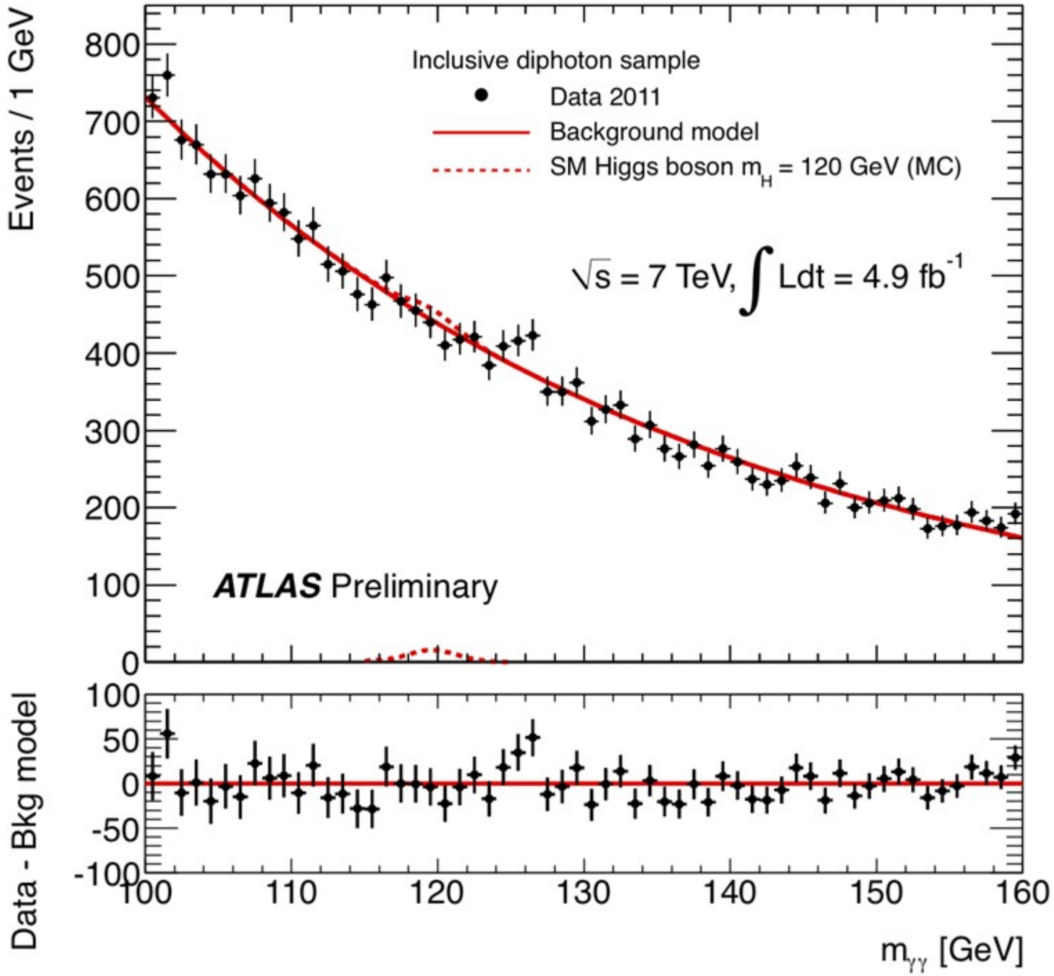
Bayesian model comparison

ICIC Data Analysis Workshop 2016

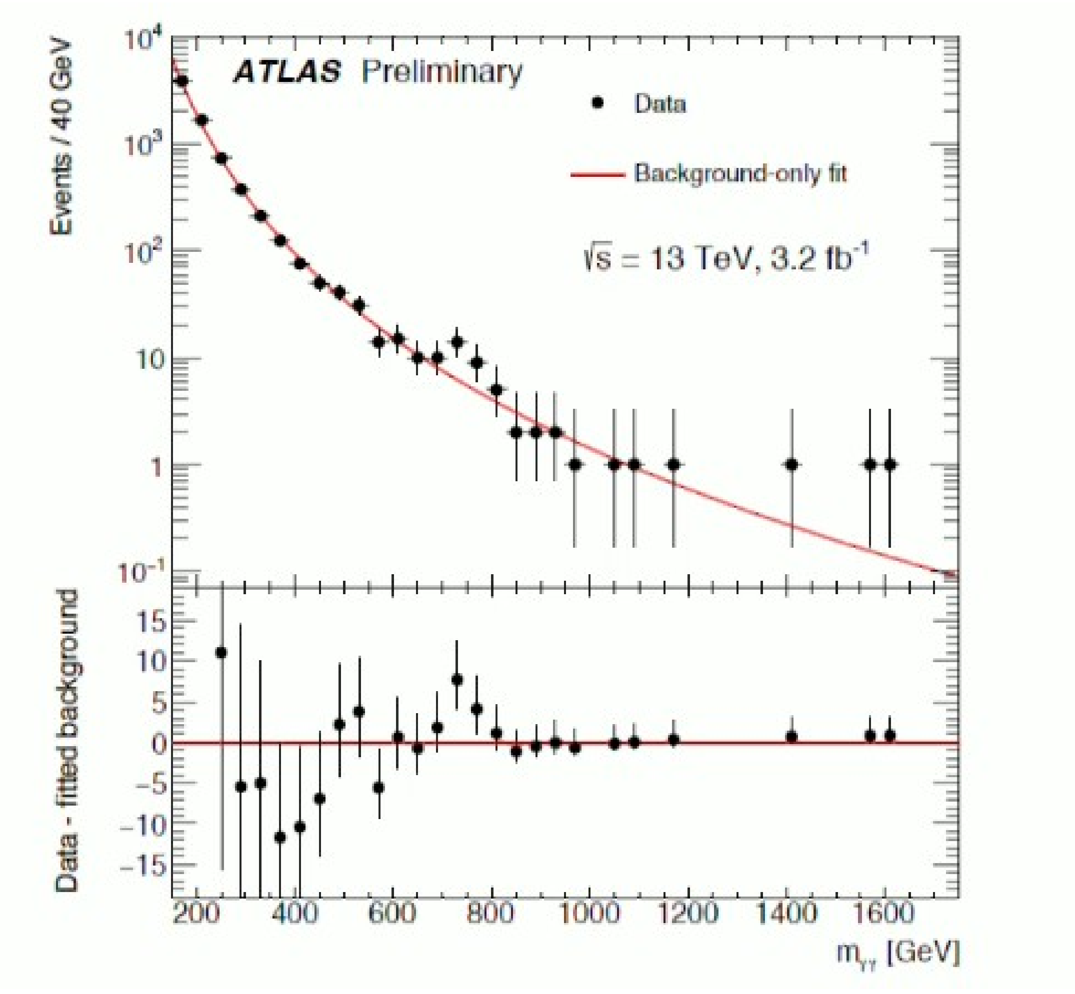
Ln(a) Sellentin
Imperial College London
&
Université de Genève

email: elena.sellentin@posteo.de

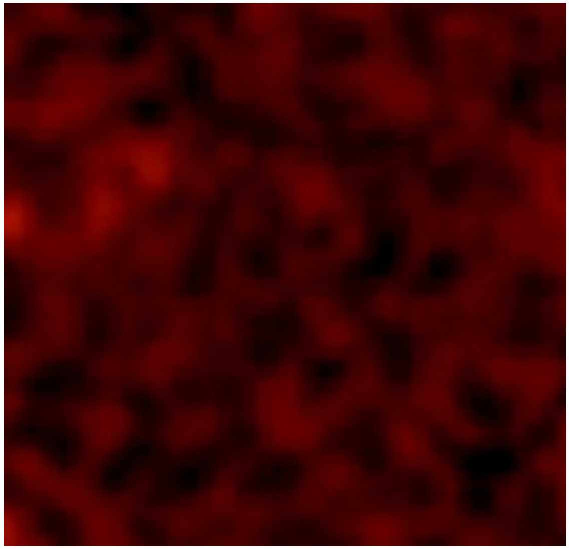
Typical questions



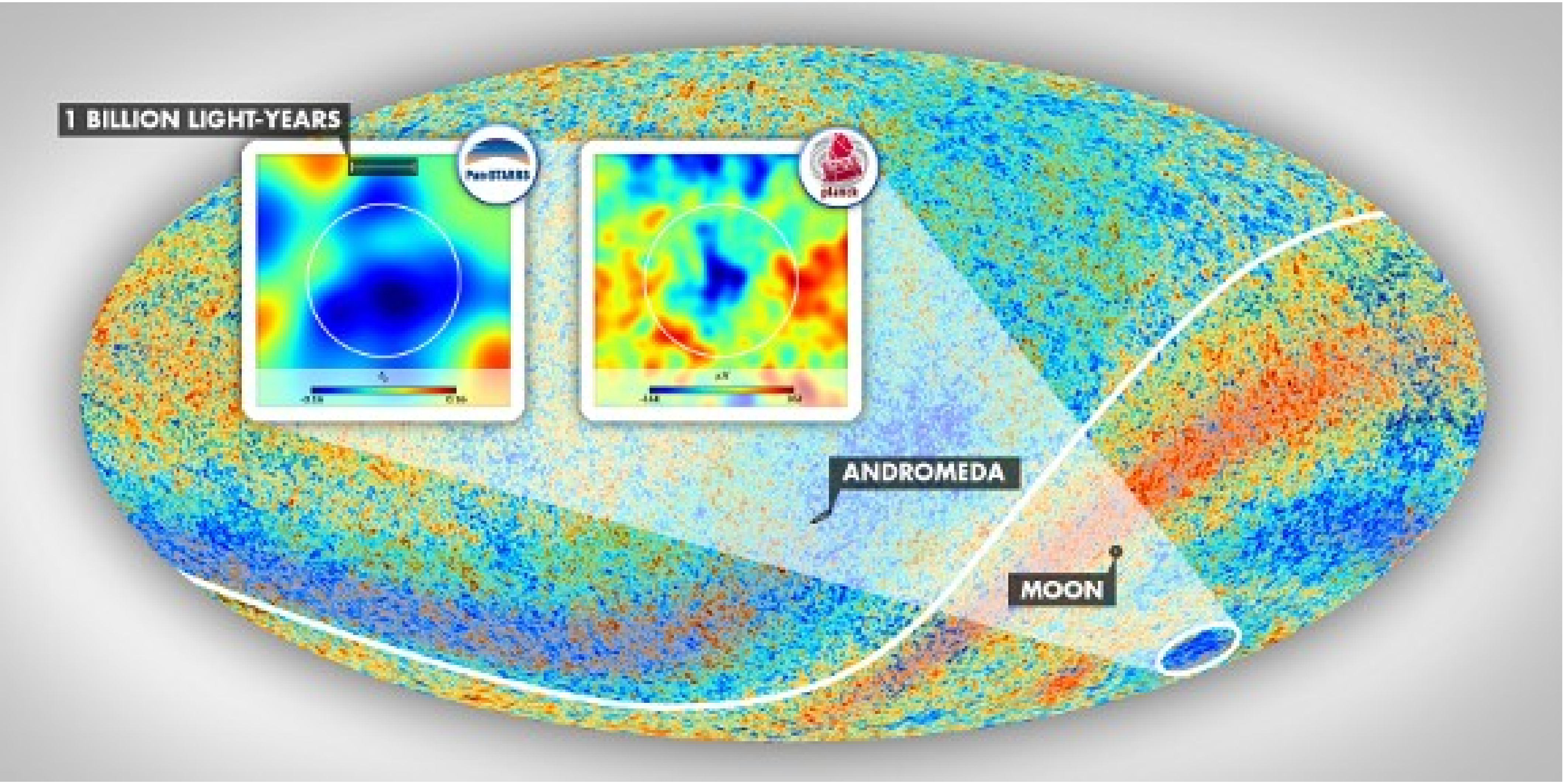
Typical questions



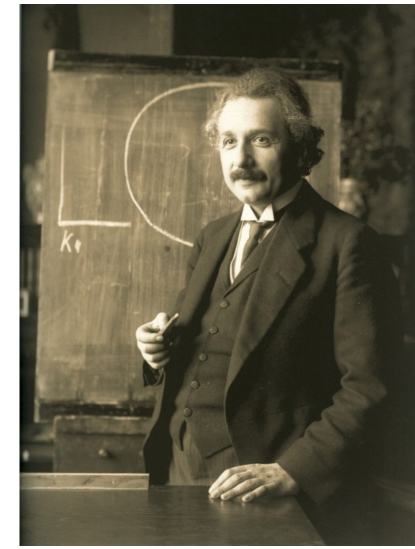
Typical questions



Typical questions



Typical questions



$$S_{EH} = \frac{1}{16\pi G} \int \sqrt{-g}\,(R - 2\Lambda)\,\mathrm{d}^4x$$

$$S = \int d^4x \sqrt{-g}\mathcal{L}_H$$

$$\mathcal{L}_H = \mathcal{L}_2 + \mathcal{L}_3 + \mathcal{L}_4 + \mathcal{L}_5$$

$$\begin{aligned}
\mathcal{L}_2 &= K(\phi, X), \\
\mathcal{L}_3 &= -G_3(\phi, X)\Box\phi, \\
\mathcal{L}_4 &= G_4(\phi, X)R + G_{4,X}[(\Box\phi)^2 - (\nabla_\mu\nabla_\nu\phi)(\nabla^\mu\nabla^\nu\phi)] \\
\mathcal{L}_5 &= G_5(\phi, X)G_{\mu\nu}(\nabla^\mu\nabla^\nu\phi) - \frac{1}{6}G_{5,X}[(\Box\phi)^3 - 3(\Box\phi)(\nabla_\mu\nabla_\nu\phi)(\nabla^\mu\nabla^\nu\phi) \\
&\quad +2(\nabla^\mu\nabla_\alpha\phi)(\nabla^\alpha\nabla_\beta\phi)(\nabla^\beta\nabla_\mu\phi)]
\end{aligned}$$

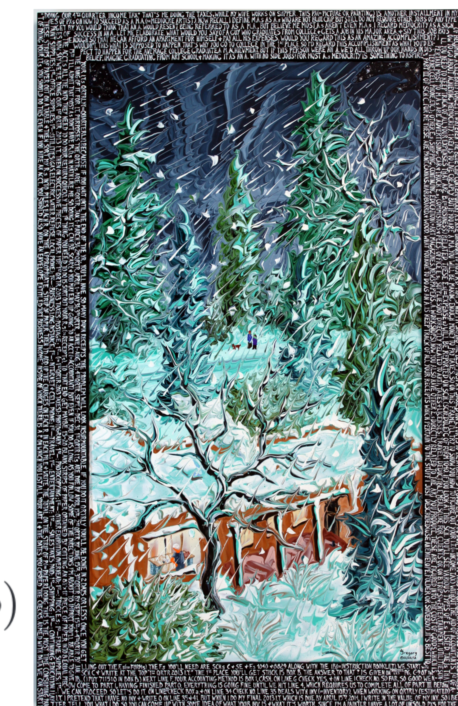


Image credit: Horndeski, Gregory W.

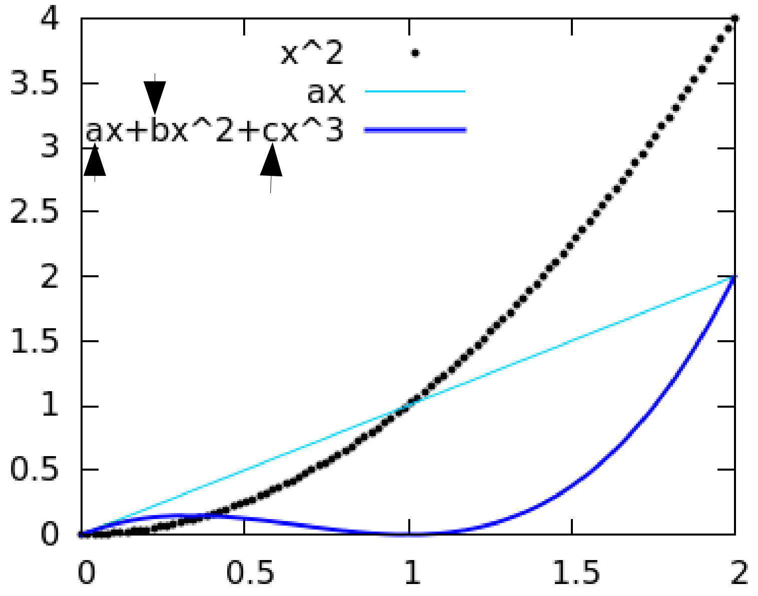
The evidence

- Normalization constant in parameter inference
- **The** quantity for model comparison

$$P(\boldsymbol{\theta}_M|\boldsymbol{X}) = \frac{\mathcal{P}(\boldsymbol{\theta}_M)L(\boldsymbol{X}|\boldsymbol{\theta}_M)}{\varepsilon}$$

$$\varepsilon = L(\boldsymbol{X}|M_1)$$

$$\varepsilon = \int L(\boldsymbol{X}|\boldsymbol{\theta}_M)\mathcal{P}(\boldsymbol{\theta}_M)\mathrm{d}^n\theta$$

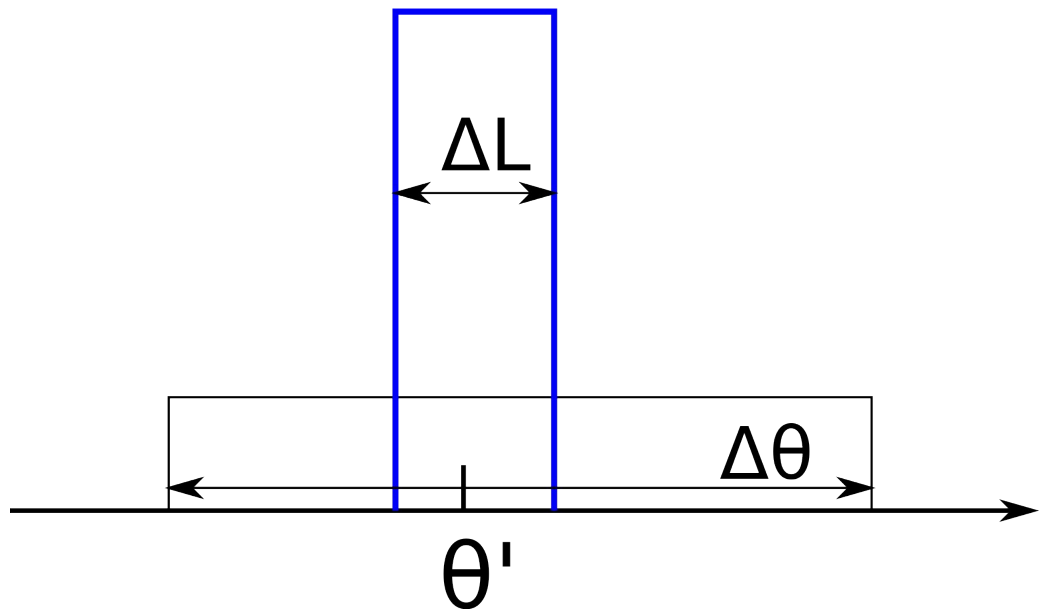


→ It balances the goodness of fit against the number of parameters.
'Occam's razor'.
→ It avoids (extreme) overfitting.

Toy Model

$$\varepsilon = \int L(\boldsymbol{X}|\boldsymbol{\theta}_M)\mathcal{P}(\boldsymbol{\theta}_M)\mathrm{d}^n\theta \quad = L(\theta')\mathcal{P}(\theta')\Delta L$$

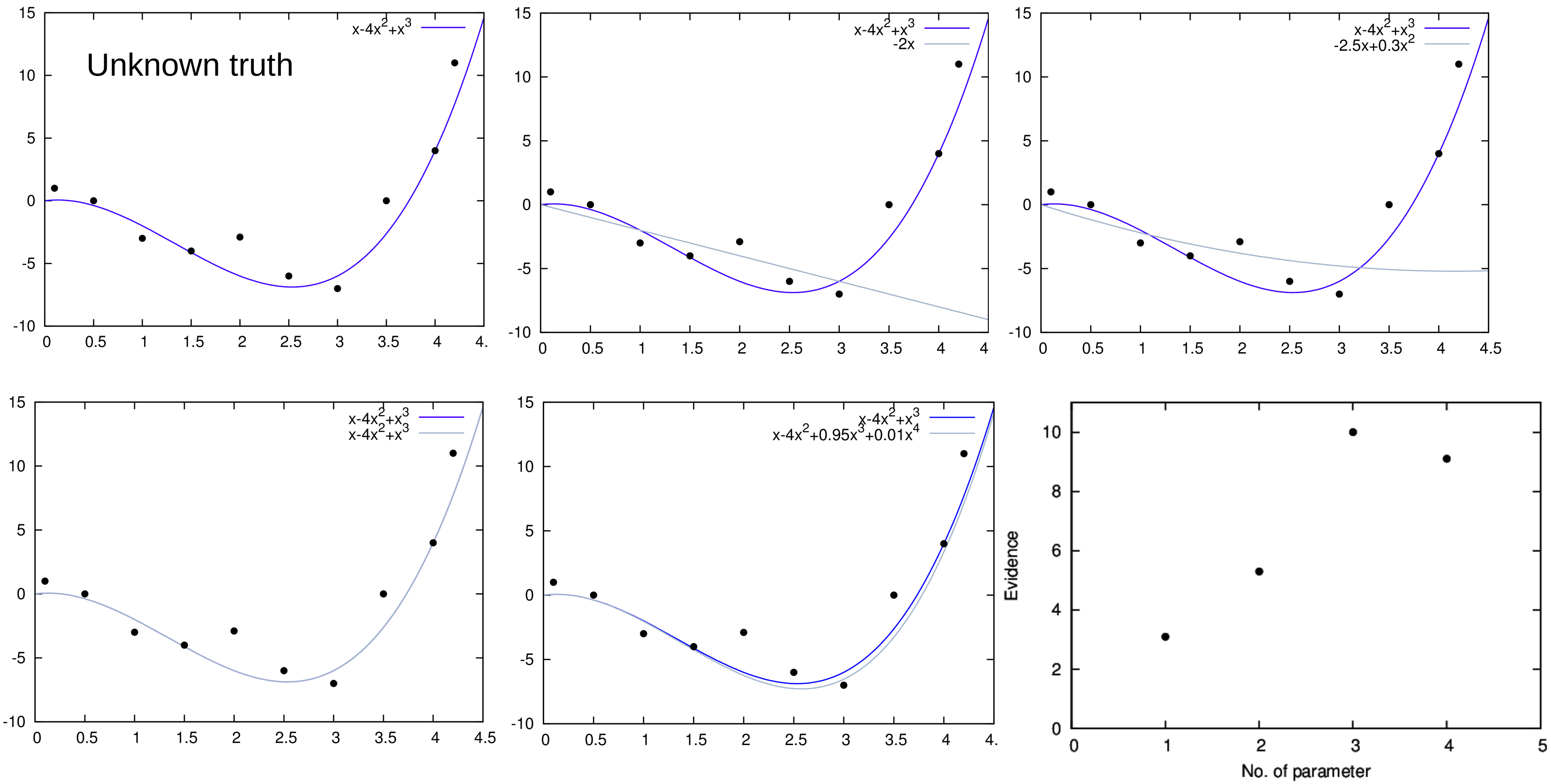
$$= L(\theta')\frac{\Delta L}{\Delta \mathcal{P}}$$



$$L \propto \exp(-\tfrac{1}{2}\chi^2)$$

Will always decrease with number of parameters.

Polynomial example

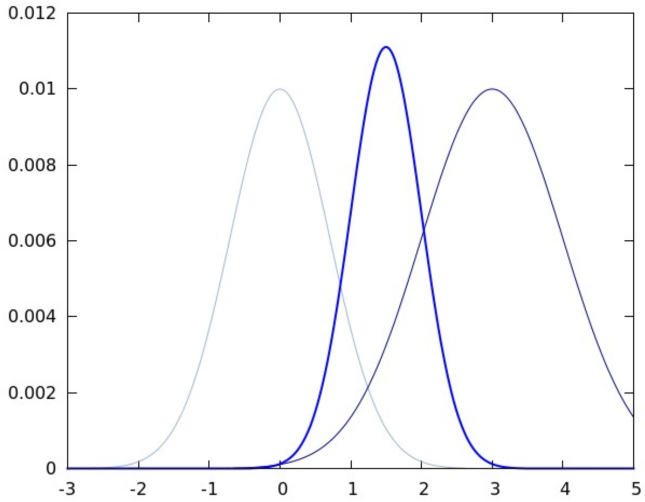


$$L(\theta')\,\frac{\Delta L}{\Delta \mathcal{P}}$$

$$L \propto \exp(-\tfrac{1}{2}\chi^2)$$

A word on priors in $\varepsilon = \int L(\boldsymbol{X}|\boldsymbol{\theta}_M)\mathcal{P}(\boldsymbol{\theta}_M)\mathrm{d}^n\theta$

- Theory or physics driven priors
  - $\Omega_m \in [0,1]$, Mass > 0
- Data driven priors & combination of experiments
  - Prior = old data
  - Likelihood = new data
  - Posterior = old and new data



- Subjective & informative priors
  - 'Only an unstated prior is a bad prior.'
- Objective & 'uninformative' priors
  - Maximize KL-divergence $D_{\mathrm{KL}}(P\|Q) = \int_{-\infty}^{\infty} p(x)\,\log\frac{p(x)}{q(x)}\,\mathrm{d}x$
  - Exploit symmetry groups: Haar-measures and invariant 'volumes'
  - Reparameterization independence (Jeffreys priors) $\pi_{IJ}(\boldsymbol{\mu},\boldsymbol{\Sigma}) = |\boldsymbol{\Sigma}|^{-(p+1)/2}$
- Frequentist matching priors $\frac{d}{d\theta}(\pi(\theta)I^{-1/2}(\theta)) = 0$

Model comparison

Have: $\varepsilon = L(\boldsymbol{X}|M_1)$

Want: $L(M_1|\boldsymbol{X})$

Bayes' theorem:

$$L(M_1|\boldsymbol{X}) = L(\boldsymbol{X}|M_1)\frac{\mathcal{P}(M_1)}{\mathcal{P}(\boldsymbol{X})}$$

?

Model comparison

Get rid off the prior probability for the data by taking a ratio:

$$\frac{L(M_1|\boldsymbol{X})}{L(M_2|\boldsymbol{X})} = \frac{\mathcal{P}(M_1)L(\boldsymbol{X}|M_1)}{\mathcal{P}(M_2)L(\boldsymbol{X}|M_2)}$$

$$= \frac{\mathcal{P}(M_1)}{\mathcal{P}(M_2)}\boxed{\frac{\varepsilon_1}{\varepsilon_2}}$$

→ Bayes factor: > 1 prefers M_1
< 1 prefers M_2

Where:

$$\varepsilon = \int L(\boldsymbol{X}|\boldsymbol{\theta}_M)\mathcal{P}(\boldsymbol{\theta}_M)\mathrm{d}^n\theta$$

Magnitude of B

- Bayes factor = evidence$_1$/evidence$_2$.
- Without loss of generality: $\epsilon_1 = b\epsilon_2$
- Then:

$$B_{12} = \frac{1}{b} \qquad and \qquad B_{21} = \frac{b}{1}$$

decisiveness asymptotes to zero vs. decisiveness grows linearly

- Ergo: Introduce ln for measure of decisiveness:

$$ln(B_{12}) = ln(1) - ln(b)$$

$$ln(B_{21}) = ln(b)$$

→ now B_{12} and B_{21} are treated equally

Calibration on the Jeffreys scale

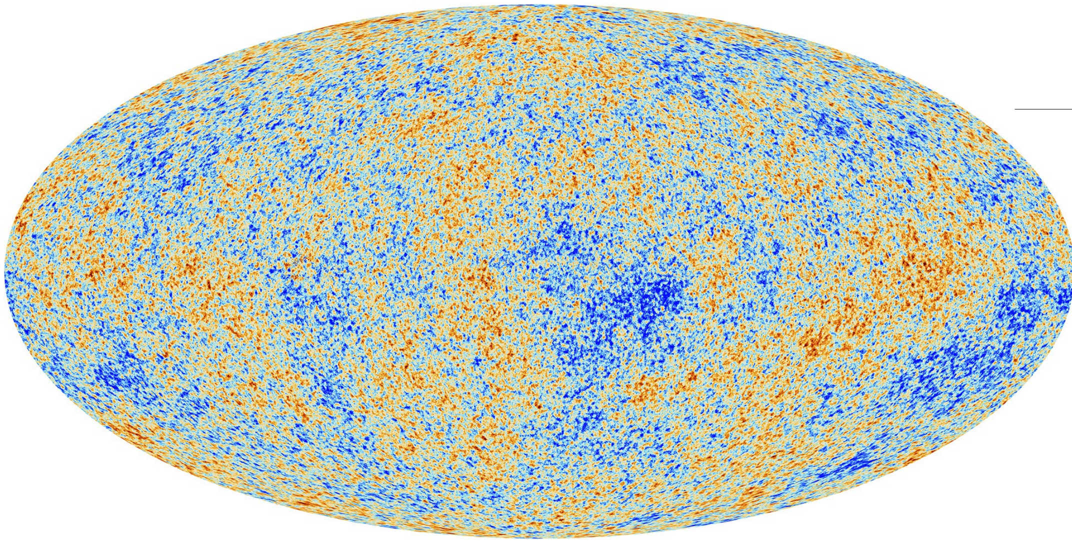
Table 6.1: Jeffreys scale

$\lvert \log\left(\frac{\varepsilon(M_1)}{\varepsilon(M_2)}\right) \rvert$	odds	interpretation	prob. of favoured model
≤ 1.0	3:1	better data is needed	≤ 0.75
≤ 2.5	12:1	weak evidence	0.923
≤ 5.0	$\leq$ 150:1	moderate evidence	0.993
≥ 5.0	$>$ 150:1	strong evidence	> 0.993

Example:

- Dark Energy Survey (DES) SV data
- WL analysis: flat LCDM vs. LCDM + curvature
- $\pi(\Omega_k) = uniform[-0.2, 0.2]$
- $ln(B) = 0.17 \pm 0.09$ Sellentin & Heavens (2016)

Model selection in the CMB



CMB = photons, in gravitational potentials of all particle species

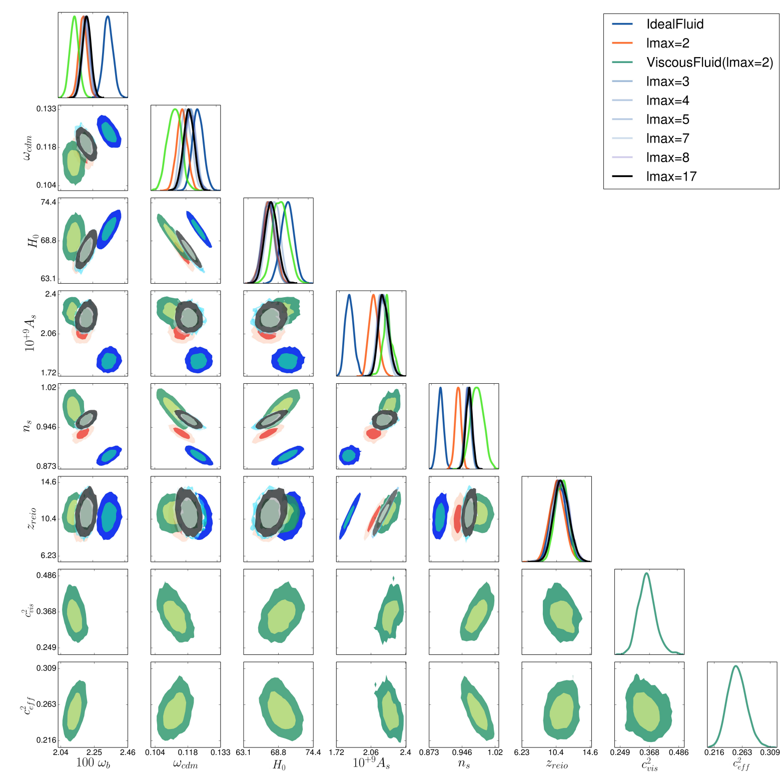
$$\gamma + p + n + e + DM + 3\xi_{rel}$$

$$Planck : \nu!$$

But are these neutrinos? Or just any relativistic fluid?

Model comparison:

Neutrinos vs. ideal fluid: $ln(B) \approx 10$
Neutrinos vs. viscous fluid: $ln(B) \approx 10.5$
+ parameter constraints as a side effect



Sellentin & Durrer (2015)

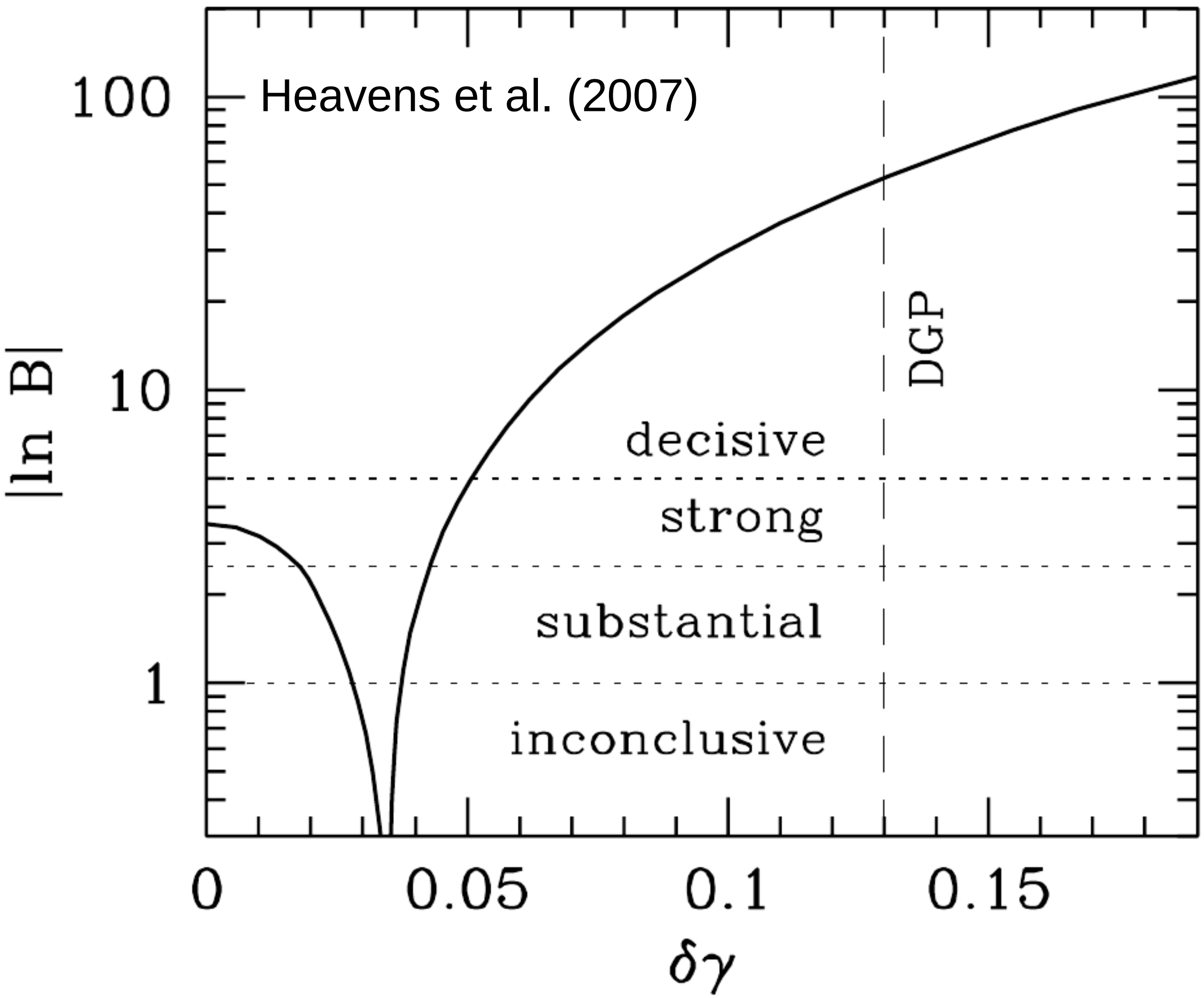
Expected support for models

- Single data realization: $B_{01} = \frac{\epsilon_0}{\epsilon_1} = \frac{\int L_0(\vec{x}|\vec{\theta}_{M_0})\mathcal{P}_0(\vec{\theta}_{M_0})d^{n_0}\theta}{\int L_1(\vec{x}|\vec{\theta}_{M_1})\mathcal{P}_1(\vec{\theta}_{M_1})d^{n_1}\theta}$

- Know statistical properties of data → calculate expected likelihood (even without having real data at all)

$$B_{01}^{expc} = \frac{\int \langle L_o(\vec{x}|\vec{\theta}_{M_0})\rangle \mathcal{P}_0(\vec{\theta}_{M_0})d^{n_0}\theta}{\int \langle L_1(\vec{x}|\vec{\theta}_{M_1})\rangle \mathcal{P}_1(\vec{\theta}_{M_1})d^{n_1}\theta}$$

Expected support for models



M_0: $g(a) = \exp\left\{\int_0^a \frac{da'}{a'}\left[\Omega_m(a')^\gamma - 1\right]\right\}$ M_1: $\Omega_m^{\gamma+\delta\gamma}$

Nested Models

- Imagine M_1 uses all parameters $\vec{\theta}$ of M_0 but introduces some extra parameters $\vec{\psi}$
- Nested model: for $\vec{\psi} = \vec{\psi}_0$ have $M_1 \rightarrow M_0$
- Examples:
  - wCDM → LambdaCDM for w = -1
  - Curved LambdaCDM → flat LambdaCDM for k = 0
  - Rainy day → sunny day for rain = 0

Savage-Dickey Density Ratio

- SDDR is an approximate Bayes factor for nested models
- The full Bayes factor is $B_{01} = \frac{\epsilon_0}{\epsilon_1} = \frac{\int L_0(\vec{x}|\vec{\theta}_{M_0})\mathcal{P}_0(\vec{\theta}_{M_0})d^{n_0}\theta}{\int L_1(\vec{x}|\vec{\theta}_{M_1})\mathcal{P}_1(\vec{\theta}_{M_1})d^{n_1}\theta}$
- For nested models: $L_0(\vec{x}|\vec{\theta}_{M_0}) = L_1(\vec{x}|\vec{\theta}_{M_0}, \vec{\psi} = \vec{\psi}_0)$
- Insert into Bayes factor:

$$B_{01} = \frac{\int L_1(\vec{x}|\vec{\theta}_{M_0}, \vec{\psi} = \vec{\psi}_0)\mathcal{P}_0(\vec{\theta}_{M_0})d^{n_0}\theta}{\int L_1(\vec{x}|\vec{\theta}_{M_0}, \vec{\psi})\mathcal{P}_1(\vec{\theta}_{M_0}, \vec{\psi})d^{n_0}\theta d^n\psi}$$

- Now need to care about the priors.

Savage-Dickey Density Ratio

- Bayes factor:

$$B_{01} = \frac{\int L_1(\vec{x}|\vec{\theta}_{M_0}, \vec{\psi} = \vec{\psi}_0)\mathcal{P}_0(\vec{\theta}_{M_0})d^{n_0}\theta}{\int L_1(\vec{x}|\vec{\theta}_{M_0}, \vec{\psi})\mathcal{P}_1(\vec{\theta}_{M_0}, \vec{\psi})d^{n_0}\theta d^n\psi}$$

- Make extra assumption for priors: $\mathcal{P}_1(\vec{\theta}_{M_0}|\vec{\psi} = \vec{\psi}_0) = a\mathcal{P}_0(\vec{\theta}_{M_0})$

- Insert into Bayes factor:

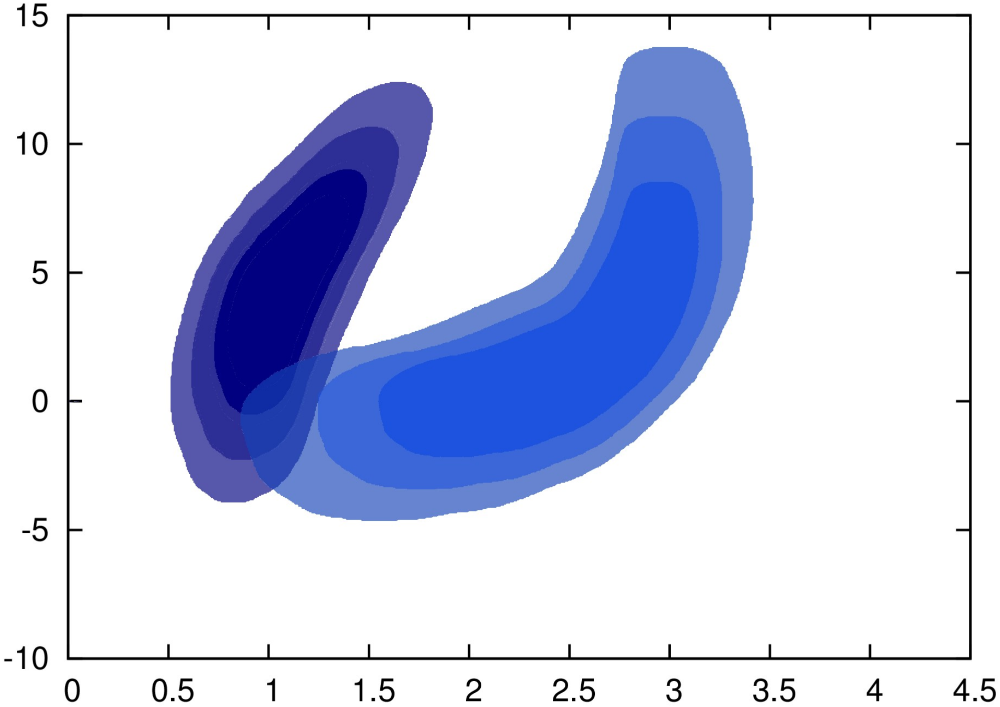
$$B_{01} \approx a\frac{\int L_1(\vec{x}|\vec{\theta}_{M_0}, \vec{\psi} = \vec{\psi}_0)\mathcal{P}_1(\vec{\theta}_{M_0}|\psi = \psi_0)d^{n_0}\theta}{\int L_1(\vec{x}|\vec{\theta}_{M_0}, \vec{\psi})\mathcal{P}_1(\vec{\theta}_{M_0}, \vec{\psi})d^{n_0}\theta d^n\psi}$$

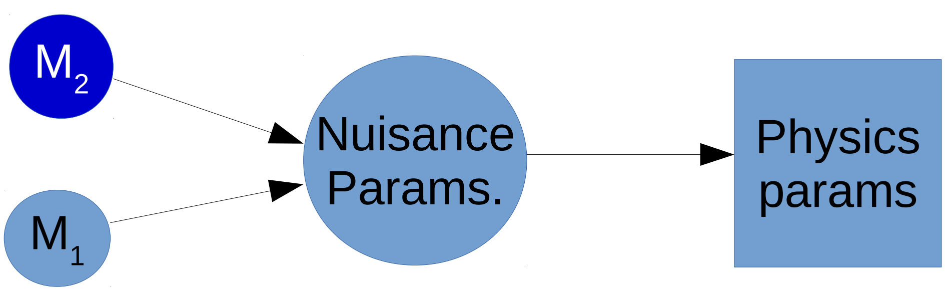
- Leading to the Savage-Dickey Density Ratio: $B_{01} \approx a\frac{P_1(\vec{\psi} = \vec{\psi}_0|\vec{x})}{P_1(\vec{\psi} = \vec{\psi}_0)}$

Example from Dirian et al.(2016): $B_{\Lambda(\Lambda+i)} \equiv \frac{P(d|\mathcal{M}_\Lambda)}{P(d|\mathcal{M}_{\Lambda+i})} = \frac{P(\Omega_{X_i} = 0|d, \mathcal{M}_{\Lambda+i})}{P(\Omega_{X_i} = 0|\mathcal{M}_{\Lambda+i})}$

→ Plan ahead, use Nested Sampling not MCMC to get B + param. constraints
→ If too late: MCMC+SDDR+importance sampling approximate B (excercise)

Model averaging





- Imagine two models explain the same effect. None is 'better' than the other, as given by B.
- Weak lensing: Intrinsic alignment model?
- Structure formation: Press-Schechter mass function or Sheth-Torman or Jenkins et al. or...?

$$P(\vec{\theta}|\vec{x}) \propto \sum_i P(\vec{\theta}|\vec{x}, M_i)P(M_i|\vec{x})$$

- Includes model uncertainty into parameter uncertainty.

Summary

- Bayesians compare models by evidence ratios
- Balance goodness of fit against number of parameters
- Samplers exist that give parameter constraints and evidences (→ JP's lecture)
- Savage-Dickey Density Ratio may or may not be of relevance to you in case of nested models...
- … depending on your attitude towards priors (subjective/objective).
- Model comparison is prior dependent.