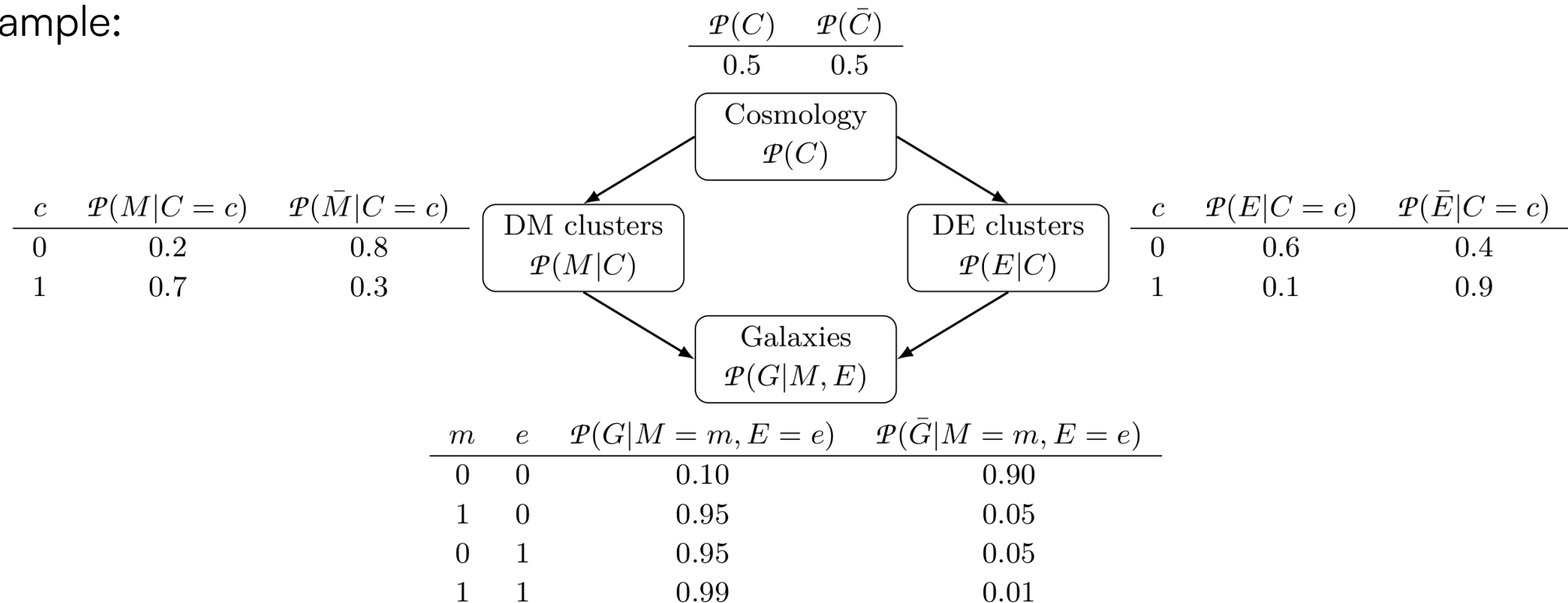
- BHMs are a way to build a statistical model of data by splitting the problem into steps.
- Decomposing into steps exposes what is needed — typically many **conditional distributions**.
- For complex experiments, this may be the only viable way to build the statistical model of the data.
- The decomposition is usually very natural and logical.
- The model allows the proper **propagation of errors** from one layer to the next, including a proper treatment of systematics.
- One can often use efficient **sampling** algorithms to sample from the posterior — precisely what one wants for a Bayesian statistical analysis.



BAYESIAN NETWORKS

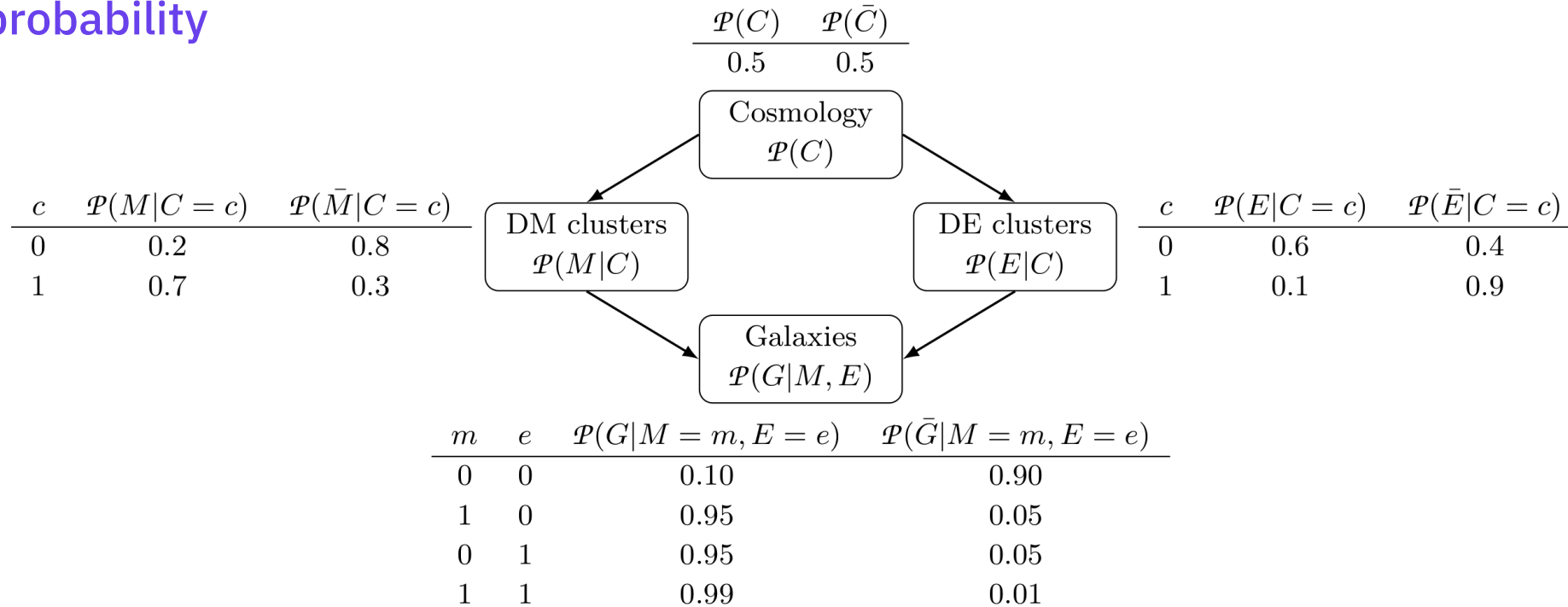
Bayesian networks

- Bayesian networks are probabilistic graphical models consisting of:
 - A directed acyclic graph (DAG)
 - At each node, conditional probabilities distributions
- Difference with Bayesian hierarchical models (for some authors): the graph can have “diamonds”
- Example:



Bayesian networks: example

Full joint probability



- The graph can be used to simplify conditional probability dependencies easily:

$$p(C, M, E, G) = p(C) p(E|C) p(M|C, \cancel{E}) p(G|\cancel{C}, M, E)$$

$$p(C, M, E, G) = p(C) p(E|C) p(M|C) p(G|M, E)$$

Bayesian networks: example

Inference and prediction

- Inference:

$$p(M|G) = \frac{p(M,G)}{p(G)} = \frac{\sum_{c,e} p(C=c, M=1, E=e, G=1)}{\sum_{c,m,e} p(C=c, M=m, E=e, G=1)} = \frac{0.4313}{0.70305} \approx 0.6135$$

$$p(E|G) = \frac{p(E,G)}{p(G)} = \frac{\sum_{c,m} p(C=c, M=m, E=1, G=1)}{\sum_{c,m,e} p(C=c, M=m, E=e, G=1)} = \frac{0.3363}{0.70305} \approx 0.4783$$

$$p(\bar{M}, \bar{E}|G) = \frac{p(\bar{M}, \bar{E}, G)}{p(G)} = \frac{\sum_c p(C=c, M=0, E=0, G=1)}{\sum_{c,m,e} p(C=c, M=m, E=e, G=1)} = \frac{0.0295}{0.70305} \approx 0.0420$$

- Prediction:

$$p(G|C) = \frac{p(G,C)}{p(C)} = \frac{\sum_{m,e} p(C=1, M=m, E=e, G=1)}{p(C=1)} = 0.7233$$

Bayesian networks: example

The “explaining away” phenomenon

$$p(E|M, G) = \frac{p(E, M, G)}{p(M, G)} = \frac{\sum_c p(C=c, M=1, E=1, G=1)}{\sum_{c,e} p(C=c, M=1, E=e, G=1)} = \frac{0.09405}{0.4313} \approx 0.2181$$

$$p(E|G) = \frac{p(E, G)}{p(G)} = \frac{\sum_{c,m} p(C=c, M=m, E=1, G=1)}{\sum_{c,m,e} p(C=c, M=m, E=e, G=1)} = \frac{0.3363}{0.70305} \approx 0.4783$$

- So we have both:

$$p(E|M) = p(E)$$

$$p(E|M, G) < p(E|G)$$

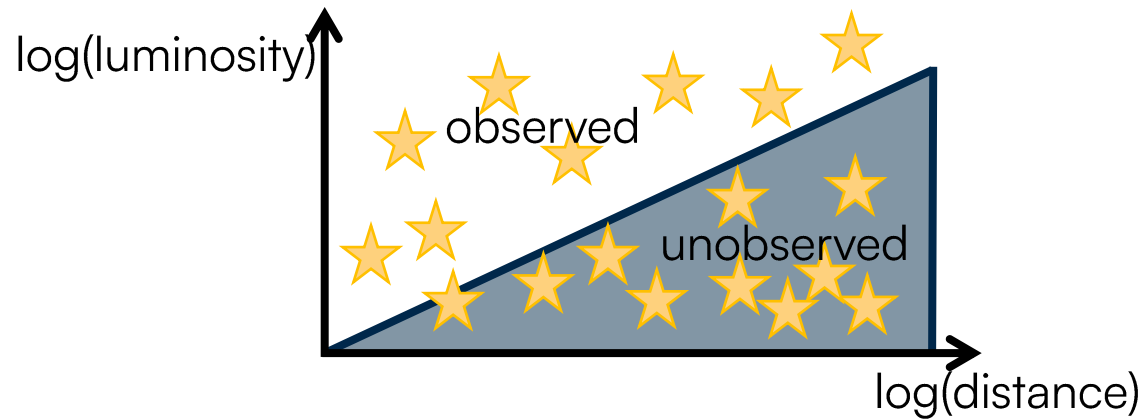
- This is “collider bias” or the “explaining away” phenomenon: two causes collide to explain the same effect.
- Particular case: “selection bias” or “Berkson’s paradox”:

$$0 < p(A) < 1; \quad 0 < p(B) < 1; \quad p(A|B) = p(A)$$

$$\Rightarrow \left\{ \begin{array}{l} p(A|B, C) < p(A|C) \\ \text{and} \\ p(A|\bar{B}, C) = 1 > p(A|C) \end{array} \right. \quad \text{with } C = A + B$$

Malmquist bias

- Malmquist bias: in magnitude-limited surveys, far objects are preferentially detected if they are intrinsically bright.



Gunnar Malmquist (1893-1982)

$$0 < p(A) < 1; \quad 0 < p(B) < 1; \quad p(A|B) = p(A)$$

$$\begin{array}{ccc} \nearrow & \nearrow & \nearrow \\ C = A + B & \xrightarrow{\text{green arrow}} & p(A|\bar{B}, C) = 1 > p(A|C) \\ \text{detected} & \text{bright} & \text{close} \end{array}$$

Malmquist (1922); Malmquist (1925)



EMPIRICAL BAYES

Empirical Bayes

An alternative to maximum entropy for choosing priors

$$p(\theta|d) \propto p(d|\theta) \overset{\text{prior}}{p(\theta|\eta)} \overset{\text{hyperprior}}{p(\eta)}$$

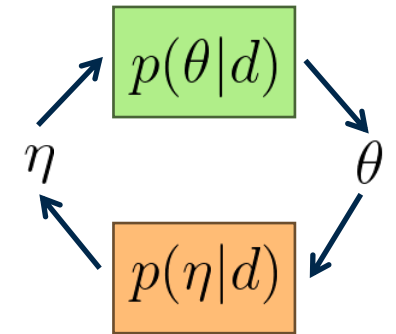
$$\underline{p(\theta|d)} = \int p(\theta|\eta, d) p(\eta|d) d\eta = \int \frac{p(d|\theta) p(\theta|\eta)}{p(d|\eta)} \underline{p(\eta|d)} d\eta$$

$$\underline{p(\eta|d)} = \int p(\eta|\theta) \underline{p(\theta|d)} d\theta$$

- Iterative scheme (“Gibbs” sampler) to calibrate the hyperprior from the data:
- [Empirical Bayes](#) is a truncation of this scheme after a few steps (often just one).
- Particular case: the [Expectation-Maximisation](#) (EM) algorithm (in machine learning, data mining)
 - η is evaluated using an estimator $\eta^*(d)$ given the data:

$$\underline{p(\eta|d)} \approx \delta_D(\eta - \eta^*(d)) \quad \Rightarrow \quad \underline{p(\theta|d)} \approx \frac{p(d|\theta) p(\theta|\eta^*)}{p(d|\eta^*)}$$

(maximisation) (expectation)



References and acknowledgements



References:

- A. Gelman et al. (2021), *Bayesian Data Analysis, Third edition*
- Trotta (2008), 0803.4089, *Bayes in the sky: Bayesian inference and model selection in cosmology*

- For their lectures, thanks to: Alan Heavens, Jonathan Pritchard, Elena Sellentin, Roberto Trotta

<https://florent-leclercq.eu/teaching.php>