Bayesian Inductive Logic

Doctoral Dissertation for Philosophy
Bayesian Inductive Logic
Jan-Willem Romeyn

University of Groningen
Facullty of Philosophy
Oude Boteringestraat 52
9712 GL Groningen

Research supported by NWO, RuG and BCN.

Published 2005.
Printed in the Netherlands by
Haveka BV Alblasserdam.
ISBN 90-367-2279-9

Explanation of cover The cover shows the physical models of two Bernoulli processes. One model consists of a pair of identical cubic dice, both with six possible outcomes. The other model consists of two different dice, a tetrahedron and an octahedron with four and eight possible outcomes respectively. If one of the pairs is selected at random and rolled repeatedly, Bayes' rule can be used to infer which of the two pairs is selected on the basis of the observations of the added outcomes. The two models can thus be told apart even though they have the same range of possible observations, namely 2 to 12.

RIJKSUNIVERSITEIT GRONINGEN

BAYESIAN INDUCTIVE LOGIC

INDUCTIVE PREDICTIONS
FROM
STATISTICAL HYPOTHESES

PROEFSCHRIFT

ter verkrijging van het doctoraat in de
Wijsbegeerte
aan de Rijksuniversiteit Groningen
op gezag van de
Rector Magnificus, dr. F. Zwarts,
in het openbaar te verdedigen op
donderdag 2 juni 2005
om 16.15 uur

door

Johannes Wilhelmus Romeijn
geboren op 13 augustus 1975
te Delft

Promotor:	Prof. dr. T. A. F. Kuipers
Copromotores:	Dr. I. Douven
	Dr. J. Peijnenburg
Beoordelingscommissie:	Prof. Dr. C. Howson
	Prof. Dr. E. C. W. Krabbe
	Prof. Dr. M. van Lambalgen

ISBN 90-367-2279-9

Preface

Writing a thesis is like gardening. Every spring of green ideas must be followed by a careful process of pruning and tripping. Accordingly, the garden that is forever fixed in this book shows only one of the many collections of ideas that I have been cultivating over the years. But I believe it contains ideas that may grow into trees one day, and perhaps even carry fruit.

When I started with the project 'Inductive Rules and the Structure of Evidence', I knew very little to nothing of statistical inference. This has proved to be an enormous advantage. Already after a few months, it became apparent that the criticisms against Bayes' rule that I had envisioned were unsound. The reorientation of the project that resulted from this insight has laid the basis for the research that I carried out in the remaining three and a half years.

The circumstances for gardening have been almost ideal. It has been a time of great intellectual freedom. This is for a large part due to the setting in which I found myself as a PhD in Groningen, and for this I am very grateful. It has made the return to academic life from an occupation in consultancy into one of the best decisions I have ever made. The intellectual freedom may further explain the fact that next to some promising plants and trees, I seem to have nurtured all sorts of weeds over the past years. There is no need for complaining. I have come to the conviction that creative development is often the result of a fearless engagement in error and confusion.

I invite the reader to wander around in the result of some four years of gardening. Many ideas in it will still appear to have only just begun growing, some others may look rather dull and dry. But luckily, as I am writing these last words, spring is in the air.

J. W. R.
April 2005

Acknowledgements

In this thesis I argue that knowledge is always a co-production of an observer and an observed world. In the same way this thesis is itself a co-production of its author and his environment. It is a pleasure to be able to thank some people in this environment for producing this thesis with me.

First of all, I thank Theo Kuipers and Jeanne Peijnenburg for their knowledge and critical reflection during all the stages of the project. Many ideas in this thesis made their first appearance in one of our discussions. Another wonderful stage for sharing ideas was provided by Igor Douven. I am grateful for his comments, his sharp mind and his good advice, from which the whole thesis has benefitted. I am further grateful to the members of the reading committee. I thank Colin Howson for his encouraging words, and for his hospitality during my visit of the London School of Economics. I thank Erik Krabbe, especially for his careful reading of the manuscript. And I thank Michiel van Lambalgen for helpful discussions on my ever diverging research plans.

Many people have been involved in writing this thesis on a more occasional basis, by looking at drafts or discussing a common interest. Alan Hajek, Alexander van den Bosch, Allard Tamminga, Atocha Aliseda, Barteld Kooi, Beerend Winkelman, Boudewijn de Bruin, Branden Fitelson, Conor Dolan, Daan Franken, David Atkinson, David Makinson, Denny Borsboom, Elliot Sober, Frank Hindriks, Franz Huber, Fred Muller, Fred Keijzer, Gabriele Contessa, Hans Rott, Hasok Chang, Hauke de Vries, Hedde Zeijlstra, Herman Philipse, Ioannis Votsis, James McAllister, Jan Albert van Laar, Jeff Seidman, Jeroen van Maanen, Jim Joyce, Johan van Benthem, John Woods, Jon Williamson, Jos Uffink, Julian Reiss, Karin de Boer, Kim van Gennip, Luc Bovens, Maarten van Dyck, Marc van Duijn, Marian Counihan, Martin van Hees, Melle Lyklema, László Pólos, Martine Prange, Menno Rol, Michael Esfeld, Miklos Redei, Nils-Eric Sahlin, Patrick Maher, Peter Grunwald, Peter Lipton, Peter Milne, Pieter Sjoerd Hasper, Prasanta Bandyopadhyay, Richard Bradley, Robert Bishop, Roberto Festa, Roman Frigg, Ruurik Holm, Stephan Hartmann, Talal Debs, and Wouter Meijs, thanks a lot.

During the project I have travelled quite a bit. Firstly, I visited the London School of Economics, which has been an absolutely wonderful experience. I want to thank all those involved in making my visit possible, and for making it

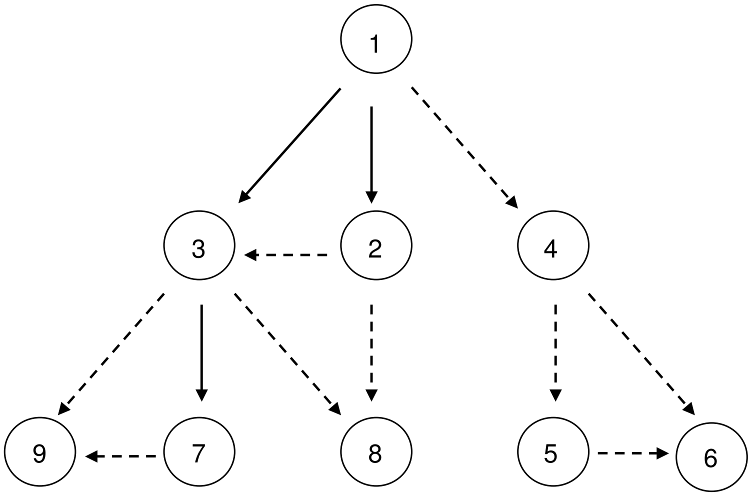
such fun. Further, I have spent a short time in Edinburgh, where Peter Milne made me feel very welcome.

Apart from these two visits I travelled to give talks. I have benefitted from discussions during and after the talks I gave for the GROningen LOGicians, for the Promotion Club Cognitive Patterns, for the ILLC Logic Tea, for the Foundations of Physics Group in Utrecht, for the PhD-seminar of the Erasmus Institute of Philosophy and Economics, for the Philosophy and Probabilistic Modelling Group in Konstanz, for the Belief Revision seminar and the Popper seminar at the LSE, for the Work in Progress seminar in Edinburgh, and for audiences at the Dutch Conference on Philosophy of Science in Leusden, the CENSS conference in Gent, the Workshop on Bayesian Epistemology at the Wittgenstein conference, the Popper Centennial conference in Vienna, the LMPS conference in Oviedo, the VlaPoLo 9 in Gent, the BSPS conference in Kent, the PSA conference in Austin Texas, and the Wiener Kreis lecture of Michael Friedman. I thank all the audiences and academic communities that allowed me to share my ideas, and thus, finally, I thank the reader of these words.

Contents

Outline

It is probably not the intention of the reader to go through this thesis from cover to cover. To accommodate this eclecticism, chapters are set up more or less independently. The technical parts of the chapters show a considerable overlap, but each chapter emphasises other aspects of the technical material relating to the current theme. On the other hand, since complete independence of the chapters leads to too many repetitions, especially in the first part of the thesis, preceding sections or chapters are sometimes presupposed.



The diagram summarises the relations of presupposition by means of arrows. For example, the arrow from chapter 1 to chapter 2 indicates that the latter presupposes the former. If a chapter presupposes only certain sections, references to these sections are included in the chapter itself. The dotted arrows, as for example from chapter 2 to chapter 3, mean that the former may be helpful for a better understanding of the latter, while it is not necessary for understanding the main line.

Two further comments are in order. Firstly, it may be noted from the diagram that the second part of this thesis is quite independent from the first and third part. Even though it may be read without any specialist knowledge, the second part concerns a rather specialist subject within inductive logic. The thesis has been organised so as to allow the reader to leave aside the second part completely, and jump from chapter 3 to chapter 7. Secondly, the introduction and conclusion are not included in the diagram. They provide a general understanding of the position of the chapters within this thesis, and of this thesis as a whole in the debate on inductive logic and its relation to the philosophy of science.

Introduction

> "You will reply that reality has not the least obligation to be interesting. And I will answer you that reality may avoid this obligation, but that hypotheses may not." – J.L. Borges in Death and the Compass

Inductive Inference. This thesis concerns inductive inferences, that is, inferences running from given observations to as yet unknown observations and general observational statements. Inductive inferences are almost everywhere, and come so naturally to us that they easily escape philosophical attention. We drink water because it has always refreshed us, we stride forward confidently because the earth has always attracted and carried us, and we give way to sleep because we have always woken up to renewed presence. In all these cases, the trust that we put in the stability of a pattern can be seen as a trust in the inherent inductive inference. A bit further removed from the backbone of life, these inferences perhaps become more easily recognisable. If the post has been delivered every morning until now, we expect it to be delivered on future mornings too, but if delivery is late on Saturday and Sunday, it will not take long before we only expect it in the morning on weekdays. Here again, the trust in the stability of the pattern is eventually a trust in an inductive inference.

The basic assumption of inductive inferences is that the world is a boring place, and that the same pattern in the observations will keep repeating itself. Usually the sameness is taken as the result of some structure in the world, which is supposed to underpin the patterns in the observations. Fortunately, the world is boring in a rather interesting way. Many observations show patterns that are occasionally violated. For example, it may be that of two equally expensive stalls at the market the fruit of one is usually better than that of the other. But this need not be the case every week. These so-called weak patterns suggest that the world contains a certain structure, but that this structure is perturbed by other structures or effects, which may be deemed noise for the occasion. The one merchant may be better in spotting good fruit at the auctions than the other, but the trade at such auctions always contains an element of chance. Nevertheless we may come to know that the fruit of the one stall is usually better than that of the other. So inductive practice is able to pick up on weak patterns as well.

Inductive inferences are abundant in daily life, and no less so in the daily life of scientists. Much of experimental science concerns the identification of, possibly weak, patterns in the observations. The interdependence between electrodynamic and magnetic forces is an example of a strong, exceptionless pattern. The correlation between vaccination and disease in a population of cows exemplifies a weak pattern, because it only shows in large herds. In general, in experimental science there is a basic trust in the stability of specific patterns, and the typical activity of theoretical science is to motivate this trust by providing a picture or story on the structure behind the stable patterns. Of course, the discussion on structures may get far removed from the observations, but this must not distract from the original intention of experimental science to select stable patterns, and to strip them of noise.

The problem of induction. Once we have realised how deeply both common sense and science are permeated with inductive inference, the destructive force of the problem of induction becomes apparent. David Hume, in his 'Treatise on Human Nature' of 1739, was the first to put his finger on the sore spot. In book 1, part 4, section 2 he writes:

> "Any degree, therefore, of regularity in our perception, can never be a foundation for us to infer a greater degree of regularity in some objects which were not perceived, since this supposes a contradiction, viz., a habit acquired by what was never present to the mind."

So a perceived regularity never allows us to infer a greater regularity, that is, a regularity that includes objects or facts that were not yet perceived. In other words, observations alone can never justify inductive inferences. Any conclusion that transcends observations is not properly inferred from these observations alone, but invokes an additional component, namely the assumption of the stability of some pattern in the observations. Moreover, according to Hume this assumption of stability is a component that cannot itself be derived from the observations. Common sense and science are both resting on nothing more than sheepish habit.

Not surprisingly, this destructive conclusion invited a lively discussion, to which the present thesis is yet another contribution. The essential characteristic of this contribution is its focus on the logical part of the problem of induction, that is, the part that concerns the inferences themselves. I claim that this thesis solves the logical problem, by presenting a scheme for valid inductive inference. It further clarifies the role of the input components of this scheme, and employs the scheme in clearing up some more specific problems concerning inductive inference. Note that the logical part is strictly separated from the epistemological part of the problem of induction, which concerns the input components of inductive inference. The latter part is dismissed as irrelevant to the task of the inductive logician. This perspective resembles the perspective of Howson (2000), but it is more explicitly focused on a specific scheme for inductive inference. The general aim is a revival of inductive logic, and a better control over inductive inference in science.

The remainder of this introduction provides a sketch of the framework within which the inductive logic of this thesis finds its place. It then introduces Bayesian inference, as it is used in this thesis, and discusses the relevance of this thesis for scientific method and the philosophy of science more generally. The introduction ends with an overview of the chapters.

Logical empiricist framework. Let me make the general perspective of this thesis precise. First, as indicated in the foregoing, it is concerned with inductive inference and the problem of induction. But in this context it has a specific aim, and it assumes a particular framework and a particular position therein. As for the aim, it is strictly normative. Inductive practice is only briefly discussed, and only to contrast practice with the norms that this thesis is concerned with. As for the framework, it is that of logical empiricism, as represented by Reichenbach (1935) and most notably Carnap (1950, 1952). In this framework, observations are considered to be clear-cut packages of information, which may be expressed in a formal observation language. Further, inductive inferences are cast in the form of probability judgements over this language. And finally, the inductive inferences are primarily concerned with predictions of single observations. On these three points, the present thesis adopts the framework of Carnap.

Before highlighting the differences with Carnap when it comes to the position of this thesis within the logical empiricist framework, it may be noted that this framework already pushes a number of philosophical positions on the problem of induction out of the picture. Some of those positions stress the theoretical content of observations over and above their empirical content. This surplus value can then be used to derive more from the observations than is warranted by their strict empirical content. The conclusion of this thesis picks up on this line of argument, but considerations on the nature of observations are not part of this thesis itself. Other positions on the problem of induction, such as the so-called structuralist position, rely not so much on the theoretical content of the observations, but on their structural aspects, which may then be connected to a theory on structures behind the observations. The present thesis relates to this possibility only indirectly. Still other positions do not employ the notion of probability, or eschew formal means altogether. Such alternative positions will not be dealt with in this thesis at all, but it may be remarked that certain forms of eliminative induction present a limiting case of the frameworks studied in this thesis.

Positioning this thesis. While the framework of this thesis is basically the one of Carnap, the position that will be developed is very different from the Carnapian position. The next few paragraphs highlight the main point of departure. For Carnapian inductive logic, valid inductive inferences are basically determined by the choice of an observation language. More specifically, probability judgements on future observations, or predictions for short, are derived by means of the notion of logical probability, where logical probability comes down to applying a principle of indifference to the observation language. It is supposed that before obtaining any observations, all the exhaustive descriptions of some system have the same epistemic status, and must therefore be assigned equal probability. On the assumption of this logical probability, both the initial predictions and the effect of accumulating observations on further predictions are determined by the structure of the observation language. The Carnapian idea is thus that the probabilistic predictions are analytic: they follow logically, namely according to logical probability, from the observation language and the preceding observations.

The inductive logic of Carnap was dealt a severe blow when Goodman (1954) proposed a new version of the problem of induction, calling it the new riddle of induction. There is hardly any need to reiterate the famed puzzle for its own sake. However, it provides a convenient way to make explicit the differences between Carnapian logic and the treatment of inductive inference in this thesis. For Carnap, all the work of induction is done by choosing the right predicates for the language, which are in the words of Goodman the projectable predicates. These projectable predicates select the weak or strong patterns that the inductive inferences focus on. If, for example, we choose to employ the predicate ‘green’ in a study on emeralds, we can derive predictions of green emeralds in the future, but if we employ ‘grue’, we can derive predictions of blue emeralds with equal force. Now the new riddle is not damaging for Carnapian inductive logic because the logic allows for crazy predictions, such as those on ‘grue’ and thus blue emeralds. Nobody has ever blamed deductive logic for generating crazy conclusions, since the responsibility for such conclusions lies in the premises. The damaging aspect is rather that Carnapian logic can only start working after a choice of language, and thus of projectability assumptions, has been made. It cannot itself express the choice of projectability assumptions as part of the inductive inference.

This is where the present treatment deviates strongly from Carnapian logic. Generally speaking, logic is concerned with the validity of arguments and not with the truth of conclusions of the arguments: if the premises are true, then so is the logically inferred conclusion, but there is no guarantee to truth if some of the premises are false, or perhaps not even well-formed statements. However, as indicated above, some substantial assumptions of inductive inferences cannot be expressed in Carnapian logic, simply because they are inherent to the observation language and its logical probability assignment. It thus seems that Carnapian logic provides not just valid inferences, but a number of implicit premises as well. Certainly, from the point of view of Carnap these premises are tautological, and therefore do not present implicit premises at all. But the new riddle makes perfectly clear that in fact they do. It is to resolve this seeming conflation of premise and inference that the present thesis presents an alternative logical scheme, following Ramsey (1921), De Finetti (1937) and Jeffrey (1984). It turns out to be perfectly possible to express projectability assumptions in an inductive logic, and thus to separate the part on valid inductive inference from the part on true inductive premises. The failure to disentangle these two aspects, so clearly separated in deductive logic, has obstructed a comparable development of inductive logic.

Other perspectives. At this point it is illustrative to consider an alternative approach to the problem of induction, which has only been touched upon implicitly so far. It is that the problem of induction presupposes a sceptical starting point that need not be accepted. It seems that the destructive conclusion of the problem is immediate once a bare language of observations is put in place: if set apart in that way, it is hardly controversial that the observations do not entail anything about each other. The answer of Carnap, if viewed from this angle, is to deny the sceptical starting point by employing a notion of logical probability over the language, thus creating an inherent dependence between observations. But note that he thereby reacts to the problem of inductive scepticism, or in other words, he is solving a problem in epistemology. Put more dramatically, Carnap seems to attack a problem in epistemology and one in logic at the same time.

Against this, I propose to consider inductive scepticism not as a problem in epistemology, but rather as a philosophical tool in logic. The tool allows us to analyze inductive knowledge in terms of observations and projectability assumptions, which must both be given a place in a scheme for inductive inferences. Moreover, after having settled the issue of valid inductive inference in a logical scheme, there is also a natural way to resolve the epistemological problem of inductive scepticism, by using an externalist theory in which inductive knowledge ultimately rests on the truth of inductive assumptions. I come back to this latter point in the conclusion.

The logical perspective of this thesis must not be mistaken for a rather different view on inductive logic, as developed in Maher (2004) and Fitelson (2005), which is in a sense closer to the initial intentions of Carnap. In this view inductive logic concerns an explication of the strength of the argument running from evidence to a hypothesis, or, in other words, the degree of confirmation that the evidence gives to the hypothesis. The position of Fitelson is that this degree of confirmation is objectively given, but further that it is a three-place function: next to evidence and hypothesis, it must include the probability model on which the confirmation relation supervenes. Unlike Maher, I agree that the probability model must be seen as a separate input component to inductive logic. However, the logic contained in this thesis does not assess the strength of arguments. Instead it simply classifies arguments as valid or invalid, and in this sense it may even be considered deductivist.

The function of Bayesian inference. Let me return to a sketch of the logical perspective of this thesis. In addition to the need for an expression of projectability assumptions as part of inductive inference, a truly logical view on induction is in need of one more thing: an innocent, or epistemically neutral, inference rule. By this I mean a rule that combines projectability assumptions and given observations to produce valid inductive predictions, or more generally, valid probability assignments, without entailing any substantial or synthetic assumptions itself. In a sense, asking for such an inference rule is equivalent to asking for a scheme that brings out all the assumptions that underlie inductive inference. The conclusions of inductive inference derive completely from the input components and the inference rule, so anything that is not implicit to the rule is driven back into the corner of the input components, and anything made explicit as input cannot hide away in the rule anymore. Employing an innocent inference rule seems to provide a natural insight into all input components of inductive inference, as conceived from within the chosen empiricist framework.

At this point, the Bayesian theory of probabilistic inference enters. Bayes' rule, or strict conditioning, prescribes how observations can be incorporated in a probability assignment over an observation language. Many arguments suggest that this rule is innocent in the required way, as long as we assign full certainty to the observations that we have made. The specific aspect of Bayes' rule that is significant here is that its use in incorporating observations induces minimal changes to the probability assignment. In other words, Bayes' rule is maximally conservative. If used to incorporate a specific observation, it takes care that no other change in the probability assignment is induced than those effected by deeming the observation itself certain. Furthermore, along the same lines, the Bayesian theory indeed determines the location of the projectability assumptions. While it is not yet clear in what form these assumptions can be stated, the assumptions must be implicit in the prior probability assignment over the observation language. Thus the Carnapian decision to choose a particular language and use logical probability is in the Bayesian scheme replaced by the decision to adopt a particular prior probability, which encodes the projectability assumptions.

Numerous Bayesianisms. It must be stressed that there is no unique Bayesian theory of inference. There are some common roots and standard texts on inductive inference and Bayesian statistics, most notably De Finetti (1936), Jeffreys (1951), Savage (1956) and more recently Howson and Urbach (1996). But there is certainly not a shared view on what it means to be a Bayesian.

Many of the quarrels among Bayesians come down to three related issues, to wit, the interpretation of epistemic probability, the origin of priors, and the basic form of the axioms of probability. Subjectivists take epistemic probability to be the expression of free personal opinion, and declare this to be the origin of all probability assignments. This point of view is associated with a further defence of the Bayesian theory and its inference rule, based on the relation between probabilities and betting contracts. Objective Bayesians, on the other hand, feel that there are certain rationality constraints on epistemic probability assignments, which may derive from physical probability or some principle of indifference. As for quarrels on the axioms of probability, subjectivism is sometimes associated with empiricist worries concerning probability assignments to opinions that cannot be expressed in finite form, while some objectivists have proposed to replace basic probability assignments with conditional probability assignments.

Bayes' Bayesianism. This thesis falls between all these positions. It employs inferences that have most in common with the inferences first put forward by the reverend Bayes himself. Statistical hypotheses occupy a central place in the original form of Bayesian inference. In the inferences, observations are first reflected in a probability assignment over statistical hypotheses, from which predictions on further observations can be derived. It may be noted that choosing a collection of such hypotheses restricts the probability assignment over the observation algebra. But more importantly, and as will be argued below, the hypotheses are related directly to the patterns in the observations that are considered to be of interest. In this way the hypotheses provide direct access to the projectability assumptions inherent in the prior probability assignment over the observations. The replacement of language in the Carnapian logic with a prior probability can therefore be made more precise. The choice of a range of projectable predicates in the Carnapian scheme can be replaced by the choice of a range of statistical hypotheses in a Bayesian scheme.

All this leads more or less to a middle position in between the above forms of Bayesianism. Statistical hypotheses are taken as so-called tail events in the observational algebra, and are defined by means of limiting relative frequencies. As for the interpretation and origin of priors, it may be noted that the use of hypotheses is connected to the dual nature of probability. On the one hand, the hypotheses pertain to weak patterns in the observations and thus to physical probability, and the restriction on the epistemic probability imposed by the hypotheses thus points to objective probability. On the other hand, the probability assignment over hypotheses is entirely free and reflects personal opinion, so it must somehow be interpreted subjectively. This is so even while it is difficult to connect the probability assignments over hypotheses to betting contracts, simply because statistical hypotheses cannot be tested with finite means. In sum, the Bayesian scheme presented in this thesis leads to a mixture of physical, epistemic, objectivist and subjectivist views on both the interpretation and origin of probability.

This blend of Bayesianism is much more natural than it may now seem. Bayes' original idea is precisely that epistemic and physical probability may be used in the very same inference, and that these two probabilities can coexist peacefully. It is a small step from this to the position that some epistemic probabilities are subjective, whereas others are restricted by physical probabilities and are thus objective, as in Jeffreys' principle of direct probability. Apart from that, the discussion on interpretation and origin loses some of its relevance once we recall that in the present thesis a prior probability assignment is an expression of a premise in an inductive inference. In view of this, both the hypotheses and the priors over them are instruments to express premises. Note that classical deductive logic does not set itself the task to clarify the exact world picture or conceptual interpretation that lays behind a premise or truth assignment. The task of logic only starts after the truth assignment has been given. Similarly, inductive logic need not fix conceptual categories for the probability assignments either. The main task is to investigate the inductive inferences themselves, and the instruments for expressing premises in them. Conceptual categories and interpretations are useful only insofar as they promote that task.

The use of hypotheses. As indicated, the central element in the Bayesian inductive schemes sketched is their use of statistical hypotheses. On this point the present treatment deviates most strongly from the empiricist and subjectivist views of respectively Carnap and De Finetti. Where Carnap localised the projectability assumptions in the choice of an observation language, this thesis makes the projectability assumptions explicit in the statistical hypotheses. Moreover, the hypotheses are introduced as an extension of the Carnapian observation language, as they are defined by means of limiting relative frequencies. Now the representation theorem of De Finetti revealed that statistical hypotheses are redundant in inductive schemes: they can be replaced by exchangeability requirements over the subjective inductive predictions. The present treatment takes the opposite view. It shows that hypotheses, even while they are redundant, are useful tools in expressing inductive assumptions and prior information. They provide a grip on a number of issues in the philosophy of science.

Philosophical import of this thesis. With these remarks on the position of the thesis and its relation to the Carnapian and Bayesian traditions in place, we can zoom out again and look at the overall relevance of this thesis. I first discuss its philosophical import.

With respect to the internal task of clarifying inductive inference, the Bayesian scheme presents a number of advantages over Carnapian inductive logic. A large part of this thesis is dedicated to making these advantages clear. As suggested, the Bayesian scheme provides a way of expressing and controlling the assumptions on the relevance of patterns in the observations. This ability allows us to solve a number of problems in Carnapian logic. A whole package of such problems relates to analogical reasoning. Here the Bayesian scheme allows us to take the package apart, and then to solve part of it. As it appears, this package is intimately related to another, seemingly different package of problems, namely that of encoding relations of probabilistic independence into inductive predictions. Furthermore, the Bayesian scheme is more readily applicable to current themes in the philosophy of science. In particular, it suggests a specific view on the problem of induction, it offers space to model dynamic changes in the projectability assumptions, and it sheds light on the role of theoretical notions in inductive inference.

This brings us to the relevance of this thesis for the philosophical discussion on induction and inductive knowledge. It is most easy to enter this discussion at the point of tension between Carnapian inductivism on the one hand, and the searchlight theory of Selz (1913) and Popper (1959) on the other. The former states that knowledge may be built up by observation alone, while the latter emphasises the importance of conjectures or theoretical starting points before collecting information. Hintikka (1966) made clear that this methodological distinction is not strict. With the perspective and scheme of this thesis, it becomes apparent that there is no methodological tension at all. The notion of conjecture may be combined with the inductivist point of view in a logical scheme, and in this scheme it is even seen to be indispensable. Put in more popular terms, Popper can finally be accepted as a member of the Vienna circle.

From this insight concerning inductivism we can move to the consequences for inductive knowledge, and the related theme of scientific realism. The proposed scheme can be used here to formalise a view that finds its roots already in Kant, and that connects my position with that of Kuipers (2000). It is the view that knowledge can only emerge on the intersection of observation, presented by a mind-independent world, and a conceptual framework, devised by, partly world-independent, minds. I hope that both radical constructivists and hardcore realists take this constructive realist message to heart.

Relevance of this thesis for science. Regarding the relevance of this thesis for science, first note that the empiricist framework accords well with the statistical inductive inferences of experimental science. In almost all experimental cases, weak patterns in observations are dealt with by means of statistics, and it is on these kind of inferences in experimental science that this thesis has its bearing. In this context, the aim of the thesis is not so much descriptive but normative, and more precisely, passively normative. The claim is not that scientists must, in all their investigations, follow the scheme laid down in this thesis. There may be practical reasons for using other procedures. However, scientists must eventually be clear on the exact inductive inference that they are making, and this they can find out by writing their procedures down in terms of the scheme provided here. In other words, they can check the validity of their procedures by writing them down in a Bayesian form.

This passive form of normativity indicates how the thesis relates to inductive inferences performed by means of classical statistics, as presented in Cramèr

(1946), Mood and Graybill (1973), Barnett (1999) and numerous other textbooks. It is well-known that classical statistics faces a number of paradoxes, some owing to the base-rate fallacy, others owing to the failure to respect the likelihood principle. But it is also well-known that procedures from classical statistics sometimes provide practical solutions where Bayesian statistics remains silent. Moreover, as far as the procedures of classical statistics are indeed inferences, they do not necessarily lead to false conclusions. It would therefore be misguided to advise scientists not to use classical statistics. On the other hand, the inferential steps in the classical procedures are often elliptic, or in other words, incomplete, and experimental scientists may not always be aware of the things they are presupposing when using these procedures. Classical statistics provides inferential shortcuts, whose applicability simply varies from case to case. The Bayesian reformulation of classical procedures can help to determine their applicability.

There is an enormous amount of literature on statistics, and many of the points made in this thesis have in some form or other been made elsewhere. Apart from the benefit of repeating the truth from time to time, the reader may wonder what innovations this thesis offers in the field of statistics. One innovative aspect is the connection of Bayesian inference with the Carnapian programme, and specifically, the use of statistical hypotheses in solving problems on analogy and inductive dependence. Another innovative aspect concerns the relation between Bayesian inference and problems with theory change and underdetermination. But perhaps the most important innovation is the use of frequentist chances in the Bayesian scheme. Statistical hypotheses can therefore be seen as part of an extended observation language, which allows for the integration of empiricist, subjectivist and frequentist views on probability.

Overview of chapters. I will now briefly run through the chapters, and indicate how their contents link up with the topics discussed in this introduction. The first three chapters form the first part of the thesis. This part concerns the reformulation of Carnapian inductive logic in terms of Bayesian logic and the improvement of the latter logic by the explicit use of hypotheses. Chapter 1 presents Carnapian inductive predictions as the conclusion of valid inductive inferences, and contrasts these with predictions deriving from the Bayesian scheme. Chapter 2 then deals with the nature and use of statistical hypotheses, and in particular with a frequentist semantics for hypotheses and the fundamental change they present to the Carnapian scheme. Chapter 3 argues that statistical hypotheses provide access to the projectability assumptions in the inductive inferences.

Chapters 4, 5 and 6 form the second part of the thesis. They concern the use of the Bayesian scheme in solving two problems in inductive logic, namely that of analogical predictions and that of causal relations or correlations between predicates. The general idea is that hypotheses provide a convenient handle on inductive dependencies between predicates, which prove hard to capture in terms of direct prediction rules. Chapter 4 shows that a natural system of prediction rules for capturing so-called explicit analogical predictions can be understood as the result of transforming a certain space of hypotheses. In chapter 5 these transformations are employed further to include analogical predictions of any kind, but unfortunately an exact match between the resulting predictions and the classification of analogical effects cannot be derived. Finally, chapter 6 employs the very same techniques to tackle the seemingly different problem of inductive inference for Bayesian networks. The mathematical structure of the problem turns out to be exactly the same.

The last part of the thesis is much smaller. It contains three short chapters on the Bayesian scheme in relation to venerable themes in the philosophy of science: the problem of induction, the problem of new theories and theory change, and the problem of underdetermination, which relates to abduction. It will be shown that these problems in methodology can be elucidated with the Bayesian scheme. In particular, chapter 7 investigates in what sense the Bayesian scheme solves the problem of induction. Chapter 8 proposes an addition to the Bayesian scheme that enables us to incorporate changes in the partition of hypotheses that are used in an inductive inference. Chapter 9 concerns the use of theoretical hypotheses in inductive inferences, and in this way provides a first sketch of a Bayesian model for abductive inference. The thesis ends with some general conclusions, and a perspective on further research.

I
Bayesian Inductive Inference

1

Carnapian and Bayesian Inductive Predictions

The first part of this thesis is concerned with a general introduction into inductive predictions, and with the development of a logical scheme for these predictions. This chapter, in particular, discusses two schemes for capturing inductive predictions, the Bayesian and the Carnapian. The focus of the discussion is on their logical nature. After an introduction into both schemes, they are reformulated to disentangle the notions of premise and inference. The resulting picture shows both schemes in the form of a Bayesian logical argument.

1.1 Introduction

A hunting example. The schemes discussed in this chapter are aimed at a conceptual clarification of inductive predictions. Inductive predictions are taken to be probabilistic statements on future events, given a set of past observations of events and further relevant knowledge. As an example, take a series of observations made of a pond in the middle of a forest. We may want to predict the animals to be observed in or around the pond, and it may be sensible to base this prediction on the record of earlier animals spotted, and on further knowledge we have of the animals in the forest. The necessity of accurate predictions is perhaps made more vivid if it is added that we can spot an empty pond, some tasty ducks, but possibly a hungry tiger, so that we must prepare for hunting or hiding.

Since we cannot be certain whether the next observation will be of a duck or a tiger, we naturally attach degrees of belief to these events. For example, if we have spotted many ducks, we may be tempted to expect spotting more ducks in the future, based on the supposition that the record of observations is representative for the numbers of ducks and tigers in the forest. Similarly, if we know the pond to be a place that is regularly visited by ducks, the initial degree of belief for spotting a duck must be high. However, because ducks usually flee from tigers, spotting a single tiger may cause us to expect no ducks for some time. The two schemes to be discussed in this chapter serve to clarify exactly such predictions of uncertain events based on a record of earlier events, data for short, and other relevant knowledge.

The logic of inductive predictions. Before presenting the schemes, I need to say a bit more on the kind of clarification that this chapter offers. This is particularly useful for those readers who have skipped the introduction of this thesis. For those who have read the introduction, it may serve to connect the introduction of the thesis to the agenda of this chapter. The two schemes discussed in this chapter look at predictions as the result of inductive inference. So clarifying these predictions involves not only a formal expression of the predictions themselves, but also of the inductive inference steps leading up to them. A scrutiny of inference is generally the task of logic, and the schemes are thus concerned with the logic of inductive predictions.

The fact that the object of the schemes is logic rather than rationality or decision theory is of importance to the kind of clarification that they offer. Let me take classical deductive logic to be the paradigmatic case, and the case that sets the standards. This logic makes a crucial distinction between validity and truth: valid reasoning need not necessarily result in a true conclusion, and a true conclusion may follow from invalid reasoning. Logic is restricted to characterising valid reasoning, and leaves the matter of truth aside. Classical deductive logic therefore enables us to make valid arguments for the most silly of claims, for example for the claim that ducks love tigers: if only we assume that ducks love furry animals, and further that tigers are furry, then the conclusion that ducks love tigers follows unproblematically. This is not to say that classical logic holds that ducks love tigers. The logic just relates the premises to a conclusion, but refrains from telling whether the conclusion is true. It only claims that if the premises are true, then so is the conclusion.

Now it may be argued that assuming the rules of classical logic to be truth-preserving or sound comes down to making substantial metaphysical claims, because it is in a sense a contingent fact about the world that these rules indeed work. But it carries us too far into the philosophy of logic to follow that line. Here it is important only to keep in mind the distinction of validity and truth. It is the particular perspective of validity that is at the core of the clarification that the schemes in this chapter offer. In the following I attempt to maintain a strict separation between premises, which the schemes take as input, and the inferences, which are inherent to the schemes. These inferences must be free of synthetic assumptions, because, parallel to classical logic, their only task is to take us from the premises to the conclusion, in this case the predictions, in a valid way.

Ampliative premises. The adoption of the logical perspective has some consequences for the way in which the schemes will be characterised and evaluated. It is sometimes claimed that inductive inference is inherently ampliative, because the conclusions say more than what the data warrant. For example, on the basis of a long morning of observing an empty pond, we are tempted to conclude that we will not see the ducks or the tiger in the afternoon either. This reasoning is ampliative because the data do not themselves contain such a conclusion. But in view of the above, I will not say that the inferences of inductive logic are ampliative, but rather that inductive logic takes more input than just the data. It also takes as input particular premises on what the data purport to tell us. In other words, inductive logic as a whole may be called ampliative, but the ampliativeness resides not in its inferences, but in the fact that the logic takes not only data as input but also, as it were, ampliative premises. In the example, the ampliative premise is that a particular regularity in past observations, namely there being no ducks or tigers, is projectable onto future observations. The main point on which the two schemes will be evaluated is precisely on how they make these ampliative premises explicit.

It may strike the reader as disappointing that a conceptual clarification of inductive predictions remains silent on the most interesting parts, namely on what a good inductive prediction is, and what kind of premises good inductive predictions take as input. I admit that a full analysis of inductive reasoning must also contain an assessment of the premises and predictive performance, and that in this sense the present chapter, and more generally this thesis, is incomplete. One of the main ideas of this thesis is, however, that these questions are separate ones, and that it makes sense to consider the logic of inductive predictions on itself, and independently from the justification of its premises or the performance of its predictions. If a scheme for clarifying inductive reasoning is to function as a logic, it must in fact be neutral towards any ampliative assumptions we may want to make. I here side with Howson (2000), who argues that to solve the problem of induction, we have to provide a logic which tells us how to relate assumptions, for example on the uniformity of nature with respect to some predicate, to the empirical data. It is not part of the task of justifying inductive inference to advise simultaneously about what premises to use.

Deviating from Carnap. Note that this perspective on inductive reasoning differs significantly from the perspective that Carnap (1950) chooses in his analysis of induction. It differs in at least two ways. First, the present discussion focuses explicitly on logical arguments, and not necessarily on a confirmation relation. It is quite natural to place the inductive arguments in a larger confirmation framework, but there is much more to confirmation than is expressed in the schemes of this chapter, particularly when it comes to the confirmation of scientific theories.

Second, we may say that Carnap deems inductive predictions tautological once an observation language is chosen. Put more carefully, he provides an observation language, and then derives so-called logical probabilities over this language from specific symmetry considerations. These logical probabilities fix the predictions relative to preceding observations, and are supposed to explicate valid inductive reasoning. Carnapian inductive predictions are therefore determined by the observation language and the preceding observations. By contrast, classical deductive logic enables us to choose premises after a language has been chosen. In the perspective of this thesis, Carnap therefore delivers the logic of inductive predictions together with the premises it takes as input.

The following is intended as an improvement on the Carnapian project in two ways. First, I define a scheme for valid inductive inference, and second, I provide tools to capture the notion of inductive premise. This latter task extends to the next two chapters. For this chapter the plan is as follows. In section 1.2 I introduce the formal notions of data, degree of belief, and predictions, after which I introduce the Carnapian scheme, and the $\lambda\gamma$ of Stegmüller rule as a typical example. The Bayesian scheme is introduced in section 1.3. Its exposition includes a discussion on hypotheses, updating over hypotheses, and priors. Section 1.4 introduces the notion of a probabilistic model, shows that both the Carnapian and the Bayesian scheme take such models as input, and thus sketches a logical picture for these schemes. The Bayesian scheme can then be seen as a generalisation of the Carnapian scheme.

1.2 Inductive predictions

This section introduces the formal framework for inductive predictions, which is used in both schemes. After that it defines the simplest of the two schemes for making inductive predictions, the Carnapian scheme. This scheme is illustrated with the so-called $\lambda\gamma$ continuum.

1.2.1 Observation framework

The framework for characterising observational data is based on set theory. Such a framework most easily accommodates a characterisation of degrees of belief in terms of the probability theory of Kolmogorov (1933). I first define a finite set-theoretical framework based on observation statements, after that a framework with so-called cylinder sets, and finally an infinite extension of this latter framework. Both frameworks using cylinder sets are employed below.

Observation language. Let K be a finite set of possible observations at any time, typically $\{0, 1, 2, \ldots, L\}$ and let q_i be a single observation q at time i, so that $q_i \in K$. Define e_t to be a finite string of indexed observations, $e_t = q_1 q_2 \ldots q_t$, and K^t the t-th Cartesian product of K, and thus the set of all strings of length t. Finally, define $e_t(i)$ to be the term q_i in the string e_t. For example, the possible observations may be an empty pond, some ducks and a tiger, encoded in the numbers $q = 0, 1, 2$, whereby it is assumed that tigers and ducks never appear together. The observations $e_6 = \langle 0, 0, 0, 1, 1, 0 \rangle$ then mean that the pond is empty for three time units, after which ducks settle in the pond during two time units, only to leave again after that. Thus we have a complete specification of the finite set-theoretical framework of observations, in which single observations or finite observation strings are elements in the sets K and K^t respectively.

This framework is finite because it does not include infinitely long sequences of observations. It may be noted that the finite framework can easily be associated with the observation language that Carnap (1950) used. The q_i can be seen as statements that the observation at time i had the result q, and strings of observations e_t can be taken as conjunctions of such observation statements. This chapter, however, does not exactly use such a statement-based framework. Instead it uses a framework with so-called cylinder sets, denoted with capital letters. The idea behind this framework is that the notions representing single observations and finite observation strings are not individual elements, but subsets in the set of all infinitely long strings of observations. In defining probabilities over sets representing subsequent observations, this framework is much more convenient.

Cylindrical algebra. Let K again be the set of all possible observations, and let q_i be a single observation at time i. Define e to be the infinite string $e = q_1 q_2 q_3 \ldots$, and K^ω the set of all such infinitely long strings. Within K^ω, define the set $E_t^{e_t}$ to be the set of all strings e that start with the string e_t and may contain any infinite string after that:

$$E_t^{e_t} = \{e : \forall i \leq t \ (e(i) = e_t(i))\}. \tag{1.1}$$

Also define $Q_i^q \subset K^\omega$, the set of strings e that start with any string e_{i-1}, have the same result q for observation i and may contain any infinite string after that:

$$Q_i^q = \{e : e(i) = q\}. \tag{1.2}$$

In the following, the sets Q_i^q and $E_t^{e_t}$ represent observations and strings of observations respectively. For the expression $E_t^{e_t}$ I usually omit reference to the string e_t for sake of brevity. The small letters q_i and e_t encode the content of the observations. Note that the numbers and the sets are different mathematical entities, just as the event of observing q at time i, denoted Q_i^q, is different from the content of that observation, namely q.

The sets Q_i^q can now be collected in an algebra $\mathcal{Q}_0$, which consists of all sets that may be constructed by any finite number of intersections and unions of these observations. The algebra $\mathcal{Q}_0$ will be called an observation algebra. Note that the sets E_t can be written down as repeated intersections of single observations Q_i^q:

$$E_t^{e_t} = \bigcap_{i=1}^{t} Q_i^{e_t(i)} \tag{1.3}$$

This also means that $E_{t+1} \subset E_t$, as is illustrated in figure 1.1. The sets E_t are sometimes called cylinder sets. As soon as we make the observation Q_{t+1}^q, the cylinder E_t, comprising of sequences e that agree on e_t and that contain any infinite string after that, is narrowed down to the smaller cylinder E_{t+1}, in which the sequences e only diverge after e_{t+1}. So the sets E_t are really like nested cylinders. It may further be noted that the observations Q_i^q, as it is called, separate $\mathcal{Q}_0$: for any pair $e \neq e'$ there is at least one Q_i^q for which $e \in Q_i^q$ while $e' \in Q_i^q$. Each infinite sequence, or possible world, e can therefore be told apart by some observation. Note finally that the sets Q_i^q and E_t can also be written down directly in terms of the finite framework, as $K^{i-1}qK^\omega$ and e_tK^ω, which is the notation used by Kuipers (1978).

It is important to be clear on how the finite and the infinite set-theoretical frameworks for observations relate. An observation result q_i is a variable which can take on a value in K, while an observation Q_i^q refers to a set of infinitely long sequences of observations e which have the value $e(i) = q$ in common. However, when it comes to expressive force, the algebra $\mathcal{Q}_0$ of sets Q_i^q is simply equivalent to the finite framework with small letters. The so-called σ-algebra

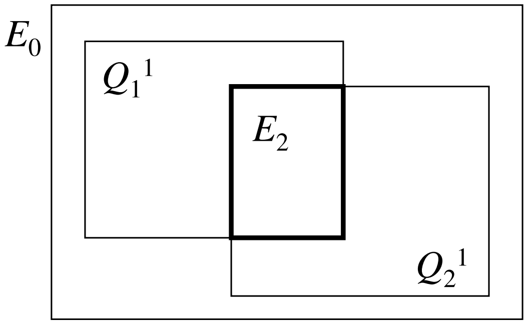


Figure 1.1: Graphical representation of the observation sets Q^1_1, Q^1_2, and E_2 as part of a cylindrical algebra.

of sets Q^q_i, denoted $\mathcal{Q} = \sigma(\mathcal{Q}_0)$, is the extension of $\mathcal{Q}_0$ that also contains sets that can only be obtained by infinitely many set-theoretical operations with elements from $\mathcal{Q}_0$. The algebra $\mathcal{Q}$ is called the extended observation algebra, because it is essentially richer than the algebra $\mathcal{Q}_0$. In this chapter I employ only the algebra $\mathcal{Q}_0$. The richer algebra $\mathcal{Q}$ will be used in subsequent chapters.

1.2.2 Carnapian scheme

Probability assignments representing beliefs. I can now define the formal notions of belief and prediction in this framework. As suggested, beliefs are represented by a probability function over the algebra of observations $\mathcal{Q}_0$. The probability function

$$p_{[e_t]} : \mathcal{Q}_0 \mapsto [0, 1] \tag{1.4}$$

represents the beliefs of some observer who is given the observations e_t, that is, whose empirical knowledge exactly comprises these observations. Every empirical knowledge base e_t is connected to a unique set of beliefs expressed in $p_{[e_t]}$. Note that the functions $p_{[e_t]}$ are defined over all elements of the algebra $\mathcal{Q}_0$, and conform to the Kolmogorov probability axioms. Popper, Renyi (1970) and more recently Hájek have argued that conditional probabilities are more natural as basic elements than unconditional ones, and that in fact the axioms must be rewritten to define probability as a two place function. In view of the later use of statistical hypotheses, this may be a useful reformulation.

It is natural to suppose that an observer who has made the observations e_t assigns a probability 1 to the corresponding set $E_t \in \mathcal{Q}_0$, that is,

$$p_{[e_t]}(E_t) = 1. \tag{1.5}$$

This is a particular regularity condition, to which I shall adhere in all of the following. As will be seen below, it is a stronger condition on the functions $p_{[e_t]}$ that they are related by Bayes' rule, as the above regularity condition follows from that. However, as it stands now, the representation leaves room for other relations between subsequent belief states. I come back to the use of Bayes' rule in sections 1.3.2 and 1.4.

Inductive predictions. With the above representation of beliefs in place, we can define inductive predictions. A prediction of some observer given some set of observations is basically any assignment of a probability to an observation that is not already entailed or excluded by the data, that is,

$$p_{[e_t]}(Q^q_{t+i}) = pr \tag{1.6}$$

for any $i > 0$. Since $p_{[e_t]}$ is a probability function, we have $0 \leq pr \leq 1$. Most of the discussion in this chapter focuses on predictions for which $i = 1$, which I call direct predictions. Direct predictions concern the observation immediately following the given observations, for example whether the next observation is of an empty pond, of ducks, or of a tiger.

With the notion of inductive prediction at hand, we can define Carnapian prediction rules. As said, the scheme uses the observation algebra $\mathcal{Q}_0$ and the probability functions $p_{[e_t]}$. The further component of a Carnapian scheme is a valuation of all direct predictions, denoted $p_{[e_t]}(Q^q_t)$:

$$p_{[e_t]}(Q^q_{t+1}) = pr(q, e_t). \tag{1.7}$$

The valuation of the direct predictions, $pr(q, e_t)$, can also be called a prediction rule. Note that pr is a function of q and e_t, which are natural numbers in K and strings of such numbers in K^t respectively. The function pr determines a probability assignment $p_{[e_t]}$, so we must have

$$\sum_{q \in K} pr(q, e_t) = 1. \tag{1.8}$$

Furthermore, the valuation must be complete, meaning that every combination of q and e_t is assigned a value by pr. In chapter 2 this restriction is discussed in more detail.

Carnapian scheme. Carnap was the first to study such prediction rules pr systematically, considering them over a finite observation language containing the terms e_t and q_{t+1}. He developed specific prediction rules in his (1950, 1952), using his notion of logical probability. Stegmüller (1973) eventually derived the so-called $\lambda\gamma$ continuum of inductive methods:

$$pr_{\lambda\gamma}(q, e_t) = \frac{t_q + \gamma_q \lambda}{t + \lambda}. \tag{1.9}$$

The parameter t_q is the number of results q in the sequence e_t. If for example $K = \{0, 1\}$, and e_3 is given by $\langle 0, 1, 1 \rangle$, then $t_0 = 1$ and $t_1 = 2$. The parameter t is the length of the sequence e_t, and can thus be called the time parameter. The special case of $\lambda = L$, the number of possible observations q, and $\gamma_q = 1$ for all q is the rule discussed by Laplace, and eventually dubbed the straight rule by Reichenbach (1948). As Zabell (1982) shows, Johnson (1921) already derived the continuum as a generalisation of the straight rule.

Rewriting the above expression provides a better understanding of the function of the parameters λ and γ:

$$pr_{\lambda\gamma}(q, e_t) = \left(\frac{t}{t + \lambda}\right) \frac{t_q}{t} + \left(\frac{\lambda}{t + \lambda}\right) \gamma_q. \tag{1.10}$$

The $\lambda\gamma$ rule can thus be seen as a mixture of an initial probability γ_q for q, and the observed relative frequency $\frac{t_q}{t}$, regulated by the factors $\frac{t}{t+\lambda}$ and $\frac{\lambda}{t+\lambda}$. At $t = 0$ we have $\frac{t}{t+\lambda} = 0$, so that the initial probability p is determined entirely by γ_q. At $t = \lambda$ the value of θ is just the mean of γ_q and $\frac{t_q}{t}$, and for $t \gg \lambda$ the import of γ_q vanishes. Thus λ is a parameter that expresses the willingness of the observer to learn from the observations.

There is much more to be said on the derivation of the Carnapian prediction rules from the notion of logical probability. But I will not deal with these derivations here. The main aim of this chapter is to put forward a rather different position on inductive logic, while using the basic Carnapian framework. Apart from Carnap (1950) and the more specific (1952), there are excellent discussions of Carnapian inductive logic which the reader may consult. The standard works are Carnap and Jeffrey (1971) and (1980). See Suppes (2002: 190-98) for a quick reference, and Schilpp (1963), Kuipers (1978), and Festa (1993) for more specific discussions.

This thesis does not use the notion of logical probability, for reasons given in 1.1 and in the introduction to this thesis. In the context of this thesis, a Carnapian scheme is therefore characterised by the function $pr(q, e_t)$. Apart from the fact that this function must be normalised and must offer a complete valuation, any function $pr(q, e_t)$ is permitted. The function pr may be a Carnapian rule as defined above, in which case repeated observations of some q enhance the probability for q's in subsequent observations. But the function pr may also express the gambler's fallacy, in which repeated observations of some q cause the probability for subsequent q's to decrease. The Carnapian scheme, as it is presented in this thesis, does not yet impose restrictions on the kind of patterns in the observations that the inductive predictions are aimed at. Considered as such, the Carnapian scheme is just one way of conceptualising inductive predictions. The next chapter introduces an alternative way.

1.3 Bayesian scheme

The Carnapian scheme is an attractive scheme for capturing inductive predictions. It is simple, and it seems that any prediction rule can be subsumed under it. The other scheme of this chapter, which I call the Bayesian scheme, is a bit more complicated. It employs the same framework for observations, but to arrive at predictions it takes a detour. First it uses Bayesian updating to assign probabilities to hypotheses on the basis of input probabilities and observations. The probability distribution over a suitable collection of hypotheses can then be used to generate inductive predictions.

1.3.1 Predictions from statistical hypotheses

Statistical hypotheses. This section concerns the use of statistical hypotheses in making predictions. Statistical hypotheses are here taken to be statements that induce a complete probability assignment over an observation algebra. In the following, whenever I speak of hypotheses, I intend to refer to statistical hypotheses of this type. To illustrate, recall the hunting example and consider the statistical hypothesis h that a tiger is hunting the ducks in the pond just like we are. This fact may be described by the following set of statements, namely that the tiger appears with a chance of $\frac{1}{2}$ if ducks are in the pond and disappears directly after that, that the tiger hides in all other cases, that the ducks do not appear for a while after the tiger appeared, and that there is otherwise a constant chance $\frac{1}{3}$ of ducks appearing or staying in the pond.

We can now construct a valuation of probabilities $pr_{[h]}$ that captures the statistical hypothesis h that a tiger is hunting ducks. We use the above statements and the encoding of empty pond, ducks and tiger in $q = 0, 1, 2$ respectively:

$$pr_{[h]}(1, e_t) \quad = \quad \begin{cases} \frac{1}{3} & \text{if } e_t(t) \neq 2, \\ 0 & \text{otherwise,} \end{cases} \tag{1.11}$$

$$pr_{[h]}(2, e_t) \quad = \quad \begin{cases} \frac{1}{2} & \text{if } e_t(t) = 1, \\ 0 & \text{otherwise,} \end{cases} \tag{1.12}$$

$$pr_{[h]}(0, e_t) \quad = \quad 1 - pr_{[h]}(1, e_t) - pr_{[h]}(2, e_t). \tag{1.13}$$

The above cases cover all sequences e_t, so that the probability assignment is indeed complete. Because of prediction (1.13), normalisation is also satisfied.

A single statistical hypothesis prescribes a single prediction rule $pr_{[h]}$ over the observation algebra $\mathcal{Q}_0$. But the Bayesian scheme is designed to deal with a collection of such statistical hypotheses, minimally two. In the example, take the hypothesis $h_1 = h$ that tigers hunt ducks and the hypothesis h_0, stating that tigers and ducks wander independently, but that ducks appear nine times more often than tigers, but that the pond is ten times more often empty than filled with ducks. The above stipulations already specify the probabilities over the observations according to the former hypothesis, namely $pr_{[h_1]} = pr_{[h]}$. For the hypothesis h_0 we must choose

$$pr_{[h_0]}(q, e_t) = \begin{cases} \frac{9}{10} & \text{if } q = 0, \\ \frac{9}{100} & \text{if } q = 1, \\ \frac{1}{100} & \text{if } q = 2. \end{cases} \tag{1.14}$$

Many other such hypotheses can be constructed. For example, hypotheses may cover the statement that tigers operate alone, that ducks are reluctant to leave the pond, that when we see a duck a tiger will appear soon, and so on. As with Carnapian prediction rules there are only two restrictions. First, the hypothesis must respect the normalisation condition (1.8). And second, they must each cover all possible preceding observations e_t, so that they determine direct predictions over the whole observation algebra.

Predictions from a partition. The Bayesian scheme considers collections of hypotheses $\mathcal{H} = \{h_0, h_1, \ldots, h_N\}$. For reasons that will become apparent later, such collections are called partitions. Instead of defining beliefs with probability functions over a single observational algebra $\mathcal{Q}_0$, the Bayesian scheme defines beliefs over the product of partition and observational algebra:

$$p_{[e_t]} : \mathcal{H} \times \mathcal{Q}_0 \mapsto [0, 1]. \tag{1.15}$$

Here every hypothesis is associated with its own observational algebra, $H_j = \{h_j\} \times \mathcal{Q}_0$, and its own direct predictions over this algebra. In this sense a Bayesian scheme is really a generalisation of the Carnapian scheme. In section 2.5 it will become apparent that the Bayesian scheme can also be seen as a special way of defining a Carnapian scheme.

In the Carnapian scheme, predictions are determined by a single prediction rule, $p_{[e_t]}(Q^q_{t+1}) = pr(q, e_t)$. In the Bayesian scheme, by contrast, the predictions are determined by the law of total probability, applied to a partition of statistical hypotheses:

$$p_{[e_t]}(Q^q_{t+1}) = \sum_j p_{[e_t]}(H_j) p_{[e_t]}(Q^q_{t+1}|H_j). \tag{1.16}$$

To make predictions, we therefore need the probabilities for all the hypotheses H_j in $\mathcal{H}$, denoted $p_{[e_t]}(H_j)$, and the direct predictions associated with these hypotheses, denoted $p_{[e_t]}(Q^q_{t+1}|H_j)$. The remainder of this section makes these two input components precise.

The direct predictions associated with statistical hypotheses are called the likelihoods of the hypotheses. These terms are given by the probability assignment over the observations according to the statistical hypotheses:

$$p_{[e_t]}(Q^q_{t+1}|H_j) = pr_{[h_j]}(q, e_t), \tag{1.17}$$

The likelihoods of the hypotheses, which are defined for certain observations, are thus the probabilities, according to the hypotheses, of these observations. Note that every hypothesis represents a separate Carnapian scheme, and that the Bayesian scheme takes a range of such schemes as input. Again, these prediction rules are not restricted to the Carnapian rules, which may be derived from applying the notion of logical probability to the observation language. Hypotheses can be chosen freely, and so can the likelihoods associated with them.

1.3.2 Probability assignments over hypotheses

The following discusses Bayesian updating over hypotheses. See Jeffrey (1984) and Howson and Urbach (1996) for more details.

Updating over a partition. The probability over hypotheses after e_t, which also figure in the prediction (1.16), require more elaborate discussion. These terms are not immediately given, but they can be worked out by means of Bayesian

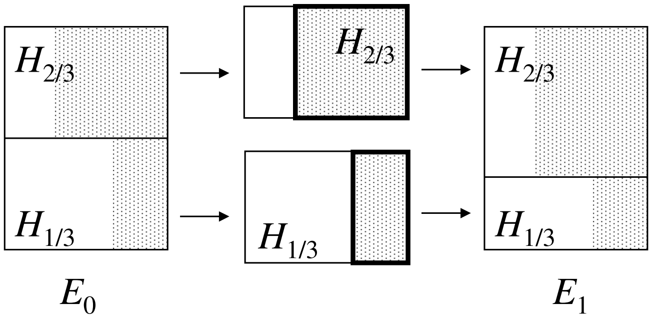


Figure 1.2: A graphical representation of an update of the probability assignment over the two hypotheses $H_{\frac{1}{3}}$ and $H_{\frac{2}{3}}$, for the observation $E_1 = Q_1^1$. The areas represent the size of the probability, the dotted areas represent those infinite sequences in which the next result is 1. On the left we start with all sequences, E_0. After observing $q_1 = 1$, we zoom in on the cylinder set $E_1 = Q_1^1$, containing all infinite sequences for which $e(1) = 1$. Within this cylinder set, the probability for the next result being 1 is different, only because the probabilities over the hypotheses have shifted.

updating, or Bayes' rule for short. In the above framework, this rule can be defined as a recursive relation between the probability functions $p_{[e_i]}$ and $p_{[e_{i+1}]}$,

$$p_{[e_{i+1}]}(H_j) = p_{[e_i]}(H_j|Q^q_{i+1}), \tag{1.18}$$

in which $q = e_{i+1}(i+1)$ is the last observation. The rule expresses that the degree of belief assigned to H_j after observing e_{i+1} must be the same as the degree of belief assigned to H_j after e_i, conditional on the further occurrence of $q_i = e_{i+1}(i+1)$. For this reason it is sometimes called conditioning. The operation is illustrated in figure 1.2.

Let me elaborate on the use of Bayesian updating for the purpose of deriving a probability assignment over statistical hypotheses. Some more general remarks on conditioning can be found in section 1.4. For now, note that the conditional probabilities in the foregoing can be written as follows:

$$p_{[e_i]}(H_j|Q^q_{i+1}) = p_{[e_i]}(H_j)\frac{p_{[e_i]}(Q^q_{i+1}|H_j)}{p_{[e_i]}(Q^q_{i+1})}. \tag{1.19}$$

This expression is known as Bayes' theorem. It follows simply from the axioms of probability. Thus, to compute the probability for some H_j relative to a data set e_{i+1}, we need the preceding probability assignment $p_{[e_i]}(H_j)$, the observation

$q = e_{i+1}(i+1)$ by means of which we can pick out the correct set Q^q_{i+1}, and the probabilities $p_{[e_i]}(Q^q_{i+1}|H_j)$ and $p_{[e_i]}(Q^q_{i+1})$ defined for that set.

This allows us to trace back the probability over the hypotheses $p_{[e_t]}(H_j)$ to their likelihoods for the specific observations e_t and an initial probability $p_{[e_0]}(H_j)$. As indicated, the probabilities of the observations $p_{[e_i]}(Q^q_{i+1}|H_j)$ are assumed to be given with the hypotheses. Furthermore, the prediction $p_{[e_i]}(Q^q_{i+1})$ can again be written in terms of the likelihoods and the probability $p_{[e_i]}(H_j)$. We can then apply the above recursive relation repeatedly and write

$$p_{[e_t]}(H_j) = p_{[e_0]}(H_j) \prod_{i=1}^{t} \frac{p_{[e_{i-1}]}(Q_i^{e_t(i)}|H_j)}{\sum_j p_{[e_{i-1}]}(H_j) p_{[e_{i-1}]}(Q_i^{e_t(i)}|H_j)}. \tag{1.20}$$

Bayesian updating thus determines the probability assignment over the hypotheses $p_{[e_t]}(H_j)$ on the basis of observations e_t, prior probability assignments $p_{[e_0]}(H_j)$ and the likelihoods $p_{[e_i]}(Q^q_{i+1}|H_j)$.

Prior probability assignment. The probability assignments $p_{[e_0]}(H_j)$ are an irreducible input component for making predictions in a Bayesian scheme. In the same way as that we assume the likelihoods of each of the hypotheses, we must therefore assume a prior probability assignment over the hypotheses themselves:

$$p_{[e_0]}(H_j) = p(h_j). \tag{1.21}$$

Here $H_j = \{h_j\} \times \mathcal{Q}_0$ is a subset of $\mathcal{H} \times \mathcal{Q}_0$. Note that we must have $p(h_j) \leq 0$ for each j, and $\sum_j p(h_j) = 1$. In other words, p is a probability function running over $\mathcal{H}$. The initial degrees of belief over the hypotheses, expressed in p, are usually referred to as priors.

It is useful to distinguish between two separate aspects of choosing priors. One aspect of it is that we allocate prior probabilities after we are given a particular partition of statistical hypotheses. As will be argued, by allocating the probabilities over a given partition we can express specific inductive knowledge. Another aspect of choosing priors is in deciding what range of hypotheses to use in the first place. In the example, the likelihoods are the prediction rules $pr_{[h_0]}$ and $pr_{[h_1]}$. Together with the prior probabilities $p_{[e_0]}(H_0)$ and $p_{[e_0]}(H_1)$, these likelihoods determine the inductive predictions that derive from the Bayesian scheme. By choosing the prior probabilities to be nonzero only for hypotheses in the partition $\mathcal{H} = \{h_0, h_1\}$, we determine which likelihoods play a role in the Bayesian scheme. This second aspect of choosing priors, which concerns the choice of possible statistical models, is therefore very important. It is discussed at length in chapter 3.

This chapter, however, need not invite discussion over the issue of choosing hypotheses or priors. Recall that its aim is to present two schemes for making inductive predictions, and to reconstruct these schemes in terms of inductive logical inferences. The issue of choosing priors is important, but as explained in the introduction of this chapter, the schemes as such cannot be expected to offer any guidelines here. The choice of priors concerns premises in the inductive logical arguments. Of course, it must be part of the development of a logical scheme to elucidate how this notion of premises relates to bits of relevant knowledge. Later chapters suggest tools for relating specific knowledge to the choice of priors. But for now I leave it at the somewhat loose remark that priors must be chosen to reflect initial beliefs and accord with relevant knowledge.

To sum up, in the Bayesian scheme the predictions $p_{[e_t]}(Q^q_{t+1})$ take as input the prior probabilities $p_{[e_0]}(H_j)$ for every $0 < j \leq N$, and the likelihoods $p_{[e_i]}(Q^q_{i+1}|H_j)$ for every $0 < j \leq N$ and $0 < i \leq t$. The beliefs attached to the hypotheses, denoted $p_{[e_t]}(H_j)$, function as an intermediate state in determining the predictions. Degrees of belief over hypotheses are updated by means of conditioning. Having obtained an expression for these degrees of belief, we can compute the predictions $p_{[e_t]}(Q^q_{t+1})$ with equation (1.16).

Infinite partitions. An important refinement in the Bayesian scheme is presented by partitions with an infinite number or even a continuum of statistical hypotheses, with which I shall now deal.

Consider the partition $\mathcal{H} = \{H_\theta\}_{\theta \in [0,1]}$ in which the index j is replaced by a variable θ over a continuum of values $[0, 1] \subset R$. The probability assignments over the hypotheses $p_{[e_t]}(H_j)$ then turn into so-called probability distributions $p_{[e_t]}(H_\theta)d\theta$, whose form is determined by the so-called density functions $p_{[e_t]}(H_\theta)$. In the predictions (1.16), the summation over hypotheses must be replaced by an integration:

$$p_{[e_i]}(Q^q_{i+1}) = \int_0^1 p_{[e_i]}(H_\theta) p_{[e_i]}(Q^q_{i+1}|H_\theta)\, d\theta. \tag{1.22}$$

Further, Bayesian updating becomes an operation which transforms the density function $p_{[e_t]}(H_\theta)$, employing the above expression for the predictions:

$$p_{[e_{i+1}]}(H_\theta)d\theta = \frac{p_{[e_i]}(Q^q_{i+1}|H_\theta)}{p_{[e_i]}(Q^q_{i+1})} p_{[e_i]}(H_\theta)d\theta. \tag{1.23}$$

In all other expressions, the index j must be replaced with the variable θ. But apart from that, there are no changes to the update machinery.

1.4 A logical picture

This section sketches a logical picture of the above schemes. It characterises inductive arguments, assuming Bayesian updating as a rule of valid inference, and introduces probability models as part of the input in both the Bayesian and the Carnapian scheme.

1.4.1 Conclusions and inference rules

In deductive logic, to put it bluntly, an argument consists of premises, inference rules, and a conclusion. Inductive logic, if it is to mimic this blunt picture of deductive logic, must also consist of these three elements.

Notion of conclusion. On the notion of conclusion I can be very brief. Since I am considering schemes for making direct predictions, the conclusions of the inductive arguments are direct predictions, that is, probability assignments of the form $p_{[e_t]}(Q^q_{t+1}) = pr$. The remainder of this subsection deals with the inference rules. The premisses of the Carnapian and the Bayesian schemes are dealt with in the next two subsections.

Inference rules. From probability assignments of the function $p_{[e_t]}$ we may derive further assignments of that function according to the axioms of probability theory. I propose to view these axioms as inference rules. The suggestion here is that the axioms generalise the rules for classical truth values over a Boolean algebra, allowing a continuum $p \in [0, 1]$ of truth values where classical truth values have $p \in \{0, 1\}$. This idea is strongly related to ideas in Ramsey (1921) and De Finetti (1937). The axiom that $p_{[e_t]}(U) = 1 - p_{[e_t]}(K^\omega \setminus U)$ generalises negation, the axiom $p_{[e_t]}(U \cup V) = p_{[e_t]}(U) + p_{[e_t]}(V)$ for $U \cap V = \emptyset$ generalises the operation of conjunction, and the axiom that $p_{[e_t]}(K^\omega) = 1 - p_{[e_t]}(\emptyset) = 1$ fixes the relation to Boolean truth valuations.

The above suggestions do not settle that we are dealing with a proper formal logic, and the remainder of this thesis will not settle that matter either. For one thing, the logical scheme is cast entirely in the language of mathematics, and complications arising from that are simply left aside. Allusions to the logical nature of the picture are here intended to convey a specific perspective on inductive inference, while the logic itself is not spelled out in a formally rigorous way. For a more elaborate treatment of these issues I refer to Cox (1961) and Scott and Kraus (1966).

The above probability assignments $p_{[e_t]}$ are always relative to the same empirical knowledge base e_t. Inductive inference must also accommodate changes in the knowledge base. It must allow for a representation of the addition of certain premisses, namely the addition of observations. In the logical scheme these additions can be seen as nonmonotonic inferential steps. The natural candidate for relating assignments with different knowledge bases is Bayes' rule in its general form:

$$p_{[e_t]}(\cdot) = p_{[e_0]}(\cdot|E_t^{e_t}). \tag{1.24}$$

In the logical picture of this chapter, Bayes' rule is the only inference rule that links probability assignments with different knowledge bases. Note that the above rule applies to all elements of the algebra $\mathcal{H} \times \mathcal{Q}_0$, both to hypotheses $H_j = \{h_j\} \times \mathcal{Q}_0$ and to observations such as Q_t^q. But apart from that, the above expression is equivalent to the expression of Bayes' rule for hypotheses, as introduce in the Bayesian scheme. Iterated conditioning of a probability function $p_{[e_t]}$ on new observations Q_{t+1}^q is just a shorthand form of conditioning the probability $p_{[e_0]}$ on a sequence $E_{t+1} = E_t \cap Q_{t+1}^q$ in one go.

There are several arguments for the validity of conditioning, for example in Birnbaum (1962) and Rosenkrantz (1992). To my biased eyes, conditioning looks very natural: after we have observed Q_{i+1}^q, the probability for hypothesis H_j becomes the probability we assigned to hypothesis H_j on the supposition that this observation occurred. But the intuitiveness of conditioning is perhaps illustrated best by considering it in the so-called cylinder algebra of section 1.2, which resembles the muddy Venn diagrams of Van Fraassen (1989). The idea is illustrated in figure 1.3. If we obtain the observation Q_{i+1}^q, we can discard all possible worlds e in which $e(i+1) \neq q$, as we are sure not to inhabit any of those worlds. We zoom in on the cylinder set $E_i \cap Q_{i+1}^q$, which contains all the possible worlds that match the knowledge base e_{i+1}. The probabilistic expression of this zooming operation is that we assign the cylinder $E_i \cap Q_{i+1}^q$ a probability 1. Within this cylinder set, however, there is no reason to change the probability assignment any further. That is, the proportions of the probabilities within the set $E_i \cap Q_{i+1}^q$ must remain invariant under the zooming operation. In more technical terms, the change in probability must respect rigidity. Conditioning is the only operation that respects both these aspects of zooming in.

Criticisms against Bayesian inference. Despite its naturalness, Bayesian updating has been subject to heavy criticism. It must here be noted that those opposing Bayesian updating do not oppose Bayes' theorem, expression (1.19), which can be derived from the axioms of probability. The discussion concerns the rule that links different probability assignments, expression (1.24), and the claim that this rule determines how we must adapt beliefs to observations if

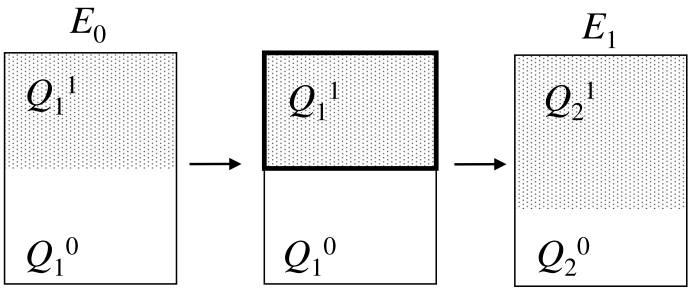


Figure 1.3: Illustration of conditionalisation. The areas represent sizes of the probabilities, the dotted areas represent possible worlds in which the next observation result is 1. After finding $q = 1$ at time $t = 1$, as represented by the observation Q_1^1, we zoom in on those possible worlds e for which $e(1) = 1$. New predictions can be derived from the probability assignment within this patch of the cylinder algebra.

we want to follow a rule at all. First I want to set aside one important line of criticism against this claim. It is sometimes supposed that conditioning cannot accommodate certain modes of inference. As argued in Bacchus (1990) and Van Fraassen (1989: 160-170), abduction is essentially at variance with Bayesian updating. For this criticism I refer to chapters 3 and 9, which both challenge the incompatibility claims.

Another criticism must be given more careful attention here. It contends that Bayesian conditioning is irrational. In particular cases it seems to prescribe unintuitive, irrational or even inconsistent probability assignments, as for example in the discussion on the reflection principle in Van Fraassen (1989), in the drinking and driving example of Maher (1993: 105-29), in epistemic logic as discussed in van Benthem (2003), and in the sleeping beauty problem as discussed by Elga (2000), Lewis (2001), Dorr (2001) and many others. Now it must first be remarked that this thesis is not concerned with rationality. It is not a problem that the proposed schemes for making inductive predictions fall short of providing rational guidelines, as long as rationality is not actively precluded by the schemes. If we derive irrational conclusions by means of Bayesian conditioning, this simply means that we had irrational starting points. Along the same line, it is not immediately problematic that the results of updating may violate intuitions.

The real worry is that conditioning is inconsistent, because an inconsistent inference rule can never yield a viable analysis of inductive inference. However, the inconsistency results in the above example cases can only be derived in an observational framework that allows for the expression of opinions about opinions, or otherwise for curious events such as memory loss or intoxication. The framework for observations does not leave room for opinions over opinions, or for such drug-related circumstances. It can only incorporate observation events and general observational hypotheses. In chapter 2 I shall come back to this, when I discuss opinions about opinions in relation to a semantics for statistical hypotheses. The general contention is that inconsistency of Bayesian conditioning can always be resolved by refining the algebra over which the subsequent probability assignments are defined.

1.4.2 Probability models as premises

With both the conclusion and the inference rule in place, we can now turn to the premises. This subsection deals with the premises in the Carnapian scheme, which prepares for a discussion of Bayesian premises in the next section.

A probability model from direct predictions. Apart from the observations, the input to the Carnapian scheme consists of a direct prediction rule. As will become apparent, such a rule can be summarised in a so-called probability model $\mathcal{M}$, which consists of an algebra $\mathcal{Q}_0$, and a probability assignment over this algebra, $p : \mathcal{Q}_0 \mapsto [0, 1]$. A probabilistic model is a 2-tuple

$$\mathcal{M} = \langle \mathcal{Q}_0, p \rangle. \tag{1.25}$$

In the following I first show that under the assumption of Bayes' rule, any prediction rule $pr(q, e_t)$ can be derived from a single probability model. These probability models can then be used as a formal notion for elaborating the Carnapian scheme in full detail.

It is easily seen that direct predictions can always be derived from a single probability model. Consider the Carnapian scheme, in which the direct predictions are given as follows:

$$p_{[e_t]}(Q^q_{t+1}) = pr(q, e_t). \tag{1.26}$$

Under the assumption of Bayes' rule, we can write

$$p_{[e_0]}(Q^q_{t+1}|E^{e_t}_t) = pr(q, e_t), \tag{1.27}$$

and with the definition of conditional probability, this is equivalent to

$$p_{[e_0]}(Q^q_{t+1} \cap E^{e_t}_t) = pr(q, e_t)p_{[e_0]}(E^{e_t}_t). \tag{1.28}$$

This is a specific restriction to the probability assignment $p_{[e_0]}$, which may be taken to underlie the direct predictions of pr.

I now show that the above restriction indeed determines a unique and complete probability $p_{[e_0]} = p$, and thus a unique probabilistic model $\mathcal{M}$ over the observation algebra $\mathcal{Q}_0$. Note first that the probability $p(E_t)$ for any e_t can be determined by mathematical induction over t. We can use $p_{[e_0]}(E_0) = p(E_0) = 1$ as induction base, and the above restriction, here written

$$p(E_{i+1}) = pr(q, e_i)p(E_i), \tag{1.29}$$

as inductive step. We can therefore determine the probability of any intersection of two observation sets Q^q_t and $Q^{q'}_{t'}$ for $t' < t$, using 1.26 and 1.29:

$$p(Q^q_t \cap Q^{q'}_{t'}) = \sum_{E_{t-1} \subset Q^{q'}_{t'}} p(Q^q_t | E_{t-1})p(E_{t-1}). \tag{1.30}$$

Here $E_{t-1} \subset Q^{q'}_{t'}$ concerns all e_{t-1} for which $e_{t-1}(t') = q'$. With this we have defined the probability for all sets in the so-called π system of observations, which consists of all sets Q^q_i and all countable intersections of them. It is then a theorem of probability theory that there is a unique extension of this probability function to the algebra $\mathcal{Q}_0$ generated by this π system, for which the reader may consult standard textbooks in measure theory, e.g., Billingsley (1995: 36-44).

Under the assumption of Bayes' rule, the Carnapian scheme can therefore be said to take a single and complete probability function p as its input probability $p_{[e_0]}$. That is, adopting a particular prediction rule $pr(q, e_t)$ is the same as equating the initial belief state $p_{[e_0]}$ with a probability function from some model $\mathcal{M}$, and updating the probability assignments according to Bayes' rule. Thus, for the Carnapian scheme a single probabilistic model and the observations are in fact all that is needed.

The Carnapian scheme. This leads up to the following reformulation of predictions that derive from the Carnapian scheme in terms of a Bayesian inductive argument:

○ 1 $p_{[e_0]}(\cdot) = p(\cdot)$, a complete prior over the observation algebra,

○ 2 e_t, some sequence of observations,

⇒ 3 $p_{[e_t]}(Q^q_{t+1}) = p_{[e_0]}(Q^q_{t+1}|E_t)$, the prediction (1, 2 and Bayes' rule).

By choosing the probability p appropriately, we may replicate any prediction rule pr. The bearing that the observations have on predictions is encoded in the probabilistic model, which is entirely under the control of the observer. In chapters 3 and 7, I will come back to this aspect of inductive arguments.

The use of Bayesian updating may cause some confusion here. Notice that Bayes' rule is now intended as an inference rule. But in the foregoing it has been presented as a rule for transforming beliefs on the occurrence of some new element of data. It may be objected that a logical inference is not at all like changing beliefs, but more like making explicit certain elements that are already contained in the beliefs. The logical picture can therefore better work with a single probability assignment to represent beliefs, and a notion of conditioning that does not involve different probability functions. There is no need for Bayes' rule, as opposed to Bayes' theorem, in such a logical picture.

In the present context, it is more appropriate to work only with conditioning on a single probability p: indexing every probability assignment with e_t is overly elaborate. The second part of this thesis is in fact organised in that way. However, it may be taken as an attractive feature of the present schemes that they express the addition of observations as a kind of nonmonotonic move. Moreover, I want the logical picture to leave room for inferential steps which cannot be captured with conditioning alone. I employ Bayes' rule in this way in order to leave open the formal possibility of deviating from conditioning in linking up probability assignments that have different knowledge bases. Chapter 8 will return to this issue.

1.4.3 Multiple probability models

The Bayesian scheme. I now deal with premises in the Bayesian scheme. Again the observations are part of the premises. The scheme further uses a whole range of probability models $p_{[h_j]}$, and a prior probability over the hypotheses H_j that are associated with these models.

The inductive argument is therefore more complicated in the Bayesian scheme. It concerns a derivation of probabilities in the algebra $\mathcal{H} \times \mathcal{Q}_0$:

○ 1 $\forall j :\ p_{[e_0]}(H_j) = p(h_j)$, a prior over hypotheses,

○ 2 $\forall j : p_{[e_0]}(\cdot|H_j) = p_{[h_j]}(\cdot)$, the likelihoods for each of the hypotheses,

○ 3 e_t, some sequence of observations,

⇒ 4 $\forall j : p_{[e_t]}(H_j) = p_{[e_0]}(H_j|E_t)$, a posterior over hypotheses (1, 2, 3, probability axioms, Bayes' rule),

⇒ 5 $\forall j : p_{[e_t]}(Q^q_{t+1}|H_j) = p_{[e_0]}(Q^q_{t+1}|H_j \cap E_t)$, the updated likelihoods (2, 3, probability axioms, Bayes' rule),

---

⇒ 6 $p_{[e_t]}(Q^q_{t+1}) = \sum_j p_{[e_t]}(Q^q_{t+1}|H_j) p_{[e_t]}(H_j)$, the prediction (4, 5, probability axioms).

Note that the likelihoods and the posterior probability over the hypotheses are both derived with Bayesian updating.

Some remarks are called for. First, note that we may use expression (1.18) and the probabilities $p_{[e_t]}(H_j)$ for computing $p_{[e_{t+1}]}(H_j)$. The observation Q^q_{t+1} is then incorporated in a new inductive argument, but this argument has posteriors from an earlier argument among its premises. Second, the Bayesian scheme of section 1.3 leaves aside updates on likelihoods in order to present the Bayesian scheme as a generalisation of the Carnapian scheme. But as part of the logical picture, the likelihoods are updated after all, both in the Carnapian and in the Bayesian scheme. Third, in the case that the hypotheses are associated with constant predictions, as for example h_0 in the hunting example, we have

$$p_{[e_0]}(Q^q_{i+1}|H_j \cap E_{i'}) = p_{[e_0]}(Q^q_{i+1}|H_j). \tag{1.31}$$

But there are also hypotheses for which the predictions do depend on earlier observations, as for example h_1. In that case the update operation affects the likelihoods.

Logical arguments. One aspect of the logical picture concerns the Carnapian and the Bayesian scheme equally: the use of probability models. It must be stressed here that I do not take probability models as objective models, and also not as complete or partial representations of beliefs actually held by some reasoner. A probability model p is a premise in an inductive argument. It is a formal tool for elucidating inductive schemes, and not more than that. Also, there is no restriction that the probability model must somehow match the world. Leaving aside reductios, the usefulness of the conclusion of an argument depends on the truth of the premises, but the argument as such can be perfectly valid independently of that. Probability models are thus similar to truth valuations or models in classical logic: it is not inherent to the use of models in classical logic that they are accepted as true by some reasoner, and they need not match the world in any way either. Premises are adopted, literally, for the sake of an argument.

Another aspect of the logical picture must be mentioned here briefly. A proper logic will advertise itself with a soundness and a completeness result, and the inductive logical schemes here may be expected to provide such results too. However, I have not developed the formal semantics of the schemes sufficiently to provide these results. Chapter 2 will deal with some aspects of semantics, and soundness and completeness results are sketched in chapter 7. But unfortunately this thesis does not contain a proper treatment of the subject.

Finally, it is again notable that apart from the observations, the Bayesian scheme consists of a range of input probabilities which are entirely free for choice. There is no further restriction on what input probabilities may be rational or acceptable. This aspect of the inductive arguments is of key importance to the rest of this thesis. It is that both in the Carnapian and in the Bayesian scheme, the observations do not determine what predictions are warranted. In choosing the input probabilities we effectively determine the patterns in the observations on which the predictions focus, but there is no restriction stemming from the observations alone.

1.5 Normative aims

The foregoing has introduced two schemes for making inductive predictions. Both of them use the observational algebra $\mathcal{Q}_0$ and probability functions $p_{[e_t]}$ to characterise observations and beliefs respectively. However, there is a lot of controversy over the framework of algebra and probability, especially when it comes to representing beliefs by means of probability functions. In this section I want to make clear how this controversy affects the current discussion. The section may be left aside without impeding further reading.

Descriptive adequacy. Many arguments that are critical of a probabilistic framework are concerned with descriptive aims. As such arguments go, a scheme for inductive predictions cannot use a framework with probability, because human reasoning cannot be adequately described in it. Experiments in cognitive psychology, as recorded in Kahneman and Tversky (1982) and also Gigerenzer (2001), show that humans do not reason in accordance with probability theory. However, the aim of the schemes discussed in this chapter is normative and not descriptive. And because the aim is not to represent human inductive reasoning as accurately as possible, it cannot be used as an argument against the framework itself or against the schemes it facilitates that real humans do not comply to it. To draw the analogy with deductive logic once again, there is strong evidence for the fact that people do not reason in compliance with the rules for material implication, which has been discussed in Wason (1968) and more recently Van Lambalgen and Stenning (2001). But most logicians do not see this as a reason to abandon classical logic as a normative theory of reasoning either.

The normative aim of this thesis is to characterise valid inductive inference. Put more carefully, it is to provide a scheme that describes inductive practice as valid inference. Clearly, the form of the inference scheme may deviate from practice, and in this sense the critical arguments mentioned above can indeed be deemed irrelevant. However, the specific normative aim does make certain considerations of descriptive adequacy relevant after all. To see this, imagine that as a normative scheme of inductive inference I presented a cookery book. This book is obviously inappropriate as a normative scheme, because inductive inference is not at all like cooking. Now it is of no help here to state that the normative theory need not be descriptively adequate. Indeed, actual inductive practice need not be described adequately, but we do want the normative scheme to provide norms for exactly those types of inferences that are exhibited by actual inductive practice. In short, the norms must still be applicable to the practice. Secondly, inspired on Earman in (1992: 56-7), it may be that inductive logic is overplaying her hand when it devises a normative scheme that is so far removed from practice that nobody knows even how to strive towards the norms. Thus, both for the applicability of norms and for their attainability, the criticism that the framework is not descriptively adequate cannot be ignored completely.

It is hardly necessary to illustrate problems that relate to the attainability of the goals laid down by the schemes. Bayesian schemes in particular are notorious for their computational intractability, and this problem is only partly solved by the use of computers and computational tools such as Bayesian networks. Nevertheless I feel that problems with the applicability of norms are potentially more harmful to the aims of this thesis. In the remainder of this section I want to illustrate two such problems. Neither can be discarded by pointing to the normative nature of the schemes, and in both cases it must simply be disclaimed that they are solved in the present study. A third problem is only briefly mentioned, and discussed more elaborately in chapter 2.

Applicability problems. First, it can be noted that the probability functions do not capture all types of uncertainty that may be involved in reasoning. This is because a belief cannot always be associated with a sharply delineated extension in the possible world semantics. As an example of this kind of uncertainty, consider the problem of logical omniscience. Imagine that we are given a Boolean algebra, and are then confronted with an expression of five pages, which as a matter of fact is a logical tautology. Since it is a tautology, probability theory prescribes a probability 1 for it. However, if we see the five-page tautology for the first time, it seems natural to feel some uncertainty over its truth. But in such cases, we are not uncertain about the probability measure that is to be allocated to the extension of the expression in the algebra. Rather, we are uncertain on what the extension of the five page expression in the algebra is, or in more common terms, we are uncertain on what the five page expression means. This kind of uncertainty in beliefs is not captured adequately in the probabilistic framework, because probability can be assigned to statements only after the extension of the statement is given. The above schemes therefore do not provide the norms for dealing with this kind of uncertainty.

Second, apart from the fact that the framework leaves a particular kind of uncertainty out of the picture, the nature of the uncertainty that the schemes are actually concerned with may not be captured adequately by the mathematical notion of a probability function. This problem concerns the fact that probabilities have sharp values within the real interval $[0, 1]$. One of the consequences of having sharp values is that the uncertainties attached to statements, or sets of possible worlds, form a complete ordering. But in actual cases, as famously discussed by Keynes (1921) and later by Kyburg (1974), it may not be true that any pair of observations can be compared with respect to the degree of belief that we attach to them, even if we have some opinion on both of them separately. It seems wrong to state this complete ordering as a norm for reasoning with uncertainty, and thus to force this ordering onto our assessment of uncertainty.

A third worry concerns the interpretation of probabilities as representations of beliefs. Clearly, inductive inferences concern degrees of belief, and are thus associated with epistemic, as opposed to physical, probability. If, for instance, on the basis of data I assign a probability of $2/3$ to the event that a tiger appears next, this means that I consider it more likely than not that the tiger appears next, and not in the first place that there is a tendency in the tiger itself to appear with that chance. For all we know, the tiger may be perfectly determined in all its hunting decisions. On the other hand, as will be seen later, this thesis also involves explicit reference to chance processes, in which the probabilities are objective and connected to physical probabilities. For example, I may assign an epistemic probability of 3/4 to the statement that the objective chance for any tiger to appear directly after a duck is smaller than $\frac{1}{2}$. Chapter 2 argues that both objective and epistemic probability can be given an unproblematic interpretation in such a setting. But in the present chapter I cannot resolve the tensions that may result from their simultaneous usage.

Disclaimers. Generally, I concede that there are mismatches between the present framework and actual reasoning, and that the schemes therefore cannot present a complete set of norms for inductive reasoning. However, it will be assumed that these mismatches are not destructive to all the aims of the schemes. I take the above considerations to show rather that the norms presented by the schemes are not detailed enough, and that they are therefore not applicable without further idealising assumptions. Such assumptions are similar to those we may make when it comes to the applicability of laws of nature: even though the conditions under which the laws function are almost never met, we still take the laws to be applicable, and we say that the description in terms of laws is incomplete rather than inapplicable.

1.6 Conclusion

Summary. The above presents a particular picture of the Carnapian and the Bayesian schemes for making predictions. According to this picture, both schemes accommodate new observations with Bayes' rule, and both schemes take probabilistic models and observations as input. For both schemes the logical picture isolates a notion of conclusion, namely the predictions, a notion of inference, namely that of the probability axioms and Bayes' rule, and finally, a notion of a set of premises, which consist of observations and probabilistic models. These latter premises bring out the assumptions needed for making inductive predictions. Thus the picture emphasises the logical nature of the schemes, but it also highlights that the observations do not entail anything by themselves.

The role of Bayesian inference. For anyone familiar with the pervasiveness of Humean criticism, the fact that the picture leaves the inductive predictions completely underdetermined will not be surprising. The import of the problem of induction is not just that the data alone do not tell everything, it is more that the data alone do not tell anything. One of the reasons for presenting the predictions as conclusions of the above arguments is exactly because these arguments reveal the assumptions underlying the predictions, and bring to the fore that these assumptions do all the inductive work. As already sketched in the introduction, this may be viewed as an advantage offered by the logical picture.

It is perhaps felt by some that not all the inductive work is done by the assumptions in the logical picture. This is related to the criticism that Bayes' rule, as used in the Carnapian scheme, cannot capture all modes of inference, and that for example abduction cannot be captured by it. In the above picture, this amounts to the claim that Bayes' rule allows to derive more than the probabilistic conclusions already implicit in the input probabilities. However, the above discussion shows that any prediction rule $pr(q, e_t)$ corresponds to adopting some prior probability assignment $p_{[e_0]} = p$, and the use of Bayes' rule for the inferences. This means that, at least on the level of inductive predictions, the rule is not restrictive at all: any prediction rule can result from it. By the same light, the rule is not restrictive in the Bayesian scheme either.

Dogmatic aspects to Bayesian inference. Once we have chosen a probability model for the Carnapian scheme, or a prior over hypotheses and a set of such models for the Bayesian scheme, Bayes' rule fully determines the predictions. This is similar to the case in deductive logic, where the inference rules do not restrict the possible conclusions, while choosing particular premises compels us to particular conclusions. Of the initial objection that Bayes' rule introduces unintended or even unacceptable restrictions, the core may very well be that the rule forces us to choose a prediction rule at the onset, and to stick to it after that. Certainly van Fraassen (1989) is concerned with this in his discussion of Bayes' rule. I must admit that it presents an unusually dogmatic aspect of the flexible theory of Bayesianism that everything must be fixed at the start.

In reaction to this criticism, let me first remark that rule following is inherent to all logical schemes. It therefore makes little sense to defend the above logical schemes against criticisms that concern this rule following aspect. However, just as in classical logic, there is no problem in deciding to start a new inductive inference if the need arises, that is, to simply drop premises that have led into useless conclusions and use the observations again in another inference. Chapter 8 proposes a formal model for a similar move, as an add-on to the Bayesian scheme, but needless to say, starting a new inference may be perfectly rational also if a formal model for such epistemic moves is lacking.

Finally, let me address a worry that is strongly related to the dogmatism that seems to be inherent in Bayesian logic. It is that in using this rule, the conceptual work may be distributed inefficiently over premises and inference: some other rule may allow for more readily applicable or easily accessible input probabilities, or for a relation between premises and conclusions that more naturally reflects inductive inference. In short, the innocence of Bayes' rule may come at a price. It is only in chapter 3 that I can present an argument to the contrary.

Carnap versus Bayes. One final remark must be made about the relation between the Carnapian and the Bayesian scheme. As indicated, the Bayesian scheme can be seen as a generalisation of the Carnapian scheme. The Carnapian scheme takes only one probabilistic model as its input, whereas the Bayesian scheme incorporates a range of models. To deal with this range, it allocates a probability assignment over them, and updates this assignment over the models, or hypotheses, just as it updates the likelihoods, or direct predictions, within the models. But the Bayesian scheme eventually leads to predictions that can also be captured in the simpler Carnapian scheme. Seen from this angle, the Bayesian scheme may be nothing more than a useless complication. Again, chapter 3 argues for the use of the Bayesian scheme in connecting relevant knowledge with prior probabilities.

A Frequentist Semantics of Hypotheses

This chapter proposes a frequentist interpretation of statistical hypotheses. In this interpretation, statistical hypotheses are associated not with probability models over a whole algebra $\mathcal{Q}_0$, but rather with strict subsets of the observation so-called σ-algebra $\mathcal{Q}$, the extension of $\mathcal{Q}_0$. The Bayesian scheme can then be taken as a further specification of the Carnapian scheme, in which hypotheses appear as convenient extensions to the observation language.

The chapter first discusses statistical hypotheses in the logical picture, and indicates how a frequentist interpretation can elucidate their use. Then it defines a specific set of statistical hypotheses, for which such an interpretation can indeed be given. Under this interpretation the Bayesian scheme is seen to be formally, and not just extensionally, equivalent to the Carnapian scheme: both schemes take a completely specified probability p over a single observational algebra as input. Some considerations on hypotheses and models complete the chapter.

This chapter presupposes chapter 1 as a whole. It is itself useful reading for chapters 3 and 8.

2.1 Statistical hypotheses

This section shows that statistical hypotheses are identical in terms of the observational algebra, and different only in the probability models associated with them. It is thus natural to view the probability over hypotheses as a second order probability, but such a probability seems at odds with the logical picture sketched in the preceding chapter. On the other hand, if this introduction of second order probability is avoided, more positive reasons for adopting the frequentist view remain.

Statistical hypotheses: algebraic or probabilistic?. The Bayesian scheme of the preceding chapter employs the algebra $\mathcal{H} \times \mathcal{Q}_0$. Hypotheses are identified with sets $H_j = \{h_j\} \times \mathcal{Q}_0$, which each consist of the same observational algebra $\mathcal{Q}_0$, while their elements are labelled h_j differently. In terms of algebraic structure, there is nothing in the hypotheses H_j to tell them apart. That is, they all have the same observational content. This reflects the fact that purely statistical hypotheses are consistent with any finite sequence of observations and therefore cannot be verified or falsified. But it may then seem rather strange that we are at the same time using observations to decide between statistical hypotheses. In technical terms, if within a sequence of observations E_t the hypotheses coincide, these observations cannot be used to distinguish between the hypotheses. The use of conditioning to decide between hypotheses seems to lack intuitive basis if the hypotheses cannot somehow be told apart in the observational algebra.

Clearly the hypotheses $H_j = \{h_j\} \times \mathcal{Q}_0$ are distinct in another aspect: the probability models defined over them are different. This difference in probability models is what connects the labels h_j to the observations, and thus provides the hypotheses with distinct observational content. In other words, hypotheses indeed overlap in the algebra, but within specific sequences of observations E_t the probabilities assigned to the hypotheses differ. However, this seems to land us in another puzzle. Recall that the Bayesian scheme offers a probability assignment $p_{[e_0]}$ over the hypotheses. If the hypotheses are only distinct because of the probability models, it seems that we are in fact assigning probabilities to these probability models. This seems to turn the probability $p_{[e_0]}(H_j)$ into a kind of second order probability assignment, ranging over models $\mathcal{M}_j$ and not just over the hypotheses H_j. And we may then wonder how this squares with the Kolmogorov definition of probability, in which probability is only assigned to elements of an algebra.

All this is portrayed as problematic a bit too eagerly. After all, in the Bayesian scheme the probabilities $p_{[e_t]}(H_j)$ are assigned to sets $H_j = \{h_j\} \times \mathcal{Q}_0$, which are elements of the algebra $\mathcal{H} \times \mathcal{Q}_0$, just as the observations Q^q_{t+1} and E_t. The sets H_j may be identical in terms of the observational algebra, but nevertheless they are different sets if only for the mere fact of their different labelling. The fact that, apart from the labelling, these sets differ solely because the probability over the observations within them is different must not distract us too much. Furthermore, even if it is conceded that the probability assignments to hypotheses are essentially of second order, there is nothing inconsistent or flatly wrong in introducing such probability assignments. The use of second order probabilities has many proponents, as for example Sahlin (1983). Moreover, second order probabilities are at the heart of so-called expert systems, as discussed by Gaifman (1986), van Fraassen (1989) and others.

Motivating a frequentist semantics. The foregoing leads up to a number of reasons for developing an alternative interpretation of statistical hypotheses.

Consider the case in which the probability over hypotheses is taken as second order. Recall that in the logical picture, the premises of inductive arguments consist of observations together with a probability assignment $p_{[e_0]}$. In this picture, the assignment may be read as a generalised truth valuation, which comprises a continuum of truth values. But the probability assignment $p(h_j)$ is taken as some kind of second order probability over the models, $p_{[e_0]}(p_{[h_j]})$, and not as a probability on the level of sets, $p_{[e_0]}(H_j)$. This move of taking probability assignments as arguments of the probability assignment runs parallel to taking propositions on truth valuations in classical logic as propositions themselves. And it is well known that this opens the door for problems such as the liar paradox. As an example, the inconsistency of Bayesian updating as revealed in Maher (1993:105-29) crucially depends on the use of probability assignments within statements that are themselves assigned probability. In the logical picture sketched above, it seems much safer, as well as more in line with classical deductive logic, to determine premisses in terms of a single probability assignment.

Hypotheses were in the preceding chapter introduced in this way: they were presented as sets in the algebra, $H_j = \{h_j\} \times \mathcal{Q}_0$, and not as probabilistic models. However, also in this presentation, a number of reasons for an alternative interpretation may be advanced. Firstly, note that the logical picture of chapter 1 stays close to the empiricist roots of inductive logic. As suggested in that chapter, it is too much to strive for the derivation of a completely analytic probability assignment from the structure of the language. But I do feel that, where possible, we must attempt to associate probability assignments to observations, or more specifically, to elements in an observational algebra. It seems to me that the Bayesian use of hypotheses as separate observational algebras, $H_j = \{h_j\} \times \mathcal{Q}_0$, removes us unnecessarily far away from the empiricist roots of inductive logic.

Secondly, there is a rather natural way in which the statistical hypotheses can be given an interpretation as statements in an observational algebra after all. This interpretation is based on frequentism. The idea is to connect statistical hypotheses h_j to sets of infinite sequences e that have the probabilities $p_{[h_j]}$ as their limiting relative frequencies. As will become apparent, not all statistical hypotheses lend themselves for such an interpretation. But for those hypotheses that do allow for a frequentist interpretation, the theoretical interpretation of hypotheses appears as conceptual decadence: it employs a multitude of algebras $\mathcal{Q}_0$, where in fact we can do with just one extended algebra, or σ-algebra,

$\mathcal{Q} = \sigma(\mathcal{Q}_0)$. In my preferred terminology: the frequentist view of von Mises can be used as a razor to cut the beard that Kolmogorov is sporting.

The general idea of this chapter is that statistical hypotheses, or probability models, need to be given some kind of empirical content. Without such a content, they cannot be given a natural place in an empiricist inductive logic. The next two sections provide this observational content for a specific class of hypotheses. The concluding section of this chapter will return to the advantages of this alternative interpretation of statistical hypotheses.

2.2 A restricted class of hypotheses

This section gives a formal definition of a class of probabilistic hypotheses, associated with a specific collection of models. Only statistical hypotheses from this restricted class can be connected to elements of the observational algebra. The definition of the class is based on a frequentist interpretation of probability.

Relation with Von Mises. Von Mises (1928) introduced the so-called frequentist interpretation of probability as part of a systematic study into statistical phenomena. The interpretation is not an attempt to derive probability models from finite or even infinite sequences of observations. Rather it is an attempt to specify what probability means by defining this notion in terms of specific infinite sequences of observations e, the so-called Kollektivs. In the following I employ the frequentist interpretation in the exact opposite direction: I start with defining a certain class of statistical hypotheses, associated with certain probability models, and after that I define the hypotheses as sets of specific sequences e by employing the frequentist interpretation of probability. I thereby leave aside many of the subtleties involved in the frequentist interpretation itself. It must be stressed that I do not attempt to justify the frequentist interpretation of probability, or to somehow prove its adequacy. Rather I am using the frequentist interpretation to provide an alternative semantics of the hypotheses h_j. This alternative semantics does not associate hypotheses with the complete observational algebra $H_j = \{h_j\} \times \mathcal{Q}_0$, but rather with strict subsets $H_j \subsetneq K^\omega$, and thus with elements in the extended algebra $\mathcal{Q}$.

Defining statistical hypotheses. Before making this restricted class of statistical hypotheses precise, let me sketch the idea behind it. Every statistical hypothesis in it is associated with a probability model, and thus prescribes, for a range of possible circumstances or states, a probability for the observations, denoted Q^q_{t+1}. The states must at every position t in the string be determined by the observations within e that are already given, $e_t = \langle e(1), e(2), \ldots, e(t)\rangle$, and they must further occur infinitely often in the infinitely long sequence of observations e. For every such e, the probability of Q^q_{t+1} in some state is associated with a relative frequency of q's occurring in this state. The subset of a hypothesis can then be identified with the set of infinitely long sequences e for which all the relative frequencies associated with the states match the probability model.

The definition of this class of statistical hypotheses has two ingredients: a set of identity functions marking the states, and a set of probability vectors, each of them associated to one selection function annex state. The identity functions serve to characterise the states in which the corresponding probability vector applies.

> DEFINITION Let $w(e_t)$ be a function assigning a natural number $\{0, 1, \ldots, M\}$ to all sequences e_t. Further let $\theta = \{\theta_1, \theta_2, \ldots, \theta_M\}$ be a set of fixed probability vectors θ_m of which the components $\theta_{qm} \in [0, 1]$ satisfy $\sum_{q \in K} \theta_{qm} = 1$ for each $0 < m \leq M$. Then the statistical hypothesis $h_{w\theta}$ determines a, possibly partial, probability model
>
> $$p_{[h_{w\theta}]}(Q^q_{t+1}|E^{e_t}_t) = \theta_{qw(e_t)}. \tag{2.1}$$
>
> If $w(e_t) = 0$ the conditional probability remains undefined.

Hypotheses that can be written down in this way are called statistical. If all $\theta_{qm} > 0$, the hypothesis is called purely statistical. Note that these hypotheses do not yet specify the class of hypotheses that may be interpreted in a frequentist manner.

Some remarks may help to clarify the foregoing. First, the hypotheses $h_{w\theta}$ distinguish different states $w(e_t) = m$, depending on the sequence of observations e_t. Further, they associate with each of these states a probability vector θ_m ranging over possible next observations $q \in K$. For any sequence of observations e_t, not more than one such probability vector is chosen by the hypotheses. Note also that the probability model can still be partial, because the function w need not assign a number $m > 0$ to all sequences e_t. Below I define two further restrictions on statistical hypotheses which clarify the point of this complication. Finally, recall that if every e_t is assigned an $m > 0$, the probability model is defined completely. From this we can derive a complete prediction rule.

Restrictions for frequentist hypotheses. I now impose two further restrictions, which, together with the above, define the intended class of hypotheses $\mathcal{F}$. Recall from section 1.4 that with the direct probabilities we can recursively derive values for all $p(E_t^{e_t})$. A sequence e_t is deemed possible by the probability model of $h_{w\theta}$ if and only if $p_{[h_{w\theta}]}(E_t^{e_t}) > 0$. The first restriction is that there may at most be finitely many sequences e_t which are deemed possible by $h_{w\theta}$, but which nevertheless have $w(e_t) = 0$. If we collect the corresponding E_t in a special set, denoted $E_{w=0}$, we can formulate this requirement in the following way:

$$\forall t > 0, \forall E_t \notin E_{w=0} : \qquad p_{[h_{w\theta}]}(E_t^{e_t}) > 0 \;\Rightarrow\; w(E_t^{e_t}) > 0. \tag{2.2}$$

This means that $h_{w\theta}$ must assign a probability to the observation Q_{t+1}^q in all those cases in which the observation set E_t that preceded the observation Q_{t+1}^q has nonzero probability and does not belong to $E_{w=0}$. Note that this presupposes that the probabilities for all observations Q_i^q for which $E_t \subset Q_i^q$ were also nonzero. This complicates the requirement, but it does not present any real difficulty.

The second restriction is that in any possible infinite string e, all states m are repeated infinitely often:

$$\forall m, \forall E_{t'} : \exists E_t^{e_t} \subset E_{t'} : \qquad p_{[h_{w\theta}]}(E_t) > 0 \wedge w(e_{t'}) = m. \tag{2.3}$$

This is to make sure that it makes sense, eventually, to talk of relative frequencies of observations in all the states. The class of frequentist statistical hypotheses comprises all statistical hypotheses, as defined with the above definition, that satisfy restrictions (2.2) and (2.3). This class of hypotheses covers a particular subset of possible statistical hypotheses. The identification of statistical hypotheses with elements $H_j \in \mathcal{Q}$ only works for this limited class.

An example hypothesis. To get to know the class of frequentist hypotheses, let me consider the example of chapter 1 again. In that example we have $K = \{0, 1, 2\}$, referring to observations of the empty pond, ducks and a tiger. Now recall the hypothesis h on tigers hunting ducks. It may be defined in the terms of the function w, given in

$$w(e_t) = e_t(t) + 1$$

and in terms of probability vectors θ, here consisting of nine components:

$$\begin{aligned} \theta_1 &= \langle 2/3, 1/3, 0\rangle, \\ \theta_2 &= \langle 1/6, 1/3, 1/2\rangle, \\ \theta_3 &= \langle 1, 0, 0\rangle. \end{aligned}$$

In words, this hypothesis states that there are three possible cases. If a tiger has not appeared in the last observation, $e_t(t) \neq 2$ and if there are no ducks in the pond, $e_t(t) \neq 1$, they may appear with a chance of $\frac{1}{3}$ while the pond may stay empty with a chance of $\frac{2}{3}$. If a tiger has not appeared in the last observation and if there are ducks, they may stay with a chance of $\frac{1}{3}$, leave with a chance of $\frac{1}{6}$, and a tiger may appear with a chance of $\frac{1}{2}$. If, finally, a tiger has appeared in the last observation, the pond stays empty for one time unit with certainty.

It must be noted that not all sequences of observations e_t are assigned a positive probability. Specifically, sequences such as $e_3 = \langle 0, 0, 2\rangle$ or $e_5 = \langle 0, 1, 2, 1, 0\rangle$ are assigned zero probability in the probability model of h_1. Thus frequentist hypotheses are not necessarily purely statistical. Note also that the probability model associated with h_1 is complete, because we have $m(e_t) > 0$ for all e_t. But this need not always be the case. For hypotheses in the frequentist class there can always be some finite number of sequences e_t for which the hypothesis does not prescribe probabilities. The above hypothesis h does not illustrate this possibility, but I return to it in chapter 3.

The reach of frequentist hypotheses. The class of frequentist hypotheses comprises many more hypotheses like h. There are hardly any restrictions on what kind of statements may be used as hypotheses. For example, they also include the formal equivalent of the statements that tigers operate alone, that ducks wander in packs, that when we see a duck a tiger is not far away, and any other such statement. The only restriction for the hypotheses is that their formal equivalent must be of the form presented above. As will be argued below, this comes down to the requirement that the probability models have an observational content, so that they can be identified with an element in the observation algebra.

On the other hand, many hypotheses cannot be deemed frequentist. As an example, consider the hypothesis that as a result of hungry tigers the number of ducks decreases:

$$\theta_{10} = 1 - \frac{1}{3}e^{-\rho t}, \tag{2.4}$$

$$\theta_{11} = \frac{1}{3}e^{-\rho t} \tag{2.5}$$

In a hypothesis with probabilities that change in such a way, there are no repeatable states with fixed probabilities. Another example is presented by the Carnapian $\lambda\gamma$ rule, conceived as a probability model. The probability for some Q^q_{t+1} given earlier observations E_t are according to this rule determined by two statistics, to wit, the index t and the fraction of the number of earlier occurrences of q in e_t, denoted t_q. With every fraction $\frac{t_q}{t}$ and value of t we can effectively associate another state m, but depending on e there may be infinitely many of such states, and moreover, some of these states are not repeated infinitely often. So the Carnapian $\lambda\gamma$ rule is not frequentist either.

While some hypotheses are not frequentist because their probabilistic models cannot be associated with limiting relative frequencies, other hypotheses are excluded because they are more specific than what limiting relative frequencies allow us to express. An example of the latter kind is presented by the so-called constituents in the $\alpha\lambda$-system of Hintikka (1966). To illustrate, consider the constituent $H_{\neg 2} = \{e : \forall i(e(i) \neq 2)\}$, which for obvious reasons may be called duck heaven. At first sight it may seem that this constituent is covered by the union of all statistical hypotheses that assign a zero probability to the observation of a tiger, $q = 2$. But the Hintikka-constituent $H_{\neg 2}$ is more specific, because it does not only mean that the limiting relative frequency for tigers in the e included in $H_{\neg 2}$ must be zero, but also that in these sequences e there are no tigers at all. The sequence $e = 012000\ldots$ has zero limiting relative frequency for 2, but it is not part of the Hintikka constituent. Because of this notorious measure-zero gap, the Hintikka-constituents are not frequentist, even while such constituents seem to be among the most basic patterns at hand.

In sum, it appears that the class of observational patterns is wider than the class covered by the notion of frequentist statistical hypothesis. Against this, one can also argue that, for example, Hintikka constituents are not observational patterns at all, or in any case much less observational than their frequentist variants.

2.3 Hypotheses as elements of $\mathcal{Q}$

This section presents an interpretation of hypotheses as elements of the extended observation algebra $\mathcal{Q}$. After that it briefly discusses the relation between these elements and the Kollektivs of Von Mises, and it elaborates on the notion of a partition.

2.3.1 Definition of the elements $H_{w\theta}$

Hypotheses as sets of infinite sequences. We are now in the position to define the set $H_{w\theta}$ that is associated with a frequentist hypothesis $h_{w\theta}$. Central to this definition is the identification of probabilities and relative frequencies. A relative frequency of some result is defined by the following:

$$W_{qi}(e) = \begin{cases} 1 & \text{if } e(i) = q, \\ 0 & \text{otherwise,} \end{cases} \tag{2.6}$$

$$f_q(e) = \lim_{t \to \infty} \frac{1}{t} \sum_{i=1}^{t} W_{qi}(e). \tag{2.7}$$

The frequentist interpretation is thus used to translate probabilities on observations, as prescribed by a hypothesis, into properties of infinite strings of observations. It is then possible to connect a probability assignment to a set of all those e for which the above relative frequencies exist, and for which they have the matching values.

To pin down the relative frequencies of observations occurring in some state determined by $w(e_t) = m$, we may define for every e the subsequence e^m of all those observations q_{t+1} following the positions t at which $w(e_t) = m$. First define $I_m(e_i) = 1$ if $w(e_i) = m$, and $I_m(e_i) = 0$ otherwise. Then define

$$s_m(e,t) = I_m(e_t) \sum_{i=1}^{t} I_m(e_i) \tag{2.8}$$

with $e_i = \langle e(1), e(2), \ldots, e(i) \rangle$ the first i entries of e. Then $s_m(e,t)$ is a sequence that has an increasing number on positions t where $w(e_t) = m$, and 0 on all other positions t. Then define

$$e^m(s_m(e,t)) = e(t+1). \tag{2.9}$$

for all $s_m(e,t) > 0$, while $e^m(0)$ remains undefined. The resulting sequence e^m contains exactly those observations made in the state m.

The set which corresponds to a hypothesis $h_{w\theta} \in \mathcal{F}$ can now be obtained by selecting all those $e \in K^\omega$ for which all the subrows e^m have exactly θ_{qm} as the relative frequencies of the observation results q. Formally,

$$H_{w\theta} = \{e : \forall m, q \, [f_q(e^m) = \theta_{qm}]\} \tag{2.10}$$

This is a subset $H_{w\theta} \subsetneq K^\omega$. Note also that only the statistical hypotheses $h_{w\theta}$ that comply to restriction (2.3) can be identified with such a strict subset. This is because the definition of the relative frequencies f_q only works for subsequences e^m that have infinite length, and because these subsequences have infinite length only if (2.3) is fulfilled.

To illustrate the foregoing, consider the hypothesis h_0 of the hunting example, which is associated with the probability model

$$p_{[h_0]}(Q^q_{t+1}|E_t) = \begin{cases} \frac{9}{10} & \text{if } q = 0, \\ \frac{9}{100} & \text{if } q = 1, \\ \frac{1}{100} & \text{if } q = 2. \end{cases} \tag{2.11}$$

For this probability model there is no need for distinguishing different states $w(e_t)$. The corresponding hypothesis may therefore be defined as an element H_0 quite easily:

$$H_0 = \{e : \ f_0(e) = \frac{9}{10} \wedge f_1(e) = \frac{9}{100}\}. \tag{2.12}$$

To this element we can now assign a probability $p_{[e_0]}(H_0)$. Furthermore, within the set of sequences H_0 the probability of the observations is fixed by the model.

Extending the observation algebra. Statistical hypotheses are in the above presented as subsets of K^ω, but they are not made part of any observational algebra yet. I now discuss the extension of the observation algebra $\mathcal{Q}_0$ that is needed for accommodating hypotheses as elements of it.

By defining frequentist hypotheses as strict subsets of the space K^ω, I am providing them with something of an observational content. But this content is not observational in the ordinary manner. Note that any finitely decidable observational hypothesis h_w can be associated with the element H_w of the finite observation algebra $\mathcal{Q}_0$. As an example, the hypothesis that more than half of the first n observations have the result q can be decided within n observations. Hypotheses $H_{w\theta}$ are not finitely decidable in this way. That is, the probability models prescribed by the frequentist hypotheses cannot be verified or falsified by any finite sequence of observations. Therefore frequentist hypotheses are not part of the finite observation algebra $\mathcal{Q}_0$.

With the above definition in place, however, we can associate the hypotheses $h_{w\theta}$ with an element of the σ-algebra $\mathcal{Q}$, the infinite extension of the observation algebra $\mathcal{Q}_0$. A hypothesis $H_{w\theta}$ is a so-called tail event in this algebra. It is an event in the observation algebra whose occurrence can only be verified or falsified at infinity. This corresponds to the fact that the hypothesis $h_{w\theta}$ is not finitely decidable, but that it is, in the vocabulary of Kelly (1996), refutable in the limit. It may be noted that hypotheses that are higher up in Kelly's hierarchy of decidability can still be associated with elements of an algebra $\mathcal{Q}$. Hypotheses may also be gradually refutable or verifiable, and they may have an even more complicated structure. Moreover, Bayesian updating can perfectly well accommodate hypotheses that are undecidable to various degrees. However, the frequentist semantics proposed here restricts hypotheses to ones that are either gradually refutable or gradually verifiable.

With frequentist hypotheses as elements in the observational algebra, the hypotheses need not be treated as probability models over the observational algebra anymore. This means that we can assign probability to hypotheses just as we can assign probability to observations, which accords well with the empiricist roots of inductive logic. Moreover, the frequentist view on hypotheses leads to a picture in which the inductive inferences are all made from a single probability assignment over a single algebra $\mathcal{Q}$, and in which there is no need for second-order probability. This is discussed further in section 2.5. Note also that the proposal to view frequentist hypotheses in terms of a partition of the σ-algebra shows similarities to the formal characterisation of Hintikka-constituents in terms of a partition of the algebra by Kuipers (1978). As made clear in the preceding section, there are differences between statistical hypotheses and these constituents, but the general idea may very well be the same.

Finally, and most importantly, note again that frequentist hypotheses are not part of the finite algebra $\mathcal{Q}_0$ that is used in the Carnapian scheme. Statistical hypotheses can be used as the result of an enrichment of the algebra, or alternatively, observation language. This enrichment accounts for a larger expressive force that the frequentist hypotheses allow us. As will become clear in chapter 3, this enlarged expressive force is one of the driving forces behind the conceptual innovations that this thesis offers.

2.3.2 Collections of Kollektivs

I now deal with the relation between the frequentist interpretation of probability and the above definition of hypotheses as elements in $\mathcal{Q}$. More specifically, I investigate the connection between the set of infinite sequences $H_{w\theta}$ and collections of so-called Kollektivs.

Infinite sequences as Kollektivs. Following the discussions in Von Plato (1994) and Van Lambalgen (1987), a Kollektiv is a specific infinite sequence of observation results e. Two properties define a Kollektiv: the limiting relative frequencies of the observations in e must exist, and it must be impossible to select, with some fixed procedure, positions within the sequence e such that the selected subsequence has different limiting relative frequencies. This latter property has become known as the so-called law of excluded gambling systems, which nicely expresses the idea behind the property: when selectively gambling on results in a Kollektiv, we cannot find a gambling procedure that changes the probabilities for any of the results. Or in again other words, apart from the patterns fixed by the relative frequencies, there is no further weak pattern in the observations.

We can use this notion of Kollektiv to elaborate the above definition of frequentist hypotheses. Recall that the hypotheses themselves already present a specific selection procedure, namely $w(e_t)$. But within the subsequences e^m created with this selection, the notion of Kollektiv becomes applicable. As a start, we can characterise the hypothesis $H_{w\theta}$ as sets containing all those sequences e of which the subsequences e^m are Kollektivs associated with the probabilities θ_{qm}. However, this is not a suitable characterisation. The definition of the hypotheses does not preclude the existence of further selections within the subsequences e^m, within which the relative frequencies are different from the probabilities θ_{qm}.

To see this, consider hypothesis h_0 above, which prescribes a single vector of probabilities for the whole sequence e. Because this is the only criterion for membership of the corresponding element H_0, the hypothesis H_0 also includes sequences e' in which every even indexed observation is of an empty pond with certainty, so that the relative frequencies for even observations are simply $\langle 1, 0, 0 \rangle$, while the relative frequencies of the results on the odd positions are $\langle \frac{8}{10}, \frac{18}{100}, \frac{2}{100} \rangle$. The resulting relative frequencies in such sequences e' is then in accordance with the probabilities prescribed in the model for h_0, while the sequences e' are not Kollektivs for these probabilities.

It is perhaps appealing to sharpen the definition of hypotheses as subsets in K^ω, and to include only the sequences whose subsequences e^m are Kollektivs for the corresponding probabilities θ_{qm}. This involves formalising the law of excluded gambling systems. The notion of admitted selection procedure can be given a proper mathematical formulation by means of recursive functions, which is here employed implicitly as domain for the function w. However, sharpening the definition of hypotheses to include only the Kollektivs involves a more detailed treatment of these recursive functions, which leads us too far away from the main line of this chapter. Moreover, there are independent reasons for preferring the looser definition given in the foregoing. The reason lies in the possibility of finding further structure in the observation results. They are made explicit in chapter 8.

In the following hypotheses $H_{w\theta}$ are sets of sequences e whose subsequences e^m, defined with a function w, have limiting relative frequencies matching the probability model, θ_{qm}. Some of these subsequences are Kollektivs for these probabilities, but others are Kollektivs of a more complicated probability structure. So frequentist hypotheses are composed of collections of Kollektivs.

Other intentions than Von Mises. As a last remark on hypotheses and frequentism, let me stress again that the use of frequentist notions here is opposite to von Mises original use of it. For von Mises the emergence of Kollektivs from sequences of actual observations was an empirical matter, which has to do with statistical phenomena. In this chapter Kollektivs and limiting relative frequencies are a purely formal tool. But more importantly, the aim of von Mises was to use these Kollektivs for understanding probabilities, that is, to interpret the notion of probability with these Kollektivs. In this chapter, by contrast, I give priority to the probability models. The use of the frequentism is only to provide an interpretation of these probability models in an observation algebra. Consequently, the interpretation is only given after the models have been specified. In view of this it is natural that, next to frequentist probabilities, we can also use the subjectively interpreted probabilities over the hypotheses.

2.3.3 Partitions

The remainder of this section discusses the notion of a partition. It may be recalled from chapter 1 that a Bayesian scheme employs collections of hypotheses, which were there called partitions. The foregoing only shows how, within a restricted class, we can construct strict subsets representing these hypotheses. But now we can also make clear in what way the collections of hypotheses form a partition. That is, the hypotheses themselves can be said to partition the observational algebra, meaning that they can form a collection of mutually exclusive and jointly exhaustive sets in K^{ω}.

A patchwork of hypotheses. To see the general idea, consider a collection of hypotheses $H_{w\theta}$ based on some fixed selection function w, but with different probability vectors. For any such collection, we may define a complete partition of the observation algebra, by adding hypotheses with the same selection function w, but with probability vectors θ that are not covered in the collection yet. In other words, the frequentist interpretation allows to cover the whole algebra with a patchwork of hypotheses.

Let me elaborate this with an example. Consider again the hypothesis h_0, which has the selection function $w(e_t) = 1$ for any e_t and a single probability vector $\theta = \langle \frac{9}{10}, \frac{9}{100}, \frac{1}{100} \rangle$. Consider the alternative hypothesis H'_0, which uses the same selection function but prescribes probabilities $\theta' = \langle \frac{1}{3}, \frac{1}{3}, \frac{1}{3} \rangle$. The two sets

H_0 and H'_0 are then mutually exclusive: sequences $e \in H_0$ have limiting relative frequencies that differ from the probabilities of θ', and vice versa. However, the probability vectors θ and θ' are just two elements from the set of all possible vectors, $C = \{\theta : \sum_q \theta_q = 1\}$. There are many more hypotheses with a uniform selection function, and the hypotheses H_0 and H'_0 are therefore not jointly exhaustive. A partition must minimally include all hypotheses $H_{w\theta}$ with the selection function $w(e_t) = 1$ that have a probability vector from C.

But we are not ready with defining the partition. Even when taken together, the hypotheses H_θ with $\theta \in C$ are not jointly exhaustive. Some sequences e do not have limiting relative frequencies for the observations at all. As an example, take the sequence $e = 01\,0011\,000000111111\ldots$, in which the result 2 does not occur, the number of consecutive 0's always equals the total number of observations preceding those 0's, and the number of consecutive 1's equals the number of consecutive 0's that precede it. The observed relative frequencies of this sequence will keep fluctuating between $\langle\frac{3}{4}, \frac{1}{4}, 0\rangle$, which is reached after each package of consecutive 0's, and $\langle\frac{1}{2}, \frac{1}{2}, 0\rangle$, which is reached after the consecutive 1's. On the whole, there is no limiting relative frequency. Therefore, only if we also provide a hypothesis $H_{\neg\theta}$ that contains all sequences e for which the limiting relative frequencies $f_q(e)$ are not all defined, we can define a real partition of the space K^ω.

In sum, a Bayesian scheme with the hypotheses H_0 and H'_0 involves a partition $\mathcal{C} = \{H_{\neg\theta}, \{H_\theta\}_{\theta\in C}\}$. In this partition we may then assign zero prior probability to all hypotheses other than H_0 and H'_0 to return to the original collection.

Generalisations. Two further remarks on partitions conclude this section. First, note that partitions, as introduced above, can easily be generalised. Every admitted selection function w leads to a specific generalised partition. Such partitions generate predictions that are, as it is called, partially exchangeable. To give an intuitive idea, a partition based on a selection function w results in predictions that are invariant under permutations of any two observations q_{t+1} and $q'_{t'+1}$ as long as we have $w(e_t) = w(e_{t'})$ for the observations preceding these observations. The notion of exchangeability is more elaborately discussed in chapter 3. For partial exchangeability I refer to De Finetti (1972) and Diaconis and Freedman (1980). It leads us too far away from the aim of this chapter to discuss it here.

Second, we may also consider partitions in which more than one selection function is used. As an example, we may want to add the hypothesis H_1, discussed at the start of subsection 1.3.1, in the partition of hypotheses $\mathcal{C}$, which is based on the uniform selection function. To do this, we must first find the vector $\theta \in C$ that results from the probability model of H_1. We must refine this specific hypothesis H_θ in the partition $\mathcal{C}$ into the hypothesis $H^*_\theta \cup H_1$. The hypothesis H^*_θ contains all those infinite sequences that have the limiting relative frequencies θ, except for those sequences e that also have the more complicated pattern described in h_1. If we employ the adapted hypothesis H^*_θ, we may then simply add H_1 to the partition $\mathcal{C}$. Clearly, many much more complicated combinations of hypotheses may be united in a single partition in the same way. I hope that the example here suffices to suggest the general idea.

2.4 Objective probability models

Traditionally the frequentist interpretation of probability serves to describe physical probability or, in one word, chance. In the context of inductive logic chances are used mainly to give an objective content to epistemic probability. Below it will be discussed how this use of the frequentist interpretation relates to hypotheses and probability models in inductive logic. It turns out that chances connect naturally to the definition of hypotheses as elements in $\mathcal{Q}$.

2.4.1 Objective probability

Probability models as models of the world. Now that we have a frequentist interpretation of statistical hypotheses $H_{w\theta}$, it is natural to link the probability models $\mathcal{M}_{w\theta} = \langle \mathcal{Q}_0, p_{[h_{w\theta}]} \rangle$ to models of the world. The probabilities are then interpreted as physical probabilities or chances. Chance models may be taken as weakened versions of deterministic models, in which the probabilities are fixed to 0 or 1 and thus specify a single string e in the observational algebra. An example of a deterministic model is that directly after the appearance of a pack of ducks a tiger appears, after the tiger the ducks hide for one time unit, after which the ducks appear again, and so on. The associated sequence is $e = 0120120120120\ldots$, assuming an empty pond at the onset.

Probability models can be viewed in very much the same way. In the hunting example, we can say that at any time the chances of there being an empty pond, a pack of ducks or a tiger are $\frac{9}{10}$, $\frac{9}{100}$ and $\frac{1}{100}$ respectively. As in the deterministic model, the numbers may here be interpreted as tendencies or propensities of events in the world, and not as expectations of observations. They pertain to something physical. However, if we subsequently take these probability models as input to an inductive argument, they cannot be interpreted as chances anymore. In that case they are objective epistemic probabilities. They are epistemic because they serve as input to an inductive scheme, which concerns opinions and expectations. But they are also objective, because they refer to and are informed by physical probabilities or, in other words, chances in the world.

Subjectivist criticisms. Subjectivists such as Ramsey, De Finetti, Savage and Howson may be opposed to an interpretation of probability models as objective. And because of the strong link between subjectivism and Bayesianism, it is certainly not a standard practice in Bayesianism conceived more broadly to interpret the likelihoods of the hypotheses in terms of objective probability models. Many Bayesians explicitly deny that probability can have objective content. They adhere to a purely subjectivist view, which states that any probability assignment is an expression of belief, and in this role cannot concern some objective aspect of the world. But it goes too far to discuss the relation between strict subjectivism and Bayesianism here. I want to leave it at two remarks to put possible criticisms of the use of objective probability in perspective.

Firstly, it must be emphasised again that the schemes of this thesis already endorse an epistemic interpretation of the probability functions $p_{[e_t]}$. They are explicitly intended to represent beliefs. The fact that, on top of that, the probabilities are objective means that these beliefs refer to and are informed by chances in the world. It accords with generally objectivist statistical inference in science to interpret the probability models associated with hypotheses in this objective epistemic manner. Furthermore, as argued in the foregoing, this objective interpretation helps us to arrive at hypotheses as elements of the observational algebra, which leads to a more natural logical picture of inductive inference. Secondly, the use of a Bayesian scheme does not make the simultaneous use of objective and subjective epistemic probabilities impossible. Pluralist views on probability are as old as Poincaré (1952) and as current as Gillies (2000). Nothing in the Bayesian scheme itself precludes the use of objective probability.

2.4.2 Deriving direct probabilities

But let me return to the main line of this section, which is to connect the probability models to the hypotheses as elements of the observation algebra.

Reformulating the Principal Principle. It may be argued that there is at least the following connection between the objective chances $p_{[h_{w\theta}]}$ in the model of some hypothesis, and the beliefs associated with that hypothesis: if we conditionally accept some model as the model of the world, we must declare its objective chances on observations to be the correct probabilities of the observations. It is a small step to develop this connection into the following principle:

$$p_{[e_0]}(E_t | p_{[h_{w\theta}]}(E_t) = f(w, \theta)) = f(w, \theta). \tag{2.13}$$

In words, conditional on some probability model $p_{[h_{w\theta}]}$ the belief we assign to observations must be the same as the chance that this model prescribes for the observations. This principle has first been formulated by Jeffreys (1939) as the principle of direct probabilities. Later it was given the name of Principal Principle by Lewis (1980), who linked it to a notion of admissible evidence and who took it not as a restriction on subjective probabilities, but rather as an implicit definition of chance. In one interpretation or another, the principle now enjoys wide popularity.

Note that the principle is formulated in terms of a second order epistemic probability $p_{[e_0]}$ over both observations and a probability assignment $p_{[h_{w\theta}]}$. With the above interpretation of hypotheses as elements in $\mathcal{Q}$, we may present an alternative formulation. Now the hypothesis $H_{w\theta}$, as opposed to the probability model $p_{[h_{w\theta}]}$, may be included in the condition:

$$p_{[e_0]}(E_t | H_{w\theta}) = f(w, \theta) \tag{2.14}$$

This accords much better with the Kolmogorovian theory of probability, in which probability can only be assigned to elements in the algebra. Moreover, in this reformulation there is a rather natural argument for adopting the principle of direct probability, as I show below. This in itself presents yet another reason for adopting the interpretation of hypotheses as elements in $\mathcal{Q}$.

From hypothesis to probability model. The natural question is whether the definition of a hypothesis, as an element $H_{w\theta} \in \mathcal{Q}$, determines the probability model associated with it, that is, whether it determines the likelihoods.

For one thing, the likelihoods are restricted by the definition of conditional probability in combination with the axioms of probability:

$$p_{[e_0]}(E_t | H_{w\theta}) = \frac{p_{[e_0]}(H_{w\theta} \cap E_t)}{p_{[e_0]}(H_{w\theta})}, \tag{2.15}$$

where I am assuming that $p_{[e_0]}(H_{w\theta}) > 0$. Whenever a hypothesis $h_{w\theta}$ deems some sequence of observations E_t to be either impossible or positively certain, this carries over to the likelihoods via the definition of $H_{w\theta}$:

$$H_{w\theta} \cap E_t = \emptyset \qquad \Rightarrow \qquad p_{[e_0]}(E_t|H_{w\theta}) = 0, \tag{2.16}$$

$$H_{w\theta} \cap E_t = H_{w\theta} \qquad \Rightarrow \qquad p_{[e_0]}(E_t|H_{w\theta}) = 1. \tag{2.17}$$

In the case of a continuum of hypotheses, it is difficult to make sense of the above expression for conditional probability. Here it is better to resort to the alternative axiomatisation of probability using conditional probability assignments.

Statistical hypotheses are not purely deterministic. So it is not exactly straightforward to link the elements $H_{w\theta}$ in the algebra with likelihoods of these hypotheses, expressed in $p_{[e_0]}(E_t|H_{w\theta})$. However, the following argues that we can employ the characteristics of the elements $H_{w\theta}$ to derive these likelihoods. Recall that the likelihoods are a function of the selection function w and the vector components θ_{qm}, as determined by the probability model of $h_{w\theta}$. For present purposes it therefore suffices to derive the probability vectors only.

Deriving a probability model. To derive these probability vectors, we need one assumption on the probability assignment within the hypotheses: the probability distribution over the possible subsequences e^m constructed from the sequences $e \in H_{w\theta}$ must be uniform. For all sequences $e \in H_{w\theta}$, the function w determines which states occur at which positions. Independently of how these states follow up on each other, we can formulate, for each state separately, the assumption that among the subsequences e^m that build up the sequences e according to the function w there are no preferred ones. That is, for each state m separately, the subsequences e^m are assumed to be equally probable. We may say that this assumption is based on some form of the principle of indifference. The uniform probability within hypotheses expresses that, apart from isolating the hypotheses $H_{w\theta}$ in the algebra and thus focusing on specific sets of sequences e, we have no further reason to prefer one sequence over another.

This uniformity can be used to derive the likelihoods θ_{qm}. I will not give a complete proof but provide a proof sketch only. The general idea in the sketch derives from ergodicity theory: on the assumption of uniform probability within the hypotheses, the long-run relative frequencies may be used as single-case probabilities. The proof has two main ingredients. First, note that for every $e \in H_{w\theta}$, the fractions of results q in the subsequences e^m are always θ_{qm}. Second, note that any possible subsequence e^m with specific relative frequencies can be constructed by permuting, possibly using an infinite number of operations, one single sequence with those frequencies. Therefore, as a further specification of the second ingredient, to state that the probability is uniform over the subsequences is the same as saying that all permutations of a single subsequence e^m are equally probable.

Now consider all the possible permutations of the subsequence e^m, and look at a specific position t within it. After one permutation, any one of the results $e^m(i)$ may end up in the specific position t. But because all permutations are equally probable, all the results $e^m(i)$ are equally probable to end up there. And since, according to the first ingredient, a fraction of θ_{qm} of the results $e^m(i)$ has the value q, the probability of subsequences e^m to have a q at position t is also exactly θ_{qm}. From the characteristics of $H_{w\theta}$ and the assumption of a uniform probability over subsequences we have thus derived the values θ_{qm} for the likelihoods $p_{[e_0]}(Q^q_{t+1}|H_{w\theta} \cap E_t)$. We thus obtain the principle of direct probabilities.

An additional reason for frequentist semantics. The foregoing provides additional reasons for adopting a picture in which hypotheses are associated with elements in the algebra $\mathcal{Q}$. First, the principle can be formulated in terms of a single probability assignment over an observational algebra. There is no need for second-order probability. Second, the principle of direct probability can be derived from an assumption of uniformity. Now it may be that hardcore subjectivists will not be impressed by this line of argument, since they may not be willing to accept such assumptions of uniformity over the probability assignment in the first place. But such subjectivists are not likely to be moved by the principle of direct probability itself either, certainly not if that principle is taken as an implicit definition of the objectivist notion of chance. For all those in favour of some form of the principle of direct probability, the derivation suggested here may be a natural motivation.

2.4.3 Physical models

The above use of objective probability models must not be confused with another such use, which originates in Polya (1954) and has also been discussed in Kuipers (1978). In this use of models, the objective interpretation does not apply to the probability of observations conditional on the hypotheses, but to the predictions that result from the inductive scheme as a whole. The remainder of this section is devoted to some observations on these so-called physical models.

Polya urn models. As a first example, consider the Carnapian prediction rule $pr_{\lambda\gamma}$ for two possible observations with $\lambda = 2$ and $\gamma_q = \frac{1}{2}$, which is sometimes called the straight rule:

$$p(Q^q_{t+1}|E_t) = \frac{t_q + 1}{t + 2}. \tag{2.18}$$

Here t_q is the number of results q in e_t. As discussed more elaborately in the following chapters, this rule is suitable for predicting the results generated by a device that produces results with a constant but initially unknown chance.

Interestingly, there is also an objective interpretation of the probabilities generated by the prediction rule. That is, we can construct a physical system that generates the results q with probabilities exactly matching the predictions. Imagine an urn consisting of one blue and one green ball, so that $q \in \{0, 1\}$. If we pick a ball from the urn at random, each of the two colours have a chance of $\frac{1}{2}$ of being picked, which matches the straight rule for $t = 0$. Now let us say that the first ball is green. We then put back this ball into the urn, and add one further green ball, so that there are 3 balls in the urn, namely two green ones and one blue. If we subsequently pick a ball from the urn, the chances of picking a green or a blue one again match the prediction rule. This time, there is a chance of $\frac{2}{3}$ on green, and of $\frac{1}{3}$ on blue. More generally, by always replacing the ball just picked and adding one of the same colour after that, we can take care that the chances on colours keep matching the predictions.

Petri dish models. This is just one example of a physical system replicating the predictions generated by some inductive scheme. Many more such systems may be constructed. For example, instead of urns with balls we may imagine drawing strings of beads from a jewellery box and adding beads to these strings, with the observations being the individual beads on the strings. Such systems are useful for replicating so-called partially exchangeable processes, using specific selections of strings of beads. Another system, which is more similar to the balls in Polya urns, is presented by colonies of bacteria in a Petri dish. This specific system is particularly suited for replicating predictions that derive from a partition of hypotheses in a Bayesian scheme, as I will now show.

Consider two colonies of bacteria mixed together in a Petri dish, and let us say that at the start of the investigations the two colonies have equal size. The colonies of bacteria differ in the proportion of certain types of cells, for example, blue and green cells. More in particular, for colony H_0 a proportion of $\frac{2}{3}$ is of the blue type, $q = 0$, while the rest is of the green type, $q = 1$. For colony H_1, on the other hand, a proportion of $\frac{2}{3}$ is of the green type while the rest is of the blue type. Now imagine that a scientist samples a single cell from the Petri dish at random, and after determining its type, removes bacteria of the other type from the Petri dish. So if the sampled cell is of the green type,

$q = 1$, a proportion of $\frac{1}{3}$ of colony H_1 is removed, while $\frac{2}{3}$ of H_0 is. After that, the scientist leaves the bacteria to grow back and fill up the Petri dish again. Since the colonies have the same growing rate, colony H_1 is twice as large as H_0 in the new mixture. Further, after the growth both colonies have restored the respective original proportions between the types, which are again $\frac{1}{3}$ and $\frac{2}{3}$ or vice versa. The scientist may then repeat the whole procedure.

As already suggested by the notation, the determination of the type of a randomly sampled cell in the above experimental setting replicates exactly the predictions resulting from a Bayesian scheme with two hypotheses, H_0 and H_1, which have likelihoods $\langle \frac{2}{3}, \frac{1}{3} \rangle$ and $\langle \frac{1}{3}, \frac{2}{3} \rangle$ for the possible results $q = 0$ and $q = 1$, respectively. It will be clear that the setting may be generalised to any number of types or colonies. Perhaps surprisingly, taking an infinity of colonies all associated with different proportions $\langle \theta, 1 - \theta \rangle$ comes down to the same system as the Polya urn described above. The next chapter returns in more detail to this equivalence, which is a special case of De Finetti's representation theorem.

2.5 Comparing inductive schemes

In this last section I consider the two schemes presented in the preceding chapter. I first show that with hypotheses as elements of the observation algebra $\mathcal{Q}$, the Bayesian and Carnapian schemes are formally equivalent. The second subsection discusses whether the two schemes allow for the same range of inductive predictions.

Carnapian scheme over a σ-algebra. Recall from chapter 1 that a Carnapian prediction rule *pr* comes down to a complete probability assignment p over the algebra $\mathcal{Q}_0$. By contrast, when using a frequentist semantics for hypotheses a Bayesian scheme minimally requires an extended algebra $\mathcal{Q} = \sigma(\mathcal{Q}_0)$. This is because the frequentist hypotheses $H_{w\theta}$ are associated with elements in $\mathcal{Q}$ that fall outside the algebra $\mathcal{Q}_0$. Fortunately, as proved in Billingsley (1995: 36-41), every probability function over an algebra $\mathcal{Q}_0$ can be extended uniquely to a probability function over the σ-algebra generated by it. So the probability function p over the algebra $\mathcal{Q}$ is already implicit to the Carnapian scheme.

There is a considerable conceptual price for the extension of the probability from $\mathcal{Q}_0$ to a unique probability over the σ-algebra $\mathcal{Q}$. For one thing, the whole idea of a σ-algebra seems at variance with empiricist views: it allows for sets such as $H_{w\theta}$, which consist of infinite disjunctions of infinite conjunctions of single observations. Such sets seem entirely unempirical. Moreover, even if we grant the use of a σ-algebra, the unique extension of the probability over it can only be derived if the first three Kolmogorov axioms are supplemented with an axiom on so-called σ-additivity. This axiom states that the probability of an infinite disjunction of disjoint sets can be written down as an infinite sum of the probabilities of these sets. Following the discussion in Williamson (1999), the axiom has a dubitable status for both empiricists and subjectivists. For example, it is impossible to justify the axiom of σ-additivity by reference to betting contracts.

Nevertheless, given the conceptual clarity offered by the use of frequentist hypotheses, I am myself more than willing to pay the conceptual price. In fact, I find the use of infinite set operations and σ-additivity only marginally more farfetched than the whole project of finding a formal framework for inductive inference, such as probabilistic inductive logic. From this point of view, the use of frequentist hypotheses is a small and profitable investment.

The razor of Von Mises. It remains to be seen that the Bayesian scheme can indeed be defined by a single probability assignment over the algebra $\mathcal{Q}$. I will now present a more detailed argument for that, and show that it presents a further reason for the frequentist semantics.

Recall that in the scheme of chapter 1, each hypothesis $h_j \in \mathcal{H}$ is associated with a labelled algebra $\{h_j\} \times \mathcal{Q}_0$, so that the space over which we define the probability functions $p_{[e_t]}$ is given by $\mathcal{H} \times \mathcal{Q}_0$. With the definition of hypotheses as elements in $\mathcal{Q}$, we can reform this scheme in two steps. As a first step, we can extend the probability assignments $p_{[h_j]}$ to a probability over the extended algebra $\mathcal{Q}$ assuming σ-additivity. This yields a probability $p_{[h_j]}$ over each algebra $\{h_j\} \times \mathcal{Q}$. We are then facing a rather awkward conceptual possibility; we can employ the hypotheses H_j, as elements in the algebra $\mathcal{Q}$, as arguments in their own probability assignments: $p_{[h_j]}(H_j)$. However, for frequentist hypotheses h_j the weak law of large numbers entails that the probability of the set H_j, according to its own probability model, is $p_{[h_j]}(H_j) = 1$. Thus, in the extension of $p_{[h_j]}$ to the algebra $\mathcal{Q}$, all probability is located within the tail event $H_j \in \mathcal{Q}$.

This brings us to the second step in reforming the scheme. If we consider the extended algebra $\mathcal{H} \times \mathcal{Q}$, it seems that we can merge the algebras along the range of hypotheses h_j. The sets in the respective algebras $\{h_j\} \times \mathcal{Q}$ that carry the probability mass, namely the hypotheses H_j, do not overlap. In each of the hypotheses $\{h_j\} \times \mathcal{Q}$ the probability $p_{[h_j]}$ is concentrated within the sets H_j, but these latter sets are mutually exclusive. Because of this, we can harmlessly compress the range of algebras $\{h_j\} \times \mathcal{Q}$ into a single algebra $\mathcal{Q}$, within which the hypotheses are simply given by the sets H_j. So the second step in reforming

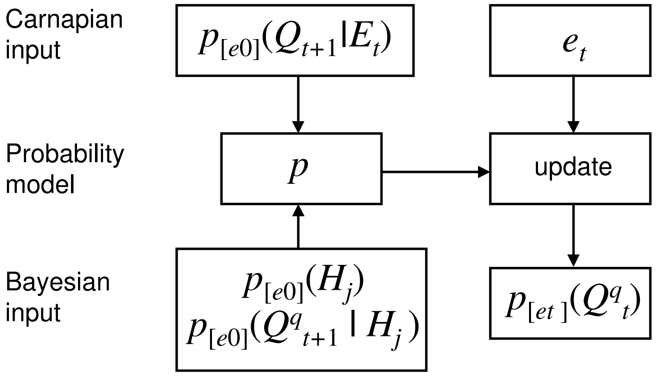


Figure 2.1: The Bayesian scheme and the Carnapian scheme come down to the same thing. Both determine a probability over the extended observation algebra $\mathcal{Q}$. The Bayesian scheme can be viewed as the microstructure of the Carnapian scheme. Prediction rules result from integral expressions over statistical hypotheses in a Bayesian scheme.

the scheme is that instead of using a range of extended observational algebras $\mathcal{H} \times \mathcal{Q}$, we can make do with a single extended algebra $\mathcal{Q}$.

This is where the razor of Von Mises finds its application. In the Kolmogorovian picture of the preceding chapter, hypotheses h_j are each associated with a separate algebra $\mathcal{Q}_0$. But following the two steps of the foregoing, it seems that we can trim away this abundance of algebras and leave only the single observational algebra $\mathcal{Q}$. This can be done by identifying the probability models with elements in the extended algebra according to the frequentist interpretation. Indeed, the razor of von Mises shaves the long Kolmogorovian beard of algebras.

It follows that the Carnapian and Bayesian schemes are formally equivalent. Both schemes may be defined by a single probability p over the observation algebra $\mathcal{Q}$. The choice of this probability p may be effected by choosing a prediction rule $pr(q, e_t)$, but also by choosing a partition of hypotheses, associated with a range of probability models, along with a prior probability assignment over these hypotheses. As depicted in figure 2.1, the Bayesian scheme thus emerges as a detailed version of the Carnapian scheme, and not as the generalisation which is presented in chapter 1. The Bayesian scheme may be said to present the microstructure underlying the Carnapian scheme.

Reasons for a frequentist semantics. The frequentist semantics makes for a more integrated picture of the Bayesian scheme. First, the definition of statistical hypotheses as elements in the observation algebra avoids the use of second-order probability, thus connecting better to the Kolmogorov axioms. Second, in view of the empiricist roots of inductive logic I think that the use of a single algebra $\mathcal{Q}$ is more natural than the use of $\mathcal{H} \times \mathcal{Q}_0$. Third, it is advantageous that both objective and subjective probabilities find a natural place in the Bayesian scheme, making the discussion on the interpretations of probability somewhat less ardent. The likelihoods may even be derived from frequentist hypotheses at the cost of a uniformity assumption. And finally, as discussed below, it is rather nice that the difference between the Carnapian and the Bayesian scheme can be traced back to an enrichment of the language or algebra that is used in these schemes.

Limits of the Bayesian scheme. The formal equivalence that is derived by means of the frequentist semantics also adds to the urgency of the question why we are considering two schemes in the first place. After all, if they can be written down in the same format, studying one of them will be enough, and simplicity considerations then lead us to the Carnapian scheme. The remainder of this section is concerned with this question. A natural answer is that the two schemes allow us to express different probability assignments over the algebra. It may seem obvious that any Bayesian scheme comes down to some Carnapian scheme: the Bayesian scheme generates predictions, and these predictions can always be summarised in a prediction rule, and thus in a Carnapian scheme. But in reality it proves very difficult to find elegant expressions for the particular prediction rule that corresponds to some Bayesian scheme. This may be a reason for employing a Bayesian scheme after all. To such reasons I will come back more elaborately in chapter 3.

Focusing only on the range of prediction rules and not on the possibility to express these rules in a convenient form, the Carnapian scheme can accommodate any Bayesian scheme. But it is not so obvious that the Bayesian scheme can accommodate any Carnapian prediction rule. Clearly, the equivalence between the Carnapian and the Bayesian scheme becomes trivial if we are allowed to consider any hypothesis we like. We may then simply use one hypothesis h in the Bayesian scheme, and give it likelihoods that correspond to the predictions of the rule that we want to replicate. But there are many prediction rules that do not coincide with a probability model associated with a frequentist hypothesis. As argued above, the class of frequentist hypotheses is rather restricted.

Frankly, I do not know whether prediction rules can always be replicated with a Bayesian scheme using frequentist hypotheses, and I also do not know what kind of argument may eventually settle that matter. But for lack of any such argument towards the affirmative, it seems that Carnapian schemes have a slight advantage over Bayesian schemes. With the use of frequentist hypotheses we run the risk of unknowingly restricting the range of possible prediction rules.

It may be argued that statistical hypotheses that are not frequentist do not deserve consideration in the first place. If indeed there are restrictions imposed by only using frequentist hypotheses, we must accept these simply because there is otherwise no proper interpretation for the probabilities in the scheme. This is indeed a very dogmatic reaction. It seems unwise to make it part and parcel of the Bayesian scheme itself.

Hypotheses as enrichment of the language. There is another difference between Carnapian and Bayesian schemes that is significant in the next chapter, and in this thesis more generally. The two schemes are equal in the sense that they both determine a single probability function over $\mathcal{Q}$, but they differ in that they determine it by specifying different sets. To return to one of the points of section 2.3.1, the introduction of hypotheses as tail events in the algebra amounts to an enrichment of the language used to express inductive predictions. With the sets $H_{w\theta}$, new events or terms are added to the observational terms already present in the algebra $\mathcal{Q}_0$. This enrichment allows us to specify the input probability of inductive arguments in a different way, which will be seen to have clear conceptual advantages. It will be argued below that the Bayesian scheme offers a more detailed description of inductive predictions, and a more natural grip on the inductive predictions themselves. In particular, the Bayesian scheme allows us to express the projectability assumptions that precede any inductive argument. The Bayesian scheme is therefore more suitable than the Carnapian for conceptualizing, applying, and adapting inductive predictions.

3

Hypotheses as Inductive Assumptions

This chapter reveals the advantages of Bayesian schemes in generating inductive predictions. It discusses Carnap-Hintikka inductive logic, and two ways in which the Bayesian schemes may expand it. Bayesian schemes are then illustrated with two partitions. One partition results in the Carnapian continuum of prediction rules, the other results in predictions typical for hasty generalisation. Following these examples I argue that choosing a partition comes down to making inductive assumptions on patterns in the data, and that by choosing appropriately any inductive assumption can be made.

The inductive predictions in this chapter are cast in the framework of section 1.2.1, and in particular in the Bayesian scheme introduced in section 1.3. Familiarity with these sections is essential to an understanding of this chapter. The introduction to this chapter is especially useful for those readers who have not read the preceding chapters, but apart from that it also indicates the main line of this chapter.

3.1 Introduction

Inductive predictions. This chapter concerns inductive predictions, taken as the result of inductive inferences. The premisses of an inductive inference include a set of observations and possibly some further assumptions. The conclusions may be predictions or general hypotheses, where predictions concern future observations, and general hypotheses are observational generalisations of some sort or other. For example, from the fact that some internet start-up has had decreasing stock price on all days until now, we may derive that the next day it will have decreasing stock price as well. This is a prediction about a single observation, namely the decrease of stock price on the next day, based on data of the stock price movements on all days until now. From the same data set we may also derive that the internet start-up will have a decreasing stock price on all future days, which is a general hypothesis about the observations.

The Bayesian scheme uses hypotheses to arrive at predictions. The data are first reflected in an opinion about a partition of hypotheses. For example, from the data on decreasing stock price we first derive an opinion about hypotheses on the state and nature of the internet start-up. The predictions on the internet start-up are subsequently derived from this opinion about hypotheses, together with the data. These predictions and opinions about hypotheses are thus expressed in terms of probability functions. In sum, this chapter concerns probabilistic inductive inferences in which hypotheses are used for making predictions. In conformity with the preceding chapters, I will say that such predictions are based on Bayesian schemes, or alternatively, based on partitions.

Organisation of chapter. The main line of the chapter is the following. First I show that the Bayesian scheme enables us to describe predictions typical for hasty generalisation. The predictions can be generated by choosing a specific partition of hypotheses for the scheme. This example triggers two separate discussions on the function of hypotheses in the Bayesian schemes. The main conclusion of the discussion in this chapter is that hypotheses are tools for making inductive assumptions. They determine which patterns are identified in the data, and subsequently projected onto future observations. Another discussion, which concerns Bayesian schemes in relation to the problem of induction, is presented in chapter 7.

In more detail, the structure of the chapter is as follows. In section 3.2, I introduce the dominant tradition in formalising inductive predictions, called Carnap-Hintikka inductive logic. The Bayesian scheme of this chapter is seen to expand the Carnap-Hintikka tradition. For the Bayesian scheme itself I refer to section 1.3. Section 3.3 considers two prediction rules working on the same data, but based on different partitions, and shows that they result in different predictions. It further elaborates the relation between inductive predictions and the Carnap-Hintikka tradition. In section 3.4 the examples are given a further philosophical interpretation. Specifically, the hypotheses are related to inductive assumptions. The conclusion connects these insights to the schemes of the preceding chapters.

3.2 Carnap-Hintikka inductive logic

This section discusses the Carnap-Hintikka tradition of inductive logic. It emphasises two characteristic features of this tradition: its focus on exchangeable prediction rules, and its reluctance, certainly within the Carnapian literature, to employ general or statistical hypotheses. The inductive predictions of this chapter, and more generally of this thesis, extend the Carnap-Hintikka tradition by departing from these two features.

Carnapian prediction rules. Let me briefly rehearse Carnapian predictions for the purpose of this chapter. Recall that predictions are probabilities over future observations based on observations and further assumptions. The observations are encoded in indexed natural numbers q_i, and collected in ordered tuples $e_t = \langle q_1, q_2, \ldots, q_t \rangle$. The further assumptions can be encoded in some collection of parameters X. We may then construct a prediction rule $pr_X(q, e_t)$ for the next observation having the result q in terms of a probability function of the preceding observations e_t, the next observation q_{t+1} and the collection of parameters X. Inductive predictions can be studied by designing and comparing classes of such probability functions, which I call inductive prediction rules.

An exemplary class of probabilistic inductive inference rules for making predictions is given by the $\lambda\gamma$ rules referred to earlier, as developed in Carnap (1950, 1952) and defined fully in Stegmüller (1973):

$$pr_{\lambda\gamma}(q, e_t) = \left(\frac{t}{t+\lambda}\right)\frac{t_q}{t} + \left(\frac{\lambda}{t+\lambda}\right)\gamma_q. \tag{3.1}$$

The function $pr_{\lambda\gamma}$, the probability for observing q at time $t+1$, is a weighted average of the observed relative frequency $\frac{t_q}{t}$ of instances of q among the ordered set of known observations e_t, and the preconceived or virtual relative frequency of observing q, denoted γ_q. The weights depend on the time t and a learning rate λ. With increasing time, the weighted average moves from the preconceived to the observed relative frequency. The learning rate λ determines the speed of this transition.

After Carnap, inductive prediction rules have been studied extensively. Axiomatisations, elaborations and synthesisations of inductive prediction rules have been developed by Kemeny (1963), Hintikka (1966), Carnap and Jeffrey (1971), Stegmüller (1973), Hintikka and Niiniluoto (1976), Kuipers (1978), Costantini (1979), Festa (1993) and Kuipers (1997). To this research tradition I refer with the names of Carnap and Hintikka.

Exchangeability. Most of the work in this tradition concerns exchangeable prediction rules. Exchangeability of a prediction rule means that the predictions do not depend on the order of the incoming observations. These exchangeable rules typically apply to settings in which the events producing the observations are independent. So they have a very wide range of application. Moreover, on the assumption that the prediction rule is exchangeable, it can be proved that the predictions eventually converge to optimal values. That is, if the observations are produced by some process with constant objective chances, the predictions of an exchangeable rule will, according to Gaifman and Snir (1982), almost always converge on these chances, whatever the further initial assumptions. Both for their range of applicability and for this convergence property, exchangeable rules are a main focus in the Carnap-Hintikka tradition.

Representation theorem. The second feature of this tradition that I want to emphasise can only be made explicit after presenting the connection, in both directions, between the exchangeability of observations and the independence of the events that may be supposed to produce these observations. For this I must first elaborate on the notion of independence. A first component of this notion is the assumption that the events producing the observations are in fact part of some underlying process. A second component is that if this underlying process generates the events with constant objective chances, then the chance of an event is independent of events occurring before or after it, so that we can speak of independent events.

The connection reaching from exchangeability to independence is then established by the representation theorem of De Finetti, as discussed in (1964). This theorem shows that any exchangeable prediction rule can be represented uniquely in a Bayesian scheme, using hypotheses that are associated with constant chance processes. Section 3.3 deals with this representation theorem in some more detail. The connection reaching from independence to exchangeability, on the other hand, is established by the fact that any Bayesian scheme using hypotheses on constant chance processes must result in an exchangeable prediction rule. This is seen most easily from the fact that updating the probability over the hypotheses for new observations is a commutative operation. The order of such updates is therefore inessential to the resulting probability assignment over the hypotheses, and thus inessential to the predictions resulting from this assignment. Again, section 3.3 discusses this in more detail. For now it is important to note that the assumption of the independence of the events producing the observations can be equated with the use of exchangeable prediction rules for these observations.

Against general hypotheses. I can now make explicit the second characteristic feature of the Carnap-Hintikka tradition. De Finetti interpreted the representation theorem as a reason to omit all reference to underlying processes, and to concentrate on exchangeable prediction rules instead. As Hintikka (1970) argues, this is not so much because of a subjectivist dislike of the objective chances featuring in the underlying processes. Rather it is because these chance processes are described by general hypotheses, which cannot be decided with finite data. De Finetti deemed such hypotheses suspect for empiricist reasons.

In a similar vein, Carnap maintained that all universal hypotheses have measure zero. Now there are large differences between De Finetti and Carnap, but both used the representation theorem to show that it is simply unnecessary to employ chance processes. We can obtain the same results using the exchangeability of the prediction rule, and in this way we stay closer to the empiricist roots of inductive logic.

In line with this, most of the Carnap-Hintikka tradition focuses on the properties of prediction rules, such as exchangeability, and eschews reference to the chance processes that may be underlying these rules. Prediction rules with this feature I call Carnapian. This terminology signals that this second feature is not fully applicable to the Hintikka part of the Carnap-Hintikka tradition. Indeed, in Hintikka (1966) and Tuomela (1966) we find a different attitude towards underlying chance processes, or at least towards the use of hypotheses in inductive logic. More in particular, Hintikka employs universal generalisations on observations in the construction of his $\alpha\lambda$ continuum of inductive prediction rules. Tuomela discusses more complicated hypotheses on ordered universes, and refers to Hintikka for the construction of prediction rules based on these universal statements. Both these authors thus employ hypotheses to inform predictions in a specific way.

Innovations of this chapter. While this already presents a valuable extension, I feel that hypotheses have not been employed with full force in the Carnap-Hintikka tradition. Perhaps some empiricist feelings have remained, which have curbed the further development of Hintikka systems. The $\alpha\lambda$ continuum of Hintikka offers little room for varying the kinds of hypotheses used, since the continuum concerns universal generalisations only. It is certainly an advantage that, in the improved versions of Hintikka and Niiniluoto (1976) and Kuipers (1978), the role of these generalisations is not entirely determined by the single parameter α. However, as Hintikka himself remarks in (1997), it is eventually much more convenient to be able to express such universal statements in terms of proper premisses, so that other kinds of universal statements can be employed too, and also controlled more naturally. However, many prediction rules in which the use of specific universal statements seems very natural do not employ such statements in their construction. Take, for example, the inductive prediction rules for Markov chains by Kuipers (1988) and Skyrms (1991), and the prediction rules describing analogical reasoning by Niiniluoto (1981) and Kuipers (1984). Here the construction of the rules is based on particular predic-

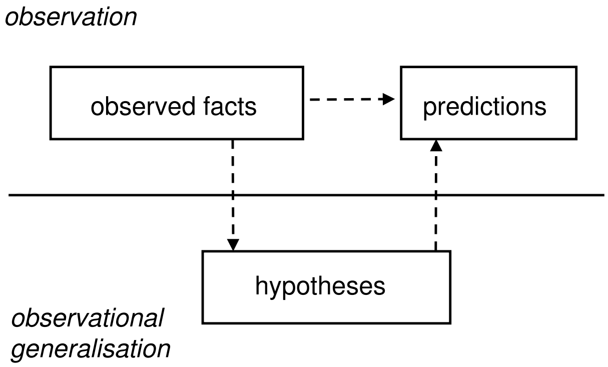


Figure 3.1: The Bayesian scheme employs hypotheses to mediate between observations and predictions.

tive properties, which, as an aside, are all non-exchangeable. Underlying chance processes are not really used in the construction of these rules.

I can now indicate more precisely the innovations that this chapter offers. It extends the Carnap-Hintikka tradition in inductive logic in two ways, connected to the two characteristic features noted above. First, this chapter proposes a prediction rule that is not exchangeable, by adding hypotheses concerning a particular deterministic pattern to an existing partition of constant chance hypotheses. Second, and following up on this, it advocates the explicit use of chance processes, or hypotheses on such processes, in the definition of inductive predictions. In accordance with the Bayesian scheme of chapter 1, and as illustrated in figure 3.1, hypotheses are used to mediate between observations and predictions. The claim following from this is that partitions of hypotheses are a tool in making assumptions about patterns in data, which widens the scope of the Carnap-Hintikka tradition. In connection with the discussions in chapters 1 and 2, it may be said that this chapter presents the main reason for preferring the rather complicated Bayesian scheme over the Carnapian.

Remark and disclaimer. In connection with chapter 2, it may be noted that the hypotheses that may be used in a Bayesian scheme belong to the class of frequentist hypotheses. One aspect of this class is in this chapter given a further illustration. Recall that the class also encompasses the hypotheses $H_{w\theta}$ employing a selection function w that sometimes assigns state $w(e_t) = 0$ to sequences e_t that are not given a zero probability by the hypotheses, but that only does so for a finite number of sequences e_t. Such hypotheses, it was argued, still allow for being connected to tail events in the algebra $\mathcal{Q}$. This chapter employs hypotheses of exactly this sort, namely the so-called crash hypotheses.

Finally, let me disclaim the treatment of some topics that otherwise complicate the discussion too much. First, it can be noted that the prediction rules of this chapter are somewhat similar to those in the paper by Tuomela on ordered universes. Both focus on predictions based on the specific patterns in the data. But, for lack of space, I will not elaborate on this similarity in the following. Second, I will not discuss the representation theorem of De Finetti in full generality, and similarly I will not touch upon the various brands of partial exchangeability. The focus of this chapter is on a particular non-exchangeable prediction rule, generated by hypotheses concerning particular chance processes, and on the moral that derives from the use of such hypotheses. There are excellent discussions of representation theorems on offer.

3.3 Examples on crash data

This section gives two applications of the scheme of section 1.3. The first application employs hypotheses on constant chances for the observations, and results in the Carnapian $\lambda\gamma$ rules. This also serves as an illustration of the representation theorem of De Finetti. The second application provides an extension of the Carnap-Hintikka tradition. Apart from the hypotheses on constant chances, it employs hypotheses concerning a particular pattern in the data. The resulting predictions are not covered by the $\lambda\gamma$ continuum, and they are not in general exchangeable.

3.3.1 Bernoulli partition

Stock market data. The example concerns stock price movements. Consider the following strings of data, representing stock prices for $t = 35$ days. In the strings, $q_i = 0$ if the stock price decreased over day i, and $q_i = 1$ if the stock price increased or remained unchanged over that day. Here are two possible histories of the prices of a stock of some internet start-up:

$$\begin{aligned} e_{35} &= 01000100000010110000010000000010000, \\ e^*_{35} &= 01001010111010000000000000000000000. \end{aligned}$$

Note that e_{35} and e^*_{35} have an equal number of trading days i with $q_i = 1$, but that the order of increase and decrease is different for the strings. In particular, e^*_{35} shows what can be called a crash: from some day onwards we only observe decreasing stock price.

Defining Bernoulli hypotheses. Now imagine a marketeer who aims to predict stock price movements based on observed price movements over foregoing trading days. Further, assume that she employs the partition $\mathcal{B}$ with a continuum of hypotheses to specify her predictions. To characterise the hypotheses, define

$$f(e) = \lim_{t\to\infty} \frac{1}{t} \sum_{i=1}^{t} e(i). \tag{3.2}$$

For any infinitely long sequence of days e, the function $f(e)$ gives the ratio of trading days i for which $e(i) = 1$. Note that $f(e)$ is undefined for some of the $e \in K^{\omega}$. Now define I_{h_θ} as follows:

$$I_{h_\theta}(e) = \begin{cases} 1 & \text{if } f(e) = \theta, \\ 0 & \text{otherwise,} \end{cases} \tag{3.3}$$

in which $\theta \in [0, 1]$. Further define $I_{h_{\neg\theta}}(e) = 1$ if $f(e)$ is not defined, and $I_{h_{\neg\theta}}(e) = 0$ otherwise. Then $\mathcal{B} = \{H_{\neg\theta}, \{H_\theta\}_{\theta\in[0,1]}\}$ is a partition including a continuum of hypotheses on relative frequencies: every e belongs to a unique hypothesis. I call $\mathcal{B}$ the Bernoulli partition, after Johan Bernoulli who first studied chance processes of this form.

Assume that the marketeer employs the following input probabilities:

$$\begin{aligned} \int_0^1 p_{[e_0]}(H_\theta) d\theta &= 1, & (3.4) \\ p_{[e_0]}(H_\theta) &= 1, & (3.5) \\ p_{[e_0]}(H_{\neg\theta}) &= 0, & (3.6) \\ \forall i > 0 : \quad p_{[e_0]}(Q^q_{i+1}|H_\theta) &= \begin{cases} \theta & \text{if } q = 1, \\ 1-\theta & \text{if } q = 0. \end{cases} & (3.7) \end{aligned}$$

where again $\theta \in [0, 1]$. Equations (3.4) and (3.5) state that the probability distribution over the hypotheses H_θ is uniform. This may be motivated with an appeal to the principle of indifference or some other symmetry principle. Equation (3.6) states that those sequences e in which the frequency of trading days with $e(i) = 1$ has no limit are negligible. This assumption is not compulsory. It is here made for computational simplicity, as it allows us to ignore hypothesis $H_{\neg\theta}$ in further calculations. Moreover, it is required if we want to illustrate the representation theorem.

Equation (3.7) can be motivated with the restriction on likelihoods as discussed in section 2.4. We may further assume that at every i the Bayesian agent

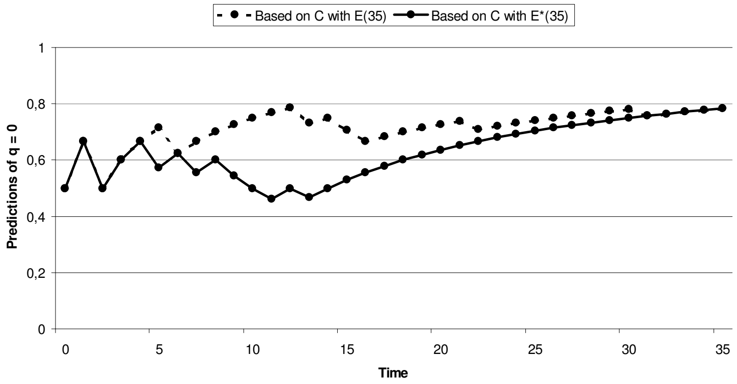


Figure 3.2: Predictions $p_{[e_t]}(Q^0_{t+1})$ against time t, based on the partition $\mathcal{B}$. The dashed line shows the predictions for the normal data e_{35}, the unbroken line shows the predictions for the crash data e^*_{35}.

updates the likelihood that is used in the next prediction to

$$p_{[e_i]}(Q^q_{i+1}|H_\theta) = p_{[e_0]}(Q^q_{i+1}|H_\theta), \tag{3.8}$$

so that, in conformity with the definition of the hypotheses H_θ, the accumulation of data e_i does not change the original likelihoods. The hypotheses H_θ on relative frequencies then have constant likelihoods $p_{[e_i]}(Q^q_{i+1}|H_\theta) = \theta$.

Resulting predictions. With the prior density $p_{[e_0]}(H_\theta)$ and the likelihoods $p_{[e_i]}(Q^q_{i+1}|H_\theta)$, we have specified all probabilities that are needed. We can compute the predictions on next observations $p_{[e_t]}(Q^q_{t+1})$ that the marketeer makes when confronted with the observations e_{35} and e^*_{35} respectively. I have calculated these predictions and depicted them in figure 3.2. In the remainder of this subsection I make some remarks on these predictions, and on the Bayesian scheme with the Bernoulli partition in general.

Note that after 32 days, the predictions in figure 3.2 are the same for both strings of data. This shows the exchangeability of the above Bayesian update procedure. Probability assignments after any e_t are invariant under the permutation of results $e_t(i)$ within that e_t, and as said, e_{35} and e^*_{35} have the same number of 1's. For both e_{35} and e^*_{35} it is further notable that the predictions $p_{[e_i]}(Q^0_{i+1})$ converge to 1. The speed of convergence, however, decreases with the addition of further instances of $e_t(i) = 0$. More precisely, the second derivative to time of the predictions, taken as a function over time, is negative. Thus the predictions do not accommodate the fact that the data e^*_{35} may be the result of a crash.

The Bayesian scheme using $\mathcal{B}$ illustrates the representation theorem of De Finetti. The hypotheses H_θ are the hypotheses on processes with constant chances that I alluded to in sections 3.2. The representation theorem is that any exchangeable prediction rule $pr_X(q, e_t)$ can be represented in a Bayesian scheme with the partition $\mathcal{B}$. Different exchangeable prediction rules may be defined by choosing different priors $p_{[e_0]}(H_\theta)$. For example, choosing a Dirichlet density for $p_{[e_0]}(H_\theta)$ results in a $\lambda\gamma$ prediction rule. As described in Festa (1993), the parameters of the Dirichlet density fix the values of γ and λ in this rule. Specifically, choosing the uniform prior of equation (3.5) results in the so-called straight rule, which has the parameters $\lambda = 2$ and $\gamma_q = \frac{1}{2}$. Note however that the range of the representation theorem is much wider than the equivalence of the $\lambda\gamma$ rules and the Dirichlet distributions over $\mathcal{B}$.

As section 3.2 indicated, the representation theorem was welcomed as a way to replace Bayesian schemes using $\mathcal{B}$ for exchangeable prediction rules. One reason for the replacement was that the Bayesian schemes committed to the assumption of underlying chance processes and the assignment of probability to universal hypotheses. Another, more immediate reason for not using the Bayesian schemes may be that they are unnecessarily roundabout: in the end they generate the same predictions as the Carnapian scheme. In the foregoing and in the following, however, I explicitly use the Bayesian schemes to design and study predictions. In section 3.4 I argue that there are independent reasons for doing so.

3.3.2 Crash hypotheses

Alternative hypotheses. Figure 3.2 shows the inductive predictions of a marketeer who is not sensitive to the possibility of a crash. Below I alter the partition in such a way that this sensitivity is modelled. This is done by adding hypotheses to the Bernoulli partition, thus implicitly altering the resulting prediction rule. In particular, I add the hypotheses $g^q_{\gamma\lambda\tau}$ to the Bernoulli partition, the meaning of which can be phrased as follows: until trading day τ, stock price behaves like the $\lambda\gamma$ rule says, but from trading day τ onwards, all stock price movements are q.

Let us denote the partition consisting of the Bernoulli hypotheses h_θ and the crash hypotheses $g^q_{\gamma\lambda\tau}$ with $\mathcal{C}$. The crash hypotheses can be associated with sets $G^q_{\gamma\lambda\tau}$ in $\mathcal{Q}$ using a characteristic function that selects for crashes:

$$I_{g^q_{\gamma\lambda\tau}}(e) = \begin{cases} 1 & \text{if } e(\tau) \neq q \wedge \forall i > \tau : e(i) = q, \\ 0 & \text{otherwise,} \end{cases} \tag{3.9}$$

$$G^q_{\gamma\lambda\tau} = \{e : I_{g^q_{\gamma\lambda\tau}}(e) = 1\}. \tag{3.10}$$

Note that the parameters γ and λ do not occur in the definition of the sets $G^q_{\gamma\lambda\tau}$. The sets can be defined solely on the basis of the crash starting at time τ.

The hypotheses $G^q_{\gamma\lambda\tau}$ can be given likelihoods that reflect the above meaning:

$$p_{[e_0]}(Q^q_{i+1}|G^{q'}_{\gamma\lambda\tau} \cap E_i) = \begin{cases} \frac{i_q+\lambda\gamma_q}{i+\lambda} & \text{if } t < \tau, \\ 1 & \text{if } i = \tau - 1,\ q \neq q', \text{ or } i \leq \tau,\ q = q', \\ 0 & \text{if } i = \tau - 1,\ q = q', \text{ or } i \leq \tau,\ q \neq q', \end{cases} \tag{3.11}$$

where i_q denotes the number of results q in the observations e_i. The last two clauses of the likelihood definition are motivated with the definition of the sets $G^q_{\gamma\lambda\tau}$. However, as there is no restriction on the first $\tau - 1$ observations in these sets, there is no restriction motivating the first clause. The likelihoods before τ may be chosen in accordance with the predictions generated by the partition $\mathcal{B}$, so that, when the hypotheses $G^q_{\gamma\lambda\tau}$ are added to that partition, they only distort the predictions insofar as there is a crash pattern in the data. Note that the likelihoods of $G^q_{\gamma\lambda\tau}$ thus depend on the actual data e_i. This means that the likelihoods change with the update of every observation before τ.

Choosing a prior. The hypotheses $G^q_{\gamma\lambda\tau}$ may be given prior probabilities of the following form:

$$p_{[e_0]}(G^0_{\gamma\lambda\tau}) = \alpha\,(1-\delta)\,\delta^\tau, \tag{3.12}$$

$$p_{[e_0]}(G^1_{\gamma\lambda\tau}) = 0, \tag{3.13}$$

where $\tau > 0$ and $0 < \delta < 1$, so that $(1-\delta)\delta^\tau$ is a discount factor, which describes how a trader slowly grows less suspicious for crashes. The factor α is the total probability that is assigned to all the crash hypotheses. From the definition of the discount factor, we have $\alpha = \sum_{\tau=0}^{\infty} p_{[e_0]}(G^0_{\gamma\lambda\tau})$, so that we must choose $0 < \alpha < 1$. Note that because of equation (3.13), booming markets, in which from some time onwards prices only go up, are not considered.

The probability $1-\alpha$ can be divided over the remaining hypotheses from $\mathcal{B}$ according to

$$\int_0^1 p_{[e_0]}(H_\theta)d\theta \quad = \quad 1-\alpha, \tag{3.14}$$

$$p_{[e_0]}(H_\theta) \quad = \quad 1-\alpha, \tag{3.15}$$

where in this case $\theta \in (0,1]$. The likelihoods of the crash hypotheses can be made to accord with this prior by setting $\lambda = 2$ and $\gamma_q = \frac{1}{2}$. Note from the domain of θ that the hypothesis H_0 is excluded from the subpartition $\mathcal{B}$. This is because all $e \in G^0_{\gamma\lambda\tau}$ have the relative frequency $f(e) = 0$, so that for each τ we have $G^0_{\gamma\lambda\tau} \subset H_0$. However, according to the original likelihoods of H_0 the hypotheses $G^0_{\gamma\lambda\tau}$ must have zero probability within H_0, because any observation of $q = 1$ is given zero probability within it. The simplest solution to all this is to exclude the hypothesis H_0 from the partition altogether. Since hypothesis H_0 had a negligible measure in the original Bayesian scheme with $\mathcal{B}$ anyway, banning it from the combined partition $\mathcal{C}$ does not affect the prediction rule that was initially derived for $\mathcal{B}$.

In sum, we have created a new partition $\mathcal{C}$, including both H_θ and $G^0_{\gamma\lambda\tau}$. As will be seen, updating over this partition generates predictions which express a sensitivity for crashes. Choosing values for α and δ determines to what extent this sensitivity influences the predictions. Admittedly, the partition $\mathcal{C}$ involves considerable idealisations, for example that a crash lasts forever and that the prior probability for a crash slowly diminishes. These idealisations are not compulsory: the Bayesian scheme offers space for further elaborations in these respects. In the following, however, I want to focus on the fundamental possibilities that the freedom in choosing partitions presents. The idealisations of $\mathcal{C}$, and the ways to avoid them, are not discussed here.

Resulting predictions. We can calculate the predictions $p_{[e_t]}(Q^q_{t+1})$ using the partition $\mathcal{C}$. Figure 3.3 shows a comparison of two marketeers confronted with the crash data e^*_{35}. The diamond curve shows the predictions based on the use of the partition $\mathcal{C}$, and the bullet curve shows the predictions of the Bernoulli partition $\mathcal{B}$. The hypotheses $G^0_{\gamma\lambda\tau}$ of this particular update have $\alpha = \frac{1}{2}$ and $\delta = \frac{4}{5}$. Note that the predictions based on $\mathcal{C}$ deviate from the predictions based on $\mathcal{B}$. As the unbroken string of $q_i = 0$ grows, the marketeer using $\mathcal{C}$ picks up on the crash regularity, and in subsequent days gives higher probability to the prediction that next days will show the result $q = 0$ as well. Further, note that the exchangeability of the observations within the data e^*_{35} is indeed violated with the use of the alternative partition $\mathcal{C}$. This is because the probability

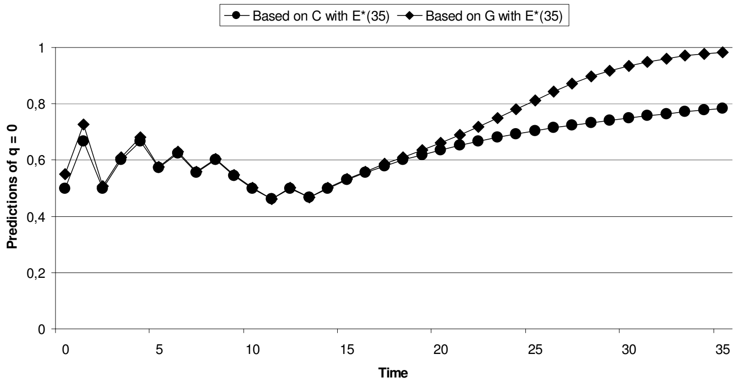


Figure 3.3: Predictions $p_{[e_t]}(Q^0_{t+1})$ against time t for the crash data e^*_{35}. The bullet curve is based on the partition $\mathcal{B}$, the diamond curve is based on the partition $\mathcal{C}$.

assignments depend directly on whether the data e_t show an unbroken string of 0's up until t. The partition $\mathcal{C}$ thus introduces a sensitivity for the occurrence of a crash pattern in the data, in addition to the usual attention that is given to the relative frequencies of the results.

It must be stressed that using the partition $\mathcal{C}$ in no way violates the Bayesian scheme. The probabilities $p_{[e_t]}(G^0_{\gamma\lambda\tau})$ are updated by conditioning just as well. They are turned to zero every time $i > \tau$ at which $q_i = 1$, or immediately if $i = \tau$ and $q_i = 0$. Further, it is not problematic to assign nonzero priors to the hypotheses in $\mathcal{C}$, even while these hypotheses had negligible or zero probability in the original partition $\mathcal{B}$. Assigning nonzero probabilities to hypotheses on specific patterns has been proposed before, for example by Jeffreys (1939) and also in the aforementioned Hintikka systems (1966). Howson (1973) provides an overview of arguments against the claim that it is inconsistent or wrong to assign strictly positive priors to generalisations.

3.4 The use of partitions

This section discusses the use of partitions in the Bayesian scheme. After some discussing some immediate insights, I develop the idea that partitions function as inductive assumptions, or projectability assumptions, in the inductive arguments. In the last subsection I discuss how this presents an advantage for the Bayesian scheme.

3.4.1 Immediate insights

Pattern recognition. Several insights may be drawn from the example that uses partition $\mathcal{C}$. Firstly, the above example shows that inductive predictions based on hypotheses can be adapted to model pattern recognition, and in this particular case, hasty generalisation. This can be done by adding hypotheses that pertain to the relevant kind of pattern. Following Putnam's critical remarks on the Carnap-Hintikka tradition in (1963a) and (1963b), this is already a useful extension of that tradition. Moreover, and as also discussed above, the modelling of hasty generalisation may convince those who consider updating on generalisations impossible due to the negligible measure of these generalisations in the observation field.

Secondly, and related to this, the example may be taken to qualify the fact that Bayesian updating is not suitable for modelling ampliative reasoning, as is argued by van Fraassen (1989). It is true that Bayesian updating cannot capture reasoning that decides between hypotheses with the same observational content, which therefore have the same likelihoods in the Bayesian schemes. But the above reasoning can nevertheless be called ampliative on the level of predictions: hasty generalisation is a typically ampliative inferential move. Thus, even though Bayesian updating is itself not ampliative, the predictions resulting from a Bayesian update can in a sense model ampliative reasoning. Note that the ampliativeness is implicit in the choice of the partition $\mathcal{C}$, and not in the inference rule of the inductive scheme.

The partition as a pair of glasses. In choosing a different partition, I implicitly alter the resulting prediction rule: the straight rule, generated by the partition $\mathcal{B}$ with uniform prior, is replaced with some other prediction rule $pr_{\alpha\delta}(q, e_t)$. Put differently, the probability assignment $p_{[e_0]}$ over the field $\mathcal{Q}$, initially determined by the partition $\mathcal{B}$ and some prior probability assignment over it, now encodes a different prediction rule, determined by the partition $\mathcal{C}$ and its prior. The probability over the added hypotheses of $\mathcal{C}$ depends on a crash pattern in the data, and the resulting predictions will therefore not be exchangeable. Thus we have defined a different prediction rule by choosing a different partition in the Bayesian scheme.

In the examples, the influence of the observations is really determined by the partition. We first choose a partition and, define a prior probability assignment over it, and via the observations determine a posterior probability assignment. The predictions can then be derived from this posterior probability and the likelihoods, which are given with the choice of partition. So the posterior probability over the partition is the only term in the predictions which depends on the observations. As Niiniluoto (1976) puts it, a partition defines a closed question, which has a limited set of possible answers, for the observations to decide over. So partitions do not provide an impartial or completely general view on the observations. Rather they are a pair of glasses for looking at the observations in a particular way.

3.4.2 Partitions as inductive assumptions

In this subsection, the function of choosing a partition is subject to further scrutiny. I shall characterise how partitions limit the view of an observer on observations, and how this connects to inductive assumptions.

Sufficient statistics. Consider the Bernoulli partition $\mathcal{B}$. The posterior probability over this partition can be computed from the prior and the observations. However, we do not need to know all the details of the observations for this computation. In fact, it suffices to know specific characteristics of the observations: for all q we must know the number of times that it occurred within the data e_t. These numbers were denoted by t_q in the above. They are the so-called sufficient statistics for computing the probability over $\mathcal{B}$ at time t, and thus for generating the predictions based on $\mathcal{B}$. The statistics t_q express those characteristics of the observations which are taken to be relevant for the predictions. Note that the exchangeability of the predictions based on $\mathcal{B}$ follows from the fact that the sufficient statistics are independent of the order of the observations.

The partition with crash hypotheses $\mathcal{C}$ limits the view on the observations in a different way. As with the Bernoulli partition, we can identify a set of sufficient statistics for it. This set includes not just the numbers t_q, but also the length of the time interval $[\tau, t]$ within which all results are 0. The numbers t_q and the number $t - \tau$ are employed together in a full determination of the probability over $\mathcal{C}$ at time t, and therefore in the generation of the predictions based on $\mathcal{C}$. It is notable that, because the value of $t - \tau$ depends on the order of the observations e_t, the resulting predictions are not exchangeable.

The above exposition shows how partitions limit the view on observations: partitions determine a set of sufficient statistics, and these statistics represent the characteristics of the observations which are taken to be relevant for further predictions. Put differently, by choosing a partition we focus on a particular set of patterns in the data, and by making predictions based on the partition we deem these patterns relevant to future observations. However, from the above discussion it is not clear what the exact function of this limitation is, or more specifically, what the nature of this relevance is. As Skyrms suggests in (1996), the answer is that sufficient statistics determine the so-called projectable characteristics of data. The function of partitions then is that they determine the projectable characteristics of the observations. They are a tool in controlling the projectability assumptions that are used in inductive predictions.

Partitions as projectable patterns. Now let me explicate in general terms how the use of a partition relates to the assumption of a projectable pattern in the observations. Recall that the hypotheses in a partition are all associated with a likelihood function. These likelihood functions may be in accordance with the actual observations to differing degrees: hypotheses that have high overall likelihoods given the observations are said to fit the data better than those with low overall average likelihoods. An update over a partition can thus be viewed as a competitive struggle among the hypotheses in the partition, in which hypotheses that fit the observations best acquire most probability. Note further that the likelihood functions associated with the hypotheses describe probabilistic patterns in the observations. An update over a partition is thus also a competition between probabilistic patterns in the observations. Choosing a particular partition thus limits the range of possible patterns that are allowed to compete in the update.

Furthermore, if we go on to employ the results of such a competition for the generation of predictions, we implicitly assume that those probabilistic patterns that fitted the observations better in the past are more likely to perform better in the future as well. This is because predictions of future observations are mainly based on the hypotheses which, relative to the chosen partition, were most successful in predicting the past observations: those hypotheses gain more probability in the update. This is exactly where the assumption on the uniformity of nature, with respect to a specific set of probabilistic patterns, is introduced into the Bayesian scheme.

These considerations show in what way the partitions are assumptions on the projectability of patterns in the observations: a partition determines a collection of probabilistic patterns, all of them patterns which may be employed for successful predictions, or projectable patterns for short. A prior probability over the hypotheses expresses how much these respective patterns are at the onset trusted with the predictive task, but the observations eventually determine which patterns perform this task best on the actual data. The predictions are subsequently derived by means of a weighing factor, the probability assignment over the partition, which favours the patterns that perform best. However, it must be stressed that the projectability assumption concerns not just these best performing patterns, but the partition as a whole, because the patterns perform better or worse only relative to a collection of patterns. The projectability assumptions are therefore implicit in the common features of the hypotheses involved. Limiting the collection of patterns to a collection with some general feature amounts to the assumption that the observations themselves exhibit this general feature, and that this general feature can therefore be projected onto future observations.

Finally, let me illustrate the projectability assumptions as general characteristics of the partitions, and link them with the sufficient statistics alluded to above. Recall once again the examples of section 3.3. Choosing the Bernoulli partition $\mathcal{B}$ means that we limit the possible probabilistic patterns to those for which the observations occur with specific relative frequencies. The projectability assumption is therefore exactly that this characteristic of the observations, namely the relative frequencies, are in fact exhibited in the observations. This is quite naturally related to the sufficient statistics for this partition, which are the observed relative frequencies t_q. Similarly, choosing to include hypotheses on crashes means that we include this particular set of crash patterns in the set of possible patterns. The projectability assumption is therefore exactly that this characteristic of a crash may be exhibited in the observations too. This additional focus of the partition is reflected in the additional statistic $t - \tau$.

3.4.3 Advantages of the Bayesian scheme

The main conclusion of the foregoing is that choosing a partition functions as a projectability assumption, by focusing on a set of sufficient statistics and by specifying how these statistics are used in the predictions. In the remainder of this section, I shall draw two further conclusions which derive from this main one.

Access to projectability assumptions. Within the inductive schemes presented in this thesis, any inductive argument must be based on some kind of projectability assumption. This can be concluded from the abundant literature on the Humean problem of induction, and the further literature on projectability, as collected in, for instance, Stalker (1996). So an inductive argument is a method that is sensitive to past observations. Prediction rules that completely ignore data and predict the same irrespectively of these data are not inductive. But in that case, any inductive argument must assume that past observations are somehow indicative of future observations, which comes down to a projectability assumption. Inductive prediction rules in a Carnapian scheme therefore employ projectability assumptions just as well as the Bayesian scheme does. For Carnap himself, the projectability assumptions are part and parcel of the choice of language. But the fact that the language provides a basis for the projectability assumptions in the Carnapian scheme must not distract from the fact that there are such assumptions in the first place.

I can now make explicit the reasons for adhering to the Bayesian schemes as opposed to Carnapian prediction rules. Recall that any update over the Bernoulli partition $\mathcal{B}$ results in exchangeable predictions. Further, the use of a Dirichlet density as prior probability assignment over this partition results in predictions that are identical to those produced by the $\lambda\gamma$ rule. As indicated, these results have been interpreted as a reason to refrain from using underlying chance processes or hypotheses, and to use the simpler prediction rules instead. However, the foregoing claims that there are good reasons for adhering to the complicated Bayesian schemes after all: these schemes provide direct insight into the projectability assumptions, as represented in the statistical hypotheses. Moreover, even though Hintikka systems did employ universal hypotheses in the construction of inductive prediction rules, we saw that these systems did not make full use of the possibilities that hypotheses offer. In short, the Bayesian scheme has an advantage over the Carnapian scheme because it provides immediate access to the projectability assumption.

Control over projectability assumptions. The advantage of Bayesian schemes is not just that they provide insight into the projectability assumptions. It may be argued that the $\lambda\gamma$ rules, for example, provide this insight just as well, because these prediction rules depend on the data e_t only via the sufficient statistics t_q. The further advantage, which perhaps discriminates more clearly between the Bayesian and the Carnapian scheme, is that Bayesian schemes provide better control of the projectability assumptions.

Let me illustrate the control over inductive assumptions with the crash example. Imagine that we already model a focus on relative frequencies, using a $\lambda\gamma$ rule, and that we want to model an additional focus on a crash pattern in the observations. Then we must somehow incorporate the statistic $t-\tau$ into the $\lambda\gamma$ rule we are using. But it is unclear how exactly to incorporate it, because we do not have insight in the projectability assumptions implicit to the form of the computation that we choose. The same problem appears for Hintikka systems, as there is no room for hypotheses on specific patterns, other than universal hypotheses. In sharp contrast with this, modelling an additional focus on a crash pattern with Bayesian schemes is straightforward: just add the hypotheses that pertain to the patterns of interest to the partition. Therefore, the Bayesian scheme may be more complicated, but in return it offers a better control of the projectability assumptions which are implicit in the predictions.

In view of the preceding chapter, it is not surprising that the Bayesian scheme improves the access to and control over the projectability assumptions. The Bayesian scheme employs an extended observational algebra, or an extended observation language, and it is only natural that this extended language offers us more expressive power.

Freedom in choosing projectable patterns. A final remark concerns the freedom in choosing partitions. Note that the choice of a partition is entirely under the control of the inductive reasoner. The only possible restriction lies in the fact that we may decide to employ frequentist hypotheses only, but this is not mandatory. Bayesian inductive logic itself provides no directions or restrictions as to what hypotheses to choose. Just as we can choose a partition which focuses on relative frequencies and crash patterns, we can choose a partition that expresses the gambler's fallacy, so that with the piling up of 0's in the crash the observation of 1 is predicted with growing confidence. The Bayesian schemes are in this sense a very general tool: any inductive prediction rule, as long as it is based on the assumption of some projectable pattern, can be captured in predictions generated with a Bayesian scheme. This shows that Bayesianism is not a particular position on inductive predictions, but rather an impartial tool for modelling predictions.

3.5 Conclusion

Summary. Sections 3.1 and 3.2 introduced inductive predictions and the tradition of Carnap-Hintikka inductive logic. The examples of section 3.3 illustrated how partitions determine the resulting predictions. In section 3.4 I argued that a partition expresses inductive assumptions concerning the projectability of particular characteristics of the observations. Partitions came out as a useful tool in defining the predictions. The main conclusion of this chapter is therefore that inductive predictions can be determined by choosing a partition in a Bayesian scheme, and that a partition expresses inductive assumptions on the projectability of particular characteristics of observations.

Further conclusions were seen to follow from this main one. One specific conclusion concerned the range of prediction rules covered by Bayesian schemes.

The example shows that the schemes enable us to model predictions typical for hasty generalisation. Now if we adopt the view of chapter 2, hypotheses must be chosen from the frequentist class, and it is not clear that just any prediction rule can be formulated in a Bayesian scheme. Note, however, that this restriction is not inherent to Bayesian logic, but rather to the frequentist add-on. Another specific conclusion was that the Bayesian scheme offers better insight in, and control over, inductive predictions than the prediction rules from the Carnap-Hintikka tradition. This tradition has focused primarily on the properties of prediction rules. It has not fully exploited the use of general hypotheses. The present chapter argues that in the construction of prediction rules, there are good reasons for employing these hypotheses after all.

Inductive logic. The general tendency in all this is in line with the main point of chapter 1. It is to view inductive logic as a proper logic: any prediction must be based on inductive assumptions, or premisses, and given these assumptions, the predictions must follow from the observations by probability axioms and Bayesian updating, which function as inference rules. So the work of induction is not done by an inference rule that implicitly contains uniformity assumptions, but by partitioning the space of possible worlds, fixing the likelihoods on the basis of that, and then choosing prior probabilities. As further discussed in chapter 7, this view on inductive predictions has consequences for the way we deal with the central problem of this thesis, the problem of induction. But the logical picture itself does not suggest any solution to the problem: the choice of a partition is not informed by the logical scheme.

Let me elaborate this latter point a bit. One of the aims specific for the Carnapian tradition is to provide a predictive scheme based solely on symmetries in the prediction rules, such as exchangeability. These symmetries are given independent justification in the notion of logical probability: the gap between past and future observations is bridged with logical means. The schemes considered here do not aim for such a justification. For the purpose of this thesis, it is enough to provide a scheme in which inductive assumptions can be expressed clearly, and in which the arguments from assumptions and observations to probabilistic predictions are valid. Certainly, the quest for plausible assumptions is an important and interesting task, but I take this task to fall outside of the logical analysis of inductive inference, and more naturally situated in epistemology. For some brief considerations on this, I refer to the conclusion of this thesis.

Analogy and independence. While the motivation of certain inductive assumptions is not included in the task of inductive logic, it may be considered part of its task to provide the translation of specific pre-formal considerations into formal premises. As an example, if there is reason to make inductive assumptions based on simplicity, there is still the task of making this simplicity formally precise. This may be done with a Bayesian or Akaike information criterion, as discussed in Akaike (1978), Sober (1998), Kieseppä (1997) and Bandyopadhyay and Boik (1999), or by means of minimal description length, as in Rissanen (1982). It is to the task of making certain considerations formally precise that I turn in the second part. It concerns translations of specific extra-logical considerations or insights into a form suitable for Bayesian inductive logic.

In particular, the second part shows that partitions may be employed to design prediction rules that incorporate analogy effects and independence assumptions. This also illustrates that the possibilities of the Bayesian scheme have not been employed fully in defining such predictions. Analogical prediction rules from the Carnap-Hintikka tradition may be combined using a Bayesian scheme with a partition that differs from $\mathcal{B}$, and interesting variations on these rules can be constructed by suitable transformations between partitions of hypotheses. Chapters 4 and 5 are concerned with these analogical predictions. Chapter 6 discusses inductive inference for Bayesian networks by means of the Bayesian scheme. As it turns out, the formal framework for these networks is exactly the same as the framework for analogy reasoning.

Philosophy of science. The third part concerns the relation between inductive assumptions in the Bayesian scheme and some main themes in the philosophy of science. It discusses the use of suppositions of underlying structure in chapters 7 and 9, and the control over changes in the assumptions within the Bayesian scheme in chapter 8. This latter research follows up on the debate over conceptual enrichment in Niiniluoto and Tuomela (1973), and also Gillies (2001), who argues that changes in the conceptual framework are a problem for the Bayesian theory. Recall that choosing a partition fixes the basic concepts that are used in the update. For example, with the hypotheses on crashes we include the phenomenon of a crash in the conceptual framework of the marketeer. We can therefore model a change in the focus on a projectable pattern by changing the partition.

II
Analogy and Independence

4

Analogical Predictions for Explicit Similarity

The above chapters have been concerned with the nature of the Bayesian scheme. The next three chapters are concerned with applications of it. The advantages of the Bayesian scheme, which are argued for in chapter 3, can now be illustrated. Apart from illustrating these advantages, the chapters also elaborate on them. It is shown that Bayesian schemes not only offer a better access to projectability assumptions, but that they also provide better control over other inductive assumptions inherent to the prior probability assignment, namely assumptions of analogy and independence. The chapters thus provide a new perspective on a well-known theme in Carnapian logic.

This chapter in particular concerns exchangeable analogical predictions based on similarity relations between predicates, and deals with a restricted class of such relations. To connect to the dominant Carnapian tradition, it first describes a system of Carnapian $\lambda\gamma$ rules on underlying predicate families to model the analogical predictions for this restricted class. But instead of the usual axiomatic definition of these rules, the system is here characterised with a Bayesian model that employs certain statistical hypotheses. Finally the paper argues that the Bayesian model can be generalised to cover cases outside the restricted class of similarity relations.

The present chapter can be read independently of the preceding ones. To see how this chapter connects to the main line of this thesis, the reader may consult the introduction of this thesis and chapter 1, in particular section 1.1.

4.1 Analogy within Carnapian rules

Analogy at the bowling alley. Imagine that the marketing director of a bowling alley is interested in the demographic composition of the crowds visiting her alley. Every evening she records the gender of some of the visitors, and whether they are married or not. Now let us say that on one evening half of the recorded visitors are male and married, and the other half are female and unmarried. Then if a newly arrived visitor is a man, the director may consider it more likely that he will be married than if this visitor were a woman. At least some of the similarity between individuals at the bowling alley is thus explicit in the observation language, namely in their gender, and in making predictions on their marital status this similarity may be employed. In such a case we speak of the inductive relevance of explicit similarity relations. Such relations are possible because individuals are categorised with predicates from multiple predicate families.

Putting G for gender and M for marital status, the analogical prediction in the example of explicit similarity may be represented in the following way:

$$\begin{array}{ccc}
G^0_1 & \cap & M^1_1 \\
G^1_2 & \cap & M^0_2 \\
 & \vdots & \\
G^0_{n-1} & \cap & M^1_{n-1} \\
G^1_n & \cap & M^0_n \\
G^0_{n+1} & & \\
\hline
\text{probably} & & M^1_{n+1}.
\end{array}$$

Here G^g_i with $g = 0$ or $g = 1$ is the record that individual i is male or female respectively, and M^m_i with $m = 0$ or $m = 1$ that this individual is not married or married respectively. The similarity between the individuals with odd index and the further individual $n+1$ is that all of them satisfy the predicate G^0 of the family G, meaning that they are all male. This similarity is used to derive, from the fact that the odd indexed individuals satisfy the predicate M^1 of the family M, meaning that they are all married, that probably also the individual $n + 1$ will be married. So the similarity of gender is made explicit in the observation language, and employed for predicting the marital status.

Carnapian continuum. Instead of the two predicate families above, we may imagine that the marketing director categorises the individuals at the bowling alley according to the division of bachelor, husband, maiden and wife, denoted with the family of predicates Q^q for $q = 0, 1, 2, 3$ respectively. This family is linked to the families G and M according to

$$Q^{2g+m} = G^g \cap M^m. \tag{4.1}$$

As illustrated in figure 4.1, $Q^0 = G^0 \cap M^0$ represents bachelors, $Q^1 = G^0 \cap M^1$ represents husbands, $Q^2 = G^1 \cap M^0$ maidens, and $Q^3 = G^1 \cap M^1$ wives. Using the single predicate family, the director may derive predictions on gender and marital status from the $\lambda\gamma$ rules of Stegmüller (1973):

$$p(Q^q_{n+1}|E_n) = \frac{n_q + \lambda\gamma_q}{n + \lambda}. \tag{4.2}$$

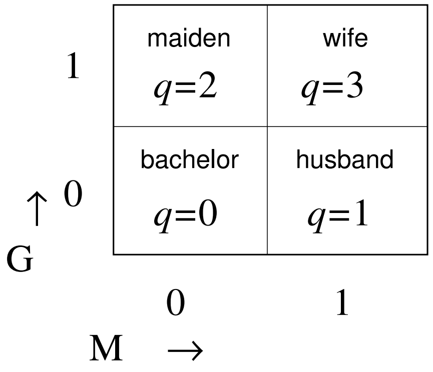


Figure 4.1: The relation between the predicate family Q and the underlying families G and M.

Here the expression E_n represents the records of Q-predicates for the first n subjects, and n_q is the number of records of category q within E_n. The parameters γ_q determine the initial expectations over the family Q, and the parameter λ determines the speed with which we change these initial expectations into the recorded relative frequencies of the predicates Q^q. With an assumption of initial symmetry we can fix $\gamma_q = {}^1\!/_4$ for all q.

With these $\lambda\gamma$ rules concerning Q-predicates we can also derive predictions on the underlying predicate families G and M, using the inverse identifications

$$G^g = Q^{2g} \cup Q^{2g+1}, \tag{4.3}$$

$$M^m = Q^m \cup Q^{2+m}. \tag{4.4}$$

With this we can derive the following expressions for predictions on marital status:

$$p(M^1_{n+1}|E_n) \;=\; \frac{(n_1+n_3)+\lambda(\gamma_1+\gamma_3)}{n+\lambda}, \tag{4.5}$$

$$p(M^1_{n+1}|E_n \cap G^0_{n+1}) \;=\; \frac{n_1+\lambda\gamma_1}{(n_0+n_1)+\lambda(\gamma_0+\gamma_1)}. \tag{4.6}$$

The prediction rules thus derived have the same format as the above $\lambda\gamma$ rules. Note that the indices of n refer to the Q-predicates.

On the evening of the example, there are, up to a certain moment, an even number n of visitors at the bowling alley, of which half are husbands and half are maidens:

$$E_n = \bigcap_{i=1}^{n/2} (Q^1_{2i-1} \cap Q^2_{2i}). \tag{4.7}$$

We therefore have $n_1 = n_2 = {}^n\!/2$ and $n_0 = n_3 = 0$. Then visitor $n+1$ parks a car, and upon entering it turns out to be a man. As already suggested in Carnap and Stegmüller (1959: 242-250), the symmetric $\lambda\gamma$ rule on family Q predicts a higher probability for this individual being married after incorporating that the visitor is a man than if the gender is unknown:

$$p(M^1_{n+1}|E_n \cap G^0_{n+1}) \;=\; \frac{n/2+\lambda/4}{n/2+\lambda/2} \;>\; \frac{n/2+\lambda/2}{n+\lambda} \;=\; p(M^1_{n+1}|E_n). \qquad (4.8)$$

The $\lambda\gamma$ rule thus shows analogy effects of explicit similarity, in the sense that the similarity of visitor $n+1$ to the present visitors with respect to the family G, the gender, affects the predictions with respect to family M, the marital status.

4.2 Analogical predictions

Similarity relations. The above analogy effects are captured in the $\lambda\gamma$ rules, but many more such effects cannot be captured. It may be the case that husbands, Q^1, and bachelors, Q^0, regularly visit the bowling alley together, and that on a particular evening the director has only recorded husbands. Then, apart from the fact that this may make further instances of husbands more probable, we may find an instance of a bachelor more probable than an instance of a maiden or a wife, Q^2 or Q^3, because husbands are more likely to hang out in the bowling alley with their bachelor friends. That is, we consider the presence of husbands more relevant to bachelors than to maidens or wives.

It is easily seen that the $\lambda\gamma$ prediction rules cannot accommodate such differences in relevance among the Q-predicates. For any instance of Q^q, the ratios between the predictions of any two other predicates Q^v and Q^w will not change, because this ratio is given by

$$\frac{p(Q^w_{n+1}|E_n)}{p(Q^v_{n+1}|E_n)} = \frac{n_w+\gamma_w\lambda}{n_v+\gamma_v\lambda}, \qquad (4.9)$$

which is independent of n_q. Therefore analogy effects that hinge on differences in inductive relevance between Q-predicates fall outside the scope of $\lambda\gamma$ rules.

The predictive relevance between Q^v and Q^w may be expressed in terms of an inductive relevance function $\rho(v,w)$. A general expression of analogy by similarity, using the relevance function, is:

$$\rho(q,w) > \rho(q,v) \quad \Rightarrow \quad \frac{p(Q^w_{n+1}|E_{n-1}\cap Q^q_n)}{p(Q^w_n|E_{n-1})} > \frac{p(Q^v_{n+1}|E_{n-1}\cap Q^q_n)}{p(Q^v_n|E_{n-1})}. \qquad (4.10)$$

It must be stressed that this is certainly not the only expression of analogy by similarity, and in particular that the focus differs from that of Carnap (1980:

46-47). The characterisation offered here is qualitatively equivalent to Kuipers' characterisation in (1984), which is associated with $K_{>G}$ inductive methods in the categorisation of Festa (1997: 232-235). The focus is therefore not on Carnap's and Maher's kind of similarity, which concerns differences between $\rho(v, q)$ and $\rho(w, q)$. On the other hand, I assume in this paper that the function is symmetric:

$$\rho(v, w) = \rho(w, v). \tag{4.11}$$

Because of this the above expression of relevance is very much akin to that of Carnap and Maher. Note finally that some authors employ a distance function instead of relevances. Strictly speaking this is inadequate, since the relevances need not comply to triangular inequalities.

Analogy models in the literature. Many models have been proposed in order to capture analogical predictions based on similarity. The main focus of these models is on an alternative prediction rule concerning Q-predicates that somehow incorporates the relevances. Some of these prediction rules are exchangeable, that is, invariant under permutations of the given Q-predicates, and some are non-exchangeable. Examples of such models are given in Kuipers (1984, 1988), Skyrms (1993), Di Maio (1995) and Festa (1997). However, to my mind analogical predictions are more easily associated with similarity in terms of underlying predicates, here called explicit similarity, than with similarity between predicates directly. Moreover, as it turns out, the use of underlying predicate families is very useful in defining analogical predictions. For these reasons I shall, in what follows, employ the underlying predicate families G and M in the construction of the analogical prediction rules for Q.

The models of Carnap, Maher and Niiniluoto do employ underlying predicates. Specifically, Niiniluoto (1981, 1988) uses the structure of underlying predicates to explicate the strengths of the similarity between the Q-predicates. As an example, husbands and bachelors are more similar than husbands and maidens, because the first two have their gender in common, where the second two do not share any underlying predicate. To the extent that this explication of similarity between Q-predicates is adopted in other models, we can say that these other models employ the underlying predicates as well. However, in all these models the relation between the similarities and the prediction rules is rather ad hoc. The predictions of Q are influenced by the similarities, but the explication of the similarity in terms of underlying predicates is itself not used in the construction of the prediction rules. The rules are defined by assigning probabilities to the Q-predicates alone. Probabilities over M and G may be derived from that, but no use is made of the possibility to assign probabilities over the families M and G.

The model of Maher (2000), which is basically an improved version of the model of Carnap and Stegmüller (1959: 251-252), makes more elaborate use of underlying predicates. Maher supposes two predicate families, such as G and M, to underly the Q-predicate over which the predictions are defined. He then formulates a hypothesis on the statistical independence of predicates G and M, and translates this hypothesis into one about Q-predicates. Conditional on this independence hypothesis, predictions of Q-predicates can be written in terms of a product of $\lambda\gamma$ rules for the predicates G and M separately. Conditional on the dependence hypothesis, on the other hand, the predictions for Q-predicates are given by a single $\lambda\gamma$ rule. The eventual prediction rule for the Q-predicates is a mixture of these two predictions, weighed with the probabilities of the dependence and independence hypotheses. Analogical considerations are then captured because these weights are themselves influenced by the observations of the Q-predicates, which implicitly convey information on the statistical dependence of the predicates G and M.

The model of this chapter. The model of the present chapter takes the use of underlying predicates a bit further. Maher provides a model in which statistical relations between underlying families like G and M are used to derive predictions over Q-predicates that capture analogy considerations. But when it comes to statistical relations between the underlying families, Maher's model considers a partition into complete independence and undifferentiated dependence, and employs a single $\lambda\gamma$ rule for Q-predicates in the latter case. By contrast, the present model employs predictions on underlying predicates in the case of statistical dependence as well. Specifically, the model employs predictions for an individual concerning the family M, conditional on the fact that this individual satisfies a particular predicate from the family G. Moreover, and perhaps more importantly, the present model elucidates the exact relation between inductive relevances ρ and the statistical dependencies between underlying predicates. Specifically, the relevance relations ρ, which are assumed at the start of the update, are related to the parameters of the model. On this point the model differs from Maher's (2000) model and the other models discussed in Maher (2001), which are not related to assumed relevance relations ρ.

The model of this chapter is restricted in a certain way. It provides analogical predictions that cannot be captured by the single $\lambda\gamma$ rule, but it considers only a limited set of relevance relations. As an example, consider the husbands and bachelors who like to go bowling together, and prefer not to have female company. In terms of relevance functions,

$$\rho(0,1) = \rho(1,0) > \rho(0,2) = \rho(1,2) = \rho(0,3) = \rho(1,3). \tag{4.12}$$

That is, the relevances of husbands and bachelors to each other are equal, and larger than relevances between individuals of different gender. Let us further say if wives visit the bowling alley, they are likely to bring their husbands, who then also invite their bachelor friends, where I am for the sake of simplicity supposing that there are no gay marriages. The wives typically do not invite their maiden friends. Similarly, if maidens visit the bowling alley, they are likely to be together with the bachelors, who in turn bring along some husbands, but the maidens do not usually invite any wives. That is,

$$\rho(2,3) = \rho(3,2) < \rho(2,0) = \rho(2,1) = \rho(3,0) = \rho(3,1), \tag{4.13}$$

or in words, the relevances of wives and maidens to each other are equal, and smaller than relevances between individuals of different gender. Note that due to the symmetry of the relevance function, the four equal relevances in expressions (4.12) and (4.13) are the same.

As said, this example is one in a set of similar cases. The common element is that the relevances between categories with different gender are all equal, and that the relevances between categories within the genders may vary. Defining

$$\begin{aligned} \forall m, m' \in \{0,1\}: \qquad \rho_{\bar{G}} &= \rho(m, 2+m'), && (4.14)\\ \forall g \in \{0,1\}: \qquad \rho_{Gg} &= \rho(2g, 2g+1), && (4.15) \end{aligned}$$

the similarity relations are in effect characterised by three relevances, $\rho_{\bar{G}}$, ρ_{G0} and ρ_{G1}, representing the relevances between individuals of different gender, the relevance between bachelors and husbands, and the relevance between maidens and wives respectively. These relevances may have any ordering in size. The subclass of cases thus defined, for which the relevance relations between categories of different gender do not differ, are exactly the cases of analogy by similarity that can be made explicit in terms of gender. In the following I therefore refer to them as cases of explicit similarity.

Summing up, the aim of this chapter is to provide a prediction rule for analogical predictions for explicit similarity, to connect the relevance relations $\rho_{\bar{G}}$, ρ_{G0} and ρ_{G1} to parameters in this prediction rule, and finally to give a proper statistical underpinning for it. The next chapter shows how the model can be used to define exchangeable analogical predictions based on symmetric relevance relations in general.

4.3 A model for explicit similarity

This section presents a system of $\lambda\gamma$ rules that models the intended analogical predictions. It is shown that the system generalises the analogy effects that are captured in single $\lambda\gamma$ rules. The function of the parameters in the system is explained, and a numerical example is provided.

A system of rules. The system of prediction rules offers separate entries for instances of the family M for individuals satisfying either of the two predicates of the family G. This is expressed in the following:

$$\begin{aligned} p(G^g_{n+1}|E_n) &= \frac{n_{Gg} + \lambda_G \gamma_{Gg}}{n_G + \lambda_G}, & (4.16) \\ p(M^m_{n+1}|E_n \cap G^g_{n+1}) &= \frac{n^g_{Mm} + \lambda^g_M \gamma^g_{Mm}}{n^g_M + \lambda^g_M}. & (4.17) \end{aligned}$$

The indexed numbers n can all be derived from E_n using the translations (4.3) and (4.4). In particular, we have the total number of records on gender $n_G = n$, the number of records of males and females, $n_{Gg} = n_{2g} + n_{2g+1}$ for $g = 0, 1$, which is the same as the number of records on marital status given a certain gender, $n^g_M = n_{Gg}$, and finally the number of records for a specific gender and marital status, $n^g_{Mm} = n_{2g+m}$ for $g, m \in \{0, 1\}$.

The above system consists of three prediction rules, one that concerns individual $n + 1$ in the family G, and two that concern the family M, conditional on the individual satisfying G^0 and G^1 respectively. With these predictions we can construct a prediction rule for Q-predicates:

$$\begin{aligned} p(Q^q_{n+1}|E_n) &= p(G^g_{n+1}|E_n) \times p(M^m_{n+1}|E_n \cap G^g_{n+1}) \\ &= \frac{n_{Gg} + \lambda_G \gamma_{Gg}}{n_G + \lambda_G} \times \frac{n^g_{Mm} + \lambda^g_M \gamma^g_{Mm}}{n^g_M + \lambda^g_M}. & (4.18) \end{aligned}$$

As will be seen below, these predictions for Q-predicates can capture explicit similarity. Note first that the above system is a generalisation of the single $\lambda\gamma$ rule for Q-predicates. By writing the numbers n in terms of the n_q, by identifying

$$\begin{aligned} \lambda_G &= \lambda, & (4.19) \\ \gamma_{2g+m} &= \gamma_{Gg} \gamma^g_{Mm}, & (4.20) \end{aligned}$$

and finally by choosing

$$\lambda^g_M = \lambda_G \gamma_{Gg}, \qquad (4.21)$$

the system of rules generates the very same predictions that are generated by the single $\lambda\gamma$ rule of equation (4.2). Note also that on the level of Q-predicates the predictions are exchangeable, whatever the values of the parameters.

Encoding explicit similarity. Analogical predictions for explicit analogy can be obtained by choosing the values of the parameters λ^g_M different from those in equation (4.21). To explain this, let me first reformulate explicit similarity in terms of probabilities over the underlying predicate families G and M. First, the higher relevance between husbands and bachelors means that the effect of updating with the male gender of an individual must be larger than the effect of updating with the marital status conditional on the individual being male. The probability for further instances of males then strongly benefits from the instance of a male, while the marital status of the husband does not make bachelors much less likely. In similar fashion, the lower relevance between wives and maidens means that the effect of updating with the female gender of an individual is smaller than the effect of updating with the marital status conditional on the individual being female. On finding a maiden, for instance, the profit that the wives derive from the fact that the maiden is female is then overcompensated by the loss that stems from the fact that contrary to the wives, the maidens are not married. A similar change in the probability assignment is effected if we find a wife.

In the $\lambda\gamma$ rules of Carnap, the reluctance to adapt probabilities to new observations is reflected in the size of λ. In the above formulation, it is exactly differences in the reluctance to adapt probability assignments that leads to analogy effects. The above paragraph therefore suggests that we can connect the differences in relevance with specific differences between the values of the parameters λ_G and λ^g_M. As it turns out, we can identify a correspondence between parameter inequalities and inequalities of relevance functions. Normalising the size of the relevances for the number of Q-predicates N, so that in this case $N = 4$, these correspondences can be translated into rather simple relations:

$$\lambda_G = \rho_{\bar{G}} N, \tag{4.22}$$

$$\lambda^g_M = \rho_{Gg} \gamma_{Gg} N. \tag{4.23}$$

In updating with observations on the family G, we may be more or less prepared to adapt our expectations concerning gender, which is reflected in a low or high value for λ_G respectively. Similarly, conditional on the observation of G^g, we may be more or less prepared to adapt our expectations on an observation concerning M, which is reflected in the value of λ^g_M. With these variations in the willingness to adapt probabilities, we can model explicit analogy.

Numerical example. Let me make explicit the relation between the values of the λs and the relevances between predicates of equal gender for the specific case of husbands and bachelors. Recall that we have chosen $\rho_{\bar{G}} < \rho_{G0}$. In the model this relevance relation is supposed to be assumed at the start of the update. Now to encode this relevance relation in the system of prediction rules, we must according to the above equations choose $\gamma_{G0}\lambda_G < \lambda^0_M$. With these parameter values, the observation of a male strongly enhances the probability for further males, while the prediction for marital status conditional on males grows much slower with the observation of the males having a specific marital status. In the resulting predictions with respect to the predicates Q, this has the combined effect that observations of husbands and bachelors are mutually beneficial. This is because the observation of the common element of these predicates, their gender, affects the expectations much more than the observation that distinguishes the one from the other, their marital status.

Let us say that one night at the bowling alley the first three visitors are husbands, after which three maidens enter:

$$E_n = Q^1_1 \cap Q^1_2 \cap Q^1_3 \cap Q^2_4 \cap Q^2_5 \cap Q^2_6. \tag{4.24}$$

For the example of equations (4.12) and (4.13) we have $\rho_{G0} > \rho_{\bar{G}} > \rho_{G1}$. We may choose $\rho_{G0} = 4$, $\rho_{\bar{G}} = 1$ and $\rho_{G1} = 1/2$, and accordingly fix the following values for the parameters in the system of prediction rules:

$$\begin{aligned}
\lambda_G &= 4, \\
\lambda^0_M &= 8, \\
\lambda^1_M &= 1, \\
\gamma_{G0} &= \frac{1}{2}, \\
\gamma^0_{M0} &= \gamma^1_{M0} = \frac{1}{2}.
\end{aligned}$$

The predictions that can be generated with the above parameter values then show the analogical effects that can be expected on the basis of the corresponding values of the relevance function:

Number n	0	1	2	3	4	5	6
Observations q	-	1	1	1	2	2	2
$p(Q^0_{n+1}\|E_n)$	0.25	0.27	0.27	0.26	0.23	0.20	0.18
$p(Q^1_{n+1}\|E_n)$	0.25	0.33	0.40	0.45	0.40	0.35	0.32
$p(Q^2_{n+1}\|E_n)$	0.25	0.20	0.17	0.14	0.28	0.37	0.44
$p(Q^3_{n+1}\|E_n)$	0.25	0.20	0.17	0.14	0.09	0.07	0.06

As can be seen from these predictions, the husbands are positively relevant to the bachelors, while the maidens are negatively relevant to the wives. As is to be expected, this effect wears off as the number of records increases, but it only reaches zero at infinity.

Let me stress once again an important aspect of the model of analogical predictions defined in this section, namely that inductive relevances serve as explicit input to the prediction rules. In this sense the model is similar to the models of Niiniluoto (1981) and Kuipers (1984), while it differs from the models of Festa (1996) and Maher (2000). In these latter models, there is no direct access, in terms of input parameters, to the inductive relevances that may be assumed. However, whether this aspect of the accessibility of inductive relevances can be considered an advantage depends on the perspective we take on inductive logic.

4.4 Problematic aspects

This section discusses the fact that the system shows two distinct asymmetries in dealing with the families G and M, and motivates a difference in methodology that distinguishes this treatment from the Carnapian tradition.

Exchangeable analogical predictions. First I discuss whether the above system of rules preselects an order in the underlying predicate families. Note that in the above system of rules, we can only directly adapt the predictions for the marital status of individuals if we already know their gender. For example, if we know that only unmarried people drive sport scars, and we see a visitor arriving at the bowling alley in such a car before having determined her or his gender, it is not immediately clear how we must adapt the prediction rules. Accordingly, we cannot directly use the system to predict the gender of the visitors on the basis of their being unmarried.

All this is not to say that the system of prediction rules breaks down if the order of the observations is reversed. The system does assign a probability to the gender of a visitor conditional on this visitor having a certain marital status, and it also assigns a probability to the marital status of a visitor unconditionally. Both can be derived from the prediction rules (4.16) and (4.17). It is just that the calculations become rather intricate if we update on marital status first, because adapting the system of rules to records of marital status independently of gender is a messy operation. Furthermore, as it turns out the numerical values of the predictions do not change under permutations in the order of the underlying predicates. That is to say, the system is still exchangeable with respect to the underlying predicates. Unfortunately, an argument for that can only be given in section 4.6. For now the main thing is that the system of rules does not necessitate a specific order in the observations to obtain numerical values for the predictions.

Asymmetry in the expressible relevance relations. Leaving the matter of order and order dependence aside for the moment, it may be noted that there is yet another way in which the above systems treat the underlying predicates differently. When it comes to expressing relevance relations, there is an irreducible asymmetry in the predicates G and M: the systems are perfectly suitable for determining the relevance relation of some Q-predicate to the other Q-predicate with the same gender, ρ_{Gg}, relative to the relevance relation of this Q-predicate to the two Q-predicates of the opposite gender, $\rho_{\bar{G}}$. But, swapping the predicate classes of G and M themselves, the system is not at all suitable to determine the relevance relations ρ_{Mm} relative to the relations $\rho_{\bar{M}}$. In short, the system models explicit similarity of gender, and not of marital status. In setting up the system, we must choose which of these two complexes of relevance relations will be allowed implementation. Therefore, while the system of rules is exchangeable in the sense of invariant under permutations of the order of the observations, it is certainly not suitable for expressing analogy effects after swapping the predicates.

Now in some cases, a natural priority is suggested by the underlying predicate families themselves. One of the two is sometimes more directly observable than the other, or epistemically prior in some other way. For the purpose of adapting the system, and for modelling explicit similarity, such considerations may guide our choice. But there remains an inherent asymmetry in the expressible relevance relations, and in this sense the present model of analogical predictions is weaker than, for example, the model of Maher (2000). It is hoped that this disadvantage is compensated by the new perspective that the model offers, and the new possibilities that may result from that. Chapter 5 shows how the asymmetry between predicates can eventually be overcome.

A probabilistic underpinning. Let me turn to the second problematic aspect of the above system, which is that so far, it lacks an axiomatic underpinning. In Carnapian inductive logic the aim is to derive, from a chosen language or algebra and the notion of logical probability, a class of probability assignments over the algebra that describes all rationally permissible predictions. But the above system of rules has been introduced without any such derivation, and in this sense seems entirely ad hoc. It is not even clear whether the probability assignment over the algebra that is entailed by the above system is internally consistent. The remainder of this chapter is aimed at solving this problem. To provide a further underpinning of the system of prediction rules, and to prove its internal consistency, the next two sections specify a class of probability models that underlie the proposed systems of rules. These probability models are cast in the Bayesian schemes developed in chapter 1. In the remainder of this section, I discuss this Bayesian perspective, and its relation to more traditional Carnapian methods.

The probability models make use of hypotheses to define the prior probability assignment over the algebra, and they employ Bayesian updating to incorporate observations into this assignment. The system of prediction rules is thus not based on the algebra or language we have chosen, or on further principles or predictive properties we may assume. Instead, we define the inductive prediction rule by partitioning the algebra into a specific set of statistical hypotheses, and by stipulating a class of prior probability assignments over them. At the start we can choose a specific configuration of relevance relations, which may be encoded in a prior probability assignment. This signals an important methodological difference between the present chapter and most papers from the Carnapian tradition. The present chapter maintains that hypotheses and priors can be chosen freely, and that there are no restrictions implicit in the statistical framework. Relatedly, in this chapter there is no attempt to provide a rationalisation for the choice of hypotheses or the prior probability. The hypotheses and prior are taken to exhibit the inductive assumptions underlying the analogical predictions, much like premises in a deductive inference. Attempting to justify hypotheses and prior falls outside the reconstruction of analogical predictions as a statistically sound, or logically valid method.

Motivating the Bayesian perspective. Adopting this perspective on analogical predictions may look like a cheap escape from a challenging problem. Surely it is much harder to give a set of axioms that have an intuitive appeal or some independent justification, from which the exact class of all rationally permissible analogical predictions can be derived. While searching for these axioms and rationalizations is a worthwhile and venerable task, I side with the criticisms towards such axiomatic methods for analogical predictions, as can be found in Spohn (1981) and Niiniluoto (1988): it may be too ambitious to aim for the definitive class of all rational probability assignments that capture analogical considerations. It is more in line with an emphasis on local inductive practice, as recently discussed in Norton (2003), to propose a collection of models only, and to decide about the exact nature of analogical predictions on a case by case basis. This perspective resembles that of Bovens and Hartmann (2003), who advocate a kind of philosophical engineering as opposed to a quest for first principles.

There are some advantages to providing probability models that underlie the analogical predictions. First of all, the models connect research in analogical prediction rules with Bayesian statistical inference. I think it is important to bring these research traditions closer together. Second, as will be seen below, extending the models to more predicate families, or to more predicates within given families, is a straightforward operation in the probability model. It thus turns out that these models are very easy to generalise. Third, the models clarify that the system of prediction rules is really invariant under permutations of the order of the underlying observations. In other words, the statistical underpinning settles the issue of the exchangeability of the underlying observations.

Finally, and perhaps most importantly, the statistical models suggest a more general model of analogical predictions, which accommodates analogical predictions based on more general relevance relations than the ones considered above. We may for example consider bachelors more relevant to maidens than to wives, and it turns out that statistical models offer a rather natural place for relevance relations of this kind. Eventually the use of the statistical model leads to a model of analogical predictions based on a completely general relevance function. This theoretical development, however, will only be dealt with in the next chapter.

4.5 Statistical underpinning for $\lambda\gamma$ rules

This section discusses an observation algebra for Q-predicates, and the statistical underpinning of the $\lambda\gamma$ rule for these predicates. It presents the Bayesian scheme of the preceding chapters in a compressed form, leaving out some of its subtleties. The treatment prepares for the statistical underpinning of the system of $\lambda\gamma$ rules in the next section, which employs the basic partition of this section in threefold.

Observational algebra. Let me first introduce a representation of records of Q-predicates in terms of a so-called observational algebra. Let K be the set of possible values for q, and let K^{ω} be the space of all infinite sequences e of such values:

$$e = q_1 q_2 q_3 \ldots \tag{4.25}$$

The observation algebra, denoted $\mathcal{Q}$, consists of all possible subsets of the space K^{ω}. If we denote the i-th element in the sequences e and e_n with $e(i)$ and $e_n(i)$ respectively, we can define an observation Q_i^q as an element of the algebra $\mathcal{Q}$ as follows,

$$Q_i^q = \{e \in K^{\omega} :\ e(i) = q\}, \tag{4.26}$$

and a finite sequence of observations $E_n^{e_n}$ as follows,

$$E_n^{e_n} = \bigcap_{i=1}^{n} Q_i^{e_n(i)}. \tag{4.27}$$

Records of visitors at the bowling alley refer to such subsets. Note that there is a distinction between the observations, which are elements of the algebra $\mathcal{Q}$, and the values of the observations, which are natural numbers.

Statistical hypotheses can also be seen as elements of the algebra. If we say of a statistical hypothesis h that its truth is determined by a function $W_h(e)$ of an infinitely long sequence of observations e, writing $W_h(e) = 1$ if h is true for the sequence e and $W_h(e) = 0$ otherwise, then we can define hypotheses as subsets of K^{ω}:

$$H = \{e \in K^{\omega} :\ W_h(e) = 1\}. \tag{4.28}$$

A partition is a collection of hypotheses, $\mathcal{D} = \{H_\theta\}_{\theta\in D}$, defined by the following condition for the indicator functions W_{h_θ}:

$$\forall e \in K^{\omega}\ \exists! \theta : \quad W_{h_\theta}(e) = 1. \tag{4.29}$$

This means that the hypotheses H_θ are mutually exclusive and jointly exhaustive sets in K^{ω}, parameterized by a vector θ in an as yet unspecified space D.

Bernoulli hypotheses and exchangeable predictions. Probability assignments are defined for all the elements of the observational algebra $\mathcal{Q}$. The probability assignment can be adapted to a sequence of observations E_n by conditioning the original probability assignment p on these observations:

$$p(\cdot) \quad \rightarrow \quad p(\cdot|E_n). \tag{4.30}$$

Both the probabilities assigned to observations, and the probabilities assigned to hypotheses can be adapted to new observations in this way.

The schemes of this chapter employ observational hypotheses for generating the predictions $p(Q_{n+1}^q|E_n)$. To calculate the predictions, we may employ a partition of hypotheses and the law of total probability:

$$p(Q_{n+1}^q|E_n) = \int_D p(H_\theta|E_n)\, p(Q_{n+1}^q|H_\theta \cap E_n)\, d\theta. \tag{4.31}$$

The probability function over the hypotheses is a so-called posterior probability, $p(H_\theta|E_n)d\theta$. This probability is obtained by conditioning a prior probability $p(H_\theta)d\theta$ on the observations E_n. The terms $p(Q^q_{n+1}|H_\theta \cap E_n)$ are called the posterior likelihoods of the hypotheses H_θ given the observation Q^q_{n+1}. The prediction is obtained by weighing these posterior likelihoods with the posterior density over the hypotheses.

To characterise the partition that renders exchangeable predictions, define the relative frequency of the observation results $q \in K$ in a sequence e:

$$f_q(e) = \lim_{n \rightarrow \omega} \frac{1}{n} \sum_{i=1}^{n} W_q(e(i)), \tag{4.32}$$

in which $W_q(e(i)) = 1$ if $e(i) = q$, and $W_q(e(i)) = 0$ otherwise. Taking θ to be a real-valued vector, we can define W_{h_θ} as follows:

$$W_{h_\theta}(e) = \begin{cases} 1 & \text{if } \forall q \in K : f_q(e) = \theta_q, \\ 0 & \text{otherwise.} \end{cases} \tag{4.33}$$

The hypotheses H_θ form a so-called simplex, associated with a hypersurface $C = \{\theta \in [0,1]^{|K|} | \sum_q \theta_q = 1\}$ in a $|K|$-dimensional space. For $|K| = 4$, this hypersurface is a tetrahedron. We can further define $W_{h_{\neg\theta}}(e) = 1$ if $f_q(e)$ is undefined for any of the $q \in K$, and $W_{h_{\neg\theta}}(e) = 0$ otherwise. The collection of hypotheses $\mathcal{B} = \{H_{\neg\theta}, \{H_\theta\}_{\theta \in B}\}$ is a partition of hypotheses concerning the relative frequencies of $q \in K$.

We can now provide the likelihoods associated with the partition that renders exchangeable predictions. First we assume that $p(H_{\neg\theta}) = 0$, which states that the observations have some convergent relative frequency. The likelihoods of $H_{\neg\theta}$ may then be left unspecified. The likelihoods of H_θ may be defined by taking the long run relative frequencies θ_q as chances on predicates Q^q at every single observation:

$$\forall n \geq 0 : \quad p(Q^q_{n+1}|H_\theta \cap E_n) = \theta_q. \tag{4.34}$$

The hypotheses H_θ may be called Bernoulli hypotheses, as they describe so-called Bernoulli processes. The likelihoods of Bernoulli hypotheses do not depend on the observations E_n. The prior probability over the hypotheses H_θ can be chosen freely. According to De Finetti's representation theorem, there is a one-to-one mapping between exchangeable prediction rules and prior probability densities over partition $\mathcal{B}$ with these likelihoods.

Because the single $\lambda\gamma$ prediction rules are exchangeable, they can also be characterised by a specific class of densities over $\mathcal{B}$. This turns out to be the class of so-called Dirichlet densities:

$$p(H_\theta) \quad \sim \quad \prod_q \theta_q^{(\lambda\gamma_q - 1)}. \tag{4.35}$$

On assuming this prior, the resulting predictions are the $\lambda\gamma$ prediction rules with the corresponding parameter values. See Festa (1993: 57-71) for further details. So the $\lambda\gamma$ rules can be generated if we assume a partition of hypotheses $\mathcal{B}$ and its associated likelihoods $p(Q^q_{n+1}|H_\theta \cap E_n) = \theta_q$, and a prior probability density $p(H_\theta)$ from the Dirichlet class.

Finding the analogy prior. We can now reformulate the problem of capturing exchangeable predictions based on analogy by similarity of gender. We are effectively looking for a prior density over partition $\mathcal{B}$ that is not from the Dirichlet class, and that somehow incorporates analogical effects. Intuitively, this is a prior over the tetrahedron B that has an internal twist, as illustrated in figure 4.2. Within the triangular segments with hypotheses that have relatively high likelihoods for Q^1, we must allocate more prior probability to those hypotheses that also have relatively high likelihoods for Q^0, and therefore less probability to hypotheses that have high likelihoods for Q^2 and Q^3. Similarly, within the triangular segments with hypotheses that have relatively low likelihoods for Q^1, we must allocate less prior probability to those hypotheses that have relatively high likelihoods for Q^0, and therefore more probability to hypotheses that have high likelihoods for Q^2 and Q^3. With such a twisted prior density, we effectively favour the probability of Q^0 over those of Q^2 and Q^3 whenever we update with Q^1.

On the level of Q-predicates, the above system of $\lambda\gamma$ rules, as defined in equations (4.16) and (4.17), is exchangeable just as the single $\lambda\gamma$ rule. It can therefore be represented as a class of prior densities over $\mathcal{B}$. To find the probability models underlying the system of rules, we must thus find the exact class of prior densities over $\mathcal{B}$ with which these systems can be represented. However, this class of priors is very hard to define if we only have recourse to the parameter components in the space B. Even if we knew what function satisfies the features sketched above, it is not easy to formulate this prior in such a way that we can actually derive the system of rules. For this reason it is worthwhile to look for an alternative framework. The following proposes a transformation of the partition $\mathcal{B}$ into the partition $\mathcal{A}$. This latter partition comprises exactly the same hypotheses, but casts these in a different parameter space. Within that space we can derive the analogical predictions of section 4.3.

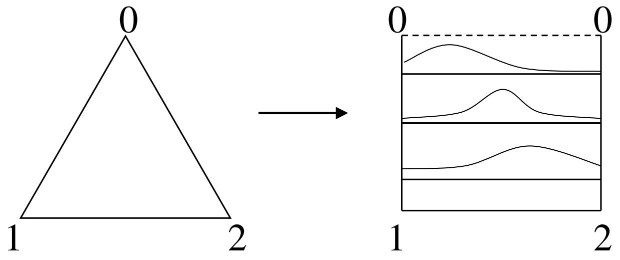


Figure 4.2: The twist in the prior is here illustrated for three predicates. To draw the probability assignment, the simplex is stretched at the side of the hypotheses with high probability for 0 to form a square. The curves over the three horizontal lines represent the probability assignment. Within the region of hypotheses with low probability for 0, most probability is assigned to hypotheses with high probability for 2, and for hypotheses with high probability for 0, most probability is assigned to hypotheses with high probability for 1. Observing 0 therefore implicitly increases the ratio of the probability of 1 over that of 2.

4.6 Analogy partition

This section proposes a transformation of the algebra $\mathcal{Q}$ into one on observations of predicates from the underlying families G and M. After that the hypotheses and densities that result in the system of $\lambda\gamma$ rules can be specified.

4.6.1 Defining the analogy partition

An observation algebra for gender and marital status. First we must define a space on which the algebra for records concerning G and M can be defined. Taking L as the set of ordered pairs $\langle g, m \rangle$, we can define the space L^ω of all infinitely long ordered sequences u of such observations:

$$u = g_1 m_1 \, g_2 m_2 \, g_3 m_3 \, \dots \tag{4.36}$$

The record that the individual i is a husband, $q_i = 1$, can then be written as two consecutive records in a sequence u, namely $g_i = 0$ and $m_i = 1$, meaning that the individual i is recorded to be male and married. More generally, we can identify all infinite strings of observations $e \in K^\omega$ with some infinite string

$u \in L^\omega$. Using $u(t)$ as the t-th number in the sequence u, we can construct

$$\begin{aligned} e(i) &= 2g_i + m_i, \\ u(2i-1) &= g_i, \\ u(2i) &= m_i. \end{aligned}$$

In this way every sequence e is mapped onto a unique sequence u, and every such u can be traced back to the original e.

We can now define the algebra $\mathcal{R}$ for records concerning the predicate families G and M on the basis of the space L^ω. The following elements generate this algebra:

$$G_i^g = \{u \in L^\omega : u(2i-1) = g\}, \tag{4.37}$$

$$M_i^m = \{u \in L^\omega : u(2i) = m\}. \tag{4.38}$$

The sets $G_i^g \cap M_i^m$ thus contain all those infinitely long sequences u that have the number g and m in the positions $2i-1$ and $2i$. The relations between the families Q, G and M are therefore as specified in equation (4.1). For future reference, sequences of records in $\mathcal{R}$ that correspond to a specific e_n are here denoted $S_n^{e_n}$.

Hypotheses for gender and marital status. The idea of the hypotheses concerning the underlying predicate families is essentially the same as for those concerning Q-predicates. We may again partition the above observational algebra into hypotheses concerning relative frequencies. However, the relative frequencies in the family M must in this case be related to the result in the family G. We may define the following relative frequencies:

$$f_g(u) = \lim_{n\to\omega} \frac{1}{n} \sum_{i=1}^{n} W_g(u(2i-1)), \tag{4.39}$$

$$f_m^g(u) = \lim_{n\to\omega} \frac{\sum_{i=1}^{n} W_g(u(2i-1)) W_m(u(2i))}{\sum_{i=1}^{n} W_g(u(2i-1))}. \tag{4.40}$$

Here $W_r(u(t)) = 1$ if $u(t) = r$ and $W_r(u(t)) = 0$ otherwise. The function f_g simply gives the relative frequency of results g within the observations with respect to G in the sequence u. But the function $f_m^g(u)$ is somewhat more complicated. It returns, for every u, the relative frequency of results m for observations with respect to the family M, conditional on the observed individual belonging to the category g within the family G. This is the relative frequency of m conditional on g within u, or the conditional relative frequency for short.

We are now in a position to define the analogy partition $\mathcal{A}$ for predictions concerning the predicate families G and M. The hypotheses in this partition employ the conditional relative frequencies in order to pick up the exact statistical dependency between the two families. Let α_g and α_{gm} be the parameters labelling these hypotheses, and define

$$W_{h_\alpha}(u) = \begin{cases} 1 & \text{if } f_g(u) = \alpha_g \text{ and } f^g_m = \alpha_{gm}, \\ 0 & \text{otherwise.} \end{cases} \tag{4.41}$$

and then define the hypotheses $H_\alpha = \{u \in L^\omega | W_{h_\alpha}(u) = 1\}$. Again define $H_{\neg\alpha}$ as the set of all u for which one of the three relative frequencies in equations (4.39) or (4.40) does not exist. The analogy partition is then given by $\mathcal{A} = \{H_{\neg\alpha}, \{H_\alpha\}_{\alpha \in A}\}$. Here the parameter $\alpha = \langle \alpha_0, \alpha_1, \alpha_{00}, \alpha_{01}, \alpha_{10}, \alpha_{11} \rangle$ lies in the set $A = \{\alpha \in [0,1]^6 \mid \alpha_0 = 1 - \alpha_1, \alpha_{00} = 1 - \alpha_{01}, \alpha_{10} = 1 - \alpha_{11}\}$.

The likelihoods of the hypotheses on the underlying predicates are given by these relative frequencies and conditional relative frequencies:

$$p(G^g_{i+1} | H_\alpha \cap S^e_i) = \alpha_g, \tag{4.42}$$

$$p(M^m_{i+1} | H_\alpha \cap S^e_i \cap G^g_{i+1}) = \alpha_{gm}. \tag{4.43}$$

It may be noted that observations in the sequence S^e_i do not influence the likelihoods of H_α, but that G^g determines which of the α_{gm} must be used as the likelihood for M^m. In this sense, the likelihoods for the family M depend on earlier observations in the family G. Note also that we can write

$$\begin{aligned} p(Q^{2g+m}_{i+1} | H_\alpha \cap S^e_i) &= p(G^g_{i+1} \cap M^m_{i+1} | H_\alpha \cap S^e_i) \\ &= \alpha_g \, \alpha_{gm}. \end{aligned} \tag{4.44}$$

The likelihoods for the separate families G and M therefore also imply likelihoods for the family Q, and with that also unconditional likelihoods for the family M.

4.6.2 Transforming partitions

Parameter transformation. It is useful to consider the parameter space A for the above partition in some more detail, and relate it to the parameter space B. First recall that pairs of the parameter components of α sum to one. The parameter space A is therefore built up from a separate simplex B_G for the two parameters α_g, and two simplexes B_{gM} for the four parameters α_{gm}. We can write

$$A = B_G \times B_{0M} \times B_{1M}. \tag{4.45}$$

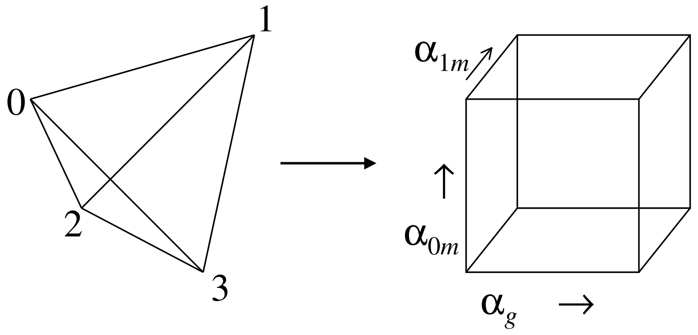


Figure 4.3: A representation of the transformation from B to A. The space B is a tetrahedron, in which the parameters are not orthogonal, the space A is a cube. Defining probability distributions over A is much easier.

Like the original simplex B for $|K| = 4$, the parameter space A therefore has three independent components. In fact, following the above expression for the likelihoods, the parameter space B can be obtained from A by a simple transformation:

$$\theta_{2g+m} = \alpha_g \alpha_{gm}. \tag{4.46}$$

When it comes to the statistical hypotheses, the partitions $\mathcal{A}$ and $\mathcal{B}$ are thus equivalent. However, they employ different parameter spaces, and therefore provide access to different classes of prior probability functions. As illustrated in figure 4.3, the space B is a tetrahedron, which is transformed in the cube A. This is the transformation intended at the end of section 4.5.

Probability over the transformed space. The prior probability assignment over $\mathcal{A}$ that generates the system of $\lambda\gamma$ rules can now be made precise. It is noteworthy that the partition $\mathcal{A}$ consists of three separate and orthogonal dimensions. However, it is not yet clear whether these three dimensions can be treated independently, let alone that they result in such prediction rules. To establish the independence, it must be determined under what conditions the updates and predictions over the parts do not affect each other. As it turns out, independence is guaranteed if we assume that the prior probability density is factorisable:

$$p(H_\alpha) = p_G(\alpha_g)\, p_{0M}(\alpha_{0m})\, p_{1M}(\alpha_{1m}). \tag{4.47}$$

In that case updating in one of the dimensions leaves the functions in the other two dimensions unchanged. The separate dimensions in A may then be associated with the separate $\lambda\gamma$ rules of 4.16 and 4.17.

In order to derive these $\lambda\gamma$ rules, we must assume more than factorisability. We must assume that the prior densities over the separate dimensions are members of the Dirichlet class:

$$p(H_\alpha) \sim \prod_g \alpha_g^{(\lambda\gamma_g - 1)} \prod_m \alpha_{0m}^{(\lambda_M^0 \gamma_{Mm}^0 - 1)} \prod_m \alpha_{1m}^{(\lambda_M^1 \gamma_{Mm}^1 - 1)}. \tag{4.48}$$

From here onwards the derivation of the separate $\lambda\gamma$ prediction rules runs entirely parallel to the derivation of a single rule. Again, the details for this derivation may be found in Festa (1993: 57-71).

The probability model shows that the predictions resulting from them are indeed exchangeable, meaning that the predictions are invariant under permutations of the order of observations. Since we can assign a likelihood for an observation M_i^m before G_i^g on every hypothesis H_α, we can adapt the probability over $\mathcal{A}$ for these unconditional observations of M_i^m in the same way as that we can adapt the probability upon observing G_i^g. Both updates are simply multiplications with the likelihood functions. In the Bayesian model, there is therefore no principled restriction on the order of the observations, and in this sense the Bayesian model offers a wider framework than the system of $\lambda\gamma$ rules. The restriction only shows up as the fact that the integrals for the predictions based on the Bayesian model cannot be solved analytically, in the form of a system of prediction rules, if the observations M_i^m occur before G_i^g.

Transformations make a prior accessible. Let me return to the relation between the parameter spaces B and A. Recall that the class of Dirichlet densities over B corresponds to a special class of Dirichlet densities over A, which results in the predictions determined by equation (4.21). This follows from the fact that for this choice of parameters, the system of rules boils down to a single $\lambda\gamma$ rule. At the level of the partitions, however, we may also transform the Dirichlet density over B by means of the relations (4.46), and multiply the transformed density with the Jacobian determinant of the transformation matrix. This results in the very same equivalence. On the other hand, there are many more Dirichlet densities over A that cannot be captured by the Dirichlet densities over B in this way. Transforming these densities over A back to B is a less clear-cut operation, and the resulting densities over B do not fall within any special class of densities. The transformation of B into A has in this way provided access to a new class of prior densities.

As suggested, the above perspective opens up the possibility of modelling many other analogical predictions. We may consider densities over A that are not Dirichlet, and more specifically, that are not factorisable. However, these latter analogical predictions can only be dealt with in the next chapter. The next section only presents a brief sketch of these possibilities. For present purposes, the main point is that the system of rules has been connected to a range of statistical models: the existence of these models ensures the consistency of the system of rules. Moreover, in view of the methodological perspective that is adopted in this chapter and in this thesis more generally, the task of an inductive logician is no more than to supply these models, in order to bring out the inductive assumptions that drive analogical predictions and provide the means to manipulate these assumptions.

4.7 Generalizing the analogy partition

This section argues that the above discussion can be generalised to cover analogical predictions based on explicit similarity more generally. It considers the extension of the foregoing to cases with more than two underlying predicate families. It further suggests how a problem for the model of Hesse, as noted in Maher (2001), can be solved. The solution opens up a number of interesting modelling possibilities.

More elaborate partitions. Until now we have been concerned with explicit similarity between predicate families M and G, but nothing precludes the use of more than two underlying predicates, or of more cells within each partition. With the same construction we can model predictions based on explicit similarity relations that are much more complex than the ones exemplified above. As an example, let us say that before recording gender and marital status, we observe the type of car C in which the individual arrives, and that we distinguish between family cars, $c = 0$, vans, $c = 1$, and sports cars, $c = 2$. We may then keep track of a dependency between marital status and gender, which can on itself be made dependent on the car type. As an example, the parameter space for that partition may be

$$A = B_C \times \prod_c B_{cG} \times \left(\prod_g B_{cgM} \right), \tag{4.49}$$

All simplexes can again be associated with separate prediction rules, leading to an extended system of prediction rules. Note that the simplex B_C is an equilateral triangle, and that the corresponding $\lambda\gamma$ rule has three possible observation results. It will be clear that in adding further underlying predicates there are no restrictions.

Varying the order of predicate families. As already discussed in section 4.4, the system of rules is suitable for expressing analogical predictions based on explicit similarity only, which means that it cannot express all possible configurations of symmetric relevance relations between the Q-predicates. To characterise the restriction on expressible relevance relations for the general case, recall first that the analogy partition always determines a certain order to the underlying predicates, such as first B, then G, and finally M. If we associate these relations with an increasing ranking number, the restriction to expressible relevance configurations may be characterised as follows: the system of prediction rules can only distinguish between the relevances of a predicate Q^q to the predicates Q^v and Q^w if the ranking numbers of the first predicate in the ranking that Q^v and Q^w do not have in common with Q^q are not the same. In other words, the system is not able to model a difference between relevances of Q^q to predicates Q^v and Q^w if the first predicate in the ranking in which the latter two differ from Q^q is the same. We may for example consider husbands more relevant to wives than to maidens. However, the system of rules cannot model these relations between the Q-predicates, because both maidens and wives differ from husbands in the first underlying predicate family in that example, namely in G.

The system of rules thus accommodates explicit similarity specifically of gender, or of marital status, but never of both. One exception to this may now be presented by slightly adapting the above example with three underlying predicates. For simplicity, the family C only concerns family cars, $c = 0$, and vans, $c = 1$. Imagine that we think that driving a van is somehow indicative of the gender, suggesting male drivers irrespective of their marital status, and further that we think family cars are indicative of the marital status, suggesting a married driver irrespective of their gender. In that case it is natural to employ the following analogy partition:

$$A = B_C \times B_{0M} \times B_{1G} \times B_{00G} \times B_{01G} \times B_{10M} \times B_{11M}. \qquad (4.50)$$

Conditional on the individual driving a family car, we make the marital status indicative of the marital status of further family car drivers. But conditional on the individual driving a van, we make the gender indicative of the gender of further van drivers. In other words, the direction of dependence relations may vary within one analogy partition, as long as these directions are themselves conditioned on different predicates from a third family.

Hesse's problem. In the remainder of this section I illustrate one further generalisation, which uses the statistical models to capture configurations of relevance relations ρ that are not covered by the system of $\lambda\gamma$ rules. To this aim I discuss an example from Maher (2001), which reveals a shortcoming in the model of analogical predictions proposed by Hesse. Contrary to that model, the statistical models can deal with the example case. It is notable that the model of Carnap and Kemeny also overcomes the difficulties of Hesse, and that their model is still more general than the models sketched here when it comes to expressible relevance relations. However, the statistical models offer a number of unexplored modelling possibilities, which may eventually solve the problems with the model of Carnap and Kemeny as well.

The example of Maher concerns the predicates of being a swan X, being Australian Y, and being white Z. The indices x, y and z are 1 or 0 for the predicate being satisfied or not. Imagine that until now we have recorded, of all animals in the world, whether they are a swan and whether they are Australian. Specifically, we have observed

$$S_{\omega}^{XY} = X_1^1 \cap Y_1^0 \cap X_2^1 \cap Y_2^1 \cap \left(\bigcap_{i=3}^{\omega} X_i^{x_i} \cap Y_i^{y_i} \right), \tag{4.51}$$

the sequence of observations of all animals with respect to being a swan and being Australian, the first animal in the sequence being a non-Australian swan, the second an Australian swan. The challenge is to define a probability over the algebra $\mathcal{Q}$ or, equivalently, $\mathcal{R}$ for which

$$p(Z_2^1 | Z_1^1 \cap S_{\omega}^{XY}) > p(Z_2^1 | S_{\omega}^{XY}). \tag{4.52}$$

That is, we want the fact that a non-Australian swan is white to be relevant to the probability of an Australian swan being white. The fact that we already know the proportions of Australian and non-Australian swans and non-swans should have nothing to do with this relevance. But unfortunately the model of Hesse cannot accommodate such a relevance.

It turns out that the above inequality can be derived by employing a restricted prior over an analogy partition. For simplicity, use the parameter space $A = B_{XY} \times B_{00Z} \times B_{01Z} \times B_{10Z} \times B_{11Z}$, which has a single, three-dimensional simplex B_{XY} for the combined predicates X and Y on being a swan and being Australian. In this tetrahedron space, conditioning on S_{ω}^{XY} forces all probability to be concentrated on one point hypothesis within the simplex B_{XY}, associated with the actual relative frequencies α_{xy} of the observations in S_{ω}^{XY}. Subsequent observations Z_i^z on being white therefore only influence the probability over the remaining spaces, B_{xyZ} for all values of xy. Now the challenge is to establish the relevance of observations of being white within the category

$xy = 10$, concerning non-Australian swans, to the probability of animals being white in the category $xy = 11$, concerning Australian swans. In other words, we must somehow couple the probability assignment over the simplex B_{11Z} to the assignment over B_{10Z}.

One way of doing this is by restricting the probability assignment to a specific subspace of the hypotheses space A, defined by $\alpha_{10z} = \alpha_{11z}$. Effectively, the marginal probability assignments over the two simplexes B_{10Z} and B_{11Z} are then identified, so that adapting the probability over B_{10Z} for the observation Z_i^1, given that $X_i^1 \cap Y_i^0$, implicitly changes the probability over B_{11Z} as well. In other words, finding a non-Australian swan to be white is immediately relevant for the probability of Australian swans being white. The probability over the remaining space on non-swans, $B_{01Z} \times B_{00Z}$, can be chosen freely.

Within the hypotheses space on swans, in which all probability is restricted to $\alpha_{10z} = \alpha_{11z}$, we may again choose a Dirichlet distribution. For the resulting predictions, this means that observations of swans, both Australian and non-Australian, are collected in the same $\lambda\gamma$ prediction rule on being white, Z:

$$p(Z_{n+1}^z | S_n^Z \cap S_\omega^{XY}) = \frac{n_{Zz}^{11} + n_{Zz}^{10} + \lambda_Z^1 \gamma_{Zz}^1}{n_Z^{11} + n_Z^{10} + \lambda_Z^1}. \tag{4.53}$$

For this prediction to be applicable, we must have that $S_\omega^{XY} \subset X_{n+1}^1$, meaning that animal $n+1$ is indeed a swan. Predictions for the case in which $S_\omega^{XY} \subset X_{n+1}^0$ are determined by the probability over the space on non-swans, $B_{01Z} \times B_{00Z}$. Note that S_n^Z denotes the sequence of observations of the first n animals with respect to Z. As in the foregoing, n_Z^{11} and n_Z^{10} are defined as the numbers of Australian and non-Australian swans in the sequence S_n^Z, and n_{Z1}^{11} and n_{Z1}^{10} as the numbers of Australian and non-Australian white swans in that sequence. The above rule further uses the abbreviations $\lambda_Z^1 = \lambda_Z^{11} = \lambda_Z^{10}$ and $\gamma_{Zz}^1 = \gamma_{Zz}^{11} = \gamma_{Zz}^{10}$. These parameters are the same for the simplexes B_{11Z} and B_{10Z}, because they are determined by one and the same probability assignment.

It can be checked easily that the inequality (4.52) is indeed satisfied in this model. It must be conceded that the above model has its own peculiarities. A much more detailed study of analogical predictions based on non-factorisable priors over the partition $\mathcal{A}$ is necessary in order to make any more general claims on its relative merits and defects. However, the above example does suggest that the use of statistical models deserves further attention.

4.8 Conclusion

This chapter presents a system of $\lambda\gamma$ rules that models analogical predictions based on analogy by explicit similarity of gender. After presenting an example of such similarity, the chapter shows how it translates to a specific subset of relevance relations between predicates in the aggregated family Q: the relevance of the predicate Q^{2g+m} for the predicate $Q^{2g+m'}$, which has predicate G^g in common with Q^{2g+m}, differs from its relevance for the two predicates $Q^{2g'+m}$ and $Q^{2g'+m'}$, that do not have G^g in common with Q^{2g+m}.

After presenting a system of rules that indeed models these relevance relations, I provide the Bayesian model that underlies the system. It is shown that analogy hypotheses treat observations M^m separately for the earlier observation G^g with $g = 0, 1$, by defining separate relative frequencies for them, and associating these frequencies with separate dimensions in the parameter space A. By assuming the prior over this space to be a product of Dirichlet marginals, the system of $\lambda\gamma$ rules can be derived. The chapter ends with some generalisations on the proposed system of rules.

The next chapter explores the possibilities of the Bayesian model that underlies the system of rules. It will turn out that this model provides the setting for a completely general model of analogical predictions based on symmetric inductive relevance relations, by employing non-factorisable probability functions over A. More generally, on the level of research programmes, I take it to be an important advantage of the present model that it seeks to integrate the rather isolated discussion on analogical predictions within Carnapian inductive logic in a wider framework of Bayesian statistical inference.

5

A General Model for Analogical Predictions

This chapter presents a general model for exchangeable analogical predictions, based on inductive relevance between four predicates. It extends the model for analogy by explicit similarity, which was dealt with in the preceding chapter. It is first shown that the predictions of this model may be captured in a more general statistical framework for exchangeable predictions. This is then used to define a model that is able to incorporate all possible symmetric inductive relevance relations. however, not all aspects of analogical predictions find a natural formalisation in the model. The chapter therefore only offers a partial explication of the model alongside some numerical approximations. At the end the model of this chapter is related to some other models in the literature.

It is not necessary to read chapter 4, or any of the other chapters, before starting on this one. But it may be noted that chapter 4 provides a different perspective on very similar inductive schemes. There is considerable overlap in the technical parts of the chapters. Reading the preceding chapter may therefore be helpful in coming to a more complete understanding of the schemes discussed in this chapter.

5.1 Introduction

Analogical predictions deviate from standard Carnapian predictions, as for example in Carnap (1952), because analogical predictions incorporate considerations of inductive relevance between predicates. This section discusses a general scheme for capturing such relevance relations. After that I make explicit the kind of analogical predictions that this chapter deals with, and contrast them with the analogical predictions of the preceding chapter.

5.1.1 Inductive relevance

Party centre example. Consider a marketing manager of a party centre who makes predictions on visitors concerning their gender and marital status. There are four aggregated predicates: bachelors, husbands, maidens, and wives. Let us say that the manager does not know the party centre yet, but that she has some general knowledge of party centres. Specifically, she knows that men like to hang out in a party centre together, so that recording a bachelor is positively relevant to recording a husband and vice versa. She further knows that married couples often go out together, and also that, while husbands tend to bring their bachelor friends if they go out with their wives, the wives tend not to invite their maiden friends. Therefore, while husbands and wives are strongly relevant to each other, as husbands and bachelors are, maidens and wives bear a much weaker relevance relation. The challenge of this chapter is to find a model for inductive predictions that incorporates complex configurations of such relevance relations.

A vector of relevance relations. Consider the notion of inductive relevance itself. We can formalise the above predicates as Q^q, with the numbers $q = 0, 1, 2, 3$ associated with the predicates bachelor, husband, maiden and wife respectively. In terms of these predicates, the example has it that Q^1 is more relevant to Q^0 than to Q^2. This inductive relevance means that the observation that individual i has predicate Q^1, or Q^1_i for short, is more favourable to the probability of the observation Q^0_{i+1} than to that of Q^2_{i+1}. Note that this may obtain quite independently of the purely inductive effect that observation Q^1_i makes observation Q^1_{i+1} more probable.

The relevance of Q^q to Q^w may be expressed in a function $\rho(q, w)$, where q and w denote predicate numbers. With this relevance function, we may specify in general terms what it means for Q^q to be more relevant to Q^w than to Q^v. Denoting the probability of observation Q^q_i with the function p, we can write

$$\rho(q, w) > \rho(q, v) \quad \Rightarrow \quad \frac{p(Q^w_{i+1}|E_{i-1} \cap Q^q_i)}{p(Q^w_i|E_{i-1})} > \frac{p(Q^v_{i+1}|E_{i-1} \cap Q^q_i)}{p(Q^v_i|E_{i-1})}. \tag{5.1}$$

It must be noted that the foregoing is not the only possible expression of inductive relevance. For a review of possible relations between inductive methods and the relevance function, see Festa (1997). The above is qualitatively equivalent to $K_{>G}$ inductive methods in his terminology.

This chapter deals with complex configurations of relevance relations, like those exemplified in the party centre example. By means of the above relevance function $\rho(v, w)$, its aim can be made a bit more specific. First, as in the preceding chapter, I restrict attention to symmetric relations:

$$\rho(v, w) = \rho(w, v). \tag{5.2}$$

Because the relevance relations are symmetric, the set of possible configurations of relevance between predicates may be captured in the following space:

$$\rho = \langle \rho(0,1),\ \rho(2,3),\ \rho(0,2),\ \rho(1,3),\ \rho(1,2),\ \rho(0,3)\rangle. \tag{5.3}$$

The more specific aim of this chapter is to provide a model of inductive predictions for all the relevance configurations that are represented in this space.

Some disclaimers. It may be remarked immediately that some important aspects of inductive relevance are not expressed in the representation ρ. First of all, expression (5.1) does not yet provide a meaning for the sizes of ρ. Until now the characterisation is entirely qualitative, only providing an interpretation for the ordering of the sizes. To a certain extend the numerical values of ρ are given a further interpretation below. Second, because I am considering exchangeable predictions, these predictions will converge onto the actual relative frequencies of the predicates Q^q. The analogy effects will therefore diminish with the accumulation of observations. But the foregoing does not specify exactly how the effects will diminish, or even the overall rate at which this happens. This aspect is simply not captured in the representation ρ. It may be that the rates are different for the different relevance relations, so that the ordering in these relations varies with the number and the nature of the observations. However, as will be argued below, in the present model the analogy effects diminish at the same rate. The relevance ordering therefore remains intact.

5.1.2 Aim of this chapter

The model for explicit similarity. It is instructive to relate it to the model of the preceding chapter, which concerned analogical predictions based on explicit similarity relations. The aggregate predicate family Q on bachelors, husbands, maidens and wives is built up from separate predicate families on gender, G, and marital status, M. We can write

$$G^g = Q^{2g} \cup Q^{2g+1}, \tag{5.4}$$

$$M^m = Q^m \cup Q^{2+m}, \tag{5.5}$$

where G^g means male for $g = 0$ and female for $g = 1$, while M^m means not married for $m = 0$ and married for $m = 1$. The idea of the model for explicit similarity is that predictions on predicates Q^q may be written down in terms of predictions for the separate families G and M:

$$p(Q^q_{n+1}|E_n) = \frac{n_{Gg} + \lambda_G \gamma_{Gg}}{n_G + \lambda_G} \ \times\ \frac{n^g_{Mm} + \lambda^g_M \gamma^g_{Mm}}{n^g_M + \lambda^g_M}. \tag{5.6}$$

Here $n_G = n$ is the total number of individuals in the preceding observations E_n, $n_{Gg} = n^g_M$ the number of individuals with gender g, and n^g_{Mm} the number of individuals with gender g and marital status m.

Recall that every point in the vector space ρ represents a different configuration of symmetric inductive relevance relations. The models based on explicit similarity can now be captured by a specific subset of configurations in this vector space, namely:

$$\rho = \langle\, \rho_{G0},\, \rho_{G1},\, \rho_{\bar{G}},\, \rho_{\bar{G}},\, \rho_{\bar{G}},\, \rho_{\bar{G}},\, \rho_{\bar{G}} \rangle. \tag{5.7}$$

The relevances $\rho_{Gg} = \rho(2g, 2g+1)$, the so-called intra-gender relevances, may be chosen independently, and the inter-gender relevances $\rho_{\bar{G}}$ must all be chosen equal. They are connected to the parameters in the above prediction rule according to

$$\lambda^g_M = \rho_{Gg}\gamma_{Gg}N, \tag{5.8}$$

$$\lambda_G = \rho_{\bar{G}}N. \tag{5.9}$$

The model of this chapter is a generalisation of the model for explicit similarity: every component of ρ can in this chapter be determined independently.

Position of the present chapter. The ideas in this chapter are strongly connected to earlier chapters. As indicated, the present model relies heavily on the model for explicit similarity, as it is discussed in the preceding chapter. The results of that chapter are here used uncritically. Furthermore, just as the preceding one, this chapter employs statistical hypotheses for generating inductive predictions. That is, the model first incorporates observations on predicates into a probability assignment over statistical hypotheses, and then employs the assignment over these hypotheses to derive predictions for new observations. As discussed in chapter 3, the use of hypotheses offers control over the assumptions underlying inductive predictions. The present chapter illustrates this. It shows how transformations between hypotheses spaces enable us to define priors that are otherwise hard to find.

The plan of this chapter is as follows. Sections 5.2 and 5.3 introduce observational algebras for the above Q-predicates and for the predicates such as G and M that underlie them, and defines the predictions based on hypotheses concerning these respective predicates. Section 5.4 elaborates on the relation between Q-predicates and underlying predicates. Specifically, it shows how the prior probability over hypotheses concerning underlying predicates that encodes analogical effects can be translated to a specific prior probability over the hypotheses concerning Q-predicates. Section 5.5 presents the general model for analogical predictions concerning Q-predicates, and discusses some of its limitations. Section 5.6 discusses the idea behind the model, and provides some numerical simulations. Finally, section 5.7 discusses the general model in light of some other models of analogical predictions. In the conclusion, the models of this chapter and the preceding one are considered from a general perspective.

5.2 Hypotheses schemes for Q-predicates

This section discusses the observational algebra for Q-predicates, and contains a short introduction to Bayesian schemes that employ hypotheses for making predictions. It also provides the partition for Q-predicates that underlies the Carnapian $\lambda\gamma$ prediction rules. The scheme and partition are elaborately discussed in preceding chapters, and in particular in 4.5. Comparable expositions can be found in Jeffrey (1984) and Howson and Urbach (1996).

Observational algebra. The expression Q_i^q refers to the observation that individual i has predicate Q^q. To characterise inductive predictions, let me represent these observations in terms of a so-called observational algebra. Let K be the set of possible values for q, so that in the case of the party centre $K = \{0, 1, 2, 3\}$. The infinite product K^ω is the space of all infinite sequences e of such values:

$$e = q_1 q_2 q_3 \dots \tag{5.10}$$

The observational algebra, denoted $\mathcal{Q}$, consists of all possible subsets of the space K^ω. If we denote the i-th element in a series $e \in K^\omega$ with $e(i)$, we can define an observation Q_i^q as an element of the algebra $\mathcal{Q}$ as follows:

$$Q_i^q = \{e : \; e(i) = q\}. \tag{5.11}$$

Note that there is a distinction between the observation Q_i^q and the result of an observation q. The values, represented with small letters, are natural numbers. The observations, denoted with large letters, are elements of the algebra $\mathcal{Q}$.

In the same way we can define an element in the algebra that represents a finite sequence of observations. If we define the ordered n-tuple $e_n = \langle q_1 q_2 \dots q_n \rangle$ and q_i as the i-th element therein, we can write

$$E_n^{e_n} = \{e : \; \forall i \leq n \, (e(i') = q_i)\}. \tag{5.12}$$

I normally suppress reference to the n-tuple e_n. The observations and sequences of observations are related to each other according to

$$E_n \cap Q^q_{n+1} = E_{n+1} \tag{5.13}$$

where $e_{n+1}(n+1) = q$. Finally, for any sequence e_n we can write down, for all numbers $q < 4$, the number of times it occurs within the sequence. These numbers are in the following denoted with n_{Qq}. Since the total number of observations $n_Q = n = \sum_q n_{Qq}$, the numbers n_{Qq} together define the observed relative frequencies $\frac{n_{Qq}}{n_Q}$ of the results q.

We can now define a probability function p over the algebra $\mathcal{Q}$. The probabilities of observations Q^q_{n+1} and E_n can then be interpreted as predictions. An important matter is how these predictions depend on observations: if the series e_n is observed, this must somehow change the predictions over the observations. I here assume that the predictions upon observing e_n are expressed by the original probability function p conditional on the observations E_n, denoted $p(\cdot|E_n)$. This dependence of predictions on observations is known as Bayesian conditioning. In the following, the initial probability is called the prior probability, and the conditional one the posterior.

Bernoulli hypotheses and exchangeable predictions. The schemes of this chapter employ partitions of statistical hypotheses to define the probability function p. A partition is a collection $\mathcal{B} = \{H_\theta\}_{\theta \in B}$ in which the hypotheses H_θ are mutually exclusive and jointly exhaustive possibilities. We can associate these hypotheses with elements of the algebra $\mathcal{Q}$, but for present purposes the less strict characterisation suffices. We can define predictions $p(Q^q_{n+1}|E_n)$ with the law of total probability over the partition:

$$p(Q^q_{n+1}|E_n) = \int_B p(H_\theta|E_n)p(Q^q_{n+1}|H_\theta \cap E_n)\, d\theta. \tag{5.14}$$

The probability over the hypotheses is determined by the probability density $p(H_\theta|E_n)$. The terms $p(Q^q_{n+1}|H_\theta \cap E_n)$ are called the likelihoods on the hypotheses H_θ, which are defined for observations Q^q_{n+1} . The prediction is obtained by weighing these likelihoods with the posterior probability over the hypotheses.

Apart from the likelihoods, the dependence of the predictions on observations are reflected in the probability assignment over the hypotheses. This probability may be determined by means of Bayesian conditioning,

$$p(H_\theta|E_{i+1})d\theta = \frac{p(Q^q_{i+1}|H_\theta \cap E_i)}{p(Q^q_{i+1}|E_i)} p(H_\theta|E_i)d\theta, \tag{5.15}$$

where $e_{i+1}(i+1) = q$. Note that the denominator $p(Q^q_{i+1}|E_i)$ can be rewritten with equation (5.14). The posterior probability over the hypotheses $p(H_\theta|E_n)d\theta$ can thus be determined recursively by the prior probability assignment $p(H_\theta)d\theta$, and the likelihoods $p(Q^q_{i+1}|H_\theta \cap E_i)$ for all times $0 \leq i < n$. With the likelihoods $p(Q^q_{n+1}|H_\theta \cap E_n)$ we can then determine the predictions.

This chapter employs specific statistical hypotheses H_θ, which have the following likelihoods for the observations Q^q_{i+1}:

$$p(Q^q_{i+1}|H_\theta \cap E_i) = \theta_q. \tag{5.16}$$

The domain of the 4-tuple θ is a simplex, $\sum_q \theta_q = 1$. Note also that the likelihoods are independent of the earlier observations E_i. The posterior likelihoods are thus identical to the prior likelihoods. Finally, the predictions resulting from the partition $\mathcal{B}$ are exchangeable, and by De Finetti's representation theorem every exchangeable prediction rule can be captured by a prior probability over the hypotheses, $p(H_\theta)d\theta$.

Carnapian predictions. One specific prior probability assignment must be given separate attention. If we assume the prior density function over the simplex to have a Dirichlet form,

$$p(H_\theta) \quad \sim \quad \prod_q \theta_q^{(c_q-1)}, \tag{5.17}$$

then the resulting predictions are of the form of Carnapian rules $p_{\lambda\gamma}$:

$$p(Q^q_{n+1}|E_n) = \frac{n_{Qq} + \lambda_Q \gamma_{Qq}}{n_Q + \lambda_Q} = pr_{\lambda\gamma}(n_{Qq}, n_Q). \tag{5.18}$$

The values of the parameters λ_Q and γ_{Qq} are determined by the exponents c_q in the Dirichlet density according to $\lambda = \sum_q c_q$ and $\gamma_q = c_q/\lambda$. In this chapter I restrict attention to natural numbers c_q. Finally, the numbers n_{Qq} and n_Q are as defined in the preceding section.

5.3 Schemes using underlying predicates

This section presents the algebra for the underlying predicates G and M. It further introduces the hypotheses partition $\mathcal{A}$ associated with this algebra, and argues that this partition leads to the system of $\lambda\gamma$ rules of section 5.1. This section shows considerable overlap with section 4.6, but it is somewhat more general.

Algebra for underlying predicates. Let me define an observation algebra for the underlying predicates G and M, as introduced in section 5.1.1. Recall the indices of these predicates, $g, m \in \{0, 1\}$. With L_{GM} as the set of ordered pairs $\langle g, m \rangle$ we can define the space $(L_{GM})^{\omega}$ of all infinitely long ordered sequences u of such index pairs:

$$u = g_1 m_1 \, g_2 m_2 \, g_3 m_3 \, \ldots \tag{5.19}$$

We can then identify all infinite strings of observations $e \in K^{\omega}$ with a unique infinite string $u \in (L_{GM})^{\omega}$:

$$\begin{aligned} e(i) &= 2g_i + m_i, & (5.20) \\ u(t) &= \begin{cases} g_i & \text{if } t = 2i - 1, \\ m_i & \text{if } t = 2i. \end{cases} & (5.21) \end{aligned}$$

So for every odd index t, the number $u(t)$ concerns an observation of predicate G, and for every even index t the number $u(t)$ concerns M. Thus every sequence e is mapped onto a unique sequence u, and every such u can be traced back to a corresponding sequence e.

Two things must be remarked on this translation of Q-predicates to underlying predicates. First, note that the order of the underlying predicates G and M is fixed in the definition of the ordered pairs L_{GM}. However, we may just as well consider the set L_{MG}, and define a space of infinite sequences u' on the basis of that. Moreover, the predicates G and M do not exhaust the possible pairwise combinations of Q-predicates. We can also employ a third partitioning of the Q-predicates:

$$W^{w} = Q^{1-w} \cup Q^{2+w}. \tag{5.22}$$

As a slightly contrived interpretation, imagine that it is a custom of the people featuring in the example that bachelors are given a traditional wedding ring at their 18th birthday. This ring is a sign that the bachelor has reached the age at which he is allowed to propose to a maiden, and it serves as a present to the bride at the wedding ceremony. Therefore, people who are in possession of this traditional ring are either bachelors or wives, and people who are not are either maidens or husbands. The important thing here is that translations of sequences e into sequences in a space based on L_{WG}, or on some other combination with the family W, may be considered just as well.

We can now define the algebra $\mathcal{R}_{GM}$ for observations of predicate families G and M in the space $(L_{GM})^{\omega}$, in the same way as we defined the algebra $\mathcal{Q}$:

$$\begin{aligned} G_i^g &= \{u \in (L_{GM})^{\omega} : \, u(2i-1) = g\}, & (5.23) \\ M_i^m &= \{u \in (L_{GM})^{\omega} : \, u(2i) = m\}. & (5.24) \end{aligned}$$

The sets $G_i^g \cap M_i^m$ contain all those infinitely long sequences $u \in (L_{GM})^\omega$ that have the number g and m in the positions $2i-1$ and $2i$ respectively. We can therefore also translate

$$\begin{aligned} Q_i^{(2g+m)} &= G_i^g \cap M_i^m, & (5.25) \\ E_n^{e_n} &= S_n^{e_n} = \bigcap_{i=1}^{n} G_i^{g_i} \cap M_i^{m_i}. & (5.26) \end{aligned}$$

In this way there is a complete mapping of the elements Q_i^q and E_n in $\mathcal{Q}$ onto elements of the algebra $\mathcal{R}_{GM}$.

Hypotheses for underlying predicates. We may define inductive predictions concerning Q-predicates by providing a probability function p over the algebra of underlying observations, $\mathcal{R}_{GM}$. Again we can employ a partition of hypotheses H_α with the parameter space $\alpha \in A_G$, so that $\mathcal{A}_G = \{H_\alpha\}_{\alpha \in A_G}$. The hypotheses H_α concern observations G_i^g and M_i^m, that is, they provide likelihoods for these observations. We may choose

$$\begin{aligned} p(G_{i+1}^g | H_\alpha \cap S_i) &= \alpha_{Gg}, & (5.27) \\ p(M_{i+1}^m | H_\alpha \cap G_{i+1}^g \cap S_i) &= \alpha_{Ggm}. & (5.28) \end{aligned}$$

These likelihoods do not depend on observations concerning other individuals: the parameters α do not depend on S_n. However, every hypothesis does have separate likelihoods for observations M_{i+1}^m conditional on either G_{i+1}^0 or G_{i+1}^1.

We can use the hypotheses H_α to generate predictions over the underlying predicate families G and M, just as we used hypotheses H_θ for direct predictions of the family Q. Since the algebra $\mathcal{R}_{GM}$ determines that we observe G_{i+1}^g before observing M_{i+1}^m, all relevant likelihoods are in this way defined. The main difference with the above discussion is in the parameter space A_G. Instead of a single simplex B with $\sum_q \theta_q = 1$, we now have a Cartesian product of three simplexes $A_G = B_G \times B_{0M} \times B_{1M}$, with components:

$$\sum_g \alpha_{Gg} = 1, \qquad \sum_m \alpha_{G0m} = 1, \qquad \sum_m \alpha_{G1m} = 1. \qquad (5.29)$$

Since in the example we have $g, m = 0, 1$, we simply have $\alpha_{G1} = 1 - \alpha_{G0}$ and $\alpha_{Gg1} = 1 - \alpha_{Gg0}$.

Carnapian rules for underlying predicates. As in the foregoing, the probability density over the hypotheses space determines the eventual predictions that derive from the hypotheses scheme. And indeed, if we assume a Dirichlet density over each separate simplex component of the parameter space A_G,

$$p(H_\alpha) \quad \sim \quad \prod_g \left(\alpha_{Gg}^{(a_{Gg}-1)} \times \prod_m \alpha_{Ggm}^{(a_{Ggm}-1)} \right), \tag{5.30}$$

we can derive Carnapian $\lambda\gamma$ rules for the predicate families G and M separately:

$$\begin{aligned} p(G^g_{n+1}|S_n) &= \frac{n_{Gg} + \lambda_G \gamma_{Gg}}{n_G + \lambda_G}, & (5.31) \\ p(M^m_{n+1}|S_n \cap G^g_{n+1}) &= \frac{n^g_{Mm} + \lambda^g_M \gamma^g_{Mm}}{n^g_M + \lambda^g_M}. & (5.32) \end{aligned}$$

The different dimensions in the parameter space are responsible for the independent prediction rules over G^g_{n+1}, and over M^m_{n+1} conditional on G^0_{n+1} and G^1_{n+1} respectively. Further, the numbers n_G, n_{Gg}, n^g_M and n^g_{Mm} are as indicated in section 5.1.

To complete the statistical underpinning of the system of $\lambda\gamma$ rules, recall the relation between the observations of families G and M on the one hand, and the observations of family Q on the other. We can write

$$p(Q^q_{n+1}|E_n) = p(G^g_{n+1}|S_n) \; \times \; p(M^m_{n+1}|S_n \cap G^g_{n+1}), \tag{5.33}$$

and thus arrive at the model for analogical predictions presented in equation (5.6). As indicated in section 5.1 and in the preceding chapter, this model allows us to express a specific subset of inductive relevance configurations.

Finally, let me provide the relations between the prior probability assignment over $\mathcal{A}_G$ and the parameters in the above prediction rules. As in the case of the prediction rule over Q-predicates, the exponents in the Dirichlet prior are directly related to these parameters:

$$\begin{aligned} \lambda_G &= \sum_g a_{Gg} & \gamma_{Gg} &= \frac{a_{Gg}}{\lambda_G}, & (5.34) \\ \lambda^g_M &= \sum_m a_{Ggm} & \gamma^g_{Mm} &= \frac{a_{Ggm}}{\lambda^g_M}. & (5.35) \end{aligned}$$

In choosing the Dirichlet priors over the simplex components of $\mathcal{A}_G$ we thus have separate command over the three Carnapian $\lambda\gamma$ rules in the system.

5.4 Transformations between partitions

The above presents a partition of statistical hypotheses for exchangeable predictions on Q-predicates. It also provides a specific partition for statistical hypotheses that, with a similar prior, results in exchangeable analogical predictions for explicit similarity. In this section we translate the prior over this latter partition into a prior over the general partition. This prepares for the general model of the next section.

5.4.1 Coordinate transformations

Equivalence of $\mathcal{B}$ and $\mathcal{A}$. Consider the two partitions $\mathcal{B}$ and $\mathcal{A}_G$. Recall that the parameter components of the partition $\mathcal{B}$, denoted θ_q, are the likelihoods for the observations Q^q, as expressed in (5.16). These likelihoods determine the nature of the partition: if we provide a prior probability over it, the predictions are determined. But it can further be noted that the partition $\mathcal{A}_G$ indirectly determines likelihoods for the Q-predicates as well:

$$\begin{aligned} p(Q^{(2g+m)}_{i+1}|H_\alpha \cap S_i) &= p(G^g_{i+1} \cap M^m_{i+1}|H_\alpha \cap S_i) \\ &= p(G^g_{i+1}|H_\theta \cap S_i)\, p(M^m_{i+1}|H_\theta \cap S_i \cap G^g_{i+1}) \\ &= \alpha_{Gg}\alpha_{Ggm}. \end{aligned} \tag{5.36}$$

This determines the update operation over $\mathcal{A}_G$ that corresponds to an update operation with $Q^{(2g+m)}_{i+1}$ over $\mathcal{B}$.

Every hypotheses H_α may now be identified with a hypothesis H_θ according to the set of transformation rules determined by the above equivalence:

$$\theta_{2g+m} = \alpha_{Gg}\alpha_{Ggm}. \tag{5.37}$$

Note that there are 4 components of θ that have to comply to 1 normalisation condition, so that $\mathcal{B}$ has 3 degrees of freedom. Since there are 6 components of α that have to comply to 3 normalisation conditions, the number of degrees of freedom in $\mathcal{A}_G$ is also 3. The mapping of equation (5.37) is in fact a bijection: for any hypothesis H_θ there is a unique H_α that has the same likelihoods for the observations Q^q_{i+1}. The partitions $\mathcal{B}$ and $\mathcal{A}_G$ are therefore essentially the same. This also means that the results on exchangeability and convergence, which may be proved for partition $\mathcal{B}$, hold for the partition $\mathcal{A}_G$ as well.

On the other hand, the structures of the parameter spaces B and A_G are certainly not identical. And as suggested above, this difference can be employed to access distributions over the hypotheses in $\mathcal{B}$ that are very difficult to come up with, or to investigate properties of, using the parameter space B itself. The access is provided by first defining the prior probability over the equivalent partition $\mathcal{A}_G$, employing the characteristics of the prediction rules defined for that partition, and by subsequently transforming this prior into one over the partition $\mathcal{B}$. The prior over $\mathcal{B}$ that is obtained in this way will result in the very same predictions as those derivable from $\mathcal{A}_G$, which, as may be recalled, incorporate analogical effects of explicit similarity. For the purpose of this chapter it is most significant that the translation reveals the characteristics of a prior probability over $\mathcal{B}$ that are responsible for the kind of analogical predictions derivable from $\mathcal{A}_G$. Eventually this leads the way to the definition of a class of priors over $\mathcal{B}$ that incorporates all possible inductive relevance relations.

Transformation rules and Jacobian. Before doing that, let me describe the transformation for the case of the predicate families Q, G and M. As for the probability function itself, we can employ the following transformation relations between the components of α and θ, which can be derived from the transformation equation (5.37):

$$\begin{aligned} \alpha_{G0} &= \theta_0 + \theta_1 \qquad \alpha_{G1} = \theta_2 + \theta_3, \\ \alpha_{G00} &= \frac{\theta_0}{\theta_0 + \theta_1} \qquad \alpha_{G01} = \frac{\theta_1}{\theta_0 + \theta_1}, \\ \alpha_{G10} &= \frac{\theta_2}{\theta_2 + \theta_3} \qquad \alpha_{G11} = \frac{\theta_3}{\theta_2 + \theta_3}. \end{aligned} \tag{5.38}$$

For any density function $p(H_\alpha)$ over A_G, we can simply write all the components of α as these fractions of components of θ. However, in order to make up for the change of the space itself, we must multiply the resulting function of components of θ with the so-called Jacobian, the determinant of the transformation matrix. This method is described in any standard textbook on vector calculus, for instance Marsden (1988).

As it turns out, it is simpler to calculate the Jacobian $J^{-1}(\alpha)$ for the inverse transformation of B to A_G first, and to derive the form of $J(\theta)$ from that. It is further simpler to employ only the three free parameter components for the space A_G:

$$\begin{aligned} \alpha_G &= \alpha_{G1} = 1 - \alpha_{G0}, \\ \alpha_{0M} &= \alpha_{01} = 1 - \alpha_{00}, \\ \alpha_{1M} &= \alpha_{11} = 1 - \alpha_{10}. \end{aligned} \tag{5.39}$$

For the space B we can simply take the θ_q with $q = 1, 2, 3$. The simpler transformation rules then are

$$\begin{aligned} \theta_1 &= (1 - \alpha_G)\alpha_{0M}, \\ \theta_2 &= \alpha_G(1 - \alpha_{1M}), \\ \theta_3 &= \alpha_G\alpha_{1M}, \end{aligned} \tag{5.40}$$

These transformations are again essentially the same as the transformation equation in (5.37).

The Jacobian $J^{-1}(\alpha)$ is now given by the determinant of the transformation matrix. The q-th row in this matrix consists of the partial derivatives of θ_q to the components of α, in the order α_G, α_{0M}, and α_{1M}. The Jacobian is thus given by

$$J^{-1}(\alpha) = \det \begin{pmatrix} -\alpha_{0M} & 1-\alpha_G & 0 \\ 1-\alpha_{1M} & 0 & -\alpha_G \\ \alpha_{1M} & 0 & \alpha_G \end{pmatrix}, \tag{5.41}$$

Writing out the determinant we find $J^{-1}(\alpha) = \alpha_G(1-\alpha_G)$. The Jacobian for a transformation from A_G to B is the inverse of this, which in terms of components of θ comes down to:

$$J(\theta) = \frac{1}{(\theta_0+\theta_1)(\theta_2+\theta_3)}. \tag{5.42}$$

This factor makes up for the change of the infinitesimal volumes $d\alpha$ into $d\theta$ during the transformation.

5.4.2 Transformed probability models

Explicit analogy in terms of Q-predicates. Recall that for the system of $\lambda\gamma$ rules, the probability over the space A_G is given by the density of equation (5.30). The density over the space B is determined by the transformation rules of equation (5.38) and the Jacobian (5.42), and thus given by

$$p(H_\theta) \sim (\theta_0+\theta_1)^{r_{01}} \times (\theta_2+\theta_3)^{r_{23}} \times \prod_{g,m} \theta_{2g+m}^{(a_{Ggm}-1)}. \tag{5.43}$$

For the exponents of the cross-terms, $r_{2g,2g+1}$, we have

$$r_{2g,2g+1} = a_{Gg} - a_{Gg0} - a_{Gg1}. \tag{5.44}$$

This prior probability over the space B results in exactly the same predictions over the Q-predicates as can be derived from the corresponding prior over the space A_G, which are expressed in the predictions (5.6).

It can be noted immediately that the prior over B deviates from a Dirichlet prior because of the cross-terms $(\theta_{2g}+\theta_{2g+1})^{r_{2g,2g+1}}$. These terms are responsible for the analogical effects in the predictions. Recall that the exponents in the above density are related to the system of prediction rules according to $\lambda_G\gamma_{Gg} = a_{Gg}$ and $\lambda_M^g = a_{Gg0} + a_{Gg1}$ with $g = 0, 1$. We can therefore write

$$r_{2g,2g+1} = \lambda_G\gamma_{Gg} - \lambda_M^g \tag{5.45}$$

Recall also that the relevance relations in the rules for explicit similarity are expressed in equations (5.8) and (5.9). The exponents of the cross-terms are thus proportional to the difference in relevance between predicates of equal gender and predicates of different gender,

$$r_{2g,2g+1} = \gamma_{Gg} N(\rho_{\bar{G}} - \rho_{Gg}), \tag{5.46}$$

where ρ_{Gg} and $\rho_{\bar{G}}$ are as indicated above. So the cross-terms in the prior probability over B have non-zero exponents precisely if there are differences in the relevances between Q-predicates of identical and different gender.

It can be seen very easily that certain systems of rules for the underlying predicate families G and M are equivalent to a single $\lambda\gamma$ rule for Q-predicates. We only need to assume $a_{Gg} = a_{Gg0} + a_{Gg1}$, or in terms of the parameters in the system of rules

$$\lambda_G \gamma_{Gg} = \lambda^g_M. \tag{5.47}$$

This choice of parameters indeed reduces the system of rules to a single $\lambda\gamma$ rule. If we identify the numbers of observations $n_G = n_Q$, $n_{Gg} = n^g_M$ and $n^g_{Mm} = n_{2g+m}$, the resulting system of rules is exactly identical to a single $\lambda\gamma$ rule with the parameters $\lambda = \lambda_G$ and $\gamma_{(2g+m)} = \gamma_{Gg}\gamma^g_{Mm}$.

The Jacobian and virtual observations. Finally, it is illustrative to connect the Jacobian determinant to the system of prediction rules for underlying predicates, in particular to the notion of virtual observations. Consider a uniform prior probability over the space B, corresponding to the exponents $\lambda\gamma_q = c_q = 1$ for all q. Sometimes these exponents are called the virtual observations of Q^q, since they are added to the number of actual observations n_q in the prediction rules. Now if we translate the uniform prior over B to a prior over A directly, the resulting exponents a_{Gg} and a_{Ggm} are all 1, so that we obtain a uniform prior again. But because of the inverse Jacobian $J^{-1}(\alpha)$, the exponents a_{Gg} are raised with 1, so that $a_{Gg} = 2$ and $a_{Ggm} = 1$, and correspondingly $p(H_\alpha) \sim \alpha_0\alpha_1$. The fact that the prior over A_G that corresponds to the uniform prior over B is not flat is thus entirely due to the Jacobian deriving from the transformation between B and A_G.

Now if we consider the exponents as resulting from virtual observations again, the correction factor given by the Jacobian may be given a very natural interpretation: the exponents must be such that all combinations of predicates G^g and M^m have exactly one virtual observation. The thing to note is that, in terms of the underlying predicate family G, virtual observations for all Q-predicates entail two observations in each predicate G^g, so that indeed we must have $a_{Gg} = 2$. More generally, we may think of the Jacobian as a function that supplements lost virtual observations after transformations of the hypotheses space. In other words, the number of virtual observations is the aspect of the prior probability that is supposed to remain intact during the transformation. This is very helpful in constructing Jacobians for more complicated parameter transformations than the one above.

5.5 General model

It is not straightforward to construct the general model from the explicit analogy models. In the first subsection, some considerations will temper the ambition of finding a general model. The second subsection, however, will develop a tentative general model, but it will also reveals a problem for this model. The last subsection speculates on a specific solution with limited parameter freedom.

5.5.1 A more modest aim

Ansatz for the general model. The preceding discussion suggests that the inductive relevance between Q^v and Q^w may be modelled by multiplying the prior probability over the partition $\mathcal{B}$ with a term $(\theta_v + \theta_w)^{r_{vw}}$. As indicated, the exponent r_{vw} expresses the difference between two relevances: the relevance between Q^v and Q^w on the one hand, and the relevance between either one of these on the one hand, and predicates Q^q with $q \neq v, w$ on the other. This suggests the following form for an overall analogy prior:

$$p(H_\theta) \sim \prod_{q<4} \theta_q^{(c_q - 1)} \prod_{v<w<4} (\theta_v + \theta_w)^{r_{vw}}. \tag{5.48}$$

The predictions are determined entirely by the exponents c_q and r_{vw}. Note that the prior can easily be generalised to settings with any number of Q-predicates.

In the Carnapian $\lambda\gamma$ rule we are able to connect the initial probabilities γ_q for $q < 4$ and a learning rate λ with the exponents of the prior probability over B. Now let us say that we are given a set of initial probabilities γ_q, a learning rate λ, and a vector of symmetric relevance relations, ρ. The challenge for the general model of analogical predictions then is to provide the exponents c_q and r_{vw} that correspond to these initial values. It can be noted immediately that there are an equal number of free components of γ, λ and ρ, namely $3 + 1 + 6 = 10$, as there are free exponents in the above probability density. This suggests that there is indeed a unique solution for the representation problem. However, as will become clear, a complete representation is too much to ask for within the context of the present chapter.

Reasons for modesty. This section has a more modest aim: it presents a prior of the above form that on certain assumptions incorporates a vector of relevance relations and a learning rate, assuming a kind of initial symmetry between the Q-predicates. This modesty is motivated by a number of reasons. First, the procedure for incorporating relevance relations cannot be generalised in any straightforward way from the model for explicit similarity. We have to make assumptions to pin down the relations between the relevances and the prior probability. Second, it is very difficult to derive an analytic expression for the initial probabilities γ_q from the analogy prior over $\mathcal{B}$. Therefore we cannot directly control the initial probabilities implicit in the general model. Third, the learning rate λ turns out to have a different role in the general model. And finally, the present model only deals with the relations between prior and relevance relations for the specific case of the party centre example. This will suggest some general guidelines for determining the exponents c_q and r_{vw} starting from a general relevance vector ρ, but full generality is not achieved.

5.5.2 Towards a general model

Encoding relevance relations. With this aim in mind, consider the relevance relations of the example on party centres. With G, M and W referring to underlying predicate families on gender, marital status and traditional wedding rings respectively, we can denote the components of the vector of equation (5.3) in the following way:

$$\rho = \langle \rho_{G0}, \rho_{G1}, \rho_{M0}, \rho_{M1}, \rho_{W0}, \rho_{W1} \rangle. \tag{5.49}$$

Here ρ_{Gg} refers to the components $\rho(2g, 2g+1)$, since Q^{2g} and Q^{2g+1} have the underlying predicate G^g in common. Other components of ρ in equation (5.3) may be explicated in similar fashion: ρ_{Mm} refers to $\rho(m, 2+m)$ and ρ_{Ww} to $\rho(1-w, 2+w)$.

Recall the basic pattern for relevance relations that derive from the partition $\mathcal{A}_G$. The relevances ρ_{G0} and ρ_{G1} can be determined separately, while the other relevances, ρ_{Mm} and ρ_{Ww} for $m, w = 0, 1$, may be fixed on some average value. As before I denote this latter average value with $\rho_{\bar{G}}$, while $\rho_G = (\rho_{G0} + \rho_{G1})/2$. Similar configurations of relevance relations may be derived from the partitions $\mathcal{A}_M$ and $\mathcal{A}_W$, enabling us to independently determine ρ_{M0} and ρ_{M1}, or ρ_{W0} and ρ_{W1}, and determine the average values $\rho_{\bar{M}}$ and $\rho_{\bar{W}}$. Now to incorporate the combination of these relevance relations into a single prior over $\mathcal{B}$, we must imagine that the single prior results from the priors over the analogy partitions, which each incorporate specific aspects of the relevance vector. The exponents

c_q in the prior thus relate to the priors over all three analogy partitions. The exponents r_{vw}, on the other hand, are related only to the priors over the analogy partitions associated with the combination of v and w.

Reducing the number of free parameters. It seems natural to combine the three models for explicit similarity by multiplying the priors over $\mathcal{B}$ corresponding to these models. This means that the exponents a_{Ggm}, a_{Mmw} and a_{Wwg} sum up to c_q. The sum of the relevance vectors for the explicit similarity models may then serve as the relevance vector associated with the resulting prior. Alternatively, we may consider the product of the relevance vectors. However, all such straightforward combination procedures result in poor representations: different general analogy models are connected to the same analogical prior, and any analogical prior may be read as the result of a multitude of relevance vectors. We have too much freedom in choosing the exponents a_{Ggm}, a_{Mmw} and a_{Wwg} on the basis of the values for c_q and r_{vw}.

The following employs a combination procedure in which the exponents c_q are not taken as built up from the exponents a_{Ggm}, a_{Mmw} and a_{Wwg} separately, but in which these latter exponents are each taken to be equal to the exponents c_q. Every pair from the predicates G^g, M^m and W^w determines a unique predicate q in the family Q. With some algebra we can obtain the relations

$$a_{Ggm} = c_{2g+m}, \qquad a_{Mmw} = c_{2-m-2w+4wm}, \qquad a_{Wwg} = c_{1-w+g+2wg}. \tag{5.50}$$

So while we were considering 4 independent exponents in all three analogy partitions, these exponents must all coincide with the same set of 4 exponents c_q. The priors over the analogy partitions that may be taken to underly the prior over $\mathcal{B}$ are thus limited in a specific way: their exponents must conform to the corresponding values of the c_q. The number of free exponents in the models for explicit similarity is thus reduced, so that the representation problems of the preceding paragraph are avoided.

Combining the explicit similarity models. Now consider the relation between the exponents in the analogy priors and the exponents r_{vw} in the prior over $\mathcal{B}$. It can be noted that the values of a_{Gg}, a_{Mm} and a_{Ww} are implicit to the values of c_q and r_{vw} in an analogical prior over $\mathcal{B}$, according to the relations (5.50) and the further relations

$$\begin{aligned} r_{2g,2g+1} &= a_{Gg} - a_{Gg0} - a_{Gg1}, && (5.51)\\ r_{m,2+m} &= a_{Mm} - a_{Mm0} - a_{Mm1}, && (5.52)\\ r_{1-w,2+w} &= a_{Ww} - a_{Ww0} - a_{Ww1}. && (5.53) \end{aligned}$$

This means that we may encode the values of a_{Gg}, a_{Mm} and a_{Ww} in a prior over $\mathcal{B}$ relative to a choice for the exponents a_{Ggm}, a_{Mmw} and a_{Wwg}.

The above equations specify how the exponents in the prior over $\mathcal{B}$ relate to the exponents in the priors over the analogy partitions. We can now concentrate on the function of the exponents in the priors over the analogy partitions in expressing relevance relations. First consider a_{Gg}, a_{Mm} and a_{Ww}. Their values are directly related to the average relevances $\rho_{\bar{G}}$, $\rho_{\bar{M}}$ and $\rho_{\bar{W}}$ according to

$$\begin{aligned}
\rho_{\bar{G}} &= a_{G0} + a_{G1}, \\
\rho_{\bar{M}} &= a_{M0} + a_{M1}, \\
\rho_{\bar{W}} &= a_{W0} + a_{W1}.
\end{aligned} \tag{5.54}$$

Any vector of relevance relations ρ thus determines the values of the sums on the right side of the equations (5.54). Moreover, by fixing these sums of exponents we also encode the given relevance vector in the prior up to averages for the pairs of relevance relations, because we can write

$$\begin{aligned}
\rho_{G} &= \rho_{\bar{M}} + \rho_{\bar{W}} - \rho_{\bar{G}}, \\
\rho_{M} &= \rho_{\bar{W}} + \rho_{\bar{G}} - \rho_{\bar{M}}, \\
\rho_{W} &= \rho_{\bar{G}} + \rho_{\bar{M}} - \rho_{\bar{W}}.
\end{aligned} \tag{5.55}$$

Thus the average size of the relevance among predicate pairs is implicit to the pairwise sums of the exponents a_{Gg}, a_{Mm} and a_{Ww}.

Overspecification. Unfortunately, if we combine the priors of the analogy partitions into a single prior over $\mathcal{B}$ we run into a problem. As indicated, the size of ρ_G is encoded in the sizes of pairwise sums of exponents, to wit $a_{G0} + a_{G1}$, $a_{M0} + a_{M1}$ and $a_{W0} + a_{W1}$. But if we follow the relations that hold in the model for explicit similarity, the sizes of ρ_{G0} and ρ_{G1}, and thereby of ρ_G, are also determined by equation (5.8). With equations (5.34) and (5.35) these are in turn determined by the exponents c_q, and the ratios ${}^{a_{G0}}/_{(a_{G0}+a_{G1})}$ and ${}^{a_{G0}}/_{(a_{G0}+a_{G1})}$. The thing to note is that these exponents and ratios are independent of the size of $a_{G0} + a_{G1}$, or of any other such pairwise sum. So the size of ρ_G seems to be doubly encoded in the combined analogy prior over $\mathcal{B}$: once by the pairwise sums, and once by the exponents and ratios. The same can be said of the average relevances ρ_M and ρ_W. So the prior cannot encode all the relevances that we wish, since there are too few free parameters left once they are restricted to agree on the sizes of the average relevances.

In other words, restriction (5.50) leads us into a problem after all. The number of free exponents in the separate analogy models matches the number of free exponents in the combined model, but somehow the restrictions chosen do not allow us to straightforwardly generalise the separate analogy models.

5.5.3 Tentative solution

Selective generalisation. The above combination procedure need not be considered as a complete failure. A more constructive reading is that the general model is not in all aspects a generalisation of the model for explicit similarity, and that only specific aspects of the explicit similarity model can be taken over into the general one. The following exposition runs along this line. It simply assumes that the pairwise sums of exponents fix the pairwise averages of the relevances according to equations (5.54) and (5.55). With this kept fixed, the analogy prior incorporates initial probabilities and learning rates as well as possible. After that, the remaining parameter freedom is used to pin down the differences between the pairs ρ_{Gg}, ρ_{Mm}, and ρ_{Ww}.

It must be stressed that this subsection is rather speculative. Within the limits set by the assumption of equations (5.54) and (5.55), it presents a model that is constructed by playing with the analogy priors and predictions in the program Mathematica™. Specifically, I have looked at the influence of varying exponents in the prior on the predictions for a number of numerical examples. The examples of the next section may give some justification for the eventual model, but I am convinced that a derivation of some general model is possible. I have unfortunately not been able to find it.

Constructing an analogy prior. As for the values of γ_q and λ in relation to the analogical prior over $\mathcal{B}$, I will make two simplifying assumptions. First, we may assume that the values of the exponents a_{Gg}, a_{Mm} and a_{Ww} are pairwise identical:

$$a_{G0} = a_{G1}, \qquad a_{M0} = a_{M1}, \qquad a_{W0} = a_{W1}. \tag{5.56}$$

With equation (5.54) and assumption (5.56), the exponents a_{Gg}, a_{Mm} and a_{Ww} are determined completely. The idea behind the assumption is that we force the resulting initial predictions to be at least close to symmetric. Specifically, in all the priors over analogy partitions that may underly the prior over $\mathcal{B}$, the exponents that concern the leading predicate are equal. At the end of the next section I return to the approximated initial symmetry resulting from that assumption.

The second assumption concerns the learning rate λ. As in the Carnapian framework, it is given by the total number of virtual observations:

$$\sum_{q} c_q = \lambda. \tag{5.57}$$

In the analogy model, the value of λ is related to the rate at which the size of the analogy effects diminish over time. But it does not straightforwardly select a specific learning rate. There is a sense in which the learning rate λ is also an expression of relevance. The larger λ is, the slower the prediction of an observation of Q^w diminishes with the observation of Q^v. It therefore seems natural to choose

$$\lambda = \frac{\rho_G + \rho_M + \rho_W}{3}. \tag{5.58}$$

At the end of the next section I shall return to the exact function of λ in the analogy prior, and consider variations on the value assumed here.

Certain aspects of the relevance relations remain to be captured in the prior probability assignment, namely the differences between the relevances within the pairs, such as between ρ_{G0} and ρ_{G1}. Further, the only freedom left in the parameters of the prior is in the division of the total number of virtual instances λ over the separate c_q. Note that we must fix three such differences, for which we have exactly three free parameters available. The above discussion makes clear that we cannot simply employ the relation (5.46) to fix the ρ_{Gg} separately. However, we may be able to employ this relation to connect the difference between the relevances ρ_{G0} and ρ_{G1} to the exponents c_q. The idea here is that instead of carrying over the relation (5.46) into the general model, we may be able to carry over a weaker relation that can be derived from it.

Taking the difference between ρ_{G0} and ρ_{G1} as the example case, and using the fact that $a_{G0} = a_{G1}$, we can write

$$\begin{aligned}\rho_{G0} - \rho_{G1} &= 2(r_{23} - r_{01}) \\ &= 2(c_2 + c_3 - c_0 - c_1),\end{aligned} \tag{5.59}$$

which exactly meets this desideratum. In the same way we can write for the other two pairs of relevances

$$\rho_{M0} - \rho_{M1} = 2(c_1 + c_3 - c_0 - c_2), \tag{5.60}$$

$$\rho_{W0} - \rho_{W1} = 2(c_0 + c_3 - c_1 - c_2). \tag{5.61}$$

With these three relations, we have completely fixed the values for the exponents c_q. And because of the relations (5.51), (5.52) and (5.53), we thereby also fix the values for the exponents r_{vw}.

Overview of the model. We have now completed the construction of the analogy prior on the basis of a vector of relevance relations, initial symmetry and a learning rate. Let me summarise the procedure. If we are given a vector of relevance relations ρ, we first calculate the averages ρ_G, ρ_M and ρ_W. With these averages, equations (5.55) and (5.54), and assumption (5.56) we can then determine the exponents a_{Gg}, a_{Mm} and a_{Ww}. The assumption ensures approximate initial symmetry. With assumption (5.58) and equations (5.59) to (5.61) we can subsequently determine the values for the exponents c_q. Here the assumption relates to the rate at which the analogy effects diminish. Finally, using equations (5.50) to (5.53) we may finally fix the exponents r_{vw}.

Before investigating some properties of the proposed model, it is important to stress again that this model is in many ways incomplete. For one thing, I have not proposed any relation between relevance relations and analogy priors for cases in which the aforementioned assumptions are violated. Because of this, many analogy priors cannot be linked to a relevance vector. However, I must leave a more complete model to future research.

5.6 Qualitative and numerical characterisations

In this section I describe the general analogy model further. First I discuss the analogy priors by considering their form on an analogy partition, after which I can relate the priors to so-called hyper-Carnapian models of analogical reasoning. Finally I provide some numerical examples of analogical predictions.

5.6.1 The form of the analogy priors

Nonfactorisable priors. It is instructive to consider the general analogy prior in its functional form over one of the analogy partitions, for example over $\mathcal{A}_G$. Recall that this partition falls into three orthogonal subpartitions: one concerning the predicate family G, associated with the exponents a_{Gg}, and two concerning the predicate family M conditional on G^0 and G^1, associated with the exponents a_{G0m} and a_{G1m} respectively. Analogy effects between the predicates Q^{2g} and Q^{2g+1} may then be captured by Dirichlet priors over the separate subpartitions: by choosing $a_{Gg0} + a_{Gg1}$ larger or smaller than a_{Gg} we can make the relevance between Q^{2g} and Q^{2g+1} larger or smaller than the relevance between pairs of predicates that do not have the gender in common. These differences between the Dirichlet priors correspond with the terms $(\theta_{2g} + \theta_{2g+1})^{r_{2g,2g+1}}$ in the general analogy prior. If the latter are the only analogical terms in the prior over $\mathcal{B}$, the prior can be factorised into separate functions over the orthogonal subpartitions of $\mathcal{A}_G$, and thus can be dealt with completely independently.

Now imagine that, starting from a product of Dirichlet priors over $\mathcal{A}_G$, we want to express additional relevance relations between predicates of different gender. Intuitively, we want the prior probability over $\mathcal{A}_G$ to be such that, if we update the probability over the hypotheses concerning the family M conditional on G^0, we implicitly update the probability over the hypotheses concerning the family M conditional on G^1, and vice versa. In other words, we must relate the independent Dirichlet parts of the prior probability. The thing to note is that such relations are exactly realized by the additional terms in the prior over the analogy partition corresponding to the terms $(\theta_v + \theta_w)^{r_{vw}}$ with $v = 0, 1$ and $w = 2, 3$. As an example, consider a relevance relation between Q^0 and Q^3. This is associated with the term $(\theta_1 + \theta_3)^{r_{13}}$ in the general analogy prior. The translated term, $(\alpha_{G0}\alpha_{G01} + \alpha_{G1}\alpha_{G11})^{r_{13}}$, relates the priors over the subpartitions in the appropriate way. Note finally that because of such terms, the resulting prior cannot be factorised into separate functions over orthogonal subpartitions anymore.

Hyper-Carnapian rules. There is yet another way to illustrate the role of the analogical terms in the general model, which connects to the hyper-Carnapian analogical prediction rules of Skyrms (1993) and Festa (1997). Leaving aside the underlying relevance vector and the initial probabilities for the moment, consider the role of the analogy term in the following prior:

$$\begin{aligned} p(H_\theta) &\sim \theta_0\theta_1\theta_2\theta_3(\theta_2+\theta_3) \\ &= \theta_0\theta_1\theta_2^2\theta_3 + \theta_0\theta_1\theta_2\theta_3^2. \end{aligned} \tag{5.62}$$

This prior consists of two parts, which can each be associated with a $\lambda\gamma$ prediction rule $pr_{\lambda\gamma}(n_{Qq}, n_Q)$ of equation (5.18). Both these rules have $\lambda = 9$, but the initial probabilities γ_q vary. For the first term we have $\gamma_2 = 1/3$ while $\gamma_q = 2/9$ for $q \neq 2$, while the second term entails $\gamma'_3 = 1/3$ while $\gamma'_q = 2/9$ for $q \neq 3$. The predictions resulting from the above prior are therefore a mixture of two $\lambda\gamma$ rules, one using γ_q and one γ'_q.

Such mixtures are called hyper-Carnapian prediction rules. For the predictions generated by the above hyper-Carnapian rule we can write

$$p(Q^q_{n+1}|E_n) = p(H_{\lambda\gamma}|E_n)pr_{\lambda\gamma}(n_{Qq}, n_Q) + p(H_{\lambda\gamma'}|E_n)pr_{\lambda\gamma'}(n_{Qq}, n_Q), \tag{5.63}$$

where the hypotheses $H_{\lambda\gamma}$ have likelihoods given by the rules $pr_{\lambda\gamma}$. It can now be seen that the hyper-Carnapian rules capture analogical considerations. On the observation Q^3, the rule is adapted in two ways: first the two $\lambda\gamma$ rules are adapted to enhance the probability for future observations of Q^3, but the update over the hypotheses also enhances the probability of the rule for which Q^3 has a higher initial probability. This latter update does not affect the probabilities for future observations of Q^0 or Q^1, but it does lower the probability for Q^2 to raise that of Q^3. This is exactly the kind of analogy effect that may be expected from the term $(\theta_2 + \theta_3)$ in an analogy prior.

5.6.2 Numerical examples

Predictions deriving from the general analogy prior cannot usually be written down in any simple analytic form. While positive exponents r_{vw} can still be captured in terms of extended hyper-Carnapian rules, negative analogy exponents seem to make analytic expressions impossible. The remainder of this section is concerned with two examples of general analogical predictions, making use of the numerical integration module of Mathematica™. I present the examples to show that the predictions generated by the general model agree with the predictions that can be expected on the basis of the relevance relations. For the sake of easy calculations I focus on examples with natural numbers.

Symmetric relevance. The first of these examples is the simpler one, as it has some inherent symmetries. Consider the following relevance vector:

$$\rho = \langle \rho_{G0}, \rho_{G1}, \rho_{M0}, \rho_{M1}, \rho_{W0}, \rho_{W1} \rangle = \langle 28, 8, 22, 22, 14, 14 \rangle. \tag{5.64}$$

To calculate the exponents of the analogy partition, note that $\rho_{\bar{G}} = 18$, $\rho_{\bar{m}} = 16$ and $\rho_{\bar{W}} = 20$, so that $a_{Gg} = 9$, $a_{Mm} = 8$ and $a_{Ww} = 10$. Furthermore, note that $\sum_q c_q = 18$, and $c_0 + c_2 - c_1 - c_3 = c_1 + c_2 - c_0 - c_3 = 0$, while $c_0 + c_1 - c_2 - c_3 = 10$. It follows that $c_0 = c_1 = 7$ and $c_2 = c_3 = 2$, and with that it follows that $r_{01} = -5$, $r_{23} = 5$, $r_{02} = r_{13} = -1$ and $r_{12} = r_{03} = 1$. The analogy prior is thus completely specified. This prior leads to the predictions given in the table. To illustrate the analogy effects, the table shows the initial probabilities, and the predictions after 10 observations of the same predicate Q^q for each q. I abbreviate $E^q_{10} = Q^q_1 \cap Q^q_2 \cap \ldots \cap Q^q_{10}$.

Predicate q	0	1	2	3
$p(Q^q_1)$	0.250	0.250	0.250	0.250
$\rho(0,q)$	-	28	22	14
$p(Q^q_{11}\|E^0_{10})$	0.482	0.197	0.175	0.146
$\rho(1,q)$	28	-	14	22
$p(Q^q_{11}\|E^1_{10})$	0.197	0.482	0.146	0.175
$\rho(2,q)$	22	14	-	8
$p(Q^q_{11}\|E^2_{10})$	0.171	0.151	0.583	0.096
$\rho(3,q)$	14	22	8	-
$p(Q^q_{11}\|E^3_{10})$	0.151	0.171	0.096	0.583

A number of things may be remarked on these results. First, the initial probabilities are only approximately symmetric. Rounded off they are the same, but they differ from each other at the fifth decimal. The analogy terms associated with G, to wit $(\theta_{2g} + \theta_{2g+1})^{r_{2g,2g+1}}$, normally cancel the differences between the exponents $c_{2g} + c_{2g+1}$ for $g = 0, 1$. But the analogy terms related to M and W interfere with these terms and cause minor imbalances. Second, it is notable that the predictions respect the ordering of the relevance relations. Moreover, the predicates Q^0 and Q^1 gain less probability from their own occurrences than the predicates Q^2 and Q^3, which is in line with the fact that the latter are on average less relevant to other predicates than the former. Unfortunately, the differences between the predictions are not in any way linear in the differences in relevance. Third, there is perfect symmetry between predicates of equal gender, Q^{2g} and Q^{2g+1}. This can be seen from the fact that the prior probability is invariant under permutation of these predicates. It is therefore not surprising that the predictions of the first and the second two lines are identical up to this permutation.

No symmetries. The second example breaks with this symmetry. Consider the following vector of relevance relations:

$$\rho = \langle 28, 8, 20, 24, 16, 12 \rangle. \tag{5.65}$$

This example also has $\rho_{\bar{G}} = 18$, $\rho_{\bar{m}} = 16$ and $\rho_{\bar{W}} = 20$, so that again $a_{Gg} = 9$, $a_{Mm} = 8$, $a_{Ww} = 10$ and $\sum_q c_q = 18$. But the example further has $c_1 + c_3 - c_0 - c_2 = 2$, $c_1 + c_2 - c_0 - c_3 = 2$, while again $c_0 + c_1 - c_2 - c_3 = 10$. It follows that $c_0 = 6$, $c_1 = 8$ and $c_2 = c_3 = 2$, and with that it follows that $r_{01} = -5$, $r_{23} = 5$, $r_{02} = 0$, $r_{13} = -2$ and $r_{12} = 2$, $r_{03} = 0$. This completely specifies the analogy prior, and we may again consider the predictions after E^q_{10}.

Predicate q	0	1	2	3
$p(Q^q_1)$	0.252	0.248	0.250	0.250
$\rho(0,q)$	-	28	20	12
$p(Q^q_{11}\|E^0_{10})$	0.499	0.193	0.169	0.139
$\rho(1,q)$	28	-	16	24
$p(Q^q_{11}\|E^1_{10})$	0.199	0.466	0.154	0.181
$\rho(2,q)$	20	16	-	8
$p(Q^q_{11}\|E^2_{10})$	0.168	0.154	0.582	0.096
$\rho(3,q)$	12	24	8	-
$p(Q^q_{11}\|E^3_{10})$	0.147	0.172	0.097	0.584

More or less the same remarks can be made on these results. The main thing is that the predictions still respect the ordering in the relevance relations, and that the probability that predicates gain from their own occurrence is still dependent on the average relevances. Note further that the symmetry in the initial probabilities is more disturbed than in the first example. Finally, note that the symmetry between predicates of equal gender is broken. With this relevance vector, Q^0 is on average less relevant to the other predicates than Q^1. In line with this, the predicate Q^0 gains more probability from its own occurrence than Q^1. This effect is almost absent for Q^2 and Q^3, but again the model is not exactly correct.

5.7 Other models of analogical predictions

Let me briefly compare the resulting models for analogical predictions with some other models in the literature. First I relate the present model to a number of alternative prediction rules from Carnap-Hintikka inductive logic. After that I consider the hyper-Carnapian systems by Skyrms and Festa, and finally I turn to the model of Maher.

Carnap-Hintikka inductive logic. The general aim in Carnap-Hintikka inductive logic is to derive a class of prediction rules from a number of natural assumptions or principles. One of these principles then is an expression of analogy by similarity, other principles may be regularity, exchangeability, the convergence to relative frequencies, initial symmetry with respect to predicates, positive probability for confirmed universal generalisations, and instantial relevance. Combinations of these principles are employed in the derivation of classes of analogical prediction rules.

It is instructive to position the present models in terms of these principles. First, the above model trivially satisfies regularity: none of the finite sequences of observations are deemed impossible from the onset. Second, the predictions deriving from the general model are by definition exchangeable, and therefore show convergence to eventual relative frequencies in the sequence of observations, if there are any. This can be seen from the fact that the models are defined by probability assignments over the partition $\mathcal{B}$, and by the standard convergence results derived for this partition. Third, initial symmetry of predicates can in principle be obtained by choosing the analogy prior over $\mathcal{B}$ appropriately. But as we have seen in the foregoing, encoding both a vector of relevance relations and initial symmetry in the analogy prior is not straightforward. Finally, the above model does not give positive probability to universal generalisations. Hypotheses H_α have infinitesimal probability, since the probability is always distributed over a continuum of hypotheses $\mathcal{A}$. The infinitesimal probability is thus also assigned to those H_α in which one or more components of α are extremal. However, there is a rather natural extension of the above schemes in which these sets are given strictly positive measure. Universal generalisations are certainly not excluded by the above models.

The principle of instantial relevance must be given separate attention. It is noteworthy that instantial relevance need not always be satisfied for predictions resulting from a prior over $\mathcal{B}$. This principle may be violated exactly because the prior probability over the hypotheses space A need not be factorisable into independent marginals. In terms of the example, we may consider the presence of wives much more probable than that of maidens if there are very few women in the party centre, while we may consider the presence of maidens much more probable than that of wives in the case that there are very few men. Now, observing a wife in the party centre has a combined effect: first of all it makes the presence of women more probable, and within the group of women it shifts the probability from maidens to wives. However, it may happen so that the former effect is much stronger than the second: after a single woman we hardly expect any further men. But because maidens are considered much more probable than wives if there are hardly any men, the probability of maidens may eventually benefit more from observing a wife than the probability of wives itself. It is not difficult to construct the prior that encodes the above effects numerically. However, I have not been able to check whether there are such priors in the class of analogy priors defined above. Moreover, because predictions defined on $\mathcal{B}$ always converge to the correct relative frequencies, the effect sketched above is necessarily a short term one.

Finally, consider the relation between the analogy model and Carnap-Hintikka inductive logic more generally. Recall that a special case of the analogy model is presented by a system of $\lambda\gamma$ rules that models explicit similarity effects. In turn, these systems of prediction rules have the single Carnapian $\lambda\gamma$ rule as special case. The model may therefore be considered as an extension of Carnap-Hintikka inductive logic. However, in contrast to the direct prediction rules of this logic, the predictions of the general analogy model can only be arrived at by numerical approximation of an integration over statistical hypotheses. Furthermore, while the model consists of a class of analogy priors, there is no claim that this class is somehow the definitive explication of analogical reasoning. In these ways the general analogy model falls outside Carnap-Hintikka logic.

Skyrms and Festa. Let me now turn to the models for analogy reasoning by Skyrms and Festa, which employ hyper-Carnapian prediction rules. The idea of such rules originates from Skyrms, but Festa explores the rules further to define a proper class of analogical prediction rules. The standard illustration involves a wheel of fortune with four equally large segments, labelled with the four quarters of the compass. At every turn in a direction chosen at random, the chance of stopping at some segment is unknown but constant. It is further given that the axis of the wheel is slightly eccentric. Finding the wheel in some segment will favour this segment in next predictions, but it will on the whole also favour the two neighbouring segments in comparison to the opposing segment. The hyper-Carnapian prediction rule proposed for this is a mixture of four $\lambda\gamma$ rules. The rules have equal λs, and for each of them the γs of the segments are chosen as $\frac{1}{2}$ for the segment favoured by the bias, $\frac{1}{5}$ for the neighbouring segments, and $\frac{1}{10}$ for the opposing one. The rules differ in that each of them has its own favoured segment. As it turns out, the resulting predictions then show similarity effects between all pairs of neighbouring segments. That is, if we find an instance of north, east and west are favoured more than south, and so on.

The hyper-Carnapian models are similar to, but also different from the present analogy models. They provide alternative ways for defining analogy priors over the partition $\mathcal{B}$, and in the specific case in which the analogy terms have positive exponents, the present analogy model is also a kind of hyper-Carnapian model. However, this correspondence fails for priors with negative analogy exponents r_{vw}. Furthermore, I see a difficulty in making sense of the prior probability proposed by hyper-Carnapian rules, which is related to the difficulties noted in Maher (2000, 2001). On the partition $\mathcal{B}$, hyper-Carnapian rules are defined as mixtures of Dirichlet priors. After an observation we must update every Dirichlet distribution separately, and apart from that we must adapt the weights assigned to these different distributions according to the predictions that the separate distributions generate. But assigning probabilities to Dirichlet priors seems rather unnatural. Such probabilities cannot be interpreted as probabilities over hypotheses, but must really be seen as probabilities assigned to different priors over the hypotheses, that is, as a kind of second-order probability. And if we can also define analogy priors by means of a single probability function over one space of hypotheses, introducing such higher order probabilities seems a high price to pay.

The model of Maher. Finally, some attention must be given to the model for analogical predictions proposed by Maher. This model is again similar to the present model in important respects. It generates predictions on Q-predicates by defining an analogy prior over the partition $\mathcal{B}$, and moreover, it employs underlying predicates such as G and M in the definition of this prior. A drawback is that the model of Maher is limited to two underlying predicate families. Maher uses these families for defining a set of hypotheses within the partition $\mathcal{B}$ for which the families are statistically independent. While a Dirichlet distribution over $\mathcal{B}$ assigns zero probability to this set, Maher assigns a positive probability to it. He shows that conditional on the independence, the predictions on Q-predicates can be represented as a product of $\lambda\gamma$ rules for the underlying predicates. He further derives a single $\lambda\gamma$ rule conditional on the dependence of the underlying predicates. The prediction rule generated by the combined prior over $\mathcal{B}$ is a mixture of this single $\lambda\gamma$ rule and the product of the two $\lambda\gamma$ rules for the underlying predicates.

As said, the model of this chapter has a lot in common with this model. However, the model of Maher does not employ the possibilities with underlying predicates completely. It uses predictions concerning such predicates on the condition that they are independent. By contrast, the model for explicit analogy also employs hypotheses concerning underlying predicates on the condition that these predicates are statistically dependent. The present model generalises this to incorporate dependencies between all three predicate families that may underlie the four Q-predicates. Moreover, there are no principal problems with defining analogy priors for predictions on larger numbers of Q-predicates.

Advantages of the present model. Let me emphasise some of the advantages of the models presented in this chapter, when compared with other models in the literature. First, the models of this chapter show analogical predictions as the result of a Bayesian scheme using hypotheses. As argued in the first part of this thesis, the Bayesian scheme ensures that the predictions are valid, and clearly reveals the inductive assumptions underlying these predictions. Second, and as shown in the preceding chapter, the present models allow for generalisations to more predicate families. Third, and as elaborated in this chapter, the models provide access to the inductive relevance vector that is inherent to a prior probability assignment over the algebra. This latter feature is also present in Kuipers (1984), but both Festa (1997) and Maher (2000) fail to provide any such connection between analogical prediction rules and relevances.

5.8 Conclusion

Summary and moral. This chapter presents a Bayesian model for exchangeable analogical predictions. First relevance relations were characterised, and related to the models for explicit similarity of the preceding chapter. Then the chapter presented a scheme that employs hypotheses on Q-predicates for generating predictions. Exchangeable predictions for Q-predicates were represented with a prior over the partition $\mathcal{B}$. It was shown how the Q-predicates may be translated in combinations of underlying predicates G, M and W. The partitions of hypotheses concerning these predicates, denoted $\mathcal{A}_G$, $\mathcal{A}_M$ and $\mathcal{A}_W$, were seen to be equivalent to the original partition $\mathcal{B}$. Dirichlet priors over the separate parts of these partitions capture the exchangeable analogical predictions for explicit similarity. Transforming these priors back to the partition $\mathcal{B}$ suggested a form for a general analogy prior over this latter partition. Finally, the chapter proposed a relation between relevance relations and this prior on the basis of the relevance relations expressed in the explicit similarity model.

In closing, let me draw a general moral from the above models. It is that some of the problems in defining analogical predictions within Carnap-Hintikka inductive logic may be solved by shifting perspective twice. The first shift concerns the use of statistical hypotheses, and the explicit use of Bayes' rule in accommodating observations. That is, we represent prediction rules with a Bayesian update over hypotheses. In this perspective, exchangeable rules can be characterised with a prior over the partition $\mathcal{B}$. But the priors that lead to analogical predictions are hard to define in the parameter space associated with $\mathcal{B}$. In the second shift, this difficulty is solved by transforming the partition $\mathcal{B}$ into the analogical partitions, which have a differently structured parameter space but are otherwise equivalent. These parameter spaces supply the conceptual means for defining priors that lead to analogical predictions.

Bayesian logic and the Carnapian programme. Let me concentrate on the first perspectival shift. It concerns the logical framework of analogical predictions. It illustrates that Carnap-Hintikka inductive logic can be viewed as part of a wider logic of inductive Bayesian inference, as developed in chapters 1 to 3. In this logic of inductive inference, validity is determined exclusively by the probability axioms, which here include Bayesian conditioning. Further, the inductive relevance of observations for each other is not inherent to the choice of language and the assumption of some further principles. Instead the inductive relevance is inherent to a partitioning of the observation algebra into statistical hypotheses and a prior probability assignment over this partition. If we are given the observational algebra, we have complete freedom in choosing these inductive assumptions. I believe that both the expression of inductive assumptions in a partition of hypotheses and the neutral way of incorporating observations into the inductive methods present valuable conceptual advantages.

While this offers a useful perspective on analogical predictions, I am aware that it also masks one of the intentions of Carnap-Hintikka inductive logic. This intention is to present a normative theory of inductive predictions from first principles. If this intention is applied to the problem of analogical predictions, it is to provide a class of rules resulting in predictions that conform to a certain characterisation of analogy, and that are logically valid. By contrast, this chapter only presents some examples of Bayesian models for analogical predictions. For those who do not share the intentions of Carnap-Hintikka inductive logic, the examples may already suffice: they are exemplars for models for analogical predictions. But from the standpoint of Carnap-Hintikka inductive logic itself, the above examples can perhaps best be taken as providing a framework and some starting points for a further normative discussion.

Using transformations between partitions. Concentrating on the second shift, note first that the use of underlying predicates follows quite naturally from the model for explicit similarity, as given in the preceding chapter. It is perhaps hard to come up with a transformation of the partition $\mathcal{B}$ into, for example, $\mathcal{A}_G$ if there is no independent reason for thinking of the partition $\mathcal{A}_G$ in the first place. On the other hand, it is strictly speaking inessential what the partition on underlying predicate families refers to. The transformation procedure from the predicate Q to underlying predicates can simply be taken as a formal tool for expressing relations between the Q-predicates. Similarly, we do not need a natural and independent description of resulting Q-predicates in order to employ explicit similarity relations in predictions over underlying predicates.

Another aspect of the second perspectival shift is more important for the general line of this thesis. In chapter 3 I have argued that the Bayesian scheme offers a better control over the inductive assumptions inherent to inductive predictions, by linking these assumptions to the choice of a partition. But this chapter and the preceding one reveal another advantage of the Bayesian scheme. It is that the scheme, once the partition has been chosen, allows us to express further inductive assumptions in a specific prior probability function over the partition. And for this we are free to transform the partition in such a way that the function becomes more easily accessible. Thus, not only the choice of a partition is a tool for making inductive assumptions, the partition itself is also a tool in defining a prior over the partition, which may express further inductive assumptions. I refer to chapter 9 for a further discussion of this idea.

6

Inductive Inference for Bayesian Networks

This chapter provides a scheme for inductive inferences concerning exchangeable observations of variables in a given Bayesian network. It presents the tools for determining the probabilities and conditional probabilities associated with the nodes and edges of a given network on the basis of these observations. It further offers simple expressions for predictions over the variables, relative to some assumptions on the observations and the dependency structure laid down in the network. Finally, the chapter signals a specific problem with representing dependencies in a Bayesian network. This leads to a distinction between so-called causal and inductive independence.

This chapter can be read independently of all preceding chapters. However, there is an intimate connection between this chapter and the two chapters on analogical predictions. Next to a model for predictions on Bayesian networks, this chapter provides a general treatment of the mathematics underlying the use of hypotheses in making analogical predictions. The resulting models for Bayesian networks are structurally similar to the models for analogical predictions. The difference is that here the observation algebra using underlying predicates is generalised, and the hypotheses are defined in terms of the selection functions of chapter 2. This chapter thus provides a deeper and more general underpinning for the models of analogical predictions.

6.1 Induction and Bayesian networks

Bayesian networks. A Bayesian network is a convenient tool in representing a multivariate probability distribution. Consider a range of variables R_k with $k = 1, 2, \ldots, n$, and a probability distribution $p(R_1, R_2, \ldots, R_n)$. Each variable R_k may be assigned values r_k, and for simplicity these values are binary numbers, $r_k = 0, 1$. A complete representation of the probability distribution over the variables is then given by a number of 2^n probabilities $p(r_1, r_2, \ldots, r_n)$, under the restriction that these probabilities sum to 1. Marginal probabilities like $p(R_k)$ and marginal conditional probabilities such as $p(R_k|R_{k'})$ can all be derived from this complete representation. However, as the number of independent probabilities grows exponentially with the number of variables, the representation of the probability distribution is rather cumbersome.

It may happen that we know of specific independencies between variables. For example, we may know that

$$p(R_1|R_2, \ldots, R_n) = p(R_1|R_2) \tag{6.1}$$

This means that if we know r_2, we do not learn anything more on r_1 by also learning r_k for any $k > 2$. Note that this is not to say that $p(R_1|R_k) = p(R_1)$ for $2 < k \leq n$, that is, R_1 is not necessarily independent of all other R_k. It may be that R_k contains information on R_2, and thus implicit information on R_1. However, if we also know that

$$p(R_k|R_1, \ldots, R_{k-1}, R_{k+1}, \ldots R_n) = p(R_k), \tag{6.2}$$

for all $k > 2$, such additional dependencies are absent. In that case, all R_k with $k > 2$ are completely independent of each other.

Independencies such as those exemplified above may be used to simplify the representation of the probability distribution over the variables R_k. Causal networks provide a systematic way of doing so. The reader may consult Lauritzen and Spiegelhalter (1988) for a clear discussion, or Pearl (2000) for references to a more detailed discussion on Bayesian networks. This chapter assumes familiarity with Bayesian networks as introduced there.

Medical example. The specific dependency structure sketched above serves as the leading example of this chapter. To illustrate this structure, let me provide a Bayesian network associated with it, and an interpretation of the variables R_1, R_2 and R_k for $k > 2$. As for the latter, we may imagine a medical doctor who is screening individuals for a symptom R_1. This variable is assigned 1 if the symptom obtains and 0 otherwise. The doctor also tests the individuals for a disease that is causally related to the symptom. The test result R_2 is fully reliable, and comes out 1 if the disease obtains and 0 otherwise. In her log the doctor further categorises each of the individuals by means of a number of binary conditions, denoted R_k with $k > 2$, all of which are known to be causally irrelevant to, and thus independent of, the symptom R_1 and the disease R_2. The probability distribution expressing her beliefs concerning the variables R_k then exactly matches the independence relations (6.1) and (6.2).

The dependency structure given in these relations may be associated with a Bayesian network as depicted in 6.1. Relative to this network, the probability distribution can be determined by independent distributions $p(R_k)$ for all

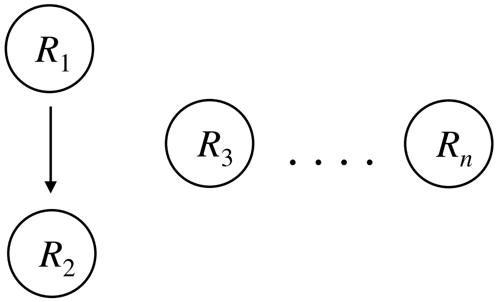


Figure 6.1: The Bayesian network for the variables R_k of the example. All variables are independent, except for R_1 whose values probabilistically depend on R_2.

$k > 1$, and by the conditional probabilities $p(R_1|R_2)$. It may be noted that the direction of the arrow between R_1 and R_2 is not determined by the dependency structure, as laid down in the probability distribution p. Rather it is determined by the order in the variables R_k that is chosen for constructing the network, as discussed in Pearl (1989: 125-126). As suggested in the foregoing, the directed arrows in the network may be interpreted as pertaining to causal relations between the variables. Depending on the interpretation of these variables, some orders in the variables may therefore be considered more natural than others.

Aim of this chapter. This chapter is concerned with repeated and exchangeable observations of complete valuations of all node variables R_k. In terms of the example, it is concerned with screening and categorising individuals. Its primary aim is to present a scheme for inductive inferences on observations of nodes, conditional on a dependency structure as expressed in a Bayesian network, but in the absence of probabilities associated with the nodes and edges. That is, when given the network and some observations, the scheme is supposed to fill in the most likely probabilities of the nodes and the conditional probabilities of the edges. A further aim is to derive predictions on observations of nodes, both unconditional and conditional on the observations of other nodes. But this can only be done sensibly after the probabilities for nodes and edges have been estimated.

It cannot be emphasised enough that the aim of this chapter differs from the aim of most other studies that concern Bayesian networks. The following does not have the object, more common in artificial intelligence, to derive the most probable network structure on the basis of observations. See Bacchus and

Lam (1994), Tong and Koller (2001), and their references for a more elaborate discussion. Also, the following does not primarily aim at predictions, for a single individual, of a node variable on the basis of valuations for other nodes, and given probabilities for nodes and edges. In the following, such probabilities are not yet given.

It seems that inductive inference on the basis of a fixed network structure but without fixed probabilities in the network has received little direct attention. On the other hand, much of more general mathematical statistics can readily be applied to this problem, and perhaps mathematical statisticians have considered the application to be too obvious to merit separate consideration. From the side of Bayesian inductive logic, however, things look a bit different. One interesting aspect is that Bayesian networks may be connected to Carnapian prediction rules. This may be interesting from a computational point of view, also outside the tradition of inductive logic. Furthermore, the Bayesian scheme of this chapter allows us to describe a specific type of inductive dependence between nodes and edges, which cannot be captured in terms of the edges in Bayesian networks. This reveals that caution is required when using a causal picture to fix screening-off relations. And finally, the present treatment signals an intimate connection between the discussion on analogy and the discussion on probabilistic causality.

Connection to Carnapian predictions. To elaborate on the aims of this chapter, let me briefly discuss Bayesian networks from the perspective of Carnapian inductive logic. Note that the nodes R_1 to R_n in a network determine a single variable Q. The possible values q thereof are associated with a binary expansion that encodes the variables $R_k = r_k(q)$ corresponding to it:

$$q = \sum_k r_k(q) 2^{k-1}. \tag{6.3}$$

Depending on the number l_x of conditional probabilities over edges that are restricted to extremal values, or in Pearl's words, restricted conceptually, we have a number of $2^n - l_x$ possible values q. We can then define a probability assignment over repeated observations of these variables that captures inductive predictions, as for example in Carnap (1952). The problem addressed in this chapter, if formulated in terms of such inductive predictions, is to incorporate the specific dependencies and independencies of the network into the probability assignment over repeated observations.

There are some problems with a Carnapian picture of inductive inferences on Bayesian networks. First, predictions defined in terms of the variable Q

cannot easily cope with observations concerning single nodes, or strict subsets of nodes. Recall that a Q-predicate determines all nodes at once. In networks that represent experimental variables, a number of the nodes in the causal picture will typically be unobservable, so that observations only specify subsets of values for Q. Formulated more positively, the Carnapian picture needs some additional conceptual tools to include these disjunctions. Second, and more importantly for present purposes, it has proved very difficult to model the dependencies, as they are expressed in Bayesian networks, with the tools of Carnapian inductive logic. The dependencies are in this logic associated with analogical predictions, as will be explained below, and the problem of capturing such predictions has been the subject of heavy debate for a long time already.

Connection to analogical predictions. I now elaborate the connection between Bayesian networks and analogical predictions. First of all, if an observation q has a varying impact on other observations q', the variations may be associated with differing relevance relations between the observations q and q'. And these relevance relations are traditionally associated with analogy. Now to link networks and analogy, note that dependencies between the variables R_k will show up in the inductive predictions over Q as differences between the impact that an observation q has on other observations q'. For example, if the underlying variables R_1 and R_2 are connected in the given network, the fact that the values r_1 and r_2, which are encoded in some q, often occur together is taken to indicate something on the dependence between these two values. The repeated observation of q then reflects positively on further observations of q, but also on those observations q' that encode the same specific values r_1 and r_2, alongside values r_k for $k > 2$ that differ from those associated with q. By contrast, if the variables R_1 and R_2 are not connected in the network, the fact that the values r_1 and r_2 repeatedly occur together is deemed a mere chance occurrence, as the network already asserts that R_1 and R_2 are independent. The observation of q then entails no additional positive impact for the other observations q' that encode the co-occurrence of r_1 and r_2.

The above considerations invite the construction of a scheme that can deal with separate observations of variables R_k, and within which dependence relations can be expressed more easily. We can in this context draw lessons from the debate on analogical predictions. Accordingly, the formal scheme of this chapter allows for observations of separate observations R_k, and is based on the model for analogical predictions developed in the preceding two chapters. The central idea is again to model inductive inferences and predictions with a Bayesian update over statistical hypotheses, and to capture the dependency relations between families R_k in terms of such hypotheses. In the following the scheme is made precise with the specific aim of deriving Carnapian predictions for Bayesian networks. The further aim is to show the limitations of Bayesian networks in expressing the dependency structure. For this latter aim, the following contains an elaborate discussion of observations and hypotheses concerning the variables R_k.

Plan of this chapter. The plan of this chapter is as follows. Section 6.2 discusses the inductive scheme in general. Section 6.3 defines specific hypotheses for exchangeable predictions on independent nodes. Then section 6.4 presents some tools for managing complicated hypotheses partitions. Section 6.5 defines a specific partition that captures a link between nodes, and shows that the tools can be employed for deriving Carnapian predictions. Section 6.6 discusses some possibilities with the scheme that reveal a shortcoming in the representation of dependencies in a Bayesian network. Finally, section 6.7 discusses how we can apply all this to Bayesian networks more generally.

6.2 Bayesian scheme

This section introduces an observational algebra for repeated observations of the variables R_k, and defines a general scheme for making predictions. The algebra and the scheme are a generalisation of the notions used in the preceding two chapters. In comparison to the schemes above, the one presented here may look unnecessarily complicated. The preceding chapters have employed a number of restrictions to simplify both notation and discussion. But I think that it is important to show that none of these restrictions is essential to the schemes themselves. In particular, the restriction on the order of the observations, as discussed in section 4.4, is here seen to be inessential.

6.2.1 Observational algebra

Observations of R-predicates. Let me first settle some notational issues. First, I refer to observations of R_k having value r with the term R_k^r. These valuations concern a specific individual or experimental system, indexed i. I will refer to the individuals by adding this index to the valuation, R_{ki}^r. Now we need not assume that these observations are collected in the order of the individuals and nodes, according to the indices i or k. For example, we may record the variable R_2 for a number of different individuals first, and record the variable R_1 only after that. To accommodate this, I add another index t which refers to the time of the observation. The expression R^r_{kit} thus refers to the observation, at time t, that variable R_k of individual i has value r. It must be stressed that the time index is not intended to keep track of the time evolutions in an individual. Each variable R_k of individual i is only observed once.

Let me introduce set theoretical representations for these observations in terms of an algebra of cylinder sets. Define the set of all possible triples (rki) as M, and consider the space M^ω of all infinite, ordered sequences u of such observations:

$$u = (rki)_1(rki)_2(rki)_3 \dots \tag{6.4}$$

The elements of the infinite sequences (rki) denote the serial number i of the individual, the number k of the variable R_k with respect to which the individual is observed, and the value r of the variable. Every such triple determines an observation at time t. Note that if we assume that there is no independent way of labelling the individuals i, permuting the individual indices in a sequence does not produce a different sequence. But in the following this redundancy is not harmful.

An algebra for the R-predicates. The observational algebra for variables R_k is given by all possible subsets of the space M^ω. This algebra is denoted with $\mathcal{R}$. Observations R^r_{kit} can be expressed as elements of this algebra. If we denote the t-th triple $(rki)_t$ in a series $u \in M^\omega$ with $u(t)$, we can write:

$$R^r_{kit} = \{u : \ u(t) = (rki)_t\}. \tag{6.5}$$

From now on observations R^r_{kit} refer to such subsets. Note that there is a formal distinction between the observations R^r_{kit} and the values of observations (rki). The values, represented with small letters, are triples of natural numbers. The observations, denoted with large letters and indexed with small ones, are subsets of M^ω, and therefore elements of the algebra $\mathcal{R}$.

In the same way we can represent sequences of observations S_t as elements of the algebra $\mathcal{R}$. These elements are determined by the ordered sequences $s_t = \langle (rki)_1, (rki)_2, \dots, (rki)_t \rangle$. Analogous to expression (6.5) we can write

$$S^s_t = \{u : \ \forall t' \leq t \, (u(t') = s_t(t'))\}, \tag{6.6}$$

in which $s_t(t') = (rki)_{t'}$. Observations of R-predicates and sequences of such observations are related to each other as follows:

$$R^r_{kit} \cap S^s_{t-1} = S^{\langle s,(rki)\rangle}_t. \tag{6.7}$$

We can therefore build up the sets S^s_t by intersecting the subsequent observation sets $R^r_{kit'}$ up until t.

Relation to Q-predicates. To map observations q of a variable Q for individual i onto elements in the algebra $\mathcal{R}$, we can employ the binary expansion of q. Writing Q^q_{it} for the observation, from time t onwards, that individual i has values according to q, we can write

$$\begin{aligned} Q^q_i &= \{u \in M^\omega : \forall k \leq N : u(t+k) = \langle r_k(q), k, i\rangle\} \\ &= \{u \in M^\omega : u(t+1) = \langle r_1(q), 1, i\rangle\} \cap \\ &\qquad \ldots \cap \{u \in M^\omega : u(t+n) = \langle r_n(q), n, i\rangle\} \\ &= \bigcap_k R^{r_k(q)}_{ki(t+k)}, \end{aligned} \tag{6.8}$$

As already suggested by the construction of sequences u, every complete observation Q^q_{it} of individual i at time t can be written down as the intersection of such observations, of object i with respect to all separate predicate families R_k. The observations Q^q_{it} can thus be integrated in the algebra for observations R^r_{kit}.

For later purposes, let me construct the following special sequences of observations in $\mathcal{R}$:

$$s_k = \langle r_1, r_2, \ldots, r_k\rangle, \tag{6.9}$$

$$S^{s_k}_t = R^{r_1}_{1i(t-k+1)} \cap R^{r_2}_{2i(t-k+2)} \ldots \cap R^{r_k}_{kit}. \tag{6.10}$$

The vectors s_k consist of k components, namely $r_{k'}$ for all $k' \leq k$. They determine sequence of k observations of R-predicates. I usually omit reference to k in the superscript, so that the string becomes S^s_{ki}. By $s_k(q)$ I mean the first k components $r_{k'}(q)$.

One further convention will prove very useful in this chapter. Given a sequence of observation results $s_t = \langle (rki)_1, (rki)_2, \ldots, (rki)_t\rangle$, we can write down, for all combinations of r and k, the number of times within the sequence that an object is observed in predicate r within predicate R_k. We can denote these numbers t_{rk}. The sum over the index r of these numbers, $t_k = \sum_r t_{rk}$, gives the number of times an observation concerned the predicate R_k. The ratios $\frac{t_{rk}}{t_k}$ are the observed relative frequencies of the triples.

6.2.2 Hypotheses and predictions

The following contains a brief introduction to Bayesian schemes that use hypotheses for making predictions. It deals with belief states, hypotheses, conditioning, and predictions based on hypotheses. Comparable expositions can again be found in Jeffrey (1984) and Howson and Urbach (1996). This section and the following specifically focus on R-predicates, but the scheme works exactly similarly for the Q-predicates.

Predictions from hypotheses. As indicated, beliefs are represented by probability functions p. These functions are defined over the observational algebra $\mathcal{R}$, and thus take observations R^r_{kit} and sequences of observations S_t as arguments. We can express full belief in the observations S_t in terms of the probability assignment p: on observing the sequence S_t we fix $p(S_t) = 1$. As a result of fixing this probability, we also have to adapt the probabilities of other elements in the observational algebra. In all of the following I assume that the probability function representing beliefs upon observing S_t can be constructed by conditioning the original probability function p on the observations S_t:

$$p(\cdot) \quad \rightarrow \quad p(\cdot|S_t). \tag{6.11}$$

Both the probabilities assigned to observations, and the probabilities assigned to hypotheses can be adapted to new observations in this way. In the following, the probability before updating is called the prior probability, and the one after updating the posterior probability.

The following employs conditioning over observational hypotheses to generate predictions of the form $p(R^r_{kit}|S_{t-1})$. Observational hypotheses can be seen as elements of the observational algebra. If we assume of some hypothesis h that its truth can be determined as a function of an infinitely long sequence of observations u, then we can define hypotheses as subsets of M^ω in the following way:

$$H = \{u \in M^\omega : W_h(u) = 1\}, \tag{6.12}$$

where $W_h = 1$ if and only if h is true of u. The function W_h is called the indicator function of h. The predictions are based on so-called partitions of such hypotheses. A partition is a collection of hypotheses $\mathcal{D} = \{H_\theta\}_{\theta \in D}$, defined by the following condition for the indicator functions W_{h_θ}:

$$\forall u \in M^\omega \; \exists! \theta : \quad W_{h_\theta}(u) = 1. \tag{6.13}$$

This means that the hypotheses H_θ are mutually exclusive and jointly exhaustive sets in M^ω. Note that the above expression refers to a hypotheses space D, while it does not specify the dimensions or even structure of D yet. In particular it must be noted that it is also possible to work only with a countable or finite number of hypotheses.

Because we are in this chapter dealing with a continuum of hypotheses, the probability function must be a probability density $p(H_\theta|S_{t-1})$. Using this density and the partition, we can define predictions by means of the law of total probability:

$$p(R^r_{kit}|S_{t-1}) = \int_D p(H_\theta|S_{t-1})\, p(R^r_{kit}|H_\theta \cap S_{t-1})\, d\theta. \tag{6.14}$$

The terms $p(R^r_{kit}|H_\theta \cap S_{t-1})$ are called the posterior likelihoods of R^r_{kit} on the hypotheses H_θ. The prediction is obtained by weighing these likelihoods with the posterior probability over the hypotheses, $p(H_\theta|S_{t-1})d\theta$.

Updating over hypotheses. The above expresses predictions in terms of the posterior probability and the posterior likelihoods, denoted $p(H_\theta|S_{t-1})d\theta$ and $p(R^r_{kit}|H_\theta \cap S_{t-1})$ respectively. Both these terms can be obtained by conditioning on the prior probability assignments $p(H_\theta)d\theta$ and $p(R^r_{kit}|H_\theta)$. In many cases the likelihoods do not change upon conditioning:

$$p(R^r_{kit}|H_\theta \cap S_{t-1}) = p(R^r_{kit}|H_\theta). \tag{6.15}$$

Whenever this is so, I do not mention the term S_{t-1} in the expression for the likelihoods. But the invariance cannot always be taken for granted.

The dependence of the predictions on observations are further reflected in the posterior probability over the hypotheses. This probability can be determined by means of conditioning:

$$p(H_\theta|S_{t'})d\theta = \frac{p(R^r_{kit'}|H_\theta \cap S_{t'-1})}{p(R^r_{kit'}|S_{t'-1})}\, p(H_\theta|S_{t'-1})d\theta, \tag{6.16}$$

Note that the denominator $p(R^r_{kit'}|S_{t'-1})$ is equivalent to equation (6.14) with t' in place of t, so that this denominator can be rewritten as a function of $p(R^r_{kit'}|H_\theta \cap S_{t'-1})$ and $p(H_\theta|S_{t'-1})$. The posterior probability over the hypotheses given S_{t-1} can thus be determined recursively by the prior probability function $p(H_\theta)d\theta$, and the likelihoods $p(R^r_{kit'}|H_\theta \cap S_{t'-1})$ for all times $t' < t$.

In sum, the predictions $p(R^r_{kit}|S_{t-1})$ can be generated if we assume some partition of hypotheses $\mathcal{D}$, the likelihoods $p(R^r_{kit}|H_\theta \cap S_{t-1})$ for all $H_\theta \in \mathcal{D}$ and at all times $t' \leq t$, and a prior probability distribution $p(H_\theta)d\theta$. First the prior and the likelihoods with $t' < t$ can be used to determine the posterior probability over the partition. The likelihoods are subsequently used with this posterior probability over the partition for generating the prediction itself.

6.3 Hypotheses on nodes

This section defines the specific partition of hypotheses that underpins exchangeable predictions for a single variable R_0, and the partition of hypotheses that concerns multiple independent variables.

6.3.1 Relative frequencies

To illustrate the above scheme, and to prepare for later sections, I here define a partition of hypotheses that results in Carnapian prediction rules. The scheme is a generalisation of the schemes used in the preceding two chapters. In the following the prediction rules are first derived for a single variable R_0. At every time t we observe R_0^r for a different individual i, so that $i = t$, and every observation is thus fully characterised with a single index r.

Bernoulli hypotheses. To characterise the hypotheses that result in exchangeable predictions, first define the notion of the relative frequency of observations R_0^1 in a sequence u,

$$f_0(u) = \lim_{t \to \omega} \frac{1}{t} \sum_{t'=1}^{t} u(t'), \tag{6.17}$$

assuming that this limit exists. For any infinitely long sequence of observations u, the function $f_0(u)$ gives the relative frequency of the observations R_{0it}^1. Note that for a given u, the frequency $f_0(u)$ need not be defined. Note further that the relative frequency for R_0^0 is defined exactly when f_0 is, and equals $1 - \theta$.

The hypotheses H_θ can be defined by the indicator function W_{h_θ} and equation (6.12):

$$W_{h_\theta}(u) = \begin{cases} 1 & \text{if } f_0(u) = \theta, \\ 0 & \text{otherwise.} \end{cases} \tag{6.18}$$

The parameter space for these hypotheses is simply $\theta \in B_0 = [0, 1]$. Further, we may define $W_{h_{\neg\theta}} = 1$ if $f_0(u)$ is undefined, and $W_{h_{\neg\theta}} = 0$ otherwise. The collection of hypotheses $\mathcal{B}_0 = \{H_{\neg\theta}, \{H_\theta\}_{\theta \in B_0}\}$ then is a partition of hypotheses concerning the relative frequencies of the observations r of variable R_0.

We can now provide likelihoods and a prior probability for this partition. First we assume that $p(H_{\neg\theta}) = 0$, which states that u has some convergent relative frequency. The prior probability over the hypotheses H_θ can otherwise be chosen freely. As for the likelihoods of H_θ, they may then be obtained by taking the long run relative frequency θ as chances on the observation R_0^1 at every single observation:

$$p(R^r_{0it}|H_\theta) = \theta^r(1-\theta)^{1-r}. \tag{6.19}$$

Since $r = 0$ or $r = 1$, either of the factors on the right is 1. The relation between hypotheses H_θ and their likelihoods is perhaps not indisputable. I refer to section 2.4.2 for a detailed discussion. Note that the likelihoods do not depend on the observations S_{t-1}. For this reason the predictions that result from the partition $\mathcal{B}_0$ are exchangeable, that is, they are the same independently of the order in the observations S_{t-1}.

Carnapian rules from Dirichlet distributions. The predictions rendered by the partition $\mathcal{B}_0$, if supplied with a so-called Dirichlet density as prior,

$$p(H_\theta) \quad \sim \quad \theta^{(a_1-1)}(1-\theta)^{(a_0-1)}, \tag{6.20}$$

can be written down in the form of the Carnapian $\lambda\gamma$ prediction rules. Using the numbers t_{rk} and t_k defined in subsection 6.2.1, it reads

$$p(R^r_{ki(t+1)}|S_t) = \frac{t_{r0} + \gamma_{r0}\lambda_0}{t_0 + \lambda_0}. \tag{6.21}$$

The term t_0 represents the number of observations with respect to R_0 within the sequence S_t, so in this case $t_0 = t$. Further, the Dirichlet prior determines the values $\lambda_0 = \sum_r a_r$ and $\gamma_{r0} = a_r/\lambda_0$. If we choose it to be the uniform distribution, the values are $\gamma_{r0} = \frac{1}{2}$, and $\lambda_0 = 2$.

The above rule is involved in all of the following. With its introduction, it almost seems that the use of hypotheses and conditioning is unnecessarily complicated. However, in line with the general import of this thesis, the hypotheses turn out to be a useful tool in laying down and controlling the dependence assumptions as expressed in Bayesian networks.

6.3.2 Unconnected nodes

The following deals with hypotheses schemes concerning a number n of statistically independent predicate families R_k. This scheme presents the point of departure for constructing hypotheses schemes associated with Bayesian networks.

Subsequences on separate nodes. We first need to define relative frequencies of R-predicates in the case that more than one such predicate occurs within a given infinite sequence u. With that aim, define, for any sequence u, a function indicating whether at time t the observation concerns the predicate R_k:

$$W_k(u,t) = \begin{cases} 1 & \text{if } \exists r, i: \ u(t) = rki, \\ 0 & \text{otherwise.} \end{cases} \tag{6.22}$$

Note that, since at any time the observation necessarily concerns exactly one such variable R_k, we have that $\sum_k W_k(u,t) = 1$ for any u and t.

We can now collect all those positions t at which the observation concerned the predicate R_k in exactly the same way as in section 2.3.1:

$$g(u,t) \quad = \quad W_k(u,t) \sum_{t'=1}^{t} W_k(u,t'), \tag{6.23}$$

$$u^k(g(u,t)) \quad = \quad u(t). \tag{6.24}$$

In words, the function $g(u,t)$ assigns a zero to all positions t at which the observation $u(t)$ does not concern the predicate R_k, and at other positions t it assigns the total number of times that observations concerned R_k up until t. The subsequence u^k is thus defined to contain only those triples in u that concern R_k, in numerical order and starting at $t = 1$. The position $u^k(0)$ remains undefined.

Bernoulli hypotheses for separate nodes. Exchangeable predictions for a single predicate can be modelled by means of hypotheses on relative frequencies. One possible extension to more predicate families is to define hypotheses H_θ that concern the relative frequencies of the cells in every R_k separately. The subsequences concerning the predicates R_k can be used for this purpose. Define

$$f_k(u) = \lim_{t \to \omega} \frac{1}{t} \sum_{t'=1}^{t} u^k(t'), \tag{6.25}$$

where u^k is the subsequence of u with respect to R_k. The f_k thus give the relative frequency of the observations R_k^1 within the sequence u_k. Again, the relative frequency of the observations R_k^0 is $1 - f_k$.

The hypotheses that generate predictions on independent predicate families can then be defined by means of these relative frequencies:

$$W_{h_\theta}(u) = \begin{cases} 1 & \text{if } \forall k \leq n: \ f_k(u) = \theta_k, \\ 0 & \text{otherwise,} \end{cases} \tag{6.26}$$

where θ is a vector in a space I of which the components $\theta_k \in I_k = [0,1]$, so that $I = [0,1]^n$. Using equation (6.12) we can define the hypotheses H_θ over the algebra $\mathcal{R}$ concerning multiple variables R_k. As in the foregoing, we can then define $H_{\neg\theta}$, and also the partition $\mathcal{I} = \{H_{\neg\theta}, \{H_\theta\}_{\theta \in I}\}$.

The schemes developed in section 6.2 can employ this partition unproblematically. To obtain predictions from it, we need the likelihoods of the predicates in the separate families, and further a prior probability over the hypotheses. The prior over the hypotheses space $\mathcal{I}$ is not restricted to any specific form, as long as we assume that $p(H_{\neg\theta}) = 0$. As for the likelihoods, they can be chosen analogous to those for the single variable R_0:

$$p(R^r_{kit}|H_\theta) = \theta^r_k (1 - \theta_k)^{1-r}. \tag{6.27}$$

For an observation in the family R_k, the relevant component of θ is θ_k. Note also that the likelihoods are independent of the observations S_t. The resulting predictions for any R^r_{kit} are therefore exchangeable.

Remarks on the partition $\mathcal{I}$. It is notable that the likelihoods over the values sum to one for all variables R_k separately. That is,

$$\sum_r p(R^r_{kit}|H_\theta) = 1. \tag{6.28}$$

The idea behind this is that the hypotheses are not concerned with the issue which variable is observed next. That is, hypotheses only determine the probabilities of values r within a given family R_k, assuming that prior to the observation the variable is fixed by external circumstances. It is of course possible to include this choice of variable into the likelihoods of the hypotheses, and to predict this aspect for the observations as well. But the present chapter does not deal with that refinement.

Note also that the parameter space I for the partition concerning multiple independent predicate families is a product of unit intervals, $[0, 1]^n$. For the case of $n = 3$, it is simply the unit cube. Updating on some R^r_{kit} means that the probability function $p(H_\theta)d\theta$ over this cube must be adapted according to expression (6.16). The predictions for R^r_{kit} can be calculated from the probability function over this cube according to expression (6.14). Now the natural question is whether there is a representation of these predictions in terms of the $\lambda\gamma$ rules, like those applicable to the simplex for a single R-predicate. This question, however, can only be answered in the next section, when we have dealt with marginal and conditional distributions.

6.4 Tools for Bayesian networks

This section introduces some tools that are helpful in understanding and managing the predictions for Bayesian networks. It deals with marginal and conditional densities over the partition $\mathcal{I}$, and derives $\lambda\gamma$ prediction rules for these partitions based on a certain class of priors. While the partition $\mathcal{I}$ is the leading example, the same tools developed here can be applied to the partition for Bayesian networks given in the next section.

6.4.1 Marginal and conditional densities

Marginal and conditional partitions. Before presenting the mathematics, let me describe marginal and conditional densities informally. Recall that the parameter space I is composed of a product of spaces I_k. A marginal probability density with respect to k is basically a projection of the probability density over I onto one such space I_k. The probability density of the hypotheses that have a certain value for θ_k are summed, and assigned to an aggregated or so-called marginal hypothesis, which is characterised by θ_k alone. A conditional density, on the other hand, is a density over all hypotheses that are included in a specific marginal hypothesis. It is a separate probability function, associated with specific values for θ_k. Thus a marginal density and the continuum of associated conditional densities together determine the density over the whole of I.

First consider the marginal partitions $\mathcal{I}_k$. Let me fix the component θ_k of θ to some specific value η:

$$\theta_k = \eta. \tag{6.29}$$

The parameter η runs over the same domain as θ_k, so $\eta \in I_k = [0, 1]$. We may now define the hypotheses H^k_η in the marginal partition $\mathcal{I}_k$ as

$$H^k_\eta = \bigcup_{\theta_k=\eta} H_\theta. \tag{6.30}$$

All hypotheses H_θ in $\mathcal{I}$ with the values $\theta_k = \eta$ are collected in the hypothesis H^k_η. So whereas the hypotheses H_θ are points in the space $[0, 1]^n$, the hypotheses $H^k_\eta \in \mathcal{I}_k$ are sets of dimension $n - 1$ in this space.

Now consider conditional partitions $\mathcal{I}^k_\eta$. Let ζ be the vector of those components of θ that are not associated with the space I_k within I:

$$\zeta = \langle \theta_1, \ldots, \theta_{k-1}, \theta_{k+1} \ldots \theta_n \rangle. \tag{6.31}$$

So ζ is an $n - 1$ dimensional vector in the parameter space $I_1 \times \ldots \times I_{k-1} \times I_{k+1} \times \ldots \times I_n$. Now the hypotheses $H^k_{\eta\zeta}$ within a conditional partition $\mathcal{I}^k_\eta$ are the same as the hypotheses H_θ, with the difference that the components of θ_k are fixed to η. The ζ are the remaining free parameters:

$$H^k_{\eta\zeta} = H_{\langle\theta_1,\ldots,\theta_{k-1},\eta,\theta_{k+1},\ldots,\theta_n\rangle}. \tag{6.32}$$

The partition $\mathcal{I}^k_\eta$ is thus a specific subpartition: each of them covers exactly one marginal hypothesis H^k_η. The hypotheses $H^k_{\eta\zeta}$ provide separate command over the vector ζ within the subspace of hypotheses H^k_η.

Probability over the marginal and conditional partitions. On the basis of a density $p(H_\theta)$ over the partition $\mathcal{I}$, we can determine the probability over the marginal partition $\mathcal{I}_k$ for any k:

$$p(H^k_\eta)d\eta = \int_{I^k_\eta} p(H_\theta)d\theta. \tag{6.33}$$

The density $p(H^k_\eta)$ is called the marginal density over $\mathcal{I}_k$. Note that marginal densities are normalised, because the density $p(H_\theta)$ is normalised. Note further that the density $p(H_\theta)$ uniquely determines the marginal densities for all k, while the converse does not hold: the marginal densities for the $\mathcal{I}_k$ do not completely specify the underlying density $p(H_\theta)$ over the partition.

Conditional densities can also be defined over the conditional partitions $\mathcal{I}^k_\eta$. Such densities are defined within the conditional partitions only, and thus present separate probability functions, here denoted p^k_η. Since $H^k_{\eta\zeta}$ is just an elaborate way of denoting H_θ conditional on H^k_η, we can write

$$p^k_\eta(H^k_{\eta\zeta})d\zeta = \frac{p(H_\theta \cap H^k_\eta)d\theta}{p(H^k_\eta)d\eta}. \tag{6.34}$$

The conditional densities $p^k_\eta(H^k_{\eta\zeta})$ are thus identical to the original density over H_θ, but they are normalised within every partition $\mathcal{I}^k_\eta$ separately. For sake of brevity, I usually suppress the indices k and η of the hypotheses if they function as argument in p^k_η.

A complete representation of the probability assignment over $\mathcal{I}$ can be obtained by using one marginal probability over $\mathcal{I}_k$, together with the associated collection of conditional probability assignments over $\mathcal{I}^k_\eta$:

$$p(H_\theta)d\theta = p(H^k_\eta)\left[p^k_\eta(H_\zeta)d\zeta\right]d\eta. \tag{6.35}$$

In the following it turns out to be very useful that the density over $\mathcal{I}$ can be written out in these terms.

Let me illustrate the marginal and conditional densities for $\mathcal{I}$ in the case $n = 3$, the example of the unit cube. First, the hypotheses H^1_η can be represented by squares in this cube that have $\theta_1 = \eta$. Further, every square $\theta_1 = \eta$ covers exactly one conditional partition $\mathcal{I}^1_\eta$. The points within these squares are parameterized by $\langle \theta_2, \theta_3 \rangle = \langle \zeta_1, \zeta_2 \rangle$. Integrating the probability over such a square yields the probability assigned to the marginal hypotheses, $p(H^1_\eta)d\eta$. The probability densities within these squares, $p^1_\eta(H_\zeta)$, are the conditional densities. Finally, the vector θ in the space $\mathcal{I}$ can indeed be reparametrized by η and ζ, and the stipulation that η is associated with $\mathcal{B}_1$.

It may seem elliptical to introduce a separate notion of conditional partitions and densities. As a terse motivation, it is useful to deal with the densities $p^k_\eta(H^k_{\eta\zeta})$ as separately normalised functions. We can then operate and calculate with these functions independently, as we do with the marginal densities.

6.4.2 Predicting and updating

The following shows that the decomposition of a density into a marginal density and a collection of conditional densities greatly simplifies predictions and update operations over $\mathcal{I}$: predictions can be derived completely from marginal distributions, and an update operations only affects one marginal at a time.

Criterion for simplification. While the partition $\mathcal{I}$ is the leading example of this section, the simplifications to be presented apply to a more general class of partitions. The criterion is that the likelihoods of the observations have the following form:

$$p(R^r_{kit} | H_\theta \cap S^s_{t-1}) = \theta_{j(s)}, \tag{6.36}$$

where $\theta_{j(s)} \in [0, 1]$. So the likelihood of observing, at some time t and for some individual i, that the variable R_k has the value r may involve only a single parameter component $\theta_{j(s)}$ in the parameter space. Partition $\mathcal{I}$ conforms to this criterion, as it simply has $j = k$. But as will be illustrated in section 6.6, meeting this criterion does not entail that the likelihoods are independent of S^s_{t-1}. The index j may depend on k and on the observations S^s_{t-1}.

Simple predictions. The prediction of R^r_{kit} can be simplified using the decomposition of partition $\mathcal{I}$ in terms of a marginal density over $\mathcal{I}_k$ and a continuum of conditional densities over $\mathcal{I}^k_\eta$:

$$\begin{aligned}
p(R^r_{kit}|S^s_{t-1}) &= \int_I p(R^r_{kit}|H_\theta \cap S^s_{t-1}) p(H_\theta|S^s_{t-1})\, d\theta \\
&= \int_I \theta_k\, p(H_\theta|S^s_{t-1})\, d\theta \\
&= \int_{I_k} \eta\, p(H^k_\eta|S^s_{t-1}) \times \left[\int_{I^k_\eta} p^k_\eta(H_\zeta|S^s_{t-1})\, d\zeta \right] d\eta \\
&= \int_{I_k} \eta\, p(H^k_\eta|S^s_{t-1})\, d\eta. \qquad (6.37)
\end{aligned}$$

That is, the prediction of observation R^r_{kit} is completely determined by the marginal density $p(H^k_\eta|S^s_{t-1})$.

Let me briefly explain this simplification in words. Note that for all the hypotheses within the conditional partition $\mathcal{I}^k_\eta$, the likelihood of R^r_{kit} is η. The hypotheses within the conditional partitions therefore contribute to the prediction in the same way, and consequently the conditional densities do not affect the predictions. They can therefore be integrated out. The hypotheses from the marginal partition, on the other hand, do have different likelihoods η for the observation R^r_{kit}, so the density over this latter partition does affect the prediction.

Simple updates. Updating the density with an observation R^r_{kit} can be simplified for the same reasons. We can write, using Bayes' rule in the first line and Bayes' theorem in the second ,

$$\begin{aligned}
p(H_\theta|S_t)d\theta &= p(H_\theta|S_{t-1} \cap R^r_{kit})d\theta \\
&= \frac{p(R^r_{kit}|H_\theta \cap S_{t-1})}{\int_I p(R^r_{kit}|H_\theta \cap S_{t-1}) p(H_\theta|S_{t-1})d\theta}\, p(H_\theta|S_{t-1})d\theta \qquad (6.38) \\
&= \frac{\theta_k}{\int_I \theta_k p(H_\theta|S_{t-1})d\theta}\, p(H_\theta|S_{t-1})d\theta \\
&= p^k_\eta(H_\zeta|S_{t-1})d\zeta \\
&\quad \times \frac{\eta}{\int_{I_k} \eta p(H^k_\eta|S_{t-1})d\eta}\, p(H^k_\eta|S_{t-1})d\eta. \qquad (6.39)
\end{aligned}$$

This shows that the update with R^r_{kit} only involves changes in the marginal density over $\mathcal{I}_k$. The conditional probability assignments $p^k_\eta(H_\zeta)d\zeta$ all remain unchanged. The only thing that changes for them is the normalisation factor, which becomes $p(H^k_\eta|S_t)d\eta$, but clearly this does not alter the density functions themselves.

It seems that the marginal densities $p(H^k_\eta|S_t)$ over $\mathcal{I}_k$ function exactly as does the density over the partition $\mathcal{B}_0$ described above. Both updating and predicting follow the same formulas. However, this is not yet the same as saying that the predictions for any variable R_k can be captured with the $\lambda\gamma$ rule that can be derived for $\mathcal{B}_0$. It may still be the case that updating with some observation $R^{r'}_{k'i't}$ within the family $k' \neq k$, adapting the marginal over $\mathcal{I}_{k'}$, implicitly alters the marginal over $\mathcal{I}_k$. The reason for this is that the densities $p^{k'}_{\eta'}$ may vary along an axis belonging to I_k. Updating the marginal density over the partition $\mathcal{I}_{k'}$ then changes the weights given to these conditional distributions, thus changing the marginal density over $\mathcal{I}_k$ implicitly. The next subsection spells out the assumption that ensures the full independence of the marginals.

6.4.3 A system of $\lambda\gamma$ rules

To derive $\lambda\gamma$ rules for the families R_k, we must assume that updates on observations concerning other predicate families $R_{k'}$ do not change the marginal densities over $\mathcal{I}_k$. This independence can be effected by assuming that the density over $\mathcal{I}$ can be decomposed into a product of marginal distributions. This subsection spells out this assumption mathematically, after which a derivation of $\lambda\gamma$ rules is given.

Decomposition into marginals. Let me first give the assumption on the decomposition, or factorisability, of the prior density into marginals:

$$p(H_\theta) \quad \sim \quad \prod_k p(H^k_\eta). \tag{6.40}$$

This means that the functional dependence of the density $p(H_\theta|S_{t-1})$ on the vectors θ_k is determined entirely by the marginal density $p(H^k_\eta|S_{t-1})$. The conditional densities associated with each of these marginal partitions, $p^k_\eta(H_\zeta|S_{t-1})$, are thus equal for all values of η.

This assumption ensures that the density over the marginal partition $\mathcal{I}_k$, for some specific value k, is affected only by the updates involving observations R^r_{kit}. That is, an observation $R^{r'}_{k'it}$ does not change the marginal probability over $\mathcal{I}_k$. To see that this is so, first note that

$$p^{k'}_{\eta'}(H_\zeta|S_{t-1})d\zeta = \prod_{k\neq k'} p(H^k_\eta|S_{t-1})d\eta. \tag{6.41}$$

But because of this, the marginal density $p(H^k_\eta)$ is completely determined by the conditional density $p^{k'}_{\eta'}(H_\zeta|S_{t-1})$. We can reconstruct any marginal distribution

$p(H^k_\eta|S_{t-1})d\eta$ by integrating out all the other factors:

$$p(H^k_\eta|S_{t-1})d\eta = \int_{I^{kk'}_{\eta\eta'}} p^{k'}_{\eta'}(H_{\zeta'}|S_{t-1})d\zeta'. \tag{6.42}$$

Here I have abbreviated $I^{kk'}_{\eta\eta'} = B_1 \times \ldots \times B_{k-1} \times \eta \times B_{k+1} \times \ldots \times B_{k'-1} \times \eta' \times B_{k'+1} \times \ldots \times B_N$. Since the conditional densities remain unchanged during an update with $R^{r'}_{k'it}$, the above expression states that the marginal density $p(H^k_\eta)$ remains unchanged during any such update.

Note that requirement (6.40) need only be assumed for the prior distribution. Since update operations only affect one marginal at the time, the requirement makes sure that at any later time the density over $\mathcal{I}$ can still be written as a product of marginal densities over $\mathcal{I}_k$.

Deriving prediction rules. The above assumption makes sure that the independence of marginals for any values of k and k', so that any marginal probability $p(H^k_\eta)d\eta$ can be updated independently. Effectively, the marginals over $\mathcal{I}_k$ can be treated as entirely separate updates within the separate partitions $\mathcal{I}_k$. Moreover, the predictions are determined entirely by the densities over these marginal partitions. We can therefore derive the same prediction rules for $\mathcal{I}_k$ as can be derived for the single predicate partition $\mathcal{B}_0$, the so-called $\lambda\gamma$ rules. This derivation hinges on two requirements: the assumption that the likelihoods for the observations involve only single parameter components, equation (6.36), and the further assumption that the prior density over $\mathcal{I}$ can be factorised into its marginal distributions, equation (6.40).

Let me finally give the resulting system of $\lambda\gamma$ rules itself. First, we have to assume a prior over $\mathcal{I}$ from the class of Dirichlet distributions:

$$p(H_\theta) \quad \sim \quad \prod_k \theta_k^{i_{1k}-1}(1-\theta_k)^{i_{0k}-1}. \tag{6.43}$$

This choice entails that assumption (6.40) holds for all predicate families. We can then derive:

$$p(R^r_{ki(t+1)}|S_t) = \frac{t_{rk} + \gamma_{rk}\lambda_k}{t_k + \lambda_k}, \tag{6.44}$$

where the numbers t_{rk} and t_k are defined as in section 6.2.1. Specifically, if we assume that the prior probability assignment over $\mathcal{I}$ is uniform, the marginals for the separate $\mathcal{I}_k$ are also uniform, so that the parameters become $\lambda_k = 2$ and $\gamma_{rk} = \frac{1}{2}$. These latter rule, known as the straight rule is in the following abbreviated by $pr^*(t_{rk}, t_k)$.

6.5 Capturing the links

This section presents a partition of hypotheses that can express dependency relations between variables R_k. The partition is given for the dependence between variables R_1 and R_2, and shown to lead to a system of $\lambda\gamma$ rules. The use of this partition is then illustrated with some numerical results.

6.5.1 Hypotheses schemes

Intuitive idea. Let me first give a general idea of the hypotheses needed for capturing links in a Bayesian network. These hypotheses must be such that the observation of a value for one variable changes the probabilities for observing values for another variable. In other words, the hypotheses that we are looking for must allow for relations of statistical dependence between two variables. In the example, after observing that individuals for which $R_2 = 1$ usually also have $R_1 = 1$, while for $R_2 = 0$ there seems to be no preference between $R_1 = 0$ and $R_1 = 1$, observing or setting $R_2 = 1$ for a new individual must enhance the probability for observing $R_1 = 1$. More generally, the hypotheses involve relative frequencies of observations $R_1 = r'$ for exactly those individuals that have $R_2 = r''$.

Conditional relative frequency. To capture this idea, I here elaborate the notion of conditional relative frequency. First define a function that returns for all sequences u the index of the individual i that is observed at time t,

$$V_i(u,t) = (001) \cdot u(t) = (001) \cdot (rki) = i. \tag{6.45}$$

Further define for individuals i and observation sequences u a function indicating whether individual i satisfies $R_k = r$ somewhere in u:

$$W_{rk}(u,i) = \begin{cases} 1 & \text{if } \exists t : u(t) = rki, \\ 0 & \text{otherwise.} \end{cases} \tag{6.46}$$

This function determines for every u whether or not we observe $R_k = r$ for individual i. If the individual i is not observed with respect to R_k within u, the function $W_{rk}(u,i)$ returns 0.

The index function $V_i(u,k,i)$ and the indicator W_{rk} can be used to define an indicator that only selects those observations concerning the variable R_1 for which the observed individual i also satisfies a specific value r of the variable R_2.

$$W_{1|r2}(u,t) = W_1(u,t)W_{r2}(u,V_i(u,t)). \tag{6.47}$$

Here $W_1(u,t)$ is defined by (6.22) with $k=1$. We can now define the sequence $u^{1|r2}$ of those observations concerning R_1 for which the individual i under consideration also satisfies $R_2 = r$:

$$\begin{aligned} g(u,t) &= W_{1|r2}(u,t) \sum_{t'=1}^{t} W_{1|r2}(u,t'), \\ u^{1|r2}(g(u,t)) &= u(t). \end{aligned} \tag{6.48}$$

This definition is the same as definition (6.23), with the indicator $W_k(u,t)$ changed for the indicator $W_{1|r2}(u,t)$. The function $g(u,t)$ is again a counter, which only selects those positions t in u that concern R_1, and that further concern an individual i for which $R_2 = r$.

With this we can define the following relative frequencies in the ordinary manner of definition (6.25), using the above definitions of subsequences:

$$f_{1|r2}(u) = \lim_{t\rightarrow\omega} \frac{1}{t} \sum_{t'=1}^{t} W_1(u^{1|r2}(t')). \tag{6.49}$$

The relative frequencies for R_k having $k > 1$ are here denoted with f_k. The above defines the relative frequency for R_1 conditional on the individuals having $R_2 = r$. Such a relative frequency can be called conditional.

Hypotheses for dependency relations. Note that the relative frequencies $f_{1|r2}$ concern only those observations with respect to R_1 for which the individuals i are also observed in terms of R_2 somewhere in the sequence u. However, there are many possible sequences in which individuals are observed in terms of R_1 somewhere down the line, but never in terms of R_2. In these sequences the overall relative frequency of results $R_1 = 1$ can differ from the relative frequency of results $R_1 = 1$ for those individuals for which R_2 is also observed. To avoid such problems, define the indicator function

$$W_{R_k}(u) = \begin{cases} 1 & \text{if } \forall i:\ W_{0k}(u,i) + W_{1k}(u,i) = 1, \\ 0 & \text{otherwise,} \end{cases} \tag{6.50}$$

and the hypotheses $H_k = \{u : W_{R_k}(u) = 1\}$. By fixing the prior to $p(H_{R_2}) = 1$, we rule out the problematic sequences alluded to above, since effectively we assume that every individual is observed with respect to R_2. Another, more elegant solution is to assume that there is at least a countable infinity of individuals that do get observed with respect to R_2, and to assume further that the relative frequencies for latter individuals are indicative for those individuals that are not observed with respect to R_2. However, for the purpose of this chapter the straightforward solution works out fine.

We are now in the position to define a partition for the Bayesian network with a single link. Imagine that we are observing a collection of variables R_k in individuals i, and that we want to find the exact statistical dependence of the family R_1 on the family R_2. We can then use the hypotheses $H_\theta = \{u \in M^\omega : W_{h_\theta}(u) W_{h_{R_2}}(u) = 1\}$ with

$$W_{h_\theta}(u) = \begin{cases} 1 & \text{if } \forall k > 1 : \; f_k = \theta_k \text{ and } f_{1|r2} = \theta_{1|r2}, \\ 0 & \text{otherwise.} \end{cases} \tag{6.51}$$

The parameter θ has components θ_k for $k > 1$, and further components $\theta_{1|02}$ and $\theta_{1|12}$. The total number of components in θ is therefore $n + 1$.

These hypotheses form the partition $\mathcal{B}$ for a Bayesian network with n nodes and a single link. Note that the components θ_k with $k > 1$ are similar to those in the partition $\mathcal{I}$. They range over spaces $B_k = [0, 1]$. The components $\theta_{1|r2}$ also range over such spaces, and may be denoted by $B_{1|r2} = [0, 1]$. So again we have a parameter space consisting of a product of subspaces:

$$B = B_{1|02} \times B_{1|02} \times B_2 \times \ldots \times B_n. \tag{6.52}$$

That is, we may define the marginal partitions $\mathcal{B}_k$ for observations concerning R_k with $k > 1$, and similarly the marginal partitions $\mathcal{B}_{1|r2}$ with $r = 0, 1$ for the observations concerning R_1 conditional on R_2.

6.5.2 Prediction rules

The above suggests that we may treat the conditional observations of R_1 given R_2^r for $r = 0$ and $r = 1$ as if they concerned separate variables, and thus allow for separate predictions. This subsection shows that the partition $\mathcal{B}$ indeed leads to a representation similar to the one elaborated in section 6.4, under the assumption that the observation sequences are restricted in a certain way, and under the further assumption that the prior density over B is Dirichlet.

Likelihoods. The relative frequencies defined above may be taken as the likelihoods for the observations. The components θ_k for $k > 1$ are the unconditional likelihoods of the predicate families R_k for any time t:

$$\forall k > 1 : \quad p(R_{kit}^r | H_\theta) = \theta_k. \tag{6.53}$$

But the likelihoods for R^r_{1it} depend on earlier observations. Specifically, the components $\theta_{1|r2}$ are the likelihoods of the variable R_1 conditional on the observation of $R_2 = r$:

$$S_{t-1} \subseteq R^{r'}_{2it'} \quad \Rightarrow \quad p(R^r_{1it}|H_\theta \cap S_{t-1}) = \theta_{1|r'2}. \tag{6.54}$$

For the likelihoods for observations concerning R_1, we cannot leave aside conditioning on the observations: the likelihood of observation R^r_{1it} depends crucially on the observation $R^{r'}_{2it'}$, for $t' < t$. For this reason the network partition $\mathcal{B}$ differs from the partition $\mathcal{I}$.

Some attention must now be given to the likelihoods for observations concerning R_1 if the individual involved is not yet observed with respect to variable R_2. There is no single parameter in the space B that corresponds to this likelihood. Instead we can write

$$S_{t-1} \not\subseteq R^{r'}_{2it'} \quad \Rightarrow \quad p(R^r_{1it}|H_\theta \cap S_{t-1}) = \theta_2\theta_{1|12} + (1-\theta_2)\theta_{1|12}. \tag{6.55}$$

This is a mixture of the parameters $\theta_{1|r'2}$ for $r' = 0$ and $r' = 1$, weighed with $1 - \theta_2$ and θ_2 respectively. Note that updating with this likelihood is a more complicated operation than updating with any of the other likelihoods: it involves updates over the spaces B_2, $B_{1|02}$ and $B_{1|12}$, and it therefore fails to meet requirement (6.36).

Conditions for deriving $\lambda\gamma$ rules. As a result, we cannot directly derive prediction rules for $\mathcal{B}$ in the manner of section 6.4. In order to derive these rules we must assume that objects are always observed in terms of R_2 before they are observed with respect to R_1. This latter assumption is a strengthened version of assumption (6.50). We must define the specific hypothesis $H_{R_1 \succ R_2}$:

$$W_{R_1 \succ R_2}(u) = \begin{cases} 1 & \text{if } \forall i : u(t) = r2i,\ u(t') = r'1i \text{ and } t < t' \text{ for} \\ & \text{some } r, t \text{ and } r', t', \\ 0 & \text{otherwise.} \end{cases} \tag{6.56}$$

In any sequence u for which $W_{R_1 \succ R_2}(u) = 1$, all individuals i are observed with respect to R_2 at some t, and with respect to R_1 at some $t' > t$. And again, by fixing $p(H_{R_1 \succ R_2}) = 1$, we consider only the observational algebra defined over the set of these sequences. We thus rule out all sequences u in which for some individual i the variable R_2 is observed only after the variable R_1, or in which for some individual i either R_1 or R_2 is not observed at all.

With this assumption in place, we only need to ensure that the prior over $\mathcal{B}$ conforms to requirement (6.40). As suggested, one way of doing so is by assuming the prior density to be a member of the Dirichlet class. The partition $\mathcal{B}$ then falls within the class of partitions for which the marginal partitions can be treated independently. Moreover, since the marginal densities are in that case Dirichlet as well, we can derive separate $\lambda\gamma$ rules for all the marginal partitions associated with the marginal partitions $\mathcal{B}_k$ for $k > 1$ and $\mathcal{B}_{1|r2}$.

Carnapian rules for Bayesian networks. Let me concentrate on the case in which the partition is uniform over $\mathcal{B}$, and therefore uniform over all marginal distributions, so that all the prediction rules are of the form pr^*. For the variables R_k with $k > 1$ this simply results in

$$p(R^r_{ki(t+1)}|S_t) = \frac{t_{rk}+1}{t_k+2} = pr^*(t_{rk}, t_k), \tag{6.57}$$

which is the same as in the partition $\mathcal{I}$. For the predicate family R_1, however, we have separate predictions rules for objects belonging to separate cells r' of R_2. But these rules are also of the form pr^*:

$$p(R^r_{1i(t+1)}|S_t) = \frac{t_{r1|r'2}+1}{t_{1|r'2}+2} = pr^*(t_{r1|r2}, t_{1|r2}), \tag{6.58}$$

where again $S_t \subset R^{r'}_{2it'}$ and $t' < t$. The numbers $t_{r1|r'2}$ and $t_{1|r'2}$ are defined analogously to t_{rk} and t_k, counting respectively the number of observations R^r_{1it} and the total number of observations concerning variable R_1, both conditional on the earlier occurrence of $R^{r'}_{2it'}$.

Finally, it is useful to determine the prediction $R^r_{1i(t+1)}$ conditional on the observations S_{t-1}, but before the observation $R^{r'}_{2it}$. To compute this prediction, we can virtually add the observation $R^{r'}_{2it}$ for all possible values of r', and then calculate the prediction for $R^r_{1i(t+1)}$ by the law of total probability, weighing the predictions concerning R_1 of (6.58) with the probabilities for $R^{r'}_2$ as given in (6.57):

$$p(R^r_{1i(t+1)}|S_{t-1}) = \sum_{r'} pr^*(t_{r'2}, t_2)\, pr^*(t_{r1|r'2}, t_{1|r'2}). \tag{6.59}$$

This expression enables us to compare the predictions of $R^r_{1i(t+1)}$ before and after the occurrence of $R^{r'}_{2it}$.

6.5.3 Illustration

Medical example. The above prediction rules are now illustrated for the example of the doctor with $n = 3$. It shows that the partition $\mathcal{B}$ indeed manages to capture the dependency structure between variables R_k. In particular it shows that while correlations of R_1 and R_2 are detected, correlations between R_3 and any other node are ignored.

Let me first specify the observations of the variables numerically. The first $t-1$ observations concern an even number $i-1$ of individuals, each observed for R_3, R_2 and R_1 respectively, and finally the individual i observed for R_3 only:

$$\begin{aligned} s_{t-1} \quad = \quad & \langle (131), (121), (111), (032), (022), (012), \ldots \\ & (13(i-2)), (12(i-2)), (11(i-2)), \\ & (03(i-1)), (02(i-1)), (01(i-1)), (13i) \rangle . \end{aligned} \tag{6.60}$$

Recall that the triples have the interpretation (rki). Note that the results $R_1 = R_2 = R_3 = 1$ and $R_1 = R_2 = R_3 = 0$ occur equally often in the observations until $t-2$. The last observation in the example is $R^1_{3i(t-1)}$, so that $t = 3i - 1$. Note that the individuals are observed in terms of R_2 before R_1, so that requirement (6.56) is met.

Assuming a uniform prior over B we can derive the prediction rule pr^* for the variables R_2 and R_3, and two conditional prediction rules pr^* for the variable R_1. Depending on the number of individuals i, we can express the predictions for the i-th individual to satisfy $R_1 = 1$ conditional on the observations S_{t-1} of expression (6.60), and conditional on the observations $S_{t-1} \cap R^2_{1it)}$ respectively. The table shows these predictions.

Number of objects i	1	3	7	15
$p(R^1_{1i(t+1)} \vert S_{t-1})$	0.500	0.500	0.500	0.500
$p(R^1_{1i(t+1)} \vert S_{t-1} \cap R^1_{2it})$	0.667	0.800	0.889	0.941

The effects of the correlation between R_2 and R_1 can be read off from the table: observing R^2_{1it} has a positive effect on the predictions of $R^1_{1i(t+1)}$. It can further be noted that the observation of variable R_3, which in the above data appears to be correlated with variable R_1 just as much as R_2, has no influence. This is because the independence structure of the Bayesian network, which deems R_1 and R_3 independent, is incorporated in the inductive predictions.

6.6 Inductive dependence

The above scheme captures the dependency structure of Bayesian networks. This section shows that the scheme also suggests specific dependencies that are not captured in such a network. These dependencies, as is explained below, may be called inductive dependencies. I do not intend to criticise the use of Bayesian networks for this failure to capture dependencies. Rather I aim to show some additional possibilities with the scheme developed in this chapter, which are not naturally captured in Bayesian networks.

6.6.1 General idea

Simplifying the partition. Let me present the case in its simplest possible form. Consider the above Bayesian network for the case of $n = 2$, consisting of R_1, R_2 and their connection. This network is associated with a partition $\mathcal{B} = \mathcal{B}_2 \times \mathcal{B}_{1|02} \times \mathcal{B}_{1|12}$, a unit cube just as $\mathcal{I}$ for $n = 3$. We may now imagine that each individual is subject to the conceptual restriction that if $R_2 = 0$ then automatically $R_1 = 0$. That is, if the disease is not found in an individual, $R_2 = 0$, we are sure that the individual does not have the symptom either, $R_1 = 0$. In terms of the probability over $\mathcal{B}$, we can express this by fixing the marginal probability $p(H^{1|02}_{\eta})d\eta = 0$ for $\eta > 0$, and by assigning all probability to the extreme case of $\eta = 0$. Effectively, we are then left with a reduced partition $\mathcal{B}' = \mathcal{B}_2 \times \mathcal{B}_{1|12}$, which reflects the fact that the medical doctor of the example is uncertain only of the probability for the occurrence of the disease R_2, and of the probability for the symptom R_1 given the occurrence of the disease, $R_2 = 1$.

Example of inductive dependence. Inductive dependence can now be illustrated in this simplified partition. Imagine that the doctor expects yet another dependence between the variables R_1 and R_2. We may imagine that there are in fact a number of subpopulations of individuals, perhaps geographically separated, and further that the doctor is sure to be sampling only from one of these, for example, because she is investigating individuals living in the same place. Imagine further that the subpopulations can be told apart in two ways. First, the incidence rate of the disease in the subpopulations differs, and second, the probability of the symptom conditional on the occurrence of the disease is proportional to this incidence rate. So she expects that the eventual relative frequency of $R_1 = 1$ conditional on $R_2 = 1$ is somewhere close to the relative frequency of $R_2 = 1$ itself. As an explanation for this, we may imagine that the same physiological mechanism that determines the chance to contract the disease for each individual is also responsible for the development of the symptom once the disease is contracted.

Such a dependence cannot be expressed in terms of an arrow of some kind, if only for the simple fact that it is a dependence that obtains between a node

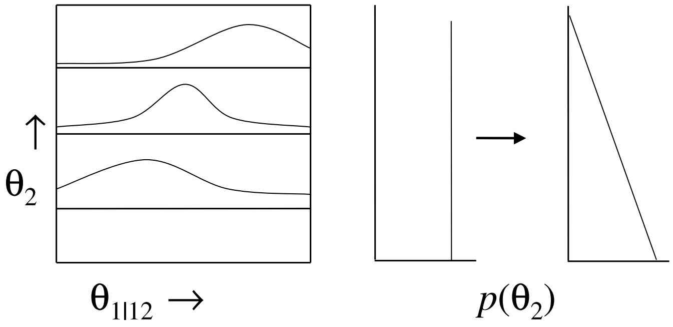


Figure 6.2: Updating the marginal probability $p(\theta_2)$ for an observation of R_2^0 also effects a shift in the marginal probability over $\theta_{1|12}$ towards lower values of $\theta_{1|12}$. This is because the conditional probability densities for lower values of θ_2 have most probability mass on lower values of $\theta_{1|12}$, while conditional probability densities for higher values of θ_2 have most probability mass on higher values of $\theta_{1|12}$.

and an arrow springing from this node. But more in general, the point is that the above dependence cannot be portrayed as a direct influence running from R_2 to R_1 in each individual. As illustrated in figure 6.2, finding an individual to be perfectly healthy, $R_2 = 0$, has the effect that the probability of developing the symptom becomes 0 for that individual. But quite apart from that effect, finding a healthy individual lowers the probability that the doctor is sampling from a subpopulation in which the incidence rate is high. And this also means that the probability to develop the symptom for those individuals that do have the disease is lower. Thus, finding healthy individuals lowers the probability of finding the symptom in two ways, both by lowering the expected incidence rate of the disease, and by lowering the probability of developing the symptom if the disease is contracted after all. This latter effect is not based on a causal dependence between disease and symptom in each individual separately, but rather on an inductive dependence at the level of subpopulations.

6.6.2 Varying conditional distributions

As noted, such a correlation cannot be expressed with the pictorial means offered in Bayesian networks. However, the scheme of this chapter does allow us to capture the dependence in terms of a prior probability density over $\mathcal{B}'$. This subsection provides the class of priors associated with the inductive dependence.

This class violates (6.40), and therefore we cannot derive a system of $\lambda\gamma$ rules for it. The predictions can only be illustrated with a discrete approximation.

Using marginal and conditional densities. The definition of the class of priors makes crucial use of the notions of marginal and conditional density. Before defining the prior itself, let me briefly specify these notions for the present case. The space B may be parameterized with θ_2 and $\theta_{1|12}$. These parameters lie in the separate spaces B_2 and $B_{1|12}$. The parameter space is therefore a unit square. If we identify $\eta = \theta_2$ and $\zeta = \theta_{1|12}$, we can write the density over this unit square as

$$p(H_\theta|S_{t-1}) = p(H^2_\eta|S_{t-1})p^2_\eta(H_\zeta|S_{t-1}). \tag{6.61}$$

Recall further that the update with $R^{r''}_{2it}$ consists in an update operation of the marginal probability $p(H^2_\eta)d\eta$ over B_2. The conditional probability assignments $p^2_\eta(H_\zeta)d\zeta$ remain unchanged during this update. But we may still change the marginal density $p(H^{1|02}_\eta)$ implicitly by updating over $p(H^2_\eta)d\eta$.

The point to note is that we can establish the dependence between these marginal densities by choosing varying conditional densities p^2_η. Specifically, we can choose the conditional density with higher values of η, signalling a higher relative frequency for $R_2 = 1$, to have more probability allocated at higher values of ζ, signalling a higher relative frequency for $R_1 = 1$. Similarly, conditional densities with lower values of η must have more probability allocated at lower values of ζ. An update with R^1_{2it}, which changes the density over B_2, then also changes the weights for the differing conditional distributions. Because the predictions for $R^1_{1i(t+1)}$ are given by a weighted average of the predictions within these conditional distributions, they are thereby altered as well.

Defining a class of twisted priors. With this idea in mind I can make the class of priors precise. For simplicity, assume that all conditional densities have a Dirichlet form dependent on η,

$$p^2_\eta(H_\zeta) \quad \sim \quad (1-\zeta)^{a_0(\eta)}\zeta^{a_1(\eta)}. \tag{6.62}$$

This means that every marginal hypothesis H^2_η is associated with a conditional density that results in predictions according to a $\lambda\gamma$ rule. As in the foregoing, the parameters λ and γ are determined by the relations $\lambda_{1|12}(\eta) = a_0(\eta)+a_1(\eta)$, and $\gamma_{r1|12} = a_r(\eta)/(a_0(\eta) + a_1(\eta))$. There are no restrictions on the marginal density $p(H^1_\eta)$ itself, or on the functions $a_r(\eta)$.

For this class of priors we can directly express how the predictions for $R^1_{1i(t+1)}$ depend on the marginal probability $p(H^2_\eta|S_t)d\eta$ for the case that $S_t = R^1_{2it} \cap$

S_{t-1}:

$$\begin{aligned} p(R^1_{1i(t+1)}|S_t) &= \int_{B'} p(R^1_{1i(t+1)}|H_\theta \cap S_t)\, p(H_\theta|S_t) d\theta \\ &= \int_{B'} \zeta\, p(H^2_\eta|S_t)\, p^2_\eta(H_\zeta|S_t)\; d\zeta d\eta \\ &= \int_{B_2} p(H^2_\eta|S_t)\, \left[\int_{B_{1|12}} \zeta\, p^2_\eta(H_\zeta|S_t) d\zeta\right] d\eta \\ &= \int_{B_2} p(H^2_\eta|S_t)\, \left[\frac{t_{11|12} + \lambda_{1|12}(\eta)\gamma_{11|12}(\eta)}{t_{1|12} + \lambda_{1|12}(\eta)}\right] d\eta. \quad (6.63) \end{aligned}$$

Note that instead of the two parameters $\lambda_{1|12}$ and $\gamma_{11|12}$, we now have a continuum of $\lambda_{1|12}(\eta)$ and $\gamma_{11|12}(\eta)$, each associated with a conditional density p^2_η. The predictions for $R^1_{1i(t+1)}$ are thus an average over a continuum of $\lambda\gamma$ rules, weighted with the marginal density over B_2.

6.6.3 Approximated predictions

I cannot give analytic expressions for the predictions that derive from the above class of priors. In particular, I cannot express the updates and predictions as simple combinations of $\lambda\gamma$ rules. To illustrate these predictions, this section works with a discrete approximation of the marginal density and its integrals.

Discrete approximation. Instead of a continuum of marginal hypotheses H^2_η, we may consider a finite number of marginal hypotheses H^2_j, with $0 < j \leq N$. These marginal hypotheses may be associated with the following likelihoods:

$$p(R^1_{2it}|H^2_j \cap S_{t-1}) = \frac{2j-1}{2N}. \quad (6.64)$$

Further take each of these marginal hypotheses H^2_j to be linked to a continuous conditional density over $B_{1|12}$ from the Dirichlet class. These densities all result in a specific prediction rules for the predicate family R_1, which are separately characterised by $\lambda_{1|12}(j)$ and $\gamma_{r1|12}(j)$. Note that these $\lambda\gamma$ rules may all be different. The resulting system then is a complete discrete approximation of the continuous scheme sketched above. Where the above equations show integrations over B_2, the discrete approximation has summations over all values of j. By choosing larger values for N we can make the discrete approximation more accurate.

With this we have implicitly defined a hyper-Carnapian prediction rule for the predicate family R_1: the predictions for this family are a mixture of the different $\lambda\gamma$ rules. Furthermore, updating with R^r_{1it} involves adapting the prediction rules themselves, but also multiplying the weights assigned to these different rules with the respective predictions that these rules gave, after which the new weights must be normalised. The difference with an ordinary hyper-Carnapian prediction rule is that the weights over the different $\lambda\gamma$ rules for R_1 are not only determined by these observations with respect to R_1, but also by the observations with respect to R_2. Specifically, updating with $R^{r'}_{2it'}$ involves multiplying the weights with the likelihoods of equation (6.64). The present scheme is therefore slightly more complicated than the hyper-Carnapian rules.

Results. Let me present some numerical results using the approximation. In the following I take the prior over the hypotheses H^2_j to conform to the density $p(H^2_\eta) = \eta$, associated with $\lambda_2 = 3$ and $\gamma_{12} = 2/3$. For $N = 20$ this entails $p(H_j) = (2j-1)/400$. Further I have chosen $\lambda_{1|12}(j) = 5$ constant. Finally, I have chosen varying $\gamma_{1|12}(j) = 1/6 + (2j-1)/80$. This linear function encodes the expectation that a larger relative frequency for $R_2 = 1$ is associated with a larger relative frequency for $R_1 = 1$. At the same time it ensures that the predictions are the same for all combinations of values for R_1 and R_2.

The table shows the predictions of $p(R^0_{2it}|S^s_{t-1})$ and $p(R^1_{1i(t+1)}|S^s_{t-1} \cap R^1_{2it})$ for the case in which earlier observations s_{t-1} consists of $i-1$ individuals that all showed $R_2 = 0$:

Number of individuals i	1	2	4	8	16
$p(R^0_{2it}\|S^s_{t-1})$	0.33	0.50	0.67	0.80	0.89
$p(R^0_{1i(t+1)}\|S^s_{t-1} \cap R^1_{2it})$	0.50	0.58	0.67	0.73	0.78

The predictions show the effects discussed in the preceding subsection. Finding $R_2 = 0$ in a number of individuals increases the probability of $R_2 = 0$ in further individuals, and also increases the probability of $R_1 = 0$ in individuals for which we do find $R_2 = 1$.

It may be clear from the above construction that many more variations on the same theme are possible. However, I cannot provide a systematic treatment of these possibilities in this chapter. Note that there are no restrictions on the number of Q-predicates, or on the combinations of R-predicates that are supposed to underly them. Moreover, there are no restrictions on the observations that serve as input to the hypotheses schemes defined above. Any resulting prediction rule is exchangeable, as will be more elaborately discussed below, and any prediction rule eventually converges on the relative frequencies exhibited in the observations.

6.7 Generalising to multiple links

The following summarises the above scheme. It further argues that the scheme can be generalised to cover predictions based on any Bayesian network. Furthermore the assumption on the order of observations, expressed in equation (6.56) is reconsidered.

Summary of the schemes. In the scheme presented, a prediction for an individual i concerning the values r of a variable R_1 may depend on the value r' of the variable R_2. Such a dependence reflects the connection between the nodes R_1 and R_2 in a Bayesian network that is supposed to underlie the observations. The foregoing suggests how we can construct hypotheses that keep track of such dependence in a systematic way: the partition $\mathcal{B}$ treats observations R^r_{1it} separately for all possible earlier observations $R^{r'}_{2it'}$, by defining separate relative frequencies for them, and associating these frequencies with separate dimensions in the parameter space B.

Under the assumption of a certain class of priors, we can treat updates over the separate marginal densities as orthogonal. That is, an update over one marginal density leaves the marginal density in any other direction unchanged. This opens up the possibility to write down the updates as separate prediction rules, and by choosing the priors to be Dirichlet, to turn the hypotheses schemes into simple systems of $\lambda\gamma$ rules. It must be noted, however, that we are not forced to assume the priors leading to these latter rules. The scheme allows for other priors leading to predictions for the variables that fall outside the $\lambda\gamma$ continuum. As has been illustrated in the preceding section, this allows us to model other kinds of dependence that are not captured by Bayesian networks.

Some generalisations. The following considers generalisations of the number of connections in the Bayesian network, and further generalisations on the nature of the observations.

First, nothing precludes the use of more than one link in the network. The scheme can model predictions based on Bayesian networks that are much more complex than the ones discussed above, also ones that involve more than one link arriving at or departing from the separate nodes. In the case of three binary variables, for example, we can keep track of a dependence of the first variable on the third and the second, while the second is itself again dependent on the results with respect to the third. The parameter space for that partition is

$$B = B_3 \times B_{2|03} \times B_{2|13} \times B_{1|02,03} \times B_{1|12,03} \times B_{1|02,13} \times B_{1|12,13}, \qquad (6.65)$$

and all these subspaces can again be associated with separate prediction rules. Any causal structure can be captured by a network partition in this way. The only restriction is that the likelihoods of a hypothesis given some variable cannot be made dependent on the observation of this variable. In terms of Bayesian networks, this restriction comes down to the network being acyclic.

Another direction of generalisation concerns not the partitions, but the observations. It must be stressed that the hypotheses scheme picks up on the correlations in the examples not just because of the rather straightforward way in which these are present in the observations. We may define s_{t-1} to be a completely messy sequence, in which observations of R_k follow on each other irregularly. The scheme will always detect the conditional dependencies that are fixed in the Bayesian network. The point here is that the data need not reveal the relations between the predicate families as clearly as in the examples.

Ordered variables. One assumption on the observations asks for special attention. The representation of the hypotheses scheme in terms of a system of $\lambda\gamma$ rules is based on assumption (6.56). This assumption takes care that those observations for which we have defined conditional relative frequencies in the partition must occur after the observations that serve as conditions. On the basis of that we can derive independent prediction rules. The first thing to note is that hypotheses schemes using partitions such as $\mathcal{B}$ are in themselves not restricted to observations with a specific order. The assumption on order is here made for computational simplicity: it makes possible the derivation of separate prediction rules. The predictions resulting from network partitions are in themselves exchangeable. This can be seen from the fact that the update operations over the partition $\mathcal{B}$ are multiplicative, and therefore commute with each other.

Nevertheless, the above prediction rules do not apply if for some set of individuals we have already observed R_1, while we have not yet observed R_2. This lack of computational tractability is a drawback. Moreover, it is unfortunate that we can only illustrate that predictions based on $\mathcal{B}$ capture Bayesian networks once we have made assumption (6.56). One possible resolution of these drawbacks is suggested by section 6.3.2, where it is noted that the order in which the variables R_k are observed is not itself subject to the hypotheses. We may suppose that the observer is free in choosing the order in which she makes the observations. If this is so, the observer is simply helping herself to an easy predictive task by collecting the observations in a convenient order. Even stronger, by encoding a set of conditional dependencies in a specific Bayesian network, the observer may decide on the direction of the dependencies largely by herself. And these directions can be made to accord with the order in which the variables are observed. The order restriction thus ties in neatly with the interpretation of the networks as causal.

Analogy and inductive dependence. One general remark on the relation between this chapter and the two preceding chapters concludes the second part of this thesis. It may be noted that there is a close link between analogical predictions for explicit similarity and predictions for Bayesian networks that employ a factorisable prior, and also between more complicated analogical predictions and nonfactorisable priors. Specifically, the more complicated models of analogical predictions discussed in chapter 5 employ priors that cannot be controlled in terms of the simple analogy partitions $\mathcal{A}$. This chapter shows that, when starting with a fixed algebra $\mathcal{R}$ of underlying predicates, these more complicated analogical predictions indeed have a different nature: they are based on a different kind of dependence, namely inductive dependence.

III
Philosophy of Science

7

Induction in the Bayesian Scheme

This chapter discusses the problem of induction in view of the Bayesian scheme developed above. Three levels of the problem are disentangled: the levels of single observations, of general patterns in the observations, and of structure behind the observations. In line with the logical view developed in chapters 1 to 3, the use of hypotheses in the Bayesian scheme does not suggest anything towards solving the problem of induction. However, we can solve the problem at the level of predictions if we make specific assumptions at the level of patterns. A similar solution can be advanced for the patterns on the basis of assumptions at the level of structure. Moreover, the conclusions about the patterns may be transferred back to refine these assumptions further. An example on the categorisation of substances by means of observations completes the chapter.

The philosophical discussion in chapter 3 is essential for understanding the main line of this chapter. The technical details of that chapter are less important. Some knowledge of chapter 1 and, to a lesser extent, chapter 2 may also be useful.

7.1 The problem of induction

The problem of induction is one of the most pervasive in philosophy since Hume posed it in 1739. It concerns the impossibility to attain knowledge of future observations on the basis of past observations. The following introduces the problem, and connects it to the Bayesian scheme of chapter 1.

Three levels of induction. The problem of induction in fact concerns problems on three different levels. As an example, take $q = 0, 1, 2, 3$ as observations of wet, cold, warm and dry respectively, and consider these observations as sensations recorded by the feet of a duck. We may imagine that the duck is confronted with the following observations,

$$001110123332301001032221100.$$

We can focus on three different levels of the apparent connection between the pairs $\{0, 1\}$ and $\{2, 3\}$ in these observations:

OBSERVATION the last observation of 0 is probably followed by a 0 or a 1;
PATTERN most observations of 0 are followed by a 0 or a 1;
STRUCTURE some structure behind the observations makes 0 and 1 occur in contiguity.

Similar considerations apply to the pair $\{2, 3\}$, and possibly also to the transitions between the pairs.

In terms of words characterising the observations, the levels concern the statements that the last observation of wet is probably followed by an observation of cold or wet, that wet is generally followed by cold or wet, and that some structure in the world connects cold and wet in this way. The first level is completely observational. The second level is observational in the sense elaborated in chapter 2. It concerns patterns in the observations that can only be checked at infinity. The last level is entirely theoretical, as it concerns structure that will always remain hidden behind the observations. It seems that claims on the theoretical level are most difficult to defend on the basis of observations. However, a structure behind the observations is usually the main reason for expecting some pattern or single observations. In the absence of an argument for some structure on that level, it may therefore be argued, on the level of patterns, that nothing can support the apparent conjunction of 0 and 1. Even at the level of observations, we may argue that after the last 0 we have no reason to expect the occurrence of 1 or 0 more than the other two results.

Focus on the observational levels. The problem of induction eventually concerns all three levels, but the focus of this chapter is on the two observational ones, and on the theoretical one in a derived sense only. More in particular, this chapter considers the relation between the problem of induction, as trisected above, and the Bayesian scheme of the preceding chapters. It is easily seen that these schemes only concern the two observational levels. They employ partitions of hypotheses to make predictions. The predictions connect naturally to the level of single observations, and the hypotheses connect to the level of patterns. But at first glance there is no obvious relation between the Bayesian scheme and the theoretical level.

At the two observational levels, this chapter argues for two claims. The first is that the Bayesian scheme itself does not provide the means for solving the problem of induction. Rather it provides a framework within which it becomes clear what means are actually required. In a metaphor, the Bayesian scheme presents a toolkit, but it does not also provide the architectural plans or the raw material for building inductive knowledge. In terms of this metaphor, the second claim concerns the relation between architectural plans and material. In the Bayesian scheme, specific roles are assigned to the observations, here presented by sequences of natural numbers, and the partitions of hypotheses. The second claim is that these partitions are comparable to the architectural plans, and that relative to these plans, the raw material of observations present definite restrictions to the building. The form of the building is thus not entirely determined by the observers, in the role of the architects. Once they have provided the plans, the raw material of observations completely determines the building. Note that inductive knowledge is thus not completely subjective, as perhaps suggested by the first claim, but rather a co-production of the observer and the observed.

Induction on the theoretical level. A further aim of this chapter is to determine whether the Bayesian scheme can also be connected to the problem of induction at the theoretical level. For this I explore the relation between assumptions concerning patterns and the structure of underlying systems. I first show that supposing some structure may justify an inductive assumption, and therefore the use of a particular partition: if we assume that a system with a specific structure generates the observations, we may expect the observations to show a pattern that is related to the structure. With this assumption on structure, and using the aforementioned function of observations, I argue that conclusions concerning patterns may be transferred back to the theoretical level, to narrow down the assumed structure to a more specific one. This is illustrated by the example: if we assume that the world consists of separate substances, we can motivate a partition that picks up on the connection between wet and cold and between warm and dry. This leads to the more specific structure of two substances, more specifically, water and air.

There is a more ambitious perspective on the claims defended in this chapter. I will not assume this general perspective in the chapter, but since it has played some role in its development, it may be of interest to mention it. According to this perspective, the inductive inferences of a Bayesian scheme can be employed in a specific tactic for building up inductive knowledge. This tactic consists in a repeated application of inductive inferences on the theoretical level. In brief, the idea is to start with a minimal assumption concerning the structure of the system under investigation, for example with the assumption that on a certain time scale the states of the system show an auto-correlation, and to narrow down this supposed structure by inductive inference. The structure of the system arrived at may then be refined with additional and tailor-made assumptions, from which further refinements can be derived, and so on. This allows us to build up inductive knowledge of the system at the cost of minimal assumptions. The details of such a tactic, however, fall outside the scope of this thesis.

Disclaimers. Let me also mention two other topics that are left aside. First, the literature provides numerous attempts to deal with the problem of induction. To give a short list, the reader may consult Popper (1959), Armstrong (1973), Papineau (1987), Howson (2000), and Norton (2004). The present chapter differs from most of these attempts because it does not provide a method of justified inductive inference together with a method of finding the correct input for the inferences. Apart from that, the present chapter focuses specifically on the Bayesian scheme presented in section 1.3. It does not relate this scheme to the conceptual schemes employed in other solutions. Consequently, the chapter does not argue that the Bayesian scheme captures the problem better than these other schemes. Mostly, the chapter clarifies the relation between the problem of induction and the Bayesian scheme in order to complete the logical picture begun in chapter 1.

Second, it must be noted that some of the idealising assumptions of the framework, which were discussed in 1.5, may here become disturbing. For one thing, the chapter talks of the raw material of observations, but such raw material does not exist. If we consider the observations of scientific experiment, it becomes apparent that a lot of effort goes into the construction of reports of observations. In this thesis such reports are simply denoted with e_t, falsely suggesting that they fall ripe from the trees. The chapter also talks of assumptions on structure which are to some extend uncertain. But this uncertainty may not be adequately expressed in a probability assignment over an observational algebra. However, I cannot deal with the problems that may arise from these two idealisations in the context of this chapter.

7.2 Hypotheses as tools

The following discusses whether the Bayesian scheme on itself suggests anything towards solving the problem of induction. I argue that it does not, and that instead it reveals the need for inductive assumptions.

7.2.1 Indispensable assumptions

Projectability assumptions. It is tempting to consider the Bayesian scheme as a solution of the problem of induction on the level of observations. It may seem that choosing a partition does not import any substantial assumptions, because the partition covers the whole space K^ω, and is therefore equal to a tautology. Moreover, because of the convergence results of Gaifman and Snir (1982) it may be argued that the prior over this partition does not present an assumption either. A Bayesian scheme that uses open-minded probability assignments can thus be seen as a fruitful cooperation of observations with a completely innocent partition of hypotheses. However, as I have argued in chapter 3, predictions based on a partition are always made at the cost of specific inductive assumptions. Whereas for deductive purposes the partitions can perhaps be deemed innocent, for inductive purposes they introduce projectability assumptions. The partitions reveal, within the inductive scheme, the assumptions needed to get to inductive predictions.

Finding weak assumptions. The present chapter elaborates in the claims of chapter 3. Note that the specific projectability assumptions alluded to in the foregoing are stronger than the unqualified assumption of the uniformity of nature. Assuming the general uniformity of nature leaves unspecified the kind of pattern with respect to which nature is uniform. But for inductive predictions, what is needed is a uniformity assumption with respect to a specific set of patterns, as revealed in statistical hypotheses. It may be noted that the Humean problem of induction is in this sense the protoversion of the so-called new riddle of induction proposed by Goodman (1955: 59-81): a general uniformity assumption may solve the problem as Hume conceived of it, but eventually such a general uniformity will not do.

On this point it may be objected that some projectability assumptions are more specific than others, and that for this reason the Bayesian scheme may solve part of the epistemological problem of induction after all. In particular, we may try to weaken the projectability assumptions as far as possible with the tools offered in the Bayesian scheme, to arrive at a formal expression of the general uniformity assumption. The first option is to use an entirely impartial partition, which does not preselect any kind of pattern in advance. If such a partition is possible, it can be argued that the predictions based on this partition assume just the overall uniformity of nature, or perhaps no uniformity at all. The second option is to consider all projectable patterns simultaneously. These options will now be investigated.

7.2.2 General uniformity

Using no projectable predicates. To assess the first option, consider the Bernoulli partition $\mathcal{B}$ of chapter 3, with the likelihoods $p_{[e_t]}(Q^q_{t+1}|H_\theta) = \theta_q$. The sufficient statistics for this partition are the numbers of times that q occurs in the sequences e_t, denoted t_q. The predictions based on the partition $\mathcal{B}$ therefore focus on a particular pattern in the observations, namely the relative frequencies of the results q, and the use of $\mathcal{B}$ comes down to assuming the projectability of this pattern. Now we may generalise this way of identifying projectability assumptions: as long as the sufficient statistics of a partition do not at all times coincide with the complete sequence of observations e_t, the partition focuses on some pattern in the observations, and therefore must employ some kind of projectability assumption.

With this way of identifying projectability assumptions, it is fairly easy to construct a partition that contains no such assumptions. There is only one partition for which the sufficient statistics always coincide with the complete sequence of observations. In this partition, here denoted with $\mathcal{E}$, the hypotheses H_e are the singletons $\{e\}$. The likelihoods of these hypotheses for observations Q^q_t are defined by $p_{[e_t]}(Q^q_{t+1}|H_e) = 1$ if $e(t+1) = q$ and $p_{[e_t]}(Q^q_{t+1}|H_e) = 0$ otherwise. With this partition I deal in detail below. The conclusion for now is that, apart from the limiting case $\mathcal{E}$, there is no partition that does not carry specific projectability assumptions.

Using all projectable predicates. The other option for weakening the strong uniformity assumption is to generalise it, by simultaneously using all conceivable partitions. An ambitious Bayesian may argue that a partition must encompass all hypotheses that can be formulated in the current language, in this case presented by the observational algebra. The idea is that, as long as nothing is known about the observations, none of the possible patterns can be excluded. The partition must therefore focus on all possible patterns, corresponding to the general uniformity assumption that there is some, as yet unspecified, pattern in the observations. However, given the observational algebra, we can always find some observation Q^q_t that tells apart any two infinite sequences e and e'. Therefore, it seems that the partition that encompasses all hypotheses that can be formulated in the given language is again the limiting case mentioned in the preceding paragraph, the singleton partition $\mathcal{E}$. I will now concentrate on this limiting case.

The uninformative singleton partition. Note first that the predictions resulting from $\mathcal{E}$ are determined entirely by the prior over the partition, $p_{[e_0]}(H_e)$. Moreover, the likelihoods of the separate singleton hypotheses are all extremal. With every new observation Q^q_t, conditioning over $\mathcal{E}$ simply means that all singleton hypotheses $H_{e'}$ for which $e'(t) \neq q$ are discarded, and that all those singleton hypotheses H_e for which $e(t) = q$ stay in. Because of this, conditioning over the hypotheses in $\mathcal{E}$ does not in itself give any comprehensive information on oncoming observations, so that the singleton partition does not carry any projectability assumption. In this sense the partition $\mathcal{E}$ meets the requirements.

On the down side, it must be noted that the task of determining the prior over the singleton partition becomes a rather laborious and obscure one: every single e must be given a probability separately, and it is not immediately clear what repercussions the assignments have for the predictions resulting from the partition. Moreover, the task of determining this assignment is very similar to the choice of the prior in a Carnapian scheme, as the singleton partition also requires us to specify the prior directly and as a assignment over the whole of the algebra $\mathcal{Q}$. It seems that we are back where we started: the Carnapian scheme employs projectability assumptions just as well as hypotheses schemes do. For the partition $\mathcal{E}$ the assumptions simply remain completely implicit in the probability over $\mathcal{E}$. In other words, the attempt to find a partition that expresses an impartial or generalised projectability assumption has pushed this assumption out of sight.

7.2.3 Logical solution

No formally motivated projectability. Thus far the discussion suggests that the tools provided in the Bayesian scheme do not offer any help in solving the problem of induction. The scheme expresses the need for assumptions underlying the predictions, but it does not suggest any natural or minimal assumption. However, in any experimental setting there may be independent reasons for specific inductive assumptions. For a realist, the projectability may be based on some suppositions on underlying structure, such as natural kinds, or on a process or mechanism that generates the observations. For an empiricist, on the other hand, the inductive assumption implicit in the use of some partition is perhaps nothing more than the empirical generalisations that they express. The thing to note is that these reasons are not supported or motivated by the tools that are offered by the Bayesian scheme. In sum, conditioning over partitions provides useful insight into the problem of induction, but we cannot solve the problem with an appeal to the formal aspects of partitions.

Projectability as premise. The above conclusions are in line with what I like to call the logical solution to the problem of induction. This solution has recently been proposed by Howson (2000), but it has its roots in Ramsey and De Finetti. The same solution is in fact implicit in many chapters arguing for local as opposed to global induction in Bogdan (1976), in the contextual approach of Festa (1993), and in a sense in Norton (2003).

The negative part of this solution is that, taken on itself, the problem of induction cannot be solved. Predictions must be based on inductive assumptions, and there is no way of deciding over these assumptions by formal or other a priori means. In the above metaphor, we cannot build a house just by buying nice tools, because we also need a building plan, and apart from that bricks, planks and mortar. The positive part of the logical solution is that once the inductive assumptions, the building plans, are made, a Bayesian logician can tell how to deal with the observations, that is, the bricks and planks. Bayesian updating functions as a consistency constraint, and generates predictions from the assumptions and observations together. It is inherent to this view that there is nothing inductive about Bayesian conditioning itself. It merely links inductive assumptions with observations to render the inductive predictions consistent with these assumptions.

This thesis can be seen as a further elaboration of the logical solution to the problem of induction. It shows how partitions provide access to inductive assumptions in a Bayesian scheme. Moreover, from this perspective this thesis is a starting point for dealing with a host of other philosophical problems. For example, ordinary life and science show that humans and other animals can be quite skilful in making inductive predictions. Peirce's suggestion that we guess efficiently, is deeply unsatisfactory as an explanation of this skill. The present discussion suggests that in a logical picture of these skills, the essential component is the selection of interesting aspects of the observations, as laid down in the choice of a partition. However, as illustrated by Chihara (1987), the complexity of actual inductive practice leaves us with little hope for a unified theory of choosing partitions.

7.3 Induction as co-production

In this section I sketch how, in combination with the Bayesian scheme, the foregoing leads to a view of inductive knowledge as partly determined by the observer, and partly by the observed. First I briefly discuss the division of labour in a Bayesian scheme between observations and inductive assumptions.

After that I discuss how inductive assumptions relate to suppositions at the level of structure. Finally, I use the conclusions of these two discussions to elucidate the respective roles of the observer and the observed in building up inductive knowledge.

7.3.1 The role of observations

Observations and inductive assumptions. To characterise the respective roles of observations and assumptions, let me first draw together some insights from the first part of this thesis.

In choosing a partition of statistical hypotheses, we select a collection of likelihood functions, which were seen to be connected to probability models for the observations. The fact that we limit attention to those probability models makes for an inductive assumption. However, with the frequentist interpretation each model also corresponds to a specific collection of infinite sequences. In light of this, the inductive assumption is simply that the eventual sequence of observations, denoted e^*, is included in the sequences of observations covered by the partition. If we employ some partition, the general characteristic in the probability models associated with that partition is assumed to be a characteristic of the actual sequence of observations. For example, hypotheses from the Bernoulli partition $\mathcal{B}$ contain sequences e in which observations have constant chances. Using the partition $\mathcal{B}$ thus amounts to assuming that this characteristic is true for the actual observations e^*. That is, the sequence e^* is assumed to have limiting relative frequencies.

I now come to the role of the observations in relation to the inductive assumptions as determined by the partition. This partition itself is chosen by the observer. Because there are no further guidelines for choosing this partition, the predictions resulting from it may be considered subjective. But once the partition is chosen, it is left to the observations to select the best fitting model from the collection of models associated with the partition.

Soundness and completeness. With the idea of partitions as premises in mind, I can briefly elaborate on the soundness and completeness of inferences in the Bayesian scheme, which is also alluded to in chapter 1.

Note first that the convergence result of Gaifman and Snir (1982) guarantees that if we assume the correct partition, that is, if the sequence of actual, real world observations e^* indeed has the appropriate limiting relative frequencies, then the observations are going to take care that the probability assignment will converge onto the correct hypothesis within this partition. In other words, on the assumption that the chosen partition contains the true hypothesis, conditioning leads to this hypothesis and to its associated predictions, or in brief, the inferences in a Bayesian scheme are such that true assumptions lead to true conclusions. This reformulation suggests that we are here dealing with an informal soundness result for conditioning in inductive Bayesian logic. Note that this soundness differs from the soundness proved in Howson (2000), which concerns Bayesian logic more generally, and which involves a subjectivist interpretation of probability. The soundness suggested here specifically concerns the inductive Bayesian logic in which the probability assignments are in part interpreted in a frequentist manner.

As for the completeness of this inductive Bayesian logic, note that any probabilistic pattern can be incorporated as a hypothesis in a Bayesian scheme. And because any such hypothesis can therefore be learned, the Bayesian scheme may be called complete. Now apart from the fact that both soundness and completeness are here treated only very sketchy, two remarks are called for. First, it is worth noticing the negative results of Putnam (1963), who showed that relative to a given learning algorithm, specific patterns in the observations can never be learned. In terms of the Bayesian scheme, the point is again that there is no single inductive assumption that covers all projectable patterns. Second, recall that the class of frequentist hypotheses covers any probabilistic pattern that corresponds to a real world structure by frequentist standards. With the frequentist restriction, the Bayesian logic may not be complete anymore.

7.3.2 The theoretical level

From structure to partition. The foregoing specifies the role of observations at the first two levels of the problem of induction. This subsection concerns the theoretical level, and more in particular the relation between inductive assumptions on the one hand, and supposition on underlying structure on the other. But let me stress first that the discussion takes for granted that there is some system, or on a larger scale, a world, from which observations originate. That is, radical forms of empiricism, as in Mach (1906), are left aside here. I simply assume that it makes sense to speak of a world that produces the observations.

Consider structures at the theoretical level in relation to patterns in the observations. The general idea in the following is that assumptions on the observational level may reflect such underlying structure. If, for example, we know that the state of the underlying system is independent of preceding states, the corresponding inductive assumption consists in a collection of models in which the observations have constant chances, which comes down to using the partition $\mathcal{B}$. If, alternatively, the state of the underlying system depends on the state of the system directly preceding it, the corresponding assumption consists in a collection of Markov models, that is, models in which the probability of observations depends on the observation directly preceding it. In each of these examples, the underlying structure is associated with some shared characteristic of the probability models that make up the inductive assumption. Suppositions on the structure of the underlying system can thus motivate a limitation of the set of probability models, or in other words, suppositions on structure can motivate the choice of a partition.

From partition back to structure. It must be noted that the above relation between patterns and structure is in many ways incomplete and idealised. At bottom it concerns the relation between observational generalisations on the one hand, and mechanisms and causal workings in nature on the other, as discussed by Cartwright (1999), Kuipers (2000), and Van Fraassen (1989) among many others. One of the key points in this discussion is that there is no clear translation that brings us from a supposition on structure to the associated observational generalisation or pattern, and that in a similar way suppositions on structure are underdetermined by generalisations. The step from a collection of observational models to a mechanism always involves additional assumptions or criteria, as provided by unification or explanatory force. Similarly, once we have imagined some structure underlying the observations, there is generally not a unique way in which this translates to a collection of possible patterns, and thus to a partition.

Nevertheless I want to suggest how inductive inferences can be employed at the theoretical level. The idea is that in taking some structure to underlie a partition, we effectively decide that the observations have a specific meaning on the theoretical level as well. Conditioning on the observations narrows down a partition to a single hypothesis, and to its corresponding probability model. The claim is that this single hypotheses can be transferred back to the theoretical level to specify the structure of the underlying system further. That is, if we have based the inductive assumptions on a supposition concerning underlying structure, we are allowed to narrow down this structure further according to the conclusions of the inductive scheme. In short, the observations are made relevant to the theoretical level. It is notable that a similar idea has recently been developed in Douven (2005), who considers realist descriptions of experimental observations besides strictly empiricist ones. The import of observations is then determined by the way in which we choose to describe these observations.

7.3.3 Externalism

Locating the suppositions. It is instructive to compare the above perspective on the problem of induction with the perspective of Hume. Hume notes that inductive inference always involves the ascription to the world of necessary or causal connections, and then claims that the ascription of such connections is based on unjustifiable habit. A first difference is that the Humean perspective focuses on necessary connections, whereas the present perspective concerns structures in general, including causal connections but also including substances, mechanisms and the like. But a more important difference concerns the location of the supposition on structure. The Humean perspective locates the supposition of structure primarily in the cognitive faculties: the causal connections are assumed to be projected onto the observations by the observer. In the perspective of this thesis, by contrast, the suppositions pertain to the structure of the outside world. Therefore, in this thesis inductive knowledge rests not on a sheepish habit in cognition, but rather on the assumed presence of a structure in the world.

Reliability and truth. This may look like a rather strange perspective. After all, getting to know the world is supposed to be the primary aim of inductive inference, and in the perspective of this thesis we seem to presuppose knowledge of this world. However, we do not exactly presuppose justified knowledge in order to set the inference machinery in motion. First of all, the inductive schemes are aimed at characterising valid inductive inference, and that the validity of the inferences can be secured independently of a justification of the suppositions on structure. The inferences can therefore be used in an externalist view on inductive knowledge. In this view, knowledge of the observations ultimately hinges on the reliability of the observation methods, which is a contingent fact whose truth depends on the world. And similarly, knowledge of the projectable patterns hinges on the truth of specific suppositions concerning structure. So in order to have inductive knowledge, we do not need to justify the suppositions on structure. They just need to be true.

7.4 Suppositions on substance

In this section I illustrate the above considerations with an example, using the observations of the duck presented in section 7.1. I first discuss the general suppositions about structure, by which I motivate a partition of hypotheses. I then use the observations to update the probability over the partition, and thus arrive at specific predictions. The updated probability is used to fill in the supposition about structure a bit further. But before that, let me stress that this example is not at all intended as an accurate description of real world inductive learning.

Motivating a Markov partition. Recall the observations of section 7.1, $q = 0, 1, 2, 3$, meaning wet, cold, warm, and dry respectively, all referring to the sensations recorded in the feet of a duck. Certainly, these kinds of stimuli never arrive as such clear-cut packages, and a complete conceptual framework is already presupposed if we take the stimuli to be captured in that way. On these presuppositions, then, it can be imagined that one fine morning the duck's feet record the series of numbers given in section 7.1. Now we may imagine that the duck is interested in the structure of the world that presents her with these sensations. In principle the duck may choose to scan her sensations for any kind of pattern, and base her expectations for further sensations on the pattern. As suggested, when it comes to the validity of the inductive inference there are no preferred patterns. In this specific case the duck supposes that the world consists of substances that are responsible for the sensations, in such a way that each substance may be associated with a cluster of sensations. Note that this supposition is still fairly general. The duck does not already preselect a specific number of substances, or the number of sensations associated with each of them.

The above suppositions on the structure of the world may now be translated into a partition on possible patterns in the sensations or observations. As indicated, it is not always a straightforward matter to connect the suppositions about structure with a general characteristic of the probability models in the partition. In this specific case, the supposed substances are associated with clusters of observations. The existence of certain substances therefore entails that after one observation, certain observations are more likely to occur than others. It is this latter kind of clustering in the observations that the partition of this example focuses on. More in particular, I employ a partition of hypotheses associated with so-called Markov processes. Before making these processes precise, it must be remarked that the inductive assumption presented by these processes is not directly derivable from the suppositions about structure, which just concern the existence of substances. As suggested in the foregoing, it is more that these suppositions are made precise in the form of the partition.

Predictions from a Markov partition. In a Markov process, the chance on an observation depends on the observation immediately preceding it. That is, the probability $p_{[h_{w\theta}]}(Q^q_{t+1}|E_t)$ of a Markov hypothesis $h_{w\theta}$ is a function of $e_t(t)$

only. If we have that $q \in \{0, 1, \ldots, N$, we can simply write

$$w(e_t) = e_t(t) + 1, \tag{7.1}$$

thus associating each Markov state with a separate value of the selection function $w(e_t)$. As discussed in chapter 2, the selection function determines which vector of likelihoods θ_m are prescribed by the hypothesis for the observation Q^q_{t+1} Since $e_0(0)$ is undefined, we may take $w(e_0) = 0$. With the selection function in place we can define a partition of statistical hypotheses for all components of the vector θ. Denoting the probability that observation q is followed by q' with $\theta_{qq'}$, we can define the Markov hypotheses as

$$p_{[h_{w\theta}]}(Q^{q'}_{t+1} | E_{t-1} \cap Q^q_t) = \theta_{qq'} \tag{7.2}$$

Note that $\theta_{qq'}$ has 4×4 components, corresponding to the fact that there are 4 transition probabilities to q' after each of the $N = 4$ possible observations q. For e_0 we may define a separate vector θ_0 of which each component is $\frac{1}{4}$.

The partition of hypotheses on Markov processes can now be used to provide predictions on the observations. Furthermore, the given observations may be used to derive a posterior probability over the hypotheses. But I will not reiterate the use of the Bayesian scheme for deriving the predictions and general conclusions here. These results, in any case, are not new. Comparable rules have been developed in Kuipers (1988) and Skyrms (1991). Instead I only present the results of the scheme. It may be noted that the hypotheses on Markov processes are a generalisation of the hypotheses on constant chances employed in chapter 3. We may therefore derive Carnapian prediction rules that apply to the separate states $w(e_t)$. On the assumption of a uniform prior probability over the Markov hypotheses, the resulting predictions are

$$p(Q^{q'}_{t+1} | E_{t-1} \cap Q^q_t) = \frac{t_{qq'} + 1}{t_q + 4}, \tag{7.3}$$

where $t_{qq'}$ denote the number of times that q' follows q in e_{t-1}, and $t_q = 1 + \sum_{q'} t_{qq'}$ is the total number of times that q occurs in e_t.

Deriving substances from data. With this uniform prior we can easily derive the predictions for the case of the duck's feet. Taking e_t as indicated in section 7.1, the predictions $p(Q^q_{t+1} | E_t)$ are given by the vector $\langle \frac{1}{3}, \frac{5}{12}, \frac{1}{12}, \frac{1}{6} \rangle$. Moreover, if we assume a uniform prior over the hypotheses $H_{w\theta}$ at the start, we can also find the hypothesis $H_{w\theta}$ for which the probability is largest after e_t very easily. It is simply the hypothesis for which the chances match the relative frequencies

$\frac{t_{qq'}}{t_q}$ in the observations. The hypothesis that fits the observations best may be summarised in the following matrix:

$$\theta_{qq'} = \begin{pmatrix} \boxed{\begin{array}{cc} \frac{1}{3} & \frac{5}{12} \\ \frac{5}{12} & \frac{1}{3} \end{array}} & \begin{array}{cc} \frac{1}{9} & \frac{2}{9} \\ \frac{2}{9} & \frac{1}{9} \end{array} \\ \begin{array}{cc} \frac{1}{12} & \frac{1}{6} \\ \frac{1}{6} & \frac{1}{12} \end{array} & \boxed{\begin{array}{cc} \frac{1}{3} & \frac{1}{3} \\ \frac{1}{3} & \frac{1}{3} \end{array}} \end{pmatrix} \tag{7.4}$$

where q labels the columns and q' the rows. It will be clear that I have deliberately chosen these observations for their relative frequencies.

If we take a closer look at the hypothesis that performs best on the observations, we notice that there is a clustering of the pairs of observations $\{0, 1\}$ and $\{2, 3\}$. Within the columns of the first pair there is a probability of $^3/_4$ that the observation $Q^{q'}_{t+1}$ is from the same pair as Q^q_t, and within the columns of the second pair this probability is $^2/_3$. On the level of observations, this may be just an interesting fact on the emerging pattern. On the level of structures, however, this fact can now be given further meaning. The best fitting hypothesis may be interpreted as telling us something on the kind of substances that may fill in the initial supposition. In particular, the hypothesis tells us that it is most likely that there are two substances, which are associated with the pair of cold and wet and the pair of dry and warm.

By way of introducing shorthand forms the duck may decide to collect the pairs into the substances water and air. Note that in the example, staying in the water is thus slightly more likely than staying in the air. The key point is that these two substances are not inherent to the observations. The starting point is just that there is an unknown number of substances, associated with unknown clusters of sensations. The Bayesian scheme and the observations then allow the duck to make this initial supposition about structure more precise.

7.5 Summary and conclusion

This chapter is the first of the chapters on inductive Bayesian logic in relation to themes in the philosophy of science. It has dealt specifically with the relation of this logic to the problem of induction.

First the three levels of the problem of induction were disentangled. In the section following up on that, I argued that the Bayesian scheme does not offer any directions for solving the problem of induction at the two observational levels. We are entirely free in choosing the inductive assumptions, and no such assumption is naturally suggested by the Bayesian scheme itself. After that I argued for a specific role, within the Bayesian scheme, for suppositions at the level of structure. Their task is to motivate the choice of specific inductive assumptions at the observational level. Furthermore, if the choice of assumptions is motivated in that way, the results of an inductive inference can in turn be used to draw further conclusions at the theoretical level. Finally, this was illustrated with an example showing the formation of the substances water and air on the basis of the weaker supposition that there are substances in the world.

The moral of the story is that in order to derive knowledge from observations, we must make assumptions from which the observations derive their meaning and impact. This is in line with the logical perspective taken in this thesis. It is notable that this perspective reflects the Kantian position that no empirical knowledge can be obtained without a theoretical scheme to organise the empirical data. In short, inductive knowledge is a co-production of the observer and the observed.

8

Bayesian Theory Change

This chapter addresses the problem that Bayesian inference cannot accommodate theory change, and proposes a framework for dealing with such changes. It first presents a Bayesian scheme for inferring predictions from observations by means of statistical hypotheses. An example shows how the hypotheses represent the theoretical structure underlying the scheme. This is followed by an example of a change of hypotheses. The chapter then presents a general framework for changing hypotheses, and proposes minimisation of the distance between hypotheses as a rationality criterion. Finally the chapter discusses the import of this for Bayesian statistical inference.

The present chapter can be read independently of the preceding chapters. There is considerable overlap with other chapters on the technical introduction of Bayesian schemes, but there is particular stress on some details that have not been given attention in the foregoing. Apart from that, chapter 1 will be helpful for situating this chapter within the general plan of this thesis. Chapter 2 may help to clarify the notion of a hypothesis employed in this chapter, in particular when it comes to the idea that changing the hypotheses amounts to a change of language. Finally, chapter 3 elaborates on the idea that partitions of hypotheses can be viewed as an expression of the theoretical structure underlying the inductive predictions.

8.1 Introduction

Fixed theoretical structure. In what follows I am concerned with Bayesian statistical inferences. These inferences are here considered in a scheme that generates predictions by means of hypotheses: Bayesian updating is used to adapt a probability over hypotheses to known observations, and this adapted probability is further used to generate predictions over unknown observations. The hypotheses in the scheme represent the theoretical structure that underlies the predictions. However, after we have chosen these hypotheses and a prior probability over them, updating fully determines the probabilities over the hypotheses at any later stage, and thus also the predictions resulting from that. There is no room for any further amendments to the hypotheses or to the prior probability assignment over them after they have been chosen. In Bayesian statistical inference, the theoretical structure is therefore fixed.

The fixity of the theoretical structure in the above schemes is a specific form of a more general problem for Bayesianism. Within the philosophy of science it has been formulated, among others by Earman (1992: 195–198), as the problem that Bayesianism fails to accommodate theory change. But the fact that Bayesian inference is in this sense dogmatic is at the origin of many other criticisms, including the criticism of Dawid (1982) that Bayesian inference is by definition calibrated. Furthermore, as hypotheses can be considered as specific terms in the observation language, changing the hypotheses in the scheme amounts to changing the language with which the predictions are made. The same problem can therefore be seen in light of the fact that Bayesianism fails to accommodate language change, as noted by Gillies (2000) and discussed elaborately by Williamson (2003).

This chapter addresses the above problems with Bayesianism. More in particular, it proposes a way of dealing with theory change within Bayesian statistical inference. The plan of the chapter is to introduce the Bayesian scheme for generating predictions from hypotheses, to present an example of such a scheme, then to show in the context of the example how hypotheses can be changed, and finally to give a general framework for such changes.

8.2 Hypotheses, conditioning and predictions

This section describes the Bayesian scheme for making predictions, as it has been presented in several of the preceding chapters. Observations and observational hypotheses are defined in terms of an observational algebra, and degrees of belief are represented by probability assignments over this algebra. The set-theoretical underpinning may seem unnecessary in the context of a short chapter. However, as will become apparent in sections 8.5 and 8.6, the underpinning is essential for a correct understanding of hypotheses change.

Observations and hypotheses. The predictions range over possible observations K, a set of consecutive natural numbers, say $\{0, 1\}$. At every time t we observe one number $q_t \in K$. We can represent these observations in an observational algebra. Let K^ω be the space of all infinite observation sequences e:

$$e = q_1 q_2 q_3 \dots \tag{8.1}$$

The observational algebra $\mathcal{Q}$, a so-called cylindrical σ-algebra, consists of all possible subsets of the space K^{ω}. If we denote the t-th element in a series e with $e(t)$, we can define an observation Q^q_t as an element of the algebra $\mathcal{Q}$ as follows:

$$Q^q_t = \{e \in K^{\omega} : \; e(t) = q\}. \tag{8.2}$$

Note that there is a distinction between the observations Q^q_t and the values of observations q. The values, represented with small letters, are natural numbers. The observations, denoted with large letters, are elements of the algebra $\mathcal{Q}$.

In the same way we can define an element in the algebra that refers to a finite sequence of observations. If we define the ordered sequence $e_t = \langle q_1 q_2 \dots q_t \rangle$, we can write

$$E^{e_t}_t = \{e \in K^{\omega} : \; \forall t' \leq t : e(t') = q_{t'}\}, \tag{8.3}$$

Again, it must be noted that the small letters e_t refer to a sequence of natural numbers, while the large letters E_t denote elements of the algebra, and carry a sequence of natural numbers as argument. The argument is sometimes omitted for sake of brevity. The observations and sequences of observations are related to each other in the natural way:

$$Q^q_{t+1} \cap E_t = E_{t+1}. \tag{8.4}$$

As in this equation, I normally refer to sequences of observations with the expression E_t, suppressing the reference to the sequence e_t.

Observational hypotheses can also be seen as elements of the observational algebra. If we say of an observational hypothesis h that its truth can be determined relative to an infinitely long sequence of observations e, then we can define hypotheses as subsets of K^{ω} in the following way:

$$H = \{e \in K^{\omega} : \; W_h(e) = 1\}. \tag{8.5}$$

Here $W_h(e) = 1$ if and only if the proposition h is true of e, and $W_h(e) = 0$ otherwise. The hypotheses can thus be arguments of the same probability functions over the observational algebra. A partition of hypotheses is a collection $\mathcal{H} = \{H_0, H_1, \dots H_N\}$ defined by the following condition for the indicator functions W_{h_n}:

$$\forall e \in K^{\omega} : \sum_n W_{h_n}(e) = 1. \tag{8.6}$$

This means that the hypotheses H_n are mutually exclusive and jointly exhaustive sets in K^{ω}.

The Bayesian scheme. Belief states are represented by probability functions over $\mathcal{Q}$. They take observations Q^q_t, sequences E_t, and hypotheses H_n as arguments. The functions are defined relative to a partition $\mathcal{H}$ and a sequence of known observations e_t: the function $p_{[\mathcal{H},e_t]}$ represents the belief state upon observing E_t under the assumption of a partition $\mathcal{H}$. It can be constructed by conditioning a prior probability function $p_{[\mathcal{H},e_0]}$ on the observations E_t:

$$p_{[\mathcal{H},e_t]}(\,\cdot\,) = p_{[\mathcal{H},e_0]}(\,\cdot\,|E_t). \tag{8.7}$$

Because of this, we have $p_{[\mathcal{H},e_t]}(E_t) = 1$. Updating the probability by simple conditioning is known as Bayes' rule. Both the probabilities assigned to observations and those assigned to hypotheses can be updated for new observations in this way. The probability before updating is called the prior probability, and the one after updating the posterior.

To calculate the predictions, we can employ a partition of hypotheses, and apply the law of total probability:

$$p_{[\mathcal{H},e_t]}(Q^q_{t+1}) = \sum_n p_{[\mathcal{H},e_t]}(H_n)\, p_{[\mathcal{H},e_t]}(Q^q_{t+1}|H_n). \tag{8.8}$$

The terms $p_{[\mathcal{H},e_t]}(Q^q_{t+1}|H_n)$ are called the posterior likelihoods of the hypotheses H_n for Q^q_{t+1}. The prediction is obtained by weighing these posterior likelihoods with the posterior probability over the hypotheses, $p_{[\mathcal{H},e_t]}(H_n)$.

Both posterior probabilities of equation (8.8) can be obtained from a Bayesian update of the prior probability $p_{[\mathcal{H},e_0]}$ according to expression (8.7). In this chapter the likelihoods do not change upon conditioning. Such likelihoods are sometimes called non-inductive.

$$p_{[\mathcal{H},e_t]}(Q^q_{t+1}|H_n) = p_{[\mathcal{H},e_0]}(Q^q_{t+1}|H_n). \tag{8.9}$$

That is, the observations influence the predictions only via the probability over the hypotheses. Part of the input probabilities for generating the predictions $p_{[\mathcal{H},e_t]}(Q^q_{t+1})$ are therefore the likelihoods $p_{[\mathcal{H},e_0]}(Q^q_{t+1}|H_n)$.

The predictions are further determined by the probability assignment over the hypotheses, $p_{[\mathcal{H},e_t]}(H_n)$. This probability can be determined by means of the relation

$$p_{[\mathcal{H},e_i]}(H_n) = p_{[\mathcal{H},e_{i-1}]}(H_n)\frac{p_{[\mathcal{H},e_{i-1}]}(Q^q_i|H_n)}{p_{[\mathcal{H},e_{i-1}]}(Q^q_i)}, \tag{8.10}$$

where q equals the last number in the sequence e_i. Note that the denominator $p_{[\mathcal{H},e_{i-1}]}(Q^q_i)$ can be rewritten with equation (8.8), substituting $t = i - 1$. Recall further that the likelihoods $p_{[\mathcal{H},e_{i-1}]}(Q^q_i|H_n)$ are in this chapter equal for all sequences e_{i-1}, as expressed in equation (8.9). The posterior probability $p_{[\mathcal{H},e_t]}(H_n)$ can therefore be determined recursively by the prior probability $p_{[\mathcal{H},e_0]}(H_n)$ for all n, and the likelihoods $p_{[\mathcal{H},e_0]}(Q_i^q|H_n)$ for all n and $i \leq t$. These are the other input probabilities for generating the predictions.

In sum, predictions can be generated if we assume hypotheses, their likelihoods, and a prior probability assignment over them. The prior and the likelihoods are first used to determine the posterior probability assignment over the partition. The likelihoods are then used together with this probability over the partition for generating the prediction itself. The whole construction that uses hypotheses to generate predictions is called the Bayesian scheme.

8.3 Contaminated cows

This section gives an example of a Bayesian scheme. The reader must be warned that the case presented falls short of actual scientific cases in many respects. The focus here is on the conceptual issues rather than on actual applications.

The example case. Consider a veterinary surgeon investigating a herd of cows during an epidemic, classifying them into contaminated and uncontaminated. The farmer claims that the herd has been treated with a drug that reduces the risk of contamination. It is an accepted fact about the epidemic that the average incidence rate among untreated cows is 0.4, as more than half of the cows show a natural resistance against contamination from other cows. The incidence rate among treated cows is 0.2 on average, because the drug is not always effective. The aim of the investigation is to decide whether the cows have been treated with the drug, and further to predict the incidence rate of the contamination in the herd. To enhance the dramatic impact, it may be imagined that the effect of the epidemic only shows in a slightly diminished milk quality, but that the fate of the cows depends on the incidence rate being lower than 0.3. For higher incidence rates the milk production fails to meet the quality criteria. Furthermore, the farmer is liable to legal prosecution if he has not treated the cows.

Setting up the inductive inference. The observations of the veterinary surgeon consist in test results concerning a number of cows. The result of testing cow t can be that it is contaminated, $q_t = 1$, or that it is not, $q_t = 0$. The test results can then be framed in the observational algebra. The vet may set up a scheme using a partition $\mathcal{D}$ of two hypotheses, which are associated with suppositions on treatment with the drug. The hypothesis D_1 is associated with the supposition that the cows are in fact treated, while D_0 means that they are not. It must be noted that the suppositions are thus not linked to observations directly, since the observations only concern contamination while the suppositions concern treatment. The relation that treatment bears to the observations is given by the incidence rates for treated and untreated cows, and this relation is laid down in the statistical hypotheses D_0 and D_1. For the observational content of the hypothesis on treatment D_1 we may take

$$W_{d_1}(e) = \begin{cases} 1 & \text{if } f(e) = 0.2, \\ 0 & \text{otherwise,} \end{cases} \tag{8.11}$$

where $f(e)$ is the relative frequency of results $q_t = 1$ in the infinite sequence e. The hypothesis D_0 may be defined in a similar way using $f(e) = 0.4$. A more precise definition is that the hypotheses comprise all so-called Von Mises Kollektivs for the given incidence rates, but for present purposes the loose definition suffices.

Being sets in the observational algebra, the hypotheses can also appear as arguments in the probability functions $p_{[\mathcal{D},e_t]}$. The fact that the veterinary surgeon is undecided on whether the farmer has treated his cows can be reflected in

$$p_{[\mathcal{D},e_0]}(D_0) = p_{[\mathcal{D},e_0]}(D_1) = 0.5. \tag{8.12}$$

Hypotheses on other relative frequencies, which are strictly speaking part of the partition, are thus given a zero probability. The likelihoods, for cow t being contaminated, of the hypotheses that it has or has not been treated are

$$p_{[\mathcal{D},e_0]}(Q_t^1|D_1) \quad = \quad 0.2, \tag{8.13}$$

$$p_{[\mathcal{D},e_0]}(Q_t^1|D_0) \quad = \quad 0.4. \tag{8.14}$$

I further assume that the estimated incidence rates are not affected by the running investigations, so that equation (8.9) holds.

Conclusions from observations and theory. With these values in place, the veterinary surgeon can start to predict the incidence rate in the herd, and decide over the treatment efforts by the farmer. Imagine that the first five test results are positive,

$$e_5 = 11111. \tag{8.15}$$

Subsequent updating on these test results yields the following probabilities and predictions:

Number of tests t	0	1	2	3	4	5
$p_{[\mathcal{D},e_5]}(D_1)$	0.50	0.33	0.20	0.11	0.06	0.03
$p_{[\mathcal{D},e_5]}(Q^1_{t+1})$	0.30	0.33	0.36	0.38	0.39	0.39

The probability that the farmer has treated his cows diminishes, and the probability that the next test result is positive tends to 0.4.

The conclusions expressed in the above values are that the farmer very probably did not treat his cows, and that a random cow from the herd has a probability close to 0.4 of being contaminated. It must be stressed, however, that these conclusions follow from the test results only if they are combined with the Bayesian scheme using $\mathcal{D}$. The scheme offers two possible hypotheses, and the observations are used to divide the probability between them. It is only relative to the partition $\mathcal{D}$ that most of the probability settles on D_0 after e_5, so that the predictions are equal to the likelihoods that D_0 prescribes for the test results. If, for example, we had also considered a hypothesis D_2 that prescribes likelihoods of 0.9 for positive test results, then this hypothesis D_2 would have been preferred over D_0, and the predictions would have followed the likelihoods of D_2. This example thus illustrates that the hypotheses in the scheme determine a range of probabilistic patterns, from which the observations may select the best fitting one. The partition of hypotheses functions as an assumption on what patterns can be picked up in the observations. The partition may therefore be called an inductive assumption.

Finally, it can be noted that the partition of hypotheses is associated with the theory underlying the scheme. In this case it concerns a classification of a state of the cows into treated and not treated. Both these concepts come with specific observational contents, which define the relevant patterns in the observations. There is no conceptual space within the Bayesian scheme, at least not as it is set up in the above discussion, to conclude anything other than that the cows are treated or not treated. In order to create this conceptual space, we must add hypotheses to the scheme.

8.4 Careless vaccination

This section shows how the hypotheses employed in the above scheme can be changed. I describe this change, and illustrate that it allows us to derive different conclusions and predictions.

Extending the example. Imagine that the veterinary surgeon becomes suspicious of the test results. After all, more than half of the cows are normally immune.

The sequence of test results must therefore be a rather unusual stochastic fluctuation on the average relative frequency of 0.4. The vet therefore decides to reconsider the inductive assumptions that underly the scheme, and to run a number of additional tests with an adapted scheme. In particular, she investigates the drug that the farmer claims to have used, and finds that it is a vaccinate with rather strict instructions for application. In most cases it works very well, even reducing the risk of contamination to 0.025, but careless use turns the vaccinate into a substance that causes a portion of 0.9 cows to be, or at least to appear, contaminated after treatment. The hypotheses that the vet wants to add to the scheme are that the drug has been used either carefully or carelessly.

Refined partition. The additional hypotheses may be collected in a separate partition $\mathcal{C}$, with C_1 associated with careful, and C_0 with careless treatment. Both hypotheses only apply to the case in which the cows have actually been treated, D_1. The combined partition is $\mathcal{B} = \{B_0, B_{10}, B_{11}\}$ in which $B_0 = D_0$, $B_{10} = D_1 \cdot C_0$, and $B_{11} = D_1 \cdot C_1$. Hypothesis B_0 is again defined with the relative frequency of 0.4, and the new hypotheses B_{10} and B_{11} can be defined with 0.9 and 0.025 respectively. These three relative frequencies define the new partition.

It is notable that the hypotheses B_{10} and B_{11} cannot be viewed as intersections $D_1 \cap C_0$ and $D_1 \cap C_1$: judged from the definition using relative frequencies, the original set D_1 and both sets B_{10} and B_{11} are disjoint. The relation between the old and the new hypotheses is a rather different one. We must imagine that within every infinite sequence $e \in D_1$, that is, within every possible world in which all cows are treated, we make a further selection of the observations q_t into those concerning cows that have been vaccinated with care, and those concerning cows that have been vaccinated carelessly. So B_{10} and B_{11} can be distilled from the old one by breaking up every $e \in D_1$, for which $f(e) = 0.2$, into two subrows e_0 and e_1 by means of a place selection, taking care that the relative frequencies of the two subrows are 0.9 and 0.025 respectively, and by grouping these subrows into B_{10} and B_{11}. Because $0.025 < 0.2 < 0.9$, such place selections can always be constructed.

The likelihoods of the hypotheses may again be equated to the relative frequencies that define the hypotheses:

$$p_{[\mathcal{B},e_0]}(Q_t^1|B_{10}) = 0.9, \tag{8.16}$$

$$p_{[\mathcal{B},e_0]}(Q_t^1|B_{11}) = 0.025. \tag{8.17}$$

In order to arrive at the overall incidence rate of 0.2 for treated cows, the veterinary surgeon may further assume that a portion of 0.2 of all farmers do not treat the vaccinate with the necessary care, as $0.2 \times 0.9 + (1-0.2) \times 0.025 = 0.2$. I come back to this choice in section 8.6. Finally, using the probability assignment after five tests, the combined probability of treatment with the drug and the lack of care is

$$p_{[\mathcal{B},e_5]}(B_{10}) = 0.03 \times 0.2 = 0.006 \tag{8.18}$$

It must be noted that with the employment of $\mathcal{B}$, the probability over the observational algebra really undergoes an external shock: instead of allocating 0.030 probability on the set D_1, we now allocate 0.006 on B_{10} and 0.024 on B_{11}.

Different conclusions. With these new hypotheses and the associated inductive assumptions, the veterinary surgeon can run a number of additional tests. Let us say that the next ten test results are all positive too,

$$e_{15} = 111111111111111. \tag{8.19}$$

Subsequent updating on these test results yields the following probabilities and predictions:

Number of tests t	5	7	9	11	13	15
$p_{[\mathcal{B},e_{15}]}(B_{10})$	0.01	0.03	0.14	0.49	0.80	0.95
$p_{[\mathcal{B},e_{15}]}(Q^1_{t+1})$	0.39	0.42	0.47	0.62	0.80	0.88

Now the probability for B_{10} approaches 1, while the predictions for a cow in the herd to be contaminated tend to 0.9. Clearly these values differ from those that were to be expected on the basis of $\mathcal{D}$.

The conclusions expressed in these values are that the farmer did treat his cows with the drug, but that he did not apply it with the necessary care. The further conclusion is that the incidence rate of the epidemic in his herd is 0.9. Again, these conclusions are drawn from the test results in combination with the inductive assumptions of partition $\mathcal{B}$. It is only when compared to the other members of the partition that the hypothesis B_{10}, which prescribes an incidence rate of 0.9, fits the test results best. For present purposes, however, it is most notable that these conclusions differ dramatically from those derivable from $\mathcal{D}$.

Note that this is again different if we further introduce the partition $\mathcal{I}$ on whether the test material is itself infected, and stipulate that in the combined partition $\mathcal{A} = \{I_0 \cdot D_0, I_0 \cdot D_1 \cdot C_0, I_0 \cdot D_1 \cdot C_1, I_1\}$ we have $p_{[\mathcal{A},e_t]}(Q^1_{t+1}|I_1) = 1$, while slightly adapting the values for the other likelihoods. Relative to the partition $\mathcal{A}$, the priors for $\mathcal{I}$ and some further observations, the conclusion may then be that the test material is infected. However, for the partition $\mathcal{B}$ and its associated inductive assumption, the conclusions must be as indicated above.

8.5 A framework for changing partitions

The above illustrates how we can change a partition of hypotheses during an update procedure. This section gives a general framework for such changes, and draws attention to the need for new criteria of rationality to guide them.

Capturing hypotheses change. On the change of partition itself, as illustrated in figure 8.1, I can be relatively brief. Let us say that the old partition $\mathcal{H} = \{H_0, H_1, \ldots, H_N\}$ consists of hypotheses H_n with likelihoods

$$p_{[\mathcal{H},e_t]}(Q^q_{t+1}|H_n) = \theta^q_n. \tag{8.20}$$

The addition of a partition $\mathcal{G} = \{F_0, F_1, \ldots, F_M\}$ to this partition generates a combined partition $\mathcal{G} = \mathcal{H} \times \mathcal{G}$, which consists of $N \times M$ hypotheses $G_{nm} = H_n \cdot F_m$. Each of these hypotheses may be associated with a relative frequency of the observation q, denoted γ^q_{nm}, so that

$$p_{[\mathcal{G},e_t]}(Q^q_{t+1}|G_{nm}) = \gamma^q_{nm}. \tag{8.21}$$

The details of the partition change may be such that for some of the H_n we have that $\gamma^q_{nm} = \theta^q_n$ for all q and m. We can then collect the hypotheses G_{nm} under the single index number n, as for example B_0 above. More in general, if two hypotheses G_{nm} and $G_{n'm'}$ are such that $\gamma^q_{nm} = \gamma^q_{n'm'}$ for all q, we can merge them into a single hypothesis. In the extreme case in which for all q the γ^q_{nm} vary only with m, the change of partition comes down to a replacement of $\mathcal{H}$ by $\mathcal{G}$.

An external shock to the probability assignment. With the introduction of new hypotheses, the probability over the observational algebra undergoes an external shock. First, the probability over the hypotheses themselves changes. But since the new hypotheses have different likelihoods, the probability over most other elements of the algebra changes as well. It is in this chapter assumed that at the time of change τ, the new probability assignment over the hypotheses observes the following restriction:

$$p_{[\mathcal{G},e_\tau]}(\cup_m G_{nm}) = \sum_m p_{[\mathcal{G},e_\tau]}(G_{nm}) = p_{[\mathcal{H},e_\tau]}(H_n). \tag{8.22}$$

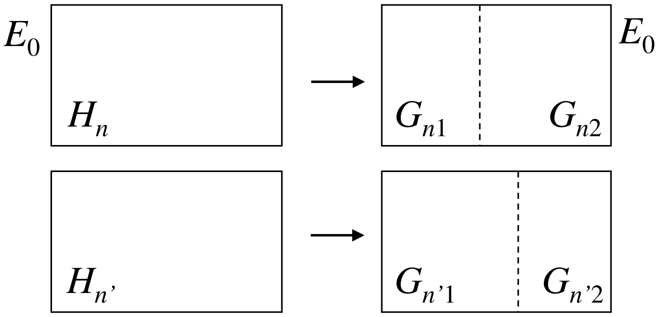


Figure 8.1: Graphical representation of a partition change. The hypotheses G_{nm} are separate patches within the hypotheses H_n. Strictly speaking, the change is therefore effected by a refinement. The dotted lines between G_{n1} and G_{n2} and between $G_{n'1}$ and $G_{n'2}$ indicate that the priors of the G_{nm} within each H_n may be chosen freely.

That is, the probability assignment arrived at by updating over $\mathcal{H}$ is taken over into the new partition $\mathcal{G}$. This restriction serves to link every collection $\cup_m G_{nm}$ to the original hypotheses H_n, but it can be dropped if further details of the partition change permit it. Finally, within the limits set by this restriction, the probabilities of the hypotheses G_{nm} can vary freely.

It can be noted that the change in probability due to partition change is not one that can be represented as Bayesian conditioning. Conditioning determines how to adapt probability assignments if for some observation Q_t^q or E_t the probability is externally fixed to 1. It is quite different to set the probability of a number of hypotheses H_n to zero, and to redistribute this probability over new hypotheses G_{nm}. A partition change is therefore an external shock to the probability assignment to which we cannot apply Bayesian updating. Now there are many arguments to the effect that Bayesian updating is the only rational way to adapt a probability assignment to new information, but these arguments do not apply in this case, since the new information can in this case not be represented, in any straightforward manner, as an element of the algebra. It seems that the possibility of partition change necessitates new criteria of rationality, and the definition of an associated update operation.

8.6 Distance between partitions

This section answers the need for a rationality criterion and an associated update operation. In particular, it elaborates on a distance function between the old and the new partition, and shows how to minimise this distance during the partition change.

Minimising cross-entropy. Partition change may be considered as an external shock to the probability assignment over the algebra. Williamson (2003) argues that changes in the probability assignment must be conservative, that is, as small as possible, and further that such conservatism can be explicated by a minimisation of the cross-entropy distance function between the old probability p_0 and the new probability p, under the restrictions imposed by the external shock. The distance function is defined by

$$\Delta(p, p_0) = \sum_U p(U) \log \frac{p(U)}{p_0(U)}, \tag{8.23}$$

where the index U runs over all sets in the finite algebra over which p_0 and p are defined. As elaborated in Kullback (1959) and Paris (1994: 120–126), minimising this distance under the external restrictions effectively minimises the information change that is induced in the probability assignment by the external shock. Interestingly, the operation of minimising cross-entropy coincides with the operation of a Bayesian update in the case that some probability $p_{[\mathcal{H},e_t]}(Q^q_t)$ is restricted to 1. It therefore accords with Bayesian statistical inference to adopt the minimisation of cross-entropy as the update operation in cases of partition change.

We are not yet done with the update operation for partition change. For one thing, the above distance function blows up if the algebra contains an infinite number of elements, as is the case for the algebra $\mathcal{Q}$. We need to select a finite collection of elements of the algebra, for which we may then minimise the distance between the old and the new probability assignment. Note that it is not desirable to minimise the difference between the old and the new predictions. The reason for the partition change is exactly that the old predictions do not match the pattern in the observations well. And note further that the probability assignment over the hypotheses is changed deliberately, so that we cannot apply the minimisation of the distance to the assignments over hypotheses either. In sum, we have to apply the minimisation of cross entropy to a collection of elements from the observational algebra that does not emphasise the predictions or the hypotheses themselves.

As already indicated in the example, it is rather intuitive to choose a minimisation of the distance between the likelihoods of the hypotheses H_n and of the associated collections $\cup_m G_{nm}$. These likelihoods fully express the hypotheses, and the distance between the likelihoods is therefore an intuitive measure for the closeness of the two partitions. A further reason for choosing the collection $\cup_m G_{nm}$ can be found in the relation between the old and the new hypotheses. Recall that the likelihoods of H_n for observations Q^q_t are determined by the relative frequencies of the observations $q \in K$ within the infinite sequences of observations, or possible worlds, for which H_n is true. With the change of hypotheses, we effectively make a further division of these possible worlds into the hypotheses G_{nm}. Specifically, each infinite sequence of observations $e \in H_n$, having a relative frequency θ^q_n, must be split into M infinite subsequences e_m, having relative frequencies γ^q_{nm}, and these subsequences can then be incorporated into separate hypotheses, $e_m \in G_{nm}$. Because the hypotheses G_{nm} are derived from the original hypotheses H_n in this way, we may expect the relative frequency associated with the aggregate $\cup_m G_{nm}$ to be the same as, or at least close to, the original relative frequency associated with H_n.

A note on Kollektivs. At this point we may recall the definition of hypotheses as sets of sequences with specific relative frequencies, which is developed in chapter 2. In the context of that chapter it seems more elegant to equate hypotheses with collections of Kollektivs. Section 2.3.2 argues that there are further reasons for the definition of hypotheses by means of relative frequencies only, and these reasons can become apparent if we consider the creation of hypotheses alluded to above. Let me first admit that this creation is not a neatly defined operation yet. However, I do think that such an operation can eventually be defined, and that it then mimics the kind of epistemic move involved in choosing a new partition. Indeed, the veterinary surgeon imagines that an infinite sequence of unobserved cows is broken up into finite segments, the herds, which are then marked as being treated carefully and carelessly. These finite segments are then concatenated to render two different infinite sequences. But if the infinite sequences $e \in H_n$ are taken to be Kollektivs, we simply cannot create these different sequences $e_m \in G_{nm}$ from the single hypothesis H_n by means of a place selection. We must therefore maintain that the $e \in H_n$ are not Kollektivs.

Calculations. Any hypothesis prescribes the likelihoods for infinitely many observations $Q^q_{\tau+t}$, associated with different times $t \geq 0$. However, these likelihoods are in this chapter constant, and it seems natural to define the distance between the partitions as the distance between the likelihoods at a single time

t. For p_0 we can use the old likelihoods $p_{[\mathcal{H},e_\tau]}(Q^q_{\tau+t}|H_n) = \theta_q$. For p we use the aggregated likelihoods, given by

$$\begin{aligned}\gamma^q_n &= p_{[\mathcal{G},e_\tau]}(Q^q_{\tau+t}|\cup_m G_{nm}) \\ &= \sum_m \frac{p_{[\mathcal{G},e_\tau]}(G_{nm})}{\sum_m p_{[\mathcal{G},e_\tau]}(G_{nm})} p_{[\mathcal{G},e_\tau]}(Q^q_{\tau+t}|G_{nm}) \qquad (8.24) \\ &= \sum_m \rho_{nm}\gamma^q_{nm}. \qquad (8.25)\end{aligned}$$

Here the ρ_{nm} are defined by the fraction in equation (8.24), so that $\sum_m \rho_{nm} = 1$. The γ^q_n are a function of these ρ_{nm}.

We can now use the distance function to find the aggregated likelihoods $p_{[\mathcal{G},e_\tau]}(Q^q_{\tau+t}|\cup_m G_{nm})$ that are closest to the likelihoods $p_{[\mathcal{H},e_\tau]}(Q^q_{\tau+t}|H_n)$, for any time t. These distances are defined for each hypothesis H_n separately:

$$\Delta_n(\rho_{nm}\gamma^q_{nm}, \theta_q) = \sum_q \gamma^q_n \log \frac{\gamma^q_n}{\theta^q_n}. \qquad (8.26)$$

The distance for H_n is thus a function only of the fractions ρ_{nm}, which determine how the probability of H_n is distributed over the G_{nm}. The update operation after a hypotheses change is to find, for every H_n separately, the values of ρ_{nm} that minimise the distance function Δ_n.

This can be employed to provide a further underpinning for the choice of the probabilities $p_{[\mathcal{B},e_5]}(B_{10})$ and $p_{[\mathcal{B},e_5]}(B_{11})$ in the example. It was stated there that the veterinary surgeon chooses these probabilities in order to arrive at the overall incidence rate of 0.2. Note that the distance between the likelihoods of $\mathcal{H}$ and the aggregated likelihoods of $\mathcal{G}$ is zero and therefore minimal if we find values for ρ_{nm} so that $\gamma^q_n = \sum_m \rho_{nm}\gamma_{nm} = \theta^q_n$. In the case of the partitions $\mathcal{D}$ and $\mathcal{B}$, the equation simply becomes $0.9 \times \rho_{10} + 0.025 \times (1 - \rho_{10}) = 0.2$, for which $\rho_{10} = 0.2$ is the solution.

Generalisations. It must be stressed that the present exposition does not comprise the full story on partition change. There are many cases of partition change that are not covered by the above framework, but that can in principle be treated in a similar way. One such case deserves separate attention here. The example presents a probability assignment that is not open-minded: almost all hypotheses on relative frequencies are given a zero probability. This may cause the impression that the framework for partition change can only be applied if the old probability assignment is not open-minded. It may be hard to see what other hypotheses can be added if, for instance, the prior probability already includes all possible hypotheses on relative frequencies. However, the above framework can also be used to change a partition of all hypotheses on relative frequencies into a partition of hypotheses that concern all Markov processes. The application of the framework for partition change is thus not limited to cases in which the prior is not open-minded.

8.7 Concluding remarks

In this chapter it has been shown how we can frame a partition change, and a procedure has been provided to render this change rational, employing a distance function between the partitions. I complete the chapter with a summary and some remarks on the proposed framework in the context of Bayesian statistical inference.

The proposed framework enables us to adapt the hypotheses that function in a scheme for making predictions. By writing down the predictions in terms of a Bayesian scheme, I locate the theoretical structure underlying the predictions inside the probability assignment. Theoretical developments can therefore be framed as external shocks to the probability assignment representing the current opinions, just as new observations. I then argue that the operation that updates the assignment for the external shock is a generalised version of Bayesian conditioning, namely cross-entropy minimisation. The framework is therefore a natural extension of Bayesian statistical inference. On the whole, the chapter proposes an answer to the problem that Bayesian statistical inference cannot accommodate theory change.

The chapter may also fulfil a role in an older discussion between inductivists and Popperians: it basically shows how we can incorporate a notion of conjecture within an inductivist setting. It is a typical feature of Carnapian inductive logic that there is no room for an explicit formulation of inductive assumptions, as such assumptions are part and parcel of the choice of language. Conjectures can therefore not be captured within a Carnapian logic. However, the above discussion associates the premisses with the hypotheses used in the Bayesian scheme, and further allows us to change them. It provides a truly nonmonotonic inductive Bayesian logic, in the sense that besides the set of available observations, also the inductive assumptions may be altered along the way. This chapter is thus a first step in generalising inductive Bayesian logic to incorporate changes in the projectability assumptions.

9

Abducted by Bayesians?

This chapter discusses the use of theoretical distinctions between hypotheses in Bayesian inductive inference. A theoretical distinction is any distinction between hypotheses that is not reflected in a difference in likelihoods. This chapter shows that under certain conditions inductive predictions may benefit from theoretical distinctions, and further that under such conditions the observations can tell theoretical hypotheses apart. Two considerations follow from this main conclusion. First, the puzzle on theoretical hypotheses can be repeated at a higher level, concerning scientific method more generally. This leads to the claim that underdetermination fulfils a function in scientific method. Second, the choice between theoretical hypotheses in science is usually associated with abductive inference. This chapter therefore contains the promise of a Bayesian model of abduction.

The present chapter can again be read entirely independently. However, the technical part of the chapter is rather concise. For a more elaborate treatment I refer to section 1.3. A deeper understanding of statistical hypotheses can be obtained from chapter 2. Chapters 3 and 7 are useful for understanding the relation between theoretical background and inductive predictions more generally.

9.1 Statistical inference using partitions

This section describes a Bayesian scheme for inductive inferences, running from observations and statistical hypotheses to predictions. The prior probability over hypotheses is first updated to the given observations, and the updated probability is subsequently used to generate predictions. The resulting predictions may be also captured directly in prediction rules, but the hypotheses are seen to be useful for expressing knowledge of underlying chance mechanisms.

Bayesian inductive inference. The inductive scheme employs a formal framework of observations and hypotheses. Consider an observation at time i with a possible result $q \in \{0, 1\}$, denoted Q_i^q, and denote sequences of observations of length t with E_t. The example of this section supposes the observations to consist of results of coin tosses. Consider a partition $\mathcal{H}$ of hypotheses H_θ concerning the chance θ on tails, $q = 1$, so that

$$p(Q^1_{i+1} | H_\theta \cap E_i) = \theta, \tag{9.1}$$

and further a prior probability over these hypotheses, $p(H_\theta)d\theta$. The partition, the likelihoods of the hypotheses in it, and the prior probability over the hypotheses together determine the Bayesian scheme.

The observations are the other component that is needed to arrive at inductive predictions. Bayes' rule can be used to update the prior probability to given observations E_t. After updating we obtain a posterior probability,

$$p(H_\theta | E_t)d\theta = \frac{p(E_t | H_\theta)}{p(E_t)} p(H_\theta)d\theta. \tag{9.2}$$

Predictions follow directly from this posterior by the law of total probability:

$$\begin{aligned} p(Q^1_{t+1} | E_t) &= \int_0^1 p(Q^1_{t+1} | H_\theta \cap E_t)\, p(H_\theta | E_t)\, d\theta \\ &= \int_0^1 \theta\, p(H_\theta | E_t)\, d\theta. \end{aligned} \tag{9.3}$$

This scheme for predictions covers a substantial part of Bayesian statistical inference, as many such inferences are made with models concerning constant chances.

The use of hypotheses. In a sense, the hypotheses H_θ are already theoretical. They concern the objective chance of an observation, and such chances cannot be translated into finite observational terms. Moreover, the hypotheses can be eliminated from the inference completely. De Finetti's representation theorem states that the above scheme of hypotheses covers exactly those prediction rules for which the order of the observations in E_t is inessential. Defining t_q as the number of Q^q_i in E_t, these rules maybe characterised with

$$p(Q^q_{t+1} | E_t) = pr(t_q, t). \tag{9.4}$$

Every rule pr corresponds to a specific prior $p(H_\theta)d\theta$ over the hypotheses in the scheme, and vice versa. In particular, if we assume the prior to be a symmetric Dirichlet distribution, we can derive the Carnapian λ rules:

$$p(Q^q_{t+1} | E_t) = \frac{t_q + \lambda/2}{t + \lambda} = pr_\lambda(t_q, t). \tag{9.5}$$

A higher peak in the Dirichlet density $p(H_\theta)$ is reflected in a larger parameter λ.

Although the hypotheses in the above inferences can thus be replaced with direct links between observations, there are good reasons for keeping the hypotheses in. First, they express the chance mechanism that is supposed to underlie the observations. For example, if the observations concern coin tosses, we know that the mechanism underlying the observations concerns constant and independent chances. Second, the hypotheses enable us to express further knowledge of the chance mechanisms in a prior probability over them. In the example, a normal coin motivates a prior over these chances that is strongly peaked at $\frac{1}{2}$, while a coin from a conjurer's box may have little probability at $\frac{1}{2}$ and more probability at 0 and 1. In the prior we can thus express knowledge of the chance mechanism that is not incorporated in the statistical hypotheses themselves. It is not always a straightforward matter to incorporate such knowledge in a direct prediction rule.

9.2 Duplicate partitions

in what follows the above scheme will be extended by a duplicate partition. The distinction between the two duplicates is therefore entirely theoretical. It is shown that this purely theoretical distinction facilitates the choice of priors. It is further shown that this move makes the two duplicate subpartitions observationally distinguishable after all. In addition, it is seen that there are computational advantages to keeping the two subpartitions distinct.

A normal or magical coin. Let me start with the example on coin tosses. Imagine that we are undecided on whether the coin is from a conjurer's box or from an ordinary wallet. Now both these kinds of coins have an unknown constant chance on tails, $q = 1$, so that we may employ the hypotheses H_θ. However, we have some further knowledge of the mechanism underlying the observations that must somehow be incorporated in the prior: either the coin is most probably fair, having a chance that is close to $\frac{1}{2}$, or the coin is most probably strongly biased, having a chance that is close to 0 or 1. To incorporate this knowledge, we can now employ an additional partition into the hypotheses G_0 concerning the normal coin and G_1 concerning the magical coin.

Both hypotheses G_j cover exactly the same subpartition, $G_j = \{g_j\} \times \mathcal{H}$. They are only labelled differently. We can use the likelihoods θ for the hypotheses $g_0 \times H_\theta$ and $g_1 \times H_\theta$ alike. In terms of these statistical hypotheses, the distinction between the magical coin and the normal coin is therefore not observable. For each hypothesis in the one subpartition, there is a hypothe-

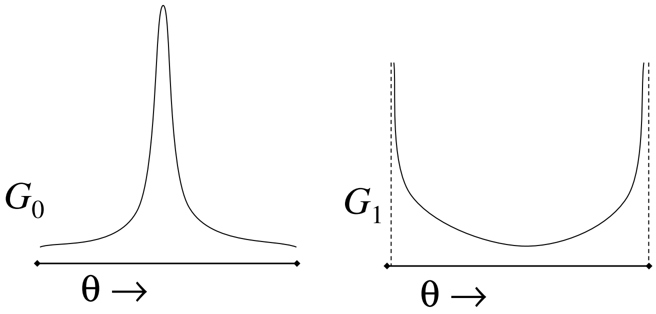


Figure 9.1: Two different priors over the two subpartitions of Bernoulli processes H_θ. The peak prior is associated with the normal coin, the valley prior with the magical coin. Both function are from the class of Dirichlet priors. For $\lambda = 2$ the prior distribution is uniform, for larger values of λ the peak gets higher, and for smaller values of $\lambda 2$ the valley gets deeper.

sis in the other subpartition that has exactly the same likelihoods for all the observations. The partition as a whole is thus underdetermined.

Advantages of a degenerate partition. There is a particular advantage, however, to employing this duplicate partition in the Bayesian scheme. We have separate control of the priors over the subpartitions on the normal and magical coin, $g_0 \times \mathcal{H}$ and $g_1 \times \mathcal{H}$ respectively. The further knowledge about the two kinds of coins motivates specific forms for the priors in both partial partitions, leading to two separate Carnapian rules, for example with $\lambda = 10$ and $\lambda = \frac{1}{4}$. The priors are illustrated in figure 9.1. Let us say that initially we are undecided between these two, $p(G_0) = p(G_1)$. The rules can then be weighed with the probabilities of the coin's origin, resulting in a so-called hyper-Carnapian prediction rule:

$$p(Q^q_{t+1}|E_t) = p(G_0|E_t)\, pr_{10}(t_q, t) \;+\; p(G_1|E_t)\, pr_{1/4}(t_q, t). \tag{9.6}$$

The idea here is that the probabilities within the two subpartitions $g_0 \cdot \mathcal{H}$ and $g_1 \cdot \mathcal{H}$ are updated separately, and that the resulting values yielded by the Carnapian rules can function as the likelihoods in an update over the hypotheses G_0 and G_1.

Interestingly, even while the subpartitions associated with G_0 and G_1 consist of pairwise identical hypotheses, the differing priors over them cause different aggregated likelihoods of G_0 and G_1, namely the different Carnapian rules. That is, the two partitions themselves are observationally indistinguishable, but the different expectations over these partitions make the partitions observationally

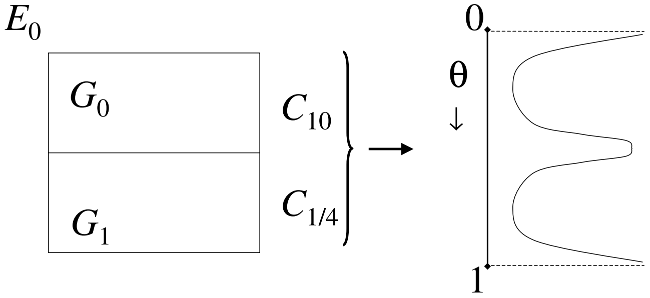


Figure 9.2: The hyper-Carnapian prediction rule can also be expressed in a single prior over one partition C. The prior is simply the sum of the two separate priors over the separate subpartitions.

distinct after all. As a side effect of updating over $g_0 \times \mathcal{H}$ and $g_1 \times \mathcal{H}$, the observations become relevant to the theoretical distinction between hypotheses G_0 and G_1.

This effect is less magical than it may seem. The distinction between the hypotheses G_0 and G_1 may not be observational relative to the subpartitions $\mathcal{H}$, but the hypotheses G_0 and G_1 do have an observational content: a magical coin is much less likely to yield an observed relative frequency of tails of close to $\frac{1}{2}$ than the normal coin. This content is exactly expressed in the differing priors over the partial partitions $g_0 \times \mathcal{H}$ and $g_1 \times \mathcal{H}$. The theoretical distinction simply facilitates the use of these differing priors over the two subpartitions. A further function of the distinction consists in keeping calculations manageable. The function that expresses the combined prior over a single partition $\mathcal{H}$ is naturally the sum of the priors defined over the above subpartitions, as expressed in figure 9.2, but it is much more convenient to update these terms separately. The resulting predictions can not be equated with a single Carnapian rule, and it is not easy to find some other exchangeable direct prediction rule that captures them.

Relations to preceding chapters. Some remarks may connect the present discussion with preceding chapters. Note first that in terms of the frequentist semantics of chapter 2, the hypotheses G_0 and G_1 are indeed identical. They consist of the very same subpartitions, and thus of the very same sequences of observations. The use of a duplicate partition reminds of the initial Kolmogorov picture of the Bayesian scheme, as presented in chapter 1. Just as the hypotheses H_j in that chapter, the hypotheses G_j are here associated with a complete observational algebra. Secondly, it is notable that the partition of hypotheses is here used to encode specific inductive assumptions in a prior probability assignment. In this sense the chapter has much in common with chapters 4 and 5. While in these chapters the partition of hypotheses is transformed, in the present chapter the partition is duplicated. But in both cases we manipulate the partition of hypotheses in order to access the appropriate prior.

In the case of the above hyper-Carnapian rule, the reader may find that the advantages of the representation of inductive predictions in terms of statistical hypotheses is entirely unhelpful, or even contrived. It may be much more natural to employ the hypotheses G_j with the Carnapian rules as likelihoods. However, in view of chapter 3 there are independent reasons for preferring the partition of statistical hypotheses H_θ, with constant likelihoods, over the single Carnapian prediction rules. Moreover, I feel that a Bayesian statistician may have some problems in making sense of the Carnapian prediction rules in the role of statistical hypotheses. According to chapter 2 they are not even included in the class of such hypotheses. Finally, and in relation to all this, I want to maintain that the use of the hypotheses H_θ allows us to disentangle two different aspects of the way we deal with the observations of the coin tosses, and that these two aspects are conflated if we use just the hyper-Carnapian rule.

Another set of considerations concerns the role of the knowledge about underlying mechanisms, in this case knowledge about the possible type of the coin. First of all, I am not sure that we can speak of knowledge of the underlying mechanism. In the example of the coin we are perhaps in that position, but in the standard case of scientific investigations the underlying mechanism can at best be a supposition. On the other hand, such suppositions may be used to inform the priors just as well. As a second consideration, and following up on this, it is not in all cases clear how exactly these suppositions determine the form of the prior probability assignment over the subpartitions. The example may suggest that this link is straightforward, but there are many cases in which this is simply not true. The second part of this thesis illustrates that finding a prior probability assignment over hypotheses that encodes certain assumptions on the underlying mechanism is a substantive and all but trivial part of the task of inductive logic.

Summary. Let me summarise the main point of this section. It directs attention to inductive inferences using two duplicate subpartitions, which differ only in the entirely theoretical property that they posit different mechanisms underlying the observations. It may seem pointless to use such duplicate partitions. However, the mechanisms underlying the observations can motivate different prior probabilities over these subpartitions. And because these priors react to the updating operations differently, the partitions can be distinguished by the observations even while they consist of statistically identical hypotheses. The reason for using duplicate partitions is thus that they facilitate the expression in the prior probability assignment of knowledge of underlying mechanisms, thereby making the duplicate partitions observationally distinguishable after all.

9.3 The use of underdetermination

One of the main messages of this thesis is that hypotheses are useful for expressing suppositions on chance mechanisms in inductive inference, by making accessible, that is, conceptually and computationally manageable, certain classes of prior probabilities. The present chapter argues that this latter usage also applies to entirely theoretical distinctions between hypotheses: the theoretical distinction motivates specific priors, and the distinction is thereby given observational content. In this last section I transfer this insight to scientific method more generally. First I propose a different perspective on the problem of underdetermination. After that I argue that the use of duplicate partitions is holding the promise of a Bayesian model of abductive inference.

Underdetermined statistical inference. The problem of underdetermination is that science, if interpreted as a realist undertaking, is dramatically underdetermined by observation: at first sight it seems that much of the theoretical superstructure of scientific theories cannot be warranted by the observational substructure. The primary challenge for realists is to show that this apparent underdetermination is not harmful to the realist objective of science, where this objective, put crudely, is to present science as an enterprise that successfully aims for the truth. A good example of this reaction is to be found in structural realism as presented in Worrall (1989), Ladyman (1998), and Votsis (2005). However, some realists take on the bigger challenge of showing that underdetermination can in some cases be avoided. They achieve this by providing inference rules such as abduction, which enable us to choose between theoretical superstructures on the basis of explanatory considerations or other theoretical virtues.

By contrast, in the following I stick to the original challenge of showing that underdetermination does not obstruct the realist aims of science. More specifically, I attempt to show that underdetermined theoretical superstructures have a specific use in statistics. I thus accept that science is underdetermined, but I go on to suggest that it is possible to explain this fact by reference to the methodological use of underdetermination. How this use of theoretical superstructures reflects back on realism I leave for future research.

Recall the claim of the preceding section that a partition that employs purely theoretical distinctions may offer a better grip on statistical analyses of experimental observations. In the example, the hypotheses G_0 and G_1 cause underdetermination, since we can never tell them apart by observations. But distinguishing them is very useful in the statistical procedure: they facilitate the expression of knowledge or suppositions on underlying mechanisms in priors, and they carve up statistical inference in manageable parts. More generally, we may tentatively say that the use of theoretical distinctions in statistical analyses reveals the advantages of underdetermination. In future research I hope to support this claim with case studies on actual experiment, in which theoretical distinctions are indeed employed to elicit specific conclusions from the observations. The idea is that enriching the observational algebra with theoretical distinctions improves the expressive force of the inductive scheme, and thus the ability to elicit answers from nature and make specific inductive inferences.

Abduction. It is important to note that, as a side effect of using theoretical distinctions, it looks as if these distinctions themselves become observational. This is where the use of theoretical distinctions in inductive inference begins to look like abduction. An abductive inference enables us to choose between a number of observationally indistinguishable, and thus theoretical, alternatives on the basis of certain theoretical virtues, for example explanatory force. Now the key insight here is that such a theoretical virtue is also presented in the fact that one of the two priors in the duplicate partition corresponds better to the observations. Recall that the observations have exactly the same impact on the separate hypotheses in each of the two subpartitions. The different impact is entirely due to the difference in the subjectively determined probability over these two subpartitions. We may therefore say that the observations reflect differently on the two subpartitions, exactly because they interact differently with our expectations.

Let me briefly explain these remarks by relating them to empirical equivalence and the nature of observations. Note first that whether two theories are empirically equivalent or not depends on the notion of theory that is employed. If we assume that the theories about the origin of the coin are determined by the statistical hypotheses that they consist of, in both cases $\mathcal{H}$, they are indeed empirically equivalent. But if we say that the prior probability assignment is an inherent part of the theories, then the two theories are not empirically equivalent. Furthermore, on the stipulation that the theories are empirically equivalent, whether two theories may or may not be told apart by observations in a sense indicates the nature of these observations. That is, if we take the content of the observations to be the effect they have on the probability assignment on the whole, it may be argued that the observations in the coin example are not entirely empirical. In that case they somehow incorporate theoretical content. If, on the other hand, we stipulate that the theories are empirically distinct in the first place, or if we stipulate that the content of the observations is given by the likelihoods for the observations and by nothing else apart from that, then there is no reason to say that observations have theoretical content.

It will be clear that I prefer the view that observations also convey theoretical content, and that they manage to do so because of the theoretical scheme in which we have chosen to frame them. My main reason for preferring this view is that I think it holds the promise of a Bayesian model of abduction. It provides a formal expression for the position that there is no sharp line between observation and theory, from which the use of observations for deciding over theoretical distinctions is seen to follow. However, the details of this position, which will elaborate the idea that the observations have different content in the context of different theoretical subpartitions, must be left to further research.

Conclusion

Summary. This thesis falls into three main parts. The first part claims that the hypotheses in the Bayesian scheme offer a better control over the inductive assumptions inherent to predictions. The second part adds to this by showing that spaces of hypotheses prove very useful in encoding specific aspects of the predictions in a prior probability. The third part illustrates the use of the Bayesian scheme in solving some problems in the philosophy of science, in particular problems concerning scientific method.

Let me run through the chapters in some more detail. As for the first part, chapter 1 presents inductive inferences as logical, using a representation of observations in a cylindrical algebra, a representation of beliefs in terms of a probability assignment over this algebra, and the probability axioms alongside Bayesian updating as the inference rules. I distinguish between a Carnapian and a Bayesian scheme for generating the inductive predictions. Chapter 2 concerns the nature of statistical hypotheses in the Bayesian scheme. The hypotheses can be associated with specific sets in the observation algebra, so that the two schemes can be treated on equal footing. In chapter 3 I argue that the Bayesian scheme has a specific advantage over the Carnapian scheme. The hypotheses offer a natural control over the inductive assumptions underlying the predictions. Where the Carnapian scheme leaves the assumptions implicit, the Bayesian scheme brings them within conceptual grasp.

The Bayesian scheme invites two different discussions in the two other parts of the thesis. The second part concerns the use of the Bayesian scheme in capturing inductive predictions that are sensitive to analogy and independence. In chapter 4, in particular, I discuss the specific class of analogical predictions based on explicit similarity. After providing a system of Carnapian prediction rules, I define the Bayesian scheme that underlies this system. This latter scheme offers some insights into the system of prediction rules, and leads up to a further exploration in chapter 5. This chapter employs the scheme to develop a general model of analogical predictions, but unfortunately it fails to achieve full generality. Chapter 6, finally, employs the same scheme to model predictions for nodes in a Bayesian network. It further shows how the notion of inductive dependence can be incorporated in the Bayesian scheme. More generally, the second part of this thesis illustrates that the Bayesian scheme, by using hypotheses, allows for a better expression of inductive assumptions in a way that stands quite apart from the advantage stressed in the first part: transformations in the hypotheses space allow for the definition of priors that are difficult to define otherwise.

The third part of the thesis considers the Bayesian scheme in relation to three venerable problems in the philosophy of science. Chapter 7 shows that the Bayesian scheme offers a solution to the logical part of the problem of induction, but also that it offers nothing on the epistemological part of the problem. Chapter 8 concerns the problem of inductive inference and theory change. It is shown that the Bayesian scheme suggests a natural place for changes in the inductive assumptions, namely in a change of the statistical hypotheses. It further develops the formal tools to ensure that such changes remain as conservative as possible. Finally, chapter 9 concerns the use of purely theoretical concepts and distinctions in inductive inference, and thus relates to the problem of underdetermination. It shows that such distinctions can indeed be useful, and suggests a further exploration of this fact in a Bayesian model of abductive inference. On the whole the third part claims that the Bayesian scheme is not only suitable for an interesting inductive logical exercise, but that it provides insight into actual scientific methods.

The bigger picture. Let me start by noting that from its conception onwards, probabilistic inductive logic has developed rather slowly, and never really picked up speed. I can see two reasons. First, without meaning to be disloyal to the old masters, there is what I call the curse of Carnap. While the Carnapian framework has certainly been a step forward in studying inductive inference as part of a formal system, both the inherent empiricist view on language and the notion of logical probability have not always been helpful in the development of this system. It may even be conjectured that a failure to disentangle logic from epistemology is the main cause for the problematic development of inductive logic, certainly in comparison to the mature discipline of deductive logic.

As a second reason for the slow development of inductive logic, it appears that Carnapian logic has never really been connected to the main use of probabilistic inductive inference in science, namely in statistical inference. An exception to this is the statistical treatment of $\lambda\gamma$ rules in Festa (1993), which has been a strong source of inspiration for the present thesis. In the larger discussion on scientific methodology, however, Popper had statisticians such as Fisher, Neyman and Pearson on his side, whereas Carnap failed to find fruitful common ground for his logical framework and the tradition of Bayesian statistics.

Apart from that, Carnapian logic has been connected to conceptual problems in the philosophy of science only to a very limited extent. Inductive logic has therefore remained a rather isolated discipline, immersed in its own problems, and at best gesturing towards applications to statistics and scientific method more generally.

This thesis hopes to improve the prospects for inductive logic, both as a separate discipline and as a formal tool for solving problems in methodology and the philosophy of science. With respect to inductive logic as a separate discipline, it proposes a reorientation of the field by pushing two points: the logical perspective, and the Bayesian scheme. The logical perspective obviates the need for a notion of logical probability, and puts strong emphasis on the fact that inductive inference must be valid inference. It further emphasises that inductive logic must make explicit the assumptions underlying inductive inference. This is where the second point becomes effective. The Bayesian scheme employs statistical hypotheses, which are seen to provide access to underlying assumptions. They enlarge the expressive force of inductive logic, and provide a new take on some well-known problems. Thus, while the first part of this thesis simply describes the reorientation of inductive logic, and shows some conceptual advantages of it, the second part shows that this reorientation also results in a better treatment of internal questions. It turns out that certain problems of traditional inductive logic can be solved more easily within a Bayesian scheme.

The use of inductive logic in philosophy of science is illustrated in the third part. It is here suggested that the Bayesian scheme can provide insight into, and to a certain extent solutions for, some problems concerning scientific method. However, much is left to be done in this last research area. First, I suspect that the Bayesian scheme connects naturally to statistical procedures as used in the sciences, but an argumentation for this has not yet been produced. Moreover, it may be noted that there is still a yawning gap between the above schemes and the abundant use of Fisherian and Neyman-Pearson statistics in most of the social sciences. Future research will be directed towards a better understanding of these statistical techniques, and where possible to a reformulation of them in terms of valid statistical inferences. As a second line of development, besides this debate on statistics, I expect that the Bayesian scheme can contribute to many more philosophical debates in the philosophy of science, apart from the ones discussed here.

The need for both observations and theory. In the philosophical debate about the theory-ladenness of observations, I expect Bayesian inductive logic to be particularly helpful. It may be noted that this thesis employs a rather naive framework for the observations, which are supposed to enter the Bayesian schemes as clear-cut and numbered packages of independently obtained information. This seems in direct opposition to the widely shared view that observations are partly determined by theory, in particular that they cannot be described or processed unless we already presuppose some theoretical framework. This point becomes all the more pressing for observations within a scientific experiment, as they are usually processed elaborately before being subjected to statistical analysis. In short, there are strong assumptions inherent to taking observations as clear-cut packages: we assume an unshakeable observation language. However, I submit that the Bayesian scheme contains the conceptual ingredients for a more nuanced view on observations than has been suggested until now. The first part of this thesis makes clear that in the Bayesian scheme, the partition functions as a pair of glasses for looking at the observations. The idea is to take the observations Q_t^q as referring to the raw material of the observation, or in other words the unrefined stimulus. The partition of hypotheses, which determines the impact of the raw observation on beliefs, then concerns the theoretical side of the observations.

This view on the Bayesian scheme, and on inductive inference within it, emphasises that there is not much that observations can convey all by themselves. They always presuppose a theoretical framework, and it is in this sense misguided to hope for objective inductive knowledge. The choice of a partition is similar to the choice of a language in an Carnapian inductive logic, and as Friedman (2004) argues, this choice may again be seen as a relativised and dynamic variant of the Kantian synthetic a priori. The difference is that within the context of Bayesian logic, the choice is within conceptual and formal grasp. On the other hand, as with the Carnapian language choice and the Kantian synthetic a priori, the assumptions underlying inductive knowledge do not convey much by themselves either: statistical partitions usually leave all possibilities open. Once one is provided with these assumptions, the observations are fully responsible for the result of the inductive inferences. It is therefore equally misguided to conclude that in inductive inference, anything goes. As elaborated in the third part of this thesis, inductive knowledge is best seen as a co-production of the observer and the observed, which interact on the strict interface of a logical scheme. The eventual value of the result, in many cases the accuracy of the predictions, thereby depends on making correct observations, on using a proper logic, and finally on the truth of the inductive assumptions.

Choosing inductive assumptions. I want to conclude with some philosophical remarks on the three elements of observation, logic and assumption, starting with the last. I must admit that it is rather disappointing that in this thesis the matter of choosing inductive assumptions has been left aside completely. As may be recalled from chapter 1, the motivation for this disregard is that choosing inductive assumptions is deemed an epistemological issue, or an issue closer to scientific practice, while this thesis is focused exclusively on the logical aspect of inductive inference. Here I want to briefly consider this epistemological aspect after all.

It can first be noted, in particular with respect to science, that choosing inductive assumptions is related to the activity of conjecturing and model building. In other words, asking for the origin of inductive assumptions leads us to the context of discovery, the realm of supposedly irrational and intuitive scientific reasoning. Now I do not think that this side of scientific reasoning is irrational, and like most other philosophers I think it is also a perfectly respectable subject for further research. However, it seems to me that this research is not served best by a restriction to philosophical methods. It also requires empirical research into actual reasoning, which may be accessed by studying the history of science, and perhaps also by performing psychological experiments. The epistemological part of inductive inference is thus moved into the domain of the sciences themselves. Note further that the findings of these empirical studies are likely to differ from the schemes presented in this thesis. After all, sailors do not use fluid mechanics to determine the optimal positions of the sails on their boat. They just follow the rules of sailing. In the same way, actual scientific reasoning will quite probably be a dense network of ad hoc rules rather than a neat logical scheme. It is only by writing down the rules in terms of a Bayesian scheme that we can reconstruct and investigate the inductive assumptions underlying the reasoning.

Spinoza resolves Cartesian doubts. As for the second element of inductive inference, namely making correct observations, I can only offer the kind of basic trust, perhaps best known from Spinoza, that human cognition is by its very nature attuned to the world. This reliabilist trust is obviously not supposed to apply to all convictions, in which case the inductive schemes of this thesis would all become irrelevant. The trust applies only to the raw material of observations, that is, the direct sensory input. And I hold that for this raw material, the reliabilist ideas are in fact rather natural. The starting point of the argument for this is that the cognitive system of a human body is part of this world, or rather, fully submerged in it just as tables and chairs are. In the way in which it interacts with the world on the level of unrefined stimuli, it does not differ in any fundamental sense from tables and chairs, although it is of course much more complex. But if that is so, then to say that human cognition, on the level of unrefined stimuli, is structurally at variance with the world is like saying that a certain type of chair cannot be fitted into space, suspends gravitation, or something of that sort. Under the assumption that human beings are nothing special, in the sense that they are as any other object part of this world, whatever this world consists in, it becomes hard to imagine what doubting the unrefined stimuli amounts to.

Logic as metaphysics? This brings me to the last element of inductive logic, namely that of using proper inference rules. Here I briefly discuss the their epistemic status. It will be clear by now that I think it sadly misguided to aim for a logical scheme that somehow also provides the correct inductive assumptions. The force of a logical scheme is exactly that it provides only the criteria for valid inference, and avoids the whole matter of truth. When it comes to probabilistic inductive logic, I am therefore emphatically against the slogan that probability is the guide of life. On itself, probability cannot tell us anything about life, if only for the simple fact that it is merely a formal tool.

It may be objected that, considered as a formal tool, the logical scheme reveals something synthetic after all. The idea behind this is that there must be something to the logical scheme that ensures its applicability to the world we live in. One may argue that logic is not just a game of symbolic manipulation, but that it really concerns the world, and that it somehow reveals invariances in its structure. Now I am not sure whether there is indeed some structure to the world that makes Bayesian updating the valid inference rule for it, or whether this validity derives only from the form that we choose for assumptions and conclusions. If the former is the case, Bayesian inductive logic does indeed not just accommodate the representation of inference, but it is also a branch of metaphysics. But this, to my mind, stretches the reach of the Bayesian scheme a bit too far.

Samenvatting

Voordat ik een overzicht geef van het proefschrift zelf, zal ik eerst het inductieprobleem, de inductieve logica, en de ideeën van Bayes inleiden. Wie zich wil beperken tot het overzicht, kan bij het kopje *Dit proefschrift* beginnen te lezen.

Eenden en tijgers. De jachtopziener van prinses Perenbloesem houdt sinds jaar en dag een overzicht bij van het aantal eenden dat zich in het voorjaar rond de vijvers ophoudt om te broeden. Daarnaast telt hij jaarlijks het aantal tijgers dat in de zomer de bossen onveilig maakt. Dat doet hij omdat elke winter van hem een prognose verlangd wordt voor het komend jachtseizoen. Het is niet ondenkbaar dat het aantal eenden en tijgers sinds zijn aantreden als volgt gefluctueerd heeft:

Jaar	1	2	3	4	5	6	7	8
Eenden	52	56	24	21	49	64	18	-
Tijgers	3	7	9	3	4	5	8	-

Wat kan hij voor het komend jaar voorspellen? En waarop mag hij zich bij die voorspellingen baseren? Hoe kan hij bijvoorbeeld het feit dat de tijgers zich jaarlijks aan de eenden tegoed doen in zijn voorspellingen verwerken?

Het inductieprobleem. De moed zinkt de jachtopziener in de schoenen wanneer hij in de lange wintermaanden de complete werken van David Hume leest. Daarin wordt uiteengezet dat tot op heden verzamelde gegevens op zichzelf niets vertellen over gegevens die nog verzameld moeten worden, en dat de verzamelde gegevens bovendien geen eenduidige relatie hebben tot de werkelijkheid waaruit de gegevens afkomstig zijn. De jachtopziener kan niet anders dan concluderen dat deze Hume gelijk heeft. Elke invuloefening van de achtste kolom is gebaseerd op een regelmaat die hij in de gegevens meent te ontwaren, maar telkens moet hij toegeven dat er ook wel andere regelmatigheden te verzinnen zijn. Nu is er wel een bepaalde regelmaat waarnaar de voorkeur van de opziener uitgaat, omdat die het beste aansluit op het verhaal dat over de gegevens verteld kan worden. Maar dat verhaal zou volgens de jachtopziener juist uit de gegevens moeten volgen, en het mag daaraan niet vooraf gaan. Het zogenaamde inductieprobleem van Hume bezorgt de opziener slapeloze nachten.

Gelukkig vindt de jachtopziener nog andere boeken in de bibliotheek van het jachthuis: een artikel van Bayes, een traktaat van Carnap en een boek van Fisher. Deze auteurs beweren dat uit de gegevens misschien geen unieke voorspelling volgt, maar dat het met de tabel in de hand wel mogelijk is om bepaalde regelmatigheden en voorspellingen waarschijnlijker te noemen dan andere. Het gat tussen de gegevens enerzijds en de regelmaat en voorspellingen anderzijds wordt dus gedicht met het begrip waarschijnlijkheid. Dat lijkt de oplossing, maar bij de opziener blijft er iets knagen. Allereerst is die tabel een nogal karige representatie van de grillige en veelzijdige werkelijkheid van eenden en tijgers. Maar nog los daarvan: als zich al een regelmaat voordoet in zijn tabel, dan zou het toch niet aan de wiskundige statistiek moeten zijn om de relevante regelmaat aan te wijzen? Waar precies wordt de keuze voor een bepaalde regelmaat gemaakt, en wat is daarna dan nog de rol van de gegevens?

Inductieve logica. Op dit punt van het verhaal was de jachtopziener erbij gebaat geweest in de bibliotheek dit proefschrift aan te treffen. Het behandelt precies de hierboven geformuleerde vragen, en geeft als antwoord daarop een redeneerschema waarin de rol van aangenomen regelmaat en die van de verzamelde gegevens helder uiteengezet wordt. Maar voordat ik dieper inga op dat redeneerschema, wil ik benadrukken dat dit proefschrift niet alleen voor jachtopzieners van belang kan zijn. Alle empirische wetenschappen verzamelen gegevens, en vormen zich op basis daarvan een oordeel over nog niet opgenomen gegevens, en over de werkelijkheid die zich achter de gegevens schuilhoudt. Veel van het empirisch werk gaat zitten in het verkrijgen van de gegevens langs experimentele weg, en in het ordenen van de gegevens in bestanden en tabellen. Daarover gaat dit toch al omvangrijke proefschrift niet. Maar zodra de gegevens eenmaal in nette nullen en enen op een schijf staan, wordt er meestal een statistische techniek op losgelaten. Het onderwerp van dit proefschrift is niets minder dan al deze statistische technieken.

Het uitgangspunt van dit proefschrift is dat het inductieprobleem in twee afzonderlijke delen uiteenvalt. Ten eerste is er het probleem dat wij uit wetenschappelijke gegevens, zoals die van de jachtopziener, niet kunnen opmaken welke regelmaat relevant is. Maar daarnaast is er het probleem dat wij, nadat we gekozen hebben voor de relevantie van een bepaald soort regelmaat, nog geen vaste procedure hebben voor de redenering die ons van die keuze brengt naar voorspellingen, en naar een oordeel over het al dan niet voorkomen van de regelmaat. Kortom, we hebben te maken met een inductieprobleem in de logica, dat de vorm van inductieve redeneringen betreft, en met een inductieprobleem in de kenleer, dat de waarheid van de uitgangspunten in die redeneringen betreft. Dit proefschrift gaat er vanuit dat het mogelijk is om deze twee problemen afzonderlijk te behandelen, en beperkt zich vervolgens tot een behandeling van het logische inductieprobleem. Het volgt hiermee de weg van de klassieke logica, die onder andere door het heldere onderscheid tussen geldigheid en waarheid tot bloei is gekomen.

Redeneren over kansspelen. Uit het bovenstaande mag duidelijk zijn waarom de titel van dit proefschrift 'Inductieve Logica' is. Wat mist, is een verklaring van de term 'Bayesiaans'. Die term heeft betrekking op het feit dat de inductieve logica in dit proefschrift zich bedient van waarschijnlijkheden en kansen. In 1763 publiceert de Royal Society een essay van wijlen dominee Bayes, waarin een aantal baanbrekende ideeën omtrent kans en waarschijnlijkheid worden uiteengezet. Kansspelen zijn dan al langer onderwerp van wiskundige bespiegelingen, maar daarbij gaat het altijd om het bepalen van kansen op bepaalde uitkomsten bij een gegeven kansspel. Het idee van Bayes is om deze relatie om te draaien: de waarschijnlijkheid dat wij meespelen in een bepaald kansspel kan worden afgeleid uit de uitkomsten die zich in het kansspel voordoen.

Neem bijvoorbeeld een dobbelspel waarbij wij alleen van de uitkomsten op de hoogte worden gebracht, zonder dat we weten of de uitkomsten een optelsom zijn van de worpen van twee zeskantige dobbelstenen, ofwel de optelsom van de worpen van een acht- en een vierkantige dobbelsteen. De regels van de kansleer, aldus Bayes, vertellen ons dan precies welke kansen wij aan deze twee mogelijke kansspelen moeten toekennen. De toepassing van dit idee van Bayes beperkt zich echter niet tot spelletjes in casino's. Ook jachtopzieners en andere empirische wetenschappers kunnen er hun voordeel mee doen. Voorwaarde is dat zij de natuur tot op zekere hoogte als een kansspel zien, en zich daarom bedienen van kansen, samengevat in statistische hypothesen. Ze kunnen dan de regels van de kansleer gebruiken als de geldige redeneerregels in inductieve redeneringen over die statistische hypothesen.

Dit proefschrift. Dit proefschrift brengt de hierboven ingeleide onderwerpen samen in de Bayesiaanse inductieve logica. Deze logica voorziet in een oplossing van het inductieprobleem, opgevat als een probleem omtrent de geldigheid van inductieve redeneringen. Deel I presenteert de Bayesiaanse logica zelf. In dit deel wordt beargumenteerd dat statistische hypothesen de formele representatie zijn van aannames omtrent de relevantie van regelmatigheden in de gegevens. Bijzondere aandacht wordt vervolgens besteed aan het gebruik van hypothesen bij het maken van aannames op het gebied van inductieve relevantie en onafhankelijkheid in deel II, en bij formele representaties van theorie en theorieverandering in deel III. Het volgende bespreekt elk van deze delen in meer detail.

Bayesiaanse inductieve logica. Om te beginnen presenteert hoofdstuk 1 de Bayesiaanse inductieve logica, en zet deze af tegen de Carnapiaanse inductieve logica. Het essentiële verschil is dat de Bayesiaanse logica gebruik maakt van statistische hypothesen, waarmee kan worden besloten welke regelmatigheden in de gegevens van belang zijn. Maar de Bayesiaanse logica zelf doet er op dit punt het zwijgen toe. Een empirische wetenschapper moet over de relevante regelmaat helemaal zelf beslissen. En bovendien, als eenmaal tot een verzameling hypothesen besloten is, dan moet daarna nog bepaald worden wat de beginwaarschijnlijkheid van elk van de alternatieven is. Het aardige van Bayesiaanse logica is nu juist dat zij alleen de redeneerregels aandraagt. Alle aannames in de redeneringen komen zodoende expliciet aan bod in de vaststelling van beginwaarschijnlijkheden, zowel die over de relevantie van regelmatigheden, zoals vastgelegd in de hypothesen, als de waarschijnlijkheid over de uitgekozen hypothesen. Het kentheoretische inductieprobleem, dat de uitgangspunten in inductieve redeneringen betreft, wordt met de Bayesiaanse logica daarom beslist niet opgelost.

Het gebruik van hypothesen. Het komt de kritische lezer misschien vreemd voor dat in het bovenstaande moeiteloos verband wordt aangebracht tussen statistische hypothesen en empirisch onderzoek. Als empirische wetenschappers zich uitsluitend willen baseren op tabellen met gegevens, is het gebruik van allerlei veronderstelde kansen achter de gegevens veel te esoterisch. Wat moeten we ons in dit verband bij een statistische hypothese voorstellen?

Hoofdstuk 2 laat zien dat statistische hypothesen in zekere zin wel met de gegevens in verband kunnen worden gebracht, door gebruik te maken van de zogenaamde frequentistische interpretatie van kansen. In die interpretatie zijn kansen altijd verbonden met herhaalbare experimenten, in het geval van de eerder opgevoerde jachtopziener bijvoorbeeld de opeenvolgende jaren. De kans op een bepaalde uitkomst kan dan worden opgevat als de proportie van metingen met die uitkomst in een groot of zelfs oneindig aantal metingen. Ik betoog dat de statistische hypothesen via deze interpretatie een min of meer natuurlijke plaats krijgen in een empiristische inductieve redenering. Bovendien kies ik daarmee voor een specifieke visie op de noties kans en waarschijnlijkheid, en op hun relatie: de eerste komt voor in statistische hypothesen, en is zodoende verbonden aan kansprocessen in de wereld, de tweede drukt de opinie over statistische hypothesen uit.

Ten slotte betoog ik in hoofdstuk 3 dat wetenschappers zichzelf een groot plezier doen wanneer zij hun empiristische reserves opzij zetten, en zich van statistische hypothesen bedienen bij het vaststellen van de beginwaarschijnlijkheden in de eerder besproken inductieve redeneringen. Omdat de hypothesen direct verbonden zijn aan veronderstelde kansprocessen in de natuur, bieden zij een helder zicht op de aannames die in deze redeneringen gemaakt worden.

Relevantie en onafhankelijkheid. Dit laatste punt vormt de opmaat voor een uitgebreide studie naar het gebruik van hypothesen in inductieve redeneringen. Zoals gezegd, de inductieve logica hoeft geen advies uit te brengen over de te kiezen aannames in inductieve redeneringen, omdat dit buiten de taakstelling van de logica valt. Zodra echter de aannames gekozen zijn, kan het wel als deel van de taakstelling worden gezien om de aannames te vertalen naar een vorm die op de logica aansluit. Met andere woorden, bij een logica hoort een handleiding waaruit duidelijk wordt hoe die logica in verschillende redeneringen kan worden toegepast.

De hoofdstukken 4 en 5 leveren die handleiding voor inductieve redeneringen waarin aannames worden gemaakt omtrent de relevantie tussen gegevens onderling, terwijl die gegevens ieder voor zich een vaste kans hebben om zich voor te doen. Een jager kan bijvoorbeeld een vaste kans veronderstellen voor de aanwezigheid van konijnen, eenden of tijgers aan de bosrand, maar niettemin menen dat de aanwezigheid van konijnen de aanwezigheid van eenden waarschijnlijker maakt, terwijl de aanwezigheid van tijgers die aanwezigheid juist minder waarschijnlijk maakt. Zulke relevanties worden in de literatuur ook wel als analogie-effecten aangeduid, omdat ze er uiteindelijk op berusten dat konijnen en eenden iets gemeen hebben dat hen tegenover tijgers plaatst. Het blijkt verrassend moeilijk om zulke aannames omtrent relevantie direct in een beginwaarschijnlijkheid onder te brengen, maar het gebruik van bepaalde statistische hypothesen biedt hierop een natuurlijke ingang. Hoofdstuk 6 gebruikt precies dezelfde statistische hypothesen voor het beschrijven van inductieve redeneringen waarin vooraf afhankelijkheidsrelaties tussen de gegevens zijn aangenomen. Causale relaties tussen aantallen eenden en tijgers kunnen op die manier ook in de inductieve redeneringen worden opgenomen.

Wetenschapsfilosofie. In het laatste deel van dit proefschrift wordt gekeken naar een drietal onderwerpen uit de wetenschapsfilosofie, en meer precies naar de relatie die zij hebben tot de Bayesiaanse inductieve logica. Allereerst wordt in hoofdstuk 7 ingegaan op de precieze functie die gegevens en aangenomen hypothesen hebben in de inductieve redeneringen. Daaruit komt het beeld naar voren van empirische wetenschap als coproductie: zowel observatie als theorie hebben een eigen en onafhankelijke inbreng op wetenschappelijke inzichten. Dit doet sterk denken aan het Kantiaanse perspectief, waarin de werkelijkheid pas in de interactie van een kenapparaat en een onbekende wereld, en dus eigenlijk onder onze ogen en handen, ontstaat. Het verschil is dat in het geval van de statistische redeneringen het kenapparaat geen transcendentaal bepaalde notie is, maar een vrij te kiezen en naar inzicht te wijzigen statistisch model.

Over het wijzigen van statistische modellen gaat hoofdstuk 8. Dit hoofdstuk keert zich tegen de opvatting dat theorieveranderingen zich onttrekken aan rationaliteitscriteria, en laat zien dat zulke veranderingen goed kunnen worden ingepast in een Bayesiaans logisch schema. Dit werkt voor zover de theorie samenvalt met de statistische hypothesen die door de theorie gemotiveerd worden, en zolang de theorie dus samenvalt met de waarschijnlijkheden die de theorie over de gegevens vastlegt. In hoofdstuk 9 behandel ik ten slotte het deel van wetenschappelijke theorieën dat voor eens en voor altijd boven die gegevens uitstijgt. Wetenschappelijke theorieën worden nu eenmaal nooit eenduidig door de gegevens vastgelegd. Het laatste hoofdstuk betoogt echter dat in sommige gevallen de onderbepaaldheid van de theorie door de gegevens een heuristisch en praktisch nut heeft voor de wetenschapper. In plaats van ons zorgen te maken over de onderbepaaldheid, doen we er misschien beter aan te proberen de functie van onderbepaaldheid in concrete gevallen te achterhalen.

Statistiek voor de wetenschappen. In deze slotparagraaf kom ik terug op het nut van dit proefschrift voor de empirische wetenschappen. Veel van die wetenschappen gebruiken sinds jaar en dag klassieke statistische methoden, zoals die van Fisher, Neyman en Pearson, en anderen. Met behulp van die methoden worden soms ongeldige redeneringen gemaakt, en in het ergste geval worden daardoor zelfs onjuiste conclusies getrokken. Toch is het onverstandig om de klassieke statistiek af te raden. In de uitgewerkte schattings- en toetsingsprocedures van de klassieke statistiek ligt vaak een schat aan situatiespecifieke kennis opgeslagen, en in veel gevallen zijn de procedures efficiënt en rekentechnisch aantrekkelijk. Het zijn in zekere zin epistemische 'shortcuts'. In plaats daarvan is de normatieve aanbeveling van dit proefschrift dat statistische procedures op hun situatiespecifieke geldigheid en toepasbaarheid kunnen worden beoordeeld door ze uit te schrijven in termen van Bayesiaanse statistische redeneringen.

Naast een kwaliteitscontrole biedt dit aan wetenschappers een beter zicht op de uitgangspunten van hun activiteiten.

Dankwoord

Dit proefschrift is naast een academisch boek ook de afsluiting van een opleiding tot zelfstandig onderzoeker. En hoe objectief en inhoudelijk die opleiding ook is, leermomenten zijn altijd persoonlijk. Ik wil hier een aantal mensen die mij de afgelopen jaren iets geleerd hebben, persoonlijk bedanken.

Allereerst ben ik een aantal filosofen uit Groningen dank verschuldigd. Ik dank in het bijzonder Theo Kuipers, voor zijn gedegen kennis, zijn vertrouwen, en zijn schat aan onderzoekservaring. Daarnaast dank ik Jeanne Peijnenburg, voor haar steun, haar enthousiasme en haar wijsheid. David Atkinson dank ik voor zijn belangstelling en zijn waardevolle commentaar, en Jeanne en David samen dank ik voor hun gastvrijheid en hun dakterras. Meer in het algemeen wil ik de Faculteit Wijsbegeerte in Groningen als geheel bedanken. Het is een bijzondere en vriendelijke omgeving. Ik dank ook alle studenten die in de afgelopen jaren mijn pad hebben gekruist. Aan de colleges en incidentele begeleiding heb ik heel veel plezier beleefd. En verder bedank ik de leesclub voor onvergetelijke discussies, Menno voor zijn eerlijkheid, Barteld omdat hij tragedies met mij las, Sjoerd voor zijn goede hart, Martine voor haar vriendschap, Casper voor alle plezier in de kroeg, en Kim voor haar gezelligheid, intelligentie en geweldige maaltijden.

In de afgelopen vier jaar bracht ik ook veel tijd buiten het facultaire leven of aan de telefoon door. Voor het laatste is met name Igor Douven verantwoordelijk. Ik dank hem hartelijk voor alle inzichten die hij mij verschafte, maar ook voor zijn bemoedigende woorden en zijn vriendschap. Met veel plezier bezocht ik voorts mijn academische vrienden Beerend, Boudewijn, Frank, Hedde, Maarten en Wouter. Het meeste leren promovendi misschien wel van elkaar. De absurdistische afleiding in mijn Groningse leven werd verzorgd door Heiner, Jasper, Willem en vooral Maarten met zijn augurken. En dan bedank ik nog mijn buurman Mike die mijn muziek moest luisteren, Loek die me veel over mezelf leert, Evert voor zijn onafhankelijke geest, Evelyne voor haar bijzondere post, Ioannis omdat hij dit niet kan lezen, Jan voor de twijfel, Hayo voor de ruimte, Chantal voor wie ze is, Martijn voor het zijn zelf, mijn zus Marleen voor alles dat we delen, Beerend voor zijn geweldige nieuwsgierigheid, Bouwe voor zijn onvoorwaardelijke vriendschap, en ten slotte Marian, voor de liefde.

Mijn paranimfen dank ik in het bijzonder. Ik dank Joep, die met mij op reis ging, en me leerde om uiteindelijk mijzelf te bedanken. En ik dank Melle, die al heel lang met mij op reis is. Hij leerde mij dat een mens het stevigst staat waar hij niets heeft om op te staan. Eindeloze dank ben ik verschuldigd aan Huib, paranimf in het leven, die mij geleerd heeft dat je door een steen te poetsen geen spiegel kunt maken.

Ten slotte dank ik mijn ouders, die mij hebben leren praten en kijken. Aan hen draag ik dit proefschrift op.

References

Achinstein, P. (1963) 'Confirmation Theory, Order, and Periodicity', *Philosophy of Science* 30, pp. 17–35.

Albert, M. (1992) 'Bayesian Learning when Chaos looms large', *Economics Letters* 65, pp. 1–7.

Albert, D. Z. (2000) *Time and Chance*, Cambridge (MA): Harvard University Press.

Akaike, H. (1978) 'A Bayesian Analysis of the Minimum AIC Procedure', *Annals of the Institute of Statistical Mathematics* 30, pp. 9–14.

Armendt, B. (1980) 'Is There a Dutch Book Argument for Probability Kinematics?', *Philosophy of Science* 47, pp. 583–588.

Armstrong, D. (1973) *Belief, Truth and Knowledge*, Cambridge: Cambridge University Press.

Bacchus, F., Kyburg jr., H. E. and Thalos, M. (1990) 'Against Conditionalisation', *Synthese* 85, pp. 475–506.

Bandyopadhyay, P. and Boik, R. (1999) 'The Curve Fitting Problem: A Bayesian Rejoinder', *Philosophy of Science* 66, pp. S390–402.

Barnet, V. (1999) *Comparative Statistical Inference*, New York: John Wiley.

Benthem, J. van (2003) 'Conditioning meets Update Logic', *Journal of Logic, Language and Information* 12, pp. 409–421.

Berger, J. O. and Wolpert, R. L. (1984) *The Likelihood Principle*, Hayward: Institute of Mathematical Statistics.

Billingsley, P. (1995) *Probability and Measure*, New York: John Wiley.

Birnbaum, A. (1962) 'On the Foundations of Statistical Inference', *Journal of the American Statistical Association* 57, pp. 269–326.

Bogdan, R. (1976) *Local Induction*, Dordrecht: Reidel.

Boolos, G. and Jeffrey, R. (1974) *Computability and Logic*, Cambridge: Cambridge University Press.

Bovens, L. and Hartmann, S. (2004) *Bayesian Epistemology*, Oxford: Oxford University Press.

Carnap, R. (1950) *The Foundations of Probability*, Chicago: University of Chicago Press.

Carnap, R. (1952) *The Continuum of Inductive Methods*, Chicago: University of Chicago Press.

Carnap, R. (1980) 'A Basic System of Inductive Logic, Part II', in Jeffrey, R.C. (ed.), Studies in Inductive Logic and Probability Vol.2, Berkeley: University of California Press, pp. 7–150.

Carnap, R., and Jeffrey, R. (eds.) (1971) *Studies in Inductive Logic and Probability*, Vol. 1, Berkeley: University of California Press.

Carnap, R., and Stegmüller, W. (1959) *Inductive Logik und Wahrscheinlichkeit*, Wien: Springer Verlag.

Cartwright, N. (1999) *The Dappled World*, Cambridge: Cambridge University Press.

Chihara, M. (1987) 'Some Problems for Bayesian Confirmation Theory', *British Journal for the Philosophy of Science* 38, pp. 551–60.

Christensen, D. (2000) 'Dynamic Coherence versus Epistemic Impartiality', *The Philosophical Review* 109(3), pp. 349–371.

Cohen, L. J. (191989) *The Philosophy of Induction and Probability*, Oxford: Clarendon Press.

Corfield, D. and Williamson, J. (2002) *Foundations of Bayesianism*, Dordrecht: Kluwer Academic Publishers.

Costantini, D. (1979) 'The Relevance Quotient', *Erkenntnis* 14, pp. 149–57.

Cox, R. T. (1961) *The Algebra of Probable Inference*, Baltimore: John Hopkins University Press.

Cramér, H. (1946) *Mathematical methods of Statistics*, Princeton: Princeton University Press.

Dawid, A. P. (1982) "The Well–Calibrated Bayesian", *Journal of the American Statistical Association* 77, pp. 605–613.

De Finetti, B. (1937) 'Foresight: its logical laws, its subjective sources' in *Studies in Subjective Probability*, eds. Kyburg, H. and Smokler, H. (1964), New York: John Wiley, pp. 97–158.

De Finetti, B. (1972) *Probability, Induction and Statistics*, New York: John Wiley.

De Finetti, B. (1974) *Theory of Probability*, New York: John Wiley.

Diaconis P. and Freedman, M. (1980) 'De Finetti's Theorem for Markov Chains', *Annals of Probability* 8, pp. 115–30.

Dijksterhuis, E. J. (1950) *De Mechanisering van het Wereldbeeld*, Amsterdam: Meulenhof.

Di Maio, M. C. (1995) 'Predictive Probability and Analogy by Similarity in Inductive Logic', *Erkenntnis* 43, pp. 369 – 394.

Doob, J. L. (1953) *Stochastic Processes*, New York: John Wiley.

DORLING, J. (1992) 'Bayesian Conditionalisation Resolves Positivist/Realist Disputes', *Journal of Philosophy* 89, pp. 362–382.

DORR, C. (2002) 'Sleeping Beauty: In Defence of Elga', *Analysis* 62, pp. 292–96.

DOUVEN, I. (1999) 'Inference to the Best Explanation is Coherent', *Philosophy of Science* 66, pp. S424–36.

DOUVEN, I. (2004) 'Empirical Equivalence, Explanatory Force, and the Inference to the Best Theory' in A. Aliseda, R. Festa and J. Peijnenburg *Logics of Scientific Cognition: Essays in Debate with Theo Kuipers*, Atlanta: Rodopi, in press.

DOUVEN, I. (2005) 'Evidence, Explanation, and the Empirical Status of Scientific Realism', *Erkenntnis*, forthcoming.

DOUVEN, I. AND MEIJS, W. (2005) 'Bootstrap Confirmation Made Quantitive' *Synthese*, forthcoming.

DUBINS, L. E. (1965) *Inequalities for Stochastic Processes, How to Gamble if You Must*, New York: Dover.

EARMAN, J. (1992) *Bayes or Bust?*, Cambridge (MA): MIT Press.

EFRON, B. (1986) 'Why Isn't Everyone a Bayesian', *American Statistician* 40(1), pp. 1–6.

ELGA, A. (2000) 'Self Locating Belief and the Sleeping Beauty Problem', *Analysis* 60, pp. 143–47.

FESTA, R. (1993) *Optimum Inductive Methods*, Dordrecht: Kluwer.

FESTA, R. (1997) 'Analogy and Exchangeability in Predictive Inferences', *Erkenntnis* 45, pp. 89–112.

FIELD , H. (1978) 'A Note on Jeffrey Conditionalisation', *Philosophy of Science* 45, pp. 361–367.

FISHER, R. A. (1956) *Statistical Methods and Scientific Inference*, Edinburgh: Oliver and Boyd.

FITELSON, B. (1999) 'The Plurality of Bayesian Measures of Confirmation and the Problem of Measure Sensitivity', *Philosophy of Science* 66, pp. S362–78.

FITELSON, B. (2005) 'Inductive Logic', *P*hilosophy of Science: an Encyclopedia, Pfeifer, J. and Sarkar, S. (eds.), London: Routledge.

FRIEDMAN, M. (2004) *The Dynamics of Reason*, Stanford: CSLI Publications.

GAIFMAN, H. (1986) 'A Theory of Higher Order Probabilities', in Halpern, J. (ed.), *Proceedings of TARK 1986*, San Mateo: Morgan–Kauffman, pp. 275–292.

Gaifman, H. and Snir, M. (1982) 'Probabilities over Rich Languages', *Journal of Symbolic Logic* 47, pp. 495–548.

Gärdenfors, P. (1979) 'Forecasts, Decisions and Uncertain Probabilities' *Erkenntnis* 14, pp. 159–181.

Gärdenfors, P. (1988) *Knowledge in Flux*, Boston: MIT Press.

Gärdenfors, P. and Sahlin, N. E. (eds.) (1988) *Decision, Probability, and Utility*, Cambridge: Cambridge University Press.

Glymour, C. (1980) 'Why I am not a Bayesian' in *Theory and Evidence*, Princeton: Princeton University Press.

Glymour, C. (1987) *Discovering Causal Structure*, London: Academic Press.

Gillies, D. (2000) *Theories of probability*, London: Routledge.

Gillies, D. (2001) 'Bayesianism and the Fixity of the Theoretical Framework' in *Foundations of Bayesianism* ed. Corfield, D. and Williamson, J., Dordrecht: Kluwer, pp. 363–379.

Giere, R.N. (1988) *Explaining Science*, University of Chicago Press, Chicago.

Gigerenzer, G. N. and Selten, R. (eds.) (2001) *Bounded Rationality: the Adaptive Toolbox*, Cambridge (MA): MIT Press.

Good, I. J. (1955) *The Estimation of Probabilities: an Essay on Modern Bayesian Methods*, Cambridge (MA): MIT press.

Goodman, N. (1955) *Fact, Fiction, and Forecast*, Cambridge (MA): Harvard University Press.

Graves, J. (1974) 'Uniformity and Induction', *British Journal for the Philosophy of Science* 24, pp. 301–318.

Hacking, I. (1965) *The Logical Foundations of Probability*, Cambridge: Cambridge University Press.

Hacking, I. (1975) *The Emergence of Probability*, Cambridge: Cambridge University Press.

Harman, G. (1986) *Change in View*, Boston: MIT Press.

Hájek, A. (1997) '"Mises Redux"–Redux', *Erkenntnis* 45, pp. 209–27.

Henkin, L. et al. *Cylindric set algebras and related structures*, Berlin: Springer.

Hintikka, J. (1966) 'A Two–dimensional Continuum of Inductive Methods' in *Aspects of Inductive Logic* ed. Hintikka, J. and Suppes, P., Amsterdam: North Holland.

Hintikka, J. (1970) 'Unknown Probabilities, Bayesianism, and De Finetti's Representation Theorem' in *Boston Studies in the Philosophy of Science*, Vol. VIII, eds. Buck, R. C. and Cohen, R. S., Dordrecht: Reidel.

HINTIKKA, J. (1997) 'Comment on Theo Kuipers' in *Knowledge and Inquiry*, ed. Sintonen, M., Amsterdam: Rodopi.

HINTIKKA, J. AND NIINILUOTO, I. (1976) 'An Axiomatic Foundation of the Logic of Inductive Generalisation' in *Formal Methods in the Methodology of the Empirical Sciences*, Przelecki, M., Szaniawski, K. and Wójcicki, R. (eds.), Synthese Library 103, Dordrecht: Kluwer.

HINTIKKA, J. AND SUPPES, P. (1966) *Aspects of Inductive Logic*, Amsterdam: North Holland.

HITCHCOCK, C. AND SOBER, E. (2004) 'Prediction versus Accommodation and the Risk of Overfitting' *British Journal for the Philosophy of Science* 55(1), pp. 1–34.

HOWSON, C. (1973) 'Must the Logical Probability of Laws be Zero?', *British Journal for the Philosophy of Science* 24, pp. 153–82.

HOWSON, C. AND URBACH, P. (1996) *Scientific Reasoning: the Bayesian Approach*, Chicago: Open Court Publishing Company.

HOWSON, C. (1997) 'Bayesian Rules of Updating', *Erkenntnis* 45, pp. 195–208.

HOWSON, C. (1997) 'A Logic of Induction', *Philosophy of Science* 64, pp. 268–90.

HOWSON, C. (2000) *Hume's Problem*, Oxford: Clarendon Press.

HUME, D. (1748) *An Enquiry Concerning Human Understanding*, Tom Beauchamp (ed.) (1999), Oxford: Oxford University Press.

JAYNES, E. T. (1998) *Probability Theory: The Logic of Science*, Cambridge: Canbridge University Press.

JEFFREY, R. C. (1965) *The Logic of Decision*, Chicago: Chicago University Press.

JEFFREY, R. C. (1973) 'Carnap's Inductive Logic', *Synthese* 25, pp. 299–306.

JEFFREY, R. (1977) 'Mises redux' in *Problems in Methodology and Linguistics*, eds. Butts, R.E. and Hintikka, J., Dordrecht: Kluwer Academic Publishers.

JEFFREY, R. C. (1984) *Probability and the Art of Judgement*, Cambridge: Cambridge University Press.

JEFFREYS, H. (1931) *Scientific Inference*, Cambridge: Cambridge University Press.

JEFFREYS, H. (1939) *Theory of Probability*, Oxford: Oxford University Press.

JOYCE, J. M. (1998) 'A non–pragmatic Vindication of Probabilism', *Philosophy of Science* 65(4), pp. 575–603.

JOHNSON, W. (1932) ‘Probability: the deductive and inductive problems’, *Mind* 49, pp. 409–423.

KAHNEMAN, D., SLOVIC, P. AND TVERSKY A. (EDS.) (1982) *Judgement under Uncertainty*, Cambridge: Cambridge University Press.

KELLY, K. (1996) *The Logic of Reliable Inquiry*, New York: Oxford University Press.

KELLY, K. (1991) ‘Reichenbach, Induction and Discovery’, *Erkenntnis* 35, pp. 123–49.

KELLY, K. (1997) ‘Learning Theory and the Philosophy of Science’, *Philosophy of Science* 64, pp. 245–67.

KELLY, K. (1999) ‘Iterated Belief Revision, Reliability, and Inductive Amnesia’, *Erkenntnis* 50, pp. 11–58.

KEMENY, J. (1963) ‘Carnap’s Theory of Probability and Induction’ in *The Philosophy of Rudolph Carnap*, ed. Schilpp, P.A., Illinois: Open Court.

KEYNES, J.M. (1921) *A Treatise on Probability*, London: MacMillan.

KIESEPPÄ, I. A. (1997) ‘Akaike Information Criterion, Curve Fitting, and the Philosophical Problem of Simplicity’, *British Journal for the Philosophy of Science* 48, pp. 21–48.

KHINCHIN, A. I. (1949) *Mathematical Foundations of Statistical Mechanics*, New York: Dover.

KOLMOGOROV, A.N. (1933) *Foundations of the Theory of Probability*, translated Morrison, N. (1950), New York: Chelsea Publishing Company.

KOOI, B. P. (2003) *Knowledge, Chance and Change*, Dissertation University of Groningen.

KUHN, T. (1962) *The Structure of Scientific Revolutions*, Chicago: University of Chicago Press.

KUIPERS, T.A.F. (1978) *Studies in Inductive Probability and Rational Expectation*, Dordrecht: Reidel.

KUIPERS, T.A.F. (1984) ‘Two Types of Inductive Analogy by Similarity’, *Erkenntnis* 21, pp. 63–87.

KUIPERS, T.A.F. (1988) ‘Inductive Analogy by Similarity and Proximity’ in D.H. Helman (ed.) *Analogical Reasoning*, pp. 299–313, Dordrecht: Kluwer.

KUIPERS, T.A.F. (1997) ‘The Carnap–Hintikka Programme in Inductive Logic’ in *Knowledge and Inquiry*, ed. Sintonen, M., Rodopi, Amsterdam.

KUIPERS, T. A. F. (2000) *From Instrumentalism to Constructive Realism*, Dordrecht: Kluwer.

KULLBACK, S. (1959) *Information Theory and Statistics*, New York: John Wiley.

KYBURG, H. AND SMOKLER, H. (EDS.) (1964) *Studies in Subjective Probability*, New York: John Wiley.

KYBURG, H. (1974) *The Logical Foundations of Statistical Inference*, Dordrecht: Reidel.

LADYMAN, J. (1998) 'What is Structural Realism', *Studies in the History and Philosophy of Science* 29, pp. 409–424.

LAKATOS, I. (1968) 'Changes in the Problem of Inductive Logic', in *The Problem of Inductive Logic*, Proceedings of the International Colloquium in the Philosophy of Science, Vol. 2, Lakatos, I. (ed.), Amsterdam: North-Holland.

LAM, W. AND BACCHUS, F. (1994) 'Using New Data to Refine a Bayesian Network' in *Proceedings of the 10th Conference on Uncertainty in Artificial Intelligence*, New York: Morgan Kaufman, pp. 383–90.

LAMBALGEN, M. VAN (1987) *Random Sequences*, Dissertation University of Amsterdam.

LAPLACE, P. S. (1951) *A Philosophical Essay on Probabilities*, transl. F.W. Truscott and F. L. Emory, New York: Dover.

LEWIS, D. (191971) 'Immodest Inductive Methods', *Philosophy of Science* 38, pp. 54–63.

LEWIS, D. (1980) 'A Subjectivist Guide to Objective Chance' in *Philosophical Papers* Vol. 2 (1986), New York: Oxford University Press, pp. 83–132.

LEWIS, D. (2001) 'Sleeping Beauty: reply to Elga' in *Analysis* 61, pp. 171–175.

MACKAY, D. (2003) *Information Theory, Inference, and Learning Algorithms*, Cambridge: Cambridge University Press.

MACH, E. (1980) *Natuurkunde, Wetenschap en Filosofie*, transl. W. de Ruiter, Amsterdam: Boom.

MAHER, P. (1993) *Betting on Theories*, Cambridge: Cambridge University Press.

MAHER, P. (1993) 'Diachronic Rationality', *Philosophy of Science* 59, pp. 120–141.

MAHER, P. (2000) 'Probabilities for Two Properties', *Erkenntnis* 52, pp. 63–91.

MAHER, P. (2001) 'Probabilities for Multiple Properties: the models of Hesse and Carnap and Kemeny', *Erkenntnis* 55, pp. 183–216.

MAHER, P. (2004) 'Probability Captures the Logic of Scientific Confirmation', preprint.

Marsden, J. E., Tromba, A. J. (1988) *Vector Calculus*, 3rd Edition, New York: Freeman.

Mayo, D.G. (1996) *Error and the growth of scientific Knowledge*, Chicago: University of Chicago Press.

Mikosh, T. (1998) *Elementary Stochastic Calculus*, London: World Scientific.

Mood, F. A. and Graybill, D. C. B. (1973) *Introduction to the Theory of Statistics*, New York: McGraw-Hill.

Murdoch, D. (2002) 'Induction, Hume and Probability', *Journal of Philosophy* 99, pp. 185–99.

Niiniluoto, I. (1976) 'Inquiries, Problems, and Questions: Remarks on Local Induction' in *Local Induction*, ed. Bogdan, R., Dordrecht: Reidel.

Niiniluoto, I. (1981) 'Analogy and Inductive Logic', *Erkenntnis* 16, pp. 1–34.

Niiniluoto, I. 1983 'Novel Facts and Bayesianism', *British Journal for the Philosophy of Science* 34, pp. 375–379.

Niiniluoto, I. and Tuomela, R. (1973) *Theoretical Concepts and Hypothetico–Deductive Inference*, Dordrecht: Reidel.

Norton, J. D. (2003) 'A Material Theory of Induction', Philosophy of Science 70, pp. 647–670.

Okasha, S. (2001) 'What did Hume Really Show about Induction?', *Philosophical Quarterly* 51, pp. 307–27.

Papineau, D. (1987) *Reality and Representation*, Oxford: Blackwell.

Paris, J. (1994) *The Uncertain Reasoner's Companion*, Cambridge: Cambridge University Press.

Paris, J. and Vencovská, A. (1997) 'In defence of the Maximum Entropy Inference Process', *International Journal of Automated Reasoning* 17, pp. 77–103

Pearl, J. (1988) *Probabilistic Reasoning in Intelligent Systems*, San Mateo: Morgan Kaufman.

Peirce, C. (1934) *Collected Papers Vol. 5*, Harvard University Press, Cambridge (MA).

Poincaré, H. (1952) *Science and Hypothesis*, New York: Dover.

Polya, G. (1954) *Patterns of Plausible Inference*, Vol. 2 of *Mathematics and Plausible Reasoning*, Princeton: Princeton University Press.

Popper, K. (1959) *The Logic of Scientific Discovery*, London: Hutchinson.

Press, S. J.(1989) *Bayesian Statistics: Principles, Models, and Applications*, New York: Wiley.

PSILLOS, S. (1995) 'Is Structural Realism the Best of Both Worlds?', *Dialectica* 49, pp. 15–46

PUTNAM, H. (1963) 'Degree of Confirmation and inductive logic' in P.A. Schilpp (ed.) *The Philosophy of Rudolf Carnap*, La Salle: Open Court, reprinted in Putnam, H., *Mathematics, Matter and Method*, 1975, Cambridge: Cambridge University Press, pp. 270–92.

PUTNAM, H. (1963) 'Probability and Confirmation', *The Voice of America* 10, reprinted in *Mathematics, Matter and Method* (1975), Cambridge: Cambridge University Press, pp. 293–304.

RAMSEY, F. (1926) 'Truth and Probability' in *Studies in Subjective Probability*, eds. Kyburg, H. and Smokler, H. (1964), New York: John Wiley, pp. 23–52.

REICHENBACH, H. (1948) *The Theory of Probability*, Stanford: University of California Press.

RÉNYI, A. (1970) *Foundations of Probability*, San Francisco: Holdaen-Day.

RISSANEN, J. (1982) 'A Universal Prior for Integers and Estimation by Minimal Description Length', *Annals of Statistics* 11, pp. 416–31.

ROMEYN, J. W. (2004) 'Hypotheses and Inductive Predictions', *Synthese* 141(3), pp. 333–64.

ROMEYN, J. W. (2005) 'Theory Change and Bayesian Statistical Inference', *Philosophy of Science*, forthcoming.

ROMEYN, J. W. (2005) 'Analogical Predictions for Explicit Similarity', *Erkenntnis*, forthcoming.

ROSENKRANTZ, R. (1977) *Inference, Method and Decision*, Dordrecht: Reidel.

ROSENKRANTZ, R. (1992) 'The Justification of Induction', *Philosophy of Science* 59, pp. 527–39.

ROTT, H. (1999) 'Coherence and Conservatism in the Dynamics of Belief', *Erkenntnis* 50, pp. 387–412.

SAHLIN, N. E. (1990) *The Philosophy of F. P. Ramsey*, Cambridge: Cambridge University Press.

SAHLIN, N. E. (1983) 'On Second–Order Probability and the Notion of Epistemic Risk' in Stigum, B. P. and Wenstrup, F. (eds.) *Foundations of Utility and Risk Theory with Applications*, Dordrecht: Reidel.

SALMON, W. (1966) *The Foundations of Scientific Inference*, Pittsburgh: Pittsburgh University Press.

SALMON, W. (1981) 'Rational Prediction', *British Journal for the Philosophy of Science* 32, pp. 115–25.

SAVAGE, L. J. (1954) *The Foundations of Statistics*, New York: Dover.

Schilpp, P. A. (1963) *The Philosophy of Rudolf Carnap*, The Library of Living Philosophers 11, Illinois: Open Court.

Scott, D. and Kraus, P. (1966) 'Assigning Probability to Logical Formulas' in *Aspects of Inductive Logic* ed. Hintikka, J. and Suppes, P., Amsterdam: North Holland, pp. 219–64.

Seidenfeld, T. and Schervish, M. J. (1983) 'A Conflict between Finite Additivity and Avoiding Dutch Book', *Philosophy of Science* 50, pp. 398–412.

Selz, O. (1913) *Über die Gesetze des geordneten Denkverlaufs*, Stuttgart: W. Spemann.

Shafer, G. (1982) 'Bayes's Two Arguments for the Rule of Conditioning', *Annals of Statistics* 10(4), pp. 1075–89.

Shore, J. E. and Johnson, R. W. (1980) 'Axiomatic Derivation of the Principle of Maximum Entropy and the Principle of Minimum Cross–Entropy', *IEEE Transactions on Information Theory*, Vol. 26(1), pp. 26–37.

Skyrms, B. (1987) 'Dynamic Coherence and Probability Kinematics', *Philosophy of Science* 54, pp. 1–20.

Skyrms, B. (1990) *The Dynamics of Rational Deliberation*, Cambridge (MA): Harvard University Press.

Skyrms, B. (1991) 'Carnapian Inductive Logic for Markov Chains', *Erkenntnis* 35, pp. 439–460.

Skyrms, B. (1993) 'Analogy by Similarity in Hyper–Carnapian Inductive Logic', in J. Earman, A.I. Janis, G. Massey, and N. Rescher (eds.), *Philosophical Problems of the Internal and External Worlds*, Pittsburgh: University of Pittsburgh Press, pp. 273–282.

Skyrms, B. (1993) P. French, T. Uehling, Jr., and H. Wettstein (eds.), Midwest Studies in Philosophy, vol. XVIII, University of Notre Dame Press, Notre Dame, pp. 78–89.

Skyrms, B. (1996) 'Bayesian Projectability' in *Grue!*, ed. Stalker, D., La Salle: Open Court.

Sober, E. (1998) 'Simplicity', *Routledge Encyclopedia of Philosophy*, Vol. 8, London: Routledge, pp. 780–3.

Spielman, S. (1977) 'Physical Probability and Bayesian Statistics', *Synthese* 36, pp. 235–69.

Spinoza, B. de (2002) *Verhandeling over de Verbetering van het Verstand*, Groningen: Historische Uitgeverij.

Spohn, W. (1981) 'Analogy and Inductive Logic: a Note on Niiniluoto', *Erkenntnis* 21, pp. 35–52.

Spohn, W. (2002) 'Laws, *Ceteris Paribus* Conditions, and the Dynamics of Belief', *Erkenntnis* 57(3), pp. 373–94.

Stalker, D. (1996) *Grue!*, La Salle: Open Court.

Stegmüller, W. (1973) *Carnap II: Normatieve Theorie des Inductiven Räsonierens*, part C (Volume iv), Berlin: Springer.

Stenning, K. and Lambalgen, M. van (2001) 'Semantics as a Foundation for Psychology: A Case Study of Wason's Selection Task', *Journal of Logic, Language and Information* 10, pp. 273–317.

Suppes, P. (1966) 'Concept Formation and Bayesian Decision' in *Aspects of Inductive Logic*, eds. Hintikka, J. and Suppes, P., Amsterdam: North Holland Publishers.

Suppes, P. (2002) *Representation and Invariance of Scientific Structures*, Stanford: CSLI Publications.

Tong, S. and Koller, D. (2001) 'Active Learning for Structure in Bayesian Network' in *Proceedings of the International Joint Conference on Artificial Intelligence*, San Francisco: Morgan Kaufman, pp. 863–69.

Tuomela, R. (1966) 'Induction in Ordered Universes' in *Aspects of Inductive Logic*, eds. Hintikka, J. and Suppes, P., Amsterdam: North Holland Publishers.

Tuomela, R. (1973) *Theoretical Concepts*, Wien: Springer.

Uffink, J. (1990) *Grondslagen van het Waarschijnlijkheidsbegrip*, manuscript.

Uffink, J. (1997) 'Can the maximum entropy principle be explained as a consistency requirement?', manuscript.

Van Fraassen, B. C. (1980) *The Scientific Image*, Oxford: Clarendon Press.

Van Fraassen, B. C. (1989) *Laws and Symmetry*, Oxford: Clarendon Press.

Von Mises, R. (1928) *Probability, Statistics, and Truth*, London: Dover.

Von Mises, R. (1964) *Mathematical Theory of Probability and Statistics*, New York: Academic Press.

Von Plato, J. (1981) 'On Partial Exchangeability as a Generalisation of Symmetry Principles' *Erkenntnis* 16, pp. 53–59.

Von Plato, J. (1994) *Creating Modern Probability*, Cambridge: Cambridge University Press.

Von Wright, G. H. (1957) *The Logical Problem of Induction*, Oxford: Blackwell.

VOTSIS, I. (2005) 'The Upward Path to Structural Realism', *Philosophy of Science*, forthcoming.

WASON, P. J. (1968) 'Reasoning about a Rule', *Quarterly Journal of Experimental Psychology* 20, pp. 273–281.

WAGNER, K. (2002) 'Probability Kinematics and Commutativity', *Philosophy of Science* 69, pp. 266–78.

WILLIAMS, P. M. (1976) 'Indeterminate Probabilities' in *Formal Methods in the Methodology of the Empirical Sciences*, Prezlecki, M., Szaniawski, K. and Wojcicki, R. (eds.), Dordrecht: Reidel.

WILLIAMS, P. M. (1980) 'Bayesian Conditionalisation and the Principle of Minimum Information', *British Journal for the Philosophy of Science* 31, pp. 131–44.

WILLIAMSON, J. (1999) 'Countable Additivity and Subjective Probability', *British Journal for the Philosophy of Science* 51(3), pp. 401–16.

WILLIAMSON, J. (2003) "Bayesianism and Language Change", *Journal of Logic, Language and Information* 12(1), pp. 53–97.

WORRALL, J. (1989) 'Structural Realism: The Best of Both Worlds?', *Dialectica* 43, pp. 99–124.

ZABELL, S. (1982) 'W. E. Johnson's "Sufficientness" Postulate', *Annals of Statistics* 10(4), pp. 1091–99.

ZABELL, S. (1989) 'The Rule of Succession', *Erkenntnis* 31, pp. 283–321.

ZABELL, S. (2002) 'It All Adds Up: the Dynamic Coherence of Radical Probabilism', *Philosophy of Science* 69, pp. S98–S103.