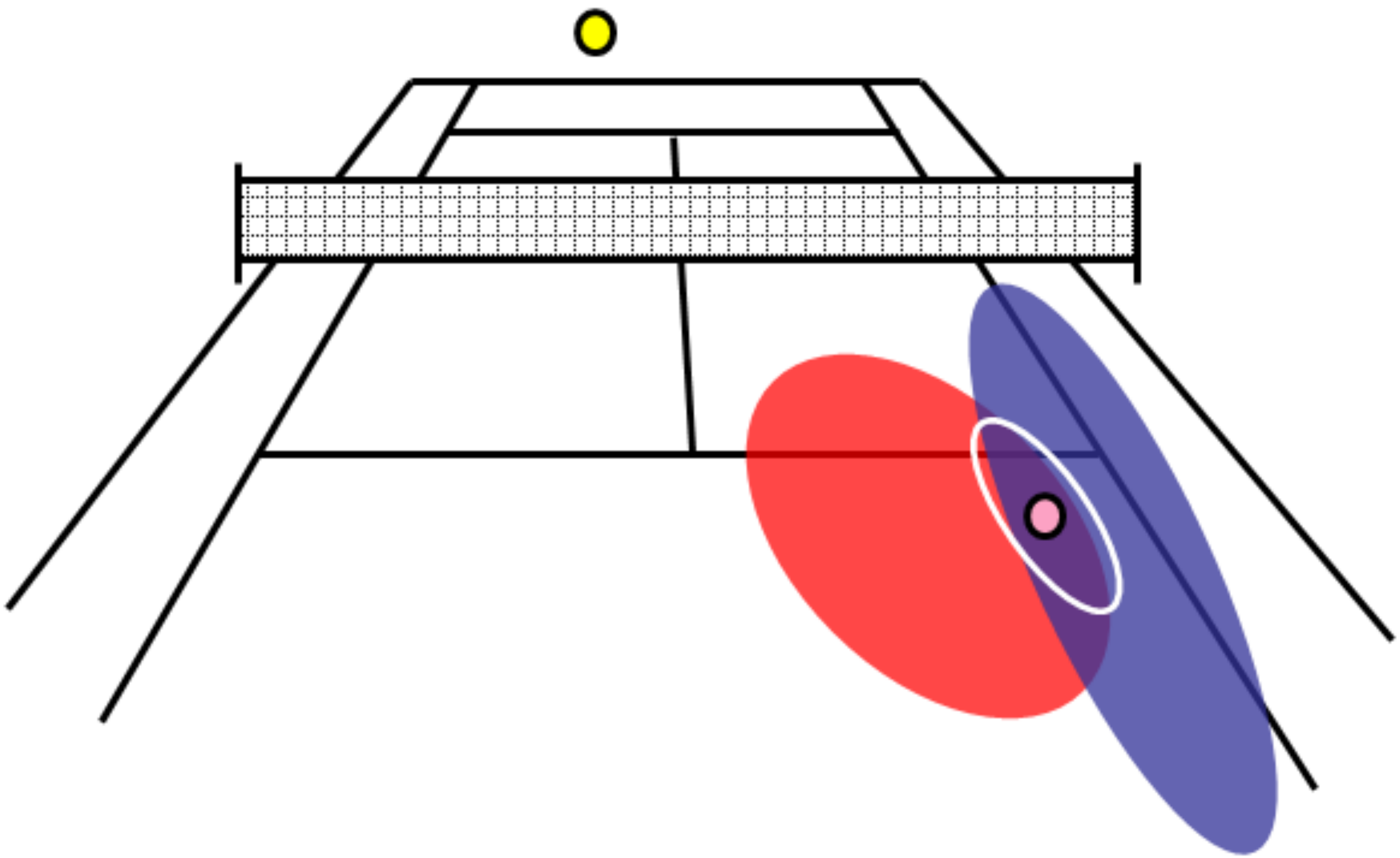
Basics of Bayesian analysis



Jorge Lillo-Box

MCMC Coffee | Season 1, Episode 5

Disclaimer

This presentation is based <u>heavily</u> on

"*Statistics, Data Mining, and Machine Learning in Astronomy: A Practical Python Guide for the Analysis of Survey Data: A Practical Python Guide for the Analysis of Survey Data*"

by

Ivezić, Connolly, Vanderplas, & Gray

Princeton University Press: 2014.

and it is only meant for internal purposes of the MCMC Coffee seassions. *Please excuse our plagiarism!*

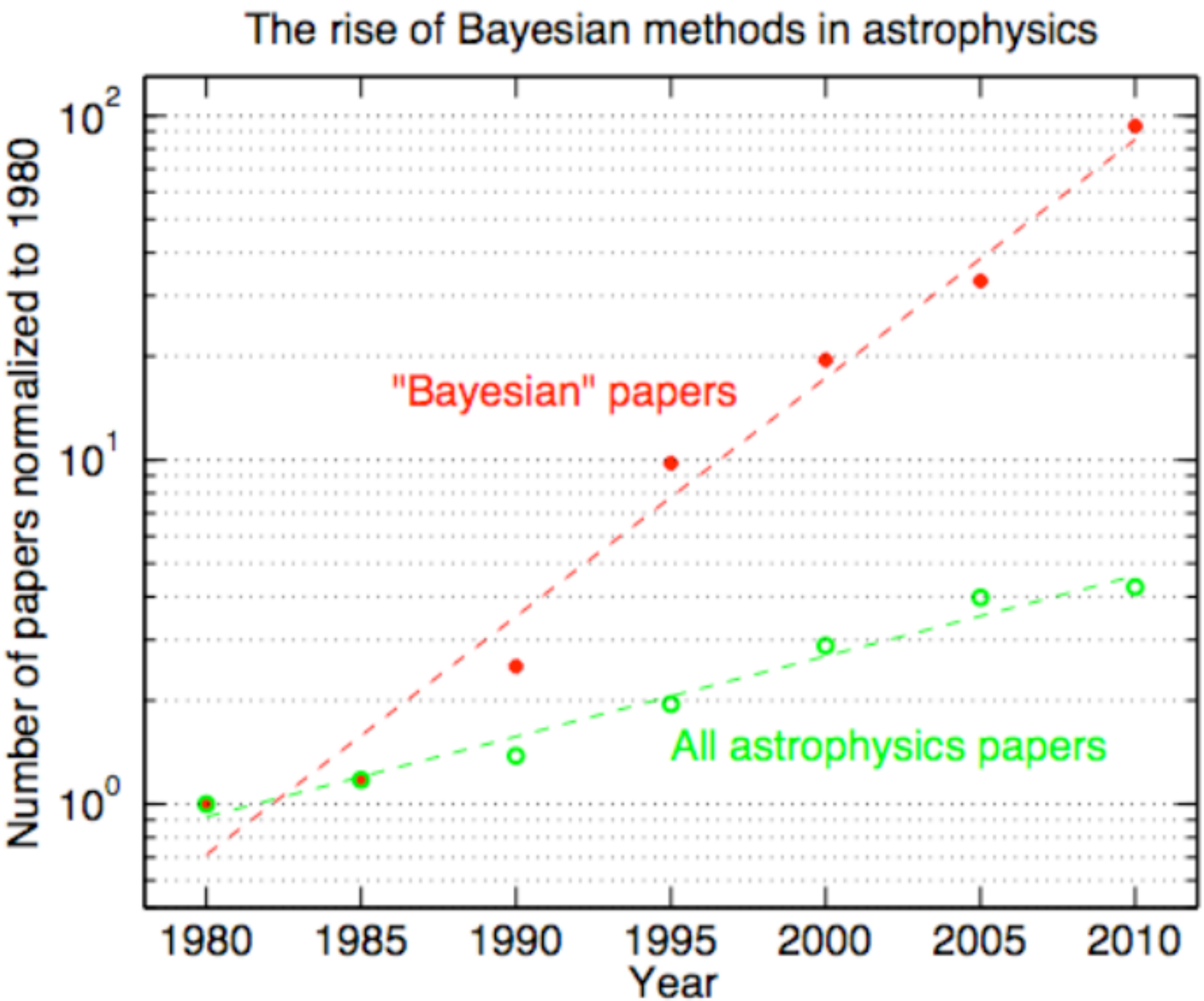
Very brief historical context

Reverend Thomas Bayes (1702 - 1761), a British amateur mathemetician, wrote about **how to combine an intial belief with new data to arrive at an improved belief**. This manuscript was published in 1763 (posthumously)

In 1774 Pierre Simon Lapalace rediscovered and clarified Bayes' principles. He applied these to astronomy, physics, population stats, jurisprudence (study of law). By the way, he estimated the mass of Saturn and its uncertainty, which remains consistent with the best current measurements.

In the early 20th century, Harold Jeffreys brought back Bayes' theorem and Laplace's work.

250 years later…



Source: Roberto Trotta (Imperial College)

The **principles** of Bayesian inference

- **Probability statements** are not limited to data, but can be made **for model parameters and models themselves**.

The **principles** of Bayesian inference

- **Probability statements** are not limited to data, but can be made **for model parameters and models themselves**.
- Inferences are made by producing **probability density functions** (PDFs)

The **principles** of Bayesian inference

- **Probability statements** are not limited to data, but can be made **for model parameters and models themselves**.
- Inferences are made by producing **probability density functions** (PDFs)
- Model **parameters** are treated as **random variables****

**A variable whose value results from the measurement of a quantity that is subject to random variations

The **principles** of Bayesian inference

- **Probability statements** are not limited to data, but can be made **for model parameters and models themselves**.
- Inferences are made by producing **probability density functions** (PDFs)
- Model **parameters** are treated as **random variables**
- Remember, Bayesian method yields optiumum results **assuming that all of the supplied info is correct(!!!)**.

The **principles** of Bayesian inference

	Classical statistics	Bayesian Methods
Likelihood function	Used to find model parameters that yield the **highest** data likelihood ↓ It **cannot** be interpreted as a probability density function for model parameters	Extends the concept of the data likelihood function by adding extra information to the analysis (so-called prior) ↓ One can assign PDF to model parameters

The **Bayes' rule**

$$p(y|x) = \frac{p(x|y)\,p(y)}{p(x)}$$

Practical exemple: the Monty Hall problem!

Section 3.1.3

The **Bayes' theorem**

Bayes' rule is not controversial

The frequentist vs. Bayesian controversy sets in when we apply Bayes' rule to the likelihood function $p(D|M)$ to obtain Bayes' theorem:

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

D = data
M = model

The **Bayes' theorem**

Bayes' rule is not controversial

The frequentist vs. Bayesian controversy sets in when we apply Bayes' rule to the likelihood function $p(D|M)$ to obtain Bayes' theorem:

Likelihood ***Prior***

Posterior probability

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

Evidence

D = data
M = model

The **Bayes' theorem**

Bayes' rule is not controversial

The frequentist vs. Bayesian controversy sets in when we apply Bayes' rule to the likelihood function $p(D|M)$ to obtain Bayes' theorem:

$$\underbrace{\boxed{p(M|D)}}_{\text{Posterior probability}} = \frac{\overbrace{\boxed{p(D|M)}}^{\text{Likelihood}}\ \overbrace{\boxed{p(M)}}^{\text{Prior}}}{\underbrace{\boxed{p(D)}}_{\text{Evidence}}}$$

D = data

M = model

An "improved belief" is proportional to the product of an "initial belief" and the probability that the "initial belief" generated the observed data.

The **Bayes' theorem**

Bayes' rule is not controversial

The frequentist vs. Bayesian controversy sets in when we apply Bayes' rule to the likelihood function $p(D|M)$ to obtain Bayes' theorem:

Likelihood **Prior**

Posterior probability

$$\boxed{p(M, \boldsymbol{\theta}|D, I)} = \frac{\boxed{p(D|M, \boldsymbol{\theta}, I)}\,\boxed{p(M, \boldsymbol{\theta}|I)}}{\boxed{p(D|I)}}$$

Evidence

D = data
M = model
$\theta = \theta_1, \ldots, \theta_k$ parameters of the model
I = any other information

An "improved belief" is proportional to the product of an "initial belief" and the probability that the "initial belief" generated the observed data.

The **Bayes' theorem**

Posterior probability

$$\boxed{p(M, \boldsymbol{\theta} | D, I)}$$

Posterior probability density function (**pdf**) for model M with parameters θ given the data D and the additional information I. It has **k+1 dimensions**

The **Bayes' theorem**

Posterior probability

$$\boxed{p(M, \boldsymbol{\theta} | D, I)}$$

Posterior probability density function (**pdf**) for model M with parameters θ given the data D and the additional information I. It has **k+1 dimensions**

Likelihood

$$\boxed{p(D | M, \boldsymbol{\theta}, I)}$$

The *likelihood* of data *given* some model *M* and given some **fixed values** of parameters θ describing it, and all other prior information *I* .

The **Bayes' theorem**

Posterior probability

$$\boxed{p(M, \boldsymbol{\theta} | D, I)}$$

Posterior probability density function (**pdf**) for model M with parameters θ given the data D and the additional information I. It has **k+1 dimensions**

Likelihood

$$\boxed{p(D | M, \boldsymbol{\theta}, I)}$$

The *likelihood* of data *given* some model *M* and given some **fixed values** of parameters θ describing it, and all other prior information *I* .

Prior

$$\boxed{p(M, \boldsymbol{\theta} | I)}$$

A *priori* joint probability for model *M* and its parameters θ in the absence of any of the data used to compute likelihood

The **Bayes' theorem**

Prior

$$\boxed{p(M, \boldsymbol{\theta} | I)}$$

A *priori* joint probability for model *M* and its parameters θ in the absence of any of the data used to compute likelihood

$$p(M, \boldsymbol{\theta} | I) = \underbrace{p(\boldsymbol{\theta} | M, I)}_{\text{Only this term is needed for parameter estimation}} \; \underbrace{p(M | I)}_{\text{Also needed for } \textit{model selection}}$$

How to **choose a prior**

Informative priors: are based on previous measurements

(e.g., when analyzing the radial velocity of a star modulated by a transiting planet, you can constrain the **period** of the planet very well from the transit periodicity as P +/- e_P)

$$RV = V_{sys} + K^{*}\cos(2\pi t/P)$$

Uninformative priors: when no other information other than the data we are analyzing is available (but note!: it can incorporate weak but objective information)

(e.g., the systemic radial velocity of the star Vsys)

How to **choose a prior**

Rules for selecting uninformative priors

➡ **Principle of indifference:** states that a set of basic, mutually exclusive possibilities need to be assigned equal probabilities (e.g., for a fair six-sided die, each of the outcomes has a prior probability of 1/6)

➡ **Principle of consistency:** based on transformation groups, demands that the prior for **a location parameter should not change with translations** of the coordinate system, and yields a *flat prior*. The prior for a scale parameter should not depend on the choice of units. The solution is $p(\sigma|I) \propto \sigma^{-1}$ (or a flat prior for $\ln\sigma$), called a *scale-invariant* prior.

➡ **Principle of maximum entropy:** when we have additional weak prior information about some parameter, such as a low-order statistic