Bayesian statistics

and some other aspects of probability theory

Florent Leclercq
www.florent-leclercq.eu
Imperial Centre for Inference and Cosmology
Imperial College London

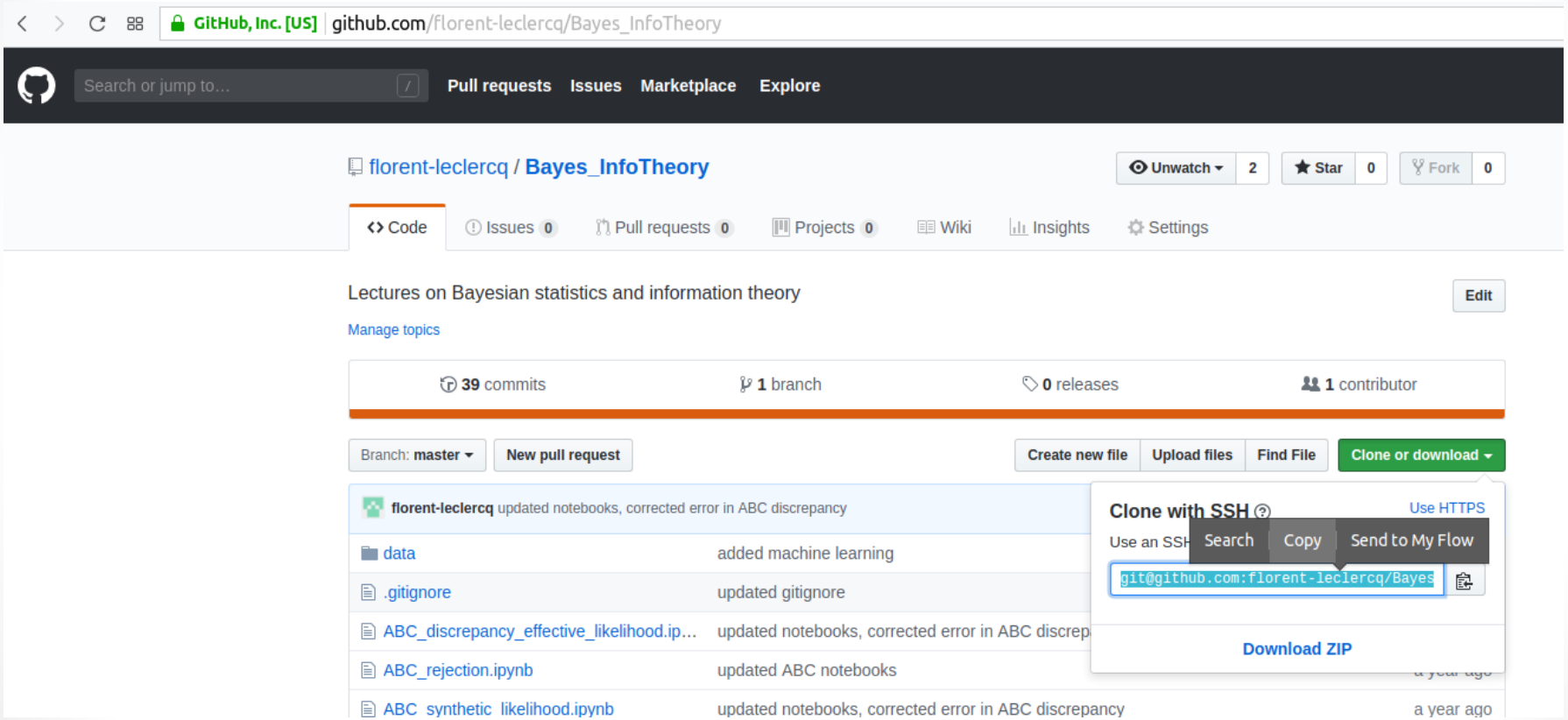
16 September 2021
STFC Summer School on Data Intensive Science

ICIC Imperial Centre for Inference & Cosmology

Imperial College London

The github repository

- https://github.com/florent-leclercq/Bayes_InfoTheory

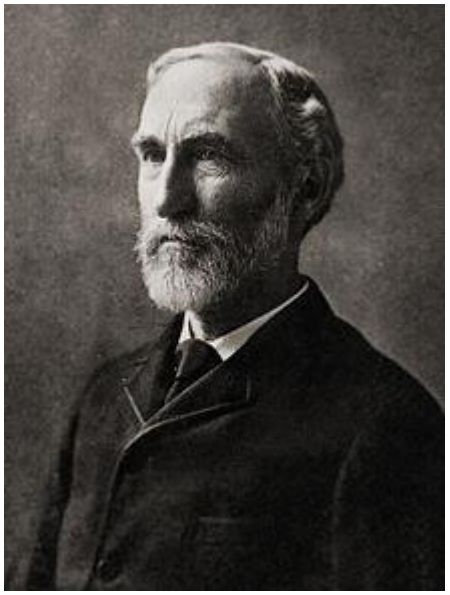


git clone https://github.com/florent-leclercq/Bayes_InfoTheory.git (or with SSH)

- Course website: http://florent-leclercq.eu/teaching.php (this lecture is actually part of a series of 3)

Introduction: why proper statistics matter

An historical example: the Gibbs paradox



J. Willard Gibbs (1839-1903)

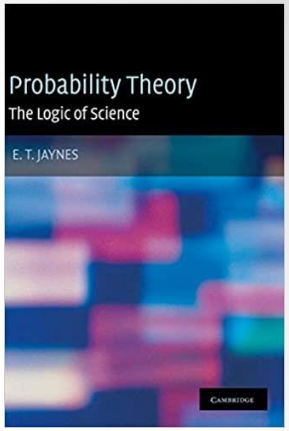
- Gibbs's canonical ensemble and grand canonical ensembles, derived from the maximum entropy principle, ***fail to correctly predict thermodynamic properties*** of real physical systems.
- The predicted entropies are always larger than the observed ones… there must exist ***additional microphysical constraints***:
  - Discreteness of energy levels: radiation: Planck (1900), solids: Einstein (1907), Debye (1912), Ising (1925), individual atoms : Bohr (1913)…
  - …Quantum mechanics: Heisenberg, Schrödinger (1927)

The first clues indicating the need for quantum physics were uncovered by seemingly "unsuccessful" application of statistics.

Outline

- Probability theory and Bayesian statistics: reminders
- Ignorance priors and the maximum entropy principle
- Gaussian random fields (and a digression on non-Gaussianity)
- Bayesian signal processing and reconstruction:
  - Bayesian de-noising
  - Bayesian de-blending
- Bayesian decision theory and Bayesian experimental design
- Bayesian networks, Bayesian hierarchical models and Empirical Bayes

Jaynes's "probability theory":
an extension of ordinary logic



Probability theory

Boolean Logic

?

Maximum Entropy

Empirical Bayes

Bayesian Inference

Decision theory & Experimental Design

Frequentist statistics

Reminders

- Product rule: $p(AB|C) = p(A|BC)\,p(B|C)$
- Sum rule: $p(A+B|C) = p(A|C) + p(B|C) - p(AB|C)$
- Bayes's formula:

posterior, likelihood, prior

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

evidence

- Bayesian model comparison: $\mathcal{B}_{12} = \frac{p(d|\mathcal{M}_1)}{p(d|\mathcal{M}_2)}$

Ignorance priors and the maximum entropy principle

Notebook 1: https://github.com/florent-leclercq/Bayes_InfoTheory/blob/master/LighthouseProblem.ipynb
Notebook 2: https://github.com/florent-leclercq/Bayes_InfoTheory/blob/master/MaximumEntropy.ipynb

Ignorance priors, functional equations and transformation groups

Ignorance priors: impose an invariant state of knowledge according to some transformation:

$$p(T(x))\mathrm{d}T(x) = p(x)\mathrm{d}x$$

- Simplest case: symmetry under the exchange of two models $\mathcal{M}_1$ and $\mathcal{M}_2$:

  $p(\mathcal{M}_1) = p(\mathcal{M}_2)$
  $p(\mathcal{M}_1) + p(\mathcal{M}_2) = 1$ ⇒ $p(\mathcal{M}_1) = p(\mathcal{M}_2) = \frac{1}{2}$ $\mathbb{Z}_2$-symmetry

- "Location parameter": $T(x) = x + s \quad \forall s$

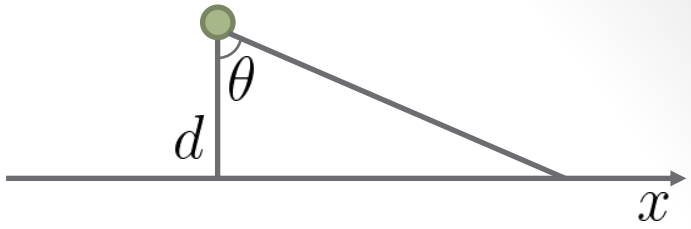
  $\mathrm{d}T = \mathrm{d}x$
  $p(x) = p(x+s) \quad \forall s$ ⇒ $p(x) = C$ Flat prior $U(1)$-symmetry

- "Scale parameter": $T(x) = ax \quad \forall a$

  $\mathrm{d}T = a\,\mathrm{d}x$
  $p(x) = a\,p(ax) \quad \forall a$ ⇒ $p(x) = C/a$ Jeffreys prior $U(1)$-symmetry

- General case: specify a **group of transformations** and solve the **functional equation**.

The lighthouse problem

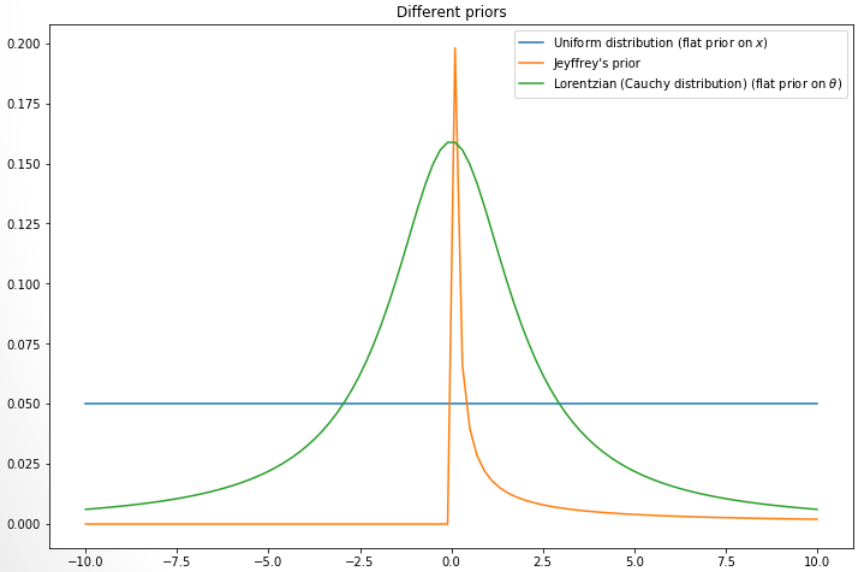


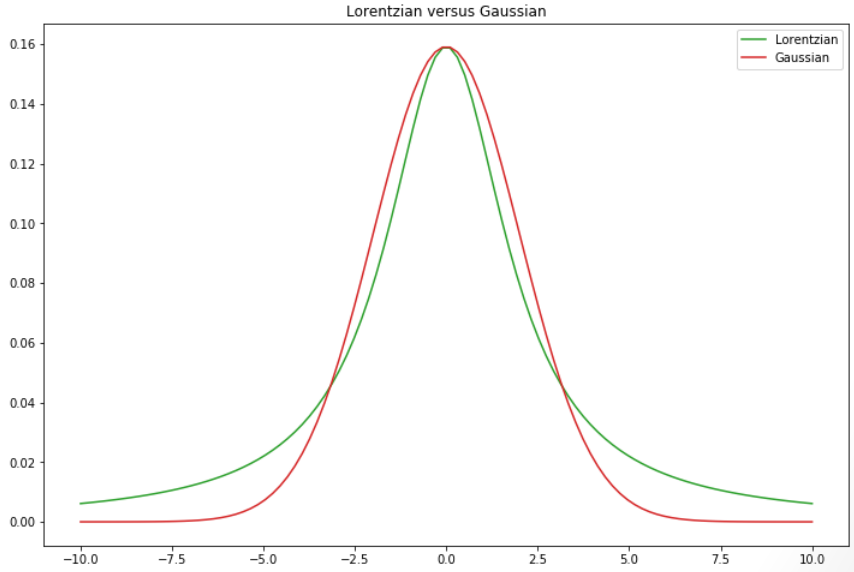
Notebook 1: https://github.com/florent-leclercq/Bayes_InfoTheory/blob/master/LighthouseProblem.ipynb

$$x = d\tan\theta$$
$$\mathrm{d}x = d(1+\tan^2\theta)\mathrm{d}\theta$$
$$p(x)\mathrm{d}x = p(\theta)\mathrm{d}\theta$$

If $p(\theta) = C$ ⇒ $p(x) \propto \frac{d}{d^2+x^2}$

Lorentzian/Cauchy distribution





Maximum ignorance for one variable is generally not the same thing as maximum ignorance for a non-linear function of that variable.

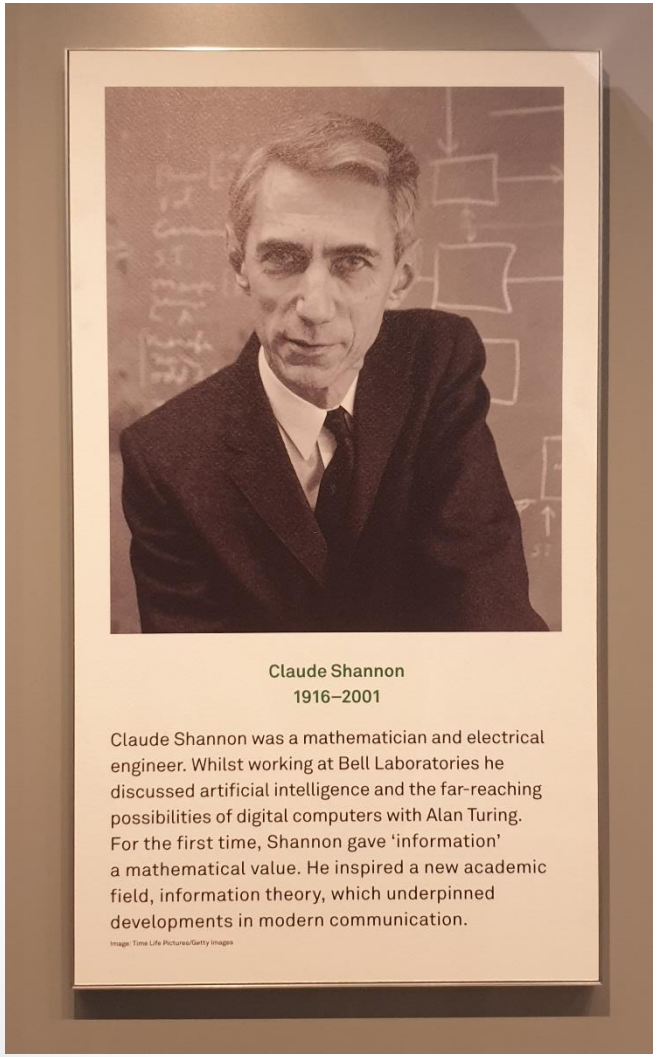
The maximum entropy principle

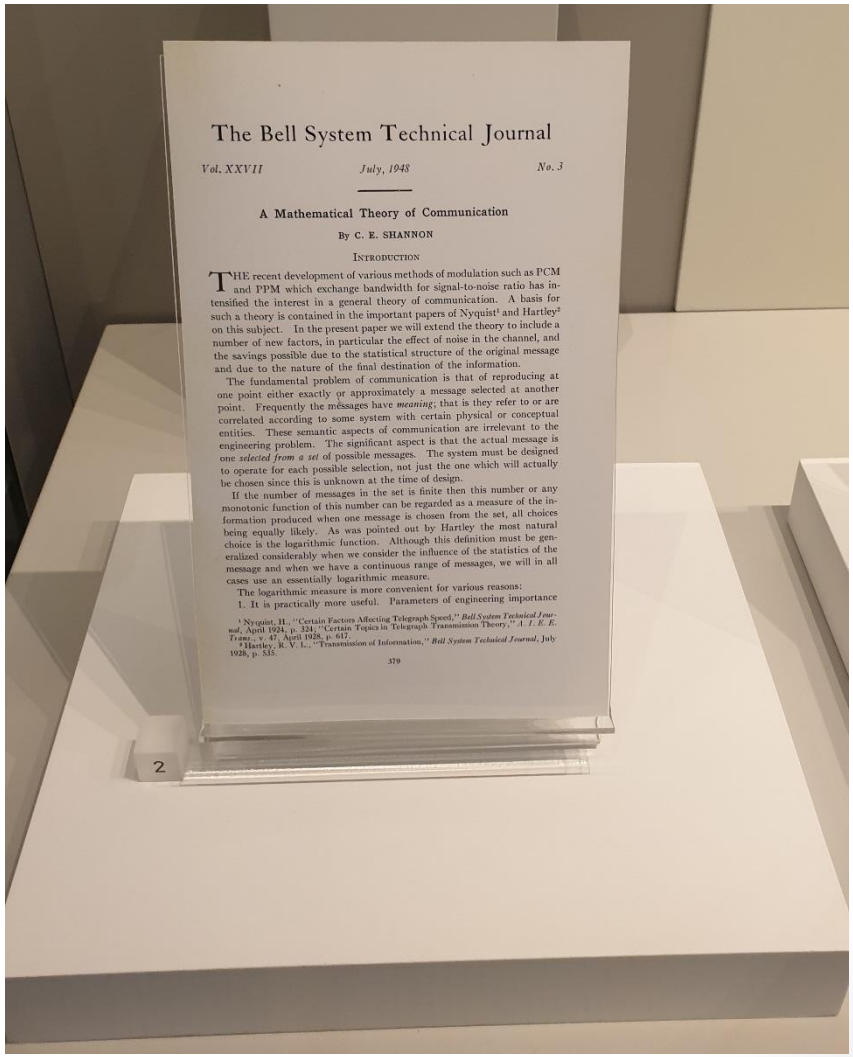
- Maximising the entropy = a general method to select priors while accounting for:
  - indifference about states of equal knowledge
  - relevant prior information
- What should $H[p]$ be for a source of information producing N finite "words" with probabilities p_n?
- Desiderata:
  - If all words are equiprobable ($\forall n,\ p_n = \frac{1}{N}$), $H[p]$ must grow with N
  - If words are generated in two steps (1- choosing a subset of words; 2- choosing a word in this subset), then the entropy is the sum of the entropy assigned to each step

⇨ Theorem (Shannon): $H[p] \propto -\sum_n p_n \log_2 p_n$

Information theory

- Pictures taken at the Science Museum in South Kensington…





The loaded dice

- For a fair dice, $p_n = \frac{1}{6} \quad \forall n \in [\![1, 6]\!]$: the principle of indifference was enough.
- Now let's say that the average value after many trials is not 3.5 but 4. What is the probability law in this case?
- We want to maximise $H[p]$ given two constraints:

$$< n >_p = \sum_{n=1}^{6} n\, p_n = 4 \qquad \text{and} \qquad \sum_{n=1}^{6} p_n = 1$$

- (1) brute force way:
  - get p_5 and p_6 as a function of p_1, p_2, p_3, p_4
  - express $H[p] = \sum_{n=1}^{6} p_n \ln p_n$ as a function of p_1, p_2, p_3, p_4
  - differentiate and solve $\frac{\partial H}{\partial p_n} = 0$ for $n \in [\![1, 4]\!]$

The loaded dice

- (2) a more elegant solution which does not break the symmetry: the method of **Lagrange multipliers**
- Lagrangian: $\mathcal{L}[\{p_n\}, \lambda, \mu] = -\sum_{n=1}^{6} p_n \ln p_n - \lambda \left(\sum_{n=1}^{6} n p_n - 4 \right) - \mu \left(\sum_{n=1}^{6} p_n - 1 \right)$
- Our two constraints are $\frac{\partial \mathcal{L}}{\partial \lambda} = 0$ and $\frac{\partial \mathcal{L}}{\partial \mu} = 0$.

  $\frac{\partial \mathcal{L}}{\partial p_n} = 0$ gives $-1 - \ln p_n - \lambda n - \mu = 0$

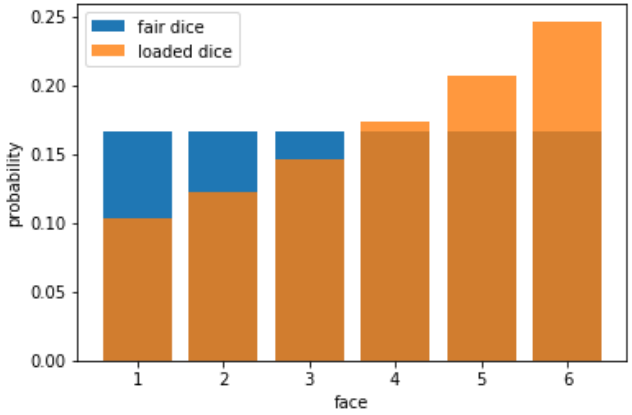
  $p_n = \frac{\mathrm{e}^{-\lambda n}}{Z}$ with $\ln Z \equiv 1 + \mu$
- The normalisation constraint fixes $Z = \sum_{n=1}^{6} \mathrm{e}^{-\lambda n} = \frac{1 - \mathrm{e}^{-6\lambda}}{\mathrm{e}^{\lambda} - 1}$
- The constraint on the mean is obtained by noting that

$$-\frac{\mathrm{d} \ln Z}{\mathrm{d}\lambda} = -\frac{1}{Z}\frac{\mathrm{d}Z}{\mathrm{d}\lambda} = \sum_{n=1}^{6} n \frac{\mathrm{e}^{-\lambda n}}{Z} = \sum_{n=1}^{6} n\, p_n = 4$$

This gives an equation for e^{λ} : $\mathrm{e}^{\lambda}/(\mathrm{e}^{\lambda} - 1) - 6/(\mathrm{e}^{6\lambda} - 1) = 4$

The loaded dice

Notebook 2: https://github.com/florent-leclercq/Bayes_InfoTheory/blob/master/MaximumEntropy.ipynb



- This is an example of probability theory beyond Bayesian statistics: we obtained a numerical probability assignment, conditional on some observations, without using Bayes' theorem.
- Thermodynamics analogy:
  - Fair dice = microcanonical ensemble: $p_n = \frac{1}{N}$
  - Loaded dice = canonical ensemble:

$$p_n = \frac{\mathrm{e}^{-\beta E_n}}{Z} \qquad \beta \equiv \frac{1}{k_\mathrm{B} T}$$

E_n = energy of different states

Z = partition function ≡ evidence in Bayesian statistics

Gaussian random fields

Notebook 3: https://github.com/florent-leclercq/Bayes_InfoTheory/blob/master/GRF_and_fNL.ipynb

Bayesian signal processing and reconstruction

Notebook 4: https://github.com/florent-leclercq/Bayes_InfoTheory/blob/master/WienerFilter_denoising.ipynb
Notebook 4bis: https://github.com/florent-leclercq/Bayes_InfoTheory/blob/master/WienerFilter_denoising_CMB.ipynb
Notebook 5: https://github.com/florent-leclercq/Bayes_InfoTheory/blob/master/WienerFilter_deblending.ipynb

Gaussian random fields

- Definition: any random vector x with pdf

$$p(x|\mu, C) = \frac{1}{\sqrt{|2\pi C|}} \exp\left[-\frac{1}{2}(x-\mu)^\intercal C^{-1}(x-\mu)\right]$$

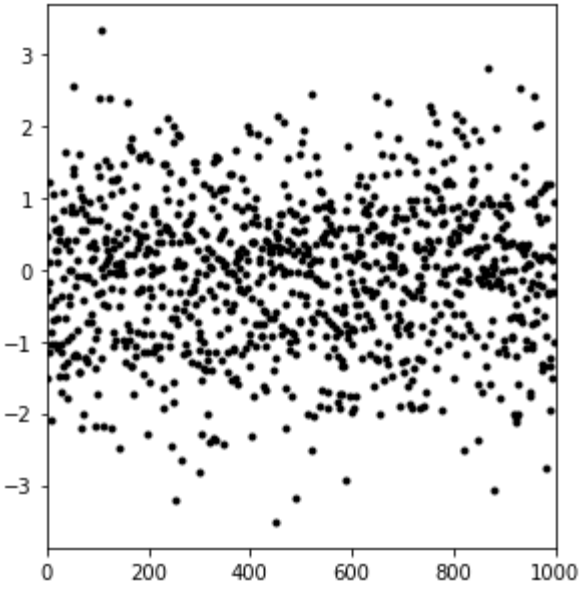
$$-2\ln p(x|\mu, C) = (x-\mu)^\intercal C^{-1}(x-\mu) + \ln|2\pi C| \qquad \text{... and that's it.}$$

- Generating a Gaussian random field:
  - Draw a white noise vector ξ (uncorrelated unit-Gaussian variables)
  - Find the matrix square-root of C: $\sqrt{C}$ (any such matrix works)
  - Compute $x = \sqrt{C}\xi + \mu$

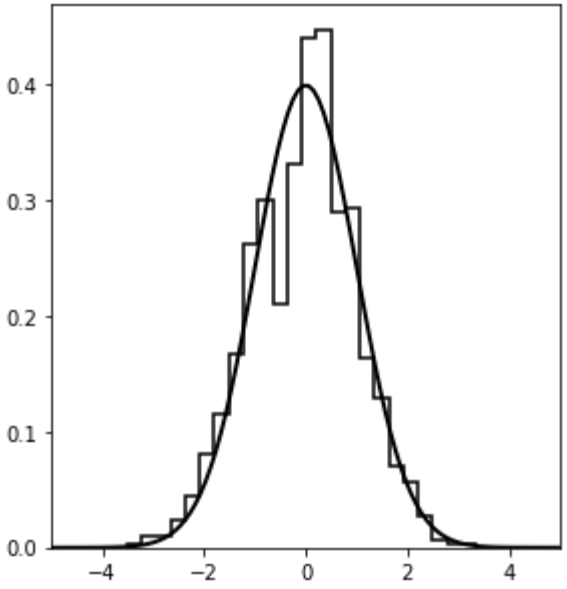
Gaussian random fields: examples

Notebook 3: https://github.com/florent-leclercq/Bayes_InfoTheory/blob/master/GRF_and_fNL.ipynb

white noise



histogram

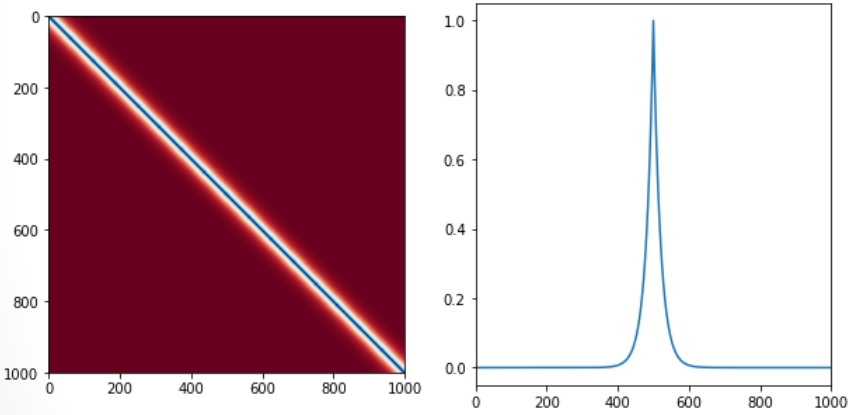


Gaussian random fields: examples

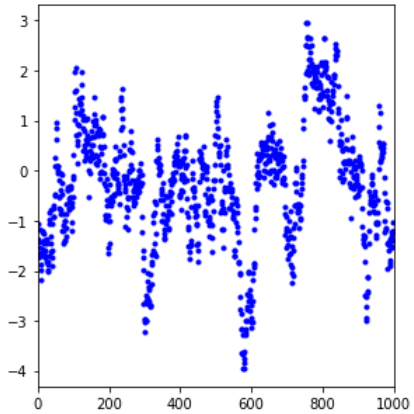
Notebook 3: https://github.com/florent-leclercq/Bayes_InfoTheory/blob/master/GRF_and_fNL.ipynb

$$C_{ij} = \exp\left(-\frac{|i-j|}{20}\right)$$

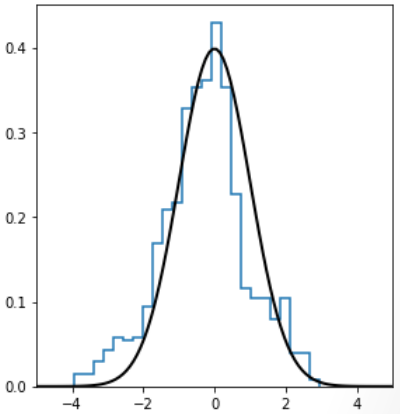
covariance matrix



GRF



histogram

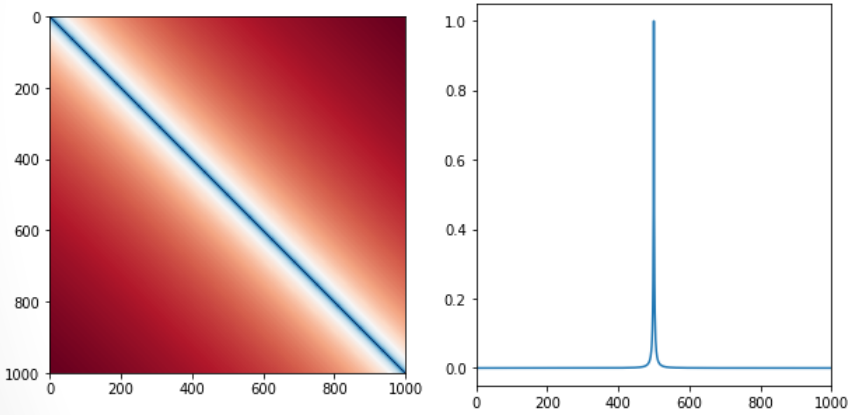


Gaussian random fields: examples

Notebook 3: https://github.com/florent-leclercq/Bayes_InfoTheory/blob/master/GRF_and_fNL.ipynb

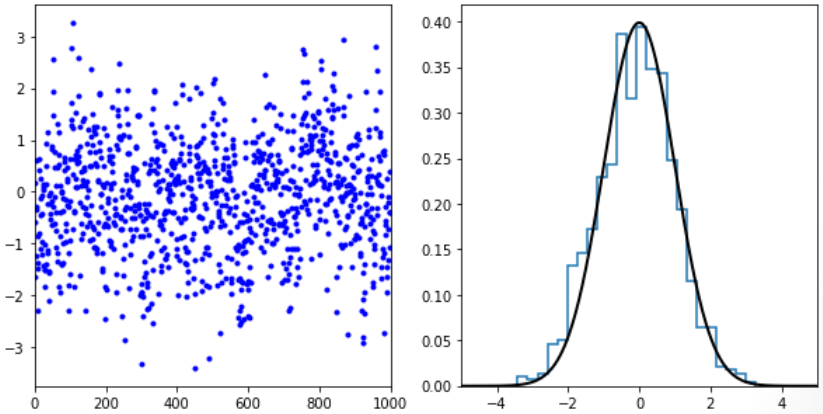
$$C_{ij} = \frac{1}{(1+|i-j|/2)^2}$$

covariance matrix



GRF

histogram

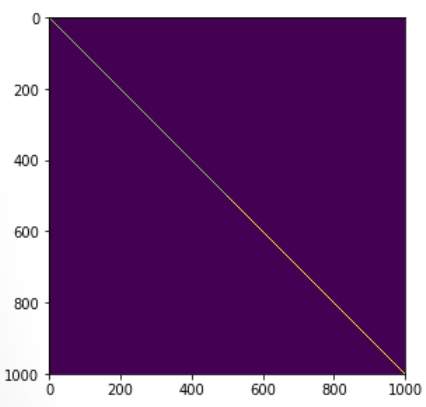


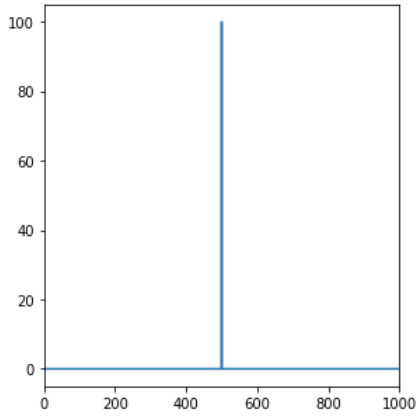
Gaussian random fields: examples

Notebook 3: https://github.com/florent-leclercq/Bayes_InfoTheory/blob/master/GRF_and_fNL.ipynb

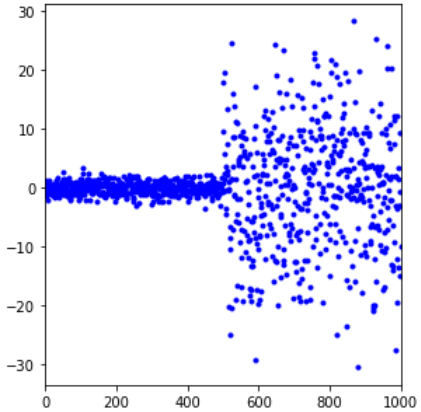
$$C_{ii} = \begin{cases} 1 & \text{if } i < N/2 \\ 100 & \text{otherwise} \end{cases}$$
$$C_{ij} = 0 \quad \text{for} \quad i \neq j$$

covariance matrix

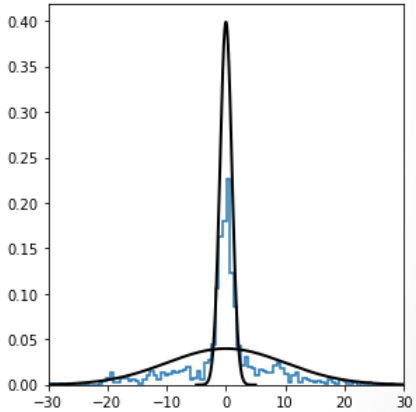




GRF



histogram

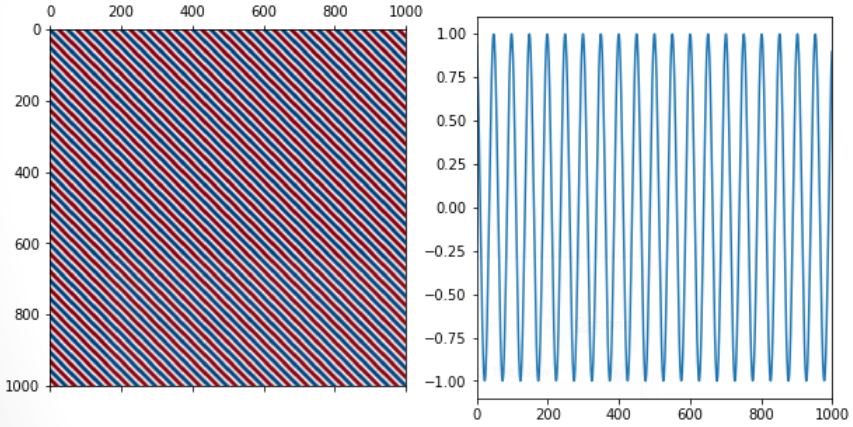


Gaussian random fields: examples

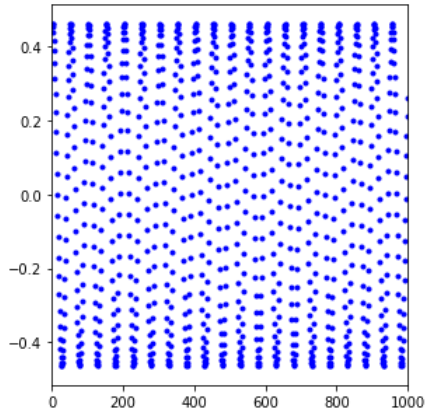
Notebook 3: https://github.com/florent-leclercq/Bayes_InfoTheory/blob/master/GRF_and_fNL.ipynb

$$C_{ij} = \cos\left(\frac{i-j}{8}\right)$$

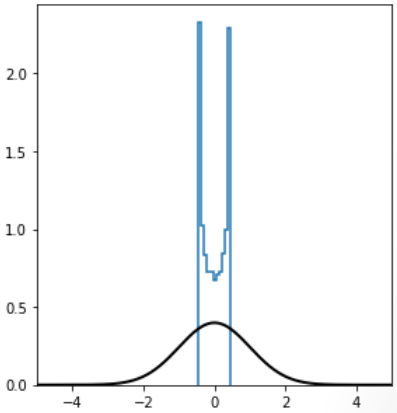
covariance matrix



GRF



histogram



- Histograms of Gaussian random fields are not always Gaussian!

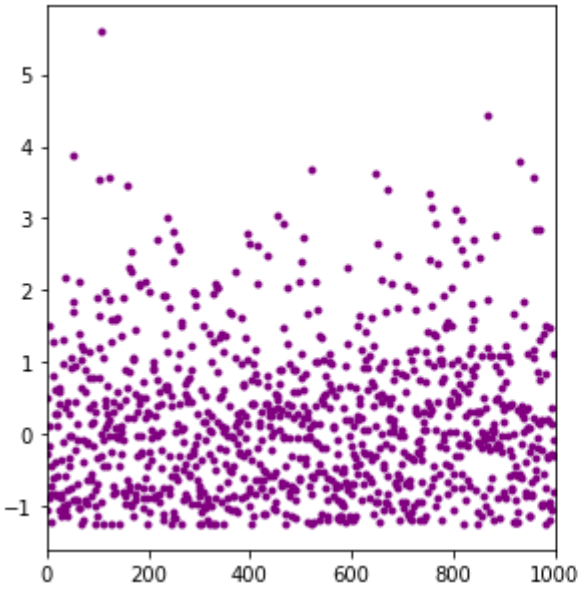
Example of a non-Gaussian signal

Notebook 3: https://github.com/florent-leclercq/Bayes_InfoTheory/blob/master/GRF_and_fNL.ipynb

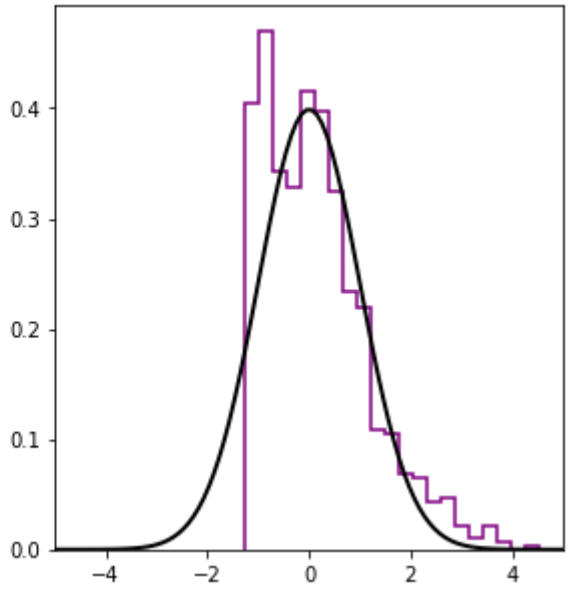
$$s = \Phi + f_{\rm NL}\Phi^2 \text{ where } \Phi \text{ is a GRF.}$$

In cosmology, this is called "local-type" non-Gaussianity

signal



histogram



- The one-point pdf is skewed.

Gaussian random fields: marginals and conditionals

- We work with a "joint" Gaussian random field $\begin{pmatrix} x \\ y \end{pmatrix}$
- Marginals:

Mean: $\mu = \begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix}$ Covariance: $C = \begin{pmatrix} C_{xx} & C_{xy} \\ C_{yx} & C_{yy} \end{pmatrix}$

The marginal means and covariances are just the corresponding parts of the joint mean and covariance.

- Conditionals:

Mean: $\mu_{x|y} = \mu_x + C_{xy} C_{yy}^{-1} (y - \mu_y)$

Covariance: $C_{x|y} = C_{xx} - C_{xy} C_{yy}^{-1} C_{yx}$

Wiener filtering: Bayesian de-noising

- Data model: $d = s + n$ where $\begin{pmatrix} s \\ d \end{pmatrix}$ is jointly Gaussian.
- Solution:

$$\mu_{s|d} = \mu_s + C_{sd} C_{dd}^{-1} (d - \mu_d)$$

$$C_{s|d} = C_{ss} - C_{sd} C_{dd}^{-1} C_{ds}$$

- Notations: $C_{ss} \equiv S$ and $C_{nn} \equiv N$.
- Assumption: $C_{sn} = C_{ns} = 0$. Then

$$C_{dd} = S + N$$

$$C_{sd} = C_{ss} + C_{sn} = C_{ss} = S$$

- Final expressions:

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

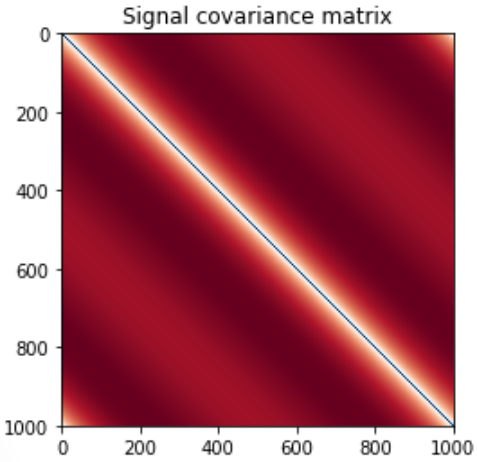
$$C_{s|d} = S - S(S+N)^{-1} S = (S^{-1} + N^{-1})^{-1}$$

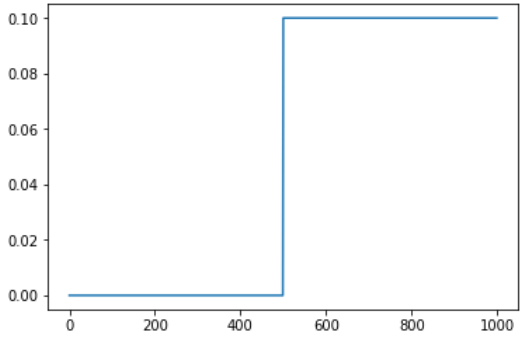
Wiener filtering: Bayesian de-noising

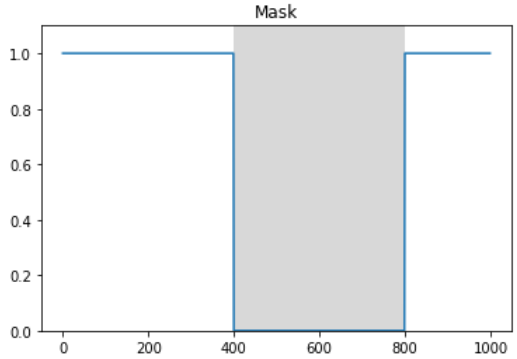
Notebook 4: https://github.com/florent-leclercq/Bayes_InfoTheory/blob/master/WienerFilter_denoising.ipynb

- Setup signal and noise covariance matrices

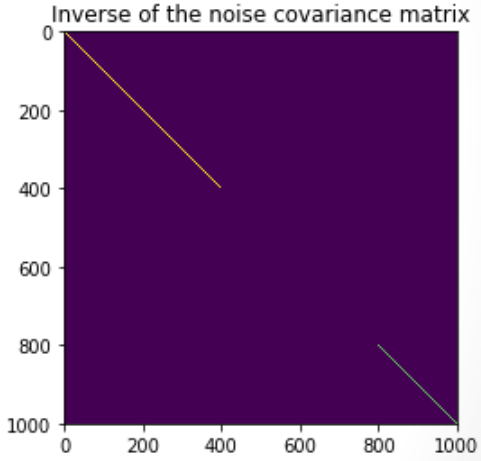
S







N^{-1}

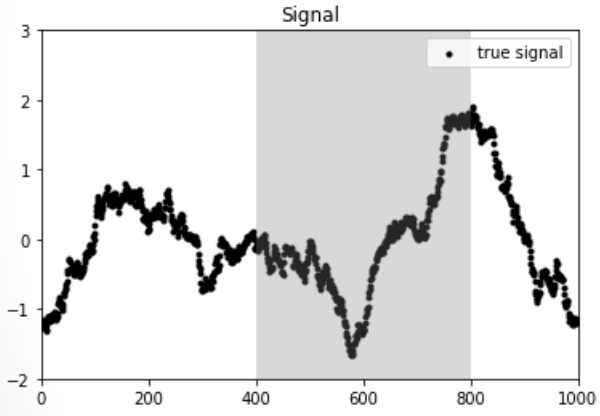


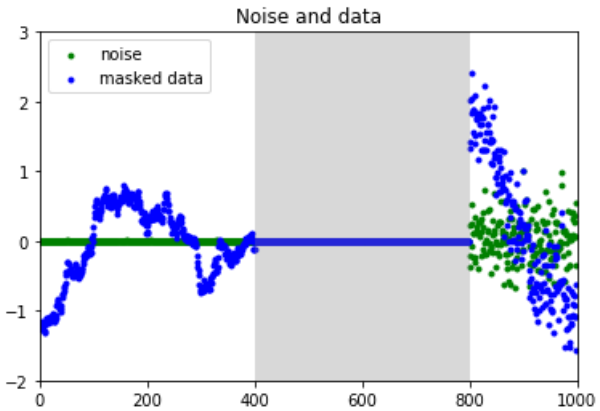
Wiener filtering: Bayesian de-noising

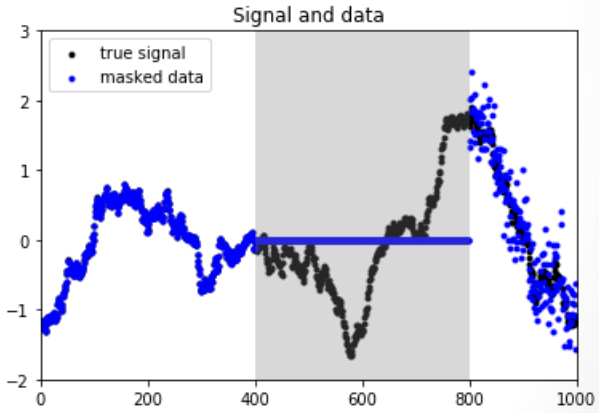
Notebook 4: https://github.com/florent-leclercq/Bayes_InfoTheory/blob/master/WienerFilter_denoising.ipynb

- Generate mock data

$$d = s + n$$





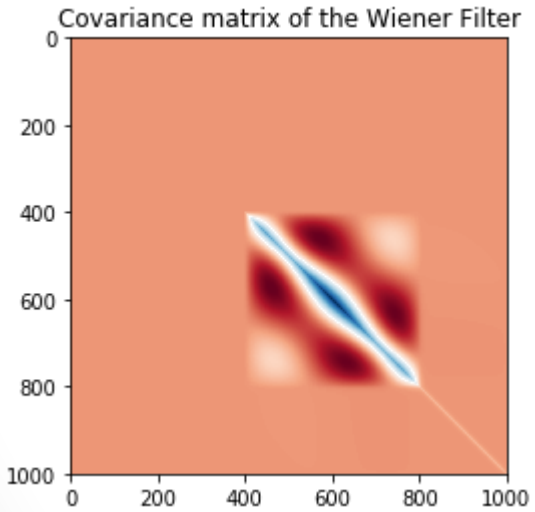


Wiener filtering: Bayesian de-noising

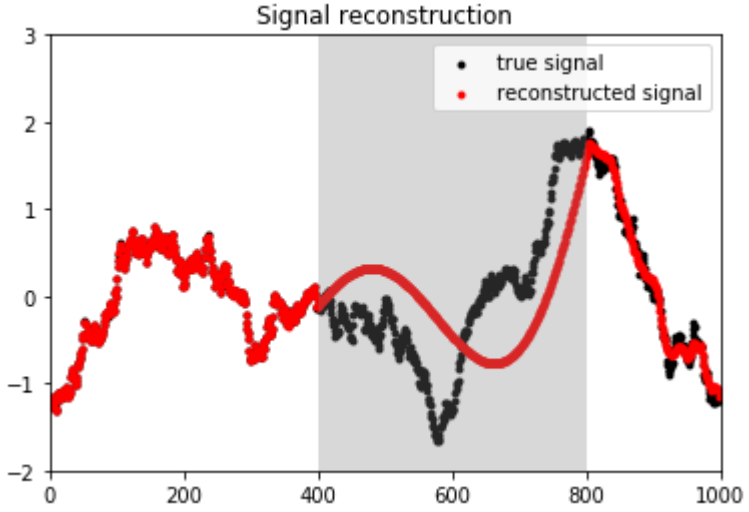
Notebook 4: https://github.com/florent-leclercq/Bayes_InfoTheory/blob/master/WienerFilter_denoising.ipynb

- Perform Wiener filtering

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



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

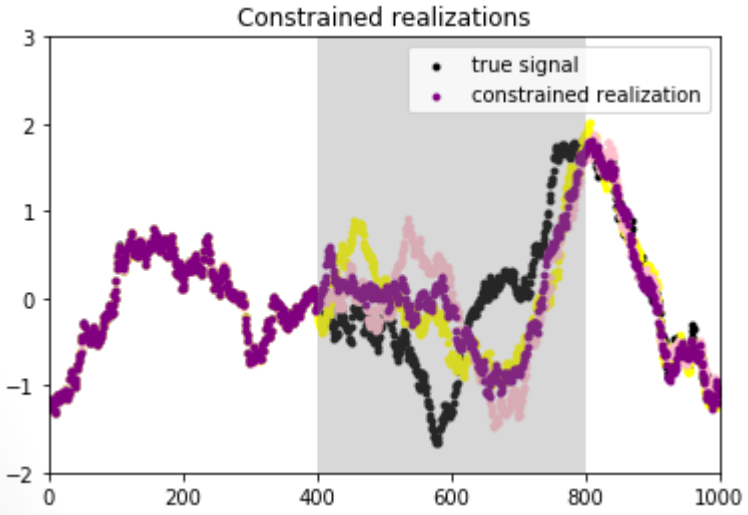


Wiener filtering: Bayesian de-noising

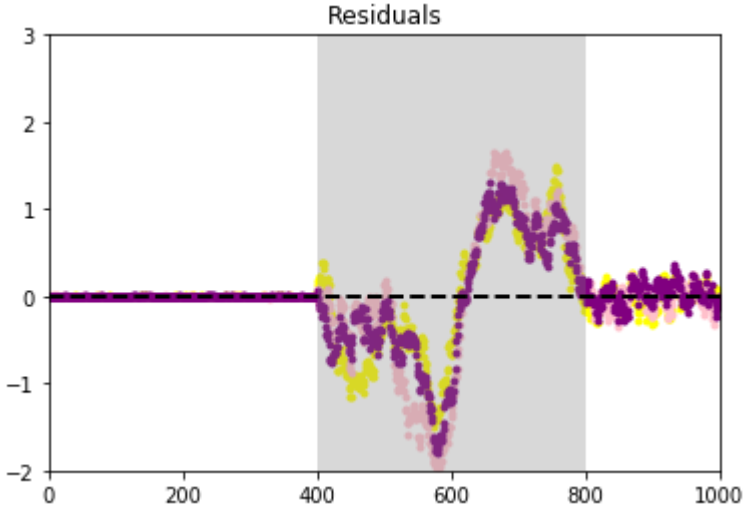
Notebook 4: https://github.com/florent-leclercq/Bayes_InfoTheory/blob/master/WienerFilter_denoising.ipynb

- Draw constrained realisations

$$s_{\mathrm{sim}} = \mu_{s|d} + \sqrt{C_{s|d}}\,\xi$$



$$s_{\mathrm{sim}} - s_{\mathrm{true}}$$



Wiener filtering: Bayesian de-blending

Notebook 5: https://github.com/florent-leclercq/Bayes_InfoTheory/blob/master/WienerFilter_deblending.ipynb

- Data model: $d = \begin{pmatrix} x_1 \\ x_2 \\ 0 \end{pmatrix} + \begin{pmatrix} 0 \\ x_1 \\ x_2 \end{pmatrix} + \begin{pmatrix} n_1 \\ n_2 \\ n_3 \end{pmatrix}$
- Assumptions: $C_{x_1 d} = \begin{pmatrix} C_{x_1 x_1} & C_{x_1 x_1} & 0 \end{pmatrix}$ $C_{x_2 d} = \begin{pmatrix} 0 & C_{x_2 x_2} & C_{x_2 x_2} \end{pmatrix}$ $C_{nn} = \begin{pmatrix} C_{n_1 n_1} & 0 & 0 \\ 0 & C_{n_2 n_2} & 0 \\ 0 & 0 & C_{n_3 n_3} \end{pmatrix}$

$$C_{dd} = \begin{pmatrix} C_{x_1 x_1} + C_{n_1 n_1} & C_{x_1 x_1} & 0 \\ C_{x_1 x_1} & C_{x_1 x_1} + C_{x_2 x_2} + C_{n_2 n_2} & C_{x_2 x_2} \\ 0 & C_{x_2 x_2} & C_{x_2 x_2} + C_{n_3 n_3} \end{pmatrix}$$

- Solution:

$$\mu_{x_1|d} = C_{x_1 d} C_{dd}^{-1} d \qquad \mu_{x_2|d} = C_{x_2 d} C_{dd}^{-1} d$$
$$C_{x_1|d} = C_{x_1 x_1} - C_{x_1 d} C_{dd}^{-1} C_{d x_1} \qquad C_{x_2|d} = C_{x_2 x_2} - C_{x_2 d} C_{dd}^{-1} C_{d x_2}$$

Bayesian decision theory

Notebook 6: https://github.com/florent-leclercq/Bayes_InfoTheory/blob/master/DecisionTheory.ipynb

Bayesian experimental design

Bayesian decision theory

- Bayesian decision theory is **optimal decision making**, given a set of possible actions and uncertain beliefs, encoded in some pdf $p(\theta|I)$ (usually the posterior of Bayesian inference)

- Notations: $\{\theta\}$= set of features (observable variables)
  $\{a\}$= set of actions

- **Expected utility hypothesis**: given a set of gain functions $G(a|\theta)$, the optimal decision rule is to take the action that maximises the expected utility $U(a|I)$, defined by

$$U(a|I) \equiv <G(a|\theta)>_{p(\theta|I)} = \int G(a|\theta)p(\theta|I)\,\mathrm{d}\theta$$

⇨ Take action $a^* = \mathrm{argmax}_a\, U(a|I)$

Example: Bayesian alerts

- We look for an event E. We have access to

$$p(E|I) \qquad p(\bar{E}|I) = 1 - p(E|I)$$

- There are two possible actions:
  - a_1 = raise the alert
  - a_2 = do nothing
- Utilities:

$$U(a_1|I) = G(a_1|E)p(E|I) + G(a_1|\bar{E})\left[1 - p(\bar{E}|I)\right]$$
$$U(a_2|I) = G(a_2|E)p(E|I) + G(a_2|\bar{E})\left[1 - p(\bar{E}|I)\right]$$

correct detection (a "hit")

false positive (a "false alarm")

false negative (a "miss")

correct rejection

- A typical choice of gain functions:

$$G(a_1|E) = G - C \qquad G(a_1|\bar{E}) = -C$$
$$G(a_2|E) = 0 \qquad G(a_2|\bar{E}) = 0$$

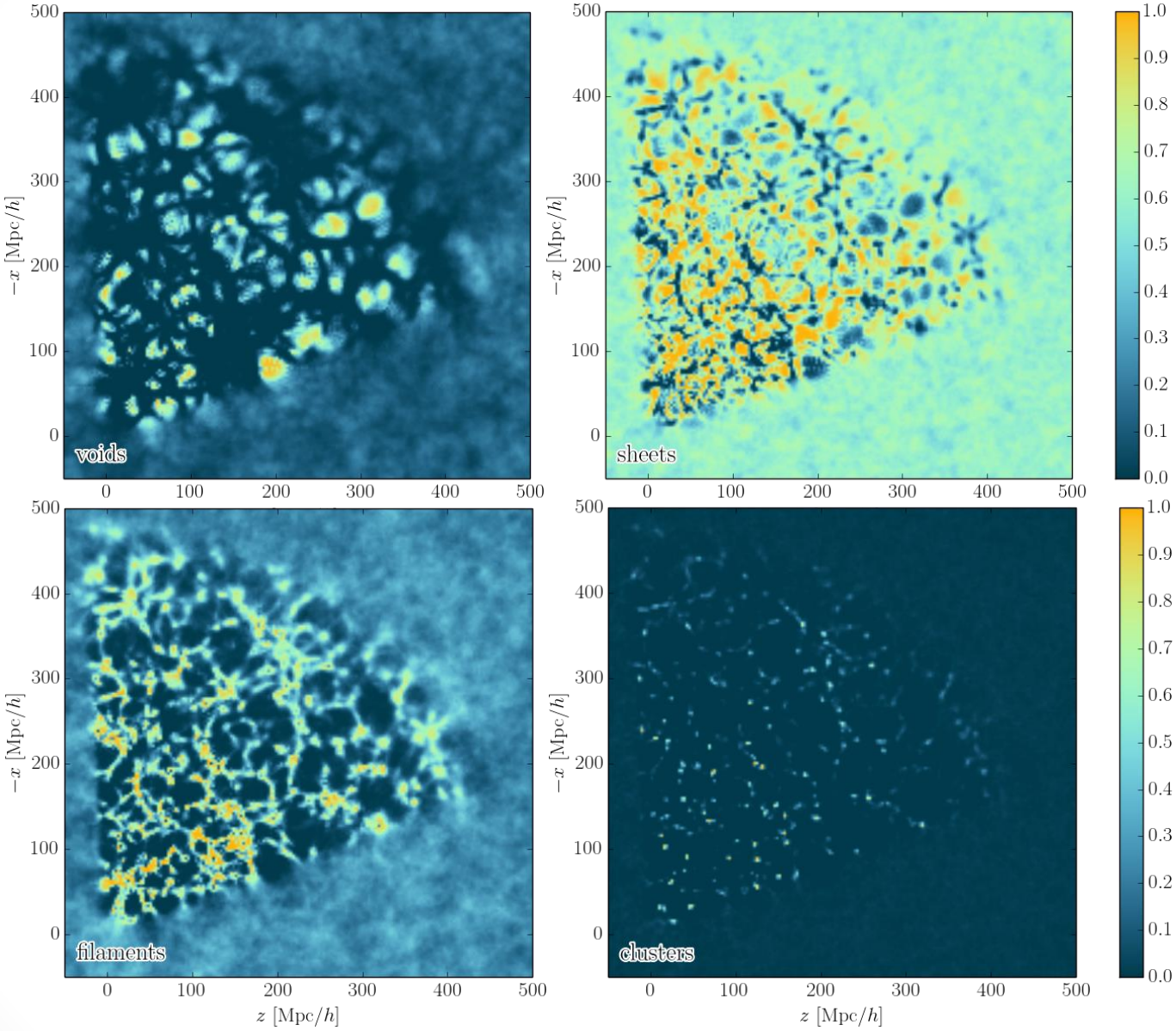
expected gain for a detection

cost of raising the alert

- So $U(a_1|I) = p(E|I)(G - C) + [1 - p(E|I)](-C)$

$U(a_2|I) = 0$ ⇒ raise the alert if and only if $p(E|I) \geq \frac{G}{C}$

Structures in the cosmic web



FL, Jasche & Wandelt 2015b, arXiv:1503.00730

A decision rule for structure classification

- Space of "input features":

$$\{\mathrm{T}_0 = \text{void}, \mathrm{T}_1 = \text{sheet}, \mathrm{T}_2 = \text{filament}, \mathrm{T}_3 = \text{cluster}\}$$

- Space of "actions":

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

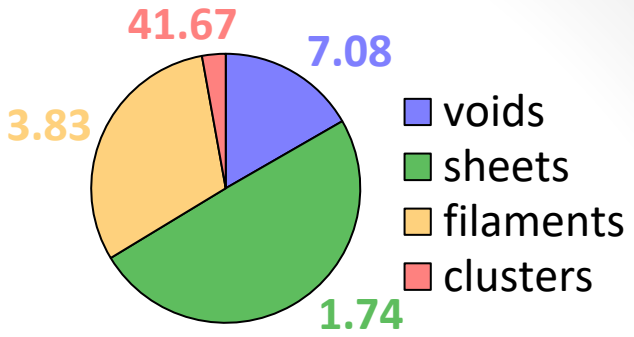
⇨ A problem of **Bayesian decision theory**:
one should take the action that maximizes the utility

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

- How to write down the gain functions?

FL, Jasche & Wandelt 2015b, arXiv:1503.00730

Gambling with the Universe



- One proposal:

$$G(a_j|\mathrm{T}_i) = \begin{cases} \dfrac{1}{\mathcal{P}(\mathrm{T}_i)} - \alpha & \text{if } j \in [\![0,3]\!] \text{ and } i = j \quad \text{“Winning”} \\ -\alpha & \text{if } j \in [\![0,3]\!] \text{ and } i \neq j \quad \text{“Losing”} \\ 0 & \text{if } j = -1. \quad \text{“Not playing”} \end{cases}$$

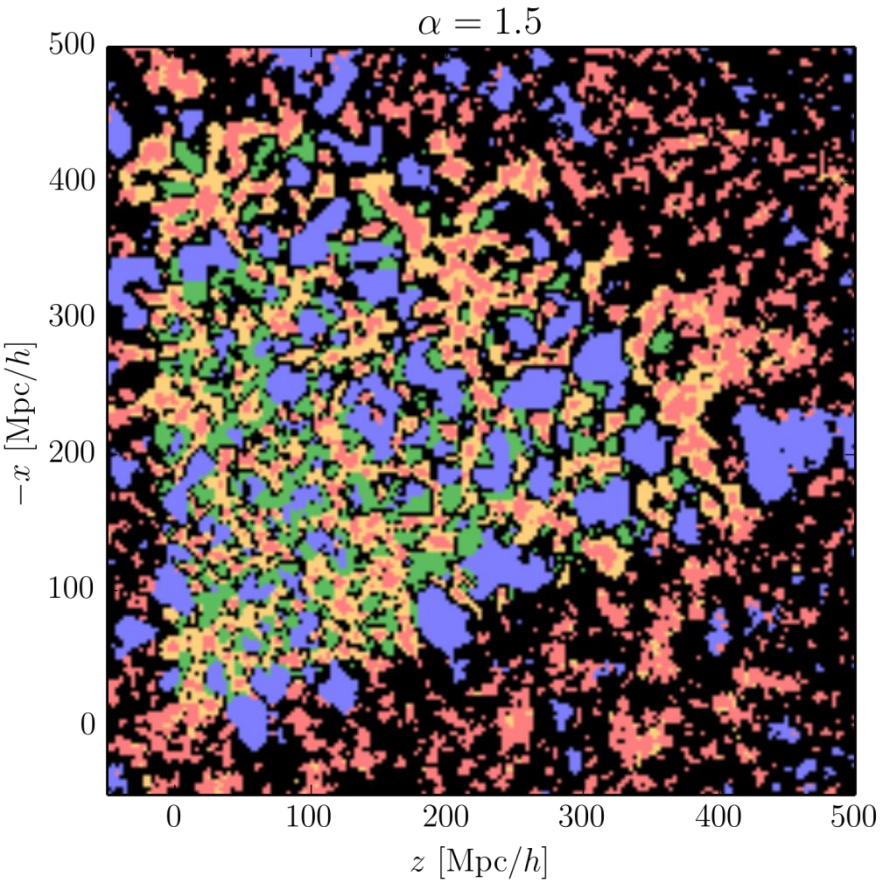
- Without data, the expected utility is

$$U(a_j) = 1 - \alpha \quad \text{if } j \neq 1 \qquad \text{“Playing the game”}$$
$$U(a_{-1}) = 0 \qquad \text{“Not playing the game”}$$

- With $\alpha = 1$, it's a *fair game* ⇨ always play ⇨ “**speculative map**” of the LSS
- Values $\alpha > 1$ represent an *aversion for risk* ⇨ increasingly “**conservative maps**” of the LSS

FL, Jasche & Wandelt 2015b, arXiv:1503.00730

Playing the game…



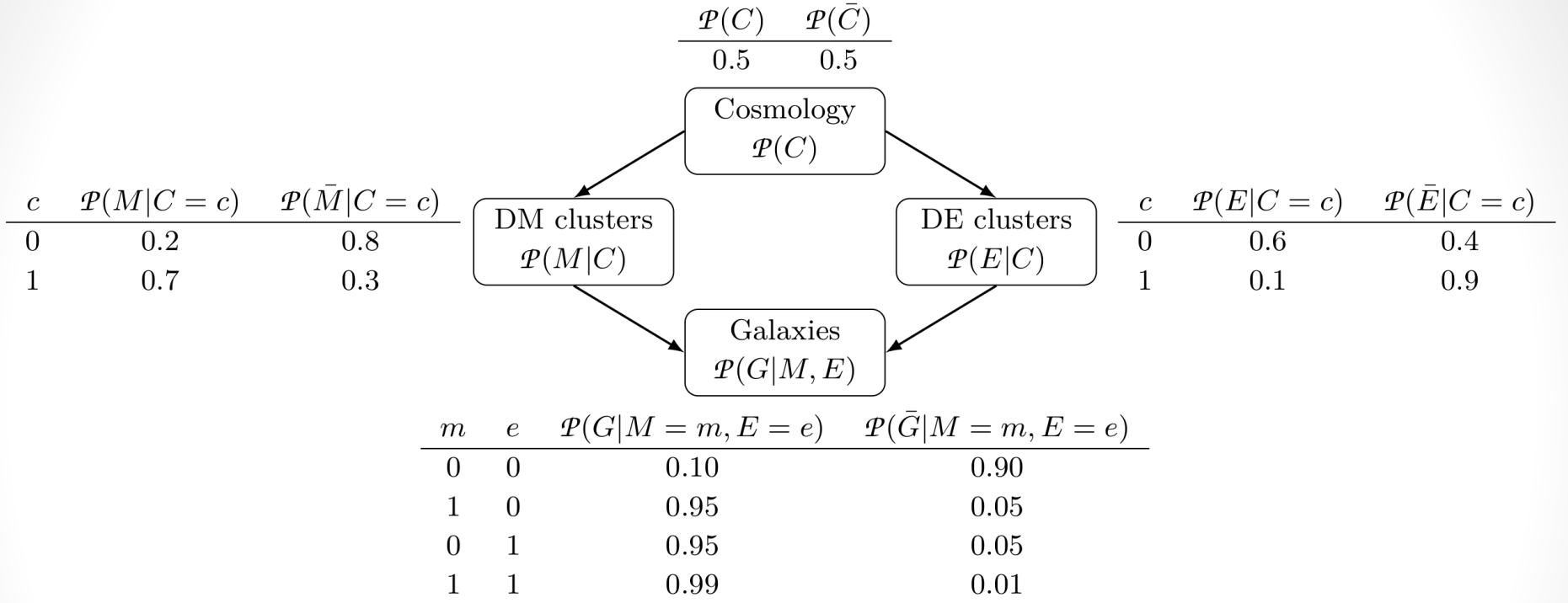
FL, Jasche & Wandelt 2015b, arXiv:1503.00730

Bayesian networks

Bayesian hierarchical models

and Empirical Bayes

Bayesian networks



c	$\mathcal{P}(M\|C=c)$	$\mathcal{P}(\bar{M}\|C=c)$
0	0.2	0.8
1	0.7	0.3

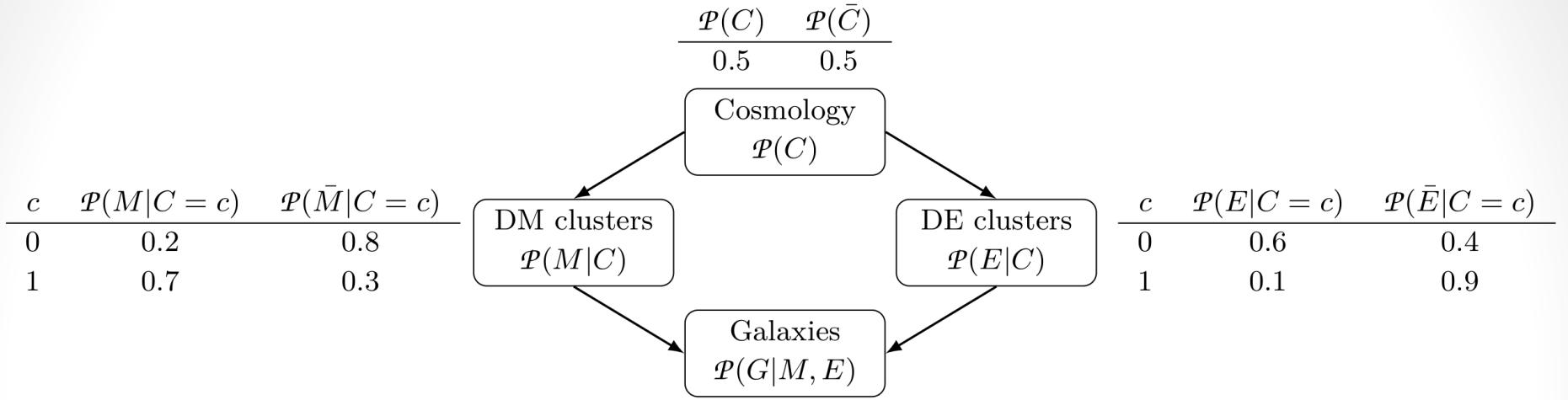
c	$\mathcal{P}(E\|C=c)$	$\mathcal{P}(\bar{E}\|C=c)$
0	0.6	0.4
1	0.1	0.9

m	e	$\mathcal{P}(G\|M=m,E=e)$	$\mathcal{P}(\bar{G}\|M=m,E=e)$
0	0	0.10	0.90
1	0	0.95	0.05
0	1	0.95	0.05
1	1	0.99	0.01

Bayesian networks are probabilistic graphical models consisting of:

- A **directed acyclic graph** (DAG)
- At each node, **conditional probabilities distributions**

Bayesian networks



c	$\mathcal{P}(M\|C=c)$	$\mathcal{P}(\bar{M}\|C=c)$
0	0.2	0.8
1	0.7	0.3

c	$\mathcal{P}(E\|C=c)$	$\mathcal{P}(\bar{E}\|C=c)$
0	0.6	0.4
1	0.1	0.9

m	e	$\mathcal{P}(G\|M=m,E=e)$	$\mathcal{P}(\bar{G}\|M=m,E=e)$
0	0	0.10	0.90
1	0	0.95	0.05
0	1	0.95	0.05
1	1	0.99	0.01

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

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

Bayesian networks
inference and prediction

- Inference:

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

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

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

- Prediction:

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

Bayesian networks

the "explaining away" phenomenon

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

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

- So we have both:

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

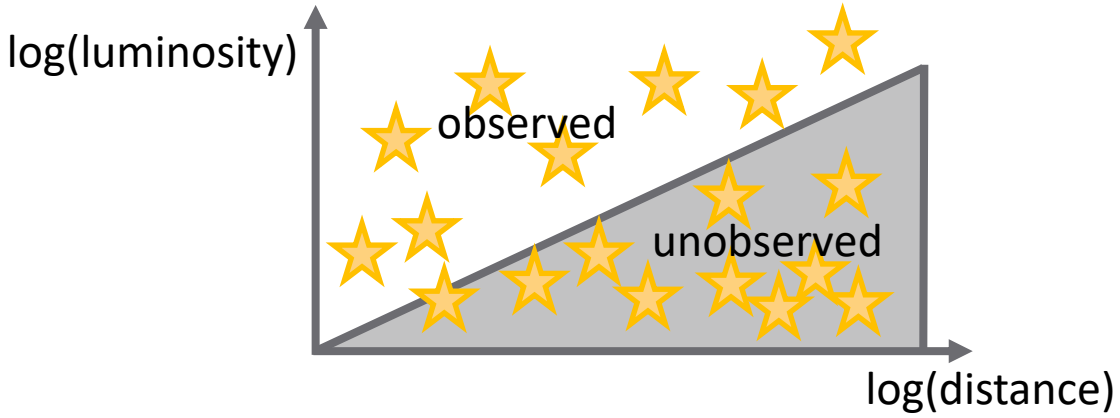
- This is "**collider bias**" or the "**explaining away**" phenomenon: two causes collide to explain the same effect.
- Particular case: "**selection bias**" or "**Berkson's paradox**"

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

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

Malmquist bias

- Malmquist (1925) bias: in magnitude-limited surveys, far objects are preferentially detected if they are intrinsically bright.



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

$$C = A + B \quad \Rightarrow \quad p(A|\bar{B}, C) = 1 > p(A|C)$$

detected bright close

Bayesian hierarchical models

- Simple inference:

prior

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

- Adaptive prior:

prior hyperprior

$$p(\theta|d) \propto p(d|\theta)\, p(\theta|\eta)\, p(\eta)$$

- … or a full hierarchy of hyperpriors.

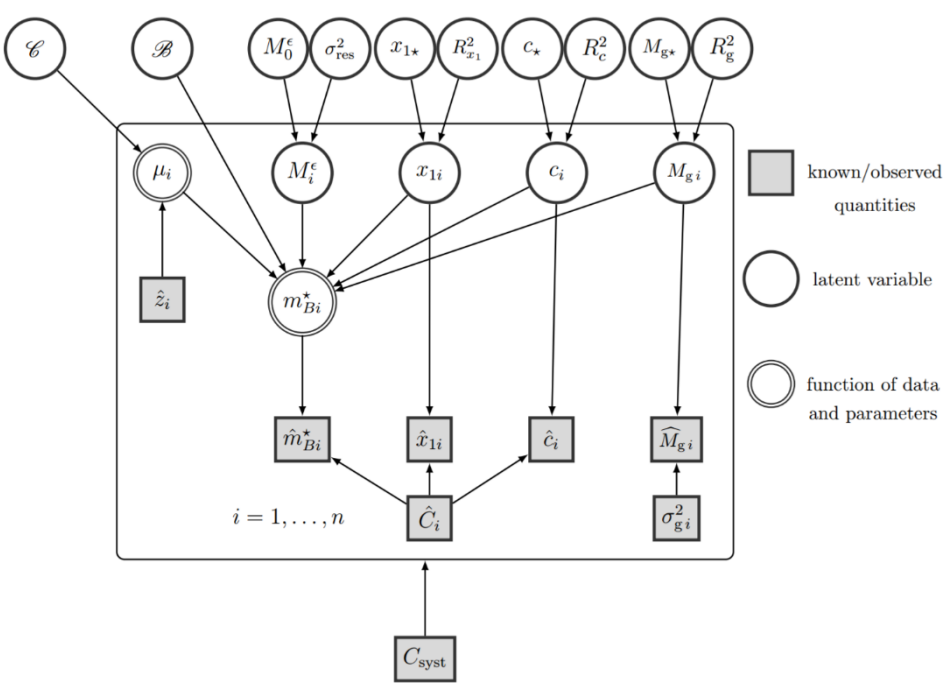
- Examples:
  - Cosmic microwave background:

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

  - Large-scale structure:

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

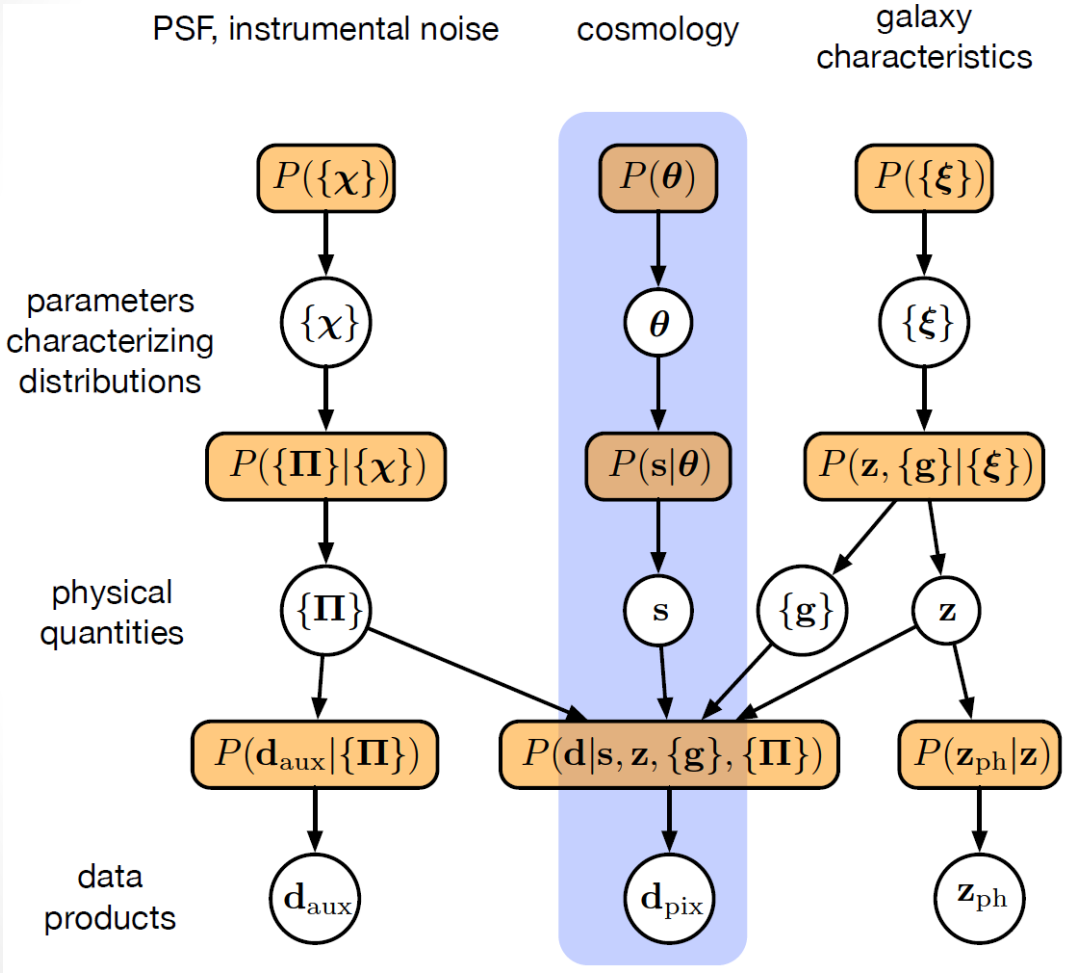
BHM example: supernovae (BAHAMAS)



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

Shariff et al. 2015, arXiv:1510.05954

BHM example: weak lensing



Can include:

Mask
Intrinsic alignments
Baryon feedback
Shape measurement
Photometric redshifts

Alsing et al. 2015, arXiv:1505.07840

Empirical Bayes

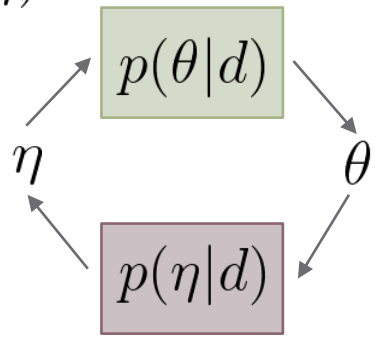
an alternative to maximum entropy for choosing priors

prior hyperprior

$$p(\theta|d) \propto p(d|\theta)\, p(\theta|\eta)\, p(\eta)$$

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

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



⇨ Iterative scheme ("Gibbs" sampler)

- **Empirical Bayes** is a truncation of this scheme after a few steps (often just one).
- Particular case: $p(\eta|d) \approx \delta_{\mathrm{D}}(\eta - \eta^\star(d))$ ⇨ $p(\theta|d) \approx \dfrac{p(d|\theta)\, p(\theta|\eta^\star)}{p(d|\eta^\star)}$

  ⇨ the **Expectation-Maximization** (EM) algorithm (machine learning, data mining).