Bayesian Inference in Cosmology

Alan Heavens

Imperial Centre for Inference and Cosmology (ICIC)
Imperial College, London

a.heavens@imperial.ac.uk

CosmoBack, Marseille

May 31, 2018

Overview

1. Parameter Inference
   - The posterior $p(parameters|data)$
   - Gaussian likelihoods
   - Data compression
   - Non-gaussian likelihoods
   - Approximate Bayesian Computation (ABC)
2. Bayesian Hierarchical Models (BHM)
3. Model Comparison
   - Do the data favour ΛCDM?
   - The future: what are the prospects to falsify ΛCDM?

Bayesian Inference

- What questions do we want to answer?

Bayesian Inference

- What questions do we want to answer?
- Given all the available data, what is the probability that cosmological parameters take certain values? [Parameter Inference]

Bayesian Inference

- What questions do we want to answer?
- Given all the available data, what is the probability that cosmological parameters take certain values? [Parameter Inference]
- What is the relative probability of ΛCDM compared with alternatives? [Model Comparison]

Notation

- Data d; Model parameters θ; Model M

Notation

- Data d; Model parameters θ; Model M
- Likelihood $\mathcal{L}(\mathrm{d}|\theta) = p(\mathrm{d}|\theta, M)$

Notation

- Data d; Model parameters θ; Model M
- Likelihood $\mathcal{L}(\mathrm{d}|\theta) = p(\mathrm{d}|\theta, M)$
- Posterior $p(\theta|\mathrm{d}, M)$

Notation

- Data d; Model parameters θ; Model M
- Likelihood $\mathcal{L}(\mathrm{d}|\theta) = p(\mathrm{d}|\theta, M)$
- Posterior $p(\theta|\mathrm{d}, M)$
- Prior $\pi(\theta) = p(\theta|M)$

Notation

- Data d; Model parameters θ; Model M
- Likelihood $\mathcal{L}(\mathrm{d}|\theta) = p(\mathrm{d}|\theta, M)$
- Posterior $p(\theta|\mathrm{d}, M)$
- Prior $\pi(\theta) = p(\theta|M)$
- Bayes theorem:

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

Notation

- Data d; Model parameters θ; Model M
- Likelihood $\mathcal{L}(\mathrm{d}|\theta) = p(\mathrm{d}|\theta, M)$
- Posterior $p(\theta|\mathrm{d}, M)$
- Prior $\pi(\theta) = p(\theta|M)$
- Bayes theorem:

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

- $p(\mathrm{d}|M)$ is the Bayesian Evidence, which is important for Model Comparison, but not for Parameter Inference.

Notation

- Data d; Model parameters θ; Model M
- Likelihood $\mathcal{L}(\mathrm{d}|\theta) = p(\mathrm{d}|\theta, M)$
- Posterior $p(\theta|\mathrm{d}, M)$
- Prior $\pi(\theta) = p(\theta|M)$
- Bayes theorem:

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

- $p(\mathrm{d}|M)$ is the Bayesian Evidence, which is important for Model Comparison, but not for Parameter Inference.
- Dropping M dependence for now (we will return to it when we discuss Model Comparison):

$$p(\theta|d) = \frac{\mathcal{L}(d|\theta)\pi(\theta)}{p(d)}$$

The Posterior

$$p(\theta|d, M)$$

If you just try long enough and hard enough, you can always manage to boot yourself in the posterior. A.J. Liebling.

It is all probability

The Posterior

Everything is focussed on getting at $p(\theta|\mathrm{d})$.

Computing the posterior

$p(\theta|\mathrm{d}) \propto \mathcal{L}(\theta)\,\pi(\theta)$.

We need to make some choices:

What are the data, d?
What is the likelihood function $\mathcal{L}(\mathrm{d}|\theta)$?
What is the prior $\pi(\theta)$?

Priors

Bayesian: prior = (usually) the state of knowledge before the new data are collected.
For parameter inference, the prior becomes unimportant as more data are added and the likelihood dominates.
For model comparison (see later), the prior remains important.
Issues:

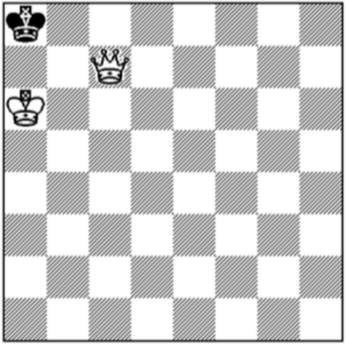
- Sometimes one wants an 'uninformative' prior, but what does this mean?

Priors

Bayesian: prior = (usually) the state of knowledge before the new data are collected.
For parameter inference, the prior becomes unimportant as more data are added and the likelihood dominates.
For model comparison (see later), the prior remains important.
Issues:

- Sometimes one wants an 'uninformative' prior, but what does this mean?
- Subjective vs Objective Bayesians

Priors



Credit: Daniel Mortlock

Subjective Bayesians, uninformative priors, and reparametrisation

- **Subjective Bayesian:** specify the prior first, independently of the experiment you are about to do.

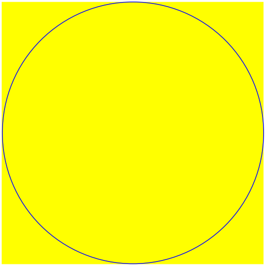
Subjective Bayesians, uninformative priors, and reparametrisation

- **Subjective Bayesian:** specify the prior first, independently of the experiment you are about to do.
- **'Flat' or 'uniform' priors:** A common and apparently reasonable 'uniformative' prior is $\pi(\theta) =$ constant.

Uninformative prior

A flat prior seems natural, but consider this problem. Imagine cartesian coordinates in N dimensions, with the prior range being $(-\frac{1}{2}, \frac{1}{2})$ for all coordinates. The prior probability of being inside the N-sphere which just fits inside the prior volume is

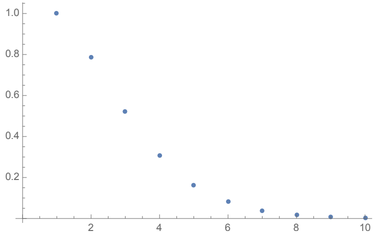
$$\frac{\pi^{N/2}}{2^N \Gamma(1 + N/2)}$$



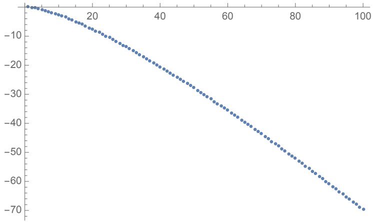
Uninformative?

Probability of being inside the N-sphere vs N:

p



$\log_{10} p$



An apparently uninformative prior may be *highly informative* when viewed a different way.

Jeffreys Prior

- Objective Bayesians would like to choose priors which influence the final outcome as little as possible, i.e. are least amount informative

Jeffreys Prior

- Objective Bayesians would like to choose priors which influence the final outcome as little as possible, i.e. are least amount informative
- For one dimension, we can find a prior that is minimally information. This is the *Reference Prior*, and in 1D is also the *Jeffreys Prior*:

Jeffreys Prior

- Objective Bayesians would like to choose priors which influence the final outcome as little as possible, i.e. are least amount informative
- For one dimension, we can find a prior that is minimally information. This is the *Reference Prior*, and in 1D is also the *Jeffreys Prior*:
- $$\pi_J(\theta) \propto \sqrt{I(\theta)}; \qquad I(\theta) = -\mathbb{E}\left[\frac{d^2 \ln \mathcal{L}(d|\theta)}{d\theta^2}\right]$$

  is the *Fisher Information*.

Jeffreys Prior

- Objective Bayesians would like to choose priors which influence the final outcome as little as possible, i.e. are least amount informative
- For one dimension, we can find a prior that is minimally information. This is the *Reference Prior*, and in 1D is also the *Jeffreys Prior*:
- $$\pi_J(\theta) \propto \sqrt{I(\theta)}; \qquad I(\theta) = -\mathbb{E}\left[\frac{d^2 \ln \mathcal{L}(d|\theta)}{d\theta^2}\right]$$

  is the *Fisher Information*.
- Note that this is against the spirit of Subjective Bayes - the prior depends on the likelihood. The expectation ($\mathbb{E}$) is taken over the data at given θ.

Jeffreys Prior

- Objective Bayesians would like to choose priors which influence the final outcome as little as possible, i.e. are least amount informative
- For one dimension, we can find a prior that is minimally information. This is the *Reference Prior*, and in 1D is also the *Jeffreys Prior*:
- $$\pi_J(\theta) \propto \sqrt{I(\theta)}; \qquad I(\theta) = -\mathbb{E}\left[\frac{d^2 \ln \mathcal{L}(d|\theta)}{d\theta^2}\right]$$

  is the *Fisher Information*.
- Note that this is against the spirit of Subjective Bayes - the prior depends on the likelihood. The expectation ($\mathbb{E}$) is taken over the data at given θ.
- Jeffreys Priors sometimes do not generalise well to multidimensional problems. But sometimes they do - e.g. neutrino masses in oscillation and cosmological experiments, $\pi(m_1, m_2, m_3) \propto m_1 m_2 + m_1 m_3 + m_2 m_3$ (Heavens & Sellentin 2018).

Data

Summary Statistics

Data compression

We do not usually compute the probability of all the measured data, since the number of these may be large (e.g. Planck has $\sim 10^{12}$ time-ordered data). We compress them, e.g. to a map, or a power spectrum.

Summary Statistics

Typical summary statistics: correlation function or power spectrum estimates. Already a massive data compression. Perhaps $10^2 - 10^4$ summary statistics for Euclid or LSST.

Likelihood

Gaussian Likelihood

Gaussian Likelihood

Often, we assume that the summary statistics are gaussian-distributed (Handwave, handwave, central limit theorem ...)

Is a Gaussian likelihood appropriate?

We rarely stop to question this, but we should. Let us run with it for now

Gaussian Likelihood

$$\mathcal{L}(\mathbf{d}|\theta) = |2\pi\Sigma|^{-1/2} \exp\left[-\frac{1}{2}(\mathbf{d}-\mu)^{\mathbf{T}}\mathbf{\Sigma}^{-\mathbf{1}}(\mathbf{d}-\mu)\right]$$

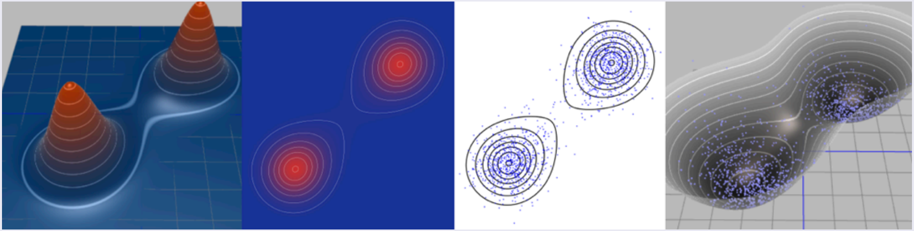
$\mu(\theta)$ is the mean signal, obtained from theory or simulation. Σ is the covariance matrix. It may depend on θ. It is a problem (except for Gaussian fields such as the CMB).

Sampling

MCMC

If we know Σ, we need to evaluate the posterior as a function of parameters θ. Not trivial if there are ~ 10 parameters. Standard technique is MCMC (Markov Chain Monte Carlo), where steps are taken in parameter space, according to a proposal distribution, and accepted or rejected according to the Metropolis-Hastings algorithm. This gives a chain of samples of the posterior (or the likelihood).

MCMC example



Covariance Matrices

1. If summary statistics are 2-point functions, Σ is a 4-point function. Hard to compute for non-gaussian fields.

Covariance Matrices

1. If summary statistics are 2-point functions, Σ is a 4-point function. Hard to compute for non-gaussian fields.
2. Either use analytic covariance matrix, or simulate (or both)

Covariance Matrices

1. If summary statistics are 2-point functions, Σ is a 4-point function. Hard to compute for non-gaussian fields.
2. Either use analytic covariance matrix, or simulate (or both)
3. For simulated covariance matrices, an estimate $\hat{\Sigma}$ can be unbiased. Note that some effects are not included - e.g. super-sample covariance.

Covariance Matrices

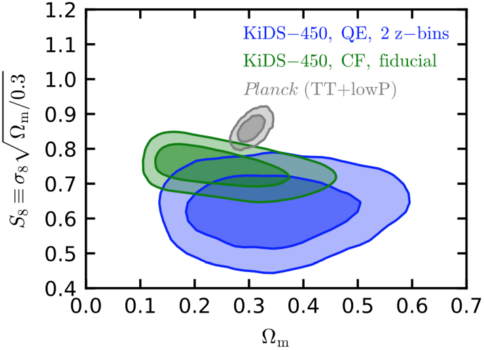
1. If summary statistics are 2-point functions, Σ is a 4-point function. Hard to compute for non-gaussian fields.
2. Either use analytic covariance matrix, or simulate (or both)
3. For simulated covariance matrices, an estimate $\hat{\Sigma}$ can be unbiased. Note that some effects are not included - e.g. super-sample covariance.
4. However, $\hat{\Sigma}^{-1}$ is not unbiased. A fix is the Hartlap et al (2007) correction: multiply by $(N-1)/(N-n-2)$, where $n =$ number of data; $N =$ no. of sims.

Covariance Matrices

1. If summary statistics are 2-point functions, Σ is a 4-point function. Hard to compute for non-gaussian fields.
2. Either use analytic covariance matrix, or simulate (or both)
3. For simulated covariance matrices, an estimate $\hat{\Sigma}$ can be unbiased. Note that some effects are not included - e.g. super-sample covariance.
4. However, $\hat{\Sigma}^{-1}$ is not unbiased. A fix is the Hartlap et al (2007) correction: multiply by $(N-1)/(N-n-2)$, where n = number of data; N = no. of sims.
5. Better: marginalise over true $\Sigma \rightarrow$ likelihood of Sellentin & Heavens (2016)

Covariance Matrices matter

e.g. KiDS weak lensing result (on $S_8 = \sigma_8\sqrt{\Omega_m/0.3}$) shifts by 1σ when changing from an analytic to a simulated covariance matrix (Hildebrandt et al 2017).



Covariance Matrices

1. Need $N > n + 2$, where $n =$ number of summary statistics

Covariance Matrices

1. Need $N > n + 2$, where $n =$ number of summary statistics
2. n could easily be 10^4 for LSST or Euclid

Covariance Matrices

1. Need $N > n + 2$, where n = number of summary statistics
2. n could easily be 10^4 for LSST or Euclid
3. If Σ varies with cosmological parameters (as it will), then it is worse. Estimating Σ would be prohibitively expensive

Covariance Matrices

1. Need $N > n + 2$, where n = number of summary statistics
2. n could easily be 10^4 for LSST or Euclid
3. If Σ varies with cosmological parameters (as it will), then it is worse. Estimating Σ would be prohibitively expensive
4. Solution: reduce n. More radical data compression

Data Compression

MOPED algorithm

MOPED

Massively Optimised Parameter Estimation and Data compression (Heavens et al. 2000). See also Zablocki & Dodelson (2016), Alsing & Wandelt (2018), Charnock et al (2018).

Linear compression (note $C = \Sigma$):

$y_\alpha = b_\alpha \cdot d$

$$\mathbf{b}_1 = \frac{\mathsf{C}^{-1}\boldsymbol{\mu}_{,1}}{\sqrt{\boldsymbol{\mu}_{,1}^T \mathsf{C}^{-1}\boldsymbol{\mu}_{,1}}}$$

and

$$\mathbf{b}_\alpha = \frac{\mathsf{C}^{-1}\boldsymbol{\mu}_{,\alpha} - \sum_{\beta=1}^{\alpha-1}(\boldsymbol{\mu}_{,\alpha}^T \mathbf{b}_\beta)\mathbf{b}_\beta}{\sqrt{\boldsymbol{\mu}_{,\alpha}^T \mathsf{C}^{-1}\boldsymbol{\mu}_{,\alpha} - \sum_{\beta=1}^{\alpha-1}(\boldsymbol{\mu}_{,\alpha}^T \mathbf{b}_\beta)^2}} \qquad 1 < \alpha \leq m,$$

1. Size of dataset reduced massively to the number of parameters.

Data Compression

MOPED algorithm

MOPED

Massively Optimised Parameter Estimation and Data compression (Heavens et al. 2000). See also Zablocki & Dodelson (2016), Alsing & Wandelt (2018), Charnock et al (2018).

Linear compression (note $C = \Sigma$):

$y_\alpha = b_\alpha \cdot d$

$$\mathbf{b}_1 = \frac{\mathsf{C}^{-1}\boldsymbol{\mu}_{,1}}{\sqrt{\boldsymbol{\mu}_{,1}^T \mathsf{C}^{-1}\boldsymbol{\mu}_{,1}}}$$

and

$$\mathbf{b}_\alpha = \frac{\mathsf{C}^{-1}\boldsymbol{\mu}_{,\alpha} - \sum_{\beta=1}^{\alpha-1}(\boldsymbol{\mu}_{,\alpha}^T \mathbf{b}_\beta)\mathbf{b}_\beta}{\sqrt{\boldsymbol{\mu}_{,\alpha}^T \mathsf{C}^{-1}\boldsymbol{\mu}_{,\alpha} - \sum_{\beta=1}^{\alpha-1}(\boldsymbol{\mu}_{,\alpha}^T \mathbf{b}_\beta)^2}} \qquad 1 < \alpha \le m,$$

1. Size of dataset reduced massively to the number of parameters.
2. Same Fisher Matrix! $F_{\alpha\beta} \equiv -\langle \partial^2 \ln \mathcal{L} / \partial\theta_\alpha \partial\theta_\beta \rangle$

Data Compression

MOPED algorithm

MOPED

Massively Optimised Parameter Estimation and Data compression (Heavens et al. 2000). See also Zablocki & Dodelson (2016), Alsing & Wandelt (2018), Charnock et al (2018).

Linear compression (note $C = \Sigma$):

$y_\alpha = b_\alpha \cdot d$

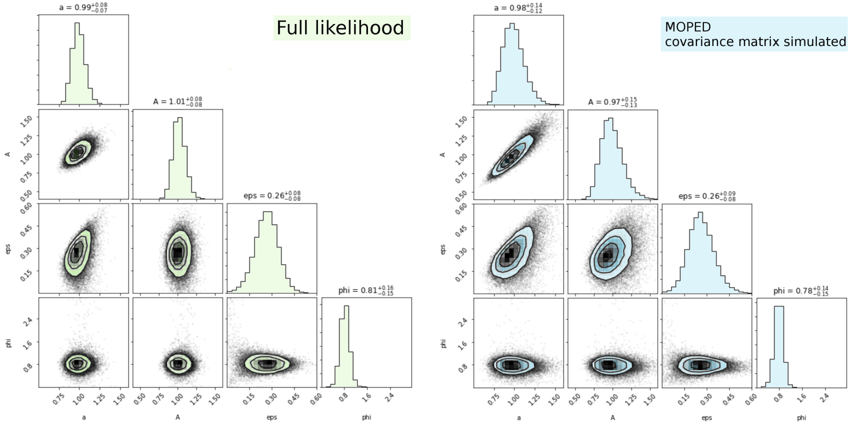
$$\mathbf{b}_1 = \frac{\mathsf{C}^{-1}\boldsymbol{\mu}_{,1}}{\sqrt{\boldsymbol{\mu}_{,1}^T \mathsf{C}^{-1}\boldsymbol{\mu}_{,1}}}$$

and

$$\mathbf{b}_\alpha = \frac{\mathsf{C}^{-1}\boldsymbol{\mu}_{,\alpha} - \sum_{\beta=1}^{\alpha-1}(\boldsymbol{\mu}_{,\alpha}^T \mathbf{b}_\beta)\mathbf{b}_\beta}{\sqrt{\boldsymbol{\mu}_{,\alpha}^T \mathsf{C}^{-1}\boldsymbol{\mu}_{,\alpha} - \sum_{\beta=1}^{\alpha-1}(\boldsymbol{\mu}_{,\alpha}^T \mathbf{b}_\beta)^2}} \qquad 1 < \alpha \le m,$$

1. Size of dataset reduced massively to the number of parameters.
2. Same Fisher Matrix! $F_{\alpha\beta} \equiv -\langle \partial^2 \ln \mathcal{L} / \partial\theta_\alpha \partial\theta_\beta \rangle$
3. MOPED (originally proposed for a different purpose) can solve the simulations problem: Heavens et al (2017) and Gualdi et al (2018).

MOPED performance



Is the likelihood Gaussian?

The data are not Gaussian-distributed, even when the CLT handwave suggests otherwise. . .

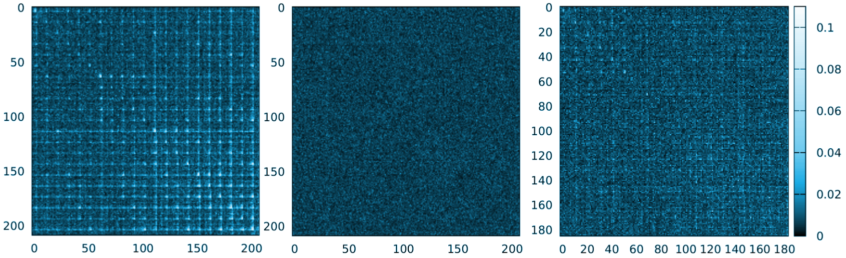


Figure: CFHTLenS sims Gaussian data Most Gaussian terms

Sellentin & Heavens (2018).

Non-Gaussian Likelihoods

A major challenge

3 approaches

Edgeworth expansion
Approximate likelihood
Bayesian Hierarchical Models

Edgeworth Expansion

Joint distribution of all Fourier coefficients $a_\mathbf{k}$

Schematically:

$$p(a_\mathbf{k}|\theta) = |\mathrm{diag}[2\pi P(k)]|^{-1/2} \exp\left[-\frac{|a_\mathbf{k}|^2}{2P(k)}\right] \{1 + B + T + B^2 + \ldots\}$$

$P(k,\theta), B(\mathbf{k},\theta), T(\mathbf{k},\theta) =$ power spectrum, bispectrum, trispectrum.

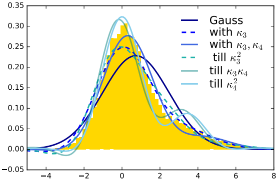


Figure: Sellentin, Jaffe, Heavens 2018

Other ideas: gaussianising transforms (e.g. Hall & Mead 2018), clipping (Simpson et al. 2011)

Approximate Likelihood and Posterior

Machine-learning techniques: Approximate Bayesian Computation

ABC

Posterior: Rejection-sampling ABC.
Run many simulations.
Keep those that match the data.
Match: not everything, but match some summary statistics.
Very expensive

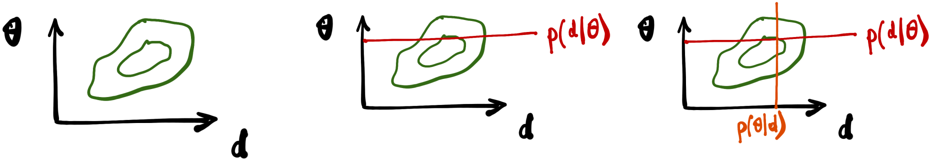
Fit the likelihood

Fit the sampling distribution $p(d|\theta)$ of mocks. e.g. Hahn et al (2018)
Feasible in relatively small numbers of dimensions
Probably impossible in very high dimensions
Data compression needed again

Fitting the joint p(data, parameters)

Fit $p(\mathrm{d}, \theta)$

Use machine learning techniques such as GMM, KDE. Alsing et al. (2018)



Bayesian Hierarchical Models

If you can, this is how to do it

BHM

- We split the inference problem into steps, where the full model is made up of a series of sub-models

Bayesian Hierarchical Models

If you can, this is how to do it

BHM

- We split the inference problem into steps, where the full model is made up of a series of sub-models
- The Bayesian Hierarchical Model (BHM) links the sub-models together, correctly propagating uncertainties in each sub-model from one level to the next.

Bayesian Hierarchical Models

If you can, this is how to do it

BHM

- We split the inference problem into steps, where the full model is made up of a series of sub-models
- The Bayesian Hierarchical Model (BHM) links the sub-models together, correctly propagating uncertainties in each sub-model from one level to the next.
- At each step ideally we will know the conditional distributions

Bayesian Hierarchical Models

If you can, this is how to do it

BHM

- We split the inference problem into steps, where the full model is made up of a series of sub-models
- The Bayesian Hierarchical Model (BHM) links the sub-models together, correctly propagating uncertainties in each sub-model from one level to the next.
- At each step ideally we will know the conditional distributions
- The aim is to build a complete model of the data

Bayesian Hierarchical Models

If you can, this is how to do it

BHM

- We split the inference problem into steps, where the full model is made up of a series of sub-models
- The Bayesian Hierarchical Model (BHM) links the sub-models together, correctly propagating uncertainties in each sub-model from one level to the next.
- At each step ideally we will know the conditional distributions
- The aim is to build a complete model of the data
- Principled way to include systematic errors, selection effects (everything, really)

Bayesian Hierarchical Models

A simple example

Often used to learn about a *population* from many *individual measurements*. e.g. we measure the fluxes $\hat{f}_i$ of a population of galaxies, but the? have errors. What are the true number counts?

- Assume (say) a power-law $N \propto f^{-\alpha}$

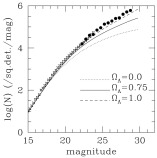


Figure: Ned Wright

Bayesian Hierarchical Models

A simple example

Often used to learn about a *population* from many *individual measurements*. e.g. we measure the fluxes $\hat{f}_i$ of a population of galaxies, but the? have errors. What are the true number counts?

- Assume (say) a power-law $N \propto f^{-\alpha}$
- Many (unobserved) true fluxes θ_i

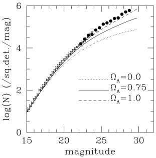


Figure: Ned Wright

Bayesian Hierarchical Models

A simple example

Often used to learn about a *population* from many *individual measurements*. e.g. we measure the fluxes $\hat{f}_i$ of a population of galaxies, but the? have errors. What are the true number counts?

- Assume (say) a power-law $N \propto f^{-\alpha}$
- Many (unobserved) true fluxes θ_i
- Add noise: $\hat{f}_i = \theta_i + n_i$

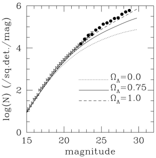
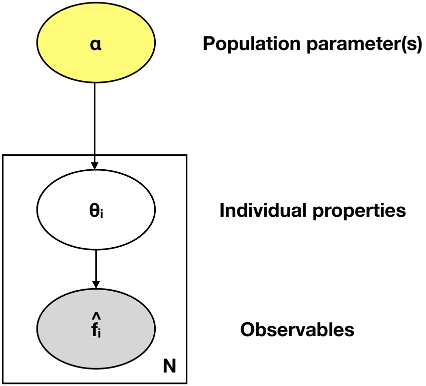
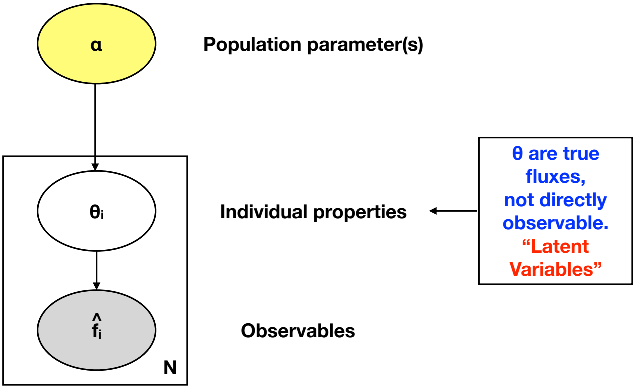


Figure: Ned Wright

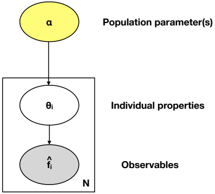
Number counts



Latent Variables



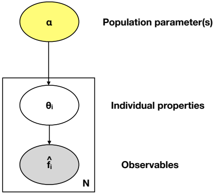
Ordinary Bayes vs Hierarchical Bayes



- Ordinary Bayes:

$$p(\alpha|\hat{f}) \propto p(\hat{f}|\alpha)\, p(\alpha)$$

Ordinary Bayes vs Hierarchical Bayes

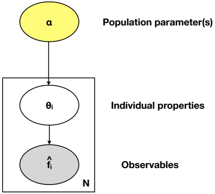


- Ordinary Bayes:

$$p(\alpha|\hat{f}) \propto p(\hat{f}|\alpha)\, p(\alpha)$$

- But we do not know $p(\hat{f}|\alpha)$!

Ordinary Bayes vs Hierarchical Bayes



- Ordinary Bayes:

$$p(\alpha|\hat{f}) \propto p(\hat{f}|\alpha)\, p(\alpha)$$

- But we do not know $p(\hat{f}|\alpha)$!
- Hierarchical Bayes:

$$\begin{aligned} p(\alpha, \theta|\hat{f}) &\propto p(\hat{f}|\theta, \alpha)\, p(\theta, \alpha) \\ &\propto p(\hat{f}|\theta, \alpha)\, p(\theta|\alpha)\, p(\alpha) \end{aligned} \quad (1)$$

Bayesian Hierarchical Models

Computing the posterior

$p(\theta|\mathrm{d})$ may be impossible to calculate directly
e.g. p(cosmology parameters θ|shapes of galaxies d)
Solution: make the problem MUCH harder:
Compute the joint probability of the cosmological parameters *and the shear map*

Joint distribution

$$\mathrm{p}(\theta\,|\,\mathrm{d}) = \int \mathrm{p}(\theta, \mathrm{map}\,|\,\mathrm{d})\,\mathrm{d}(\mathrm{map})$$

$$\mathrm{p}(\theta, \mathrm{map}\,|\,\mathrm{d}) \propto \mathcal{L}(\mathrm{d}\,|\,\theta, \mathrm{map})\,\mathrm{p}(\mathrm{map}|\theta)\,\pi(\theta)$$

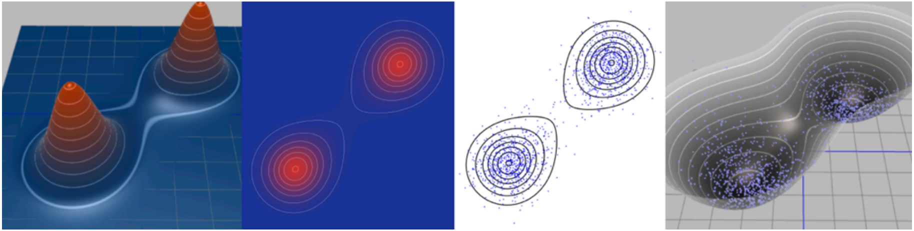
Joint map, parameter sampling

Latent parameters

Each pixel in the map is a parameter
10 cosmological parameters, plus 1,000,000 shear values
One million-dimensional probability distribution to calculate

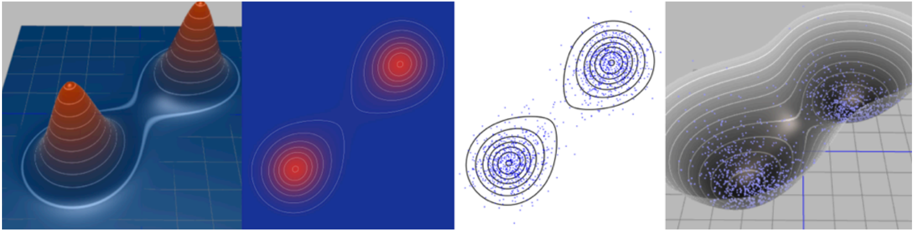
Sampling

- MCMC: Metropolis-Hastings fails since it is very hard to devise an efficient proposal distribution



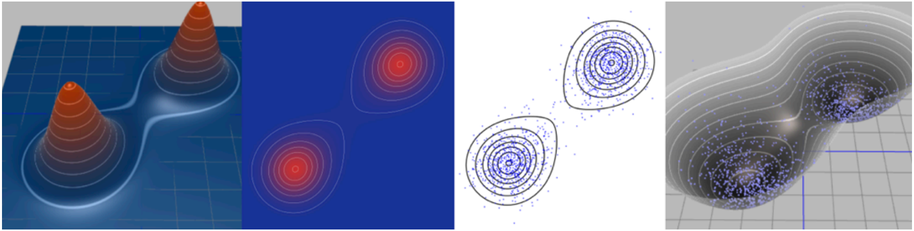
Sampling

- MCMC: Metropolis-Hastings fails since it is very hard to devise an efficient proposal distribution
- Gibbs sampling: effective if conditional distributions are known

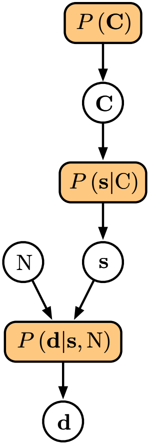


Sampling

- MCMC: Metropolis-Hastings fails since it is very hard to devise an efficient proposal distribution
- Gibbs sampling: effective if conditional distributions are known
- Hamiltonian Monte Carlo (HMC) works in very high dimensions (e.g. STAN)



Weak Lensing BHM: Forward Model or Generative Model



C = Power Spectrum

s = shear map

N = noise variance in each pixel

d = noisy shear estimates in each pixel

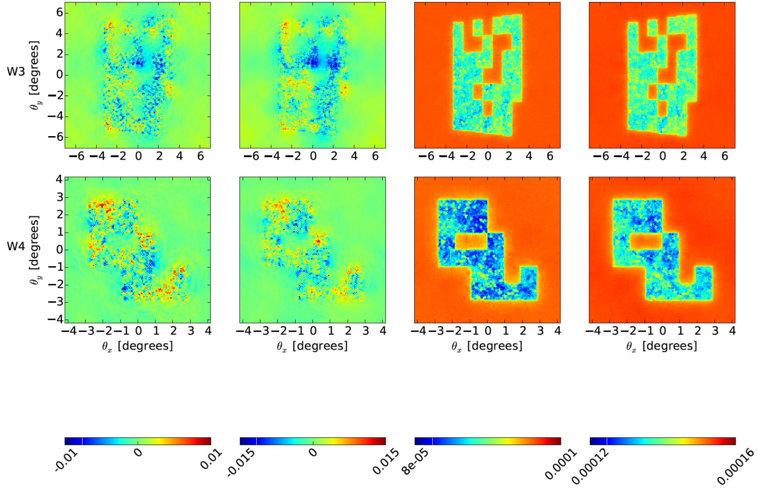
CFHTLenS

Alsing, AFH et al (2016). $\sim 130,000$ parameters; Gibbs sampling

BORG and SDSS

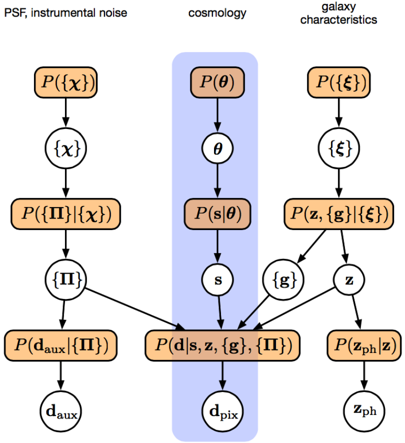
Weak lensing: CFHTLenS maps

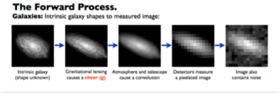
Alsing, Heavens & Jaffe (2017). $\sim 250,000$ parameters; Gibbs sampling

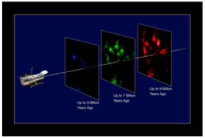


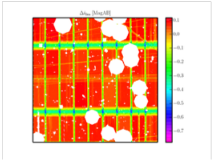
Weak Lensing BHM: Forward Model or Generative Model

Add in elements: uncertainties in redshifts, intrinsic alignments, etc









CFHTLenS weak lensing

Band powers

Make EE, BB, EB mode bandpowers the parameters (cosmology-independent)

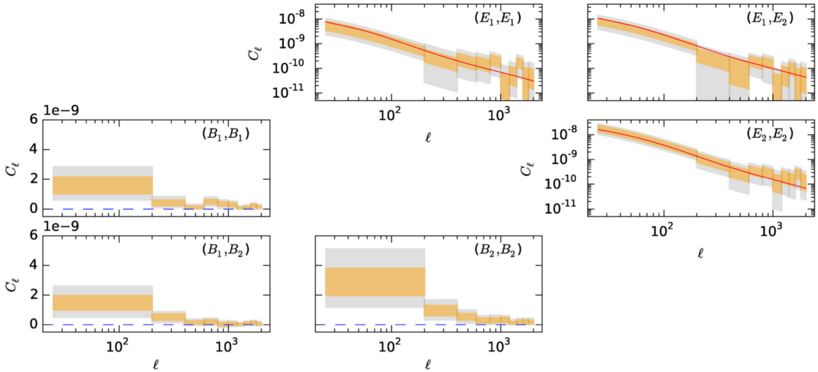
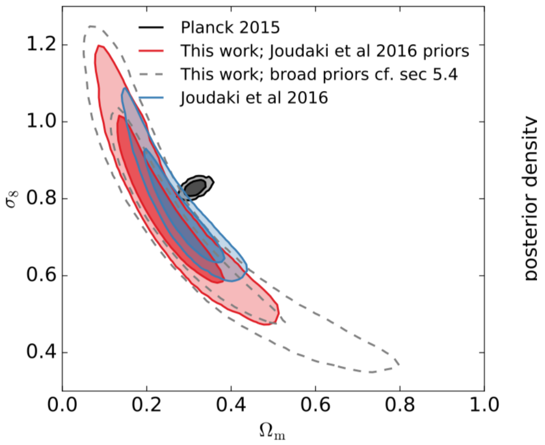


Figure: Alsing, AFH et al (2016)

CFHTLenS cosmological parameters



Summary of BHM

- Bayesian Hierarchical Models are a way to build a statistical model of the data by splitting into steps

Summary of BHM

- Bayesian Hierarchical Models are a way to build a statistical model of the data by splitting into steps
- Typically, decomposing into steps exposes what is needed - typically many conditional distributions

Summary of BHM

- Bayesian Hierarchical Models are a way to build a statistical model of the data by splitting into steps
- Typically, decomposing into steps exposes what is needed - typically many conditional distributions
- For complex data, this may be the *only* viable way to build the statistical model

Summary of BHM

- Bayesian Hierarchical Models are a way to build a statistical model of the data by splitting into steps
- Typically, decomposing into steps exposes what is needed - typically many conditional distributions
- For complex data, this may be the *only* viable way to build the statistical model
- The decomposition is usually very natural and logical

Summary of BHM

- Bayesian Hierarchical Models are a way to build a statistical model of the data by splitting into steps
- Typically, decomposing into steps exposes what is needed - typically many conditional distributions
- For complex data, this may be the *only* viable way to build the statistical model
- The decomposition is usually very natural and logical
- The model allows the proper propagation of errors from one layer to the next,

Summary of BHM

- Bayesian Hierarchical Models are a way to build a statistical model of the data by splitting into steps
- Typically, decomposing into steps exposes what is needed - typically many conditional distributions
- For complex data, this may be the *only* viable way to build the statistical model
- 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

Summary of BHM

- Bayesian Hierarchical Models are a way to build a statistical model of the data by splitting into steps
- Typically, decomposing into steps exposes what is needed - typically many conditional distributions
- For complex data, this may be the *only* viable way to build the statistical model
- 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 from a Bayesian statistical analysis

Model Comparison

- A higher-level question than parameter inference, in which one wants to know which theoretical framework ('model') is preferred, given the data (regardless of the parameter values)

Model Comparison

- A higher-level question than parameter inference, in which one wants to know which theoretical framework ('model') is preferred, given the data (regardless of the parameter values)
- The models may be completely different (e.g. compare Big Bang with Steady State, to use an old example),

Model Comparison

- A higher-level question than parameter inference, in which one wants to know which theoretical framework ('model') is preferred, given the data (regardless of the parameter values)
- The models may be completely different (e.g. compare Big Bang with Steady State, to use an old example),
- or variants of the same idea. E.g. comparing a simple cosmological model where the Universe is assumed to be flat, with a more general model where curvature is allowed to vary

Model Comparison

- A higher-level question than parameter inference, in which one wants to know which theoretical framework ('model') is preferred, given the data (regardless of the parameter values)
- The models may be completely different (e.g. compare Big Bang with Steady State, to use an old example),
- or variants of the same idea. E.g. comparing a simple cosmological model where the Universe is assumed to be flat, with a more general model where curvature is allowed to vary
- The sort of question asked here is often 'Do the data favour a more complex model?'

Model Comparison

- A higher-level question than parameter inference, in which one wants to know which theoretical framework ('model') is preferred, given the data (regardless of the parameter values)
- The models may be completely different (e.g. compare Big Bang with Steady State, to use an old example),
- or variants of the same idea. E.g. comparing a simple cosmological model where the Universe is assumed to be flat, with a more general model where curvature is allowed to vary
- The sort of question asked here is often 'Do the data favour a more complex model?'
- Clearly in the latter type of comparison the likelihood itself will be of no use - it will always increase if we allow more freedom.

Model Comparison

- Assuming uninformative priors for the models (i.e. the same a priori probability), the probability of the models given the data is simply proportional to the **Bayesian Evidence**.

Model Comparison

- Assuming uninformative priors for the models (i.e. the same a priori probability), the probability of the models given the data is simply proportional to the **Bayesian Evidence**.
- The **Bayesian Evidence**, or **Marginal Likelihood**, is the denominator in Bayes' theorem

$$p(\theta|d) = \frac{p(d|\theta)\pi(\theta)}{p(d)}$$

Model Comparison

- Assuming uninformative priors for the models (i.e. the same a priori probability), the probability of the models given the data is simply proportional to the **Bayesian Evidence**.
- The **Bayesian Evidence**, or **Marginal Likelihood**, is the denominator in Bayes' theorem

$$p(\theta|d) = \frac{p(d|\theta)\pi(\theta)}{p(d)}$$

- It is much more obvious if we include the model dependence as a condition:

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

Model Comparison

- Assuming uninformative priors for the models (i.e. the same a priori probability), the probability of the models given the data is simply proportional to the **Bayesian Evidence**.
- The **Bayesian Evidence**, or **Marginal Likelihood**, is the denominator in Bayes' theorem

$$p(\theta|d) = \frac{p(d|\theta)\pi(\theta)}{p(d)}$$

- It is much more obvious if we include the model dependence as a condition:

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

- The Bayesian Evidence normalises the posterior, so is

$$p(d|M) = \int d\theta \, p(d|\theta, M)\pi(\theta|M)$$

Model Comparison

$p(d|M)$ may not be large (but it is for ΛCDM)

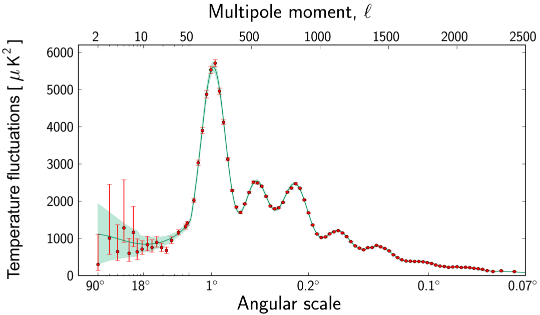


Figure: The Planck power spectrum, with the theoretical model with best fitting cosmological parameters.

Bayesian Evidence or Marginal Likelihood

- We denote two competing models by M and M'.

Bayesian Evidence or Marginal Likelihood

- We denote two competing models by M and M'.
- We denote by d the data vector, and by θ and θ' the parameter vectors (of length n and n').

Bayesian Evidence or Marginal Likelihood

- We denote two competing models by M and M'.
- We denote by d the data vector, and by θ and θ' the parameter vectors (of length n and n').
- Rule 1: Write down what you want to know.

Bayesian Evidence or Marginal Likelihood

- We denote two competing models by M and M'.
- We denote by d the data vector, and by θ and θ' the parameter vectors (of length n and n').
- Rule 1: Write down what you want to know.
- Here it is $p(M|d)$ - the probability of the model, given the data.

Bayesian Evidence or Marginal Likelihood

- We denote two competing models by M and M'.
- We denote by d the data vector, and by θ and θ' the parameter vectors (of length n and n').
- Rule 1: Write down what you want to know.
- Here it is $p(M|d)$ - the probability of the model, given the data.
- Use Bayes' theorem:

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

Bayesian evidence

- The Bayesian Evidence is

$$p(d|M) = \int d\theta\, p(d|\theta, M)\pi(\theta|M),$$

Bayesian evidence

- The Bayesian Evidence is

$$p(d|M) = \int d\theta \, p(d|\theta, M)\pi(\theta|M),$$

- If a model has no parameters, then the integral is simply replaced by $p(\mathrm{d}|M)$

Bayesian evidence

- The Bayesian Evidence is

$$p(d|M) = \int d\theta\, p(d|\theta, M)\pi(\theta|M),$$

- If a model has no parameters, then the integral is simply replaced by $p(\mathrm{d}|M)$
- The relative probabilities of two models is

$$\frac{p(M'|d)}{p(M|d)} = \frac{\pi(M')}{\pi(M)} \frac{\int d\theta'\, p(d|\theta', M')\pi(\theta'|M')}{\int d\theta\, p(d|\theta, M)\pi(\theta|M)}$$

Bayesian evidence

- The Bayesian Evidence is

$$p(d|M) = \int d\theta \, p(d|\theta, M)\pi(\theta|M),$$

- If a model has no parameters, then the integral is simply replaced by $p(\mathrm{d}|M)$
- The relative probabilities of two models is

$$\frac{p(M'|d)}{p(M|d)} = \frac{\pi(M')}{\pi(M)} \frac{\int d\theta' \, p(d|\theta', M')\pi(\theta'|M')}{\int d\theta \, p(d|\theta, M)\pi(\theta|M)}$$

- With uninformative priors on the models, $p(M') = p(M)$, this ratio simplifies to the ratio of evidences, called the **Bayes Factor**,

$$B \equiv \frac{\int d\theta' \, p(d|\theta', M') \, \pi(\theta'|M')}{\int d\theta \, p(d|\theta, M) \, \pi(\theta|M)}$$

The Kass & Raftery scale

$\|\ln B\|$	Interpretation
< 1	not worth more than a bare mention
1 to 3	positive
3 to 5	strong
> 5	very strong

But better to stick with probabilities rather than descriptions.

Nested models

- We assume that M' is a simpler model, which has fewer parameters in it ($n' < n$)

Nested models

- We assume that M' is a simpler model, which has fewer parameters in it ($n' < n$)
- We further assume that it is *nested* in Model M', i.e. the n' parameters of model M' are common to M, which has $p \equiv n - n'$ extra parameters in it. These parameters are fixed to fiducial values in M'.

Nested models

- We assume that M' is a simpler model, which has fewer parameters in it ($n' < n$)
- We further assume that it is *nested* in Model M', i.e. the n' parameters of model M' are common to M, which has $p \equiv n - n'$ extra parameters in it. These parameters are fixed to fiducial values in M'.
- Note that the more complicated model M will (if M' is nested) inevitably lead to a higher likelihood (or at least as high), but the evidence may favour the simpler model if the fit is nearly as good, through the smaller prior volume.

Nested models

- We assume uniform (and hence separable) priors in each parameter, over ranges $\Delta\theta$ (or $\Delta\theta'$). Hence $p(\theta|M) = (\Delta\theta_1 \ldots \Delta\theta_n)^{-1}$

Nested models

- We assume uniform (and hence separable) priors in each parameter, over ranges $\Delta\theta$ (or $\Delta\theta'$). Hence $p(\theta|M) = (\Delta\theta_1 \ldots \Delta\theta_n)^{-1}$
-
$$B = \frac{\int d\theta'\, p(d|\theta', M')}{\int d\theta\, p(d|\theta, M)} \frac{\Delta\theta_1 \ldots \Delta\theta_n}{\Delta\theta'_1 \ldots \Delta\theta'_{n'}}.$$

Nested models

- We assume uniform (and hence separable) priors in each parameter, over ranges $\Delta\theta$ (or $\Delta\theta'$). Hence $p(\theta|M) = (\Delta\theta_1 \ldots \Delta\theta_n)^{-1}$
- $$B = \frac{\int d\theta'\, p(d|\theta', M')}{\int d\theta\, p(d|\theta, M)} \, \frac{\Delta\theta_1 \ldots \Delta\theta_n}{\Delta\theta'_1 \ldots \Delta\theta'_{n'}}.$$
- Assume prior range includes (virtually) all the likelihood.

Nested models

- We assume uniform (and hence separable) priors in each parameter, over ranges $\Delta\theta$ (or $\Delta\theta'$). Hence $p(\theta|M) = (\Delta\theta_1 \ldots \Delta\theta_n)^{-1}$
- $$B = \frac{\int d\theta'\, p(d|\theta', M')}{\int d\theta\, p(d|\theta, M)}\, \frac{\Delta\theta_1 \ldots \Delta\theta_n}{\Delta\theta'_1 \ldots \Delta\theta'_{n'}}.$$
- Assume prior range includes (virtually) all the likelihood.
- In the nested case, the ratio of prior hypervolumes simplifies to

$$\frac{\Delta\theta_1 \ldots \Delta\theta_n}{\Delta\theta'_1 \ldots \Delta\theta'_{n'}} = \Delta\theta_{n'+1} \ldots \Delta\theta_{n'+p},$$

where $p \equiv n - n'$ is the number of extra parameters in the more complicated model.

Bayesian Evidence

Challenges: The evidence requires a multidimensional integration over the likelihood and prior, and this may be *very* expensive to compute.

- Nested sampling (multinest, polychord), where one tries to sample the likelihood in an efficient way.

Bayesian Evidence

Challenges: The evidence requires a multidimensional integration over the likelihood and prior, and this may be *very* expensive to compute.

- Nested sampling (multinest, polychord), where one tries to sample the likelihood in an efficient way.
- Approximations: e.g., AIC and BIC may be unreliable as they are based on the best-fit χ^2, and from a Bayesian perspective we want to know how much parameter space would give the data with high probability. Also don't include the prior. Not Bayesian.

Bayesian Evidence

Challenges: The evidence requires a multidimensional integration over the likelihood and prior, and this may be *very* expensive to compute.

- Nested sampling (multinest, polychord), where one tries to sample the likelihood in an efficient way.
- Approximations: e.g., AIC and BIC may be unreliable as they are based on the best-fit χ^2, and from a Bayesian perspective we want to know how much parameter space would give the data with high probability. Also don't include the prior. Not Bayesian.
- MCEvidence may be useful for computing Evidence from pre-existing MCMC chains

Gaussian Example

Assume everything is gaussian
Let M_0 be $d \sim \mathcal{N}(0, \sigma^2)$, and M_1 be $d \sim \mathcal{N}(\mu, \sigma^2)$, where the prior on μ is gaussian with zero mean and variance Σ^2. Let the measurement be $d = \lambda\sigma$.

$$p_1(d|\mu) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(d-\mu)^2/(2\sigma^2)}$$

and

$$p_1(\mu|d) = \frac{p_1(d|\mu)\,\pi_1(\mu)}{p_1(d)} = \frac{p_1(d|\mu)\,\pi_1(\mu)}{\int p_1(d|\mu)\,\pi_1(\mu) d\mu}$$

Gaussian Example

Hence

$$BF_{01} = \frac{p_1(d|\mu = 0)\,\pi_1(\mu = 0)}{p_1(d)}$$

i.e.,

$$BF_{01} = \frac{\frac{1}{\sqrt{2\pi}\sigma} e^{-d^2/(2\sigma^2)} \cdot \frac{1}{\sqrt{2\pi}\Sigma}}{\frac{1}{\sqrt{2\pi}\sigma} \frac{1}{\sqrt{2\pi}\Sigma} \int_{-\infty}^{\infty} e^{-(d-\mu)^2/(2\sigma^2)} e^{-\mu^2/(2\Sigma^2)} d\mu}$$

so

$$BF_{01} = \sqrt{1 + \frac{\Sigma^2}{\sigma^2}} \exp\left[-\frac{\lambda^2}{2(1 + \frac{\sigma^2}{\Sigma^2})}\right]$$

Gaussian Example

$$BF_{01} = \sqrt{1 + \frac{\Sigma^2}{\sigma^2}} \exp\left[-\frac{\lambda^2}{2(1+\frac{\sigma^2}{\Sigma^2})}\right]$$

If $\lambda \gg 1$, then $B_{01} \ll 1$ and M_1 is favoured. If $\lambda \simeq 1$ and $\sigma \ll \Sigma$, then M_0 is favoured (Occam's razor). If likelihood is much broader than prior, $\sigma \gg \Sigma$ then $BF_{01} \simeq 1$ and nothing has been learned.

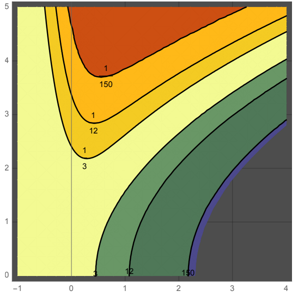


Figure: The Bayes Factor for a gaussian likelihood (variance σ^2), and a gaussian prior (variance Σ^2). The x axis $=\log_{10}(\Sigma/\sigma)$; the y axis is datum$/\sigma$. From Trotta (2008).

Summary

- Bayesian formalism can easily be generalised to model comparison

Summary

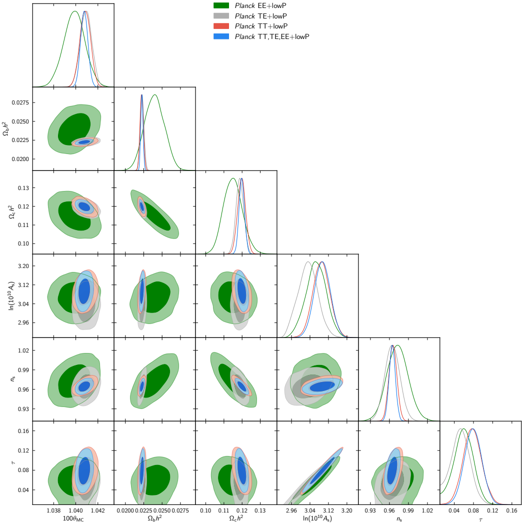
- Bayesian formalism can easily be generalised to model comparison
- Resulting integrals over parameter space may be challenging to compute

Summary

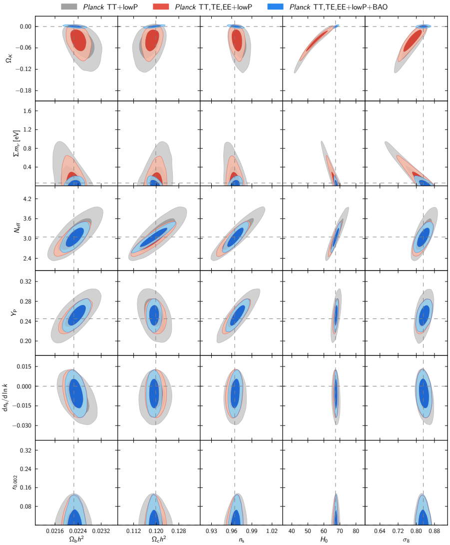
- Bayesian formalism can easily be generalised to model comparison
- Resulting integrals over parameter space may be challenging to compute
- Evidence ratios have sensitivity to the prior, even asymptotically

Planck parameter inference

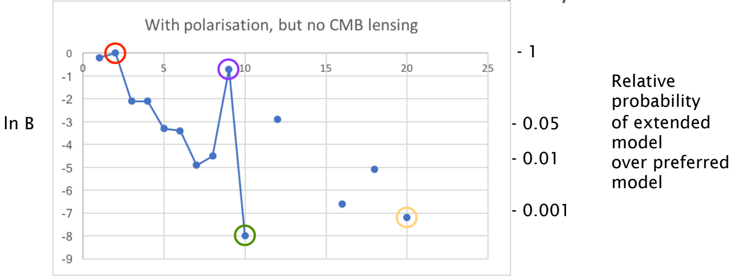
Assuming ΛCDM



Extensions to ΛCDM



BAYESIAN EVIDENCE: WHICH MODELS ARE PREFERRED? PLANCK TT,TE,EE DATA

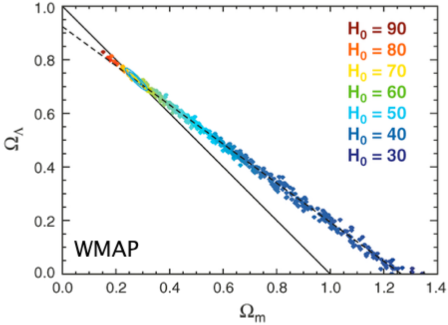


1. LCDM
2. Curvature
3. Non-standard lensing amp
4. Non-standard lensing fid
5. Neutrino number
6. Neutrino mass
7. Running of n_s
8. Tensor-to-scalar ratio
9. Dark Energy not Λ
10. Isocurvature modes
12. Non-standard BBN
16. Neutrino number and mass
18. Neutrino number and BBN
20. Sterile neutrinos

ICIC

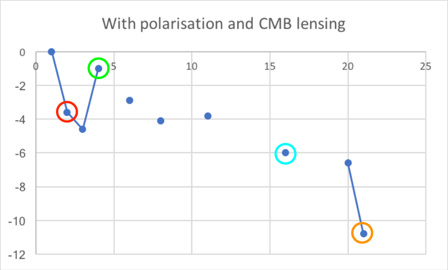
Figure: Heavens et al 2107

Planck tensions



Extensions to ΛCDM

WITH CMB LENSING



ln B

- 1
- 0.05
- 0.01
- 0.001
- 0.0001
- 0.00001

Relative probability of extended model over base ΛCDM model

1. LCDM
2. Curvature
3. Non-standard lensing amp
4. Non-standard lensing fid
5. Neutrino number
6. Neutrino mass
7. Running of n_s

8. Tensor-to-scalar ratio
9. Dark Energy not Λ
10. Isocurvature modes
12. Non-standard BBN
16. Neutrino number and mass
18. Neutrino number and BBN
20. Sterile neutrinos
21. Sterile neutrinos and r

State of Play

Analysis of Planck chains using MCEvidence (Heavens et al 2017)

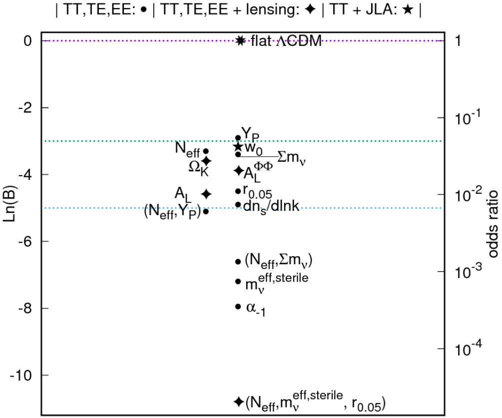
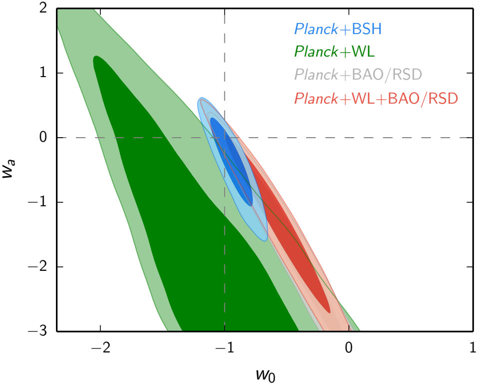


Figure: The Bayes Factor for Planck

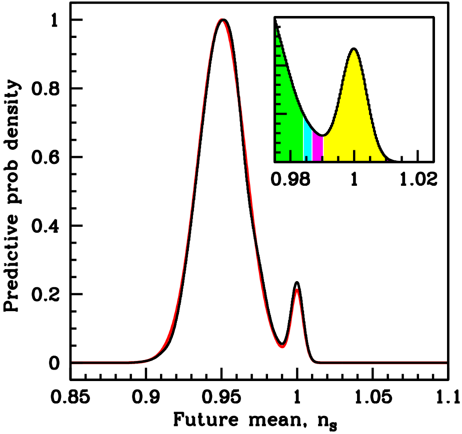
H_0

Planck Dark Energy Equation of State

$$w(a) = w_0 + w_a(a - 1).$$



Forecasting the future



Conclusions

- Assuming that data are gaussian-distributed will almost certainly not be good enough
- For likelihood-free parameter inference, or for approximating sampling distributions, massive data compression will also be necessary
- MOPED offers a way to do this without loss of information
- Bayesian Hierarchical Modelling is the principled solution to the analysis challenge
- For models that make subtly different predictions from ΛCDM, a very careful analysis will be necessary, including careful treatment of systematics and full propagation of errors
- Marginalising over uncertain parameters weakens sensitivity to new physics
- Model comparison may struggle to prefer non-ΛCDM models in future with high probability, unless we make new types of observation