Statistical Analysis in Cosmology

Francisco-Shu Kitaura

September 20, 2017

1 Introduction to information
- Notions of information
- Information and entropy
- Probability theory axioms
- Fisher information

2 Bayesian inference steps
- Introduction
- Bayes theorem
- Bayesian evidence and Bayes factors
- Bayesian inference steps

3 Bayesian modelling
- Basic data model: link between prior and likelihood
- Informative priors
- Non-informative priors

4 Bayesian inference: posterior sampling
- Gibbs sampling
- Hamiltonian Monte Carlo sampling

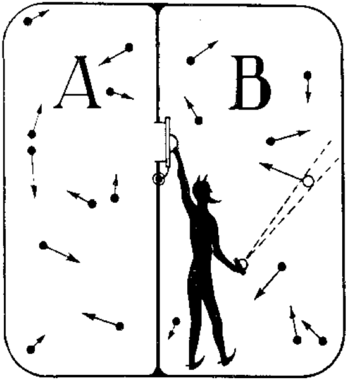
5 Data modelling

Notions of information

- What is information?
- How do we quantify it?
- Etymology: (*Latin*) informare: *give form to the mind*
- Systems theory: *information is any type of pattern that influences the formation of other patterns*
- J. D. Bekenstein 03: *the physical world is made of information itself*
- Relation between entropy and information: Maxwell's demon

Information and entropy: Maxwell's demon

- Container in thermal equilibrium divided into 2 parts A and B with a trapdoor
- Demon lets faster molecules pass from B to A
- Kinematic energy is reduced in B
- Violation of second law of thermodynamics?
- The information on the molecule velocities increases the overall entropy!
- Information and entropy are tightly related.



Information and entropy

- Boltzmann entropy, density of equal probable microstates $W = N!/\Pi_i N_i!$, with N particles in N_i microstate of position and momentum

$$S = k_B \ln W \tag{1}$$

- Gibbs entropy, microstates with different probabilities

$$S = -k_B \sum_i P_i \ln P_i \tag{2}$$

Shannon's entropy

- Claude Elwood Shannon (April 30, 1916 February 24, 01), father of information theory, electronic ingeneer worked for Bell Labs (Shannon 48 *A mathematical theory of communication* link)
- Shannon's entropy

$$H(\mathbf{x}) = -\sum_i P(x_i) \log_b P(x_i) \tag{3}$$

units of entropy: b=2: bits; b=e: nats; b=10: dits

- conditional entropy

$$H(\mathbf{y}|\mathbf{x}) = \sum_{i,j} P(x_i, y_j) \log_b \frac{P(y_j)}{P(x_i, y_j)} \tag{4}$$

Shannon's entropy

- Kullback Leibler distance (or divergence) between two distributions

$$D_{\mathrm{KL}}(P||Q) = \sum_i P_i \log_b \frac{P_i}{Q_j} \tag{5}$$

- Mutual information

$$I(x,y) = \sum_{i,j} P(x_i, x_j) \log_b \frac{P(x_i, x_j)}{P(x_i)P(y_j)} \tag{6}$$

Probability theory axioms

- Sum rule: *OR* (Venn diagram)

$$P(\mathbf{a}_1 + \mathbf{a}_2|\mathbf{c}) = P(\mathbf{a}_1|\mathbf{c}) + P(\mathbf{a}_2|\mathbf{c}) - P(\mathbf{a}_1, \mathbf{a}_2|\mathbf{c}) \quad (7)$$

- Product rule: *AND*

$$P(\mathbf{a}, \mathbf{b}|\mathbf{c}) = P(\mathbf{a}|\mathbf{b}, \mathbf{c})P(\mathbf{b}|\mathbf{c}) \quad (8)$$

- Invariance under permutation of arguments

$$P(\mathbf{s}, \mathbf{d}|\mathbf{p}) = P(\mathbf{d}, \mathbf{s}|\mathbf{p}) \quad (9)$$

where **s** is some signal (or set of model parameters $\{s_1, s_2, \dots\}$), **d** some data and **p** some prior information. Probability distribution functions are always conditioned on some prior information!, although sometimes we skip **p** for simplicity.

Cramer-Rao inequality/lower bound

- unbiased estimator

$$\langle \hat{s} \rangle = s \tag{10}$$

$$\langle \hat{s} - s \rangle \equiv \int \mathrm{d}d\, P(d|s)(\hat{s} - s) = 0, \tag{11}$$

with $P(d|s)$ being the likelihood of the data given the model or signal s.

- $\partial/\partial s \rightarrow$

$$\int \mathrm{d}d\,(\hat{s} - s)\frac{\partial P(d|s)}{\partial s} - \int \mathrm{d}d\, P(d|s) = 0 \tag{12}$$

- $\partial P(d|s)/\partial s = P(d|s)\partial \ln P(d|s)/\partial s \rightarrow$

$$\int \mathrm{d}d \left[\frac{\partial \ln P(d|s)}{\partial s}\sqrt{P(d|s)}\right]\left[(\hat{s} - s)\sqrt{P(d|s)}\right] = 1 \tag{13}$$

Cramer-Rao inequality/lower bound

- Cauchy Schwarz inequality, inner product:

$$|\langle x, y\rangle|^2 \leq \langle x, x\rangle\langle y, y\rangle \tag{14}$$

$$\rightarrow \left[\int \mathrm{d}d\, P(d|s)\left(\frac{\partial \ln P(d|s)}{\partial s}\right)^2\right]\left[\int \mathrm{d}d\, P(d|s)(\hat{s}-s)^2\right] \geq 1 \tag{15}$$

- Mean Squared Error (MSE) of the model parameters or signal as

$$e^2(s) \equiv (\Delta s)^2 \equiv \langle(\hat{s}-s)^2\rangle = \int \mathrm{d}d\, P(d|s)(\hat{s}-s)^2 \tag{16}$$

- Fisher information

$$\mathcal{F}(s) \equiv \int \mathrm{d}d\, P(d|s)\left(\frac{\partial \ln P(d|s)}{\partial s}\right)^2 \equiv \left\langle\left(\frac{\partial \ln P(d|s)}{\partial s}\right)^2\right\rangle$$

Cramer-Rao inequality/lower bound

- $$\Delta s \geq \sqrt{\mathcal{F}^{-1}} \tag{18}$$
- If $e^2\mathcal{F} = 1 \rightarrow$ Minimum Variance Unbiased (MVU) estimator
- In general there is a statistical bias $\mathrm{B}(\hat{s}) \equiv \langle \hat{s} \rangle - s$

$$\mathrm{MSE}(\hat{s}) = \mathrm{VAR}(\hat{s}) + \mathrm{B}^2(\hat{s}) \tag{19}$$

$$\mathrm{VAR}(\hat{s}) \equiv \sigma^2(s) \equiv \langle (\hat{s} - \langle \hat{s} \rangle)^2 \rangle \tag{20}$$

Fisher information

- Score

$$\mathcal{S} \equiv \frac{\partial}{\partial s} \ln P(d|s) = \frac{1}{P(d|s)} \frac{\partial P(d|s)}{\partial s} \tag{21}$$

- Fisher information: variance of the score

$$\mathcal{F}(s) \equiv \langle \left(\frac{\partial \ln P(d|s)}{\partial s} \right)^2 \rangle \tag{22}$$

- if the regularity condition is fulfilled

$$\int \mathrm{d}d \, \frac{\partial^2 P(d|s)}{\partial s^2} = 0 \tag{23}$$

- we have

$$\mathcal{F}(s) = -\langle \frac{\partial^2 \ln P(d|s)}{\partial s^2} \rangle \tag{24}$$

Fisher information

- proof:

$$\mathcal{F}(s) = -\langle \frac{\partial^2 \ln P(d|s)}{\partial s^2} \rangle \quad (25)$$

$$= -\int \mathrm{d}d\, P(d|s) \left[\frac{\partial}{\partial s} \left(\frac{1}{P(d|s)} \frac{\partial P(d|s)}{\partial s} \right) \right]$$

$$= -\int \mathrm{d}d\, P(d|s) \left[-\frac{1}{P(d|s)^2} \left(\frac{\partial P(d|s)}{\partial s} \right)^2 + \frac{1}{P(d|s)} \frac{\partial^2 P(d|s)}{\partial s^2} \right]$$

$$= \int \mathrm{d}d\, P(d|s) \frac{1}{P(d|s)^2} \left(\frac{\partial P(d|s)}{\partial s} \right)^2 - \int \mathrm{d}d\, \frac{\partial^2 P(d|s)}{\partial s^2} \quad (26)$$

Fisher information

- Generalization to Fisher matrix

$$\mathcal{F}(s)_{ij} = -\langle \frac{\partial^2}{\partial s_i \partial s_j} \ln P(d|s) \rangle \tag{27}$$

- Information may be seen to be a measure of the *sharpness*/curvature of the support curve ($\ln P(d|s)$) near the Maximum Likelihood estimate of **s**.
- the Cramer-Rao inequality: $\Delta s_i \geq \sqrt{\mathcal{F}_{ii}^{-1}}$.
- examples: a) Gaussian likelihood $-\ln P(d|s) \simeq -\ln P(s_0|s) = 1/2(s-s_0)^2/\sigma^2 \rightarrow F = 1/\sigma^2 \rightarrow \Delta s = \sigma$ based on some fiducial model s_0. b) Forecast for redshift surveys (White et al 08): link. c) You can also use e.g. icosmo: link. d) See also notes from a Winter School in the Canary islands from Licia Verde link. See Yun Wang's talk!

Bayesian inference

Bayesian inference is based on (Thomas) Bayes theorem and in particular Bayesian probability theory (further developed by Pierre-Simon Laplace). We focus on the objectivists view, in which probability is a reasonable expectation that represents the state of knowledge. It relies on the (Richard T.) Cox theorem (how to construct a probability theory from a set of logical postulates). The Bayesian view is a different description than the frequentist view, which relies on the frequency of a phenomenon (how frequently something happens in an infinite number of trials). In the Bayesian framework we work with degrees of belief or credences. It is a quantitative approach to the abductive inference vs the deductive or inductive approaches.

Bayes theorem: the posterior/inference

- We use the product and the invariance rules

$$P(\mathbf{s},\mathbf{d}|\mathbf{p}) = P(\mathbf{s}|\mathbf{d},\mathbf{p})P(\mathbf{d}|\mathbf{p}) \tag{28}$$

$$P(\mathbf{d},\mathbf{s}|\mathbf{p}) = P(\mathbf{d}|\mathbf{s},\mathbf{p})P(\mathbf{s}|\mathbf{p}), \tag{29}$$

and get

$$P(\mathbf{s}|\mathbf{d},\mathbf{p}) = \frac{P(\mathbf{s}|\mathbf{p})P(\mathbf{d}|\mathbf{s},\mathbf{p})}{P(\mathbf{d}|\mathbf{p})} \tag{30}$$

$$\mathrm{posterior} = \mathrm{prior} \times \mathrm{likelihood}/\mathrm{evidence}$$

- Prior: $P(\mathbf{s}|\mathbf{p})$
- Likelihood: $\mathcal{L}(\mathbf{s}|\mathbf{d},\mathbf{p}) = P(\mathbf{d}|\mathbf{s},\mathbf{p})$
- Bayesian notion of information: information is encoded in conditional probability distribution functions.
- Machine learning: update of prior with posterior

Evidence

- normalization of the posterior
- marginalization over the signal

$$P(\mathbf{d}|\mathbf{p}) = \int \mathrm{d}\mathbf{s}\, P(\mathbf{d},\mathbf{s}|\mathbf{p}) = \int \mathrm{d}\mathbf{s}\, P(\mathbf{s}|\mathbf{p})P(\mathbf{d}|\mathbf{s},\mathbf{p}) \tag{31}$$

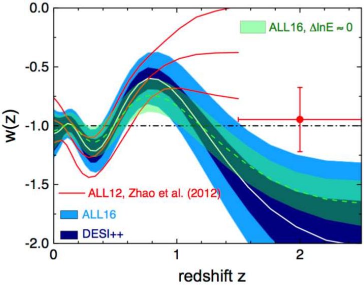
- Bayesian model comparison, Bayes factor for models M_1 and M_2 with model parameters s_1 and s_2

$$K = \frac{P(d|M_1) = \int \mathrm{d}s_1\, P(d, s_1|M_1) = \int \mathrm{d}s_1\, P(s_1|M_1)P(d|s_1, M_1)}{P(d|M_2) = \int \mathrm{d}s_2\, P(d, s_2|M_2) = \int \mathrm{d}s_2\, P(s_2|M_2)P(d|s_2, M_2)} \tag{32}$$

One can use Jeffreys criteria. $K > 0$ supports M_1, $K < 0$ supports M_2 (see e.g. Kass and Raftery 94 link).

Example: Dynamical Dark Energy, Zhao et al 17. Nat.Astr.

There is tension within Λ-CDM framework: in matter density fraction $\Omega_{\rm M}$ from Ly-α, and in the Hubble constant H_0 from SN, both as compared to CMB. This can be alleviated with a dynamical dark energy: link (see also Bernal et al 16 link; and Miguel Zumalacárregui's talk with Galileon gravity models link).



While the surprise (difference between the expected and actual Kullback-Leibler distances) favours dynamical dark energy with 3.5 σ significance, Bayesian evidence is insufficient to favour it over Λ-CDM.

Bayesian inference steps

- Definition of the prior: knowledge of the underlying signal
- Definition of the likelihood: nature of the observed data
- Linking the prior to the likelihood: link signal to the data
- Bayes theorem: the posterior

and then

- Evidence computation: e.g. nested sampling
- Maximization of the posterior: Maximum a posteriori: MAP
- Sampling the posterior: MCMC, importance sampling, Metropolis-Hasting, Hamiltonian Monte Carlo, Gibbs-sampling, population Monte Carlo, etc.

See Edwin Thompson Jaynes ▸ link based on statistical mechanics methods by J. Willard Gibbs, and for modern MCMC reviews Neal 93 ▸ link, 12 ▸ link.

Data model: signal degradation model

Let us start with simple data models.

- Nonlinear data model: data $\mathbf{d}$ m-vector, signal $\mathbf{s}$ n-vector, $n \gg m$

$$\mathbf{d} = R(\mathbf{s}) + \epsilon \tag{33}$$

- Linear data model

$$\mathbf{d} = \mathbf{R}\mathbf{s} + \epsilon \tag{34}$$

$$d_i = \sum_j R_{ij} s_j + n_i \tag{35}$$

- $\mathbf{R}$ response operator ($m \times n$ matrix) may include: mask, selection fnct, foregrounds, blurring fnct, PSF, pixel window ...
- ϵ noise: random component, white noise, colored noise, attention: mask, selection fct, pixel window etc.

Informative priors

- Gaussian prior (Wiener 49, Rybicki & Press 92, Zaroubi et al 95) $\rightarrow$ Thikonov regularization
- Lognormal prior/nonlinear transformation (Tarantola & Valette 82, FSK et al 10)
- Expanded Gaussian prior (Juszkiewiz et al 95; Bernardeau & Kofman 95; Colombi 94, FSK 10)

Gaussian prior+Gaussian likelihood

- Gaussian prior

$$P(\mathbf{s}|\mathbf{p}) \propto \exp\left(-\frac{1}{2}\mathbf{s}^{+}\mathbf{S}^{-1}\mathbf{s}\right) \tag{36}$$

with the covariance matrix of the signal $\mathbf{s}$ (e.g. the power-spectrum in Fourier space of the overdensity field) $\mathbf{S} = \langle \mathbf{s}\mathbf{s}^{+}\rangle_{(\mathbf{s}|\mathbf{p})}$.

- Gaussian likelihood

$$P(\mathbf{d}|\mathbf{s},\mathbf{p}) \propto \exp\left(-\frac{1}{2}(\mathbf{d}-\mathbf{R}\mathbf{s})^{+}\mathbf{N}^{-1}(\mathbf{d}-\mathbf{R}\mathbf{s})\right) \tag{37}$$

- with the noise covariance matrix (e.g. the variance of the Poisson likelihood) $\mathbf{N} \equiv \langle \boldsymbol{\epsilon}\boldsymbol{\epsilon}^{+}\rangle_{(\boldsymbol{\epsilon}|\mathbf{p})}$
- Wiener-filter (Wiener 49, Rybicki & Press 92, Zaroubi et al 95, $\rightarrow$ Thikonov regularization, see also FSK& Ensslin 08 link) $\mathbf{F}$: $\hat{\mathbf{s}} = \mathbf{F}\mathbf{d}$ with 2 equivalent formulations (exercise: demonstrate it) see Jenny Sorce talk!
$\mathbf{F} = (\mathbf{S}^{-1} + \mathbf{R}^{+}\mathbf{N}^{-1}\mathbf{R})^{-1}\mathbf{R}^{+}\mathbf{N}^{-1} = \mathbf{S}\mathbf{R}^{+}(\mathbf{R}\mathbf{S}\mathbf{R}^{+} + \mathbf{N})^{-1}$

Non-informative priors

- Flat prior
- Jeffrey's prior (see e.g. prior for the power-spectrum FSK & Ensslin 08)
- Entropic prior (Jaynes 63, see also Narayan & Nityananda 86, Skilling 89, FSK & Ensslin 08)

Flat prior+Gaussian likelihood

- improper prior: integral diverges to infinity
- maximization leads to maximum likelihood ML
- COBE-filter or CMB map making algorithm:
  $\mathbf{F} = (\mathbf{R}^{+}\mathbf{N}^{-1}\mathbf{R})^{-1}\mathbf{R}^{+}\mathbf{N}^{-1}$

Flat prior+Poissonian likelihood

- Poissonian likelihood

$$P(\mathbf{N}|\boldsymbol{\lambda},\mathbf{p}) = \Pi_i \exp(-\lambda_i)\frac{\lambda_i^{N_i}}{N_i} \quad (38)$$

- Richardson-Lucy deconvolution algorithm (Richardson 72, Lucy 74, Shepp & Vardi 82)

Lognormal prior

See Coles and Jones 91 and FSK & Angulo 12.
Let us consider only dark matter and assume the single stream limit. The first moment of the Vlasov equation describing the phase-space dynamics yields the continuity equation:

$$\frac{\partial \rho}{\partial \tau} + \nabla_{\mathbf{r}}(\rho \mathbf{v}) = 0, \tag{39}$$

which can be expanded

$$\frac{\partial \rho}{\partial \tau} + (\mathbf{v} \cdot \nabla_{\mathbf{r}})\rho + \rho \nabla_{\mathbf{r}} \cdot \mathbf{v} = 0, \tag{40}$$

We can write this equation in Lagrangian coordinates introducing the total derivative

$$\frac{1}{\rho}\frac{\mathrm{d}\rho}{\mathrm{d}\tau} + \nabla_{\mathbf{r}} \cdot \mathbf{v} = 0. \tag{41}$$

As long as we can follow particles (no shell-crossings) we can also write the continuity equation as $\ln(1+\delta) = -\int \mathrm{d}\tau \nabla_{\mathbf{r}} \cdot \mathbf{v}$. see Bridget Falk's, Rien van de Weygaert, and Raul Angulo's talk!

Lognormal prior

Linear expansion in the velocity yields a linear overdensity field term and higher order contributions can be summarised by δ^+:

$$\ln(1+\delta) = \delta_{\rm L} + \delta^+, \tag{42}$$

Taking the ensemble average of the previous equation we find that $\mu \equiv \langle \delta^+ \rangle$, since $\langle \delta_{\rm L} \rangle$ and thus

$$\delta_{\rm L} = \ln(1+\delta) - \mu, \tag{43}$$

For that reason we can assume a Gaussian distribution for $\delta_{\rm L}$ in a certain range of scales

$$P(\boldsymbol{\delta}_{\rm M}|\mathbf{S}) = \frac{1}{\sqrt{(2\pi)^{N_{\rm cells}}\det(\mathbf{S})}} \prod_k \frac{1}{1+\delta_{{\rm M}k}} \tag{44}$$

$$\times \exp\left(-\frac{1}{2}\sum_{ij}\left(\ln(1+\delta_{{\rm M}i})-\mu_i\right) S^{-1}_{ij} \left(\ln(1+\delta_{{\rm M}j})-\mu_j\right)\right) ,$$

Lognormal prior/nonlinear transformation (Tarantola & Valette 82, FSK et al 10; Jasche, FSK et al 10)

Maximum a posteriori

- Let us define the *energy* $E(\mathbf{s})$

$$E(\mathbf{s}) \equiv -\ln\left(P\left(\mathbf{s}|\mathbf{d},\mathbf{S}\right)\right), \tag{45}$$

- MAP

$$\frac{\partial E(\mathbf{s})}{\partial s_l} = 0, \tag{46}$$

- Krylov conjugate gradient schemes (FSK & Ensslin 08; FSK, Jasche, & Metcalf 09, for Wiener Filter also SVD, Cholesky-D, or messenger approach Elsner & Wandelt 13; Jasche & Lavaux 15)

$$s_i^{j+1} = s_i^j - \sum_k T_{ik}\frac{\partial E(\mathbf{s})}{\partial s_k}, \tag{47}$$

Markov Chains

- Andrey Markov (June 14, 1856 N.S. July 20, 1922) was a Russian mathematician
- Famous papers:
  - A.A. Markov. "Extension of the limit theorems of probability theory to a sum of variables connected in a chain". reprinted in Appendix B of: R. Howard. Dynamic Probabilistic Systems, volume 1: Markov Chains. John Wiley and Sons, 1971. (original in Russian 1906)
  - applied to language and vowels: link
- Gibbs-sampling, Metropolis-Hastings, Hybrid MCMC, Hamiltonian MCMC, etc

Gaussian distributions: Gibbs sampling

The Gibbs algorithm (Geman & Geman 84) samples from the joint PDF by repeatedly replacing each component with a value drawn from its distribution conditional on the current values of all other components. The Gibbs sampler starts with some initial values $\boldsymbol{\theta}^{(0)} = (\theta_1^{(0)}, ..., \theta_n^{(0)})$ and obtains new updates $\boldsymbol{\theta}^{(j)} = (\theta_1^{(j)}, ..., \theta_n^{(j)})$ from the previous step $\boldsymbol{\theta}^{(j-1)}$ through successive generation of values

$$\begin{aligned} \theta_1^{(j)} &\sim P(\theta_1 \mid \{\theta_i^{(j-1)} : i \neq 1\}) \\ \theta_2^{(j)} &\sim P(\theta_2 \mid \theta_1^{(j)}, \{\theta_i^{(j-1)} : i > 2\}) \\ &\dots \\ \theta_n^{(j)} &\sim P(\theta_n \mid \{\theta_i^{(j)} : i \neq n\}) \end{aligned} \tag{48}$$

In this way a random walk on the vector $\boldsymbol{\theta}$ is performed by making subsequent steps in low-dimensional subspaces, which span the full product space. This is similar to individual collisions of particles in a mechanical system that drives a many-body system to an equilibrium distribution for all degrees of freedom.

Gaussian distributions: Gibbs sampling

- Example Power spectrum and map sampling (Jewell et al 04; Wandelt et al 04, Eriksen et al 07, FSK & Ensslin 08; Jasche, FSK, Wandelt & Ensslin 10 link; Granett et al 15; Jasche & Lavaux 17)

$$\mathbf{s}^{j+1} \sim P(\mathbf{s}|\mathbf{S}^j, \mathbf{d}) \tag{49}$$

$$\mathbf{S}^{j+1} \sim P(\mathbf{S}|\mathbf{s}^{j+1}) \tag{50}$$

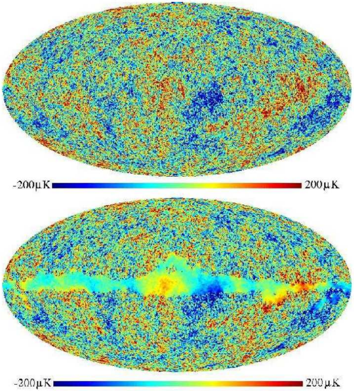
- Wiener filter

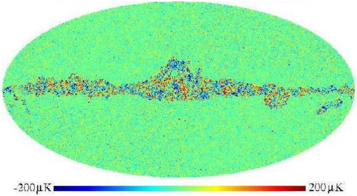
$$\mathbf{s}^j = \hat{\mathbf{s}}^j + \mathbf{y}^j \tag{51}$$

$\mathbf{y}^j = ((\mathbf{S}^j)^{-1} + \mathbf{R}^{+}\mathbf{N}^{-1}\mathbf{R})^{-1}((\mathbf{S}^j)^{-1/2}\mathbf{x}_1 + \mathbf{R}^{+}\mathbf{N}^{-1/2}\mathbf{x}_2)$ with Gaussian random variates $\mathbf{x}_1$ and $\mathbf{x}_2$

Another example is with the peculiar velocity field (see FSK & Ensslin 08; FSK, Ata, Angulo et al 16; and Ata, FSK et al 17).

Gaussian distributions: Gibbs sampling





Eriksen et al 07

Sampling the posterior

- Hamiltonian sampling (Duane et al 87; Taylor et al 08, Jasche & FSK 10 ▸ link)

$$H(\mathbf{s},\mathbf{p}) = K(\mathbf{p}) + E(\mathbf{s}), \tag{52}$$

- kinetic term with a given mass as the variance for the momenta

$$K(\mathbf{p}) = \frac{1}{2}\mathbf{p}^{+}\mathbf{M}^{-1}\mathbf{p}, \tag{53}$$

- Marginalization over the momenta

$$P(\mathbf{s},\mathbf{p}) = \frac{e^{-H}}{Z_H} = \frac{e^{-K}}{Z_K}\frac{e^{-E}}{Z_E} = P(\mathbf{p})P(\mathbf{s}), \tag{54}$$

- Please note, that the kinetic PDF is a Gaussian
- Marginalization occurs by drawing momenta from a Gaussian and throwing them away after each step

Sampling the posterior

- Hamiltonian evolution equations: $(\mathbf{s},\mathbf{p}) \rightarrow (\mathbf{s}',\mathbf{p}')$

$$\frac{\mathrm{d}\mathbf{p}}{\mathrm{d}t} = -\frac{\partial H}{\partial \mathbf{s}} = -\frac{\partial E}{\partial \mathbf{s}}, \tag{55}$$

$$\frac{\mathrm{d}\mathbf{s}}{\mathrm{d}t} = \frac{\partial H}{\partial \mathbf{p}} = \mathbf{M}^{-1}\mathbf{p}, \tag{56}$$

- Metropolis-Hastings acceptance step

$$p_a = \min(1, e^{-\delta H}), \tag{57}$$

$\delta H = H(\mathbf{s}',\mathbf{p}') - H(\mathbf{s},\mathbf{p}) \rightarrow$ we do not care about the evidence!

Expansions of the Bayesian models

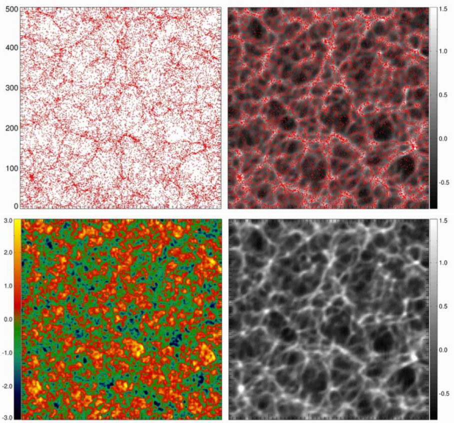
- Expand the prior
  When structures start to virialize the peculiar velocity field changes and shell crossing becomes important. One can relax the Gaussian assumption: Colombi 94: skewed lognormal model with the 1D Edgeworth exansion (based on the skewed Gaussian Edgeworth expansion: Juszkiewicz, Bouchet & Colombi 95; Bernardeau & Kofman 95; and the multivariate case: FSK 10 link). It is complicated and you need models for the higher order correlation functions.

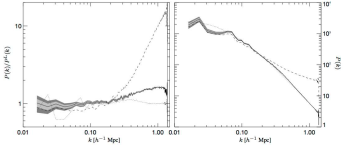
Expansions of the Bayesian models

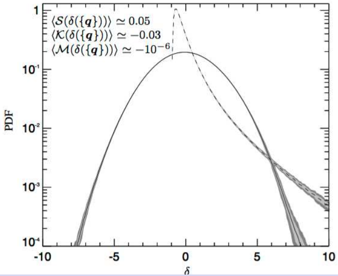
- Expand the likelihood.
  This requires that the Gaussian assumption applies for the density field at early cosmic times. This leads to forward Bayesian methods. Jasche et al 13, 15; FSK 12, 13; Wang et al 13, 14; . . . . One way is to explictely write the connection between the initial Gaussian density field and the final one with some kind of gravity model (2LPT: Jasche& Wandelt; LPT with corrections and PM: Wang et al 13,14).
  The other strategy is to introduce a new variable, the distribution of dark matter tracers at initial cosmic times and iteratively solve within a Gibbs-sampling procedure for the Gaussian initial field and the initial tracers which are compatible with the final ones given a structure formation model. Pioneering techniques to obtain a Gaussian field from a set of constraints: Bertschinger 87; Hoffman & Ribak 91; van de Weygaert & Bertschinger 96.

Sampling the initial conditions

Example of forward models on mock galaxy catalogs (FSK 13) ▸ link

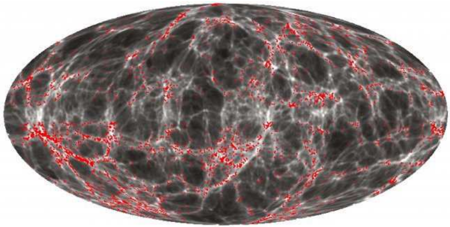


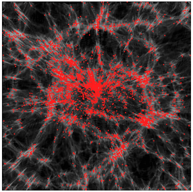


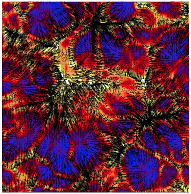


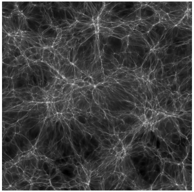
Sampling the initial conditions

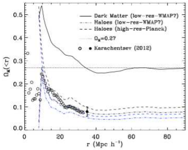
Example of forward models on data (FSK 12 ▸ link; Hess, FSK et al 13 ▸ link). Bottom left: redshift-space; middle: cosmic flows right: real-space (Abel, Kaehler, Hess, FSK phase space mapping 15 NatGeo). Is all this useful? Dipole FSK et al 12; Very right plot: are we living in a special place? Nuza, FSK et al 14 ▸ link. Reduce H_0 tensions Hess & FSK 16 ▸ link.











Data modelling

Effective galaxy bias is complex and has different components

- deterministic,
  - nonlinear,
  - nonlocal,
  - threshold bias: peak-background split (loss of information)
- stochastic (non-Poisson, shot noise).

Implementing part of this within a Bayesian framework: Ata, FSK & Mueller 15
▸ link

General bias modelling references

See some works below. Basics on bias, linear, nonlinear, peak-background split: Press Schechter 74; Peebles 80, 93; Kaiser 84; Peacock & heavens 85; Bardeen, Bond, Kaiser & Szalay 86; Cen & Ostriker 93; Fry & Gaztanaga 93; Lacey & Cole 93; Mo & White 96, 02; Sheth & Tormen 99; And list of works (not complete): *Stochastic bias*: Dekel & Lahav 99; Sheth & Lemson 99; Somerville et al 01. *Semi-analytic models*: Cole 00; Haton 03; Somerville & Primack 99; Cole et al. 00; Somerville et al 01b; Croton et al. 06; De Lucia & Blaizot 06; Cattaneo et al. 07; Somerville et al. 08; Bower 06; Baugh 06; Monaco 07; Guo et al 10, 16. White & Frenk 91; Kauffmann et al. 93; Cole et al. 94. *Halo model*: Seljak 00; Cooray & Sheth; *Halo occupation distribution*: Berlinde & Weinberg 02; Zhehavi et al 11; *Abundance matching*: Klypin et al 99; Kravtsov et al 04; Vale & Ostriker 04,06,08; Nagai & Kravtsov 05; Conroy & Wechsler 06; Behroozi 10; Trujillo-Gomez et al 11; Nuza et al 12; Rodriguez et al 16; *Perturbative expansions*: McDonald & Roy 09; Baldauf et al 12, ..., 16; Hearin et al 13; Saito et al 14, 15; Desjacques et al 16.

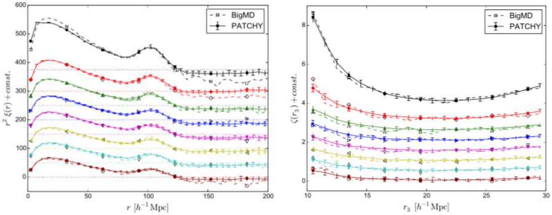
Some primers and reviews (not complete list): on nonlinear and nonlocal expansions of bias with perturbation theory Desjacques et al 16 link, on semi-analytic models Baugh 06 link, on the halo model Cooray & Sheth link, on the Halo occupation distribution (HOD), e.g. Zheng et al 05 link, on abundance matching Klypin et al 14 link.

Modelling effective bias

- What do we need to populate halos on a large scale dark matter field? nonlinear, stochastic and peak-background split (de La Torre & Peacock 12; FSK et al 14; Angulo et al 14; Neyrinck et al 14; Ahn et al 15; Chuang, FSK et al 15). How where the BOSS mocks constructed? FSK et al 16 link
- What information do we need to assign masses to a distribution of halos?

$$M_i \curvearrowleft P(M_i|r, \rho_{\mathrm{M}}, T, \Delta_r, \{p_{\mathrm{c}}\}, z) \qquad (58)$$

Zhao, FSK, Chuang et al 15 link We have to check up to 3-point statistics (and 4-point). We need nonlinear and nonlocal contributions! This is only partially solved for massive objects!



Determining effective bias parameters with MCMC

Vakili, FSK, Feng et al 17 ▸ link using PM and emcee ▸ link (very different results for different populations!)

$$\lambda_h \equiv \rho_h \equiv \langle N_{\rm h}\rangle_{{\rm d}V} = f_{\rm h}\, B(\rho_{\rm h}|\rho_{\rm m})\,, \tag{59}$$

$$f_{\rm h} = \frac{\rho_{\rm h}}{\langle B(\rho_{\rm h}|\rho_{\rm m})\rangle_V}\,, \tag{60}$$

$$\begin{aligned} B(\rho_{\rm h}|\rho_{\rm m}) &= \underbrace{\rho_{\rm m}^{\alpha}}_{\text{nonlinear bias}} \\ &\times \underbrace{\theta\left(\rho_{\rm m}-\rho_{\rm th}\right)}_{\text{threshold bias}} \times \underbrace{\exp\left(-(\rho_{\rm m}/\rho_{\epsilon})^{\epsilon}\right)}_{\text{exponential cutoff}}, \end{aligned} \tag{61}$$

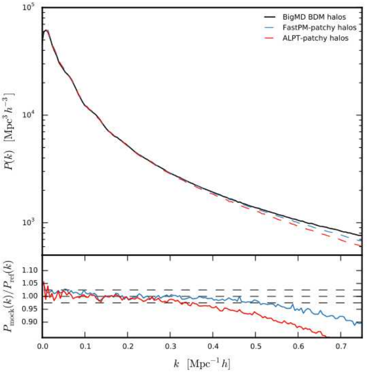
$$\begin{aligned} P(N_{\rm h}|\lambda_{\rm h},\beta) &= \underbrace{\frac{\lambda_{\rm h}^{N_{\rm h}}}{N_{\rm h}!}\, e^{-\lambda_{\rm h}}}_{\text{Poisson distribution}} \\ &\times \underbrace{\frac{\Gamma(\beta+N_{\rm h})}{\Gamma(\beta)(\beta+\lambda_{\rm h})^{N_{\rm h}}} \times \frac{e^{\lambda_{\rm h}}}{(1+\lambda_{\rm h}/\beta)^{\beta}}}_{\text{Deviation from Poissonity}} \end{aligned} \tag{62}$$

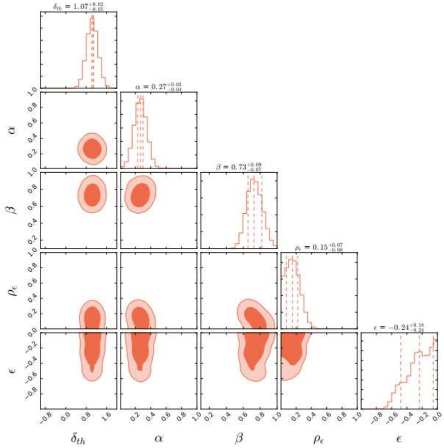
we need also the PDF to fit the 3-point statistics (FSK, Gil-Mar'in et al 15)

$$\begin{aligned} -2\ln p({\rm ref}|\theta) &= \sum_k \Bigg[\frac{\left(P_{\rm ref}(k)-P_{\rm mock}(k)\right)^2}{\sigma_k^2} \\ &+ \ln(2\pi\sigma_k^2)\Bigg] \\ &+ \sum_n \Bigg[\frac{\left(\rho_{\rm ref}(n)-\rho_{\rm mock}(n)\right)^2}{\sigma_n^2} \\ &+ \ln(2\pi\sigma_n^2)\Bigg] \end{aligned} \tag{63}$$

Determining effective bias parameters with MCMC

Vakili, FSK, Feng et al 17 ▸ link using PM and emcee ▸ link (percentage accuracy up to $k \sim 0.4$!)





Conclusions

- Statistical methods in cosmology can be very powerful.
- Active field with exciting developments.
- We need to improve the connection between the likelihoods and the priors, or equivalently between the data and the models. This is still very challenging!

Appendix

- Cosmic voids are dangerous (see Andreu Font's talk) Chuang, FSK (Font-Ribera) et al 17 ▸ link
- but useful, we showed that we can get the BAO from voids! FSK, Chuang et al 16 ▸ link; Yu et al 16 ▸ link; Zhao et al 16 ▸ link (including always Charling Tao)