



UNIVERSITY OF EXETER
DEPARTMENT OF MATHEMATICS

Score Decompositions in Forecast Verification

Keith Mitchell

July 2019

Supervised by

Dr Christopher A.T. Ferro
Prof David B. Stephenson

Submitted by Keith Mitchell, to the University of Exeter in July 2019, as a thesis for the degree of
Doctor of Philosophy in Mathematics.

This thesis is available for Library use on the understanding that it is copyright material and that no quotation from the thesis may be published without proper acknowledgement.

I certify that all material in this thesis which is not my own work has been properly identified and acknowledged and that no material herein has previously been submitted and approved for the award of a degree by this or any other University.

Signed:

Abstract. A forecast for an event should be based on a probability distribution of the possible outcomes of the event. In assessing the forecast, the forecast is given a penalty according to the outcome that occurs. The penalty, or score, is determined by a scoring rule and proper scoring rules are preferred; under a proper scoring rule, there is no incentive for a forecaster to issue a forecast that differs from the forecast they believe is appropriate. For proper scoring rules, the accuracy of a forecaster is defined to be their expected score over all possible forecasts and outcomes. Other measures of forecaster performance can be obtained by expressing accuracy as a sum of several terms, a process known as decomposing the accuracy. Each term of a decomposition measures a quality of the issued forecasts; qualities considered important are those that represent features of the joint distribution of the forecasts and outcomes. In the main decomposition, which we call the URR decomposition, the terms are uncertainty, resolution and reliability. The form of the URR decomposition is known for scoring rules of discrete events and their precise-probabilistic forecasts, which are forecasts issued as precisely-known probability distributions. We extend this URR decomposition to events with outcomes in any space and forecasts that may be functions of a probability distribution. We also determine in general form a second decomposition of accuracy, the RDC decomposition, in which the terms refer to the qualities of refinement, discrimination and correctness of the forecasts; the RDC decomposition has previously only been calculated in the specific instance of a binary event under the Brier scoring rule (Brier, 1950). The URR and RDC decompositions must be modified if the issued forecasts or recorded outcomes are separated into groups or bins before being assessed, and we give these amended decompositions. In a different setting, the URR and RDC decompositions we derive can also be used to examine the properties of interval-probabilistic forecasts. Interval-probabilistic forecasts are specific to binary events and issue a range of probabilities that the event will occur. There is little previous work on interval-probabilistic forecasts and to apply the URR and RDC decompositions we first define and characterise proper scoring rules for interval-probabilistic forecasts, known as interval-proper scoring rules, before establishing the URR and RDC decompositions of a particular interval-proper scoring rule, the interval-Brier scoring rule.

Acknowledgements

I'd like to thank Dr C. Ferro, for taking on an unknown applicant as a part-time student and for his assiduity.

To my fellow PhD students, postdocs and faculty of the Department of Mathematics, thank you for your warm welcome.

The need to devote time to this thesis has meant changes, delays, interruptions and postponements for my family, to which they never did object. Without their support this work would not have been possible. Thank you seems inadequate.

Contents

- Acknowledgements ii
- List of Figures vi
- List of Tables vii
- Publications viii
- 1 Introduction 1**
 - 1.1 Some Background 1
 - 1.2 Plan of Thesis 14
- 2 Decompositions and Dual Decompositions for Proper Scoring Rules 16**
 - 2.1 Introduction 16
 - 2.2 Notation 20
 - 2.3 Limitations of the Decompositions 21
 - 2.4 Towards General Decompositions 25
 - 2.4.1 URR Decomposition 25
 - 2.4.2 RDC Decomposition 27
 - 2.4.3 Problems and Aims 28
 - 2.5 Most General Decompositions 29
 - 2.5.1 URR Decomposition 30
 - 2.5.2 RDC Decomposition 34
 - 2.6 Examples 39
 - 2.6.1 Brier Scoring Rule 40
 - 2.6.1.1 URR Decomposition 40
 - 2.6.1.2 RDC Decomposition 41
 - 2.6.2 Ranked Probability Scoring Rule 41
 - 2.6.2.1 URR Decomposition 43
 - 2.6.2.2 RDC Decomposition 44
 - 2.6.2.3 Other Decompositions 45
 - 2.6.3 Ignorance Scoring Rule 49
 - 2.6.3.1 URR Decomposition 50
 - 2.6.3.2 RDC Decomposition 51
 - 2.6.4 α -Quantile Scoring Rule 53
 - 2.6.4.1 URR Decomposition 53
 - 2.6.4.2 RDC Decomposition 54
 - 2.7 Computing the Decompositions 57

2.8	Decompositions Under Binning	62
2.8.1	URR Decomposition	63
2.8.1.1	Example: Brier Scoring Rule	65
2.8.2	RDC Decomposition	67
2.8.3	Classification of the New Terms	69
2.8.4	An Illustration	70
2.9	Discussion and Conclusion	79
3	Proper Scoring Rules for Interval-Probabilistic Forecasts	82
3.1	Introduction	82
3.2	Problems Obtaining Interval-Propriety	85
3.3	Characterisation of Interval-Proper Scoring Rules	92
3.3.1	Choosing f and g	94
3.4	Examples	103
3.4.1	Interval-Brier scoring rule	103
3.4.2	Interval-Ignorance scoring rule	104
3.4.3	Pseudo-spherical scoring rule	105
3.5	Interval-Proper Scoring Rules in Practice	106
3.6	An Application	108
3.6.1	The Data	109
3.6.1.1	UK Met Office (UKMO) Data	109
3.6.1.2	Australian Bureau of Meteorology (ABOM) Data	109
3.6.2	Simulating Interval-Probabilistic Forecasts	110
3.6.3	Calculating Forecaster Skill	110
3.6.4	Example: Interval-Brier scoring rule	111
3.7	Discussion and Conclusion	120
4	Decompositions and Dual Decompositions for Interval-Proper Scoring Rules	123
4.1	Introduction	123
4.2	Interval-Proper Scoring Rules	124
4.2.1	Interval-Brier Scoring Rule	126
4.3	URR Decomposition	127
4.4	RDC Decomposition	129
4.5	Computational Formulae for the Attributes	133
4.5.1	Illustration: Simulated Data	137
4.5.2	Application: Precipitation Data	139
4.6	Discussion and Conclusion	151

5 Possible Directions for Future Work 153
5.1 Introduction 153
5.2 Applying the Decompositions 153
5.3 Interval-Probabilistic Forecasts 154
References 158

List of Figures

2.1	Illustration of the ‘ideal’ forecast distribution	37
2.2	Histograms of Attributes of the CRPS	61
2.3	Histograms of Attributes Under Binning of the CRPS	78
2.4	Binned and Unbinned URR Decompositions	79
2.5	Binned and Unbinned RDC Decompositions	80
3.1	Propriety of the Brier λ -scoring rule for an equally-spaced partition.	90
3.2	Total impropriety of the Brier λ -scoring rule	91
3.3	Comparison of Interval-Proper and Interval-Improper Scoring Rules	114
3.4	Hedging Propensity (UKMO Data)	115
3.5	Hedging Propensity (ABOM Data)	116
3.6	Hedging Propensity (UKMO Data) Under a Different Partition	117
3.7	Hedging Propensity: Comparing Sites Across Lead-times	118
3.8	Hedging Propensity: Comparing Lead-times Across Sites	119
4.1	Distributions of Probabilistic Forecasts Under Simulated Model	138
4.2	Estimates of URR and RDC Decomposition Attributes (Model)	140
4.3	Trend in Bias of Estimates of URR and RDC Decomposition	141
4.4	Estimates of URR Attributes (UKMO Data, Precise-Probabilistic Forecasts)	143
4.5	Estimates of URR Attributes (ABOM Data, Precise-Probabilistic Forecasts)	144
4.6	Estimates of RDC Attributes (UKMO Data, Precise-Probabilistic Forecasts)	145
4.7	Estimates of RDC Attributes (ABOM Data, Precise-Probabilistic Forecasts)	146
4.8	Estimates of URR Attributes (UKMO Data, Interval-Probabilistic Forecasts)	147
4.9	Estimates of URR Attributes (ABOM Data, Interval-Probabilistic Forecasts)	148
4.10	Estimates of RDC Attributes (UKMO Data, Interval-Probabilistic Forecasts)	149
4.11	Estimates of RDC Attributes (ABOM Data, Interval-Probabilistic Forecasts)	150

List of Tables

3.1 Optimal Partition: Interval-Brier Scoring Rule 122
4.1 Estimators for URR and RDC Attributes 135
4.2 Partitions for Precipitation Data 142

Publications

Material from this thesis has appeared in the following publication:

K. Mitchell and C.A.T. Ferro. Proper scoring rules for interval probabilistic forecasts. *Quarterly Journal of the Royal Meteorological Society*, 143(704):1597–1607, 2017.

1

Introduction

“And he said also to the people, When ye see a cloud rise out of the west, straightway ye say, There cometh a shower; and so it is. And when *ye see* the south wind blow, ye say, There will be heat; and it cometh to pass.” (Luke 12:54-55 [Carroll and Prickett, 1997](#))

“Consider the medical practice of John Mirfield, a priest and adviser at St Bartholomew’s Hospital, London, at the end of the [14th] century. He advises his fellow physicians that, if they wish to know whether a patient might survive or not, they should follow this procedure: Take the name of the patient, the name of the messenger sent to summon you, and the name of the day upon which the messenger first came to you; join all their letters together, and if an even number result, the patient will not escape; if the number be odd, he will recover. Such numerological practices are not peculiar to Mirfield: variant methods include the Sphere of Apuleis, wherein one assigns a numerical value to each letter of a name and subtracts thirty from the total to determine whether the patient will live or die. Considering that the standardised spelling of names has yet to be introduced, this is extraordinary: a practice of introducing sufficient random variables so that the whole matter is one of chance, or, as Mirfield would rather have it, divine providence. Other diagnostic advice of Mirfield’s includes this: take the herb cinquefoil and, while collecting it, say a paternoster on behalf of the patient. Then boil it in a new jar with some of the water which the patient is destined to drink; if the water be red in colour after this boiling, then the patient will die.” [Mortimer \(2009, pages 191-192\)](#)

“All men at all times have been wanting to know the future.” [Schumacher \(1974, page 186\)](#)

[1.1] Some Background

A *forecast* is a statement about an event, the outcome of which is, at the time of the forecast, unknown, but which *can* be known at some future time. A forecast can, therefore, be *verified*, or checked against the actual outcome when it occurs. If the forecast statement makes mention in some way of the likelihood or chance of an outcome then the forecast is referred to

as a probabilistic forecast; for example, ‘there is a 20% chance of rain today’, ‘there is a 40% likelihood of a recession next year’, or ‘in 9 out of 10 cases, patients admitted to this ward will recover in a week’. If we allow probabilities of 0 or 1, then categorical forecasts can be regarded as a special case of probabilistic forecasting, making the verification of probabilistic forecasts a broad subject.

Indeed, probabilistic forecasts now appear in many different areas of the extended sciences. [Gneiting and Katzfuss \(2014\)](#) provide examples from hydrology, geology, medicine, epidemiology, economics and politics that refer to probabilistic forecasting. Yet, it has been in meteorology that much of the study of probabilistic forecasting has been made. Some of the reasons why meteorology has been such fertile soil for the growth of probabilistic forecasting theory and practise are “experience in making forecasts for many variables and locations and in quantifying the uncertainty in forecasts, and the timely availability of formal and/or informal outcome (and other) feedback. Moreover, meteorological outcomes (i.e. future weather events) are not influenced by the forecasts. Although some of these characteristics are shared by forecasting in other fields, few if any fields share all of the characteristics.” ([Murphy and Winkler, 1992](#)).

Even early weather forecasters were considered in their pronouncements. As [Sheynin \(1984\)](#) describes, Dalton in his *Meteorological Essays* (1793), writes that the “probability of a fair day ... to that of a wet one, ... is as ten to one”, while at the turn of the 19th century, Lamarck issued forecasts in which “he indicated the probable states of the weather” for France, although Lamarck seems to have reported probabilities in qualitative terms (using, at least on one occasion “a scale of ... probabilities *trés-grandes, grandes, moyennes, and indéterminables*”). Such weather forecasts, in which possible future weather conditions were accompanied by qualitative probabilities or odds-based numerical assessments appeared throughout the 19th century, with forecasters “[feeling] a need to characterise the uncertainty in their forecasts explicitly” ([Murphy, 1998](#)).

Issuing different possible events together with their probabilities was not necessarily a defensive act; indeed, [Murphy \(1998\)](#) quotes a meteorologist of the time, W.S. Nichols’s view, that “knowledge of the degree of certainty with which an event may be expected, increases the value of the information”, a feeling seconded by [Cooke \(1906a\)](#) who advocated “[i]t is more scientific and honest to be allowed occasionally to say ‘I feel very doubtful about the weather for to-morrow, but to the best of my belief it will be so-and-so’”, although [Cooke \(1906b\)](#) mentions that the additional advantage of probabilistic forecasts over exact forecasts is that probabilistic forecasts “eliminate beforehand the adverse opinion which a great number of incorrect [exact] forecasts must produce”. Probabilistic forecasts empower not only the forecaster, for “weather forecasts expressed in probability terms ... enable users of

the forecasts to make the best possible decisions” (Murphy, 1998).

A pivotal change occurred with the introduction, by Hallenbeck (1920), of quantitative probabilities to express the uncertainty of each possible outcome of an event. The immediate question is how the probabilities should be determined. In Hallenbeck (1920) “maps . . . were designed to show the frequency, in percentages, of [the event, but] . . . these maps are used only as a basis at most, and the probability of [the event] . . . is not often accepted at its face value. General and local conditions usually are such as to justify modifying the indicated probability, sometimes materially”.

In broad terms, probabilities “can be accorded at least two different interpretations. Firstly, the probability can be interpreted as a (limiting) relative frequency. . . . Alternatively, the probability(ies) . . . can be interpreted as the forecaster’s degree of belief” (Murphy, 1998). Yet, there are many subtleties (see, for example de Eliá and Laprise, 2005). Sanders (1963) proposed the “healthier [interpretation] that a [probabilistic] forecast is a fallible judgment which can use all the objectively processed help it can get. The objective technique provides a probability reference point which the forecaster ‘sharpens’ by critical appraisal with the use of additional information . . . the subjectively modified objective [probabilistic forecast] is the best product.” More conclusively, to Epstein (1966), “the conceptual basis . . . is the ‘subjectivist’ view of probability . . . The term ‘subjectivist’ as used here should not be confused with the term ‘subjective’ as used in contradistinction to ‘objective’ . . . the latter term refers to the manner in which the probabilities are *derived*; the former to the manner in which they are *used* or *interpreted*. From a subjectivist point of view, a probability must be *believed* to be meaningful. In general, I am far more likely to believe the probability suggested by a well-tested objective system than one based on an ill-defined subjective evaluation. On this basis my own personal probability . . . will in general agree more readily with objectively rather than subjectively, derived forecasts. In the same vein, the [forecaster’s probability] should be his personal probability whether or not it is wholly subjectively or objectively derived. Indeed, this is a good argument for the normal procedure of providing the forecaster with objective probabilities as guidance material, and allowing him to issue modified statements according to his further subjective judgment and belief.”

While the innovation of Hallenbeck (1920) was to produce a directional shift in the discipline of probabilistic forecasting, probabilistic forecasting was not overwhelmingly accepted. For example, Dexter (1962) thought that one should “not have probabilities in the forecast” and especially “not even try to pin [the forecaster] down to a strict expression of probabilities”. The dissent was sufficiently widespread for Brown (1970) to note, “[the] advantages seem so overwhelming that it is somewhat surprising how few forecasts are actually put in terms of explicit probabilities. There must, therefore, be factors which militate against putting

forecasts in such terms.” Brown goes on to suggest that possible reasons for this reticence might include “(a) Forecasters may be ... ignorant of the language of probabilities ... (b) Desire for vagueness may afflict some forecasters who ... [would like to] be ‘proved correct’ independently of the actual course of events. Vagueness may also be a bureaucratic necessity at times.... (c) Users may pressure forecasters to come out flatly behind one alternative or another. ... (d) Epistemological difficulties arise when [assigning probabilities].... (e) Desire for credit, in a sense the opposite to the desire for vagueness, may influence forecasters to make specific rather than probability predictions.”. There was the additional reason, “with the evolving view of meteorology in the late nineteenth and early twentieth centuries as an exact physical science, key segments of the community adopted the view that the weather forecasting problem was best understood as a deterministic problem”, and “[r]elatively few if any developments related to probability forecasts can be identified during the period between the mid-1920s and the early 1940s” (Murphy, 1998). Whether the complexities of meteorological phenomena can be modelled or there is “uncertainty inherent in the forecasting process” (Murphy, 1998), the exact forecasts offered by physical models are not currently perfect, and were not so in the 1940s, prompting (Brier, 1944) to argue that “1. The scientific and economic (or, military) value of weather forecasts can be enhanced by increased use of probability statements, and 2. The verification problem (both scientific and economic) can be simplified if forecasts are stated in terms of probabilities.”.

Although the debate about whether (and how) to issue probabilistic forecasts was to continue (and remains a point of discussion), Brier (1950) was to ask the crucial question: if a probabilistic forecast *is* published, how should such a forecast be assessed? Whereas a categorical forecast has the advantage that once an outcome for the event being forecast is observed, the forecast can be unequivocally said to have been correct or incorrect, the verification of a probabilistic forecast is less clear: if a binary event is forecast to have a positive outcome with probability $p \in (0, 1)$ and then the outcome observed is negative, was the forecast right or wrong? Brier (1950) was the first to propose a “verification score”, a formula or rule, for computing how close a probabilistic forecast is to the outcome that occurs; shortly thereafter, Good (1952) proposed an alternative scoring rule. Given the scoring rule, the accuracy of the forecaster is their expected score: the average of their scores.

A scoring rule allows one to assess a forecaster’s skill. “Skill is the forecaster’s ability to analyze and classify the antecedent weather so that, ... the [forecast] probability of a subsequent event is increased above, or decreased below, the relative frequency of that event in all weather situations (hereafter referred to as the ‘climatic frequency’).” (Gringorten, 1951). And, “[s]kill must be measured in relation to something ... Then the forecaster’s skill lies in his ability to recognize factors in his array of synoptic information which, to him, make the likelihood of occurrence of meteorological events in a particular instance different from

the climatological likelihood.” (Sanders, 1963). Sanders (1963) defined skill in an absolute sense: as the difference between the expected score of the forecaster and the expected score if every forecast followed the climatology. Such absolute skill was to be distinguished from relative skill, in which the absolute skill was divided by the expected score under climatology. This distinction was first made by Murphy (1974): “[t]wo basic types of skill scores can be identified: 1) difference skill scores and 2) ratio skill scores. A difference skill score simply represents the difference between a measure of the ‘accuracy’ ... of the relevant climatological probabilities ... and the ‘accuracy’ of the forecasts according to the same measure. Difference skill scores, then, measure the amount by which the ‘accuracy’ of the forecasts improves upon the ‘accuracy’ of the climatological probabilities... A ratio skill score, on the other hand, consists of a ratio in which a difference skill score appears in the numerator and one of the terms in this difference, in general the measure for [climatology], appears in the denominator. Ratio skill scores, then, measure the *fractional* amount by which the ‘accuracy’ ... of the forecasts according to a particular measure improves upon the ‘accuracy’ of the climatological probabilities ... according to the same measure.”

Before forecaster skill can be calculated, a scoring rule must be chosen. The forecaster should be told the scoring rule (Gneiting, 2011a), however, once aware of the scoring rule, the natural response of the forecaster is to attempt to maximise their performance as measured by their score. Brier (1950) referred to this distortion of behaviour as ‘hedging’ or ‘playing the system’ and advised that “one essential criterion for satisfactory verification is that the verification scheme [i.e. scoring rule] should influence the forecaster in no undesirable way”. Epstein (1966) was more forthright, arguing that a forecaster “should *believe* his statement ... [otherwise] the forecaster is not being honest ... [and] this is not proper behavior for a forecaster. We must and shall expect our forecasters to be honest, both to themselves and to the public.” This sentiment was repeated in Murphy and Epstein (1967a) who advocated that the scoring rule “possesses the property that the best ‘hedge’ is no ‘hedge’ i.e. which possesses the property that the [forecaster] can expect to obtain the best score if and only if the forecasts he prepares express his true beliefs. In other words, ... make his behavior correspond with the proper ... behavior.” Murphy and Epstein (1967a) define such scoring rules as *proper scoring rules*. Brown (1970), using the same terminology as Shuford et al. (1966), called such scoring rules “which are free of distorting incentives ... ‘reproducing scoring systems’”, because such scoring rules report or reproduce the true beliefs of the forecaster. Proper scoring rules are occasionally also referred to as admissible scoring rules (Brown, 1974; Shuford et al., 1966). Brown (1970) provides some interesting examples of scoring rules that appear reasonable but are, in fact, improper.

Of course, scoring rules that promoted the publication of a forecaster’s actual view have what Winkler and Murphy (1968) refer to as normative goodness: the scoring rule encour-

ages the forecaster to publish what they *ought* to forecast (this includes the requirements that the probabilistic forecasts are coherent i.e. satisfy the conditions for being a probability). But, normative goodness is not the only criterion that a scoring rule must satisfy. As [Winkler and Murphy \(1968\)](#) note, a scoring rule must also faithfully capture substantive goodness (that is, give greater value to forecasts that demonstrate scientific knowledge and less value to forecasts that are whimsical or arbitrary).

Even before the terminology of ‘proper scoring rule’ had been established, [McCarthy \(1956\)](#) had provided a mathematical definition of proper scoring rules, stating that a “payoff rule is said to ‘keep the forecaster honest’” if the expected score of a forecaster is optimised if and only if the forecaster’s issued probability matches the forecaster’s true belief in the chance of the outcome. The specific scoring rules of both [Good \(1952\)](#) and (as confirmed by [Murphy and Epstein \(1967a\)](#)) [Brier \(1950\)](#) satisfy this definition. [McCarthy \(1956\)](#) goes on to prove that any proper scoring rule must have a particular functional form.

Numerous refinements of McCarthy’s characterisation theorem for proper scoring rules have been given since. McCarthy’s statement was limited to the case in which the observation can take finitely many different values and the forecast probabilities are discrete; McCarthy did not offer a proof of his theorem. Apparently independently, [Shuford et al. \(1966\)](#) stated and proved the characterisation theorem for a binary observation. This result of [Shuford et al. \(1966\)](#) is referred to and proved again in [Brown \(1970\)](#) and an advanced, probability-theoretic generalisation of the result is given by [Schervish \(1989\)](#). [Brown \(1970\)](#) also reconsiders the original setting of McCarthy’s theorem, wherein the observation is a multi-category random variable, and provides the proof missing in McCarthy’s original work. Using a similar approach, [Savage \(1971\)](#) also states and proves, in McCarthy’s framework, the conditions a proper scoring rule must satisfy. [Hendrickson and Buehler \(1971\)](#) prove McCarthy’s theorem in the more general case in which the observation is a random variable with a probability density function. [Gneiting and Raftery \(2007\)](#) have proved a general version of McCarthy’s theorem in which the probability forecasts are restricted only to be probability measures, and more recently [Frongillo and Kash \(2014\)](#) have given a wholly general characterisation theorem.

Although the characterisation theorem narrowed the class of functions that could be considered when seeking a proper scoring rule, there remained many different functions that could be selected as proper scoring rules. This introduced the complication that different proper scoring rules could give different and inconsistent measures of accuracy ([Winkler and Murphy, 1968](#)). To address this problem, the characterisation theorem was supplemented by an axiomatic directive: desirable mathematical properties that a proper scoring rule should satisfy were listed as axioms and from these axioms the list of possible proper scoring rules was de-

rived; and, to certain axioms, there corresponds a unique proper scoring rule, so adherence to the axioms removes all choice from the selection of a scoring rule. Examples of particular meta-features of scoring rules are: (i) locality (the scoring rule depends only on the forecast for the outcome that does occur) (see, for example [Bernado, 1979](#); [Shuford et al., 1966](#)), (ii) sensitivity-to-distance (more favourable scores are achieved if higher probabilities are assigned to outcomes close to the outcome that does occur) ([Epstein, 1969](#); [Murphy, 1970](#); [Staël von Holstein, 1970](#)). The axioms associated with particular scoring rules are discussed in [Shuford et al. \(1966\)](#), [Selten \(1998\)](#) and more recently [Jose \(2009\)](#).

The (mathematical) definition of a proper scoring rule is, in fact, a consequence of the behavioural assumption that the score a forecaster receives provides the forecaster with a utility and that the forecaster “chooses his [forecasts] in such a way as to maximize his expected utility” [Winkler \(1969\)](#). It is further assumed that the utility function is a linear function of the score. As [Winkler \(1969\)](#) then notes, “[s]ince his utility function is linear, this is equivalent to maximising his expected score in the probability assessment task [provided the scoring rule awards a higher score for a better forecast]”.

The assumption that utility is a *linear* function of the score is contentious. Even in the very early developments of the theory of probabilistic forecast verification, mention was made of the need to consider nonlinear utilities ([Good, 1952](#)). If the score is interpreted as a non-monetary award, for example promotion, extended holiday or some other such award, then nonlinear utility has some credence. Less obvious is that an objection to a linear score can be made when the score is interpreted as a monetary reward. But, it is easy to appreciate that a certain sum of cash is of proportionally greater utility to an individual when they have no money, than it is to the same individual should they have much more money. As [Brown \(1970\)](#) reasons “there is the well-known nonlinear utility of money ... if you make awards too substantial you may find yourself in a situation where some of your respondents become overcautious, and prefer to go for small but relatively certain gains in preference to maximizing their expected gain through a riskier strategy.” And, so a “point of some interest is the possibility of relaxing the assumption that the assessor’s utility function is linear with respect to money. If this assumption is not met, then it might not be optimal for the assessor to honestly report his judgments.” ([Winkler, 1969](#)). “A nonlinear utility function implies that ‘risk’ considerations may be relevant ...” ([Murphy and Winkler, 1970](#)). The utility function may be nonlinear by being either convex, in which case the forecaster is a risk taker, or concave, whereupon the forecaster is a risk-avoider ([Murphy and Winkler, 1970](#)).

“What are the implications of nonlinear utility functions for the process of probability assessment ...? If the [forecaster] is able to specify his utility function, then this function can, and should, be incorporated into the assessment process.” ([Murphy and Winkler, 1970](#)). The

details of how to adapt the criterion of maximising the expected score when the utility function is nonlinear (but invertible) is given in detail in [Winkler \(1969\)](#). Assuming that the forecaster continues to maximise their expected utility, [Winkler and Murphy \(1970\)](#) examine the consequences of a nonlinear utility when the Brier score is adopted and find that “the utility function of the [forecaster] may cause him to ‘hedge’ in some manner. Generally, a risk-taker will ‘hedge’ toward a categorical forecast, and a risk-avoider will ‘hedge’ away from a categorical forecast (i.e. toward a forecast in which all the probabilities are equal).” But, as [Murphy and Winkler \(1970\)](#) continue, “(on) the other hand, if the [forecaster’s] utility function is not known, then the function cannot, of course, be incorporated into the assessment process. Therefore, the [forecaster’s] statements may differ from his judgments. For some utility functions the differences might be quite large . . .”. Nonetheless, the position has been, generally, to assume that the forecaster’s utility function is a linear function of their score, for, as [Savage \(1971\)](#) notes “such linearity assumptions . . . are presumably tolerable if only moderate sums of money are involved. In the purely mathematical formulation . . . no precautions have been, or need be, taken to keep these sums moderate, but it is fairly clear how such precautions could be taken in applying the theory. . .”. Recently, however, [Carvalho \(2015\)](#) has reopened the debate.

[Brier \(1950\)](#) had introduced scoring rules in response to perceived failures of earlier methods to verify probabilistic forecasts. Both [Cooke \(1906a\)](#) and [Hallenbeck \(1920\)](#) had measured the performance of their probabilistic forecasts, by comparing (in a table), each probabilistic forecast of a binary event with the frequency of occurrence: a low-probability forecast suggests occurrence is unlikely and so the frequency of occurrence should be small, while a high-probability forecast suggests occurrence is likely, and the number of occurrences should be high; numerically, the forecast should be close to the actual frequency of occurrence for the forecasts to be regarded as acceptable. But, such a comparison can be misleading, for “knowledge of a good relationship between forecast and observed probabilities is not sufficient to indicate how useful the forecasts are, for it is also necessary to know the frequency with which forecasts are made in the various categories.” ([Brier, 1950](#)). To assess the relationship between forecast and observations, it is the *joint* distribution of the forecasts and observations that is important, so not only is it necessary to have the conditional distribution of the observations given the forecasts, but the marginal distribution of the forecasts must also be known, to allow the joint distribution to be inferred. Herein, [Brier \(1950\)](#) anticipates both the distributions-oriented approach to forecast verification that would later be advocated by [Murphy and Winkler \(1987\)](#) and the forecast evaluation approach of [Gneiting et al. \(2007\)](#) (more of both approaches shall be mentioned later).

The scoring rule offers a clear and elegant method of determining accuracy, or average score. Since [Brier \(1950\)](#)’s scoring rule, many alternative proper scoring rules have been proposed.

In part, this proliferation of scoring rules has been because the assessment of probabilistic forecasts by their accuracy necessarily overlooks the evidently subtle ways in which probabilistic forecasts could be in error. “Beyond the question of how good such probabilistic judgments are in a general sense, the issue of how good they are in very specific ways is of interest, too” (Yates, 1982). In an attempt to capture different aspects of forecast performance, one response is to use different scoring rules. Another, is to define the forecasts’ properties that are regarded as important and then try to measure them.

As far as can be ascertained, Bross (1953) was the first to consider different *qualities* of a “predicting system which leads to probabilities . . . a Probability Predicting System (or PPS for short)” specifically to “tell whether the scheme is doing a good job . . . by examining the *record* of the Predicting System. First, let us consider those events which were classed as ‘very likely’. Suppose that this classification was used in a hundred cases and in ninety of these cases, the event actually occurred. This would indicate that the classification ‘very likely’ was justified.” (Bross, 1953, page 39). “More generally, if we were to consider cases in which the PPS gave some other numerical value of the probability, which I will represent symbolically by p , then if the system is valid it should be true that

$$p = \frac{\text{Number of cases in which event occurred}}{\text{Total number of cases where the probability assigned was } p} \quad [\text{sic}].” \quad (1.1)$$

(Bross, 1953, page 47). The term *reliability* or *calibration* is now used in place of Bross’s validity. And, while validity is still regarded as a fundamental property of forecasts, Bross is quick to qualify that validity should not be the sole consideration, for, while “validity is a desirable property of a probability predicting system, the fact that a system has this property does not automatically insure that it will be a useful system in practice. A second important auxiliary requirement of a PPS is that it be ‘sharp.’” (Bross, 1953, page 48). Bross does not give a definition for sharpness, stating “[i]t is not very easy to give a clear-cut definition of sharpness Perhaps the closest that we can come to a mathematical formulation is one similar to the quantity *information* in . . . information theory.” (Bross, 1953, page 52). Nonetheless, Bross does give several examples of different levels of sharpness and these examples suggest that Bross regards sharpness as the extent to which forecasts are “discriminatory” (Bross, 1953, page 49) i.e. how distinct the forecasts are for different outcomes. (Later, Matheson and Winkler (1976) would refer to validity as the ‘honesty’ of the forecaster and sharpness as the ‘expertise’ of the forecaster.)

The quality defined by equation 1.1 is a distributional quality: it refers directly to the conditional distribution of the observations. Additional distributional features were introduced (see, for example Murphy and Epstein, 1967b) and methods were developed to enhance the assessment of these distributional features.

There were now two seemingly disparate methods of evaluating probabilistic forecasts: measuring accuracy using proper scoring rules, and measuring forecast qualities (specifically, validity and sharpness) through the features of the distributions of the forecasts and outcomes.

At the same time, there emerged the concern that “[o]ne of the factors that has contributed to the difficulties and controversies of forecast verification is the failure to distinguish carefully between the scientific and practical objectives of forecasting” (Brier, 1944). Assessing the proximity of forecasts to the true outcome “will not necessarily be appropriate” (Brier, 1944) as an evaluation for the users of forecasts who make decisions and take actions for which there are economic, social and environmental consequences. Each decision or action by a user will result in a benefit to the user if conditions match the forecast, but a reduction in this benefit should conditions differ from the forecast; the loss to the user can be mitigated for a cost (for example, insurance). An operationally optimal forecast is one which maximises the benefit to the user net of costs and losses. So, Gringorten (1951) argued that the usefulness “of a forecast should be judged by . . . the value of [the] forecast for given operational needs” later adding “that on one and the same day, two forecasters might make probability statements, each statement as valid as the other, and both forecasters equally skillful or performing with the same degree of accuracy or both. Yet one forecaster’s probability statements might serve the operations better” (Gringorten, 1958). This view was supported by Thompson and Brier (1955) who added that “it is not an uncommon experience to find that a group of forecasts which may show a high verification score need not necessarily be economically useful predictions . . .”.

That forecasts designed to promote scientific accuracy tend not to recognise the costs and losses associated with users’ actions based on these forecasts, is often a pragmatic choice rather than dismissiveness by the forecasters, because costs and losses are often highly individual and the forecast is aimed at the public as a whole. A perfect scientific forecast, one that it is both exact and correct (Winkler and Murphy, 1968), is also perfect from an operational perspective. But if a forecast is imperfect, the nature of the imperfections, that is the qualities of the forecast, will influence the operational value of the forecast and operational and scientific measures of value will diverge. In the main, operationally optimal and scientifically accurate forecasts are not regarded as coincident. Indeed, Murphy and Epstein (1967b) persuasively argued that inasmuch as forecast verification consisted of the evaluation of forecasts, there were two distinct evaluation tasks: “the process of assessing the [usefulness] of predictions . . . [or] operational evaluation . . . [for which] only a single attribute, or ‘desirable’ property, of the predictions is of concern, their [usefulness]. The ‘value’ of a prediction is measured in terms of the [use], to the decision maker, of the consequence that

results. However, since the measure of the attribute [usefulness] will be different for different decision makers and for different decision situations, *a universal measure for operational evaluation does not exist.*” For this reason attention is usually focused on the “process of assessing the absolute and/or relative perfection of predictions, [that is] empirical evaluation or verification” (Murphy and Epstein, 1967b) and assumes that “the forecaster’s duty ends with providing accurate and unbiased estimates of the probabilities” (Brier, 1944). More so, addressing the (scientific, or, empirical) verification of forecasts should expose the qualities of the forecasts, which will allow users of the forecasts to perform (to some degree) their own operational verification.

What is clear is that whether the concern is about agreement between the accuracy of a forecaster and the forecasts’ properties, or about the disagreement between operational and scientific value of the forecasts, establishing a measure of the qualities of the forecasts and finding their relationship to the accuracy of the forecaster issuing the forecasts was of fundamental importance if these concerns were to be meaningfully addressed.

The first attempt to determine a relationship between the properties of forecasts and the accuracy of a forecaster was by Sanders (1963), with the informal, but insightful, recommendations of Bross (1953) serving as a foundation. Sanders (1963) showed that the expected value of the proper scoring rule of Brier (1950) could be written as the sum of two terms; this separation of accuracy into parts is a process known as *decomposing* or *factorising* the expected score/accuracy. Sanders called the first term of his decomposition sharpness and the second term validity. Sanders’s choice of terminology was deliberate as he regarded the two components as direct quantitative representatives of Bross’s forecast properties of the same names, interpreting Bross’s sharpness as the ability to separate different outcomes into distinct groups (or, “sorting ability”) and validity as the ability to correctly assign probabilities to the different groups (or, “labeling ability”) (Sanders, 1963, page 197). Different forecasters will have different values of sharpness and validity and Sanders argued that “[t]he effectiveness of the forecasts in contrast to the climatological control can be assessed by calculating the gain due to sharpness and the penalty due to lack of validity.” These conclusions echo the remarks of Bross (1953, page 61) that when examining the “validity versus sharpness” of different forecasters “[t]he place to draw [the] line is where the advantage of increasing the sharpness is offset by the decreasing validity, but the actual determination of this point is no easy matter.” Later, Gneiting et al. (2007) were to reintroduce the “paradigm of maximizing the sharpness . . . subject to calibration”.

Epstein and Murphy (1965) give a geometric interpretation of Sanders’s decomposition and elaborate on Sanders’s description of the two components, adding “[v]alidity refers to the resemblance, on an independent collection of predictions, between the probabilities assigned

to, and the observed relative frequencies of, the events, while sharpness refers to the discrimination among events exhibited by the probabilities.” Changing the balance between validity and sharpness in Sanders’s factorisation for each fixed level of accuracy, generates a surface which was examined by [Blattenberger and Lad \(1985\)](#) (although they used the terms reliability for validity and refinement for sharpness, a convention used by several other authors). [DeGroot and Fienberg \(1983\)](#) generalised the factorisation of Sanders to a broader group of scoring rules for probabilistic forecasts of a *binary* observation.

[Murphy \(1973\)](#) advanced Sanders’s decomposition, by factorising Sanders’s sharpness term into two further terms: the first term, which is the variance of the observation, was called “the *uncertainty* inherent in the events”, the second term, was given the name resolution. For an event with only two possible outcomes, resolution is “the variability of the observed frequency of [the] event, when the forecast probability of [the] event varies” ([Atger, 2003](#)). In general, resolution is the extent to which the forecasts contribute to reducing uncertainty. [Murphy \(1973\)](#) referred to validity as reliability and so gave the first uncertainty-resolution-reliability (URR) decomposition.

[Yates \(1982\)](#) (see also [Yates and Curley, 1985](#)) made an additional extension to the decomposition of [Murphy \(1973\)](#), to arrive at a “covariance decomposition”. The uncertainty term remains and “reflect(s) aspects of forecasting performance *not* under the forecaster’s control. . . the remaining terms in the decomposition index aspects that *are* under the forecaster’s influence”. The resolution term of [Murphy \(1973\)](#) is also present. But Yates factorises Murphy’s reliability term into four terms, of which one is the negation of Murphy’s resolution term, and introduces three new terms. Of these new terms, the first is proportional to the covariance between the forecasts and the outcomes (a higher covariance increasing the accuracy). [Yates \(1982\)](#) refers to this term as the “heart of forecasting skill”, a notion to be repeated later by [Murphy and Winkler \(1987\)](#). The second of the three new terms of [Yates \(1982\)](#) is the variance of the forecasts. For Yates, “the aim of the forecaster should be to minimize the variance of his or her forecasts . . . [but] the only way [the variance of the forecasts] can take on its absolute minimum possible value of zero is when the forecaster offers constant forecasts. This strategy would make the covariance term zero, too. So the proper objective of the forecaster should be to minimize [the variance of the forecasts], *given* that he or she exercises his or her fundamental forecasting abilities as represented by [the covariance between the observation and the forecasts]”. Such a conditional minimisation results when, for each different outcome of the observation, all probabilistic forecasts preceding the outcome are the same. “If, under these conditions, [there are distinct forecasts preceding every pair of different outcomes], one has the situation in which the forecaster has perfect foresight, in that he or she exhibits perfect discrimination of instances in which [each different outcome occurs]. The only thing that would possibly mar the forecaster’s per-

formance is mislabeling; the forecaster’s numerical [probabilities are] inappropriate” (Yates, 1982). The third of the new terms of Yates (1982) is reliability-in-the-large (which follows Murphy and Epstein, 1967b), being the (squared) difference between the average forecast probability and the average outcome.

The importance of Yates (1982)’s decomposition was that it incorporated alternative qualities that might be used to assess forecasts, some of which had been suggested previously by Murphy and Epstein (1967b) as desirable. All previous decompositions had established quantities for components that represented qualities derived from philosophical premises about what a good forecast is. But, Yates (1982)’s decomposition created components that suggested qualities that the forecasts should have. Two similar decompositions to that of Yates (1982) had been proposed earlier by Theil (1966, pages 29,33).

Some of the attributes of the decompositions of Yates (1982) and Theil (1966), and their formal expressions, reappear in a second decomposition of the Brier score for probabilistic forecasts, introduced by Murphy (1986) and complementary to the decomposition of Murphy (1973). This alternative decomposition of Murphy (1986) presented the accuracy of the Brier scoring rule as the sum of two new terms, which we shall call (Murphy (1986) did not name the two terms) correctness and excess. Excess quantifies the ability of the forecaster to vary their forecasts sufficiently so that different forecasts are offered for different outcomes, *but* without too much variability of the forecasts so as to make the forecasts equivocal. Correctness considers how close the average forecast for an observation is to the subsequent outcome of the observation. Murphy and Winkler (1987) (see too, Murphy and Winkler, 1992) extended the decomposition of Murphy (1986), introducing two further attributes, refinement and discrimination, and expressed excess as refinement less discrimination, giving, under the Brier scoring rule, the first presentation of the refinement-discrimination-correctness (RDC) decomposition; the RDC decomposition of Murphy and Winkler (1987) reinforced the qualities promoted by Yates’s decomposition demonstrating even more clearly that optimality is obtained when the probabilistic forecast preceding each outcome is a constant, where this constant preferably matches the outcome (recalling that the above decompositions all refer to the scoring rule of Brier (1950) for which the observation has a value of 0 or 1).

The key decompositions of Sanders (1963), Murphy (1973), and Yates (1982) are all specific to the Brier scoring rule (Brier, 1950) (or, scoring rules derived therefrom, known as quadratic scoring rules). But, as Winkler (1996) notes “[i]nvestigations of decompositions have concentrated on quadratic rules, perhaps because they are the most commonly-used scoring rules in practice and because they are easy to decompose. An exploration of decompositions of other rules would increase our storehouse of measures of various attributes of ‘goodness’ of probabilities.” Winkler (1996) did make partial steps in this direction (gen-

eralising [Matheson and Winkler \(1976\)](#)) and, previously, [DeGroot and Fienberg \(1983\)](#) had established a general sharpness-reliability decomposition of any proper scoring rule for probability forecasts of a binary observation. But, little other progress was made until [Bröcker \(2009\)](#) established the general URR decomposition under any (proper) scoring rule for probabilistic forecasts of an observation with finitely many values (for an intuitive derivation of Bröcker’s result see [Siegert \(2017\)](#)). URR decompositions that had, for other scoring rules, previously required extensive derivations, for example, scoring rules related to the ignorance scoring rule of [Good \(1952\)](#) and [Roulston and Smith \(2002\)](#) and their decompositions given in [Weijs et al. \(2010\)](#) and [Tödter and Ahrens \(2012\)](#) (and even for quadratic scoring rules more general than the Brier scoring rule, (see, for example [Young, 2010](#))) could now be more routinely established. And, by rewriting scoring rules for point forecasts as scoring rules for probabilistic forecasts (see, [Gneiting, 2011a](#)), it was also now possible to determine URR decompositions for the accuracy of scoring rules for point forecasts; for example, the URR decomposition for the quantile scoring rule obtained in [Bentzien and Friederichs \(2014\)](#). However, Bröcker’s generalisation was restricted to the URR decomposition and no comparable generalisation for the RDC decomposition has, up to now, been published.

[1.2] Plan of Thesis

In the next chapter, chapter 2, we develop entirely general URR and RDC decompositions for proper scoring rules of forecasts that refer, either explicitly or implicitly, to the probability distribution of the observation. The URR decomposition extends the result of [Bröcker \(2009\)](#) for proper scoring rules of probabilistic forecasts. The RDC decomposition generalises the only known earlier work in this direction, that of [Murphy and Winkler \(1987\)](#) (see also [Murphy, 1986](#)). When computing the decompositions, especially from small data sets, it may be necessary to group forecasts and/or outcomes into subgroups and, in each subgroup, replace the value of each member with a representative value, thereby increasing the number of times each forecast and/or outcome value appears in the data. If data is subgrouped or binned in this manner, the decompositions need to be amended: the resolution and reliability terms of the URR decomposition and/or the discrimination and correctness terms of the RDC decomposition must be augmented to account for the binning. At the end of chapter 2 we derive general decompositions when data is binned.

Throughout chapter 2, probabilities associated with forecasts are considered exact (either numerically or in distributional terms). In chapters 3 and 4, we study the verification of forecasts where each forecast is issued as an interval of probabilities, or, interval-probabilistic forecasts; in these chapters the observation is binary. In chapter 3 we are concerned with the definition and characterisation of scoring rules for interval-probabilistic forecasts that are

proper; we call such scoring rules interval-proper scoring rules and give several examples. A particular example of an interval-proper scoring rule, the interval-Brier scoring rule, is the focus of chapter 4, in which we derive a URR decomposition and RDC decomposition for the interval-Brier scoring rule, and in so doing, introduce expressions for the attributes of interval-probabilistic forecasts.

The final chapter briefly discusses topics which lead on from the work of the earlier chapters.

In the thesis, later chapters do use results from earlier chapters and these dependencies are an important factor in determining the order in which the chapters have been presented, but each chapter presents a separate idea and is as self-contained as possible, save some references to the ideas and results of earlier chapters in an attempt to avoid repetition.

Decompositions and Dual Decompositions for Proper Scoring Rules

Summary. A scoring rule is a rule for measuring how close a forecast for an event is to the outcome for the event. Before the outcome is known, the expected value of the scoring rule, referred to as the expected score, can be used as a measure of forecast performance. It is now well-known that the expected score of a proper scoring rule can be decomposed into several terms, each term reflecting a particular performance property of the forecasts. In what is arguably the most well-known decomposition there are three terms: uncertainty (the difficulty of the forecasting environment), resolution (the extent to which the forecasts reduce the uncertainty of the event's outcome) and reliability (how close, on average, the outcomes following a repeated forecast are to the forecast). This primary decomposition has two limitations. *Firstly*, a general form of the decomposition, applicable to any proper scoring rule, is known only when the forecasts are probabilistic forecasts. For other types of forecast, the primary decomposition must be determined specifically for each scoring rule. We address this shortfall by giving a general primary decomposition for proper scoring rules for all types of forecast. *Secondly*, the primary decomposition does not account for all the important properties of the forecasts, in particular, the properties of discrimination (whether different outcomes are preceded by distinct forecasts) and correctness (how close the average forecast for an outcome is to the outcome). We derive a dual decomposition, the terms of which are precisely these properties; our dual decomposition is applicable to all proper scoring rules of all types of forecast. The terms of both primary and dual decompositions are readily evaluated, although too few data can make estimating the terms difficult. A common technique to counter sparse data, is binning. But, binning changes the forecast properties and must be accompanied by changes to the decompositions. We give general primary and dual decompositions when binning is used.

[2.1] Introduction

A forecast may be described as any statement that presupposes the unknown outcome of an event. The problem of how to assess whether forecasts are good, and how 'good' should be interpreted, has led to so abundant a literature that any comprehensive overview is all but impossible. Nonetheless, we have in chapter 1 attempted to outline the main historical developments.

By definition, the event that is the subject of the forecast will have an uncertain value and, therefore, the forecast should (implicitly or explicitly) refer to a probability distribution over the possible values of the event (Cooke, 1906a; Lindley, 1982). It is the case then that although, when the outcome for the event, when it occurs, it will be known precisely (e.g. 'rain' or 'no rain'), a forecast for the event can be less definite (e.g. 'chance of rain 20%'); we do not consider, here, the problem of error in observing the event's outcome (see Ferro,

2017). How then is it possible to evaluate how close the forecast is to the outcome?

One approach is to use a *scoring rule* (Winkler, 1996). A scoring rule is a function that assigns to each forecast-outcome pair, a value or score, reflecting how close the forecast and outcome are to each other. Before the outcome is known (in which case it may be called the observation), the expected score can be calculated for each forecast by averaging the scores the forecast obtains for each possible outcome (i.e. each value of the observation).

Some care needs to be taken when choosing a scoring rule. For, having been presented with a scoring rule, a forecaster may attempt to improve their expected score by issuing a forecast that differs from the forecast they believe to be best (a practice referred to as ‘hedging’ (Brier, 1950; Murphy and Epstein, 1967a)). A scoring rule that allows for such distorted behaviour is considered improper and should be discarded. Consequently, only proper scoring rules will be considered.

We define the accuracy of a forecaster to be the average, over all possible forecasts, of the forecasts’ expected scores. As such, accuracy is an example of a *verification measure*: a (often simple) numerical summary of the correspondence between the forecasts and observation. While verification measures have the attraction of being definitive and simple, Murphy and Winkler (1987) note that the ‘joint probability distribution [of forecast and observation] contains all information that is relevant to the evaluation of a ... forecast system’ and therefore, ‘this [joint] distribution – or, equivalently, the factors involved in the factorizations of the joint distribution – should play a central role in verification studies.’ This prompts Murphy and Winkler (1987) to ‘ask the following question: What, if any, are the relationships between [verification] measures and the general framework for verification [based on] basic indicators of performance such as the joint distribution’?

In short, the answer is the factorisation of verification measures into components that are ‘directly related either to the [joint, conditional and marginal] distributions themselves or to summary measures of these distributions’ (Murphy and Winkler, 1987). These factorisations provide, in part, a response to the call by Murphy and Winkler (1987) for a ‘“diagnostic” approach to verification ... [for although common] verification measures are quite useful ... to compare ... forecasters in some overall sense ... they are not particularly helpful when it comes to obtaining a more detailed understanding of the strengths and weaknesses in forecasts’.

In other words, with reference to the more prominent verification measure of accuracy, better accuracy does not, on its own, indicate a better forecaster (Murphy, 1993; Yates and Curley, 1985). Accuracy, as an average, subsumes many of the finer features of the scores that con-

tribute to it, and it is these finer features that inform about the qualities (Murphy, 1993) or attributes (Epstein and Murphy, 1965) of the forecasts.

Forecasts are important not as an end in themselves, but because they are used to make decisions. Each attribute will have differing relevance depending on the decision, and knowing the attribute profile of the forecasts determines the forecasts' suitability (or, 'goodness') for the decision at hand and helps to inform where forecasts can be improved. It is, therefore, the different attributes that determine how 'good' the forecasts are.

Historically (see chapter 1 and, for example, Murphy and Winkler (1987)), the attributes of most interest, are: *uncertainty* (the variability of the observation), which indicates the difficulty of the forecasting environment and places the accuracy obtained in context; *resolution* (whether the forecasts decrease doubt about the outcome); *reliability* (whether, having issued a particular forecast several times, the outcomes that follow the forecast are, on average, close to the forecast); *refinement* (the variability of the forecasts), which is suggestive of the adaptability of the forecasts; *discrimination* (the extent to which different outcomes are preceded by distinct forecasts); *correctness* (after a specific outcome has occurred on a number of occasions, how close are the forecasts preceding it, on average, to the outcome). We attempt here to settle the meanings of these attributes, warning that in past works by many different authors some of these attributes have been given different names, or the same names have been used with different meaning.

The aim, therefore, is to factorise, or decompose, accuracy into a sum of several terms carefully so that each term represents a particular attribute. Because the joint distribution can be factorised into the product of a conditional and a marginal distribution in two different ways, there are two decompositions of accuracy.

Of the aforementioned two decompositions, the more familiar is the uncertainty-resolution-reliability (URR) decomposition, with component terms that are measures of the attributes of uncertainty, resolution and reliability of the forecasts. The URR decomposition relates to the conditional distribution of the observation given the forecast. To date, the most general form of the URR decomposition has been derived by Bröcker (2009) (see also Siegert, 2017). However, this version of the URR decomposition is limited to proper scoring rules for a particular type of forecast: (precise-)probabilistic forecasts i.e. forecasts that are issued as statements about the exact probability distribution of the event's outcomes. URR decompositions for other types of forecasts have been determined, but only for specific scoring rules (see, for example Bentzien and Friederichs, 2014; Christensen, 2015). In this chapter, we develop a URR decomposition that is entirely general, being applicable to proper scoring rules of all types of forecast.

A second decomposition of accuracy, the refinement-discrimination-correctness (RDC) decomposition, has component terms that are measures of the refinement, discrimination and correctness of the forecasts. The RDC decomposition is dual to the URR decomposition in that it relates to the conditional distribution of the forecast given the observation. The RDC decomposition has only been determined for the Brier scoring rule (Brier, 1950) by Murphy (1986) and Murphy and Winkler (1987). We construct a general RDC decomposition applicable to proper scoring rules of all forecast types.

Although the URR decomposition has been pre-eminent in the literature, the RDC decomposition is just as important: the two ‘factorizations [of the joint distribution] constitute *complementary* rather than alternative ways to approach the verification problem ... the two factorizations are concerned with different attributes of the forecasts and/or observations. Thus, a complete verification study would necessarily involve the evaluation of factors associated with both factorizations.’ Murphy and Winkler (1987). Nonetheless, because the URR decomposition is both the earlier and more widely studied decomposition, we may refer to it as the primary decomposition, with the RDC decomposition being called the dual (or, occasionally, secondary) decomposition. However, we emphasise that these labels mark the relative familiarity of the decompositions rather than their relative importance.

The outline for the remainder of this chapter is as follows. In the next section, we establish notation. Following immediately thereafter, in section 2.3, we clarify the role of decompositions and their limitations. Having established the importance of decompositions we consider, in section 2.4 one method of generalising the URR and RDC decompositions and highlight the problems that restrict this approach from being completely general. The approach in section 2.4 is not, however, fruitless, and gives a good deal of direction to our derivation, in section 2.5, of the entirely general URR and RDC decompositions. Section 2.6 presents examples of both decompositions for different scoring rules: specifically, we consider probabilistic forecasts (see, for example Dawid, 1986, for details) and point forecasts (where one or more summary measures of a probability distribution, for example the mean, median or quantiles are issued (see, for example Gneiting, 2011a,b)). The computation of the component terms of the decompositions is discussed in section 2.7. In calculating the attributes, a problem can arise when no, or few, forecasts or outcomes in the sample data are the same; one mitigating solution is to classify the data into groups and in each group replace all data with the same value, an approach known as *binning*. If binning is performed, adjustments to the decompositions are required, adjustments which we make in section 2.8. We conclude briefly in section 2.9.

[2.2] Notation

Let X be the unknown value of an event. Often, X is called the *observation*, and the value taken by X , x , the *outcome* (of the event) or *verification*; the set of all possible outcomes is \mathcal{X} . Denote by \mathcal{P} the set of all probability distributions for X ; we let P be the random quantity whose value is a probability distribution for X i.e. an element of \mathcal{P} . The elements of \mathcal{P} will be denoted by lower case letters such as p , q , and r . To each element of \mathcal{P} there corresponds a distribution function; we denote the distribution functions by upper case letters such as F , G , and H .

An important probability distribution for X will be the point-mass distribution at x , δ_x , for which, the corresponding distribution function is H_x (the Heaviside function at x (Riley et al., 1998)) defined by,

$$H_x(x') = \begin{cases} 0 & \text{if } x' < x \\ 1 & \text{if } x \leq x'. \end{cases} \quad (2.1)$$

A *probabilistic forecast* is a choice of $p \in \mathcal{P}$. To evaluate the match between a probabilistic forecast and observation, we can use a *scoring rule*, which is a function that assigns to each probabilistic forecast, p , and outcome, x , a real-number, $S(p, x)$, known as a *score*. We take S to be negatively-oriented so that a lower value for $S(p, x)$ indicates a *better* match between p and x (Winkler and Murphy, 1968). The values of a scoring rule have, primarily, an ordinal scale; negative scores are acceptable and, in general, a zero-score is not a perfect score.

Before the true outcome of the observation is known, we can, for any forecast, p , calculate its expected score, being the average of the scores $S(p, x)$ over the possible outcomes of the observation (i.e. values of x). Then, the expected score of forecast p when X has probability distribution q and distribution function G is written

$$S(p, q) \stackrel{\text{def}}{=} \mathbb{E}[S(p, X)] = \int_{\mathcal{X}} S(p, x) dG(x). \quad (2.2)$$

We adopt the convention of presenting all expectations as integrals even when \mathcal{X} is a discrete set.

A rational forecaster, for whom the correct probability distribution for X is q , will issue the forecast p_{\min} that minimises $S(p, q)$. If $p_{\min} \neq q$, the forecaster has issued a forecast different to the forecast they consider correct, q , the reason being that they will achieve a better expected score, *under* S , by doing so, a practice known as hedging (Brier, 1950; Murphy and Epstein, 1967a). Any scoring rule under which hedging can occur, is considered inappropriate or improper (Murphy and Epstein, 1967a). To avoid the perversities of improper scoring rules, we consider only *proper* scoring rules. The formal definition of a proper

scoring rule is (see [Bröcker, 2009](#); [Gneiting and Raftery, 2007](#); [McCarthy, 1956](#); [Winkler, 1996](#); [Winkler and Murphy, 1968](#), and others): a (negatively-oriented) scoring rule is proper if and only if

$$S(q, q) \leq S(p, q) \quad \forall p, q \in \mathcal{P}. \quad (2.3)$$

For proper scoring rules, we define the accuracy of the forecaster (or, their forecasts) to be the expected score over all forecasts and outcomes, $\mathbb{E}[S(P, X)]$.

[2.3] Limitations of the Decompositions

As has been discussed in the introduction to this chapter, under the general framework for verification advocated by [Murphy and Winkler \(1987\)](#), the fundamental task is to assess the properties of the joint, conditional and marginal distributions of the forecasts and outcomes. In line with this purpose, there are firm, probability-based, definitions of the attributes of probabilistic forecasts. For clarity we consider the case when X is a discrete random variable. Then, following, for example, [Bröcker \(2012, 2015\)](#), forecasts are

(i) said to have no resolution, if

$$\Pr(X = x|P = p) = \Pr(X = x) \quad \text{for all } p \in \mathcal{P}, x \in \mathcal{X}, \quad (2.4)$$

(ii) perfectly reliable, if

$$\Pr(X = x|P = p) = \Pr_p(X = x) \quad \text{for all } p \in \mathcal{P}, x \in \mathcal{X}, \quad (2.5)$$

(iii) classified as having no discrimination, if

$$\Pr(P = p|X = x) = \Pr(P = p) \quad \text{for all } x \in \mathcal{X}, p \in \mathcal{P}. \quad (2.6)$$

These definitions allow one to establish certain formal theorems regarding the attributes. For example, as [Bröcker \(2015\)](#) shows, an application of Bayes' Rule proves that forecasts have no resolution if and only if forecasts have no discrimination. But, whether the equations (2.4) to (2.6) are satisfied can be difficult to assess.

In the absence of a direct, distributional assessment of forecast quality, the URR and RDC decompositions offer an indispensable method of *indirectly* evaluating the attributes of the forecasts. However, some caution must be exercised because the measures of the attributes given by the decompositions are not wholly commensurate with the exact, distributional

measures of equations (2.4) to (2.6).

To see that the decompositional measures are imperfect evaluations of the attributes let X be a binary observation, with values 0 and 1, and define the forecasts to be probabilities $\Pr_p(X = 1)$ for all $p \in \mathcal{P}$; for ease of exposition, and as no confusion should arise, we write p for $\Pr_p(X = 1)$. If we adopt the (half-)Brier scoring rule (Brier, 1950), $S(p, x) = (p - x)^2$, then as shown in Murphy and Winkler (1987), the URR decomposition of accuracy is

$$\begin{aligned} \mathbb{E}[S(P, X)] &= \mathbb{E}[(P - X)^2] \\ &= \underbrace{\mathbb{E}[(X - \mathbb{E}[X])^2]}_{\text{Uncertainty}} - \underbrace{\mathbb{E}[(\mathbb{E}[X|P] - \mathbb{E}[X])^2]}_{\text{Resolution}} + \underbrace{\mathbb{E}[(P - \mathbb{E}[X|P])^2]}_{\text{Reliability}}. \end{aligned} \quad (2.7)$$

Murphy and Winkler (1987) also derive a second, dual decomposition for S ,

$$\begin{aligned} \mathbb{E}[S(P, X)] &= \mathbb{E}[(P - X)^2] \\ &= \underbrace{\mathbb{E}[(P - \mathbb{E}[P])^2]}_{\text{Refinement}} - \underbrace{\mathbb{E}[(\mathbb{E}[P|X] - \mathbb{E}[P])^2]}_{\text{Discrimination}} + \underbrace{\mathbb{E}[(X - \mathbb{E}[P|X])^2]}_{\text{Correctness}}. \end{aligned} \quad (2.8)$$

Although Murphy and Winkler (1987) did not give names to the component terms of this dual decomposition, we have labelled them as indicated (terminology which we justify when considering the decomposition in more general terms) and refer to the decomposition as the RDC decomposition.

For X binary we may restate distributional condition (2.4) as

$$\text{no resolution if: } \mathbb{E}[X|P = p] = \mathbb{E}[X] \text{ for all } p \in \mathcal{P} \quad (2.9)$$

Similarly, distributional condition (2.5) becomes

$$\text{perfect reliability if: } \mathbb{E}[X|P = p] = p \text{ for all } p \in \mathcal{P} \quad (2.10)$$

and distributional condition (2.6) becomes

$$\text{no discrimination if: } \mathbb{E}[P|X = x] = \mathbb{E}[P] \text{ for all } x \in \mathcal{X}. \quad (2.11)$$

From equations (2.9) to (2.11), it is seen that for resolution, reliability and discrimination, their terms in the decompositions (2.7) and (2.8) measure the extent of the difference between the left- and right-hand sides of the equations (2.4) to (2.6).

In particular, from equations (2.9) to (2.11), we see immediately that if forecasts have no

resolution then the resolution term of the URR decomposition is $\mathbb{E}[(\mathbb{E}[X|P] - \mathbb{E}[X])^2] = 0$, forecasts that are perfectly reliable have a reliability term in the URR decomposition of $\mathbb{E}[(P - \mathbb{E}[X|P])^2] = 0$, and forecasts that are totally non-discriminating have a discrimination term in the RDC decomposition equal to $\mathbb{E}[(\mathbb{E}[P|X] - \mathbb{E}[P])^2] = 0$.

So, a *necessary* condition for perfect reliability is a zero reliability term in the URR decomposition. Similarly, a zero resolution term in the URR decomposition is a necessary condition for forecasts with no resolution, and a zero discrimination term in the RDC decomposition is a necessary condition for non-discriminatory forecasts. From the contrapositive we have that if the resolution term of the URR decomposition is non-zero, then the resolution of the forecasts must be non-zero too and similar statements may be made about reliability and discrimination. (If the decompositions are being evaluated from data, statistical tests are required to determine whether any term is different to zero.).

If X is binary, it is evident that from $\mathbb{E}[X|P = p] = \mathbb{E}[X]$ we have $\Pr(X = x|P = p) = \Pr(X = x)$ for $x = 0, 1$, so in this case, zero decompositional resolution implied no distributional resolution. However, in general, the terms of the decomposition are *not sufficient* quantities for the distributional attributes: if the resolution term of the URR decomposition *is* zero it does not imply that the resolution attribute is zero, and the same is true for reliability and discrimination. In the following simple example we illustrate that a zero decompositional discrimination does not imply a zero distributional discrimination even if X is binary.

Consider tossing a fair coin to give a value X equal to either 0 (for tails) or 1 (for heads). As the coin is fair $\Pr(X = 0) = \Pr(X = 1) = 1/2$. A forecaster, before the outcome of the coin toss, issues a probabilistic forecast, p for the probability that $X = 1$. The forecasts are issued so that if the outcome is $X = 0$, the preceding forecast is either $p = 1/4$ or $p = 3/4$ with equal likelihood: $\Pr(P = 1/4|X = 0) = \Pr(P = 3/4|X = 0) = 1/2$. If the outcome is $X = 1$, the preceding forecast is always $p = 1/2$: $\Pr(P = 1/2|X = 1) = 1$. From the definition of resolution, the forecaster has resolution, because by Bayes' Theorem,

$$\begin{aligned}\Pr(X = 0|P = 1/4) &= 1 \neq \Pr(X = 0) = \frac{1}{2} \\ \Pr(X = 0|P = 1/2) &= 0 \neq \Pr(X = 0) = \frac{1}{2} \\ \Pr(X = 0|P = 3/4) &= 1 \neq \Pr(X = 0) = \frac{1}{2}\end{aligned}\tag{2.12}$$

and,

$$\begin{aligned}
\Pr(X = 1|P = 1/4) &= 0 \neq \Pr(X = 1) = \frac{1}{2} \\
\Pr(X = 1|P = 1/2) &= 1 \neq \Pr(X = 1) = \frac{1}{2} \\
\Pr(X = 1|P = 3/4) &= 0 \neq \Pr(X = 1) = \frac{1}{2}.
\end{aligned} \tag{2.13}$$

If the (half-)Brier scoring rule (Brier, 1950) is chosen, then the resolution term of the URR decomposition is (from equation (2.7)),

$$\mathbb{E}[(\mathbb{E}[X|P] - \mathbb{E}[X])^2] = \frac{1}{4} \tag{2.14}$$

in agreement with the non-zero resolution of equations (2.12) and (2.13).

The forecaster also has a level of discrimination, because

$$\begin{aligned}
\Pr(P = 1/4|X = 0) &= \frac{1}{2} \neq \Pr(P = 1/4) = \frac{1}{4} \\
\Pr(P = 1/2|X = 0) &= 0 \neq \Pr(P = 1/2) = \frac{1}{2} \\
\Pr(P = 3/4|X = 0) &= \frac{1}{2} \neq \Pr(P = 3/4) = \frac{1}{4}
\end{aligned} \tag{2.15}$$

and,

$$\begin{aligned}
\Pr(P = 1/4|X = 1) &= 0 \neq \Pr(P = 1/4) = \frac{1}{4} \\
\Pr(P = 1/2|X = 1) &= 1 \neq \Pr(P = 1/2) = \frac{1}{2} \\
\Pr(P = 3/4|X = 1) &= 0 \neq \Pr(P = 3/4) = \frac{1}{4}.
\end{aligned} \tag{2.16}$$

Calculating the discrimination term in the RDC decomposition given by equation (2.8), we have

$$\mathbb{E}[(\mathbb{E}[P|X] - \mathbb{E}[P])^2] = 0 \tag{2.17}$$

which is contrary to the non-zero discrimination of the forecast shown by equations (2.15) and (2.16).

Although, as this small example demonstrates, the decompositional measures are not sufficient evaluations of the distributional attributes of a forecasting system, nonetheless they provide an assessment of the forecasts' qualities that is (i) contextual: through the decompo-

sitions, all qualities are expressed in terms of the chosen scoring rule used to assess accuracy, so that quality and accuracy are measured on equivalent bases, and (ii) relative: the decompositions emphasise the contribution of each quality to the accuracy of the forecasting system and allow these contributions to be contrasted.

Decompositions, therefore, have a part in all verification analyses and it is of interest for decompositions to be as widely available as possible.

[2.4] Towards General Decompositions

Less than perfect correspondence between forecast and observation can be the consequence of bias (failure of the forecast and observation to agree on average) and/or dissociation (wide variation in the *difference* between forecast and outcome across forecast-outcome pairs). However, variation of the forecasts in alignment with variation in the observation is essential for correspondence. We isolate these positive and negative contributions to the correspondence between forecast and observation by separating accuracy (i.e. overall correspondence) into components reflecting bias, aligned variation (or, co-variation) and dissociation.

2.4.1 || URR Decomposition

To begin, we simplify the discussion by supposing X takes only the values 0 and 1, and forecasts are probabilities $\Pr_p(X = 1)$ for all $p \in \mathcal{P}$; p is used instead of $\Pr_p(X = 1)$. Under the (half-)Brier scoring rule (Brier, 1950), $S(p, x) = (p - x)^2$, the URR decomposition given in equation (2.7) can be rewritten in terms of S as

$$\mathbb{E}[S(P, X)] = \underbrace{\mathbb{E}[S(\mathbb{E}[X], X)]}_{\text{Uncertainty}} - \underbrace{(\mathbb{E}[S(\mathbb{E}[X], X)] - \mathbb{E}[\mathbb{E}[S(\mathbb{E}[X|P], X)|P]])}_{\text{Resolution}} + \underbrace{(\mathbb{E}[S(P, X)] - \mathbb{E}[\mathbb{E}[S(\mathbb{E}[X|P], X)|P]])}_{\text{Reliability}}. \quad (2.18)$$

Equation (2.18) is helpful in two-ways. Firstly, it will suggest a general interpretation to the terms of the URR decomposition (see below) and, second, it will permit a ready extension of the URR decomposition to more general observations and other scoring rules (which we also discuss below).

Underlying the construction of the URR decomposition is the preliminary step of stratifying the outcomes by forecast. With this stratification of the outcomes, we have the summary

measures: (i) $\mathbb{E}[X]$, the overall mean of the outcomes (determined from the marginal distribution of the outcomes; the marginal distribution is also known as the *base-rate* or, in meteorology, the *climatology*), (ii) $\mathbb{E}[X|P]$, the within-stratum means of the outcomes (determined from the conditional distribution of the observation given the forecast). Both $\mathbb{E}[X]$ and $\mathbb{E}[X|P]$ are, in our simple binary setting, probabilities and are, therefore, themselves acceptable forecasts for X , and it is meaningful to write $S(\mathbb{E}[X], x)$ and $S(\mathbb{E}[X|P = p], x)$ for all p, x .

Having established $S(\mathbb{E}[X], x)$ and $S(\mathbb{E}[X|P = p], x)$ as being well-defined, we now take $\mathbb{E}[X]$ and $\mathbb{E}[X|P]$ as mean values rather than probabilities. The quantity $\mathbb{E}[S(\mathbb{E}[X], X)]$ is the overall correspondence of the observation with its mean value and, as such, is a measure of the total variation of the observation i.e. of *uncertainty*. Similarly, for each p , $\mathbb{E}[S(\mathbb{E}[X|P], X)|P = p]$ is the overall correspondence between the observation and its within-stratum mean value and is a measure of the within-stratum variation of the observation (for the stratum determined by $P = p$); because the forecast is fixed for each stratum, the variation of the observation within a stratum is equally variation in the difference between the forecast and observation and indicative of dissociation. We take, therefore, as a measure dissociation the average within-stratum variation of the observation across all strata, $\mathbb{E}[\mathbb{E}[S(\mathbb{E}[X|P], X)|P]]$. The difference between the total variation of the observation (uncertainty) and the within-stratum variation of the observation (dissociation) is the between-strata variation of the observation. Between strata, the forecast also changes so the between-strata variation of the observation reflects the alignment between forecast and observation (whether different forecasts are followed by different outcomes); the quantity $\mathbb{E}[S(\mathbb{E}[X], X)] - \mathbb{E}[\mathbb{E}[S(\mathbb{E}[X|P], X)|P]]$ is the *resolution* of the forecasts.

If one accounts for the negative impact of dissociation on total correspondence (i.e. accuracy), the remaining correspondence above perfect correspondence represents bias. The bias $\mathbb{E}[S(P, X)] - \mathbb{E}[\mathbb{E}[S(\mathbb{E}[X|P], X)|P]]$ is the *reliability* of the forecasts.

Second, if we treat $\mathbb{E}[X]$ and $\mathbb{E}[X|P]$ as probabilities and write $q = \mathbb{E}[X]$ and $q_p = \mathbb{E}[X|P = p]$, then we can express equation (2.18) as

$$\begin{aligned}
\mathbb{E}[S(P, X)] &= \mathbb{E}[S(q, X)] - (\mathbb{E}[S(q, X)] - \mathbb{E}[\mathbb{E}[S(q_P, X)|P]]) \\
&\quad + (\mathbb{E}[S(P, X)] - \mathbb{E}[\mathbb{E}[S(q_P, X)|P]]) \\
&= S(q, q) - (S(q, q) - \mathbb{E}[S(q_P, q_P)]) + (\mathbb{E}[S(P, X)] - \mathbb{E}[S(q_P, q_P)]) \\
&= \underbrace{S(q, q)}_{\text{Uncertainty}} - \underbrace{\mathbb{E}[S(q, q_P) - S(q_P, q_P)]}_{\text{Resolution}} + \underbrace{\mathbb{E}[S(P, q_P) - S(q_P, q_P)]}_{\text{Reliability}}.
\end{aligned} \tag{2.19}$$

While we have arrived at equation (2.19) because $\mathbb{E}[X]$ and $\mathbb{E}[X|P]$ are probabilities, as presented the URR decomposition in equation (2.19) does not explicitly depend on the form of S nor on the nature of X . Therefore, equation (2.19) suggests an entirely general URR decomposition for scoring rules of probabilistic forecasts (see Bröcker, 2009). We also see that, for proper scoring rules, both the resolution term and the reliability term of the URR decomposition are non-negative.

2.4.2 || RDC Decomposition

Continuing in the setting of X a binary 0/1 observation with each forecast given as probability p (see above), for (half-)Brier scoring rule $S(p, x) = (p - x)^2$, the dual RDC decomposition of equation (2.8) is, in terms of S ,

$$\begin{aligned}
\mathbb{E}[S(P, X)] &= \underbrace{\mathbb{E}[S(P, \mathbb{E}[P])]}_{\text{Refinement}} - \underbrace{(\mathbb{E}[S(P, \mathbb{E}[P])] - \mathbb{E}[\mathbb{E}[S(P, \mathbb{E}[P|X])|X]])}_{\text{Discrimination}} \\
&\quad + \underbrace{(\mathbb{E}[S(P, X)] - \mathbb{E}[\mathbb{E}[S(P, \mathbb{E}[P|X])|X]])}_{\text{Correctness}}. \tag{2.20}
\end{aligned}$$

Equation (2.20) is permissible because although the Brier scoring rule (Brier, 1950) is defined for a binary secondary argument, there is clearly no mathematical difficulty in extending the definition $S(p, x) = (p - x)^2$ to a second argument that takes values in the interval $[0, 1]$ making $S(p, \mathbb{E}[P])$ and $S(p, \mathbb{E}[P|X = x])$ well-defined for all p, x .

The RDC decomposition is dual to the URR decomposition in the sense that to prepare for the RDC decomposition, the forecasts are stratified by outcome. We can then consider the summary measures: (i) $\mathbb{E}[P]$, the overall mean of the forecasts (determined from the marginal distribution of the forecasts), (ii) $\mathbb{E}[P|X]$, the within-stratum means of the forecasts (determined by the conditional distribution of the forecasts given the observation).

A measure of the total variation of the forecasts is the overall correspondence, $\mathbb{E}[S(P, \mathbb{E}[P])]$, between the forecast and its mean value, which is the *refinement* of the forecasts. And, within the stratum determined by $X = x$, the overall correspondence between the forecast and the mean forecast, $\mathbb{E}[S(P, \mathbb{E}[P|X])|X = x]$, is a measure of the variation of the forecasts in the stratum. With the observation fixed for a stratum, all variation in the forecasts in a stratum is variation in the difference between the forecasts and observation and so is a measure of dissociation for the stratum. We define the total measure of dissociation to be the average within-stratum forecast variation, $\mathbb{E}[\mathbb{E}[S(P, \mathbb{E}[P|X])|X]]$, over all strata. Between the strata the observation will take different values and the variation in the forecasts accompanying a change in strata, and not assignable to dissociation, is the variation of the forecasts to align with the observation. This variation of the forecasts for alignment, $\mathbb{E}[S(P, \mathbb{E}[P])] - \mathbb{E}[\mathbb{E}[S(P, \mathbb{E}[P|X])|X]]$ is forecast *discrimination* (the extent to which different outcomes will be preceded by different forecasts). As above, removing the decremental effect of dissociation on accuracy, any inaccuracy that remains is bias in the forecasts; we call the bias, $\mathbb{E}[S(P, X)] - \mathbb{E}[\mathbb{E}[S(P, \mathbb{E}[P|X])|X]]$, the (in)*correctness* of the forecasts.

2.4.3 || Problems and Aims

The URR decomposition (2.19) refers to probabilistic forecasts. There is no immediate modification that would permit it to be applied to other types of forecast. For point forecasts, for example, with scoring rule $S(y, x)$, $x \in \mathcal{X}$, $y \in \mathcal{Y} \subset \mathcal{X}$, it is not clear what to propose for, say, the uncertainty term; one possibility for the uncertainty term is, following the above derivation, $S(\mathbb{E}[X], X)$, but this requires that $\mathbb{E}[X] \in \mathcal{Y}$, which need not be the case, as we may have $\mathbb{E}[X] \in \mathcal{X} \setminus \mathcal{Y}$.

A scoring rule for point forecasts can always be re-expressed in terms of probabilistic forecasts (see Gneiting, 2011a) (and this is the approach taken by Bentzien and Friederichs (2014)), but this approach, when considering decompositions, has both theoretical and practical difficulties. Theoretically, the URR decomposition is derived by stratifying on the *issued* forecast (i.e. the forecast that is the argument of the scoring rule). Re-expressing a scoring rule for point forecasts in terms of the probabilistic forecast underlying the point forecast, requires conditioning on the *underlying probability forecast* and, because there may be a number of probability distributions that map to the same point forecast, this may lead to a different decomposition to that obtained when the scoring rule is expressed in terms of point forecasts and for which the conditioning is with respect to point forecasts. There is the additional difficulty that the probability distribution underlying the point forecast may be latent, so that even if it is theoretically justifiable to condition on a single probability distribution, in practice it may not be possible to stratify the outcomes according to the underlying prob-

ability distribution, a prerequisite to computing the decomposition terms. These problems are not specific to point forecasts, but extend to non-probabilistic forecasts in general.

Different difficulties apply to the RDC decomposition. Subject to being able to extend the chosen scoring rule to allow for a second argument that lies in the space of forecasts, the RDC decomposition (2.20) is entirely general and can be applied to all types of forecast. The extension is often readily done. However, even if S is proper, we cannot, in general, conclude that the discrimination and correctness terms are non-negative. There is a further, overarching, criticism of RDC decomposition (2.20): the summary measures $\mathbb{E}[P]$ and $\mathbb{E}[P|X]$ discount many material characteristics of the marginal and conditional distributions; indeed, entirely different distributions may have the same summary measures. This is not to say that RDC decomposition (2.20) is without all merit. Indeed, [Murphy and Winkler \(1987\)](#) counsel that ‘the fundamental role played by [the joint and conditional] distributions suggests that common summary measures (means, variances, etc.) of such distributions should be useful verification measures in many situations.’ But, if we are to tend closer to the general framework envisaged by [Murphy and Winkler \(1987\)](#) of the ‘careful and reasoned evaluation’ of the conditional, marginal and joint distributions of forecast and observation, we must attempt to establish decompositions with attributes defined by quantities that refer more intimately to the distributions of the forecast and observation. Such decompositions will not offer complete measures of the forecasts qualities as we have shown in the previous section. Nonetheless, the decompositions are at times the *only* assessment of forecast quality available and in section 2.5 we derive entirely general decompositions that, for some classes of forecast and observation and for some scoring rules, reduce to the decompositions of equation (2.19) and (2.20) above.

[2.5] Most General Decompositions

Define a forecast for the observation X to be a value, $f(p)$, for some $p \in \mathcal{P}$, of a function f on \mathcal{P} . The image space of f , $\mathcal{F}_f \stackrel{\text{def}}{=} \{f(p') | p' \in \mathcal{P}\}$, is the set of possible forecasts for X . We need not, at this stage, be more specific about the form of f , but we make it clear that the forecast $f(p)$ depends on $p \in \mathcal{P}$, although this dependence may be implicit. We do not place any constraints on the image space of the observation, \mathcal{X} .

To allow for such an open description of a forecast, it is necessary to define a scoring rule in more general terms.

Definition 2.5.1 (Scoring Rule). A scoring rule with respect to the class of forecasts \mathcal{F}_f is a real-valued function, $S : \mathcal{F}_f \times \mathcal{X} \rightarrow \mathbb{R}$, with the value, for forecast $f(p)$ and outcome

x , $S(f(p), x)$, referred to as a *score*. A scoring rule need *not* be symmetric in its arguments. \square

We continue to regard all scoring rules as negatively-oriented. If X has probability distribution q and distribution function G , let

$$S(f(p), q) \stackrel{\text{def}}{=} \mathbb{E}[S(f(p), X)] = \int_{\mathcal{X}} S(f(p), x) dG(x). \quad (2.21)$$

We now generalise the definition of a proper scoring rule.

Definition 2.5.2 (Proper Scoring Rule). A scoring rule, S , with respect to the class of forecasts \mathcal{F}_f , is proper if and only if

$$S(f(q), q) \leq S(f(p), q) \quad \forall p, q \in \mathcal{P}. \quad (2.22)$$

S is *strictly* proper only if for $p \neq q$,

$$S(f(q), q) < S(f(p), q). \quad (2.23)$$

\square

2.5.1 || URR Decomposition

Fix S to be a proper scoring rule. A forecast, f , and outcome, x , correspond inasmuch as they are close (in some sense) to one another. The score $S(f, x)$ may be viewed as such a ‘distance’ between f and x . We use the word ‘distance’ in a general, descriptive sense and S need not be a metric nor a quasi-distance (or, divergence) (Bröcker and Smith, 2007a): (a) the score evaluated by S has (primarily) an ordinal scale and need not be greater than or equal to zero; (b) S need not be symmetric in its arguments, and is typically not so because of the differing identities of the forecast and the verification; (c) nor is there an *a priori* imperative on any scoring rule that it satisfy the triangle inequality. (See Deza and Deza (2014, page 4,5) for the distinction between a metric and quasi-distance.)

Given that, using S , we have a measure of the correspondence between forecast and observation, we ask what the best forecast is, i.e. what forecast has the best correspondence to the observation? Because each forecast is a function of a probability distribution of the observation, this question can be rephrased as: what is the best probability distribution

(considering the forecasting system) to propose for the observation? Since, for $r \in \mathcal{P}$, the correspondence between the forecast $f(r)$ and X is $\mathbb{E}[S(f(r), X)]$, i.e. the average ‘distance’ between $f(r)$ and the possible outcomes, the best probability distribution is the probability distribution, r_{\min} , that minimises this ‘distance’ i.e. satisfies $r_{\min} = \arg \min_r \mathbb{E}[S(f(r), X)]$, where the minimum depends on what we know about X (which determines the evaluation of the expectation). We are reminded by [Murphy and Winkler \(1987\)](#) that in the context of forecast evaluation ‘the joint distribution [of forecast and observation] contains *all* of the relevant information’. In other words, the information about X will be either (i) the marginal (or, unconditional) distribution of the observation, or (ii) the conditional distribution of the observation given the forecast.

Let q be the unconditional probability distribution of X (that is, the marginal, base-rate, climatology or historical probability distribution of X), and let the conditional distribution of X given the extant forecast, $f(p)$, be $q_{f(p)}$. If we admit only the unconditional information about X (in the case we have no forecast information) then the best forecast is, by the propriety of S , $f(q)$ and we have

$$\min_r \mathbb{E}[S(f(r), X)] = \min_r S(f(r), q) = S(f(q), q). \quad (2.24)$$

However, if we know the extant forecast, $f(p)$, then we have $X \sim q_{f(p)}$ and the best forecast is, by the propriety of S , $f(q_{f(p)})$, with

$$\min_r \mathbb{E}[S(f(r), X)] = \min_r S(f(r), q_{f(p)}) = S(f(q_{f(p)}), q_{f(p)}). \quad (2.25)$$

As $q_{f(p)}$ is the optimal (here, minimising) probability distribution, and, by propriety

$$S(f(q_{f(p)}), q_{f(p)}) \leq S(f(p), q_{f(p)}) \quad (2.26)$$

the difference

$$S(f(p), q_{f(p)}) - S(f(q_{f(p)}), q_{f(p)}) \quad (2.27)$$

is a measure of the sub-optimality of the extant forecast, $f(p)$, relative to the optimal forecast, $f(q_{f(p)})$, and this difference is greater than or equal to 0 by the propriety of S .

To calculate the sub-optimality of the forecasts overall, we take the average,

$$\mathbb{E}[S(f(P), q_{f(P)}) - S(f(q_{f(P)}), q_{f(P)})], \quad (2.28)$$

which we define to be the *reliability* of the forecasts. The greater the value of the reliability term, the more *unreliable* the forecasts. The forecasts are perfectly reliable (i.e. optimal)

if $f(p) = f(q_{f(p)})$ for all $f(p)$. The first term in equation in (2.28) is simply accuracy, $\mathbb{E}[S(f(P), X)]$. We refer to the second term of equation (2.28), $\mathbb{E}[S(f(q_{f(P)}), q_{f(P)})]$, as the *sharpness* of the forecasts. The reason for this terminology is as follows. The forecast $f(p)$ is identified with a particular a stratum of outcomes (those outcomes that follow the forecast $f(p)$). Were the outcomes limited to those within the stratum, the best forecast would be $f(q_{f(p)})$. If within the stratum, the outcomes correspond closely to their best forecast, then they will be concentrated around their best forecast. If this is true for all strata determined by the different forecasts, then the stratification of outcomes by the forecasts leads to a well-delineated, clear, or sharp, partition of the outcomes.

Forecasts may be sharp but unreliable if in the strata determined by the forecasts, there is a narrow spread of outcomes, but the outcomes in each stratum show little match to the published forecast associated with the stratum (i.e. the stratum's published forecast is different to the stratum's optimal forecast). For example, consider the repeated toss of a fair coin. Suppose that forecasts are either $f(p) = 1$ or $f(p) = 0$ (i.e. stating with certainty that the coin will land on heads or tails) but on every occasion that $f(p) = 0$ a head occurs and on every occasion that $f(p) = 1$ a tail occurs; the forecasts are perfectly sharp because for every forecast it is clear what outcome will occur, but totally unreliable because the forecast of each stratum is maximally different from the optimal forecast for the stratum (the optimal forecast being the outcome itself).

Equally forecasts may be reliable but unsharp: the issued forecasts are as close a match as possible to the outcomes collectively, but provide little clarity about the outcome that will occur (as the outcomes are broadly dispersed about the forecast and there are large overlaps between the strata). For example, consider again the repeated toss of a fair coin, for which, on every toss, the forecast is always $f(p) =$ 'probability of heads is $1/2$ '. The forecasts are reliable because they are as close as possible to the outcomes collectively (i.e. both head and tail), but are completely unsharp giving no guidance about the value that will occur (given the forecast, the outcomes are dispersed over the entire observation space).

Sharpness, therefore, refers to how well the forecasts separate the outcomes into groups and reliability refers to how well the outcomes in each group match their classifying forecast.

The level of sharpness must be assessed relative to the difficulty of separating the outcomes: if it is difficult to divide the outcomes into groups, then low levels of sharpness may nonetheless be acceptable. To evaluate the difficulty of grouping the outcomes, we consider the grouping of the forecasts in the absence of any issued forecasts. In this case, X has the unconditional distribution, q , and there is a single stratum of all the outcomes with optimal reference forecast $f(q)$. How tightly concentrated all the outcomes are around $f(q)$ is

given by $\mathbb{E}[S(f(q), X)] = S(f(q), q)$, which we call the *uncertainty* of X . We can compare uncertainty to the sharpness of the forecasts. Noting that,

$$S(f(q), q) = \mathbb{E}[S(f(q), q_{f(P)})] \quad (2.29)$$

and by propriety $S(f(q), q_{f(p)}) \geq S(f(q_{f(p)}), q_{f(p)})$ for all $f(p)$, we have

$$S(f(q), q) \geq \mathbb{E}[S(f(q_{f(P)}), q_{f(P)})]. \quad (2.30)$$

From the above definitions, we can reciprocally, consider sharpness as uncertainty in the presence of forecasts, so that in moving from not having forecasts to having forecasts, uncertainty always decreases (improves). The amount of improvement is

$$S(f(q), q) - \mathbb{E}[S(f(q_{f(P)}), q_{f(P)})] \geq 0 \quad (2.31)$$

which is the *resolution* of the forecasts: with the introduction of forecasts, uncertainty about the possible outcome of X resolves from $S(f(q), q)$ to the lower level $\mathbb{E}[S(f(q_{f(P)}), q_{f(P)})]$.

It follows from the relationship between sharpness and resolution that forecasts can be unresolved but reliable, or unreliable and resolved. For example, it is clear that a forecasting system that issues only the climatological forecast, $f(q)$, is both totally unresolved and perfectly reliable (both resolution and reliability are zero).

We have, in summary, that accuracy is a combination of reliability and sharpness and that sharpness can be expressed as uncertainty less resolution. This deconstruction of accuracy may be expressed more formally as the Uncertainty-Resolution-Reliability (URR) decomposition of the forecasts, given by the identity

$$\begin{aligned} \mathbb{E}[S(f(P), X)] = & \underbrace{S(f(q), q)}_{\text{Uncertainty}} - \underbrace{(S(f(q), q) - \mathbb{E}[S(f(q_{f(P)}), q_{f(P)})])}_{\text{Resolution}} \\ & \underbrace{\hspace{10em}}_{\text{Sharpness}} \\ & + \underbrace{\mathbb{E}[S(f(P), q_{f(P)}) - S(f(q_{f(P)}), q_{f(P)})]}_{\text{Reliability}}. \quad (2.32) \end{aligned}$$

Recently, [Siegert \(2017\)](#) has given an elegant derivation of [Bröcker \(2009\)](#)'s general URR decomposition for proper scoring rules of precise-probabilistic forecasts. As [Siegert \(2017\)](#)

emphasises, the URR decomposition is, fundamentally, a simple arithmetic identity, in which the accuracy of the extant forecasts is equal to the sum of the differences between accuracy and the accuracies of two other comparative sets of forecasts (which [Siegert \(2017\)](#) calls the recalibrated forecasts and the reference forecasts). However, [Siegert \(2017\)](#) allows these comparative forecasts to be freely chosen (although [Siegert \(2017\)](#) does insist that these choices be justified). It is this freedom that limits [Siegert \(2017\)](#)'s approach and the extension of this approach to other types of forecast (and is contrary to the constrained choices of the general framework of [Murphy and Winkler \(1987\)](#)). For, not only does the allowance of any choices for the comparative forecasts detach the resulting terms from the attributes that they are intended to represent (for example, in [Siegert \(2017\)](#)'s decomposition, it is feasible to swap the comparative forecasts for one another, confusing the meaning of the decomposition's terms), it is also the case that, as [Siegert \(2017\)](#) acknowledges, free choice of the comparative forecasts, can lead to the terms of the decomposition being negative, which further complicates their interpretation.

Although our URR decomposition satisfies the arithmetic identity highlighted by [Siegert \(2017\)](#), our approach is the inverse. We derive, from meta-principles, mathematical quantities that satisfy the descriptive definitions of the attributes of uncertainty, resolution and reliability and then show that these mathematical quantities form a decomposition of accuracy.

2.5.2 || RDC Decomposition

Fundamentally, the URR decomposition compares three probability distributions of the *observation*: the unconditional, or marginal, distribution of the observation, the conditional distribution of the observation given the forecast and the distribution underlying the forecast. Of these three probability distributions, the probability distribution underlying the forecast is, from the perspective of the forecasting system, the ideal distribution for the observation (the probability distribution underlying the forecast is the distribution the forecasting system believes *should be* the probability distribution of the observation); if the conditional distribution of the observation given the forecasts is ideal (i.e. matches the ideal distribution) then the forecasts are perfectly reliable (if X is discrete, for example, we have $\Pr(X = x|f(P) = f(p)) = \Pr_p(X = x)$ for all p such that $\Pr(P = p) > 0$). In contrast, the unconditional distribution of the observation is, to the forecasting system, the base probability distribution (for the forecasting system to improve upon). The URR decomposition shows where the conditional distribution of the observation (given the forecasts) lies between the base and ideal distributions of the observation.

However, the URR decomposition, by referring to just one member (the observation) of the forecast-observation pair, provides only partial insight into the properties of the forecasting system. To complete the evaluation of the forecasting system, the features of the forecasts that precede the outcomes must also be analysed.

For *verification purposes*, within the general framework of [Murphy and Winkler \(1987\)](#), the extent of our knowledge of the features of the forecasts is summarised by: (i) the unconditional (i.e. marginal, or historical) distribution of the forecasts, (ii) the conditional distribution of the forecasts given the observation.

We would like to develop a decomposition that refers to the forecasts and is dual to the URR decomposition in that it places the conditional distribution of the forecasts given the observation between a *base* distribution and an *ideal* distribution. The base distribution is the unconditional distribution of the forecasts. It remains to define an ideal distribution for the forecasts: the probability distribution that, were the outcome known, the forecasting system would most like the forecasts to have.

It seems reasonable to propose that forecasts are ideal or perfect if, for each outcome x , the forecast preceding x is categorical that x will occur. In other words, given $X = x$, the ideal distribution of the forecasts, which we denote by δ_x^* , is a point-mass distribution on the forecasts at $f(P) = f(\delta_x)$.

In broad terms, therefore, the correspondence between forecast and observation will lie on a scale between ‘no correspondence’ and ‘perfect correspondence’. A forecasting system with forecasts that correspond well with the observation will have a distribution of forecasts, conditional on the observation, that is close to ‘ideal’, while a forecasting system with forecasts poorly aligned with the observation will have a conditional distribution of the forecasts given the observation that is closer to the ‘base’ unconditional distribution of the forecasts. We can assess the forecasts by determining how close the conditional distribution of the forecasts given the observation is to both the base and ideal distributions of the forecasts.

To position the conditional distribution of the forecasts given the observation on the scale from base distribution to ideal distribution, we require the notion of ‘distance’ between any two distributions of forecasts. We choose to define such a ‘distance’ between a nominated distribution of forecasts p^* and a second distribution of forecasts r^* as the average, under r^* , of the distances between p^* and each of the forecasts. This definition requires, in turn, a notion of distance between a distribution of forecasts and a single forecast. We define the distance between a distribution of forecasts and a forecast using a scoring rule, S^* .

A particular use of the scoring rule S^* will be to determine the distance between the ideal distribution given $X = x$, δ_x^* , and the extant forecast preceding $X = x$, say, $f(p)$; we write this distance as $S^*(\delta_x^*, f(p))$ following the conventions of all scoring rules. Of note, however, is that the distance $S^*(\delta_x^*, f(p))$ also expresses a correspondence between $f(p)$ and x , for which we have, under the scoring rule S , the measure $S(f(p), x)$. To be consistent in our evaluation in quantifying correspondence between a forecast and outcome, we would like to choose S^* so that $S^*(\delta_x^*, f(p)) = S(f(p), x)$. To distinguish S^* from S , we refer to S^* as an *extended* scoring rule. Let \mathcal{P}^* be the set of probability distributions over the set of forecasts, \mathcal{F}_f .

Definition 2.5.3 (Extended Scoring Rule). A real-valued function $S^* : \mathcal{P}^* \times \mathcal{F}_f \rightarrow \mathbb{R}$ is called an extended scoring rule (for S) with respect to the class of forecasts \mathcal{F}_f if, for each $f(p) \in \mathcal{F}_f$, $x \in \mathcal{X}$,

$$S^*(\delta_x^*, f(p)) = S(f(p), x). \quad (2.33)$$

(We bring to attention the specific order of the arguments in both S^* and S .) □

For interpretation purposes we are concerned only with proper extended scoring rules. The definition of a proper extended scoring rule follows naturally from equation (2.3): letting $q^* \in \mathcal{P}^*$ be the probability distribution of $f(P)$, define

$$S^*(p^*, q^*) \stackrel{\text{def}}{=} \mathbb{E}[S^*(p^*, f(P))] \text{ for all } p^* \in \mathcal{P}^*. \quad (2.34)$$

Definition 2.5.4 (Proper Extended Scoring Rule). The (negatively-oriented) extended scoring rule, $S^* : \mathcal{P}^* \times \mathcal{F}_f \rightarrow \mathbb{R}$ is said to be proper if and only if

$$S^*(q^*, q^*) \leq S^*(p^*, q^*) \text{ for all } p^*, q^* \in \mathcal{P}^*. \quad (2.35)$$

□

With S^* a proper extended scoring rule, let q^* be the unconditional probability distribution of the forecasts, and q_x^* the distribution of the forecasts given $X = x$. Given $X = x$, we would like conditional distribution, q_x^* , to be as close as possible to the ideal distribution, δ_x^* , as Figure 2.1 illustrates.

The difference

$$S^*(\delta_x^*, q_x^*) - S^*(q_x^*, q_x^*) \quad (2.36)$$

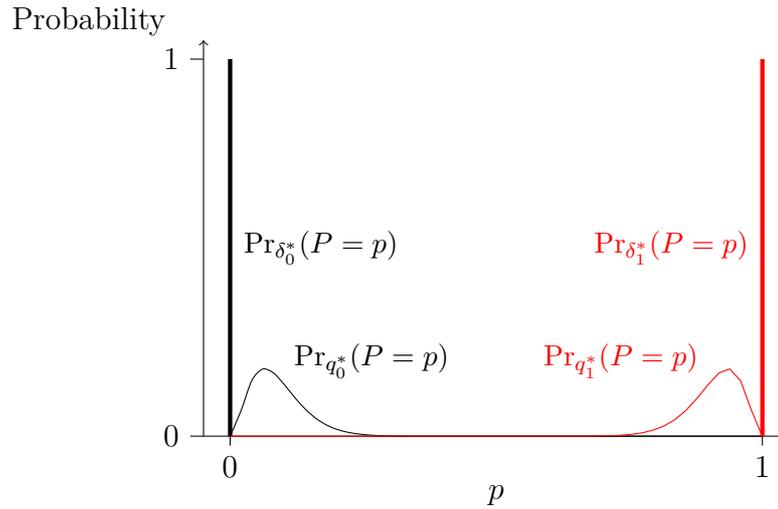


Figure 2.1: X is a binary 0/1 observation, with forecasts $\Pr_p(X = 1)$ for different probability distributions p . For simplicity, for each distribution p , write p for $\Pr_p(X = 1)$ and allow for $p \in [0, 1]$. Each distribution of forecasts is a distribution for the values of p . The ideal distributions of the forecasts are shown as thick vertical lines: the ideal distribution of forecasts preceding $x = 1$ assigns probability of 1 to $p = 1$ (vertical red line) i.e. ideally all forecasts preceding $x = 1$ should be $p = 1$; similarly, the ideal distribution of forecasts preceding $x = 0$ assigns a probability of 1 to $p = 0$ (vertical black line). The closer the actual distribution of forecasts preceding $x = 1$, q_1^* , is to the ideal distribution, δ_1^* , the better, and the closer the actual distribution of forecasts preceding $x = 0$, q_0^* , is to the ideal distribution, δ_0^* , the better.

(which is always positive by the propriety of S^*) conveys how far from ideal the conditional distribution of the forecasts is. Because the ideal distribution δ_x^* identifies with perfectly correct forecasts (for $X = x$), the further the conditional distribution q_x^* is from δ_x^* the more incorrect the forecasts may be deemed to be. Averaging the difference in equation (2.36) over all outcomes gives

$$\mathbb{E}[S^*(\delta_X^*, q_X^*) - S^*(q_X^*, q_X^*)] \quad (2.37)$$

which we refer to as the (in)correctness of the forecasts. We note that the first term of equation (2.37) is

$$\begin{aligned} \mathbb{E}[S^*(\delta_X^*, q_X^*)] &= \mathbb{E}[\mathbb{E}[S^*(\delta_X^*, f(P))|X]] \\ &= \mathbb{E}[S^*(\delta_X^*, f(P))] \\ &= \mathbb{E}[S(f(P), X)] \end{aligned} \quad (2.38)$$

which is the accuracy of the forecasts.

To interpret the second term of equation (2.37), $\mathbb{E}[S^*(q_X^*, q_X^*)]$, we begin with some general reasoning. If r^* is the distribution of a group of forecasts, then the average ‘distance’ of the forecasts in the group from r^* , $S^*(r^*, r^*)$, reflects the variation of the forecasts in the group: if the forecasts in the group change, becoming more or less varied, r^* , being the distribution of these forecasts, will change and so too then will the distance from each forecast in the group to r^* change, changing the average ‘distance’ $S^*(r^*, r^*)$; each value of $S^*(r^*, r^*)$ can be associated with a particular spread of the forecasts.

Returning to the second term of equation (2.37), it is expected that forecasts will vary in an attempt to adjust to changes in the observations. Forecasts will also (to a greater or lesser extent) exhibit exogenous variations, which are variations for reasons beyond those related to a need to adapt to movements in the values of the observations. Suppose that the value of X is known i.e. we have $X = x$ for some x . There is no reason for the forecasts that precede the value x of X to show any variation: with x fixed, it is not necessary for the corresponding forecasts to differ between themselves. Any variation in the forecasts preceding $X = x$, therefore, must be exogenous variation. The distribution of the forecasts preceding $X = x$ is, by definition, q_x^* , and by the general reasoning above, the variation of the forecasts within the stratum determined by the outcome $X = x$ can be measured by $S^*(q_x^*, q_x^*)$. Therefore, we define $\mathbb{E}[S^*(q_X^*, q_X^*)]$ as the *excess* variation of the forecasts (a term which we re-appropriate from Yates (1982)).

(In)correctness, therefore, is accuracy less excess variation. The degree of (in)correctness, or, whether forecasts are near to or far from ideal, depends on the scale. To determine the scale, we must also assess the distance of the conditional distribution of the forecasts to the unconditional distribution of the forecasts.

In direct analogy with equation (2.37), how far the conditional distribution of the forecasts is from the unconditional distribution of the forecasts, is

$$\mathbb{E}[S^*(q^*, q_X^*) - S^*(q_X^*, q_X^*)] = S^*(q^*, q^*) - \mathbb{E}[S^*(q_X^*, q_X^*)] \quad (2.39)$$

Applying our earlier general reasoning, $S^*(q^*, q^*)$ is the total variation of the forecasts as a whole (with unconditional distribution q^*). We refer to $S^*(q^*, q^*)$ as the *refinement* of the forecasts (confusingly, the term refinement is sometimes used differently to refer to sharpness as in, for example, DeGroot and Fienberg (1983) and Blattenberger and Lad (1985)). Whether excess variation is relatively high depends on how the excess variation compares to the total variation (i.e. refinement).

If the excess variation of the forecasts is close to the forecasts’ refinement then most of the

forecasts' variation is excessive, exogenously driven, and we conclude that excess variation is high. The difference between refinement and excess,

$$S^*(q^*, q^*) - \mathbb{E}[S^*(q_X^*, q_X^*)], \quad (2.40)$$

is known as the *discrimination* of the forecasts and, as the non-exogenous, or necessary, component of the forecasts' variation, indicates the ability of the forecasts to adjust to X , i.e. to anticipate different outcomes for X . When excess variation is high, and thereby close to refinement, discrimination is low and close to zero.

Equations (2.37) and (2.39) together, position the forecasts on the scale between 'no correspondence' and 'perfect correspondence' with the observation. Collecting the refinement, discrimination and correctness terms we have

$$\begin{aligned} \mathbb{E}[S(f(P), X)] &= \mathbb{E}[S^*(\delta_X^*, f(P))] \\ &= S^*(q^*, q^*) - (S^*(q^*, q^*) - \mathbb{E}[S^*(q_X^*, q_X^*)]) \\ &\quad + \mathbb{E}[S^*(\delta_X^*, q_X^*) - S^*(q_X^*, q_X^*)] \\ &= \underbrace{S^*(q^*, q^*)}_{\text{Refinement}} - \underbrace{(S^*(q^*, q^*) - \mathbb{E}[S^*(q_X^*, q_X^*)])}_{\text{Discrimination}} \\ &\quad + \underbrace{(\mathbb{E}[S(f(P), X)] - \mathbb{E}[S^*(q_X^*, q_X^*)])}_{\text{Correctness}} \quad (2.41) \\ &\quad \underbrace{\hspace{10em}}_{\text{Excess}} \end{aligned}$$

giving a second, dual decomposition for accuracy, the refinement-discrimination-correctness (RDC) decomposition.

From equation (2.41) and the arguments preceding it, it can be seen that for *any* proper scoring rule, $S^* : \mathcal{P}^* \times \mathcal{F}_f \rightarrow \mathbb{R}$, $S^*(q^*, q^*)$ is a measure of refinement and the difference $S^*(q^*, q^*) - \mathbb{E}[S^*(q_X^*, q_X^*)]$ is a measure of discrimination. However for these measures to be comparable and consistent with the measures, under S , of uncertainty, resolution and reliability, *and* for the difference $\mathbb{E}[S(f(P), X)] - \mathbb{E}[S^*(q_X^*, q_X^*)]$ to be a true reflection of (in)correctness, the proper scoring rule S^* must be an extension of S . We consider this point further in the examples that now follow.

[2.6] Examples

We demonstrate the evaluation of the URR and RDC decompositions, with several examples beginning with the (half-)Brier scoring rule (Brier, 1950) and revisiting the results of section 2.4. The simplicity and familiarity of the (half-)Brier scoring rule will be helpful in illustrat-

ing the more abstract ideas proposed so far in this chapter. In several of the examples it will be necessary to indicate the probability distribution that is referred to by an expectation or distribution function. In such cases, the probability distribution is shown as a subscript. For example, $\mathbb{E}_q[X]$ is the expected value of X when $X \sim q$, and F_q is the distribution function corresponding to q .

2.6.1 || Brier Scoring Rule

Let 0 and 1 be the only possible values that the observation, X , can take. For each $p \in \mathcal{P}$, the forecast is $f(p) = \Pr_p(X = 1)$. Because each probability distribution of X is completely identified by the value that the probability distribution assigns to the probability that $X = 1$, knowing p is equivalent to knowing $\Pr_p(X = 1)$, and so for simplicity of notation, we write p for $\Pr_p(X = 1)$. Then the (half-)Brier scoring rule (Brier, 1950) is defined by

$$S(f(p), x) = S(\Pr_p(X = 1), x) = S(p, x) = (p - x)^2. \quad (2.42)$$

Brier (1950) notes that S is a proper scoring rule. Because $p = \Pr_p(X = 1) = \mathbb{E}_p[X]$, we write equation (2.42) as

$$S(p, x) = (\mathbb{E}_p[X] - x)^2. \quad (2.43)$$

2.6.1.1 | URR Decomposition

Taking the expectation of the equation (2.43) with respect to X under a probability distribution $r \in \mathcal{P}$, we have

$$S(p, r) \stackrel{\text{def}}{=} \mathbb{E}_r[S(p, X)] = \mathbb{E}_p^2[X] - 2\mathbb{E}_p[X]\mathbb{E}_r[X] + \mathbb{E}_r[X] \quad (2.44)$$

from which, letting q be the unconditional (i.e climatological) probability distribution of X ,

$$S(q, q) = \mathbb{E}[X] - \mathbb{E}^2[X] \quad (2.45)$$

and (recalling the notation defined in equation (2.2))

$$S(f(q_{f(p)}), q_{f(p)}) \equiv S(q_p, q_p) = \mathbb{E}[X|P = p] - \mathbb{E}^2[X|P = p]. \quad (2.46)$$

By simple calculation, from equation (2.32) the URR decomposition of the Brier scoring rule is equal to the decomposition in equation (2.7).

2.6.1.2 | RDC Decomposition

Define the scoring rule S^* by (recalling that $f(p) = p$)

$$S^*(r^*, p) = (\mathbb{E}_{r^*}[P] - p)^2 \text{ for all } r^* \in \mathcal{P}^*, p \in \mathcal{P}. \quad (2.47)$$

Then S^* is proper, for

$$\begin{aligned} S^*(r^*, q^*) &= \mathbb{E}_{q^*}[(\mathbb{E}_{r^*}[P] - P)^2] \\ &= \mathbb{E}_{r^*}^2[P] - 2\mathbb{E}_{r^*}[P]\mathbb{E}_{q^*}[P] + \mathbb{E}_{q^*}[P^2] \\ &= (\mathbb{E}_{r^*}[P] - \mathbb{E}_{q^*}[P])^2 + \mathbb{E}_{q^*}[(P - \mathbb{E}_{q^*}[P])^2] \end{aligned} \quad (2.48)$$

which is minimised for $r^* = q^*$. And, noting that it is valid under the (half-)Brier scoring rule for $p \in \mathcal{P}$ to be 0 or 1, the ideal distribution δ_x^* is defined as the point-mass distribution on $f(\delta_x) = \Pr_{\delta_x}(X = 1) = x$, so that

$$\begin{aligned} S^*(\delta_x^*, p) &= (\mathbb{E}_{\delta_x^*}[P] - p)^2 \\ &= (\Pr_{\delta_x}(X = 1) - p)^2 \\ &= (x - p)^2 \\ &= S(p, x). \end{aligned} \quad (2.49)$$

Therefore, S^* is a proper extended scoring rule for S . From equation (2.48), we have

$$S^*(q^*, q^*) = \mathbb{E}[(P - \mathbb{E}[P])^2], \quad (2.50)$$

the variance of the forecasts under their historical distribution, and for all x ,

$$S^*(q_x^*, q_x^*) = \mathbb{E}[(P - \mathbb{E}[P|X = x])^2 | X = x]. \quad (2.51)$$

Referring to equation (2.41), the RDC decomposition for the (half-)Brier scoring rule is, with a little calculation, given by the decomposition in equation (2.8).

2.6.2 || Ranked Probability Scoring Rule

The *quadratic* scoring rule generalises the (half-)Brier scoring rule of the previous section (the quadratic scoring rule is also referred to as the probability scoring rule (Murphy and Epstein, 1967b); the use of the terminology ‘quadratic scoring rule’ seems to have appeared first in Winkler and Murphy (1968)). The most general form of the quadratic scoring rule is given

by [Gneiting and Raftery \(2007, equation \(18\)\)](#), which in our negatively-oriented context is, letting $\varphi_p(\cdot)$ be the probability density function of X under the probability distribution p ,

$$S(p, x) = \frac{1}{2} \left(\mathbb{E}_p[\varphi_p(X)] + 1 \right) - \varphi_p(x) \quad (2.52)$$

(where we have also multiplied by a factor of one-half). An alternative generalisation of the (half-)Brier scoring rule [Brier \(1950\)](#) is to consider the *cumulative* distribution function of the forecast relative to the *cumulative* outcome (the cumulative outcome being equivalent to the *cumulative* distribution function for the point-mass distribution at the outcome). Letting $f(p) = F_p$ for each $p \in \mathcal{P}$, $A_x = \{x' \in \mathcal{X} | x' \geq x\}$ be a subset of \mathcal{X} for each $x \in \mathcal{X}$ and define (where the indicator function $\mathbb{1}_A(y) = 1$ if $y \in A$ and 0 otherwise)

$$S(f(p), x) = S(F_p, x) = \int_{\mathcal{X}} (F_p(y) - \mathbb{1}_{A_x}(y))^2 dy \quad (2.53)$$

The scoring rule defined by equation (2.53) is referred to as the *ranked* probability scoring rule. Particular instances of the ranked probability scoring rule are the following.

- (i) (Half-)Brier Scoring Rule ([Brier, 1950](#)), for dichotomous events: $X \in \{0, 1\} = \mathcal{X}$, $A_0 = \{0, 1\}$, $A_1 = \{1\}$, $p = \Pr(X = 1)$, $F_p(0) = 1 - p$, $F_p(1) = 1$, then

$$S(f(p), x) = (p - x)^2. \quad (2.54)$$

- (ii) (Discrete) Ranked Probability Score ((D)RPS) ([Epstein, 1969](#); [Murphy, 1971](#); [Wilks, 2006](#)) for events that can take one of a discrete number, $n \geq 2$, different possible values, one of which must occur: letting \mathbf{e}_j be an n -component vector with every entry 0 except the j th component, which is 1 i.e. the j th natural basis vector of \mathbb{R}^n , $X \in \{\mathbf{e}_j | j = 1, \dots, n\} = \mathcal{X}$. Let the ordering, or *ranking* on \mathcal{X} be $\mathbf{e}_j \leq \mathbf{e}_k$ if and only if $j \leq k$. Then, $A_{\mathbf{e}_j} = \{\mathbf{e}_k \in \mathcal{X} | j \leq k\}$. Setting $p_j = \Pr(X = \mathbf{e}_j)$, we have under the given ordering $F_p(\mathbf{e}_k) = \Pr_p(X \leq \mathbf{e}_k) = \sum_{j=1}^k \Pr_p(X = \mathbf{e}_j) = \sum_{j=1}^k p_j$. The discrete ranked probability scoring rule is

$$S(f(p), x) = \sum_{i=1}^n \left(\sum_{j=1}^i p_j - \sum_{j=1}^i x_j \right)^2. \quad (2.55)$$

A usual circumstance in which the DRPS is applicable is when an event has a finite number of possible outcomes (either values or classes), only one of which may occur. Then the observation X is a vector for which the j th component is 1 if and only if the j th outcome (value or class) occurs for the event.

- (iii) Continuous Ranked Probability Score (CRPS) (Brown, 1974; Gneiting and Raftery, 2007; Matheson and Winkler, 1976) for events, the outcome of which takes a value in the subset of the real-numbers: $X \in \mathcal{X} \subset \mathbb{R}$, $A_x = [x, \infty)$, $F_p(x) = \Pr_p(X \leq x)$, then

$$S(f(p), x) = \int_{\mathcal{X}} (F_p(y) - \mathbf{1}_{[x, \infty)}(y))^2 dy. \quad (2.56)$$

As shown in Murphy (1969) and Gneiting and Raftery (2007) the ranked probability scoring rule is proper.

It will be helpful (particularly when determining the RDC decomposition), to note that the ranked probability scoring rule may be written equivalently as

$$S(f(p), x) = S(\mathbb{E}_p[\mathbf{1}_{A_x}], x) = \int_{\mathcal{X}} (\mathbb{E}_p[\mathbf{1}_{A_x}(y)] - \mathbf{1}_{A_x}(y))^2 dy \quad (2.57)$$

where we have $f(p) = \mathbb{E}_p[\mathbf{1}_{A_x}]$.

In the interests of clarity, we restrict attention to one particular case of the quadratic scoring rule, the CRPS (all other cases of the quadratic scoring rule follow similarly). From above, for the CRPS, $\mathcal{X} = \mathbb{R}$, $A_x = [x, \infty)$ and $f(p) = F_p = \mathbb{E}_p[\mathbf{1}_{[X, \infty)}]$.

2.6.2.1 | URR Decomposition

We have, assuming sufficiently regular conditions for the interchange of expectation and integration, for $r \in \mathcal{P}$,

$$\begin{aligned} S(f(p), r) &= \int_{-\infty}^{\infty} (\mathbb{E}_p[\mathbf{1}_{[X, \infty)}(y)] - \mathbb{E}_r[\mathbf{1}_{[X, \infty)}(y)])^2 dy \\ &\quad + \int_{-\infty}^{\infty} \mathbb{E}_r[\mathbf{1}_{[X, \infty)}(y)](1 - \mathbb{E}_r[\mathbf{1}_{[X, \infty)}(y)]) dy. \end{aligned} \quad (2.58)$$

From equation (2.58), with q the unconditional (i.e. climatological) probability distribution of X ,

$$S(f(q), q) = \int_{-\infty}^{\infty} \mathbb{E}[\mathbf{1}_{[X, \infty)}(y)](1 - \mathbb{E}[\mathbf{1}_{[X, \infty)}(y)]) dy \quad (2.59)$$

and

$$S(f(q_{f(p)}), q_{f(p)}) = \int_{-\infty}^{\infty} \mathbb{E}[\mathbf{1}_{[X, \infty)}(y)|f(p)](1 - \mathbb{E}[\mathbf{1}_{[X, \infty)}(y)|f(p)]) dy, \quad (2.60)$$

from which (by equation (2.32)) the URR decomposition for the CRPS is

$$\begin{aligned}
 \mathbb{E}[S(f(P), X)] = & \\
 & \left. \begin{aligned}
 & \underbrace{\int_{-\infty}^{\infty} \mathbb{E}[\mathbf{1}_{[X, \infty)}(y)](1 - \mathbb{E}[\mathbf{1}_{[X, \infty)}(y)]) dy}_{\text{Uncertainty}} \\
 & - \underbrace{\mathbb{E} \left[\int_{-\infty}^{\infty} \left(\mathbb{E}[\mathbf{1}_{[X, \infty)}(y) | f(P)] - \mathbb{E}[\mathbf{1}_{[X, \infty)}(y)] \right)^2 dy \right]}_{\text{Resolution}}
 \end{aligned} \right\} \text{Sharpness} \\
 & + \underbrace{\mathbb{E} \left[\int_{-\infty}^{\infty} \left(\mathbb{E}[\mathbf{1}_{[X, \infty)}(y) | f(P)] - f(P)(y) \right)^2 dy \right]}_{\text{Reliability}}.
 \end{aligned} \tag{2.61}$$

As the information in p is the same as the information in F_p , expectations conditional on $f(p) = F_p$ may be written as expectations conditional on P .

2.6.2.2 | RDC Decomposition

Motivated by the representation of the quadratic scoring rule given in equation (2.57), define, for $r^* \in \mathcal{P}^*$,

$$S^*(r^*, F_p) = \int_{-\infty}^{\infty} (\mathbb{E}_{r^*}[F_P(y)] - F_p(y))^2 dy. \tag{2.62}$$

Taking expectations of equation (2.62) with respect to F_P , we have

$$\begin{aligned}
 S^*(r^*, q^*) &= \mathbb{E}_{q^*}[S^*(r^*, F_P)] \\
 &= \int_{-\infty}^{\infty} (\mathbb{E}_{r^*}[F_P(y)] - \mathbb{E}_{q^*}[F_P(y)])^2 dy \\
 &\quad + \int_{-\infty}^{\infty} (\mathbb{E}_{q^*}[F_P^2(y)] - \mathbb{E}_{q^*}^2[F_P(y)]) dy, \tag{2.63}
 \end{aligned}$$

which is minimised for r^* such that

$$\mathbb{E}_{r^*}[F_P(y)] = \mathbb{E}_{q^*}[F_P(y)] \tag{2.64}$$

for all $y \in (-\infty, \infty)$. Because $r^* = q^*$ satisfies equation (2.64), S^* is a proper scoring rule (but not strictly proper, because equation (2.64) may be satisfied by a distribution of F_P other than q^*). Noting that under S , point-mass distributions underlying the forecast are permissible, the ideal distribution, δ_x^* , is defined as the point-mass distribution at $f(\delta_x) =$

$F_{\delta_x} = \Pr_{\delta_x}(X \leq \cdot) = \mathbf{1}_{[x, \infty)}(\cdot)$. Consequently,

$$\begin{aligned}
 S^*(\delta_x^*, F_p) &= \int_{-\infty}^{\infty} \left(\mathbb{E}_{\delta_x^*}[F_P(y)] - F_p(y) \right)^2 dy \\
 &= \int_{-\infty}^{\infty} (F_{\delta_x}(y) - F_p(y))^2 dy \\
 &= \int_{-\infty}^{\infty} \left(\mathbf{1}_{[x, \infty)}(y) - F_p(y) \right)^2 dy \\
 &= S(F_p, x)
 \end{aligned} \tag{2.65}$$

and S^* is a proper extended scoring rule for S . We have, from equation (2.63),

$$\begin{aligned}
 S^*(q^*, q^*) &= \int_{-\infty}^{\infty} \left(\mathbb{E}[F_P^2(y)] - \mathbb{E}^2[F_P(y)] \right) dy \\
 &= \int_{-\infty}^{\infty} \mathbb{E} \left[(F_P(y) - \mathbb{E}[F_P(y)])^2 \right] dy
 \end{aligned} \tag{2.66}$$

and for $x \in \mathbb{R}$,

$$S^*(q_x^*, q_x^*) = \int_{-\infty}^{\infty} \left(\mathbb{E}[F_P^2(y)|X = x] - \mathbb{E}^2[F_P(y)|X = x] \right) dy. \tag{2.67}$$

It is then readily shown (referring to equation (2.41)), that the RDC decomposition for the CRPS is

$$\begin{aligned}
 \mathbb{E}[S(F_P, X)] &= \\
 & \left. \begin{aligned}
 & \underbrace{\int_{-\infty}^{\infty} \mathbb{E} \left[(F_P(y) - \mathbb{E}[F_P(y)])^2 \right] dy}_{\text{Refinement}} \\
 & - \underbrace{\mathbb{E} \left[\int_{-\infty}^{\infty} (\mathbb{E}[F_P(y)|X] - \mathbb{E}[F_P(y)])^2 dy \right]}_{\text{Discrimination}}
 \end{aligned} \right\} \text{Excess} \\
 & \quad + \underbrace{\mathbb{E} \left[\int_{-\infty}^{\infty} (\mathbb{E}[F_P(y)|X] - \mathbf{1}_{[X, \infty)}(y))^2 dy \right]}_{\text{Correctness}}.
 \end{aligned} \tag{2.68}$$

2.6.2.3 | Other Decompositions

The decompositions in equations (2.61) and (2.68) are not unique: there are other decompositions for the ranked probability scoring rule. The most evident of these decompositions

is arrived at by noting that the CRPS can be expressed as the aggregate of a collection of Brier scores (Brier, 1950). From equation (2.57), the expected CRPS can be written as

$$\mathbb{E}[S(f(P), X)] = \int_{\mathcal{X}} \mathbb{E}[(\mathbb{E}_P[\mathbf{1}_{[X, \infty)}(y)] - \mathbf{1}_{[X, \infty)}(y))^2] dy \quad (2.69)$$

where, noting that for all $y \in \mathbb{R}$, $\mathbf{1}_{[X, \infty)}(y) \in \{0, 1\}$ is a binary observation with probability of being equal to 1 given by $\mathbb{E}_P[\mathbf{1}_{[X, \infty)}(y)]$. It is, therefore, the case that, writing S_B for the Brier scoring rule (Brier, 1950),

$$(\mathbb{E}_P[\mathbf{1}_{[X, \infty)}(y)] - \mathbf{1}_{[X, \infty)}(y))^2 = S_B(\mathbb{E}_P[\mathbf{1}_{[X, \infty)}(y)], \mathbf{1}_{[X, \infty)}(y)) \quad (2.70)$$

and the expected CRPS is

$$\mathbb{E}[S(f(P), X)] = \int_{\mathcal{X}} \mathbb{E}[S_B(\mathbb{E}_P[\mathbf{1}_{[X, \infty)}(y)], \mathbf{1}_{[X, \infty)}(y))] dy. \quad (2.71)$$

Each expected Brier score (Brier, 1950) has both URR and RDC decompositions, which when substituted for the integrand of equation (2.71) give the following alternative URR decomposition for the CRPS,

$$\begin{aligned} \mathbb{E}[S(f(P), X)] = & \left. \begin{aligned} & \underbrace{\int_{-\infty}^{\infty} \mathbb{E}[\mathbf{1}_{[X, \infty)}(y)](1 - \mathbb{E}[\mathbf{1}_{[X, \infty)}(y)]) dy}_{\text{Uncertainty}} \\ & - \underbrace{\int_{-\infty}^{\infty} \mathbb{E}[(\mathbb{E}[\mathbf{1}_{[X, \infty)}(y)] - \mathbb{E}[\mathbf{1}_{[X, \infty)}(y)|F_P(y)])^2] dy}_{\text{Resolution}} \end{aligned} \right\} \text{Sharpness} \\ & + \underbrace{\int_{-\infty}^{\infty} \mathbb{E}[(F_P(y) - \mathbb{E}[\mathbf{1}_{[X, \infty)}(y)|F_P(y)])^2] dy}_{\text{Reliability}} \end{aligned} \quad (2.72)$$

and alternative RDC decomposition,

$$\begin{aligned}
 \mathbb{E}[S(f(P), X)] = & \\
 & \left. \begin{aligned}
 & \underbrace{\int_{-\infty}^{\infty} \mathbb{E}[(F_P(y) - \mathbb{E}[F_P(y)])^2] dy}_{\text{Refinement}} \\
 & - \underbrace{\int_{-\infty}^{\infty} \mathbb{E}[(\mathbb{E}[F_P(y)|\mathbb{1}_{[X,\infty)}(y)] - \mathbb{E}[F_P(y)])^2] dy}_{\text{Discrimination}}
 \end{aligned} \right\} \text{Excess} \\
 & + \underbrace{\int_{-\infty}^{\infty} \mathbb{E}[(\mathbb{1}_{[X,\infty)}(y) - \mathbb{E}[F_P(y)|\mathbb{1}_{[X,\infty)}(y)])^2] dy}_{\text{Correctness}}. \quad (2.73)
 \end{aligned}$$

The URR decomposition of equation (2.72) has been derived by Candille and Talagrand (2005) (using the same approach, Candille and Talagrand (2005) also derive an alternative URR decomposition for the RPS, reproducing a result first arrived at by Murphy (1972)). We will refer to the decomposition of equation (2.72) as the URR_B decomposition to distinguish it from the URR decomposition of equation (2.61).

The singular difference between the URR and the URR_B decompositions is the conditioning variables (with uncertainty, therefore, being the same in both decompositions). The conditional expectations of the URR decomposition depend on the full distribution function (or, P), a stronger dependency than for the conditional expectations of the URR_B, which depend only on the distribution function at a point (i.e. $F_P(y)$). Under the weaker conditioning of the URR_B decomposition, a larger group of outcomes can be associated with each forecast, resulting, for each y , in a conditional expectation closer to $\mathbb{E}[\mathbb{1}_{[X,\infty)}(y)]$, giving a lower (i.e. poorer) resolution than under the URR decomposition; to maintain the same expected score, the reliability of the URR_B must then too be lower (i.e. better) than under the URR decomposition. Because the URR decomposition is determined by minimisation, the URR_B decomposition can be said to under-report resolution and overstate reliability.

The RDC decomposition of equation (2.73), which we call the RDC_B decomposition, too has weaker conditioning than the RDC decomposition of equation (2.67): for each y , many different outcomes will have the same value for $\mathbb{1}_{[x,\infty)}(y)$. Therefore, for any y , $\mathbb{E}[F_P(y)|\mathbb{1}_{[X,\infty)}(y)]$ will be closer to $\mathbb{E}[F_P(y)]$ than $\mathbb{E}[F_P(y)|X] \equiv F_{\mathbb{E}[P|X]}(y)$. It follows that the discrimination attribute of the RDC_B decomposition will be lower (i.e. worse) than the discrimination attribute of the RDC decomposition. And as both decompositions must equate to the same expected score, the RDC_B decomposition will have a lower (i.e. better) correctness attribute than the RDC decomposition. In other words, the RDC_B decomposition will give too low an assessment of discrimination and too good an evaluation of correctness. Refinement is the

same in both decompositions.

Still further URR decompositions have been proposed (for example for the RPS, [Murphy \(1972\)](#) proposes a scalar decomposition in addition to the URR_B decomposition; [Murphy \(1972\)](#) calls the URR_B decomposition the vector decomposition of the RPS). In the specific context of the CRPS, [Hersbach \(2000\)](#) gives a thoroughly different URR decomposition. Although [Hersbach \(2000\)](#)'s original presentation is in terms of ensemble forecasts, it is easily reframed in terms of probabilistic forecasts and such a reworking of [Hersbach \(2000\)](#)'s URR decomposition is given in the appendix to [Candille and Talagrand \(2005\)](#), where it is labelled the Hersbach-Lalauette URR decomposition.

[Hersbach \(2000\)](#)'s URR decomposition differs principally because of [Hersbach \(2000\)](#)'s interpretation of reliability, which stems from the type of forecasts considered i.e. ensemble forecasts. For ensemble forecasts, a measure of reliability is the uniformity of the rank histogram. Rank histograms were proposed by [Anderson \(1996\)](#), [Hamill and Colucci \(1997\)](#), [Talagrand et al. \(1997\)](#) (see also [Weigel, 2012](#); [Wilks, 2006](#)); [Hamill \(2001\)](#) shows that suitable (i.e. uniform, or, flat) rank histograms are a necessary condition of the reliability of ensemble forecasts but not a sufficient condition and this has led to other assessments of ensemble forecast reliability (see, for example [Mason et al., 2007](#)) and the discussions [Bröcker et al. \(2011\)](#) and [Mason et al. \(2011\)](#) (see also [Weigel, 2012](#)); however, we do not review the properties of rank histograms here, and mention them only to place in context the motivation for [Hersbach \(2000\)](#)'s decomposition.

To translate [Hersbach \(2000\)](#)'s arguments to probabilistic forecasts, suppose that an independent and identically distributed sample is drawn from the forecast (a probability distribution). The members of this sample together with the observation can be ranked (in ascending order). The forecast may then be considered reliable if the probability that the observation has rank i , is the same for all i : in this sense, the observation is (statistically) indistinguishable from any other random variable with the same probability distribution as was forecast. In other words, here reliability refers to the *position* of the observation in the sample. To quantify this concept of reliability, we note that if each rank of the observation is equally likely, then the probability distribution of the observation's rank is a uniform distribution (or, equivalently, the cumulative distribution of the observation's rank is a straight line with unit slope). What amounts to the same assessment is that (for large sample sizes), the observation is equally likely to fall on the α -quantile for all values of α . Therefore, if we were to construct a histogram with class boundaries set to the different values of $\alpha \in [0, 1]$ and determine the frequency with which the observation fell between the α -quantiles (with respect to the sample from the forecast), i.e. the frequency with which the observation had rank equal to α , then the resulting frequency histogram would be flat for reliable forecasts.

In this histogram construction, we note that the classes have the same width (equal to $d\alpha$) so the density and frequency histogram are equivalent.

Rather than examining the reliability criterion in position-(or, rank-)space of the observation, we can assess this same criterion in the value-space of the observation by transforming the histogram. To do so, the α -axis is mapped to the range of the observation and the corresponding density graph expressed against these values. But, the equally-spaced classes of α will map, in general, to unequally-spaced classes on the range of the observation (i.e. α -quantiles, for different α , are not necessarily equally-spaced). So to maintain the transformed densities as densities, they must be divided by the ratio of the new class width to the previous class width. Having completed the transformation, the new densities can be compared to the similarly transformed uniform densities: an aggregate measure of their difference (under [Hersbach \(2000\)](#) the squared difference), gives a measure of reliability as expressed in [Hersbach \(2000\)](#)'s decomposition.

Being of a fundamentally different perspective, [Hersbach \(2000\)](#)'s URR decomposition does not bear direct comparison to our URR decomposition (nor to the URR_B decomposition). It remains open how [Hersbach \(2000\)](#)'s approach could lead to a comparable RDC decomposition.

2.6.3 || Ignorance Scoring Rule

Following the work of [Bernado \(1979\)](#) and [Benedetti \(2010\)](#), the ignorance scoring rule for precise-probabilistic forecasts has become a strong alternative to the quadratic scoring rule(s). The ignorance scoring rule was first proposed by [Good \(1952\)](#) for binary events and later re-introduced by [Winkler and Murphy \(1968\)](#) (as the logarithmic scoring rule) and, more recently, by [Roulston and Smith \(2002\)](#) for observations that can take a discrete number of values. Its most general form, for continuous observations, is given by [Gneiting et al. \(2005\)](#). The ignorance scoring rule is proper (see, for example [Gneiting and Raftery, 2007](#)). Several extensions to the ignorance scoring rule have also been proposed, for example, the ranked divergence/ignorance score of [Weijts et al. \(2010\)](#) and the continuous ranked ignorance score of [Tödter and Ahrens \(2012\)](#).

For each $p \in \mathcal{P}$, let $\varphi_p(\cdot)$ be the probability density function (with respect to an appropriate measure, for example, the Lebesgue measure for X continuous) associated with p ; write $\varphi_p(\cdot|\cdot)$ for a conditional probability density. Set $f(p) = \varphi_p(\cdot)$. The ignorance scoring rule is

defined by

$$S(f(p), x) = -\log \varphi_p(x). \quad (2.74)$$

Here ‘log’ indicates the natural logarithm (although, logarithm to the base 2 and the common logarithm have also been used; such details will not affect the results that follow).

2.6.3.1 | URR Decomposition

Let $r \in \mathcal{P}$. We then have (with the measure implicit in the term dx),

$$S(p, r) = - \int_{\mathcal{X}} \varphi_r(x) \log \varphi_p(x) dx. \quad (2.75)$$

If we let q be the unconditional (i.e. climatological) probability distribution of X , then it is immediate from equation (2.75) that

$$S(q, q) = - \int_{\mathcal{X}} \varphi_q(x) \log \varphi_q(x) dx \quad (2.76)$$

and

$$S(q_p, q_p) = - \int_{\mathcal{X}} \varphi_q(x|p) \log \varphi_q(x|p) dx. \quad (2.77)$$

Noting that $\varphi_q(x) = \mathbb{E}[\varphi_q(x|P)]$, a little algebra shows that from equation (2.32) the URR decomposition of the ignorance scoring rule is

$$\begin{aligned} \mathbb{E}[S(P, X)] = & \left. \begin{aligned} & - \int_{\mathcal{X}} \varphi_q(x) \log \varphi_q(x) dx \\ & - \mathbb{E} \left[\underbrace{\int_{\mathcal{X}} \log \left(\frac{\varphi_q(x|P)}{\varphi_q(x)} \right) \varphi_q(x|P) dx}_{\text{Resolution}} \right] \end{aligned} \right\} \text{Sharpness} \\ & + \underbrace{\mathbb{E} \left[\int_{\mathcal{X}} \log \left(\frac{\varphi_q(x|P)}{\varphi_P(x)} \right) \varphi_q(x|P) dx \right]}_{\text{Reliability}}. \end{aligned} \quad (2.78)$$

This URR decomposition matches that given by [Weijs et al. \(2010\)](#) and [Tödter and Ahrens \(2012\)](#).

2.6.3.2 | RDC Decomposition

For $p^* \in \mathcal{P}^*$, let $\psi_{p^*}(\cdot)$ be the probability density function for p^* with respect to a suitable measure. Define the scoring rule S^* by

$$S^*(r^*, p) = -\log \psi_{r^*}(p). \quad (2.79)$$

Then, we can write (subsuming the measure in the term dp)

$$\begin{aligned} S^*(r^*, q^*) &= \mathbb{E}_{q^*}[S^*(r^*, P)] \\ &= -\int_{\mathcal{P}} \psi_{q^*}(p) \log \psi_{r^*}(p) dp \\ &\stackrel{\text{def}}{=} \phi_{\psi_{q^*}}(\psi_{r^*}) \end{aligned} \quad (2.80)$$

To consider the propriety of S^* , define the function. We can show the propriety of S^* by demonstrating that

$$\psi_{q^*} = \arg \min_{\psi_{r^*}} \phi_{\psi_{q^*}}(\psi_{r^*}) \quad \text{subject to} \quad \int_{\mathcal{P}} \psi_{r^*}(p) dp = 1. \quad (2.81)$$

To prove that (2.81) holds, form the Lagrangian

$$L(h) = \phi_g(h) + \lambda \left(1 - \int_{\mathcal{P}} h(p) dp\right); \quad \lambda \in \mathbb{R} \quad (2.82)$$

on the linear-space of functions on \mathcal{P} , with inner-product

$$\langle g, h \rangle = \int_{\mathcal{P}} g(p)h(p) dp. \quad (2.83)$$

We have (see, for example [Luenberger, 1969](#)), for all h ,

$$\begin{aligned} \langle \nabla_h L(h), h \rangle &= \left. \frac{d}{d\alpha} L(h + \alpha h) \right|_{\alpha=0} \\ &= -\int_{\mathcal{P}} \left(\frac{g(p)}{h(p)} + \lambda \right) h(p) dp \\ &= \langle -(g/h) - \lambda, h \rangle \end{aligned} \quad (2.84)$$

so that $\nabla_h L(h) = -(g/h) - \lambda$. A necessary condition that g_{\min} be a minimum of $L(h)$ is $\nabla_h L(h)|_{h=g_{\min}} = 0$ i.e. $-g = \lambda g_{\min}$. But, g_{\min} is a value for h and so satisfies the constraint $\int_{\mathcal{P}} g_{\min}(p) dp = 1$, and we have

$$-1 = -\int_{\mathcal{P}} g(p) dp = \lambda \int_{\mathcal{P}} g_{\min}(p) dp = \lambda \quad (2.85)$$

giving $g_{\min} = g$ i.e. equation (2.81) holds and S^* is proper.

As a proper scoring rule on $\mathcal{P}^* \times \mathcal{F}_f$, S^* may be used to determine measures of refinement and discrimination. From equation (2.80), we have

$$S^*(q^*, q^*) = - \int_{\mathcal{P}} \psi_{q^*}(p) \log \psi_{q^*}(p) dp, \quad (2.86)$$

which gives refinement, and, for each $x \in \mathcal{X}$,

$$S^*(q_x^*, q_x^*) = - \int_{\mathcal{P}} \psi_{q_x^*}(p|x) \log \psi_{q_x^*}(p|x) dp, \quad (2.87)$$

so that discrimination is equal to

$$\mathbb{E} \left[\int_{\mathcal{P}} \psi_{q^*}(p|X) \log \left(\frac{\psi_{q^*}(p|X)}{\psi_{q^*}(p)} \right) dp \right]. \quad (2.88)$$

We note that the discrimination term (2.88) is precisely the discrimination quantity put forward by Bröcker (2015, equation (10)).

However, the scoring rule S^* is *not* an extension of S , as δ_x^* places all probability on the point-mass δ_x so that

$$\begin{aligned} S^*(\delta_x^*, p) &= - \log \psi_{\delta_x^*}(p) \\ &\neq - \log \varphi_p(x) \quad \text{unless } p = \delta_x. \end{aligned} \quad (2.89)$$

It is not necessarily the case, therefore, that discrimination (2.88) is comparable with the resolution term of equation (2.78). Moreover, while we can *define* (in)correctness as

$$\mathbb{E}[S(f(P), X)] - \mathbb{E}[S^*(q_X^*, q_X^*)], \quad (2.90)$$

which here simplifies to

$$\mathbb{E} \left[- \int_{\mathcal{P}} \psi_{q^*}(p|X) \log \left(\frac{\varphi_p(X)}{\psi_{q^*}(p|X)} \right) dp \right] \quad (2.91)$$

this measure may not be a faithful guide to the correctness of the forecasts under S , and may not be non-negative.

However, Bröcker (2015) has shown that discrimination (2.89) is commensurable with resolution in equation (2.78), which suggests that correctness (2.91) is a reasonable reflection of the correctness of the forecasts.

While a proper extended scoring rule would give a complete RDC decomposition, there is no clear proper extended scoring rule for the ignorance scoring rule (2.73).

2.6.4 || α -Quantile Scoring Rule

Fix $\alpha \in (0, 1)$. The α -quantile under probability distribution $p \in \mathcal{P}$ of the observation, X , taking values in \mathbb{R} , is defined by

$$F_p^{-1}(\alpha) = \inf\{y \in \mathbb{R} | \alpha \leq F_p(y)\}. \quad (2.92)$$

Often the infimum is obtained, in which case $F_p^{-1}(\alpha)$ is that value x_α such that $\alpha = F_p(x_\alpha)$. An α -quantile forecast is a *point* forecast: $f(p) = F_p^{-1}(\alpha) \in \mathcal{X}$. We follow [Gneiting \(2011a\)](#) and define the (negatively-oriented) α -quantile scoring rule by

$$\begin{aligned} S(f(p), x) &= S(F_p^{-1}(\alpha), x) \\ &= (F_p^{-1}(\alpha) - x)(\mathbf{1}_{[x, \infty)}(F_p^{-1}(\alpha)) - \alpha) \\ &= \begin{cases} \alpha(x - F_p^{-1}(\alpha)) & \text{if } F_p^{-1}(\alpha) \leq x, \\ (1 - \alpha)(F_p^{-1}(\alpha) - x) & \text{if } F_p^{-1}(\alpha) \geq x. \end{cases} \end{aligned} \quad (2.93)$$

The propriety of the α -quantile scoring rule is demonstrated in [Gneiting and Raftery \(2007\)](#) (see also the appendix of [Friederichs and Hense \(2008\)](#)).

2.6.4.1 | URR Decomposition

For any $r \in \mathcal{P}$,

$$S(F_p^{-1}(\alpha), r) = \alpha(\mathbb{E}_r[X] - F_p^{-1}(\alpha)) + F_p^{-1}(\alpha)F_r(F_p^{-1}(\alpha)) - \mathbb{E}_r[X\mathbf{1}_{[X, \infty)}(F_p^{-1}(\alpha))]. \quad (2.94)$$

Two instances of equation (2.94) will be of importance. Fixing q to be the unconditional (i.e. climatological) probability distribution of X ,

$$S(F_q^{-1}(\alpha), q) = \alpha\mathbb{E}[X] - \mathbb{E}[X\mathbf{1}_{[X, \infty)}(F_q^{-1}(\alpha))] \quad (2.95)$$

and for $q_{F_p^{-1}(\alpha)}$ the conditional probability distribution of X given the forecast $F_p^{-1}(\alpha)$,

$$S(f(q_{F_p^{-1}(\alpha)}), q_{F_p^{-1}(\alpha)}) = \alpha\mathbb{E}[X|F_p^{-1}(\alpha)] - \mathbb{E}[X\mathbf{1}_{[X, \infty)}(F_q^{-1}(\alpha|F_p^{-1}(\alpha))|F_p^{-1}(\alpha)]. \quad (2.96)$$

Using equations (2.95) and (2.96) in equation (2.32), the URR decomposition for the α -quantile scoring rule is

$$\begin{aligned}
 \mathbb{E}[S(F_P^{-1}(\alpha), X)] = & \\
 & \left. \begin{aligned} & \underbrace{\alpha \mathbb{E}[X] - \mathbb{E}[X \mathbb{1}_{[X, \infty)}(F_q^{-1}(\alpha))]}_{\text{Uncertainty}} \\ & - \underbrace{\mathbb{E} \left[\mathbb{E}[X \mathbb{1}_{[X, \infty)}(F_q^{-1}(\alpha) | F_P^{-1}(\alpha)) | F_P^{-1}(\alpha)] - \mathbb{E}[X \mathbb{1}_{[X, \infty)}(F_q^{-1}(\alpha))] \right]}_{\text{Resolution}} \end{aligned} \right\} \text{Sharpness} \\
 + & \left. \begin{aligned} & \left(\mathbb{E} \left[F_P^{-1}(\alpha) \left(F_q(F_P^{-1}(\alpha) | F_P^{-1}(\alpha)) - \alpha \right) \right] \right. \\ & \left. + \mathbb{E} \left[\mathbb{E} \left[X \left(\mathbb{1}_{[X, \infty)}(F_q^{-1}(\alpha) | F_P^{-1}(\alpha)) - \mathbb{1}_{[X, \infty)}(F_P^{-1}(\alpha)) \right) | F_P^{-1}(\alpha) \right] \right] \right) \end{aligned} \right\} \text{Reliability.}
 \end{aligned} \tag{2.97}$$

Some algebra shows that equation (2.97) agrees with the URR decomposition of [Bentzien and Friederichs \(2014\)](#). Note too, that an α -quantile forecast is perfectly reliable if $F_p^{-1}(\alpha) = F_q^{-1}(\alpha | F_p^{-1}(\alpha))$ for all $p \in \mathcal{P}$.

2.6.4.2 | RDC Decomposition

Define the scoring rule S^* by

$$S^*(r^*, F_p^{-1}(\alpha)) = \begin{cases} (1 - \alpha)(F_p^{-1}(\alpha) - F_{r^*}^{-1}(1 - \alpha)) & \text{if } F_{r^*}^{-1}(1 - \alpha) \leq F_p^{-1}(\alpha), \\ \alpha(F_{r^*}^{-1}(1 - \alpha) - F_p^{-1}(\alpha)) & \text{if } F_{r^*}^{-1}(1 - \alpha) \geq F_p^{-1}(\alpha). \end{cases} \tag{2.98}$$

Note that for each $r^* \in \mathcal{P}^*$, $F_{r^*}^{-1}(1 - \alpha)$ is a value for $F_P^{-1}(\alpha)$ (for some value of P) and is, therefore, also in \mathbb{R} .

For S^* to be proper it is required that for all $r^* \in \mathcal{P}^*$, $S^*(q^*, q^*) \leq S^*(r^*, q^*)$. To show this, we adapt the argument of [Gneiting and Raftery \(2007\)](#). First, for $r^* \in \mathcal{P}^*$,

$$\begin{aligned}
 S^*(r^*, q^*) & \\
 & = \mathbb{E}_{q^*}[S^*(r^*, F_P^{-1}(\alpha))] \\
 & = \mathbb{E}_{q^*} \left[F_P^{-1}(\alpha) \left\{ (1 - \alpha) \mathbb{1}_{[F_{r^*}^{-1}(1 - \alpha), \infty)}(F_P^{-1}(\alpha)) - \alpha \mathbb{1}_{(-\infty, F_{r^*}^{-1}(1 - \alpha)]}(F_P^{-1}(\alpha)) \right\} \right. \\
 & \quad \left. + F_{r^*}^{-1}(1 - \alpha) \left\{ \alpha \mathbb{1}_{(-\infty, F_{r^*}^{-1}(1 - \alpha)]}(F_P^{-1}(\alpha)) - (1 - \alpha) \mathbb{1}_{[F_{r^*}^{-1}(1 - \alpha), \infty)}(F_P^{-1}(\alpha)) \right\} \right]. \tag{2.99}
 \end{aligned}$$

Because $F_{q^*}(F_{q^*}^{-1}(1 - \alpha)) = 1 - \alpha$,

$$\begin{aligned}
& S^*(r^*, q^*) - S^*(q^*, q^*) \\
&= \mathbb{E}_{q^*} \left[F_P^{-1}(\alpha) \left\{ (1 - \alpha) \left(\mathbb{1}_{[F_{r^*}^{-1}(1-\alpha), \infty)}(F_P^{-1}(\alpha)) - \mathbb{1}_{[F_{q^*}^{-1}(1-\alpha), \infty)}(F_P^{-1}(\alpha)) \right) \right. \right. \\
&\quad \left. \left. - \alpha \left(\mathbb{1}_{(-\infty, F_{r^*}^{-1}(1-\alpha)]}(F_P^{-1}(\alpha)) - \mathbb{1}_{(-\infty, F_{q^*}^{-1}(1-\alpha)]}(F_P^{-1}(\alpha)) \right) \right\} \right] \\
&\quad + F_{r^*}^{-1}(1 - \alpha)(\alpha - 1) + F_{r^*}^{-1}(1 - \alpha)F_{q^*}(F_{r^*}^{-1}(1 - \alpha)). \quad (2.100)
\end{aligned}$$

Consider two cases:

Case (i): $F_{q^*}^{-1}(1 - \alpha) > F_{r^*}^{-1}(1 - \alpha)$.

$$\begin{aligned}
& S^*(r^*, q^*) - S^*(q^*, q^*) \\
&= \mathbb{E}_{q^*} \left[F_P^{-1}(\alpha) \left\{ (1 - \alpha) \mathbb{1}_{[F_{r^*}^{-1}(1-\alpha), F_{q^*}^{-1}(1-\alpha)]}(F_P^{-1}(\alpha)) \right. \right. \\
&\quad \left. \left. + \alpha \mathbb{1}_{[F_{r^*}^{-1}(1-\alpha), F_{q^*}^{-1}(1-\alpha)]}(F_P^{-1}(\alpha)) \right\} \right] \\
&\quad + F_{r^*}^{-1}(1 - \alpha)(\alpha - 1) + F_{r^*}^{-1}(1 - \alpha)F_{q^*}(F_{r^*}^{-1}(1 - \alpha)) \\
&= \mathbb{E}_{q^*} \left[F_P^{-1}(\alpha) \mathbb{1}_{[F_{r^*}^{-1}(1-\alpha), F_{q^*}^{-1}(1-\alpha)]}(F_P^{-1}(\alpha)) \right] \\
&\quad + F_{r^*}^{-1}(1 - \alpha)(\alpha - 1) + F_{r^*}^{-1}(1 - \alpha)F_{q^*}(F_{r^*}^{-1}(1 - \alpha)) \\
&\geq F_{r^*}^{-1}(1 - \alpha) \left(F_{q^*}(F_{q^*}^{-1}(1 - \alpha)) - F_{q^*}(F_{r^*}^{-1}(1 - \alpha)) \right) \\
&\quad + F_{r^*}^{-1}(1 - \alpha)(\alpha - 1) + F_{r^*}^{-1}(1 - \alpha)F_{q^*}(F_{r^*}^{-1}(1 - \alpha)) \\
&= F_{r^*}^{-1}(1 - \alpha)(1 - \alpha) - F_{r^*}^{-1}(1 - \alpha)F_{q^*}(F_{r^*}^{-1}(1 - \alpha)) \\
&\quad + F_{r^*}^{-1}(1 - \alpha)(\alpha - 1) + F_{r^*}^{-1}(1 - \alpha)F_{q^*}(F_{r^*}^{-1}(1 - \alpha)) \\
&= 0.
\end{aligned} \tag{2.101}$$

Case (ii): $F_{q^*}^{-1}(1 - \alpha) < F_{r^*}^{-1}(1 - \alpha)$.

$$\begin{aligned}
& S^*(r^*, q^*) - S^*(q^*, q^*) \\
&= \mathbb{E}_{q^*} \left[F_P^{-1}(\alpha) \left\{ (1 - \alpha) \mathbb{1}_{[F_{q^*}^{-1}(1-\alpha), F_{r^*}^{-1}(1-\alpha))}(F_P^{-1}(\alpha)) \right. \right. \\
&\quad \left. \left. + \alpha \mathbb{1}_{[F_{q^*}^{-1}(1-\alpha), F_{r^*}^{-1}(1-\alpha))}(F_P^{-1}(\alpha)) \right\} \right] \\
&\quad + F_{r^*}^{-1}(1 - \alpha)(\alpha - 1) + F_{r^*}^{-1}(1 - \alpha)F_{q^*}(F_{r^*}^{-1}(1 - \alpha)) \\
&= \mathbb{E}_{q^*} \left[F_P^{-1}(\alpha) \mathbb{1}_{[F_{q^*}^{-1}(1-\alpha), F_{r^*}^{-1}(1-\alpha))}(F_P^{-1}(\alpha)) \right] \\
&\quad + F_{r^*}^{-1}(1 - \alpha)(\alpha - 1) + F_{r^*}^{-1}(1 - \alpha)F_{q^*}(F_{r^*}^{-1}(1 - \alpha)) \\
&\geq F_{q^*}^{-1}(1 - \alpha) \left(F_{q^*}(F_{r^*}^{-1}(1 - \alpha)) - F_{q^*}(F_{q^*}^{-1}(1 - \alpha)) \right) \\
&\quad + F_{r^*}^{-1}(1 - \alpha)(\alpha - 1) + F_{r^*}^{-1}(1 - \alpha)F_{q^*}(F_{r^*}^{-1}(1 - \alpha)) \\
&= \left(F_{q^*}^{-1}(1 - \alpha) + F_{r^*}^{-1}(1 - \alpha) \right) \left(F_{q^*}(F_{r^*}^{-1}(1 - \alpha)) + \alpha - 1 \right) \\
&\geq \left(F_{q^*}^{-1}(1 - \alpha) + F_{r^*}^{-1}(1 - \alpha) \right) (1 - \alpha + \alpha - 1) \\
&= 0
\end{aligned} \tag{2.102}$$

where we have used the fact that $F_{q^*}(\cdot)$ is, as a distribution function, non-decreasing so $F_{r^*}^{-1}(1 - \alpha) > F_{q^*}^{-1}(1 - \alpha)$ gives $F_{q^*}(F_{r^*}^{-1}(1 - \alpha)) \geq F_{q^*}(F_{q^*}^{-1}(1 - \alpha)) = 1 - \alpha$.

In summary, for $F_{r^*}^{-1}(1 - \alpha) \neq F_{q^*}^{-1}(1 - \alpha)$ i.e. $r^* \neq q^*$, $S^*(r^*, q^*) \geq S^*(q^*, q^*)$, as required.

Further, it is the case under the α -quantile scoring rule that a zero probability for some values of X is permissible. Consequently, we define the ideal distribution of the forecasts given $X = x$, δ_x^* as the point-mass distribution at the forecast $f(\delta_x) = F_{\delta_x}^{-1}(\alpha) = x$ for $\alpha \in (0, 1)$. Then $F_{\delta_x^*}^{-1}(1 - \alpha) = F_{\delta_x}^{-1}(\alpha) = x$ too. It follows that

$$\begin{aligned}
S^*(\delta_x^*, F_p^{-1}(\alpha)) &= \begin{cases} (1 - \alpha)(F_p^{-1}(\alpha) - F_{\delta_x^*}^{-1}(1 - \alpha)) & \text{if } F_{\delta_x^*}^{-1}(1 - \alpha) \leq F_p^{-1}(\alpha) \\ \alpha(F_{\delta_x^*}^{-1}(1 - \alpha) - F_p^{-1}(\alpha)) & \text{if } F_{\delta_x^*}^{-1}(1 - \alpha) \geq F_p^{-1}(\alpha) \end{cases} \\
&= \begin{cases} (1 - \alpha)(F_p^{-1}(\alpha) - x) & \text{if } x \leq F_p^{-1}(\alpha) \\ \alpha(x - F_p^{-1}(\alpha)) & \text{if } x \geq F_p^{-1}(\alpha) \end{cases} \\
&= S(F_p^{-1}(\alpha), x).
\end{aligned} \tag{2.103}$$

From the propriety of S^* and equation (2.103), S^* is a proper extended scoring rule for S .

To determine the attributes of the RDC decomposition, two particular instances of equation (2.99) are important:

$$\begin{aligned}
S^*(q^*, q^*) &= (1 - \alpha) \mathbb{E} \left[F_P^{-1}(\alpha) \right] - \mathbb{E} \left[F_P^{-1}(\alpha) \mathbf{1}_{(-\infty, F_{q^*}^{-1}(1-\alpha)]} (F_P^{-1}(\alpha)) \right] \\
&= (1 - \alpha) \mathbb{E} \left[F_P^{-1}(\alpha) \right] - \mathbb{E} \left[F_P^{-1}(\alpha) \mathbf{1}_{[F_P^{-1}(\alpha), \infty)} (F_{q^*}^{-1}(1 - \alpha)) \right]
\end{aligned} \tag{2.104}$$

and, for $x \in \mathbb{R}$,

$$\begin{aligned}
S^*(q_x^*, q_x^*) &= (1 - \alpha) \mathbb{E} \left[F_P^{-1}(\alpha) | X = x \right] \\
&\quad - \mathbb{E} \left[F_P^{-1}(\alpha) \mathbf{1}_{[F_P^{-1}(\alpha), \infty)} (F_{q^*}^{-1}(1 - \alpha | X = x)) | X = x \right].
\end{aligned} \tag{2.105}$$

From equations (2.104) and (2.105), using equation (2.41) we obtain after some simplification, the RDC decomposition

$$\begin{aligned}
\mathbb{E}[S(F_P^{-1}(\alpha), X)] &= \\
&\quad \underbrace{(1 - \alpha) \mathbb{E} \left[F_P^{-1}(\alpha) \right] - \mathbb{E} \left[F_P^{-1}(\alpha) \mathbf{1}_{[F_P^{-1}(\alpha), \infty)} (F_{q^*}^{-1}(1 - \alpha)) \right]}_{\text{Refinement}} \left. \vphantom{\mathbb{E}[S(F_P^{-1}(\alpha), X)]} \right\} \text{Excess} \\
&\quad - \left(\mathbb{E} \left[\mathbb{E} \left[F_P^{-1}(\alpha) \mathbf{1}_{[F_P^{-1}(\alpha), \infty)} (F_{q^*}^{-1}(1 - \alpha | X)) \right] | X \right] \right. \\
&\quad \quad \left. - \mathbb{E} \left[F_P^{-1}(\alpha) \mathbf{1}_{[F_P^{-1}(\alpha), \infty)} (F_{q^*}^{-1}(1 - \alpha)) | X \right] \right) \left. \vphantom{\mathbb{E}[S(F_P^{-1}(\alpha), X)]} \right\} \text{Discrimination} \\
&\quad + \left(\mathbb{E} [X \{ F_{q^*}(X | X) - (1 - \alpha) \}] \right. \\
&\quad \quad \left. + \mathbb{E} \left[\mathbb{E} \left[F_P^{-1}(\alpha) \left\{ \mathbf{1}_{[F_P^{-1}(\alpha), \infty)} (F_{q^*}^{-1}(1 - \alpha | X)) - \mathbf{1}_{[F_P^{-1}(\alpha), \infty)} (X) \right\} | X \right] \right] \right) \left. \vphantom{\mathbb{E}[S(F_P^{-1}(\alpha), X)]} \right\} \text{Correctness}
\end{aligned} \tag{2.106}$$

[2.7] Computing the Decompositions

Analytical formulae for the terms of the URR and RDC decompositions for several examples are given in section 2.6. Computational formulae for the decompositional attributes are also required if they are to be estimated. We consider a common forecasting scheme in which the observation, X , takes values on the real line, and each probabilistic forecast, P , is a distribution function for X . The CRPS is then a suitable scoring rule (see equation (2.56)). We present expressions, based on the empirical distribution function, that can be used to calculate the attributes of the CRPS. We choose the empirical distribution as the basis for our computational formulae because it is usual in applied forecast verification to adopt empirical distributions when evaluating the expected score (see, for example, [Brier \(1950\)](#), [Ferro and Fricker \(2012\)](#) and [Bröcker \(2012, Section 7.5.3\)](#)). For the circumstances

just described, the wide range of outcomes and the large choice of probabilistic forecasts has the consequence that in any sample of forecast-outcome pairs it is rare for values of either forecast or outcome to be repeated. This makes the computation of the attributes of the decompositions unstable as we now illustrate with a toy model (for similar models, see [Weigel and Bowler \(2009\)](#) and [Bouallègue et al. \(2015\)](#)).

In the toy model, the forecast for X is the distribution $P = \mathcal{N}(\Theta, \xi^2)$, i.e. a Gaussian/normal distribution with mean Θ and variance ξ^2 , where ξ is a fixed, predetermined value and Θ is a random variable such that X and Θ have joint distribution,

$$(X, \Theta)^\top \sim \mathcal{N}_2((\mu_X, \mu_\Theta)^\top, (\sigma_X^2, \sigma_\Theta^2)^\top; \rho) \quad (2.107)$$

a bi-variate normal distribution with mean vector $(\mu_X, \mu_\Theta)^\top$, and variance-covariance matrix

$$\begin{pmatrix} \sigma_X^2 & \rho\sigma_X\sigma_\Theta \\ \rho\sigma_X\sigma_\Theta & \sigma_\Theta^2 \end{pmatrix}. \quad (2.108)$$

We have, therefore (see, for example [Mood et al., 1974](#)) that

$$\begin{aligned} X &\sim \mathcal{N}(\mu_X, \sigma_X^2) \\ \Theta &\sim \mathcal{N}(\mu_\Theta, \sigma_\Theta^2) \\ X|_{\Theta=\theta} &\sim \mathcal{N}(\mu_X + (\rho\sigma_X/\sigma_\Theta)(\theta - \mu_\Theta), \sigma_X^2(1 - \rho^2)) \\ \Theta|_{X=x} &\sim \mathcal{N}(\mu_\Theta + (\rho\sigma_\Theta/\sigma_X)(x - \mu_X), \sigma_\Theta^2(1 - \rho^2)). \end{aligned} \quad (2.109)$$

Although the model is Gaussian, it is not unnecessarily restrictive, for as [Weigel and Bowler \(2009\)](#) note “given that skewed distributions can be normalized by appropriate transformations, the toy model . . . can in principle be generalized to non-normal variables”.

The attributes of the URR and RDC decompositions of the CRPS can, under the toy model, be evaluated analytically by the formulae in (2.110) and (2.111) ($\Phi(z)$ is the cumulative standard normal distribution function at z , $\phi(\cdot)$ is the density function of the standard normal distribution; the following abbreviations are made: UNC is uncertainty, RES is resolution, REL is reliability, REF is refinement, DIS is discrimination, COR is correctness, SCR is the expected score).

$$\begin{aligned}
\text{UNC} &= \int_{-\infty}^{\infty} a(y)(1 - a(y)) dy \\
\text{RES} &= \int_{-\infty}^{\infty} \left\{ \int_{-\infty}^{\infty} (b(y, w) - a(y))^2 \phi(w) dw \right\} dy \\
\text{REL} &= \int_{-\infty}^{\infty} \left\{ \int_{-\infty}^{\infty} (b(y, w) - c(y, w))^2 \phi(w) dw \right\} dy \\
\text{REF} &= \int_{-\infty}^{\infty} \left\{ \int_{-\infty}^{\infty} \left(c(y, w) - \left[\int_{-\infty}^{\infty} c(y, v) \phi(v) dv \right] \right)^2 \phi(w) dw \right\} dy \\
\text{DIS} &= \int_{-\infty}^{\infty} \left\{ \int_{-\infty}^{\infty} \left[\int_{-\infty}^{\infty} d(y, u, z) \phi(u) du - \int_{-\infty}^{\infty} c(y, v) \phi(v) dv \right]^2 \phi(z) dz \right\} dy \\
\text{COR} &= \int_{-\infty}^{\infty} \left\{ \int_{-\infty}^{\infty} \left[\int_{-\infty}^{\infty} d(y, u, z) \phi(u) du - \mathbf{1}_{[z\sigma_X + \mu_X, \infty)}(y) \right]^2 \phi(z) dz \right\} dy \\
\text{SCR} &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \left(d(y, u, z) - \mathbf{1}_{[z\sigma_X + \mu_X, \infty)}(y) \right)^2 \phi(u) du \phi(z) dz dy
\end{aligned} \tag{2.110}$$

where

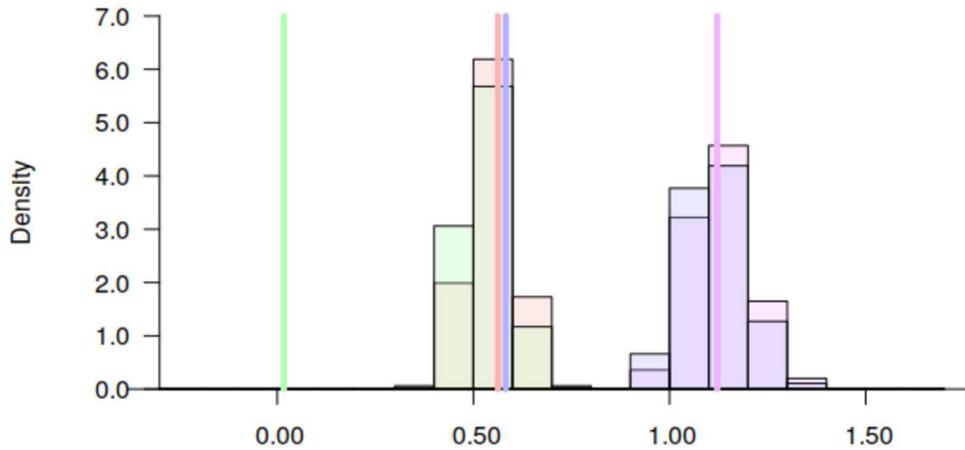
$$\begin{aligned}
a(y) &= \Phi \left(\frac{y - \mu_X}{\sigma_X} \right) \\
b(y, w) &= \Phi \left(\frac{y - (\mu_X + \rho\sigma_X w)}{\sigma_X \sqrt{1 - \rho^2}} \right) \\
c(y, w) &= \Phi \left(\frac{y - (w\sigma_\Theta + \mu_\Theta)}{\xi} \right) \\
d(y, w, z) &= \Phi \left(\frac{y - [(w\sqrt{1 - \rho^2} + \rho z)\sigma_\Theta + \mu_\Theta]}{\xi} \right)
\end{aligned} \tag{2.111}$$

With ξ fixed, each forecast is uniquely determined by the value of Θ . Therefore, in a sample of size T of forecast-outcome pairs, each outcome, x_t , is preceded by a forecast given as a value for Θ , θ_t , for $t = 1, \dots, T$. Under the assumption of empirical distributions for the joint and marginal distributions of the outcomes and forecasts, with a little work it can be shown that the expectations in equations (2.61) and (2.68) of the URR and RDC decompositions of the CRPS, reduce to the estimators given in (2.112).

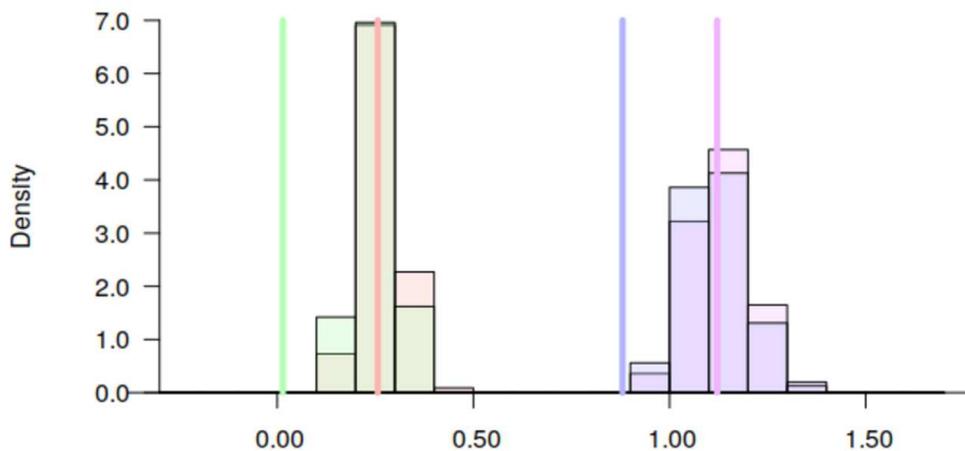
$$\begin{aligned}
\widehat{\text{SCR}} &= \frac{1}{T} \sum_{r=1}^T \int_{-\infty}^{\infty} \left(\Phi \left(\frac{y - \theta_r}{\xi} \right) - \mathbb{1}_{[x_r, \infty)}(y) \right)^2 dy \\
\widehat{\text{UNC}} &= \int_{-\infty}^{\infty} \left(\frac{1}{T} \sum_{t=1}^T \mathbb{1}_{[x_t, \infty)}(y) \right) \left(1 - \frac{1}{T} \sum_{r=1}^T \mathbb{1}_{[x_r, \infty)}(y) \right) dy \\
\widehat{\text{RES}} &= \frac{1}{T} \sum_{r=1}^T \int_{-\infty}^{\infty} \left(\frac{\sum_{t=1}^T \mathbb{1}_{[x_t, \infty)}(y) \mathbb{1}_{\{\theta_t\}}(\theta_r)}{\sum_{u=1}^T \mathbb{1}_{\{\theta_u\}}(\theta_r)} - \frac{1}{T} \sum_{t=1}^T \mathbb{1}_{[x_t, \infty)}(y) \right)^2 dy \\
\widehat{\text{REL}} &= \frac{1}{T} \sum_{r=1}^T \int_{-\infty}^{\infty} \left(\frac{\sum_{t=1}^T \mathbb{1}_{[x_t, \infty)}(y) \mathbb{1}_{\{\theta_t\}}(\theta_r)}{\sum_{u=1}^T \mathbb{1}_{\{\theta_u\}}(\theta_r)} - \Phi \left(\frac{y - \theta_r}{\xi} \right) \right)^2 dy \\
\widehat{\text{REF}} &= \frac{1}{T} \sum_{r=1}^T \int_{-\infty}^{\infty} \left(\Phi \left(\frac{y - \theta_r}{\xi} \right) - \frac{1}{T} \sum_{t=1}^T \Phi \left(\frac{y - \theta_t}{\xi} \right) \right)^2 dy \\
\widehat{\text{DIS}} &= \frac{1}{T} \sum_{r=1}^T \int_{-\infty}^{\infty} \left(\frac{\sum_{t=1}^T \Phi \left(\frac{y - \theta_t}{\xi} \right) \mathbb{1}_{\{x_t\}}(x_r)}{\sum_{u=1}^T \mathbb{1}_{\{x_u\}}(x_r)} - \frac{1}{T} \sum_{t=1}^T \Phi \left(\frac{y - \theta_t}{\xi} \right) \right)^2 dy \\
\widehat{\text{COR}} &= \frac{1}{T} \sum_{r=1}^T \int_{-\infty}^{\infty} \left(\frac{\sum_{t=1}^T \Phi \left(\frac{y - \theta_t}{\xi} \right) \mathbb{1}_{\{x_t\}}(x_r)}{\sum_{u=1}^T \mathbb{1}_{\{x_u\}}(x_r)} - \mathbb{1}_{[x_r, \infty)}(y) \right)^2 dy
\end{aligned} \tag{2.112}$$

To illustrate the evaluation of the attributes, we choose parameter values $\mu_{\Theta} = 1$, $\sigma_{\theta} = 2$, $\mu_X = 0$, $\sigma_X = 1$, $\rho = 0.25$ and $\xi = 3$ and simulate samples in \mathbf{R} (Ihaka and Gentleman, 1996; R Core Team, 2017) using the `mvrnorm` function in the MASS library (Venables and Ripley, 2002). Specifically, for each $m = 1, \dots, M = 1000$, a sample of $T = 50$ forecast-outcome pairs, (θ_t, x_t) , $t = 1, \dots, T$ was generated and used to estimate the decomposition terms using the computational formulae in (2.112) (although, in general, these estimates are biased (see Ferro and Fricker, 2012)); each x_t and θ_t are rounded to 2 decimal places. From the M estimates of each attribute a histogram for the value of the attribute is constructed on which the exact value of the attribute, calculated using the formulae in (2.110) and (2.111), was displayed. All integrals were computed using Monte Carlo integration (where the Monte Carlo samples used to compute the integrals were common to all integrals to obviate their effect on the variation of the estimates of the attributes). The histograms, each augmented with the exact attribute value, are presented in Figures 2.2a and 2.2b.

In Figure 2.2, it can be seen that for the resolution, reliability, discrimination and correct-



(a) URR Decomposition.



(b) RDC Decomposition.

Figure 2.2: Histograms of the attributes of the CRPS. For each attribute, a set of $M = 1000$ estimates were calculated, each estimate based on a sample of $T = 50$ forecast-outcome pairs that was simulated from the toy model of equations (2.107) to (2.109) with parameter values $\mu_{\Theta} = 1$, $\sigma_{\theta} = 2$, $\mu_X = 0$, $\sigma_X = 1$, $\rho = 0.25$ and $\xi = 3$. The URR attributes are: expected score (purple), uncertainty (red), resolution (green) and reliability (blue). The RDC attributes are: expected score (purple), refinement (red), discrimination (green) and correctness (blue). The exact values of the attributes are marked by vertical lines.

ness attributes there is disagreement between the empirical estimates and the exact attribute value. A principal cause of this disagreement is sparsity: most forecasts and outcomes appear infrequently in the sample.

Sparsity affects those attributes that have conditional expectations in their expressions. Consider the calculation of the resolution and reliability attributes of the URR decomposition: the sample of forecast-outcome pairs must be stratified according to the forecast values. If a

particular instance of a forecast does not appear in a sample, no outcomes will be recorded in this forecast's stratum and the expectation of the observation conditioned on this forecast will have a computed value that is spuriously null; even if the forecast is issued, but only on a few occasions, the small number of times the forecast is issued, reduces the sample size of the corresponding outcomes to a value too small for the conditional expectation to be estimated with any stability. A similar difficulty can arise on computing the terms of the RDC decomposition when many different values of the observation are recorded with each value appearing on only a few occasions at most. We address the sparsity problem in the next section.

[2.8] Decompositions Under Binning

As the illustration at the end of the previous section showed, under-representation, or sparsity, of forecasts or outcomes can lead to false estimates of the terms of the decompositions.

Previous authors (see, for example [Atger, 2003](#); [Bröcker and Smith, 2007b](#)) have considered the sparsity problem for the attributes of the URR decomposition. One approach taken by these authors has been to divide the set of *possible* forecasts into a small number of groups, or categories, or *bins* and then to allocate each sample forecast-outcome pair to the bin in which the forecast lies. For each bin, a representative forecast is nominated (examples include the mid-forecast and average forecast for the bin) and all forecasts in the bin are changed to the bin's nominated forecast value and these redefined forecasts used to evaluate forecast quality. If the bins are chosen in such a way as to ensure that each bin has a suitably large number of the sample forecast-outcome pairs, then the problem of sparsity is alleviated.

However, two difficulties arise from the method of binning. Firstly, while the attributes are evaluated using binning, the accuracy of the forecasts (i.e. the expected score) is still calculated from the original sample forecast-outcome pairs (as sparsity is not a problem for the computation of accuracy); consequently, the left-hand and right-hand sides of the URR decomposition are no longer equal ([Atger, 2003](#)). Secondly, the choice of bins, and the choice of nominated forecast for each bin, will determine how the forecasts are relabelled and will, therefore, affect the calculated values of the attributes: different choices of bins and nominated forecasts, may lead to two different sets of values for the attributes of the URR decomposition.

[Atger \(2004\)](#) and [Bröcker \(2012\)](#) discuss the problem of how best to choose bins and the nominated forecasts when computing the attributes of the URR decomposition. In this section, our emphasis is on correcting the URR decomposition under binning so as to ensure that

the left-hand side (accuracy) is equal to the right-hand side (attributes). [Stephenson et al. \(2008\)](#) provide such a correction for the URR decomposition of the Brier scoring rule ([Brier, 1950](#)) by introducing two further terms to the right-hand side of the URR decomposition. Here, we give a general corrected URR decomposition that holds for all scoring rules of all types of forecast under any binning. We also derive a corrected RDC decomposition to adjust for the binning of outcomes.

2.8.1 || URR Decomposition

Having chosen a set of bins, let $\beta : \mathcal{F}_f \rightarrow \mathcal{F}_f$ be a function that maps the original issued forecasts to their nominated bin values. We refer to such a function as a binning of the issued forecasts. It is the issued forecasts, $f(p)$, $p \in \mathcal{P}$ that are used to evaluate the accuracy of the forecasts, but it is the binned forecasts, $\beta(f(p))$, $p \in \mathcal{P}$ that are used to determine the attributes of the forecasts.

If we regard the binned forecasts as the forecasts that *are* issued, then by substituting $\beta(f(p))$ for $f(p)$ in equation (2.28), we have that for binned forecasts the reliability term is

$$\mathbb{E}[S(\beta(f(P)), q_{\beta(f(P))}) - S(f(q_{\beta(f(P))}), q_{\beta(f(P))})] \geq 0 \quad (2.113)$$

which agrees with the definition given by [Bröcker and Smith \(2007b\)](#).

Replacing $f(p)$ with $\beta(f(p))$ in equation (2.31), the resolution term is

$$S(f(q), q) - \mathbb{E}[S(f(q_{\beta(f(P))}), q_{\beta(f(P))})] \geq 0. \quad (2.114)$$

Uncertainty, which does not depend on the forecasts (and is, therefore, unaffected by binning) remains as

$$S(f(q), q). \quad (2.115)$$

By definition, the variability of the outcomes around the binned forecast $\beta(f(p))$ is $S(\beta(f(p)), q_{\beta(f(p))})$, in contrast with the variability of the outcomes about the issued forecast $f(p)$, which is $S(f(p), q_{f(p)})$. Taking the expectation over all forecasts, the difference

$$\mathbb{E}[S(f(P), q_{f(P)})] - \mathbb{E}[S(\beta(f(P)), q_{\beta(f(P))})] \quad (2.116)$$

is a measure of the change in variability of the outcomes caused by binning. It is easily seen that the first term of equation (2.116) is the accuracy of the issued forecasts and the second term in equation (2.116) is the accuracy of the binned forecasts.

The difference in variability quantified by equation (2.116) arises from the presence of additional outcomes corresponding to each binned forecast (i.e. nominated forecast of each bin). If all forecasts in each bin were the same, then there would be no additional outcomes because the outcomes attached to the bin would be the same outcomes associated with the singular forecast of the bin. But, if there are distinct forecasts in the same bin, the outcomes associated with a binned forecast will be a union of all outcomes paired with the forecasts in the bin.

The totality of the outcomes associated with a bin will have a variability that depends on the covariance between the forecasts and outcomes: if there is little correspondence between outcomes and forecasts, then the spread of different forecasts in the bin will not lead to a more diverse set of outcomes for the bin (although the number of different outcomes may increase, the distribution of all outcomes in the bin will differ little from the distribution of the outcomes corresponding to each forecast in the bin), but if there is good correspondence between outcomes and forecasts, having a range of forecasts in the bin will increase the variability of the outcomes of the bin.

We may, therefore, factorise the difference in variability defined by equation (2.116) into two parts: (i) the within-bin variance of the forecasts, which, recalling that q^* is the unconditional, i.e. historical, probability distribution of the forecasts, and $q_{\beta(f(p))}^*$ the conditional distribution of the forecasts given the forecasts fall in the bin $\beta(f(p))$, we define as

$$\mathbb{E}[S^*(q_{\beta(f(P))}^*, q_{\beta(f(P))}^*)] \quad (2.117)$$

and (ii) the remaining part which must represent the extent to which, within each bin, on average, the outcomes co-vary with the forecasts,

$$\mathbb{E}[S(f(P), q_{f(P)})] - \mathbb{E}[S(\beta(f(P)), q_{\beta(f(P))})] - \mathbb{E}[S^*(q_{\beta(f(P))}^*, q_{\beta(f(P))}^*)]. \quad (2.118)$$

Equations (2.113) to (2.118) give the general URR decomposition for binned forecasts,

$$\begin{aligned}
\mathbb{E}[S(f(P), X)] = & \\
& \underbrace{S(f(q), q)}_{\text{Uncertainty}} - \underbrace{(S(f(q), q) - \mathbb{E}[S(f(q_{\beta(f(P))), q_{\beta(f(P))})])]}_{\text{Resolution|Bins}} \\
& \underbrace{\hspace{10em}}_{\text{Sharpness|Bins}} \\
& + \underbrace{\mathbb{E}[S(\beta(f(P)), q_{\beta(f(P))}) - S(f(q_{\beta(f(P))}), q_{\beta(f(P))})]}_{\text{Reliability|Bins}} \\
& + \underbrace{\mathbb{E}[S^*(q_{\beta(f(P))}^*, q_{\beta(f(P))}^*)]}_{\text{Within-Bin Forecast Variation}} \\
& + \underbrace{\mathbb{E}[S(f(P), q_{f(P)})] - \mathbb{E}[S(\beta(f(P)), q_{\beta(f(P))})] - \mathbb{E}[S^*(q_{\beta(f(P))}^*, q_{\beta(f(P))}^*)]}_{\text{Within-Bin Covariation}}. \quad (2.119)
\end{aligned}$$

2.8.1.1 | Example: Brier Scoring Rule

Referring to section 2.6.1, for the Brier scoring rule (Brier, 1950),

$$\begin{aligned}
S(p, x) &= (p - x)^2 \\
S(p, r) &= p^2 - 2p\mathbb{E}_r[X] + \mathbb{E}_r[X] \quad (2.120)
\end{aligned}$$

and

$$\begin{aligned}
S^*(r^*, p) &= (\mathbb{E}_{r^*}[P] - p)^2 \\
S^*(r^*, q^*) &= (\mathbb{E}_{r^*}[P] - \mathbb{E}_{q^*}[P])^2 + \mathbb{E}_{q^*}[(P - \mathbb{E}_{q^*}[P])^2]. \quad (2.121)
\end{aligned}$$

Suppose there are m bins B_1, \dots, B_m . For the forecast $p \in \mathcal{P}$, define $\beta(p)$ to be the expected value of the forecasts in the bin in which p falls. Then

$$\beta(p) = \sum_{i=1}^m \mathbb{E}[P|P \in B_i] \mathbb{1}_{\{B_i\}}(p). \quad (2.122)$$

Noting that the value of $\beta(p)$ indicates the bin containing p (i.e. the expected value of the forecasts in a bin will differ from the expected value of the forecasts of every other bin), conditioning on the bin containing a forecast p is equivalent to conditioning on the value of $\beta(p)$, so that if $p \in B_i$, then

$$\beta(p) = \mathbb{E}[P|P \in B_i] = \mathbb{E}[P|\beta(P) = \beta(p)] \quad (2.123)$$

and we can write, more generally,

$$\beta(P) = \mathbb{E}[P|\beta(P)]. \quad (2.124)$$

We evaluate the terms of the binned decomposition (2.119) individually. Uncertainty is equal to

$$S(q, q) = q(1 - q) = \mathbb{E}[X](1 - \mathbb{E}[X]). \quad (2.125)$$

Resolution under binning is

$$\begin{aligned} S(q, q) - \mathbb{E}[S(q_{\beta(P)}, q_{\beta(P)})] \\ &= \mathbb{E}[X](1 - \mathbb{E}[X]) - \mathbb{E}[\mathbb{E}[X|\beta(P)](1 - \mathbb{E}[X|\beta(P)])] \\ &= \mathbb{E}[(\mathbb{E}[X|\beta(P)] - \mathbb{E}[X])^2]. \end{aligned} \quad (2.126)$$

The last of the standard three terms of the URR decomposition, reliability, is, given binning,

$$\begin{aligned} \mathbb{E}[S(\beta(P), q_{\beta(P)}) - S(q_{\beta(P)}, q_{\beta(P)})] \\ &= \mathbb{E}[\beta^2(P) - 2\beta(P)\mathbb{E}[X|\beta(P)] + \mathbb{E}[X|\beta(P)] - \mathbb{E}[X|\beta(P)] + \mathbb{E}^2[X|\beta(P)]] \\ &= \mathbb{E}[(\beta(P) - \mathbb{E}[X|\beta(P)])^2]. \end{aligned} \quad (2.127)$$

The effect of binning is to introduce two additional terms to the URR decomposition. The first, within-bin forecast variation, is equal to

$$\begin{aligned} \mathbb{E}[S^*(q_{\beta(P)}^*, q_{\beta(P)}^*)] \\ &= \mathbb{E}[\mathbb{E}_{q_{\beta(P)}^*}[(P - \mathbb{E}_{q_{\beta(P)}^*}[P])^2]] \\ &= \mathbb{E}[\mathbb{E}[(P - \mathbb{E}[P|\beta(P)])^2|\beta(P)]] \\ &= \mathbb{E}[\mathbb{E}[P^2|\beta(P)] - \mathbb{E}^2[P|\beta(P)]]. \end{aligned} \quad (2.128)$$

The second of the new terms, within-bin covariation, is

$$\begin{aligned} \mathbb{E}[\mathbb{E}[S(P, X)|\beta(P)]] - \mathbb{E}[S(\beta(P), q_{\beta(P)})] - \mathbb{E}[S^*(q_{\beta(P)}^*, q_{\beta(P)}^*)] \\ &= \mathbb{E}[-2\mathbb{E}[PX|\beta(P)] - \beta^2(P) + 2\beta(P)\mathbb{E}[X|\beta(P)] + \mathbb{E}^2[P|\beta(P)]] \end{aligned} \quad (2.129)$$

and, substituting $\beta(P) = \mathbb{E}[P|\beta(P)]$, we have $-\beta^2(P) + \mathbb{E}^2[P|\beta(P)] = 0$, leaving the within-bin covariation as

$$\begin{aligned}
& \mathbb{E}[-2\mathbb{E}[PX|\beta(P)] + 2\beta(P)\mathbb{E}[X|\beta(P)]] \\
&= -2\mathbb{E}[\mathbb{E}[(P - \beta(P))(X - \mathbb{E}[X|\beta(P)])|\beta(P)]] \\
&= -2\mathbb{E}[\mathbb{E}[(P - \mathbb{E}[P|\beta(P)])(X - \mathbb{E}[X|\beta(P)])|\beta(P)]]. \tag{2.130}
\end{aligned}$$

The five attributes under binning, agree with those of the empirical result of [Stephenson et al. \(2008, equation \(7\)\)](#).

2.8.2 || RDC Decomposition

In those circumstances in which the range space of the observation is sufficiently large that it is unlikely for any outcome to occur more than a few times, binning can be applied to the outcomes too. We are unaware of any discussion in the literature on the binning of outcomes, it being a topic relevant to computing the terms of the less studied RDC decomposition. Here, we give a general RDC decomposition when outcomes are binned.

Let $\beta^* : \mathcal{X} \rightarrow \mathcal{X}$ be a mapping of the actual outcomes to the set of values nominated for the outcome bins i.e. $\beta^*(x)$ is the nominated value of the bin in which outcome x lies. If the binned outcomes are treated as *the* outcomes, a re-statement of the RDC decomposition (2.41) under binning, gives discrimination (compare with equation (2.40)) to be

$$S^*(q^*, q^*) - \mathbb{E}[S^*(q_{\beta^*(X)}^*, q_{\beta^*(X)}^*)] \tag{2.131}$$

and correctness (compare with equation (2.37)) is

$$\mathbb{E}[S(f(P), \beta^*(X))] - \mathbb{E}[S^*(q_{\beta^*(X)}^*, q_{\beta^*(X)}^*)]. \tag{2.132}$$

Refinement, being a property of the forecasts alone, is unaffected by the binning applied to the outcomes and is unchanged as

$$S^*(q^*, q^*). \tag{2.133}$$

Equations (2.131) and (2.132) summarise the variability of forecasts preceding the binned outcomes. For a particular outcome, x , with bin, $\beta^*(x)$, we can compare the total variability of the forecasts preceding x to the variability of the forecasts preceding the binned outcome $\beta^*(x)$, by calculating the difference

$$\mathbb{E}[S(f(P), X)|X = x] - \mathbb{E}[S(f(P), \beta^*(X))|\beta^*(X) = \beta^*(x)]. \tag{2.134}$$

Aggregating the difference in equation (2.134) over all possible outcomes, the change in the

variability of the forecasts as a result of binning is

$$\mathbb{E}[\mathbb{E}[S(f(P), X)|X]] - \mathbb{E}[\mathbb{E}[S(f(P), \beta^*(X))|\beta^*(X)]] \quad (2.135)$$

or, the accuracy of the forecasts with respect to the original outcomes less the accuracy of the forecasts with respect to the binned outcomes.

The aggregate difference (2.135) is brought about by the set of forecasts associated with the binned outcome being larger than the set of forecasts associated with each individual outcome in the bin. If there is one outcome in a bin then the forecasts of the binned outcome are necessarily the same as the forecasts for the sole outcome of the bin and the difference of (2.134) for the bin is zero. Should there be a multitude of different outcomes in a bin, however, then the variability of the forecasts for the bin will differ from the variability of the forecasts preceding any outcome in the bin provided that the forecasts for each outcome in the bin have either different locations or different dispersions, or both. We identify, therefore, two components to equation (2.135).

The first component is the change in forecast variability under binning that results from different outcomes being preceded by forecasts of differing spreads: we approximate this effect by the variability of the outcomes in the bin (the greater the outcome variability in a bin, each outcome accompanied by differing forecasts, the more dispersed the forecasts of the bin). The within-bin variance of the outcomes is, from our definitions,

$$\mathbb{E}[S(f(q_{\beta^*(X)}), q_{\beta^*(X)})]. \quad (2.136)$$

The second component is the change in forecast variability arising from differing outcomes being preceded by distinct forecasts: as outcomes change there is a shift in the forecasts preceding the outcomes, contributing to a more diverse set of forecasts. This co-movement of the forecasts with the outcomes is the quantity

$$\mathbb{E}[\mathbb{E}[S(f(P), X)|X]] - \mathbb{E}[\mathbb{E}[S(f(P), \beta^*(X))|\beta^*(X)]] - \mathbb{E}[S(f(q_{\beta^*(X)}), q_{\beta^*(X)})]. \quad (2.137)$$

which we refer to as the within-bin covariance.

The terms (2.131) to (2.137), give, in combination, the binned RDC decomposition

$$\begin{aligned}
\mathbb{E}[S(f(P), X)] = & \\
& \underbrace{S^*(q^*, q^*)}_{\text{Refinement}} - \underbrace{(S^*(q^*, q^*) - \mathbb{E}[S^*(q_{\beta^*}^*(X), q_{\beta^*}^*(X))])}_{\text{Discrimination|Bins}} \\
& \underbrace{\hspace{10em}}_{\text{Excess|Bins}} \\
& + \underbrace{(\mathbb{E}[S(f(P), \beta^*(X))] - \mathbb{E}[S^*(q_{\beta^*}^*(X), q_{\beta^*}^*(X))])}_{\text{Correctness|Bins}} \\
& + \underbrace{\mathbb{E}[S(f(q_{\beta^*}^*(X)), q_{\beta^*}^*(X))]}_{\text{Within-Bin Variance}} \\
& + \underbrace{(\mathbb{E}[\mathbb{E}[S(f(P), X)|X]] - \mathbb{E}[\mathbb{E}[S(f(P), \beta^*(X))|\beta^*(X)]] - \mathbb{E}[S(f(q_{\beta^*}^*(X)), q_{\beta^*}^*(X))])}_{\text{Within-Bin Covariance}}. \quad (2.138)
\end{aligned}$$

in parallel with the binned URR decomposition of equation (2.119).

2.8.3 || Classification of the New Terms

In their discussion of the binned URR decomposition for the Brier scoring rule, [Stephenson et al. \(2008\)](#) argue that the two new terms of the binned URR decomposition should be combined with the resolution term to give a generalised resolution term (and reduce the binned URR decomposition to three terms). In the same context, [Siegert \(2017\)](#) suggests that the two new terms be merged with the reliability term. We propose a compromise.

For the URR decomposition: (i) the within-bin variance of the forecasts should be combined with the resolution term, because, the more variable the forecast within a bin, the greater the increase in the number of outcomes attached to the bin, affecting the resolution of the forecasts; (ii) the within-bin covariance be combined with the reliability term, because the within-bin covariance conveys how much the outcomes are in alignment with the forecasts, an element of reliability. Making these allocations,

$$\begin{aligned}
 \mathbb{E}[S(f(P), X)] = & \\
 & \left. \begin{array}{l} \underbrace{S(f(q), q)}_{\text{Uncertainty}} \\ - \left(\underbrace{S(f(q), q) - \mathbb{E}[S(f(q_{\beta(f(P))), q_{\beta(f(P))})]}_{\text{Binned Resolution}}}_{\text{Binned Sharpness}} \right) \end{array} \right\} \\
 & + \mathbb{E} \left[S(\beta(f(P)), q_{\beta(f(P))}) - S(f(q_{\beta(f(P))}), q_{\beta(f(P))}) \right] \\
 & + \mathbb{E}[S(f(P), q_{f(P)})] - \mathbb{E}[S(\beta(f(P)), q_{\beta(f(P))})] \\
 & - \mathbb{E}[S^*(q_{\beta^*(f(P))}^*, q_{\beta^*(f(P))}^*)]. \end{array} \left. \right\} \text{Binned Reliability} \quad (2.139)
 \end{aligned}$$

As a consequence of this allocation, *neither* binned resolution nor binned reliability need be *positive*.

With similar justification, for the RDC decomposition we join the within-bin variance of the outcomes with the binned discrimination term and the within-bin covariance with the correctness term, giving the modified RDC decomposition to be

$$\begin{aligned}
 \mathbb{E}[S(f(P), X)] = & \\
 & \left. \begin{array}{l} \underbrace{S^*(q^*, q^*)}_{\text{Refinement}} \\ - \left(\underbrace{S^*(q^*, q^*) - \mathbb{E}[S^*(q_{\beta^*(X)}^*, q_{\beta^*(X)}^*)]}_{\text{Binned Discrimination}} \right) \end{array} \right\} \text{Binned Excess} \\
 & + \mathbb{E} \left[S(f(P), \beta^*(X)) \right] - \mathbb{E}[S^*(q_{\beta^*(X)}^*, q_{\beta^*(X)}^*)] \\
 & + \mathbb{E}[\mathbb{E}[S(f(P), X)|X]] - \mathbb{E}[\mathbb{E}[S(f(P), \beta^*(X))|\beta^*(X)]] \\
 & - \mathbb{E}[S(f(q_{\beta^*(X)}), q_{\beta^*(X)})]. \end{array} \left. \right\} \text{Binned Correctness} \quad (2.140)
 \end{aligned}$$

2.8.4 || An Illustration

To demonstrate the results of binning, consider again the CRPS, for which (from equations (2.56) and (2.62))

$$S(f(p), x) = \int_{\mathbb{R}} (\mathbb{E}_p[\mathbf{1}_{[X, \infty)}(y)] - \mathbf{1}_{[x, \infty)}(y))^2 dy \quad (2.141)$$

and

$$S^*(r^*, F_p) = \int_{-\infty}^{\infty} (\mathbb{E}_{r^*}[F_P(y)] - F_p(y))^2 dy. \quad (2.142)$$

With binning we have the CRPS for the binned forecasts

$$S(\beta(f(p)), x) = \int_{\mathbb{R}} (\mathbb{E}_{\beta(f(p))}[\mathbf{1}_{[x, \infty)}(y)] - \mathbf{1}_{[x, \infty)}(y))^2 dy. \quad (2.143)$$

The URR and RDC decompositions for unbinned forecasts and observations are given in equations (2.61) and (2.68). Under binning, the modified and additional terms of the URR decomposition are

$$\begin{aligned} \text{Uncertainty (UNC)} &= \int_{-\infty}^{\infty} \mathbb{E}[\mathbf{1}_{[X, \infty)}(y)](1 - \mathbb{E}[\mathbf{1}_{[X, \infty)}(y)]) dy \\ \text{Resolution|Bins (RES|B)} &= \mathbb{E} \left[\int_{-\infty}^{\infty} (\mathbb{E}[\mathbf{1}_{[X, \infty)}(y)|\beta(f(P))] - \mathbb{E}[\mathbf{1}_{[X, \infty)}(y)])^2 dy \right] \\ \text{Reliability|Bins (REL|B)} &= \mathbb{E} \left[\int_{-\infty}^{\infty} (\mathbb{E}[\mathbf{1}_{[X, \infty)}(y)|\beta(f(P))] - \beta(f(P))(y))^2 dy \right] \\ \text{Within-Bin Variation (WBV)} &= \\ &\mathbb{E} \left[\int_{-\infty}^{\infty} \mathbb{E}[f(P)^2(y)|\beta(f(P))] - \mathbb{E}^2[f(P)(y)|\beta(f(P))] dy \right] \end{aligned} \quad (2.144)$$

giving the binned URR decomposition to be

$$\begin{aligned} \mathbb{E}[S(f(P), x)] &= \text{UNC} - \underbrace{(\text{RES|B} - \text{WBV})}_{\text{Binned Resolution}} \\ &+ \underbrace{(\text{REL|B} + (\mathbb{E}[S(f(P), X)] - \mathbb{E}[S(\beta(f(P)), X)] - \text{WBV}))}_{\substack{\text{Within-Bin Covariance} \\ \text{Binned Reliability}}} \end{aligned} \quad (2.145)$$

and the modified and additional terms for the RDC decomposition are

$$\begin{aligned} \text{Refinement (REF)} &= \int_{-\infty}^{\infty} \mathbb{E}[(f(P)(y) - \mathbb{E}[f(P)(y)])^2] dy \\ \text{Discrimination|Bins (DIS|B)} &= \mathbb{E} \left[\int_{-\infty}^{\infty} (\mathbb{E}[f(P)(y)] - \mathbb{E}[f(P)(y)|\beta^*(X)])^2 dy \right] \\ \text{Correctness|Bins (COR|B)} &= \mathbb{E} \left[\int_{-\infty}^{\infty} (\mathbb{E}[f(P)(y)|\beta^*(X)] - \mathbf{1}_{[\beta^*(X), \infty)}(y))^2 dy \right] \\ \text{Within-Bin Variation (WBV(O))} &= \\ &\mathbb{E} \left[\int_{-\infty}^{\infty} \mathbb{E}[\mathbf{1}_{[X, \infty)}(y)|\beta^*(X)](1 - \mathbb{E}[\mathbf{1}_{[X, \infty)}(y)|\beta^*(X)]) dy \right] \end{aligned} \quad (2.146)$$

from which the binned RDC decomposition can be written

$$\begin{aligned} \mathbb{E}[S(f(P), x)] = & \text{UNC} - \underbrace{(\text{DIS|B} - \text{WBV(O)})}_{\text{Binned Discrimination}} \\ & + \underbrace{(\text{COR|B} + \underbrace{(\mathbb{E}[S(f(P), X)] - \mathbb{E}[S(f(P), \beta^*(X))] - \text{WBV(O)})}_{\text{Within-Bin Covariance}})}_{\text{Binned Correctness}}. \end{aligned} \quad (2.147)$$

To illustrate the evaluation of the binned attributes, we use again the toy model of section 2.7. Introducing binning, let B_Θ be the number bins for the parameter Θ and define the bins for Θ to be $(b_{i-1}, b_i]$, $i = 1, \dots, B_\Theta$, where

$$b_i = \begin{cases} -\infty & i = 0, \\ \mu_\Theta - 3\sigma_\Theta + (i - 1) \left(\frac{6\sigma_\Theta}{B_\Theta - 2} \right) & 1 \leq i \leq B_\Theta - 1, \\ \infty & i = B_\Theta. \end{cases} \quad (2.148)$$

Each forecast cumulative distribution function of X , $f(p) = F_p$, is, under the toy model, associated with a unique value for Θ and the forecasts are grouped according to the bins in which their Θ -parameter values fall. We choose for the representative forecast of the bin to which the forecast F_p is allocated, the cumulative distribution function $\beta(F_p)$ equal to the normal distribution with variance ξ^2 and mean $\beta(\Theta)$, the mean of Θ in the bin. With this binning there are exactly B_Θ different binned forecasts; the mean of Θ given bin $(b_{i-1}, b_i]$ will be denoted β_i , $i = 1, \dots, B_\Theta$.

Similarly, let B_X be the number of bins for the observation and set the bins to be $(b_{i-1}^*, b_i^*]$, $i = 1, \dots, B_X$, with

$$b_i^* = \begin{cases} -\infty & i = 0, \\ \mu_X - 3\sigma_X + (i - 1) \left(\frac{6\sigma_X}{B_X - 2} \right) & 1 \leq i \leq B_X - 1, \\ \infty & i = B_X. \end{cases} \quad (2.149)$$

Define the representative outcome for the bin containing the observed value $X = x$, $\beta^*(x)$, to be the mean of the outcomes in the bin containing x . There are then B_X binned outcomes; denote the representative outcome for the i th bin $(b_{i-1}^*, b_i^*]$ by β_i^* .

Under the toy model, the analytical formulae for the attributes of the URR decomposition (2.145) and the RDC decomposition (2.147) are listed in (2.150) to (2.153).

$$\begin{aligned}
\text{SCR} &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \left(d(y, u, z) - \mathbf{1}_{[z\sigma_X + \mu_X, \infty)}(y) \right)^2 \phi(u) \, du \phi(z) \, dz \, dy \\
\text{UNC} &= \int_{-\infty}^{\infty} a(y)(1 - a(y)) \, dy \\
\text{RES|B} &= \int_{-\infty}^{\infty} \sum_{i=1}^{B_{\Theta}} \left\{ \left(\int_{-\infty}^{\infty} \mathbf{1}_{[z\sigma_X + \mu_X, \infty)}(y) \left[\frac{f(b_i, z) - f(b_{i-1}, z)}{e(b_i) - e(b_{i-1})} \right] dz - a(y) \right)^2 \right. \\
&\quad \left. \times [e(b_i) - e(b_{i-1})] \right\} dy \\
\text{REL|B} &= \int_{-\infty}^{\infty} \sum_{i=1}^{B_{\Theta}} \left\{ \left(\int_{-\infty}^{\infty} \mathbf{1}_{[z\sigma_X + \mu_X, \infty)}(y) \left[\frac{f(b_i, z) - f(b_{i-1}, z)}{e(b_i) - e(b_{i-1})} \right] dz - g(y, i) \right)^2 \right. \\
&\quad \left. \times [e(b_i) - e(b_{i-1})] \right\} dy \\
\text{WBV} &= \int_{-\infty}^{\infty} \sum_{i=1}^{B_{\Theta}} \left\{ \left(c(y, u) - \int_{-\infty}^{\infty} c(y, v) \varphi(v) \, dv \right)^2 \varphi(u) \, du \right\} [e(b_i) - e(b_{i-1})] \, dy \\
\text{SCR|B} &= \int_{-\infty}^{\infty} \sum_{i=1}^{B_{\Theta}} \int_{-\infty}^{\infty} \left(g(y, i) - \mathbf{1}_{[z\sigma_X + \mu_X, \infty)}(y) \right)^2 [f(b_i, z) - f(b_{i-1}, z)] \phi(z) \, dz \, dy \\
\text{REF} &= \int_{-\infty}^{\infty} \left\{ \int_{-\infty}^{\infty} \left(c(y, w) - \left[\int_{-\infty}^{\infty} c(y, v) \phi(v) \, dv \right] \right)^2 \phi(w) \, dw \right\} dy \\
\text{DIS|B} &= \int_{-\infty}^{\infty} \sum_{i=1}^{B_X} \left\{ \left(\int_{-\infty}^{\infty} c(y, w) \left[\frac{h(b_i^*, w) - h(b_{i-1}^*, w)}{a(b_i^*) - a(b_{i-1}^*)} \right] \phi(w) \, dw - \int_{-\infty}^{\infty} c(y, w) \phi(w) \, dw \right)^2 \right. \\
&\quad \left. \times [a(b_i^*) - a(b_{i-1}^*)] \right\} dy \\
\text{COR|B} &= \int_{-\infty}^{\infty} \sum_{i=1}^{B_X} \left\{ \left(\int_{-\infty}^{\infty} c(y, w) \left[\frac{h(b_i^*, w) - h(b_{i-1}^*, w)}{a(b_i^*) - a(b_{i-1}^*)} \right] \phi(w) \, dw - \mathbf{1}_{[\beta_i^*, \infty)}(y) \right)^2 \right. \\
&\quad \left. \times [a(b_i^*) - a(b_{i-1}^*)] \right\} dy \\
\text{WBV(O)} &= \int_{-\infty}^{\infty} \sum_{i=1}^{B_X} \left(\frac{a(\min(\max(b_{i-1}^*, y), b_i^*)) - a(b_{i-1}^*)}{a(b_i^*) - a(b_{i-1}^*)} \right) \\
&\quad \times \left(1 - \frac{a(\min(\max(b_{i-1}^*, y), b_i^*)) - a(b_{i-1}^*)}{a(b_i^*) - a(b_{i-1}^*)} \right) dy \\
\text{SCR|B(O)} &= \int_{-\infty}^{\infty} \sum_{i=1}^{B_X} \left\{ \int_{-\infty}^{\infty} \left(c(y, w) - \mathbf{1}_{[\beta_i^*, \infty)}(y) \right)^2 [h(b_i^*, w) - h(b_{i-1}^*, w)] \phi(w) \, dw \right\} dy
\end{aligned} \tag{2.150}$$

$$\begin{aligned}
a(y) &= \Phi\left(\frac{y - \mu_X}{\sigma_X}\right) \\
b(y, w) &= \Phi\left(\frac{y - (\mu_X + \rho\sigma_X w)}{\sigma_X \sqrt{1 - \rho^2}}\right) \\
c(y, w) &= \Phi\left(\frac{y - (w\sigma_\Theta + \mu_\Theta)}{\xi}\right) \\
d(y, w, z) &= \Phi\left(\frac{y - [(w\sqrt{1 - \rho^2} + \rho z)\sigma_\Theta + \mu_\Theta]}{\xi}\right) \\
e(y) &= \Phi\left(\frac{y - \mu_\Theta}{\sigma_\Theta}\right) \\
f(y, z) &= \Phi\left(\frac{\left(\frac{y - \mu_\Theta}{\sigma_\Theta}\right) - \rho z}{\sqrt{1 - \rho^2}}\right) \\
g(y, i) &= \Phi\left(\frac{y - \beta_i}{\xi}\right) \\
h(y, w) &= \Phi\left(\frac{\left(\frac{y - \mu_X}{\sigma_X}\right) - \rho w}{\sqrt{1 - \rho^2}}\right) \\
&\text{(note: } h(y, w) \equiv b(y, w)\text{)}
\end{aligned} \tag{2.151}$$

using

$$\begin{aligned}
\varphi(w) &= \frac{\mathbb{1}\left(\frac{b_{i-1} - \mu_\Theta}{\sigma_\Theta}, \frac{b_i - \mu_\Theta}{\sigma_\Theta}\right](w)\phi(w)}{\Phi\left(\frac{b_i - \mu_\Theta}{\sigma_\Theta}\right) - \Phi\left(\frac{b_{i-1} - \mu_\Theta}{\sigma_\Theta}\right)} \\
\beta_i &= \mu_\Theta + \frac{\frac{\sigma_\Theta}{\sqrt{2\pi}} \left\{ \exp\left(-\frac{1}{2}\left(\frac{b_{i-1} - \mu_\Theta}{\sigma_\Theta}\right)^2\right) - \exp\left(-\frac{1}{2}\left(\frac{b_i - \mu_\Theta}{\sigma_\Theta}\right)^2\right) \right\}}{\Phi\left(\frac{b_i - \mu_\Theta}{\sigma_\Theta}\right) - \Phi\left(\frac{b_{i-1} - \mu_\Theta}{\sigma_\Theta}\right)} \\
\beta_i^* &= \mu_X + \frac{\frac{\sigma_X}{\sqrt{2\pi}} \left\{ \exp\left(-\frac{1}{2}\left(\frac{b_{i-1}^* - \mu_X}{\sigma_X}\right)^2\right) - \exp\left(-\frac{1}{2}\left(\frac{b_i^* - \mu_X}{\sigma_X}\right)^2\right) \right\}}{\Phi\left(\frac{b_i^* - \mu_X}{\sigma_X}\right) - \Phi\left(\frac{b_{i-1}^* - \mu_X}{\sigma_X}\right)}
\end{aligned} \tag{2.152}$$

and

$$\begin{aligned}
&\text{Binned Resolution} = \text{RES|B} - \text{WBV} \\
&\text{Binned Reliability} = \text{REL|B} + \text{SCR} - \text{SCR|B} - \text{WBV} \\
&\hspace{15em} \tag{2.153}
\end{aligned}$$

$$\text{Binned Discrimination} = \text{DIS|B} - \text{WBV(O)}$$

$$\text{Binned Correctness} = \text{COR|B} + \text{SCR} - \text{SCR|B(O)} - \text{WBV(O)}$$

In practice, the URR and RDC decompositions must be estimated from a sample of forecast-outcome pairs. Following the approach of section 2.7, each forecast is uniquely determined by the value of Θ ; a sample of size T of forecast-outcome pairs has the form (θ_t, x_t) for $t = 1, \dots, T$. Estimators for the attributes of the decompositions are derived by using the empirical distribution to model the joint and marginal distributions of the forecasts and outcomes. Under the empirical distributions, the representative forecast for each forecast-bin is, by our choice, the arithmetic average of the forecasts in the bin and we write $\beta(\theta_t) = \bar{\theta}_t$ for the representative forecast of the bin containing the forecast θ_t . Similarly, under our choice of using the average of the outcomes in a bin as the representative outcome for the outcome-bin, we write $\beta^*(x_t) = \bar{x}_t$ for the representative outcome of the bin containing the outcome x_t . Then the empirical estimators for the attributes of the binned URR decomposition (under which the forecasts alone are binned; for each t , the forecast θ_t is replaced with the simple arithmetic average, $\bar{\theta}_t$, of the forecasts in the bin in which θ_t falls) are presented in equations (2.154) and (2.155) (with the additional abbreviations: $\widehat{\text{RES}}|\text{B}$ is resolution|bins, $\widehat{\text{REL}}|\text{B}$ is reliability|bins and $\widehat{\text{SCR}}|\text{B}$ is the expected score|(forecast) bins).

$$\begin{aligned}
\widehat{\text{SCR}} &= \frac{1}{T} \sum_{r=1}^T \int_{-\infty}^{\infty} \left(\Phi \left(\frac{y - \theta_r}{\xi} \right) - \mathbb{1}_{[x_r, \infty)}(y) \right)^2 dy \\
\widehat{\text{SCR}}|\text{B} &= \frac{1}{T} \sum_{r=1}^T \int_{-\infty}^{\infty} \left(\Phi \left(\frac{y - \bar{\theta}_r}{\xi} \right) - \mathbb{1}_{[x_r, \infty)}(y) \right)^2 dy \\
\widehat{\text{UNC}} &= \int_{-\infty}^{\infty} \left(\frac{1}{T} \sum_{t=1}^T \mathbb{1}_{[x_t, \infty)}(y) \right) \left(1 - \frac{1}{T} \sum_{r=1}^T \mathbb{1}_{[x_r, \infty)}(y) \right) dy \\
\widehat{\text{RES}}|\text{B} &= \frac{1}{T} \sum_{r=1}^T \int_{-\infty}^{\infty} \left(\frac{\sum_{t=1}^T \mathbb{1}_{[x_t, \infty)}(y) \mathbb{1}_{\{\bar{\theta}_t\}}(\bar{\theta}_r)}{\sum_{u=1}^T \mathbb{1}_{\{\bar{\theta}_u\}}(\bar{\theta}_r)} - \frac{1}{T} \sum_{t=1}^T \mathbb{1}_{[x_t, \infty)}(y) \right)^2 dy \\
\widehat{\text{REL}}|\text{B} &= \frac{1}{T} \sum_{r=1}^T \int_{-\infty}^{\infty} \left(\frac{\sum_{t=1}^T \mathbb{1}_{[x_t, \infty)}(y) \mathbb{1}_{\{\bar{\theta}_t\}}(\bar{\theta}_r)}{\sum_{u=1}^T \mathbb{1}_{\{\bar{\theta}_u\}}(\bar{\theta}_r)} - \Phi \left(\frac{y - \bar{\theta}_r}{\xi} \right) \right)^2 dy \\
\widehat{\text{WBV}} &= \frac{1}{T} \sum_{r=1}^T \int_{-\infty}^{\infty} \left(\frac{\sum_{t=1}^T \Phi^2 \left(\frac{y - \theta_t}{\xi} \right) \mathbb{1}_{\{\bar{\theta}_t\}}(\bar{\theta}_r)}{\sum_{u=1}^T \mathbb{1}_{\{\bar{\theta}_u\}}(\bar{\theta}_r)} - \left[\frac{\sum_{t=1}^T \Phi \left(\frac{y - \theta_t}{\xi} \right) \mathbb{1}_{\{\bar{\theta}_t\}}(\bar{\theta}_r)}{\sum_{u=1}^T \mathbb{1}_{\{\bar{\theta}_u\}}(\bar{\theta}_r)} \right]^2 \right)^2 dy
\end{aligned} \tag{2.154}$$

and

$$\begin{aligned} \widehat{\text{Binned Resolution}} &= \widehat{\text{RES}}|\text{B} - \widehat{\text{WBV}} \\ \widehat{\text{Binned Reliability}} &= \widehat{\text{REL}}|\text{B} + \widehat{\text{SCR}} - \widehat{\text{SCR}}|\text{B} - \widehat{\text{WBV}} \end{aligned} \quad (2.155)$$

The empirical estimators for the attributes of the binned RDC decomposition (for which the outcomes alone are binned; for each t , the outcome x_t is replaced with the simple arithmetic average, \bar{x}_t , of the outcomes in the bin in which x_t lies) have the expressions in equations (2.156) and (2.157) (with abbreviations: $\widehat{\text{DIS}}|\text{B}$ is discrimination|bins, $\widehat{\text{COR}}|\text{B}$ is correctness|bins, $\widehat{\text{SCR}}|\text{B}(\text{O})$ is the expected score|(observation) bins).

$$\begin{aligned} \widehat{\text{SCR}} &= \frac{1}{T} \sum_{r=1}^T \int_{-\infty}^{\infty} \left(\Phi \left(\frac{y - \theta_r}{\xi} \right) - \mathbf{1}_{[x_r, \infty)}(y) \right)^2 dy \\ \widehat{\text{SCR}}|\text{B}(\text{O}) &= \frac{1}{T} \sum_{r=1}^T \int_{-\infty}^{\infty} \left(\Phi \left(\frac{y - \theta_r}{\xi} \right) - \mathbf{1}_{[\bar{x}_r, \infty)}(y) \right)^2 dy \\ \\ \widehat{\text{REF}} &= \frac{1}{T} \sum_{r=1}^T \int_{-\infty}^{\infty} \left(\Phi \left(\frac{y - \theta_r}{\xi} \right) - \frac{1}{T} \sum_{t=1}^T \Phi \left(\frac{y - \theta_t}{\xi} \right) \right)^2 dy \\ \widehat{\text{DIS}}|\text{B} &= \frac{1}{T} \sum_{r=1}^T \int_{-\infty}^{\infty} \left(\frac{\sum_{t=1}^T \Phi \left(\frac{y - \theta_t}{\xi} \right) \mathbf{1}_{\{\bar{x}_t\}}(\bar{x}_r)}{\sum_{u=1}^T \mathbf{1}_{\{\bar{x}_u\}}(\bar{x}_r)} - \frac{1}{T} \sum_{t=1}^T \Phi \left(\frac{y - \theta_t}{\xi} \right) \right)^2 dy \\ \widehat{\text{COR}}|\text{B} &= \frac{1}{T} \sum_{r=1}^T \int_{-\infty}^{\infty} \left(\frac{\sum_{t=1}^T \Phi \left(\frac{y - \theta_t}{\xi} \right) \mathbf{1}_{\{\bar{x}_t\}}(\bar{x}_r)}{\sum_{u=1}^T \mathbf{1}_{\{\bar{x}_u\}}(\bar{x}_r)} - \mathbf{1}_{[\bar{x}_r, \infty)}(y) \right)^2 dy \\ \widehat{\text{WBV}}(\text{O}) &= \frac{1}{T} \sum_{r=1}^T \int_{-\infty}^{\infty} \left(\frac{\sum_{t=1}^T \mathbf{1}_{[x_t, \infty)}(y) \mathbf{1}_{\{\bar{x}_t\}}(\bar{x}_r)}{\sum_{u=1}^T \mathbf{1}_{\{\bar{x}_u\}}(\bar{x}_r)} \right) \left(1 - \frac{\sum_{t=1}^T \mathbf{1}_{[x_t, \infty)}(y) \mathbf{1}_{\{\bar{x}_t\}}(\bar{x}_r)}{\sum_{u=1}^T \mathbf{1}_{\{\bar{x}_u\}}(\bar{x}_r)} \right) dy \end{aligned} \quad (2.156)$$

and

$$\begin{aligned} \widehat{\text{Binned Discrimination}} &= \widehat{\text{DIS}}|\text{B} - \widehat{\text{WBV}}(\text{O}) \\ \widehat{\text{Binned Correctness}} &= \widehat{\text{COR}}|\text{B} + \widehat{\text{SCR}} - \widehat{\text{SCR}}|\text{B}(\text{O}) - \widehat{\text{WBV}}(\text{O}) \end{aligned} \quad (2.157)$$

The formulae of (2.154) to (2.157) were applied to each of $M = 1000$ samples, a single sample being $T = 50$ forecast-outcome pairs simulated for the toy model of section 2.7 with

parameter values $\mu_{\Theta} = 1$, $\sigma_{\Theta} = 2$, $\mu_X = 0$, $\sigma_X = 1$, $\rho = 0.25$ and $\xi = 3$ (the simulated samples were generated in **R** (Ihaka and Gentleman, 1996; R Core Team, 2017) using the `mvrnorm` function in the **MASS** library (Venables and Ripley, 2002) and both outcomes and forecasts were rounded to 2 decimal places). All integrals in the empirical estimators were computed using Monte Carlo integration (where the Monte Carlo samples used were common to all integrals so that any variation in the estimates of the attributes may be attributed to variation in the samples of forecast-outcome pairs).

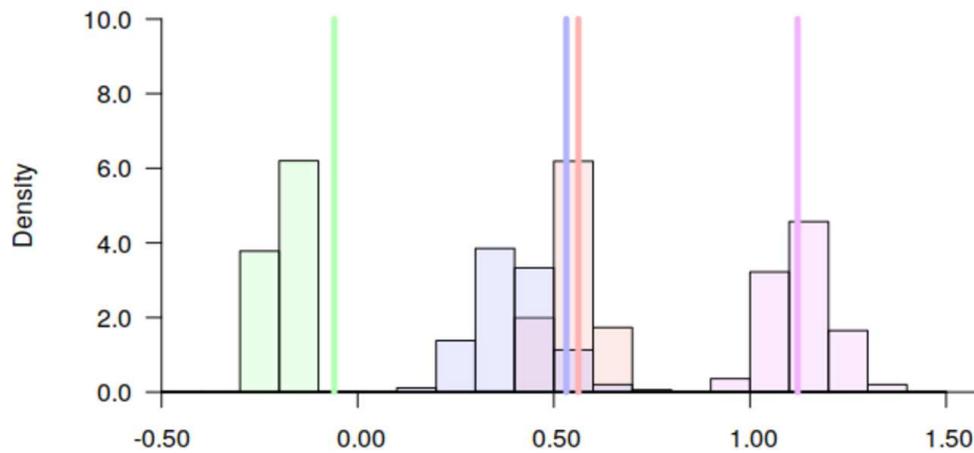
The $M = 1000$ estimates of each attribute are presented in the histograms of Figure 2.3; on each histogram the exact value of the attribute under binning is also displayed.

Binning forecasts reduces sample sparsity by increasing the number of occasions on which each (representative) forecast is repeated in the sample; in like manner, when the outcomes are binned, each (representative) outcome occurs more often in the sample. If representative forecasts and outcomes are repeated often in the sample, data sparsity is reduced and each conditional expectation can be calculated from a larger number of data.

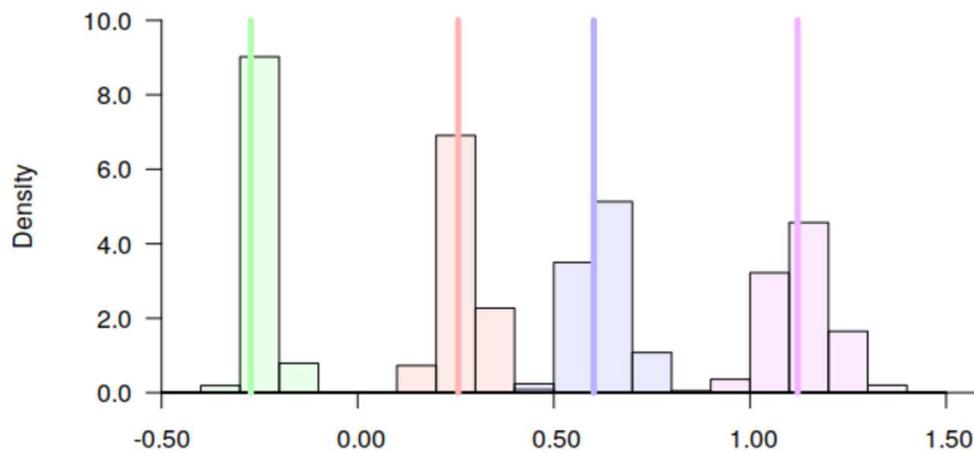
We see from Figure 2.3 that there is closer agreement between the sample estimates of the binned attributes and their exact values than was evident for the sample estimates of the unbinned attributes in Figure 2.2. However, binned resolution remains poorly estimated. A possible cause of this underestimation is that the within-bin variation of the forecasts has a larger effect on binned resolution than on binned reliability (in binned reliability there are several terms that can partly mute the contribution of the within-bin variation). That discrimination and correctness do not display the same contrast, may be explained by the variability of the forecasts: by choice of model parameters in our illustration, the forecasts have larger variation than the outcomes, therefore, it is more likely there will be values in the extreme bins for forecasts than for outcomes, and the few values in the extreme bins of the forecasts will destabilise the estimation of the expectation that is conditional on the forecasts.

It is also of interest to examine whether the binned attribute values do converge to the true (i.e. unbinned) attribute values as the number of bins increases indefinitely. Figures 2.4 and 2.5 show the values of the attributes of the URR and RDC decompositions with and without binning for different numbers of bins. The values calculated are the *exact* values for the attributes evaluated using the expressions in (2.110), (2.111) and (2.150) to (2.153) (the integrals are evaluated using Monte Carlo integration). The parameters chosen for the toy model are: $\mu_X = 0$, $\sigma_X = 1$, $\mu_{\Theta} = 1$, $\sigma_{\Theta} = 2$, $\rho = 0.25$ and $\xi = 3$.

We see that the attributes of both decompositions under binning converge, as the number of bins increases, to the attributes of the unbinned decomposition. The attributes of the



(a) URR Decomposition.



(b) RDC Decomposition.

Figure 2.3: Histograms of the attributes of the CRPS when binning is used. For the URR decomposition the forecasts alone are binned, for the RDC decomposition the outcomes alone are binned. For each attribute, a set of $M = 1000$ estimates were calculated, each estimate based on a sample of $T = 50$ forecast-outcome pairs simulated from the toy model of equations (2.107) to (2.109) with parameter values $\mu_\theta = 1$, $\sigma_\theta = 2$, $\mu_X = 0$, $\sigma_X = 1$, $\rho = 0.25$ and $\xi = 3$. The URR attributes are: expected score (■), uncertainty (■), resolution (■) and reliability (■). The RDC attributes are: expected score (■), refinement (■), discrimination (■) and correctness (■). The exact values of the attributes are marked by vertical lines.

binned RDC decomposition display a slower rate of convergence than the attributes of the binned URR decomposition, although this may be a model artefact rather than a more general pattern.

Figures 2.4 and 2.5 also confirm that the binned attributes can be negative (when, as is often

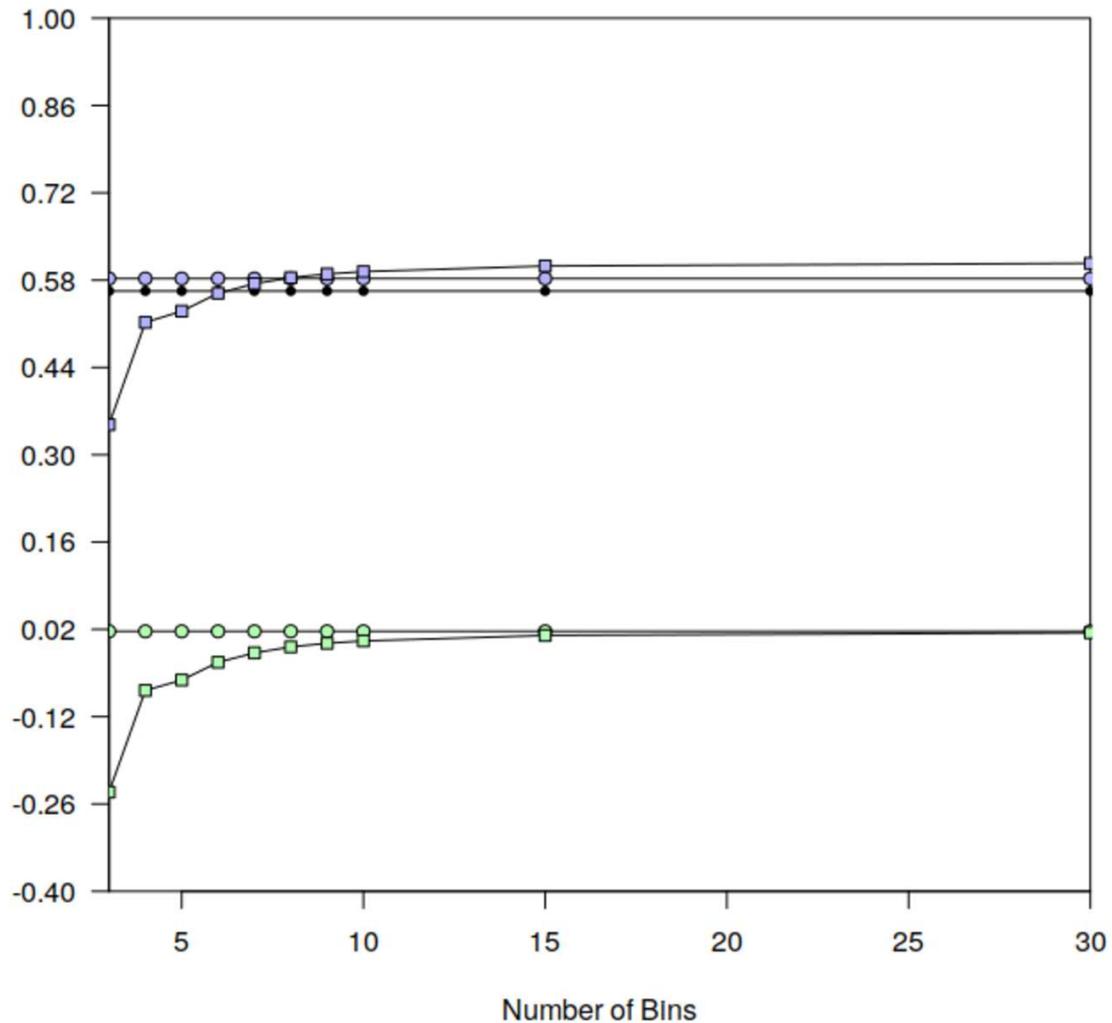


Figure 2.4: Comparison of the attribute values of the unbinned and binned URR decompositions of the CRPS for forecasts and outcomes that follow the toy model in section 2.7; parameters: $\mu_{\Theta} = 1$, $\sigma_{\Theta} = 2$, $\mu_X = 0$, $\sigma_X = 1$, $\rho = 0.25$, and $\xi = 3$. The attributes are: uncertainty (—●—), resolution (—○—), reliability (—●—), binned resolution (—■—), and binned reliability (—■—).

a practical necessity, the number of bins is small).

[2.9] Discussion and Conclusion

Formally, the qualities of forecasts are expressed in terms of the probability distributions of forecast and observation, but it can be difficult to translate these definitions into equivalent computable quantities. Quantities that indirectly measure the qualities of the forecasts can

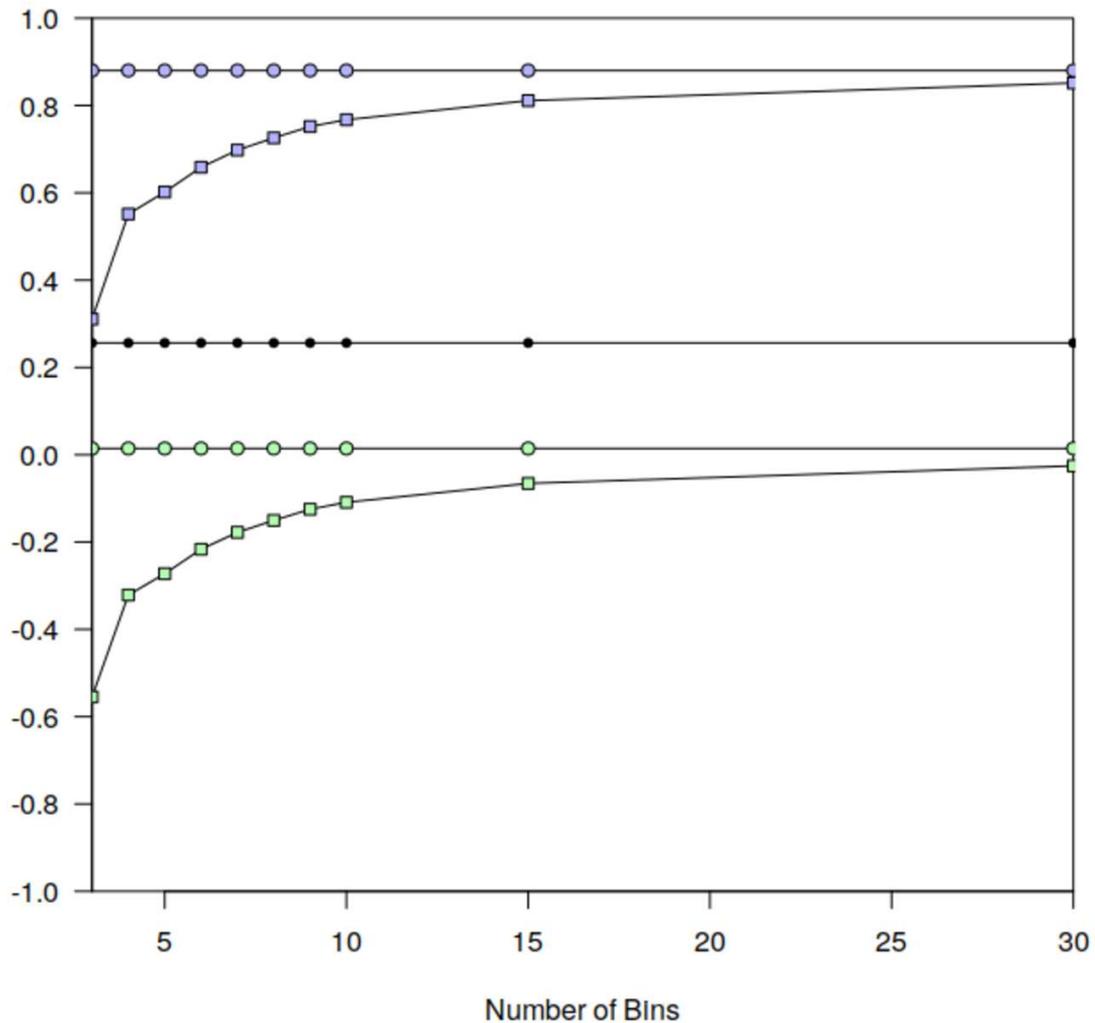


Figure 2.5: Comparison of the attribute values of the unbinned and binned RDC decompositions of the CRPS for forecasts and outcomes that follow the toy model in section 2.7; parameters: $\mu_{\Theta} = 1$, $\sigma_{\Theta} = 2$, $\mu_X = 0$, $\sigma_X = 1$, $\rho = 0.25$, and $\xi = 3$. The attributes are: refinement (—●—), discrimination (—○—), correctness (—●—), binned discrimination (—■—), and binned correctness (—■—).

be established by decomposing the expected score, or accuracy, of the forecasts into the sum of terms, each term quantifying a particular quality, or attribute, of the forecasts.

The attributes of uncertainty, resolution and reliability can be assessed from the URR decomposition, while quantities for the attributes of refinement, discrimination and correctness are determined from the separate RDC decomposition.

Bröcker (2009) gave a general approach for establishing the URR decomposition for all scor-

ing rules of probabilistic forecasts of discrete observations. But Bröcker (2009)'s result did not extend beyond probabilistic forecasts of events with finitely many values. Here, we give a general URR decomposition that is applicable to probabilistic and point forecasts of any observation (continuous or discrete).

We also derive a similarly general RDC decomposition, knowing of only one prior version of the RDC decomposition for the Brier scoring rule (Brier, 1950), which is derived by Murphy and Winkler (1987) (see also Murphy, 1986). By developing a general RDC decomposition we give a score-related measure to, in particular, discrimination. A score-related discrimination measure provides a firm basis for the comparison of discrimination with the values of the URR attributes and has the benefit of further enabling the extension of the assessment of discrimination beyond categorical forecasts, in which discrimination is a prominent quality.

Both URR and RDC decompositions are typically expressed in terms of the issued forecasts and recorded outcomes. But, in cases where it is rare for any forecast to be issued more than once or a particular outcome to be recorded more than once, the sparsity of data in samples reduces the stability and accuracy of the computed attributes. One response to this problem has been to give all forecasts and outcomes that are close to one another (i.e. are in the same *bin*) the same value. But, this approach makes the standard URR and RDC decompositions only approximate. We quantify the approximation error and give modified URR and RDC decompositions that take account of binning and which are applicable to all scoring rules of probabilistic and point forecasts.

The decomposition of accuracy is not the only method of deriving quantitative expressions for the qualities of the forecasts. For some attributes, many methods have been proposed for how to quantify them (see chapter 1), and new methods continue to be suggested. But, as Murphy and Winkler (1987) note, such 'measures have tended to proliferate, with relatively little effort being made to develop general concepts and principles [but] it should not be necessary to approach the verification problem from different perspectives in different situations; all situations should yield to a common approach'. We believe that a thorough study of decompositions offers this common approach.

3

Proper Scoring Rules for Interval-Probabilistic Forecasts of Binary Events¹

Summary. For a binary event with outcome $X = 0$ (non-occurrence) or $X = 1$ (occurrence), probabilistic forecasts can be issued as precise-probabilistic forecasts (an exact value for the probability that $X = 1$, $\Pr(X = 1)$ e.g. ‘chance of rain = 0.14%’), or as interval-probabilistic forecasts (a range of values for $\Pr(X = 1)$ e.g. ‘chance of rain: 10-20%’). A common instance of interval-probabilistic forecasts is rounded precise-probabilistic forecasts. To evaluate any probabilistic forecast a scoring rule can be used. An important requirement for scoring rules, if they are to provide a faithful assessment of a forecaster, is that they be proper, by which is meant that they encourage forecasters to issue their true beliefs as their forecasts. Proper scoring rules for precise-probabilistic forecasts i.e. precise-proper scoring rules, have been studied extensively. But, if a precise-proper scoring rule is adapted to interval-probabilistic forecasts, propriety can be lost. Complementing parallel work by other authors, we derive a general characterisation of scoring rules that are proper for interval-probabilistic forecasts i.e. interval-proper scoring rules. From the general characterisation of interval-proper scoring rules, we determine particular scoring rules that correspond to the familiar scoring rules used for precise-probabilistic forecasts. Of these, the most inviting is the interval-Brier scoring rule, which we use to assess interval-forecasts simulated from actual precise-probabilistic forecasts.

[3.1] Introduction

As was emphasised in the previous chapter, any forecast of an uncertain event should be based on a probability distribution of the outcome for the event. A forecast that makes reference to the full probability distribution of the event’s outcome is known as a probabilistic forecast.

In this chapter we restrict attention to binary events: events for which the outcome can take one of two values, 0 and 1 (here, 0 and 1 may be representative); we refer to the unknown outcome of the event, X , as the observation. As X is binary, a full probability distribution of X can be expressed by a single number: for the probability distribution p of X we also write $p = \Pr_p(X = 1)$ for the probability that X takes the value 1 under p . A probabilistic forecast, therefore, may state a value in the interval $[0, 1]$ for p . In this case, because a precise value for p is given, we shall call such a probabilistic forecast a *precise*-probabilistic forecast. One approach to evaluating precise-probabilistic forecasts is to use a scoring rule, which is a function, S , of p and the outcome x of X (Winkler, 1996). The value $S(p, x)$ is the score of the forecast p when the outcome is x . We assume that lower scores are better

¹A version of this chapter was published as [Mitchell and Ferro \(2017\)](#).

(i.e. that scoring rules are negatively oriented (Winkler and Murphy, 1968)). The expected score for a forecast p is the average of $S(p, x)$ over all possible values for x . If the probability distribution of X is q , write $S(p, q) \stackrel{\text{def}}{=} \mathbb{E}[S(p, X)]$ for the expected score. Proper scoring rules are scoring rules for which the expected score $S(p, q)$ is optimised (i.e. minimised) when the forecast matches the assessed distribution of X (i.e. $p = q$). For ease of reference, we shall refer to proper scoring rules for precise-probabilistic forecasts as *precise-proper scoring rules*.

Rather than issuing a precise-probabilistic forecast, a probabilistic forecast of a binary observation is often expressed as a range of probabilities (for example, “Chance of rain: 25-30%”), a practice adopted by meteorological offices around the world. The forecaster will compute their forecast probability precisely but must *issue* a range of probabilities. We call a probabilistic forecast issued as a range of probabilities, an *interval-probabilistic forecast*.

Suppose that the forecaster can articulate their precise true belief, q , that $X = 1$, but must issue an interval of probabilities for the outcome $X = 1$. Let $0 = a_0 < a_1 < \dots < a_{n-1} < a_n = 1$ be a partition of the interval $[0, 1]$, with subintervals $I_i = [a_{i-1}, a_i]$ for $i = 1, \dots, n$. An interval-probabilistic forecast is a choice of I_i for some $0 < i \leq n$.

A scoring rule, s , for an interval-probabilistic forecast, I_i , gives a value $s(I_i, x)$ when the value of X is x . Let q be the actual (and precise) belief q that $X = 1$. Having issued the interval I_i , the forecaster’s expected score, with respect to q , is $s(I_i, q) \stackrel{\text{def}}{=} \mathbb{E}_q[s(I_i, X)]$. A scoring rule for interval-probabilistic forecasts is proper if the interval containing q optimises the expected score and is strictly proper if the only interval that optimises the forecaster’s expected score is the interval that contains q . Assuming that lower values of s indicate better scores, we say, formally, the following.

Definition 3.1.1. (Propriety for Interval Probabilistic Forecasts) Let $X \in \{0, 1\}$, be a random variable, and let the probability q be the forecaster’s actual belief that $X = 1$. The scoring rule, s , is defined to be proper if

$$s(I_i, q) \leq s(I_j, q) \text{ for all } i, j \text{ and } q \in I_i. \quad (3.1)$$

The scoring rule s is strictly proper only if

$$s(I_i, q) < s(I_j, q) \text{ for all } i, j \text{ and } q \in I_i, q \notin I_j. \quad (3.2)$$

□

We refer to scoring rules that are proper for interval-probabilistic forecasts as *interval-proper scoring rules*.

The following lemma shows that an interval-proper scoring rule can be obtained from another interval-proper scoring rule.

Lemma 3.1.2. *Let s be an interval-proper scoring rule and define the interval-probabilistic scoring rule s^* by*

$$s^*(I_k, x) = s(I_k, x) + \psi(x) \quad (3.3)$$

for ψ any function of x not depending on k . Then s^* is an interval-proper scoring rule. \square

Proof. With s^* defined by equation (3.3), and q the probability distribution of X ,

$$s^*(I_k, q) = s(I_k, q) + \mathbb{E}[\psi(X)]. \quad (3.4)$$

Therefore, as ψ does not depend on k , if $q \in I_i$, we have s interval-proper if and only if

$$\begin{aligned} & s(I_i, q) \leq s(I_j, q) \quad \forall j \\ \Leftrightarrow & s^*(I_i, q) - \mathbb{E}[\psi(X)] \leq s^*(I_j, q) - \mathbb{E}[\psi(X)] \quad \forall j \\ \Leftrightarrow & s^*(I_i, q) \leq s^*(I_j, q) \quad \forall j \end{aligned} \quad (3.5)$$

which holds if and only if s^* is interval-proper. \blacksquare

An application of Lemma 3.1.2 requires that an interval-proper scoring rule already be known. There are many possible ways of constructing an interval-probabilistic scoring rule, s , from a given precise-proper scoring rule S (e.g. maximum of S over an interval, average of S over an interval). However, an illustration in the next section shows that even when S is precise-proper and for each i , $s(I_i, X)$ is defined simply as the value of S at the mid-point of I_i , s need not be an interval-proper scoring rule. In response to this difficulty, we present in section 3.3 a general characterisation for any interval-proper scoring rule. This result is a special case of more general results that have been proved by Lambert et al. (2008), Lambert and Shoham (2009), Lambert (2013) and Frongillo and Kash (2014). But, their results, while powerful, are abstract and this has prompted us to offer a short new proof of the characterisation of interval-proper scoring rules. From this general characterisation, we derive, in section 3.4, particular interval-proper scoring rules that are analogues of some familiar precise-proper scoring rules. All interval-proper scoring rules are determined under the assumption that all the forecasts are closed intervals i.e. $I_i = [a_{i-1}, a_i]$ for all i . However, to avoid difficulties arising in practice if a forecaster's precise-probabilistic forecast lies on a boundary of two sub-intervals (e.g. $q = a_i$ for some i), it is more natural for the sub-intervals to be semi-open intervals i.e. $I_1 = [a_0, a_1]$, $I_i = (a_{i-1}, a_i]$ for $i > 1$. We show in section 3.5 that the same characterisation theorem (and, therefore, interval-proper scoring rules) apply

for semi-open forecast intervals. The need for interval-proper scoring rules is illustrated in section 3.6 where we apply our results to the verification of probability of precipitation (PoP) forecasts issued by the Australian Bureau of Meteorology and the United Kingdom Meteorological Office: I'd like to thank the Australian Bureau of Meteorology, in particular, Dr. D. Griffiths and Ms. I. Ioannou, and the UK Meteorological Office, specifically Dr. M. Mittermaier, for their help in making the data available. Section 3.7 concludes, with a brief discussion on which partitions are optimal for a given interval-proper scoring rule.

As a brief aside we note that interval-probabilistic forecasts differ from forecasts expressed as *imprecise-probabilities* (see, for example Coolen (2004) and related works). With imprecise-probabilities a forecaster issues two separate precise probabilities representing the lower and upper bounds of the range of their belief. In contrast, with interval-probabilities a forecaster is presented with a predetermined set of intervals and chooses the interval containing their precise belief.

[3.2] Problems Obtaining Interval-Propriety

Fix the partition $0 = a_0 < a_1 < \dots < a_n = 1$, with $I_i = [a_{i-1}, a_i]$ for $i = 1, \dots, n$. Given a precise-probabilistic scoring rule, S , the scoring rules $s(I_i, x) = \mathbb{E}_P[S(P, x) | P \in I_i]$, and $\hat{s}(I_i, x) = \min_{p \in I_i} S(p, x)$ are immediate suggestions for an interval-proper scoring rule. However, such interval scoring rules are not necessarily proper even if S is a precise-proper scoring rule.

To demonstrate, given $\lambda \in [0, 1]$, let $\hat{p}_i = (1 - \lambda)a_{i-1} + \lambda a_i$ and consider the interval scoring rule

$$s(I_i, x) = S(\hat{p}_i, x). \quad (3.6)$$

It suffices to set $S(p, x) = (p - x)^2$ the (half-)Brier scoring rule (Brier, 1950) (which is proper (see, for example Winkler and Murphy, 1968)); in this case, we refer to s as the Brier λ -scoring rule. Defining $\Delta_i = a_i - a_{i-1}$, Theorem 3.2.3 gives conditions under which the Brier λ -scoring rule is proper. In proving Theorem 3.2.3, the following two lemmas will be useful.

Lemma 3.2.1. *Let*

$$\alpha_{k,j} \stackrel{\text{def}}{=} \frac{\Delta_k + a_k - a_{j-1}}{\Delta_k + \Delta_j}. \quad (3.7)$$

Then $\alpha_{k,k+1} \geq \alpha_{k,j}$ for all $j > k$. □

Proof. If $j = k + 1$, then clearly $\alpha_{k,k+1} \geq \alpha_{k,j}$. So, let $j > k + 1$. Then

$$\begin{aligned}
 & \alpha_{k,k+1} \geq \alpha_{k,j} \\
 \Leftrightarrow & \frac{\Delta_k}{\Delta_k + \Delta_{k+1}} \geq \frac{\Delta_k + a_k - a_{j-1}}{\Delta_j + \Delta_k} \\
 \Leftrightarrow & \frac{\Delta_k(\Delta_j + \Delta_k)}{\Delta_k + \Delta_{k+1}} - \Delta_k \geq a_k - a_{j-1} \\
 \Leftrightarrow & \frac{\Delta_k(\Delta_j - \Delta_{k+1})}{\Delta_k + \Delta_{k+1}} \geq -\Delta_{k+1} + a_{k+1} - a_{j-1} \\
 \Leftrightarrow & \Delta_{k+1} - a_{k+1} + a_{j-1} \geq \frac{\Delta_k(\Delta_{k+1} - \Delta_j)}{\Delta_k + \Delta_{k+1}}. \tag{3.8}
 \end{aligned}$$

Because $\Delta_i > 0$ for all $i = 1, \dots, n$,

$$\begin{aligned}
 \frac{\Delta_k(\Delta_{k+1} - \Delta_j)}{\Delta_k + \Delta_{k+1}} & \leq \frac{\Delta_k \Delta_{k+1}}{\Delta_k + \Delta_{k+1}} \\
 & \leq \Delta_{k+1}. \tag{3.9}
 \end{aligned}$$

If $j > k + 1$, then $j - 1 \geq k + 1$ so $a_{j-1} \geq a_{k+1}$ and $\Delta_{k+1} \leq \Delta_{k+1} - a_{k+1} + a_{j-1}$, so from (3.9),

$$\frac{\Delta_k(\Delta_{k+1} - \Delta_j)}{\Delta_k + \Delta_{k+1}} \leq \Delta_{k+1} - a_{k+1} + a_{j-1} \tag{3.10}$$

and (3.8) is true, from which, $\alpha_{k,k+1} \geq \alpha_{k,j}$. As this holds for all $j > k$, $\max_{j>k} \alpha_{k,j} = \alpha_{k,k+1}$. ■

Lemma 3.2.2. *Let*

$$\beta_{k,j} \stackrel{\text{def}}{=} \frac{a_{k-1} - a_{j-1}}{\Delta_k + \Delta_j}. \tag{3.11}$$

Then $\beta_{k,k-1} \leq \beta_{k,j}$ for all $j < k$. □

Proof. If $j = k - 1$, then there is nothing further to show. So let $j < k - 1$. Then, because

$$\beta_{k,k-1} = \frac{\Delta_{k-1}}{\Delta_k + \Delta_{k-1}} \tag{3.12}$$

we have

$$\beta_{k,k-1} \leq \beta_{k,j} \quad (3.13)$$

$$\begin{aligned} \Leftrightarrow & \frac{\Delta_{k-1}(\Delta_k + \Delta_j)}{\Delta_k + \Delta_{k-1}} \leq a_{k-1} - a_{j-1} \\ \Leftrightarrow & \frac{\Delta_{k-1}(\Delta_k + \Delta_j)}{\Delta_k + \Delta_{k-1}} \leq \Delta_{k-1} + \dots + \Delta_j \end{aligned} \quad (3.14)$$

which holds if and only if

$$0 \leq \Delta_{k-1}^2 + (\Delta_{k-2} + \dots + \Delta_{j+1})(\Delta_k + \Delta_{k-1}) + \Delta_j \Delta_k. \quad (3.15)$$

Inequality (3.15) always holds, as the right-hand side is always positive and therefore $\beta_{k,k-1} \leq \beta_{k,j}$ for all $j < k$. ■

We now state and prove the theorem.

Theorem 3.2.3. *The Brier λ -scoring rule is proper if and only if*

$$\left\{ \begin{array}{ll} \frac{\Delta_1}{\Delta_1 + \Delta_2} \leq \lambda \leq 1 & \text{if } q \in [a_0, a_1], \\ \frac{\Delta_k}{\Delta_k + \Delta_{k+1}} \leq \lambda \leq \frac{\Delta_{k-1}}{\Delta_k + \Delta_{k-1}} & \text{if } q \in [a_{k-1}, a_k] \text{ for } 1 < k < n, \\ 0 \leq \lambda \leq \frac{\Delta_{n-1}}{\Delta_n + \Delta_{n-1}} & \text{if } q \in [a_{n-1}, a_n]. \end{array} \right. \quad (3.16)$$

□

Proof. The Brier λ -scoring rule is proper if and only if $s(I_i, q) \leq s(I_j, q)$ for $q \in I_i$ i.e. $S(\hat{p}_i, q) \leq S(\hat{p}_j, q)$ for $q \in I_i$. For S the Brier scoring rule, we have

$$S(\hat{p}_i, q) = \hat{p}_i^2 - 2\hat{p}_i q + q \quad \text{for all } i. \quad (3.17)$$

Therefore, the Brier λ -scoring rule is proper for $i \neq j$, $q \in I_i$, if and only if, substituting $\hat{p}_i = (1 - \lambda)a_{i-1} + \lambda a_i$, and noting that $\hat{p}_i < \hat{p}_j \Leftrightarrow i < j$,

$$q \leq \frac{1}{2}(a_{i-1} + a_{j-1}) + \frac{\lambda}{2}(\Delta_i + \Delta_j) \quad i < j, \quad (3.18)$$

$$q \geq \frac{1}{2}(a_{i-1} + a_{j-1}) + \frac{\lambda}{2}(\Delta_i + \Delta_j) \quad i > j. \quad (3.19)$$

Equation (3.18) must hold for all $q \in I_i = [a_{i-1}, a_i]$, in particular, for $q = a_i$. It follows that, for propriety, when $i < j$,

$$\begin{aligned}
& a_i \leq \frac{1}{2}(a_{i-1} + a_{j-1}) + \frac{\lambda}{2}(\Delta_i + \Delta_j) \\
\Leftrightarrow & \Delta_i + a_i - a_{j-1} \leq \lambda(\Delta_i + \Delta_j) \\
\Leftrightarrow & \alpha_{i,j} \stackrel{\text{def}}{=} \frac{\Delta_i + a_i - a_{j-1}}{\Delta_i + \Delta_j} \leq \lambda. \tag{3.20}
\end{aligned}$$

By Lemma 3.2.1, $\lambda \geq \alpha_{i,i+1}$. We note that if equation (3.20) holds, then for $q < a_i$, equation (3.18) is true.

Similarly, equation (3.19) must also hold for $q = a_{i-1}$. Then when $i > j$,

$$\begin{aligned}
& a_{i-1} \geq \frac{1}{2}(a_{i-1} + a_{j-1}) + \frac{\lambda}{2}(\Delta_i + \Delta_j) \\
\Leftrightarrow & \beta_{i,j} \stackrel{\text{def}}{=} \frac{a_{i-1} - a_{j-1}}{\Delta_i + \Delta_j} \geq \lambda. \tag{3.21}
\end{aligned}$$

Applying lemma 3.2.2, $\lambda \leq \beta_{i,i-1}$. Again, we note that if equation (3.21) holds, then for $q > a_{i-1}$, equation (3.19) is valid.

If $1 < i < n$ then both $i > j$ and $i < j$ are possible, and propriety exists if and only if $\alpha_{i,i+1} \leq \lambda \leq \beta_{i,i-1}$. If $i = n$, there is no $j > i$, and we can conclude only that $\lambda \leq \beta_{i,i-1} = \beta_{n,n-1}$ (the lower bound on λ being 0). If $i = 1$, there is no $j < i$, set and we can only state that $\lambda \geq \alpha_{i,i+1} = \alpha_{1,2}$ (the upper bound on λ being 1). ■

From equation (3.16), values of q lying in different intervals impose different restrictions on λ . But, for propriety, the *same* value of λ must apply for *every* value of q . A circumstance in which this can occur is when the partition consists of equally-spaced points, as the following corollary to Theorem 3.2.3 proves.

Corollary 3.2.4. *If $\lambda = 1/2$, then the Brier λ -scoring rule is proper if and only if the intervals are determined by equally spaced points.*

Proof. Suppose $\lambda = 1/2$. Then, from Theorem 3.2.3, the Brier λ -scoring rule is proper if and only if

$$\frac{\Delta_k}{\Delta_k + \Delta_{k+1}} \leq \frac{1}{2} \quad \Rightarrow \quad a_k - a_{k-1} \leq a_{k+1} - a_k; \quad k = 1, \dots, n-1; \quad (3.22)$$

$$\frac{\Delta_{k-1}}{\Delta_k + \Delta_{k-1}} \geq \frac{1}{2} \quad \Rightarrow \quad a_k - a_{k-1} \leq a_{k-1} - a_{k-2}; \quad k = 2, \dots, n. \quad (3.23)$$

For $1 \leq k < n$, we have, from equation (3.22)

$$a_k - a_{k-1} \leq a_{k+1} - a_k \quad (3.24)$$

and $k+1$ lies in the set $\{2, \dots, n\}$ so, from equation (3.23),

$$a_{k+1} - a_k \leq a_k - a_{k-1}. \quad (3.25)$$

Equations (3.24) and (3.25) together give $a_{k+1} - a_k = a_k - a_{k-1}$, $1 \leq k < n$, showing that the a_k are equally-spaced. If the a_k are equally spaced then

$$\frac{\Delta_k}{\Delta_k + \Delta_{k+1}} \leq \frac{1}{2} \leq \frac{\Delta_{k-1}}{\Delta_k + \Delta_{k-1}} \quad (3.26)$$

■

As an example, refer to Figures 3.1a and 3.1b. In each figure, the horizontal axis shows q (the forecaster's true belief). The forecaster may issue one of 10 equally-spaced intervals $\left[\frac{i-1}{10}, \frac{i}{10}\right]$ for $i = 1, \dots, 10$. The bounds of the probabilistic intervals are shown on the vertical axis. The value of λ differs between Figures 3.1a and 3.1b. For a given value of λ , the optimal interval, that is, the interval that achieves the lowest expected Brier λ -score, is shown for each value of q . In each figure an optimal interval is displayed as light-grey if q does *not* lie in the interval and dark-grey if q does lie in the interval. The Brier λ -scoring rule is proper if and only if every value of q lies in the optimal interval (which occurs if and only if the 45°-line passes through every optimal interval). Therefore, the Brier λ -scoring rule is proper if and only if no intervals are light-grey. It is evident that only for $\lambda = \frac{1}{2}$ (Figure 3.1b) is the Brier λ -scoring rule proper.

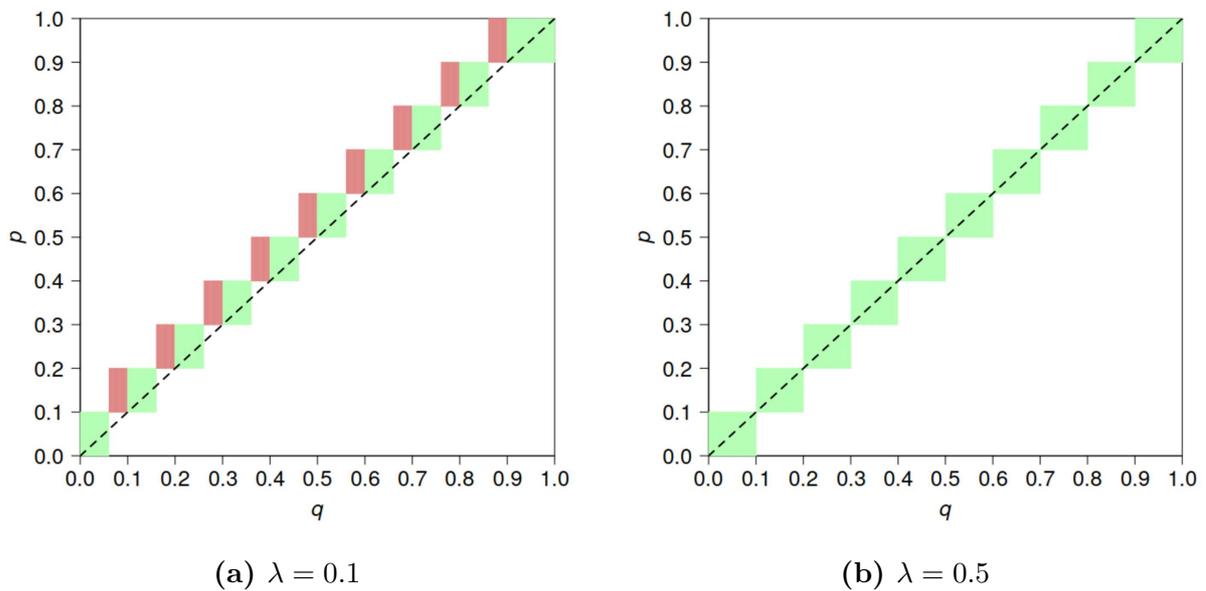


Figure 3.1: For each value of q , the interval forecast that gives the lowest expected Brier λ -score is shown. If q does not lie in an interval (indicating impropriety), the interval is coloured red (■); if q does lie in the interval (demonstrating propriety), the interval is coloured green (□). The Brier λ -scoring rule is, therefore, proper if and only if there are no red intervals. Only for $\lambda = \frac{1}{2}$ is the Brier λ -scoring rule proper.

But, for non-equally spaced intervals, a satisfactory value for λ may not exist; whether the λ -Brier scoring rule is proper, depends on the choice of the partition a_i , $i = 0, \dots, n$, which the forecaster is presented with.

Is there a partition such that for every value of λ , the Brier λ -scoring rule is improper? Yes. Consider a partition such that for some $1 < k < n$, it is the case that $\frac{\Delta_{k-1}}{\Delta_k} < \frac{\Delta_k}{\Delta_{k+1}}$. Then $\frac{\Delta_{k-1}}{\Delta_k + \Delta_{k-1}} < \frac{\Delta_k}{\Delta_{k+1} + \Delta_k}$, so no matter which value is chosen for λ , for each $q \in [a_{k-1}, a_k]$, condition (3.16) is violated and so the Brier λ -scoring rule is improper. There are infinitely many partitions for which the λ -Brier scoring rule is improper for every value of λ . Figure 3.2 gives an example.

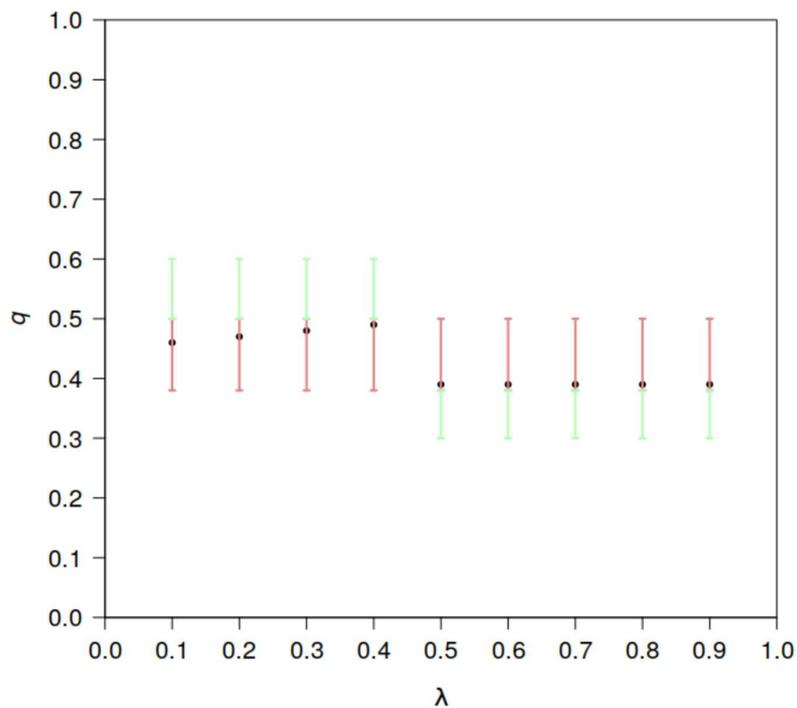


Figure 3.2: The partition is $a_i = \frac{i}{10}$ for $i = 0, \dots, 10$, $i \neq 4$, $a_4 = 0.38$. For each value of λ , the point \bullet indicates a value of $q \in [a_4, a_5]$ for which the interval containing q (red interval) differs from the interval for which the expected Brier λ -score is a minimum (green interval). For the given partition, for every value of λ the interval at which the expected Brier λ -scoring rule is a minimum (---) is different to the interval containing q (---); the Brier λ -scoring rule is improper for all values of λ .

[3.3] Characterisation of Interval-Proper Scoring Rules

We would like to be able to write down the general form of those scoring rules that are proper for interval-probabilistic forecasts. Interval-probabilistic forecasts are a particular example of the wider class of statistical functionals (functions of probability distributions) (see for example, [Gneiting, 2011a](#)). Recently, [Lambert et al. \(2008\)](#), [Lambert and Shoham \(2009\)](#), [Lambert \(2013\)](#) and [Frongillo and Kash \(2014\)](#) have derived a general expression for scoring rules that are proper for statistical functionals (scoring rules that are proper for some particular statistical functionals are given in [Gneiting \(2011a\)](#)). To arrive at a form for scoring rules that are proper for interval-probabilistic forecasts, we can therefore, contextualise these general results, in particular those of [Lambert \(2013\)](#), to our setting. However, for *binary* observations, it is possible to give a straightforward derivation of the form that an interval-proper scoring rule must have.

Let s be a negatively-oriented *strictly* proper scoring rule for interval-probabilistic forecasts. Recall that the expected value of $s(I_i, X)$, with respect to the observation, X , under q , is

$$s(I_i, q) \stackrel{\text{def}}{=} \mathbb{E}[s(I_i, X)] = s(I_i, 0)(1 - q) + s(I_i, 1)q. \quad (3.27)$$

By the propriety of s ,

$$s(I_i, q) \leq s(I_j, q) \quad \text{for all } i, j \text{ and } q \in I_i. \quad (3.28)$$

Condition (3.28) must be satisfied for every $q \in I_i = [a_{i-1}, a_i]$, in particular, for $q = a_i$. Setting $i = k$, $j = k + 1$ and $q = a_k$ in equation (3.28), we have

$$s(I_k, a_k) \leq s(I_{k+1}, a_k). \quad (3.29)$$

Similarly, setting $i = k + 1$, $j = k$ and $q = a_k$ (noting that $a_k \in I_{k+1}$), in equation (3.28),

$$s(I_{k+1}, a_k) \leq s(I_k, a_k). \quad (3.30)$$

From inequalities (3.29) and (3.30),

$$s(I_k, a_k) = s(I_{k+1}, a_k). \quad (3.31)$$

Using equation (3.27), equation (3.31) may be written as

$$\{s(I_k, 0) - s(I_{k+1}, 0)\}(1 - a_k) + \{s(I_k, 1) - s(I_{k+1}, 1)\}a_k = 0 \quad (3.32)$$

and this must hold for every $k = 1, \dots, n - 1$.

One possible solution to (3.32) is the trivial solution $s(I_k, x) = 0$ for all values of k and x . But such a solution violates the *strict* propriety of s : suppose that $i < j$ and choose $q \in I_i$; by strict propriety we should have $s(I_i, q) < s(I_j, q)$, but because $s(I_k, x) = 0$ for all k and x we have $s(I_i, q) = s(I_j, q)$, a contradiction. So, the trivial solution is inadmissible.

Excluding the trivial solution, for each $k = 1, \dots, n - 1$, the solution must then have the form,

$$\begin{aligned} s(I_k, 0) - s(I_{k+1}, 0) &= -a_k \gamma_k \\ s(I_k, 1) - s(I_{k+1}, 1) &= (1 - a_k) \gamma_k \end{aligned} \tag{3.33}$$

where γ_k is a constant.

It is also the case that, by the propriety of s ,

$$s(I_k, a_{k-1}) \leq s(I_{k+1}, a_{k-1})$$

from which

$$\{s(I_k, 0) - s(I_{k+1}, 0)\}(1 - a_{k-1}) + \{s(I_k, 1) - s(I_{k+1}, 1)\}a_{k-1} \leq 0. \tag{3.34}$$

Substituting the relationships in (3.33) into equation (3.34) gives

$$\begin{aligned} -a_k \gamma_k (1 - a_{k-1}) + (1 - a_k) \gamma_k a_{k-1} &\leq 0 \\ \Leftrightarrow \gamma_k (a_{k-1} - a_k) &\leq 0 \\ \Leftrightarrow \gamma_k &\geq 0. \end{aligned} \tag{3.35}$$

We can, therefore, write

$$s(I_k, x) - s(I_{k+1}, x) = \gamma_k (x - a_k) \quad \text{for } k = 1, \dots, n - 1 \tag{3.36}$$

for non-negative constants γ_k . The difference equation (3.36) has a solution

$$s(I_k, x) = f(x) - \sum_{i=1}^{k-1} \gamma_i (x - a_i) \tag{3.37}$$

with f an arbitrary function of x . Defining the function g by $g(i) - g(i - 1) = \gamma_i$, $i = 1, \dots, n - 1$, we have proved the following theorem.

Theorem 3.3.1. (*Characterisation for Interval-Proper Scoring Rules*) Let $X \in \{0, 1\}$ be a

binary observation. Given a partition $0 = a_0 < a_1 < \dots < a_{n-1} < a_n = 1$, let s be a strictly interval-proper scoring rule for interval-probabilistic forecasts $I_i = [a_{i-1}, a_i]$ and $i = 1, \dots, n$ of the outcome $X = 1$. Then s has the form

$$s(I_k, x) = f(x) - \sum_{i=1}^{k-1} (g(i) - g(i-1))(x - a_i) \quad (3.38)$$

where f is an arbitrary function and g is a non-decreasing function. \square

Under s given by equation (3.38), interval-probabilistic forecasts that are closer to the outcome for X receive a lower (that is, better) score than interval-probabilistic forecasts that are further from the outcome for X . For, suppose that $X = 0$. We have

$$s(I_k, 0) = f(0) + \sum_{i=1}^{k-1} (g(i) - g(i-1))a_i \quad (3.39)$$

and the summation term increases as k increases (g being a non-decreasing function) so that as I_k moves further away from 0 (as k increases) $s(I_k, 0)$ increases. Similarly, if $X = 1$,

$$s(I_k, 1) = f(1) - \sum_{i=1}^{k-1} (g(i) - g(i-1))(1 - a_i) \quad (3.40)$$

and the summation term is always positive and decreases in size as k decreases so that $s(I_k, 1)$ increases as I_k moves away from 1.

3.3.1 || Choosing f and g

In equation (3.38), each choice for the function f and for the non-decreasing function g , will give a new proper scoring rule for interval-probabilistic forecasts. While *any* real-valued function may be chosen for f and *any* non-decreasing real-valued function may be chosen for g , it is helpful to have some method to guide these choices. Here we suggest one such method.

To begin, choose a non-decreasing function $h : [0, 1] \rightarrow \mathbb{R}$, and define the function g by $g(k) = h(\xi_{k+1})$ for $\xi_k \in I_k$ for $k = 1, \dots, n$; each choice of ξ_k may lead to a different g . Replacing g by h in (3.38),

$$s(I_k, x) = f(x) - \sum_{i=1}^{k-1} (h(\xi_{i+1}) - h(\xi_i))(x - a_i) \quad (3.41)$$

from which,

$$s(I_{k+1}, x) - s(I_k, x) = (h(\xi_{k+1}) - h(\xi_k))(a_k - x). \quad (3.42)$$

We now show that for a suitable precise-proper scoring rule S , we can conclude from equation

(3.42) that in the limit of ever finer partitions, the differential equation

$$\frac{\partial S(p, x)}{\partial p} = (p - x) \frac{dh(p)}{dp} \quad (3.43)$$

holds. By choosing S , equation (3.43) can be solved for h to allow g to be found, by setting $g(k) = h(\xi_{k+1})$ (for some predetermined choice for the ξ_{k+1}).

To derive differential equation (3.43) from equation (3.42), a formal notion of convergence is required. Some of the terminology that follows will be familiar from standard analysis (see, for example [Browder, 1996](#)).

Definition 3.3.2. For each $n \in \mathbb{N}$, let $[a]_n$ be the partition $0 = a_{n,0} < a_{n,1} < \dots < a_{n,n} = 1$. For $n > m$, the partition $[a]_n$ is a refinement of the partition $[a]_m$, written $[a]_m \leq [a]_n$, if $\{a_{m,0}, \dots, a_{m,m}\} \subset \{a_{n,0}, \dots, a_{n,n}\}$. The partitions $\{[a]_n\}_{n \in \mathbb{N}}$ are said to be increasingly refined if $[a]_n \leq [a]_{n+1}$ for all n and the mesh, $\mu_n = \max\{a_{n,i} - a_{n,i-1} \mid i = 1, \dots, n\}$ tends to 0 as $n \rightarrow \infty$. \square

We will use the notation $I_{n,i} = [a_{n,i-1}, a_{n,i}]$, $i = 1, \dots, n$.

Lemma 3.3.3. Let $p \in [0, 1]$. If the partitions $\{[a]_n\}_{n \in \mathbb{N}}$ are increasingly refined then $\forall \epsilon > 0$, $\exists N \geq 0$ such that for each $n > N$, there is a k (depending on n) such that $|a_{n,k} - p| < \epsilon$. \square

Proof. Fix $\epsilon > 0$. Since the partitions $\{[a]_n\}_{n \in \mathbb{N}}$ are increasingly refined, there is an $N \geq 0$ such that $\forall n > N$, $\mu_n < \epsilon$. Let $n > N$ so that $\mu_n < \epsilon$. If $p \in [0, 1]$ then there is some k such that $p \in I_{n,k}$. Therefore, $|a_{n,k} - p| \leq |a_{n,k} - a_{n,k-1}| \leq \mu_n < \epsilon$. So for all $n > N$, there exists a k (depending on n) such that $|a_{n,k} - p| < \epsilon$. \blacksquare

Definition 3.3.4. The interval-proper scoring rule s converges in the Lipschitz sense to the precise-proper scoring rule S at p if and only if $\forall \epsilon > 0$, $\exists N \geq 0$ such that for all $n \geq N$, $|s(I_{n,k}, x) - S(p, x)| < \epsilon \min\{|a_{n,k-1} - p|, |a_{n,k} - p|\}$ for all x , and for all $p \in I_{n,k}$. If s converges to S in the Lipschitz sense at every $p \in [0, 1]$, then we shall say simply that s converges to S in the Lipschitz sense. \square

With these definitions we can state and prove the following proposition,

Proposition 3.3.5. Let s be an interval-proper scoring rule satisfying equation (3.42), with h continuously differentiable. Suppose that s converges to the precise-proper scoring rule S in the Lipschitz sense, where S is continuously partially differentiable with respect to p . If the partitions $\{[a]_n\}_{n \in \mathbb{N}}$ are increasingly refined then

$$\frac{\partial S(p, x)}{\partial p} = \frac{dh(p)}{dp} (p - x). \quad (3.44)$$

\square

Proof. From equation (3.42), for any $\xi_{n,k} \in I_{n,k}$,

$$\frac{s(I_{n,k+1}, x) - s(I_{n,k}, x)}{\xi_{n,k+1} - \xi_{n,k}} = \left(\frac{h(\xi_{n,k+1}) - h(\xi_{n,k})}{\xi_{n,k+1} - \xi_{n,k}} \right) (a_{n,k} - x). \quad (3.45)$$

Let $\epsilon > 0$, $p \in [0, 1]$. Because S is partially differentiable with respect to p , $\exists \delta_* > 0$ such that for $|\xi_{n,k+1} - \xi_{n,k}| < \delta_*$,

$$\left| \frac{S(\xi_{n,k+1}, x) - S(\xi_{n,k}, x)}{\xi_{n,k+1} - \xi_{n,k}} - \frac{\partial S(\xi_{n,k}, x)}{\partial \xi_{n,k}} \right| < \frac{\epsilon}{4} \quad (3.46)$$

and as S is continuously partial differentiable with respect to p , $\exists \delta' > 0$ such that if $|\xi_{n,k} - p| < \delta'$,

$$\left| \frac{\partial S(\xi_{n,k}, x)}{\partial \xi_{n,k}} - \frac{\partial S(p, x)}{\partial p} \right| < \frac{\epsilon}{4}. \quad (3.47)$$

Further, since h is differentiable, $\exists \delta_{**}$ such that $\forall |\xi_{n,k+1} - \xi_{n,k}| < \delta_{**}$,

$$\left| \frac{h(\xi_{n,k+1}) - h(\xi_{n,k})}{\xi_{n,k+1} - \xi_{n,k}} - \frac{dh(\xi_{n,k})}{d\xi_{n,k}} \right| < \frac{\epsilon}{2} \quad (3.48)$$

and, by the continuous differentiability of h , $\exists \delta'' > 0$ such that if $|\xi_{n,k} - p| < \delta''$, then

$$\left| \frac{dh(\xi_{n,k})}{d\xi_{n,k}} - \frac{dh(p)}{dp} \right| < \frac{\epsilon}{2}. \quad (3.49)$$

Let $\delta = \min\{\delta_*, \delta', \delta_{**}, \delta'', \epsilon\}$.

As the partitions $\{[a]_n\}_{n \in \mathbb{N}}$ are increasingly refined, $\exists N_* \geq 0$ such that for $n \geq N_*$, $\mu_n < \frac{\delta}{2}$. The interval scoring rule s converges to S in the Lipschitz sense, so $\exists N' \geq 0$ such that $\forall n > N'$,

$$|s(I_{n,j}, x) - S(p, x)| < \frac{\epsilon}{4} \min\{|a_{n,j-1} - p|, |a_{n,j} - p|\} \quad (3.50)$$

for $p \in I_{n,j}$.

Considering the left-hand side of equation (3.50), for any p ,

$$\begin{aligned}
& \left| \frac{s(I_{n,k+1}, x) - s(I_{n,k}, x)}{\xi_{n,k+1} - \xi_{n,k}} - \frac{\partial S(p, x)}{\partial p} \right| \\
&= \left| \frac{s(I_{n,k+1}, x) - S(\xi_{n,k+1}, x)}{\xi_{n,k+1} - \xi_{n,k}} - \left(\frac{s(I_{n,k}, x) - S(\xi_{n,k}, x)}{\xi_{n,k+1} - \xi_{n,k}} \right) + \frac{S(\xi_{n,k+1}, x) - S(\xi_{n,k}, x)}{\xi_{n,k+1} - \xi_{n,k}} \right. \\
&\quad \left. - \frac{\partial S(\xi_{n,k}, x)}{\partial \xi_{n,k}} + \frac{\partial S(\xi_{n,k}, x)}{\partial \xi_{n,k}} - \frac{\partial S(p, x)}{\partial p} \right| \\
&\leq \left| \frac{s(I_{n,k+1}, x) - S(\xi_{n,k+1}, x)}{\xi_{n,k+1} - \xi_{n,k}} \right| + \left| \frac{s(I_{n,k}, x) - S(\xi_{n,k}, x)}{\xi_{n,k+1} - \xi_{n,k}} \right| \\
&\quad + \left| \frac{S(\xi_{n,k+1}, x) - S(\xi_{n,k}, x)}{\xi_{n,k+1} - \xi_{n,k}} - \frac{\partial S(\xi_{n,k}, x)}{\partial \xi_{n,k}} \right| + \left| \frac{\partial S(\xi_{n,k}, x)}{\partial \xi_{n,k}} - \frac{\partial S(p, x)}{\partial p} \right|. \quad (3.51)
\end{aligned}$$

Let $N = \max\{N_*, N'\}$, $n \geq N$ and k (depending on n) satisfy $p \in I_{n,k}$. We have $\xi_{n,k+1} - \xi_{n,k} \leq a_{n,k+1} - a_{n,k-1} \leq 2\mu_n < \delta$, $|\xi_{n,k} - p| < \mu_n < \delta$,

$$\frac{\min\{|a_{n,k+1} - \xi_{n,k+1}|, |a_{n,k} - \xi_{n,k+1}|\}}{\xi_{n,k+1} - \xi_{n,k}} \leq \frac{|a_{n,k} - \xi_{n,k+1}|}{\xi_{n,k+1} - \xi_{n,k}} \leq 1 \quad (3.52)$$

and

$$\frac{\min\{|a_{n,k} - \xi_{n,k}|, |a_{n,k-1} - \xi_{n,k}|\}}{\xi_{n,k+1} - \xi_{n,k}} \leq \frac{|a_{n,k} - \xi_{n,k}|}{\xi_{n,k+1} - \xi_{n,k}} \leq 1. \quad (3.53)$$

It follows from (3.45), (3.46), (3.49) and (3.51) to (3.53) that

$$\begin{aligned}
& \left| \frac{s(I_{n,k+1}, x) - s(I_{n,k}, x)}{\xi_{n,k+1} - \xi_{n,k}} - \frac{\partial S(p, x)}{\partial p} \right| \\
&< \frac{\epsilon \min\{|a_{n,k+1} - \xi_{n,k+1}|, |a_{n,k} - \xi_{n,k+1}|\}}{4 \xi_{n,k+1} - \xi_{n,k}} \\
&\quad + \frac{\epsilon \min\{|a_{n,k} - \xi_{n,k}|, |a_{n,k-1} - \xi_{n,k}|\}}{4 \xi_{n,k+1} - \xi_{n,k}} + \frac{\epsilon}{4} + \frac{\epsilon}{4} \\
&< \frac{\epsilon}{4} + \frac{\epsilon}{4} + \frac{\epsilon}{4} + \frac{\epsilon}{4} \\
&= \epsilon. \quad (3.54)
\end{aligned}$$

Similarly, using (3.47) and (3.48), from the right-hand side of equation (3.50),

$$\begin{aligned}
& \left| \frac{h(\xi_{n,k+1}) - h(\xi_{n,k})}{\xi_{n,k+1} - \xi_{n,k}} - \frac{dh(p)}{dp} \right| \\
&= \left| \frac{h(\xi_{n,k+1}) - h(\xi_{n,k})}{\xi_{n,k+1} - \xi_{n,k}} - \frac{dh(\xi_{n,k})}{d\xi_{n,k}} + \frac{dh(\xi_{n,k})}{d\xi_{n,k}} - \frac{dh(p)}{dp} \right| \\
&\leq \left| \frac{h(\xi_{n,k+1}) - h(\xi_{n,k})}{\xi_{n,k+1} - \xi_{n,k}} - \frac{dh(\xi_{n,k})}{d\xi_{n,k}} \right| + \left| \frac{dh(\xi_{n,k})}{d\xi_{n,k}} - \frac{dh(p)}{dp} \right| \\
&< \frac{\epsilon}{2} + \frac{\epsilon}{2} \\
&= \epsilon
\end{aligned} \tag{3.55}$$

and,

$$|(x - a_{n,k}) - (x - p)| = |p - a_{n,k}| \leq \mu_n < \frac{\delta}{2} \leq \epsilon. \tag{3.56}$$

Combining the limit results (3.54) to (3.56), equation (3.50) gives

$$\frac{\partial S(p, x)}{\partial p} = -\frac{dh(p)}{dp}(x - p). \tag{3.57}$$

■

On integrating both sides of (3.57) with respect to p , we obtain

$$S(p, x) + a(x) = h(p)(p - x) - \int h(p) dp \tag{3.58}$$

where $a(\cdot)$ is a function of x alone. Taking the expectation of equation (3.58) in X under p gives

$$S(p, p) + \mathbb{E}[a(X)] = - \int h(p) dp. \tag{3.59}$$

With $X \in \{0, 1\}$, we can write $a(X) = a(0)(1 - X) + a(1)X$ so that $\mathbb{E}[a(X)] = a(0)(1 - p) + a(1)p$. The function $e_S(p) = -S(p, p)$ is known as the entropy of p associated with S (Bröcker, 2009; Gneiting and Raftery, 2007). We have

$$\int h(p) dp = e_S(p) - a(0)(1 - p) - a(1)p. \tag{3.60}$$

Differentiating both sides of (3.60) with respect to p ,

$$h(p) = \frac{de_S(p)}{dp} - (a(1) - a(0)). \tag{3.61}$$

Equation (3.61) indicates that $h(p)$ is (up to a constant), the derivative of the entropy of p associated with S . This property of h was suggested to us by an anonymous referee of the version of this chapter submitted for publication, and, in light of this property, the referee

proposed we consider the limit of equation (3.41), which we now do, showing that equation (3.41) is a discrete version of a well-known result from the theory of precise-proper scoring rules. The core of our argument is contained in the following lemma.

Lemma 3.3.6. *Suppose that h is continuously differentiable on $(0, 1)$ and $(x - q)dh(q)/dq$ is bounded on $[0, 1]$. Given $p \in [0, 1]$, for each n , let k satisfy, $p \in I_{n,k}$. Then*

$$\lim_{n \rightarrow \infty} \sum_{i=1}^{n-1} \mathbb{1}_{[0, a_{n,k}]}(a_{n,i})(x - a_{n,i})(h(a_{n,i}) - h(a_{n,i-1})) = \int_0^1 \mathbb{1}_{[0,p]}(r)(x - r)dh(r) \quad (3.62)$$

($\mathbb{1}_A(\cdot)$ is the indicator function of the set A , taking the value 1 if its argument is in A , and 0 otherwise). \square

Proof. Let $p \in [0, 1]$. If $p = 0$ then both sides of equation (3.62) are zero. Suppose, therefore, that $p > 0$, and for each n , fix k such that $p \in (a_{n,k-1}, a_{n,k}]$. We have $p \leq a_{n,k}$ and if $a_{n,i} \geq p$ then $a_{n,i} \geq a_{n,k}$ and if $a_{n,i} < p$ then $a_{n,i} < a_{n,k}$. Therefore, $\mathbb{1}_{[0,p]}(a_{n,i}) = \mathbb{1}_{[0, a_{n,k}]}(a_{n,i})$ and this holds for all $i = 1, \dots, n-1$. For k , therefore

$$\sum_{i=1}^{n-1} \mathbb{1}_{[0, a_{n,k}]}(a_{n,i})(x - a_{n,i})(h(a_{n,i}) - h(a_{n,i-1})) = \sum_{i=1}^{n-1} \mathbb{1}_{[0,p]}(a_{n,i})(x - a_{n,i})(h(a_{n,i}) - h(a_{n,i-1})). \quad (3.63)$$

Consider the absolute difference

$$\begin{aligned} & \left| \sum_{i=1}^{n-1} \mathbb{1}_{[0,p]}(a_{n,i})(x - a_{n,i})(h(a_{n,i}) - h(a_{n,i-1})) - \int_0^1 \mathbb{1}_{[0,p]}(r)(x - r) dh(r) \right| \\ &= \left| \sum_{i=1}^{n-1} \mathbb{1}_{[0,p]}(a_{n,i})(x - a_{n,i}) \left(\frac{h(a_{n,i}) - h(a_{n,i-1})}{a_{n,i} - a_{n,i-1}} - \frac{dh(a_{n,i})}{da_{n,i}} + \frac{dh(a_{n,i})}{da_{n,i}} \right) (a_{n,i} - a_{n,i-1}) \right. \\ & \quad \left. - \int_0^1 \mathbb{1}_{[0,p]}(r)(x - r) dh(r) \right| \\ &\leq \left| \sum_{i=1}^{n-1} \mathbb{1}_{[0,p]}(a_{n,i})(x - a_{n,i}) \left(\frac{h(a_{n,i}) - h(a_{n,i-1})}{a_{n,i} - a_{n,i-1}} - \frac{dh(a_{n,i})}{da_{n,i}} \right) (a_{n,i} - a_{n,i-1}) \right| \\ & \quad + \left| \sum_{i=1}^{n-1} \mathbb{1}_{[0,p]}(a_{n,i})(x - a_{n,i}) \frac{dh(a_{n,i})}{da_{n,i}} (a_{n,i} - a_{n,i-1}) - \int_0^1 \mathbb{1}_{[0,p]}(r)(x - r) \frac{dh(r)}{dr} dr \right|. \quad (3.64) \end{aligned}$$

Fix $\epsilon > 0$. From the continuous differentiability of h on $(0, 1)$ (and since $a_{n,i} \in (0, 1)$), $\exists N$ such that for $n > N$,

$$\left| \frac{h(a_{n,i}) - h(a_{n,i-1})}{a_{n,i} - a_{n,i-1}} - \frac{dh(a_{n,i})}{da_{n,i}} \right| < \frac{\epsilon}{2}. \quad (3.65)$$

Therefore, for $n > N$,

$$\begin{aligned}
 & \left| \sum_{i=1}^{n-1} \mathbb{1}_{[0,p)}(a_{n,i})(x - a_{n,i}) \left(\frac{h(a_{n,i}) - h(a_{n,i-1})}{a_{n,i} - a_{n,i-1}} - \frac{dh(a_{n,i})}{da_{n,i}} \right) (a_{n,i} - a_{n,i-1}) \right| \\
 & \leq \sum_{i=1}^{n-1} \left| \mathbb{1}_{[0,p)}(a_{n,i})(x - a_{n,i}) \left(\frac{h(a_{n,i}) - h(a_{n,i-1})}{a_{n,i} - a_{n,i-1}} - \frac{dh(a_{n,i})}{da_{n,i}} \right) (a_{n,i} - a_{n,i-1}) \right| \\
 & = \sum_{i=1}^{n-1} \left| \mathbb{1}_{[0,p)}(a_{n,i})(x - a_{n,i})(a_{n,i} - a_{n,i-1}) \right| \left| \frac{h(a_{n,i}) - h(a_{n,i-1})}{a_{n,i} - a_{n,i-1}} - \frac{dh(a_{n,i})}{da_{n,i}} \right| \\
 & < \frac{\epsilon}{2} \sum_{i=1}^{n-1} \left| \mathbb{1}_{[0,p)}(a_{n,i})(x - a_{n,i})(a_{n,i} - a_{n,i-1}) \right| \\
 & \leq \frac{\epsilon}{2} \sum_{i=1}^{n-1} |(a_{n,i} - a_{n,i-1})| \\
 & \leq \frac{\epsilon}{2} \sum_{i=1}^n |(a_{n,i} - a_{n,i-1})| \\
 & = \frac{\epsilon}{2}.
 \end{aligned} \tag{3.66}$$

Also, setting

$$\phi(a_{n,i}) = \mathbb{1}_{[0,p)}(a_{n,i})(x - a_{n,i}) \frac{dh(a_{n,i})}{da_{n,i}} \tag{3.67}$$

by the continuous differentiability of h , ϕ is continuous except at the single point p and the points 0 and 1, and, as $(x - q)dh(q)/dq$ is bounded on $[0, 1)$ and $\phi(1) = 0$, then ϕ is bounded on $[0, 1]$. It is the case, then, that ϕ is Riemann-integrable on $[0, 1]$, and $\exists N'$ such that for $n > N'$,

$$\left| \sum_{i=1}^n \phi(a_{n,i})(a_{n,i} - a_{n,i-1}) - \int_0^1 \mathbb{1}_{[0,p)}(r)(x - r) \frac{dh(r)}{dr} dr \right| < \frac{\epsilon}{2}. \tag{3.68}$$

But, because $\phi(a_{n,n}) = \phi(1) = 0$,

$$\sum_{i=1}^n \phi(a_{n,i})(a_{n,i} - a_{n,i-1}) = \sum_{i=1}^{n-1} \phi(a_{n,i})(a_{n,i} - a_{n,i-1}) \tag{3.69}$$

so that

$$\left| \sum_{i=1}^{n-1} \phi(a_{n,i})(a_{n,i} - a_{n,i-1}) - \int_0^1 \mathbb{1}_{[0,p)}(r)(x - r) \frac{dh(r)}{dr} dr \right| < \frac{\epsilon}{2}. \tag{3.70}$$

Letting $n > \max\{N, N'\}$, from equation (3.64),

$$\left| \sum_{i=1}^{n-1} \mathbb{1}_{[0,p)}(a_{n,i})(x - a_{n,i})(h(a_{n,i}) - h(a_{n,i-1})) - \int_0^1 \mathbb{1}_{[0,p)}(r)(x - r) dh(r) \right| < \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon \tag{3.71}$$

giving the result (3.62). ■

Lemma 3.3.6 does not consider the case that $(x - q)dh(q)/dq$ is bounded on $(0, 1]$, for which a separate lemma is required, and is now given.

Lemma 3.3.7. *Suppose that h is continuously differentiable on $(0, 1)$ and $(x - q)dh(q)/dq$ is bounded on $(0, 1]$. Given $p \in [0, 1]$, for each n , let k satisfy, $p \in I_{n,k}$. Then*

$$\lim_{n \rightarrow \infty} \sum_{i=1}^{n-1} \mathbf{1}_{[a_{n,k}, 1)}(a_{n,i})(x - a_{n,i})(h(a_{n,i}) - h(a_{n,i-1})) = \int_0^1 \mathbf{1}_{[p, 1)}(r)(x - r)dh(r). \quad (3.72)$$

□

Proof. Let $p \in [0, 1]$. If $p = 1$ then both sides of equation (3.72) are zero. If $p = 0$ then the result follows from

$$\lim_{n \rightarrow \infty} \sum_{i=1}^{n-1} (x - a_{n,i})(h(a_{n,i}) - h(a_{n,i-1})) = \int_0^1 (x - r)dh(r). \quad (3.73)$$

Consider, then the case of $0 < p < 1$. For each n , fix k such that $p \in (a_{n,k-1}, a_{n,k}]$. If $a_{n,i} \geq a_{n,k}$, then $a_{n,i} > p$ and if $a_{n,i} < p$, then $a_{n,i} < a_{n,k}$. It follows that $\mathbf{1}_{[a_{n,k}, 1)}(a_{n,i}) = \mathbf{1}_{[p, 1)}(a_{n,i})$ for all $i = 1, \dots, n - 1$, and we have

$$\sum_{i=1}^{n-1} \mathbf{1}_{[a_{n,k}, 1)}(a_{n,i})(x - a_{n,i})(h(a_{n,i}) - h(a_{n,i-1})) = \sum_{i=1}^{n-1} \mathbf{1}_{[p, 1)}(a_{n,i})(x - a_{n,i})(h(a_{n,i}) - h(a_{n,i-1})). \quad (3.74)$$

The remainder of the proof is the same as that of Lemma 3.3.6, with the exception that we define ϕ by

$$\phi(a_{n,i}) = \mathbf{1}_{[p, 1)}(a_{n,i})(x - a_{n,i}) \frac{dh(a_{n,i})}{da_{n,i}}. \quad (3.75)$$

By assumption, h is continuously differentiable on $(0, 1)$ so that ϕ is too other than perhaps at 0 and 1. It is also the case that $(x - q)dh(q)/dq$ is bounded on $(0, 1]$ and because $p > 0$, $\phi(0) = 0$ so ϕ is bounded on $[0, 1]$. With these properties, ϕ is Riemann-integrable on $[0, 1]$. With the additional property that $\phi(a_{n,n}) = \phi(1) = 0$ by construction, the result follows as in the proof of Lemma 3.3.6. ■

From equation (3.41), for the partition $[a]_n$ and for all $k = 1, \dots, n$, with $\xi_{n,k} \in I_{n,k}$ we have

$$\begin{aligned}
 s(I_{n,k}, x) &= f(x) - \sum_{i=1}^{k-1} \left(h(\xi_{n,i+1}) - h(\xi_{n,i}) \right) (x - a_{n,i}) \\
 &= f(x) - \sum_{i=1}^{n-1} (x - a_{n,i}) \mathbb{1}_{[0, a_{n,k}]}(a_{n,i}) \left(h(\xi_{n,i+1}) - h(\xi_{n,i}) \right). \tag{3.76}
 \end{aligned}$$

Equation (3.76) holds for all choices of $\xi_{n,k} \in I_{n,k}$ and so holds for $\xi_{n,k} = a_{n,k-1}$, allowing us to write

$$s(I_{n,k}, x) = f(x) - \sum_{i=1}^{n-1} (x - a_{n,i}) \mathbb{1}_{[0, a_{n,k}]}(a_{n,i}) \left(h(a_{n,i}) - h(a_{n,i-1}) \right). \tag{3.77}$$

By Lemma (3.3.6), the right-hand side of equation (3.77) converges, for suitable h , given $p \in I_{n,k}$, to

$$f(x) - \int_0^1 \mathbb{1}_{[0,p]}(r) (x - r) dh(r) \tag{3.78}$$

and, for S such that s converges to S in the Lipschitz sense, $s(I_{n,k}, x) \rightarrow S(p, x)$ as $n \rightarrow \infty$. Therefore, taking the limit of equation (3.41) as $n \rightarrow \infty$,

$$S(p, x) = f(x) - \int_0^1 \mathbb{1}_{[0,p]}(r) (x - r) dh(r). \tag{3.79}$$

Similarly, from equation (3.41),

$$\begin{aligned}
 s(I_{n,k}, x) &= f(x) - \sum_{i=1}^{k-1} \left(h(\xi_{n,i+1}) - h(\xi_{n,i}) \right) (x - a_{n,i}) \\
 &= f(x) - \left(\sum_{i=1}^{n-1} \left(h(\xi_{n,i+1}) - h(\xi_{n,i}) \right) (x - a_{n,i}) - \sum_{i=k}^{n-1} \left(h(\xi_{n,i+1}) - h(\xi_{n,i}) \right) (x - a_{n,i}) \right) \\
 &= f(x) - s(I_{n,n}, x) + \sum_{i=k}^{n-1} \left(h(\xi_{n,i+1}) - h(\xi_{n,i}) \right) (x - a_{n,i}). \tag{3.80}
 \end{aligned}$$

Replacing $\xi_{n,i}$ with $a_{n,i-1}$ in equation (3.80), we obtain

$$\begin{aligned}
 s(I_{n,k}, x) &= f(x) - s(I_{n,n}, x) + \sum_{i=k}^{n-1} \left(h(a_{n,i}) - h(a_{n,i-1}) \right) (x - a_{n,i}) \\
 &= f(x) - s(I_{n,n}, x) + \sum_{i=1}^{n-1} \mathbb{1}_{[a_{n,k}, 1]}(a_{n,i}) (x - a_{n,i}) \left(h(a_{n,i}) - h(a_{n,i-1}) \right). \tag{3.81}
 \end{aligned}$$

Letting $n \rightarrow \infty$, equation (3.81) gives (referring to Lemma 3.3.7 with appropriate h),

$$S(p, x) = f(x) - S(1, x) + \int_0^1 \mathbb{1}_{[p,1]}(r) (x - r) dh(r). \tag{3.82}$$

We can compare equations (3.79) and (3.82) to the Schervish-representation of a proper scoring rule for a binary event (Schervish (1989), Theorem 4.2, page 1861; see also Gneiting and Raftery (2007), page 364),

$$S(p, x) = \begin{cases} S(0, 0) + \int_0^1 \mathbb{1}_{[0,p)}(r)r \, dh(r) & \text{if } x = 0, \\ S(1, 1) + \int_0^1 \mathbb{1}_{[p,1)}(1-r) \, dh(r) & \text{if } x = 1. \end{cases} \quad (3.83a)$$

$$(3.83b)$$

Equations (3.79) and (3.83a) together and equations (3.82) and (3.83b) together suggest that f should be set to

$$f(x) = (x + 1)S(x, x). \quad (3.84)$$

[3.4] Examples

As examples of this method we take some familiar precise-proper scoring rules and derive the corresponding analogues that are interval-proper. In all cases, we assume that the precise probabilistic forecast that $X = 1$ is $p \in [0, 1]$ and that the interval $[0, 1]$ has partition $[a]_n : 0 = a_0 < a_1 < \dots < a_n = 1$.

3.4.1 || Interval-Brier scoring rule

The (half-)Brier scoring rule (Brier, 1950) is $S(p, x) = (p - x)^2$. Substituting for S in equation (3.43), we have

$$2(p - x) = (p - x) \frac{dh(p)}{dp} \quad (3.85)$$

giving $h(p) = 2p$. Identify points $\xi_k \in I_k = [a_{k-1}, a_k]$ for all $k = 1, \dots, n$. Then $g(k) = h(\xi_{k+1}) = 2\xi_{k+1}$. Choose $f(x) = (x + 1)S(x, x) = 0$.

With these choices of f and g , equation (3.41) gives the following scoring rule for interval-probabilistic forecasts

$$s(I_k, x) = 0 - \sum_{i=1}^{k-1} (2\xi_{i+1} - 2\xi_i) (x - a_i) \quad (3.86)$$

which may be rewritten as

$$\begin{aligned}
s(I_k, X) &= -(\xi_1 - x)^2 + (\xi_k - x)^2 - \sum_{i=1}^{k-1} \{(\xi_{i+1} - a_i)^2 - (\xi_i - a_i)^2\} \\
&= S(\xi_k, x) - S(\xi_1, x) - \sum_{i=1}^{k-1} \{S(\xi_{i+1}, a_i) - S(\xi_i, a_i)\}.
\end{aligned} \tag{3.87}$$

Choosing $\xi_k = (a_k + a_{k-1})/2$ (the mid-point of each subinterval), then

$$s(I_k, x) = \left(\frac{1}{2}(a_{k-1} + a_k) - x\right)^2 - \frac{1}{4}(a_k - a_{k-1})^2 + (a_1 x - x^2) \tag{3.88}$$

$$= (x - a_{k-1})(x - a_k) + (a_1 x - x^2). \tag{3.89}$$

Define

$$s^*(I_k, x) = s(I_k, x) - (a_1 x - x^2) = (x - a_{k-1})(x - a_k). \tag{3.90}$$

Then, by Lemma 3.1.2, s^* is an interval-proper scoring rule, which we refer to as the interval-Brier scoring rule.

We note too, that if the partition is *equally-spaced*, then, letting $\Delta = a_k - a_{k-1}$ for all $k = 1, \dots, n$, equation (3.88) becomes

$$\begin{aligned}
s(I_k, x) &= \left(\frac{1}{2}(a_{k-1} + a_k) - x\right)^2 - \frac{1}{4}\Delta^2 + \Delta x - x^2 \\
&= \left(\frac{1}{2}(a_{k-1} + a_k) - x\right)^2 - \left(x - \frac{\Delta}{2}\right)^2
\end{aligned} \tag{3.91}$$

so that, defining

$$s^\circ(I_k, x) = s(I_k, x) + \left(x - \frac{\Delta}{2}\right)^2 = \left(\frac{1}{2}(a_{k-1} + a_k) - x\right)^2 \tag{3.92}$$

by Lemma 3.1.2, s° is interval-proper, in agreement with Corollary 3.2.4.

3.4.2 || Interval-Ignorance scoring rule

The Ignorance scoring rule (Good, 1952) is defined by

$$S(p, x) = -x \log(p) - (1 - x) \log(1 - p). \tag{3.93}$$

Substituting into equation (3.43) gives

$$\frac{p-x}{p(1-p)} = (p-x) \frac{dh(p)}{dp}. \quad (3.94)$$

We have, therefore, that for $p \in (0, 1)$, $h(p) = \log\{p/(1-p)\}$, from which $g(k) = h(\xi_{k+1}) = \log\{\xi_{k+1}/(1-\xi_{k+1})\}$, for $1 \leq k \leq n$. Following the guidance of setting $f(x) = (x+1)S(x, x)$, $f(x) = 0$ and the expression (3.41) for $s(I_k, x)$ becomes

$$\begin{aligned} s(I_k, x) &= - \sum_{i=1}^{k-1} \left(\log \left(\frac{\xi_{i+1}}{1-\xi_{i+1}} \right) - \log \left(\frac{\xi_i}{1-\xi_i} \right) \right) (x - a_i) \\ &= S(\xi_k, x) - S(\xi_1, x) - \sum_{i=1}^{k-1} \{S(\xi_{i+1}, a_i) - S(\xi_i, a_i)\}. \end{aligned} \quad (3.95)$$

We note in passing, that this is of the same form as was obtained for the Brier scoring rule (see equation (3.87)). No simple choice of ξ_k reduces equation (3.95) to an intuitive scoring rule.

3.4.3 || Pseudo-spherical scoring rule

Fix $\alpha > 1$. The α -pseudo-spherical scoring rule (Roby, 1964) is

$$S(p, x) = \frac{\{-xp + (x-1)(1-p)\}^{\alpha-1}}{\{p^\alpha + (1-p)^\alpha\}^{\frac{\alpha-1}{\alpha}}}. \quad (3.96)$$

Replacing S in equation (3.43) gives

$$\frac{dh(p)}{dp} = \frac{(\alpha-1)\{(p-1)p\}^{\alpha-2}}{\{p^\alpha + (1-p)^\alpha\}^{2-\frac{1}{\alpha}}}. \quad (3.97)$$

Solving for h , we have

$$h(p) = \frac{-(-p)^{\alpha-1} - (p-1)^{\alpha-1}}{\{p^\alpha + (1-p)^\alpha\}^{\frac{\alpha-1}{\alpha}}}. \quad (3.98)$$

Set $g(k) = h(\xi_{k+1})$ and choose

$$f(x) = (x+1)S(x, x) = \left[(x+1)\{-x^2 - (1-x)^2\}^{\alpha-1} \right] / \left[\{x^\alpha + (1-x)^\alpha\}^{\frac{\alpha-1}{\alpha}} \right] \quad (3.99)$$

i.e. $f(0) = (-1)^{\alpha-1}$, $f(1) = 2(-1)^{\alpha-1}$. (If $\alpha = 2$, the pseudo-spherical scoring rule is referred to as the spherical scoring rule.)

[3.5] Interval-Proper Scoring Rules in Practice

Theorem 3.3.1 gives the characteristic form of all interval-proper scoring rules, under the condition that the forecasts are *closed* intervals: given the partition $0 = a_0 < a_1 < \dots < a_n = 1$, the i th interval is $I_i = [a_{i-1}, a_i]$. If a forecaster's actual belief for the probability that $X = 1$, q , falls in the (open) interval (a_{i-1}, a_i) , the forecaster will issue $I_i = [a_{i-1}, a_i]$ as their forecast and will receive a score of $s(I_i, x)$ if $X = x$ (and have an expected score of $s(I_i, q)$).

In practice, closed intervals have an inherent problem: if $q = a_i$ then it is unclear whether the forecaster should issue $I_{i-1} = [a_{i-1}, a_i]$ or $I_i = [a_i, a_{i+1}]$, either interval being acceptable as both intervals contain q . To avoid this difficulty, the intervals from which the forecaster must select a forecast are presented as semi-open intervals i.e. $[a_0, a_1]$ and $(a_{i-1}, a_i]$ for all $i = 2, \dots, n$.

However, for semi-open intervals a *strictly* proper scoring rule does not exist. This follows from previous work by Lambert and Shoham (2009, Definition 4, Theorem 4.1), Lambert (2013, Theorem 1) and the more general result of Frongillo and Kash (2014, Definition 3.3, Theorem 5.2)², which when applied in our context proves that a strictly proper scoring rule for forecasts issued as intervals, exists only if the intervals are closed and intersect at their boundaries: the intervals must have the form $[a_{i-1}, a_i]$, $i = 1, \dots, n$. The following proposition gives a direct proof.

Proposition 3.5.1. *For a binary observation, X , taking the values 0 or 1, there is no strictly proper scoring rule for interval-probabilistic forecasts expressed as semi-open intervals. \square*

Proof. Let the intervals from which the interval-probabilistic forecasts are chosen be the semi-open intervals $I'_1 = [a_0, a_1]$, $I'_k = (a_{k-1}, a_k]$ for $k = 2, \dots, n$. Assume there exists a *strictly* proper scoring rule, s' , for the semi-open interval-probabilistic forecasts. Then for r a precise probability that $X = 1$,

$$s'(I'_k, r) < s'(I'_j, r) \quad \text{for all } r \in I'_k \text{ and } r \notin I'_j. \quad (3.100)$$

In particular,

$$s'(I'_{j-1}, a_{j-1}) < s'(I'_j, a_{j-1}) \quad \text{for all } j = 1, \dots, n-1. \quad (3.101)$$

Further, given that X is a binary observation with a value of either 0 or 1, we have for any

²Also highlighted in a personal communication from Professor R.M. Frongillo of the University of Colorado, Boulder, U.S.A., in response to Mitchell and Ferro (2017).

interval I'_j ,

$$s'(I'_j, r) = (1 - r)s'(I'_j, 0) + rs'(I'_j, 1) \quad (3.102)$$

from which we note that $s'(I'_j, r)$ is continuous in r ; in particular, if $r \in I'_j$, $\lim_{r \downarrow a_{j-1}} s'(I'_j, r) = s'(I'_j, a_{j-1})$.

Therefore, fix a semi-open interval I'_j . For any $r \in I'_j$, because the semi-open intervals are disjoint, $r \notin I'_{j-1}$ and $s'(I'_j, r) < s'(I'_{j-1}, r)$, because by assumption s' is a strictly proper scoring rule. But then

$$s'(I'_j, a_{j-1}) = \lim_{r \downarrow a_{j-1}} s'(I'_j, r) \leq \lim_{r \downarrow a_{j-1}} s'(I'_{j-1}, r) = s'(I'_{j-1}, a_{j-1}). \quad (3.103)$$

But inequality (3.103) contradicts inequality (3.101). Therefore our initial assumption that s' is strictly proper is false and we have shown that there does not exist a strictly proper scoring rule for interval-probabilistic forecasts of semi-open intervals. ■

In other words, for a precise belief that equals an interval boundary any scoring rule will regard each interval adjacent to this boundary as being an equally reasonable forecast and assign both adjacent intervals the same expected score (there are no strong grounds for preferring either interval). Therefore, if a scoring rule is to be strictly proper, the boundary must be shared by both intervals (otherwise the strictly proper scoring rule would necessarily assign the interval containing the boundary a strictly lower expected score in contravention of the scoring rule's assignment of equal expected scores to both intervals).

However, as we prove in chapter 4 with the interval-Brier scoring rule, proper scoring rules do exist for forecasts that are semi-open intervals. More generally, we now show that scoring rules that are strictly proper for closed intervals remain proper (but cannot be strictly proper) if the intervals are made semi-open.

Let s be a (negatively-oriented) interval-proper scoring rule with respect to the *closed* intervals I_i and for $X \sim q$, set $q = a_i$. There are two approaches to consider.

(a) The forecaster predetermines always to forecast the lower of the two intervals, i.e. $I_i = [a_{i-1}, a_i]$. The forecaster's expected score is $s(I_i, q)$. But, because s is interval-proper, if $q = a_i$, then $q \in I_i$ and $q \in I_{i+1}$ so $s(I_i, q) = s(I_{i+1}, q)$. Therefore, the forecaster receives the same expected score as they would do if they had predetermined always to forecast the upper of the two intervals, I_{i+1} . Whether the forecaster decides always to report I_i or always to report I_{i+1} , when their precise-probabilistic belief falls on the boundary of the two intervals, will make no difference to the forecaster's expected score.

(b) The forecaster adopts a randomised approach: the interval forecast is chosen according to the value of an independent binary random variable, Y , where Y has a Bernoulli distribution with parameter θ . For example, on each occasion that $q = a_i$, the forecaster flips a coin for which the probability of obtaining a head is θ and chooses to forecast either $I_i = [a_{i-1}, a_i]$ if the coin shows tails ($Y = 0$), or $I_{i+1} = [a_i, a_{i+1}]$ if the coin shows heads ($Y = 1$) (it is assumed that the coin has no chance of landing on its edge). The score for the forecast will be $(1 - Y)s(I_i, x) + Ys(I_{i+1}, x)$ when $X = x$. But, the forecaster's expected score is

$$\mathbb{E}[(1 - Y)s(I_i, X) + Ys(I_{i+1}, X)] \quad (3.104)$$

By construction, Y and X are independent, so the expected score in equation (3.104) is equal to

$$\begin{aligned} & (1 - \theta)\mathbb{E}[s(I_i, X)] + \theta\mathbb{E}[s(I_{i+1}, X)] \\ & \equiv (1 - \theta)s(I_i, q) + \theta s(I_{i+1}, q) \end{aligned} \quad (3.105)$$

and, for $q = a_i \in I_i, I_{i+1}$, because s is interval-proper, $s(I_i, q) = s(I_{i+1}, q)$, so that the expected score is

$$(1 - \theta)s(I_i, q) + \theta s(I_{i+1}, q) = s(I_i, q) = s(I_{i+1}, q). \quad (3.106)$$

Under a randomised policy, the forecaster's expected score is the same as it would be under a non-randomised policy (as in (a)).

In summary, if $q = a_i$, the forecaster may choose to report either I_i or I_{i+1} , or vary (randomly) between these choices without changing their expected score. It follows that if a forecaster is presented with a collection of semi-open intervals, $[a_0, a_1], (a_1, a_2], \dots, (a_{n-1}, 1]$, the forecaster can behave as though the intervals are closed and adopt the approach of always selecting the lower interval of $[a_{i-1}, a_i]$ and $[a_i, a_{i+1}]$ when $q = a_i$. By such means, the scoring rule will be proper for forecasts that are semi-open intervals.

[3.6] An Application

Equation (3.38) presents the form that an interval-proper scoring rule must have, with section 3.4 giving specific examples. We now investigate the implications of using an improper scoring rule for interval forecasts. We use actual precise-probabilistic forecasts provided by two separate meteorological offices, from which we simulate interval-probabilistic forecasts.

From these synthetic, yet representative, interval-probabilistic forecasts, we can establish empirical measures of the influence of impropriety.

3.6.1 || The Data

Two separate data sets, in both cases precipitation data, were used. The amount of precipitation per day (the 24-hour period beginning at midnight local time) is converted into a binary variable, X , by choosing a threshold rainfall level (in mm) and defining $X = 1$ if the recorded amount of precipitation is greater than the threshold level; otherwise $X = 0$. Forecasts are precise-probabilistic forecasts as described below. Missing data (either observed precipitation or precise-probabilistic forecast) were omitted not imputed.

3.6.1.1 | UK Met Office (UKMO) Data

The UKMO provided data for two locations: Heathrow Airport (Station Number 03772) and Eskdalemuir (Station Number 03162). For each location, the observation was precipitation level (in mm) and the forecasts were precise-probabilities issued as a set of nodes $(z_j, F(z_j))$ $j = 1, \dots, m$ of the cumulative distribution function (F) of the precipitation level in mm (z), from which the precise probability of the precipitation level exceeding a threshold of 1mm was calculated (if necessary, the nodes were linearly interpolated and the tails were linearly extrapolated, subject to a maximum possible value for $F(z)$ of 1 and a minimum possible value for $F(z)$ of 0). Daily forecasts for a period of approximately two years were available, and on each date there were forecasts for 58 lead-times (from 6 to 348 hours at 6-hourly intervals). Climatology was provided as a single cumulative distribution function which had been determined empirically from daily observations between 1 January 1983 to 31 December 2012 (a total of 10951 observations).

3.6.1.2 | Australian Bureau of Meteorology (ABOM) Data

The ABOM provided data for 18 locations around Australia for a period of 290 consecutive days. On each day, the observation at each location was rainfall (in mm) in the 24 hours following midnight of the given day and the forecast was the precise-probabilistic forecast of receiving more than 0.2mm of rain in the same 24-hour time-interval. On each day and for each location 7 forecasts were provided: a forecast issued 12, 36, 60, 84, 108, 132 and 156 hours before the day. Forecasts were provided for two different forecasters, of which only one forecaster will be considered ³.

³Results for the forecaster not used in this chapter appear in [Mitchell and Ferro \(2017\)](#).

3.6.2 || Simulating Interval-Probabilistic Forecasts

We assume that each precise-probabilistic forecast, p , is determined under a precise-proper scoring rule and so represents the forecaster's true belief that $X = 1$. A partition is fixed and the (negatively-oriented) scoring rule for the corresponding interval forecasts is selected (see for example, [Gneiting \(2011a\)](#) on the need for the forecaster to be made aware of the scoring rule).

The simulated interval forecast is the interval which minimises the expected score the forecaster receives under the precise-probabilistic forecast: $I_{\text{issued}} = \arg \min_{I_k} s(I_k, p)$. If s is an interval-proper scoring rule, the interval issued by the forecaster will be the interval containing p . Under an interval-improper scoring rule, the forecast interval will not necessarily contain the forecaster's true belief p .

UKMO precise probabilistic forecasts were translated into interval-probabilistic forecasts (see below) using the following partition of the interval $[0, 1]$ used by the UKMO: $a_0 = 0, 0.025, 0.05, 0.10, 0.20, 0.25, 0.30, 0.40, 0.50, 0.60, 0.70, 0.75, 0.80, 0.90, 0.95, 1 = a_{15}$.

The ABOM precise probabilistic forecasts were converted to interval-probabilistic forecasts (see below) using the following partition of the interval $[0, 1]$: $a_0 = 0, 0.025, 0.075, 0.15, 0.25, 0.35, 0.45, 0.55, 0.65, 0.75, 0.85, 0.925, 0.975, 1 = a_{13}$; a similar partition is used by the ABOM.

For each precise-probabilistic forecast, therefore, two interval-probabilistic forecasts are simulated: one when s is an interval-improper scoring rule and one when s is an interval-proper scoring rule. We emphasise that all interval-probabilistic forecasts so calculated are hypothetical and are not actual interval-probabilistic forecasts provided by either the UKMO or the ABOM.

3.6.3 || Calculating Forecaster Skill

Let x_i be the i th outcome (with value either 0 or 1) and I_{k_i} be the interval-probabilistic forecast associated with x_i , $i = 1, \dots, N$. The forecaster's *accuracy* is calculated as the average score,

$$\bar{s}_N = \frac{1}{N} \sum_{i=1}^N s(I_{k_i}, x_i). \quad (3.107)$$

For large N , assuming that the forecast-outcome pairs are independent, \bar{s}_N is approximately

normally distributed with mean $\mathbb{E}[s(I, X)]$ and variance $\hat{\sigma}^2/N$ where

$$\hat{\sigma}^2 = \frac{1}{N-1} \sum_{i=1}^N (s(I_{k_i}, x_i) - \bar{s}_N)^2. \quad (3.108)$$

We denote the accuracy of the climatological forecaster by

$$\hat{\mu}_{\text{clim}} = \bar{s}_{N,\text{clim}} \approx \mathbb{E}_{\text{clim}}[s(I, X)] \quad (3.109)$$

and the accuracy of the perfect forecaster (who always issues the interval I_n when the outcome is 1 and the interval I_1 when the outcome is zero), by

$$\hat{\mu}_{\text{perf}} = \bar{s}_{N,\text{perf}} \approx \mathbb{E}_{\text{perf}}[s(I, X)]. \quad (3.110)$$

We define (Wilks, 2006, page 259), the forecaster's *skill* by

$$\hat{\psi} \stackrel{\text{def}}{=} \frac{\bar{s}_N - \hat{\mu}_{\text{clim}}}{\hat{\mu}_{\text{perf}} - \hat{\mu}_{\text{clim}}}. \quad (3.111)$$

A skill of 1 for a forecaster demonstrates a perfect forecast record for the forecaster, while a skill of 0 indicates that the forecaster is no more skilful than a climatological forecaster. Assuming that accuracy of the climatological and perfect forecasters is constant, it follows that a forecaster's skill is approximately normally distributed with mean

$$\psi \stackrel{\text{def}}{=} \frac{\mathbb{E}[s(I, X)] - \hat{\mu}_{\text{clim}}}{\hat{\mu}_{\text{perf}} - \hat{\mu}_{\text{clim}}} \quad (3.112)$$

and variance

$$\zeta^2 \stackrel{\text{def}}{=} \frac{\hat{\sigma}^2}{N(\hat{\mu}_{\text{perf}} - \hat{\mu}_{\text{clim}})^2} \quad (3.113)$$

from which we have an approximate $100(1 - \alpha)\%$ confidence interval for forecaster skill of

$$(\hat{\psi} - z_{1-\frac{\alpha}{2}}\zeta, \hat{\psi} - z_{\frac{\alpha}{2}}\zeta) \quad (3.114)$$

where $z_{\frac{\alpha}{2}}$ is the $\alpha/2$ -quantile of the standard normal distribution.

3.6.4 || Example: Interval-Brier scoring rule

Let s be the interval-proper interval-Brier scoring rule (equation (3.90)) and \tilde{s} be the interval-improper (adjusted) Brier midpoint-scoring rule

$$\begin{aligned}\tilde{s}(I_k, x) &= \left(\frac{1}{2}(a_{k-1} + a_k) - x\right)^2 - \frac{1}{4}a_1^2 \\ &= (x - a_{k-1})(x - a_k) + \frac{1}{4}(a_k - a_{k-1})^2 - \frac{1}{4}a_1^2.\end{aligned}\tag{3.115}$$

(For *equally-spaced* intervals s and \tilde{s} are equivalent and \tilde{s} is interval-proper. But, here, unequally-spaced intervals are supposed and were \tilde{s} to be interval-proper, then by Lemma 3.1.2, the Brier midpoint-scoring rule would be proper, contradicting Corollary 3.2.4.)

Figures 3.3a and 3.3b compare forecaster skill under s (proper) and \tilde{s} (improper). In Figure 3.3a, the skill of interval-probabilistic forecasts at Heathrow Airport for different lead-times is shown, while Figure 3.3b displays the skill of the 12-hour lead-time forecasts at each of 18 different locations around Australia.

The immediate conclusion from Figures 3.3a and 3.3b, is that there appears to be no material difference in skill measured under the interval-proper and interval-improper scoring rules.

But, there is a insidious danger from impropriety: impropriety permits hedging, wherein the forecaster chooses to publish an interval-probabilistic forecast that differs from the interval they truly believe is appropriate. In such cases, a forecaster's accuracy (or skill) does not measure their true forecasts but measures their given forecasts, thereby misrepresenting their ability. In the presence of hedging, decisions based on the forecaster's ability, in particular whether one forecaster is better than another, are invalid.

Denoting the forecaster's true (precise-probabilistic) belief that $X = 1$ by q , whether a forecaster is induced to hedge depends on the values of $\tilde{s}(I, q)$ for different intervals I , and therefore, given \tilde{s} , hedging depends only on the partition from which the interval forecasts are selected. In Figures 3.4 to 3.6, we demonstrate the effect of the choice of partition on a forecaster's hedging profile.

Each of Figures 3.4 to 3.6 shows a bar-graph. Every bar is centred over the upper bound of an interval, the height of a bar being the proportion of times the interval is issued as a forecast. For each bar, the white area (if any) is the proportion of times the interval is forecast and is a hedge that understates the forecaster's true interval belief; the brown area (if any) is the proportion of times the interval is forecast and is a hedge that overstates the forecaster's true interval belief. The points marked by \bullet , are the proportion of times the interval is a hedge (either an overstated or understated interval) given the interval is forecast, that is, the points marked \bullet show the propensity to hedge.

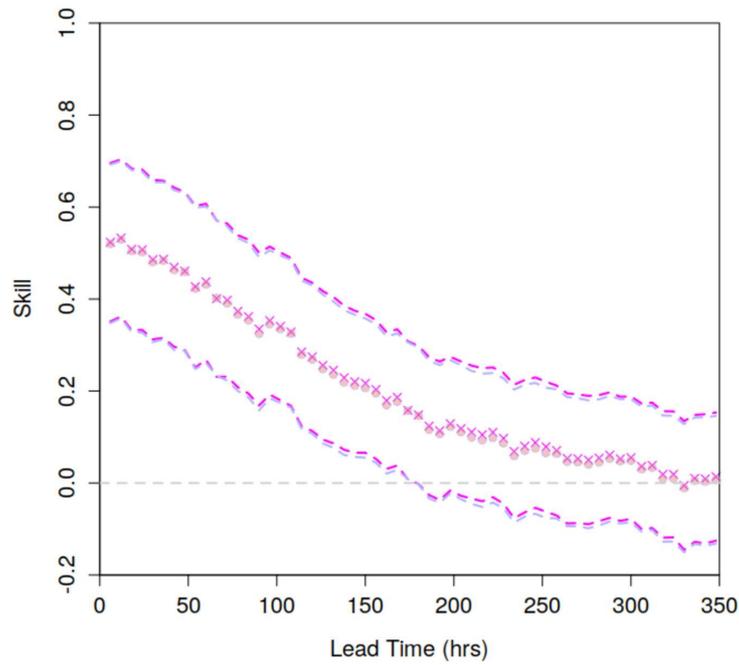
In Figure 3.4, the distribution of the 12-hour lead-time forecasts at Heathrow Airport is shown. The relative frequency of hedging is about 5% with hedging existing in both the lower and upper mid-ranges of the $[0, 1]$ interval. A hedge in the lower mid-ranges of the $[0, 1]$ interval may be either an understatement or an overstatement, as too a hedge in the upper mid-ranges may be. There is no simple trend in the propensity to hedge across the subintervals.

An altogether different set of features is displayed in Figure 3.5, a bar-graph of the 12-hour lead-time interval-probabilistic forecasts at Perth Airport. Here, the relative frequency of hedging is approximately 6% and the forecaster only tends to hedge when issuing forecasts in the extremities of the $[0, 1]$ interval. Hedges in the lower ranges of the $[0, 1]$ interval will understate the forecaster's beliefs while hedges in the upper ranges of the $[0, 1]$ interval will overstate the forecaster's beliefs.

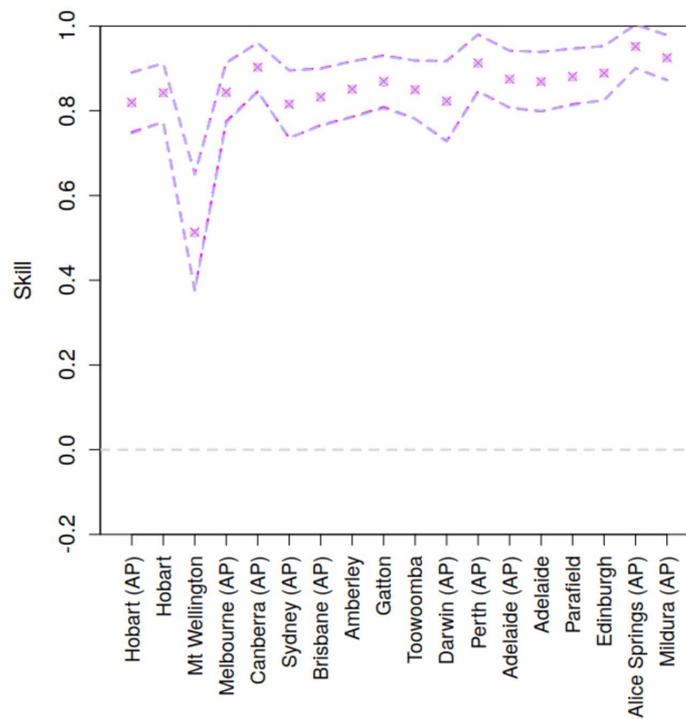
To examine the impact of the form of the partition on the hedging profile, Figure 3.6 presents a bar-graph of the 12-hour lead-time interval-probabilistic forecasts at Heathrow Airport assuming the same partition that was applied at Perth Airport. The relative frequency of hedging is about 6% and hedging behaviour is now much more similar to the hedging behaviour seen for the Perth Airport forecasts: the partition appears to affect both the type and propensity to hedge.

An examination of single site and lead-time forecasts, for a predetermined partition and preselected interval-improper scoring rule, as in Figures 3.4 to 3.6, is helpful in assessing *local* properties of a forecaster's hedges. Of interest too, is aggregate hedging behaviour, the proportion of times the forecaster hedges (either understates or overstates their true beliefs) as the site and forecast lead-time changes. In Figure 3.7, the relative frequency of hedging is shown for different lead-times at two sites: Heathrow Airport and Eskdalemuir. Hedging is, on the whole, higher for Heathrow Airport than for Eskdalemuir, although the pattern of hedging is similar over the different lead-times: hedging occurs on no more than 12% or so of occasions, tending to peak shortly before the 150-hour lead-time forecast and is lowest for the longest lead-times.

In Figure 3.3b, the relative frequency of hedging when issuing interval-probabilistic forecasts is compared for a number of lead-times across different locations in Australia. Here, hedging occurs in between approximately 0% and 20% of forecasts. Hedging levels are similar for sites that are geographically close. For many sites, hedging is higher for shorter duration lead-times with hedging decreasing as the lead-time increases.



(a)



(b)

Figure 3.3: In each figure, estimated forecaster skill (equation (3.111)) and 95% confidence intervals (equation (3.114);) are shown under the interval-proper Brier scoring rule (\times and ---) and the interval-improper adjusted Brier midpoint-scoring rule (\bullet and —). (a) Skill of interval-probabilistic forecasts at Heathrow Airport for different forecast lead-times. (b) Skill of interval-probabilistic forecasts for the 12-hour lead-time at different locations in Australia.

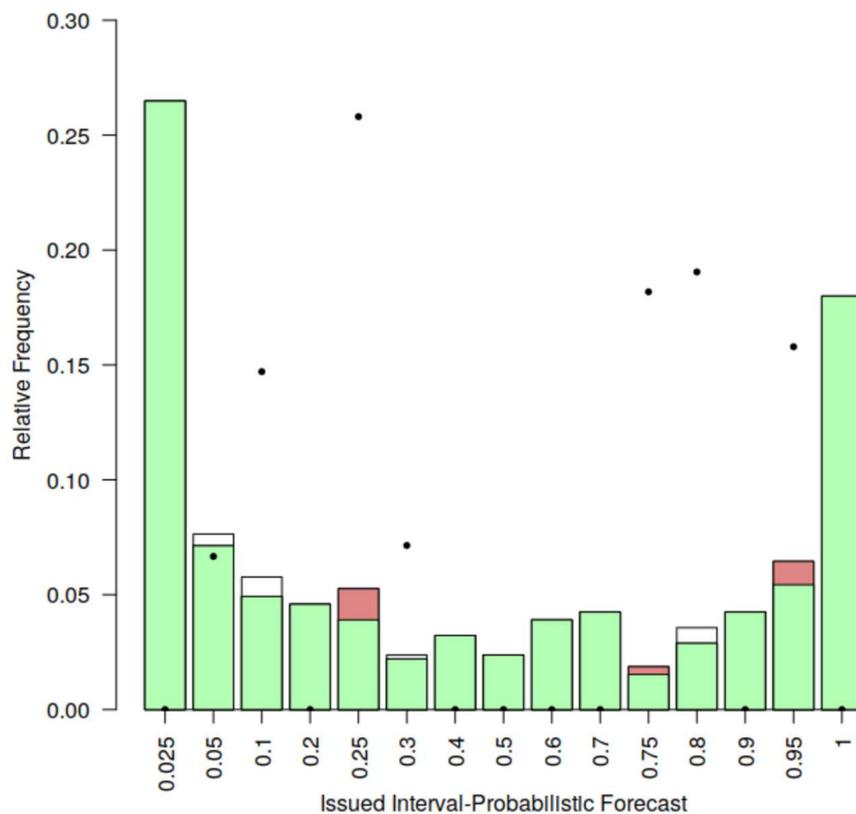


Figure 3.4: The relative frequency of 12-hour lead-time interval-probabilistic forecasts issued for Heathrow Airport. The height of each *entire* bar is an estimate of the probability that the interval is the forecast. The white portion of each bar (\square) is an estimate of the probability that the interval is the forecast published *and* is a hedge that understates the forecaster’s true belief. The brown portion of each bar (\blacksquare) is an estimate of the probability that the interval is the published forecast *and* is a hedge that overstates the forecaster’s true belief. The \bullet points are estimates of the conditional probability that when the interval is forecast, it is a hedge. (The tick-labels on the horizontal axis are the upper end-points of each subinterval.)

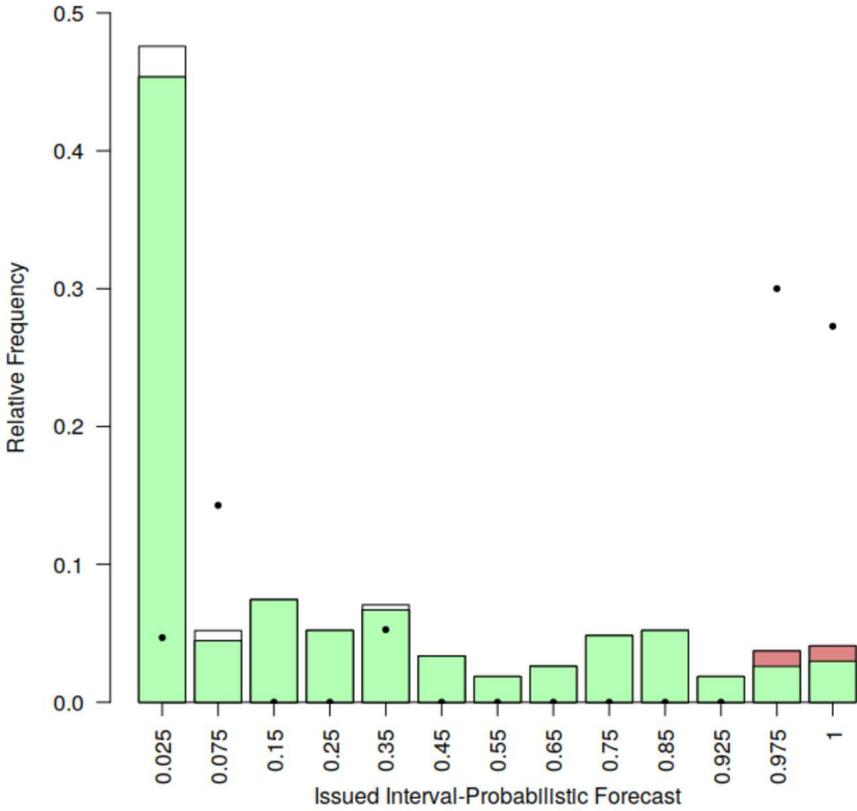


Figure 3.5: The relative frequency of 12-hour lead-time interval-probabilistic forecasts issued for Perth Airport. For an interpretation of the bars see the caption to Figure 3.4.

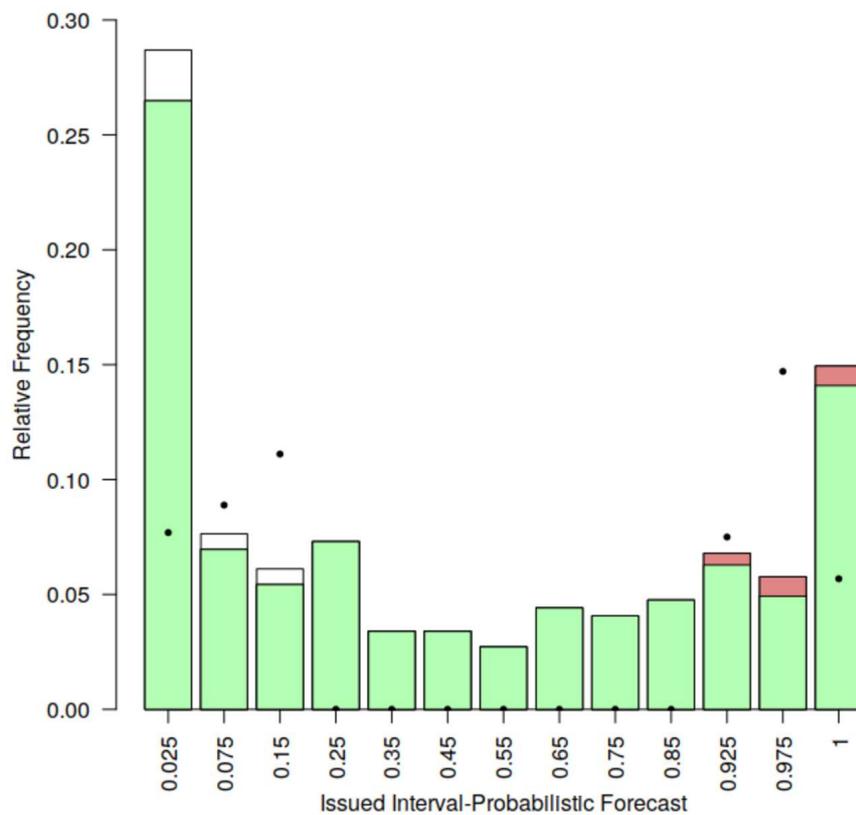


Figure 3.6: The relative frequency of 12-hour lead-time interval-probabilistic forecasts issued for Heathrow Airport *under the same partition used for the forecasts issued for Perth Airport in Figure 3.5*. For an interpretation of the bars see the caption to Figure 3.4.

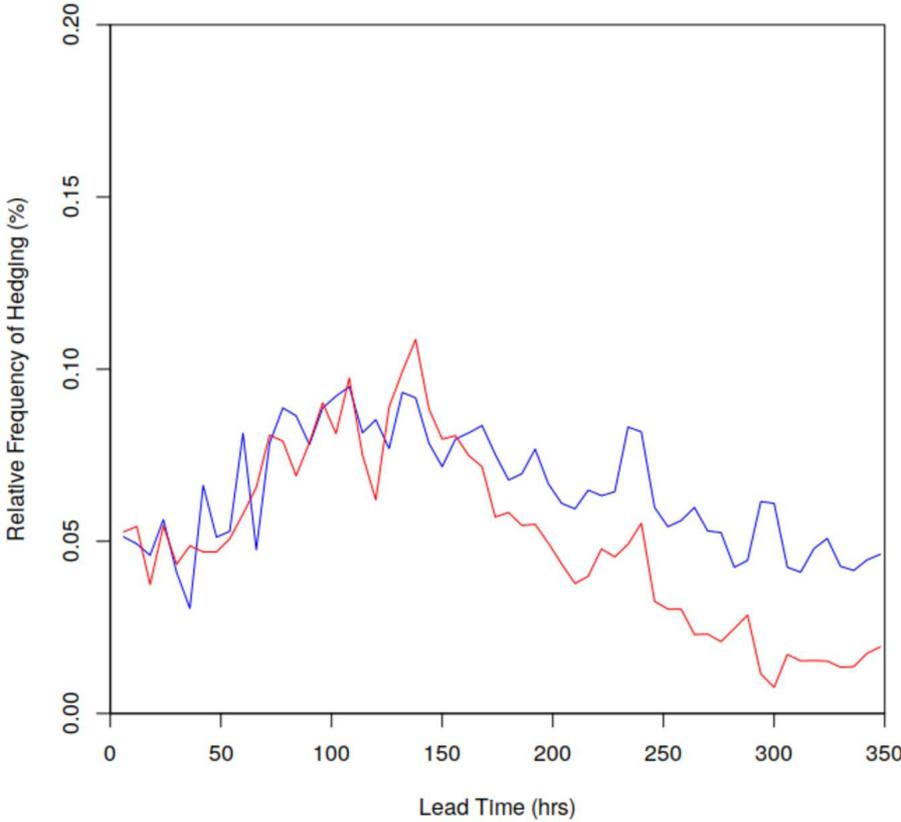


Figure 3.7: Relative frequency of hedging when issuing interval-probabilistic forecasts for different lead-times (hours). Hedging profiles for two different locations are shown: Heathrow Airport (—) and Eskdalemuir (—).

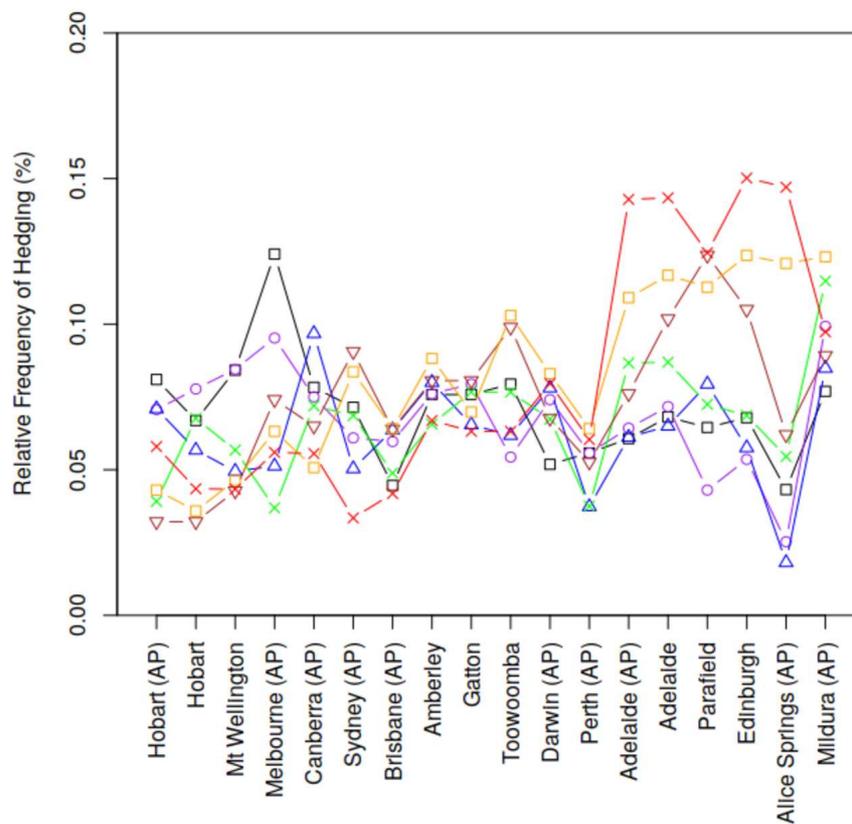


Figure 3.8: Relative frequency of hedging when publishing interval-probabilistic forecasts, for different locations in Australia. Each line shows the hedging profile for forecasts of a particular lead-time: 12-hour (\circ), 36-hour (\triangle), 60-hour (\times), 84-hour (∇), 108-hour (\square), 132-hour (\times), 156-hour (\square).

[3.7] Discussion and Conclusion

If the observation, X , for an event can have one of only two outcomes (e.g. 0 or 1), a probabilistic forecast is a statement about the probability that the outcome is 1, $\Pr(X = 1)$. There are two types of probabilistic forecast: a precise-probabilistic forecast, which is an exact value for $\Pr(X = 1)$, or an interval-probabilistic forecast, which is an interval in which $\Pr(X = 1)$ lies, the interval being selected from a set of predetermined and exhaustive subintervals of the unit interval $[0, 1]$.

An important instance of interval-probabilistic forecasts is *rounded*-probabilistic forecasts. A rounded-probabilistic forecast occurs when a precise-probabilistic forecast is rounded before being issued. Each rounded probability represents a range of probabilities, namely those probabilities that, when rounded, reduce to the forecast probability. For example, if precise-probabilistic forecasts are rounded to the nearest 10%, a rounded probabilistic forecast of 20% can be represented as the interval-probabilistic forecast [15%, 25%).

A forecaster hedges if they issue a forecast that differs from their true belief, because by doing so they improve their expected score. To assess any probabilistic forecast a proper scoring rule must be used if hedging is to be precluded. There is a well-established characterisation of scoring rules that are proper for precise-probabilistic forecasts (see, for example [Gneiting and Raftery, 2007](#)). Scoring rules for interval-probabilistic forecasts can be constructed by applying a scoring rule for precise-probabilistic forecasts to some probability selected from the interval forecast. However, even if the precise-probabilistic scoring rule is proper, the derived scoring rule for interval-probabilistic forecasts need not, in general, be proper.

In this chapter, we show that in order for a scoring rule for interval-probabilistic forecasts to be proper, the scoring rule must have a particular form. From this general form specific scoring rules can be determined that are counterparts to the familiar scoring rules for precise-probabilistic forecasts (e.g. Brier scoring rule ([Brier, 1950](#)), ignorance scoring rule ([Good, 1952](#)) and pseudo-spherical scoring rule ([Roby, 1964](#))).

The need to have scoring rules that are proper for interval forecasts, which we refer to as interval-proper scoring rules, is not that interval-proper scoring rules will give substantially different values of forecaster skill than interval-improper scoring rules (we give an example in which the differences in forecaster skill between an interval-proper and a similar but improper scoring rule are very small). Rather, as we demonstrate using synthetic interval forecasts generated from precise probability of precipitation (PoP) forecasts provided by The Australian Bureau of Meteorology and the UK Met Office, under an improper scoring rule,

a forecaster may be induced to hedge (the relative frequency of hedging in the cases we consider lies approximately in the range of 0 – 15%). Skill is calculated from *published* forecasts, so if hedging occurs, the published forecasts do not completely reflect the forecaster’s true beliefs and the measured skill is not a faithful representation of their forecasting ability (i.e. their substantive insight). Propriety insures against such misrepresentation.

Interval-proper scoring rules depend explicitly on the partition that determines the intervals. For the same precise-probabilistic forecast held by the forecaster, different partitions will lead to different interval-probabilistic forecasts being issued and a different expected score for the forecaster. A natural question arises as to whether there is an optimal set of intervals. We define optimal as follows: let n be the number of *non-zero* nodes in each possible partition. As n increases, we assume that the interval-probabilistic scoring rule, s , tends to a precise-probabilistic scoring rule, S . For fixed n , we propose that the optimal partition is the partition $[a]_n : 0 = a_0 < a_1 < \dots < a_n = 1$ that minimises the difference between $S(p, q)$ and $s(I_i, q)$ for all $p \in I_i$, for each possible q , for all i i.e. minimises

$$\sum_{i=1}^n \int_{a_{i-1}}^{a_i} \int_0^1 d(S(p, q), s([a_{i-1}, a_i], q)) dq dp \quad (3.116)$$

where d is a metric; specifically we choose the metric $d(u, v) = (u - v)^2$, and the optimal partition then minimises

$$D([a]_n) = \sum_{i=1}^n \int_{a_{i-1}}^{a_i} \int_0^1 (S(p, q) - s([a_{i-1}, a_i], q))^2 dq dp. \quad (3.117)$$

To compute the optimal partition, noting that $a_0 = 0$ and $a_n = 1$, the points a_1, \dots, a_{n-1} must be chosen to minimise $D([a]_n)$ subject to the conditions

$$a_{i-1} < a_i \quad i = 1, \dots, n \quad (3.118)$$

which is a $(n - 1)$ -dimensional non-linear constrained optimisation problem.

As an illustration, suppose that S is the (half-)Brier scoring rule ([Brier, 1950](#)) and s is the interval-Brier scoring rule. Then

$$\begin{aligned} W(I_i, p) &\stackrel{\text{def}}{=} \int_0^1 [S(p, q) - s(I_i, q)]^2 dq \\ &= \int_0^1 [(p^2 - 2pq + q) - (q - q(a_{i-1} + a_i) + a_{i-1}a_i)]^2 dq \\ &= (p^2 - a_{i-1}a_i)^2 - (2p - a_{i-1} - a_i)(p^2 - a_{i-1}a_i) + \frac{1}{3}(2p - a_{i-1} - a_i)^2. \end{aligned} \quad (3.119)$$

Integrating equation (3.119), gives

$$\begin{aligned}
\int_{a_{i-1}}^{a_i} W(I_i, p) dp = & \frac{1}{5}a_i^5 - \frac{1}{5}a_{i-1}^5 - \frac{2}{3}a_i^4 a_{i-1} + \frac{2}{3}a_{i-1}^4 a_i + a_{i-1}^2 a_i^3 - a_{i-1}^3 a_i^2 \\
& - \frac{1}{2}a_i^4 + \frac{1}{2}a_{i-1}^4 + \frac{1}{3}a_i^3 a_{i-1} - \frac{1}{3}a_{i-1}^3 a_i + \frac{1}{3}a_i^4 - \frac{1}{3}a_i a_{i-1}^3 \\
& + \frac{4}{9}a_i^3 - \frac{4}{9}a_{i-1}^3 - \frac{1}{3}a_i^2 a_{i-1} - \frac{1}{3}a_{i-1}^2 a_i + \frac{1}{3}a_i^3 + \frac{1}{3}a_i a_{i-1}^2
\end{aligned} \tag{3.120}$$

from which $D([a]_n)$ can be immediately calculated by summation over i .

To find the optimal partition for a given n , subject to constraints (3.118) and $a_0 = 0$, $a_n = 1$, we adopt the following approach.

Set $a_0 = 0$ and define the function D^* by

$$\begin{aligned}
D^*(z) = D([a]_n) \quad \text{for} \\
a_i = a_{i-1} + \Delta_i, \quad \Delta_i = e^{z_i} / \sum_{i=1}^n e^{z_i}, \quad \text{for } i = 1, \dots, n.
\end{aligned} \tag{3.121}$$

Using the `optim` function (with the Nelder-Mead method) in **R** (Ihaka and Gentleman, 1996; R Core Team, 2017), we find those z_i for which $D^*(z)$ is minimised; denote the minimising z_i by z_i^* . The optimal partition is then determined by transforming the z_i^* to the a_i using (3.121).

As an example, Table 3.1 shows the optimal partition for $n = 10$.

a_0	a_1	a_2	a_3	a_4	a_5	a_6	a_7	a_8	a_9	a_{10}
0	0.082	0.172	0.271	0.381	0.501	0.621	0.730	0.829	0.918	1

Table 3.1: The values of the optimal partition for the interval-Brier scoring rule, where optimal is defined as minimising $D([a]_n)$ in equation (3.117). The preset value of n is 10. The optimal nodes of the partition are given to 3 decimal places (minimum value of $D([a]_n)$ is 0.0005 to four decimal places).

Decompositions and Dual Decompositions of Proper Scoring Rules for Interval-Probabilistic Forecasts of Binary Events

Summary. The previous chapter defined and characterised proper scoring rules for interval-probabilistic forecasts of binary observations. For a given scoring rule, accuracy, the expected score over the joint distribution of the forecasts and outcomes, is a measure of the overall fit of the forecasts to the outcomes. But accuracy is an incomplete measure of forecast performance. A more comprehensive evaluation of forecast performance requires the measurement of additional properties of the joint distribution of forecasts and outcomes. Indirect measurements of the properties of the distributions of forecasts and outcomes can be obtained by decomposing accuracy into several components, each component quantifying a feature or attribute of the forecasts. Decompositions of proper scoring rules for precise-probabilistic forecasts are well-established, but no decompositions have yet been proposed for proper scoring rules of interval-probabilistic forecasts. In this chapter, we derive the uncertainty-resolution-reliability (URR) and refinement-discrimination-correctness (RDC) decompositions for the interval-Brier scoring rule. These decompositions are compared to the known URR and RDC decompositions of the Brier scoring rule for precise-probabilistic forecasts. Estimators for the attributes of the decompositions of the interval-Brier scoring rule are proposed, and used to compute values for the attributes of interval-probabilistic forecasts of precipitation.

[4.1] Introduction

As was noted in chapter 2, [Murphy and Winkler \(1987\)](#) advocate that forecast verification should be wholly directed to the examination of the properties of the joint distribution of forecast and observation, this joint distribution representing the total information available for the purposes of evaluating the forecasting system.

Although qualitative analysis of the joint distribution's properties provides some insight into the relationship between forecast and observation, to make definitive statements it is necessary to quantify these properties. One approach to measuring the properties of the joint distribution is to specify a scoring rule ([Winkler, 1996](#)) which assigns, to each forecast and outcome, a score according to how closely the forecast and outcome correspond; as in earlier chapters we continue to assume that all scoring rules are negatively-oriented, so lower scores indicate better correspondence ([Winkler and Murphy, 1968](#)), and we consider only proper scoring rules (see chapters 2, 3). Having selected a proper scoring rule, the expectation of the scoring rule under the joint distribution gives a summary measure of forecaster performance, accuracy.

But, as has been discussed (see, for example, chapter 2), accuracy, by its definition, cannot convey many of the finer details of the correspondence between forecast and obser-

vation. These finer attributes of the forecasting system can be indirectly evaluated by decomposing accuracy into the sum of components, each component representing a particular attribute. A comprehensive analysis must consider two distinct decompositions: the uncertainty-resolution-reliability (URR) decomposition (which is related to the joint distribution through the conditional distribution of the observation given the forecast) and the refinement-discrimination-correctness (RDC) decomposition (related to the joint distribution by the conditional distribution of the forecast given the observation).

In chapter 2, general forms of the URR and RDC decompositions were obtained for any type of forecast, but examples of their application were given only for point-forecasts and precise-probabilistic forecasts, and not for interval-probabilistic forecasts, which had not, at that stage, been introduced.

Interval-probabilistic forecasts (for binary observations) were described in the previous chapter and interval-proper scoring rules, i.e. proper scoring rules for interval-probabilistic forecasts, were defined and their general characteristic form derived. Several examples of interval-proper scoring rules were also given.

With the setting for interval-probabilistic forecasts established, it is now possible to consider the URR and RDC decompositions for interval-probabilistic forecasts. We do so by determining these decompositions for the most amenable of the interval-proper scoring rules given in chapter 3, the interval-Brier scoring rule.

In the section that follows we re-introduce the interval-Brier scoring rule in preparation for determining its URR decomposition in section 4.3 and RDC decomposition in section 4.4. Computational formulae for the attributes of both decompositions are outlined in section 4.5 and having considered their statistical properties, we estimate the attributes of the interval-forecasts of chapter 3, which were synthesised from actual forecasts of the probability of precipitation (PoP) provided by the Australian Bureau of Meteorology and the United Kingdom Met Office. We re-extend our thanks to, Dr. D. Griffiths, Ms. I. Ioannou, and Dr. M. Mittermaier, for their help in making this data available. The chapter ends with a short review in section 4.6.

[4.2] Interval-Proper Scoring Rules

We consider a binary observation, X , with possible outcomes in $\mathcal{X} = \{0, 1\}$, and the set \mathcal{P} of probability distributions for X . A precise-probabilistic forecast for X is the publication of an *exact* value for the probability that $X = 1$, $\Pr(X = 1)$. Because X is binary, the value

$\Pr(X = 1)$ completely identifies a probability distribution for X i.e. a member of \mathcal{P} . In other words, for each $p \in \mathcal{P}$, the probability that $X = 1$ under p , $\Pr_p(X = 1)$ is known if and only if p is known. In this case, therefore, we will write, for ease of exposition, $p = \Pr_p(X = 1)$ for the precise-probabilistic forecast.

All precise-probabilistic forecasts $p \equiv \Pr_p(X = 1)$, will be values in the interval $[0, 1]$. A partition of the interval $[0, 1]$, with nodes $0 = a_0 < a_1 < \dots < a_{n-1} < a_n = 1$, divides the interval $[0, 1]$ into a set of sub-intervals

$$\begin{aligned} I_1 &= [a_0, a_1], \\ I_k &= (a_{k-1}, a_k] \quad k = 2, \dots, n \end{aligned} \tag{4.1}$$

(see chapter 3 for a discussion on semi-closed versus closed sub-intervals). *Given* this set of sub-intervals, the interval-probabilistic forecast corresponding to the precise-probabilistic forecast p , is $f(p) = I(p)$, the interval containing p ; if $p \in I_k$ then $I(p) = I_k$. Let \mathcal{I} be the set of possible interval-probabilistic forecasts $\{I_k | k = 1, \dots, n\}$.

Note that for a published interval forecast, the underlying precise-probabilistic forecast (i.e. probability distribution of X) need not be known. When reference is made to the underlying distribution of an interval forecast it is understood to be latent and not an indication that the probability distribution is available.

A scoring rule for interval-probabilistic forecasts is a function, $s : \mathcal{I} \times \mathcal{X} \rightarrow \mathbb{R}$, with value $s(I(p), x)$, the score when the forecast $I(p)$ is accompanied by the outcome x for X . For r any probability distribution of X , we write

$$s(I(p), r) \stackrel{\text{def}}{=} \mathbb{E}_r[s(I(p), X)] \tag{4.2}$$

and define a (negatively-oriented) scoring rule for interval forecasts as proper if and only if

$$s(I(r), r) \leq s(I(p), r) \quad \text{for all } p, r \in \mathcal{P} \tag{4.3}$$

(compare equation (2.22) of Definition 2.5.2 on page 30, setting $f(p) = I(p)$) that is, if and only if the optimal expected score assuming $X \sim r$ is obtained when the interval forecast is the interval containing r . If a scoring rule for interval forecasts, s , is proper, we say that s is interval-proper.

An example of a negatively-oriented interval-proper scoring rule is the interval-Brier scoring rule, which we now describe.

4.2.1 || Interval-Brier Scoring Rule

For the set of possible interval-probabilistic forecasts in (4.1), the interval-Brier scoring rule is defined in chapter 3, equation (3.90), to be the scoring rule

$$s(I_k, x) = (x - a_{k-1})(x - a_k). \quad (4.4)$$

We can rewrite (4.5) without direct reference to the partition $0 = a_0 < a_1 < \dots < a_n = 1$. For $p \in \mathcal{P}$ and interval-probabilistic forecast $f(p) = I(p)$, let $I^-(p)$ be the infimum of $I(p)$ and let $I^+(p)$ be the supremum of $I(p)$ (for example, if $I(p) = I_k = (a_{k-1}, a_k]$, then $I^-(p) = a_{k-1}$ and $I^+(p) = a_k$). Then the interval-Brier scoring rule can be defined (compare with (4.5)) as the scoring rule

$$s(I(p), x) = (x - I(p)^-)(x - I(p)^+) \quad \text{for } x \in \{0, 1\}, I(p) \in \mathcal{I}. \quad (4.5)$$

The propriety of the interval-Brier scoring rule follows from the general characterisation result of chapter 3, but we give a specific proof of this propriety here.

Lemma 4.2.1. *If s is the interval-Brier scoring rule defined by equation (4.3) then s is interval-proper.* \square

Proof. By (4.3), it is required to show that if r is a probability distribution of X , then

$$s(I(r), r) \leq s(I(p), r) \quad \text{for all } p, r \in \mathcal{P} \quad (4.6)$$

For each $p \in \mathcal{P}$, $I(p) \in \mathcal{I} = \{I_k | k = 1, \dots, n\}$ where the intervals I_k correspond to some finite partition of $[0, 1]$. Therefore, the requirement (4.6) may be restated as

$$s(I_k, r) \leq s(I_j, r) \quad \text{for } r \in I_k, \text{ and all } j \quad (4.7)$$

By definition (4.5) of the interval-Brier scoring rule, for $x \in \mathcal{X} = \{0, 1\}$, and any $j \in \{1, \dots, n\}$,

$$s(I_j, x) = (x - I_j^-)(x - I_j^+) = x - x(I_j^- + I_j^+) + I_j^- I_j^+. \quad (4.8)$$

Therefore,

$$s(I_j, r) = r - r(I_j^- + I_j^+) + I_j^- I_j^+ = r(1 - r) + (r - I_j^-)(r - I_j^+). \quad (4.9)$$

If $I_k = I_j$ then equation (4.7) is immediate. So, suppose that $I_k \neq I_j$. There are two cases to consider.

Case (i): $I_j > I_k$ i.e. $I_j^+ > I_k^+$ and $I_j^- > I_k^-$. For $r \in I_k \neq I_j$, $I_j^+ > I_k^+ \geq r$ and $I_j^- \geq r > I_k^-$ (as r cannot be in both I_k and I_j , the intervals being disjoint). So,

$$\begin{aligned} (r - I_j^+) < (r - I_k^+) \leq 0 \quad \text{and} \quad (r - I_j^-) \leq 0 < (r - I_k^-) \\ \Rightarrow (r - I_j^+)(r - I_j^-) \geq 0 \geq (r - I_k^+)(r - I_k^-). \end{aligned} \quad (4.10)$$

Case (ii): $I_j < I_k$ i.e. $I_j^+ < I_k^+$ and $I_j^- < I_k^-$. For $r \in I_k \neq I_j$, then $I_j^+ < r \leq I_k^+$ and $I_j^- < I_k^- < r$, giving

$$\begin{aligned} (r - I_k^+) \leq 0 < (r - I_j^+) \quad \text{and} \quad 0 < (r - I_k^-) < (r - I_j^-) \\ \Rightarrow (r - I_j^-)(r - I_j^+) > 0 \geq (r - I_k^-)(r - I_k^+). \end{aligned} \quad (4.11)$$

In either case, equation (4.7) is satisfied. ■

[4.3] URR Decomposition

For interval-probabilistic forecasts, $f(p) = I(p)$ for all $p \in \mathcal{P}$ and the URR decomposition of an interval-proper scoring rule s becomes (following from equation (2.32) with $f(p) = I(p)$),

$$\begin{aligned} \mathbb{E}[s(I(P), X)] = & \underbrace{s(I(q), q)}_{\text{Uncertainty}} - \underbrace{(s(I(q), q) - \mathbb{E}[s(I(q_{I(P)}), q_{I(P)})])}_{\text{Resolution}} \\ & \underbrace{\hspace{10em}}_{\text{Sharpness}} \\ & + \underbrace{\mathbb{E}[s(I(P), q_{I(P)}) - s(I(q_{I(P)}), q_{I(P)})]}_{\text{Reliability}} \quad (4.12) \end{aligned}$$

where, as before, q is the unconditional (or, marginal, or climatological) distribution of X and $q_{I(p)}$ is the conditional distribution of X given that the interval forecast is $I(p)$.

It is clear that the URR decomposition is expressed in terms of the (distribution of the) *issued* forecasts (which are available) and not the underlying probability distribution of X (which, in general, is not available).

To determine the URR decomposition for the interval-Brier scoring rule, we require expressions for $s(I(q), q)$ and $s(I(q_{I(p)}), q_{I(p)})$ for each interval forecast $I(p)$ (i.e. each $p \in \mathcal{P}$).

From equation (4.7), setting $r = q$ and $I_j = I(q)$ gives

$$s(I(q), q) = q(1 - q) + (q - I^-(q))(q - I^+(q)) \quad (4.13)$$

and, on setting $r = q_{I(p)}$ and $I_j = I(q_{I(p)})$ we have

$$s(I(q_{I(p)}), q_{I(p)}) = q_{I(p)}(1 - q_{I(p)}) + (q_{I(p)} - I^-(q_{I(p)}))(q_{I(p)} - I^+(q_{I(p)})). \quad (4.14)$$

For X binary, $q = \mathbb{E}[X]$ and $q_{I(p)} = \mathbb{E}[X|I(P) = I(p)]$. Using equations (4.13) and (4.14), it is a short calculation to show that

$$\mathbb{E}[s(I(P), X)] = \underbrace{\text{UNC}_{\mathcal{I}} - \text{RES}_{\mathcal{I}}}_{\text{Sharpness}} + \text{REL}_{\mathcal{I}} \quad (4.15)$$

where, noting the dependency on the set of possible interval forecasts, \mathcal{I} , uncertainty for the interval forecasts is

$$\text{UNC}_{\mathcal{I}} = \mathbb{E}[X](1 - \mathbb{E}[X]) + (\mathbb{E}[X] - I^-(\mathbb{E}[X]))(\mathbb{E}[X] - I^+(\mathbb{E}[X])), \quad (4.16)$$

resolution for the interval forecasts is

$$\begin{aligned} \text{RES}_{\mathcal{I}} = \mathbb{E} \left[(\mathbb{E}[X|I(P)] - \mathbb{E}[X])^2 \right] &+ (\mathbb{E}[X] - I^-(\mathbb{E}[X]))(\mathbb{E}[X] - I^+(\mathbb{E}[X])) \\ &- \mathbb{E} \left[(\mathbb{E}[X|I(P)] - I^-(\mathbb{E}[X|I(P)]))(\mathbb{E}[X|I(P)] - I^+(\mathbb{E}[X|I(P)])) \right] \end{aligned} \quad (4.17)$$

and reliability for the interval forecasts is

$$\begin{aligned} \text{REL}_{\mathcal{I}} = \mathbb{E} \left[(\mathbb{E}[X|I(P)] - I^-(P))(\mathbb{E}[X|I(P)] - I^+(P)) \right. \\ \left. - (\mathbb{E}[X|I(P)] - I^-(\mathbb{E}[X|I(P)]))(\mathbb{E}[X|I(P)] - I^+(\mathbb{E}[X|I(P)])) \right]. \end{aligned} \quad (4.18)$$

Equation (4.18) shows that interval-probabilistic forecasts are perfectly reliable if $\mathbb{E}[X|I(P)] \in I(P)$ for then $I(\mathbb{E}[X|I(P)]) = I(P)$, reducing the reliability term to zero.

The intervals in \mathcal{I} are determined by the partition $a_0 < a_1 < \dots < a_n$. As n changes the partition will change. For any n , write $a_{n,0} < a_{n,1} < \dots < a_{n,n}$ for the associated partition. If, as n tends to infinity, the mesh of the partition, $\mu_n = \max\{a_{n,i} - a_{n,i-1} \mid i = 1, \dots, n\}$, tends to 0, so that for all $q \in \mathcal{P}$, $I(q) \rightarrow q$, it can be seen from equations (4.16) to (4.18) that the URR decomposition for the interval-Brier score, tends to the URR decomposition in equation (2.7) for the Brier scoring rule (for precise-probabilistic forecasts) (Brier, 1950).

[4.4] RDC Decomposition

Our overriding concern in decomposing accuracy is to quantify the features of the joint distribution of forecast and observation. The joint distribution governs the distributions (marginal and conditional) of the observation and the distributions of the forecasts. So by examining, separately, the distributional properties of the observation and the distributional properties of the forecasts, we can reflect on the features of the joint distribution.

The URR decomposition quantifies the distributional properties of the observation. The distributional properties of the forecasts are measured by the RDC decomposition, which examines the marginal, conditional (on the observation) and ideal distributions of the forecasts.

The ideal distribution, given $X = x$, of the interval-probabilistic forecasts in \mathcal{I} , is the distribution δ_x^* that assigns probability 1 to the interval in \mathcal{I} containing x , $I(x)$, i.e. $\Pr_{\delta_x^*}(I = I(x)) = 1$, where

$$I(x) = \begin{cases} I_1 = [a_0, a_1] & \text{if } x = 0 \\ I_n = (a_{n-1}, a_n] & \text{if } x = 1. \end{cases} \quad (4.19)$$

The marginal, conditional and ideal distributions of the interval-probabilistic forecasts are elements of the set \mathcal{I}^* , the set of all probability distributions over \mathcal{I} . To find the RDC decomposition for the interval-Brier scoring rule, s , it is necessary to establish a (negatively-oriented) scoring rule $s^* : \mathcal{I}^* \times \mathcal{I} \rightarrow \mathbb{R}$, with score $s(r^*, I)$ for $r^* \in \mathcal{I}^*$ and $I \in \mathcal{I}$, such that s^* is proper *and* s^* extends s in the sense that $s^*(\delta_x^*, I) = s(I, x)$.

To compose a scoring rule s^* extending s , we begin by noting that the interval-Brier scoring rule (4.3) may be written in terms of $I(x)$ as

$$s(I, x) = \left((1-x)I^-(x) + xI^+(x) - I^- \right) \left((1-x)I^-(x) + xI^+(x) - I^+ \right). \quad (4.20)$$

As a small generalisation of s , to a scoring rule, s' , on two intervals in \mathcal{I} , define s' by

$$s'(I, I_k) = \left(\left[1 - \left(\frac{k-1}{n-1} \right) \right] I_k^- + \left(\frac{k-1}{n-1} \right) I_k^+ - I^- \right) \left(\left[1 - \left(\frac{k-1}{n-1} \right) \right] I_k^- + \left(\frac{k-1}{n-1} \right) I_k^+ - I^+ \right). \quad (4.21)$$

By definition of $I(x)$, it is seen that

$$\begin{aligned}
s'(I, I(x)) &= \begin{cases} s'(I, I_1) & \text{if } x = 0 \\ s'(I, I_n) & \text{if } x = 1 \end{cases} \\
&= \begin{cases} (I_1^- - I^-)(I_1^- - I^+) & \text{if } x = 0 \\ (I_n^+ - I^-)(I_n^+ - I^+) & \text{if } x = 1 \end{cases} \\
&= \begin{cases} (0 - I^-)(0 - I^+) & \text{if } x = 0 \\ (1 - I^-)(1 - I^+) & \text{if } x = 1 \end{cases} \\
&= (x - I^-)(x - I^+) \\
&= s(I, x). \tag{4.22}
\end{aligned}$$

We extend s' to a scoring rule s^* , by associating with each $r^* \in \mathcal{I}^*$ an interval $I_{r^*} \in \mathcal{I}$ (see below) and defining

$$s^*(r^*, I) = s'(I, I_{r^*}). \tag{4.23}$$

For $r^* \in \mathcal{I}^*$, the interval I_{r^*} is set to be the interval in \mathcal{I} closest to

$$\mathbb{E}_{r^*}[I] = (\mathbb{E}_{r^*}[I^-], \mathbb{E}_{r^*}[I^+]) \tag{4.24}$$

that is, I_{r^*} is the interval closest to the expected interval when the interval-probabilistic forecasts are distributed according to r^* . Here we use the term closest in the sense that, for $j \in \{1, \dots, n\}$ letting

$$I_j^\circ = \left[1 - \left(\frac{j-1}{n-1}\right)\right] I_j^- + \left(\frac{j-1}{n-1}\right) I_j^+ \tag{4.25}$$

then

$$\begin{aligned}
I_{r^*} &= I_{j_{\min}}, \quad \text{where} \\
j_{\min} &= \arg \min_j (I_j^\circ - \mathbb{E}_{r^*}[I^-])(I_j^\circ - \mathbb{E}_{r^*}[I^+]). \tag{4.26}
\end{aligned}$$

Because $I_1^\circ = I_1^-$ and $I_n^\circ = I_n^+$, and by definition of the ideal distribution, δ_x^* , $\mathbb{E}_{\delta_x^*}[I^-] = I^-(x)$ and $\mathbb{E}_{\delta_x^*}[I^+] = I^+(x)$, then

$$\begin{aligned}
I_{\delta_x^*} &= I_{j_{\min}}, \quad \text{where} \\
j_{\min} &= \arg \min_j (I_j^\circ - I^-(x))(I_j^\circ - I^+(x)) \\
&= \begin{cases} 1 & \text{if } x = 0 \\ n & \text{if } x = 1 \end{cases}
\end{aligned} \tag{4.27}$$

giving $I_{\delta_x^*} = I(x)$. It follows from equations (4.22) and (4.23), that

$$s^*(\delta_x^*, I) = s'(I, I(x)) = s(I, x). \tag{4.28}$$

Therefore, s^* extends s . For s^* to be an proper extension of s it remains to show that s^* is proper, that is, that

$$s^*(r^*, r^*) \leq s^*(p^*, r^*) \quad \text{for all } p^*, r^* \in \mathcal{I}^*. \tag{4.29}$$

Proposition 4.4.1. *The scoring rule s^* defined by equation (4.23), satisfies equation (4.29) and is, therefore, a proper scoring rule. \square*

Proof. For $p^* \in \mathcal{I}^*$, define I_{p^*} according to equation (4.26), so that $I_{p^*} \in \mathcal{I}$. Let $\lfloor I_{p^*} \rfloor$ be the index of I_{p^*} in \mathcal{I} . For example, if $I_{p^*} = I_k$ then $\lfloor I_{p^*} \rfloor = k$.

Following equation (4.25), with a slight abuse of notation, write

$$I_{p^*}^\circ = \left[1 - \left(\frac{\lfloor I_{p^*} \rfloor - 1}{n - 1} \right) \right] I_{p^*}^- + \left(\frac{\lfloor I_{p^*} \rfloor - 1}{n - 1} \right) I_{p^*}^+. \tag{4.30}$$

Then

$$\begin{aligned}
s^*(p^*, r^*) &\stackrel{\text{def}}{=} \mathbb{E}_{r^*}[s^*(p^*, I)] \\
&= \mathbb{E}_{r^*}[s'(I, I_{p^*})] \\
&= \mathbb{E}_{r^*}[(I_{p^*}^\circ - I^-)(I_{p^*}^\circ - I^+)] \\
&= (I_{p^*}^\circ)^2 - I_{p^*}^\circ \left(\mathbb{E}_{r^*}[I^-] + \mathbb{E}_{r^*}[I^+] \right) + \mathbb{E}_{r^*}[I^- I^+] \\
&= \left(I_{p^*}^\circ - \mathbb{E}_{r^*}[I^-] \right) \left(I_{p^*}^\circ - \mathbb{E}_{r^*}[I^+] \right) + \text{cov}_{r^*}(I^-, I^+).
\end{aligned} \tag{4.31}$$

Therefore, $s^*(p^*, r^*)$ is minimised for that probability distribution $p^* \in \mathcal{I}^*$ for which

$$(I_{p^*}^\circ - \mathbb{E}_{r^*}[I^-])(I_{p^*}^\circ - \mathbb{E}_{r^*}[I^+]) \tag{4.32}$$

is smallest. But, by definition of I_{p^*} , expression (4.32) is smallest for $p^* = r^*$ and we can conclude that s^* is proper. \blacksquare

For q^* the marginal (i.e. unconditional historical) distribution for the interval-probabilistic forecasts and q_x^* the conditional historical distribution of the forecasts given $X = x$, from equation (4.31) we have

$$s^*(q^*, q^*) = (I_{q^*}^\circ)^2 - I_{q^*}^\circ \left(\mathbb{E}[I^-] + \mathbb{E}[I^+] \right) + \mathbb{E}[I^- I^+] \quad (4.33)$$

and for $x \in \{0, 1\}$,

$$s^*(q_x^*, q_x^*) = (I_{q_x^*}^\circ)^2 - I_{q_x^*}^\circ \left(\mathbb{E}[I^- | X = x] + \mathbb{E}[I^+ | X = x] \right) + \mathbb{E}[I^- I^+ | X = x]. \quad (4.34)$$

Referring to equation (2.41) the RDC decomposition for the interval-Brier scoring rule is, in terms of s^* ,

$$\begin{aligned} \mathbb{E}[s(I(P), X)] &= \underbrace{s^*(q^*, q^*)}_{\text{Refinement}} - \underbrace{(s^*(q^*, q^*) - \mathbb{E}[s^*(q_X^*, q_X^*)])}_{\text{Discrimination}} \\ &\quad \underbrace{\hspace{10em}}_{\text{Excess}} + \underbrace{(\mathbb{E}[s(I(P), X)] - \mathbb{E}[s^*(q_X^*, q_X^*)])}_{\text{Correctness}} \end{aligned} \quad (4.35)$$

which, using equations (4.33) and (4.34), reduces in a few lines to

$$\mathbb{E}[s(I(P), X)] = \underbrace{\text{REF}_{\mathcal{I}} - \text{DIS}_{\mathcal{I}}}_{\text{Excess}} + \text{COR}_{\mathcal{I}} \quad (4.36)$$

where, with the dependency on the intervals \mathcal{I} being made explicit, refinement is equal to

$$\text{REF}_{\mathcal{I}} = (I_{q^*}^\circ)^2 - I_{q^*}^\circ \left(\mathbb{E}[I^-] + \mathbb{E}[I^+] \right) + \mathbb{E}[I^- I^+] \quad (4.37)$$

which is a measure of the variance of the interval-probabilistic forecasts. Discrimination is given by

$$\text{DIS}_{\mathcal{I}} = \mathbb{E} \left[(I_{q^*}^\circ - I_{q_x^*}^\circ) \left(I_{q^*}^\circ + I_{q_x^*}^\circ - \mathbb{E}[I^- | X] - \mathbb{E}[I^+ | X] \right) \right] \quad (4.38)$$

which measures the extent to which different outcomes are preceded by different forecasts, and can be quantified by measuring how much less the variance of the interval-probabilistic forecasts preceding each distinct outcome is, than the overall variance of the interval-probabilistic forecasts. The final term, correctness, is

$$\text{COR}_{\mathcal{I}} = \mathbb{E} \left[(X - I_{q_x^*}^\circ) \left(X + I_{q_x^*}^\circ - \mathbb{E}[I^- | X] - \mathbb{E}[I^+ | X] \right) \right] \quad (4.39)$$

which measures on average, how close the average of the forecasts preceding each outcome is to the outcome.

As was the case for the URR decomposition, let $n \rightarrow \infty$, such that $\mu_n \rightarrow 0$ and the intervals tend to singletons (precise-probabilistic forecasts for X). In this case, each $p^* \in \mathcal{I}^*$ becomes a distribution over the elements of \mathcal{P} (the precise-probabilistic forecasts for X). In addition, for any $r^* \in \mathcal{I}^*$, $\mathbb{E}_{r^*}[I]$ becomes the single value $\mathbb{E}_{r^*}[P]$ and I_{r^*} , the closest value to $\mathbb{E}_{r^*}[P]$, must equal $\mathbb{E}_{r^*}[P]$. It follows that $I_{r^*}^o = I_{r^*} = \mathbb{E}_{r^*}[P]$. With these limits, it is seen that the RDC decomposition for the interval-Brier scoring rule tends to the RDC decomposition of the Brier scoring rule (Brier, 1950) for precise-probabilistic forecasts (given in equation (2.8)).

[4.5] Computational Formulae for the Attributes

We will be interested in computing not only the decompositions of the interval-Brier scoring rule, but also the decompositions of the Brier scoring rule. We consider the Brier scoring rule both for comparative purposes, and because, as far as we are aware, evaluating the RDC decomposition of the Brier scoring rule has not been studied before in the literature.

One approach to computing the attributes of the URR and RDC decompositions from a sample of forecast-outcome pairs, is to estimate the marginal, conditional and joint probability distribution functions of the forecasts and observation using the empirical distribution function. For precise-probabilistic forecasts, the distribution function over the continuum of values $[0, 1]$ can be approximated in many ways; we shall assume that the precise-probabilistic probabilities that can be *issued* form a (perhaps large) finite set of distinct points.

For a sample size of T , for each $t = 1, \dots, T$, denote by $x(t)$ the outcome (which may be either 0 or 1), $p(t) \in [0, 1]$ the precise-probabilistic forecast preceding $x(t)$ and $I(t) \in \mathcal{I} = \{I_1, \dots, I_n\}$ the interval-probabilistic forecast preceding the outcome $x(t)$.

Under the empirical distribution function, the estimates of the (unconditional) means are

$$\begin{aligned} \hat{\mathbb{E}}[X] &= \hat{q} = \frac{1}{T} \sum_{t=1}^T x(t), & \hat{\mathbb{E}}[P] &= \bar{p} = \frac{1}{T} \sum_{t=1}^T p(t), \\ \hat{\mathbb{E}}[I] &= \bar{I} = \frac{1}{T} \sum_{t=1}^T I(t) = \left(\frac{1}{T} \sum_{t=1}^T I^-(t), \frac{1}{T} \sum_{t=1}^T I^+(t) \right] \end{aligned} \quad (4.40)$$

(the semi-open interval in equation (4.40) being a closed interval, should $I^-(t) = 0$ for all t). Estimates of the conditional means are

$$\begin{aligned}
\hat{\mathbb{E}}[X|P = p] &= \hat{q}_p = \sum_{t=1}^T x(t) \mathbf{1}_{\{p\}}(p(t)) / \sum_{t=1}^T \mathbf{1}_{\{p\}}(p(t)) \\
\hat{\mathbb{E}}[X|I = I_j] &= \hat{q}_{I_j} = \sum_{t=1}^T x(t) \mathbf{1}_{\{I_j\}}(I(t)) / \sum_{t=1}^T \mathbf{1}_{\{I_j\}}(I(t)) \\
\hat{\mathbb{E}}[P|X = x] &= \bar{p}_x = \sum_{t=1}^T p(t) \mathbf{1}_{\{x\}}(x(t)) / \sum_{t=1}^T \mathbf{1}_{\{x\}}(x(t))
\end{aligned} \tag{4.41}$$

and

$$\begin{aligned}
\hat{\mathbb{E}}[I|X = x] &= \bar{I}_x \\
&= \left(\sum_{t=1}^T I^-(t) \mathbf{1}_{\{x\}}(x(t)) / \sum_{t=1}^T \mathbf{1}_{\{x\}}(x(t)), \sum_{t=1}^T I^+(t) \mathbf{1}_{\{x\}}(x(t)) / \sum_{t=1}^T \mathbf{1}_{\{x\}}(x(t)) \right). \tag{4.42}
\end{aligned}$$

For the RDC decomposition of the interval-Brier scoring rule, we also require estimates for the values $I_{q^*}^\circ$ and $I_{q_x^*}^\circ$, x a value of X . For each $j = 1, \dots, n$, find I_j° and define

$$\begin{aligned}
\hat{I}_{q^*}^\circ &= I_{j_{\min}}^\circ, \quad j_{\min} = \arg \min_j (I_j^\circ - \bar{I}^-)(I_j^\circ - \bar{I}^+), \\
\hat{I}_{q_x^*}^\circ &= I_{j_{\min, x}}^\circ, \quad j_{\min, x} = \arg \min_j (I_j^\circ - \bar{I}_x^-)(I_j^\circ - \bar{I}_x^+). \tag{4.43}
\end{aligned}$$

Using equations (4.40) to (4.43), estimators for accuracy and the attributes of the URR and RDC decompositions for both the interval-Brier scoring rule and (for comparison) the (precise-, half-)Brier scoring rule are given in Table 4.1.

	Brier	Interval-Brier
SCR	$\frac{1}{T} \sum_{t=1}^T (x(t) - p(t))^2$	$\frac{1}{T} \sum_{t=1}^T (x(t) - I^-(t))(x(t) - I^+(t))$
UNC	$\hat{q}(1 - \hat{q})$	$\hat{q}(1 - \hat{q}) + (\hat{q} - I^-(\hat{q}))(\hat{q} - I^+(\hat{q}))$
RES	$\frac{1}{T} \sum_{t=1}^T (\hat{q} - \hat{q}_{p(t)})^2$	$\frac{1}{T} \sum_{t=1}^T (\hat{q} - \hat{q}_{I(t)})^2 + (\hat{q} - I^-(\hat{q}))(\hat{q} - I^+(\hat{q}))$ $-\frac{1}{T} \sum_{t=1}^T (\hat{q}_{I(t)} - I^-(\hat{q}_{I(t)}))(\hat{q}_{I(t)} - I^+(\hat{q}_{I(t)}))$
REL	$\frac{1}{T} \sum_{t=1}^T (p(t) - \hat{q}_{p(t)})^2$	$\frac{1}{T} \sum_{t=1}^T (\hat{q}_{I(t)} - I^-(t))(\hat{q}_{I(t)} - I^+(t))$ $-\frac{1}{T} \sum_{t=1}^T (\hat{q}_{I(t)} - I^-(\hat{q}_{I(t)}))(\hat{q}_{I(t)} - I^+(\hat{q}_{I(t)}))$
REF	$\frac{1}{T} \sum_{t=1}^T (p(t) - \bar{p})^2$	$(\hat{I}_{q^*}^\circ)^2 - \hat{I}_{q^*}^\circ(\bar{I}^- + \bar{I}^+) + \frac{1}{T} \sum_{t=1}^T I^-(t)I^+(t)$
DIS	$\sum_{x=0}^1 (\bar{p} - \bar{p}_x)^2 \hat{q}^x (1 - \hat{q})^{1-x}$	$\sum_{x=0}^1 (\hat{I}_{q^*}^\circ - \hat{I}_{q_x^*}^\circ)(\hat{I}_{q^*}^\circ + \hat{I}_{q_x^*}^\circ - \bar{I}_x^- - \bar{I}_x^+) \hat{q}^x (1 - \hat{q})^{1-x}$
COR	$\sum_{x=0}^1 (x - \bar{p}_x)^2 \hat{q}^x (1 - \hat{q})^{1-x}$	$\sum_{x=0}^1 (x - \hat{I}_{q_x^*}^\circ)(x + \hat{I}_{q_x^*}^\circ - \bar{I}_x^- - \bar{I}_x^+) \hat{q}^x (1 - \hat{q})^{1-x}$

Table 4.1: Estimators for the accuracy (SCR), uncertainty (UNC), resolution (RES), reliability (REL), refinement (REF), discrimination (DIS) and correctness (COR) of the (precise, half-)Brier scoring rule and the interval-Brier scoring rule. Estimators are derived by substituting, into the theoretical formulae of the accuracy and attributes, the empirical distribution function for the marginal, conditional and joint distribution functions.

Assume that the sample pairs $(p(t), x(t))$ for $t = 1, \dots, T$, and separately $(I(t), x(t))$, $t = 1, \dots, T$ are instances of independently and identically distributed random pairs. Then, [Bröcker \(2012\)](#) proves that the empirical estimators for the URR attributes in Table 4.1, of the Brier scoring rule, are biased; and, while the estimator for uncertainty has a negative bias, the biases for the empirical estimators of resolution and reliability are positive i.e. the estimates of resolution and reliability are, on average, higher than the actual attribute values (so, resolution is worse and reliability better than the estimates suggest). In our notation, letting, for each $p \in [0, 1]$,

$$N_p \stackrel{\text{def}}{=} \sum_{t=1}^T \mathbf{1}_{\{p\}}(p(t)) \quad (4.44)$$

and denoting by \sum_p the sum over distinct values of p , the expressions for the biases are (see also [Ferro and Fricker, 2012](#))

$$\begin{aligned} \text{bias}(\widehat{\text{UNC}}) &= \mathbb{E}[\widehat{\text{UNC}}] - \text{UNC} = -\frac{1}{T} \text{var}[X], \\ \text{bias}(\widehat{\text{RES}}) &= \mathbb{E}[\widehat{\text{RES}}] - \text{RES} = \frac{1}{T} \sum_p \Pr(N_p > 0) \text{var}[X|P = p] - \frac{1}{T} \text{var}[X], \\ \text{bias}(\widehat{\text{REL}}) &= \mathbb{E}[\widehat{\text{REL}}] - \text{REL} = \frac{1}{T} \sum_p \Pr(N_p > 0) \text{var}[X|P = p]. \end{aligned} \quad (4.45)$$

It can be proved in a manner similar to the analysis of [Bröcker \(2012\)](#) and [Ferro and Fricker \(2012\)](#), that letting for each $x \in \{0, 1\}$,

$$N_x = \sum_{t=1}^T \mathbf{1}_{\{x\}}(x(t)) \quad (4.46)$$

the empirical estimators of the attributes of the RDC decomposition of the Brier scoring rule in Table 4.1 are also biased, with biases,

$$\begin{aligned} \text{bias}(\widehat{\text{REF}}) &= \mathbb{E}[\widehat{\text{REF}}] - \text{REF} = -\frac{1}{T} \text{var}[P], \\ \text{bias}(\widehat{\text{DIS}}) &= \mathbb{E}[\widehat{\text{DIS}}] - \text{DIS} = \frac{1}{T} \sum_{x=0}^1 \Pr(N_x > 0) \text{var}[P|X = x] - \frac{1}{T} \text{var}[P], \\ \text{bias}(\widehat{\text{COR}}) &= \mathbb{E}[\widehat{\text{COR}}] - \text{COR} = \frac{1}{T} \sum_{x=0}^1 \Pr(N_x > 0) \text{var}[P|X = x]. \end{aligned} \quad (4.47)$$

Refinement has a negative bias and correctness has positive bias. But, noting that for all x , for $T \geq 1$, $\Pr(N_x > 0) \geq \Pr(X = x)$ so that

$$\begin{aligned} \frac{1}{T} \sum_{x=0}^1 \text{var}[P|X = x] \Pr(N_x > 0) &\geq \frac{1}{T} \sum_{x=0}^1 \text{var}[P|X = x] \Pr(X = x) \\ &= \frac{1}{T} (\text{var}[P] - \text{var}[\mathbb{E}[P|X]]), \end{aligned} \quad (4.48)$$

the bias of the empirical estimator for discrimination may be positive or negative. In other words, the estimate of uncertainty is, on average, below the true value of uncertainty and the estimate of correctness is, on average, above the actual attribute value (so, correctness is better – incorrectness is less – than the estimates indicate). If the estimator for discrimination has positive bias, then estimated discrimination will overstate true discrimination; similarly, if the estimator for discrimination has negative bias, the estimates will, on average, undervalue actual discrimination.

Unfortunately, the biases of the empirical estimators of the attributes of the URR and RDC decompositions of the interval-Brier scoring rule are less amenable to analytical examination. Nonetheless, if the intervals determined by the partition grow increasingly narrow as n increases, it is possible to conclude that the degree of absolute bias of the interval-estimators will increase (although the increase is an overall increase and not necessarily a monotonic trend). This can be seen by considering, first, the case of only a single interval ($n = 1$). For a single interval, a little algebra shows that the empirical estimators of all the attributes (of both URR and RDC decompositions) are unbiased (here, because I_k^o is undefined when $n = 1$, without loss of generality we set $I_k^o = 1$). As n increases, the bias of the empirical estimators of the (interval-Brier) attributes of the interval-probabilistic forecasts tends to the bias of the empirical estimators of the (Brier) attributes of the precise-probabilistic forecasts (i.e. the difference between the two biases decreases). This follows from noting that in the limit of ever finer partitions, the empirical estimators of the attributes of the interval-Brier scoring rule tend to the empirical estimators of the attributes of the Brier scoring rule, which are biased. However, the nature of the increase of absolute bias will depend on the structure of the partition.

4.5.1 || Illustration: Simulated Data

We demonstrate the computation of the attributes and the level of bias using simulated data. To generate the data, let

$$\begin{aligned} X &\sim \text{Ber}(q) \\ P|_{X=x} &\sim \text{Beta}(\alpha_x, \beta_x) \quad x \in \{0, 1\}. \end{aligned} \quad (4.49)$$

We choose $q = 0.225$, $\alpha_0 = 2, \beta_0 = 10$ and $\alpha_1 = 5, \beta_1 = 2$. Each sample of precise-probabilistic forecasts and associated outcomes has size $T = 50$ and was generated from the joint distribution determined by model (4.49). In the sample, all precise-probabilistic forecasts are rounded to 2 decimal places; the rounded precise-probabilistic forecasts are considered the *issued* forecasts. For each simulated precise-probabilistic forecast in the sample, the value of the corresponding interval-probabilistic forecast was determined by finding the interval in which the precise-probabilistic forecast fell. By this method, to each sample of precise-probabilistic forecasts and outcomes, a corresponding sample of interval-probabilistic forecasts and outcomes was constructed.

Figures 4.1a and 4.1b show the unconditional and conditional probability distributions of the precise-probabilistic forecasts and interval-probabilistic forecasts respectively, for model (4.49).

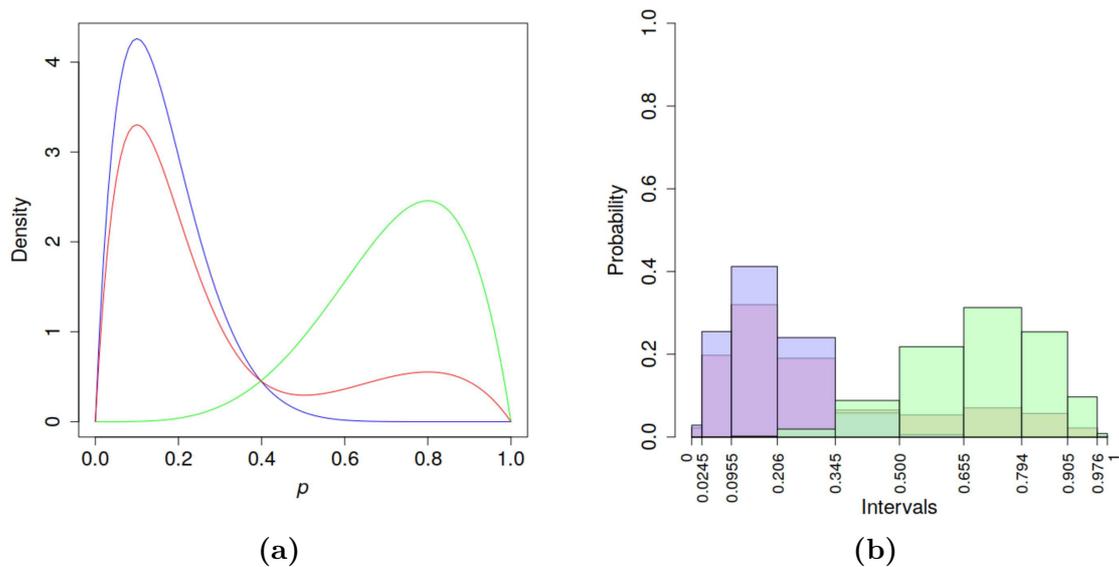


Figure 4.1: Figure 4.1a shows the unconditional (—) and conditional densities of P (given $X = 1$, —; given $X = 0$ —). For the interval-probabilistic forecasts, Figure 4.1b shows the unconditional (■) and conditional probability distributions (given $X = 1$, ■; given $X = 0$, ■). Both Figure 4.1a and Figure 4.1b correspond to model (4.49).

To estimate the attributes it remains to define the partition $0 = a_0 < a_1 < \dots < a_n = 1$. We use the following approach which allows for the easy generation of customised partitions by setting the nodes of the partition to be the quantiles of a chosen probability distribution (different probability distributions then determining different partitions): (a) choose a distribution function, F_{part} , the probability mass function of which has support $[0, 1]$, (b) having fixed the number of non-zero nodes in the partition, n , set the nodes to the $100/n\%$

percentiles of the distribution function F_{part} i.e. $a_i = F_{\text{part}}^{-1}(i/n)$ for $i = 0, \dots, n$. In our illustration, we choose to have a partition for which the intervals nearer the end points of the interval $[0, 1]$ are narrower (in other words, the more certain the forecasting system, the more precise it is required to be), by selecting F_{part} to be the distribution function of a Beta(1/2, 1/2) random variable.

In Figure 4.2, the calculated attributes are displayed. The boxplot in Figure 4.2a compares the values of the attributes of the precise-probabilistic and interval-probabilistic forecasts with respect to the Brier and interval-Brier scoring rules. Also displayed in Figure 4.2a is the mean estimate and the actual (model) value for each attribute. It is seen that the mean estimate is (up to sampling error) less than the actual value for uncertainty and refinement, but the mean estimate is greater than or equal to the actual value for the remaining attributes, with the bias in reliability and resolution being much greater than the bias in discrimination and correctness. In Figure 4.2b the ratio of the true attribute values under the precise-probabilistic forecasts to the true attribute values under the interval-probabilistic forecasts is displayed, and shows the degradation in attribute values that accompanies the loss of precision in the forecasts: for the interval-probabilistic forecasts, uncertainty is lower, resolution is poorer, as is reliability, while there is less refinement, weaker discrimination and more incorrectness.

In Figure 4.3 the absolute difference between the level of bias of the empirical estimators of the attributes under precise-probabilistic forecasts ('precise-bias') and the level of bias of the empirical estimators of the attributes under interval-probabilistic forecasts ('interval-bias') is shown. As the number of intervals increases, the interval-bias converges to the precise-bias (i.e. the absolute difference decreases); note, however, that this convergence is particular to partitions for which all intervals narrow as n increases (a property of a partition defined by our approach above).

These results serve to offer some guidelines when interpreting the results of applying the estimators to real data. An example of such an application is presented now.

4.5.2 || Application: Precipitation Data

We evaluate estimates of the attributes of the URR and RDC decompositions of the Brier and interval-Brier scoring rules for precipitation data provided by the UKMO and the ABOM. A detailed description of the data is given in chapter 3. Briefly, the raw data provided by the respective meteorological offices is transformed first into a set of precise-probabilistic forecast and outcome pairs, where each outcome is that of the dichotomous observation rain/no rain.

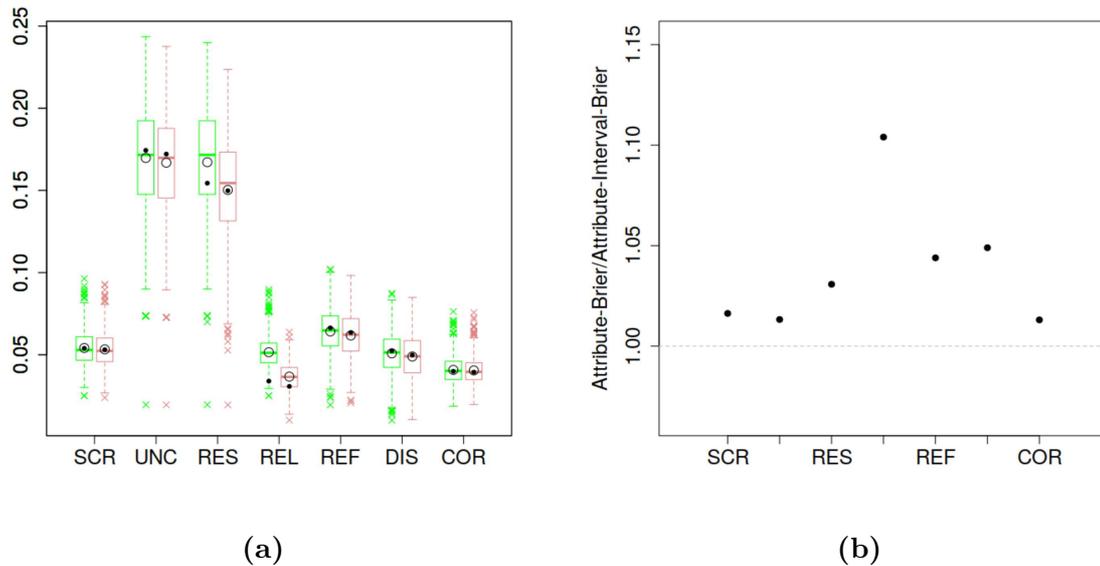


Figure 4.2: The values of the attributes under the theoretical model (4.49) are shown in Figure 4.2a (values for precise-probabilistic forecasts are shown in green, values for interval-probabilistic forecasts are shown in brown). Estimates of the true values were calculated by applying the formulae in Table 4.1 to samples of size $T = 50$ simulated from the model. A total of $M = 1000$ samples were simulated, giving M estimates of each attribute, which were used to calculate a boxplot and a mean estimate (marked by \circ) of each attribute. Also shown are the true values of the attributes (marked by \bullet ; where necessary in computing the true values, integrals were evaluated numerically using the `integrate` function in **R** (Ihaka and Gentleman, 1996; R Core Team, 2017)). The difference between the true and mean-estimates of each attribute are approximations of the bias of the empirical estimators. The ratio of the true attribute values under the Brier scoring rule and the interval-Brier scoring rule are shown in Figure 4.2b, demonstrating, for the given partition, the relative differences between the quality of the interval and precise forecasting systems.

The precise-probabilistic forecasts are used to synthesise interval-probabilistic forecasts by assigning to each precise-probabilistic forecast the interval in which it falls. The partitions used are given in Table 4.2 (as we have emphasised in chapter 3, the interval-forecasts are artificial and are not forecasts that have been issued by either meteorological office).

The UKMO data was used to examine how the estimated attribute values change with lead-time (hours). In contrast, the ABOM data was used to determine how the estimated attribute values varied at different locations for the same lead-time (hours). The behaviour of the estimates is shown in Figures 4.4 to 4.11.

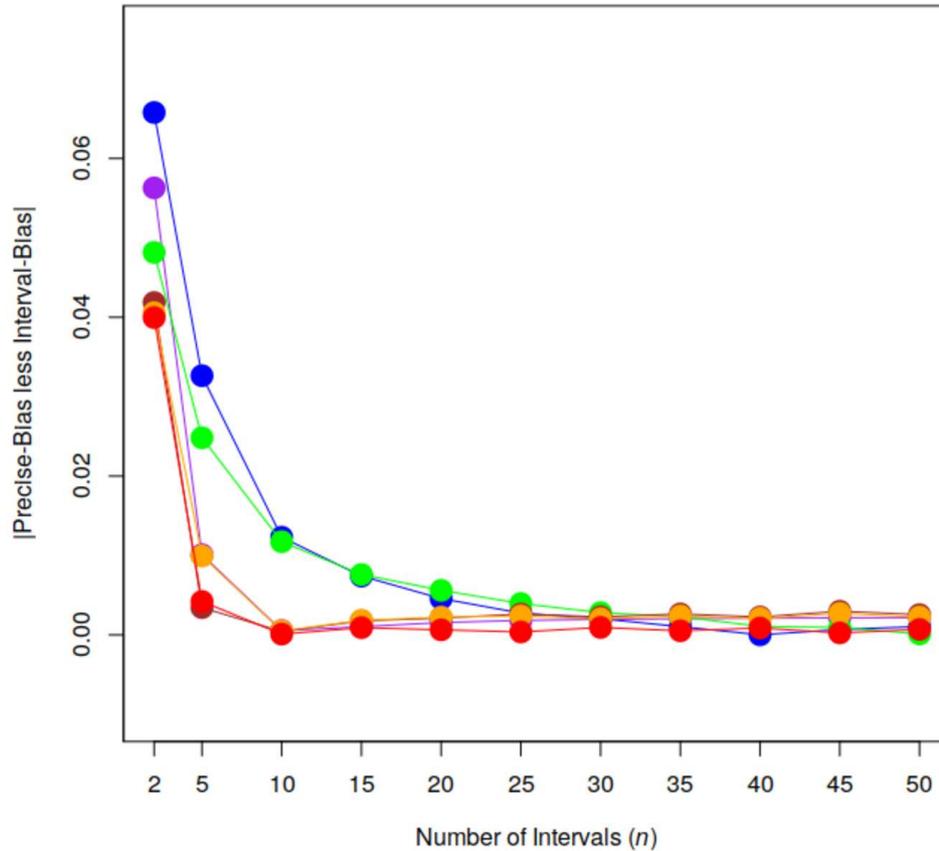


Figure 4.3: The model is (4.49). For each value of n , the intervals from which the interval-probabilistic forecasts can be selected are determined by setting the nodes to be the $(100/n)$ -percentiles of the $\text{Beta}(1/2, 1/2)$ probability distribution (as described above). For each set of intervals (i.e. each n), $M = 1000$ samples of $T = 50$ forecast-outcome pairs were generated for both precise-probabilistic and interval-probabilistic forecasts (according to model (4.49)). For a given n , the bias of each attribute's estimator was calculated as the mean estimate (over the M samples) for the attribute less its actual (model) value; the bias of estimators under precise-probabilistic forecasts is 'precise-bias', the bias of estimators under interval-probabilistic forecasts is 'interval-bias'. The absolute *difference* between the precise-bias and interval-bias of each attribute is plotted against n (uncertainty ●, resolution ●, reliability ●, refinement ●, discrimination ● and correctness ●). As n increases, the interval-bias converges to the precise-bias.

	a_0	a_1	a_2	a_3	a_4	a_5	a_6	a_7	a_8	a_9	a_{10}	a_{11}	a_{12}	a_{13}	a_{14}	a_{15}
UKMO	0	0.025	0.05	0.10	0.20	0.25	0.30	0.40	0.50	0.60	0.70	0.75	0.80	0.90	0.95	1
ABOM	0	0.025	0.075	0.15	0.25	0.35	0.45	0.55	0.65	0.75	0.85	0.925	0.975	1		

Table 4.2: The nodes of the partitions used to define the interval-forecasts synthesised from the precise-probabilistic forecasts provided by the UKMO and the ABOM.

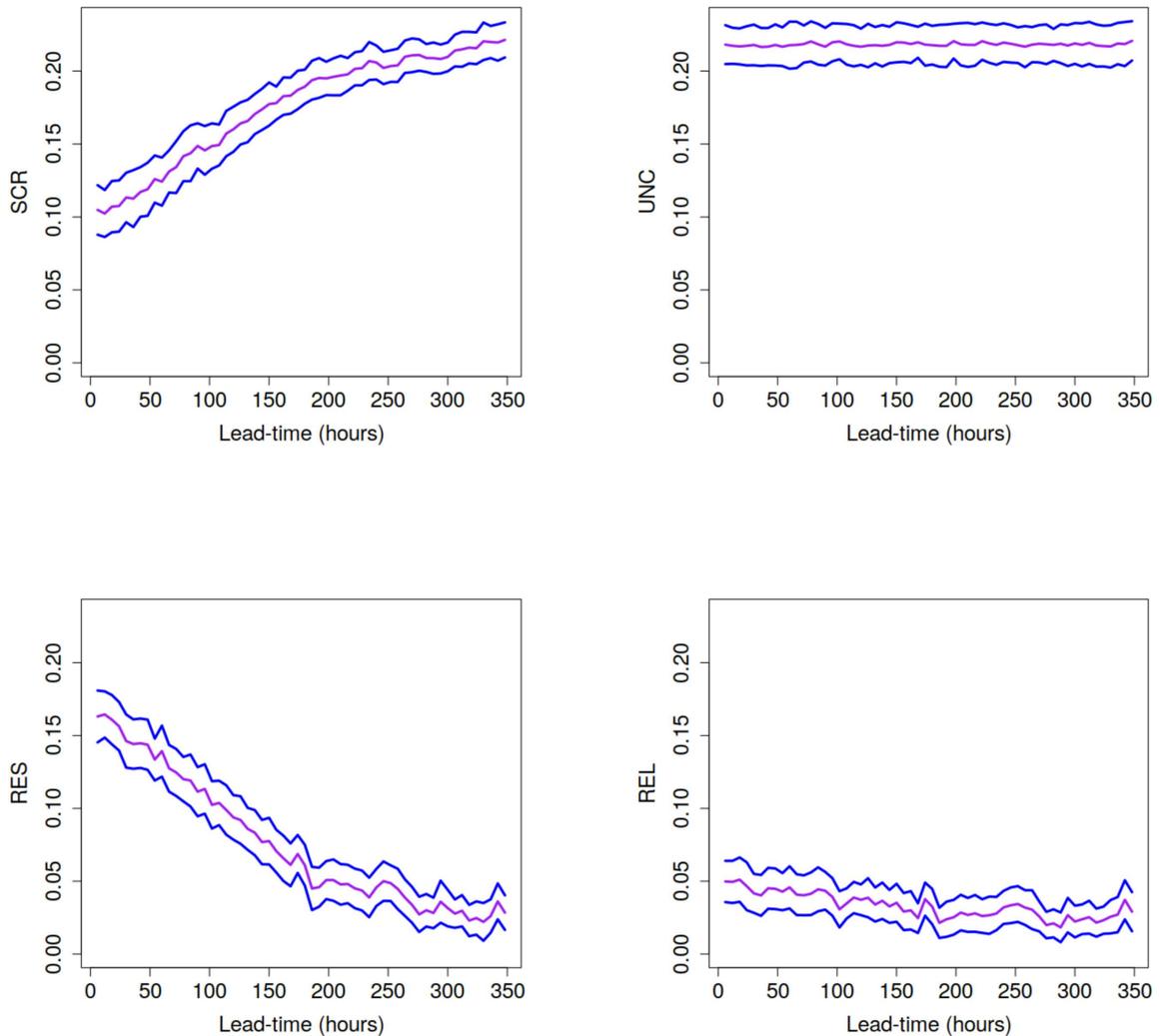


Figure 4.4: URR Attributes for *precise*-probabilistic forecasts. The figures display estimates of each attribute at different lead-times (hours); the estimates are computed from UKMO precipitation forecasts at Heathrow Airport (for each lead-time the sample size was $T = 590$, less one or two missing data for some lead-times). The purple line (—) shows the attribute estimate (\hat{a}) calculated from that sample associated with the value on the horizontal axis, the blue lines (—) show the 95% confidence interval calculated as $(\hat{a} - z_{0.975}\hat{\sigma}_B, \hat{a} + z_{0.025}\hat{\sigma}_B)$, where $\hat{\sigma}_B$ is the standard deviation of the estimate approximated from $B = 100$ bootstrapped samples. The abbreviations used are: SCR for expected score, UNC for uncertainty, RES for resolution, and REL for reliability.

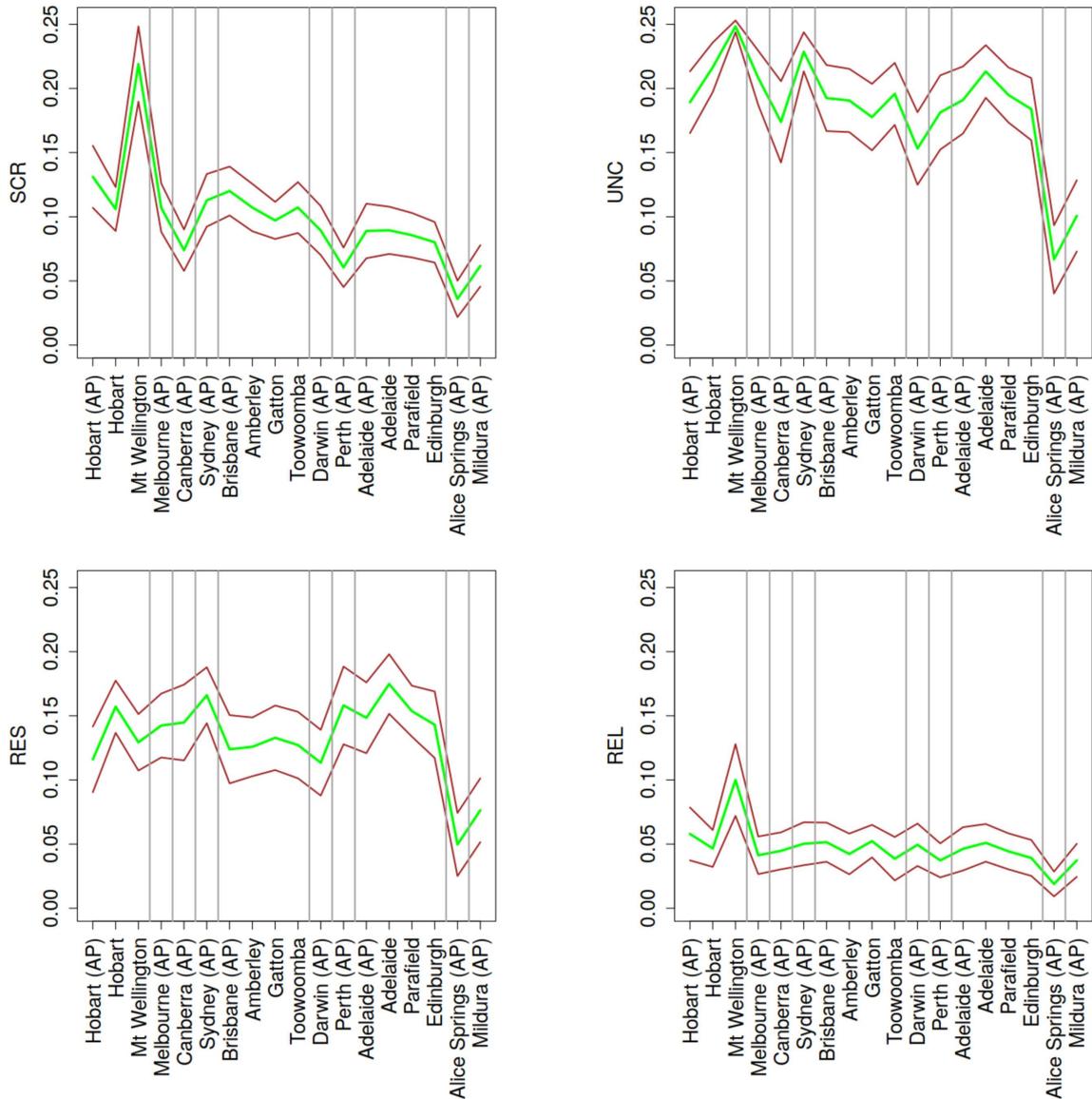


Figure 4.5: URR Attributes for *precise*-probabilistic forecasts. The figures show estimates of each attribute. The estimates are calculated using ABOM *precise*-probabilistic forecasts for precipitation, at different locations around Australia for a 12-hour lead-time; all sites between two vertical grey lines are neighbouring locations (for each site the sample size was $T = 285$, except in the case of missing data). The green line (—) shows the attribute estimate (\hat{a}) calculated from that sample associated with the value on the horizontal axis, the brown lines (—) show the 95% confidence interval calculated as $(\hat{a} - z_{0.975}\hat{\sigma}_B, \hat{a} - z_{0.025}\hat{\sigma}_B)$, where $\hat{\sigma}_B$ is the standard deviation of the estimate approximated from $B = 100$ bootstrapped samples. The abbreviations used are: SCR for expected score, UNC for uncertainty, RES for resolution, and REL for reliability.

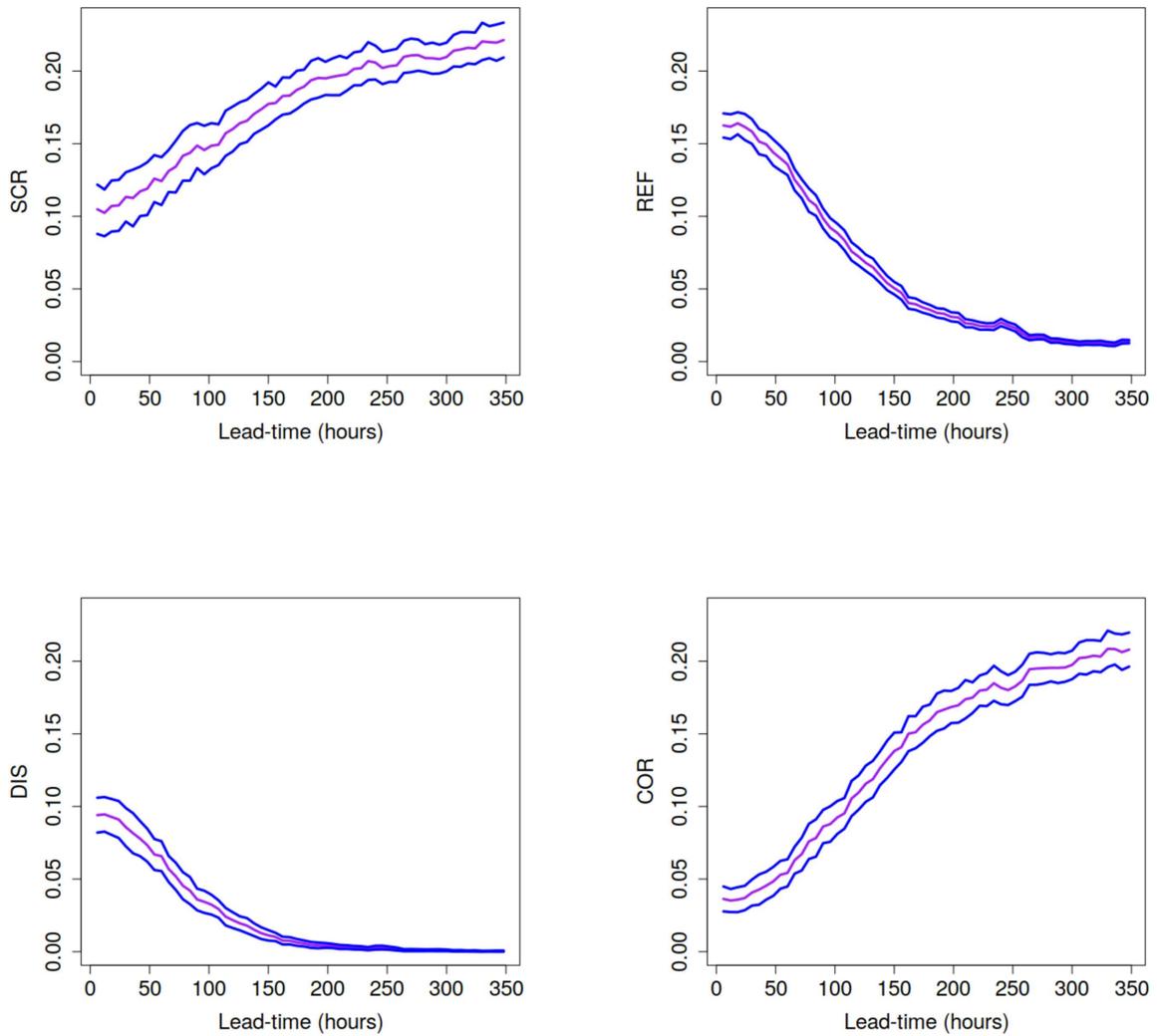


Figure 4.6: RDC Attributes for (UKMO) *precise*-probabilistic forecasts of different lead-times (hours) for precipitation at Heathrow Airport. For a description of the data and a key to the figures, see the caption for Figure 4.4. Abbreviations used in the above figures are: SCR for expected score, REF for refinement, DIS for discrimination, and COR for correctness.

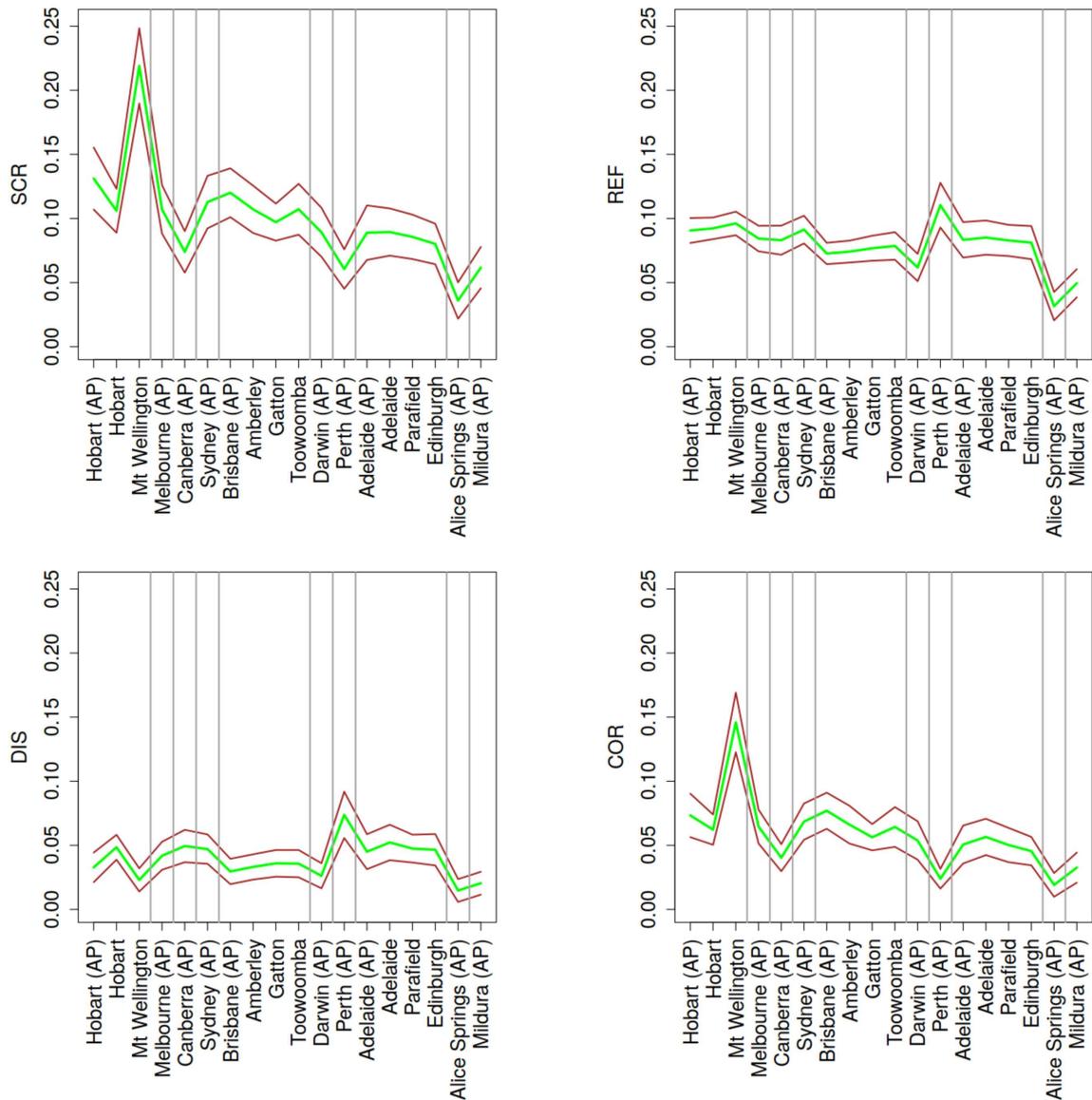


Figure 4.7: RDC Attributes for (ABOM) *precise*-probabilistic forecasts with a lead-time of 12 hours at different locations in Australia. For a description of the data see the caption for Figure 4.5. Abbreviations used in the above figures are: SCR for expected score, REF for refinement, DIS for discrimination, and COR for correctness.

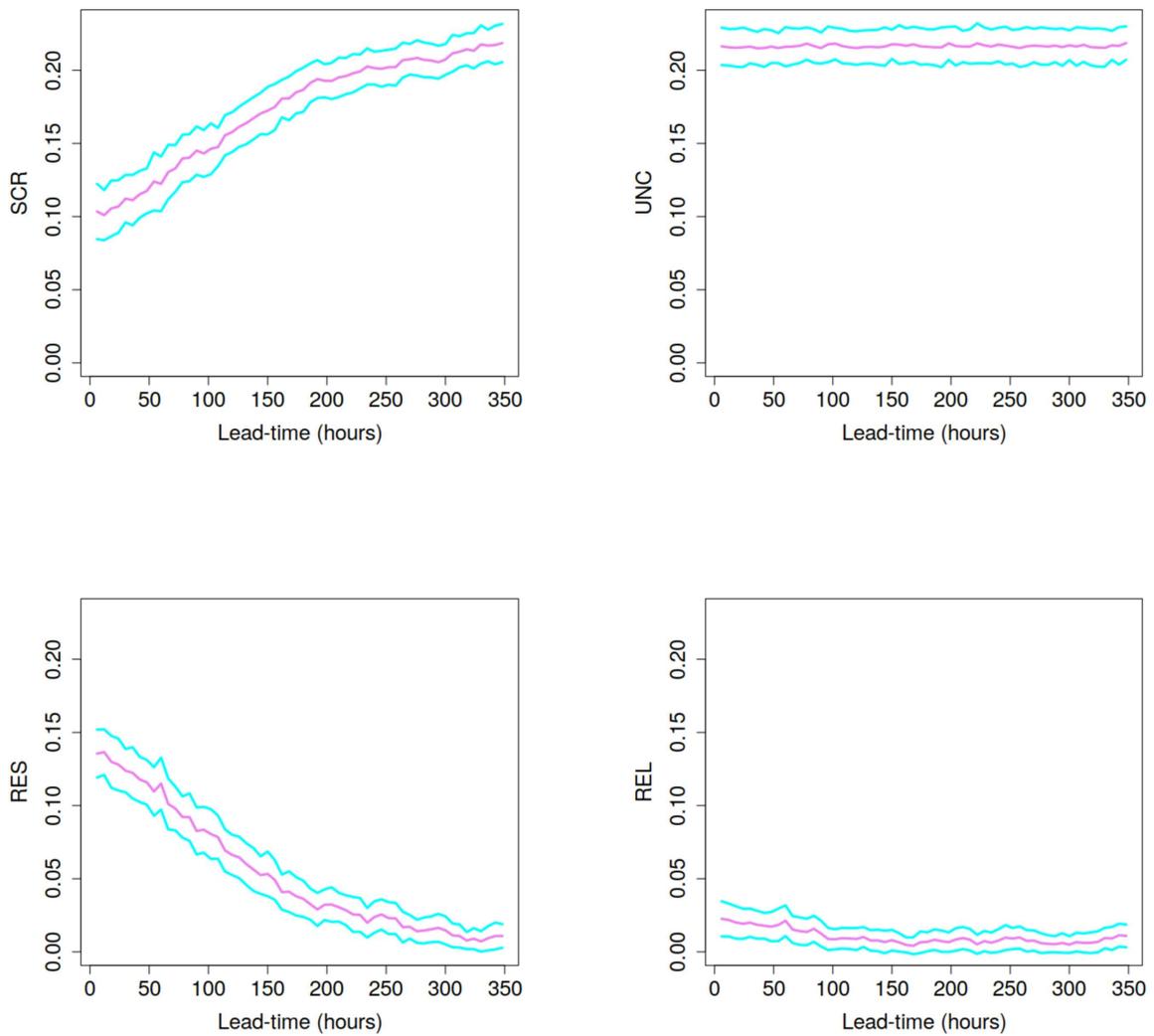


Figure 4.8: URR Attributes for *interval*-probabilistic forecasts. The figures display estimates of each attribute at different lead-times (hours); the estimates are computed from interval-probabilistic forecasts constructed from UKMO precise-probabilistic forecasts for precipitation at Heathrow Airport (for each lead-time the sample size was $T = 590$, less one or two missing data for some lead-times). The magenta line (—) shows the attribute estimate (\hat{a}) calculated from that sample associated with the value on the horizontal axis, the cyan lines (—) show the 95% confidence interval calculated as $(\hat{a} - z_{0.975}\hat{\sigma}_B, \hat{a} - z_{0.025}\hat{\sigma}_B)$, where $\hat{\sigma}_B$ is the standard deviation of the estimate approximated from $B = 100$ bootstrapped samples. The abbreviations used are: SCR for expected score, UNC for uncertainty, RES for resolution, and REL for reliability.

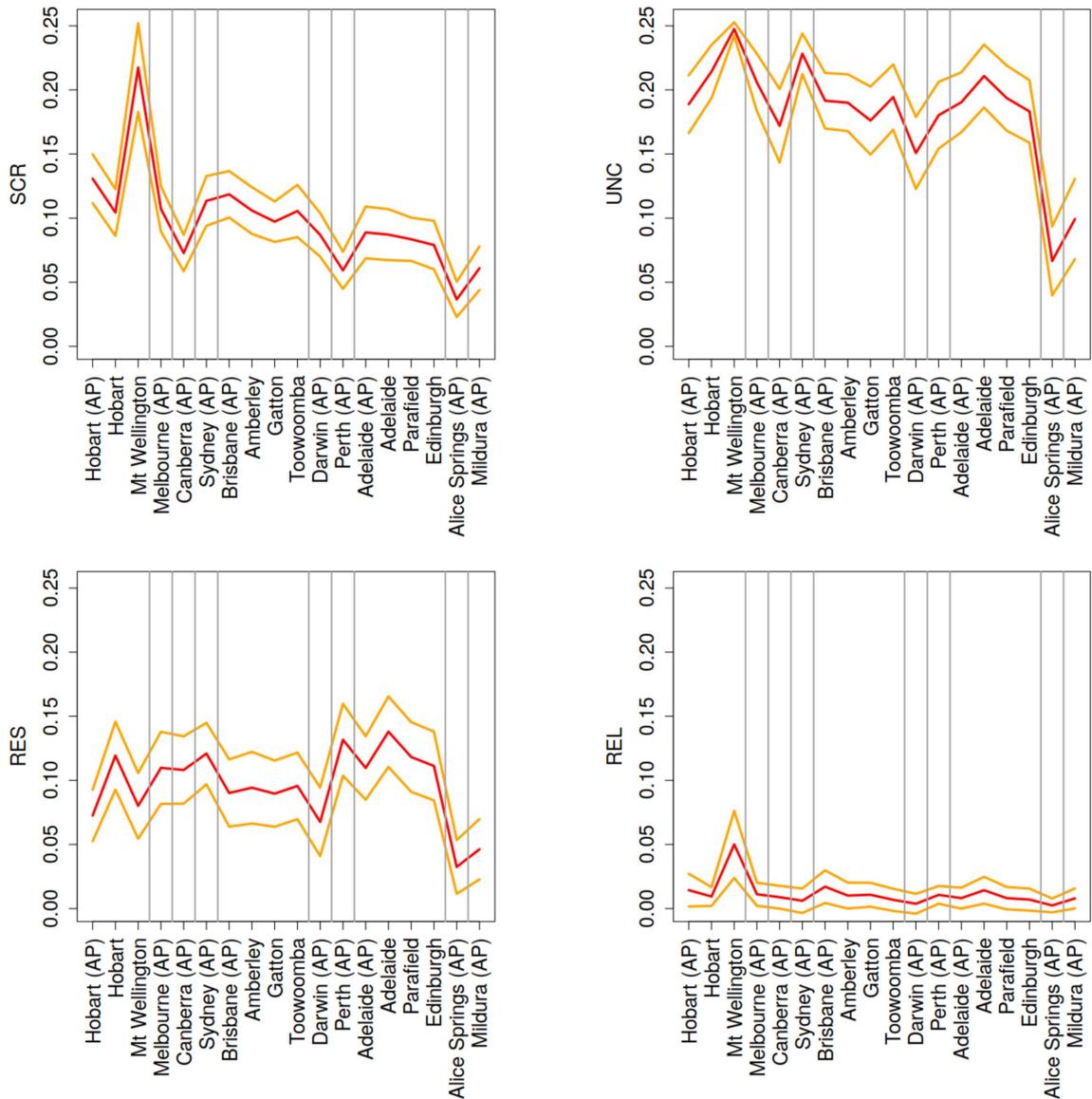


Figure 4.9: URR Attributes for *interval*-probabilistic forecasts. The figures show estimates of each attribute. The estimates are calculated using interval-probabilistic forecasts synthesised from ABOM precise-probabilistic forecasts for precipitation, at different locations around Australia for a 12-hour lead-time; all sites between two vertical grey lines are neighbouring locations (for each site the sample size was $T = 285$, except in the case of missing data). The red line (—) shows the attribute estimate (\hat{a}) calculated from that sample associated with the value on the horizontal axis, the orange lines (—) show the 95% confidence interval calculated as $(\hat{a} - z_{0.975}\hat{\sigma}_B, \hat{a} - z_{0.025}\hat{\sigma}_B)$, where $\hat{\sigma}_B$ is the standard deviation of the estimate approximated from $B = 100$ bootstrapped samples. The abbreviations used are: SCR for expected score, UNC for uncertainty, RES for resolution, and REL for reliability.

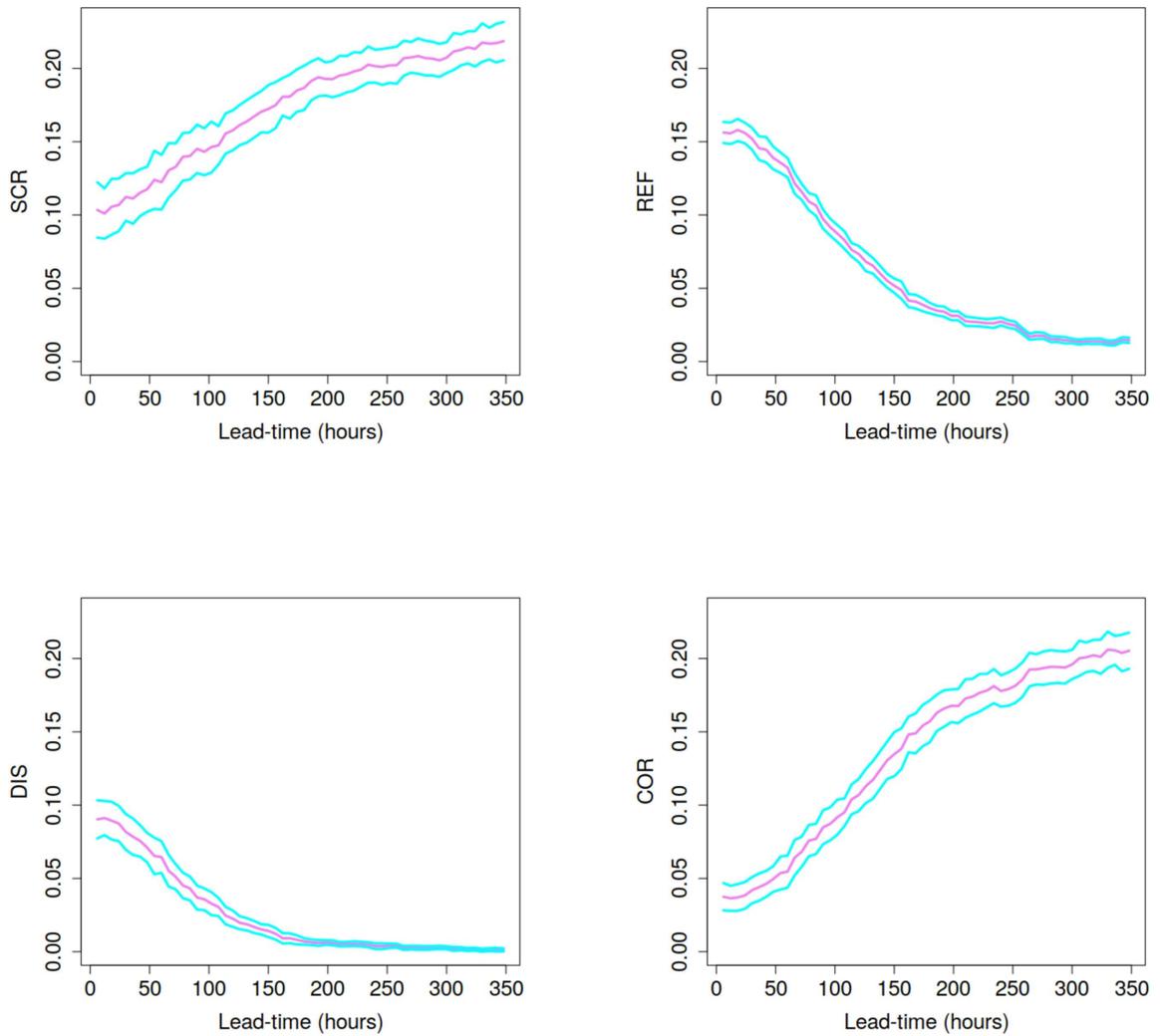


Figure 4.10: RDC Attributes for (UKMO) *interval*-probabilistic forecasts of different lead-times (hours) for precipitation at Heathrow Airport. For a description of the data and a key to the graphs, see the caption for Figure 4.8. Abbreviations used in the above figures are: SCR for expected score, REF for refinement, DIS for discrimination, and COR for correctness.

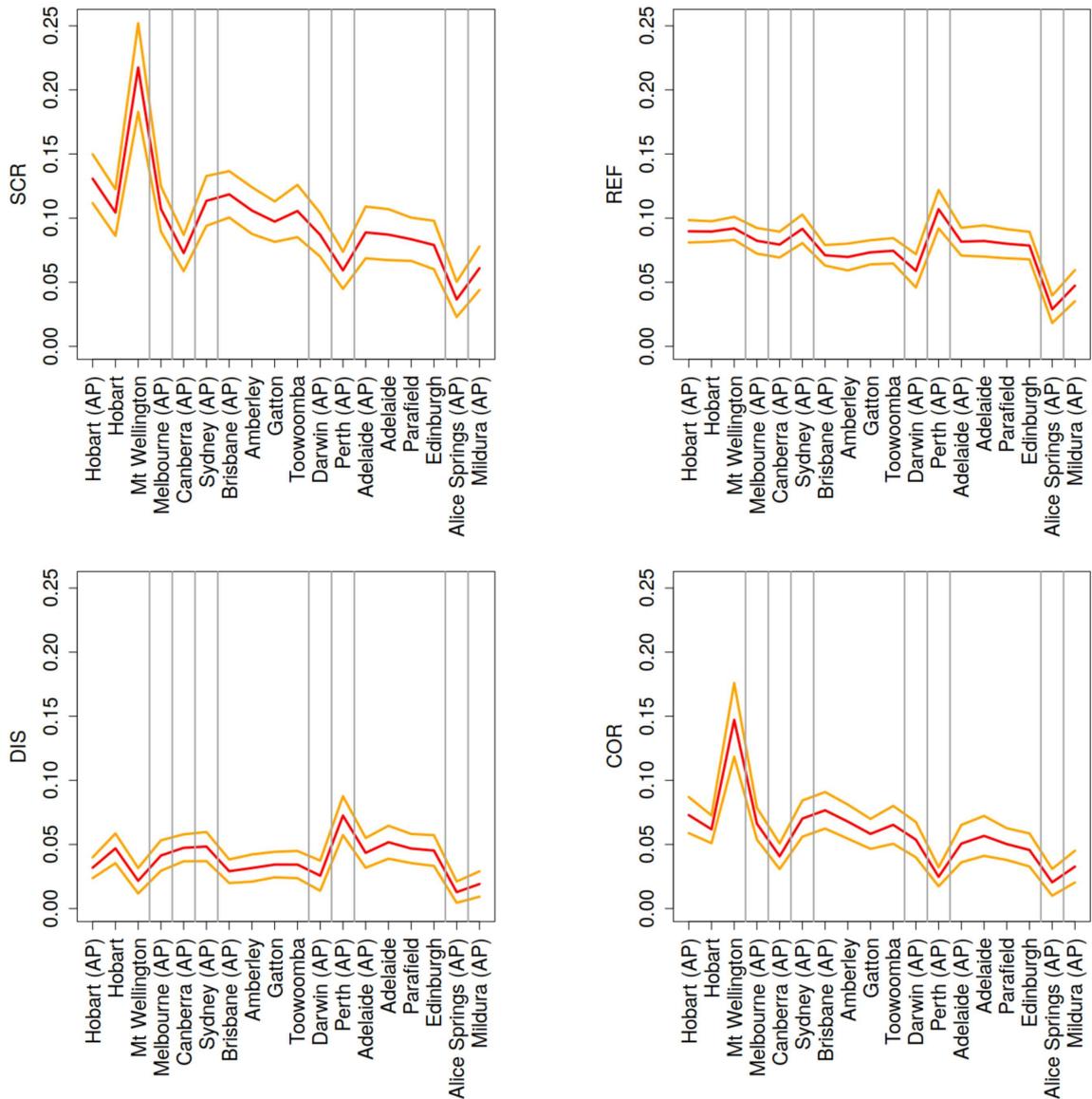


Figure 4.11: RDC Attributes for (ABOM) *interval*-probabilistic forecasts with a lead-time of 12 hours at different locations in Australia. For a description of the data and the caption, see the caption for Figure 4.9. Abbreviations used in the above figures are: SCR for expected score, REF for refinement, DIS for discrimination, and COR for correctness.

The attributes of both the precise-probabilistic forecasts and the interval-probabilistic forecasts deteriorate as forecast lead-time increases, with precise- and interval-probabilistic forecasts displaying similar patterns. While the environmental conditions are stable (i.e. uncertainty is constant across lead-times), forecasts with longer lead-times are less capable than forecasts with short lead-times of distinguishing the outcomes that will eventually occur (i.e. resolution decreases or worsens as lead-time increases), although there is only a small difference between the forecasts of different lead-times in their ability to identify the proportion of occasions on which precipitation is recorded. The RDC attributes show that as lead-time increases, the forecasts become more similar (i.e. refinement decreases), but because the environment will not change with lead-time, this similarity reflects that at longer lead-times the forecasts are more uniform, with different outcomes being preceded by much the same forecasts (shown by discrimination decreasing with increasing lead-time). It is also seen from the correctness attribute that the forecasts with longer lead-times are more incorrect than the forecasts of shorter lead-times, an intuitive trend.

In comparison, the attributes are broadly stable across different locations. For sites with lower uncertainty and less changeable environmental conditions (e.g. Alice Springs (AP)), forecasts need not vary and this gives lower resolution and will show lower refinement and discrimination, but these same forecasts may, nonetheless show good (i.e. low) reliability and correctness. Interestingly, in the first group of neighbouring sites (on the left of each graph), the uncertainty differs markedly within the group (a consequence of the contrast between Hobart on the coast of Tasmania to the nearby yet inland and elevated Mount Wellington) and forecasts perform quite differently (forecasts for Hobart show better resolution and reliability than Mount Wellington and better discrimination and correctness, although across the group of sites forecasts are similarly refined, suggesting that the forecasting system does not account materially for the substantive factors affecting precipitation at Mount Wellington that are additional to those factors affecting precipitation at Hobart).

[4.6] Discussion and Conclusion

Probabilistic forecasts for a binary observation, X (with values 0, for non-occurrence, and 1, for occurrence) are typically studied as precise-probabilistic forecasts, forecasts expressed as precise probabilities that $X = 1$. Often, however, probabilistic forecasts for X are published as interval-probabilistic forecasts, forecasts expressed as an interval of values for the probability that $X = 1$. In the previous chapter, proper scoring rules for interval-probabilistic forecasts, referred to as interval-proper scoring rules, were introduced.

While interval-proper scoring rules advance the evaluation of interval-probabilistic forecasts by allowing for the accuracy, i.e. expected score, to be calculated, a thorough evaluation of interval-probabilistic forecasts requires additional measures of the qualities of the forecasts. For precise-probabilistic forecasts, standard measures of forecast quality have historically been computed by decomposing accuracy. There are two common decompositions: the uncertainty-resolution-reliability (URR) decomposition and the refinement-discrimination-correctness (RDC) decomposition.

In this chapter we have derived versions of the URR and RDC decompositions for one particular interval-proper scoring rule, the interval-Brier scoring rule. These decompositions allow, in principle, a complete evaluation of interval-probabilistic forecasts for binary observations evaluated using the interval-Brier scoring rule.

For such a comprehensive analysis to be practical, estimators of the components of the decompositions are required. Empirical estimators, estimators based on the empirical joint distribution function of forecasts and outcomes, were proposed and their statistical properties considered. The estimators were then applied to interval-probabilistic forecasts of precipitation synthesised from actual precise-probabilistic forecasts. The estimates suggested that as the lead-time of interval-probabilistic forecasts increases, the quality of the forecasts decreases. Estimates were also computed for different sites at the same lead-time. Although, in general the estimated attributes were similar across all sites, there was enough deviation in the estimates to argue for a review of the forecasts at some sites, so demonstrating the usefulness of the decompositions in the verification of interval-probabilistic forecasts.

Throughout this chapter, comparison is made between the attributes of the interval-Brier scoring rule and the attributes of the Brier scoring rule for precise-probabilistic forecasts. While theoretically, the attributes of the interval-Brier scoring rule generalise the attributes of the Brier scoring rule, it is in evaluating estimates for both sets of attributes that it is most clearly seen that precise-probabilistic forecasts, if available, have a higher quality than their corresponding interval-probabilistic forecasts, but that the interval-probabilistic forecasts, although of lower quality, display the same general patterns of performance.

5 Possible Directions for Future Work

[5.1] Introduction

From the position reached in earlier chapters, we now consider some extensions that either reinforce the results we have obtained or build on them. We look separately at: *(i)* topics more closely aligned with decompositions of proper scoring rules, and *(ii)* the smaller but promising area of interval-probabilistic forecasts.

[5.2] Applying the Decompositions

The general URR and RDC decompositions developed in chapter 2 have been applied to several of the more well-known proper scoring rules for (precise-)probabilistic forecasts. There are other prominent proper scoring rules, for example the Dawid-Sebastiani scoring rule and kernel scoring rules ([Gneiting and Raftery, 2007](#)), for which the URR and RDC decompositions are, as far as we are aware, unknown and to which the URR and RDC decompositions of chapter 2 could be applied.

Deriving the URR and RDC decompositions for a proper scoring rule is only one aspect determining the qualities of a forecasting system. It is also necessary to estimate the attributes established by the URR and RDC decompositions. We have briefly discussed empirical estimators for the attributes of the different decompositions we have examined, but a more thorough consideration of estimation and statistical inference for the attributes is needed (we mention this again briefly when reviewing interval-probabilistic forecasts below).

Our derivations of the URR and RDC decompositions of the expected score have given regard only to the marginal and conditional distributions of the forecasts and outcome. This approach has been led by the work of [Murphy and Winkler \(1987\)](#) who emphasise that it is only the joint distribution of the forecasts and outcomes that need be taken into consideration for evaluating a forecasting system. However, the framework of [Murphy and Winkler \(1987\)](#) can be criticised as being too narrow.

[Ehm and Ovcharov \(2017\)](#) (see also the references therein) have recently proposed that there may be “auxiliary information”, other than the information provided by the forecasts and outcomes, that is material to the evaluation of a forecasting system. By grouping forecasts according to the supplementary information a more detailed and nuanced assessment of the forecasts is possible. For example, in assessing weather forecasts, the geographical region of each forecast may enhance the forecasts’ evaluation. As another example, separating global economic forecasts according to the size of the economies to which the forecasts relate

(or some similar division) can present a more subtle analysis of the forecasts' performance. Such detail is more pressing when the supplementary information by which the forecasts are grouped concerns extremes: how well do hydrological forecasts perform during a drought, or how well do public health forecasts anticipate the reach of a serious virus.

One approach to incorporating supplementary information in the verification of forecasts is by using scoring rules that include a weighting function to amplify the contribution of a subgroup of forecasts (see, for example, [Lerch et al. \(2017\)](#)). But [Ehm and Ovcharov \(2017\)](#) have recently considered an alternative approach: to adopt a conventional scoring rule, but change the decomposition to recognise the additional information available. [Ehm and Ovcharov \(2017\)](#) derive a URR decomposition in which the resolution and reliability terms are 'localised' by explicitly conditioning on a random variable representing the supplementary information. The result of [Ehm and Ovcharov \(2017\)](#) is a quite general URR decomposition, but has the disadvantage that the attributes of their URR decomposition can be negative, which they admit "makes interpretations difficult" ([Ehm and Ovcharov, 2017](#)). To avoid negative attributes, one route may be to develop conditional-proper scoring rules. In this regard earlier work by [Lindley \(1982\)](#) (see also, [Gilio and Sanfilippo \(2011\)](#)) who proposes a definition for conditional scoring rules may be of some help. Should a definition of conditional-proper scoring rules be agreed and, together with the ideas of [Ehm and Ovcharov \(2017\)](#), allow for a conditional-URR decomposition with everywhere-positive attributes, there is the further potential for the derivation of a conditional-RDC decomposition.

[5.3] Interval-Probabilistic Forecasts

Interval-probabilistic forecasts are, in the form in which we have introduced them in this thesis, new to verification theory and there is no other published work in this area that we are aware of, apart from the more general and abstract studies of [Lambert and Shoham \(2009\)](#); [Lambert et al. \(2008\)](#); [Lambert \(2013\)](#) and [Frongillo and Kash \(2014\)](#) that we have mentioned. This leaves many interesting questions still to be answered, some of which we now touch upon.

Proper scoring rules for interval-probabilistic forecasts, which we have called interval-proper scoring rules, were introduced in chapter 3. In chapter 4, the URR and RDC decompositions of one particular interval-proper scoring rule, the interval-Brier scoring rule, were derived. Yet as was discussed in chapter 2, the attributes in the decompositions, although important, are *indirect* measures of the properties of forecasts. For precise-probabilistic forecasts, as was originally emphasised by [Murphy and Winkler \(1987\)](#), the direct measures of forecast

quality are distributional i.e. defined in terms of the joint probability distribution of the forecasts and outcomes. So far, no definitions, in distributional terms, of the properties of interval-probabilistic forecasts have been proposed or investigated. In the context of chapters 3 and 4, for which the observation, X , is binary, the forecaster's belief is held as a precise-probability, p , and the forecast associated with this belief is the interval containing p , $I(p)$, then following similar definitions in chapter 2 for precise-probabilistic forecasts, natural definitions for some distributional qualities are,

(i) no resolution, if

$$\Pr(X = x|I(P) = I(p)) = \Pr(X = x) \quad \text{for all } p \in \mathcal{P}, x \in \mathcal{X}, \quad (5.1)$$

(ii) perfect reliability, if

$$\Pr(X = 1|I(P) = I(p)) \in I(p) \quad \text{for all } p \in \mathcal{P}, \quad (5.2)$$

(iii) no discrimination, if

$$\Pr(I(P) = I(p)|X = x) = \Pr(I(P) = I(p)) \quad \text{for all } x \in \mathcal{X}, p \in \mathcal{P}. \quad (5.3)$$

The corresponding attributes of the URR and RDC decompositions (see equations (4.17), (4.18) and (4.38)) are necessarily zero if equations (5.1) to (5.3) hold.

Having proposed definitions of the distributional attributes, approaches must be developed that allow these distributional definitions to be applied. A useful method of assessing reliability as defined in equation (5.2), would be a reliability diagram analogous to the fundamental reliability diagram for precise-probabilistic forecasts. Label the intervals that can be forecast as I_k , $k = 1, \dots, n$, for some n , with interval I_k having infima I_k^- and suprema I_k^+ . One possible construction of a reliability diagram for interval-probabilistic forecasts would be to choose, for each k , a point $p_k \in I_k$ and plot three interpolated curves: (a) (p_k, I_k^-) , $k = 1, \dots, n$, (b) $(p_k, \Pr(X = 1|I_k))$, $k = 1, \dots, n$, and (c) (p_k, I_k^+) , $k = 1, \dots, n$. In the absence of any information about the distribution of the forecaster's beliefs within each interval, we could choose $p_k = (I_k^- + I_k^+)/2$. We refer to the curves (a), (b) and (c) collectively as an interval-reliability diagram. As the number of intervals, n , increases, curves (a) and (c) should converge to the line of perfect reliability and the interval-reliability diagram becomes a standard reliability diagram for precise-probabilistic forecasts. The interpretation of the interval-reliability diagram mirrors the interpretation of the reliability diagram for precise-probabilistic forecasts (see, for example Wilks, 2006, pages 288-289): if curve (b) lies entirely below curve (a) then the forecaster is overforecasting ($\Pr(X = 1|I_k) < I_k^-$ for all k); if curve

(*b*) lies entirely above curve (*c*) then the forecaster is underforecasting ($\Pr(X = 1|I_k) > I_k^+$ for all k); the forecaster is overconfident if curve (*b*) lies above curve (*c*) for small k and below curve (*a*) for large k , while an underconfident forecaster has an interval-reliability diagram in which curve (*b*) is below curve (*a*) for small k but above curve (*c*) for large k .

An interval-reliability diagram (of any form) does not amount to a statistical assessment of the reliability of interval-probabilistic forecasts. For precise-probabilistic forecasts, there are statistical tests of reliability (see, for example, [Bröcker \(2012\)](#); [Lemeshow and Hosmer \(1982\)](#); [Lemeshow et al. \(1988\)](#); [Spiegelhalter \(1986\)](#); [Stallard \(2009\)](#) as well as [Redelmeier et al. \(1991\)](#); see also [Bröcker and Smith \(2007b\)](#) for a discussion of the statistical assessment of reliability through reliability diagrams). Currently no comparable statistical analysis exists for the reliability of interval-probabilistic forecasts.

To evaluate the discrimination (in distributional terms) of interval-probabilistic forecasts, an approach based on the receiver operating characteristic (ROC) curve can be employed. Conceptually, interval-probabilistic forecasts are no different to responses to a rating-scale experiment of signal-detection theory (see, for example, [McNicol \(1972, pages 25,99\)](#), [Green and Swets \(1966, pages 40,86\)](#)) in which the recipient of data (of some form) must chose from a scale, their degree-of-belief that the data represented a signal (rather than noise); rating-scale experiments are familiar in medicine (e.g. [Hanley and McNeil, 1982](#)). The conceptual equivalence of interval-probabilistic forecasts to rating-scale experiments allows the large body of theory developed for investigating discriminatory performance in rating-scale experiments, the core of which is ROC curve theory, to be applied to evaluating the discrimination of interval-probabilistic forecasts; ROC analysis has its criticisms (see, for example, [Hand \(2009\)](#) and [Hand and Anagnostopoulos \(2013\)](#)) and an alternative is the Lorenz curve approach of [Lee \(1999\)](#). In either case, the pivotal step is to index the intervals that can be forecast, $I_k, k = 1, \dots, n$, and then define a random variable κ with values in the set of indices $\{k|k = 1, \dots, n\}$. Either a ROC curve or Lorenz curve can be constructed from estimates of the conditional probabilities $\Pr(\kappa = k|X = 0)$ and $\Pr(\kappa = k|X = 1)$ for $k = 1, \dots, n$.

The distributional attributes described above are concerned with the detailed features of interval-probabilistic forecasts. Here, the interval-probabilistic forecasts are selected from a fixed collection of exhaustive intervals by the forecaster whose belief is held as a precise probability. This leads readily to two particular extensions.

First, the interpretation of an interval-probabilistic forecast is that every probability in the interval is equally likely as the probability of a 1-outcome for X . To address this limitation, the forecaster could be allowed to issue probabilistic-probabilistic forecasts in which the intervals remain fixed, but the forecaster issues a distribution over the interval as their forecast

(the current interval-probabilistic forecasts are a special case in which the forecaster chooses a uniform distribution over an interval as their probabilistic-probabilistic forecast).

Second, a singular premise in our development, in chapters 3 and 4, of verification results for interval-probabilistic forecasts, is that the forecaster's belief for the outcome $X = 1$, is held as a precise probability (but must be issued by choosing an interval of probabilities). It is interesting to consider how verification might be performed when the forecaster's belief is also an interval; in other words, the forecaster is only able to form their belief as a range of probabilities. The forecaster's belief should then also be expressed as an interval, either from a predetermined set, or if greater freedom is to be permitted as imprecise probabilities (for a brief overview of imprecise probabilities, see [Troffaes et al., 2001](#)).

References

- J.L. Anderson. A method for producing and evaluating probabilistic forecasts from ensemble model integrations. *Journal of Climate*, 9(7):1518–1530, 1996. [48](#)
- F. Atger. Spatial and interannual variability of the reliability of ensemble-based probabilistic forecasts: Consequences for calibration. *Monthly Weather Review*, 131:1509–1523, 2003. [12](#), [62](#)
- F. Atger. Estimation of the reliability of ensemble-based probabilistic forecasts. *Quarterly Journal of the Royal Meteorological Society*, 130:627–646, 2004. [62](#)
- R. Benedetti. Scoring rules for forecast verification. *Monthly Weather Review*, 138:203–211, 2010. [49](#)
- S. Bentzien and P. Friederichs. Decomposition and graphical portrayal of the quantile score. *Quarterly Journal of the Royal Meteorological Society*, 140(683):1924–1934, 2014. [14](#), [18](#), [28](#), [54](#)
- J.M. Bernardo. Expected information as expected utility. *The Annals of Statistics*, 7(3):686–690, 1979. [7](#), [49](#)
- G. Blattenberger and F. Lad. Separating the Brier score into calibration and refinement components: A graphical exposition. *The American Statistician*, 39(1):26–32, 1985. [12](#), [38](#)
- Z.B. Bouallègue, P. Pinson, and P. Friederichs. Quantile forecast discrimination ability and value. *Quarterly Journal of the Royal Meteorological Society*, 141:3415–3424, 2015. [58](#)
- G.W. Brier. Verification of a forecaster’s confidence and the use of probability statements in weather forecasting. Research report no. fr-16, U.S. Department of Commerce Weather Bureau, 1944. [4](#), [10](#), [11](#)
- G.W. Brier. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1–3, 1950. [i](#), [4](#), [5](#), [6](#), [8](#), [11](#), [13](#), [17](#), [19](#), [20](#), [22](#), [24](#), [25](#), [27](#), [39](#), [40](#), [42](#), [46](#), [57](#), [63](#), [65](#), [81](#), [85](#), [103](#), [120](#), [121](#), [128](#), [133](#)
- J. Bröcker. Reliability, sufficiency, and the decomposition of proper scores. *Quarterly Journal of the Royal Meteorological Society*, 135:1512–1519, 2009. [14](#), [18](#), [21](#), [27](#), [33](#), [80](#), [81](#), [98](#)
- J. Bröcker. Estimating reliability and resolution of probability forecasts through decomposition of the empirical score. *Climate Dynamics*, 39(3-4):655–667, 2012. [136](#)
- J. Bröcker. Probability forecasts. In [Jolliffe and Stephenson \(2012\)](#), chapter 7, pages 119–139. [21](#), [57](#), [62](#), [156](#)
- J. Bröcker. Resolution and discrimination – two sides of the same coin. *Quarterly Journal of the Royal Meteorological Society*, 141:1277–1282, 2015. [21](#), [52](#)
- J. Bröcker and L.A. Smith. Scoring probabilistic forecasts: The importance of being proper. *Weather and Forecasting*, 22:382–388, 2007a. [30](#)
- J. Bröcker and L.A. Smith. Increasing the reliability of reliability diagrams. *Weather and Forecasting*, 22: 651–661, 2007b. [62](#), [63](#), [156](#)
- J. Bröcker, S. Siebert, and H. Kantz. Comments on “conditional exceedance probabilities”. *Monthly Weather Review*, 139:3322–3324, 2011. [48](#)
- I.D.J Bross. *Design for Decision*. The Macmillan Company, 1953. [9](#), [11](#)
- A. Browder. *Mathematical Analysis: An Introduction*. Springer, 1996. [95](#)
- T.A. Brown. *Probabilistic Forecasts and Reproducing Scoring Systems*. RAND Corporation, 1970. [3](#), [5](#), [6](#), [7](#)
- T.A. Brown. Admissible scoring systems for continuous distributions. Technical report, RAND Corporation, <http://www.rand.org/pubs/papers/2008/P5235.pdf>, 1974. [5](#), [43](#)
- G. Candille and O. Talagrand. Evaluation of probabilistic prediction systems for a scalar variable. *Quarterly Journal of the Royal Meteorological Society*, 131:2131–2150, 2005. [47](#), [48](#)
- A. Carvalho. Tailored proper scoring rules elicit decision weights. *Judgment and Decision Making*, 10(1): 86–96, 2015. [8](#)
- R. Carroll and S. Prickett, editors. *The Bible: Authorized King James Version*. Oxford University Press,

1997. [1](#)
- H.M. Christensen. Decomposition of a new proper score for verification of ensemble forecasts. *Monthly Weather Review*, 143(5):1517–1532, 2015. [18](#)
- W.E. Cooke. Forecasts and verifications in Western Australia. *Monthly Weather Review*, 34(1):23–24, 1906a. [2](#), [8](#), [16](#)
- W.E. Cooke. Weighting forecasts. *Monthly Weather Review*, 34(6):274–275, 1906b. [2](#)
- F.P.A. Coolen. On the use of imprecise probabilities in reliability. *Quality and Reliability Engineering International*, 20:193–202, 2004. [85](#)
- A.P. Dawid. Probability forecasting. In S. Kotz, N.L. Johnson, and C.B. Read, editors, *Encyclopedia of Statistical Sciences*, volume 7, pages 210–218. John Wiley & Sons, 1986. [19](#)
- R. de Eliá and R. Laprise. Diversity in interpretations of probability: Implications for weather forecasting. *Monthly Weather Review*, 133:1129–1143, 2005. [3](#)
- M.H. DeGroot and S.E. Fienberg. The comparison and evaluation of forecasters. *Journal of the Royal Statistical Society. Series D (The Statistician)*, 32:12–22, 1983. [12](#), [14](#), [38](#)
- R.V. Dexter. Confidence factors are fictional. *Weather*, 17(4):132–135, 1962. [3](#)
- M.M. Deza and E. Deza. *Encyclopedia of Distances*. Springer, third edition, 2014. [30](#)
- W. Ehm and E.Y. Ovcharov. Bias-corrected score decomposition for generalized quantiles. *Biometrika*, 104(2):473–480, 2017. [153](#), [154](#)
- E.S. Epstein. Quality control for probability forecasts. *Monthly Weather Review*, 94(8):487–494, 1966. [3](#), [5](#)
- E.S. Epstein. A scoring system for probability forecasts of ranked categories. *Journal of Applied Meteorology*, 8:985–987, 1969. [7](#), [42](#)
- E.S. Epstein and A.H. Murphy. A note on the attributes of probabilistic predictions and the probability score. *Journal of Applied Meteorology*, 4:297–299, 1965. [11](#), [18](#)
- C.A.T. Ferro. Measuring forecast performance in the presence of observation error. *Quarterly Journal of the Royal Meteorological Society*, 143(708):2665–2676, 2017. [16](#)
- C.A.T. Ferro and T.E. Fricker. A bias-corrected decomposition of the Brier score. *Quarterly Journal of the Royal Meteorological Society*, 138:1954–1960, 2012. [57](#), [60](#), [136](#)
- P. Friederichs and A. Hense. A probabilistic forecast approach for daily precipitation totals. *Weather and Forecasting*, 23:659–673, 2008. [53](#)
- R.M. Frongillo and I.A. Kash. General truthfulness characterizations via convex analysis. arXiv:1211.3043v3, 2014. [6](#), [84](#), [92](#), [106](#), [154](#)
- A. Gilio and G. Sanfilippo. Coherent conditional probabilities and proper scoring rules. In F. Coolen, G. de Cooman, T. Fetz, and M. Oberguggenberger, editors, *Proceedings of the Seventh International Symposium on Imprecise Probability: Theories and Applications*, pages 199–208, July 25–28 2011. [154](#)
- T. Gneiting. Making and evaluating point forecasts. *Journal of the American Statistical Association*, 106:746–762, 2011a. [5](#), [14](#), [19](#), [28](#), [53](#), [92](#), [110](#)
- T. Gneiting. Quantiles as optimal point forecasts. *International Journal of Forecasting*, 27:197–207, 2011b. [19](#)
- T. Gneiting and M. Katzfuss. Probabilistic forecasting. *Annual Review of Statistics and Its Application*, 2014(1):125–151, 2014. [2](#)
- T. Gneiting and A.E. Raftery. Strictly proper scoring rules, prediction and estimation. *Journal of the American Statistical Association*, 102:359–378, 2007. [6](#), [21](#), [42](#), [43](#), [49](#), [53](#), [54](#), [98](#), [103](#), [120](#), [153](#)
- T. Gneiting, A.E. Raftery, Westveld III A.H., and T. Goldman. Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. *Monthly Weather Review*, 133(5):1098–1118, 2005. [49](#)

References

- T. Gneiting, F. Balabdaoui, and A.E. Raftery. Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society B*, 69:243–268, 2007. [8](#), [11](#)
- I.J. Good. Rational decisions. *Journal of the Royal Statistical Society. Series B (Methodological)*, 14(1): 107–114, 1952. [4](#), [6](#), [7](#), [14](#), [49](#), [104](#), [120](#)
- D.M. Green and J.A. Swets. *Signal Detection Theory and Psychophysics*. John Wiley and Sons Inc., 1966. [156](#)
- I.I. Gringorten. The verification and scoring of weather forecasts. *Journal of the American Statistical Association*, 46(255):279–296, 1951. [4](#), [10](#)
- I.I. Gringorten. On the comparison of one or more sets of probability forecasts. *Journal of Meteorology*, 15: 283–287, 1958. [10](#)
- C. Hallenbeck. Forecasting precipitation in percentages of probability. *Monthly Weather Review*, pages 645–647, 1920. [3](#), [8](#)
- T.M. Hamill. Interpretation of rank histograms for verifying ensemble forecasts. *Monthly Weather Review*, 129:550–560, 2001. [48](#)
- T.M. Hamill and S.J. Colucci. Verification of Eta-RSM short-range ensemble forecasts. *Monthly Weather Review*, 125:1312–1327, 1997. [48](#)
- D.J. Hand. Measuring classifier performance: A coherent alternative to the area under the ROC curve. *Machine Learning*, 77:103–123, 2009. [156](#)
- D.J. Hand and C. Anagnostopoulos. When is the area under the receiver operating characteristic curve an appropriate measure of classifier performance? *Pattern Recognition Letters*, 34:492–495, 2013. [156](#)
- J.A. Hanley and B.J. McNeil. The meaning and use of the area under the receiver operating characteristic (roc) curve. *Radiology*, 143(1):29–36, 1982. [156](#)
- A.D. Hendrickson and R.J. Buehler. Proper scores for probability forecasters. *The Annals of Mathematical Statistics*, 42:1916–1921, 1971. [6](#)
- H. Hersbach. Decomposition of the continuous ranked probability score for ensemble prediction systems. *Weather and Forecasting*, 15:559–570, 2000. [48](#), [49](#)
- R. Ihaka and R. Gentleman. R: A language for data analysis and graphics. *Journal of Computational and Graphical Statistics*, 5:299–314, 1996. [www.r-project.org](#). [60](#), [77](#), [122](#), [140](#)
- I.T. Jolliffe and D.B. Stephenson, editors. *Forecast Verification: A Practitioner’s Guide in Atmospheric Science*. John Wiley & Sons, Ltd., second edition, 2012. [158](#), [162](#)
- V.R. Jose. A characterisation for the spherical scoring rule. *Theory and Decision*, 66:263–281, 2009. [7](#)
- N. Lambert and Y. Shoham. Eliciting truthful answers to multiple-choice questions. *EC ’09 Proceedings of the 10th ACM Conference on Electronic Commerce*, pages 109–118, 2009. [84](#), [92](#), [106](#), [154](#)
- N. Lambert, D.M. Pennock, and Y. Shoham. Eliciting properties of probability distributions. *EC ’08 Proceedings of the 9th ACM Conference on Electronic Commerce*, pages 129–138, 2008. [84](#), [92](#), [154](#)
- N.S. Lambert. Elicitation and evaluation of statistical forecasts. Preprint. Stanford University. ([web.stanford.edu/~nlambert/papers/elicitation.pdf](#)), 2013. [84](#), [92](#), [106](#), [154](#)
- W-C. Lee. Probabilistic analysis of global performances of diagnostic tests: Interpreting the Lorenz curve-based summary measures. *Statistics In Medicine*, 18:455–471, 1999. [156](#)
- S. Lemeshow and D.W. Hosmer. A review of goodness of fit statistics for use in the development of logistic regression models. *American Journal of Epidemiology*, 115(1):92–106, 1982. [156](#)
- S. Lemeshow, D. Teres, J.S. Avrunin, and H. Pastides. Predicting the outcome of intensive care unit patients. *Journal of the American Statistical Association*, 83(402):348–356, 1988. [156](#)
- S. Lerch, T.L. Thorarinsdottir, F. Ravazzolo, and T. Gneiting. Forecaster’s dilemma: Extreme events and forecast evaluation. *Statistical Science*, 32(1):106–127, 2017. [154](#)

References

- D.V. Lindley. Scoring rules and the inevitability of probability. *International Statistical Review*, 50(1):1–11, 1982. [16](#), [154](#)
- D.G. Luenberger. *Optimization by Vector Space Methods*. John Wiley & sons, Inc., 1969. [51](#)
- S.J. Mason, J.S. Galpin, L. Goddard, N.E. Graham, and B. Rajaratnam. Conditional exceedance probabilities. *Monthly Weather Review*, 135:363–372, 2007. [48](#)
- S.J. Mason, M.K. Tippett, A.P. Weigel, L. Goddard, and B. Rajaratnam. Reply to comments on “conditional exceedance probabilities”. *Monthly Weather Review*, 139:3325–3327, 2011. [48](#)
- J.E. Matheson and R.L. Winkler. Scoring rules for continuous probability distributions. *Management Science*, 22(10):1087–1096, 1976. [9](#), [14](#), [43](#)
- J. McCarthy. Measures of the value of information. *Proceedings of the National Academy of Sciences of the United States of America*, 42:654–655, 1956. [6](#), [21](#)
- D. McNicol. *A Primer of Signal Detection Theory*. George Allen & Unwin Ltd., 1972. [156](#)
- K. Mitchell and C.A.T. Ferro. Proper scoring rules for interval probabilistic forecasts. *Quarterly Journal of the Royal Meteorological Society*, 143(704):1597–1607, 2017. [82](#), [106](#), [109](#)
- A.M. Mood, F.A. Graybill, and D.C. Boes. *Introduction to the Theory of Statistics*. McGraw-Hill, international edition, 1974. [58](#)
- I. Mortimer. *The Time Traveller’s Guide to Medieval England*. Vintage Books, 2009. [1](#)
- A.H. Murphy. On the ranked probability score. *Journal of Applied Meteorology*, 8:988–989, 1969. [43](#)
- A.H. Murphy. The ranked probability score and the probability score: A comparison. *Monthly Weather Review*, 98(12):917–924, 1970. [7](#)
- A.H. Murphy. A note on the ranked probability score. *Journal of Applied Meteorology*, 10:155–156, 1971. [42](#)
- A.H. Murphy. Scalar and vector partitions of the ranked probability score. *Monthly Weather Review*, 100(10):701–708, 1972. [47](#), [48](#)
- A.H. Murphy. A new vector partition of the probability score. *Journal of Applied Meteorology*, 12:595–600, 1973. [12](#), [13](#)
- A.H. Murphy. A sample skill score for probability forecasts. *Monthly Weather Review*, 102:48–55, 1974. [5](#)
- A.H. Murphy. A new decomposition of the Brier score: Formulation and interpretation. *Monthly Weather Review*, 114:2671–2673, 1986. [13](#), [14](#), [19](#), [81](#)
- A.H. Murphy. What is a good forecast? an essay on the nature of goodness in weather forecasting. *Weather and Forecasting*, 8:281–293, 1993. [17](#), [18](#)
- A.H. Murphy. The early history of probability forecasts: Some extensions and clarifications. *Weather and Forecasting*, 13:5–15, 1998. [2](#), [3](#), [4](#)
- A.H. Murphy and E.S. Epstein. A note on probability forecasts and ‘hedging’. *Journal of Applied Meteorology*, 6:1002–1004, 1967a. [5](#), [6](#), [17](#), [20](#)
- A.H. Murphy and E.S. Epstein. Verifications of probabilistic predictions: A brief review?. *Journal of Applied Meteorology*, 6:748–755, 1967b. [9](#), [10](#), [11](#), [13](#), [41](#)
- A.H. Murphy and R.L. Winkler. Scoring rules in probability assessment and evaluation. *Acta Psychologica*, 34:273–286, 1970. [7](#), [8](#)
- A.H. Murphy and R.L. Winkler. A general framework for forecast verification. *Monthly Weather Review*, 115:1330–1338, 1987. [8](#), [12](#), [13](#), [14](#), [17](#), [18](#), [19](#), [21](#), [22](#), [29](#), [31](#), [34](#), [35](#), [81](#), [123](#), [153](#), [154](#)
- A.H. Murphy and R.L. Winkler. Diagnostic verification of probability forecasts. *International Journal of Forecasting*, 7:435–455, 1992. [2](#), [13](#)
- R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2017. URL <https://www.R-project.org/>. [60](#), [77](#), [122](#), [140](#)
- D.A. Redelmeier, D.A. Bloch, and D.H. Hickam. Assessing predictive accuracy: How to compare Brier

References

- scores. *Journal of Clinical Epidemiology*, 44(11):1141–1146, 1991. [156](#)
- K.F. Riley, M.P. Hobson, and S.J. Bence. *Mathematical Methods for Physics and Engineering*. Cambridge University Press, 1998. [20](#)
- T.B. Roby. Belief states: A preliminary empirical study. Technical documentary report no. ESD-TDR-64-238, Decision Sciences Laboratory, Electronic Systems Division, Air Force Systems Command, United States Air Force, 1964. [105](#), [120](#)
- M.S. Roulston and L.A. Smith. Evaluating probabilistic forecasts using information theory. *Monthly Weather Review*, 130:1653–1660, 2002. [14](#), [49](#)
- F. Sanders. On subjective probability forecasting. *Journal of Applied Meteorology*, 2(2):191–201, 1963. [3](#), [5](#), [11](#), [13](#)
- L.J. Savage. Elicitation of personal probabilities and expectations. *Journal of the American Statistical Association*, 66:783–801, 1971. [6](#), [8](#)
- M.J. Schervish. A general method for comparing probability assessors. *The Annals of Statistics*, 17(4):1856–1879, 1989. [6](#), [103](#)
- E.F. Schumacher. *Small is Beautiful: A Study of Economics as if People Mattered*. Abacus, 1974. [1](#)
- R. Selten. Axiomatic characterization of the quadratic scoring rule. *Experimental Economics*, 1:43–62, 1998. [7](#)
- O.B. Sheynin. On the history of the statistical method in meteorology. *Archive for History of Exact Sciences*, 31(1):53–95, 1984. [2](#)
- E.H. Shuford, Jr., A. Albert, and H.E. Massengill. Admissible probability measurement procedures. *Psychometrika*, 31(2):125–145, 1966. [5](#), [6](#), [7](#)
- S. Siegert. Simplifying and generalising Murphy’s Brier score decomposition. *Quarterly Journal of the Royal Meteorological Society*, 143(703):1178–1183, 2017. [14](#), [18](#), [33](#), [34](#), [69](#)
- D.J. Spiegelhalter. Probabilistic prediction in patient management and clinical trials. *Statistics in Medicine*, 5:421–433, 1986. [156](#)
- C-A.S. Staël von Holstein. A family of strictly proper scoring rules which are sensitive to distance. *Journal of Applied Meteorology*, 9(3):360–364, 1970. [7](#)
- N. Stallard. Simple tests for the external validation of mortality prediction scores. *Statistics in Medicine*, 28(3):377–388, 2009. [156](#)
- D.B. Stephenson, C.A.S. Coelho, and I.T. Jolliffe. Two extra components in the Brier score decomposition. *Weather and Forecasting*, 23:752–757, 2008. [63](#), [67](#), [69](#)
- O. Talagrand, R. Vautard, and B. Strauss. Evaluation of probabilistic prediction systems. In *Workshop on Predictability, 20-22 October 1997*, pages 1–26. ECMWF, 1997. <https://www.ecmwf.int/en/elibrary/12555-evaluation-probabilistic-prediction-systems>. [48](#)
- H. Theil. *Applied Economic Forecasting*. North-Holland Publishing Company, 1966. [13](#)
- J.C. Thompson and G.W. Brier. The economic utility of weather forecasts. *Monthly Weather Review*, 83(11):249–254, 1955. [10](#)
- J. Tödter and B. Ahrens. Generalization of the ignorance score: continuous ranked version and its decomposition. *Monthly Weather Review*, 140(6):2005–2017, 2012. [14](#), [49](#), [50](#)
- M.C.M. Troffaes, G. de Cooman, and D. Aeyels. Imprecise probabilities — discussion and open problems. In *Proceedings of the 2nd FTW PhD Symposium*, 2001. [157](#)
- W.N. Venables and B.D. Ripley. *Modern Applied Statistics with S*. Springer, New York, fourth edition, 2002. URL <http://www.stats.ox.ac.uk/pub/MASS4>. [60](#), [77](#)
- A.P. Weigel. Ensemble forecasts. In [Jolliffe and Stephenson \(2012\)](#), chapter 8, pages 141–166. [48](#)
- A.P. Weigel and N.E. Bowler. Comment on ‘can multi-model combination really enhance the prediction skill

References

- of probabilistic ensemble forecasts?'. *Quarterly Journal of the Royal Meteorological Society*, 135:535–539, 2009. [58](#)
- S.V. Weijts, R. van Nooijen, and N. van de Giesen. Kullback-leibler divergence as a forecast skill score with classic reliability-resolution-uncertainty decomposition. *Monthly Weather Review*, 138(9):3387–3399, 2010. [14](#), [49](#), [50](#)
- D.S. Wilks. *Statistical Methods in the Atmospheric Sciences*. Elsevier, second edition, 2006. [42](#), [48](#), [111](#), [155](#)
- R.L. Winkler. Scoring rules and the evaluation of probability assessors. *Journal of the American Statistical Association*, 64(327):1073–1078, 1969. [7](#), [8](#)
- R.L. Winkler. Scoring rules and the evaluation of probabilities. *Test*, 5(1):1–60, 1996. [13](#), [17](#), [21](#), [82](#), [123](#)
- R.L. Winkler and A.H. Murphy. ‘Good’ probability assessors. *Journal of Applied Meteorology*, 7:751–758, 1968. [5](#), [6](#), [10](#), [20](#), [21](#), [41](#), [49](#), [83](#), [85](#), [123](#)
- R.L. Winkler and A.H. Murphy. Nonlinear utility and the probability score. *Journal of Applied Meteorology*, 9:143–148, 1970. [8](#)
- J.F. Yates. External correspondence: Decompositions of the mean probability score. *Organizational Behavior and Human Performance*, 30:132–156, 1982. [9](#), [12](#), [13](#), [38](#)
- J.F. Yates and S.P. Curley. Conditional distribution analyses of probabilistic forecasts. *Journal of Forecasting*, 4:61–73, 1985. [12](#), [17](#)
- R.M.B. Young. Decomposition of the Brier score for weighted forecast-verification pairs. *Quarterly Journal of the Royal Meteorological Society*, 136(650):1364–1370, 2010. [14](#)