OXFORD

Bayesian Theory and Applications

EDITED BY
PAUL DAMIEN, PETROS DELLAPORTAS
NICHOLAS G. POLSON, & DAVID A. STEPHENS

BAYESIAN THEORY AND APPLICATIONS

This page intentionally left blank

Bayesian Theory and Applications

Edited by

PAUL DAMIEN
University of Texas, Austin

PETROS DELLAPORTAS
Athens University of Economics and Business

NICHOLAS G. POLSON
University of Chicago

DAVID A. STEPHENS
McGill University

OXFORD
UNIVERSITY PRESS

OXFORD
UNIVERSITY PRESS

Great Clarendon Street, Oxford, OX2 6DP,
United Kingdom

Oxford University Press is a department of the University of Oxford.
It furthers the University's objective of excellence in research, scholarship,
and education by publishing worldwide. Oxford is a registered trade mark of
Oxford University Press in the UK and in certain other countries

First Edition published in 2013

Impression: 1

British Library Cataloguing in Publication Data
Data available

ISBN 978–0–19–969560–7

Printed in Great Britain by
CPI Group (UK) Ltd, Croydon, CR0 4YY

Dedication

This volume is dedicated to Sir Adrian Smith, F.R.S.

This page intentionally left blank

Contents

Contributors x
Introduction xii

Part I Exchangeability

1 Observables and models: exchangeability and the inductive argument 3
Michael Goldstein

2 Exchangeability and its ramifications 19
A. Philip Dawid

Part II Hierarchical Models

3 Hierarchical modelling 33
Alan E. Gelfand and Souparno Ghosh

4 Bayesian hierarchical kernel machines for nonlinear regression and classification 50
Sounak Chakraborty, Bani K. Mallick and Malay Ghosh

5 Flexible Bayesian modelling for clustered categorical responses in developmental toxicology 70
Athanasios Kottas and Kassandra Fronczyk

Part III Markov Chain Monte Carlo

6 Markov chain Monte Carlo methods 87
Siddhartha Chib

7 Advances in Markov chain Monte Carlo 104
Jim E. Griffin and David A. Stephens

Part IV Dynamic Models

8 Bayesian dynamic modelling 145
Mike West

9 Hierarchical modelling in time series: the factor analytic approach 167
Dani Gamerman and Esther Salazar

10 Dynamic and spatial modelling of block maxima extremes 183
Gabriel Huerta and Glenn A. Stark

Part V Sequential Monte Carlo

11 Online Bayesian learning in dynamic models: an illustrative introduction to particle methods 203
Hedibert F. Lopes and Carlos M. Carvalho

12 Semi-supervised classification of texts using particle learning for probabilistic automata 229
Ana Paula Sales, Christopher Challis, Ryan Prenger and Daniel Merl

Part VI Nonparametrics

13 Bayesian nonparametrics 249
Stephen G. Walker

14 Geometric weight priors and their applications 271
Ramsés H. Mena

15 Revisiting Bayesian curve fitting using multivariate normal mixtures 297
Stephen G. Walker and George Karabatsos

Part VII Spline Models and Copulas

16 Applications of Bayesian smoothing splines 309
Sally Wood

17 Bayesian approaches to copula modelling 336
Michael Stanley Smith

Part VIII Model Elaboration and Prior Distributions

18 Hypothesis testing and model uncertainty 361
M. J. Bayarri and J. O. Berger

19 Proper and non-informative conjugate priors for exponential family models 395
E. Gutiérrez-Peña and M. Mendoza

20 Bayesian model specification: heuristics and examples 409
David Draper

21 Case studies in Bayesian screening for time-varying model structure: the partition problem 432
Zesong Liu, Jesse Windle and James G. Scott

Part IX Regressions and Model Averaging

22 Bayesian regression structure discovery 451
Hugh A. Chipman, Edward I. George and Robert E. McCulloch

23 Gibbs sampling for ordinary, robust and logistic regression with Laplace priors 466
Robert B. Gramacy

24 Bayesian model averaging in the M-open framework 483
Merlise Clyde and Edwin S. Iversen

Part X Finance and Actuarial Science

25 Asset allocation in finance: a Bayesian perspective 501
Eric Jacquier and Nicholas G. Polson

26 Markov chain Monte Carlo methods in corporate finance 516
Arthur Korteweg

27 Actuarial credibility theory and Bayesian statistics—the story of a special evolution 546
Udi Makov

Part XI Medicine and Biostatistics

28 Bayesian models in biostatistics and medicine 557
Peter Müller

29 Subgroup analysis 576
Purushottam W. Laud, Siva Sivaganesan and Peter Müller

30 Surviving fully Bayesian nonparametric regression models 593
Timothy E. Hanson and Alejandro Jara

Part XII Inverse Problems and Applications

31 Inverse problems 619
Colin Fox, Heikki Haario and J. Andrés Christen

32 Approximate marginalization over modelling errors and uncertainties in inverse problems 644
Jari Kaipio and Ville Kolehmainen

33 Bayesian reconstruction of particle beam phase space 673
C. Nakhleh, D. Higdon, C. K. Allen and R. Ryne

Adrian Smith's research supervision (PhD) 687
Adrian Smith's publications 689
Index 697

Contributors

- C. K. Allen, Oak Ridge National Laboratory
- M. J. Bayarri, Universitat de Valencia
- J. O. Berger, Duke University
- Carlos M. Carvalho, University of Texas in Austin
- Sounak Chakraborty, University of Missouri-Columbia
- Christopher Challis, Duke University
- Siddhartha Chib, Washington University
- Hugh A. Chipman, Acadia University
- J. Andrés Christen, CIMAT, Mexico
- Merlise Clyde, Duke University
- A. Philip Dawid, University of Cambridge
- David Draper, University of California in Santa Cruz
- Colin Fox, University of Auckland
- Kassandra Fronczyk, M.D. Anderson Cancer Institute
- Dani Gamerman, UFRJ, Brazil
- Alan E. Gelfand, Duke University
- Edward I. George, University of Pennsylvania
- Malay Ghosh, University of Florida
- Souparno Ghosh, Duke University
- Michael Goldstein, Durham University
- Robert B. Gramacy, University of Chicago
- Jim E. Griffin, University of Kent
- E. Gutiérrez-Peña, IIMAS, UNAM
- Heikki Haario, Lappeenranta University of Technology
- Timothy E. Hanson, University of South Carolina
- D. Higdon, Los Alamos National Laboratory
- Gabriel Huerta, Indiana University
- Edwin S. Iversen, Duke University
- Eric Jacquier, MIT
- Alejandro Jara, Pontifica Universidad Catolica de Chile

- Jari Kaipio, University of Auckland
- George Karabatsos, University of Illinois
- Ville Kolehmainen, University of Eastern Finland
- Arthur Korteweg, Stanford University
- Athanasios Kottas, University of California in Santa Cruz
- Purushottam W. Laud, Medical College of Wisconsin
- Zesong Liu, University of Texas in Austin
- Hedibert F. Lopes, University of Chicago
- Udi Makov, University of Haifa
- Bani K. Mallick, Texas A & M University
- Robert E. McCulloch, University of Chicago
- Ramsés H. Mena, IIMAS, UNAM
- M. Mendoza, ITAM
- Daniel Merl, Lawrence Livermore National Laboratory
- Peter Müller, University of Texas in Austin
- C. Nakhleh, Sandia Labs
- Nicholas G. Polson, University of Chicago
- Ryan Prenger, Lawrence Livermore National Laboratory
- R. Ryne, Lawrence Berkeley National Laboratory
- Esther Salazar, Duke University
- Ana Paula Sales, Lawrence Livermore National Laboratory
- James G. Scott, University of Texas in Austin
- Siva Sivaganesan, University of Cincinnati
- Michael Stanley Smith, Melbourne Business School
- Glenn A. Stark, University of New Mexico
- David A. Stephens, McGill University
- Stephen G. Walker, University of Kent
- Mike West, Duke University
- Jesse Windle, University of Texas in Austin
- Sally Wood, Melbourne Business School

Introduction

At the outset, we would like to thank all the authors that have contributed to this volume. This book is dedicated to a statistician whose work in Bayesian statistics has forever changed the way in which statistical research and practice has been and will be carried out. Adrian Smith's accomplishments are documented at the end of this volume. Here, we simply note that three key ideas in this volume—hierarchical models, Markov chain Monte Carlo and sequential Monte Carlo—that have revolutionized Bayesian statistics are in large measure due to Adrian's contributions. These concepts are now ubiquitous wherever Bayesian models are used. In this volume, we have selected broad topic areas where these ideas come into play in a significant manner. Of course these topics are by no means exhaustive, but they serve to illustrate the impact that Adrian's research has had on Bayesian statistics in the last four decades or so.

When we conceived this volume, we wanted to position it somewhat differently from other tribute volumes. To accomplish this, based on our collective experiences, we felt that some of the basic ideas in modern Bayesian statistics with which Bayesian statisticians are familiar are foreign to some (if not many) colleagues and practitioners in other disciplines. Therefore, we felt that a volume that had a 'Bayesian textbook' flavour to it, and which also included application papers would prove useful in spreading modern Bayesian ideas. This is the modus operandi adopted in most of the chapters. We now discuss each part in turn.

Part I: Exchangeability Dawid and Goldstein explore the fundamental notion of Bayesian statistics, namely exchangeability.

Part II: Hierarchical Models The first key idea in modern Bayesian statistics is hierarchical models. Gelfand and Ghosh discuss the elementary ideas underlying such models. This is then followed up by Chakraborty, Mallick and Ghosh, and Kottas and Fronczyk's papers.

Part III: Markov Chain Monte Carlo The second key idea in modern Bayesian statistics is Markov chain Monte Carlo (MCMC). Chib reviews the key MCMC approaches to implementing full Bayesian analysis. Griffin and Stephens' contribution further describes and exemplifies advanced MCMC notions.

Part IV: Dynamic Models West describes the fundamentals of dynamic linear and nonlinear models. Papers by Gamerman and Salazar, and Huerta and Stark elaborate on these ideas via some novel applications.

Part V: Sequential Monte Carlo Carvalho and Lopes describe the use of SMC in a variety of Bayesian models, which is then followed up by an applications paper by Sales, Challis, Prenger and Merl.

Part VI: Nonparametrics Bayesian nonparametrics is embedded in the exchangeability ideas found in Part I. Walker discusses Bayesian nonparametric models and argues that to perform proper data analysis one must adopt nonparametric models at the outset. Two papers, one by Karabatsos and Walker, and the second by Ména complete this part.

Part VII: Spline Models and Copulas Part VI considers Bayesian nonparametrics using exchangeability as the basis. There are related approaches to nonparametrics but with some key

differences. Two such classes of models are discussed in this part: Bayesian splines by Wood, and Bayesian copulas by Smith.

Part VIII: Model Elaboration and Prior Distributions Bayarri and Berger describe the fundamentals of Bayesian hypothesis testing, followed by three research papers by Draper; Liu, Windle and Scott; and Gutiérrez-Peña and Mendoza.

Part IX: Regressions and Model Averaging Chipman, George and McCulloch describe the correct way of doing regressions. This is further elaborated on in two papers by Clyde and Iversen, and Gramacy.

Part X: Finance and Actuarial Science Jacquier and Polson discuss the role of Bayes in financial applications. This is followed by a comprehensive review of Bayesian models in corporate finance by Korteweg. One area where Bayesian methods are only now beginning to gain popularity is actuarial science. Makov describes Bayesian models in this context.

Part XI: Medicine and Biostatistics It is safe to say that Bayesian methods have found most widespread use in biostatistics and bio-informatics. Mueller details the Bayesian models in these areas, followed by two key papers in biostatistics by Laud, Müller and Sivaganesan, and Hanson and Jara.

Part XII: Inverse Problems and Applications This is an exciting area of science where Bayesian methods are fast gaining in popularity. Fox, Haario and Christen provide a complete description of Bayesian ideas in this field, followed by two practical papers: one by Kaipio and Kolehmainen, and a second by Nakhleh, Higdon, Allen and Ryne.

Special thanks to Carlos Carvalho, Marcin Kacperczyk, Bani Mallick, Tom Shively, and Daniel Zantedeschi for helping review some of the papers.

Finally, we would like to thank Clare Charles, Elizabeth Hannon, Keith Mansfield, Viki Mortimer, Subramaniam Vengatakrishnan and their colleagues at Oxford University Press for their tireless efforts in ensuring that this book was completed in a timely and efficient manner.

This page intentionally left blank

Part I
Exchangeability

This page intentionally left blank

1 Observables and models: exchangeability and the inductive argument

MICHAEL GOLDSTEIN

1.1 Introduction

When quantifying uncertainty for large and complex systems, it is often considered helpful to regard such uncertainty as being of two kinds, epistemic and aleatory. Epistemic uncertainty is that which relates to our lack of knowledge, and could be reduced by receipt of further information. Aleatory uncertainty is that which relates to intrinsic chance variation in the system, and cannot be resolved except by direct observation. The distinction between aleatory and epistemic uncertainty is informal rather than precise, particularly within the view that all uncertainty stems from a lack of knowledge and understanding. Indeed, a basic activity in much of science is searching for explanatory structure within apparently random events, which corresponds to moving uncertainty from the aleatory to the epistemic form, where it can be better understood and, possibly, reduced.

The aleatory/epistemic distinction has a natural counterpart in much statistical analysis, where aleatory uncertainty is expressed through the likelihood function for the data given the population parameters, while epistemic uncertainty is expressed through the prior distribution over the parameters, within the Bayesian formulation, and is treated less formally within relative frequency based approaches. This division between uncertain model parameters and likelihoods conditional on the values of the parameters is helpful and constructive when modelling our uncertainty about a physical system. However, as with any other form of modelling, this does raise fundamental questions when we seek to apply the results of the model based analysis to actual real world inferences. All that we actually observe are individual measurements of real things. The parametric forms that we introduce to describe intrinsic chance variation are simply models whose meaning and justification remains to be established.

Within the subjectivist approach, there is a precise answer to the question of meaning for many statistical models. This meaning is rooted in the judgement of exchangeability. Exchangeability allows us to construct parametric statistical models purely on the basis of the uncertainty statements that we make about observable random quantities. Indeed, in many cases, the argument shows that we have no choice but to behave as though we consider that we are sampling from a parametric model (the aleatory uncertainty) given the true but unknown values of some population distribution (the epistemic uncertainty). Therefore, exchangeability is the logical bedrock to a large part of

current statistical analysis. Beyond this, the distinction between aleatory and epistemic uncertainty pervades so much of current scientific analysis that the notion of exchangeability is a necessary conceptual tool to provide the underpinnings of meaning for uncertainty quantification in general and the inductive argument, namely the reasoning from particular cases to general principles, in particular.

Our aims in this chapter are two-fold. Firstly, we shall give an elementary and self-contained account of the notion of exchangeability and the derivation of de Finetti's representation theorem, which shows how we may construct operational statistical models based strictly on our judgements over observables. Secondly, we shall consider the relevance of this representation to real world inferences, and introduce a second collection of exchangeability judgements which are necessary in order that the inductive argument, when applied to inferences over models so constructed, also has an operational real world counterpart.

1.2 Finite population sampling

Finite population sampling gives a concrete illustration of the distinction between aleatory and epistemic uncertainty. Consider a simple version of this problem. We have a bucket, which contains a known large number, N, of counters, of which an unknown proportion q are red, and the remaining proportion $(1 - q)$ are blue. We intend to draw a counter at random from the bucket. (Here, and below, we use the term 'at random' as shorthand for the subjective judgement that each counter currently in the bucket is equally likely to be selected at each stage.) Let $Z = 1$ if this draw is red, and let $Z = 0$, otherwise. We are uncertain as to the value that Z will take. This uncertainty has two components. Firstly, we do not know the value of q. This is epistemic uncertainty. It can be quantified by consideration of what we know about the way that the population was formed, and will be further reduced if we take samples from the bucket. Different people will have different states of knowledge and so their epistemic uncertainty may differ. Secondly, even if we did know q, we still would not know the value of Z. This value would now be the realization of a Bernoulli random variable, parameter q, and this irreducible uncertainty is aleatory. The distinction between aleatory and epistemic uncertainty is most useful when there is a general consensus as to the representation of aleatory uncertainty, e.g. here, to the extent that there is general agreement that the draw from the bucket will be random, and no obvious way to impose more structure upon this variation.

The possible values of q are $q_i = i/n, i = 0, 1, \ldots, N$. In the subjective Bayes view, we may quantify our epistemic uncertainty for q by specifying our collection of probabilities $p_i = \mathrm{P}(q = q_i)$. Therefore, we can assess our probability that $Z = 1$, by the law of total probability, as

$$\mathrm{P}(Z = 1) = \sum_{i=0}^{N} p_i q_i \tag{1.1}$$

A useful way to rewrite (1.1) is

$$\mathrm{P}(Z = 1) = \int_0^1 q \, dF(q) \tag{1.2}$$

where F is the probability measure on $[0,1]$ which assigns probability p_i to the point i/N.

Now suppose, instead, that we are going to take a random sample of size n, without replacement, from the bucket. Let X denote the number of red counters in the sample. Epistemic uncertainty is as before. Our aleatory uncertainty relates to the probability distribution for X if we know q, the proportion of red counters in the bucket. This distribution is hypergeometric, so that

$$\mathrm{P}(X = k|q) = \frac{\binom{Nq}{k}\binom{N(1-q)}{n-k}}{\binom{N}{n}} \tag{1.3}$$

Therefore, the corresponding version of (1.2) is

$$\mathrm{P}(X = k) = \int_0^1 \frac{\binom{Nq}{k}\binom{N(1-q)}{n-k}}{\binom{N}{n}}\, dF(q) \tag{1.4}$$

If n is small compared to N, then there is little difference between sampling with and without replacement, and so, approximately, we can rewrite (1.3) as

$$\mathrm{P}(X = k|q) \approx \binom{n}{k} q^k (1-q)^{(n-k)} \tag{1.5}$$

so that (1.4) may be approximated as

$$\mathrm{P}(X = k) \approx \int_0^1 \binom{n}{k} q^k (1-q)^{(n-k)}\, dF(q) \tag{1.6}$$

Representation (1.6) is familiar in the Bayesian context, and is often described by saying that X has a binomial likelihood, parameters n, q, where our prior measure for q is given by F. In the above examples, F was a discrete measure placing probabilities on each value i/N. As N increases, it is often helpful to approximate this discrete measure by a continuous pdf $f(q)$, so that

$$\mathrm{P}(X = k) \approx \int_0^1 \binom{n}{k} q^k (1-q)^{(n-k)} f(q)\, dq \tag{1.7}$$

For example, most introductory treatments for Bayesian statistics deal with (1.7) by discussing the special case where $f(q)$ is a beta distribution, as this case has simplifying conjugacy properties. However, in our development, it is helpful to retain the possibility that F could have any form at all; for example F could be a mixture of discrete and continuous components if there were certain special choices for q. (Suppose, for example, that our bucket had been chosen by a coin flip between two buckets, for one of which we knew that q was 0.5, but we had no information about the value for q in the other bucket.)

As we increase the size of N as compared to n, the approximation (1.5), and so also (1.6), becomes increasingly precise. We can see this informally as (1.5) would be exact if we were sampling with replacement, and only removing a small number of counters from a large bucket will only change the proportion of red counters by a small amount. We can support this intuition by showing that the right-hand side of (1.3) tends uniformly to the right-hand side of (1.5) with N; for example, the most extreme change to the proportions in the bucket is to draw all counters of the same colour, and for any N, q and $n < Nq$, the probability of n successes when sampling without replacement is less than the probability when sampling with replacement, but greater than the probability for sampling with replacement if we first remove n red counters from the bucket, so that

$$(\frac{q-f}{1-f})^n \leq \mathrm{P}(X = n|q) \leq q^n$$

where f is the sampling fraction $f = n/N$, and so the approximation is very close for f near zero.

We have described this sampling problem from a Bayesian viewpoint. In the common situation where we have a large sample, n, from a much larger population, N, most inferential approaches will reach the same conclusion, namely that the proportion of red counters in the sample estimates the population proportion with high accuracy. When the sampling fraction f is not small, we must take more care in approximating the hypergeometric distribution, and if n is small, then our representation of epistemic uncertainty through F will be important. However, in all cases, the meaning of the analysis will be clear, in the sense that there is a true but unknown population parameter q, which is, in principle, observable, and a generally agreed aleatory description as to how the sample is drawn, given q.

Most statistical problems lack this logical bedrock. For example, if we spin a coin repeatedly, and would like to use our observed spins to revise our judgements about future spins, then we might represent our uncertainty by means of a model in which, given the value of q, the 'the true but unknown' value of the probability that the spun coin will land heads, the coin spins are independent Bernoulli variables with parameter q. This model is formally similar to the finite population problem that we have been discussing, but with the fundamental distinction that the quantity q over which we now express our uncertainty is only a model quantity, which is not observable even in principle and lacks even a real world definition. However, there is a bridge between such uncertainty models and the problem of finite population sampling and this comes through the concept of exchangeability, as we shall now describe.

1.3 Exchangeable samples

In the problem of sampling counters from the bucket that we described above, consider making an ordered series of draws, $X_1, X_2, \ldots, X_N$ from the bucket, without replacement, where $X_i = 1$ if the ith draw is red, $X_i = 0$ otherwise. For us, the sequence is not independent. Observing each draw alters both the aleatory uncertainty (each time we observe a red counter, then this reduces the proportion of red counters available for the next draw) and the epistemic uncertainty (each time we observe a red counter, this changes our state of knowledge about the true proportion of red counters in the bucket). However, the sequence does have certain probabilistic properties which are important for the general account that we shall develop.

Consider first a single draw X_i. For each draw i, given the initial proportion q of red counters, the probability of drawing a red counter is the same, namely q, as, on each draw, each individual counter has the same probability of being selected. Therefore, each X_i has the same probability distribution, namely Bernoulli, with parameter given by (1.2). Now consider any pair of draws, X_i, X_j. Given q, the ith draw has probability q of being red. If the ith draw is red, then the jth draw is a random draw from a bucket, size $N - 1$, with $qN - 1$ red counters. Therefore, the probability distribution of the number of red counters in two draws is given by (1.4), for the case where $n = 2$, and this is true for all pairs $i \neq j$. Continuing in this way, we have that the probability distribution of any collection of n elements $(X_{i_1}, \ldots, X_{i_n})$ from the series has the same probability distribution, however we select and permute the indices $i_1, \ldots, i_n$, as given by (1.4). This notion, that the probability distribution of any collection of n of the quantities depends only on the value of n, and not on the individual quantities selected, or the order in which they are arranged, is termed **exchangeability** and is fundamental to the subjectivist representation for epistemic and aleatory uncertainty.

While we have only discussed simple two-valued scalar quantities so far, the concept of exchangeability is quite general. We make the following definition.

Definition *A sequence* $(Y_1, Y_2, \ldots)$ *of random vectors* $Y_i = (Y_{i1}, Y_{i2}, \ldots, Y_{im})$ *taking values in some space* Ω *is said to be* **exchangeable** *if the joint probability distribution of each subcollection of n quantities* $(Y_{i_1}, Y_{i_2}, \ldots, Y_{i_n})$ *is the same.*

In our account of picking counters from the bucket, we deduced exchangeability of the sequence of selections from our views as to the physical description of the problem. The notion of exchangeability reverses the logic of this argument and allows us to deduce the structure of the problem directly from the judgement of exchangeability. This is usually termed **de Finetti's representation theorem for exchangeable sequences**. We will introduce this representation by discussing the example of spinning coins and then use a more general form of this argument to give the general form for the theorem. Our account builds on the treatment in [10].

1.4 The representation theorem for exchangeable binary sequences

Suppose that we spin a coin repeatedly. The familiar model, which treats coin spins as independent Bernoulli random variables each with a true but unknown probability q of landing heads, lacks an operational physical basis. However, we can retrieve a version of this model if we make the judgement that the coin spins are exchangeable.

Let $U_i = 1$ if the ith spin is heads, $U_i = 0$ otherwise. Suppose that we view the sequence $(U_1, U_2 \ldots)$ as exchangeable. To simplify our account, we will treat this sequence of coin spins as, in principle, infinite, which we can informally interpret as saying that we are able to consider as large a number of spins as we need when we construct our uncertainty judgements over the outcomes. (If we are restricted to a finite exchangeable sequence, then the results that we obtain will correspond to those deducible from finite population sampling, see for example, [3].)

Consider the following thought experiment. Imagine that we have an empty bucket, and a pile of counters, numbered, sequentially, 1 to N. Consider spinning the coin N times. We shall mark the outcome of the ith spin on the ith counter, so each counter is either marked 1 or 0. Each counter is added to the bucket.

As the spins are exchangeable, we must assign the same probability, q say, to the event that the first spin is heads, i.e. that $U_1 = 1$, as we do to the event that a randomly chosen counter from the bucket has value 1 (as the probability that the randomly chosen counter has value 1 is the average of the probabilities that each counter has value 1, which, by the exchangeability judgement, must all be equal to q). Again we may make a division into a notional epistemic uncertainty as to the value of the proportion of counters marked 1 in the bucket and an aleatory uncertainty for the value on the single counter that we pick, given this proportion. Therefore the probability, for the randomly selected counter in the thought experiment, can be constructed as in (1.1), by first considering the possible values $q_i = i/N, i = 0, 1, \ldots, N$ for the proportion of counters labelled 1, in the bucket, and assigning probabilities p_i to the outcomes for q as above. We have

$$\mathrm{P}(U_1 = 1) = \int_0^1 q \, dF_N(q) \tag{1.8}$$

in the same way, where F_N assigns probability p_i to point i/N.

We can extend this argument to our judgement about the outcome of n tosses in the same way. If W_n is the number of heads in the first n spins, then our probability for observing $W_n = k$ is the same as the probability that we assign for this event in any sample of n spins, and so this probability must be equal to the probability of drawing k counters labelled 1 from the bucket in n random picks, which, by relation (1.4), is given, $\forall N \geq n$, as

$$\mathrm{P}(W_n = k) = \int_0^1 \frac{\binom{Nq}{k}\binom{N(1-q)}{n-k}}{\binom{N}{n}} \, dF_N(q) \tag{1.9}$$

The simplest way to consider what happens as N increases is to invoke Helly's theorem (see, for example, [4], which also contains an insightful discussion of exchangeability) and a quite different derivation of the exchangeability representation theorem), which states that any infinite sequence of probability distributions G_N on a bounded interval contains a subsequence which converges in distribution to a limit, say G.

(Helly's theorem is a consequence of the result that any infinite sequence of numbers $a_1, a_2, \ldots$ on a bounded interval has a uniformly convergent subsequence. The result for number sequences can be shown as follows, where we suppose all numbers lie in [0,1]. Divide the sequence into ten subsequences, according to the first decimal place. At least one subsequence must be infinite. Keep one such subsequence and discard the rest. Let b_1 be the element a_{i_1} with the smallest subscript in this subsequence. Now divide this subsequence into ten subsequences according to the value of the second decimal place. At least one subsequence must be infinite. Keep one such subsequence and discard the rest. Let b_2 be the element a_{i_2} with the smallest subscript in this subsequence with $i_2 > i_1$. Continue in this way and the sequence $b_1, b_2, \ldots$ converges uniformly to a limit, as all values $b_j, b_{j+1}, \ldots$ agree in the first j decimal places, for each j. Helly's theorem follows by repeated application of this method. We first select an infinite subsequence of probability distributions which agree in the first decimal place for the probabilities that they assign to the intervals $[0, 0.5)$ and $[0.5, 1]$. From this subsequence, we select a subsequence which agrees to two decimal places for the probability assigned to intervals $[0, 0.25)$, $[0.25, 0.5)$, $[0.5, 0.75)$, $[0.75, 1]$ and so forth. Choosing an element from each subsequence constructed in this way, we arrive at Helly's theorem.)

Applying Helly's theorem to the sequence F_N, there is a subsequence which converges in distribution to a limit F. Letting N tend to infinity, F_N tends to F and the hypergeometric integrand tends uniformly to the binomial, so that we have, for each k, n

$$\mathrm{P}(W_n = k) = \int_0^1 \binom{n}{k} q^k (1-q)^{(n-k)}\, dF(q) \tag{1.10}$$

(1.10) is de Finetti's theorem for an infinite exchangeable sequence of binary outcomes, derived in [1]. The uniqueness of the distribution F satisfying (1.10) follows as a probability distribution on a bounded interval is uniquely determined by its moments: this is the Hausdorff moment problem (see [4] which contains a direct derivation of the exchangeability representation theorem based on this property). The theorem shows that the judgement of exchangeability, alone, is sufficient to ensure that our beliefs about the sequence are just as if we consider that there is a true but unknown quantity q given the value of which we view the sequence as a series of independent Bernoulli trials with probability q.

The convergence of the sequence F_N to F is uniform, as for any $N_1 < N_2 < N_3$ we may view F_{N_1}, F_{N_2} respectively as the distribution of q in buckets formed by draws of size N_1, N_2 respectively from a bucket formed by N_3 spins of the coin, so that F_{N_2} will be probabilistically closer than F_{N_1} to F_{N_3}, corresponding to the intuition that there are no features of a population that are better estimated by a small sample than by a large sample. In this way, we see that the exchangeability representation (1.10) is really a statement about our judgements over large finite collections of coin spins We invoke infinity simply to allow us to make a continuous approximation to the discrete process, to any order of accuracy that we require.

Notice, in particular, that we have constructed the measure $F(q)$ as the limit of the measures $F_N(q)$, namely the measures for the proportion of heads, $q_{[N]}$ in the first N tosses. This is another way of saying that the relative frequencies $q_{[N]}$ tend to a limit q in distribution. This is the subjectivist formulation of the notion of limiting relative frequency. The relative frequency approach to statistics uses the limiting relative frequency as the definition for the notion of probability but is unable to give a proper justification for this definition, or even a satisfactory explanation as to the

way in which the limit should be understood. In contrast, the subjectivist approach constructs the limiting relative frequency as a subjective judgement which is implied by subjective exchangeability judgements over the sequence and deduces the limit as a necessary consequence of this judgement.

1.5 The general form for the exchangeability representation

The argument of the preceding section relates to coin spins, but it applies similarly to any infinite random exchangeable sequence of vectors, $(Y_1, Y_2, \ldots)$, over a space Ω. Just as before, we carry out the thought experiment of constructing a bucket with N counters, where the ith counter is marked with the value of Y_i. Let Q_N denote the empirical distribution of the counters in the bucket, so that Q_N assigns probability $1/N$ to the value on each counter. As the sequence Y is exchangeable, the first value, Y_1, has the same probability distribution as a draw according to the distribution Q_N. Therefore, we can split up our uncertainty as to the outcome of Y_1 into two parts. Firstly, we are uncertain as to the distribution Q_N, and secondly, given Q_N, we are uncertain as to the value of the observation Y_1. Denote our probability distribution for Q_N by F_N (so F_N assigns probabilities for all possible empirical distributions consisting of N selections from the space Ω). Then, analogously to (1.8), we have, for any $A_1 \in \Omega$,

$$\mathrm{P}(Y_1 \in A_1) = \int Q_N(A_1)\, dF_N(Q_N) \tag{1.11}$$

where $Q_N(A_1)$ is the probability assigned to A_1 by the distribution Q_N (i.e. the proportion of the first N outcomes that are within A_1).

Now consider our probability distribution for the first n outcomes $(Y_1, Y_2, \ldots, Y_n)$. We can assess this distribution in two stages as above. First, we make a random choice for the empirical distribution Q_N according to F_N. Given the choice of Q_N, we now make n draws, without replacement, from the bucket consisting of N counters with this empirical distribution. We can evaluate this distribution exactly by a counting argument. If n is small compared to N, then each draw will only change the composition of the remaining counters in the bucket by a small amount, so that the draws will be almost independent. Therefore, we have that

$$P(Y_1 \in A_1, \ldots, Y_n \in A_n) \approx \int Q_N(A_1) \ldots Q_N(A_n) dF_N(Q_N) \tag{1.12}$$

As we let N increase, keeping n fixed, the exact form of the integrand in (1.12) tends uniformly to the product integrand. The distribution F_N tends to a limiting distribution F over the probability distributions Q over the space Ω. (The details of the limiting argument are technically more complicated than for the coin flips, due to the generality of the formulation, but the argument is the same, namely that the empirical distribution of a large sample, size N, from a much larger population, size M say, is close to the population distribution, by the standard arguments of finite population sampling, and therefore the sequence of distributions F_N must converge.)

Proceeding in this way, we have the generalization of de Finetti's result given by Hewitt and Savage, [11], which is as follows.

Theorem *Let $(Y_1, Y_2, \ldots)$ be an infinite exchangeable sequence of random quantities with values in Ω. Then there exists a probability measure F on the set of probability measures $Q(\Omega)$ on Ω, such that, for each n, and subsets $A_1, \ldots, A_n$ of Ω,*

$$P(Y_1 \in A_1, \ldots, Y_n \in A_n) = \int Q(A_1) \ldots Q(A_n) dF(Q) \tag{1.13}$$

F is the limiting distribution of the empirical measure, i.e. the probability assigned to any set A by F is given by the limit of the probability assigned to the proportion of the first N trials whose outcome is in A.

The exchangeability representation theorem is both surprising and prosaic. It is surprising, in the sense that the simple and natural symmetry judgement of exchangeability over observable quantities leads to such a strong result, namely that our beliefs must be as though we considered that we were making independent draws from a 'true but unknown' distribution Q for which we had assigned a prior measure F. This can be thought of as a version of the separation of our uncertainty into aleatory and epistemic components. Observation of a sample $Y_{[n]} = (Y_1, \ldots, Y_n)$ reduces our uncertainty about future elements of the sequence by applying the Bayes theorem to the prior measure F to update judgements about Q, as

$$P(Y_{i_1} \in A_1, \ldots, Y_{i_m} \in A_m | Y_{[n]}) = \int Q(A_{i_1}) \ldots Q(A_{i_m}) dF(Q|Y_{[n]}) \tag{1.14}$$

for all subsets $A_1, \ldots, A_m$ of Ω, and indices $i_1, \ldots, i_m$ all greater than n. Increasingly large samples tend to a 'relative frequency limit' eventually resolving all of our epistemic uncertainty, leaving the unresolvable aleatory uncertainty as to the outcomes of future draws from a known distribution, a posteriori. Compared to the conceptual confusion at the heart of traditional descriptions of statistical inference, this formulation is clear, unambiguous and logically compelling, building everything on natural belief statements about quantities which are, in principle, observable. The theory of exchangeability is rich and elegant and also of great practical and conceptual importance. This article has only focused on the most basic form for the representation. A characteristic example of the type of results that follow when we impose more structure on the exchangeability specifications is [12] which derives the additive model for log-odds in a two way table from natural exchangeability statements over rows and columns. (The discussion following that article contains some comments from me on the links between this result and the types of limiting finite population representations that we have described above.)

However, the representation theorem is also prosaic, as the population distribution is nothing more than the outcomes of all the possible future observations in the sequence, and the division into aleatory and epistemic components of uncertainty based on this structuring is just a partitioning of our judgements about such future observations. The bucket representation simply gives a concrete form to this identification with finite population sampling, and makes clear the role of exchangeability judgements in equating the observation of the members of the sequence with the random samples from the bucket.

1.6 Expectation as primitive

While the exchangeability representation is highly revealing, the real world implementation of the representation faces two difficulties, one in the construction of the representation and one in its inferential application.

The first difficulty is implicit in the derivation that we have described for the representation theorem. To construct the measure F in representation (1.13), we need to be able to quantify our beliefs for the outcome of the thought experiment comprising the composition of the large bucket with counters indexed by the vector outcomes of the first N members of the sequence. Specifying prior beliefs over the possible choices for this collection is both scientifically difficult,

as we must consider questions at a level of detail beyond our ability to give scientifically meaningful answers, and technically difficult, because of the complexity of the objects over which we are aiming to develop a meaningful probabilistic representation. Therefore, one of the key advantages of the exchangeability representation, namely that it provides a method for us to restrict our belief statements to those related to observable quantities, in practice is usually unfeasible, and so the representation is rarely used in this way.

The second difficulty is as follows. The representation theorem appears to retrieve for us the familiar division into epistemic and aleatory uncertainties, but this division is itself based on an epistemic judgement, which is therefore subject to revision. We aim to use relation (1.14) to update beliefs about future outcomes given a current sample $Y_{[n]}$ by constructing the update $F(Q|Y_{[n]})$, and then deriving beliefs over future outcomes with respect to this distribution. However, the meaning that we ascribe to $F(Q)$ only holds for as long as we judge the sequence as exchangeable. We may change this judgement at any time. Bayesian statistics describes how to make inferences about quantities which have true but unknown values. There is no provision within the Bayesian approach (or any other approach to inference that I know of) for making operationally meaningful inferences about quantities which, at the time when we come to make the inference, may simply cease to exist.

What we need is both to simplify the specification requirements for the exchangeability representation, so that we may use it in practice, as well as in principle, and also to sharpen our formulation for inference to make sense of the issues raised when we make conditioning statements over evanescent quantities. We may address the first issue by changing the primitive for our theory from probability to expectation. To address the second issue requires us to augment our collection of exchangeability specifications, in ways that we shall describe below.

These are larger issues than we can do justice to in the space of this article. All that we will do here is to sketch the key steps that we must take to establish an operational meaning for our inferences over exchangeable quantities, building on ideas first outlined in [6] and [7].

Firstly, we shall discuss a simpler form of exchangeability, based on a different choice of primitive for the theory. Typically, the primitive of choice for the subjectivist theory is probability, but this is largely for historical reasons and to align the theory as closely as possible with its non-subjectivist counterparts. However, we do have a choice and we can, instead, choose expectation as the primitive for quantifying uncertainty. With this choice, we can make as many, or as few, expectation judgements as we wish, when treating a problem of uncertainty, including as many probability statements as we wish—these are simply expectation statements for the corresponding indicator variables. However, when probability is the primitive, we must make all possible probability statements before we can make any expectation statements. (For this reason, expectation was de Finetti's choice of primitive for the theory and the work which best summarized his views, [2], is actually a theory of expectation or, as he terms it, prevision.)

This is not an issue for non-subjectivist approaches—the probabilities all somehow exist separately from us and it is simply our task to learn about them. In the subjectivist theory, we are much more involved. Each uncertainty is a statement that we make, expressing our best judgements as to the likely outcomes. This is exactly the problem that we identified with the exchangeability representation. We need to specify so many probability judgements over observable quantities before we can construct the representation theorem that it is rarely used in this way. The theorem is drained of much of its power by the excessive demands of the probabilistic formalism. We shall now describe the second order version of the representation theorem, which does not suffer from this problem.

De Finetti makes expectation primitive under the operational definition in which $\mathrm{E}(X)$ is the value of x that you would choose if confronted with the penalty

$$L = k(X - x)^2$$

where k is a constant defining the units of loss and the penalty is paid in probability currency (i.e. tickets in a lottery with a single prize). The value of $\mathrm{E}(X)$ is chosen directly, as a primitive, as is probability in the standard Bayesian account. De Finetti shows, under this definition, that $\mathrm{E}(X)$ satisfies the usual properties of expectation, such as linearity. With this penalty scale, expectations are consistent with preferences, in the sense that preferring penalty A to B is equivalent to assigning $\mathrm{E}(A) < \mathrm{E}(B)$, as expectation for the penalty is equal to the probability of the reward.

Bayes linear analysis is a version of Bayesian analysis which follows when we take expectation as primitive; for a detailed account, see [9]. The particular features that are of concern for this article are the practical alternative for the exchangeability representation, which can actually be specified by judgements over observables in practice as well as in principle, and the linkage between this representation and an operationally meaningful form of inference for the evanescent model quantities expressed through the representation theorem. This formalism allows us to address the twin concerns that we have raised about current approaches to statistical induction (and we know of no alternative approach for so doing).

1.7 Second-order exchangeability representation theorem

We say that the sequence of random vectors $X_1, X_2, \ldots$, where $X_j = (X_{1j}, \ldots, X_{rj})$, is **Second-Order Exchangeable** (SOE), if each vector has the same mean and variance matrix and all pairwise covariance matrices are the same, i.e. if

$$\mathrm{E}(X_i) = \mu,\ \mathrm{Var}(X_i) = \Sigma,\ \mathrm{Cov}(X_i, X_j) = \Gamma,\ \forall i \neq j \tag{1.15}$$

We suppose that all quantities in (1.15) are finite. We may separate our uncertainty about each X_i into aleatory and epistemic components, with corresponding second order specifications, according to the following representation theorem, derived in [5].

Theorem (Second-order exchangeability representation theorem) *If $X_1, X_2, \ldots$. is an infinite Second-Order Exchangeable sequence of random vectors, then, for each i,*

$$X_i = \mathcal{M}(X) \oplus \mathcal{R}_i(X) \tag{1.16}$$

where $\mathcal{R}_1(X), \mathcal{R}_2(X), \ldots$ is a mutually uncorrelated second-order exchangeable sequence, each with mean zero and uncorrelated with $\mathcal{M}(X)$.

(The notation $U \oplus W$ expresses the condition that all of the elements of the vector U are uncorrelated with all of the elements of the vector W.)

The proof of the representation theorem is similar to that for the full exchangeability representation. Our thought experiment is to construct a bucket containing N counters, marking the ith counter with the outcome for the ith case. Instead of considering the whole probability distribution of the counters in the bucket, we consider a single quantity, the average of the counters in the bucket, $\overline{X}_N = (\overline{X}_{1N}, \ldots, \overline{X}_{rN})$ where

$$\overline{X}_N = \frac{1}{N} \sum_{i=1}^{N} X_i$$

For the general exchangeability representation, we construct the population distribution from our beliefs relating to the limit of the sample distributions. For the second order theorem, we construct

the population mean $\mathcal{M}(X)$ from our limiting beliefs about the sample means. We can do this because, from the specifications (1.15), the sequence $\overline{X}_i$ is a Cauchy sequence in mean square, as for $n < m$ and each i,

$$\mathrm{E}((\overline{X}_{im} - \overline{X}_{in})^2) = (\frac{1}{n} - \frac{1}{m})(\Sigma_i - \Gamma_i) \tag{1.17}$$

where Σ_i, Γ_i are the ith diagonal terms of Σ, Γ. Therefore the sequence $\overline{X}_n$ tends to a limit, and this limit is the mean quantity $\mathcal{M}(X)$. The properties of the sequence $\mathcal{R}_i(X)$ follow by evaluating $\mathrm{Cov}(\mathcal{R}_i(X), \mathcal{M}(X))$ as the limit of terms $\mathrm{Cov}(X_i - \overline{X}_n, \overline{X}_n)$ and checking that this limit is zero, and similarly for $\mathrm{Cov}(\mathcal{R}_i(X), \mathcal{R}_j(X))$.

We can formalize the construction of this limit, treating expectation as primitive, by constructing the inner product space $\mathcal{I}(X)$ whose vectors are linear combinations of all of the elements X_{ij}, with covariance as the inner product (we identify as equivalent all quantities which differ by a constant) and squared norm given by variance. $\mathcal{I}(X)$ is a pre-Hilbert space for which we may construct the minimal closure by adding limit points for all Cauchy sequences whose limits are not already elements of the space. The inner product over limit points is equal to the limit of the inner product for the associated Cauchy sequence. By (1.17), the sample means form Cauchy sequences, and therefore our specification is consistent with the existence of such limit points, which we identify with the elements of $\mathcal{M}(X)$.

The second-order exchangeability representation theorem is concerned with population mean quantities. It is our choice as to what elements we introduce into our base vectors and therefore what we may learn about from such specifications. For example, we may want to learn about population variation, in which case we must introduce appropriate squared terms into our base vectors, and make exchangeability statements over the corresponding fourth-order quantities. Details as to how to make the appropriate exchangeability specifications, the technicalities of the resulting inferences and the inter-relationship between the adjustment of means and variances are provided in [9].

1.8 Adjusted beliefs

The inner product space described above is the fundamental geometric construct underpinning the Bayes linear approach. The general form of this construction takes a collection U of random quantities, with covariance inner product, and constructs the closure of the inner product space $\mathcal{I}(U)$, denoted $[\mathcal{I}(U)]$. For any quantity $Y \in [\mathcal{I}(U)]$, the adjusted mean and variance of Y, given a data vector D, are defined to be, respectively, the orthogonal projection of Y into the subspace spanned by the elements of D and the orthogonal squared distance between Y and this subspace.

The explicit form for the *adjusted expectation* for a vector B given D, where $D = (D_0, D_1, \ldots, D_s)$, with $D_0 = 1$ is the linear combination $\overline{a}^T D$ where $\overline{a}$ is the value of a that you would choose if faced with the penalty

$$L = (B - a^T D)^2$$

It is given by

$$\mathrm{E}_D(B) = \mathrm{E}(B) + \mathrm{Cov}(B, D)(\mathrm{Var}(D))^{-1}(D - \mathrm{E}(D))$$

(We may use an appropriate generalized inverse if $\mathrm{Var}(D)$ is not invertible.)

The *adjusted variance matrix* for D given D, is

$$\begin{aligned}\mathrm{Var}_D(B) &= \mathrm{Var}(B - \mathrm{E}_D(D)) \\ &= \mathrm{Var}(B) - \mathrm{Cov}(B, D)(\mathrm{Var}(D))^{-1}\mathrm{Cov}(D, B)\end{aligned}$$

An important special choice for the belief adjustment occurs when D comprises the indicator functions for the elements of a partition, i.e. where each D_i takes value one or zero and precisely one element D_i will equal one. In this case adjusted expectation is equivalent to conditional expectation, e.g. if B is the indicator for an event, then

$$\mathrm{E}_D(B) = \sum_i \mathrm{P}(B|D_i)D_i$$

Therefore, the general inferential properties of belief adjustment that we shall describe below are inherited by full Bayes analysis, and this offers a formal interpretation of the real world inferential content of conditional probability arguments.

1.9 Temporal rationality

To understand how subjectivist theory can treat evanescent quantities such as population means, we must first discuss the inferential content of the standard Bayesian argument for observable quantities. This is a large and fundamental issue, which deserves far more space than we can give it here, where all that we will do is to sketch the outline of what is, in my view, the heart of the subjectivist argument.

Firstly, recall the precise meaning of a formal Bayesian inference. If A and B are both events, then $\mathrm{P}(B)$ is your betting rate on B (e.g. your fair price for a ticket that pays 1 if B occurs, and pays 0 otherwise) and $\mathrm{P}(B|A)$ is your current 'called off' betting rate on B (e.g. your fair price now for a ticket that pays 1 if B occurs, and pays 0 otherwise, if A occurs. If A doesn't occur your price is refunded).

This is not the same as the posterior probability that you will have for B if you find out that A occurs. There is no obvious relationship between the called off bet and the posterior judgement at all, and, in my view, no one has advanced an intellectually compelling argument as to why the two concepts should be conflated. The called off bet formulation can, however, be understood within the subjectivist theory as a model for the inference that you will make at the future time.

Models describe how system properties influence system behaviour. They involve two types of simplification, firstly, the description of system properties and secondly the rules by which system properties influence system behaviour. Good models capture enough features of the system that the insight and guidance they provide is sufficient to reduce our actual uncertainty as to system behaviour. This is valuable, provided that we do not commit the modeller's fallacy of considering that the analysis of the model is the same as the analysis of the system. A crucial condition for making good use of a model is to establish the relationship between the model and the actual system, as a basis for making real world inferences.

To derive such a relationship for the Bayesian model, we must make a link between our conditional judgements now and our actual future posterior judgements. This requires a meaningful notion of 'temporal rationality'. Our description is operational, based on preferences between random penalties, as assessed at different time points, considered as payoffs in probability currency.

Current preferences, even when constrained by current conditional preferences given possible future outcomes, cannot require you to hold certain future preferences; for example, you may

obtain further, hitherto unsuspected, information or insights into the problem before you come to make your future judgements, and, always, the way in which you come to learn the information contained in any conditioning event will convey additional information that was not part of the formal conditioning.

These difficulties have no such force when considering whether future preferences should determine prior preferences. Suppose that you must choose between two random penalties, J and K. For your future preferences to influence your current preferences, you must know what your future preference will be. You have a **sure preference** for J over K at (future) time t, if you know now, as a matter of logic, that at time t you will not express a strict preference for penalty K over penalty J.

Our (extremely weak) temporal consistency principle is that future sure preferences are respected by preferences today. We call this

> **The temporal sure preference (TSP) principle** *Suppose that you have a sure preference for J over K at (future) time t. Then you should not have a strict preference for K over J now.*

At first sight, the temporal sure preference principle seems so weak that it can never be invoked, because we will never have a temporal sure preference. However, we actually have many such sure preferences and these are sufficient to determine the inferential content of the Bayesian model, provided we accept the temporal sure preference principle for the problem at hand. It is an interesting philosophical and practical question as to whether and when even this principle is too strong, but for our purposes here it is sufficient to note that this is the weakest principle which is sufficient to give a meaningful account of the content of a Bayesian inference. We will construct the argument for adjusted expectation, the argument for conditional expectation following as a special case, and then consider inference for exchangeable quantities under this formalism.

1.10 Prior inference

For a particular random vector B, suppose that you specify a current expectation $\mathrm{E}(B)$ and you intend to express a revised expectation $\mathrm{E}_t(B)$ at time t. As $\mathrm{E}_t(B)$ is unknown to you, you may express current beliefs about this quantity. Suppose that you will observe the vector D by time t. What information does the adjusted expectation, $\mathrm{E}_D(B)$, offer to you now about the posterior assessment $\mathrm{E}_t(B)$ that you will make having observed D?

We argue as follows. For any random quantity, Z, you can specify a current expectation for $(Z - \mathrm{E}_t(Z))^2$. Suppose that F is any further random quantity whose value you will surely know by time t. Suppose that you assess a current expectation for $(Z - F)^2$. From the definition of expectation, at future time t you will certainly prefer to receive penalty $(Z - \mathrm{E}_t(Z))^2$ to penalty $(Z - F)^2$. Therefore, by temporal sure preference, you should hold this preference now, and so you must now assign

$$\mathrm{E}((Z - \mathrm{E}_t(Z))^2) \leq \mathrm{E}((Z - F)^2) \tag{1.18}$$

Let D be a vector whose elements will surely be known by time t. Let $I(D, \mathrm{E}_t(Y))$ be the inner product space formed by adding the elements of $\mathrm{E}_t(B)$ to $I(D)$. From (1.18), $\mathrm{E}_t(B)$ is the orthogonal projection of B into $I(D, \mathrm{E}_t(B))$ and $\mathrm{E}_D(B)$ is the orthogonal projection of $\mathrm{E}_t(B)$ into $I(D)$.

Therefore, the temporal sure preference principle implies that your actual posterior expectation, $\mathrm{E}_t(B)$, at time t when you have observed D, satisfies the following prior assessments:

$$B = \mathrm{E}_t(B) \oplus S, \mathrm{E}_t(B) = \mathrm{E}_D(B) \oplus R \tag{1.19}$$

where S, R each have, a priori, zero expectation and are uncorrelated with each other and with D.

Therefore, evaluation of $\mathrm{E}_D(B)$ resolves some of your current uncertainty for $\mathrm{E}_t(B)$ which resolves some of your uncertainty for B. The actual amount of variance resolved is

$$\mathrm{Cov}(B, D)(\mathrm{Var}(D))^{-1}\mathrm{Cov}(D, B)$$

We say that $\mathrm{E}_D(B)$ is a **prior inference** for $\mathrm{E}_t(B)$, and therefore also for B. Relation (1.19) holds whatever the context in which the future judgements will be made. Adjusted expectation may be viewed as a model for such judgements which reduces, but does not eliminate, uncertainty about what those judgements should be. This argument is no different than that for the relationship between any real world quantity and a model for that quantity, except that, within a subjectivist analysis, we can rigorously derive the basis for this relationship, under very weak, plausible and testable assumptions.

Note that, if D represents a partition, then conditional and posterior judgements are related as

$$\mathrm{E}_t(B) = \mathrm{E}(B|D) \oplus R$$

where

$$\mathrm{E}(R|D_i) = 0, \forall i$$

with interpretation as above.

1.11 Prior inferences for exchangeable quantities

We now extend the notion of prior inference to the model quantities arising in the second order exchangeability representation, and thus provide an account of the inductive argument relating inferences about the population model and inferences about members of the population. To do this, we must first construct an operational meaning for posterior judgements over model quantities.

Suppose that the sequence of vectors $(X_1, X_2, \ldots)$ is infinite SOE. Suppose that you will observe a sample $X_{[n]} = (X_1, \ldots, X_n)$, by time t. You don't know whether you will still consider $(X_{n+1}, X_{n+2}, \ldots)$ to be SOE at time t. We would like to apply the posterior expectation operator $\mathrm{E}_t(.)$ directly to the exchangeability representation $X_i = \mathcal{M}(X) + \mathcal{R}_i(X)$, by the decomposition $\mathrm{E}_t(X_i) = \mathrm{E}_t(\mathcal{M}(X)) + \mathrm{E}_t(\mathcal{R}_i(X))$. In order to do this, we need to give a meaningful construction for the quantity $\mathrm{E}_t(\mathcal{M}(X))$. This cannot be done directly, as by time t there may be no vector $\mathcal{M}(X)$ to attach the posterior expectation to.

We construct an operational meaning for $\mathrm{E}_t(\mathcal{M}(X))$ by extending the thought experiment in which we construct a bucket marked with counters corresponding to the individual X_i values. For each $i > n$, we additionally record, on the counter marked with X_i, the value $\mathrm{E}_t(X_i)$, so that each counter is marked with a vector $U_i = (X_i, \mathrm{E}_t(X_i))$. Let us suppose that we currently view the sequence U_i as SOE, for $i > n$. This is a comparatively weak constraint. We do not now consider that, at time t, the sequence will necessarily still be exchangeable, but we cannot yet identify any future subsequences about which we already have reason to believe that our future judgements will be systematically different from our judgements over the rest of the sequence.

Therefore, we have the representation

$$U_i = \mathcal{M}(U) + \mathcal{R}_i(U)$$

The first half of the components of U_i consist of the elements of X_i. The remaining components consist of the elements of $\mathrm{E}_t(X_i)$, giving the representation

$$\mathrm{E}_t(X_i) = \mathcal{M}(\mathrm{E}_t(X)) + \mathcal{R}_i(\mathrm{E}_t(X))$$

For any $N > n$,

$$\frac{1}{N-n}\sum_{i=n+1}^{N}\mathrm{E}_t(X_i) = \mathrm{E}_t(\frac{1}{N-n}\sum_{i=n+1}^{N}X_i) \tag{1.20}$$

Taking the limit, in N, of the left hand side of (1.20) gives the quantity that we identify with $\mathcal{M}(\mathrm{E}_t(X))$. The corresponding limit in N of the right hand side of (1.20) is the limit of $\mathrm{E}_t(\overline{X}_n)$, which, as $\overline{X}_n$ tends to $\mathcal{M}(X)$, we equate with $\mathrm{E}_t(\mathcal{M}(X))$. Therefore, we can equate $\mathrm{E}_t(\mathcal{M}(X))$ with $\mathcal{M}(\mathrm{E}_t(X))$. By this construction, we can identify $\mathrm{E}_t(\mathcal{M}(X))$ as a quantity, derived through natural exchangeability judgements, which has the same logical status as the quantity $\mathcal{M}(X)$ itself.

We are now able to integrate model based assessments into our prior inference structure. We have the following theorem.

Theorem (Prior inferences for exchangeable models) *Suppose that, by time t, we will observe a sample $X_{[n]} = (X_1 \ldots, X_n)$ from an infinite SOE sequence of vectors. Suppose, also, that the sequence $U_i = (X_i, \mathrm{E}_t(X_i)), i = n+1, n+2, \ldots$ is a SOE sequence. We can construct the further vector, $\mathrm{E}_t(\mathcal{M}(X))$, which, given temporal sure preference, decomposes our judgements about any future outcome $X_j, j > n$ as*

$$\begin{aligned} X_j - \mathrm{E}(X) = \quad & [\mathcal{M}(X) - \mathrm{E}_t(\mathcal{M}(X))] && (1.21)\\ \oplus\,& [\mathrm{E}_t(\mathcal{M}(X)) - \mathrm{E}_{X_{[n]}}(\mathcal{M}(X))] && (1.22)\\ \oplus\,& [\mathrm{E}_{X_{[n]}}(\mathcal{M}(X)) - \mathrm{E}(\mathcal{M}(X))] && (1.23)\\ \oplus\,& [\mathcal{R}_j(X) - \mathrm{E}_t(\mathcal{R}_j(X))] && (1.24)\\ \oplus\,& [\mathrm{E}_t(\mathcal{R}_j(X))] && (1.25) \end{aligned}$$

(The orthogonal decomposition of (1.21), (1.22) and (1.23) follows by combining the construction for $(\mathcal{M}(X), \mathrm{E}_t(\mathcal{M}(X))$, as the limit of partial means of the quantities U_i, with the relationship (1.19) between each $X_i, \mathrm{E}_t(X_i)$ and $\mathrm{E}_{[n]}(X_i)$, derived from TSP. The orthogonal decomposition (1.24), (1.25) follows from TSP and the orthogonality between the two residual terms and the three mean terms follows as each covariance between an individual residual term and the mean terms must have the same value, by the SOE property of the sequence, and this covariance must therefore be zero, as the limiting average of the residual terms is equivalent to the zero random quantity.)

The above theorem shows that we may treat the vector $\mathcal{M}(X)$ as though it were, in principle, observable, allowing us to decompose our current uncertainty about each $X_j, j > n$, into five uncorrelated components of variation, as follows.

Firstly, our epistemic uncertainty is resolved into three components. We will be uncertain about the value of $\mathcal{M}(X)$, at time t, as expressed by the difference between the expectation, $\mathrm{E}_t(\mathcal{M}(X))$,

that we will express for this quantity and the quantity itself, from (1.21). Secondly, part of our uncertainty (corresponding to (1.23)) about $\mathrm{E}_t(\mathcal{M}(X))$ (and thus about $\mathcal{M}(X)$) will be resolved by the adjusted expectation for $\mathcal{M}(X)$ given $X_{[n]}$, but a part corresponding to (1.22), will be unresolved. Thirdly, this adjusted expectation given $X_{[n]}$ is uninformative for the uncertainty currently treated as aleatory, namely each $\mathcal{R}_j(X)$, about which our future expectation will reduce variation according to (1.25), leaving variation according to (1.24). Whether we will hold this variation to be aleatory at time t will be a subjective judgement that can only be made at that future time.

Each term in this decomposition raises basic practical, methodological, foundational and computational issues. As with the exchangeability representation itself, the prior inference theorem for the representation should be viewed as a starting point, establishing that such a formulation for inductive inference has a natural and operational meaning, based on the careful treatment of each of the five components of variation that we must account for. This is a part of the much wider issue as to the extent to which a Bayesian uncertainty analysis based on a complex scientific model may be informative for actual judgements about the real world; see the discussion in [8].

References

[1] de Finetti, B. (1931). Funzione caratteristica di un fenomeno aleatorio. Atti della R. Academia Nazionale dei Lincei, Serie 6. Memorie, Classe di Scienze Fisiche, Mathematice e Naturale, 4:251–299.

[2] de Finetti, B (1974, 1975). *Theory of Probability*, vol 1, 2, Wiley.

[3] Diaconis, P. (1977). Finite forms of de Finetti's theorem on exchangeability, *Synthese*, **36**, 271–281.

[4] Feller, W. (1966). *An Introduction to Probability Theory and its Applications*, vol II, Wiley, New York.

[5] Goldstein, M (1986). Exchangeable belief structures, *JASA*, **81**, 971–976.

[6] Goldstein, M. (1994). Revising exchangeable beliefs: subjectivist foundations for the inductive argument, in *Aspects of Uncertainty, A Tribute to D.V. Lindley*, 201–222.

[7] Goldstein, M. (1997). Prior inferences for posterior judgements in *Structures and Norms in Science*, M. C. D. Chiara *et al.*, eds, Kluwer, 55–71.

[8] Goldstein, M. (2011). External Bayesian analysis for computer simulators. In *Bayesian Statistics* 9. Bernardo, J. M. *et al.*, eds, Oxford University Press.

[9] Goldstein, M. and Woolf, D. (2007). *Bayes Linear Statistics: Theory and Methods*, Wiley.

[10] Heath, D. and Sudderth, W. (1976). de Finetti's theorem on exchangeable variables, *American Statistician*, 188–189.

[11] Hewitt, E. and Savage, L. J. (1955). Symmetric measures on Cartesian products. *Transactions of the American Mathematical Society*, **80**: 470–501.

[12] Lauritzen, S. L. (2003). Rasch models with exchangeable rows and columns, *Bayesian Statistics*, 7, Bernardo, J. M. *et al.*, eds, Oxford University Press.