BAYESIAN METHODS IN COSMOLOGY

Edited by Michael P. Hobson, Andrew H. Jaffe, Andrew R. Liddle, Pia Mukherjee and David Parkinson

$$P(\theta|D) = \frac{P(D|\theta)P(\theta)}{P(D)}$$

CAMBRIDGE

CAMBRIDGE www.cambridge.org/9780521887946

This page intentionally left blank

BAYESIAN METHODS IN COSMOLOGY

In recent years, cosmologists have advanced from largely qualitative models of the Universe to precision modelling using Bayesian methods in order to determine the properties of the Universe to high accuracy. This timely book is the only comprehensive introduction to the use of Bayesian methods in cosmological studies, and is an essential reference for graduate students and researchers in cosmology, astrophysics and applied statistics.

The first part of the book focuses on methodology, setting the basic foundations and giving a detailed description of techniques. It covers topics including the estimation of parameters, Bayesian model comparison, and separation of signals. The second part explores a diverse range of applications, from the detection of astronomical sources (including through gravitational waves), to cosmic microwave background analysis and the quantification and classification of galaxy properties. Contributions from 24 highly regarded cosmologists and statisticians make this an authoritative guide to the subject.

MICHAEL P. HOBSON is Reader in Astrophysics and Cosmology at the Cavendish Laboratory, University of Cambridge, where he researches theoretical and observational cosmology, Bayesian statistical methods, gravitation and theoretical optics.

ANDREW H. JAFFE is Professor of Astrophysics and Cosmology at Imperial College, London, and a member of the Planck Surveyor Satellite collaboration, which will create the highest-resolution and most sensitive maps of the cosmic microwave background ever produced.

ANDREW R. LIDDLE is Professor of Astrophysics at the University of Sussex. He is the author of over 150 journal articles and four books on cosmology, covering topics from early Universe theory to modelling astrophysical data.

PIA MUKHERJEE is a Postdoctoral Research Fellow in the Astronomy Centre at the University of Sussex, specializing in constraining cosmological models, including dark energy models, from observational data.

DAVID PARKINSON is a Postdoctoral Research Fellow in the Astronomy Centre at the University of Sussex, working in the areas of cosmology and the early Universe.

BAYESIAN METHODS IN COSMOLOGY

MICHAEL P. HOBSON
Cavendish Laboratory, University of Cambridge

ANDREW H. JAFFE
Imperial College, London

ANDREW R. LIDDLE
University of Sussex

PIA MUKHERJEE
University of Sussex

DAVID PARKINSON
University of Sussex

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE UNIVERSITY PRESS
Cambridge, New York, Melbourne, Madrid, Cape Town, Singapore,
São Paulo, Delhi, Dubai, Tokyo

Cambridge University Press
The Edinburgh Building, Cambridge CB2 8RU, UK

Published in the United States of America by Cambridge University Press, New York

www.cambridge.org
Information on this title: www.cambridge.org/9780521887946

First published in print format 2009

ISBN-13 978-0-511-76400-4 eBook (Adobe Reader)

ISBN-13 978-0-521-88794-6 Hardback

Contents

Contributors

Filipe B. Abdalla
Department of Physics and Astronomy,
University College London,
Gower Street,
London WC1E 6BT, UK

Stefano Andreon
INAF–Osservatorio Astronomico di
Brera via Brera 28, 20121 Milano, Italy

M. A. J. Ashdown
Astrophysics Group, Cavendish
Laboratory, JJ Thomson Avenue,
Cambridge CB3 0HE, UK

Manda Banerji
Department of Physics and Astronomy,
University College London,
Gower Street, London WC1E 6BT, UK

Sarah Bridle
Department of Physics and Astronomy,
University College London,
Gower Street, London WC1E 6BT, UK

Neil Cornish
Department of Physics, Montana
State University, Bozeman,
MT 597717, USA

Martin A. Hendry
Department of Physics and Astronomy,
University of Glasgow,
Glasgow G12 8QQ, UK

M. P. Hobson
Astrophysics Group, Cavendish
Laboratory, JJ Thomson Avenue,
Cambridge CB3 0HE, UK

Andrew H. Jaffe
Astrophysics Group, Imperial College
London, Blackett Laboratory,
London SW7 2AZ, UK

Martin Kunz
Astronomy Centre, University of
Sussex, Brighton BN1 9QH, UK

Ofer Lahav
Department of Physics and Astronomy,
University College London, Gower Street,
London WC1E 6BT, UK

Antony Lewis
Institute of Astronomy and Kavli
Institute for Cosmology,
Madingley Road,
Cambridge CB3 0HA, UK

Andrew R. Liddle
Astronomy Centre, University of Sussex, Brighton BN1 9QH, UK

Thomas J. Loredo
Department of Astronomy, Cornell University, Ithaca, NY 14853, USA

Daniel Mortlock
Astrophysics Group, Imperial College London, Blackett Laboratory, London SW7 2AZ, UK

Pia Mukherjee
Astronomy Centre, University of Sussex, Brighton BN1 9QH, UK

David Parkinson
Astronomy Centre, University of Sussex, Brighton BN1 9QH, UK

Steve Rawlings
Astrophysics, Department of Physics, Oxford University, Keble Road, Oxford OX1 3RH, UK

Graça Rocha
California Institute of Technology, 1200 East California Boulevard, Pasadena, CA 91125, USA

Richard S. Savage
Astronomy Centre, University of Sussex, Brighton BN1 9QH, UK, and Systems Biology Centre, University of Warwick, Coventry CV4 7AL, UK

D. S. Sivia
St John's College, St. Giles, Oxford OX1 3JP, UK

John Skilling
Maximum Entropy Data Consultants Ltd, Kenmare, County Kerry, Ireland

V. Stolyarov
Astrophysics Group, Cavendish Laboratory, JJ Thomson Avenue, Cambridge CB3 0HE, UK

Roberto Trotta
Astrophysics, Department of Physics, Oxford University, Keble Road, Oxford OX1 3RH, UK
and Astrophysics Group, Imperial College London, Blackett Laboratory, London SW7 2AZ, UK

Preface

A revolution is underway in cosmology, with largely qualitative models of the Universe being replaced with precision modelling and the determination of Universe's properties to high accuracy. The revolution is driven by three distinct elements – the development of sophisticated cosmological models and the ability to extract accurate predictions from them, the acquisition of large and precise observational datasets constraining those models, and the deployment of advanced statistical techniques to extract the best possible constraints from those data.

This book focuses on the last of these. In their approach to analyzing datasets, cosmologists for the most part lie resolutely within the Bayesian methodology for scientific inference. This approach is characterized by the assignment of probabilities to all quantities of interest, which are then manipulated by a set of rules, amongst which Bayes' theorem plays a central role. Those probabilities are constantly updated in response to new observational data, and at any given instant provide a snapshot of the best current understanding. Full deployment of Bayesian inference has only recently come within the abilities of high-performance computing.

Despite the prevalence of Bayesian methods in the cosmology literature, there is no single source which collects together both a description of the main Bayesian methods and a range of illustrative applications to cosmological problems. That, of course, is the aim of this volume. Its seeds grew from a small conference 'Bayesian Methods in Cosmology', held at the University of Sussex in June 2006 and attended by around 60 people, at which many cosmological applications of Bayesian methods were discussed. CUP editor Vince Higgs, who attended the conference, saw the need for a comprehensive volume covering these topics, and suggested that we put together an edited volume of articles. And here it is!

The book is divided into two part. The first part, 'Methods', concentrates on the formalism, methods and algorithms, with only limited illustrative examples. The focus is very much on those aspects that have proven valuable in cosmological

studies, complementing the much more complete treatments of Bayesian inference given in the excellent books by MacKay (2003; see page 35), Gregory (2005; see page 97), and Sivia and Skilling (2006; see page 35), all of which we recommend to the interested reader. The second part, 'Applications', studies a wide range of cosmological applications in detail. Many of the codes used in these applications are publicly available.

Part I

Methods

1

Foundations and algorithms

John Skilling

Why and how – simply – that's what this chapter is about.

1.1 Rational inference

Rational inference is important. By helping us to understand our world, it gives us the predictive power that underlies our technical civilization. We would not function without it. Even so, rational inference only tells us *how* to think. It does not tell us *what* to think. For that, we still need the combination of creativity, insight, artistry and experience that we call intelligence.

In science, perhaps especially in branches such as cosmology, now coming of age, we invent models designed to make sense of data we have collected. It is no accident that these models are formalized in mathematics. Mathematics is far and away our most developed logical language, in which half a page of algebra can make connections and predictions way beyond the precision of informal thought. Indeed, one can hold the view that frameworks of logical connections are, by definition, mathematics. Even here, though, we do not find absolute truth. We have conditional implication: 'If axiom, then theorem' or, equivalently, 'If not theorem, then not axiom'. Neither do we find absolute truth in science.

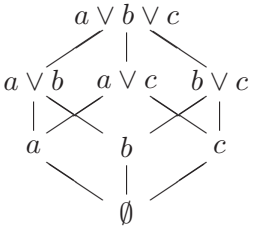
Our question in science is not 'Is this hypothetical model true?', but '*Is this model better than the alternatives?*'. We could not recognize absolute truth even if we stumbled across it, for how could we tell? Conversely, we cannot recognize absolute falsity. If we believe dogmatically enough in a particular view, then no amount of contradictory data will convince us otherwise, if only because the data could be dismissed as evidence of conspiracy to deceive. Yet even a determined sceptic might be sufficiently charitable to acknowledge that a model with demonstrable ability to predict future effects could have practical value.

Let us, then, avoid the philosophical minefields of belief and truth, and pay attention to what we really need, which is predictive ability. We anticipate the Sun will rise tomorrow, not just because it always has done so far, but because this is predicted by models of stellar structure and planetary dynamics, which accord so well with such a variety of data that perceived failure of the Sun to rise might more likely be hallucination.

Rational assessment of different models is the central subject of Bayesian methods, so called after Revd Thomas Bayes, the eighteenth century clergyman generally associated with the beginnings of formal probability theory. We will find that probability calculus is forced upon us as the only method which lets us learn from data irrespective of their order – surely a required symmetry. We will also discover how to use it properly, with the aid of modern computers and algorithms. Inference was held back for a century by technical inability to do the required sums, but that sad era has closed.

1.2 Foundations

Suppose we are given a choice of basic models, $a = apple$, $b = banana$, $c = cherry$, which purport to explain some data. Such models can be combined, so that $apple\text{-}\mathtt{OR}\text{-}banana$, written $a \vee b$, is also meaningful. Data that excluded cherries would, in fact, bring us down from the original $apple\text{-}\mathtt{OR}\text{-}banana\text{-}\mathtt{OR}\text{-}cherry$ combination to just that choice. With n basic models, there are 2^n possible combinations. They form the elements of a lattice, ranging from the absurdity in which none of the models is allowed, up to the provisional truism in which all of them remain allowed. In inference, we need to be able to navigate these possibilities as we refine our knowledge.



The absurdity $\emptyset$ is introduced merely because analysis is cleaner with it than without it, rather as 0 is often included with the positive integers.

The three core concepts of *measure*, *information* and *probability* all have wider scope than inference alone. They apply to lattices in general, whether or not the lattice fills out all 2^n possibilities. By exposing just the foundation that we need, and no more, we can allow wider application, as well as clarifying the basis so that

alternative formulations of these concepts become even less plausible than they may have been before.

1.2.1 Lattices

The critical idea we need is 'partial ordering'. We always have "=": every element equals itself, $x = x$. Sometimes, we have "<", as in $x < y$, meaning that y includes x. In inference, we say that *apple* is included in *apple*–OR–*banana*, because the scope of the latter is wider and includes all of the former, but we would not try to include *apple* within *banana*. We don't need that particular motivation, though. All we need is "<" in the abstract. Ordering is to be transitive,

$$x < y \text{ and } y < z \text{ implies } x < z, \tag{1.1}$$

otherwise it would not make sense.

The other idea we need is 'least upper bound'. The upper bounds to elements x and y are those elements at or including both x and y. If there is a least such bound, we write it as $x \vee y$ and call it the least upper bound:

$$\left\{ \begin{array}{l} x \leq x \vee y \\ y \leq x \vee y \end{array} \right\}, \text{ and } x \vee y \leq u \text{ for all } u \text{ obeying } \left\{ \begin{array}{l} x \leq u \\ y \leq u \end{array} \right\}. \tag{1.2}$$

In inference, the unique least upper bound $x \vee y$ is that element including all the components of x and y, but no more. There, the existence of least upper bound is obvious.

Technically, a *lattice* is a partially ordered set with least upper bound, so that "<" and "$\vee$" are defined. Any pair of elements x and y also has lower bounds, being all those elements at or beneath both. There is a unique greatest lower bound, written $x \wedge y$. (If there were alternatives u and v, then $u \vee v$ could be ambiguously x or y, contradicting uniqueness of their least upper bound.) Mathematicians (Klain & Rota 1997) traditionally define a lattice in terms of $\vee$ and $\wedge$, but our use of $<$ and $\vee$ is equivalent, and (with $=$) underlies their traditional axioms of reflexivity, antisymmetry, transitivity, idempotency, commutativity, associativity and absorption. Of these, the associativity property

$$(x \vee y) \vee z = x \vee (y \vee z) \tag{1.3}$$

is of particular importance to us.

What we now seek is a numerical *valuation* $v(x)$ on our lattice of models, so that we can rank the possibilities. Remarkably, there is only one way of conforming to lattice structure, and this leads us to measure theory, thence to information and probability. Though modernized following Knuth (2003), the approach dates back to Cox (1946, 1961).

1.2.2 Measure

Addition

For a start, we want valuations to conform to "$\leq$", so we require

$$x \leq y \quad \Longrightarrow \quad v(x) \leq v(y). \tag{1.4}$$

Moreover, whatever our valuations were originally, we can shift them to give a standard value 0 to the ubiquitous absurdity $\emptyset$, so that the range of value becomes $0 = v(\emptyset) \leq v(x)$.

We next assume that if x and y are disjoint, so that $x \wedge y = \emptyset$ and they have nothing in common, then the valuation $v(x \vee y)$ should depend only on $v(x)$ and $v(y)$. Write this relationship as a binary operation $\oplus$,

$$v(x \vee y) = v(x) \oplus v(y) \;\text{ when }\; x \wedge y = \emptyset\,. \tag{1.5}$$

To conform with associativity (1.3), we require

$$\big(v(x) \oplus v(y)\big) \oplus v(z) = v(x) \oplus \big(v(y) \oplus v(z)\big). \tag{1.6}$$

This has to hold for arbitrary values $v(x)$, $v(y)$, $v(z)$, and the *associativity theorem* (Azcél 2003) then tells us that there must be some invertable function F of our valuations v such that

$$F\big(v(x \vee y)\big) = F\big(v(x)\big) + F\big(v(y)\big). \tag{1.7}$$

That being the case, we are free to discard the original valuations v and use $m = F(v)$ instead, for which $\vee$ is simple addition: for disjoint x and y we have the *sum rule*

$$\boxed{m(x \vee y) = m(x) + m(y)} \tag{1.8}$$

In other words, valuation can without loss of generality be taken to be what mathematicians call a '*measure*'. They traditionally define measures from the outset as additive over infinite sets, but offer little justification. Mathematicians just do it. Physicists want to know why. Here we see that there's no alternative, and although we start finite we can extend to arbitrarily many elements; it is the same structure. This is *why* measure theory works – it is because of associativity – and we physicists don't have to worry about the infinite.

Assignment

As for actual numerical values, we can build them upwards by addition – except for foundation elements that are not equal to any least upper bound of different elements. Those values alone cannot be determined by the sum rule.

Thus, in the inference example, we can value an apple, a banana and a cherry arbitrarily, but there is a scale on which combinations add. On that scale, if an apple costs $3¢$ and a cherry costs $4¢$, then their combination costs $3 + 4 = 7¢$, not $3^2 + 4^2 = 25$ or other non-linear construction. Associativity underlies money. Personal assignments may be on a different scale. In economics, for example, personal benefit is sometimes held to be logarithmic in money, $m = \log(\$)$, to reflect the asymmetry between devastating downside risk and comforting upside reward. On that scale, money combines non-linearly, as $\log \$(x \vee y) = \log \$(x) + \log \$(y)$. That's permitted, the point being that there *is* a scale on which one's numbers add. Thus quantification is intrinsically linear – because of associativity.

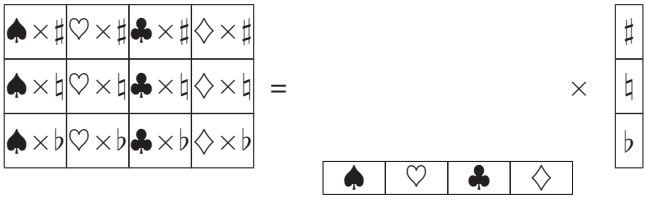
In inference, $\vee$ behaves as logical OR and $\wedge$ as logical AND, obeying the extra property of distributivity:

$$\begin{aligned} (x \texttt{ OR } y) \texttt{ AND } z &= (x \texttt{ AND } z) \texttt{ OR } (y \texttt{ AND } z)\,, \\ (x \texttt{ AND } y) \texttt{ OR } z &= (x \texttt{ OR } z) \texttt{ AND } (y \texttt{ OR } z)\,. \end{aligned} \tag{1.9}$$

Equivalently, they behave as set union and set intersection of the foundation elements, which can therefore be assigned arbitrary values. In other applications, $\vee$ and $\wedge$ might not be distributive, and the foundation assignments become restricted by the non-equality of combinations that would otherwise be identical. But their calculus would still be additive.

Multiplication

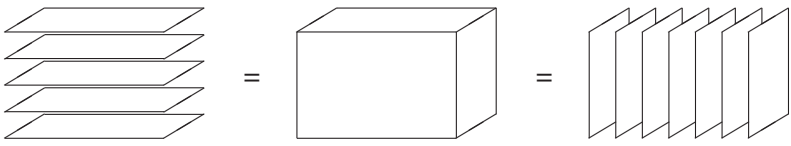
As well as by addition, measures can also combine by multiplication. Here, we consider a direct product of lattices. For example, one lattice might have playing-card foundation elements (♠, ♡, ♣, ♢) while the other has music-key foundations (♭, ♮, ♯). The direct-product lattice treats both together, here with 12 foundation elements like $\heartsuit \times \natural$ and 2^{12} elements overall:



We now assume that the measure $m(x \times y)$ should depend only on $m(x)$ and $m(y)$. Write this relationship as a binary operation

$$m(x \times y) = m(x) \otimes m(y). \tag{1.10}$$

Now the direct-product operator is associative, $(x \times y) \times z = x \times (y \times z)$,



so

$$\big(m(x) \otimes m(y)\big) \otimes m(z) = m(x) \otimes \big(m(y) \otimes m(z)\big). \tag{1.11}$$

This has to hold for arbitrary values $m(x)$, $m(y)$, $m(z)$, and the associativity theorem then tells us that there must be some invertable function Φ of the measures m such that

$$\Phi(m(x \times y)) = \Phi(m(x)) + \Phi(m(y)). \tag{1.12}$$

We cannot now re-grade to $\Phi(m)$ and ignore m because we have already fixed the behaviour of m to be additive. What we can do is require consistency with that behaviour by requiring the sum rule (1.8) to hold for composite elements, $m(x \times t) + m(y \times t) = m((x \vee y) \times t)$ for any t. The *context theorem* (Knuth and Skilling, in preparation) then shows that Φ has to be logarithmic, so that

$$m(x \times y) = m(x)\, m(y). \tag{1.13}$$

While $\oplus$ is addition, $\otimes$ is multiplication. Combination is intrinsically multiplicative – because of associativity. There is no alternative.

Commutativity

Technically, we have not used the commutative property $x \vee y = y \vee x$ of a lattice. However, the sum rule automatically generates values that are equal, $v(x \vee y) = v(y \vee x)$. So real valuations cannot capture non-commutative behaviour. Quantum mechanics is an example, where states lack ordering so do not form a lattice, and the calculus is complex. Inference is not an example. There, *apple*–`OR`–*banana* is the same as *banana*–`OR`–*apple* so $\vee$ is commutative for us, and we are allowed to use the real values that we need.

1.2.3 Information

Different measures can legitimately be assigned to the same foundation elements, as when different individuals value apples, bananas and cherries differently. The difference between source measure μ and destination measure m can be quantified, consistently with lattice structure, as '*information*' $H(m \mid \mu)$.

One way of deriving the form of H is as a variational potential, in which destination m is obtained at the extremal (minimum, actually) of H, subject to whatever constraints require the change from μ. Suppose the playing-card example above has

source measure μ, with destination m obtained by some constraint on card suits. Independently, the music-key example has source measure ν, with destination n obtained by some constraint on music keys.

Equivalently, we must be able to analyze the problems jointly. Measures multiply, so the joint element 'card suit i and music key j' has source measure $\mu_i\nu_j$ and destination m_in_j. The latter is to be obtained at the extremal of $H(m_in_j \mid \mu_i\nu_j)$, under one constraint acting on i and another on j. Temporarily suppressing the fixed source $\mu\nu$, the variational equation for the destination measure is

$$H'(m_in_j) = \lambda_1(i) + \lambda_2(j), \tag{1.14}$$

where the λ's are the Lagrange multipliers of the i and j constraints. Writing $x = m_i$ and $y = n_j$, and differentiating $\partial^2/\partial x\partial y$, the right-hand side is annihilated, leaving

$$xyH'''(xy) + H''(xy) = 0, \tag{1.15}$$

whose solution is

$$H(z) = A - Bz + Cz\log z. \tag{1.16}$$

Setting $C = 1$ an as arbitrary scale (positive to ensure a minimum), $B = 1$ to place that minimum correctly at $m = \mu$, and $A = \mu$ to make the minimum zero, we reach (Skilling 1988)

$$\boxed{H(m \mid \mu) = \mu - m + m\log\tfrac{m}{\mu}} \tag{1.17}$$

This obeys (1.14), so the potential we seek exists, and is required to be of this unique form. The difference between measures plays a deep rôle in Bayesian analysis.

1.2.4 Probability

Acquiring data involves a reduction of possibilities. Some outcomes that might have happened, did not. In terms of the lattice of possibilities, the all-encompassing top element moves down. To deal with this, we seek a *bi*-valuation $p(x \mid t)$, in which the context t of model x can shrink. Within any fixed context, p is to be a measure, being non-negative and obeying the sum rule. But we want to change the numbers when the context changes.

To find the dependence on context, take ordered elements $x \le y \le z \le t$. As before, we require conformity with lattice ordering, here

$$x \le y \le z \quad\Longrightarrow\quad p(x \mid z) \le p(x \mid y) \tag{1.18}$$

so that a wider context dilutes the numerical value. Ordering such as $x \leq z$ can be carried out in two steps, $x \leq y$ and $y \leq z$. Our bi-valuation should conform to this, meaning that we require a "$\odot$" operator combining the two steps into one:

$$p(x \mid z) = p(x \mid y) \odot p(y \mid z). \tag{1.19}$$

Extending this to three steps and considering passage $p(x \mid t)$ from x to t, via y and z, gives another associativity relationship,

$$\big(p(x \mid y) \odot p(y \mid z)\big) \odot p(z \mid t) = p(x \mid y) \odot \big(p(y \mid z) \odot p(z \mid t)\big), \tag{1.20}$$

representing $(((x \leq y) \leq z) \leq t) = (x \leq (y \leq (z \leq t)))$. As before, this induces some invertable function Φ of our valuations p such that

$$\Phi\big(p(x \mid z)\big) = \Phi\big(p(x \mid y)\big) + \Phi\big(p(y \mid z)\big). \tag{1.21}$$

Again, we require consistency with the sum rule $p(x \vee y \mid t) = p(x \mid t) + p(y \mid t)$ for arbitrary context t. A variant of the context theorem (Knuth and Skilling, in preparation) then shows that Φ has to be logarithmic as before, so $\odot$ was multiplication. Specifically, we recognize p as *probability*, hereafter "pr",

$$\left.\begin{array}{ll} 0 = \mathrm{pr}(\emptyset) \leq \mathrm{pr}(x) \leq \mathrm{pr}(t) = 1 & \text{Range} \\ \mathrm{pr}(x \vee y) = \mathrm{pr}(x) + \mathrm{pr}(y) & \text{Sum rule for disjoint } x, y \\ \mathrm{pr}(x \wedge y) = \mathrm{pr}(x \mid y)\,\mathrm{pr}(y) & \text{Product rule} \end{array}\right\} \quad \| \; t \tag{1.22}$$

(The "$\| \; t$" notation means that all probabilities are conditional on t, and avoids proliferation of "$\mid t$" without introducing ambiguity.)

Just as measure theory was forced for valuations, so probability theory is forced for bi-valuations. We need not be distracted by claimed alternatives because they conflict with very general requirements. It is all very simple. There's only this one calculus for numerical bi-valuations on a lattice. If, say, we seek a calculus for conditional beliefs, then this has to be it. But the calculus itself is abstract and motive-free. We don't have to subscribe to an undefined idea like 'belief' in order to use it. In fact, the reverse holds. It is probability, with its defined properties, that would underpin belief, not the other way round.

Most simply of all, probability calculus can be subsumed in the single definition of probability as a ratio definition of measures:

$$\boxed{\mathrm{pr}(x \mid t) = \frac{m(x \wedge t)}{m(t)}} \tag{1.23}$$

This is the original discredited frequentist definition, as the ratio of number of successes to number of trials, now retrieved at an abstract level, which bypasses the catastrophic difficulties of literal frequentism when faced with isolated non-reproducible situations. The calculus of probability is no more than the calculus of proportions.

1.3 Inference

Henceforward, in accordance with traditional accounts, we take all foundation elements to be disjoint, and work in terms of these. The OR operator $\vee$ can be replaced by the summation to which it reduces, while the AND operator $\wedge$ can be written as the traditional comma. It is also usual to use I for context, and allow the discrete choice x to be continuous θ. The rules of probability calculus then reduce to

$$\left.\begin{aligned} \mathrm{pr}(\theta) &\geq 0 && \text{Positivity} \\ \int \mathrm{pr}(\theta)\,\mathrm{d}\theta &= 1 && \text{Sum rule} \\ \mathrm{pr}(\phi,\theta) &= \mathrm{pr}(\phi \mid \theta)\,\mathrm{pr}(\theta) && \text{Product rule} \end{aligned}\right\} \quad \| \, I \tag{1.24}$$

1.3.1 Bayes' theorem

In inference, we need to consider both parameter(s) θ and data D, all in the overarching context I of all possibilities we are currently considering. By the product law, the joint probability of model and data factorizes:

$$\begin{array}{ccccc} \mathrm{pr}(\theta)\,\mathrm{pr}(D \mid \theta) & = \mathrm{pr}(\theta, D) = & \mathrm{pr}(D)\,\mathrm{pr}(\theta \mid D) & \| \, I \\ \text{Prior} \times \text{Likelihood} & = \text{Joint} = & \text{Evidence} \times \text{Posterior} & \\ \pi(\theta)\,\mathcal{L}(\theta) & = \cdots\cdots = & E\,\mathcal{P}(\theta) & \\ \text{Inputs} & \Longrightarrow & \text{Outputs} & \end{array} \tag{1.25}$$

On the left lies the prior probability $\pi(\theta) = \mathrm{pr}(\theta \mid I)$, representing how we originally distributed the parameters' unit mass of probability. This assignment has provoked legendary argumentation, and we discuss it below. Also on the left is the likelihood $\mathcal{L}(\theta) = \mathrm{pr}(D \mid \theta)$, representing the probability distribution of the data for each allowed input θ. This is less controversial. The instrument acquiring the data can usually be calibrated with known inputs θ to find how often it produces specific outputs D, which effectively fixes the likelihood to any desired precision. If there remain any unknown calibration parameters in the likelihood, they can be incorporated in θ as extra parameters to be determined, leading to extra computation but no difficulty of principle.

On the far right is the posterior $\mathcal{P}(\theta) = \mathrm{pr}(\theta \mid D, I)$, representing our inferred distribution of probability among the models, after using the data. The difference between prior and posterior is the information (1.17)

$$H(\mathcal{P} \mid \pi) = \int \mathcal{P}(\theta) \log\big(\mathcal{P}(\theta)/\pi(\theta)\big)\,\mathrm{d}\theta \tag{1.26}$$

gleaned about θ. Also on the right is the evidence $E = \mathrm{pr}(D \mid I)$, representing how well our original assignments managed to predict the data. E is also known as 'prior predictive' (how it is often used), 'marginal likelihood' (how it is often

computed), and various similar terms. However, there ought to be a simple moniker (what it is) for this key quantity in Bayesian analysis, and 'evidence' (not to be confused with dataset) is that name (MacKay 2003).

Of course, the terminology is for convenience only. It is not hard and fast. A posterior to a first analyst may become a prior to a second with new data. Evidence values become likelihoods if the context is widened, so that I becomes merely a provisional model within a wider analysis, and so on. There is really just one quantity, *probability*.

The two outputs, evidence and posterior, can be disentangled by noting that the posterior, being a probability, sums to 1. Here, then, is the complete calculus of inference:

$$\boxed{\begin{aligned} \int \pi(\theta)\,\mathrm{d}\theta = 1 \qquad & \text{Prior} \\ E = \int \pi(\theta)\,\mathcal{L}(\theta)\,\mathrm{d}\theta \qquad & \text{Evidence} \\ \mathcal{P}(\theta) = \frac{\pi(\theta)\,\mathcal{L}(\theta)}{E} \qquad & \text{Bayes' theorem} \end{aligned}} \tag{1.27}$$

Bayes' theorem shows how the prior is modulated into posterior through the likelihood/evidence ratio. The same ratio $\mathcal{L}/E = \mathcal{P}/\pi$ shows up in the information H, alternatively known as the negative entropy.

1.3.2 Prior probability

Before using the data, we need to assign a distribution of prior probability. Probability calculus tells us how to manipulate probability values, but not what they should be in the first place. Neither does the world tell us. The only restriction is that, by the sum rule, all the possibilities must add to 1. Beyond that, we are free to invent any model we want. In that sense, anything goes. It is a challenge. However, the world does give its opinion through data. Better hypotheses predict the data better, through having high values of evidence. That and that alone is what we get, and it is all we need for understanding and for technology. The quest for certainty is mistaken and naive.

Guidelines have been developed for assigning priors, but *beware the dangers!* The first step is to decide on a model – which parameters are to be used to predict the data? The range of these parameters defines what is known as the 'hypothesis space', over which unit mass of prior probability is to be distributed.

Informal

No matter how sophisticated the methodology, there is in the end no escape from an informal assessment of what is judged reasonable in the light of whatever background knowledge is available. Your author proceeds by contemplating perhaps ten points, each representing 10% of the prior, and assigning plausible θ to them. This introspection gives a rough range and indication of shape for the prior, which is then assigned some algebraic form conforming to these. Instead of putting prior mass onto θ, this procedure puts θ onto prior mass, which seems more sympathetic to the basic equations. It is also more sympathetic to the computational requirements, because the prior is uniform by definition when prior mass is the underlying coordinate. Either way, the prior doesn't have to be 'right' in some undefinable sense – it just has to be reasonable.

For a location parameter, an informal centre c and width w might suggest a Cauchy distribution,

$$\pi(\theta) = \frac{w/3.14159\ldots}{(\theta - c)^2 + w^2}\,, \quad (-\infty < \theta < \infty), \tag{1.28}$$

which comfortably tolerates quite wide excursions from the guessed centre if the data demand them. Note that we don't need to interpret c and w as moments, and indeed we may be wiser not to. Probability calculus requires normalization, but not the existence of mean and standard deviation. To some extent, moments are a holdover from the days of manual paper-and-pencil calculation.

A necessarily positive intensity parameter of plausible magnitude a might be assigned either a truncated Cauchy or an exponential distribution,

$$\pi(\theta) = a^{-1}\mathrm{e}^{-\theta/a}\,, \quad (\theta > 0)\,. \tag{1.29}$$

If magnitude was accompanied by width, then a Gamma distribution

$$\pi(\theta) = \frac{\theta^{-1+\mu}\mathrm{e}^{-\theta/\lambda}}{\Gamma(\mu)\lambda^{\mu}}\,, \quad (\theta > 0) \tag{1.30}$$

might be appropriate. Whatever the choice, the prior has to be normalized because it is a probability.

Symmetry

Sometimes, our knowledge of some or all of the states is invariant to exchange. The classic example is a six-sided die, for which θ can be 1 or 2 or 3 or 4 or 5 or 6. Given this knowledge and nothing more, we can only assign equal probability to each: $\pi(1) = \pi(2) = \cdots = \pi(6) = \frac{1}{6}$. If we did anything else, say $\pi(1) > \pi(2)$, then we could exchange the labels 1 and 2 and reach a different assignment with $\pi(2) > \pi(1)$ on the basis of a null change to our prior knowledge. So the same state

of knowledge would be coded two different ways, which is unlikely to be helpful. Prior assignments should conform to any symmetry in our prior knowledge. This does *not* mean that the object being investigated need be symmetric. Indeed, data may well tell us it is not.

Symmetry arguments can be over-played. Here, the classic example is a location parameter θ for which the hypothesis space is unbounded, $-\infty < \theta < \infty$. Given this, and nothing more, one's prior knowledge would be invariant to offset of origin, implying $\pi(\theta) = \text{constant}$. After acquiring data, the posterior distribution $\mathcal{P} = \pi\mathcal{L}/E$ would be independent of whatever constant was chosen. $\mathcal{P}$ would be proportional to the likelihood, which would plausibly prohibit infinite values. With non-zero constant, the prior would become un-normalized (the dreaded 'improper prior'), but otherwise all might be well.

Actually, no. The posterior is the lesser half of Bayesian inference. The evidence comes first. As the allowed range W of θ increases indefinitely, the prior $\pi = 1/W$ decreases indefinitely, and so does the evidence. This means that the model with $W \to \infty$ loses by an infinite factor when compared with any prior that includes even the slightest knowledge of the expected range. In the limit, the posterior becomes $\mathcal{P}(\theta) = 0 \times \mathcal{L}(\theta) \,/\, 0$, and total ignorance is seen to be total stupidity. Moreover, the improper prior extending arbitrarily far fails the sanity check of informal assessment. Are you *really* almost certain that $|\theta| > 10^{100}$?

Approximate invariance to small offsets of origin suggests that the prior should be smooth, but that's as far as the argument should go.

Maximum entropy

Sometimes, informal background knowledge is accompanied by 'testable' constraints in the form of known means $\langle Q \rangle = \int Q(\theta) p(\theta)\, \mathrm{d}\theta$. Here, the informal prior π can be modified to a revised p by minimizing the information $H(p \mid \pi)$. 'Entropy' is just the traditional word for the negative of information, so that maximum entropy just means minimum information. Note that maximum entropy is a method of assignment, not inference. It assigns a single p to be used in later inference. It does not infer a probabilistic distribution of plausible p's.

The variational equation is

$$\delta\Big(\int p(\theta) \log \frac{p(\theta)}{\pi(\theta)}\, \mathrm{d}\theta + \lambda_0 \int p(\theta)\, \mathrm{d}\theta + \lambda_1 \int Q(\theta) p(\theta)\, \mathrm{d}\theta + \cdots \Big) = 0, \quad (1.31)$$

where the first term is H, λ_0 is the Lagrange multiplier for normalization, λ_1 is the multiplier for the constraint, and so on for as many constraints as are given. The solution is

$$p(\theta) = \pi(\theta) \exp(-1 - \lambda_0 - \lambda_1 Q(\theta) - \cdots), \quad (1.32)$$

where the λ's fit the constraints to their required values.

One standard example is a location parameter subject to first and second moments:

$$\mu = \int \theta\, p(\theta)\, \mathrm{d}\theta\,, \qquad \mu^2 + \sigma^2 = \int \theta^2 p(\theta)\, \mathrm{d}\theta. \tag{1.33}$$

The original π becomes modified by a Gaussian,

$$p(\theta) = \pi(\theta) \exp(-1 - \lambda_0 - \lambda_1\theta - \lambda_2\theta^2). \tag{1.34}$$

If the original informal knowledge was weaker than the new constraints, so that π was effectively constant within a few σ of μ, the Gaussian modification would dominate, leading to the standard normal (or Gaussian) distribution:

$$p(\theta) = \frac{\exp\left[-(\theta-\mu)^2/2\sigma^2\right]}{\sqrt{2\pi\sigma^2}}\,, \qquad (\pi = 3.14159\ldots)\,. \tag{1.35}$$

Another standard example is an intensity parameter subject to mean value μ. Here, the original π becomes modified by an exponential,

$$p(\theta) = \pi(\theta) \exp(-1 - \lambda_0 - \lambda_1\theta)\,, \qquad (\theta > 0)\,. \tag{1.36}$$

Again, if the original knowledge was appropriately weaker than the new constraints, so that π was effectively constant for small or moderate θ, this result becomes of standard exponential form:

$$p(\theta) = \mu^{-1} \exp(-\theta/\mu)\,, \qquad (\theta > 0)\,. \tag{1.37}$$

On the other hand, if the original knowledge was weak but different, perhaps effectively uniform over θ^2 so that $\pi \propto \theta$, the result

$$p(\theta) = (\theta/\mu^2) \exp(-\theta/\mu)\,, \qquad (\theta > 0) \tag{1.38}$$

would be different too. Maximum entropy refines prior knowledge, but does not replace it.

Continuous problems

Here, we want to infer a measure m_i defined over so many states i that we may as well use continuum notation $m(x)$. Such a model might represent a spectrum or image, whose intensity is distributed across frequency or spatial coordinate(s) x. In practice, x is digitized into cells i, with the continuum limit merely meaning that macroscopic results stop changing when the digitization gets arbitrarily fine.

We commonly want cell boundaries to be invisible. This means that, in the absence of data saying otherwise, we have the same expectation for the intensity accumulated in a domain Δx whether the domain is treated as a whole, or as the

sum of two or more subdivisions. In symbols, with cell $k = i \cup j$ decomposed into i and j (known as 'stick-breaking'), we require

$$\pi(m_k) = \iint \delta(m_k - m_i - m_j)\, \pi(m_i, m_j)\, \mathrm{d}m_i\, \mathrm{d}m_j \tag{1.39}$$

so that the intensities add correctly as $m_k = m_i + m_j$. Additionally, we often suppose that each cell is to behave independently,

$$\pi(m_i, m_j) = \pi(m_i)\pi(m_j), \tag{1.40}$$

so that there is no prior expectation of internal correlation.

These conditions are actually quite restrictive on the form of prior. As an example of a prior that does not work, take the candidate $\pi(m_k) = \delta(m_k{-}1){+}\delta(m_k{-}2)$ with two-point support (1 and 2) for the values. This cannot be subdivided at all, let alone infinitely. In any subdivision into symmetric halves, each half would need at least two points of support, because giving each only one would be insufficient. But the combination would then cover at least three points, which is too many: QED. Another prior that does not work is $\pi(m) \propto \exp(-H(m))$, proposed in the hope that maximum entropy assignment of a single m might be promoted to a distribution of m's. That cannot be subdivided either.

In technical parlance, a prior that behaves consistently on all scales right down to the infinitesimal limit is called a 'process', and the property of such consistency is called 'infinite divisibility' (Steutel 1979). One prior that does work, common in physics, is the Poisson process. Each small cell i is usually empty, but has small probability λ_i of receiving a quantum, and negligible chance of more than one. By construction, this works in the infinitesimal limit. Occupancies r (usually 0, occasionally 1, negligibly more) of small cells are distributed as

$$\pi(r_1, r_2, \ldots, r_n) = \prod_{i=1}^{n} \big((1 - \lambda_i)\delta(r_i - 0) + \lambda_i \delta(r_i - 1)\big). \tag{1.41}$$

If the quanta are allowed to have individually variable intensity, say exponential

$$\pi(m \mid \text{quantum}) = \mathrm{e}^{-m} \tag{1.42}$$

in suitable units, the intensity pattern among the small cells is

$$\pi(m_1, m_2, \ldots, m_n) = \prod_{i=1}^{n} \big((1 - \lambda_i)\delta(m_i) + \lambda_i \mathrm{e}^{-m_i}\big), \tag{1.43}$$

with each small cell having a small chance of holding a macroscopic intensity. If micro-cells are combined, their Poisson rates λ add, so that $\lambda(x)$ is itself a measure on x. In fact, there is an exact macroscopic formula,

$$\pi(m) = \mathrm{e}^{-\lambda}\big(\delta(m) + \mathrm{e}^{-m}\sqrt{\lambda/m}\, I_1(2\sqrt{\lambda m})\big), \tag{1.44}$$

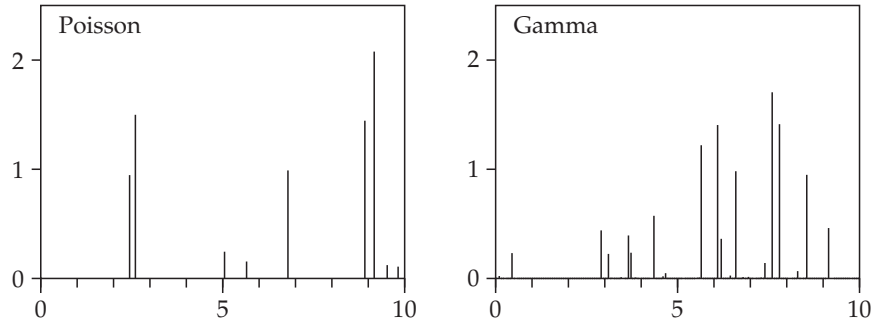


Fig. 1.1. (Left) Sample of Poisson process averaging 10 spikes of mean intensity 1. (Right) Sample of Gamma process of same mean and variance.

for the Poisson model. It is parameterized by the unit of quantum intensity (here 1) and by the production measure λ, and is not too difficult to program in terms of its constituent quanta.

Another prior that works, more popular among statisticians, is the Gamma process:

$$\pi(m_1, m_2, \ldots, m_n) = \prod_{i=1}^{n} \frac{m_i^{-1+\lambda_i}}{\Gamma(\lambda_i)} \, \mathrm{e}^{-m_i}, \tag{1.45}$$

where, as before, $\lambda(x)$ is a measure over x. Although the formula is arguably simpler algebraically, it is less interpretable and more expensive to program because every micro-cell enters the prior (instead of a limited number of quanta). Not that random samples look very different. As calculated at high resolution, both give spiky results (Figure 1.1). The difference is that the Gamma process produces a lot of extremely low-level grass which the Poisson process cuts away. If the Gamma measure m is normalized, the formula reduces to the Dirichlet process (Ferguson 1973)

$$\pi(p_1, p_2, \ldots, p_n) = \delta\big(1 - \textstyle\sum p\big)\, \Gamma\big(\sum \lambda\big) \prod_{i=1}^{n} \frac{p_i^{-1+\lambda_i}}{\Gamma(\lambda_i)} \tag{1.46}$$

for inferring a probability distribution p.

Geometry

When two measures m and $m + \delta m$ are close, their information H becomes approximately symmetric,

$$H(m + \delta m \mid m) \approx \sum_{j=1}^{n} \frac{(\delta m^j)^2}{m^j}, \tag{1.47}$$

and behaves as a distance-squared in parameter space, here digitized to n points for notational clarity, and with coordinates written as contravariant superscripts because the space is about to become Riemannian. The metric describing this distance is diagonal,

$$g_{jk} = \begin{cases} 1/m^j & \text{if } j = k, \\ 0 & \text{otherwise,} \end{cases} \tag{1.48}$$

and, locally, $H = (\mathrm{d}s)^2 = \sum g_{jk}\,\mathrm{d}m^j\,\mathrm{d}m^k$. Measures within a small fixed distance ϵ of m fill an ellipsoid of 'radius' ϵ and volume proportional to $(\det g)^{-1/2}$. By supposing that ϵ-ellipsoids all contain the same mass, regardless of central location, the m's induce their own natural density proportional to $(\det g)^{1/2}$. In the absence of any better guidance, one might try to use this as the prior on m.

For example, suppose parameter space has just two intensities (a 2-cell image, perhaps), over which we seek a prior $\pi(m^1, m^2)$. The proposal is

$$(\det g)^{1/2} = \sqrt{\det \begin{pmatrix} 1/m^1 & 0 \\ 0 & 1/m^2 \end{pmatrix}} = \frac{1}{\sqrt{m^1 m^2}}. \tag{1.49}$$

An immediate objection is that this expression is not normalizable, leading to an improper prior for m. However, the components of m could represent proportions p and $1-p$ of some fixed total, and the proposed prior for p (the *shape* of the 2-cell image) would then be

$$\pi(p) = \frac{1/3.14159\ldots}{\sqrt{p(1-p)}}, \tag{1.50}$$

which is normalized, and might well be acceptable.

The generalization to a larger number n of proportions p is a Dirichlet distribution (1.46), but with all indices λ equal to $\frac{1}{2}$. This would not be acceptable for inference about an image digitized to arbitrarily many cells. The reason is that macroscopic structure is washed out. For example, the total proportion in any $n/2$ cells is almost certain to be very close to the uniform, featureless $\frac{1}{2}$. Dirichlet indices λ need to be a measure, thereby getting individually smaller as cells are subdivided. The geometrical indices of $\frac{1}{2}$ do not get smaller. Here, the geometric proposal fails to be infinitely divisible.

The geometrical formulation can be generalized to a parameterized subspace $m^i = m(i \mid \theta)$ restricted to the range of r parameters $\theta^1, \ldots, \theta^r$. The information between neighbours becomes

$$H(\theta + \mathrm{d}\theta \mid \theta) = \sum_{i=1}^{n} \frac{(\mathrm{d}m^i)^2}{m^i} = \sum_{i=1}^{n} \frac{1}{m^i} \Big(\sum_{j=1}^{r} \frac{\partial m^i}{\partial \theta^j}\,\mathrm{d}\theta^j \Big) \Big(\sum_{k=1}^{r} \frac{\partial m^i}{\partial \theta^k}\,\mathrm{d}\theta^k \Big), \tag{1.51}$$