Lecture 1: Probability theory and signal processing

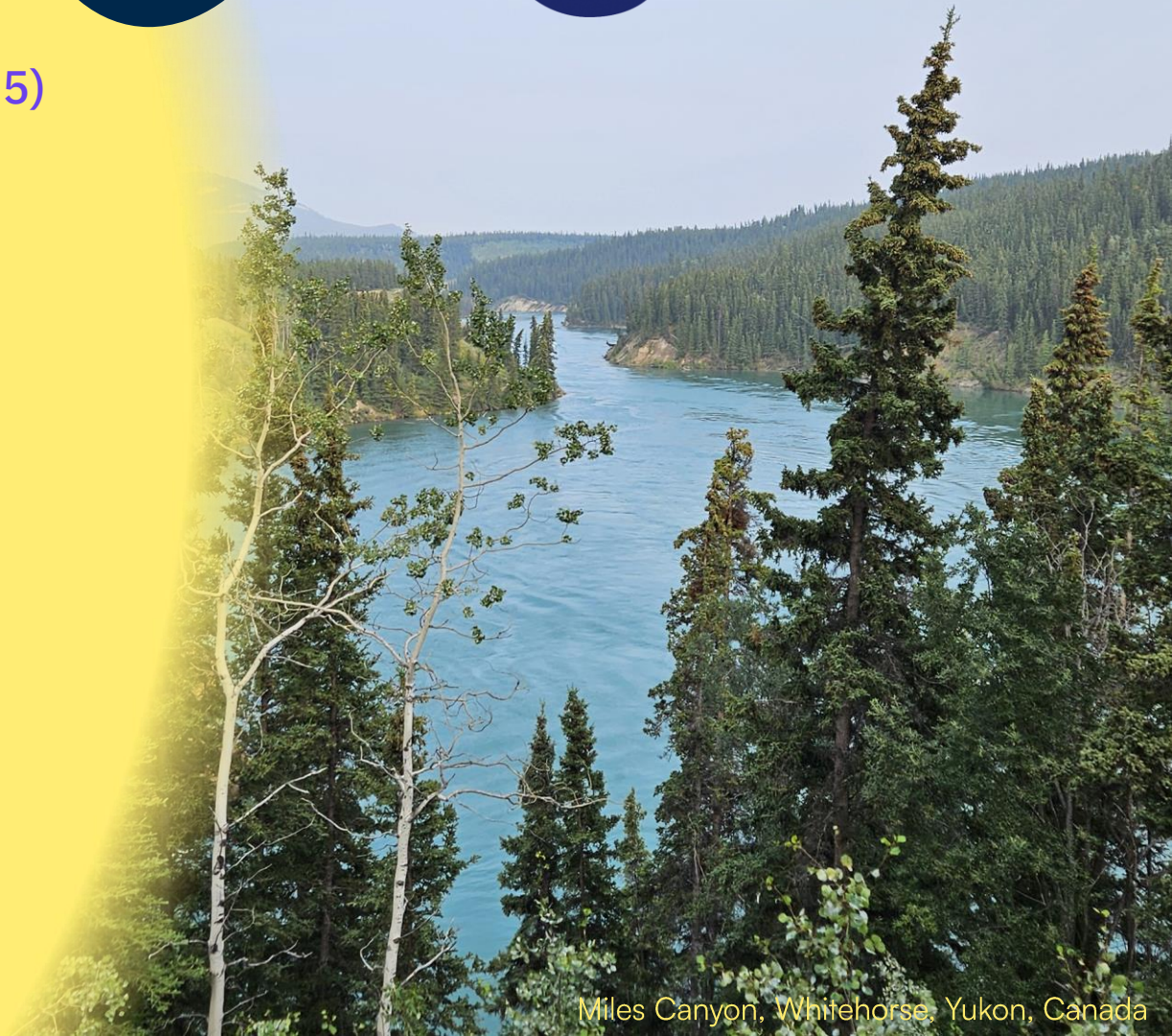
Data Science and Information Theory, ED127 course (2025)

Florent Leclercq

www.florent-leclercq.eu

Institut d'Astrophysique de Paris
CNRS & Sorbonne Université

31 March 2025



Miles Canyon, Whitehorse, Yukon, Canada

Programme

01 Probability theory and signal processing

02 Monte Carlo techniques

03 Advanced Bayesian topics

04 Forecasts, perspectives, simulations

05 Information theory

06 Machine learning theory

Course overview

Homepage: http://florent-leclercq.eu/teaching.php

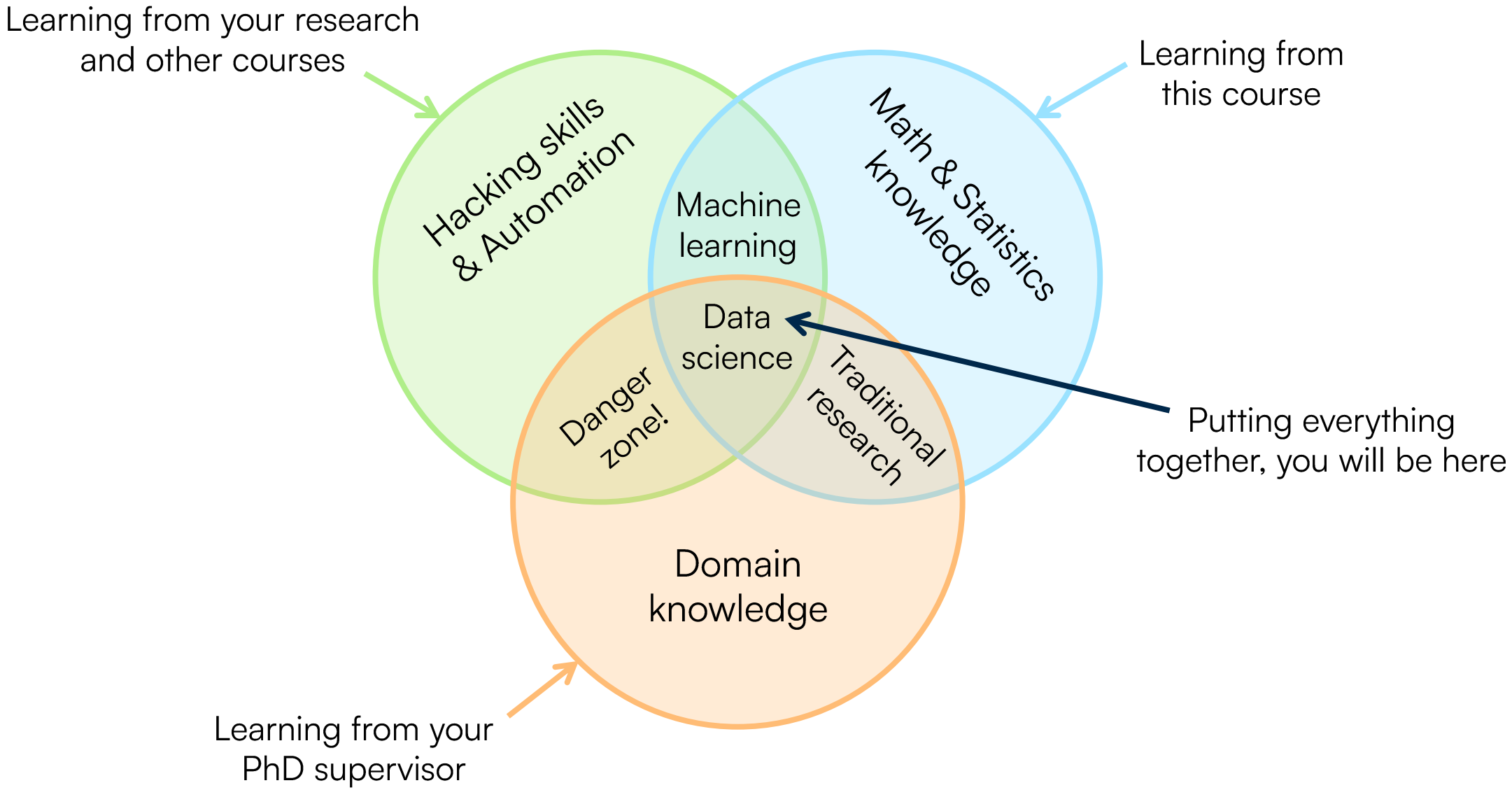
Schedule	Monday 31 March	Tuesday 1 April	Wednesday 2 April	Monday 7 April	Tuesday 8 April	Wednesday 9 April
09:45-11:00	Overview	Monte Carlo	Model Comparison	Fisher Information	Information Theory	Machine Learning Theory
11:00-11:15	Break	Break	Break	Break	Break	Break
11:15-12:30	Probability Theory	Monte Carlo	Bayesian Decision Theory	Bayesian vs Frequentist statistics	Information Theory	Machine Learning Theory
12:30-14:00	Lunch	Lunch	Lunch	Lunch	Lunch	Lunch
14:00-15:15	Bayesian Signal Processing	Monte Carlo	Bayesian Hierarchical Models	Implicit Likelihood Inference	Information Theory	Machine Learning
15:15-15:30	Break	Break	Break	Break	Break	Break
15:30-17:45	Bayesian Signal Processing	Monte Carlo	Data science for your research project	Caveats	Thermodynamics and Inference	Machine Learning

Colour code	Lecture	Hands-on session	Discussion or Q/A session

01 PROBABILITY THEORY

INTRODUCTION

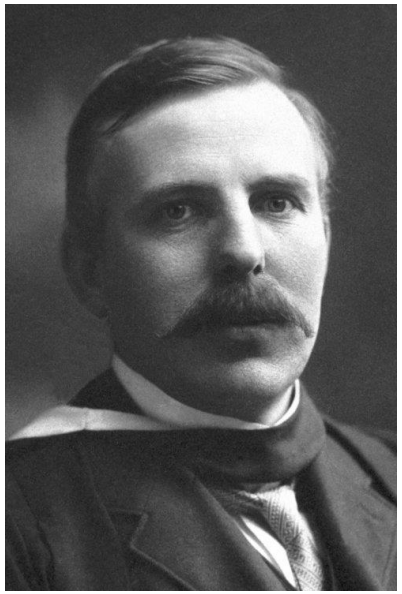
Why you are here



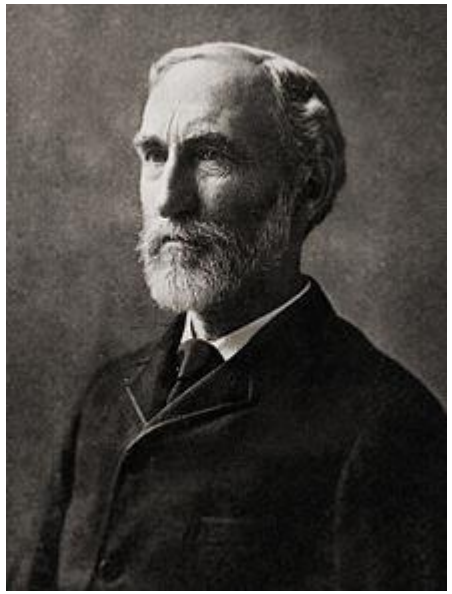
Why proper statistics matter

An historical example: the Gibbs paradox

If your experiment needs statistics, you ought to have done a better experiment.
Ernest Rutherford



Ernest Rutherford
(1871-1937)



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)

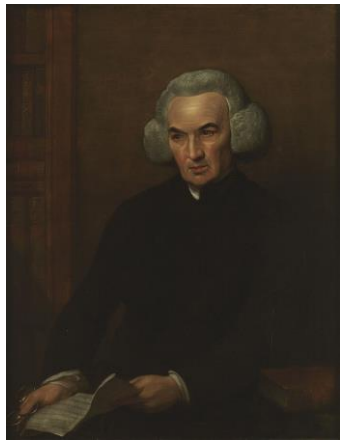
The first clues indicating the need for quantum physics were uncovered by seemingly "unsuccessful" application of statistics.

How to reason rationally in the presence of uncertainty: Bayes' Theorem

- How do we measure a quantity? How do we verify a theory? More broadly, how does knowledge progress?
- Bayes' Theorem (1763): a mathematical statement about how we analyse evidence and change our minds when we obtain new information.
- But why should we use it?
  - Bayes' theorem is trivial and outdated.
  - It measures belief. It says we can learn even from missing or incomplete data, from approximations, from ignorance. It runs counter to the conviction that science requires objectivity and precision.
  - After Laplace's death, it was pronounced dead and buried.

This is (probably!) not the right person

Thomas Bayes
(1701-1761)



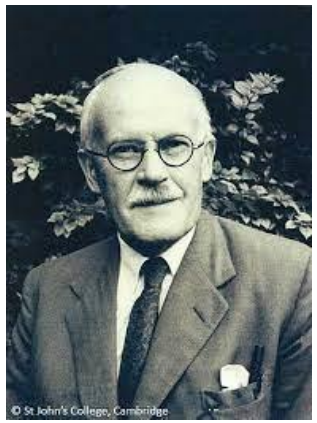
Richard Price
(1723-1791)

Pierre-Simon de Laplace
(1749-1827)

Pictures taken at Bunhill Fields Burial Ground, City of London, 2021

Frequentism versus Bayesianism in the 19th and 20th century

- Two conceptions of the nature of probabilities and scientific questions:
  - "Objective" probabilities linked to the frequency of repetitive random phenomena. Questions related to determined and reproducible experiments.
  - "Subjective" probabilities linked to the certainty given to a measurement or a theory. Questions related to phenomena and choices that do not involve the idea of repetition.

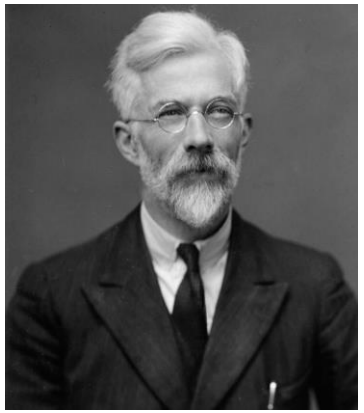


Harold Jeffreys
(1891-1989)

Leonard J. Savage
(1917-1971)

> [Fisher] sometimes published insults that only a saint could entirely forgive.
> Savage 1976, Annals of Statistics

Karl Pearson
(1857-1936)



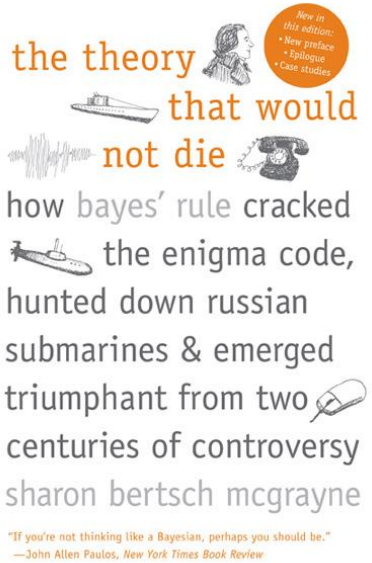
Ronald Aylmer Fisher
(1890-1962)

Jerzy Neyman
(1894-1981)

- Frequentist and Bayesian techniques yield the same results when working with large samples. It is only with small numbers and low occurrences that frequentist estimation and Bayesian induction differ.

The theory that would not die

- And yet, Bayes' theorem helped in many practical situations:
  - Exonerate Alfred Dreyfus from miscarriage of justice (Henri Poincaré, 1899-1906),
  - Save the Bell Telephone system from financial panic (Edward C. Molina, 1907),
  - Predict earthquakes and tsunamis (Harold Jeffreys, 1930-1940),
  - Direct Allied artillery fire and locate German submarines (1939-1945),
  - Break the German navy's Enigma cipher (Alan Turing, 1940-1944),
  - Prove that smoking causes lung cancer (Jerome Cornfield, 1951)
  - Search for a lost H-bomb and then a submarine at sea (John P. Craven, 1966-1968)...

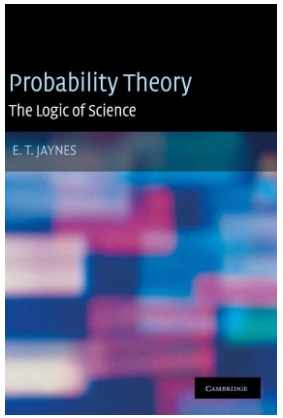


Sharon Bertsch McGrayne (2012)

- The scientific battle lasted for 150 years, until computers arrived.

> The superiority of Bayesian methods is now a thoroughly demonstrated fact in a hundred different areas. One can argue with a philosophy; it is not so easy to argue with a computer printout, which says to us: "Independently of all your philosophy, here are the facts of actual performance."
> Jaynes (2002), Probability Theory — The logic of science

- Cox-Jaynes theorem (1946): Any system to manipulate "plausibilities", consistent with Cox's desiderata, is isomorphic to Bayesian probability theory.

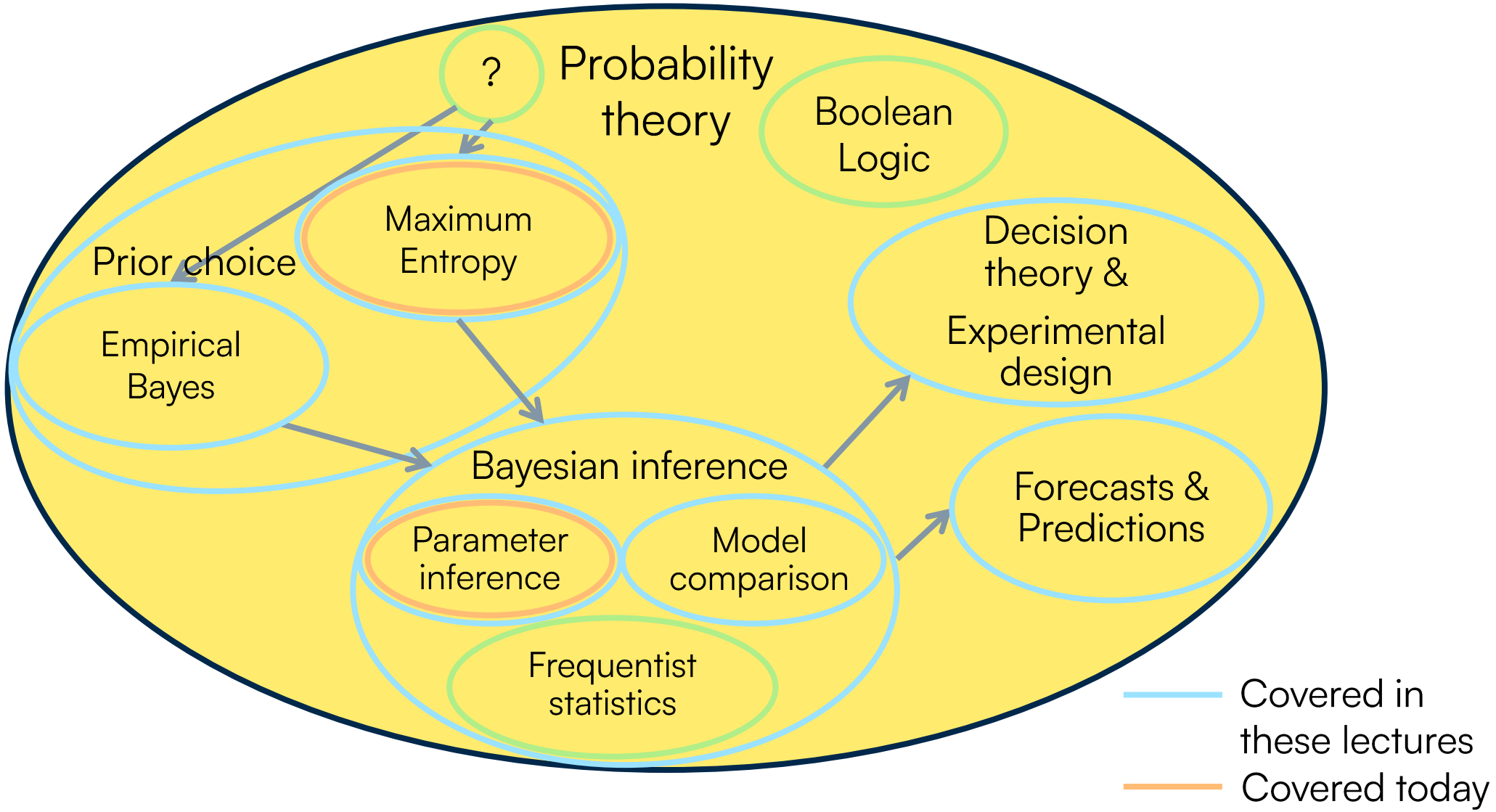


Jaynes (2002)

Richard Threlkeld Cox (1898-1991)

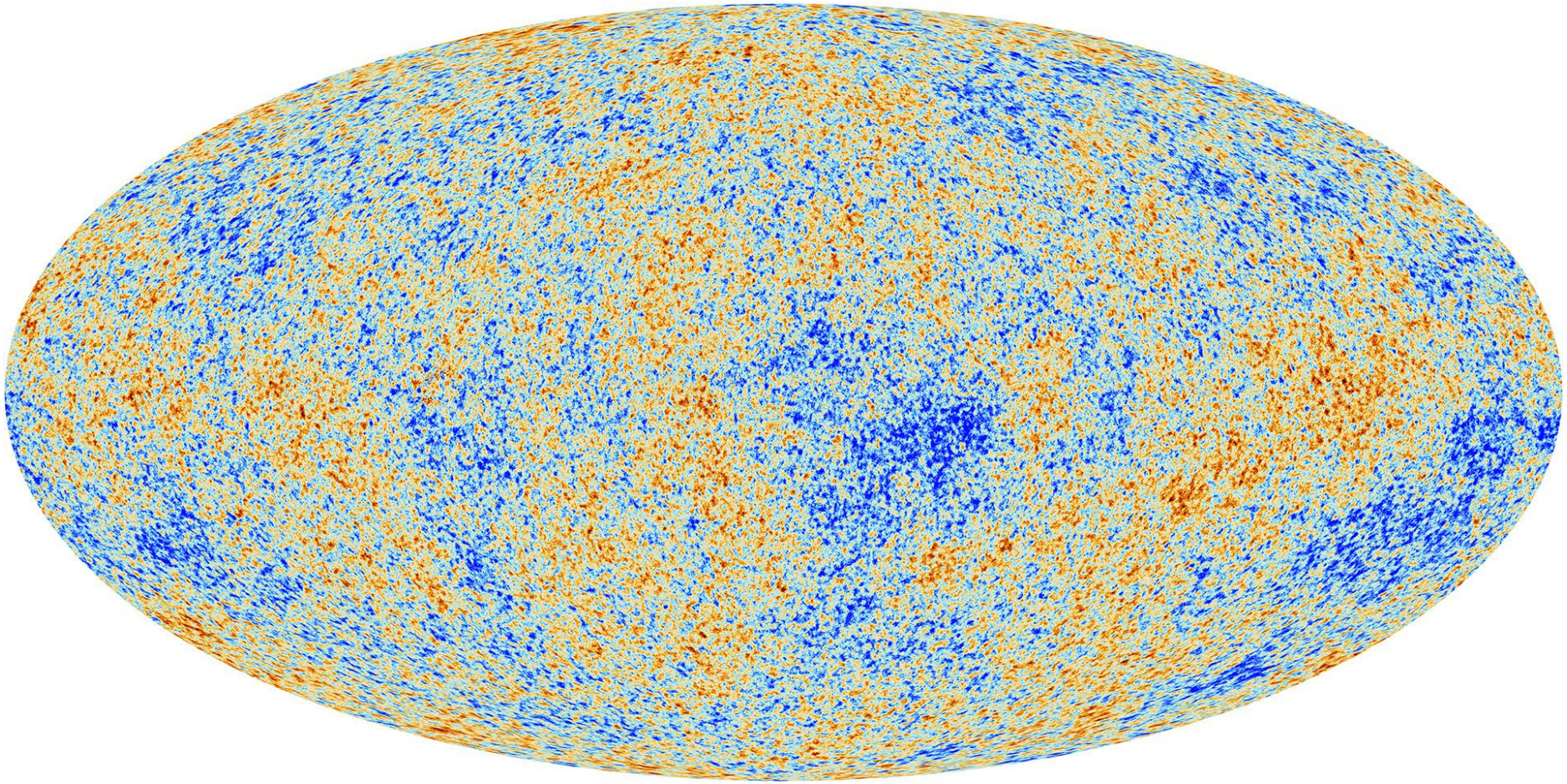
Edwin Thompson Jaynes (1922-1998)

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



BASIC PRINCIPLES OF PROBABILITY THEORY

Typical problem: analysis of the Cosmic Microwave Background

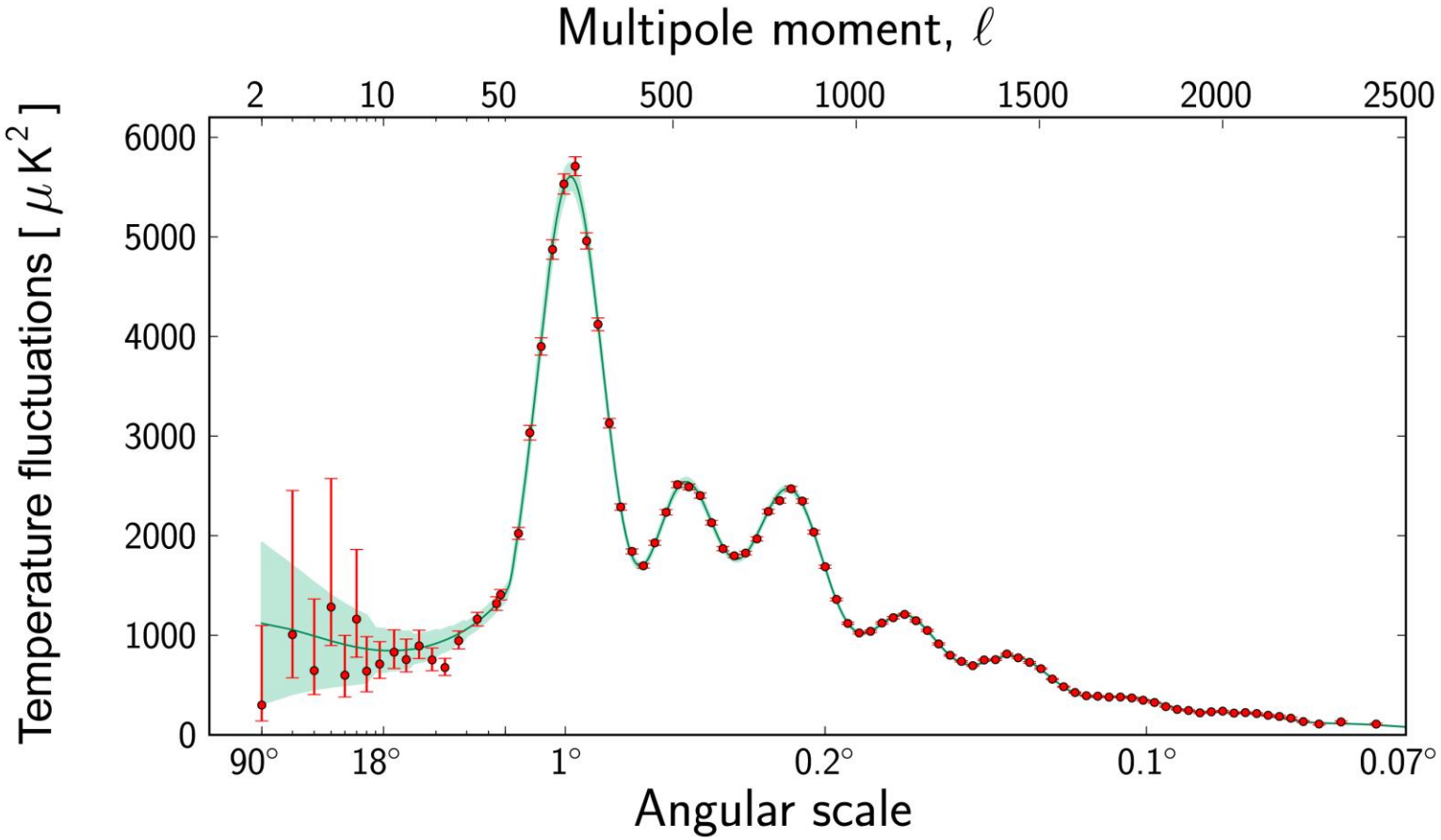


Usually we compress the data into some “summary statistics”, such as the correlation function of temperature fluctuations, or the power spectrum

Planck 2018

Typical problem: analysis of the Cosmic Microwave Background

ΛCDM fits the Planck data well.



Planck 2018 V, 1907.12875

Vocabulary: direct and inverse problems

- “Synthesis problems” are direct problems: given some theory/model of a physical process (a generative model), we want to predict the result of an experiment. This is also called forward modelling.
- “Analysis problems” are inverse problems: given some data, we want to infer something about the process that generated the data.
- Inverse problems are generally harder than predicting an outcome, given a physical process.
- Typical classes of inverse problems:
  - Parameter inference
  - Hypothesis testing
  - Model comparison

What kind of questions do we want to answer?

- Parameter inference:
  - I have a set of (x, y) pairs, with errors. If I assume $y = mx + c$, what are the values of m and c?
  - I have detected 5 X-ray photons from a source at known distance in my lab. What is the luminosity of the source and its uncertainty?
  - By analysing the velocity distribution of stars in a galaxy, what can be inferred about the galaxy's mass distribution and the presence of dark matter?
  - Given LIGO/Virgo gravitational wave data, what are the masses of the inspiralling objects?

What kind of questions do we want to answer?

- Hypothesis testing:
  - Is the cosmic microwave background radiation consistent with (initially) Gaussian fluctuations?
  - Do high-energy cosmic rays exhibit correlations with astrophysical sources?
  - Are neutron star mergers the dominant source of heavy elements (e.g., r-process elements) in the universe?
  - Are the statistical properties of exoplanet populations consistent with standard planet formation models?

What kind of questions do we want to answer?

- Model comparison:
  - Do observations of young stellar objects with protoplanetary disks better support core accretion or disk instability models for planet formation?
  - Based on historical climate data, which climate model more accurately predicts global temperature changes—those incorporating strong feedback mechanisms or those with moderate feedback?
  - Do LHC data support the existence of the Higgs boson, or no Higgs boson?
  - Do observations of Type Ia supernovae better support the cosmological constant model of dark energy or dynamic dark energy models such as quintessence?

The meaning of probability

- Frequentist view: probability describes the relative frequency of outcomes in infinitely long trials.
- Bayesian view: probability describes a degree of belief.
- But!
  - The first view (regarding repetitive phenomena) can be included in the second one.
  - It is misleading when, in 2025, the debate between frequentists and Bayesians is still referred to as if it were ongoing, since it was settled in the second half of the 20th century.
- The Bayesian view expresses what we often want to know, e.g.
  - Given the Planck CMB data, what is the probability that the density parameter of cold dark matter Ω_{m} is between 0.3 and 0.4?
  - and not: given a fictitious infinite population of universes, what is the probability that 95% of them have Ω_{m} between 0.3 and 0.4?
- Logical proposition: a statement that could be true or false.
- Conditional probability: $p(A|B)$ is the degree to which truth of a logical proposition B implies that A is also true.

Basic principles of probability theory

- Joint probabilities and conditional probabilities: $p(A, B) = p(A|B)p(B)$
- Marginal probabilities: $p(A) = \int p(A, B)\,\mathrm{d}B$
- Normalisation rule: $p(A) + p(\bar{A}) = 1$ or $\int p(A)\mathrm{d}A = 1$
- Product rule (logical "and"): $p(AB|C) = p(A|BC)\,p(B|C)$
- Sum rule (logical "or"): $p(A + B|C) = p(A|C) + p(B|C) - p(AB|C)$
- Bayes Theorem (parameter inference):

posterior, likelihood, prior

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

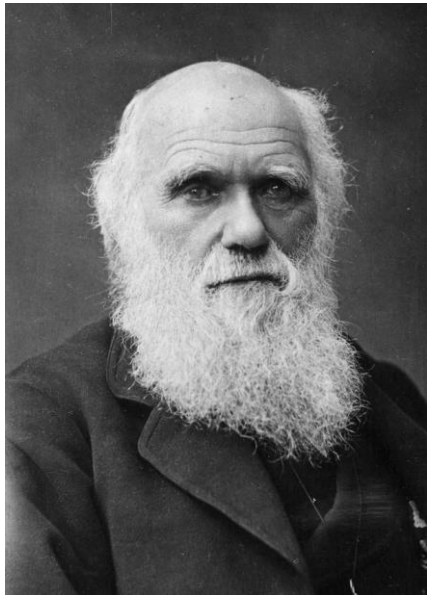
evidence

- Bayes factor (model comparison): $\mathcal{B}_{12} = \frac{p(d|\mathcal{M}_1)}{p(d|\mathcal{M}_2)}$ where $p(d|\mathcal{M}_i) = \int p(d|s_i, \mathcal{M}_i)p(s_i|\mathcal{M}_i)\,\mathrm{d}s_i$

  The ratio of evidences (not of likelihoods) takes into account the effect of "Occam's razor".

Conditional probabilities

- Avoid the “probability 101” mistake: $p(B|A)$ is not the same as $p(A|B)$!
- Example:



Charles Darwin (1809-1882), photographed by Herbert Rose Barraud (1881)

  - A = is male; B = has beard
  - $p(B|A) \approx 0.1$ (?)
  - $p(A|B) \approx 1$

Conditional probabilities: warm-up exercises

- Exercise: Bayesian inference with virology tests
- Exercise: The Monty Hall problem

Parameter inference in practice

- Notations: data d, model $\mathcal{M}$, parameters θ
- First rule: write down what you want to know. Usually, it is the probability distribution for the parameters, given the data, and assuming a model: $p(\theta|d, \mathcal{M})$. This is the posterior.
- To compute it we use Bayes' theorem: $p(\theta|d, \mathcal{M}) = \dfrac{p(d|\theta, \mathcal{M})p(\theta|\mathcal{M})}{p(d|\mathcal{M})}$
  - where the likelihood is $L(\theta) = p(d|\theta, \mathcal{M})$ (an unnormalised function of θ, while d is fixed)
  - and the prior is $\pi(\theta) = p(\theta|\mathcal{M})$
- Dropping the dependence on $\mathcal{M}$: $p(\theta|d) = \dfrac{L(\theta)\pi(\theta)}{p(d)} \propto L(\theta)\pi(\theta)$
  - The evidence $p(d|\mathcal{M})$ is irrelevant for parameter inference.
  - In practice, in your code, you will write: $\ln p(\theta|d) = \ln L(\theta) + \ln \pi(\theta) + \text{const.}$
- So you need to analyse the problem:
  - What is the data? → d
  - What is the model for the data? → $L(\theta)$
  - What are the parameters?
  - What is the prior distribution? → $\pi(\theta)$
- After the experiment, the posterior may act as the prior for the next experiment: we "update the prior" with the information from the experiment

Some conceptual considerations regarding Bayesian inference

- Self-consistency? Yes.
  - Consider data from 2 experiments. We can do one of 3 things:
    - Define prior; obtain posterior from dataset 1; update the prior, then analyse dataset 2
    - As above, but swap datasets 1 and 2
    - Define prior; obtain posterior from datasets 1 and 2 combined
  - These have to (and do) give the same answer.

- Subjective? Yes and No. Jaynes (2002), Probability Theory — The logic of science, Chap 2, p. 39
  - Any probability assignment is necessarily subjective in the sense that it describes only a state of knowledge, and not anything that could be measured in a physical experiment. But *whose* state of knowledge? Answer is always: any rational thinker. Anyone who has the same information but comes to a different conclusion is necessarily violating one of the Cox-Jaynes desiderata.
  - Probability assignments completely objective in the sense that they are independent of the personality of the user (e.g. in parameter inference the posterior is determined unambiguously from the prior and the likelihood). It is objectivity in this sense that is needed for a "scientifically respectable theory of inference".

Some conceptual considerations regarding Bayesian inference

- Priors? Yes, they are needed.
  - The No-Free-Lunch theorem for optimisation problems:
    - When searching for the local extremum of a target function (the likelihood in our case) in a finite space, the average performance of algorithms (that do not resample points) across all possible problems is identical (Wolpert & Macready 1997).
    - In other words: if no assumptions are made about the data, there is NO reason to prefer one model over any other. This means, without prior knowledge we have to test all models!
    - Important implication: there exists no algorithm that performs equally well on all inference tasks (Ho & Pepyne 2002); prior information should always be used to match procedures to problems.

> It appears to be a quite general principle that, whenever there is a randomized way of doing something, then there is a nonrandomized way that delivers better performance but requires more thought.
> Jaynes (2002), Probability Theory — The logic of science

  - Prior specification is for model comparison a key ingredient of the model building step. If the prior cannot be meaningfully set, then the physics in the model is probably not good enough.

Wolpert & Macready (1997); Ho & Pepyne (2002)

PRIOR CHOICE AND MAXIMUM ENTROPY PRINCIPLE

Priors

- Any probability is always conditional on some background information I (usually omitted in the notation): $p(\cdot) = p(\cdot|I)$
- Bayesian prior: (usually) the state of knowledge before the new data are collected. You never "know nothing" about a problem!
- For parameter inference, the prior becomes unimportant as more data are added and the likelihood dominates (Bernstein-von Mises theorem).
- For model comparison, the prior remains important.
- Issues: One usually wants an "uninformative"* prior, but what does this mean?
- Typical choices:
  - Uniform (flat) prior: $\pi(\theta) = \text{const.}$ (within some range)
  - Scale-invariance prior (Jeffreys prior): $\pi(\theta) \propto 1/\theta$ (an improper prior: not normalisable when $\theta \to 0$ or $\theta \to +\infty$)

Exercise: The dominance of the likelihood in Bayesian inference

Exercise: Ignorance priors for an urn problem

* Actually, it's better not to use these terms — other people use them to mean different things — just say what your prior is! Uniform priors can in fact be very informative.

Conjugate priors

- A conjugate prior is a prior distribution that, when combined with the likelihood function, results in a posterior distribution that belongs to the same family as the prior. This property simplifies the process of updating beliefs in Bayesian statistics.
- Conjugate priors are particularly useful because they allow for straightforward analytical solutions, when possible. They were very popular in Bayesian statistics before computers arrived.
- Some examples:

Model parameters	Likelihood	Prior	Posterior
Probability of a Bernoulli experiment	Binomial	Beta	Beta
Poisson rate	Poisson	Gamma	Gamma
Variance	Gaussian with known mean	Inverse-Gamma	Inverse-Gamma
Mean vector	Multivariate Gaussian with fixed covariance matrix	Multivariate Gaussian	Multivariate Gaussian

For expression of the hyperparameters and more examples, see e.g. https://en.wikipedia.org/wiki/Conjugate_prior.

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 $\mathcal{M}_2$:

$$\left.\begin{array}{l} p(\mathcal{M}_1) = p(\mathcal{M}_2) \\ p(\mathcal{M}_1) + p(\mathcal{M}_2) = 1 \end{array}\right. \Rightarrow \quad p(\mathcal{M}_1) = p(\mathcal{M}_2) = \frac{1}{2} \qquad \mathbb{Z}_2\text{-symmetry}$$

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

$$\left.\begin{array}{l} \mathrm{d}T = \mathrm{d}x \\ p(x) = p(x+s) \quad \forall s \end{array}\right. \Rightarrow \quad p(x) = C \qquad \text{Uniform prior} \quad U(1)\text{-symmetry}$$

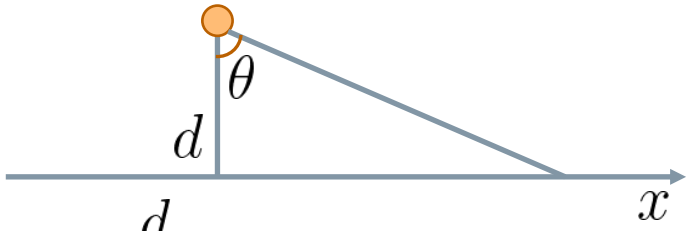
- "Scale parameter": $T(x) = ax \quad \forall a$

$$\left.\begin{array}{l} \mathrm{d}T = a\,\mathrm{d}x \\ p(x) = a\,p(ax) \quad \forall a \end{array}\right. \Rightarrow \quad p(x) = C/a \qquad \text{Jeffreys prior} \quad U(1)\text{-symmetry}$$

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

The lighthouse problem

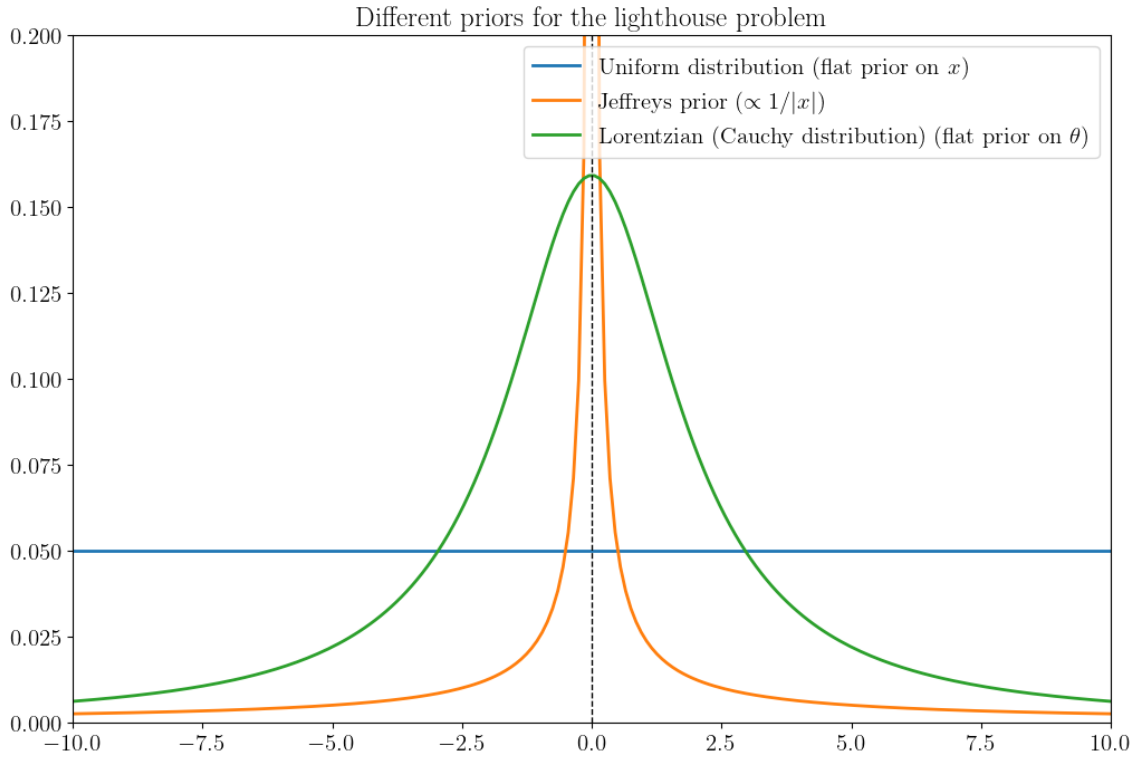
Exercise: The lighthouse problem

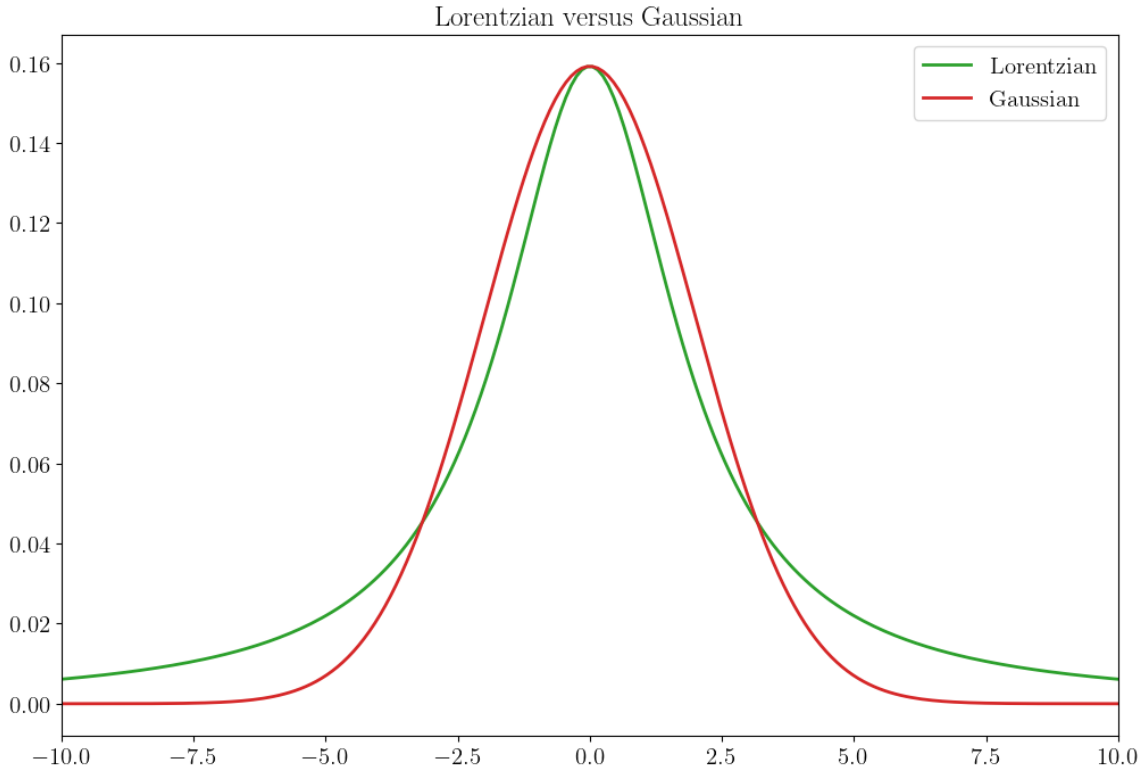


$$x = d\tan\theta$$
$$\mathrm{d}x = d(1+\tan^2\theta)\mathrm{d}\theta$$
$$p(x)\mathrm{d}x = p(\theta)\mathrm{d}\theta$$

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

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.

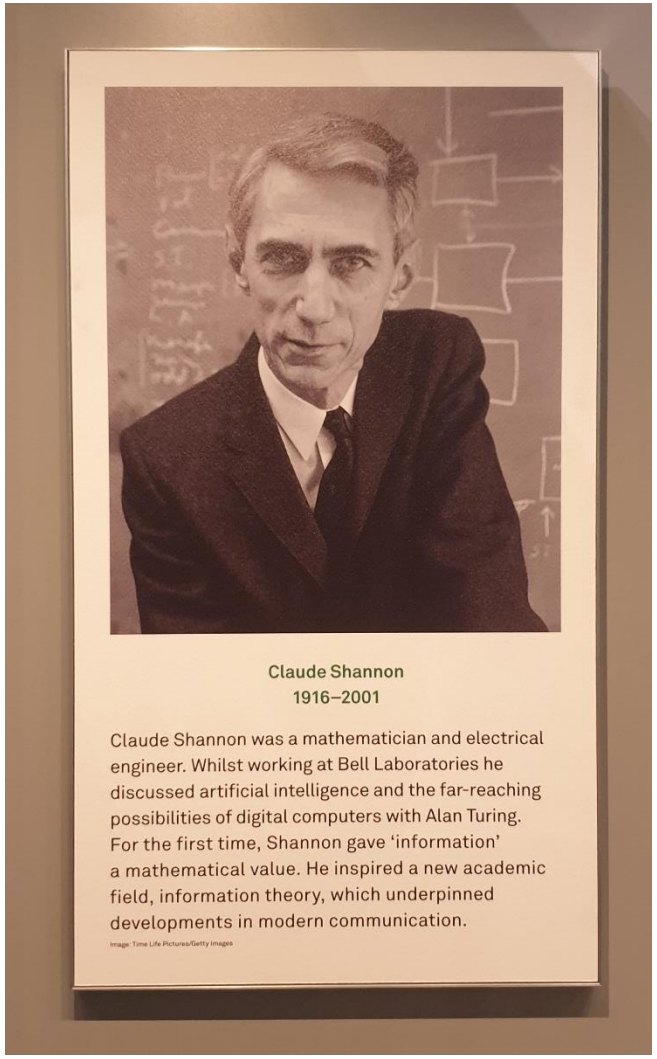
The maximum entropy principle

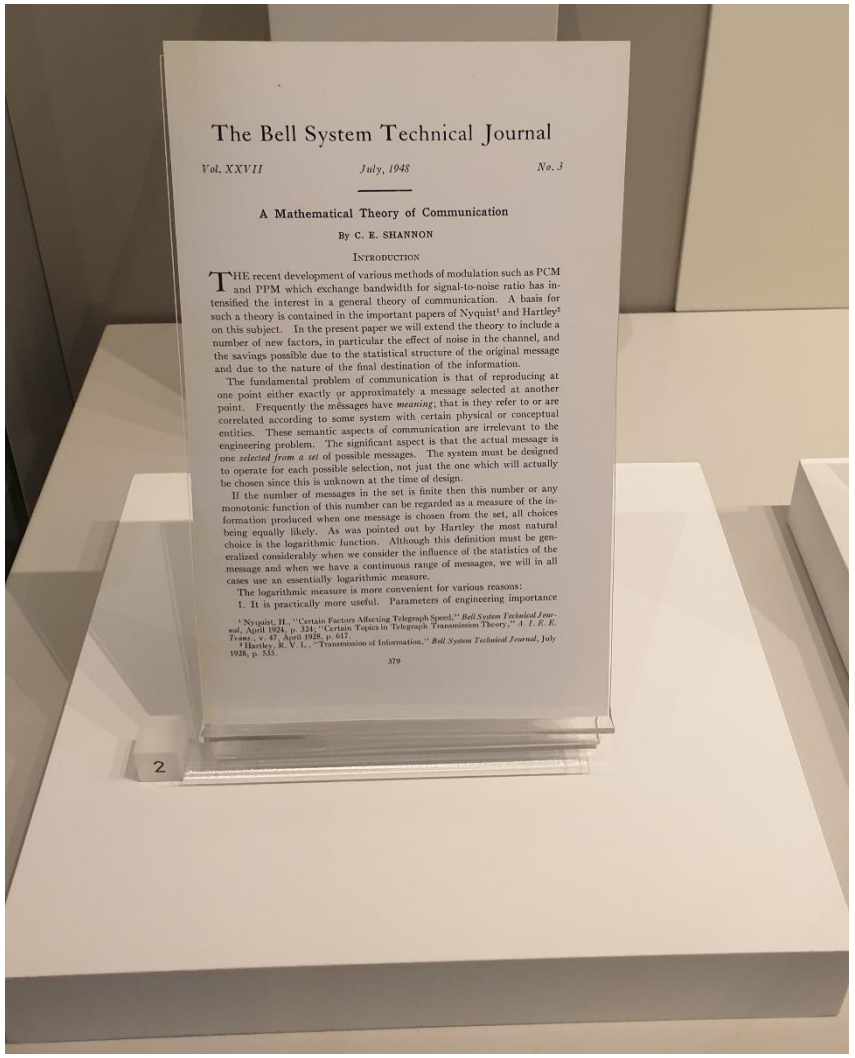
- Maximising the entropy: a general method to select priors while accounting for:
  - indifference about states of equal knowledge
  - relevant prior information
- What should $H[p]$ be for a source of information producing N finite "words" with probabilities p_n ?
- Desiderata:
  - If all words are 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): $$H[p] \propto -\sum_n p_n \log_2 p_n$$

The birth of information theory

- Pictures taken at the Science Museum, South Kensington, 2021.





Information entropy
Shannon (1948)

> Why don't you call it entropy? In the first place, a mathematical development very much like yours already exists in Boltzmann's statistical mechanics, and in the second place, no one understands entropy very well, so in any discussion you will be in a position of advantage.
> von Neumann to Shannon, about a name for "missing information"

The loaded dice

Exercise: 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 distribution 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 in the variables: the method of Lagrange multipliers:
   - We write the 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 $\frac{\partial \mathcal{L}}{\partial \mu} = 0$.

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

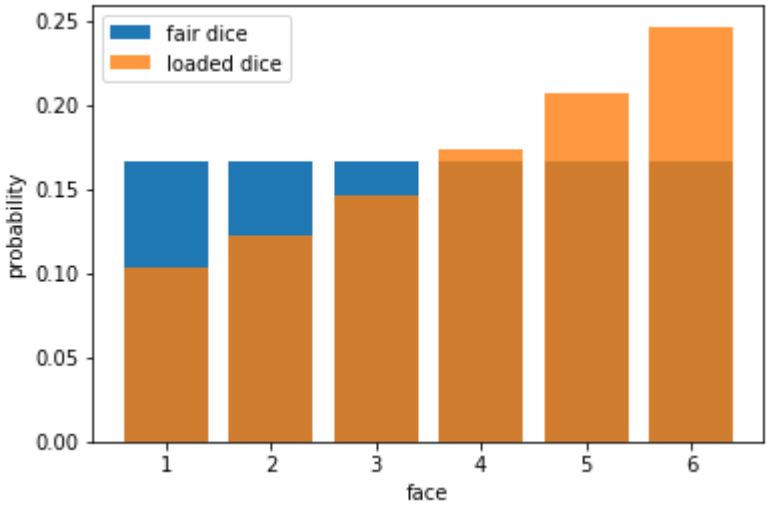
$$p_n = \frac{\mathrm{e}^{-\lambda n}}{Z} \qquad \text{with} \qquad \ln Z \equiv 1 + \mu$$

   - 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



- 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: $p_n = \dfrac{1}{N}$
  - 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

01 SIGNAL PROCESSING

GAUSSIAN RANDOM FIELDS

Gaussian random fields

Exercise: Gaussian random fields

- Definition: any random vector x with pdf

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

  for any vector μ (the mean) and any symmetric positive-definitive matrix C (the covariance matrix).

- Property: one can check that the mean is actually μ and the covariance matrix is actually C, i.e.

$$\langle x\rangle = \int_{-\infty}^{+\infty} x p(x|\mu, C)\,\mathrm{d}x = \mu \quad \text{and} \quad \langle (x-\mu)(x-\mu)^\intercal\rangle = \int_{-\infty}^{+\infty} (x-\mu)(x-\mu)^\intercal p(x|\mu, C)\,\mathrm{d}x = C$$

- Trick for doing Gaussian integrals: integration by differentiation.
  - Since the Gaussian pdf has only one global maximum, the mean of the pdf is its mode (its maximum) and it can be found by maximising the exponent. Therefore, μ is found by maximising $\partial_x \ln p(x|\mu, C)$:

$$-\partial_x \ln p(x|\mu, C)|_{x_{\max}} = \partial_x \left[\frac{1}{2}(x-\mu)^\intercal C^{-1}(x-\mu)\right]\Bigg|_{x_{\max}} = C^{-1}(x-\mu)|_{x_{\max}} = 0 \quad \text{gives} \quad x_{\max} = \mu$$

  - It's easy to verify that $\partial_x \partial_{x^\intercal} \ln p(x|\mu, C) = -C^{-1}$. Therefore, C is found by $C = -\left[\partial_x \partial_{x^\intercal} \ln p(x|\mu, C)\right]^{-1}$ (look at the exponent, select the coefficient matrix of all quadratic terms in the variable, invert, and multiply by -1).

Gaussian random fields

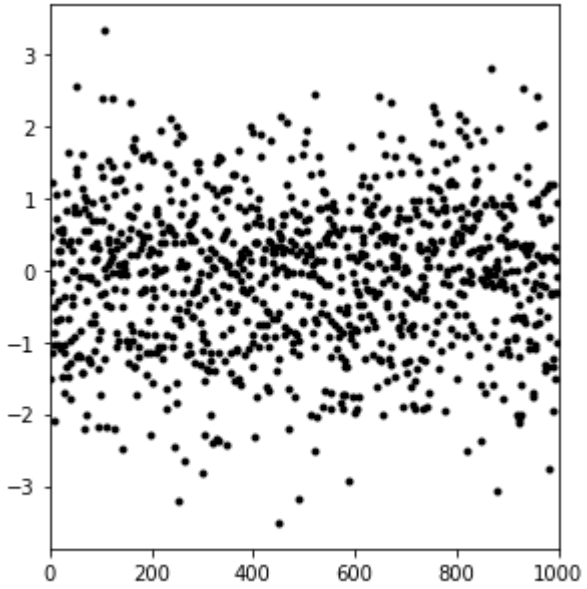
- Generating a Gaussian random field x:
  - Draw a white noise vector ξ (uncorrelated unit-Gaussian variables)
  - Find the matrix square-root of C: $\sqrt{C}$ (any such matrix works)
  - Compute $x = \sqrt{C}\xi + \mu$
- Then x will be a sample from the desired pdf:

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

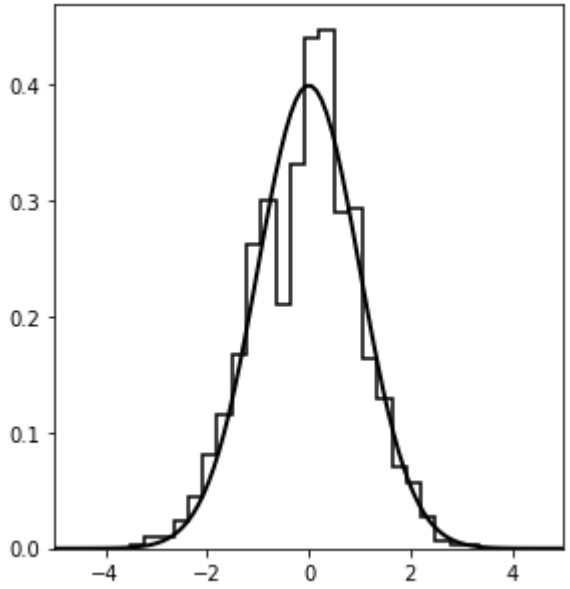
$$-2\ln p(x|\mu, C) = (x-\mu)^\intercal C^{-1}(x-\mu) + \ln|2\pi C|$$

Gaussian random fields: examples

white noise



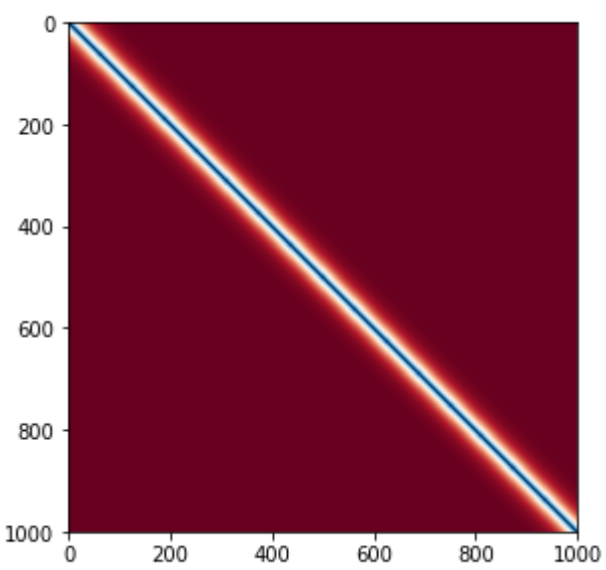
histogram

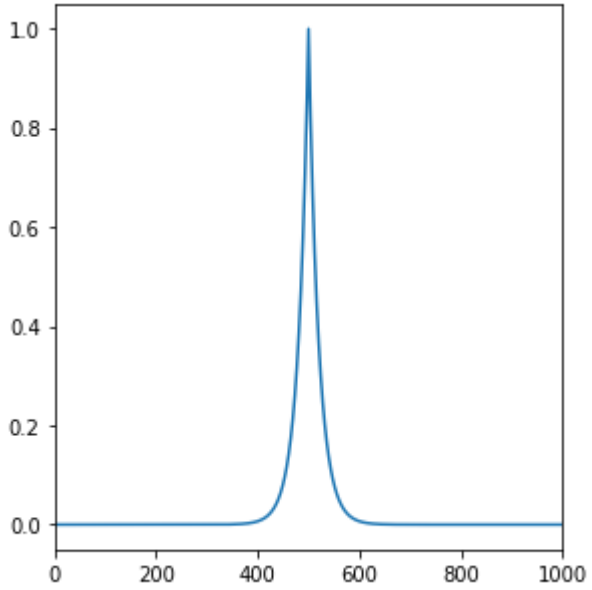


Gaussian random fields: examples

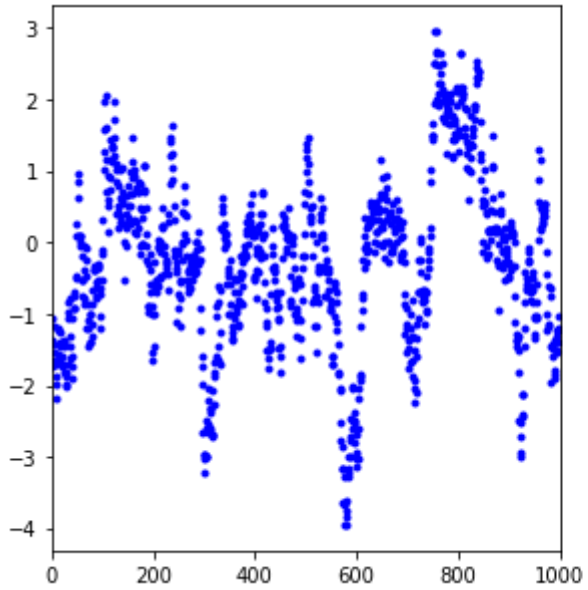
$$C_{ij} = \exp\left(-\frac{|i-j|}{20}\right)$$

covariance matrix

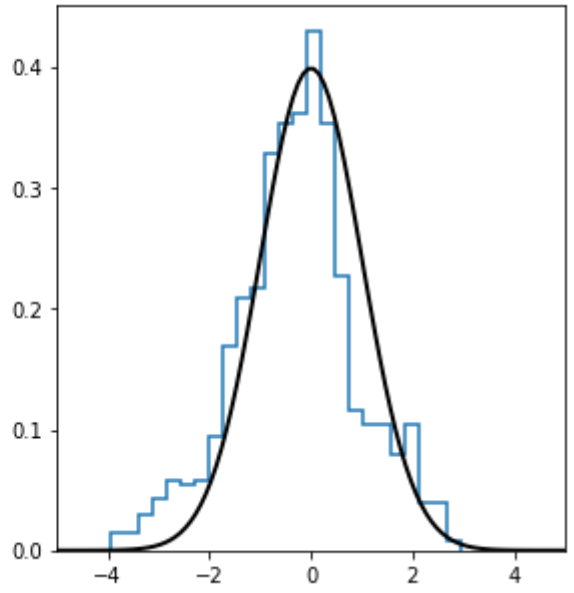




GRF



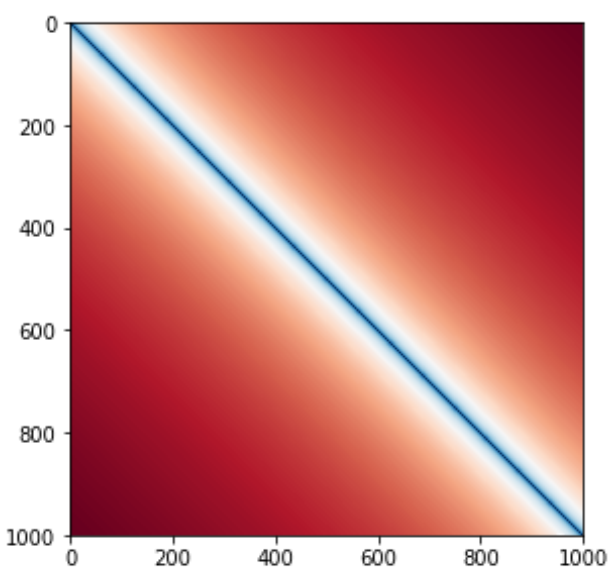
histogram

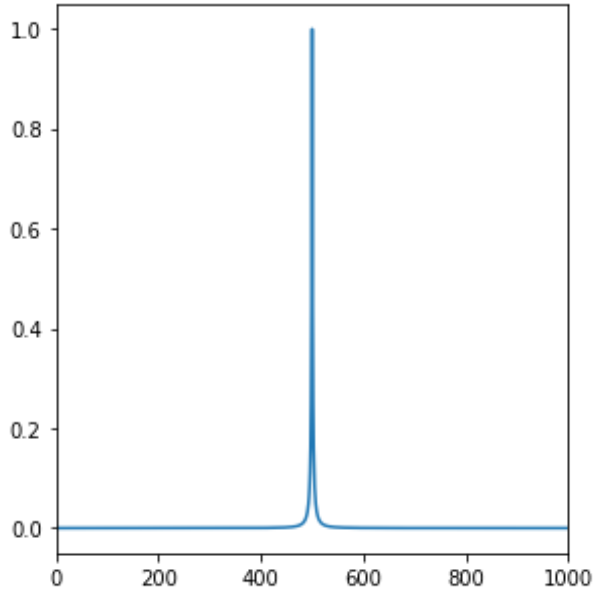


Gaussian random fields: examples

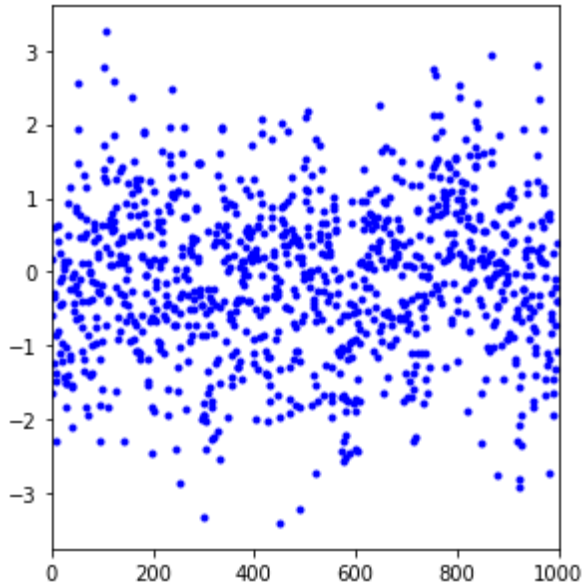
$$C_{ij} = \frac{1}{(1 + |i - j|/2)^2}$$

covariance matrix

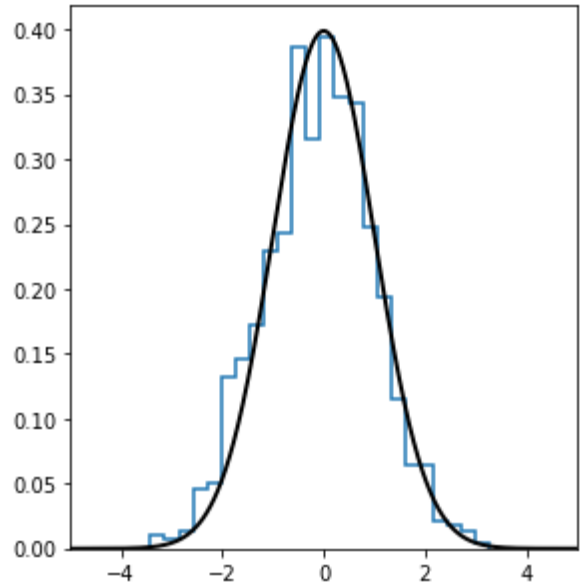




GRF



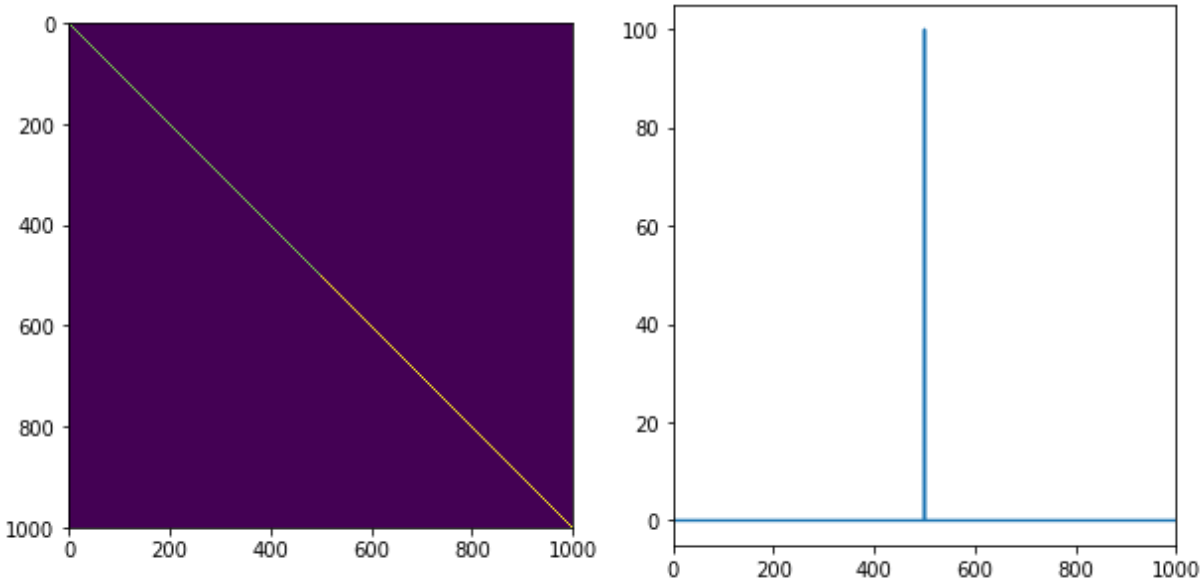
histogram



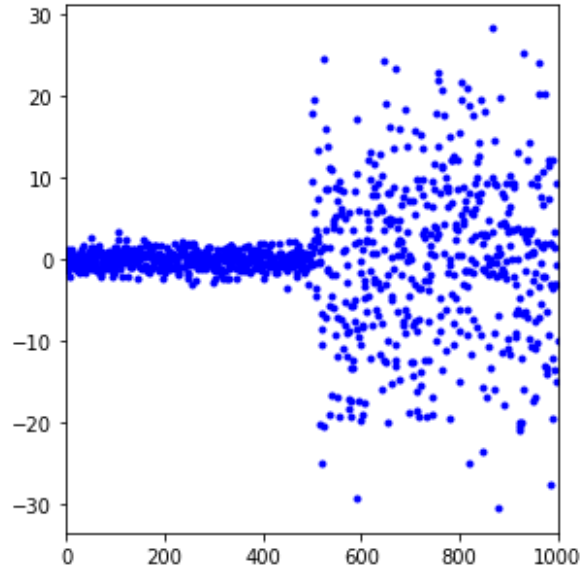
Gaussian random fields: examples

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

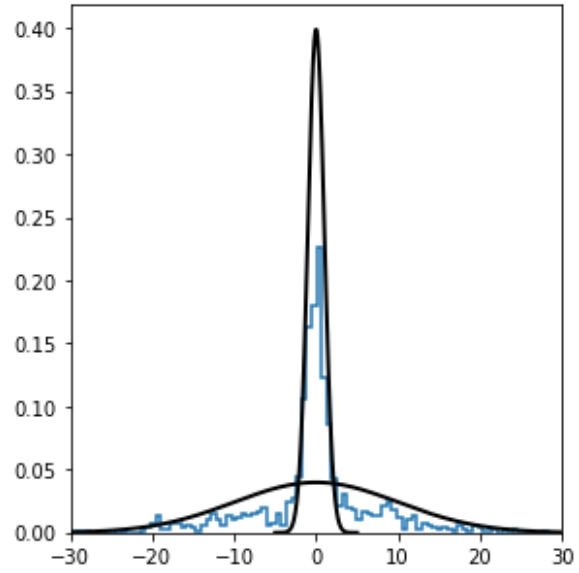
covariance matrix



GRF



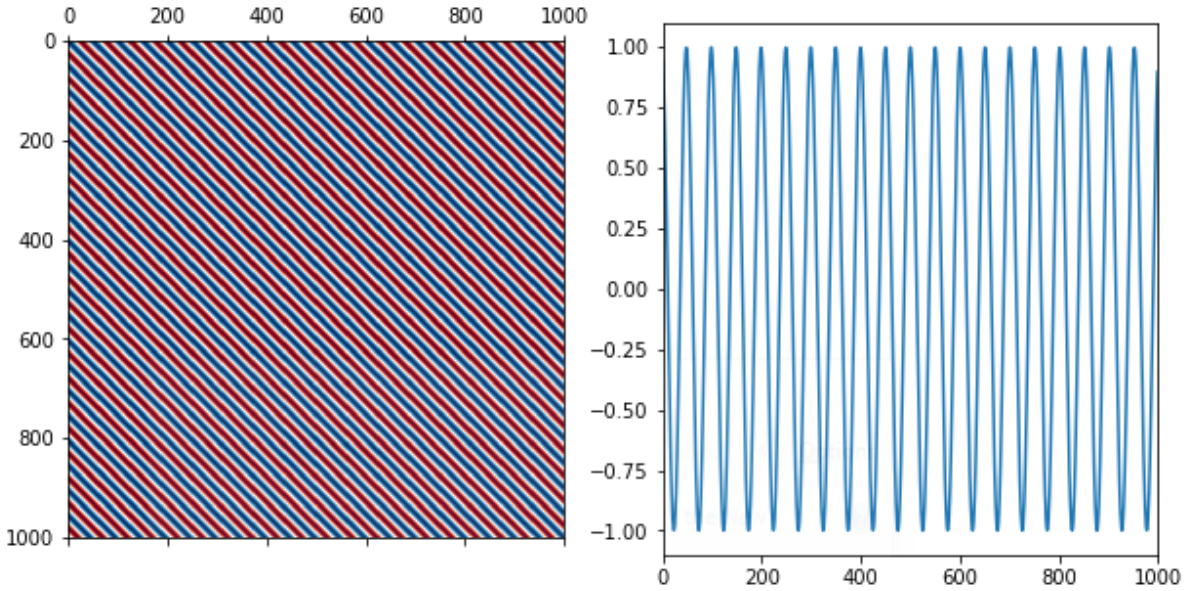
histogram



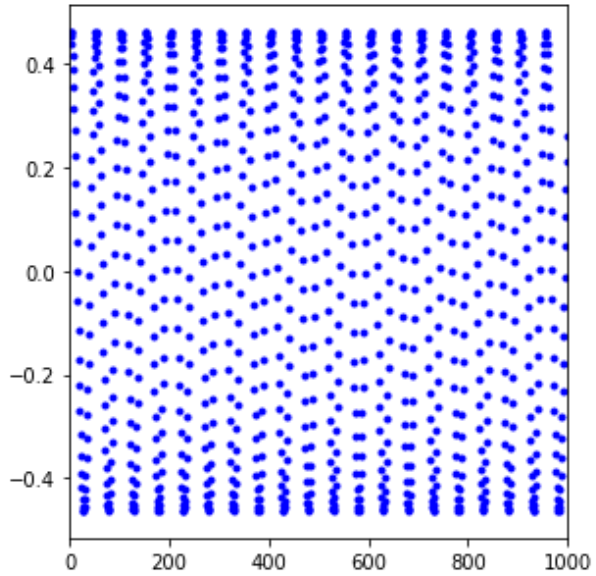
Gaussian random fields: examples

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

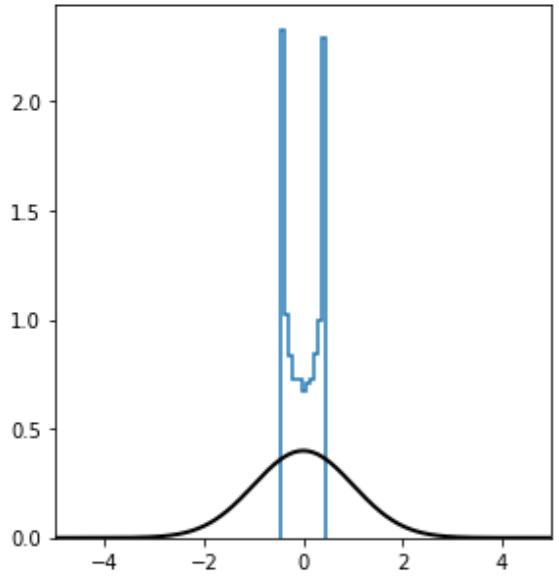
covariance matrix



GRF



histogram

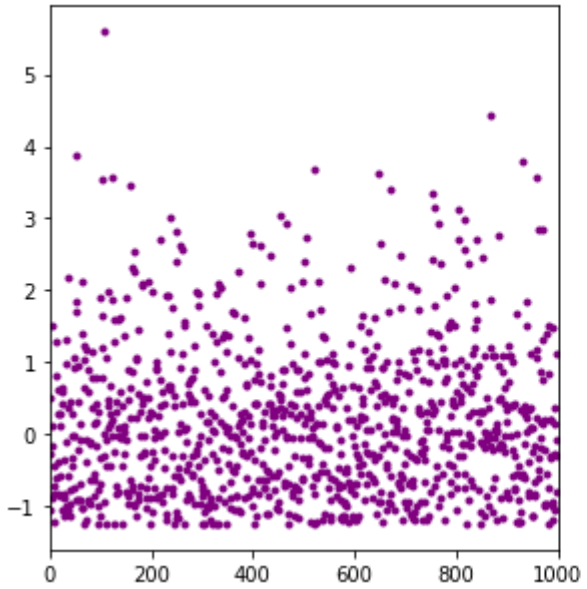


Histograms of Gaussian random fields are not always Gaussian!

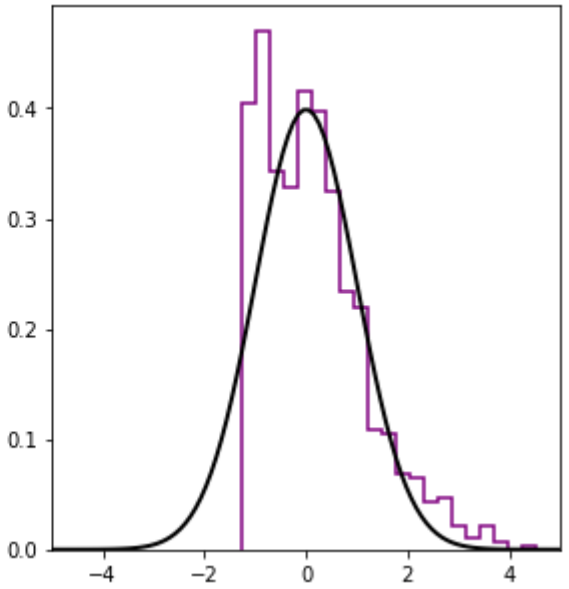
Example of a non-Gaussian signal

- $s = \Phi + f_{\mathrm{NL}}\Phi^2$ where Φ is a GRF.
- In cosmology, this is called “local-type” non-Gaussianity
- Primordial non-Gaussianity contains information on cosmological inflation.

signal



histogram



- The one-point pdf is skewed.

Marginals and conditionals of Gaussian random fields

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

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

- Theorem: The marginal ($p(x)$) and the conditional ($p(x|y)$) pdfs are also Gaussian, with means and variances given below.

- Marginals:
$$\langle x \rangle_{p(x)} = \int x p(x,y)\,\mathrm{d}y = \mu_x$$
$$\langle (x-\mu_x)(x-\mu_x)^\intercal \rangle_{p(x)} = \int (x-\mu_x)(x-\mu_x)^\intercal p(x,y)\,\mathrm{d}y = C_{xx}$$

i.e. the marginal mean and covariance are just the corresponding parts of the joint mean and covariance.

- Conditionals:
$$\langle x \rangle_{p(x|y)} \equiv \mu_{x|y} \quad \text{where}$$
$$\langle (x-\mu_x)(x-\mu_x)^\intercal \rangle_{p(x|y)} \equiv C_{x|y}$$

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

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

For a proof, see e.g. Leclercq (2015), appendix A.

BAYESIAN SIGNAL PROCESSING & WIENER FILTERING

Bayesian denoising (Wiener filtering)

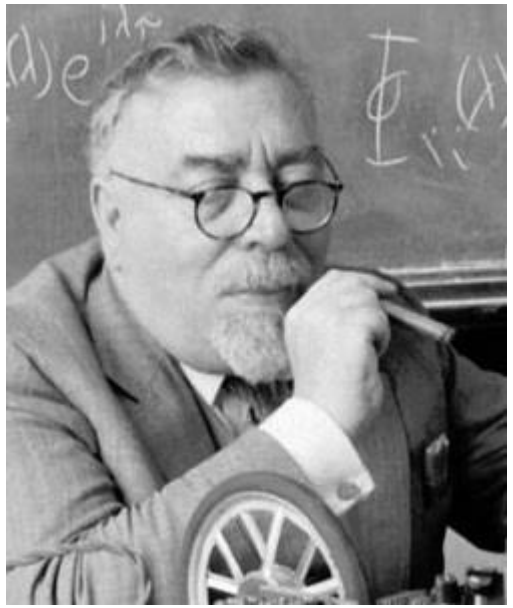
Exercise: Bayesian denoising

- 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

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



Norbert Wiener
(1894-1964)

- Final expressions:

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

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

Wiener filtering: derivation

- The canonical expression for a Gaussian is:

$$-2\ln p(x|\mu,C) = \ln|2\pi C| + (x-\mu)^\intercal C^{-1}(x-\mu) = \ln|2\pi C| + \eta^\intercal \Lambda^{-1}\eta - 2\eta^\intercal x + x^\intercal \Lambda x$$

  where $\Lambda \equiv C^{-1}$ and $\eta \equiv C^{-1}\mu$

- Assuming $\mu_s = \mu_d = 0$ (it's easy to put the mean back in when necessary), we have for the Wiener filtering problem:

$$-2\ln p(s) = \ln|2\pi S| + s^\intercal S^{-1} s$$

$$-2\ln p(d|s) = \ln|2\pi N| + (d-s)^\intercal N^{-1}(d-s)$$

$$= \ln|2\pi N| + \eta^\intercal N \eta - 2\eta^\intercal s + s^\intercal N^{-1} s \qquad \text{with} \quad \eta \equiv N^{-1}d$$

- Therefore:

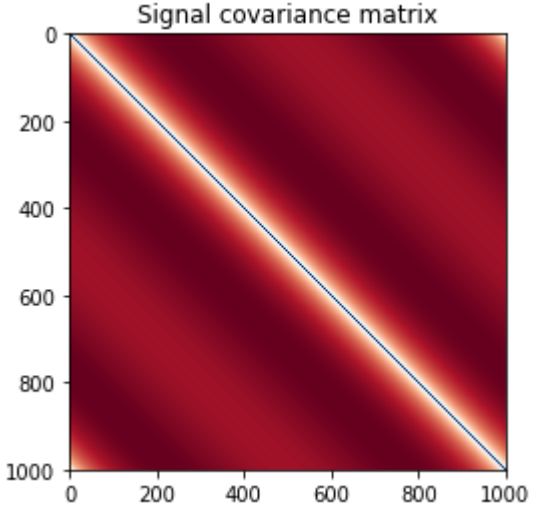
$$-2\ln p(s|d) = \text{const} - 2\eta^\intercal s + s^\intercal (S^{-1} + N^{-1}) s$$

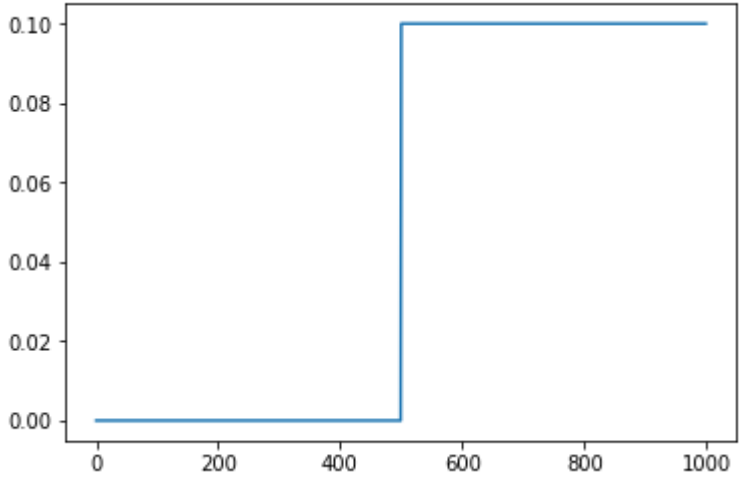
- The posterior has the canonical expression of a Gaussian with covariance matrix $W = (S^{-1} + N^{-1})^{-1}$, and its mean is $W\eta = (S^{-1} + N^{-1})^{-1} N^{-1} d$.

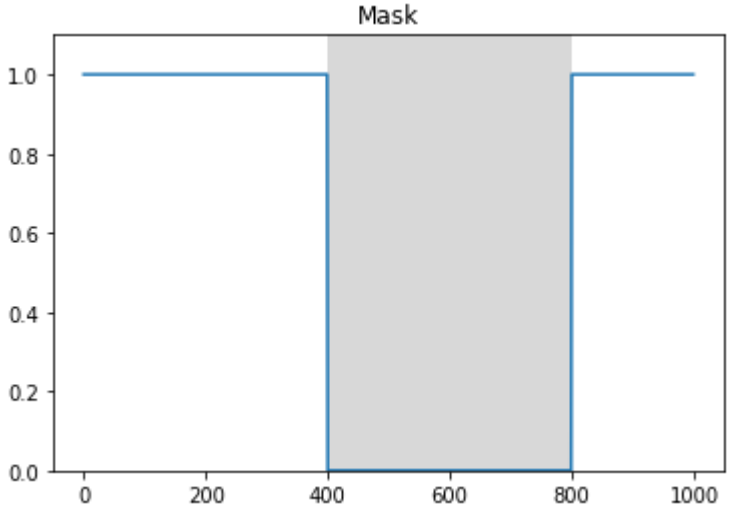
Bayesian denoising (Wiener filtering): example

- Setup signal and noise covariance matrices

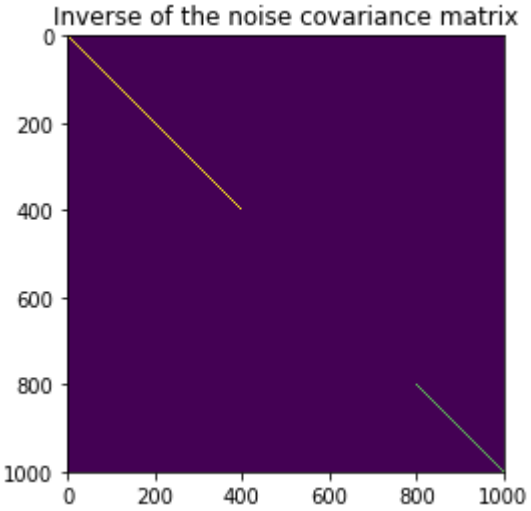
S







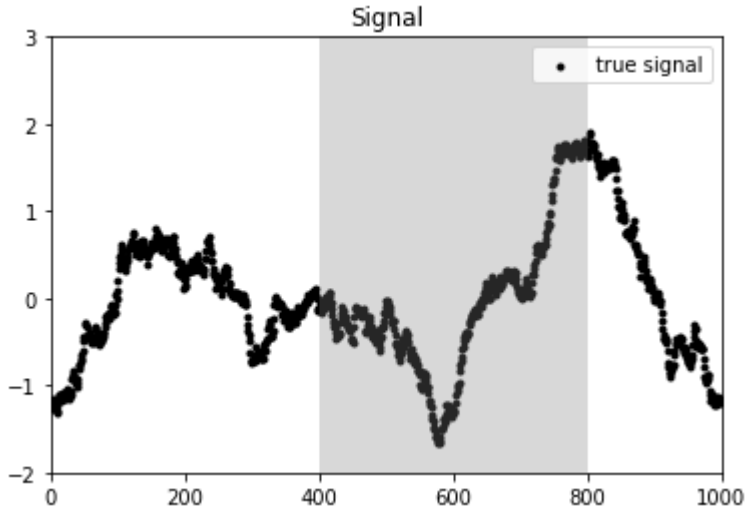
N^{-1}

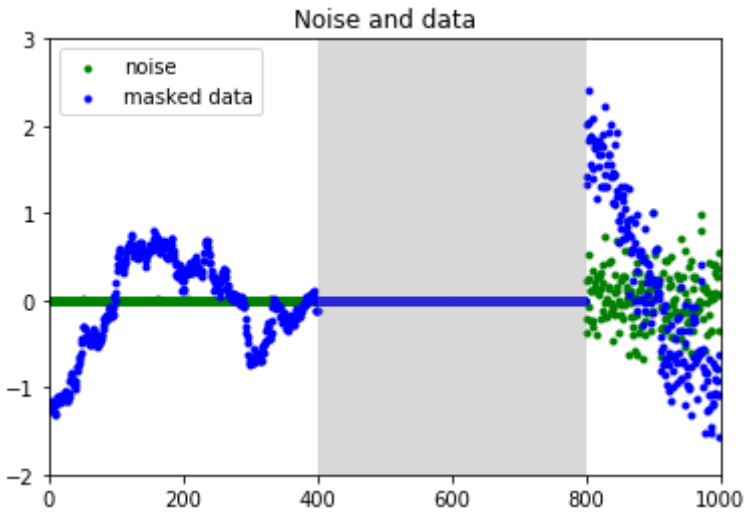


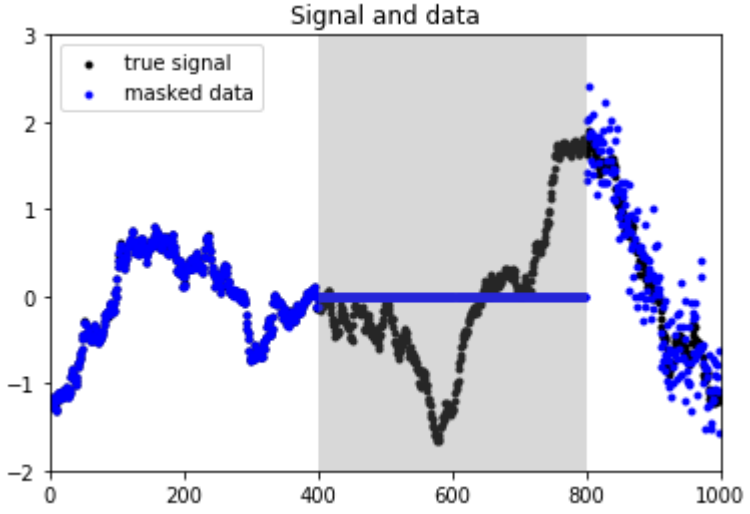
Bayesian denoising (Wiener filtering): example

- Generate signal and mock data

$$d = s + n$$



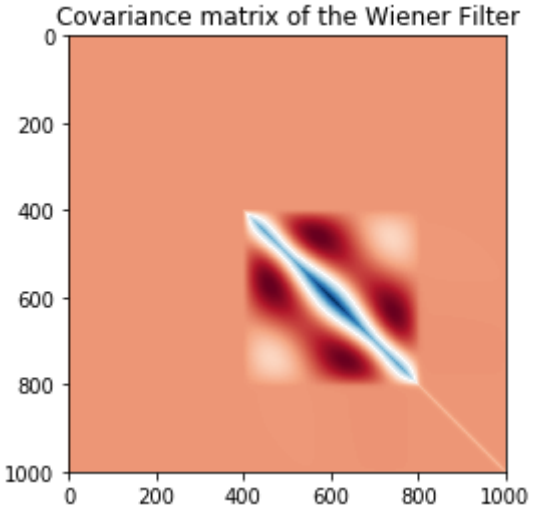




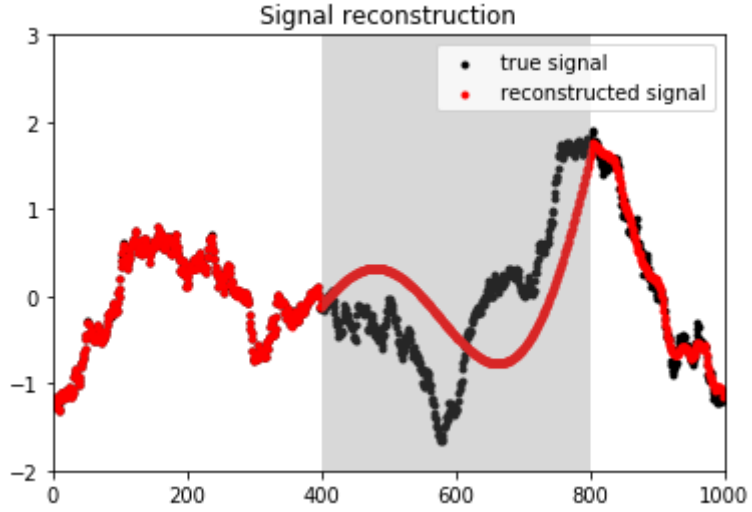
Bayesian denoising (Wiener filtering): example

- Perform Wiener filtering
- The mean of the reconstruction corresponds to the maximum a posteriori

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



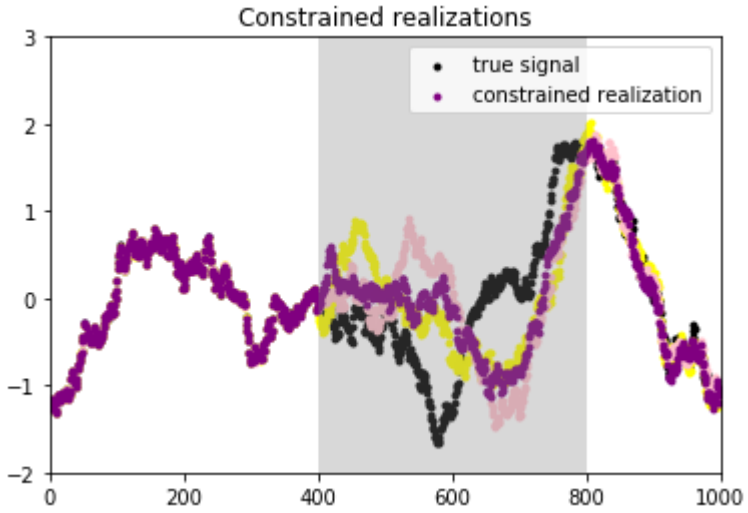
$$\mu_{s|d} = \mu_s + C_{s|d} N^{-1} (d - \mu_d)$$



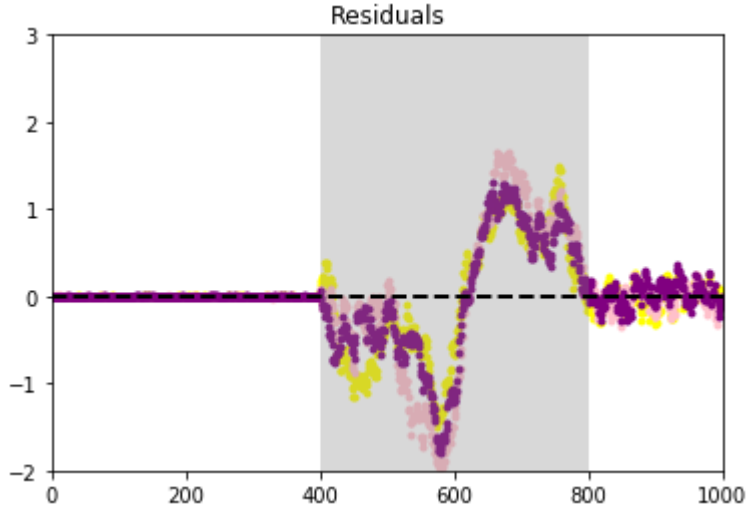
Bayesian denoising (Wiener filtering): example

- Draw constrained realisations of the denoised signal

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

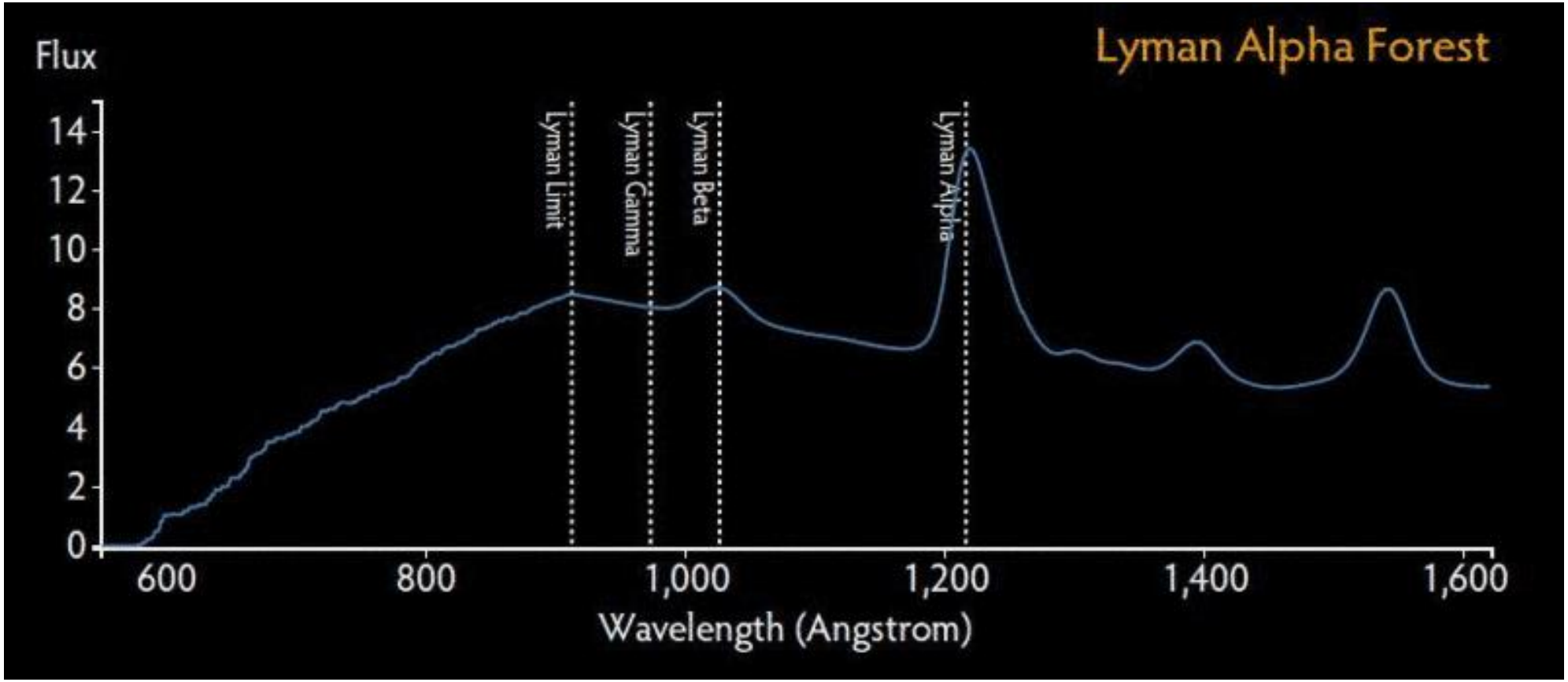


$$s_{\mathrm{sim}} - s_{\mathrm{true}}$$



Blending of signals: an astrophysical example

- Absorption of light in the spectrum of distant quasars:



Bayesian deblending (Wiener filtering)

Exercise: Bayesian deblending

- 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:

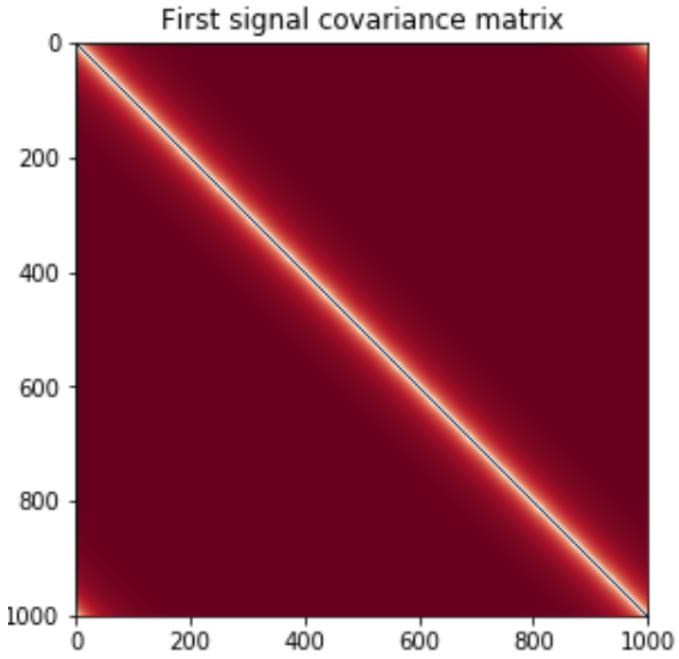
$$\mu_{x_1|d} = C_{x_1 d} C_{dd}^{-1} d$$
$$C_{x_1|d} = C_{x_1 x_1} - C_{x_1 d} C_{dd}^{-1} C_{d x_1}$$

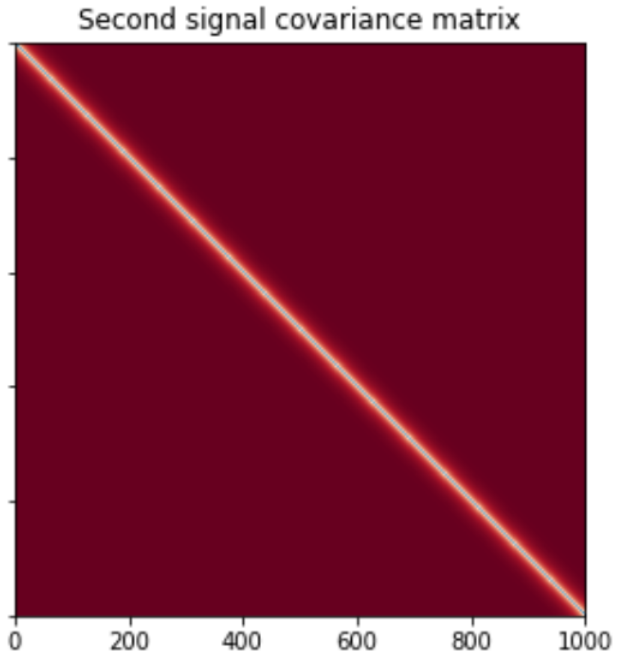
$$\mu_{x_2|d} = C_{x_2 d} C_{dd}^{-1} d$$
$$C_{x_2|d} = C_{x_2 x_2} - C_{x_2 d} C_{dd}^{-1} C_{d x_2}$$

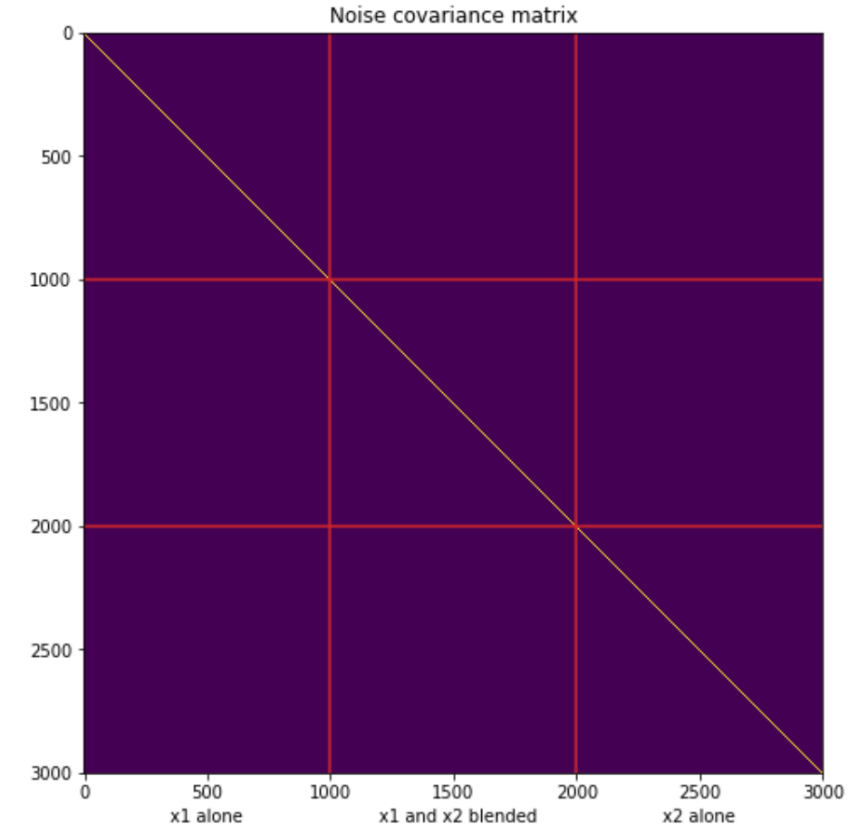
Bayesian deblending (Wiener filtering): example

Easy case: non-blended regions are not masked

- Setup first signal covariance, second signal covariance, and noise covariance matrices



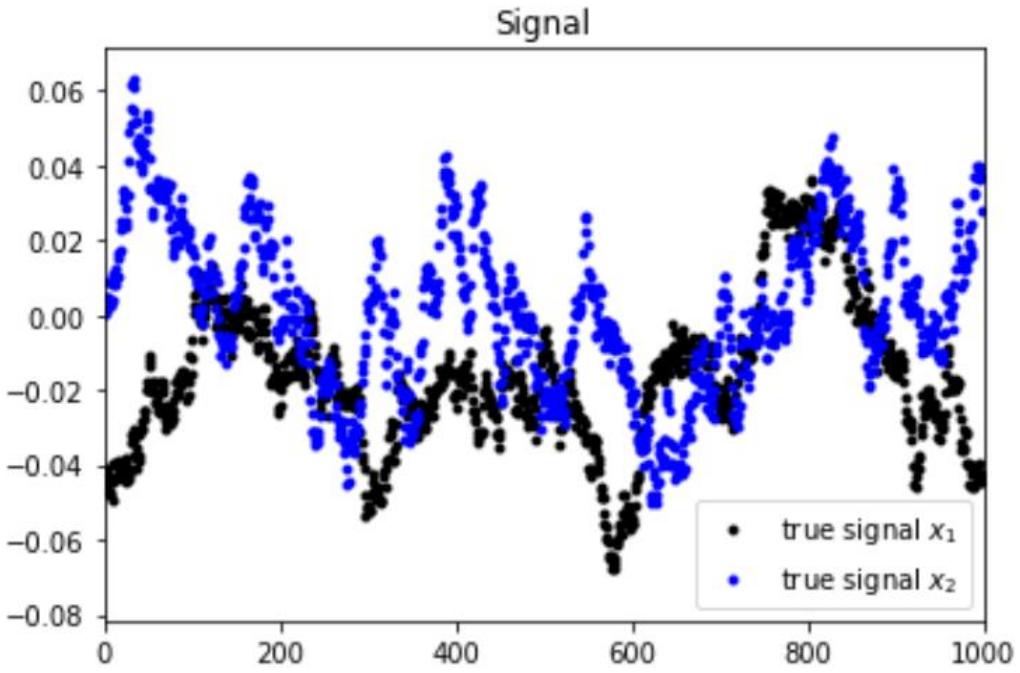


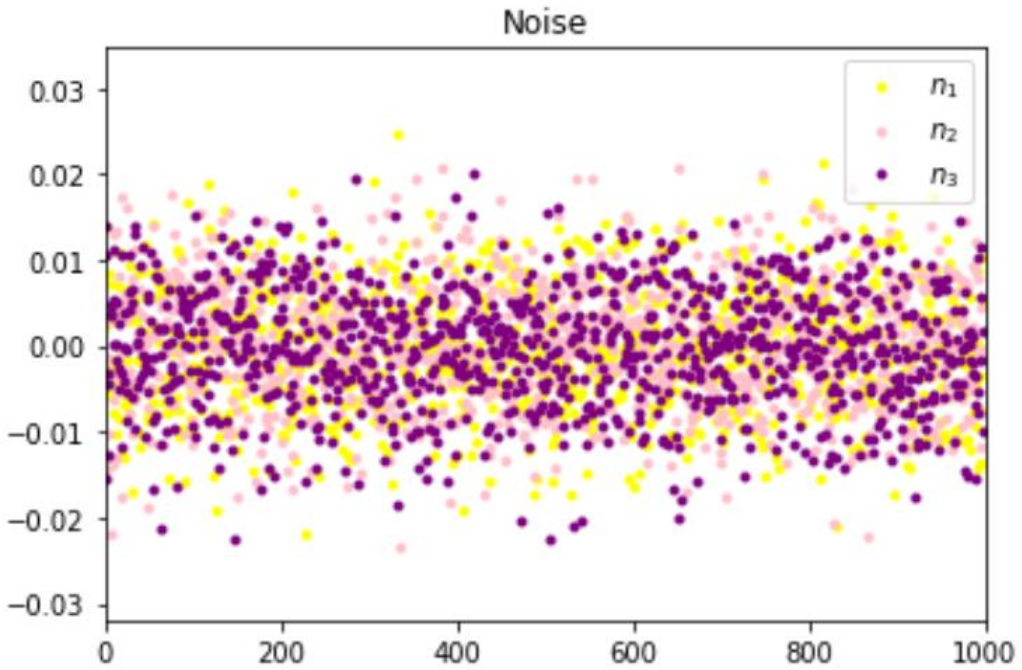


Bayesian deblending (Wiener filtering): example

Easy case: non-blended regions are not masked

- Generate signal and noise

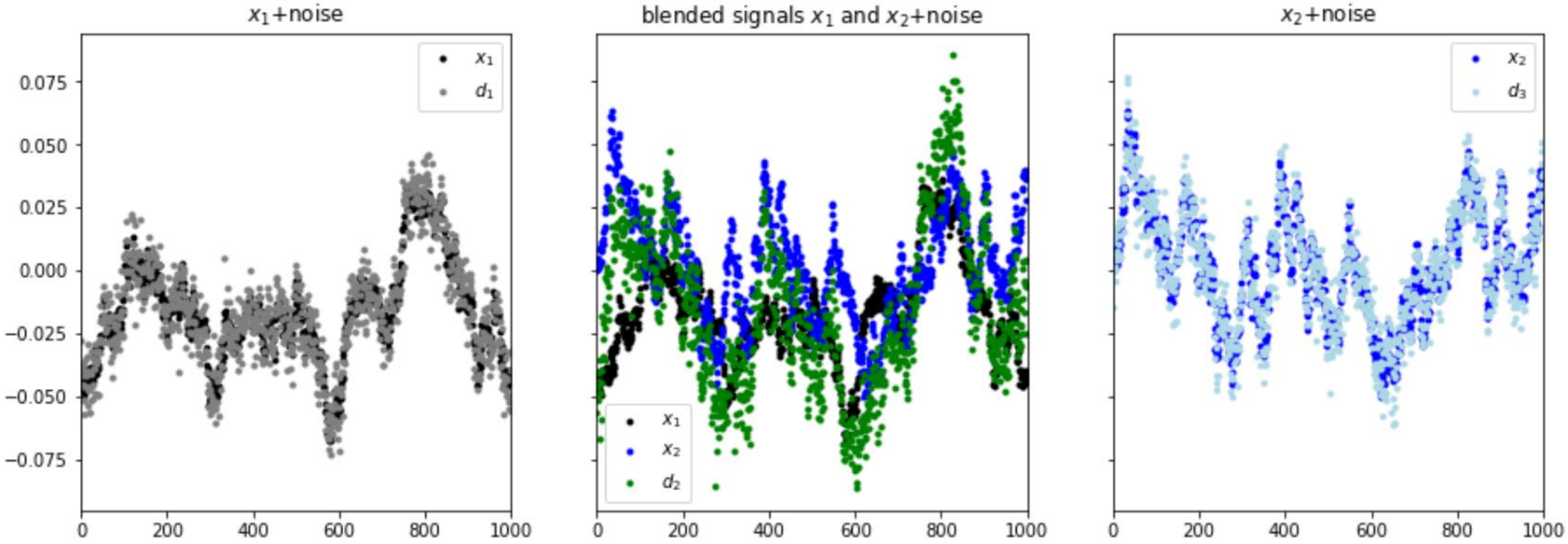




Bayesian deblending (Wiener filtering): example

Easy case: non-blended regions are not masked

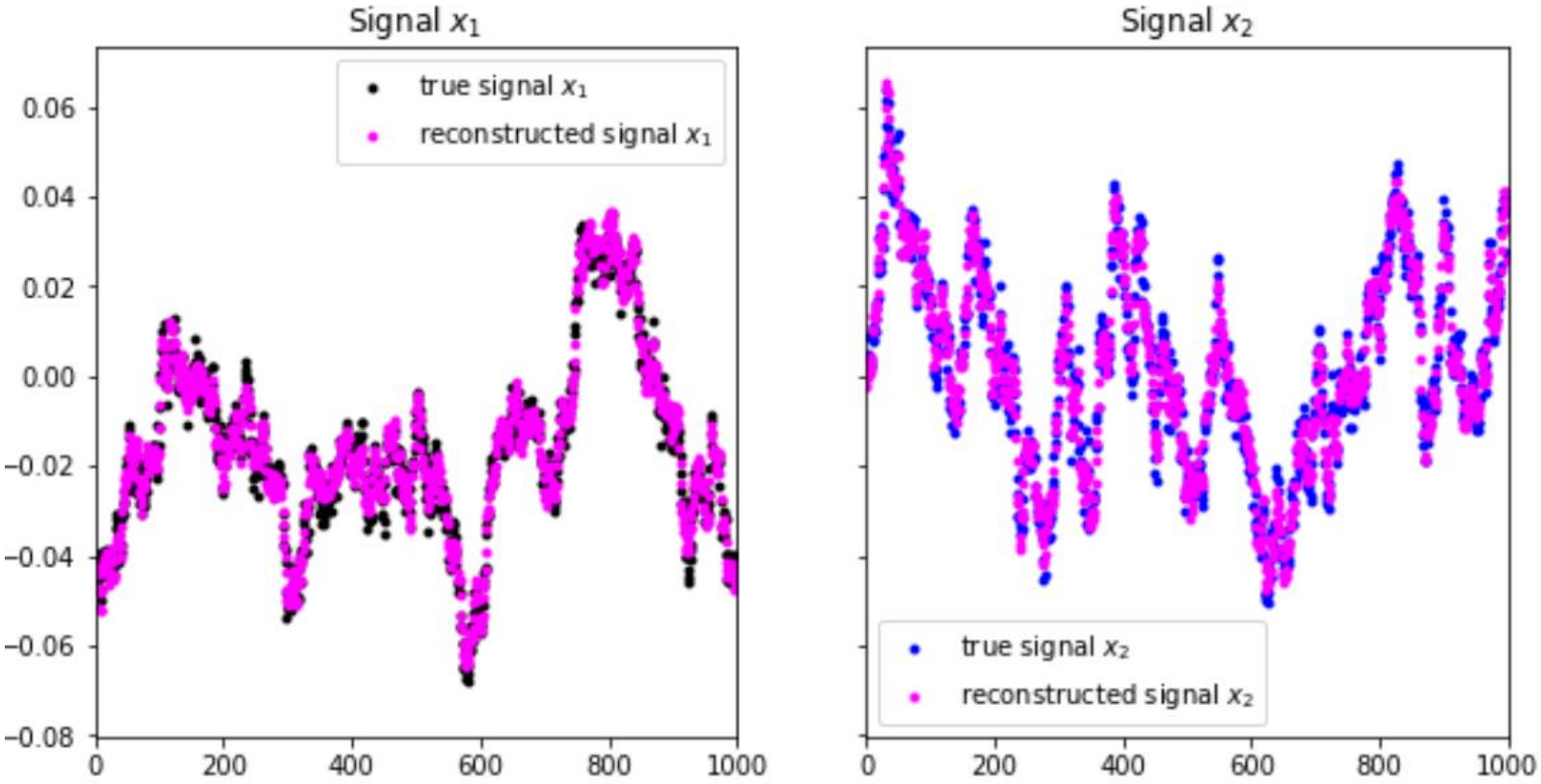
- Generate mock data



Bayesian deblending (Wiener filtering): example

Easy case: non-blended regions are not masked

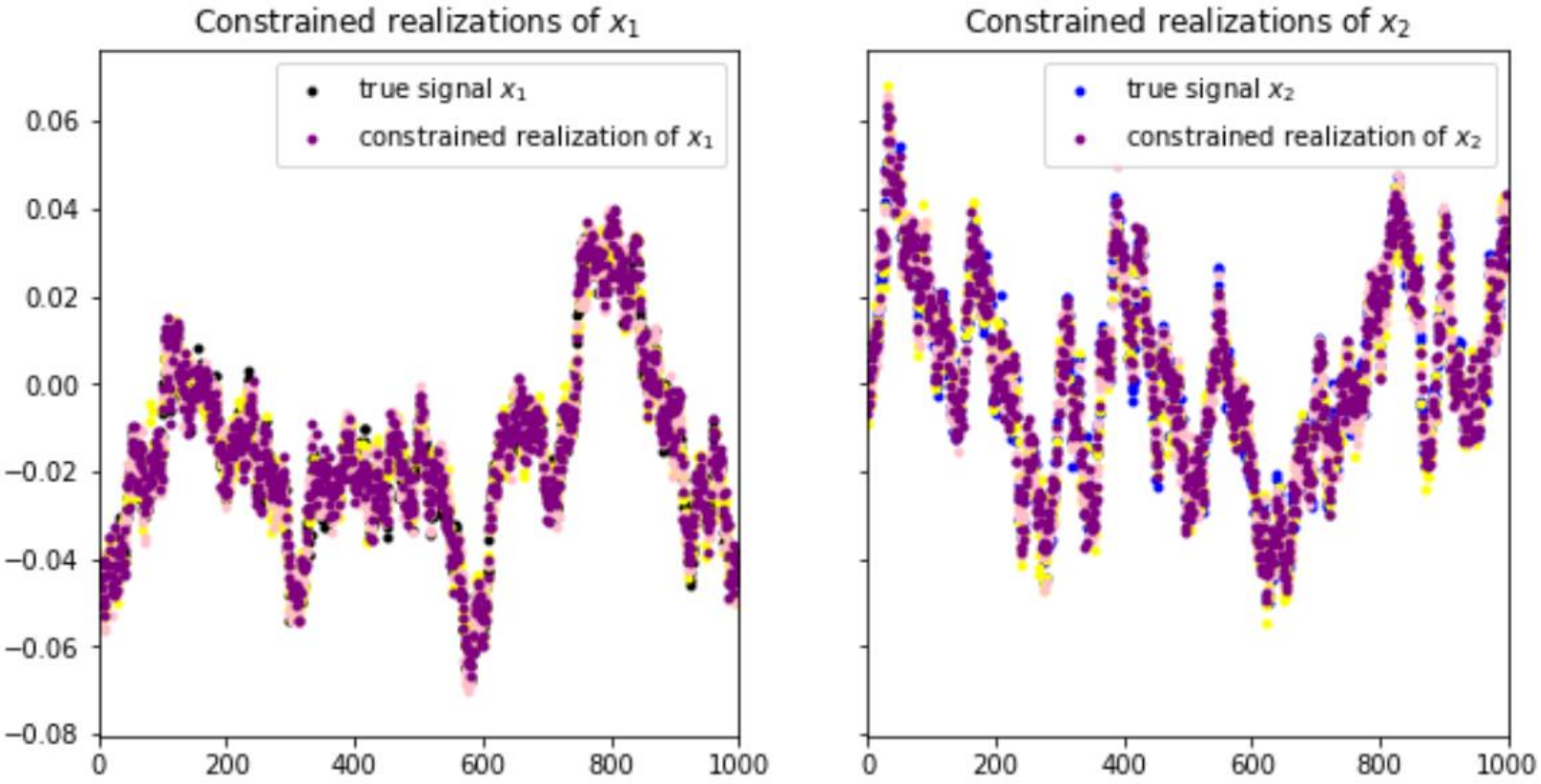
- Setup Wiener deblender and perform reconstruction of the two signals



Bayesian deblending (Wiener filtering): example

Easy case: non-blended regions are not masked

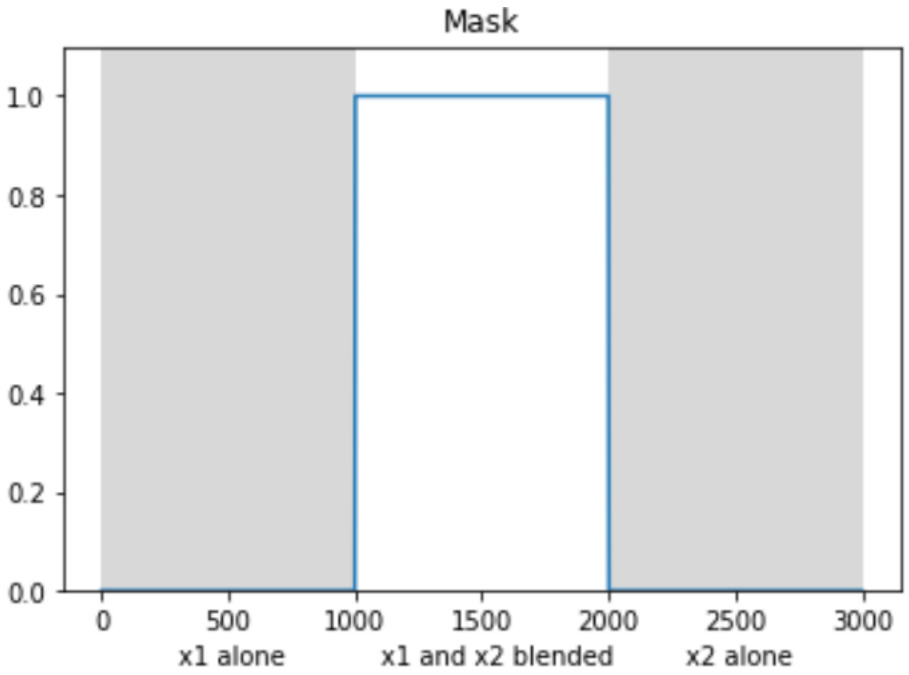
- Generate constrained realisations of the two signals

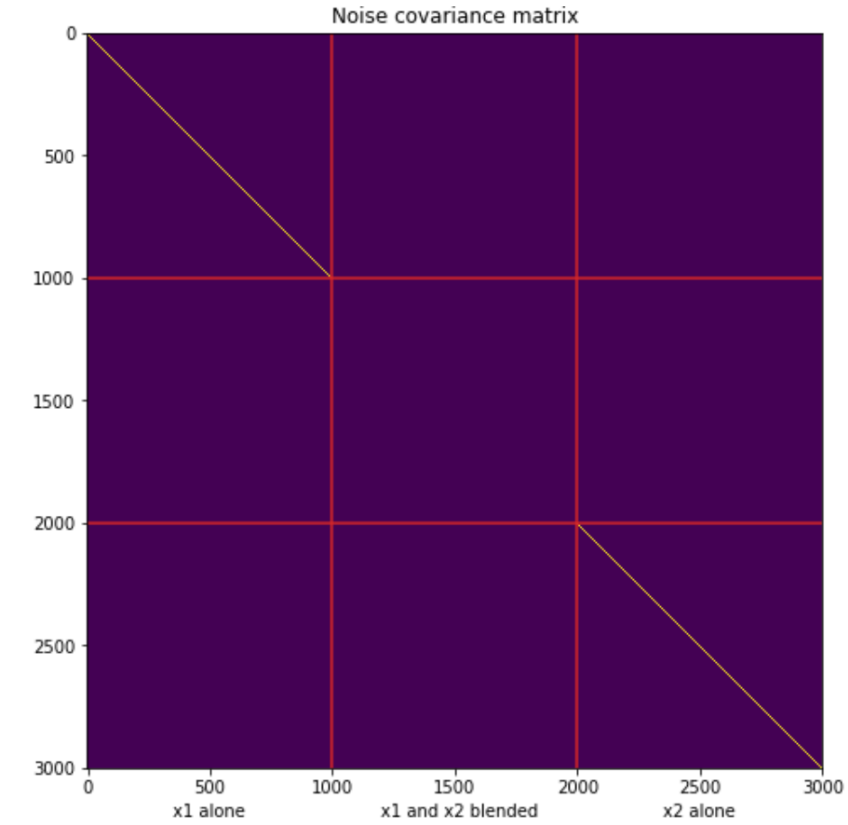


Bayesian deblending (Wiener filtering): example

Difficult case: non-blended regions are masked

- Setup noise covariance matrices (covariance matrices for the two signals stay the same)

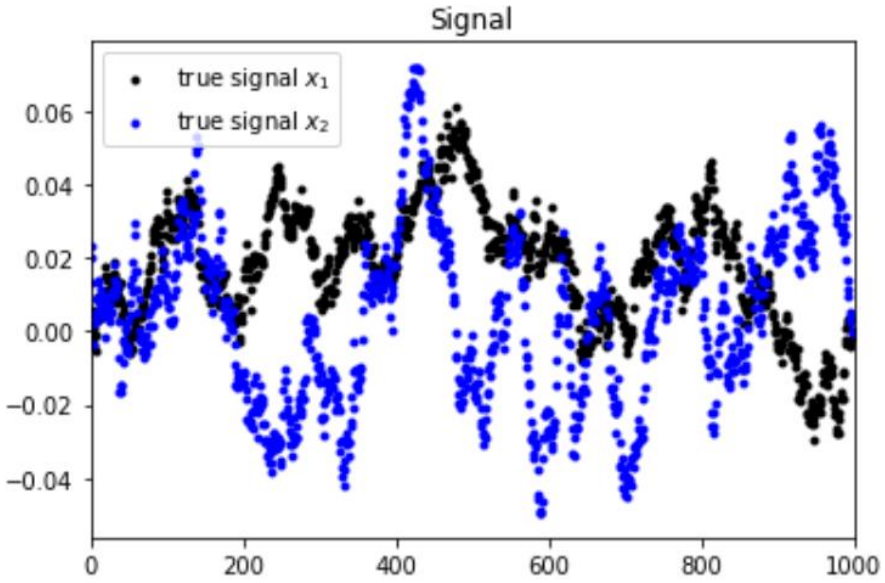


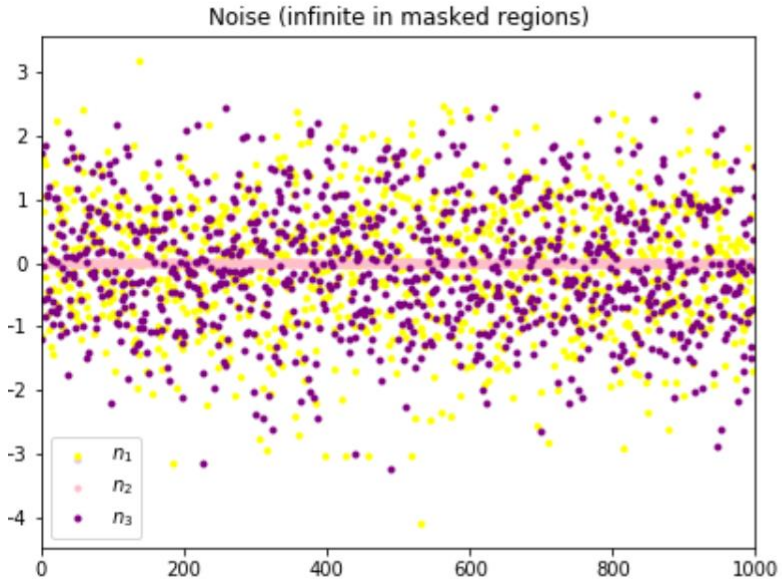


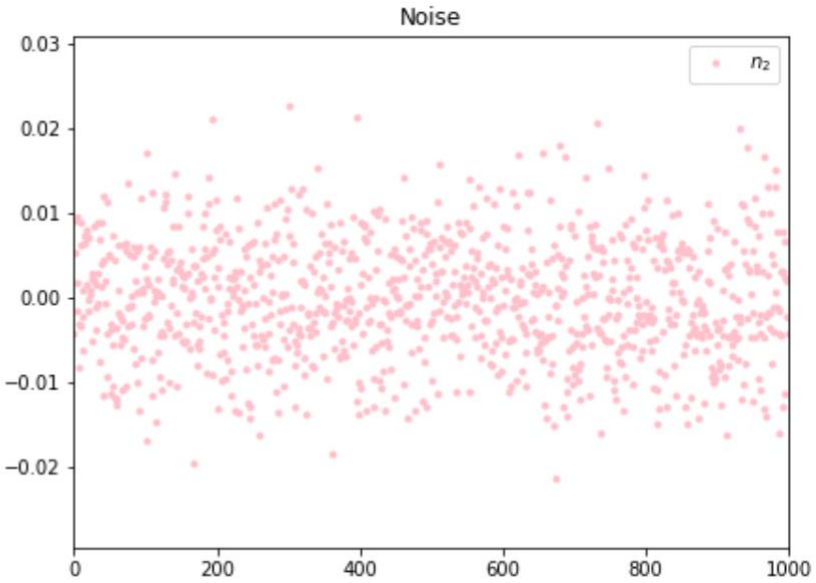
Bayesian deblending (Wiener filtering): example

Difficult case: non-blended regions are masked

- Generate signal and noise



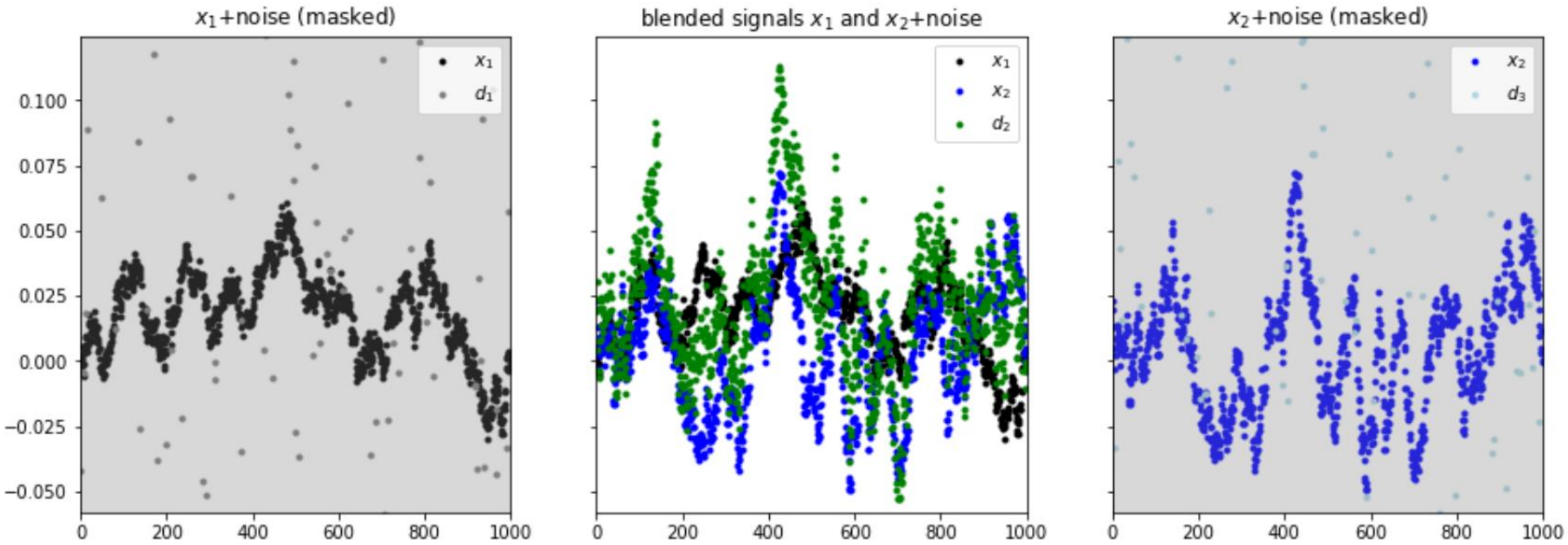




Bayesian deblending (Wiener filtering): example

Difficult case: non-blended regions are masked

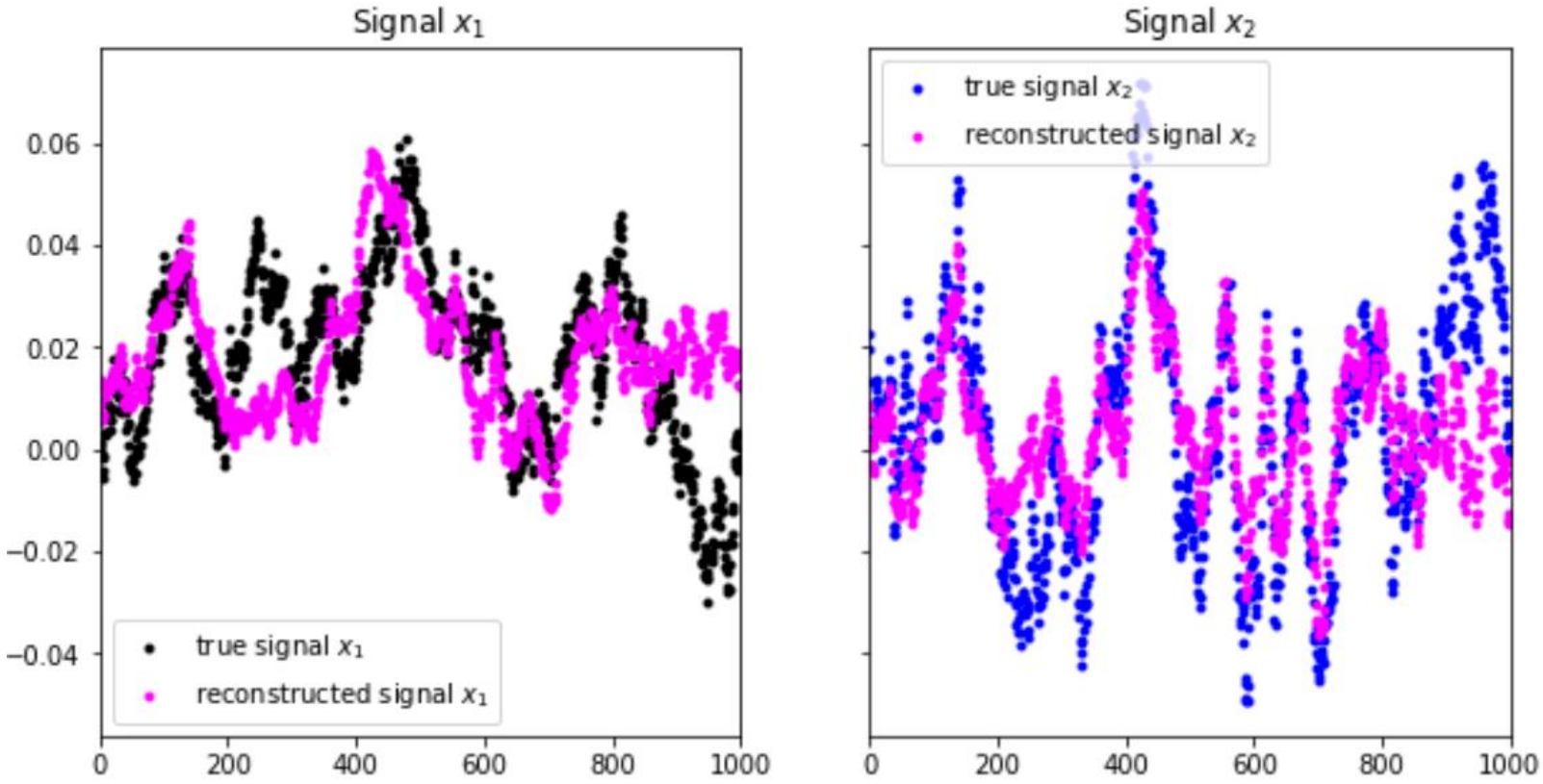
- Generate mock data



Bayesian deblending (Wiener filtering): example

Difficult case: non-blended regions are masked

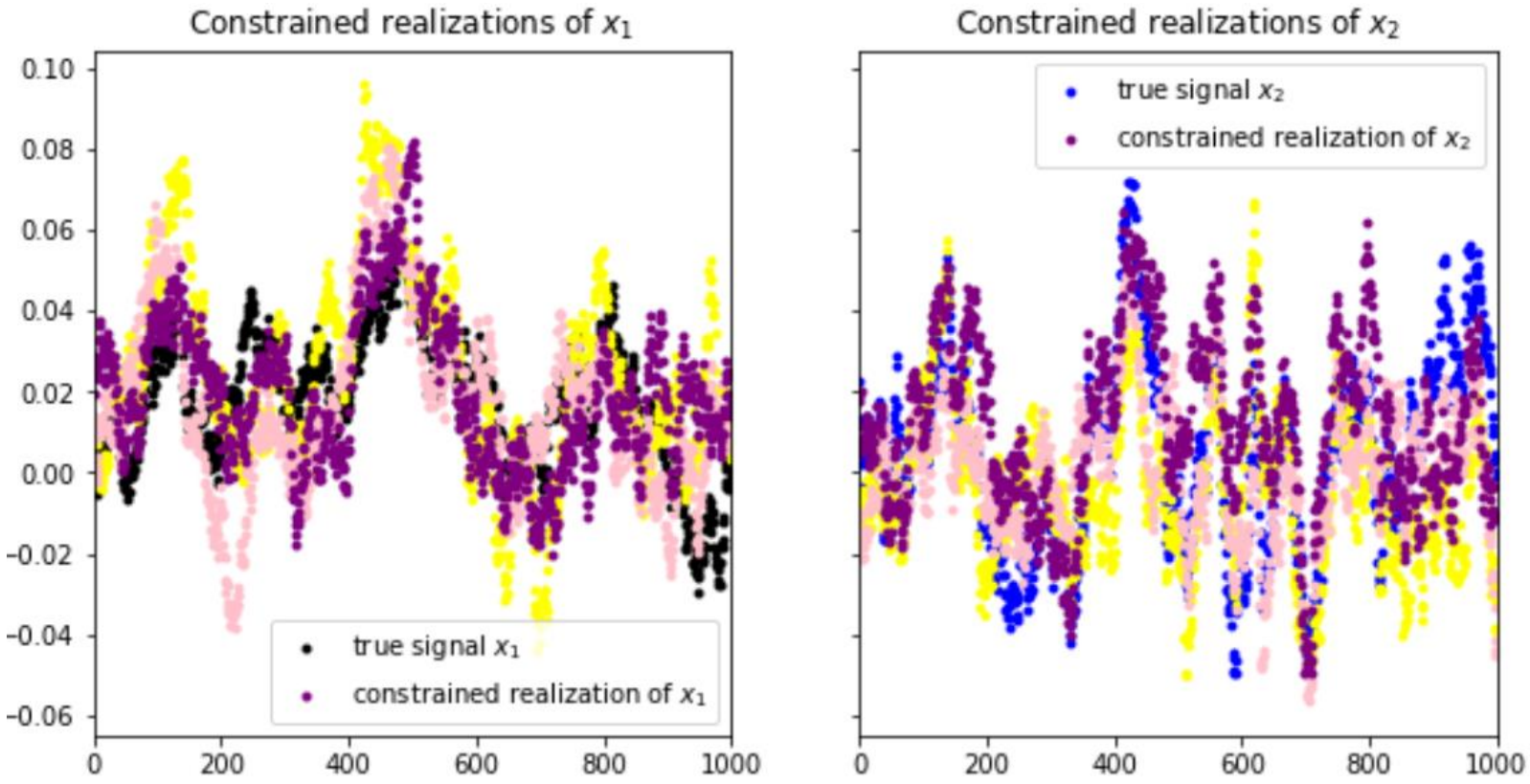
- Setup Wiener deblender and perform reconstruction of the two signals



Bayesian deblending (Wiener filtering): example

Difficult case: non-blended regions are masked

- Generate constrained realisations of the two signals



References and acknowledgements

References:

- S. B. McGrayne (2012), *The Theory That Would Not Die: How Bayes' Rule Cracked the Enigma Code, Hunted Down Russian Submarines, and Emerged Triumphant from Two Centuries of Controversy*
- E. T. Jaynes (2002), *Probability Theory: The Logic of Science*

- For their lectures, thanks to: Alan Heavens, Benjamin Wandelt

https://florent-leclercq.eu/teaching.php