cosmic inference with partition functions

from cosmological data to information on fundamental physics

Björn Malte Schäfer

30.Oct.2023

Faculty for Physics and Astronomy
Heidelberg University, Germany

Teilchentee, Heidelberg

UNIVERSITÄT
HEIDELBERG
ZUKUNFT
SEIT 1386

cosmology: what one knows and what not (yet)

- on large scales, spacetime exhibits the FLRW-symmetries, i.e. homogeneity and isotropy
- gravity is repulsive on large scales, well described by the cosmological constant Λ
- most of the matter is dark and interacts only gravitationally
- there are small, Gaussian and adiabatic fluctuations from an inflationary epoch
- one begins to understand galaxy formation quantitatively

but...

- new gravitational phenomena on large scales?
- particle physics connection for dark matter? axions?
- shape of the inflationary potential $V(\varphi)$, reheating and particle generation?
- dark ages and reionisation? AGN activity and regulation?

why cosmology is peculiar

- incredibly well-working standard model ΛCDM
- all parameters are of order unity or close to zero
- self-view of the community: era of precision cosmology (what about accuracy, though?)
- 6+ parameters, difficult/impossible to measure independently
- hierarchy of precision:

$$\Omega_m, \Omega_\Lambda, \sigma_8, \Omega_b \quad \dots \quad h, n_s, w_0 \quad \dots \quad \alpha, \beta, w_a, f_{\mathrm{NL}} \quad \dots \quad \tau_{\mathrm{NL}}, g_{\mathrm{NL}}$$

- tensions at marginal significance generate a lot of discussion lately
- non-parametric questions: cosmological principle
- fundamental physics in astrophysical observations: biases
- many sources of information: CMB, galaxy clustering, weak lensing, 21cm, supernova distance-redshift, time delays, cluster counts

weird questions that are difficult to ask (incomplete list)

- when do we stop believing in and old model and start believing in a new model?
  $\rightarrow$ **(Bayesian) evidence**
- how much information on a model is contained in data?
  $\rightarrow$ **information entropy**
- experimental design: decision between competing models or optimised discovery potential? or simply maximum precision in the standard model?
  $\rightarrow$ **surprise statistic**
- likelihoods, priors and posteriors independent from a given parameterisation? treatment of nuisance parameters?
  $\rightarrow$ **information geometry**
- quantification of tensions in posterior distributions and of data compatibility?
  $\rightarrow$ **Fisher-metric? Wasserstein-metric?**

Bayes' law for statistical inference

- Bayes' law makes a statement about conditional probabilities

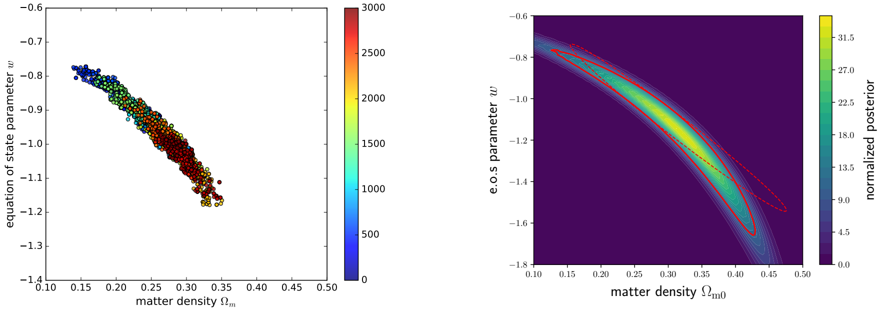
$$p(\theta|y) = \frac{\mathcal{L}(y|\theta)\pi(\theta)}{p(y)}$$

  with data y and parameter choice θ, specifically:
  - prior distribution $\pi(\theta)$: state of knowledge before experiment
  - likelihood $\mathcal{L}(y|\theta)$: probability that y is observed if θ is true
  - posterior $p(\theta|y)$: state of knowledge after the experiment
  - evidence as normalising factor, probability to obtain the data averaged over the prior

$$p(y) = \int \mathrm{d}^n\theta\, \mathcal{L}(y|\theta)\pi(\theta)$$

- for a given model and a known error process, $\mathcal{L}(y|\theta)$ is computable
- thermodynamics is a theory of information (E.T. Jaynes)... but can one use thermodynamics to understand information (of physical experiments) better?

two standard techniques in cosmology: Fisher-matrix and MCMC



supernova likelihood in (Ω_m, w): analytics vs. MCMC

- MCMC: random walk on the $\ln\mathcal{L}$-surface: samples from the posterior
- Fisher-matrix: approximate $\ln\mathcal{L}$ as $F_{\mu\nu}\theta^\mu\theta^\nu$
- more accurate parameterisations for the non-Gaussian case exist: DALI

Gauss-Markov-theorem and Gaussian posteriors

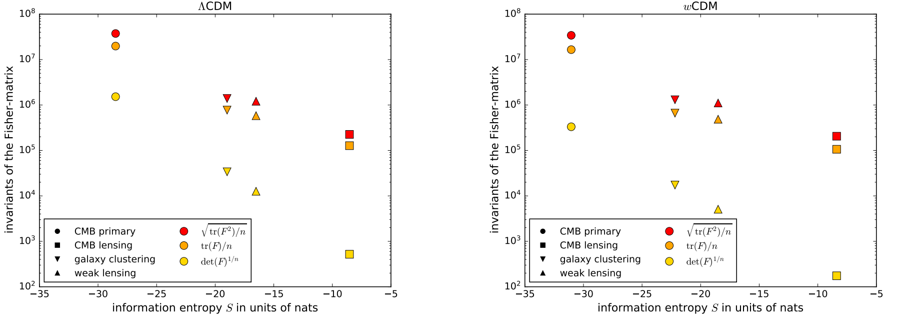
- Gauss-Markov-theorem: if the model is **linear** and the error process **Gaussian**, then
  - model parameters can be estimated with a linear estimator
  - estimate is unbiased
  - smallest error realised: Cramér-Rao inequality
  - likeihood is a Gaussian-function in the parameters, and the posterior a Gaussian distribution (for a flat prior)

  basis of all least squares estimation
- posterior is necessarily Gaussian

$$p(\theta^\mu) = \frac{1}{\sqrt{(2\pi)^n \det C}} \exp\left(-\frac{1}{2}\theta^\mu (C^{-1})_{\mu\nu}\theta^\nu\right)$$

  with covariance matrix $C^{\mu\nu} = \langle\theta^\mu\theta^\nu\rangle$
- Fisher matrix $F_{\mu\nu} = (C^{-1})_{\mu\nu}$ as inverse covariance
- good measurement with small uncertainty: small values in $C^{\mu\nu}$, large values in $F_{\mu\nu}$

absolute entropy: ΛCDM versus wCDM



absolute entropy of the posterior distribution

- information content: entropies are commensurate with Fisher-invariants

relative entropies

- relative entropies measure the relative amount of randomness in two distributions
- invariant under reparameterisation, in contrast to absolute entropies
- Kullback-Leibler divergence: relative entropy

$$\Delta S = \int \mathrm{d}^n\theta\, p(\theta) \ln \frac{p(\theta)}{q(\theta)}$$

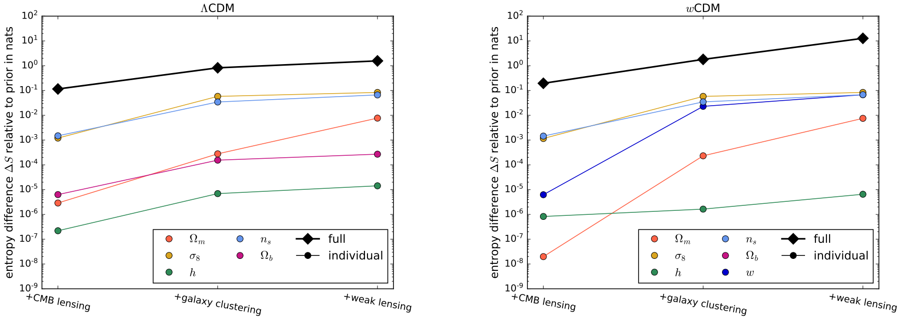
- α-divergence

$$\Delta S_\alpha = \frac{1}{\alpha - 1} \ln \int \mathrm{d}^n\theta\, p(\theta) \left(\frac{p(\theta)}{q(\theta)}\right)^{\alpha-1}$$

  incompatible with Bayes' law if $\alpha \neq 1$
- non-symmetric, exception $\alpha = 1/2$: Battacharyya-entropy

$$\Delta S_{1/2} = -2 \ln \int \mathrm{d}^n\theta\, \sqrt{p(\theta) q(\theta)}$$

relative entropy: ΛCDM versus wCDM



relative entropy of the posterior distribution

- reduction in uncertainty by probe combination

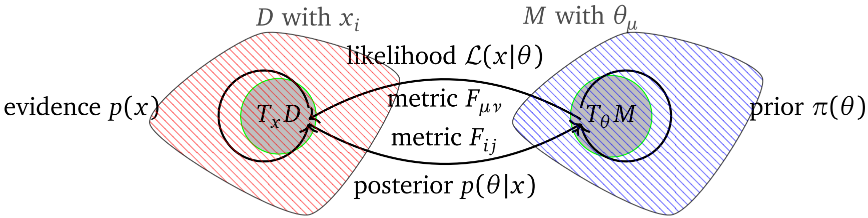
Fisher-metric: distance between distributions

- likelihood $\mathcal{L}(y|\theta) = p(y|\theta)$: conditional probability of the data y for a parameter set θ
- depending on the random process generating the experimental error, there is an entire distribution of y
- difference of two distributions $p(y|\theta)$ and $p(y|\theta+\epsilon)$:

$$\Delta S = \int \mathrm{d}^n y\, p(y|\theta) \ln \frac{p(y|\theta)}{p(y|\theta+\epsilon)} \simeq \frac{1}{2} F_{\mu\nu}\epsilon^\mu\epsilon^\nu$$

  quantify with Kullback-Leibler-divergence
- symmetric to lowest non-vanishing order
- quadratic distance measure with a positive definite $F_{\mu\nu}$
- curvature on the manifold associated with non-Gaussianity in distributions, Gaussian distribution equivalent with flatness
- $F_{\mu\nu}$: Fisher-metric, for Gaussians: $F_{\mu\nu}\epsilon^\mu\epsilon^\nu = \Delta\chi^2$

Bayes' manifolds



- Bayes' law links up two conditional probabilities:

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

  with prior $p(\theta)$ and evidence $p(y)$
  - Fisher-metric $F_{\mu\nu}$ in parameter space derived from $\mathcal{L}(y|\theta)$
  - Fisher-metric F_{ij} in data space derived from posterior $p(\theta|y)$
- invariants exist on the manifolds, quantify uncertainty

Bayes' manifolds

- Kullback-Leibler divergence in data space between posteriors $p(\theta|x)$ and $p(\theta|y)$: two data sets x and y, identical prior $p(\theta)$

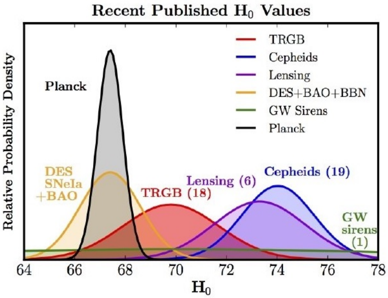
$$\Delta S(x,y) = -\int \mathrm{d}\theta \; p(\theta|x) \ln\left(\frac{\mathcal{L}(x|\theta)}{\mathcal{L}(y|\theta)}\frac{p(y)}{p(x)}\right) = \ldots$$

  substitute

$$\ldots = -\int \mathrm{d}\theta \; p(\theta|x) \ln\left(\frac{\mathcal{L}(x|\theta)}{\mathcal{L}(y|\theta)}\right) + \ln\frac{p(x)}{p(y)} = -\int \mathrm{d}\theta \; p(\theta|x) r + B$$

  identify
  - likelihood ratio r
  - evidence ratio B
- really weird: entropy = average logarithmic likelihood ratio + logarithmic evidence ratio

tensions in cosmology



posteriors for h (arXiv:1911.06456)

- many tensions in cosmology
- origin unclear: could be a signature of "new physics" or a badly understood systematic
- even quantification is unclear, entropies would be asymmetric
- perhaps it's rather a question of data consistency? F_{ij}?

Bayes' evidence and partition function

- Bayes-evidence

$$p(y) = \int \mathrm{d}^n\theta \; \mathcal{L}(y|\theta)p(\theta)$$

- distributions are members of the exponential family

$$p(y) = \int \mathrm{d}^n\theta \; \exp\left(-\frac{1}{2}\chi^2(y|\theta)\right)\exp(-\phi(\theta))$$

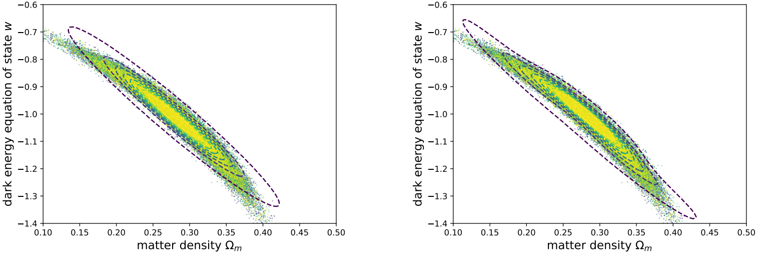
- build in Laplace transform and an inverse temperature β:
  mathematical structure of a partition function, $\chi^2/2+\phi$ plays the role of a potential

$$Z[\beta, J_\mu] = \int \mathrm{d}^n\theta \; \exp\left(-\beta\left[\chi^2(y|\theta)/2+\phi(\theta)\right]\right)\exp(\beta J_\mu\theta^\mu)$$

- cumulants of the posterior

$$\kappa_n = \left.\frac{\partial^n \ln Z}{\partial J^n}\right|_{J=0,\beta=1}$$

Bayes' evidence and partition function



approximations to the supernova likelihood in (Ω_m, w)

- Gauß-Markov: a linear model leads to a parabolic χ^2 and therefore to a Gaussian integral
- in general: factorisation of the partition function into a Gaussian and a non-Gaussian part
- series expansion of the non-Gaussian part: recover Gram-Charlier-series

weird and interesting relations

- define Helmholtz-free energy $F(\beta)$ from $Z[\beta, J_\mu]$

$$F(\beta) = -\frac{1}{\beta} \ln Z[\beta, J_\mu] \quad \text{with} \quad Z = \int \mathrm{d}^n\theta \; [\mathcal{L}(y|\theta)\pi(\theta)]^\beta$$

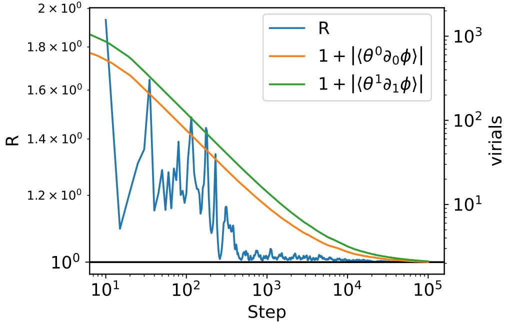
- derive **information entropy** of the posterior distribution at $\beta = 1$

$$S = \beta^2 \frac{\partial F}{\partial \beta} = -\int \mathrm{d}^n\theta \; p(\theta|y) \ln p(\theta|y)$$

- related partition with $1/\beta$ weighting yields **Kullback-Leibler divergence** ΔS between posterior $p(\theta|y)$ and prior $\pi(\theta)$

$$Z = \int \mathrm{d}^n\theta \; \mathcal{L}(y|\theta)^\beta \pi(\theta)^{\beta+1/\beta} \quad \rightarrow \quad \Delta S = \int \mathrm{d}^n\theta \; p(\theta|y) \ln \frac{p(\theta|y)}{\pi(\theta)}$$

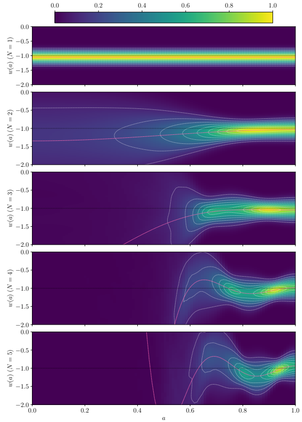
virialisation, equipartition and thermalisation of MCMC



Gellmann-Rubin R versus virialisation

- Gellmann-Rubin criterion R for convergence of MCMC sampling: sampling fair and stationary
- Hamilton Monte-Carlo: degrees of freedom should separate and carry same amounts of total energy $\propto 1/\beta$

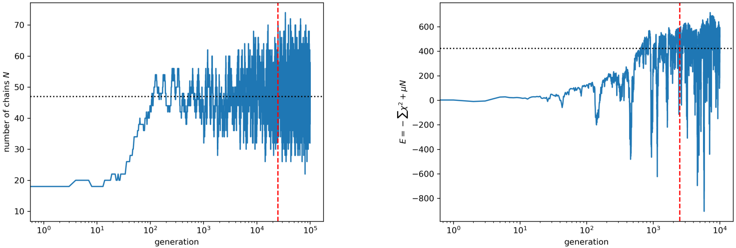
statistical inference in more complicated cases: functional inference



error band around the dark energy eos $w(a)$

- physical models might depend on a function $w(a)$ instead of (countably many) paramters θ
- partition function becomes a path integral over the space of functions $w(a)$
- Gaussian path integrals feature the Fisher-functional $F_{w(a),w(a')}$ as a generalisation to the Fisher-matrix $F_{\mu\nu}$
- example from cosmology: dark energy equation of state $w(a)$, $a = 0\ldots1$
- generalisation to partition functions for differential equations? to Lagrange functions?

sampling with the macrocanonical ensemble: *Kill & Spawn*-algorithm



sampling from the macrocanonical ensemble

- ensemble of samplers, controlled by a chemical potential μ:

$$\Xi[\beta, J_\alpha, \mu] = \sum_n \frac{1}{n!} Z[\beta, J_\alpha]^n \exp(n\beta\mu)$$

- samplers are produced in regions of low χ^2, increase sampling efficiency

summary

- thermodynamics as a theory of information extracted from data in the inference process
- central element: partition functions, suitable for analytical work
- analogous quantities in thermodynamics and information theory
- thermodynamics of MCMC-sampling, burn-in and equilibration
- generalisation to functional spaces
- ensembles of samplers: macrocanonical sampling

many thanks to: Lennart Röver, Benedikt Schosser, Rebecca Maria (Maria) Kuntz, Maxi Herzog, Heinrich von Campe