Karlsruhe Institute of Technology

Facets of Forecast Evaluation

Zur Erlangung des akademischen Grades eines

DOKTORS DER NATURWISSENSCHAFTEN

von der Fakultät für Mathematik des
Karlsruher Instituts für Technologie (KIT)
genehmigte

DISSERTATION

von

Dipl.-Math. Alexander Jordan
aus
Heidelberg

Tag der mündlichen Prüfung: 23. November 2016

1. Referent: Prof. Dr. Tilmann Gneiting
2. Referent: Prof. Dr. Norbert Henze

Abstract

Forecasts are fundamental to decision making. The paradigm shift from deterministic to probabilistic forecasts in recent decades facilitates accounting for the forecaster's uncertainty, and simultaneously poses new challenges in forecast evaluation.

The first part of this thesis is concerned with elementary measures of predictive performance, and studies how these can be combined into measures with relevance in real-world applications. We show that any scoring function that is consistent for a quantile or an expectile functional can be represented as a mixture of elementary or extremal scoring functions that form a linearly parameterized family. These elementary scores also relate to the more involved risk measure of expected shortfall. Furthermore, first steps are made towards mixture representations of interesting subclasses of proper scoring rules used to evaluate probabilistic forecasts of categorical events.

Mixture representations based on linearly parameterized families suggest graphical tools which we term Murphy diagrams after the late Allan H. Murphy. Forecast rankings may change with respect to the choice of performance measure, and Murphy diagrams allow for the graphical inspection of all elementary scoring functions simultaneously. They also confirm the presence or absence of dominance relations, and provide checks of whether a claim of a forecaster's superiority over another is in line with personal preferences.

We move on to statistical significance tests for the hypothesis of equal predictive performance. Again, we observe a distinct effect on the test performance with respect to the choice of scoring criterion. Under certain conditions, the sign test, which has commonly associated with poor discriminative ability of null hypothesis violations, proves to be an effective choice.

And lastly, we address the issue of calculating a particular proper scoring rule, the continuous ranked probability score. This scoring rule is of great interest due to its broad applicability, but closed-form expressions can be difficult to find. Often numerical approximation is necessary, or sampling methods need to be employed as a last resort.

Zusammenfassung

Vorhersagen spielen im Hinblick auf die Entscheidungsfindung eine grundlegende Rolle. Im Laufe der letzten Jahrzehnte hat ein Wandel von deterministischen hin zu probabilistischen Vorhersagen stattgefunden, und dies erlaubt, die Unsicherheit einer Vorhersage miteinzubeziehen, stellt aber gleichzeitig neue Herausforderungen an die Bewertung.

Der erste Teil dieser Dissertation beschäftigt sich mit elementaren Maßen für Vorhersagegüte und damit, wie sie zu Maßen mit praktischer Relevanz verbunden werden können. Wir zeigen, dass sich jede für ein Quantil- oder Expektilfunktional konsistente Bewertungsfunktion als Mischung von elementaren bzw. extremalen Bewertungsfunktionen, welche eine linear parametrisierte Familie bilden, darstellen lässt, und untersuchen ferner, welche Beziehung zu Bewertungsfunktionen für das Risikomaß *Expected Shortfall* besteht. Darüber hinaus werden erste Schritte unternommen, um Mischungsdarstellungen für interessante Unterklassen von Bewertungsregeln für probabilistische Vorhersagen kategorischer Ereignisse zu finden.

Mischungsdarstellungen, die auf linear parametrisierten Familien beruhen, legen ein grafisches Werkzeug nahe, das wir nach Allan H. Murphy als *Murphy Diagramm* bezeichnen. Die Rangordnung von Vorhersagen kann sich abhängig von der Wahl des Bewertungsmaßes ändern, und Murphy Diagramme erlauben die gleichzeitige, grafische Untersuchung aller elementaren Bewertungsfunktionen. Die An- oder Abwesenheit von Dominanzbeziehungen kann direkt abgelesen werden, und Murphy Diagramme stellen ein Mittel dar, um zu überprüfen, ob eine Schlussfolgerung zugunsten einer von zwei konkurrierenden Vorhersagen mit den eigenen Präferenzen übereinstimmt.

Von dort gehen wir über zu statistischen Signifikanztests für die Hypothese gleicher Vorhersagegüte. Wiederholt können wir beobachten, dass die Wahl des Bewertungskriteriums einen deutlichen Einfluss auf das Verhalten der Tests hat. Unter bestimmten Bedingungen zeigt sich der Vorzeichentest als besonders effektiv, obwohl er üblicherweise mit schlechter Erkennungsrate einer Abweichung von der Nullhypothese assoziiert wird.

Zuletzt beschäftigten wir uns mit dem Problem, eine bestimmte Bewertungsregel, den *continuous ranked probability score*, zu berechnen. Diese Bewertungsregel ist von besonderem Interesse aufgrund ihrer breiten Anwendbarkeit. Häufig sind analytische Formen jedoch schwer zu finden, sodass es nötig wird numerisch zu approximieren, oder gar Methoden zu benutzen, die auf dem Ziehen einer Stichprobe beruhen.

Acknowledgements

I express my thankfulness to Tilmann Gneiting for his support, guidance, collaboration, and the opportunity to work in such a stimulating environment. Furthermore, Norbert Henze has my gratitude for agreeing to act as a reviewer of this thesis containing my accumulated work over the last few years.

This work has been funded by the European Union Seventh Framework Programme under grant agreement 290976. I also thank the Klaus Tschira Foundation for infrastructural support at the Heidelberg Institute for Theoretical Studies (HITS) and Heidelberg University for infrastructural support at its Institute of Applied Mathematics. I acknowledge the opportunities provided by the Deutsche Forschungsgemeinschaft through Research Training Group (RTG) 1953 ”Statistical Modeling of Complex Systems and Processes”.

In the course of my doctoral studies, I have greatly benefited from the experience and expertise of many colleagues through discussions at conferences, workshops, invited lectures and during the peer review process. Specifically, I would like to thank Werner Ehm, Fernando Fasciati, Johanna Fasciati-Ziegel, Fabian Krüger, and Sebastian Lerch for the fruitful scientific collaboration, Donald Richards for generously sharing his knowledge of integrals and special mathematical functions, and Kira Feldmann, Stephan Hemri, Evgeni Ovcharov, Roman Schefzik, Patrick Schmidt and Peter Vogel for countless stimulating discussions. I also appreciate the pleasant experience as an associate member of Research Training Group 1953 and the welcoming environment during my employment at the Heidelberg Institute of Theoretical Studies, the Institute of Applied Mathematics at Heidelberg University, and the Institute of Stochastics at Karlsruhe Institute of Technology.

Finally, I am grateful and feel very fortunate in light of the great support of my family and friends.

Declaration

This thesis contains ideas and excerpts from the following previous publications or ongoing collaborations with significant contributions by myself.

Ehm, W., Gneiting, T., Jordan, A. and Krüger, F. (2016). Of Quantiles and Expectiles: Consistent Scoring Functions, Choquet Representations, and Forecast Rankings. *Journal of the Royal Statistical Society: Series B*, 78, 505–562.

Fasciati, F., Jordan, A., Krüger, F., and Ziegel, J. F. (2016). Murphy Diagrams for Evaluating Forecasts of Value-at-Risk and Expected Shortfall. *Working paper.*

Jordan, A. and Krüger, F. (2016). *murphydiagram: Murphy Diagrams for Forecast Comparisons.* R package version 0.11, `https://cran.r-project.org/package=murphydiagram`.

Jordan, A., Krüger, F. and Lerch, S. (2016). *scoringRules: Scoring Rules for Parametric and Simulated Distribution Forecasts.* R package version 0.9.1, `https://cran.r-project.org/package=scoringRules`.

Ehm et al. (2016) and the software package by Jordan and Krüger (2016) lend content to Sections 2.1, 3.1.1, 3.2.2, 3.3, 4.1, 4.2, 4.3, 4.3.1, 4.4, 4.4.1, 4.4.2, 4.4.3, 4.5 and 7. **Note:** Any material in these sections may be subject to copyright owned by the publisher *John Wiley and Sons.*

Section 3.1.2 is based on Fasciati et al. (2016). Jordan et al. (2016) is a software package for the statistical programming language R (R Core Team 2016) lending content to Sections 6.1, 6.1.1, 6.1.2, 6.2, and 6.3.

Contents

1 Introduction

The conquest to make predictions about the inherently uncertain future has been a driving factor in many scientific endeavours. Today, we have access to a myriad of physical models describing various processes in the real world to astonishing degrees of accuracy. However, sources of uncertainty persist for any moderately complex system, e.g. missing details in the model or lacking ability to collect necessary data. As a result, forecasts can be issued in two main ways: point forecasts with only limited information about the corresponding uncertainty, and probabilistic forecasts which provide full information in the form of probability distributions.

Of course, forecasts by themselves are never the ultimate goal. A purely statistical approach can, on some occasion, produce models and forecasts which may allow some inference of underlying structures. One may argue that statistics – fitting functions to data, making and evaluating forecasts – plays an integral part in the development of physical models. However, the main purpose of forecasts is their fundamental contribution to decision making. Believing in the persistence of the status quo may often result in reasonable actions, but correctly guessing the direction of developments should allow for improvement. This is true in weather prediction, but also in economic forecasting, where the market-driving processes are inextricably linked to previously made predictions. In either case, expert opinions and assessments are in perpetual demand and the question of trustworthiness, or genuine ability, arises naturally. So, how do we correctly measure predictive performance? Are the criteria that determine a good forecast universal or subjective, and how do they incorporate into a performance measure? Lastly, if I base my choice of expert judgment on the only available data, their past performance, how reliable are my conclusions? We will answer these questions to some extent in the course of this thesis, bearing in mind that the complexity of these questions may never lead to complete answers.

Chapter 2 introduces the probabilistic framework and mathematical objects, i.e. performance measures, that have become state of the art in forecast evaluation. By now, we have an ample range of performance measures for the above mentioned types of forecasts. Point forecasts should be evaluated using consistent scoring functions (Gneiting 2011), while probabilistic forecasts call for the use of proper scoring rules (Gneiting and Raftery 2007).

A fundamental problem is the multiplicity of valid performance measures when evaluating the quality of a forecast. Chapter 3 is concerned with the provision of structure in the form of mixture representations. Typically, proper scoring rules, and as a special case consistent scoring functions, are compositions of more basic, elementary proper scoring rules. If we can identify these elementary scores, learn their properties, and find interpretations in terms of decision making, then we may be able to increase our understanding of the more complex measures and provide guidance in choosing the appropriate one for a given set of preferences. While results from convex analysis by Johansen (1974) and Bronshtein (1978) suggest that it may be futile to attempt a full classification of extremal scoring rules, the restriction to certain subclasses, most prominently classes of consistent scoring functions, seems a promising undertaking.

Chapter 4 builds on these results by focusing on classes of elementary scores that can be linearly parameterized. This property allows simultaneous comparison of all elementary scores in graphical representations which we call Murphy diagrams, after the meteorologist Allan H. Murphy. These diagrams facilitate considerations regarding the information sets and dominance relations of competing forecasters. A related question is the stability of forecast rankings with respect to the choice of performance measure.

Chapter 5 is concerned with a different type of a forecast ranking's stability. Statistical significance tests can be employed to determine if the accumulated data exhibit sufficient indication of one forecaster's superiority over another with respect to sample size. While tests of this type have been developed a long time ago, the choice of performance measure has an impact on the score distribution, which in turn can sway the decision for or against the use of a certain test.

Chapter 6 deals with the problem of calculating a particular scoring rule, the continuous ranked probability score. It is one of the most popular scoring rules in practice and allows for the evaluation of predictive cumulative distribution functions. This implies broad applicability across multiple disciplines where predictive distributions can be discrete, continuous, or mixtures thereof. As a special case, it can be used to evaluate point forecasts where it becomes equivalent to the absolute error. In view of a mixture representation of the continuous ranked probability score in terms of consistent scoring functions for quantiles, it is our hope that considerations regarding the numerical approximation are transferable to more general classes of mixtures.

We conclude with a recapitulation of the most important results and an outline of possible avenues for further research.

2 Preliminaries on Forecasting

"What is a forecast? And when is it good?"

The importance of forecasting in many areas of decision making warrants careful assessment of the predictive performance. To this end, concepts from probability theory, decision theory, and convex analysis have been combined into a theoretical framework.

It is important that forecasters are encouraged to make careful and honest predictions, and this needs to be reflected by the performance measure. Furthermore, we need to have the capability of evaluating subjective beliefs of individual forecasters, i.e. probability distributions, against a realizing observation. In this context, point forecasts are special cases of probabilistic forecasts where information about the uncertainty is discarded for ease of communication.

In a nutshell, we consider proper scoring rules that evaluate predictive distributions, and consistent scoring functions that assess point forecasts, i.e. extractions from probabilistic forecasts, and both types of performance measures need to encourage honest reporting under the assumption of expectation minimization. This requires a general probabilistic framework where forecasts and observations arise as elementary events in an underlying probability space.

2.1 Prediction spaces

A *prediction space* is a probability space tailored to the study of forecasting problems. Following the seminal work of Murphy and Winkler (1987), the prediction space setting of Gneiting and Ranjan (2013) considers the joint distribution of forecasts and observations. We focus on real-valued point forecasts, X, or CDF-valued probabilistic forecasts, F, for a real-valued outcome, Y. For point forecasts, the elements of the respective sample space Ω can be identified with tuples of the form

$$(X_1, \ldots, X_k, Y)\,, \tag{2.1}$$

where the point forecasts $X_1, \ldots, X_k$ utilize information sets $\mathcal{A}_1, \ldots, \mathcal{A}_k \subseteq \mathcal{A}$, respectively, with $\mathcal{A}$ being a σ-field on the sample space Ω. In measure theoretic language, the information sets correspond to sub-σ-fields, and X_j is a real-valued random quantity measurable with respect to $\mathcal{A}_j$. The joint distribution of forecasts and observation is encoded by a probability measure $\mathbb{Q}$ on $(\Omega, \mathcal{A})$. In a common generalization, we allow the forecasts in (2.1) to be CDF-valued random quantities $F_1, \ldots, F_k$. An extended, more realistic notion of prediction space that

Forecaster	σ-field	Predictive Distribution	Mean	α-Quantile
Climatological	$\sigma(\emptyset)$	$\mathcal{N}(0,2)$	0	$\sqrt{2}z_\alpha$
Perfect	$\sigma(\mu)$	$\mathcal{N}(\mu,1)$	μ	$\mu + z_\alpha$
Sign-reversed	$\sigma(\mu)$	$\mathcal{N}(-\mu,1)$	$-\mu$	$-\mu + z_\alpha$
Unfocused	$\sigma(\mu,\tau)$	$\frac{1}{2}(\mathcal{N}(\mu,1) + \mathcal{N}(\mu+\tau,1))$	$\mu + \frac{\tau}{2}$	$\mu + z_{\alpha,\tau}$

Table 2.1: An example of a prediction space with four competing forecasters, where those labeled "Climatological" and "Perfect" issue ideal predictions with respect to their σ-field. The outcome is generated as $Y \mid \mu \sim \mathcal{N}(\mu,1)$, where $\mu \sim \mathcal{N}(0,1)$. The real-valued random variable τ is independent of μ and Y. For $\alpha \in (0,1)$ and $\tau \in \mathbb{R}$, we let $z_\alpha = \Phi^{-1}(\alpha)$, $\Phi_\tau(x) = (\Phi(x) + \Phi(x-\tau))/2$, and $z_{\alpha,\tau} = \Phi_\tau^{-1}(\alpha)$, where Φ denotes the CDF of the standard normal distribution.

allows for serial dependence between forecast-observation tuples has recently been introduced by Strähl and Ziegel (2015).

A forecaster utilizes the information available to them optimally if the forecast matches the conditional distribution of the outcome given the information set. For point forecasts, this means matching a real-valued summary of the conditional distribution, e.g. the expectation, and we refer to the mapping of a distribution to its summary as functional.

Definition 2.1 (ideal forecasts). *Let $G_j = \mathcal{L}(Y|\mathcal{A}_j)$ be the conditional distribution of the outcome Y given forecaster j's sigma field $\mathcal{A}_j$, and let $G_j \mapsto \mathrm{T}(G_j) \in \mathbb{R}$ be a well-defined, single-valued functional.*

(a) A probabilistic forecast F_j is ideal *relative to $\mathcal{A}_j$ if $F_j = G_j$ almost surely.*

(b) A point forecast X_j is ideal for T *relative to $\mathcal{A}_j$ if $X_j = \mathrm{T}(G_j)$ almost surely.*

Any predictive distribution F can be reduced to a point forecast by extracting the sought functional, $\mathrm{T}(F)$. Generalizations to set-valued functionals, or multivariate functionals with appropriate adjustments to the prediction space, are straight-forward.

To give an example, Table 2.1 revisits a scenario studied by Gneiting et al. (2007) and Gneiting and Ranjan (2013). Here, the outcome is generated as $Y \mid \mu \sim \mathcal{N}(\mu,1)$ where $\mu \sim \mathcal{N}(0,1)$. Issuing the unconditional distribution of the outcome Y as prediction, the climatological forecaster uses the uninformative sigma field $\{\emptyset, \Omega\}$ ideally. The perfect forecaster is ideal relative to the sigma field generated by the random variable μ, and despite equivalent knowledge, the sign-reversed prediction fails to be ideal. Lastly, even though the unfocused forecaster makes use of his knowledge about μ, he also incorporates spurious information from an independent variable τ. Unless τ equals zero almost surely, this leads to non-ideal forecasts. Concerning functionals and corresponding point forecasts, we focus on quantiles and the mean or expectation functional. The respective point forecasts for the climatological, perfect, sign-reversed, and unfocused forecaster are shown in Table 2.1.

2.2 Proper scoring rules

Probabilistic forecasts can be evaluated using proper scoring rules. They assign a numerical score based on a forecast-observation pair, where the expected score measures the predictive distribution's divergence from the true distribution. Issuing the true data-generating distribution as probabilistic forecast minimizes the expected score (Gneiting and Raftery 2007).

For the most common scenario[1], a scoring rule is defined as a function $\mathrm{S} : \mathcal{P} \times \mathbb{R} \to \bar{\mathbb{R}} = \mathbb{R} \cup \{\infty\}$, where $\mathcal{P}$ is a class of probability measures on $\mathbb{R}$. As a minimal regularity condition, we require

$$\mathbb{E}_G \mathrm{S}(P, Y) - \mathbb{E}_G \mathrm{S}(G, Y) \in \bar{\mathbb{R}} \tag{2.2}$$

for all $P, G \in \mathcal{P}$. The difference in (2.2) is central to the concept of propriety and gives a first insight to the problem of choosing a scoring rule in practice.

Definition 2.2 (proper scoring rules). *A scoring rule* $\mathrm{S} : \mathcal{P} \times \mathbb{R} \to \bar{\mathbb{R}}$ *is* proper *relative to* $\mathcal{P}$ *if*

$$\mathbb{E}_G \mathrm{S}(P, Y) \geq \mathbb{E}_G \mathrm{S}(G, Y) \tag{2.3}$$

for all $P, G \in \mathcal{P}$. *It is* strictly proper *if equality in* (2.3) *implies* $P = G$.

Proper scoring rules encourage truth-telling. If a forecaster genuinely believes that the observation follows the distribution G, then reporting G as his predictive distribution minimizes his expected score. Propriety also directly relates to the concept of an ideal forecast. If G is the conditional distribution of Y given some information set $\mathcal{A}$, then the best $\mathcal{A}$-measurable prediction a forecaster can make is G. Hence, a forecaster should report his genuine belief and take care to use the available information optimally.

The minimized expected score $\mathbb{E}_G \mathrm{S}(G, Y)$ is a functional that maps $\mathcal{P}$ to $\mathbb{R}$. We call this functional the *entropy function* E, and following Theorem 2.2 in Gneiting and Raftery (2007) we have

$$\mathrm{S}(G, y) = E(G) - \int_{\mathbb{R}} E^*(G, x) \, \mathrm{d}G(x) + E^*(G, y), \tag{2.4}$$

where $E^*(G, \cdot) : \mathbb{R} \mapsto \bar{\mathbb{R}}$, called a supertangent in $G \in \mathcal{P}$, satisfies

$$E(P) - E(G) \leq \mathbb{E}_P E^*(G, Y) - \mathbb{E}_G E^*(G, Y) \in \bar{\mathbb{R}}, \quad \text{for all } P \in \mathcal{P}.$$

[1] This definition can be generalized to work on any measurable space, e.g. categorical, multivariate, or spherical sample spaces. Here, we focus on $(\mathbb{R}, \mathcal{B})$ and a class $\mathcal{P}$ of corresponding probability measures.

This theorem connects the class of proper scoring rules to the class of concave, real-valued functions on $\mathcal{P}$, thus giving an idea of the wealth of available proper scoring rules.

The logarithmic score (LS) is arguably the most popular proper scoring rule, due to its connections to maximum-likelihood and information theory, and its ease of implementation. Another reason is that its entropy function is finite for all probability measures on $\mathcal{B}(\mathbb{R})$ with a Lebesgue density. The logarithmic score is defined as

$$\mathrm{LS}(P, y) = -\log p(y),$$

where p denotes the density function of P and y is the observation. An appealing property of the logarithmic score is *locality*, as motivated by Bernardo (1979, p. 689),

> *"[...] when assessing the worthiness of a scientist's final conclusions, only the probability he attaches to small interval containing the true value should be taken into account."*

Intuitively, locality requires the existence of a probability density function which can be a limitation for the logarithmic score in practice.

Another frequently used proper scoring rule is the continuous ranked probability score (CRPS) introduced by Matheson and Winkler (1976). It can be given in terms of the cumulative distribution function, in terms of the quantile function, or as a difference of expectations. Hence, it allows for the evaluation of distributions without a density function. This comes at the cost of difficulties in the implementation (see Chapter 6) and the requirement of a finite first moment. The three equivalent representations (Gneiting and Raftery 2007; Matheson and Winkler 1976; Laio and Tamea 2007) are as follows,

$$\begin{aligned}
\mathrm{CRPS}(P, y) &= \mathbb{E}_P|Y - y| - \frac{1}{2}\mathbb{E}_P|Y - Y'| && (2.5)\\
&= \int_{-\infty}^{\infty} (F(x) - \mathbb{1}(y \le x))^2 dx && (2.6)\\
&= 2\int_0^1 (\mathbb{1}\{y < Q(\alpha)\} - \alpha)(Q(\alpha) - y) d\alpha, && (2.7)
\end{aligned}$$

where F denotes the cumulative distribution function and Q denotes the quantile function of P. The symbols Y and Y' denote independent random variables with distribution P, and y is the observation.

2.3 Consistent scoring functions

In point prediction problems, simply requesting a value representing the forecaster's best guess is an ill-posed approach. Without additional information this leads to some forecasters reporting the subjectively most likely outcome while others report their expectation, and in the worst case a forecaster's unique preferences may lead to predictions with little value to anyone. This obscure amalgamation of forecasts makes it difficult to draw meaningful conclusions.

Proper guidance facilitates interpretation when we know that reported values are subjective expectations, for example. There are two options to ensure predictions are consistent with the desired objective. We can communicate which loss function will be used to evaluate forecast performance and thus encourage reporting of the Bayes act, or we can promise to use a loss function where the Bayes act corresponds to a specified functional. These loss functions are called *consistent scoring functions.*

The two most commonly used measures of performance for point forecasts are the *absolute error* (AE)

$$\mathrm{AE}(x, y) = |x - y|, \tag{2.8}$$

and the *squared error*(SE)

$$\mathrm{SE}(x, y) = (x - y)^2, \tag{2.9}$$

where the optimal point forecast,

$$\hat{x} = \arg\min_x \mathbb{E}_F \mathrm{S}(x, Y), \quad \mathrm{S} \in \{\mathrm{AE}, \mathrm{SE}\}, \tag{2.10}$$

that minimizes the expected loss, corresponds to the predictive distribution's median and mean, respectively. This property makes the absolute error a consistent scoring function for the median functional, and the squared error a consistent scoring function for the mean or expectation functional.

A typical regularity assumption, mostly motivated by ease of interpretation, is that consistent scoring functions assign a non-negative score with a zero value attained for equality in the forecast-observation pair. In the following definition of consistent scoring functions, we consider the possibility that functionals are set-valued or multivariate.

Definition 2.3 (consistent scoring function). *A scoring function* $\mathrm{S}^{\mathrm{T}} : \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}_0^+$ *is* consistent *for the functional* $\mathrm{T} : \mathcal{P} \to \mathbb{R}^n$ *if*

$$\mathbb{E}_G \mathrm{S}^{\mathrm{T}}(x, Y) \geq \mathbb{E}_G \mathrm{S}^{\mathrm{T}}(t, Y) \tag{2.11}$$

for all $G \in \mathcal{P}$*, all* $t \in \mathrm{T}(G)$*, and all point forecasts* $x \in \mathbb{R}^n$*. It is* strictly consistent *if equality in* (2.11) *implies* $x \in \mathrm{T}(G)$*.*

Given a functional T and a corresponding scoring function S^{T}, we can construct a proper scoring rule S using the representation

$$\mathrm{S}(F, y) = \mathrm{S}^{\mathrm{T}}(t_F, y), \tag{2.12}$$

where $t_F \in \mathrm{T}(F)$ is a point forecast, extracted from the probabilistic forecast F using the functional T. However, not every proper scoring rule admits a representation (2.12), and not every functional admits a corresponding consistent scoring function. We call a functional T *elicitable* when it does admit a strictly consistent scoring function.

Example 2.1. The Brier score,

$$\mathrm{BS}_\theta(F, y) = (F(\theta) - \mathbb{1}(y \leq \theta))^2 \tag{2.13}$$

evaluates exceedance probabilities with respect to an event threshold θ. It can be written in terms of the functional $\mathrm{T} : F \mapsto F(\theta)$ and the following version of the squared error,

$$\mathrm{SE}_\theta(x, y) = \mathrm{SE}(x, \mathbb{1}(y \leq \theta)).$$

In Chapter 3, we examine the scoring functions that are consistent for quantile and expectile functionals, and we also investigate those for the non-elicitable functional *expected shortfall*. Remarkably, the bivariate functional of value-at-risk, the finance literature's name for the left-most quantile, and expected shortfall is elicitable (Fissler and Ziegel 2016; Fissler et al. 2016).

3 Mixture Representations

"What are the elementary measures of predictive performance?"

When evaluating probabilistic forecasts one chooses a proper scoring rule, thereby incentivizing honest reporting of the forecaster's subjective belief. However, the class of proper scoring rules is uncountably large, and the choice affects which forecasters are credited with superior ability relative to others. The current practice involves using multiple proper scoring rules for evaluation, each of which is known to put an emphasis on a certain characteristic, e.g. one that focuses on the location and one that evaluates spread or tail-behavior.

Ideally, we would find a class $\mathcal{S}_\Theta$ of elementary or extremal proper scoring rules S_θ, parameterized by $\theta \in \Theta$, which fulfills two properties. Firstly, any proper scoring rule S should admit a mixture representation over $\mathcal{S}_\Theta$,

$$\mathrm{S}(F, y) = a(y) + \int_\Theta \mathrm{S}_\theta(F, y)\,\mathrm{d}H(\theta), \tag{3.1}$$

for some nonnegative measure H on Θ, and some function $a : \mathbb{R} \to \mathbb{R}$. And secondly, the class $\mathcal{S}_\Theta$ should only contain elementary scoring rules S_θ in the sense that the mixture representation (3.1) implies $H = \lambda\delta_\theta$, where $\lambda > 0$ and δ_θ denotes the Dirac measure in some uniquely determined $\theta \in \Theta$.

First steps have been made to provide such a structure, mostly for proper scoring rules used in the evaluation of forecasts for binary events (Schervish 1989). Recently, Ehm et al. (2016) have provided representations for scoring functions which are consistent for the quantile and expectile functionals, which nests the earlier discovery.

However, an extension from probability forecasts of binary to ternary or general discrete variables does not appear to be feasible, due to results by Johansen (1974) and Bronshtein (1978) in convex analysis. In a nutshell, Savage (1971) showed that in the case of k categories, the proper scoring rules for probability forecasts essentially are parameterized by the convex functions on the unit simplex in $\mathbb{R}^k$, a $(k-1)$-dimensional subspace. Johansen (1974) and Bronshtein (1978) proved that in the class of convex functions defined on a bounded domain in $\mathbb{R}^k$, $k \geq 2$, the extremal functions lie dense, i.e. any non-extremal entropy function can be approximated to arbitrary accuracy by an extremal member in the class of entropy functions.

This immediately limits the results we can expect. Our main goal is find meaningful subclasses of proper scoring rules that allow representations of the form in (3.1). Prime candidates in terms of interpretability are classes of consistent scoring functions due to their connection with certain functionals.

3.1 Consistent scoring functions

3.1.1 Quantiles and expectiles

The functionals

Let $\mathcal{F}_0$ denote the class of the probability measures on the Borel-Lebesgue sets of the real line, $\mathbb{R}$. For simplicity, we do not distinguish between a measure $F \in \mathcal{F}_0$ and the associated cumulative distribution function (CDF). We follow standard conventions and assume CDFs to be right-continuous.

In the case of quantiles, the functional might be set-valued. Specifically, the *quantile* functional at level $\alpha \in (0,1)$ maps a probability measure $F \in \mathcal{F}_0$ to the closed interval $[q^-_{\alpha,F}, q^+_{\alpha,F}]$, with lower limit $q^-_{\alpha,F} = \sup\{s : F(s) < \alpha\}$ and upper limit $q^+_{\alpha,F} = \sup\{s : F(s) \le \alpha\}$. The two limits differ only when the level set $F^{-1}(\alpha)$ contains more than one point, so typically the functional is single-valued. Any number between $q^-_{\alpha,F}$ and $q^+_{\alpha,F}$ represents an α-quantile and will be denoted $q_{\alpha,F}$.

The expectation functional is well-defined with respect to the class $\mathcal{F}_1$ of the probability measures with finite first moment. More generally, the *expectile* at level $\alpha \in (0,1)$ of a probability measure $F \in \mathcal{F}_1$ is the unique solution t to the equation

$$(1-\alpha)\int_{-\infty}^{t} (t-y)\,\mathrm{d}F(y) = \alpha \int_{t}^{\infty} (y-t)\,\mathrm{d}F(y),$$

where $\alpha = 1/2$ corresponds to the mean functional (Newey and Powell 1987).

Scoring functions

The classes of the consistent scoring functions for quantiles and expectiles have been described by Savage (1971), Thomson (1979), and Gneiting (2011), and we review the respective characterizations in the setting of the latter paper, where further detail is available.

Up to mild regularity conditions, a scoring function S is consistent for the quantile functional at level $\alpha \in (0,1)$ relative to the class $\mathcal{F}_0$ if and only if it is of the form

$$\mathrm{S}(x,y) = (\mathbb{1}(y < x) - \alpha)\,(g(x) - g(y)), \tag{3.2}$$

where g is non-decreasing. The most prominent example arises when $g(t) = t$, which yields the asymmetric piecewise linear scoring function,

$$\mathrm{S}(x,y) = \begin{cases} (1-\alpha)\,(x-y), & y < x, \\ \alpha\,(y-x), & y \ge x, \end{cases} \tag{3.3}$$

that lies at the heart of quantile regression (Koenker and Bassett 1978; Koenker 2005). Similarly, a scoring function is consistent for the expectile at level $\alpha \in (0,1)$ relative to the class $\mathcal{F}_1$ if and only if it is of the form

$$\mathrm{S}(x,y) = |\mathbb{1}(y < x) - \alpha|\,(\phi(y) - \phi(x) - \phi'(x)(y-x)), \tag{3.4}$$

where ϕ is convex with subgradient ϕ'. The key example arises when $\phi(t) = t^2$, where

$$\mathrm{S}(x,y) = \begin{cases} (1-\alpha)\,(x-y)^2, & y < x. \\ \alpha\,(x-y)^2, & y \geq x. \end{cases} \tag{3.5}$$

This is the loss function used for estimation in expectile regression (Newey and Powell 1987; Efron 1991), including the ubiquitous case $\alpha = 1/2$ of ordinary least squares regression.

In view of the representations (3.2) and (3.4), the scoring functions that are consistent for quantiles and expectiles are parameterized by the non-decreasing functions g, and the convex functions ϕ with subgradient ϕ', respectively. In general, neither g nor ϕ and ϕ' are uniquely determined. We therefore select special versions of these functions. Let $\mathcal{I}$ denote the class of all left-continuous non-decreasing real functions g, and let $\mathcal{C}$ denote the class of all convex real functions ϕ with subgradient $\phi' \in \mathcal{I}$. This last condition is satisfied when ϕ' is chosen to be the left-hand derivative of ϕ, which exists everywhere and is left-continuous by construction.

Mixture representations

In what follows, we use the symbol $\mathcal{S}_\alpha^{\mathrm{Q}}$ to denote the class of the scoring functions S of the form (3.2) where $g \in \mathcal{I}$. Similarly, we write $\mathcal{S}_\alpha^{\mathrm{E}}$ for the class of the scoring functions S of the form (3.4) where $\phi \in \mathcal{C}$. For all practical purposes, the families $\mathcal{S}_\alpha^{\mathrm{Q}}$ and $\mathcal{S}_\alpha^{\mathrm{E}}$ can be identified with the classes of the scoring functions that are consistent for quantiles and expectiles, respectively. These classes appear to be rather large. However, in either case the apparent multitude can be reduced to a one-dimensional family of elementary scoring functions, in the sense that every consistent scoring function admits a representation as a mixture of elementary elements.

Theorem 3.1a (quantiles). *Any member of the class $\mathcal{S}_\alpha^{\mathrm{Q}}$ admits a representation of the form*

$$\mathrm{S}(x,y) = \int_{-\infty}^{\infty} \mathrm{S}_{\alpha,\theta}^{\mathrm{Q}}(x,y)\,\mathrm{d}H(\theta) \qquad (x, y \in \mathbb{R}), \tag{3.6}$$

where H is a nonnegative measure and

$$\begin{aligned} \mathrm{S}_{\alpha,\theta}^{\mathrm{Q}}(x,y) &= (\mathbb{1}(y < x) - \alpha)\ (\mathbb{1}(\theta < x) - \mathbb{1}(\theta < y)) \\ &= \begin{cases} 1-\alpha, & y \leq \theta < x, \\ \alpha, & x \leq \theta < y, \\ 0, & \text{otherwise.} \end{cases} \end{aligned} \tag{3.7}$$

The mixing measure H is unique and satisfies $\mathrm{d}H(\theta) = \mathrm{d}g(\theta)$ for $\theta \in \mathbb{R}$, where g is the nondecreasing function in the representation (3.2)*. Furthermore, we have $H(x) - H(y) = \mathrm{S}(x,y)/(1-\alpha)$ for $x > y$.*

Theorem 3.1b (expectiles). *Any member of the class $\mathcal{S}_\alpha^{\mathrm{E}}$ admits a representation of the form*

$$\mathrm{S}(x,y) = \int_{-\infty}^{\infty} \mathrm{S}_{\alpha,\theta}^{\mathrm{E}}(x,y)\,\mathrm{d}H(\theta) \qquad (x,y \in \mathbb{R}), \tag{3.8}$$

where H is a nonnegative measure and

$$\begin{aligned} \mathrm{S}_{\alpha,\theta}^{\mathrm{E}}(x,y) &= |\mathbb{1}(y<x)-\alpha|\,((y-\theta)_+ - (x-\theta)_+ - (y-x)\,\mathbb{1}(\theta<x)) \\ &= \begin{cases} (1-\alpha)\,|y-\theta|, & y \le \theta < x, \\ \alpha\,|y-\theta|, & x \le \theta < y, \\ 0, & \text{otherwise.} \end{cases} \end{aligned} \tag{3.9}$$

The mixing measure H is unique and satisfies $\mathrm{d}H(\theta) = \mathrm{d}\phi'(\theta)$ for $\theta \in \mathbb{R}$, where ϕ' is the left-hand derivative of the convex function ϕ in the representation (3.4). Furthermore, we have $H(x) - H(y) = -\partial_2 \mathrm{S}(x,y)/(1-\alpha)$ for $x > y$, where ∂_2 denotes the left-hand derivative with respect to the second argument.

Proof. The specific structure of the scoring functions in (3.2) and (3.4) permits us to focus on the case $\alpha = 1/2$, with the general case $\alpha \in (0,1)$ then being immediate.

For Theorem 3.1a, the mixture representation (3.6), the fact that $\mathrm{d}H(\theta) = \mathrm{d}g(\theta)$ for $\theta \in \mathbb{R}$, and the relationship $H(x) - H(y) = \mathrm{S}(x,y)/(1-\alpha)$ for $x > y$, are straightforward consequences of the fact that for every $g \in \mathcal{I}$ and $x, y \in \mathbb{R}$,

$$g(x) - g(y) = \int_{-\infty}^{\infty} \{\mathbb{1}(\theta < x) - \mathbb{1}(\theta < y)\}\,\mathrm{d}g(\theta).$$

As the increments of H are determined by S, the mixing measure is unique.

For Theorem 3.1b, we associate with any function $\phi \in \mathcal{C}$ the Bregman type function of two variables

$$\Phi(x,y) = \phi(y) - \phi(x) - \phi'(x)(y-x) \qquad (x,y \in \mathbb{R}). \tag{3.10}$$

Then the mixture representation (3.8), the fact that $\mathrm{d}H(\theta) = \mathrm{d}\phi'(\theta)$ for $\theta \in \mathbb{R}$, and the relationship $H(x) - H(y) = -\partial_2 \mathrm{S}(x,y)/(1-\alpha)$ for $x > y$, are immediate consequences of the fact that for all $\phi \in \mathcal{C}$ and $x < y$,

$$\begin{aligned} \Phi(x,y) &= (y-\theta)\phi'(\theta)\Big|_{\theta=x}^{y} + \int_x^y \phi'(\theta)\,\mathrm{d}\theta \\ &= \int_x^y (y-\theta)\,\mathrm{d}\phi'(\theta) = 2\int_{-\infty}^{\infty} \mathrm{S}_{1/2,\theta}^{\mathrm{E}}(x,y)\,\mathrm{d}\phi'(\theta). \end{aligned}$$

The case $x > y$ is handled analogously, and the case $x = y$ is trivial. Finally, as the increments of H are determined by S, the mixing measure is unique. □

Note that the relations in (3.6) and (3.8) hold pointwise. In particular, the respective integrals are pointwise well-defined. This is because for $(x, y) \in \mathbb{R}^2$ the functions $\theta \mapsto \mathrm{S}^{\mathrm{Q}}_{\alpha,\theta}(x, y)$ and $\theta \mapsto \mathrm{S}^{\mathrm{E}}_{\alpha,\theta}(x, y)$ are right-continuous, non-negative, and uniformly bounded with bounded support, and because the non-decreasing functions g and ϕ' define non-negative measures $\mathrm{d}g$ and $\mathrm{d}\phi'$ that assign finite mass to any finite interval. In particular, given any non-negative measure H that assigns finite mass to any finite interval, the representations (3.6) and (3.8) generate members of the classes $\mathcal{S}^{\mathrm{Q}}_{\alpha}$ and $\mathcal{S}^{\mathrm{E}}_{\alpha}$, respectively. Strict consistency is obtained in case H assigns positive mass to any finite interval.

In the case of quantiles, the asymmetric piecewise linear scoring function corresponds to the choice $g(t) = t$ in (3.2), so the mixing measure H in the representation (3.6) is the Lebesgue measure. The elementary scoring function $\mathrm{S}^{\mathrm{Q}}_{\alpha,\theta}$ arises when $g(t) = \mathbb{1}(\theta < t)$, i.e., when H is a one-point measure in θ.

In the case of expectiles, the mixing measure for the asymmetric squared error scoring function is twice the Lebesgue measure. The choice $\alpha = 1/2$ recovers the mean or expectation functional, for which existing parametric subfamilies emerge as special cases of our mixture representation. Patton's (2015) exponential Bregman family,

$$\mathrm{S}_a(x, y) = \frac{1}{a^2}\left(\exp(ay) - \exp(ax)\right) - \frac{1}{a}\exp(ax)(y - x) \qquad (a \neq 0),$$

which nests the squared error loss in the limit as $a \to 0$, corresponds to the choice $\phi(t) = a^{-2}\exp(at)$ in (3.4). The mixing measure H in the representation (3.8) then has Lebesgue density $h(\theta) = \exp(a\theta)$ for $\theta \in \mathbb{R}$. For Patton's (2011) family

$$\mathrm{S}_b(x, y) = \begin{cases} \frac{y^b - x^b}{b(b-1)} - \frac{x^{b-1}}{b-1}(y - x), & b \notin \{0, 1\}, \\ \frac{y}{x} - \log\frac{y}{x} - 1, & b = 0, \\ y\log\frac{y}{x} - (y - x), & b = 1, \end{cases}$$

of homogeneous scoring functions on the positive half line the mixing measure has Lebesgue density $h(\theta) = \theta^{b-2}\mathbb{1}(\theta > 0)$, remarkably with no case distinction being required. The elementary scoring function $\mathrm{S}^{\mathrm{E}}_{\alpha,\theta}$ emerges when $\phi(t) = (t - \theta)_+$ in (3.4); here the mixing measure in (3.8) is a one-point measure in θ.

From a theoretical perspective, a natural question is whether the mixture representations (3.6) and (3.8) can be considered *Choquet representations* in the sense of functional analysis (Phelps 2001). A Choquet representation is a special, non-redundant type of mixture representation. Specifically, a member S of a convex class $\mathcal{S}$ is an *extreme point* of $\mathcal{S}$ if it cannot be written as an average of two other members, i.e., if $\mathrm{S} = (\mathrm{S}_1 + \mathrm{S}_2)/2$ with $\mathrm{S}_1, \mathrm{S}_2 \in \mathcal{S}$ implies $\mathrm{S}_1 = \mathrm{S}_2 = \mathrm{S}$. Our mixture representations qualify as Choquet representations if the elementary scores $\mathrm{S}^{\mathrm{Q}}_{\alpha,\theta}$ and $\mathrm{S}^{\mathrm{E}}_{\alpha,\theta}$ form extreme points of the underlying classes of scoring functions. This cannot possibly be true for our classes $\mathcal{S}^{\mathrm{Q}}_{\alpha}$ and $\mathcal{S}^{\mathrm{E}}_{\alpha}$ because they are invariant under dilations, hence admit trivial average representations built with multiples of one and the same scoring function. Therefore, the families $\mathcal{S}^{\mathrm{Q}}_{\alpha}$ and $\mathcal{S}^{\mathrm{E}}_{\alpha}$ need to be restricted suitably. Specifically, let the class $\mathcal{I}_1$ consist of all functions $g \in \mathcal{I}$

such that $\lim_{x\to -\infty} g(x) = 0$ and $\lim_{x\to +\infty} g(x) = 1$. Similarly, let $\mathcal{C}_1$ denote the family of all $\phi \in \mathcal{C}$ such that $\phi(0) = 0$ and $\phi' \in \mathcal{I}_1$. These classes are convex, and so are the associated subclasses of the families $\mathcal{S}_\alpha^{\mathrm{Q}}$ and $\mathcal{S}_\alpha^{\mathrm{E}}$, which we denote by $\mathcal{S}_\alpha^{\mathrm{Q},1}$ and $\mathcal{S}_\alpha^{\mathrm{E},1}$, respectively. The elementary scores $\mathrm{S}_{\alpha,\theta}^{\mathrm{Q}}$ and $\mathrm{S}_{\alpha,\theta}^{\mathrm{E}}$ evidently are members of these restricted families.

Proposition 3.1a (quantiles). *For every $\alpha \in (0,1)$ and $\theta \in \mathbb{R}$, the scoring function $\mathrm{S}_{\alpha,\theta}^{\mathrm{Q}}$ is an extreme point of the class $\mathcal{S}_\alpha^{\mathrm{Q},1}$.*

Proposition 3.1b (expectiles). *For every $\alpha \in (0,1)$ and $\theta \in \mathbb{R}$, the scoring function $\mathrm{S}_{\alpha,\theta}^{\mathrm{E}}$ is an extreme point of the class $\mathcal{S}_\alpha^{\mathrm{E},1}$.*

Proof. The specific structure of the scoring functions in (3.2) and (3.4) permits us to focus on the case $\alpha = 1/2$, with the general case $\alpha \in (0,1)$ then being immediate.

In the case of the elementary quantile scoring function (3.7), suppose that $\mathrm{S}_{\alpha,\theta}^{\mathrm{Q}} = (\mathrm{S}_1 + \mathrm{S}_2)/2$, where S_1 and S_2 are of the form (3.2) with associated functions $g_1, g_2 \in \mathcal{I}_1$. Then

$$g_1(x) - g_1(y) + g_2(x) - g_2(y) = \begin{cases} 2, & y \le \theta < x, \\ -2, & x \le \theta < y, \\ 0, & \text{otherwise.} \end{cases}$$

As $g_1, g_2 \in \mathcal{I}_1$ we have $g_j(x) - g_j(y) \in [0,1]$ if $y \le x$, and $g_j(x) - g_j(y) \in [-1,0]$ if $x \le y$, where $j = 1,2$. It follows that

$$g_1(x) - g_1(y) = g_2(x) - g_2(y) = \begin{cases} 1, & y \le \theta < x, \\ -1, & x \le \theta < y, \\ 0, & \text{otherwise.} \end{cases}$$

This coincides with the value distribution of $g(x) - g(y)$ when $g(x) = \mathbb{1}(\theta < x)$, whence indeed $\mathrm{S}_1 = \mathrm{S}_2 = \mathrm{S}_{\alpha,\theta}^{\mathrm{Q}}$.

In the case of the elementary expectile scoring function (3.9), suppose that $\mathrm{S}_{\alpha,\theta}^{\mathrm{E}} = (\mathrm{S}_1 + \mathrm{S}_2)/2$, where S_1 and S_2 are of the form (3.4) with associated functions $\phi_1, \phi_2 \in \mathcal{C}_1$. Let Φ_1, Φ_2 be defined as in (3.10). Then

$$\Phi_1(x,y) + \Phi_2(x,y) = 4\mathrm{S}_{1/2,\theta}^{\mathrm{E}}(x,y).$$

Taking left-hand derivatives with respect to y, we obtain

$$\phi_1'(y) - \phi_1'(x) + \phi_2'(y) - \phi_2'(x) = \begin{cases} -2, & y \le \theta < x, \\ 2, & x \le \theta < y, \\ 0, & \text{otherwise.} \end{cases}$$

As $\phi_1', \phi_2' \in \mathcal{I}_1$, we may apply the same argument as in the quantile case to show that $\phi_1'(x) - \phi_1'(y) = \phi_2'(x) - \phi_2'(y) = \mathbb{1}(\theta < x) - \mathbb{1}(\theta < y)$, whence $\mathrm{S}_1 = \mathrm{S}_2 = \mathrm{S}_{\alpha,\theta}^{\mathrm{E}}$. □

We thus have furnished Choquet representations for subclasses of the consistent scoring functions for quantiles and expectiles. In the extant literature, such Choquet representations have been known in the binary case only, where $y = 1$ corresponds to a success and $y = 0$ to a non-success, so that the mean, $p \in [0, 1]$, of the predictive distribution provides a probability forecast for a success. In this setting, the Savage representation (3.4) for the members of the respective class $\mathcal{S}_{1/2}^{\mathrm{E},1}$ reduces to

$$\mathrm{S}(p, 0) = \frac{1}{2}\,(p\phi'(p) - \phi(p)), \quad \mathrm{S}(p, 1) = \frac{1}{2}\,(\phi(1) - \phi(p) - (1 - p)\phi'(p)).$$

The mixture representation (3.8) can then be written as

$$\mathrm{S}(p, y) = \int_0^1 \mathrm{S}_\theta^{\mathrm{B}}(p, y)\,\mathrm{d}H(\theta), \tag{3.11}$$

where H is a nonnegative measure and

$$\mathrm{S}_\theta^{\mathrm{B}}(p, y) = 2\mathrm{S}_{1/2,\theta}^{\mathrm{E}}(p, y) = \begin{cases} \theta, & y = 0,\ p > \theta, \\ 1 - \theta, & y = 1,\ p \le \theta, \\ 0, & \text{otherwise.} \end{cases} \tag{3.12}$$

Up to unimportant conventions regarding coding, scaling, and gain-loss orientation, this recovers the well known mixture representation of the proper scoring rules for probability forecasts of binary events (Shuford et al. 1966; Schervish 1989). Different choices of the mixing measure yield the standard examples of scoring rules in this case; see Buja et al. (2005) and Table 1 in Gneiting and Raftery (2007). The widely used Brier score,

$$\mathrm{S}(p, 0) = p^2, \quad \mathrm{S}(p, 1) = (1 - p)^2, \tag{3.13}$$

arises when H is twice the Lebesgue measure.

We close the section by noting a fundamental connection between the extremal scoring rules for quantiles, expectiles, and probabilities in (3.7), (3.9), and (3.12), respectively. Specifically, given any predictive CDF, F, and outcome, $y \in \mathbb{R}$,

$$\mathrm{S}_{\alpha,\theta}^{\mathrm{Q}}\big(q_{\alpha,F}^{-}, y\big) = 2\mathrm{S}_{1/2,1-\alpha}^{\mathrm{E}}(1 - F(\theta), \mathbb{1}(y > \theta)) \tag{3.14}$$

for every $\alpha \in (0, 1)$ and $\theta \in \mathbb{R}$. This relation can facilitate computations, particularly in synthetic settings, as we exemplify in Example 4.2.

Order sensitivity

The extremal scoring functions $\mathrm{S}_{\alpha,\theta}^{\mathrm{Q}}$ and $\mathrm{S}_{\alpha,\theta}^{\mathrm{E}}$ are not only consistent for their respective functional, they in fact enjoy the stronger property of order sensitivity. Generally, a scoring function S is *order sensitive* for the functional $F \mapsto \mathrm{T}(F)$ relative to the class $\mathcal{F}$ if, for all $F \in \mathcal{F}$, all $t \in \mathrm{T}(F)$, and all $x_1, x_2 \in \mathbb{R}$,

$$x_2 \le x_1 \le t \implies \mathbb{E}_F[\mathrm{S}(x_2, Y)] \ge \mathbb{E}_F[\mathrm{S}(x_1, Y)],$$

and

$$t \le x_1 \le x_2 \implies \mathbb{E}_F[\mathrm{S}(x_1, Y)] \le \mathbb{E}_F[\mathrm{S}(x_2, Y)].$$

The order sensitivity is *strict* if the above continues to hold when the inequalities involving x_1 and x_2 are strict. As before, we denote the class of the Borel probability measures on $\mathbb{R}$ by $\mathcal{F}_0$, and we write $\mathcal{F}_1$ for the subclass of the probability measures with finite first moment.

Proposition 3.2a (quantiles). *For every $\alpha \in (0,1)$ and $\theta \in \mathbb{R}$, the extremal scoring function $\mathrm{S}^{\mathrm{Q}}_{\alpha,\theta}$ is order sensitive for the α-quantile functional relative to $\mathcal{F}_0$.*

Proposition 3.2b (expectiles). *For every $\alpha \in (0,1)$ and $\theta \in \mathbb{R}$, the extremal scoring function $\mathrm{S}^{\mathrm{E}}_{\alpha,\theta}$ is order sensitive for the α-expectile functional relative to $\mathcal{F}_1$.*

Proof. In the case of the elementary quantile scoring function $\mathrm{S}^{\mathrm{Q}}_{\alpha,\theta}$ in (3.7) suppose first that $x_2 < x_1 \le q_{\alpha,F}$. Since

$$\mathrm{S}^{\mathrm{Q}}_{\alpha,\theta}(x_2, y) - \mathrm{S}^{\mathrm{Q}}_{\alpha,\theta}(x_1, y) = (\mathbb{1}(y \le \theta) - \alpha)\,(\mathbb{1}(\theta < x_2) - \mathbb{1}(\theta < x_1)),$$

we have

$$\mathbb{E}_F[\mathrm{S}^{\mathrm{Q}}_{\alpha,\theta}(x_2, Y)] - \mathbb{E}_F[\mathrm{S}^{\mathrm{Q}}_{\alpha,\theta}(x_1, Y)] = (F(\theta) - \alpha)\,(\mathbb{1}(\theta < x_2) - \mathbb{1}(\theta < x_1)).$$

The second factor on the right-hand side vanishes unless $\theta \in [x_2, x_1)$, and under this latter condition we have $F(\theta) \le \alpha$ and $\mathbb{1}(\theta < x_2) - \mathbb{1}(\theta < x_1) = -1$, whence the desired expectation inequality. The case $q_{\alpha,F} \le x_1 < x_2$ is handled analogously.

In the case of the elementary expectile scoring function $\mathrm{S}^{\mathrm{E}}_{\alpha,\theta}$ in (3.9) we assume first that $x_2 < x_1 \le t$, where t denotes the α-expectile of F. Since

$$\mathrm{S}^{\mathrm{E}}_{\alpha,\theta}(x_2, y) - \mathrm{S}^{\mathrm{E}}_{\alpha,\theta}(x_1, y) = ((1-\alpha)(\theta - y)_+ - \alpha(y - \theta)_+)\,(\mathbb{1}(\theta < x_2) - \mathbb{1}(\theta < x_1)),$$

we get

$$\begin{aligned}&\mathbb{E}_F[\mathrm{S}^{\mathrm{E}}_{\alpha,\theta}(x_2, Y)] - \mathbb{E}_F[\mathrm{S}^{\mathrm{E}}_{\alpha,\theta}(x_1, Y)]\\ &\qquad = ((1-\alpha)\mathbb{E}_F(\theta - Y)_+ - \alpha\mathbb{E}_F(Y - \theta)_+)\,(\mathbb{1}(\theta < x_2) - \mathbb{1}(\theta < x_1)).\end{aligned}$$

As the first term on the right-hand side is strictly increasing in θ and has a unique zero at the α-expectile of F, the proof can be completed in the same way as above. □

Owing to the mixture representations (3.6) and (3.8), the order sensitivity of the extremal scoring functions transfers to all consistent scoring functions. Strict order sensitivity applies if the functions g and ϕ' in (3.2) and (3.4), respectively, are strictly increasing. For suitably large classes $\mathcal{F}$, the respective condition is also necessary. Analogous relationships hold in regard to (strict) consistency.

Recent studies of elicibility have revealed that (strict) order sensitivity and (strict) consistency are in fact equivalent in quite general settings (Nau 1985; Lambert 2013; Steinwart et al. 2014; Bellini and Bignozzi 2015). These results rely on continuity conditions on the scoring function, and do not readily apply in our framework.

Economic interpretation

Our results in the previous section give rise to natural economic interpretations of the extremal scoring functions $\mathrm{S}^{\mathrm{Q}}_{\alpha,\theta}$ and $\mathrm{S}^{\mathrm{E}}_{\alpha,\theta}$, along with the quantile and expectile functionals themselves. In either case, the interpretation relates to a binary betting or investment decision with random outcome, y.

In the case of the extremal quantile scoring function $\mathrm{S}^{\mathrm{Q}}_{\alpha,\theta}$ in (3.7), the payoff takes on only two possible values, relating to a bet on whether or not the outcome y will exceed the event threshold θ. Specifically, consider the following payoff scheme, which is realized in spread betting in prediction markets (Wolfers and Zitzewitz 2008):

- If Quinn refrains from betting, his payoff will be zero, independently of the outcome y.
- If Quinn enters the bet and $y \leq \theta$ realizes, he loses his wager, $\rho_L > 0$.
- If Quinn enters the bet and $y > \theta$ realizes, his winnings are $\rho_G > \rho_L$, for a gain of $\rho_G - \rho_L$.

The top left matrix in Table 3.1 summarizes Quinn's payoff under the decision rule *enter the bet if and only if* $x > \theta$, where x is Quinn's point forecast. This payoff scheme is equivalent to the extremal scoring function $\mathrm{S}^{\mathrm{Q}}_{\alpha,\theta}$. To demonstrate this, we shift attention from positively oriented payoffs to negatively oriented regrets, which we define as the difference between the payoff for an oracle and Quinn's payoff. Here the term oracle refers to a (hypothetical) omniscient bettor who enters the bet if and only if $y > \theta$ realizes, which would yield an ideal payoff $\rho_G - \rho_L$ if $y > \theta$, and zero otherwise. Quinn's regret equals the extremal score $\mathrm{S}^{\mathrm{Q}}_{\alpha,\theta}(x, y)$ except for an irrelevant multiplicative factor. This is illustrated in the bottom left matrix in the table and corresponds to the classical, simple cost-loss decision model (Richardson 2012). In decision theoretic terms, the distinction between payoff and regret is inessential, because the difference depends on the outcome, y, only. In either case, the optimal strategy is to choose $x = q_{\alpha,F}$, where

$$\alpha = \frac{\rho_G - \rho_L}{\rho_G} \in (0, 1), \tag{3.15}$$

and the quantile $q_{\alpha,F}$ is computed from Quinn's predictive CDF, F, for the future outcome y.[1] In summary, Quinn is willing to accept the bet if $q_{\alpha,F} > \theta$.

[1] For simplicity, we assume that F is strictly increasing.

Quantiles

Monetary Payoff

	$y \leq \theta$	$y > \theta$
$x \leq \theta$	0	0
$x > \theta$	$-\rho_L$	$\rho_G - \rho_L$

Score (Regret)

	$y \leq \theta$	$y > \theta$
$x \leq \theta$	0	$\rho_G - \rho_L$
$x > \theta$	ρ_L	0

Expectiles

Monetary Payoff

	$y \leq \theta$	$y > \theta$
$x \leq \theta$	0	0
$x > \theta$	$-(1-\kappa_L)(\theta - y)$	$(1-\kappa_G)(y - \theta)$

Score (Regret)

	$y \leq \theta$	$y > \theta$
$x \leq \theta$	0	$(1-\kappa_G)(y - \theta)$
$x > \theta$	$(1-\kappa_L)(\theta - y)$	0

Table 3.1: Overview of payoff structures for decision rules of the form *enter the bet/invest if and only if $x > \theta$*. Monetary payoffs are positively oriented, whereas scores are negatively oriented regrets relative to an oracle. In the left column, the regret equals the extremal score $\mathrm{S}^{\mathrm{Q}}_{\alpha,\theta}(x,y)$, where $\alpha = (\rho_G - \rho_L)/\rho_G$, up to a multiplicative factor. In the right column, the regret is $\mathrm{S}^{\mathrm{E}}_{\alpha,\theta}(x,y)$, where $\alpha = (1-\kappa_G)/(2-\kappa_G-\kappa_L)$, again up to a multiplicative factor.

In the case of the extremal expectile scoring function $\mathrm{S}^{\mathrm{E}}_{\alpha,\theta}$ in (3.4), the payoff is real-valued. Specifically, suppose that Eve considers investing a fixed amount θ into a start-up company, in exchange for an unknown, future amount y of the company's profits or losses. The payoff structure then is as follows:

- If Eve refrains from the deal, her payoff will be zero, independently of the outcome y.
- If Eve invests and $y \leq \theta$ realizes, her payoff is negative, at $-(1-\kappa_L)(\theta - y)$. Here, $\theta - y$ is the sheer monetary loss, and the factor $1 - \kappa_L$ accounts for Eve's reduction in income tax, with $\kappa_L \in [0,1)$ representing the deduction rate.[2]
- If Eve invests and $y > \theta$ realizes, her payoff is positive, at $(1-\kappa_G)(y-\theta)$, where $\kappa_G \in [0,1)$ denotes the tax rate that applies to her profits.

The top right matrix in Table 3.1 shows Eve's payoff under the decision rule *enter the deal if and only if $x > \theta$*, where x is Eve's point forecast. In order to show that the payoff is equivalent to the extremal scoring function $\mathrm{S}^{\mathrm{E}}_{\alpha,\theta}$, we again shift attention to regrets relative to an omniscient investor or oracle who

[2] In financial terms, the loss acts as a tax shield. The linear functional form assumed here is not unrealistic, even though it is simpler than many real-world tax schemes, where nonlinearities may arise from tax exemptions, progression, etc.

enters the deal if and only if $y > \theta$ occurs, which would yield the ideal payoff $(1-\kappa_G)(y-\theta)_+$. As seen in the bottom right matrix, Eve's regret equals the extremal score $\mathrm{S}^{\mathrm{E}}_{\alpha,\theta}(x,y)$, up to a multiplicative factor. This implies that Eve's optimal decision rule is to enter the deal if and only if the expectile at level

$$\alpha = \frac{1-\kappa_G}{2-\kappa_G-\kappa_L} \in (0,1) \tag{3.16}$$

of her predictive CDF, F, exceeds θ.

Therefore, expectiles induce optimal decision rules in investment problems with fixed costs and differential tax rates for profits versus losses. The mean arises in the special case when $\alpha = 1/2$ in (3.16). It corresponds to situations in which losses are fully tax deductible ($\kappa_G = \kappa_L$) and nests situations without taxes ($\kappa_G = \kappa_L = 0$). Tough taxation settings where $\kappa_L < \kappa_G$ shift Eve's incentives toward not entering the deal and correspond to expectiles at levels $\alpha < 1/2$. For example, if losses cannot be deducted at all ($\kappa_L = 0$), whereas profits are taxed at a rate of $\kappa_G = 1/2$, Eve will invest only if the expectile at level $\alpha = 1/3$ of her predictive CDF, F, exceeds the deal's fixed costs, θ. Note that we permit the case $\theta < 0$, which may reflect subsidies or tax credits, say.

The elementary score $\mathrm{S}^{\mathrm{B}}_{\theta}$ for probability forecasts of a binary event in (3.12) is obtained as the further special case that arises when $\alpha = 1/2$ and $y \in \{0,1\}$. Then $|y-\theta| \in \{\theta, 1-\theta\}$, so that the payoffs in the bottom right matrix of Table 3.1 attain only two possible values. Hence, θ can be interpreted as a cost-loss ratio. We emphasize that this latter interpretation is specific to the binary case. In the general setting where y is continuous, θ takes the role of an event threshold, whereas α governs the costs of under- versus overprediction relative to this threshold.

The above interpretation of expectiles attaches an economic meaning to this class of functionals, which thus far seems to have been missing; e.g., Schulze Waltrup et al. (2015, p. 434) note that "expectiles lack an intuitive interpretation".[3] The foregoing may also bear on the debate about the revision of the Basel protocol for banking regulation, which involves contention about the choice of the functional of in-house risk distributions that banks are supposed to report to regulators (Embrechts et al. 2014). Recently, expectiles have been put forth as potential candidates, as it has been proved that expectiles at level $\alpha \geq 1/2$ are the only elicitable law-invariant coherent risk measures (Ziegel 2016; Bellini and Bignozzi 2015; Delbaen et al. 2016). See McNeil et al. (2015, Chapters 8 and 9) for a recent textbook treatment of these concepts and Fissler et al. (2016) for a discussion of the use of consistent scoring functions in financial regulation.

[3] In very recent work, Bellini and Di Bernardino (2015) offer a succinct financial interpretation of expectiles, which is of the same spirit as ours.

3.1.2 Expected shortfall

The functional

Expected shortfall (ES), being a conditional expectation, is well-defined with respect to the class $\mathcal{F}_1$ of the probability measures with finite first moment. At level $\alpha \in (0,1)$, it is defined as the lower-tail expectation,

$$\mathrm{ES}_\alpha(F) = \frac{1}{\alpha}\int_0^\alpha \mathrm{VaR}_u(F)\,\mathrm{d}u,$$

in close dependence on Value-at-Risk (VaR),

$$\mathrm{VaR}_\alpha(F) = \inf\{z \in \mathbb{R} : F(z) \geq \alpha\},$$

which is equivalent to $q^-_{\alpha,F}$, the left-most α-quantile.

As popular measures of tail risk in the modeling of a financial asset's return, VaR_α and ES_α will typically have negative values and $\mathrm{VaR}_\alpha \geq \mathrm{ES}_\alpha$ always holds. This corresponds to the sign convention of utility functions as used in Delbaen (2012). Both VaR_α and ES_α are functionals of the return distribution F. Following Fissler and Ziegel (2016), one can stack the two functionals to obtain the two-dimensional functional

$$\mathrm{T}_\alpha(F) = (\mathrm{VaR}_\alpha, \mathrm{ES}_\alpha)(F).$$

Scoring functions

Fissler and Ziegel (2016) have recently characterized the family of scoring functions which are consistent for T_α. These scoring functions take the form $\mathrm{S}(x_1, x_2, y)$, where x_1 is the forecast of VaR_α, x_2 is the forecast of ES_α, and y is the realization. By imposing the restriction $x_1 \geq x_2$, we rule out irrational forecasts that violate the logical necessity that $\mathrm{VaR}_\alpha \geq \mathrm{ES}_\alpha$. Again, we consider normalized scores for which $\mathrm{S}(y,y,y) = 0$ holds true, even though other normalizations can easily be accommodated. Corollary 5.5 of Fissler and Ziegel (2016) implies that scoring functions S of the form

$$\begin{aligned} \mathrm{S}(x_1, x_2, y) = {} & (\mathbb{1}(y < x_1) - \alpha)(g(x_1) - g(y)) \\ & + \frac{\phi'(x_2)}{\alpha}(\mathbb{1}(y < x_1) - \alpha)(x_1 - y) \\ & + \phi(y) - \phi(x_2) - \phi'(x_2)(y - x_2), \end{aligned} \tag{3.17}$$

are consistent for T_α with respect to $\mathcal{F}_1$, where $g \in \mathcal{I}$ and $\phi \in \mathcal{C}^+$. The symbol $\mathcal{C}^+$ denotes the class of convex functions $\phi \in \mathcal{C}$ with subgradient $\phi' \in \mathcal{I}$ such that $\lim_{x\to-\infty} \phi'(x) = 0$. If ϕ' is strictly increasing and strictly positive, we obtain strict consistency. For example, the choice $g(z) = 0, \phi(z) = \exp(z)$ satisfies all of these requirements. Subject to regularity assumptions, all normalized consistent scoring functions are of the form (3.17).

Note that the equation (3.17) is a combination of equations (3.2), (3.3), and (3.4). Keeping in mind that the choice of ϕ is restricted compared to (3.4), we take two consistent scoring functions, one for the α-quantile and one for the mean, and then add a mixed term consisting of $\frac{\phi'(x_2)}{\alpha}$ and the asymmetric piecewise linear score (3.3).

Mixture representation

In what follows, we use the symbol $\mathcal{S}_\alpha^{\mathrm{ES}}$ to denote the class of the scoring functions S of the form (3.17) where $g \in \mathcal{I}$ and $\phi \in \mathcal{C}^+$. The family of consistent scoring functions described in (3.17) is rich, and it is hard to justify specific choices of g and ϕ on economic grounds. Furthermore, there is no empirical guidance as to the choice of g and ϕ in practice. Motivated by these concerns, the following result provides a mixture representation for consistent scoring functions in the class $\mathcal{S}_\alpha^{\mathrm{ES}}$.

Theorem 3.2 (VaR, ES). *Any member of the class $\mathcal{S}_\alpha^{\mathrm{ES}}$ admits a representation of the form*

$$\mathrm{S}(x_1, x_2, y) = \mathrm{S}_\alpha^{\mathrm{Q}}(x_1, y) + \int_{-\infty}^{\infty} \mathrm{S}_{\alpha,\theta}^{\mathrm{ES}}(x_1, x_2, y)\, \mathrm{d}H(\theta) \qquad (x_1, x_2, y \in \mathbb{R}, x_1 \geq x_2), \tag{3.18}$$

where $\mathrm{S}_\alpha^{\mathrm{Q}} \in \mathcal{S}_\alpha^{\mathrm{Q}}$, H is a nonnegative measure which is finite on all intervals of the form $(-\infty, x]$, $x \in \mathbb{R}$, and

$$\mathrm{S}_{\alpha,\theta}^{\mathrm{ES}}(x_1, x_2, y) = \frac{\mathbb{1}(\theta < x_2)}{\alpha}(\mathbb{1}(y < x_1) - \alpha)(x_1 - y) + 2\, \mathrm{S}_{1/2,\theta}^{\mathrm{E}}(x_2, y).$$

The mixing measure H is unique and satisfies $\mathrm{d}H(\theta) = \mathrm{d}\phi'(\theta)$ for $\theta \in \mathbb{R}$, where ϕ' is the left-hand derivative of the convex function ϕ in the representation (3.17).

Proof. The mixture representation (3.18), the uniqueness of H, and the fact that $\mathrm{d}H(\theta) = \mathrm{d}\phi'(\theta)$ for $\theta \in \mathbb{R}$, are immediate consequences of Theorems 3.1a and 3.1b and the fact that for all $\phi \in \mathcal{C}^+$,

$$\phi'(x_2) = \int_{-\infty}^{x_2} \mathrm{d}\phi'(\theta). \qquad \square$$

Note that the first term $\mathrm{S}_\alpha^{\mathrm{Q}}$ in (3.18) has its own mixture representation (3.6). The subsequent integral corresponds to the evaluation of ES_α, conditional on VaR_α, where the associated elementary score $\mathrm{S}_{\alpha,\theta}^{\mathrm{ES}}(x_1, x_2, y)$ takes a more complex form in that it depends on both forecasts x_1 and x_2, and the realization y.

Relationship to option pricing

In recent years, the connection between Value-at-Risk, expected shortfall, and European options has been stated explicitly (Mitra 2015; Barone Adesi 2016).

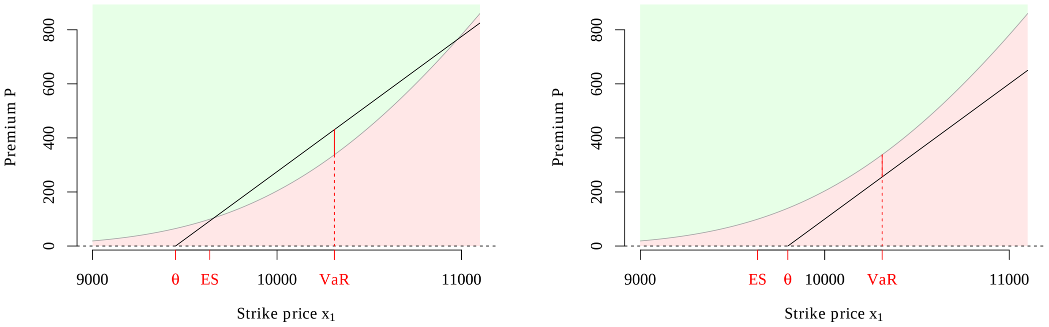


Figure 3.1: The convex curve described by (3.20) separates contracts with positive and negative expected profit. The straight line indicates available contracts. On the left, a positive writing decision $x_2 > \theta$ should be made due to the availability of contracts with positive expected profit, where the optimal strike price x_1 equals $\mathrm{VaR}_{1/2}(F)$. On the right, all contracts offer negative expected profit, so an optimal decision is to not sign any contract, i.e. $x_2 \leq \theta$.

Our extremal score $\mathrm{S}^{\mathrm{ES}}_{\alpha,\theta}$ is equivalent in decision-theoretic terms to the European short put option with its profit described by

$$\pi = P - (K - S)_+,$$

where P is the put option's price, K is the strike price, and S is the spot price. We see the extremal scores' relation to π by identifying the spot price S with y, the strike price K with x_1, and by imposing $\alpha(x_1 - \theta)$ as the premium P's structure, such that

$$\begin{aligned} \pi &= \alpha(x_1 - \theta) - (x_1 - y)_+ \\ &= \alpha(y - \theta)_+ - \alpha \mathrm{S}^{\mathrm{ES}}_{\alpha,\theta}(x_1, x_2, y), \end{aligned} \tag{3.19}$$

conditional on a positive writing decision $x_2 > \theta$. Actions are limited to the choice of x_1 and x_2, corresponding to the strike price and the writing decision, respectively. The first term, $\alpha(y - \theta)_+$, describes the best case scenario without playing a role in the decision-making problem, while the second term can be interpreted as the regret, solely determining the best course of action.

Let F denote a market participant's belief about the spot price of a given asset. We can find equilibria where a participant is indifferent between buying and writing a put option. As an example, the Black-Scholes pricing model calculates

$$P = \alpha(\mathrm{VaR}_\alpha(F) - \mathrm{ES}_\alpha(F)) \tag{3.20}$$

under a geometric Brownian motion propagation of the current stock price. For simplicity, we omit the multiplicative factor related to the risk-free interest rate. Figure 3.1 mimics the investment decision for an European short put option on the DAX (German stock index) with one month maturity, where the available contracts are restricted by the premium's price structure as in (3.19) with $\alpha = 1/2$. The convex curve described by (3.20) separates contracts with positive and negative expected profit based on the subjective belief F, i.e. considerations such as the distribution model, volatility, and trend. The figure illustrates that under the extremal score $\mathrm{S}^{\mathrm{SE}}_{1/2,\theta}$, the functional $\mathrm{VaR}_{1/2}$ always leads to an optimal choice for x_1, while x_2 can be anything that lies on the same side of θ as $\mathrm{ES}_{1/2}(F)$. Clearly, when θ is sufficiently small, suboptimal choices of x_1 still lead to acceptable contracts with positive expected profit.

3.2 Proper scoring rules

An interesting question is whether there might be mixture representations in terms of interpretable elementary scores for proper scoring rules. As noted, a scoring rule $\mathrm{S}(F, y)$ assigns a loss or penalty when we issue the predictive CDF F and y realizes, and for a scoring rule to be proper, the expectation inequalities in (2.3) need to hold.

In the following, we investigate the existence of mixture representations for proper scoring rules via the equivalent representations using corresponding entropy functions.

Proposition 3.3a. *If a mixture representation of a proper scoring rule* S *in terms of proper scoring rules* S_θ*,*

$$\mathrm{S}(F, y) = \int_\Theta \mathrm{S}_\theta(F, y)\,\mathrm{d}H(\theta), \tag{3.21}$$

exists, for all F *and* y*, then a mixture representation of an entropy function* G *corresponding to* S *in terms of entropy functions* G_θ *corresponding to* S_θ*,*

$$G(F) = \int_\Theta G_\theta(F)\,\mathrm{d}H(\theta), \tag{3.22}$$

exists, for all F*.*

Proposition 3.3b. *If a mixture representation of an entropy function* G *in terms of entropy functions* G_θ*, as in* (3.22)*, exists for all* F*, then a mixture representation of a proper scoring rule* S *corresponding to* G *in terms of proper scoring rules* S_θ *corresponding to* G_θ*, as in* (3.21)*, exists for all* F *and* y*.*

Proof. For Proposition 3.3a, if a representation as in (3.21) exists, then

$$G(F) = \int_\mathbb{R} \mathrm{S}(F, x)\,\mathrm{d}F(x) = \int_\Theta \int_\mathbb{R} \mathrm{S}_\theta(F, x)\,\mathrm{d}F(x)\,\mathrm{d}H(\theta) = \int_\Theta G_\theta(F)\,\mathrm{d}H(\theta)$$

using Fubini's Theorem. For Proposition 3.3b, let a representation as in (3.22) exist, and let S_θ be proper scoring rules with entropy functions G_θ. Then, S defined as in (3.21) is a proper scoring rule corresponding to G. □

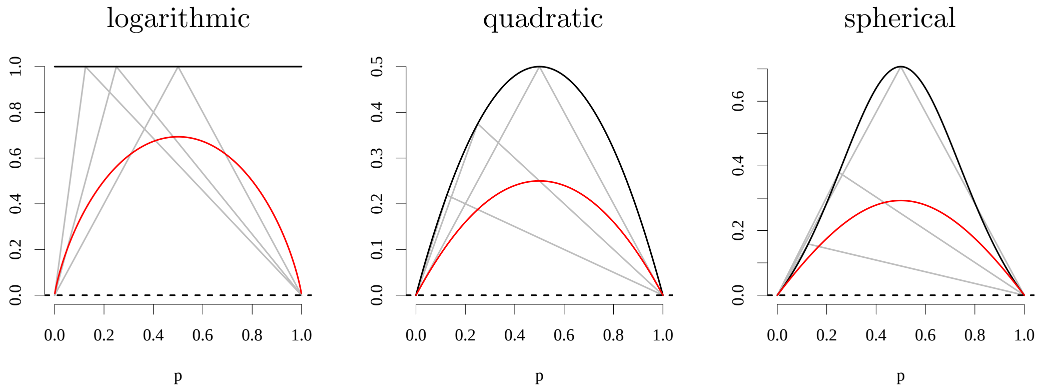


Figure 3.2: Illustration of the mixture representations in the binary case for quadratic, logarithmic, and spherical score as defined in equations (3.34), (3.35), and (3.36). The respective entropy functions are displayed in red, and the corresponding weighted extremal entropy functions (3.23), for $\theta = 0.125, 0.25, 0.5$, are shown in grey. Maxima of the weighted extremal entropy functions are indicated in black.

The classical zero-one score for binary events, where a forecast is made as a single value $p \in [0,1]$ for the probability of success, is defined by the scoring rule

$$\mathrm{S}^{\mathrm{B}}_{\theta}(p,y) = \begin{cases} \theta, & y = 0, p > \theta, \\ 1-\theta, & y = 1, p \leq \theta, \\ 0, & \text{otherwise}, \end{cases}$$

where $\theta \in (0,1)$ can be interpreted as a cost-loss ratio. The corresponding entropy function is given by

$$G_{\theta}(p) = \min(p,\theta) - \theta p. \tag{3.23}$$

In the previous section, and earlier by Schervish (1989), it was shown that under mild regularity conditions any proper scoring rule for binary events can be constructed by integration over the family of asymmetric zero-one scores. According to Proposition 3.3a, given the entropy function G of a proper scoring rule, we can find a non-negative measure H on the open interval $(0,1)$ such that

$$G(p) = \int_{(0,1)} G_{\theta}(p)\,\mathrm{d}H(\theta)$$

for all $p \in (0,1)$. Figure 3.2 provides a visual representation for some of the scores reviewed by Gneiting and Raftery (2007).

3.2.1 Categorical events

We start with some notation that is specific to categorical events described by the sample space $\Omega^m = \{1,\dots,m\}$. Various subsets of $\mathbb{R}^m$ are relevant in our

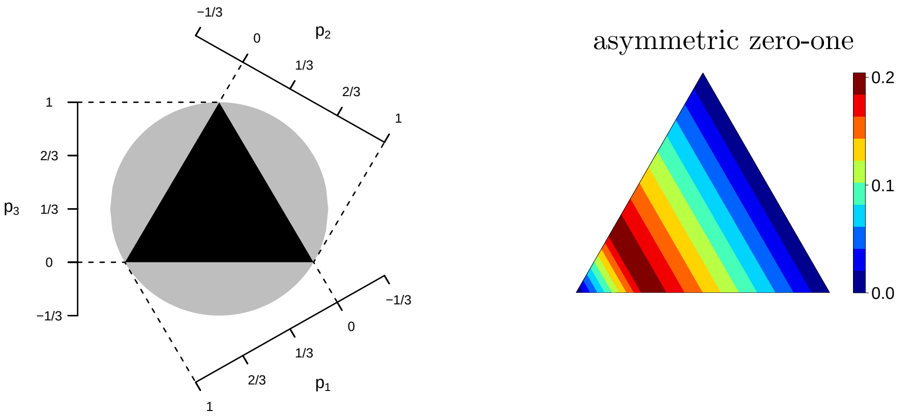


Figure 3.3: At left, we have a two-dimensional representation of Δ^2 (and B^2) which we subsequently use for levelplots, as exemplified at right. Here, we visualize the entropy function (3.31) with parameter choices $\theta = 5/7$ and $j = 1$.

considerations,

$$\begin{aligned}
\mathcal{H}^{m-1} &= \{(p_1, \ldots, p_m) \in \mathbb{R}^m \ : \ p_1 + \ldots + p_m = 1\}, && (3.24)\\
\Delta^{m-1} &= \{(p_1, \ldots, p_m) \in \mathcal{H}^{m-1} \ : \ p_1, \ldots, p_m \geq 0\}, && (3.25)\\
\mathbb{S}^{m-1} &= \{(p_1, \ldots, p_m) \in \mathcal{H}^{m-1} \ : \ \|p - c_m\| = r_m\}, && (3.26)\\
B^{m-1} &= \{(p_1, \ldots, p_m) \in \mathcal{H}^{m-1} \ : \ \|p - c_m\| < r_m\}, && (3.27)
\end{aligned}$$

where $c_m = (\frac{1}{m}, \ldots, \frac{1}{m})$, $r_m = \left(1 - \frac{1}{m}\right)^{1/2}$, and $\|\cdot\|$ denotes the Euclidean distance. The most general is the hyperplane $\mathcal{H}^{m-1}$ such that the sum of all components equals 1. An additional restriction of non-negativity in each dimension gives the standard $(m-1)$-simplex Δ^{m-1} which can be identified with the class of probability distributions for m categories. As a regular polytope, the standard simplex has a circumscribed sphere $\mathbb{S}^{m-1}$ with center c_m and radius r_m. The corresponding convex hull without its extremal points is denoted by B^{m-1}.

In the representation by Savage (1971), any concave function $G : \Delta^{m-1} \to \mathbb{R}$ with super-gradient G' induces a proper scoring rule $\mathrm{S} : \Delta^{m-1} \times \Omega^m \to \mathbb{R}$ by the relationship

$$\mathrm{S}(p, i) = G(p) + \langle G'(p), e_i - p \rangle, \tag{3.28}$$

where $\langle \cdot, \cdot \rangle$ denotes the standard scalar product and e_i is the i-th unit vector in $\mathbb{R}^m$. The entropy function G is called *regular* if it attains the same value (usually zero) for all extremal probability distributions. Formally, this can be expressed as

$$G(e_j) = G(e_k) \tag{3.29}$$

for all $j, k = 1, \ldots, m$.

Separable entropy functions

An entropy function $G : \Delta^{m-1} \to \mathbb{R}$ is separable if there exist m concave functions $g_1, \ldots, g_m : [0,1] \to \mathbb{R}$ such that

$$G(p) = \sum_{j=1}^{m} g_j(p_j). \tag{3.30}$$

A multivariate generalization of the entropy function for binary events in (3.23), is the concave function

$$G^{\mathrm{A}}_{\theta,\,j}(p) = \min(p_j, \theta) - \theta\, p_j, \tag{3.31}$$

which belongs to a family that is parameterized by the domain $\Theta^{\mathrm{A}} = (0,1) \times \Omega^m$. Figure 3.3 illustrates this entropy function in a level plot for $m = 3$. The entropy function in (3.31) is separable, and also extremal in the class of concave functions on $\mathcal{H}^{m-1}$ as a result of describing the minimum of two affine functions (Johansen 1974; Bronshtein 1978). It defines a multivariate version of the asymmetric zero-one score for binary events,

$$\mathrm{S}^{\mathrm{A}}_{\theta,\,j}(p,i) = \begin{cases} \theta, & i \neq j,\ p_j > \theta, \\ 1-\theta, & i = j,\ p_j \leq \theta, \\ 0, & \text{otherwise.} \end{cases} \tag{3.32}$$

The separable mixtures of entropy functions take the form

$$G(p) = \sum_{j=1}^{m} \int_0^1 G^{\mathrm{A}}_{\theta,j}(p)\,\mathrm{d}H_j(\theta), \tag{3.33}$$

where each H_j is a non-negative measure on $(0,1)$. Thus, the class of multivariate asymmetric zero-one scores spans the class of proper scoring rules for categorical events with regular separable entropy functions, and furthermore, any mixture of separable entropy functions is again separable.

Example 3.1. We give three examples of entropy functions, visualized in Figure 3.4, only two of which are separable for $m > 2$.

(a) The entropy function of the logarithmic score,

$$G(p) = -\sum_{j=1}^{m} p_j \log p_j, \tag{3.34}$$

admits a representation as in (3.33), where $H_j(\theta) = \log(\theta)$ for all $j = 1, \ldots, m$.

(b) The entropy function of the quadratic score,

$$G(p) = 1 - \sum_{j=1}^{m} p_j^2, \tag{3.35}$$

admits a representation as in (3.33), where $H_j(\theta) = 2\theta$ for all $j = 1, \ldots, m$.

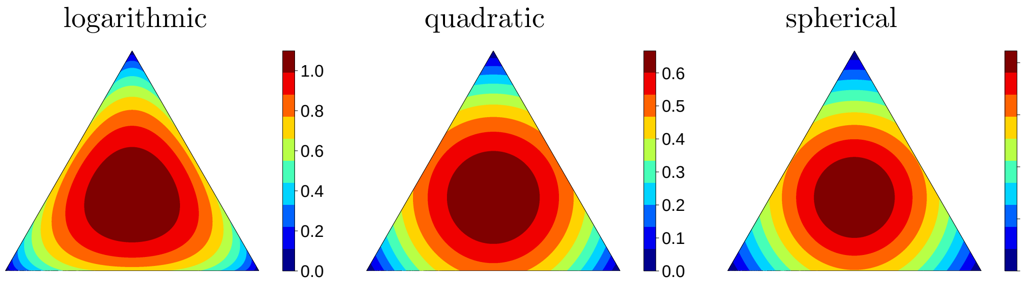


Figure 3.4: Levelplots of the entropy functions (3.34), (3.35), and (3.36), from Example 3.1 on Δ^2.

(c) The entropy function of the spherical score is given by

$$G(p) = 1 - \left(\sum_{j=1}^{m} p_j^2 \right)^{1/2}, \tag{3.36}$$

and therefore only admits a representation as in (3.33) for $m = 2$. In the binary case, all entropy functions are separable since they can always be fully described as a function of p_1.

Directional scoring rules

We are now interested in non-separable generalizations of the binary asymmetric zero-one score,

$$\mathrm{S}_\theta^{\mathrm{B}}(p, y) = \begin{cases} \theta, & y = 0, p > \theta, \\ 1 - \theta, & y = 1, p \leq \theta, \\ 0, & \text{otherwise.} \end{cases}$$

The parameter θ in this scoring rule divides the interval $[0, 1]$ into subintervals $[0, \theta]$ and $(\theta, 1]$ on which the score of a binary distribution described by $p \in [0, 1]$ is constant. This introduces the notion of *directionality*. All distributions p to the left of θ receive the same score, and analogously for all distributions to the right of θ.

Definition 3.1 (directional scoring rules). *A proper scoring rule* S_θ *for events with* m *categories is* directional *with respect to* $\theta \in \mathcal{H}^{m-1}$ *if*

$$\mathrm{S}_\theta(p, i) = \mathrm{S}_\theta(p_0, i) \tag{3.37}$$

for all $i = 1, \dots, m$*, whenever* $p_0 \in \Delta^{m-1}$ *and* p *lies in the intersection of* Δ^{m-1} *and the ray* $\overrightarrow{\theta p_0}$ *generated by* p_0 *with origin* θ*,*

$$\overrightarrow{\theta p_0} = \Big\{ \theta + \lambda(p_0 - \theta) : \lambda > 0 \Big\}.$$

We denote the respective class of proper scoring rules by $\mathcal{S}_\theta^{\mathrm{D}}$*.*

Proper scoring rules are essentially supergradients of their corresponding entropy functions. For a given ray $\overrightarrow{\theta p_0}$, the entropy function of a directional proper scoring rule $\mathrm{S}_\theta \in \mathcal{S}_\theta^{\mathrm{D}}$ satisfies the functional equation

$$G_\theta(p) = G_\theta(\theta) + \langle G'_\theta(p), p - \theta \rangle \tag{3.38}$$

for all $p \in \overrightarrow{\theta p_0}$, where the supergradient G'_θ is constant along the ray $\overrightarrow{\theta p_0}$, namely

$$G'_\theta(p) = \Big(\mathrm{S}_\theta(p_0, 1), \ldots, \mathrm{S}_\theta(p_0, m)\Big)'.$$

Naturally, the functional equation (3.38) needs to hold for any ray with origin $\theta \in \mathcal{H}^{m-1}$. Entropy functions that satisfy (3.38) for all rays admit a representation of the form

$$G_\theta(p) = g(p - \theta) + a_\theta(p), \tag{3.39}$$

where a_θ is an affine function in p, and $g : \{x \in \mathbb{R}^m : x_1 + \cdots + x_m = 0\} \to \mathbb{R}$ is a concave function that satisfies

$$g(\lambda x) = \lambda g(x)$$

for all $\lambda > 0$, i.e. that is positive homogeneous of degree 1. We use the symbol $\mathcal{G}_\theta^{\mathrm{D}}$ to denote the class of entropy functions that are of the form (3.39). This class is a generalization of the extremal entropy functions (3.23) for the binary case to an arbitrary number of categories m. Note that the entropy functions of directional scoring rules are typically non-separable.

We can now give the following version of the Savage (1971) representation,

$$\mathrm{S}(p, i) = G(p) + \langle G'(p), e_i - p \rangle,$$

for the special case of directional proper scoring rules. A proper scoring rule S_θ is a member of the class $\mathcal{S}_\theta^{\mathrm{D}}$ if and only if there exists a representation

$$\mathrm{S}_\theta(p, i) = G_\theta(\theta) + \langle G'_\theta(p), e_i - \theta \rangle, \tag{3.40}$$

where G_θ is a member of $\mathcal{G}_\theta^{\mathrm{D}}$ with corresponding supergradient G'_θ.

Natural questions regarding this class are for the extremal members and possible mixture representations. Results by Johansen (1974) and Bronshtein (1978) suggest that the members of $\mathcal{G}_\theta^{\mathrm{D}}$ that are of the form (3.41) are indeed extremal. However, it remains to be shown whether all extremal members are of that form.

Conjecture 3.1 (extremal members of $\mathcal{G}_\theta^{\mathrm{D}}$). *Any extremal member of the class $\mathcal{G}_\theta^{\mathrm{D}}$ with $\theta \in \mathcal{H}^{m-1}$ is of the form*

$$\min_{i=1,\dots,j} \alpha_i(p), \tag{3.41}$$

for some $j \in \{1, \dots, m\}$, where the $\alpha_i : \mathcal{H}^{m-1} \to \mathbb{R}$ are affine functions with an intersection

$$\mathbb{I}_j = \{p : \alpha_1(p) = \ldots = \alpha_j(p)\},$$

such that $\theta \in \mathbb{I}_j$ and $\dim(\mathbb{I}_j) = m - j$*.*

We leave further investigations into the class of extremal members, and the resulting mixture representations, to future research. For now, we confine ourselves to the illustration of select directional proper scoring rules.

Example 3.2. We give three examples of directional proper scoring rules and their entropy functions, which are visualized in Figure 3.5.

(a) Consider the regular entropy functions of proper scoring rules that are directional with respect to $\theta \in \Delta^{m-1}$, which correspond to extremal members with $k = 2$. Subject to rescaling, these entropy functions can be written as

$$G_{\theta,\phi}(p) = \min\left(\langle p, \phi\rangle, \langle \theta, \phi\rangle\right) - \sum_{j=1}^{m} p_j \min\left(\langle e_j, \phi\rangle, \langle \theta, \phi\rangle\right) \tag{3.42}$$

where $\phi \in \mathbb{S}^{m-1}$. One proper scoring rule corresponding to $G_{\theta,\phi}$ is given by

$$\mathrm{S}_{\theta,\phi}(p, i) = \begin{cases} \langle \theta\,, \phi\rangle - \min(\langle e_i, \phi\rangle, \langle \theta, \phi\rangle), & \langle p, \phi\rangle > \langle \theta, \phi\rangle, \\ \langle e_i, \phi\rangle - \min(\langle e_i, \phi\rangle, \langle \theta, \phi\rangle), & \langle p, \phi\rangle \le \langle \theta, \phi\rangle, \end{cases} \tag{3.43}$$

with multiple options for the supergradient $G'_{\theta,\phi}$ on the intersection $\langle p, \phi\rangle = \langle \theta, \phi\rangle$ of the two affine functions.

The proper scoring rule $\mathrm{S}_{\theta,\phi}$ is a generalization of the multivariate asymmetric zero-one score $\mathrm{S}^{\mathrm{A}}_{\theta,j}$ from (3.32), where the additional instances of entropy functions are all non-separable. The multivariate asymmetric zero-one score can be recovered by choosing $\phi = e_j$.

(b) The *directional oblique elliptical score* is generated by

$$g(x) = -\|x\|, \tag{3.44}$$

and choosing the affine function a_θ in (3.39) to ensure regularity of the resulting entropy function. Adding an affine function distorts the spherical level sets of $\|x\|$ depending on the choice of $\theta \in \mathcal{H}^{m-1}$ leading to the entropy function and corresponding proper scoring rule

$$G_\theta(p) = -\|p - \theta\| + \sum_{j=1}^{m} p_j \|e_j - \theta\|, \tag{3.45}$$

$$\mathrm{S}_\theta(p, i) = \begin{cases} \left\langle -\frac{p-\theta}{\|p-\theta\|}, e_i - \theta \right\rangle + \|e_i - \theta\|, & p \neq \theta, \\ G_\theta(\theta), & p = \theta. \end{cases} \tag{3.46}$$

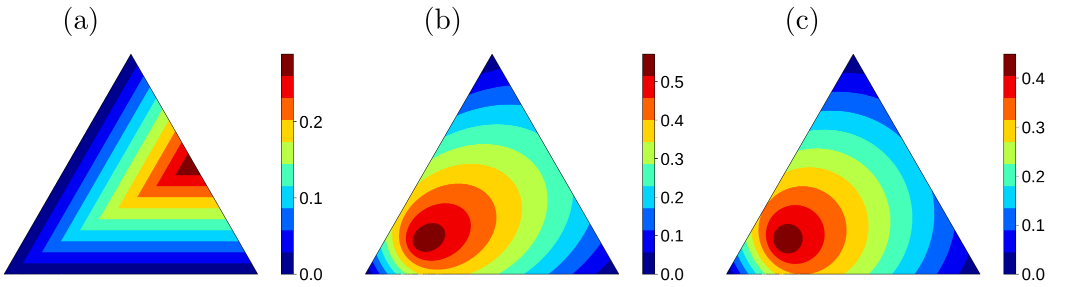


Figure 3.5: Levelplots of the entropy functions (3.42), (3.45), and (3.47), from Example 3.2 on Δ^2. The parameters are $\theta = \left(\frac{5}{7}, \frac{1}{7}, \frac{1}{7}\right)$ and ϕ such that $\phi - c_m$ is orthogonal to $\theta - c_m$.

The parameter space for entropy functions of the form (3.45) can readily be extended to $\mathbb{R}^m$, thus losing the directionality property but generating strictly proper scoring rules when θ does not lie in the hyperplane $\mathcal{H}^{m-1}$. A common choice is $\theta = 0$ which yields the spherical score (3.36).

(c) Another interesting proper scoring rule, the *directional oblique spherical score*, is generated by a variation of the function g in (3.44),

$$g^{\mathrm{C}}(x) = -\sqrt{\|x\|^2 + \langle x, \theta\rangle^2 - \|x\|^2\|\theta\|^2},$$

where g^{C} is a real-valued function when $\theta \in B^{m-1}$. In contrast to the previous example, adding a regularizing affine function leads to spherical level sets of the corresponding regular entropy function, in that

$$G^{\mathrm{C}}_\theta(p) = 1 - \langle p, \theta\rangle - \sqrt{\|p-\theta\|^2 + \langle p, \theta\rangle^2 - \langle p, p\rangle\langle\theta, \theta\rangle}, \tag{3.47}$$

$$\mathrm{S}^{\mathrm{C}}_\theta(p, i) = \begin{cases} 1 - \theta_i + \frac{\langle p-\theta, e_i-\theta\rangle + \theta_i\langle p,\theta\rangle - p_i\langle\theta,\theta\rangle}{g^{\mathrm{C}}(p-\theta)}, & p \neq \theta, \\ 1 - \|\theta\|^2, & p = \theta. \end{cases} \tag{3.48}$$

Incidentally, the entropy function G^{C}_θ describes the surface of an oblique circular cone, i.e. suitably rescaled it is the solution in $\lambda \geq 0$ to the equation

$$1 - \lambda = \|p - \lambda\theta\|, \tag{3.49}$$

which requires $\theta \in B^{m-1}$ for a unique solution.

Quadratic score

As noted, the quadratic score (3.35) has a separable entropy function and thus admits a mixture representation of the form (3.33). In this section, we give an alternative representation as a mixture of non-separable entropy functions G^{C}_θ in (3.47). This demonstrates that many alternative representations may exist and that mixtures of non-separable entropy functions can be separable. The integral in Proposition 3.4 is a surface integral over B^{m-1} in m-dimensional space. Since we integrate with respect to m times the uniform probability measure on B^{m-1}, we require the formula for the volume of an n-dimensional ball with radius r, given by

$$V_n(r) = \frac{\pi^{n/2} r^n}{\Gamma(\frac{n}{2}+1)}.$$

Proposition 3.4. *Let $p \in B^{m-1}$, $m \geq 2$. Then the mixture representation*

$$\int_{B^{m-1}} G_\theta^{\mathrm{C}}(p)\,\mathrm{d}H(\theta) = 1 - \|p\|^2 \tag{3.50}$$

holds, where H is m times the uniform probability measure on B^{m-1}.

Proof. We reparameterize G_θ^{C} to allow a representation of the surface integral over B^{m-1} as a double integral of $(m-2)$-dimensional slices along the axis given by p and c_m. Then, we find recurrence relations of order 1 in odd and even m, respectively, using hypergeometric summation theory. We start with the reparameterization

$$\begin{aligned}
G_\theta^{\mathrm{C}}(p) &= 1 - \langle p, \theta\rangle - \sqrt{\|p-\theta\|^2 + \langle p,\theta\rangle^2 - \langle p,p\rangle\langle\theta,\theta\rangle}, \\
&= 1 - \langle p,\theta\rangle - \sqrt{(1-\langle p,\theta\rangle)^2 - (1-\langle p,p\rangle)(1-\langle\theta,\theta\rangle)} \\
&= r_m^2\left(1 - uv - \sqrt{(1-uv)^2 - (1-u^2)(1-v^2-w^2)}\right) \\
&= r_m^2(1-uv)\left(1 - \sqrt{1 - \tfrac{(1-u^2)(1-v^2-w^2)}{(1-uv)^2}}\right)
\end{aligned}$$

where

$$\begin{aligned}
u &= \sqrt{\langle \tfrac{p-c_m}{r_m}, \tfrac{p-c_m}{r_m}\rangle}, \\
v &= \begin{cases} 0, & p = c_m, \\ u^{-1}\langle \tfrac{p-c_m}{r_m}, \tfrac{\theta-c_m}{r_m}\rangle, & \text{otherwise}, \end{cases} \\
w &= \sqrt{\langle \tfrac{\theta-c_m}{r_m}, \tfrac{\theta-c_m}{r_m}\rangle - v^2},
\end{aligned}$$

such that $u \in [0,1)$, $v \in (-1,1)$, and $w \in [0,1)$. For $m = 2$, we choose a representation of the integral that allows embedding into the results for $m \geq 3$,

$$\begin{aligned}
&\int_{B^1} G_\theta^{\mathrm{C}}(p)\,\mathrm{d}H(\theta) = \\
&\quad = \frac{2r_2^2}{V_1(r_2)} \int_{-1}^{1} r_2(1-uv)\left(1 - \sqrt{1 - \tfrac{(1-u^2)(1-v^2)}{(1-uv)^2}}\right)\mathrm{d}v \\
&\quad = \frac{2r_2^2}{B(\frac{1}{2},\frac{2}{2})} \int_{-1}^{1} (1-uv)\left(1 - {}_2F_1\left(-\tfrac{1}{2}, 1; 1; \tfrac{(1-u^2)(1-v^2)}{(1-uv)^2}\right)\right)\mathrm{d}v.
\end{aligned}$$

For higher dimensions, $m \geq 3$, we get the following representation

$$\begin{aligned}
&\int_{B^{m-1}} G_\theta^{\mathrm{C}}(p)\,\mathrm{d}H(\theta) = \\
&\quad = \frac{m r_m^2}{V_{m-1}(r_m)} \int_{-1}^{1} r_m(1-uv) \int_0^{\sqrt{1-v^2}} V_{m-2}(r_m) \left(1 - \sqrt{1 - \tfrac{(1-u^2)(1-v^2-w^2)}{(1-uv)^2}}\right) \mathrm{d}w^{m-2}\,\mathrm{d}v \\
&\quad = \frac{m r_m^2}{B\left(\frac{1}{2}, \frac{m}{2}\right)} \int_{-1}^{1} (1-uv) \int_0^{\sqrt{1-v^2}} \left(1 - \sqrt{1 - \tfrac{(1-u^2)(1-v^2-w^2)}{(1-uv)^2}}\right) \mathrm{d}w^{m-2}\,\mathrm{d}v \\
&\quad = \frac{m r_m^2}{B\left(\frac{1}{2}, \frac{m}{2}\right)} \int_{-1}^{1} (1-uv)(1-v^2)^{(m-2)/2} \left(1 - {}_2F_1\left(-\tfrac{1}{2}, 1; \tfrac{m}{2}; \tfrac{(1-u^2)(1-v^2)}{(1-uv)^2}\right)\right) \mathrm{d}v,
\end{aligned}$$

using the substitution $w^2 = (1-v^2)(1-w')$ and Euler's integral representation of hypergeometric functions, i.e.

$$\begin{aligned}
&\int_0^{\sqrt{1-v^2}} \sqrt{1 - \tfrac{(1-u^2)(1-v^2-w^2)}{(1-uv)^2}}\,\mathrm{d}w^{m-2} \\
&\qquad = (1-v^2)^{(m-2)/2} \int_0^1 (\tfrac{m}{2}-1)(1-w')^{\frac{m}{2}-2} \sqrt{1 - w' \tfrac{(1-u^2)(1-v^2)}{(1-uv)^2}}\,\mathrm{d}w' \\
&\qquad = (1-v^2)^{(m-2)/2}\ {}_2F_1\left(-\tfrac{1}{2}, 1; \tfrac{m}{2}; \tfrac{(1-u^2)(1-v^2)}{(1-uv)^2}\right).
\end{aligned}$$

A further substitution, $v = 2v' - 1$, in combination with the change of variable $x = \frac{1-u}{1+u}$ and the binomial theorem, results in a hypergeometric double sum,

$$\begin{aligned}
&\frac{m}{B\left(\frac{1}{2}, \frac{m}{2}\right)} \int_{-1}^{1} (1-uv)(1-v^2)^{(m-2)/2} \left(1 - {}_2F_1\left(-\tfrac{1}{2}, 1; \tfrac{m}{2}; \tfrac{(1-u^2)(1-v^2)}{(1-uv)^2}\right)\right) \mathrm{d}v \\
&\quad = \frac{1-u^2}{B\left(\frac{1}{2}, \frac{m}{2}\right)} \int_{-1}^{1} \frac{(1-v^2)^{m/2}}{1-uv}\ {}_2F_1\left(\tfrac{1}{2}, 1; \tfrac{m}{2}+1; \tfrac{(1-u^2)(1-v^2)}{(1-uv)^2}\right) \mathrm{d}v \\
&\quad = \frac{1-u^2}{B\left(\frac{1}{2}, \frac{m}{2}\right)} 2^{m+1} \int_0^1 \frac{(v'(1-v'))^{m/2}}{1+u-2uv'}\ {}_2F_1\left(\tfrac{1}{2}, 1; \tfrac{m}{2}+1; \tfrac{(1-u^2)4v'(1-v')}{(1+u-2uv')^2}\right) \mathrm{d}v' \\
&\quad = \frac{x}{(1+x)} \frac{2^{m+2}}{B\left(\frac{1}{2}, \frac{m}{2}\right)} \int_0^1 \frac{(v'(1-v'))^{m/2}}{(1-(1-x)v')}\ {}_2F_1\left(\tfrac{1}{2}, 1; \tfrac{m}{2}+1; \tfrac{x4v'(1-v')}{(1-(1-x)v')^2}\right) \mathrm{d}v' \\
&\quad = \frac{x}{(1+x)} \frac{2^{m+2}}{B\left(\frac{1}{2}, \frac{m}{2}\right)} \sum_{k=0}^{\infty} \frac{(\frac{1}{2})_k}{(\frac{m}{2}+1)_k} 2^{2k} x^k \int_0^1 \frac{(v'(1-v'))^{m/2+k}}{(1-(1-x)v')^{1+2k}}\,\mathrm{d}v'
\end{aligned}$$

$$
\begin{aligned}
&= \frac{x}{1+x}\sum_{k=0}^{\infty} m\frac{(\frac{1}{2})_k\Gamma(\frac{1}{2}+\frac{m}{2})}{\Gamma(1+\frac{m}{2}+k)\Gamma(\frac{1}{2})}2^{m+2k+1}x^k \\
&\qquad\qquad\qquad \sum_{l=0}^{\infty}\frac{(2k+1)_l}{(1)_l}(1-x)^l\int_0^1 (v')^{m/2+k+l}(1-v')^{m/2+k}\,\mathrm{d}v' \\
&= \frac{x}{1+x}\sum_{k=0}^{\infty}\sum_{l=0}^{\infty} m\left(2^{2k}\frac{(\frac{1}{2})_k(1)_{2k+l}}{(1)_{2k}(1)_l}\right)\left(2^{m+1}\frac{\Gamma(1+\frac{m}{2}+k+l)\Gamma(\frac{1}{2}+\frac{m}{2})}{\Gamma(2+m+2k+l)\Gamma(\frac{1}{2})}\right)x^k(1-x)^l \\
&= \frac{x}{1+x}\sum_{k=0}^{\infty}\sum_{l=0}^{\infty}F(m,k,l;x),
\end{aligned}
$$

where we use the definition $(a)_n = \frac{\Gamma(a+n)}{\Gamma(a)}$ for the rising factorial. We now have a representation of the surface integral as a hypergeometric double sum,

$$
\int_{B^{m-1}} G_\theta^{\mathrm{C}}(p)\,\mathrm{d}H(\theta) = \frac{xr_m^2}{1+x}\sum_{k=0}^{\infty}\sum_{l=0}^{\infty}F(m,k,l;x),
$$

and the following representation of $1-\|p\|^2$ in terms of x,

$$
1-\|p\|^2 = r_m^2(1-u^2) = r_m^2\left(1-\left(\frac{1-x}{1+x}\right)^2\right) = \frac{4xr_m^2}{(1+x)^2}.
$$

This leaves us with the following representation of the equation we need to prove, namely

$$
\sum_{k=0}^{\infty}\sum_{l=0}^{\infty}F(m,k,l;x) = \frac{4}{1+x}. \tag{3.51}
$$

The sequence of summands simplifies using the duplication formula for the gamma function and splits along odd and even m,

$$
\begin{aligned}
F(m,k,l;x) &= m\,\frac{(1)_{2k+l}}{(1)_k(1)_l}\left(2^{m+1}\frac{\Gamma(\frac{1}{2}+\frac{m}{2})\Gamma(1+\frac{m}{2}+k+l)}{\Gamma(\frac{1}{2})\Gamma(2+m+2k+l)}\right)x^k(1-x)^l \\
&= \begin{cases} F_{\text{even}}(n,k,l;x), & m=2n, \\ F_{\text{odd}}(n,k,l;x), & m=2n+1, \end{cases} \qquad \text{for } n=1,2,\dots,
\end{aligned}
$$

$$
\begin{aligned}
F_{\text{even}}(n,k,l;x) &= 4n\frac{(1)_{2n}(1)_{2k+l}(1)_{n+k+l}}{(1)_n(1)_k(1)_l(2)_{2n+2k+l}}x^k(1-x)^l, \\
F_{\text{odd}}(n,k,l;x) &= (2n+1)\frac{(1)_{2n}(1)_{2k+l}(\frac{3}{2})_{n+k+l}}{(\frac{1}{2})_n(1)_k(1)_l(3)_{2n+2k+l}}x^k(1-x)^l.
\end{aligned}
$$

In the following, we suppress x in the notation $F(n,k,l;x)$ since it is treated as an indeterminate in hypergeometric summation theory (Apagodu and Zeilberger 2006). Separately for F_{even} and F_{odd}, we follow theorem (mZ) in Apagodu and

Zeilberger (2006) to find an integer L and polynomials $e_0, ..., e_L$ in n, and two rational functions Rk and Rl of (n, k, l), leading to terms

$$\begin{aligned} Gk(n,k,l) &= Rk(n,k,l)F(n,k,l), \\ Gl(n,k,l) &= Rl(n,k,l)F(n,k,l), \end{aligned}$$

such that

$$\begin{aligned} \sum_{i=0}^{L} e_i(n)F(n+i,k,l) = Gk(n,k+1,l) - Gk(n,k,l) \\ + Gl(n,k,l+1) - Gl(n,k,l). \end{aligned}$$

Summation over k and l then yields, due to the telescoping structure,

$$\sum_{i=0}^{L}\sum_{k=0}^{\infty}\sum_{l=0}^{\infty} e_i(n)F(n+i,k,l) + \sum_{l=0}^{\infty} Gk(n,0,l) + \sum_{k=0}^{\infty} Gl(n,k,0) = 0. \qquad (3.52)$$

Applying the Maple package `MultiZeilberger` accompanying Apagodu and Zeilberger (2006) reveals that for F_{even},

$$\begin{aligned} L &= 1, \\ e_0(n) &= -1, & Rk_{\text{even}}(n,k,l) &= \frac{4k}{n}, \\ e_1(n) &= 1, & Rl_{\text{even}}(n,k,l) &= \frac{(8k+4l+4)l}{n(2n+2k+l+2)}, \end{aligned}$$

as shown in Figure 3.6, equation (3), and for F_{odd},

$$\begin{aligned} L &= 1, \\ e_0(n) &= -1, & Rk_{\text{odd}}(n,k,l) &= \frac{4k}{2n+1}, \\ e_1(n) &= 1, & Rl_{\text{odd}}(n,k,l) &= \frac{(8k+4l+4)l}{(2n+1)(2n+2k+l+3)}, \end{aligned}$$

as shown in Figure 3.6, equation (4). This means that the recurrence relation (3.52) reduces to

$$\sum_{k=0}^{\infty}\sum_{l=0}^{\infty} F(n+1,k,l) = \sum_{k=0}^{\infty}\sum_{l=0}^{\infty} F(n,k,l),$$

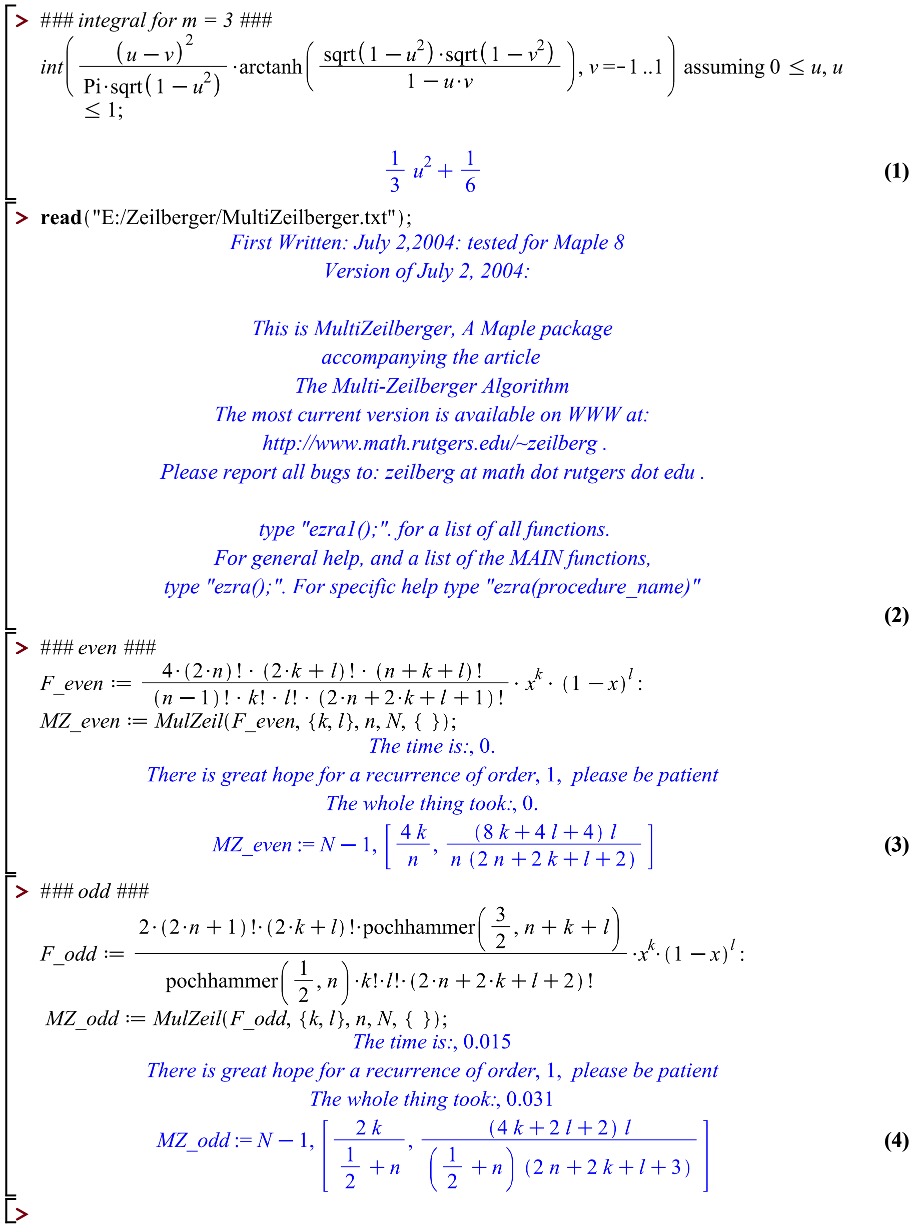
for both F_{even} and F_{odd}. This resulting recurrence relation of order 1 has the desired term $\frac{4}{1+x}$ in (3.51) as a potential solution, leaving us with the task to prove (3.51) for the initial values $m = 2, 3$. Equivalently, we can prove (3.50) for $m = 2$, in that

$$\frac{2r_2^2}{B\left(\frac{1}{2},1\right)} \int_{-1}^{1} 1 - uv - \sqrt{(u-v)^2}\,\mathrm{d}v = r_2^2(2-(1+u^2)) = 1 - \|p\|^2$$

and for $m = 3$, in that

$$\begin{aligned}
&\frac{3r_3^2}{B\left(\frac{1}{2},\frac{3}{2}\right)}\int_{-1}^{1}(1-uv)\sqrt{1-v^2}\left(1-\ {}_2F_1\left(-\tfrac{1}{2},1;\tfrac{3}{2};\tfrac{(1-u^2)(1-v^2)}{(1-uv)^2}\right)\right)\,\mathrm{d}v\\
&\quad= 3r_3^2\int_{-1}^{1}\frac{(1-uv)\sqrt{1-v^2}}{\pi}\left(1-\frac{(u-v)^2\mathrm{arctanh}\left(\frac{\sqrt{1-u^2}\sqrt{1-v^2}}{1-uv}\right)}{(1-uv)\sqrt{1-u^2}\sqrt{1-v^2}}\right)\,\mathrm{d}v\\
&\quad= 3r_3^2\left(\frac{1}{2}-\int_{-1}^{1}\frac{(u-v)^2}{\pi\sqrt{1-u^2}}\mathrm{arctanh}\left(\tfrac{\sqrt{1-u^2}\sqrt{1-v^2}}{1-uv}\right)\,\mathrm{d}v\right)\\
&\quad= 3r_3^2\left(\frac{1}{2}-\frac{1}{6}-\frac{u^2}{3}\right) = 1-\|p\|^2,
\end{aligned}$$

where the integral containing the inverse hyperbolic tangent is calculated by Maple in Figure 3.6, equation (1). □



> *### integral for m = 3 ###*

$$int\left(\frac{(u-v)^2}{\mathrm{Pi}\cdot\mathrm{sqrt}(1-u^2)}\cdot\mathrm{arctanh}\left(\frac{\mathrm{sqrt}(1-u^2)\cdot\mathrm{sqrt}(1-v^2)}{1-u\cdot v}\right), v=-1..1\right) \text{ assuming } 0 \le u, u \le 1;$$

$$\frac{1}{3}u^2+\frac{1}{6} \tag{1}$$

> **read**("E:/Zeilberger/MultiZeilberger.txt");

First Written: July 2,2004: tested for Maple 8
Version of July 2, 2004:

This is MultiZeilberger, A Maple package
accompanying the article
The Multi-Zeilberger Algorithm
The most current version is available on WWW at:
http://www.math.rutgers.edu/~zeilberg .
Please report all bugs to: zeilberg at math dot rutgers dot edu .

type "ezra1();". for a list of all functions.
For general help, and a list of the MAIN functions,
type "ezra();". For specific help type "ezra(procedure_name)"

(2)

> *### even ###*

$$F\_even := \frac{4\cdot(2\cdot n)!\cdot(2\cdot k+l)!\cdot(n+k+l)!}{(n-1)!\cdot k!\cdot l!\cdot(2\cdot n+2\cdot k+l+1)!}\cdot x^k\cdot(1-x)^l:$$

$MZ\_even := MulZeil(F\_even, \{k, l\}, n, N, \{\ \});$

The time is:, 0.
There is great hope for a recurrence of order, 1, *please be patient*
The whole thing took:, 0.

$$MZ\_even := N-1, \left[\frac{4\,k}{n}, \frac{(8\,k+4\,l+4)\,l}{n\,(2\,n+2\,k+l+2)}\right] \tag{3}$$

> *### odd ###*

$$F\_odd := \frac{2\cdot(2\cdot n+1)!\cdot(2\cdot k+l)!\cdot\mathrm{pochhammer}\left(\frac{3}{2}, n+k+l\right)}{\mathrm{pochhammer}\left(\frac{1}{2}, n\right)\cdot k!\cdot l!\cdot(2\cdot n+2\cdot k+l+2)!}\cdot x^k\cdot(1-x)^l:$$

$MZ\_odd := MulZeil(F\_odd, \{k, l\}, n, N, \{\ \});$

The time is:, 0.015
There is great hope for a recurrence of order, 1, *please be patient*
The whole thing took:, 0.031

$$MZ\_odd := N-1, \left[\frac{2\,k}{\frac{1}{2}+n}, \frac{(4\,k+2\,l+2)\,l}{\left(\frac{1}{2}+n\right)(2\,n+2\,k+l+3)}\right] \tag{4}$$

>

Figure 3.6: Maple code and results in support of the proof of Proposition 3.4.

3.2.2 Real-valued events

In the previous section, we investigated entropy functions on the unit simplex in $\mathbb{R}^m$, where the extremal members lie dense in the respective class, once the number of categories m increases beyond two. Presumably, a similar result will also hold for the concave functions on the class of cumulative distribution functions. However, investigations in this direction are beyond the scope of this dissertation. Instead, we revisit the earlier results for consistent scoring functions and combine them to obtain mixture representations for classes of proper scoring rules for real-valued events.

Let $\mathcal{T}$ be a set of elicitable functionals, then for any $\mathrm{T} \in \mathcal{T}$ we can find a class $\mathcal{S}^{\mathrm{T}}$ of consistent scoring functions S^{T}, which in turn induce proper scoring rules. Clearly these can be part of a mixture, thus generating a class of proper scoring rules of the form

$$\mathrm{S}(F, y) = \int_{\mathcal{T}} \mathrm{S}^{\mathrm{T}}(\mathrm{T}(F), y) \, \mathrm{d}\mu(\mathrm{T}),$$

where μ is a σ-finite measure on an appropriate measurable space generated by $\mathcal{T}$, and each S^{T} enjoys its own mixture representation in its respective class of consistent scoring functions. Dealing with real-valued events, we usually assume that the class $\mathcal{T}$ can be parameterized by $\Theta \subseteq \mathbb{R}^n$, and that μ is defined on the corresponding Borel σ-algebra.

Example 3.3 (continuous ranked probability score). The quadratic score for categorical events derived from the entropy function (3.35) is typically defined as the sum of Brier scores over all categories,

$$\mathrm{S}(p, y) = \sum_{i=1}^{m} (p_i - \mathbb{1}(y = i))^2.$$

In the construction, Brier (1950) summed over scoring functions that are strictly consistent for the functionals T_i that map onto the expected value of the binary event $\mathbb{1}(y = i)$, i.e. the probability of $y = i$. Due to summation over all categories, we get a strictly proper scoring rule, but summation may be restricted to a subset without losing anything but the strictness of propriety.

In the late 1960s, meteorologists were searching for proper scoring rules on a sample space of ordered categories that are sensitive to distance, i.e. when category 1 is observed then a forecast assigning high probability to category 2 is better than one with a high probability on category 3. The quadratic score, also known as probability score in the meteorological literature, clearly does not satisfy this property. Their solution was the ranked probability score (Epstein 1969; Murphy 1970) which can be given in the form

$$\mathrm{RPS}(p, y) = \sum_{i=1}^{m} (P(X \leq i) - \mathbb{1}(y \leq i))^2,$$

where X is a random variable taking values in $\{1, \ldots, m\}$ with distribution P corresponding to the probability mass function p. Here, the functionals map onto the expected value of the binary event $\mathbb{1}(y \leq i)$.

Shortly after the ranked probability score's introduction, the continuous version (Matheson and Winkler 1976),

$$\mathrm{CRPS}(F,y) = \int_{-\infty}^{\infty} (F(\theta) - \mathbb{1}(y \leq \theta))^2 \,\mathrm{d}\theta, \tag{3.53}$$

for CDF-valued forecasts was discovered. This form of the continuous ranked probability score is now known as the threshold decomposition of the CRPS, which equals the integral over all $\theta \in \mathbb{R}$ of the Brier score (2.13, 3.13) for exceedence probabilities with respect to an event threshold. We may then invoke the mixture representation (3.11) along with the relationships (3.12) and (3.14) to yield

$$\begin{aligned} \mathrm{CRPS}(F,y) &= 2\int_{-\infty}^{\infty}\int_0^1 \mathrm{S}_\alpha^{\mathrm{B}}(F(\theta), \mathbb{1}(\theta \geq y)) \,\mathrm{d}\alpha \,\mathrm{d}\theta \\ &= 2\int_{-\infty}^{\infty}\int_0^1 \mathrm{S}_{\alpha,\theta}^{\mathrm{Q}}(q_{\alpha,F}, y) \,\mathrm{d}\alpha \,\mathrm{d}\theta. \end{aligned} \tag{3.54}$$

Changing the order of integration, we recover the representation by Laio and Tamea (2007),

$$\mathrm{CRPS}(F,y) = 2\int_0^1 (\mathbb{1}(y < q_{\alpha,F}) - \alpha)(q_{\alpha,F} - y) \,\mathrm{d}\alpha, \tag{3.55}$$

which is commonly known as the quantile decomposition. This is an integral over all $\alpha \in (0,1)$ of the asymmetric piecewise linear score (3.3), a strictly consistent scoring function for the α-quantile. Gneiting and Ranjan (2011) investigate the possibility of inserting a threshold-depending weighting function $u(\theta)$ into the integral in (3.53), or inserting a probability-depending weighting function $w(\alpha)$ into the integral in (3.55), for increased flexibility in combining the strictly consistent scoring functions. Our representations given by (3.54) allow weighting using a function $u(\alpha, \theta)$ for a general family of proper scoring rules that can be economically motivated and justified. Related ideas have recently been put forward in the hydrologic and meteorological literatures (Laio and Tamea 2007; Bradley and Schwartz 2011; Smet et al. 2012).

3.3 Discussion

We have studied mixture representations for the scoring functions that are consistent for quantiles and expectiles, including the ubiquitous case of the mean or expectation functional, and nesting probability forecasts for binary events as a further special case. These results reappear in the mixture representation for scoring functions that are consistent for the multivariate functional of value-at-risk and expected shortfall. A particularly interesting aspect of these results is that they allow an economic interpretation of consistent scoring functions in terms of betting and investment problems.

The developed interpretations can help design economically or societally relevant criteria in more general settings. For example, the elementary expectile score $\mathrm{S}^{\mathrm{E}}_{\alpha,\theta}(x, y)$ in (3.9) depends on x and y via the absolute deviation between the event threshold θ and the observation y only, and therefore might be interpreted in terms of the original unit in any applied problem. Owing to the mixture representation (3.8), any consistent scoring function $\mathrm{S}^{\mathrm{E}}_{\alpha}$ can be associated with a weighting of thresholds, as encoded by the mixing measure H. If $H(\theta) = \theta^n$ for odd $n \in \mathbb{N}$, then the scoring function $\mathrm{S}^{\mathrm{E}}_{\alpha}$ is denoted in the same unit as y^{n+1}. The choice of the mixing measure requires careful consideration of the decision problem at hand, and it seems hard to provide general guidance. As noted, squared error corresponds to Lebesgue measure. In applications, non-uniform measures with finite mass may provide more realistic descriptions.

Our results also bear on estimation problems, in that scoring functions connect naturally to M-estimation (Huber 1964; Koltchinskii 1997). An interesting observation is that the loss functions that have traditionally been employed for estimation in quantile regression, ordinary least squares regression, and expectile regression, namely the asymmetric piecewise linear and squared error scoring functions (3.3) and (3.5), correspond to the choice of the Lebesgue measure in the mixture representations (3.6) and (3.8), respectively. This is in contrast to binary regression, where estimation is typically based on the logarithmic score, which corresponds to the choice of the infinite measure with density $h(\theta) = (\theta(1-\theta))^{-1}$ in the mixture representation (3.11), rather than the Lebesgue or uniform measure that yields (half) the Brier score (3.13). Quite generally, this raises the question of the optimal choice of the loss or scoring function to be used for estimation in regression problems. Focusing on the binary case, Hand and Vinciotti (2003), Buja et al. (2005), Lieli and Springborn (2013) and Elliott et al. (2016) have considered the use of economically justifiable criteria.

Mixture representations of Choquet type can be found for other, more general classes of consistent scoring functions. For instance, our results extend to the class of functionals known as generalized quantiles or M-quantiles (Breckling and Chambers 1988; Koltchinskii 1997; Bellini et al. 2014; Steinwart et al. 2014), which subsume both quantiles and expectiles. Related, but more complex mixture representations apply in the case of scoring functions that are consistent for multi-dimensional functionals, as recently studied by Fissler and Ziegel (2016) and illustrated in Section 3.1.2.

In the context of categorical events, we have illustrated the difficulties in deriving mixture representations for proper scoring rules. The elementary scores lie dense in the class of proper scoring rules and the parameter spaces of meaningful subclasses will typically be multi-dimensional. Regular directional proper scoring rules for probabilistic forecasts of an observation with m categories are presumably generated by elementary scores which are parameterized by a collection of $m + 1$ parameters, one for the apex and m parameters for directions, which are themselves m-dimensional. The corresponding surface integrals can be painfully difficult to calculate, as illustrated in Section 3.2.1. For predictive distributions of real-valued variables this complexity increases further. Still, we can construct families of meaningful proper scoring rules by combination of consistent scoring functions as outlined in Section 3.2.2. The family of weighted versions of the CRPS determined by a non-negative mixing density $u(\alpha, \theta)$ in

$$\int_{-\infty}^{\infty} \int_{0}^{1} u(\alpha, \theta)\, \mathrm{S}^{\mathrm{Q}}_{\alpha,\theta}(q_{\alpha,F}, y)\, \mathrm{d}\alpha\, \mathrm{d}\theta,$$

where $q_{\alpha,F}$ denotes an α-quantile of the distribution F, could analogously be constructed using the elementary scores $\mathrm{S}^{\mathrm{E}}_{\alpha,\theta}$. As a special case, further research might consider looking into the continuous expectile score (CES), the CRPS' analog using the asymmetric squared error (3.5),

$$\mathrm{S}(F, y) = 2 \int_{0}^{1} (\mathbb{1}(y < \tau_{\alpha,F}) - \alpha)(\tau_{\alpha,F} - y)^2\, \mathrm{d}\alpha,$$

where $\tau_{\alpha,F}$ denotes the α-expectile of the distribution F. The CRPS is known for a certain robustness, or insensitivity with respect to spread. In this regard, one would expect the CES to have increased sensitivity in comparison. A caveat remains in that closed form expressions should be even more difficult to find than they are for the CRPS, which is already known to be elusive (see Chapter 6).

4 Murphy Diagrams

> *"What happens to the forecast ranking if I change the performance measure?"*

The mixture representations developed in Chapter 3 encourage questions regarding the choice of performance measure. Slight changes in the mixing measure can potentially lead to different conclusions, and yet, sometimes the choice is completely irrelevant. Fortunately, the identification of elementary scores facilitates considerations in this direction. It is sufficient only to evaluate the rankings induced by this reduced subclass of performance measures, which may still be large. However, comparisons via linearly parameterized families of elementary scores can be efficiently summarized in a plot, a tool we call *Murphy diagram.* Especially in empirical examples, they provide reliable first impressions due to the support's boundedness. One glance can confirm the presence or absence of dominance relations.

4.1 Forecast dominance

We now define notions of forecast dominance, starting with probabilistic forecasts that take the form of predictive CDFs, and then turning to point forecasts.

Definition 4.1a (dominance). *Let F_1 and F_2 be probabilistic forecasts, and let Y be the outcome, in a prediction space. Then F_1 dominates F_2 relative to a class $\mathcal{P}$ of proper scoring rules if $\mathbb{E}_{\mathbb{Q}}\mathrm{S}(F_1, Y) \leq \mathbb{E}_{\mathbb{Q}}\mathrm{S}(F_2, Y)$ for every $\mathrm{S} \in \mathcal{P}$.*

As outlined in (2.12), a scoring function S that is consistent for a single-valued functional T relative to a class $\mathcal{F}$ induces a proper scoring rule.

Definition 4.1b (dominance). *Let X_1 and X_2 be point forecasts, and let Y be the outcome, in a point prediction space. Then X_1 dominates X_2 relative to a class $\mathcal{S}^{\mathrm{T}}$ of scoring functions that are consistent for T if $\mathbb{E}_{\mathbb{Q}}\mathrm{S}^{\mathrm{T}}(X_1, Y) \leq \mathbb{E}_{\mathbb{Q}}\mathrm{S}^{\mathrm{T}}(X_2, Y)$ for every $\mathrm{S}^{\mathrm{T}} \in \mathcal{S}^{\mathrm{T}}$.*

It is important to note that the expectations in the definitions are taken with respect to the joint distribution of the probabilistic forecasts and the outcome. The dominance notions provide partial orderings for the predictive distributions $F_1, \ldots, F_k$, or point predictions $X_1, \ldots, X_k$, in (2.1) respectively.[1] Essentially, a

[1] In the special case of probability forecasts of a binary event, related notions of sufficiency and dominance have been studied by DeGroot and Fienberg (1983), Vardeman and Meeden (1983), Schervish (1989), Feuerverger and Rahman (1992), Krämer (2005), and Bröcker (2009).

probabilistic forecast that dominates another is preferable, or at least not inferior, in any type of decision that involves the respective predictive distributions.[2] In the case of the functional T, a point forecast that dominates another is preferable, or at least not inferior, in any type of decision problem that depends on the respective predictive distributions via the considered functional only.

Under which conditions does a forecast dominate another? Holzmann and Eulert (2014) recently showed that if two predictive distributions are ideal, then the one with the richer information set dominates the other. Furthermore, the result carries over to ideal forecasters' induced point predictions. To give an example in the setting of Table 2.1, the perfect and the climatological forecasters are ideal relative to the sigma fields generated by μ, and generated by the empty set, respectively. Therefore, the perfect forecaster dominates the climatological forecaster, in any of the above senses.

Tsyplakov (2014) went on to show that if a predictive distribution is ideal relative to a certain information set, then it dominates any predictive distribution that is measurable with respect to the information set. Again, the result carries over to the induced point forecasts. In the setting of Table 2.1, the perfect forecaster is ideal relative to the sigma field generated by the random variables μ and τ. The climatological, unfocused, and sign-reversed forecasters are measurable with respect to this sigma field, and so they are dominated by the perfect forecaster, in any of the above senses.

Sometimes order sensitivity can be invoked to prove dominance. For example, consider a mixed prediction space setting with tuples (F, X_1, X_2, Y). Suppose that the CDF-valued random quantity F is ideal relative to the sigma field $\mathcal{A}$, and that X_1 and X_2 are measurable with respect to $\mathcal{A}$. For a single-valued, univariate functional T, the forecast X_1 dominates X_2 as a T forecast if with probability one either

$$X_2 \leq X_1 \leq \mathrm{T}(F) \qquad \text{or} \qquad \mathrm{T}(F) \leq X_1 \leq X_2$$

holds true. By Corollary 4.1a and 4.1b in concert with Proposition 3.2a and 3.2b and a conditioning argument, this argument applies in the case of α-quantiles and α-expectiles. In the scenario of Table 2.1, the argument can be put to work in the case $\alpha = 1/2$ that corresponds to median and mean forecasts, respectively. Specifically, let F be the perfect forecast, which has median and mean μ, let $\mathcal{A}$ be the sigma field generated by μ, and let $X_1 = 0$ and $X_2 = -\mu$. Invoking the order sensitivity argument, we see that the climatological forecaster dominates the sign-reversed forecaster for both median and mean predictions.

In the practice of forecasting, predictive distributions are hardly ever ideal, and information sets may not be nested, as emphasized by Patton (2015). Therefore, the above theoretical results are not readily applicable, and distinct soring rules, or distinct consistent scoring functions, may yield distinct forecast rankings, as in

[2] To see this, note that any utility function induces a proper scoring rule via the respective Bayes act. For details, see Section 3 of Dawid (2007) and Section 2.2 of Gneiting and Raftery (2007).

empirical examples given by Schervish (1989), Merkle and Steyvers (2013), and Patton (2015), among others. Furthermore, in general it is not feasible to check the validity of the expectation inequalities in Definitions 4.1a and 4.1b for every proper scoring rule $\mathrm{S} \in \mathcal{P}$, or consistent scoring function $\mathrm{S}^{\mathrm{T}} \in \mathcal{S}^{\mathrm{T}}$.

Fortunately, in the case of quantile and expectile forecasts, the mixture representations in Theorems 3.1a and 3.1b reduce checks for dominance to the respective one-dimensional families of elementary scoring functions.

Corollary 4.1a (dominance - quantiles). *In a point prediction space, X_1 dominates X_2 relative to the class $\mathcal{S}_\alpha^{\mathrm{Q}}$ if*

$$\mathbb{E}_{\mathbb{Q}} \mathrm{S}_{\alpha,\theta}^{\mathrm{Q}}(X_1, Y) \le \mathbb{E}_{\mathbb{Q}} \mathrm{S}_{\alpha,\theta}^{\mathrm{Q}}(X_2, Y)$$

for every $\theta \in \mathbb{R}$.

Corollary 4.1b (dominance - expectiles). *In a point prediction space, X_1 dominates X_2 relative to the class $\mathcal{S}_\alpha^{\mathrm{E}}$ if*

$$\mathbb{E}_{\mathbb{Q}}\, \mathrm{S}_{\alpha,\theta}^{\mathrm{E}}(X_1, Y) \le \mathbb{E}_{\mathbb{Q}}\, \mathrm{S}_{\alpha,\theta}^{\mathrm{E}}(X_2, Y)$$

for every $\theta \in \mathbb{R}$.

4.2 Diagnostic tools

The reduction to a one-dimensional problem suggests graphical comparisons of quantile and expectile forecasts via Murphy diagrams, including the special cases of the mean or expectation functional, and the further special case of probability forecasts of a binary event. We describe these diagnostic tools in the setting of a point prediction space, where $X_1, \ldots, X_k$ denote point forecasts for the outcome Y, and the probability measure $\mathbb{Q}$ represents their joint distribution. In the case of probability forecasts, we use the more suggestive notation $p_1, \ldots, p_k$ for the forecasts.

- For quantile forecasts at level $\alpha \in (0,1)$, we plot the graph of the expected elementary quantile score $\mathrm{S}_{\alpha,\theta}^{\mathrm{Q}}$,

$$\begin{aligned} \theta \mapsto s_j(\theta) &= \mathbb{E}_{\mathbb{Q}}\, \mathrm{S}_{\alpha,\theta}^{\mathrm{Q}}(X_j, Y) \\ &= (1-\alpha)\mathbb{Q}(Y \le \theta) + \alpha\mathbb{Q}(X_j \le \theta) - \mathbb{Q}(X_j \le \theta, Y \le \theta), \end{aligned} \tag{4.1}$$

  for $j = 1, \ldots, k$. By Corollary 4.1a, the forecast X_i dominates X_j if and only if $s_i(\theta) \le s_j(\theta)$ for every $\theta \in \mathbb{R}$. The area under $s_j(\theta)$ equals the respective expected asymmetric piecewise linear score (3.3).

- For expectile forecasts at level $\alpha \in (0,1)$, we plot the graph of the expected elementary expectile score $\mathrm{S}_{\alpha,\theta}^{\mathrm{E}}$,

$$\begin{aligned} \theta \mapsto s_j(\theta) &= \mathbb{E}_{\mathbb{Q}}\, \mathrm{S}_{\alpha,\theta}^{\mathrm{E}}(X_j, Y) \\ &= (1-\alpha)\, \mathbb{E}_{\mathbb{Q}}\, \mathbb{1}\{Y \le \theta\}(\theta - Y) - \alpha\, \mathbb{E}_{\mathbb{Q}}\, \mathbb{1}\{X_j \le \theta\}(\theta - Y) \\ &\quad - (1-2\alpha)\, \mathbb{E}_{\mathbb{Q}}\, \mathbb{1}\{Y \le \theta, X_j \le \theta\}(\theta - Y), \end{aligned} \tag{4.2}$$

for $i = 1, \ldots, k$. By Corollary 4.1b, the forecast X_i dominates X_j if and only if $s_i(\theta) \le s_j(\theta)$ for every $\theta \in \mathbb{R}$. The area under $s_j(\theta)$ equals half the respective expected asymmetric squared error (3.5).

- For probability forecasts of a binary event, we plot the graph of the expected elementary score $\mathrm{S}_\theta^{\mathrm{B}}$,

$$\begin{aligned} \theta \mapsto s_j(\theta) &= \mathbb{E}_{\mathbb{Q}}\, \mathrm{S}_\theta^{\mathrm{B}}(p_j, Y) \\ &= \theta \mathbb{Q}(Y = 0) + (1 - \theta)\mathbb{Q}(p_j \le \theta) - \mathbb{Q}(Y = 0, p_j \le \theta) \end{aligned} \tag{4.3}$$

  for $i = 1, \ldots, k$. By Corollary 4.1b, the probability forecast p_i dominates p_j if and only if $s_i(\theta) \le s_j(\theta)$ for every $\theta \in (0, 1)$. The area under $s_j(\theta)$ equals half the expected Brier score (3.13).

In the context of probability forecasts for binary weather events, displays of this type have a rich tradition that can be traced to Thompson and Brier (1955) and Murphy (1977). More recent examples include the papers by Schervish (1989), Feuerverger and Rahman (1992), Richardson (2000), Wilks (2001), Mylne (2002), and Berrocal et al. (2010), among many others. Murphy (1977) distinguished three kinds of diagrams that reflect the economic decisions involved. The negatively oriented *expense diagram* shows the mean raw loss or expense of a given forecast scheme; the positively oriented *value diagram* takes the unconditional or climatological forecast as reference and plots the difference in expense between this reference forecast and the forecast at hand, and lastly, the *relative-value diagram* plots the ratio of the utility of a given forecast and the utility of an oracle forecast. The displays introduced above are similar to the value diagrams of Murphy, and we refer to them as *Murphy diagrams.* Our Murphy diagrams are by default negatively oriented and plot the expected elementary score for competing forecasters. If interest focuses on binary comparisons, it is natural to consider Murphy diagrams for the difference,

$$\theta \mapsto D(\theta) = \mathbb{E}_{\mathbb{Q}}\mathrm{S}_\theta(X_1, Y) - \mathbb{E}_{\mathbb{Q}}\mathrm{S}_\theta(X_2, Y), \tag{4.4}$$

between the expected elementary scores of two point forecasters. For better visual appearance, we generally connect the left- and right-hand limits at the jump points of the empirical score curves.

Example 4.1 (climatology). If the forecast is constant, i.e. $\mathbb{Q}(X = x_0) = 1$ or $\mathbb{Q}(p = p_0) = 1$, we get the simpler forms,

$$s^Q(\theta) = \begin{cases} (1 - \alpha)\mathbb{Q}(Y \le \theta), & x_0 > \theta, \\ \alpha \mathbb{Q}(Y > \theta), & x_0 \le \theta, \end{cases}$$

$$s^E(\theta) = \begin{cases} (1 - \alpha)\mathbb{E}_{\mathbb{Q}}\, \mathbb{1}\{Y \le \theta\}(\theta - Y), & x_0 > \theta, \\ \alpha\, \mathbb{E}_{\mathbb{Q}}\, \mathbb{1}\{Y > \theta\}(Y - \theta), & x_0 \le \theta, \end{cases}$$

$$s^B(\theta) = \begin{cases} \theta \mathbb{Q}(Y = 0), & p_0 > \theta, \\ (1 - \theta)\mathbb{Q}(Y = 1), & p_0 \le \theta, \end{cases}$$

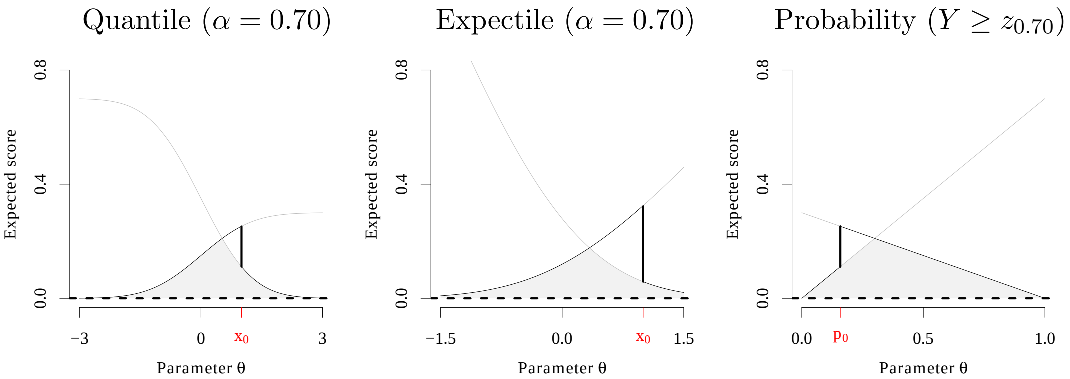


Figure 4.1: Murphy diagrams for constant quantile or expectile forecasts $x_0 = 1$, and a constant probability forecast $p_0 = 1 - \Phi(x_0)$, when $Y \sim \mathcal{N}(0, 1)$. The area under the Murphy diagram of the optimal forecast is colored in grey.

which are minimized pointwise in θ when x_0 equals the climatological α-quantile or α-expectile, respectively, or when p_0 equals the climatological probability $\mathbb{Q}(Y = 1)$. This is illustrated in Figure 4.1. Note that the climatological forecast is ideal with respect to the minimal information set, i.e. the σ-algebra $\{\emptyset, \Omega\}$. Then, if the Murphy diagram of a forecast is below that of the climatological forecaster for any θ, this forecast must be based on a superior information set. If, additionally, this forecast's Murphy diagram is above that of the climatological forecaster for any θ, then the forecast cannot be ideal because an ideal forecast dominates all other forecasts based on smaller information sets.

Example 4.2. Figure 4.2 shows Murphy diagrams for the perfect, climatological, unfocused (where $\tau = \pm 2$ with equal probability), and sign-reversed forecasters in Table 2.1. We compare point predictions for the mean or expectation functional, and the quantile at level $\alpha = 0.90$, along with probability forecasts for the binary event that the outcome exceeds the threshold value 2. Analytic expressions for the respective expected scores are given in Table 4.1, which in view of the relationships (3.12) and (3.14) implicitly covers the case for event probabilities, too. As proved in the previous section, the perfect forecaster dominates the other forecasters for all functionals considered. The expected score curves for the climatological and the unfocused, and for the unfocused and the sign-reversed forecasters, intersect in all three cases, so there are no order relations between these forecasters. Finally, the Murphy diagrams suggest that the climatological forecaster dominates the sign-reversed forecaster for all three functionals, and in the case of the mean functional, the order sensitivity argument in the previous section confirms the visual impression. In the cases of the quantile and probability forecasts, final confirmation would need to be based on tedious analytic investigations of the asymptotic behavior of the expected score functions.

Forecast	α-Quantile	Mean
F	$\mathbb{E}_{\mathbb{Q}}\, \mathrm{S}^{\mathrm{Q}}_{\alpha,\theta}(q_{\alpha,F},Y)$	$2\,\mathbb{E}_{\mathbb{Q}}\, \mathrm{S}^{\mathrm{E}}_{1/2,\theta}(\mu_F,Y)$
Perfect	$a_{\alpha,\theta} + \alpha\Phi(\theta - z_\alpha) + \int_{\theta-z_\alpha}^{\infty} A_\theta(x)\,\mathrm{d}x$	$c_\theta - \theta\Phi(\theta) - \varphi(\theta)$
Climatological	$a_{\alpha,\theta} + \min\Big(\Phi(\frac{\theta}{\sqrt{2}}),\alpha\Big)$	$c_\theta - \theta\mathbb{1}(\theta \geq 0)$
Unfocused	$a_{\alpha,\theta} + \mathbb{E}_\tau\Big[\alpha\Phi(\theta - z_{\alpha,\tau}) + \int_{\theta-z_{\alpha,\tau}}^{\infty} A_\theta(x)\,\mathrm{d}x\Big]$	$c_\theta - \mathbb{E}_\tau\big[\theta\Phi\big(\theta - \frac{\tau}{2}\big) + \varphi\big(\theta - \frac{\tau}{2}\big)\big]$
Sign-reversed	$a_{\alpha,\theta} + \alpha\Phi(\theta - z_\alpha) + \int_{-\infty}^{z_\alpha-\theta} A_\theta(x)\,\mathrm{d}x$	$c_\theta - \theta\Phi(\theta) + \varphi(\theta)$

Table 4.1: Expected extremal scores in the prediction space example of Table 2.1. For $\alpha \in (0,1)$ and $\theta \in \mathbb{R}$, we let $a_{\alpha,\theta} = -\alpha\Phi(\theta/\sqrt{2})$, $A_\theta(x) = \Phi(\theta - x)\varphi(x)$, and $c_\theta = \theta\Phi(\theta/\sqrt{2}) + \sqrt{2}\varphi(\theta/\sqrt{2})$, where Φ and φ denote the CDF and the probability density function of the standard normal distribution, respectively.

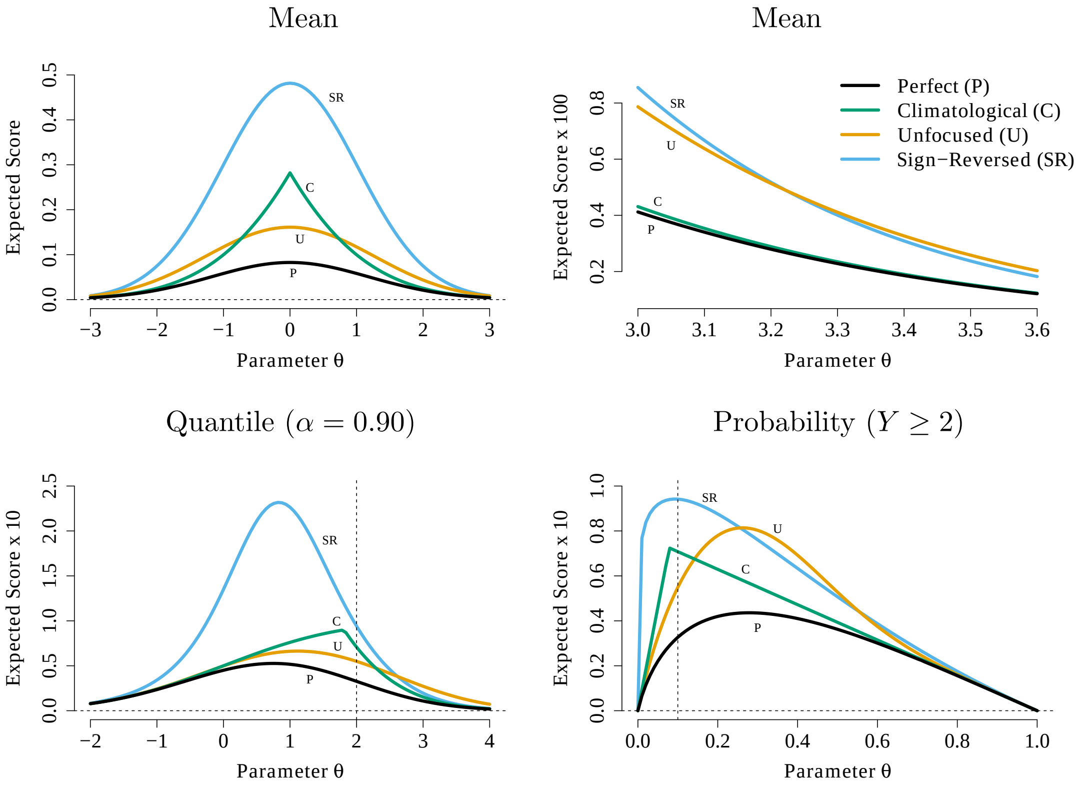


Figure 4.2: Murphy diagrams for the forecasters in Table 2.1 where $\tau = \pm 2$ with equal probability. The functionals considered are the mean, the quantile at level $\alpha = 0.90$, and the probability of the binary event $Y \geq 2$. The vertical dashed lines in the bottom panels indicate the extremal scores $\mathrm{S}^{\mathrm{Q}}_{\alpha,\theta}$ and $\mathrm{S}^{\mathrm{B}}_{\theta}$ that relate to each other as in (3.12) and (3.14).

4.3 Empirical forecasts

We now turn to the comparison and ranking of empirical forecasts. Specifically, we consider tuples

$$(x_{i1}, \ldots, x_{ik}, y_i), \qquad i = 1, \ldots, n, \tag{4.5}$$

where $x_{1j}, \ldots, x_{nj}$ are the jth forecaster's point predictions, for $j = 1, \ldots, k$, and $y_i, \ldots, y_n$, are the respective outcomes. Thus, we have k competing forecasters, and each of them issues a set of n point predictions. A convenient interpretation of the empirical setting is as a special case of a point prediction space, in which the tuples $(X_1, \ldots, X_k, Y)$ in (2.1) attain each of the values in (4.5) with probability $1/n$. Then the probability measure $\mathbb{Q}$ is the corresponding empirical measure, and with this identification, the (average) empirical scores

$$s_j(\theta) = \frac{1}{n} \sum_{i=1}^{n} \mathrm{S}_\theta(x_{ij}, y_i),$$

where S_θ is either $\mathrm{S}^{\mathrm{Q}}_{\alpha,\theta}$, $\mathrm{S}^{\mathrm{E}}_{\alpha,\theta}$, or $\mathrm{S}^{\mathrm{B}}_{\theta}$, become the expected elementary scores from (4.1), (4.2), and (4.3), respectively. Accordingly, we say that forecaster X_1 *empirically dominates* forecaster X_2 if $s_1(\theta) \leq s_2(\theta)$ for all $\theta \in \mathbb{R}$. When comparing the two forecasters X_1 and X_2, it is convenient to show a Murphy plot of the equivalent of the difference (4.4), namely

$$\theta \mapsto D_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} d_i(\theta),$$

where

$$d_i(\theta) = \mathrm{S}_\theta(x_{i1}, y_i) - \mathrm{S}_\theta(x_{i2}, y_i) \tag{4.6}$$

for $i = 1, \ldots, n$, and again S_θ is either $\mathrm{S}^{\mathrm{Q}}_{\alpha,\theta}$, $\mathrm{S}^{\mathrm{E}}_{\alpha,\theta}$, or $\mathrm{S}^{\mathrm{B}}_{\theta}$, respectively.

4.3.1 Empirical dominance

Murphy diagrams can be used efficiently to show a lack of domination when forecasters' expected elementary score curves intersect. However, in general it is not possible to conclude domination, unless the visual impression is supported by tedious analytic investigations of the behavior of the expected score functions as $\theta \to \pm\infty$. Fortunately, these complications do not arise in the empirical case, where dominance can be established by comparing the empirical score functions at a well-defined, finite set of arguments only, as follows.

Corollary 4.2a (empirical dominance - quantiles). *The forecast X_1 empirically dominates X_2 for α-quantile predictions if*

$$\frac{1}{n} \sum_{i=1}^{n} \mathrm{S}^{\mathrm{Q}}_{\alpha,\theta}(x_{1i}, y_i) \leq \frac{1}{n} \sum_{i=1}^{n} \mathrm{S}^{\mathrm{Q}}_{\alpha,\theta}(x_{2i}, y_i)$$

for $\theta \in \{x_{11}, x_{12}, y_1, \ldots, x_{n1}, x_{n2}, y_n\}$.

Corollary 4.2b (empirical dominance - expectiles). *The forecast X_1 empirically dominates X_2 for α-expectile predictions if*

$$\frac{1}{n}\sum_{i=1}^{n} \mathrm{S}^{\mathrm{E}}_{\alpha,\theta}(x_{1i}, y_i) \leq \frac{1}{n}\sum_{i=1}^{n} \mathrm{S}^{\mathrm{E}}_{\alpha,\theta}(x_{2i}, y_i)$$

for $\theta \in \{x_{11}, x_{12}, y_1, \ldots, x_{n1}, x_{n2}, y_n\}$ and in the left-hand limit as $\theta \uparrow \theta_0 \in \{x_{11}, x_{12}, \ \ldots, x_{n1}, x_{n2}\}$. In the case $\alpha = 1/2$ evaluations at $\theta \in \{y_1, \ldots, y_n\}$ can be omitted.

To see why these results hold, note that in either case the score differential $d_i(\theta)$ is right-continuous, and that it vanishes unless $\min(x_{i1}, x_{i2}) \leq \theta < \max(x_{i1}, x_{i2})$. Furthermore, in the case of quantiles $d_i(\theta)$ is piecewise constant with no other jump points than x_{i1}, x_{i2}, or y_i. Similarly, in the case of expectiles $d_i(\theta)$ is piecewise linear with no other jump points than x_{i1} and x_{i2}, and no other change of slope than at y_i. The change of slope disappears when $\alpha = 1/2$. Figure 4.3 illustrates the behavior of $d_i(\theta)$ in the cases of the median and the mean, respectively.

To give an example, we consider the 10 forecasters in Table A.1 of Merkle and Steyvers (2013), each of whom issues probability forecasts for 21 binary events. The data are artificial but mimic forecasters in the Aggregate Contingent Estimation System (ACES), a web based survey that solicited probability forecasts for world events from the general public. The Murphy diagram in the left-hand panel of Figure 4.4 shows the empirical score curves

$$\theta \mapsto s_j(\theta) = \frac{1}{21}\sum_{i=1}^{21} \mathrm{S}^{\mathrm{B}}_{\theta}(p_{ij}, y_i),$$

where $p_{ij} \in [0, 1]$ is forecaster j's stated probability for world event i to materialize, and $y_i \in \{0, 1\}$ is the respective binary realization. By Corollary 4.2b, dominance relations can be inferred by evaluating $s_j(\theta)$ at the forecasters' stated probabilities. We note that ID 3 empirically dominates IDs 6 and 8, and that ID 5 empirically dominates ID 10. The remaining pairwise comparisons do not give rise to dominance relations. The induced partial order between the IDs applies to comparisons under any proper scoring rule, as reflected by the rankings in Table 1 of Merkle and Steyvers (2013). It can also be represented in the form of a directed graph, which we call a dominance graph, as illustrated in Figure 4.5. It is the forecast IDs at the top which are not dominated. In big surveys with possibly redundant forecasts, dominance graphs may become much more complex and much more informative. They give rise to simple pruning algorithms for the identification of forecasts that are to be combined or considered further. In the simplest case, one might restrict attention to the subset of the forecasts which are dominated by at most two other forecasters. The right-hand panel in Figure 4.4 considers joint comparisons. We see that ID 3 attains the lowest score over a wide range of θ. However, IDs 2, 5, 7, and 9 show the unique best empirical score under $\mathrm{S}^{\mathrm{B}}_{\theta}$ for other values of θ and, therefore, have superior economic utility under the associated cost-loss ratios.

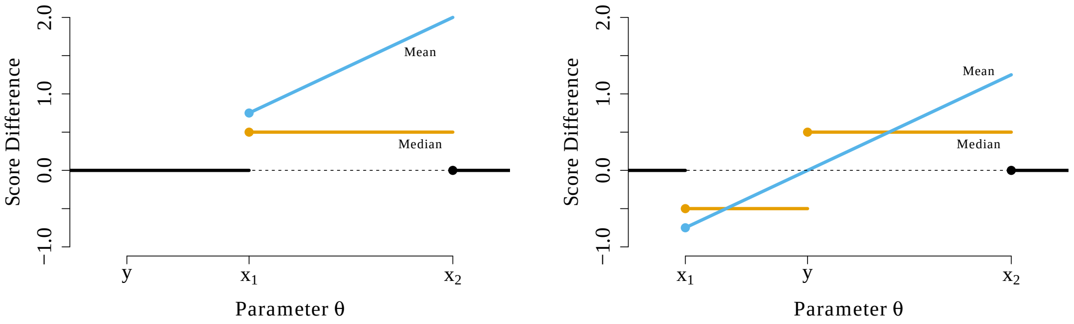


Figure 4.3: The general shape of the score differential $d_i(\theta)$ in (4.6) for the median and mean functionals.

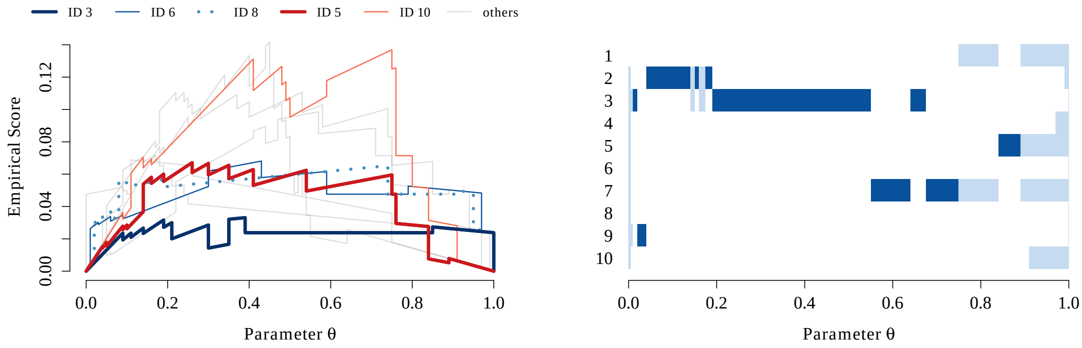


Figure 4.4: Left: Murphy diagram for the probability forecasters in Table A.1 of Merkle and Steyvers (2013). Right: The best forecast ID(s) under $\mathrm{S}^{\mathrm{B}}_{\theta}$, with dark blue indicating a unique best score, and light blue a shared best score. For example, ID 9 attains the unique best score for $\theta \in [0.02, 0.04)$, and ID 10 attains the shared best score for $\theta \in [0.91, 1)$.

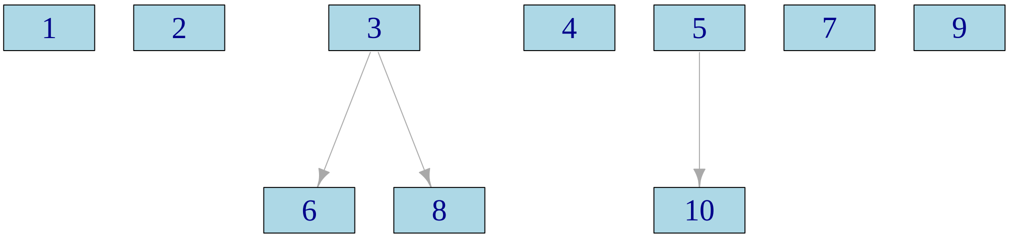


Figure 4.5: Empirical dominance graph in the synthetic example of Merkle and Steyvers (2013).

4.3.2 Stability of forecast rankings

Forecast dominance is a strong property and may not often be encountered in empirical examples. Holzmann and Klar (2016) and Ventura and Nugent (2016) note that sometimes it is clear or obvious which of two forecasters is the better one even though dominance does not hold. This encourages the search for a type of measure that determines the effort necessary to change a given forecast ranking. If it cannot be changed we have a dominance relation.

In the remainder of this section, we consider quantile or expectile forecasts and the corresponding mixture representations (3.6) and (3.8) for the respective consistent scoring functions. However, the ideas in Definitions 4.2 and 4.3, and the result in Proposition 4.1, hold in general.

Definition 4.2 (forecast ranking). *Let $s_1(\theta), \dots, s_n(\theta)$ be the empirical scores of n forecasts and let H be a mixing measure for the extremal scores. The measure H induces a forecast ranking by the order of the set of integrated scores*

$$\int_{\mathbb{R}} s_i(\theta)\,\mathrm{d}H(\theta), \quad i = 1, \dots, n. \tag{4.7}$$

Given a forecast ranking induced by some default mixing measure H_d we are interested in mixing measures that change the ranking, specifically in the one that is closest to the default measure in some sense. With focus on the ranking rather than the scores, we have no interest in the set of θ with equal empirical score for all forecasters. After discarding that subset, the remaining domain Δ is always bounded in empirical settings, and we use $\lambda(\Delta)$ to denote the Lebesgue measure of Δ. Furthermore, we assume that H and H_d are normalized to $\lambda(\Delta)$ on Δ with Lebesgue densities h and h_d, and use the Kullback-Leibler divergence as distance measure,

$$\mathrm{KL}_\Delta(H, H_d) = \frac{1}{\lambda(\Delta)} \int_\Delta h(\theta) \log \frac{h(\theta)}{h_d(\theta)}\,\mathrm{d}\theta. \tag{4.8}$$

Due to the dependence of Δ on the realized forecast-observation pairs, a normalization with respect to Δ facilitates comparing the stability of forecast rankings across applications. The number given as the stability in Definition 4.3 should match the visual impression from a Murphy diagram, which is typically displayed to show the relevant range of θ, i.e. Δ.

The Kullback-Leibler divergence and its relevance in information theory enables an attractive interpretation of the distance from H to H_d as the deviation of an individual's preference from a predefined reference. Specifically, when H is the mixing measure that corresponds to the loss structure of a real-world application and H_d is uninformative, then $\mathrm{KL}_\Delta(H, H_d)$ describes the amount of application-specific information regarding the loss structure that is lost by using the measure H_d instead of H. Information can only enter by weighting thresholds, given the nature of our mixture representations for consistent scoring functions for quantiles and expectiles. Consequently, for an uninformative mixing measure the weighted score $h(\theta)\mathrm{S}_{\alpha,\theta}(x,y)$ should be translation-invariant in the vector (θ, x, y). A glance at the extremal scores confirms that they already satisfy this requirement, hence only a constant mixing density h retains translation-invariance. This suggests the Lebesgue measure as default mixing measure H_d for the broadest range of applications.

We are now interested in the minimal amount of information that can change a predefined ranking.

Definition 4.3 (stability). *Let $s_1(\theta), \ldots, s_n(\theta)$ be the empirical scores of n forecasts and let Δ be the union of the pairwise defined supports $\Delta_{ij} = \operatorname{supp}(s_i - s_j)$. The class $\mathcal{H}_\Delta$ consists of the normalized mixing measures that admit a Lebesgue density on Δ. Let $\mathcal{H}^{\mathrm{e}}_\Delta \subseteq \mathcal{H}_\Delta$ be the class of measures H that admit indices $i \neq j$ such that*

$$\int_\Delta (s_i - s_j)\,\mathrm{d}H = 0. \tag{4.9}$$

A measure $H_c \in \mathcal{H}^{\mathrm{e}}_\Delta$ is called critical with respect to the default measure $H_d \in \mathcal{H}_\Delta$ if

$$\mathrm{KL}_\Delta(H_c, H_d) \leq \mathrm{KL}_\Delta(H, H_d), \tag{4.10}$$

for all $H \in \mathcal{H}^{\mathrm{e}}_\Delta$. The stability of the default forecast ranking induced by H_d is the distance $\mathrm{KL}_\Delta(H_c, H_d)$.

Note that there is no lower bound that guarantees a different ranking. The default ranking can be enforced with mixing measures that are arbitrarily far away from the default. However, when a critical measure H_c exists, we have an upper bound such that the ranking remains stable. A default forecast ranking is stable for the class

$$B(H_c, H_d) = \{H \in \mathcal{H}_\Delta : \mathrm{KL}_\Delta(H, H_d) < \mathrm{KL}_\Delta(H_c, H_d)\}. \tag{4.11}$$

Proposition 4.1 (critical mixing density). *The critical measure $H_c \in \mathcal{H}^{\mathrm{e}}_\Delta$ for two forecasts with score difference $D(\theta) = s_1(\theta) - s_2(\theta)$ and a forecast ranking induced by $H_d \in \mathcal{H}_\Delta$ is given by*

$$h_c \propto h_d \exp(aD)$$

where $a \in \mathbb{R}$ such that $\int_\Delta D\,\mathrm{d}H_c = 0$, and h_c and h_d are the Lebesgue densities of H_c and H_d, respectively.

Proof. We seek to minimize the functional

$$J(h) = \int_\Delta h \log \frac{h}{h_d}\,\mathrm{d}\theta, \quad \text{subject to} \quad \int_\Delta hD\,\mathrm{d}\theta = 0, \quad \int_\Delta (h-1)\,\mathrm{d}\theta = 0.$$

The associated Lagrangian with Lagrange multipliers $a, b \in \mathbb{R}$ is

$$L(\theta, h) = h(\theta) \log \frac{h(\theta)}{h_d(\theta)} - ah(\theta)D(\theta) - b(h(\theta) - 1).$$

Informally, we have

$$\begin{aligned} \frac{\partial L}{\partial h} &= 1 + \log \frac{h}{h_d} - aD - b \overset{!}{=} 0, \\ \iff \qquad\qquad & h = h_d \exp(aD + b - 1). \end{aligned}$$

Choosing $a, b \in \mathbb{R}$ such that the two constraints are satisfied gives a density h_c as potential solution. To prove that h_c is a minimizer of $J(h)$ in $\mathcal{H}^e_\Delta$, let

$$f : \mathbb{R}^+ \to \mathbb{R}, \quad f(x) = x \log \frac{x}{d}, \quad \text{with} \quad f'(x) = 1 + \log \frac{x}{d},$$

where d is a real-valued constant. Since f is convex it satisfies the inequality

$$f(y) - f(x) \geq f'(x)(y - x)$$

for all $x, y \in \mathbb{R}^+$. This allows the conclusion that

$$\begin{aligned} J(h) - J(h_c) &\geq \int_\Delta \left(1 + \log \frac{h_c}{h_d}\right)(h - h_c)\,\mathrm{d}\theta \\ &= \int_\Delta (aD + b)\,(h - h_c)\,\mathrm{d}\theta = 0 \end{aligned}$$

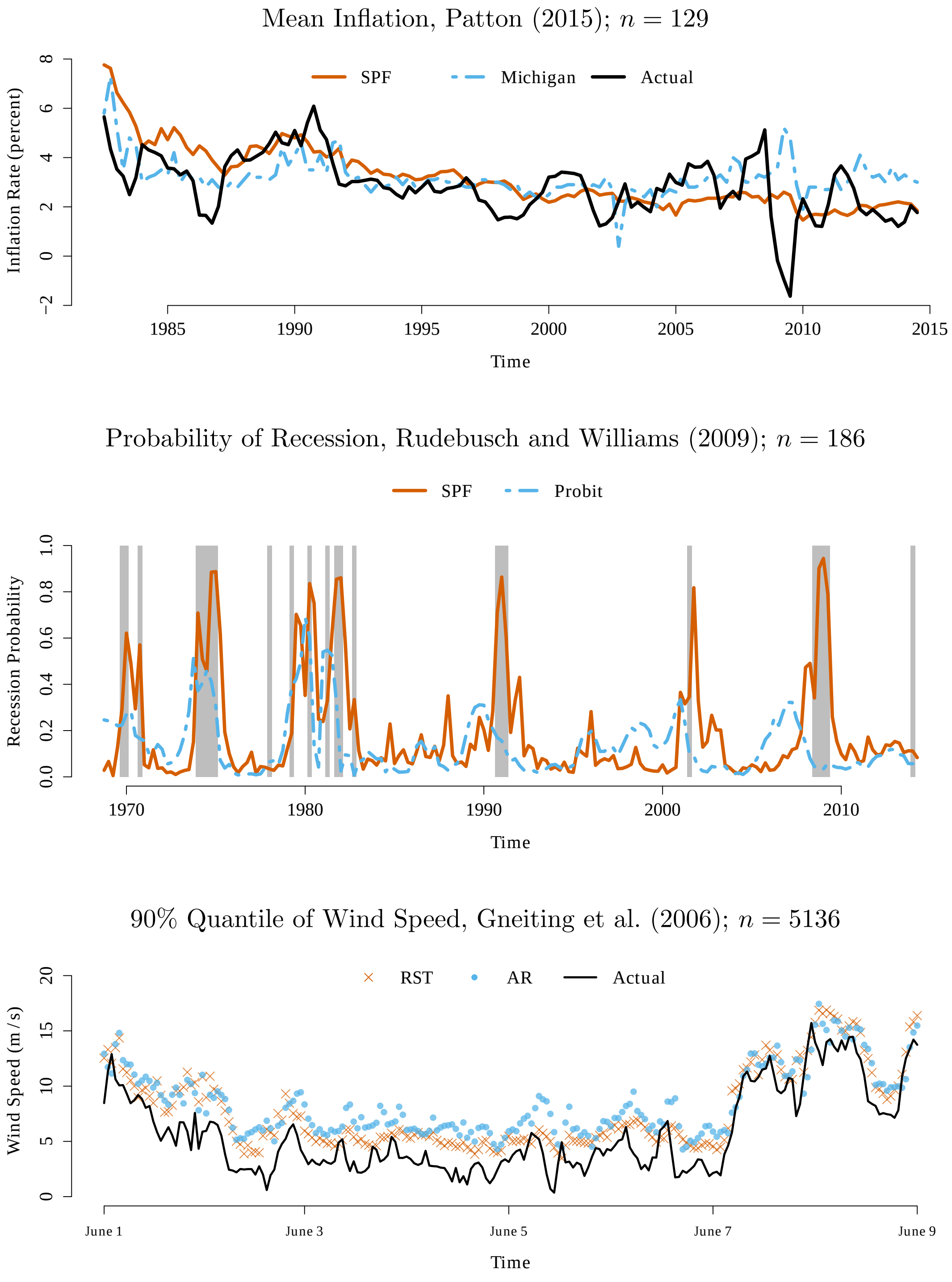
for any Lebesgue density h corresponding to a measure $H \in \mathcal{H}^e_\Delta$. □

4.4 Empirical examples

We now demonstrate the use of Murphy diagrams in economic and meteorological case studies in time series settings. In each example, interest is in a comparison of two forecasts, and so we show Murphy diagrams for the empirical scores and their difference. The jagged visual appearance stems from the behavior of the empirical score functions just explained and depends on the number n of forecast cases. We supplement the Murphy diagrams for a difference by confidence bands based on Diebold and Mariano (1995) tests with a heteroscedasticity and autocorrelation robust variance estimator (Newey and West 1987). The approach of Diebold and Mariano (1995) views empirical data of the form (4.5) as a sample from an underlying population and tests the hypothesis of equal expected scores. The confidence bands are pointwise and have a nominal level of 95%. For details on the data and their sources, we refer to the respective prior work.

4.4.1 Mean forecasts of inflation

In macroeconomics, subjective expert forecasts often compare favorably to statistical forecasting approaches; see Faust and Wright (2013) for evidence and discussion. For the United States, the Survey of Professional Forecasters (SPF) run by the Federal Reserve Bank of Philadelphia is a key data source; see, e.g.,

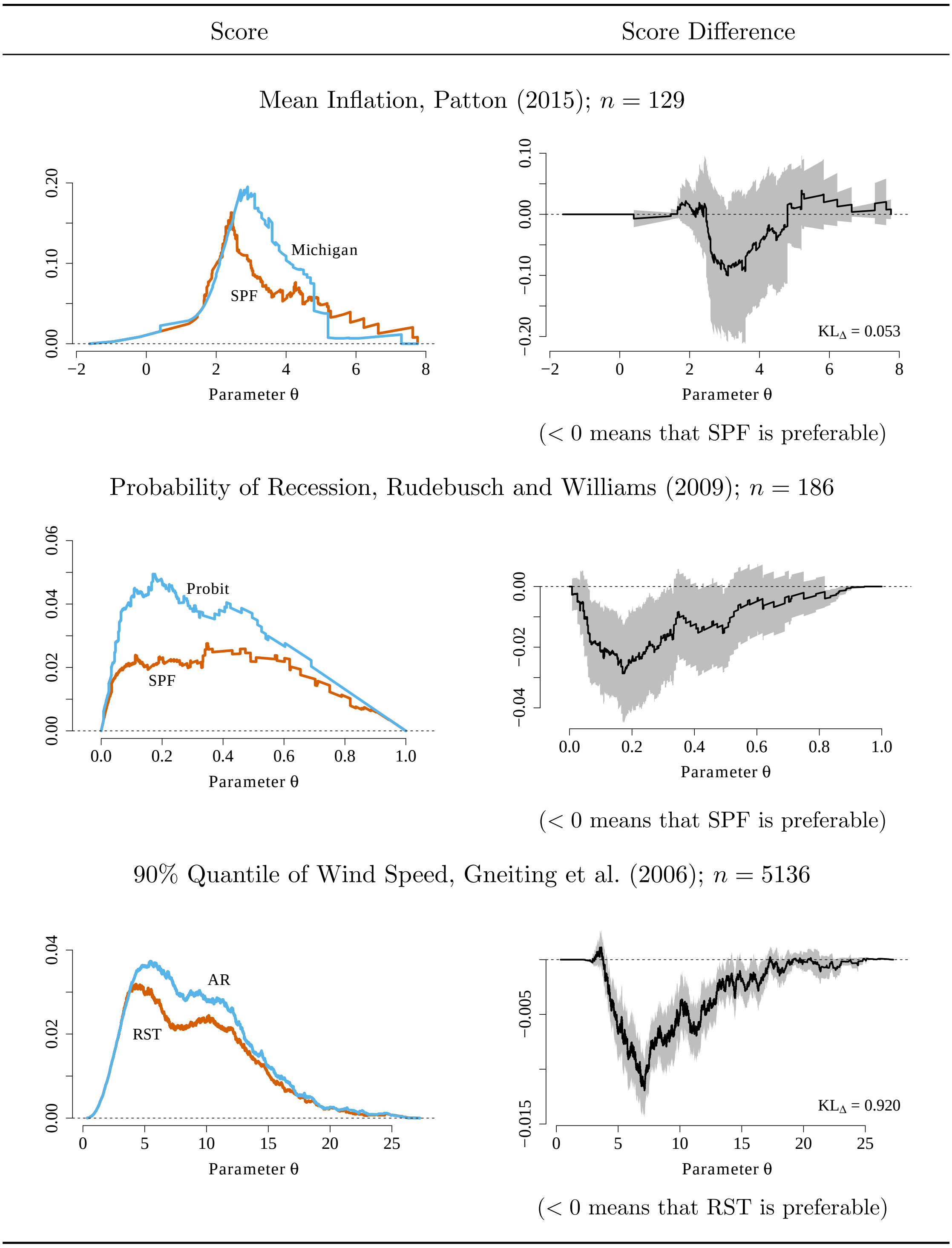


Mean Inflation, Patton (2015); $n = 129$

Probability of Recession, Rudebusch and Williams (2009); $n = 186$

90% Quantile of Wind Speed, Gneiting et al. (2006); $n = 5136$

Figure 4.6: Point forecasts and realizations in the empirical examples. In the middle plot, shaded areas correspond to actual recessions. The plot at bottom is restricted to a subperiod in summer 2003.



Score

Score Difference

Mean Inflation, Patton (2015); $n = 129$

(< 0 means that SPF is preferable)

Probability of Recession, Rudebusch and Williams (2009); $n = 186$

(< 0 means that SPF is preferable)

90% Quantile of Wind Speed, Gneiting et al. (2006); $n = 5136$

(< 0 means that RST is preferable)

Figure 4.7: Murphy diagrams in the empirical examples.

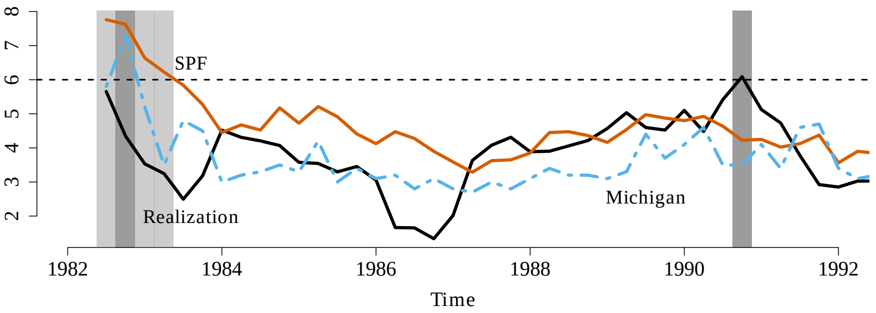


Figure 4.8: Mean forecasts of inflation for the third quarter of 1982 to the first quarter of 1992. The dark shade indicates a nonzero extremal score $\mathrm{S}^{\mathrm{E}}_{1/2,6}$ for both the SPF and the Michigan forecast, the light shade for the SPF forecast only.

Engelberg et al. (2009). Patton (2015) uses SPF data to illustrate the use of various scoring functions that are consistent for the mean functional.

Motivated by Patton's analysis, we analyze quarterly SPF mean forecasts for the annual inflation rate of the Consumer Price Index (CPI) over the next 12 months in the United States. We compare the SPF forecasts to forecasts from another survey, the Michigan Survey of Consumers, based on data from the third quarter of 1982 to the third quarter of 2014, for a test period of 129 quarters. Our implementation choices are as in Section 5 of Patton (2015), except that we update the data set to cover the observations for the second and third quarters in 2014, and that we use the slightly newer fourth quarter of 2014 vintage for the CPI realizations. The top panel of Figure 4.6 shows the forecasts along with the realizing values.

The respective Murphy diagrams are shown in the top panel of Figure 4.7. At left, the curves for the empirical elementary score $\mathrm{S}^{\mathrm{E}}_{1/2,\theta}$ of the SPF and the Michigan survey intersect prominently, suggesting that neither of the two surveys empirically dominates the other. At right, the confidence bands for the score differences are fairly broad and include zero for all values of θ. Furthermore, the stability of the default forecast ranking, induced by the Lebesgue mixing measure, is a low value of 0.053. (For a reference value, we refer to the data example of quantile forecasts for wind speed.) Note that the SPF is preferred for smaller values, whereas the Michigan forecast is preferred for larger values of θ. To interpret these results, consider the event threshold $\theta = 6$. A forecast x_t attains a nonzero extremal score $\mathrm{S}^{\mathrm{E}}_{1/2,6}$ in (3.4) if $x_t \leq 6 < y_t$ or $y_t \leq 6 < x_t$. The top panel of Figure 4.6 and the more detailed display in Figure 4.8 identify five quarters when the SPF incurs a nonzero penalty, as compared to two quarters only for the Michigan survey. Interestingly, the threshold $\theta = 6$ has become less relevant over time, in that forecasts and realizations have remained below six percent from 1991 onwards.

4.4.2 Probability forecasts of recession

We now relate to the rich literature on binary regression and prediction and analyze probability forecasts of United States recessions, as proxied by negative real gross domestic product (GDP) growth. The SPF covers probability forecasts for this event since the fourth quarter of 1968. Following Rudebusch and Williams (2009), we compare current quarter probability forecasts from the SPF to forecasts from a probit model based on the term spread, i.e., the difference between long and short term interest rates. We follow Rudebusch and Williams (2009) in all data and implementation choices, except that we update their sample through the second quarter of 2014, for a test period of 186 quarters. Detailed economic and/or statistical justification of these choices can be found in the original paper.

The middle row of Figure 4.6 shows the SPF and probit model based probability forecasts for a recession, with the gray vertical bars indicating actual recessions. During recessionary periods, the SPF tends to assign higher forecast probabilities than the probit model. Also, the SPF tends to assign lower forecast probabilities during non-recessionary periods. The respective Murphy diagrams in the middle row of Figure 4.7 show that the SPF attains lower empirical elementary scores $\mathrm{S}_\theta^{\mathrm{B}}$ at all thresholds $\theta \in (0, 1)$. The confidence bands for the score differences exclude zero for small values of the cost-loss ratio θ and confirm the superiority of the SPF over the probit model for current quarter forecasts. This can partly be attributed to the fact that SPF panelists have access to timely within-quarter information that is not available to the probit model. As demonstrated by Rudebusch and Williams (2009), the relative performance of the probit model improves at longer forecast horizons, where within-quarter information plays a lesser role.

4.4.3 Quantile forecasts for wind speed

We turn to a meteorological example and consider quantile forecasts at level $\alpha = 0.90$. We compare the regime-switching space-time (RST) approach introduced by Gneiting et al. (2006) to a simple autoregressive (AR) benchmark for two-hour ahead forecasts of hourly average wind speed at the Stateline wind energy center in the Pacific Northwest of the United States. The original paper refers to the specifications considered here as RST-D-CH and AR-D-CH, respectively. This terminology indicates that the methods account for the diurnal cycle and conditional heteroscedasticity. The data set, evaluation period, estimation and forecast methods for this example are identical to those in Gneiting et al. (2006), and we refer to the original paper for detailed descriptions. Both methods yield predictive distributions, from which we extract the quantile forecasts. The evaluation period ranges from May 1 through November 30, 2003, for a total of 5,136 hourly forecast cases.

The bottom panel in Figure 4.6 shows the quantile forecasts and realizations. The quantile forecasts exceed the outcomes at about the nominal level, at 89.7% for the RST forecast and 90.9% for the AR forecast, respectively, indicating good calibration. However, the RST forecasts are sharper, in that the average forecast value over the evaluation period is 9.2 meters per second, as compared to 9.7 meters per second in the case of the AR forecast. To see why the sharpness interpretation applies here, note that wind speed is a nonnegative quantity, so the lower prediction interval at level $\alpha \in (0,1)$ ranges from zero to the α-quantile, whence smaller quantiles translate into shorter, more informative prediction intervals and sharper predictive distributions. These observations suggest the superiority of the RST forecasts over the benchmark AR forecasts, and the Murphy diagrams for the empirical elementary scores $\mathrm{S}^{\mathrm{Q}}_{0.90,\,\theta}$ in the bottom row of Figure 4.7 confirm this intuition.

A visual comparison of the Murphy diagrams in the top and bottom rows of Figure 4.7 suggest a more stable ranking in the latter case. This is confirmed by the stability values corresponding to the default forecast rankings induced by the Lebesgue mixing measure. In the data example of inflation forecasts, where the Murphy diagrams at left intersect prominently and the confidence intervals at right contain 0 for all θ, we report a stability value of 0.053. Here, for the comparison of the RST and AR forecasts of wind speed, we observe barely intersecting Murphy diagrams at left, confidence intervals at right that do not contain 0 for a large region of θ, and a stability of 0.920.

4.5 Discussion

From a general applied perspective, Gneiting (2011, p. 757) had argued that if point forecasts are to be issued and evaluated,

> *"it is essential that either the scoring function be specified ex ante, or an elicitable target functional be named, such as the mean or a quantile of the predictive distribution, and scoring functions be used that are consistent for the target functional."*

Patton (2015, p. 1) took this argument a step further, by positing that

> *"rather than merely specifying the target functional, which narrows the set of relevant loss functions only to the class of loss functions consistent for that functional [...] forecast consumers or survey designers should specify the single specific loss function that will be used to evaluate forecasts."*

This is a very valid point. Whenever forecasters are to be compensated for their efforts in one way or another, the scoring function ought to be disclosed. To give an example of this best practice, the participants of forecast competitions hosted on the Kaggle platform (`www.kaggle.com`) are routinely informed about the relevant scoring function prior to the start of the competition. See, e.g., Hong et al. (2014) for a description of the Global Energy Forecasting Competition 2012.

However, there remain many situations in which point forecasters receive directives in the form of a functional, without an accompanying scoring function being available. This might be, because the forecasts are utilized by a myriad of communities, a situation often faced by national and international weather centers, because costs and losses are unknown or confidential, because the goal is general methodological development, as opposed to a specific applied task, because interest centers on an understanding of forecasters' behaviors and performance, or simply because of negligence of best practices. In such settings, our findings suggest the routine use of new diagnostic tools in the evaluation and ranking of forecasts, which we call Murphy diagrams. Interest sometimes centers on decompositions of expected or empirical scores into uncertainty, resolution, and reliability components, as studied by DeGroot and Fienberg (1983), Bröcker (2009), and Bentzien and Friederichs (2014), among others. Extensions of Murphy diagrams in these directions may be worthwhile.

As discussed in Section 4.1, nested information sets are sufficient for forecast dominance. However, the converse is not true, in that, if a forecaster dominates another, the respective information sets need not be nested. Specifically, if a forecaster has access to a highly informative explanatory variable, but not to a weakly informative one, then she may dominate a competitor who can access the weakly informative variable only, even though the information sets are not nested. Explicit examples of this type can readily be constructed. From a broader perspective, it would be of interest to study any implications of forecast dominance on information sets. As a caveat, the dominance relation appears to be strong, and empirical dominance may not be very commonly observed in practice. In such cases, Murphy diagrams can still provide informal clues to critical threshold values θ, which can then be investigated in detail, as illustrated in our inflation example. An R package (Jordan and Krüger 2016) accompanying Ehm et al. (2016) facilitates the implementation of Murphy diagrams in a wide range of applications.

5 Statistical Tests

"How stable is the ranking when I collect more data?"

Thus far we have investigated how the choice of performance measure can influence the perceived predictive ability of competing forecasters. In this chapter, we are interested in the question whether enough data has been collected to draw meaningful conclusions.

The importance of forecasting is undeniable. Knowledge about future developments allows making better decisions at the present time, an everyday process both consciously and subconsciously. Whenever we cannot predict the future ourselves, we turn to experts for their opinion and subsequently trust their judgment. Thus, we would like to identify forecasters with a predictive ability that is genuinely superior to that of others, which brings us to the notion of statistical significance. This entails the following questions:

How do we correctly measure predictive performance?
How big is the expected difference in performance?
How strongly do the observed differences fluctuate?
How many observations are available?

The first question has largely been answered in the previous chapters, but reappears in the formulation of this chapter's statistical tests. The remaining questions can be used to determine superior predictive performance, but as shown in the context of the Weak Law of Large Numbers, only in combination.

One may argue that tests of *statistical significance* are primarily interesting for small sample sizes, where *small* is relative to the second and third questions' answers. It could easily apply to a situation with hundreds or even thousands of observations when the forecasters are almost equally skilled. The main arguments why the sample size should be small are that all models are wrong and that the probability for two nonidentical forecasts to have equal predictive performance is essentially zero. As a result, we will always find a significant difference asymptotically and the task of assigning superior predictive ability boils down to the question of a sufficiently large sample size. Our focus will be macroeconomic data, which are usually reported quarterly which leads to sample sizes on the order of a few dozen up to maybe two hundred observations, corresponding to time periods up to fifty years.

5.1 Equal predictive performance

As discussed in the previous chapters, we evaluate predictive performance by mapping forecast-observation pairs to real numbers. Depending on the scenario a forecast may be in the form of a predictive distribution, requiring the use of a proper scoring rule, or in the form of a point forecast, where evaluation is performed using consistent scoring functions. In either case, we obtain two series of scores $(s_{i1})_{i=1}^n$ and $(s_{i2})_{i=1}^n$ when comparing two forecasters across n forecast cases, which can then be merged into one series of score differentials $(d_i)_{i=1}^n$, where $d_i = s_{i1} - s_{i2}$. Depending on the sign of the sample mean

$$\bar{d}_n = \frac{1}{n}\sum_{i=1}^n d_i,$$

we determine which of the two forecasts is superior. Importantly, the mean difference is the only valid measure to compare two forecasts, because the properties of propriety and consistency are defined by minimization of the expected score. We begin by assuming that $\bar{d}_n \sim F \in \mathcal{F}_1$, i.e. the sample mean of the score differentials follows a distribution with finite first moment. Hence, we express the null hypothesis of significance tests for equal predictive performance in the following way,

$$\bar{d}_n \sim F \in \{F \in \mathcal{F}_1 : \mathbb{E}_F X = 0\}\,. \tag{5.1}$$

Implicitly, we assume the existence of a prediction space with tuples $(F_{i1}, F_{i2}, Y_i)_{i=1}^n$ for the chosen proper scoring rule such that the expectation of $\bar{d}_n$ is zero, or the existence of a point prediction space for a given consistent scoring function with the same restriction.

Testing other functionals for a deviation from zero

Alternative test statistics become available when we restrict the class $\mathcal{F}_1$ of valid distributions for $\bar{d}_n$. We consider substitute statistics of the form

$${}_\psi\bar{d}_n = \frac{1}{n}\sum_{i=1}^n \operatorname{sgn}(d_i)\,\psi(|d_i|),$$

using a non-negative function ψ that is not constant zero and the sign function as defined by

$$\operatorname{sgn}(x) = \begin{cases} -1, & x < 0, \\ 0, & x = 0, \\ 1, & x > 0. \end{cases}$$

To test for $\mathbb{E}_\mathbb{Q}\,\bar{d}_n = 0$ using the substitute statistic ${}_\psi\bar{d}_n$, we require

$$\mathbb{E}_\mathbb{Q}\,\bar{d}_n = 0 \quad \Longleftrightarrow \quad \mathbb{E}_\mathbb{Q}\,{}_\psi\bar{d}_n = 0, \tag{5.2}$$

where $\mathbb{Q}$ denotes a joint distribution of the score differentials. Let $\mathcal{F}_1^\psi$ be the subclass of distributions $F \in \mathcal{F}_1$ induced by distributions $\mathbb{Q}$ satisfying (5.2) for the non-negative function ψ that is not constant zero. We can then use ${}_\psi\bar{d}_n$ to detect a deviation from zero, and under the restriction (5.2) we can write the null hypothesis (5.1) as

$$
{}_\psi\bar{d}_n \sim F \in \left\{ F \in \mathcal{F}_1^\psi : \mathbb{E}_F X = 0 \right\}. \tag{5.3}
$$

If ψ is the identity function $\psi(x) = x$, then (5.1) and (5.3) are equivalent. If the family of distributions $\mathbb{Q}$ is restricted to satisfy (5.2) for all fathomable functions ψ, then the null distribution in (5.3) needs to be symmetric. For $\psi \equiv 1$, we only consider distributions where a deviation from the null hypothesis implies a non-zero value both for the mean and the median. We may then test for a non-zero median to answer the question of a sufficiently large sample size, i.e. when we have enough data to determine that the median is non-zero then so must be the mean. Clearly, the restriction (5.2) is closely tied to the scenario of a two-sided test. One-sided tests require additional sign assumptions under a deviation from zero.

The following example illustrates how distributions of $\bar{d}_n$ can be symmetric under the true null hypothesis, and be potentially asymmetric otherwise.

Example 5.1 (exchangeable normalized forecast errors). Let there be two forecasters issuing predictions from the same location-scale family with parameters $\mu \in \mathbb{R}$ and $\sigma > 0$. Furthermore, let the normalized forecast errors

$$
\frac{y - \mu_1}{\sigma_1} \quad \text{and} \quad \frac{y - \mu_2}{\sigma_2}
$$

of the two forecasters be exchangeable. Then, for proper scoring rules that admit a representation

$$
\mathrm{S}(F_{\mu,\sigma}, y) = a(\sigma)\, \mathrm{S}\left(F_{0,1}, \tfrac{y-\mu}{\sigma}\right) + b(\sigma), \tag{5.4}
$$

using non-decreasing functions $a : \mathbb{R}^+ \to \mathbb{R}^+$ and $b : \mathbb{R}^+ \to \mathbb{R}$, the distribution of the score difference

$$
\mathrm{S}(F_{\mu_1,\sigma_1}, y) - \mathrm{S}(F_{\mu_2,\sigma_2}, y)
$$

satisfies assumption (5.2) for a large class of functions ψ. The idea is that two forecasters use different information sets determining their uncertainty, while their location forecasts are correctly specified. If the information sets contain equal amount of knowledge, the forecasters will be equally skilled. Incidentally, the four most commonly used performance measures admit a representation as in (5.4) to isolate the uncertainty,

$$
\begin{aligned}
\mathrm{AE}(\mu, y) &= \sigma\, \mathrm{AE}\left(0, \tfrac{y-\mu}{\sigma}\right) \\
\mathrm{SE}(\mu, y) &= \sigma^2\, \mathrm{SE}\left(0, \tfrac{y-\mu}{\sigma}\right) \\
\mathrm{CRPS}(F_{\mu,\sigma}, y) &= \sigma\, \mathrm{CRPS}\left(F_{0,1}, \tfrac{y-\mu}{\sigma}\right) \\
\mathrm{LS}(F_{\mu,\sigma}, y) &= \mathrm{LS}\left(F_{0,1}, \tfrac{y-\mu}{\sigma}\right) + \log(\sigma).
\end{aligned}
$$

5.1.1 p-values

Randomization

Uniformity of the probability integral transform $F(\bar{d_n})$, when $\bar{d_n} \sim F$, is a classical result for continuous distributions F. For a correctly specified (continuous) F, we have

$$F(\bar{d_n}) \sim \mathcal{U}(0,1). \tag{5.5}$$

Null distributions with point masses can be a nuisance, requiring an additional randomization step using a uniform random variable U,

$$F(\bar{d_n}) - UP_{\bar{d_n}} \sim \mathcal{U}(0,1),$$

where $P_x = F(x) - F(x^-)$ and $F(x^-)$ is the left-sided limit of F in x. Furthermore, this translates to p-values. We define the randomized p-value as a convex combination of the lower limit p^- and the upper limit p^+ depending on the realization of U,

$$p = p^+ - U(p^+ - p^-),$$

where p^- and p^+ are equal for a continuous F. Measuring deviations from the null hypothesis (5.1) in both directions, we define

$$\begin{aligned} p^- &= \mathbb{P}_{X\sim F}(|X| > |\bar{d_n}|) \\ p^+ &= \mathbb{P}_{X\sim F}(|X| \geq |\bar{d_n}|). \end{aligned}$$

Habiger and Peña (2011) discuss randomized p-values and tests, suggesting the augmentation of a dataset with a realization from the uniform random variable U in the instant the data set is obtained. The augmentation is treated as being part of the data generating process, a concept we will use in the simulation studies as it facilitates comparison of non-parametric tests. One single randomization variable as part of the dataset is preferable to multiple ones for the individual tests. For the empirical real data examples, we argue that the communication of an interval $[p^-, p^+]$ retains full information and is only mildly more complicated than reporting a single p-value and the corresponding realization of U. When we give only a single p-value it will be the upper bound p^+ by default.

Adjustments in multiple testing

In the process of finding a suitable null distribution F for $\bar{d_n}$, we will often assume independence in the sequence of score differentials. Reasons for this could be the small sample behavior of autocorrelation estimators or the argument that ideal one-step ahead forecasts lead to independent scores. However, ideal multi-step ahead forecasts will typically be correlated.

Using the intuition that ideal h-step ahead forecasts are at most $(h-1)$-dependent (Diebold and Mariano 1995), we can split up the series of score differentials in h sub-series, each of which containing every hth member. Discarding all but one of those series is the safest choice to ensure a uniform distribution of the

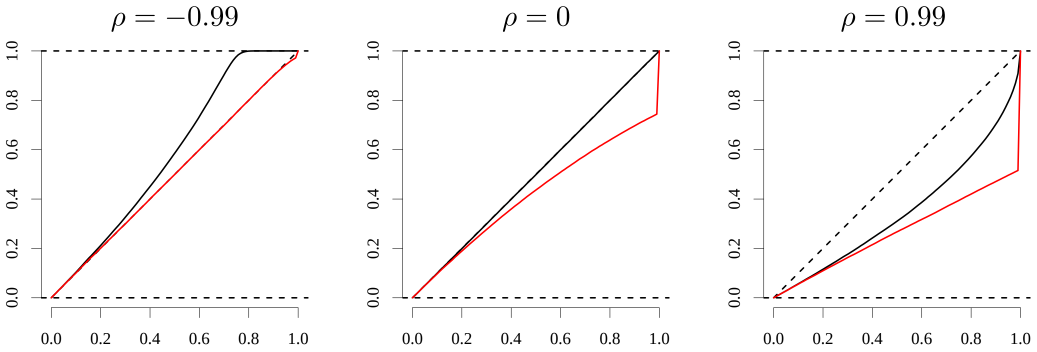


Figure 5.1: Cumulative distribution functions of p_B in red and p_S in black, for $h = 2$ with $p_1, p_2 \sim \mathcal{U}(0,1)$ and correlation ρ. Values above the diagonal indicate a tendency to overstate significance.

p-values under a correctly specified null distribution, but reducing the sample size in this way hampers the ability to detect a deviation from the null hypothesis.

Bonferroni's and Šidàk's corrections (Abdi 2007) for testing a null hypothesis against an alternative hypothesis allow the combination of multiple test results, thus using the full data. When performing a hypothesis test at level α, the Bonferroni correction performs h tests at level α/h, one on each partial series, and rejects the null hypothesis when any one of the subtests reject the null hypothesis. This is a very conservative rejection rule that generally does not reach level α if the null hypothesis is true. The Šidàk correction tests each partial series at the level of $1 - (1-\alpha)^{1/h}$, thus maintaining the proper rejection rate in favor of an alternative hypothesis in the case of independence.

Significance tests with p-values as the focus of interest require a method to combine p-values from multiple tests while maintaining a uniform distribution of the resulting p-value under a correctly specified null distribution. The adjusted p-values derived from Bonferroni's and Šidàk's considerations can be defined as

$$\begin{aligned} p_B &= \min\left\{1, h\, p_{\min}\right\} \\ p_S &= 1 - \left(1 - p_{\min}\right)^h, \end{aligned}$$

where $p_{\min} = \min_{k=1,\dots,h} p_k$ is the smallest observed p-value from the collection. Figure 5.1 illustrates the effects of correlation. We see that a negative correlation creates a tendency to overstate significance for Šidàk's approach which is what the Bonferroni correction protects against. However, this tendency seems to be almost negligible for p-values smaller than 0.2. In practice, positive autocorrelation of the score differentials, resulting in positively correlated p-values, is more commonly observed than negative correlation. For these reasons we prefer the adjustment based on the assumption of independence based on Šidàk's correction.

All these considerations are mainly important when the p-values come from non-parametric tests that assume zero autocorrelation at all lags in the series of score differentials. Since these tests are often based on sign permutations at each point in time, the resulting null distributions are discrete and the pair of smallest

h	1	2	3	4	5	6	7	8
$\alpha = 10\%$	5	11	18	26	33	42	50	59
$\alpha = 5\%$	6	13	21	30	39	48	57	67
$\alpha = 1\%$	8	18	28	39	50	62	74	86

Table 5.1: Lowest number of observations necessary for Šidàk-adjusted p-values of two-sided tests as defined in equation (5.6) for varying forecast horizons h and levels α.

lower and upper p-values is $p^- = 0$ and $p^+ = \left(\frac{1}{2}\right)^{n-1}$ for two-sided tests. Any subchain has length n/h for h-step ahead forecasts, so that under a two-sided alternative we require

$$\left(\tfrac{1}{2}\right)^{n/h-1} \leq 1 - (1-\alpha)^{1/h},$$

in order to ensure that the smallest possible p_S^+ is below some level α. The necessary number of observations is given by

$$n_{\min} = \text{ceiling}\left(h \left[\frac{\log(1-(1-\alpha)^{1/h})}{\log(1/2)} + 1 \right] \right). \tag{5.6}$$

In Table 5.1 we provide a selection of values for $n_{\min}$ with different forecast horizons h and typical levels α.

5.1.2 Tests

Diebold and Mariano (1995) introduced a z-test which has become the standard test for the null hypothesis in (5.1), and is closely related to the paired t-test. It has the beauty of imposing only one single assumption on the series of score differentials, namely wide-sense stationarity, and is capable to account for autocorrelation. The test is designed to take forecasts as primitives, given to the econometrician without further knowledge about the way they were issued (Diebold 2015). If, on the other hand, a researcher has developed a forecasting technique and wishes to assess its predictive performance in comparison to a reference forecast, uncertainty in the parameter estimation may lead to different asymptotic results (West 1996). In the case of a degenerate variance of the null distribution, e.g. when the series of score differentials is constant, the normality assumption is violated but we can use randomized p-values. Many other modifications of the basic DM test followed, who addressed the issues of comparing more than just two forecasters (White 2000) and comparing nested models (Clark and McCracken 2001). The DM principle is also used in tests for forecast encompassing (Harvey et al. 1998). A recent paper by Busetti and Marcucci (2013) provides an extensive Monte Carlo investigation with the most popular tests for equal mean squared error and forecast encompassing in the context of nested models. For comprehensive surveys, we refer to West (2006) and Clark and McCracken (2013).

We also investigate three nonparametric tests which are based on the permutation principle. This means that the small sample assumptions for the nonparametric tests are distribution-free, whereas the DM test requires normality of the score differential distribution. The trade-offs are stronger assumptions on the correlation structure, i.e. assuming independence of the score differentials which is unrealistic for multi-step ahead predictions. Some ways to circumvent this restriction were discussed in Section 5.1.1.

Diebold-Mariano test

The DM test (Diebold and Mariano 1995) is widely used due to its versatility and ease of implementation. It has only one simple assumption which is the wide-sense stationarity, with short memory, of the score-differential series $\{d_i\}_{i=1}^n$. This allows the use of a version of the central limit theorem

$$\sqrt{n}(\bar{d}_n - \mu) \xrightarrow{d} \mathcal{N}(0, \sigma^2), \tag{5.7}$$

where $\sigma^2 = \sum_{k=-\infty}^{\infty} \gamma_k$ is the variance of the wide-sense stationary process with γ_k denoting the autocovariance of the score differential series at lag k. The finite-sample test in the form of (5.1) is then defined via

$$\text{DM} : \bar{d}_n \sim \mathcal{N}(0, \hat{\sigma}^2/n), \tag{5.8}$$

where $\hat{\sigma}^2$ is a plug-in estimate of σ^2 using the available score differential series. Diebold and Mariano (1995) argue that the autocovariance depends on the forecast horizon $h \geq 1$ and employ a lag window to truncate the infinite sum,

$$\begin{aligned} \hat{\sigma}^2 &= \hat{\gamma}_0 + 2 \sum_{k=1}^{\infty} \hat{\gamma}_k \mathbb{1}\{k \leq h-1\} \\ \hat{\gamma}_k &= \frac{1}{n} \sum_{i=|k|+1}^{n} (d_{i-|k|} - \bar{d}_n)(d_i - \bar{d}_n), \end{aligned}$$

under the assumption that h is small with respect to n, but necessarily $h \leq n$. Clearly, this ignores the possibility of autocorrelation in the case of $h = 1$. However, this estimator for the auto-covariance is not unbiased and size distortions are common for small sample sizes. This issue has been addressed by Harvey et al. (1997) who propose a bias correction in finite-sample scenarios,

$$\begin{aligned} \text{DM}_{\text{HLN}} : \bar{d}_n &\sim \mathcal{N}(0, \hat{\sigma}^2/n') \\ n' &= n + 1 - 2h + h(h-1)/n, \end{aligned}$$

explicitly requiring $n \geq 2$. Nevertheless, the normality assumption for $\bar{d}_n$ is still based on asymptotic theory, or the assumption of normality of the score differentials in small samples. For the latter case, Harvey et al. (1997) propose to use the critical values of a t-distribution with $n - 1$ degrees of freedom.

Permutation test

We can also address the small sample behavior with permutation tests, which can be traced back to at least Fisher (1935) and have since been in wide-spread use. Essentially, based on a series $\{B_i\}_{i=1}^n$ of independent random variables with equal probability for -1 and 1, we consider the test defined via

$$\text{PM} : \bar{d_n} \sim \mathcal{L}\left(\frac{1}{n}\sum_{i=1}^{n} B_i d_i\right). \tag{5.9}$$

This distribution can be computed exactly with computational complexity $\mathcal{O}(2^n)$, or approximated via Monte Carlo methods. Asymptotically, when assuming wide-sense stationarity with zero autocorrelation at all lags, we have

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{B_i d_i}{\sqrt{\sigma_n^2}} \xrightarrow{d} \mathcal{N}(0,1), \tag{5.10}$$

where $\sigma_n^2 = \frac{1}{n}\sum_{i=1}^n d_i^2$.

Sign test

The sign test (Conover 1999) dates back to the 18th century, and is mentioned as an alternative to their proposed test by Diebold and Mariano (1995) when testing for the null hypothesis of equal predictive performance. We use a modified version defined via

$$\text{SN} : {}_{\psi_{\text{SN}}}\bar{d}_n \sim \mathcal{L}\left(\frac{1}{n}\sum_{i=1}^{n} \operatorname{sgn}(B_i d_i)\right), \tag{5.11}$$

where $\psi_{\text{SN}} \equiv 1$, and using a series $\{B_i\}_{i=1}^n$ of random variables with equal probability for -1 and 1. This means we require the assumption ${}_{\psi_{\text{SN}}}\bar{d}_n \sim F \in \mathcal{F}_1^{\psi_{\text{SN}}}$ in addition to assuming that it follows the proposed distribution in (5.11) when its expected value is zero. The latter is typically argued to be the result of independent identically distributed draws, but can be slightly weakened to independent draws of random variables with a zero median under the true null hypothesis. It is a test which determines superiority in predictive performance by checking which of the two forecasters performs better more often, irrespective of the magnitude of difference. When every score differential d_i is non-zero, the test's definition is equivalent to that of the standard sign test,

$$\sum_{i=1}^{n} \mathbb{1}(d_i > 0) \sim B(n, 0.5),$$

where B denotes the binomial distribution.

Wilcoxon signed-rank test

The Wilcoxon signed-rank test (Wilcoxon 1945; Siegel 1956) was also discussed by Diebold and Mariano (1995). Again, based on a series $\{B_i\}_{i=1}^n$ of random variables with equal probability for -1 and 1, we use the version defined as

$$\mathrm{W} : {}_{\psi_\mathrm{W}}\bar{d}_n \sim \mathcal{L}\left(\frac{1}{n}\sum_{i=1}^{n} \operatorname{sgn}(B_i d_i)\psi_\mathrm{W}(|d_i|)\right), \tag{5.12}$$

where ψ_W maps the absolute value of d_i to its fractional rank, i.e.

$$\psi_\mathrm{W}(x) = n\,\mathbb{P}_n(|X| < x) + \frac{n\,\mathbb{P}_n(|X| = x) + 1}{2},$$

where $\mathbb{P}_n$ is the empirical distribution of the realized series $(d_i)_{i=1}^n$. Regarding assumptions, we suppose that the distribution F from which ${}_{\psi_\mathrm{W}}\bar{d}_n$ is drawn lies in $\mathcal{F}_1^{\psi_\mathrm{W}}$, and furthermore follows the distribution given in (5.12) under the null hypothesis. Again, the latter is typically argued to be the result of independent identically distributed draws of the score differentials, yet with the additional assumption of symmetry around zero.

5.2 Simulation study

Let us start our simulation study with a simple example that illustrates some properties of the four tests. We simulate a series of 64 independent identically distributed score differentials following either a normal distribution or a Laplace distribution, both with mean 0.2 and scale parameter 1. Then, we add two contaminations with values of 30 and 60 successively. Repeating this experiment 10,000 times yields distributions of p-values which are shown in Figure 5.2.

Comparing the uncontaminated distributions in the left-most column, we make an interesting observation. Typically, the sign test is associated with a low ability to detect deviations from the null hypothesis compared to the signed-rank test or the paired t-test. However, this rule of thumb is based on a normality assumption. For leptokurtic distributions this statement can be false, as we observe for the Laplace distribution, and the sign test would be the preferred test, subject to the appropriate assumptions. The book by Conover (1999) on nonparametric statistics discusses this relationship for the signed-rank test and the paired t-test in its chapter dedicated to the sign test.

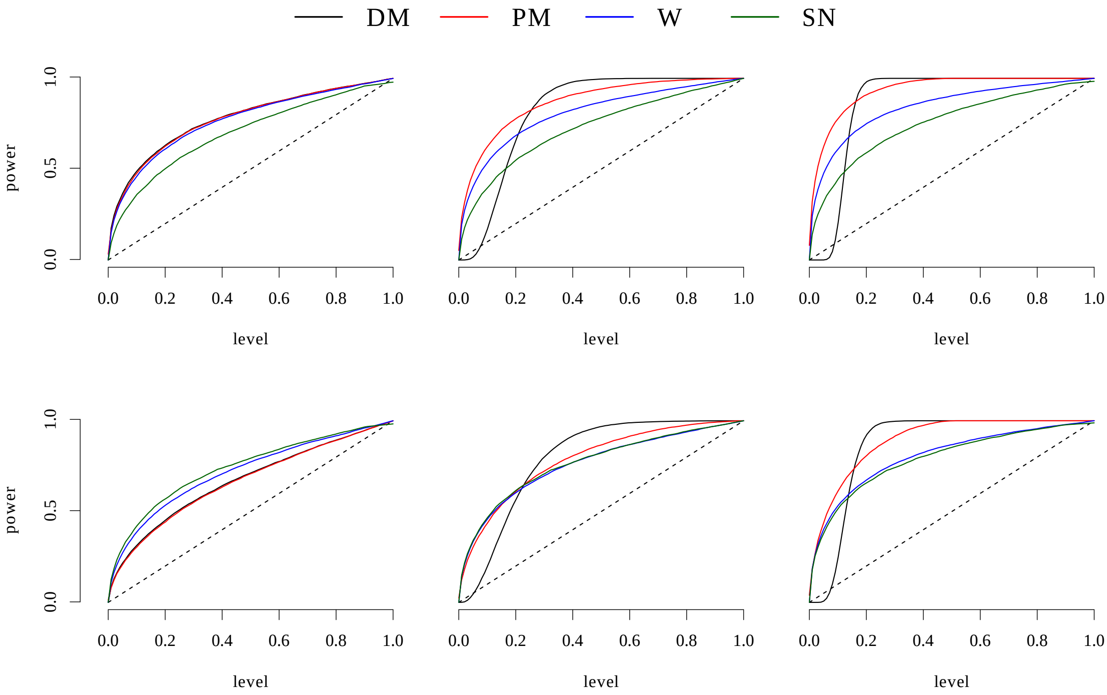


Figure 5.2: Power with respect to the level, i.e. the p-value's cumulative distribution function. Top row: $\mathcal{N}(0.2, 1)$. Bottom row: Laplace$(0.2, 1)$. Sample size: 64. From left to right, we start with no contamination and append values of 30 and 60 to the series.

Regarding contaminations, we observe a stable behavior of all three nonparametric tests, whereas the DM test behaves in unexpected ways. Our data generating process violates the null hypothesis with a positive expected value for the sample mean, and the contaminations reinforce this deviation. However, in this scenario it is highly unlikely, even less likely than under the true null hypothesis, to observe very small p-values for the DM test. Depending on the level at which we define a result to be significant, and whether the outlying observations are really contaminations or just extreme results, the DM test can exhibit the best or the worst ability of the considered tests to detect a null hypothesis violation.

This inflation of p-values by extreme contaminations has previously been reported by Dell'Aquila and Ronchetti (2004), and the reason for this is a comparatively larger inflation of the sample variance compared to the sample mean's inflation. Removal by hand could be feasible for small samples but how do we distinguish between a misreported prediction and an extreme opinion of a forecaster? Adding to this conundrum, decision theoretic considerations demand that forecasters should have no incentive to deviate from honest reporting given they have knowledge of the evaluation criteria. However, these results suggest that forecasters can possibly hide their relatively weaker performance by issuing an exceptionally bad prediction and then dropping out of the panel, when the DM test is used to test for significance at a low level.

Simulation model

The previous section gave a first impression how the tests behave, yet the score differentials' data generating process was chosen arbitrarily without justification or explanation. In this section, we generate forecasts and observations, transform the pairs into scores, and then investigate the tests' behaviors for the realized score differentials. The considerations include various choices with respect to the data generating process, the employed performance measure, and the forecast horizon. All Monte Carlo simulations are run in 10,000 replicates.

In terms of performance measures, we consider the two most common consistent scoring functions, the absolute error (AE) and the squared error (SE), and the two most common proper scoring rules, the logarithmic score (LS) and the continuous ranked probability score (CRPS). Additionally, we consider a member of Patton's (2015) exponential Bregman family,

$$\mathrm{S}_a(x,y) = \frac{1}{a^2}(\exp(ay) - \exp(ax)) - \frac{1}{a}\exp(ax)(y-x),$$

which can also be written in the form

$$\begin{aligned}\mathrm{S}_a(x,y) &= \frac{1}{a^2}\exp(ax)\mathrm{L}_a(y-x),\\ \mathrm{L}_a(z) &= \exp(az) - az - 1,\end{aligned}$$

where L_a is a version of the LINEX loss function introduced by Varian (1975). In the following, we will use the term mLINEX for the scoring function S_a with the parameter choice $a = 1$. The mLINEX punishes overprediction linearly and underprediction exponentially, while being a consistent scoring function for the expectation functional. And lastly, we consider a threshold-weighted version of the CRPS,

$$\mathrm{tCRPS}(F,y) = \int_{-\infty}^{\infty} (F(\theta) - \mathbb{1}\{\theta \geq y\})^2 \left(1 - \tfrac{\varphi(\theta)}{\varphi(0)}\right) \mathrm{d}\theta,$$

where φ is the density function of the standard normal distribution. This proper scoring rule reduces the importance of thresholds close to zero.

We generate the observations as a sum of two independent AR(1) processes with the same parameter, but different variance in the innovations. Our two forecasters issue ideal predictions with respect to their information sets, each containing perfect knowledge about the past and future of one of the two processes. Specifically,

$$\begin{aligned}
Y_t &= X_{1,t} + X_{2,t} && \\
X_{1,t} &= \theta X_{1,t-1} + \epsilon_{1,t} && \epsilon_{1,t} \sim \mathcal{N}(0, 1+s)\\
X_{2,t} &= \theta X_{2,t-1} + \epsilon_{2,t} && \epsilon_{2,t} \sim \mathcal{N}(0, 1-s),
\end{aligned}$$

where $\theta < 1$ and $s \in [0,1)$. We sample the chain's starting points from their equilibrium distributions

$$X_{1,1} \sim \mathcal{N}\left(0, \tfrac{1+s}{1-\theta^2}\right) \qquad X_{2,1} \sim \mathcal{N}\left(0, \tfrac{1-s}{1-\theta^2}\right).$$

This setup allows us to control the information content of the forecasters' knowledge bases and thus the predictions' quality. The first forecaster has perfect knowledge of X_1 and the second forecaster of X_2, resulting in the following predictive distributions for h-step ahead forecasts

$$\begin{aligned} F_{1,t+h} &= \mathcal{L}(Y_{t+h}|X_{1,t+h}, X_{2,t}) = \mathcal{N}\left(\mu_{1,t+h}, \sigma^2_{1,t+h}\right) \\ F_{2,t+h} &= \mathcal{L}(Y_{t+h}|X_{1,t}, X_{2,t+h}) = \mathcal{N}\left(\mu_{2,t+h}, \sigma^2_{2,t+h}\right), \end{aligned}$$

where $F_{1,t+h}$ is the probabilistic forecast of the first forecaster and $F_{2,t+h}$ of the second, and the parameters are given by

$$\begin{aligned} \mu_{1,t+h} &= X_{1,t+h} + \theta^h X_{2,t} \qquad & \sigma^2_{1,t+h} &= \sum_{k=0}^{h-1} \theta^{2k}(1-s) \\ \mu_{2,t+h} &= \theta^h X_{1,t} + X_{2,t+h} \qquad & \sigma^2_{2,t+h} &= \sum_{k=0}^{h-1} \theta^{2k}(1+s). \end{aligned}$$

The forecast horizons we will investigate are one-step, two-step, and four-step ahead, corresponding to forecasts of quarterly reported variables predicted up to one year into the future. For one-step ahead forecasts, we also consider innovations following a Student's t-distribution with six degrees of freedom. Since the predictive distributions are symmetric, the location parameters $\mu_{1,t+h}$ and $\mu_{2,t+h}$ correspond to the respective forecasters' optimal point prediction under AE, SE, and mLINEX.

The whole data generating process is symmetric when $s = 0$, which leads to equal mean scores corresponding to the null hypothesis, and the proposed tests have the task to detect deviations $s > 0$. As we are now considering the power with respect to the parameter s, we have to choose a level at which the results are deemed significant. For the plots in Figures 5.3, 5.4, and 5.5, each test receives an individual level based on the Monte Carlo distribution of the p-values for $s = 0$, so that the level equals 5% under the true null hypothesis. However, we also state the actual level each test achieves for $s = 0$ under a nominal level of 5%.

One-step ahead forecasts

One-step ahead forecasts are different from multi-step ahead forecasts in that we can assume score differentials to be independent for ideal forecasts. Figure 5.3 shows plots of the 5%-adjusted significance rate for one-step ahead forecasts. The results differ substantially between scoring functions for point forecasts (AE, SE, mLINEX) and proper scoring rules for probabilistic forecasts (CRPS, tCRPS, LS). As heavy-tailed distributions are common in economics and finance, we also show some results for Student's t-distribution with six degrees of freedom. The auto-regressive parameter in the model is set to 0, since ideal forecasts would not be serially correlated for one-step ahead forecasts, anyway.

For the AE, the score differentials' distribution is only slightly leptokurtic. Not much excess kurtosis is introduced so the DM test has the highest power, while maintaining the appropriate rate under the true null hypothesis. The permutation test performs equally well, closely followed by the signed-rank test, and the sign test performs worst. Similar results can be observed for the SE, except that for t_6-distributed innovations the weakness with respect to excess kurtosis of the DM test, and also to a lesser extent of the permutation test, are starting to show. This effect under strongly leptokurtic score differential distributions is further amplified when using the mLINEX scoring function, to the point where even the sign test outperforms the DM test.

Moving to probabilistic forecasts evaluated with CRPS, tCRPS, and LS, we can observe a superior performance of the sign test in all scenarios, while the DM and the permutation test exhibit the lowest power. Apparently, the score-differential distribution is strongly leptokurtic, so that signed-rank and sign test outperform the other tests. Especially for the LS, the score differentials will be symmetric (see Example 5.1) for any value of $s \in [0, 1)$, while putting a strong emphasis on the tails. Size distortions can be observed for the DM test in the presence of high excess kurtosis under mLINEX and tCRPS, especially for t-distribution innovations.

In the presented scenario, it seems that the permutation test is always at least as good as the DM test for one-step ahead forecasts. When faced with strongly leptokurtic distributions it can leverage its robustness compared to plug-in estimates of mean and variance in the DM test. Of course, the advantages of the DM test are its computational efficiency and reproducibility, whereas Monte Carlo simulation becomes necessary for the permutation test somewhere between 15 and 30 observations. The suggestion of Harvey et al. (1997) to use a t-distribution as null distribution seems to be appropriate when the score differentials suggest a low excess kurtosis, but for some performance measures that induce high excess kurtosis this would exacerbate the problem of inflated p-values.

Serial correlation and multi-step ahead forecasts

While it is expected for the non-parametric tests to be competitive in a one-step ahead forecasting scenario, this cannot be said for multi-step ahead predictions. The DM test incorporates serial correlation by using a modified version of the Central Limit Theorem, while the discussed non-parametric tests are unable to take this into account. The loss of power from multiple testing on subchains using the Šidàk correction is substantial, yet the superiority of non-parametric tests in certain situations may sustain.

Figure 5.4 shows the results for two-step ahead forecasts with autoregressive parameter values 0 and 0.9, the latter of which introduces serial correlation. For one-step ahead forecasts, we observed that the DM and the permutation test are the ones performing best for the AE and the SE. As expected, the non-parametric tests fall off because of multiple testing. Only for the mLINEX, presumably in the presence of extreme outliers, will the permutation test be able to outperform the DM test. Evaluation using the CRPS also favors the DM test, whereas the results remain similar for the tCRPS and the LS. The superiority of the sign

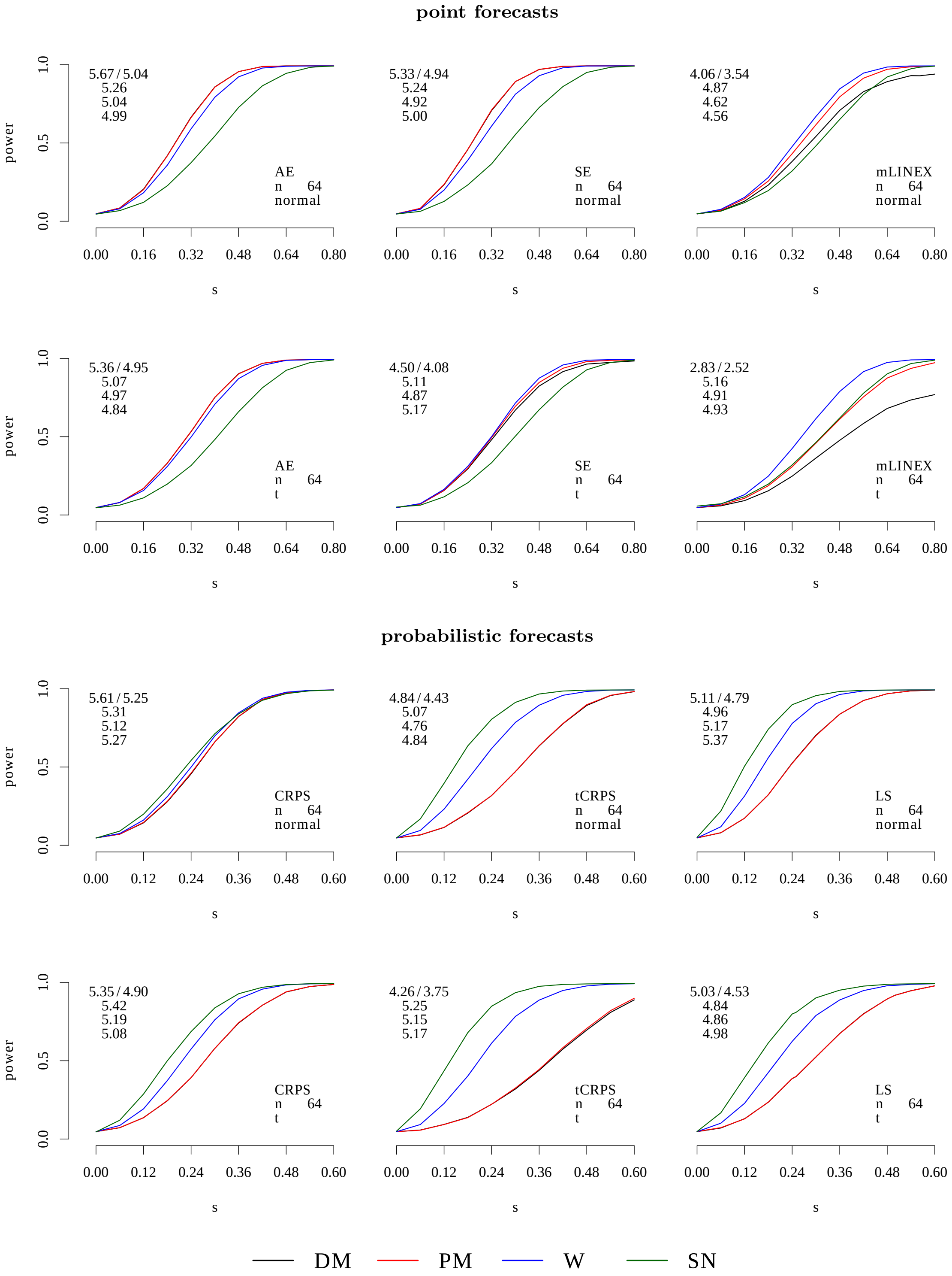


Figure 5.3: Plots of the adjusted power (5% level) for one-step ahead forecasts and normal/t_6 distributed innovations. Numbers in the top left corner indicate the actual size at the 5% level with two values for DM (based on normal/t-critical values). Black and red line may overlap completely.

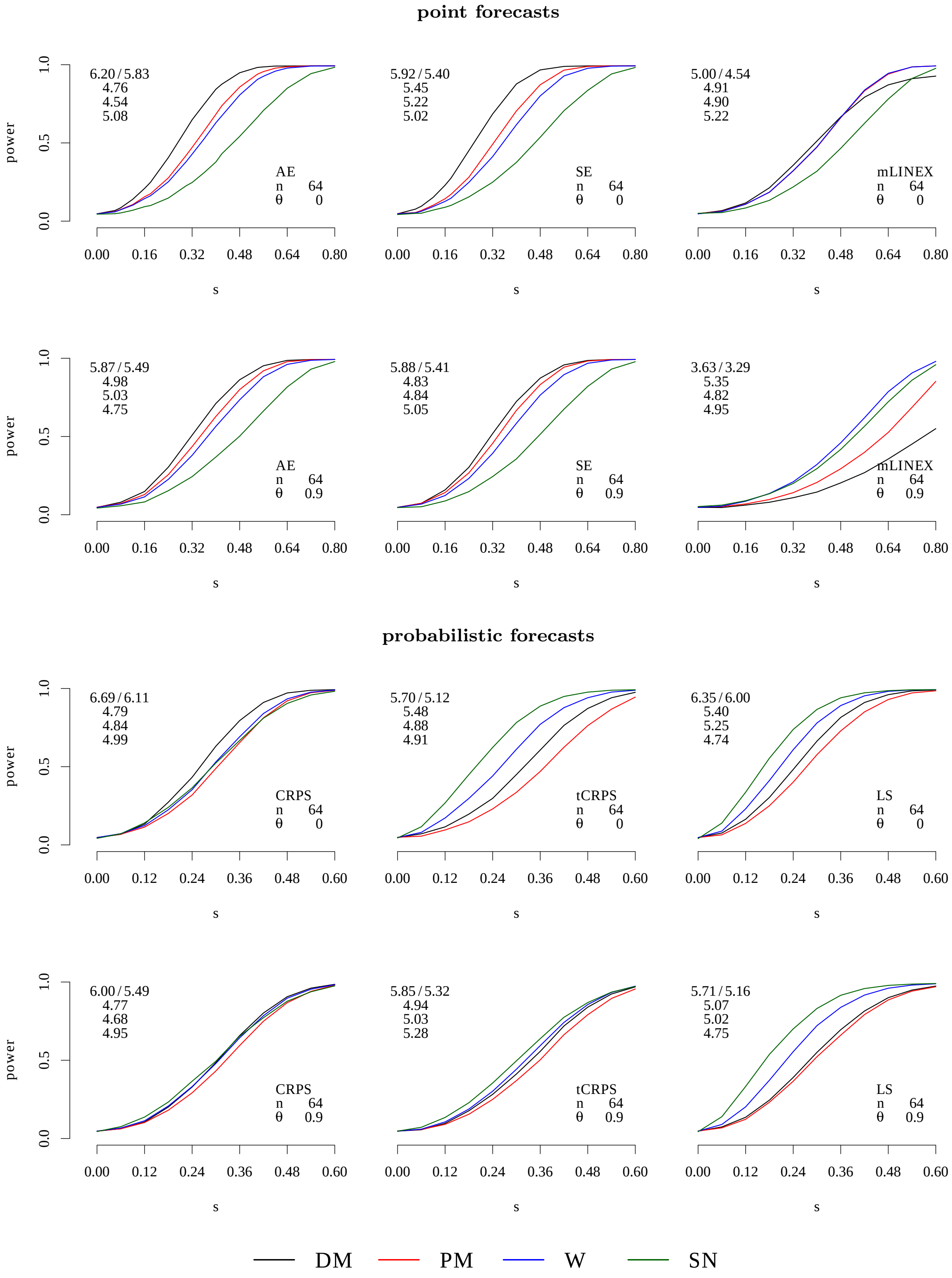


Figure 5.4: Plots of the adjusted power (5% level) for two-step ahead forecasts and normal distributed innovations. Numbers in the top left corner indicate the actual size at the 5% level with two values for DM (based on normal/t-critical values).

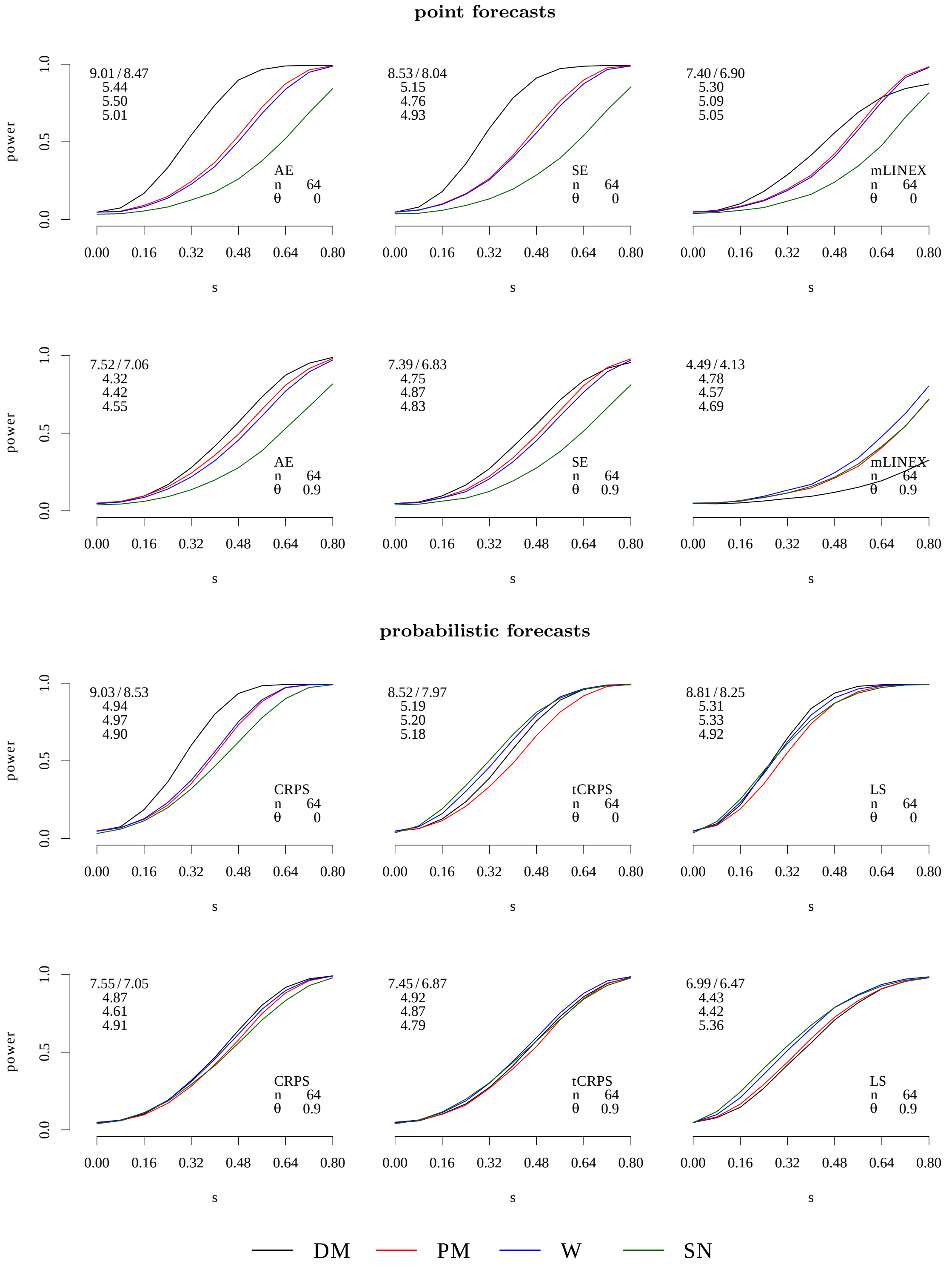


Figure 5.5: Plots of the adjusted power (5% level) for four-step ahead forecasts and normal distributed innovations. Numbers in the top left corner indicate the actual size at the 5% level with two values for DM (based on normal/t-critical values).

and signed-rank tests over the DM and permutation tests is big enough that the loss by multiple testing is not too detrimental. We also have to keep in mind that t-distributed innovations would generally sway considerations in favor of the nonparametric tests.

In Figure 5.5, for four-step ahead forecasts, the size-adjusted power of the DM test relative to the non-parametric tests increases again. With this further increase in size-adjusted power the DM test is clearly the most powerful for the AE, the SE, and the CRPS when $\theta = 0$. The mLINEX remains challenging for the DM test, and for the tCRPS and LS making out any differences in performance between the four tests becomes difficult.

Serial correlation plays an interesting role when testing for equal performance of multi-step ahead predictions. Apparently, the DM test is affected heavily by autocorrelation in comparison to the p-value corrections for multiple testing, which seem to be more robust against correlation. It seems we could make the point that the nonparametric tests stay competitive for multi-step ahead forecasts under high excess kurtosis and high autocorrelation. However, we do start to observe distinct distortions of the p-value distribution under the true null hypothesis. While the DM test detects significance too frequently, the non-parametric Šidàk-corrected tests are slightly conservative under high serial correlation ($\theta = 0.9$).

5.3 Empirical example: Bank of England projections of quarterly inflation rates

In the years from October 1992 until December 2003 the United Kingdom's target inflation rate was defined in terms of the RPIX index which is equivalent to the Retail Price Index (RPI) excluding mortgage interest payments. Over the entire period, the Bank of England had provided quarterly forecasts of the RPIX index. Quarters extend from March to May (Q1), June to August (Q2), September to November (Q3) and December to February (Q4). In 1997, the Bank of England's Monetary Policy Commitee (MPC) was handed the control over interest rates and this structural change also brought predictive densities as inflation projections following the second quarter. Previous forecasts have since been converted to implied probability distributions and added to the online data base (`http://www.bankofengland.co.uk/publications/Pages/inflationreport/irprobab.aspx`).

We revisit the results from Gneiting and Ranjan (2011) who compared the Bank of England's forecasts with those from a Gaussian AR(1) model which uses a rolling time window of length six for estimation. The inflation data exhibit skewness which the MPC's forecasts attempt to capture using a two-piece normal distribution with predictive density

$$f(y) = \begin{cases} \frac{\sqrt{2}}{\sqrt{\pi}(\sigma_1+\sigma_2)} \exp\left(-\frac{(y-\mu)^2}{2\sigma_1^2}\right) & \text{for } y \leq \mu \\ \frac{\sqrt{2}}{\sqrt{\pi}(\sigma_1+\sigma_2)} \exp\left(-\frac{(y-\mu)^2}{2\sigma_2^2}\right) & \text{for } y \geq \mu \end{cases}$$

where $\mu \in \mathbb{R}$ and $\sigma_1, \sigma_2 > 0$. For the comparison of predictive performance, we are

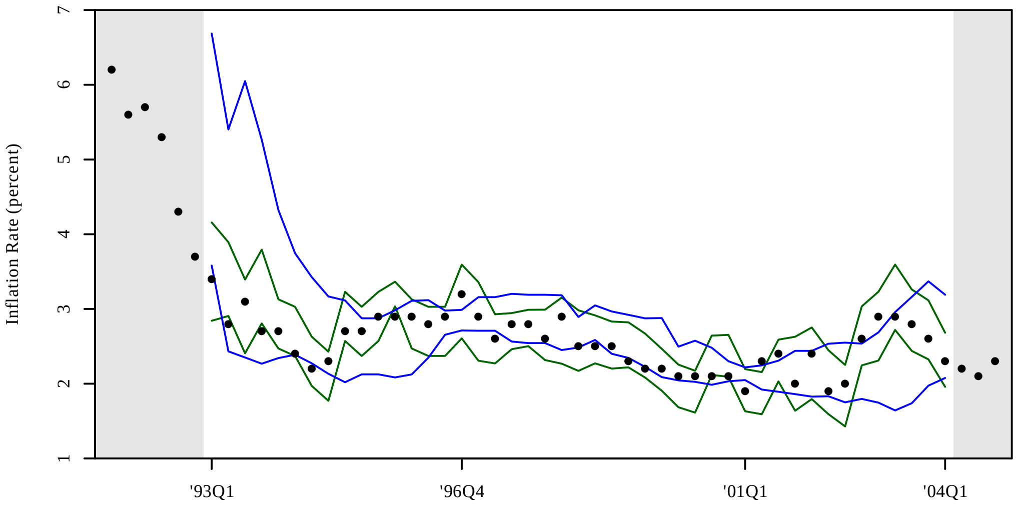


Figure 5.6: Inflation data and central 90% prediction intervals for one-step ahead forecasts of the Bank of England (green) and the Gaussian AR (1) model (blue).

provided with a data set of 45 such forecasts. Figure 5.6 shows the inflation data from 1991Q3 until 1994Q4 with the evaluation period for one-step ahead forecast starting 1993Q1 and ending 2004Q1. In the evaluation period, we provide the central 90 % prediction intervals of the forecasts from the Bank of England and the Gaussian AR(1) model. In the beginning, the Gaussian AR(1) model strongly overpredicts inflation as a result of unusually high inflation reports in the last 6 quarters before 1993Q1. Over the course of the evaluation period, the forecasts from this reference model are slightly lagging behind and as a result are often too sharp, leading to 13 (29%) out of 45 values falling outside the 90 % prediction intervals. The Bank of England's uncertainty is adequately appraised, and only 4 (9%) times an inflation value is within the outlying 10 % of their predictions. For longer forecast horizons as illustrated in Figure 5.7, we see that the Bank of England issued increasingly cautious forecasts with respect to their uncertainty in the period from 1993 until 2004, up to the point where no observations are within the outlying 10% of their predictions.

In Figure 5.7, the series of score-differentials $d_t = s_t^{\mathrm{BoE}} - s_t^{\mathrm{AR}}$ using the CRPS, tCRPS, and LS are provided for forecast horizons of one through four. Using a normal density with standard deviation 1 and centering that around the target inflation rate of 2.5%, we get $u(x) = 1 - \varphi_{2.5,1}(x)/\varphi(0)$ as the weighting function for the tCRPS. None of the score-differential series look very symmetric, with outlying negative scores favoring the Bank of England predictions especially for the tCRPS and the LS. Looking at the corresponding inflation data and the prediction intervals for these outlying values in Figure 5.6, we see that they are always the result of a sudden change which could not be predicted by the AR model, while the Bank of England forecasts adjust appropriately. The big difference in

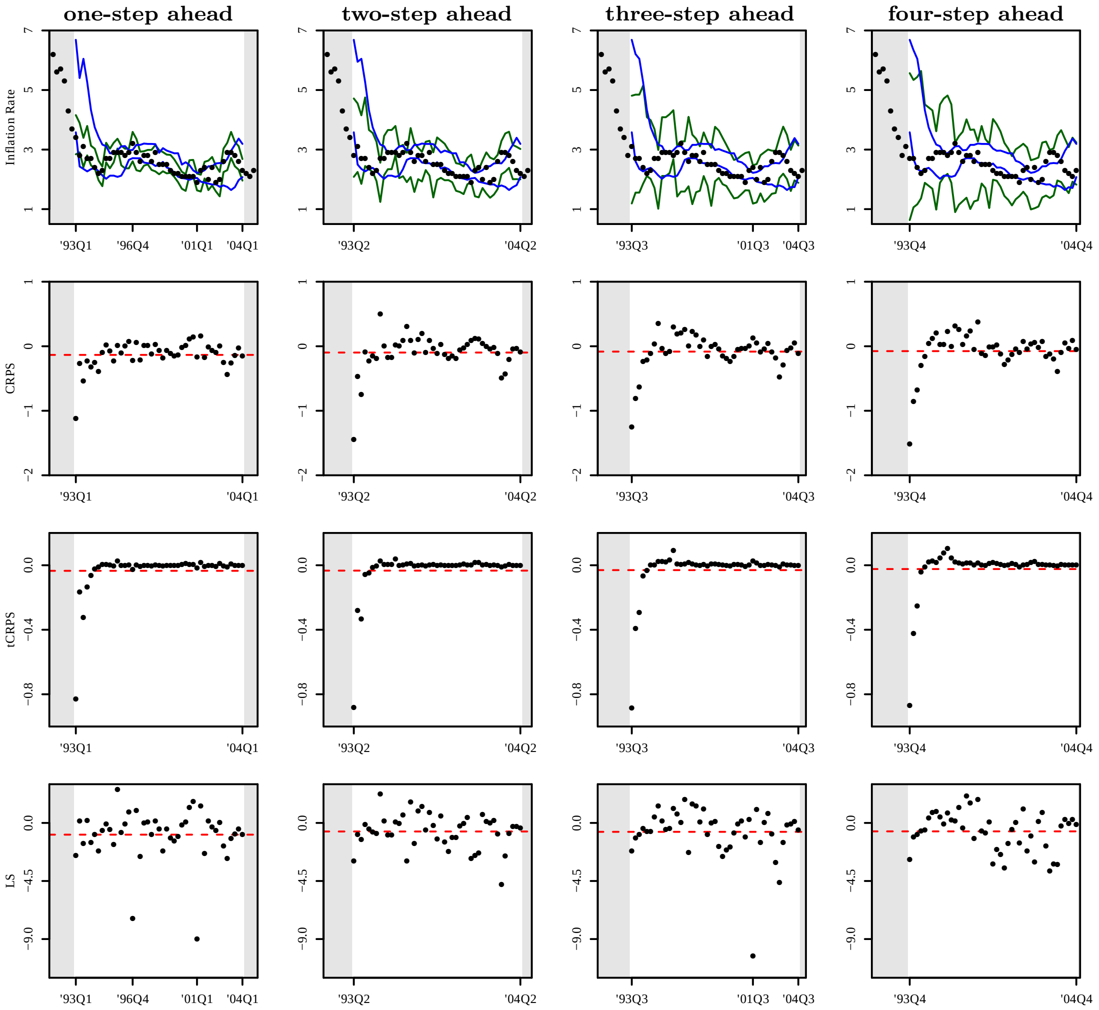


Figure 5.7: Observations, forecasts and score differences for the Bank of England and Gaussian AR(1) model forecasts with forecast horizons one through four, evaluated using CRPS, tCRPS, and LS. The score differentials' sample mean is indicated by the dashed red line, and negative values favor the Bank of England.

h	scoring rule	ex. kurt.	p_{DM}	p_{PM}	p_{W}	p_{SN}
1	CRPS	8.33	0.000	[0.000, 0.000]	[0.000, 0.000]	[0.002, 0.007]
	tCRPS	24.83	0.083	[0.015, 0.015]	[0.029, 0.030]	[0.007, 0.016]
	LS	6.38	0.002	[0.001, 0.001]	[0.001, 0.001]	[0.007, 0.016]
2	CRPS	8.52	0.100*	[0.093, 0.093]	[0.106, 0.114]	[0.250, 0.491]
	tCRPS	23.57	0.259*	[0.350, 0.350]	[0.812, 0.830]	[0.491, 0.773]
	LS	0.60	0.006*	[0.009, 0.009]	[0.009, 0.011]	[0.034, 0.102]
3	CRPS	5.42	0.274*	[0.252, 0.252]	[0.389, 0.426]	[0.102, 0.315]
	tCRPS	20.93	0.423*	[0.866, 0.866]	[0.426, 0.464]	[0.315, 0.660]
	LS	10.53	0.057*	[0.205, 0.205]	[0.320, 0.354]	[0.660, 0.939]
4	CRPS	8.27	0.403*	[0.851, 0.852]	[0.787, 0.838]	[0.642, 0.959]
	tCRPS	20.19	0.547*	[0.859, 0.859]	[0.243, 0.293]	[0.046, 0.237]
	LS	-0.78	0.115*	[0.237, 0.240]	[0.348, 0.409]	[0.642, 0.959]

* denotes the use of t-distribution critical values

Table 5.2: Comparison of Bank of England projections against Gaussian AR(1) process. The excess kurtosis is with respect to the score differential series. Weight function for tCRPS $u(x) = 1 - \phi_{2.5,1}(x)/\phi(0)$.

performance of the forecasts issued in '92Q4 for the respective forecast horizon arises only under scoring with CRPS and tCRPS and is a result of the strong absolute overprediction of the AR model in the beginning of the evaluation period. The only other extreme differences in forecast performance can be observed in '96Q4 and '01Q1 for one-step ahead forecasts, and in '01Q3 for three-step ahead forecasts, and only under scoring with the LS. These are strong relative under- and overpredictions of the AR model at a time when it issues particularly sharp forecasts, which illustrates nicely how tCRPS and LS emphasize tail-behavior. The tCRPS is static in that it focuses on performance in the region of interest, while the LS is more adaptive and penalizes events that are extreme with respect to the predictive density. This variability is why particularly poor scores under the LS may appear or disappear at different forecast horizons.

Table 5.2 shows the excess kurtosis of the score-differential series, and p-values for all four considered tests. For longer forecast horizons, we use t-distribution critical values for the DM test due to the tendency to overstate significance. In a nutshell, we observe the typical result that experts are significantly better than simple statistical models for one-step ahead forecasts, whereas their comparative skill fades for forecast horizons of multiple quarters. They may still be superior but the required sample size for such a deduction is distinctly higher.

We observe that the p-value from the DM test is typically smaller than those from the nonparametric tests with one notable exception. For one-step ahead forecasts that have been evaluated using the tCRPS, we observe an outlier in '93Q1 that is extreme enough that we find ourselves in the scenario discussed in Figure 5.2. This observation inflates the p-values of the DM test from a hypothetical value of 0.059, when the observation in question is excluded, to the actually observed value of 0.083. The influence of this extreme observation on the permutation test is a decrease from a hypothetical value of 0.029 to the observed 0.015, as we would expect from an observation that pulls the score differentials' sample mean further away from 0. In Figure 5.5 we have learned that the DM test will overstate significance for longer forecast horizons, so it is difficult to distinguish whether the comparatively small p-values are an artifact of this tendency or a result of being a more suitable test. However, we note an indication that the permutation test performs at least as good as the DM test for one-step ahead forecasts.

5.4 Discussion

In this chapter, we have discussed significance tests for equal predictive performance. The question these tests seek to answer is not which of two forecaster is superior, nor do they tell us how much better a forecast is in comparison to its competitor. If we test for significance, we are interested in the question if we have enough data to draw meaningful conclusions. Hence, the choice of level at which we deem a result significant is connected to our perception of meaningful. In a decision making scenario where waiting for additional data is unacceptable and one of two forecasters has to be chosen, performing a significance test will not help in making better decisions. The choice of forecaster must always be based on the score differentials' sample mean, but does not depend on whether the sample size is sufficiently large. In a way, we can interpret the level at which we test for significance as a measure of urgency with which a forecaster has to be credited with superior predictive ability.

In contrast, we could consider a situation where we have one operational forecasting method and the question is whether it should be replaced by an alternative. Performing a significance test for equal predictive performance would not answer the posed question. An appropriate testing procedure is an hypothesis test, e.g.

$$\begin{aligned} H_0 &: Y_i \sim F_{1,i} \\ H_1 &: Y_i \sim F_{2,i} \end{aligned} \qquad \text{for all } i = 1, \dots, n, \tag{5.13}$$

where the null hypothesis is that the observations can be interpreted as being drawn from the operational forecaster issuing probabilistic predictions $(F_{1,i})_{i=1}^n$, and the alternative is that the competitor provides a superior model. The likelihood ratio test (Neyman and Pearson 1933) is the most prominent example of such a test, and interestingly, the log-transform of its test statistic corresponds to the score differentials' sample mean under the logarithmic score. This means that a one-sided version of the discussed significance tests can be used as a hypothesis test of the scenario in (5.13). However, these tests should prove to be very conservative since the critical value for rejection of the null hypothesis would be based on a distribution with a mean value of zero, i.e. a distribution that lies somewhere between H_0 and H_1. Extensive investigation of these considerations lies outside the scope of this thesis.

We have discussed one asymptotic (Diebold-Mariano) and three non-parametric (permutation, Wilcoxon signed-rank, sign) tests, and have stated the necessary assumptions that allow the use of alternative test statistics in significance tests for equal predictive performance. We observed that the tests' performance is influenced by many factors, where the choice of scoring function is a central one that has not been studied extensively to our knowledge. It is just as important for the distribution of score differentials as the underlying distribution of the observations, since certain performance measures may introduce a large amount of excess kurtosis. For one-step ahead forecasts in small sample sizes it seems preferable to rely on nonparametric tests due to the sample variance's volatility. In particular, the permutation test seems to perform at least equally well with distinct advantages under heavy tails and high excess kurtosis. Of course, the DM test retains its broad applicability especially for multi-step ahead predictions and score differentials with platykurtic distributions.

For the nonparametric tests, the discreteness of p-values and the presence of serial correlation have to be accommodated. Further investigations could be made regarding the viability of DM type tests under wide-sense stationarity of the transformed score differentials leading to statistics of the type ${}_{\psi}\bar{d}_n$. Considerations in this direction have already been made by Dell'Aquila and Ronchetti (2004).

Lastly, significance tests based on Murphy diagrams using the bootstrap or the Westfall-Young method (Westfall and Young 1993; Cox and Lee 2008) deserve further investigation. Also, the possibility of hypothesis tests for forecast dominance seems intriguing.

6 Score Computation

“Computing the continuous ranked probability score is challenging.”

Primarily, the choice of performance measure is determined by the type of forecast, and secondly, ease of implementation and interpretability. In this context, the continuous ranked probability score (Matheson and Winkler 1976) is particularly interesting because it is admissable for any forecast that can be represented as a univariate cumulative distribution function. This includes the case of dichotomous events, some categorical events, and any event measured by a real-valued linear variable. It belongs to the class of kernel scores (Eaton 1982; Gneiting and Raftery 2007) which work for forecasts valued in any metric space. However, they are notoriously difficult to compute for non-discrete distributions. Often the only remaining options for computing the scores are numerical methods, particularly for multivariate forecasts.

6.1 Closed form expressions for the continuous ranked probability score

Using the CRPS, probabilistic forecasts may be issued as predictive cumulative distribution functions, allowing mixtures of discrete and continuous distributions as predictions. For a Dirac measure in a single point, the CRPS collapses to the absolute error. There are three well-known representations (Gneiting and Raftery 2007; Matheson and Winkler 1976; Laio and Tamea 2007) that are interesting for their interpretability and diagnostic value,

$$\begin{aligned}\mathrm{CRPS}(F,y) &= \mathbb{E}_F|X-y| - \frac{1}{2}\mathbb{E}_F|X-X'| && (6.1)\\ &= \int_{-\infty}^{\infty} (F(\theta) - \mathbb{1}(y \le \theta))^2 \,\mathrm{d}\theta && (6.2)\\ &= 2\int_0^1 (\mathbb{1}(y < Q(\alpha)) - \alpha)(Q(\alpha) - y)\,\mathrm{d}\alpha, && (6.3)\end{aligned}$$

where X and X' are independent random variables with cumulative distribution function F and finite first moment, and Q is any generalized inverse to F, typically mapping α to $q^{-}_{\alpha,F}$ as defined in section 3.1.1. Based on the elementary scores for quantile forecasts, we have given another representation in (3.54), which can be rewritten in terms of Lebesgue-Stieltjes integrals in the following ways,

$$\begin{aligned}\mathrm{CRPS}(F,y) &= 2\int_{-\infty}^{\infty}\int_0^1 \mathrm{S}^{\mathrm{Q}}_{\alpha,\theta}(Q(\alpha),y)\,\mathrm{d}\alpha\,\mathrm{d}\theta \\ &= \int_{-\infty}^{y}\int_0^{F(y)} \mathbb{1}(Q(\alpha)\le\theta)\,\mathrm{d}\alpha^2\,\mathrm{d}\theta - \int_y^{\infty}\int_{F(y)}^1 \mathbb{1}(\theta<Q(\alpha))\,\mathrm{d}(1-\alpha)^2\,\mathrm{d}\theta \quad (6.4)\\ &= \int_{-\infty}^{y}\int_{-\infty}^{y} \mathbb{1}(x\le\theta)\,\mathrm{d}F^2(x)\,\mathrm{d}\theta - \int_y^{\infty}\int_y^{\infty} \mathbb{1}(\theta<x)\,\mathrm{d}(1-F)^2(x)\,\mathrm{d}\theta. \quad (6.5)\end{aligned}$$

Integration of either equation with respect to α and x, respectively, recovers the threshold decomposition of the CRPS. If we integrate with respect to θ first, then (6.4) results in the quantile decomposition, while (6.5) gives a representation in between the threshold decomposition, the quantile decomposition, and the kernel representation. Due to the representation as a mixture of elementary scores, we recover representations in terms of the cumulative distribution function and the quantile function in a straight-forward manner.

A representation of the CRPS in terms of characteristic functions can be given using the kernel representation (6.1) and the identity

$$\mathbb{E}_F|X| = \frac{2}{\pi}\int_0^{\infty}\frac{1-\Re\,\phi_X(t)}{t^2}\,\mathrm{d}t,$$

where $\Re$ denotes the real part, and $\phi_X(t)=\mathbb{E}_F[e^{itX}]$ is the characteristic function of the random variable X with distribution F. This is a special case of a result by Brown (1972) based on Hsu (1951) and von Bahr (1965). We obtain

$$\begin{aligned}\mathrm{CRPS}(F,y) &= \frac{1}{\pi}\int_0^{\infty}\frac{\Re\,[1-2\phi_{X-y}(t)+\phi_{X-X'}(t)]}{t^2}\,\mathrm{d}t,\\ &= \frac{1}{\pi}\int_0^{\infty}\frac{|\phi_X(t)-\phi_y(t)|^2}{t^2}\,\mathrm{d}t, \quad (6.6)\end{aligned}$$

where the same formula can be recovered as a special case of the energy score (Gneiting and Raftery 2007) with reference to Székely and Rizzo (2013). Sometimes we may need to calculate the CRPS for distributions where the cumulative distribution function cannot be given in closed form, and even numerical approximation may be challenging. These scenarios can arise for distributions of sums of independent random variables, e.g., for multi-step ahead forecasts in auto-regressive time-series models.

6.1.1 Transformations

Here, we will introduce the unconventional notation

$$\mathrm{CRPS}_F(X,y) = \mathrm{CRPS}(F,y),$$

where X has distribution F, which helps clarify why the CRPS can be described as translation invariant and homogeneous. The most common tools to introduce additional parameters for a given kernel are the location-scale transformation and the truncation of a distribution's support. Let $\mu \in \mathbb{R}$ and $\sigma > 0$ be location and scale parameters in the transformation $h(x) = \frac{x-\mu}{\sigma}$, and let $I = [a, b)$ be a half-open interval with $a, b \in \mathbb{R} \cup \{-\infty, \infty\}$, $a < b$. This restriction on the support leads to truncated versions of the cumulative distribution function and the observation,

$$F_I(x) = \begin{cases} 0, & x < a, \\ F(x), & x \in I, \\ 1, & x \geq b, \end{cases} \qquad \text{and} \qquad y_I = \begin{cases} a, & y < a, \\ y, & y \in I, \\ b, & y \geq b. \end{cases}$$

Combining the two transformations, $h(I)$ denotes the transformed interval with endpoints $h(a)$ and $h(b)$.

Properties 6.1. All of the following properties follow straight-forwardly from the threshold decomposition.

(a) The CRPS is invariant under translation,

$$\mathrm{CRPS}_F(X + \mu, y + \mu) = \mathrm{CRPS}_F(X, y). \tag{6.7}$$

(b) The CRPS is homogeneous of order 1,

$$\mathrm{CRPS}_F(\sigma X, \sigma y) = \sigma \mathrm{CRPS}_F(X, y) \tag{6.8}$$

(c) The CRPS for a location-scale family generated by F and h is given by

$$\mathrm{CRPS}(F \circ h, y) = \sigma \mathrm{CRPS}(F, h(y)). \tag{6.9}$$

(d) Let $\mathcal{I} = (I_i)_{i=1,2,\ldots}$ be a cover of $\mathbb{R}$, such that the $I_i = [a_i, b_i)$ are half-open, pairwise disjoint intervals. The CRPS can then be calculated as

$$\mathrm{CRPS}(F, y) = \sum_{I \in \mathcal{I}} \mathrm{CRPS}(F_I, y_I). \tag{6.10}$$

(e) Introducing a location-scale transformation on the underlying cumulative distribution function F without changing the cover $\mathcal{I}$ of $\mathbb{R}$ can be reformulated as keeping F and transforming the cover and observation,

$$\mathrm{CRPS}(F \circ h, y) = \sigma \sum_{I \in \mathcal{I}} \mathrm{CRPS}(F_{h(I)}, h(y_I)). \tag{6.11}$$

Properties (a) through (c) allow us to derive the CRPS for members of location-scale families from the standardized versions, e.g. the non-standardized Student's t-distribution or a transformed beta distribution for variables on an interval of finite length. Properties (d) and (e) suggest how to piece together distributions from censored versions, e.g., cumulative distribution functions that are step functions, or the two-piece normal distribution.

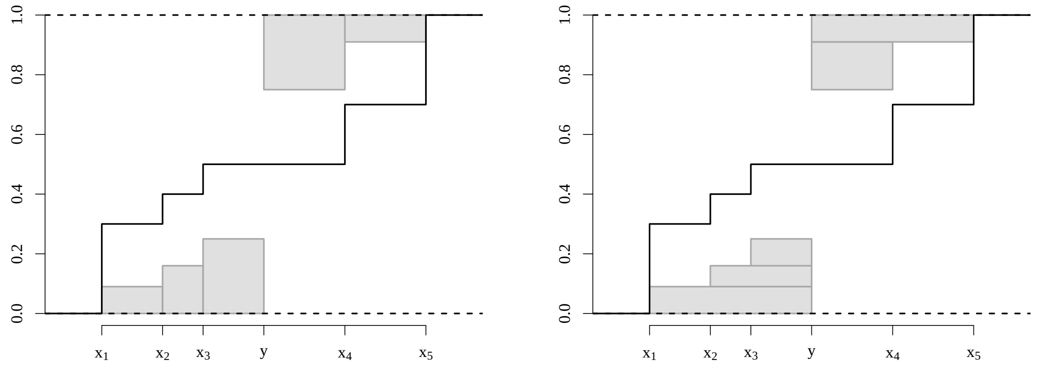


Figure 6.1: Cumulative distribution function with jumps in $\{x_1, \dots x_5\}$. The grey area corresponds to the CRPS with respect to the observation y. Left: Summation (6.12) as given by Hersbach (2000). Right: Summation (6.13) as given by Murphy (1970).

6.1.2 Closed form expressions

One important case of a closed-form expression is for discrete distributions with finitely many events $\{x_1, ..., x_m \in \mathbb{R} : x_i < x_{i+1}\}$, i.e., when the cumulative distribution function is a step function with a finite number of jumps. The class $\mathcal{P}$ of probability distributions comprises the cumulative distribution functions of the type

$$P(x) = \sum_{i=1}^{m} p_i \, \mathbb{1}\{x \geq x_i\},$$

where $p_1, ..., p_m \geq 0$ are the probabilities corresponding to $x_1, .., x_m$. Hersbach (2000) and Murphy (1970) give two convenient representations[1] that can be computed exactly,

$$\begin{aligned} \mathrm{CRPS}(P, y) &= \sum_{i|x_i<y} l_i \Big(\sum_{j \leq i} p_j \Big)^2 + \sum_{i|x_i>y} l_i \Big(\sum_{j \geq i} p_j \Big)^2 && (6.12) \\ &= \sum_{i|x_i<y} (y - x_i) \Big(p_i^2 + 2p_i \sum_{j<i} p_j \Big) + \sum_{i|x_i>y} (x_i - y) \Big(p_i^2 + 2p_i \sum_{j>i} p_j \Big) && (6.13) \end{aligned}$$

where l_i is the length of the interval generated by the i-th component and its successor in the sorted vector comprising of entries $x_1, ..., x_m, y$. Figure 6.1 visualizes these two summation formulas of the CRPS, which are Riemann sums corresponding to the treshold and the quantile decomposition, respectively.

Clearly, some representations of the CRPS require less effort than others when the goal is to find closed-form expressions for a given distribution. The kernel representation is useful if the Gini coefficient of the predictive distribution is

[1] Murphy (1970, Section 4) gives (6.13) in the context of the ranked probability score (RPS), but the generalization to the continuous RPS (CRPS) is straight-forward.

already known, and the threshold decomposition may be the easiest if the square interacts naturally with the specific form of the cumulative distribution function. Otherwise, we will usually use one of the following representations, which are again in the terms of Lebesgue-Stieltjes integrals,

$$\begin{aligned}
\mathrm{CRPS}(F, y) &= y(2F(y) - 1) + 2\int_{F(y)}^{1} Q(\alpha)\,\mathrm{d}\alpha - \int_0^1 Q(\alpha)\,\mathrm{d}\alpha^2 && (6.14)\\
&= y(2F(y) - 1) + 2\int_y^{\infty} x\,\mathrm{d}F(x) \; - \int_{-\infty}^{\infty} x\,\mathrm{d}F^2(x) && (6.15)\\
&= y(2F(y) - 1) + 2\int_y^{\infty} x\,\mathrm{d}F(x) \; - \int_{-\infty}^{\infty} \Big(F(x) + F(x^-)\Big)\,\mathrm{d}G(x) && (6.16)\\
&= y(2F(y) - 1) - 2\int_{-\infty}^{y} x\,\mathrm{d}F(x) \; + \int_{-\infty}^{\infty} \Big(G(x) + G(x^-)\Big)\,\mathrm{d}F(x), && (6.17)
\end{aligned}$$

where $G(x) = \int_{-\infty}^{x} t\,\mathrm{d}F(t)$, Q is a generalized inverse to F, and $F(x^-)$ denotes the left-sided limit of F in x. Equations (6.14) and (6.15) follow directly from (6.4) and (6.5) by integration with respect to θ. Alternatively, calculation of the quantile decomposition, or using integration by parts for Stieltjes integrals on the threshold decomposition yields equivalent results. Using the definitions of the Stieltjes integral and the function G, we can transform (6.15) into (6.16), which can then be transformed into (6.17) using integration by parts. Again, the alternative route of calculating the convolution in the kernel representation leads to the same result. Equation (6.14) was already stated by Friederichs and Thorarinsdottir (2012), and for the case of continuous cumulative distribution functions Taillardat et al. (2016) give a version of equations (6.15) and (6.16), namely,

$$\mathrm{CRPS}(F, y) = \mathbb{E}_F\Big(|Y - y| + Y - 2YF(Y)\Big).$$

Example 6.1 (Beta distribution). The cumulative distribution function of a beta distribution with parameters $\alpha, \beta > 0$ is given by

$$F_{\alpha,\beta}(x) = \begin{cases} 0 & x < 0 \\ I_x(\alpha, \beta) & 0 \le x < 1 \\ 1 & x \ge 1 \end{cases},$$

where I_x denotes the regularized incomplete beta function. We can find a closed form expression for the CRPS using the beta function B,

$$\mathrm{CRPS}(F_{\alpha,\beta}, y) = y(2F_{\alpha,\beta}(y) - 1) + \tfrac{\alpha}{\alpha+\beta}\Big(1 - 2F_{\alpha+1,\beta}(y) - \tfrac{2B(2\alpha,2\beta)}{\alpha B(\alpha,\beta)^2}\Big).$$

Proof. We use a simplified version of (6.17), namely

$$\mathrm{CRPS}(F, y) = y(2F(y) - 1) - 2\int_{-\infty}^{y} x\,\mathrm{d}F(x) \; + 2\int_{-\infty}^{\infty} G(x)\,\mathrm{d}F(x).$$

For the beta distribution we have

$$G(y) = \int_{-\infty}^{y} x f_{\alpha,\beta}(x)\,\mathrm{d}x = \tfrac{\alpha}{\alpha+\beta} F_{\alpha+1,\beta}(y),$$

and

$$\begin{aligned}\int_{-\infty}^{\infty} G(x)\,\mathrm{d}F(x) &= \tfrac{\alpha}{\alpha+\beta}\int_0^1 F_{\alpha+1,\beta}(x) f_{\alpha,\beta}(x)\,\mathrm{d}x \\ &= \tfrac{\alpha}{\alpha+\beta} - \tfrac{\alpha}{\alpha+\beta}\int_0^1 F_{\beta,\alpha+1}(x) f_{\beta,\alpha}(x)\,\mathrm{d}x \\ &= \tfrac{\alpha}{\alpha+\beta} - \tfrac{1}{B(\alpha,\beta)^2}\int_0^1 x^{\beta-1}(1-x)^{\alpha-1} B_x(\beta,\alpha+1)\,\mathrm{d}x \\ &= \tfrac{\alpha}{\alpha+\beta} - \tfrac{\alpha}{2(\alpha+\beta)} - \tfrac{B(2\alpha,2\beta)}{(\alpha+\beta)B(\alpha,\beta)^2}.\end{aligned}$$

The last step can be seen using equation (8.17.8) from the NIST Digital Library of Mathematical Functions (Olver et al. 2016) and equation (7.512.5) in Gradshteyn and Ryzhik (2007),

$$\begin{aligned}&\int_0^1 x^{\beta-1}(1-x)^{\alpha-1} B_x(\beta,\alpha+1)\,\mathrm{d}x \\ &= \tfrac{1}{\beta}\int_0^1 x^{2\beta-1}(1-x)^{2\alpha}\ {}_2F_1(1,\alpha+\beta+1;\beta+1;x)\,\mathrm{d}x \\ &= \tfrac{\alpha}{\alpha+\beta}\tfrac{B(2\alpha,2\beta)}{\beta}\ {}_3F_2(1,2\beta,\alpha+\beta+1;\beta+1,2\alpha+2\beta+1;1).\end{aligned}$$

Equation (1) in Lavoie (1987) gives an explicit formula for the hypergeometric function in terms of gamma functions, which can be further simplified using the duplication formula,

$$\begin{aligned}&{}_3F_2(1,2\beta,\alpha+\beta+1;\beta+1,2\alpha+2\beta+1;1) \\ &= \frac{2^{1+2\beta}\Gamma(\beta+1)\Gamma(\alpha+\beta+\frac{1}{2})\Gamma(\alpha)}{\Gamma(\frac{1}{2})\Gamma(2\beta+1)} \\ &\quad\times\left(\frac{\Gamma(\frac{3}{2})\Gamma(\beta+1)}{\Gamma(\alpha+\beta)\Gamma(\alpha+\frac{1}{2})} + \frac{2\beta\Gamma(\beta+\frac{1}{2})}{4\Gamma(\alpha+\beta+\frac{1}{2})\Gamma(\alpha+1)}\right) \\ &= \frac{\beta B(\alpha,\beta)^2}{2B(2\alpha+2\beta)} + \frac{\beta}{\alpha}.\end{aligned}$$

□

Example 6.2 (truncation/censoring). Let F be the cumulative distribution and f be the Lebesgue density of a continuous distribution. Restricting the support to $I = [a, b)$, in combination with an affine transformation using parameters $c, d \in \mathbb{R}$, yields the cumulative distribution function

$$F_I(x) = \begin{cases} 0, & x < a, \\ cF(x) + d, & x \in I, \\ 1, & x \geq b, \end{cases}$$

where the parameters c and d must satisfy

$$0 \leq cF(a) + d \leq cF(b) + d \leq 1.$$

Let $P_x = F_I(x) - F_I(x^-)$ and $G(x) = \int_{-\infty}^{x} tf(t)\,\mathrm{d}t$, then we can find the following representation of the CRPS,

$$\begin{aligned} \operatorname{CRPS}(F_I, y) = {} & y(2F_I(y) - 1) - aP_a^2 + bP_b^2 + 2cG(a)P_a + 2cG(b)P_b \\ & - 2 \begin{cases} cG(a) - aP_a, & y < a, \\ cG(y), & a \leq y < b, \\ cG(b) + bP_b, & y \geq b, \end{cases} \\ & + 2c^2 \int_a^b G(x)f(x)\,\mathrm{d}x. \end{aligned} \tag{6.18}$$

Proof. This result can be shown by isolation of the point masses in a and b from the representation (6.17),

$$\operatorname{CRPS}(F_I, y) = y(2F_I(y) - 1) - 2\int_{-\infty}^{y} x\,\mathrm{d}F_I(x) + \int_{-\infty}^{\infty} G_I(x) + G_I(x^-)\,\mathrm{d}F_I(x),$$

using the definition of Stieltjes integrals, where

$$\begin{aligned} G_I(x) &= \int_{-\infty}^{x} t\,\mathrm{d}F_I(t) \\ &= aP_a - cG(a) + \begin{cases} cG(a) - aP_a, & x < a, \\ cG(y), & a \leq x < b, \\ cG(b) + bP_b, & x \geq b. \end{cases} \end{aligned}$$

and

$$\begin{aligned} \int_{-\infty}^{\infty} G_I(x) + G_I(x^-)\,\mathrm{d}F_I(x) = {} & aP_a^2 + P_b(2aP_a - 2cG(a) + 2cG(b) + bP_b) \\ & + 2(1 - P_b - P_a)(aP_a - cG(a)) \\ & + 2c^2 \int_a^b G(x)f(x)\,\mathrm{d}x. \end{aligned}$$

□

Equation (6.18) can be leveraged for numerous distributions including the normal distribution and Student's t-distribution. For the standard normal distribution with CDF Φ and density φ, we know that $G(x) = -\varphi(x)$ and

$$\int_a^b G(x)\varphi(x)\,\mathrm{d}x = -\int_a^b \varphi^2(x)\,\mathrm{d}x = -\left[\frac{\Phi(x\sqrt{2})}{2\sqrt{\pi}}\right]_a^b.$$

Similarly, the cumulative distribution function and probability density function for Student's t-distribution with $\nu > 0$ degrees of freedom are given by

$$\begin{aligned} F_\nu(x) &= \frac{1}{2} + \frac{x\;{}_2F_1(\frac{1}{2}, \frac{\nu+1}{2}; \frac{3}{2}; -\frac{x^2}{\nu})}{\sqrt{\nu}B(\frac{1}{2}, \frac{\nu}{2})}, \\ f_\nu(x) &= \frac{1}{\sqrt{\nu}B(\frac{1}{2}, \frac{\nu}{2})}\left(1 + \frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}, \end{aligned}$$

where the symbol B denotes the beta function and ${}_2F_1$ denotes the hypergeometric function. For $\nu > 1$, we have

$$G_\nu(y) = \int_{-\infty}^y x f_\nu(x)\,\mathrm{d}x = -\frac{\nu}{\nu - 1}\left(1 + \tfrac{y^2}{\nu}\right) f_\nu(y),$$

leading to

$$\begin{aligned} \int_a^b G_\nu(x) f_\nu(x)\,\mathrm{d}x &= -\frac{\nu}{\nu - 1}\int_a^b \left(1 + \tfrac{x^2}{\nu}\right) f_\nu(x)^2\,\mathrm{d}x \\ &= -\frac{1}{(\nu - 1)B(\frac{1}{2}, \frac{\nu}{2})^2}\int_a^b \left(1 + \tfrac{x^2}{\nu}\right)^{-\nu}\mathrm{d}x \\ &= -\frac{\sqrt{\nu}B(\frac{1}{2}, \nu - \frac{1}{2})}{2(\nu - 1)B(\frac{1}{2}, \frac{\nu}{2})^2}\left[\operatorname{sgn}(t) I_{t^2/(\nu+t^2)}(\tfrac{1}{2}, \nu - \tfrac{1}{2})\right]_a^b, \end{aligned}$$

where I_x denotes the regularized incomplete beta function. The identity can be shown by splitting the integral at zero (if $a < 0$ and $b > 0$) and using integration by substitution with

$$1 - t = \left(1 + \tfrac{x^2}{\nu}\right)^{-1}.$$

While closed form expressions of the CRPS for the normal distribution and derivations using truncation and censoring are well known (Gneiting et al. 2005, 2006; Thorarinsdottir and Gneiting 2010; Gneiting and Thorarinsdottir 2010), formula (6.18) covers truncation and censoring with arbitrary redistribution of the tail probabilities, including the special cases

(a) of the original distribution, i.e. $a = -\infty, b = \infty, c = 1, d = 0$,

(b) of the censored version with concentration of the tail probabilities into point masses in the respective boundary, i.e. $-\infty \le a < b \le \infty, c = 1, d = 0$,

(c) of the truncated version with proportional redistribution of the total tail probability, i.e. $-\infty \leq a < b \leq \infty$,

$$c = \frac{1}{F(b) - F(a)}, \qquad d = -\frac{F(a)}{F(b) - F(a)},$$

(d) and of the two-piece distributions with location parameter μ and scale parameters $\sigma_1, \sigma_2 > 0$,

$$F(x) = \begin{cases} F_1(x), & x < \mu, \\ F_2(x), & x \geq \mu, \end{cases}$$

where, for appropriate c_1, c_2, d_2,

$$F_1(x) = \begin{cases} c_1 F(\frac{x-\mu}{\sigma_1}), & x < \mu, \\ 1, & x \geq \mu, \end{cases}$$

$$F_2(x) = \begin{cases} 0, & x < \mu, \\ c_2 F(\frac{x-\mu}{\sigma_2}) + d_2, & x \geq \mu, \end{cases}$$

the CRPS can be calculated as

$$\mathrm{CRPS}(F, y) = \begin{cases} \mathrm{CRPS}(F_1, y) + \mathrm{CRPS}(F_2, \mu), & y < \mu, \\ \mathrm{CRPS}(F_1, \mu) + \mathrm{CRPS}(F_2, y), & y \geq \mu, \end{cases}$$

using property 6.1(e).

6.2 Approximations for the continuous ranked probability score

Approximation of the CRPS becomes necessary if a closed form expression is unattainable. While analytical solutions can be found for many of the standard distributions, this is not the case in general. We will gradually move from almost-closed forms to cases where the CRPS needs to be estimated from a sample. It stands to reason that as much analytical information as possible should be used at all times.

Numerical integration

In the best case scenario, we are still able to numerically approximate the CRPS using one of the representations in this chapter. For simplicity we will assume that the predictive distribution is continuous. We consider several functions which can be used to fully describe a distribution. The most standard ones are the cumulative distribution function F and its inverse or quantile function Q, but also the probability density function f and the quantile density function q. The cumulative expectation functions $G(x) = \int_{-\infty}^{x} tf(x)\,\mathrm{d}x$ and $\mathcal{Q}(\alpha) = \int_0^\alpha Q(t)\,\mathrm{d}t$, or the characteristic function $\phi_X(t) = \mathbb{E}_F[e^{itX}]$ may also be used. This gives us the two following groups of CRPS representations which are useful for numerical integration,

$$\begin{aligned} \mathrm{CRPS}(F, y) &= \int_{-\infty}^{y} F(x)^2\,\mathrm{d}x + \int_{y}^{\infty} (1 - F(x))^2\,\mathrm{d}x && (6.19) \\ &= y(2F(y) - 1) - 2\int_{-\infty}^{y} xf(x)F(x)\,\mathrm{d}x + 2\int_{y}^{\infty} xf(x)(1 - F(x))\,\mathrm{d}x && (6.20) \\ &= y(2F(y) - 1) - 2G(y) + 2\int_{-\infty}^{\infty} f(x)G(x)\,\mathrm{d}x && (6.21) \\ &= \frac{1}{\pi}\int_0^\infty \frac{|\phi_X(t) - \phi_y(t)|^2}{t^2}\,\mathrm{d}t, && (6.22) \end{aligned}$$

and

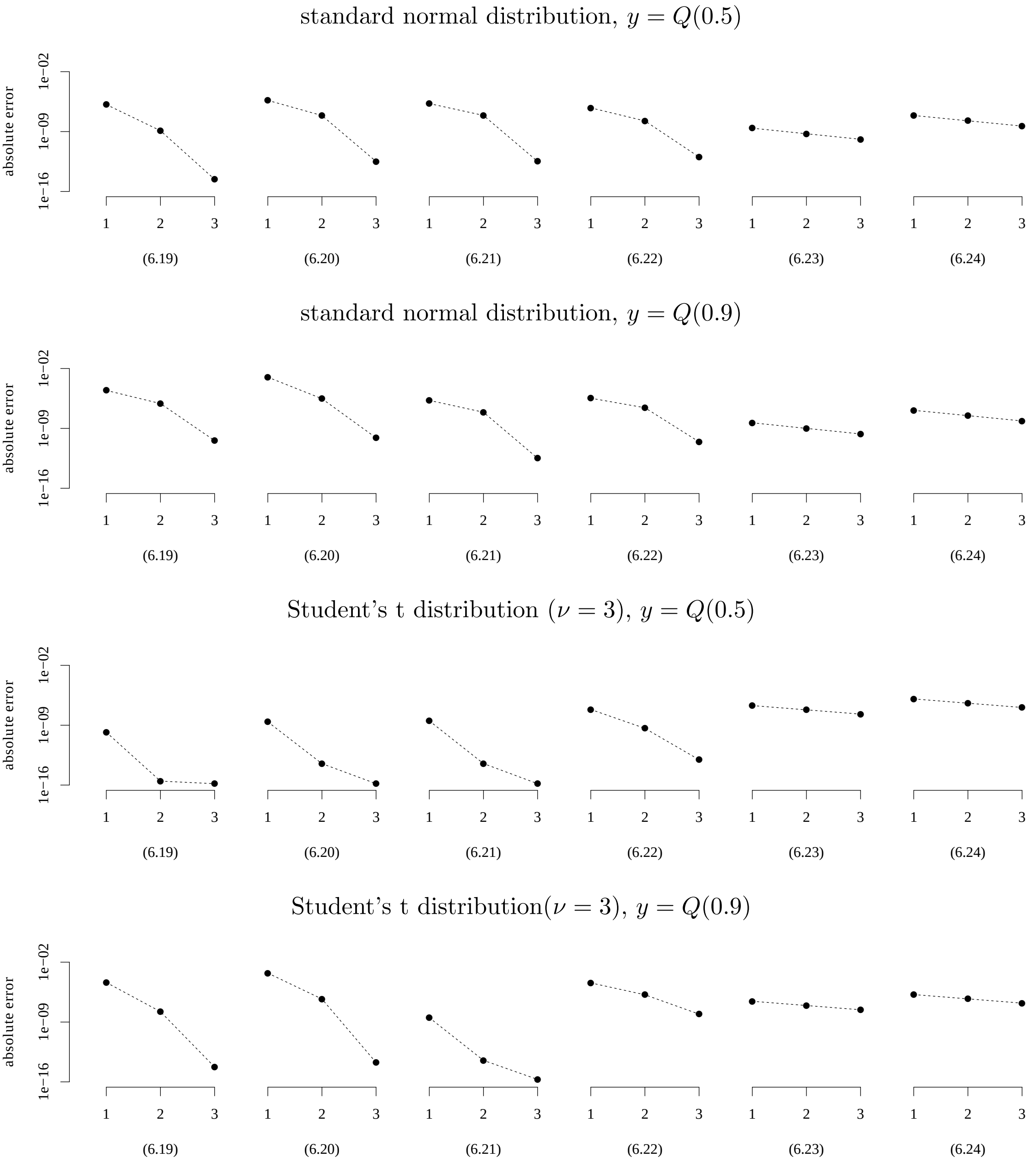
$$\begin{aligned} \mathrm{CRPS}(F, y) &= \int_0^{F(y)} \alpha^2 q(\alpha)\,\mathrm{d}\alpha + \int_{F(y)}^{1} (1 - \alpha)^2 q(\alpha)\,\mathrm{d}\alpha && (6.23) \\ &= y(2F(y) - 1) - 2\int_0^{F(y)} \alpha Q(\alpha)\,\mathrm{d}\alpha + 2\int_{F(y)}^{1} (1 - \alpha)Q(\alpha)\,\mathrm{d}\alpha && (6.24) \\ &= y(2F(y) - 1) - 2\mathcal{Q}(F(y)) + 2\int_0^1 \mathcal{Q}(\alpha)\,\mathrm{d}\alpha. && (6.25) \end{aligned}$$

Nearly all representations require a closed form of the cumulative distribution function F, except for equation (6.22) using solely the characteristic function. Apart from that, the first group contains only the functions f, F, and G with the corresponding infinite integrals, whereas the second group requires calculation of merely finite integrals using the functions q, Q, and $\mathcal{Q}$. By means of integration by parts we can navigate through the representations (6.19) to (6.21) and (6.23) to (6.25), and the substitution $\alpha = F(x)$ allows moving between the groups. The representations (6.21) and (6.25) facilitate storage of the integral's numerical result when calculating the CRPS for different observations, or when the class of predictive distributions happens to be a location-scale family (see 6.1(c)).

In Figure 6.2, we illustrate the behavior of representations (6.19) to (6.24) in a non-comprehensive way. We choose the normal distribution and Student's t-distribution as examples, which means passing over representation (6.25) for lack of a closed form or numerical implementation. The quantile density function, on the other hand, can be represented as

$$q(\alpha) = \frac{1}{f(Q(\alpha))}.$$

Naturally, these results depend strongly on the distribution and numerical integration routine, which in this case is the QUADPACK implementation used in the `integrate` function of the statistical computing language R. As a proxy for



standard normal distribution, $y = Q(0.5)$

standard normal distribution, $y = Q(0.9)$

Student's t distribution ($\nu = 3$), $y = Q(0.5)$

Student's t distribution($\nu = 3$), $y = Q(0.9)$

Figure 6.2: Absolute errors (logarithmic scale) of the QUADPACK numerical integration routines as implemented in the `integrate` function of the statistical computing language R. The x-axis shows the number of iterations of the adaptive integration procedure for each integral in the corresponding representation.

the number of function evaluations in the numerical integration, we restrict the number of iterations for the adaptive procedure. We observe a distinct leverage of the implementation's adaptive nature when using the first group of representations with infinite integrals. Additionally, as the integral in representation (6.21) does not depend on y, it proves to be stable with respect to outlying observations. In comparison, the second group of representations exhibits great stability, but also slower convergence with additional subdivisions. Although, the initial estimate seems to be already quite accurate.

As a conclusion, numerical integration gives accurate approximations with all of the given representations. Often the choice will be determined by the functions which are available in closed forms. We note that it is important to avoid points of discontinuity and singularities. Regarding singularities, the form of representation (6.24) is chosen to leverage the fact that

$$\begin{aligned} \alpha\, Q(\alpha) &\to 0 \qquad \text{as } \alpha \to 0, \\ (1-\alpha)Q(\alpha) &\to 0 \qquad \text{as } \alpha \to 1, \end{aligned}$$

when the first moment exists.

Sampling

Sampling procedures are characterized by yielding different results, or estimates, on repeated calculations. They become necessary when none of the aforementioned representations can be evaluated numerically. However, we may still be able to leverage analytical information.

In the classical Markov chain Monte Carlo setting, our predictive distribution F has the form of a continuous mixture with respect to the distribution P of some parameter vector θ. The cumulative distribution function typically does not admit a closed form expression, but can be written as

$$F(x) = \int_\Theta F(x|\theta)\,\mathrm{d}P(\theta),$$

where $F(\cdot|\theta)$ is the conditional cumulative distribution function with a known closed form for a given $\theta \in \Theta$. Krüger et al. (2016) investigate this scenario systematically using four different estimators. The first one is based on the empirical mixture with $P(\theta) = \mathrm{ecdf}((\theta_i)_{i=1}^n)$, where $(\theta_i)_{i=1}^n$ is the parameter draw from the MCMC algorithm. The remaining three ignore all analytical information and are based solely on the sample $(x_i)_{i=1}^n$ of the variable of interest, i.e., the empirical cumulative distribution function, a kernel density estimator without degenerate kernel, and a normal approximation based on the sample mean and variance. In all cases, the CRPS can then be calculated using either a closed form expression or a numerical approximation. Their results clearly suggest that using additional analytical information is beneficial.

We now consider the fringe case where no analytical information can be leveraged. The only available information comes from a sample $(x_i)_{i=1}^n$, where the size

n may be chosen arbitrarily and even be increased during the evaluation. Based on equation (6.13), we can estimate the CRPS as

$$\mathrm{QD}(F, y) = \frac{2}{n^2} \sum_{i=1}^{n} \left(\mathbb{1}\{y < x_{(i)}\} n - i + \frac{1}{2} \right) (x_{(i)} - y),$$

where $x_{(i)}$ denotes the i-th order statistic. This is the equivalent to the algorithm by Hersbach (2000), but instead using the quantile decomposition (6.3). Alternatively, we can estimate the kernel representation (6.1) using

$$\mathrm{AKR}(F, y) = \frac{1}{n} \sum_{i=1}^{n} |x_i - y| - \frac{1}{2n} \sum_{i=1}^{n} |x_i - x_{\sigma(i)}|, \tag{6.26}$$

where $\sigma(i)$ is a cyclic permutation with no fixed points. The first option uses the complete information in the given sample but requires sorting, while the second option is computationally cheap but does not use all available information for the estimation of the second term.

However, if we can draw additional samples, we are not interested in the accuracy for a given sample, but the accuracy for a given amount of computation time. The sampling variance of the two terms in the threshold decomposition (6.2),

$$\mathbb{E}|X - y| \quad \text{and} \quad \mathbb{E}|X - X'|,$$

is equivalent in order. This means that if we spend computation time for estimation more evenly between the two components and if the computational cost for the sampling procedure is low, we can potentially draw a much bigger sample and effectively reduce the overall error variance of the CRPS estimate for a given computation time.

For independent samples, the choice of the cyclic permutation in (6.26) does not matter. However, when dependencies are present we can minimize their effect on the estimation by choosing an appropriate permutation. For example, the circular shift operation

$$\sigma(i) = (i + j) \mod n,$$

with $j = \left\lfloor \frac{n-1}{2} \right\rfloor$, maximizes the distance $|i - \sigma(i)|$ while keeping it constant, for all $i = 1, ..., n$, and thus minimizes time dependencies in MCMC methods.

Figure 6.3 shows the average computation time for a given sample size of four different methods, the kernel representation with the double sum

$$\mathrm{KR}(F, y) = \frac{1}{n} \sum_{i=1}^{n} |x_i - y| - \frac{1}{2n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} |x_i - x_j|,$$

the approximative kernel representation with cyclic permutation (AKR), the algorithm by Hersbach (2000) based on the threshold decomposition (TD), and the quantile decomposition (QD). For reference, the additional black line shows

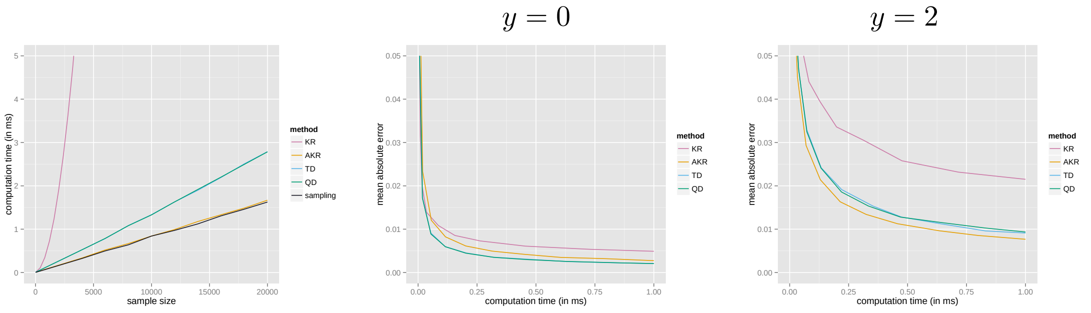


Figure 6.3: Left: Average computation time for a given a sample size with normal i.i.d. samples. Right: Mean absolute error for a given mean computation time (1000 iterations). Standard normal i.i.d. samples are used to estimate $\mathrm{CRPS}(\mathcal{N}_{0,1}, 0)$ and $\mathrm{CRPS}(\mathcal{N}_{0,1}, 2)$.

how much time is spent drawing the sample, effectively giving a lower bound for the CRPS estimation methods. We observe that computation of the AKR barely costs anything in addition to the sampling procedure, and that the KR is substantially more expensive than any other method. Between the TD and the QD, we cannot observe any difference in computation time, as the dominating factor is the sorting algorithm. Obviously, this is dependent on the implementation. Our methods are written in C++ and integrated into R via the Rcpp package.

Equipped with fairly efficient implementations for the four methods, we are interested in their accuracy as measured by the mean absolute error (MAE) for a given computation time. Figure 6.3 shows results for CRPS estimates derived from 1000 iterations of independent standard normal samples with different realized observations. As we can see in the figure, for realizations which lie in the center of the predictive distribution, it is easier to compute accurate estimates of the CRPS than for realizations which can be considered outlying with respect to the predictive distribution. Something else we can observe, is the superior performance of the AKR method for outlying realizations, which is not surprising considering this method is designed to increase the emphasis on the estimation of $\mathbb{E}|X - y|$. The variance of $|X - y|$ is largest when the realization y shifts the distributions so far in one direction that $X - y$ is either positive or negative almost surely, making those the more difficult cases to estimate the CRPS accurately.

Ideally, we would choose a method and a corresponding sample size according to the realization y to achieve a predetermined accuracy. Choosing the best method may be difficult, but choosing an appropriate sample size is certainly possible. Let us take a closer look at the AKR method in combination with the circular shift operation as choice of permutation. The shift does not have to be $\left\lfloor \frac{n-1}{2} \right\rfloor$ when we know that some smaller shift j already produces approximately independent pairs $(x_i, x_{\sigma(i)})$. And if that is the case, we may as well divide the sample into sub samples of size at least $2j + 1$, adjust the modulo of the shift operator, and take the average of the CRPS values on those sub samples. This does not change the accuracy of the final result on average, because it simply corresponds to a different cyclic permutation, which also produces approximately independent pairs. However, we can then estimate the accuracy of the CRPS approximation from this collection. If we have m of these CRPS values, then the standard deviation of the final CRPS estimate is $m^{-1/2}$ times the standard deviation derived from the collection. When the total sample size is large enough we can assume a normal distribution as a result of the Central Limit Theorem, which allows us to compute a variety of accuracy metrics like the MAE, or a confidence interval. Of course, we need an initiation period to create a CRPS value collection of size at least 20, say, after which we would determine in regular intervals whether additional sampling is necessary or the desired accuracy of the CRPS estimate has been achieved. The same can be done in combination with the TD and QD methods, but in these cases the accuracy will go down compared to the accuracy of an estimation from the full sample. As trade-off we can expect a lower computation time, because splitting up the sample reduces the work for the sorting algorithm.

6.3 R-package: `scoringRules`

Collaborative work with Fabian Krüger and Sebastian Lerch has lead to the development of a software package for the statistical programming language R (R Core Team 2016). In the R-package `scoringRules`, we aim to lower barriers to the use of proper scoring rules in practice, and give a dictionary-like reference for computing the continuous ranked probability score and the logarithmic score. Catering to the most common scenarios in practice, we allow predictive distributions to arise from parametric families or draws from sampling methods.

Commonly known, as well as previously unavailable, closed form expressions have been implemented for numerous distributions. An overview of the analytic representations that are available as of version 0.9.2 published on The Comprehensive R Archive Network (CRAN) is given in Table 6.1. Whenever the `scoringRules` package is referenced in Table 6.1, we refer to a vignette that is soon to be included in the package. A preliminary version is currently available at `https://github.com/FK83/scoringRules`.

For predictive distributions given as draws from Markov chain Monte Carlo or other sampling methods, the scores are computed from closed form expressions that are associated with some approximation method. We offer default choices based on the results presented in Krüger et al. (2016). Beyond that, the previous section on CRPS approximation discusses considerations which are currently under development for implementation.

To the best of our knowledge, `scoringRules` is the most comprehensive library of closed form expressions for the CRPS. However, other packages offer comparable functionality in their respective domains of application. Inseparably linked with model estimation in ensemble settings, the R packages `ensembleBMA` (Fraley et al. 2016) and `ensembleMOS` (Yuen et al. 2013) provide for the computation of the CRPS for normal and gamma distributions, as well as normal and gamma mixtures. Evaluation in terms of the CRPS for predictive distributions issued

Distributions on the entire real line

Distribution	Reference
Laplace	Jordan et al. (2016)
logistic	Jordan et al. (2016); Taillardat et al. (2016)
normal	Gneiting et al. (2005)
t	Jordan et al. (2016)
mixture of normals	Grimit et al. (2006)
two-piece-exponential	Jordan et al. (2016)
two-piece-normal	Gneiting and Thorarinsdottir (2010)

Distributions on the positive half-line

Distributions	Reference
exponential	Jordan et al. (2016)
gamma	Scheuerer and Möller (2015)
log-Laplace	Jordan et al. (2016)
log-logistic	Jordan et al. (2016); Taillardat et al. (2016)
log-normal	Baran and Lerch (2015)

Distributions on bounded intervals

Distribution	Reference
beta	Jordan et al. (2016); Taillardat et al. (2016)
uniform	Jordan et al. (2016)

Distributions with variable support

Distribution	Reference
censored/truncated logistic	Scheuerer and Möller (2015); Jordan et al. (2016)
censored/truncated normal	Gneiting et al. (2006); Thorarinsdottir and Gneiting (2010); Jordan et al. (2016)
censored/truncated *t*	Jordan et al. (2016)
generalized Pareto	Friederichs and Thorarinsdottir (2012)
generalized extreme value	Friederichs and Thorarinsdottir (2012)

Distributions on the natural numbers

Distribution	Reference
negative-binomial	Wei and Held (2014)
Poisson	Wei and Held (2014)

Table 6.1: Parametric families, with closed form expressions for the CRPS and LS, that are implemented in `scoringRules` as of version 0.9.2 published on CRAN.

in the form of samples is provided by the `verification` (National Center for Atmospheric Research 2015) and `SpecsVerification` (Siegert 2015) packages. Base R allows the computation of the LS for all its native distributions via the log-likelihood, and the `crch` (Messner et al. 2016) package extends to censored and truncated versions of the logistic, the normal, and the t-distribution.

7 Conclusion

In this thesis, we have discussed various facets of forecast evaluation. Fundamentally, proper measures of predictive performance reconcile a forecaster's subjective belief with their decision-theoretically optimal prediction in reporting. This means that individual preferences may affect the evaluation of predictive performance, but forecasters should not feel encouraged to deviate from their honest opinion. As forecast rankings depend on subjective criteria, we have investigated some classes of elementary performance measures and have developed tools to determine a ranking's stability. Furthermore, the reliability of an observed ranking also depends on the chosen performance measure. Lastly, computing scores can be challenging, and we have provided an overview of closed form solutions and approximative procedures for the continuous ranked probability score.

In Chapter 3, we have investigated the classes $\mathcal{S}_\alpha^{\mathrm{Q}}$ and $\mathcal{S}_\alpha^{\mathrm{E}}$ of scoring functions that are consistent for the quantile and expectile functionals at level $\alpha \in (0,1)$, respectively. The class $\mathcal{S}_\alpha^{\mathrm{ES}}$ comprising the consistent scoring functions for the bivariate combination of quantile and expected shortfall at level $\alpha \in (0,1)$ has also been considered. The elementary members of the respective classes are linearly parameterized by $\theta \in \mathbb{R}$,

$$\mathrm{S}_{\alpha,\theta}^{\mathrm{Q}}(x,y) = \begin{cases} 1-\alpha, & y \le \theta < x, \\ \alpha, & x \le \theta < y, \\ 0, & \text{otherwise}, \end{cases}$$

$$\mathrm{S}_{\alpha,\theta}^{\mathrm{E}}(x,y) = \begin{cases} (1-\alpha)|y-\theta|, & y \le \theta < x, \\ \alpha|y-\theta|, & x \le \theta < y, \\ 0, & \text{otherwise}, \end{cases}$$

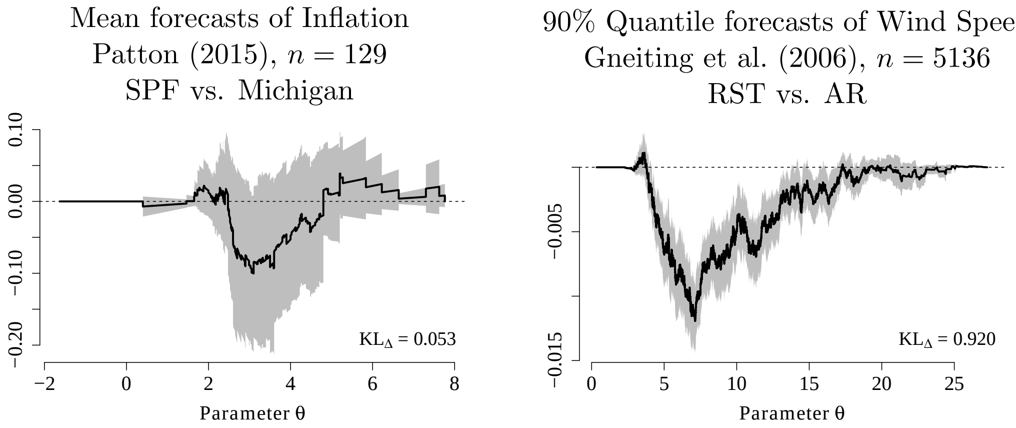
$$\mathrm{S}_{\alpha,\theta}^{\mathrm{ES}}(x_1,x_2,y) = 2\,\mathrm{S}_{1/2,\theta}^{\mathrm{E}}(x_2,y) + \frac{\mathbb{1}(\theta \le x_2)}{\alpha} \int_{-\infty}^{\infty} \mathrm{S}_{\alpha,t}^{\mathrm{Q}}(x_1,y)\,\mathrm{d}t.$$

Theorems 3.1a and 3.1b state the existence of a mixture representation for any member $\mathrm{S}_\alpha^{\mathrm{T}}$ in the respective class $\mathcal{S}_\alpha^{\mathrm{T}}$ of consistent scoring functions,

$$\mathrm{S}_\alpha^{\mathrm{T}}(x,y) = \int_{-\infty}^{\infty} \mathrm{S}_{\alpha,\theta}^{\mathrm{T}}(x,y)\,\mathrm{d}H(\theta),$$

where T designates the quantile or expectile functional, respectively, and H is a nonnegative mixing measure. For the joint evaluation of quantile and expected shortfall predictions, a similar result applies,

$$\mathrm{S}_\alpha^{\mathrm{ES}}(x_1,x_2,y) = \mathrm{S}_\alpha^{\mathrm{Q}}(x_1,y) + \int_{-\infty}^{\infty} \mathrm{S}_{\alpha,\theta}^{\mathrm{ES}}(x_1,x_2,y)\,\mathrm{d}H(\theta),$$


Mean forecasts of Inflation

Patton (2015), $n = 129$

SPF vs. Michigan

90% Quantile forecasts of Wind Speed

Gneiting et al. (2006), $n = 5136$

RST vs. AR

(< 0 means that SPF is preferable) (< 0 means that RST is preferable)

Figure 7.1: Murphy diagrams as in Figure 4.7 for the score difference of two forecasters.

where $\mathrm{S}_{\alpha}^{\mathrm{Q}} \in \mathcal{S}_{\alpha}^{\mathrm{Q}}$, as presented in Theorem 3.2. Finding mixture representations gets considerably more complicated for higher-dimensional multivariate point predictions. Probability forecasts for categorical events can be interpreted as such, and we have demonstrated that the parameterization of elementary members gets exceedingly difficult. If the class of proper scoring rules is too large, it will even become impossible to fully characterize the subclass of elementary members. Hence, suitable restrictions allowing for linear parameterizations serve multiple purposes. First, it is likely that an economic interpretation of the elementary scores and the corresponding parameterization exists. All elementary members of the classes $\mathcal{S}_{\alpha}^{\mathrm{Q}}$, $\mathcal{S}_{\alpha}^{\mathrm{E}}$, and $\mathcal{S}_{\alpha}^{\mathrm{ES}}$ correspond to betting and investment scenarios where the parameter θ corresponds to a threshold with implications for the decision making process. Second, the existence of a linear parameterization allows graphical inspection of all elementary members simultaneously by *Murphy diagrams* as proposed in Chapter 4.

An examiner has the ability to influence the forecast evaluation, even when choosing only from the *proper* performance measures. *Murphy diagrams* illustrate the robustness of a conclusion with respect to the choice of scoring function. In particular, the presence or absence of dominance relations can be confirmed with a single glance. Whenever empirical Murphy diagrams intersect, a dominance relation does not exist, and the forecast ranking depends on the mixing measure. For such cases, we provide a means of quantifying the stability of the forecast ranking. Figure 7.1 recapitulates the results from Figure 4.7, and illustrates the correspondence between visual impression of stability and the numerical quantification using the results from Section 4.3.2. At left, the line describing the average score difference for the elementary members crosses 0 prominently, whereas at right, the mean score differences are mostly in favor of the RST method and only barely favor the AR model otherwise. This is reflected by stability values of 0.053 and 0.920, respectively. Further investigations in multiple avenues may

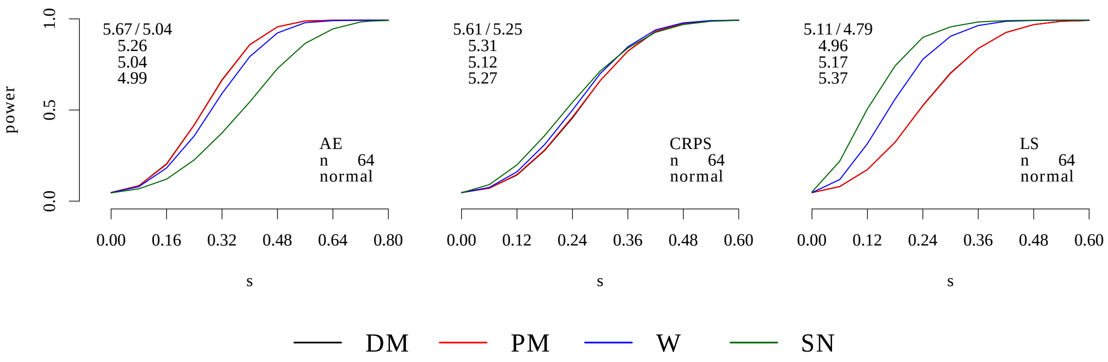


Figure 7.2: Plots of the adjusted power (5% level) for one-step ahead forecasts and normal innovations as in Figure 5.3. For details we refer to Section 5.2.

be warranted. First, the considerations regarding default mixing measures bear a strong resemblance to the topic of uninformative priors in Bayesian statistics. While translation-invariance is a natural criterion in the context of quantile and expectile predictions, the choice for probability forecasts of binary events is less obvious. Second, the results in Section 4.3.2 assume normalized mixing measures, which, for bounded domains, correspond to probability measures, and regularity conditions may be required for a well-defined Kullback-Leibler divergence (4.8) otherwise.

Another factor that plays a distinct role in forecast evaluation is sampling variation. Choosing a forecaster's opinion to rely on, and then reevaluating that decision as further data arrive, is common practice. Significance tests do not influence the choice of forecaster, but describe a means of quantifying the reliability of a forecast ranking. At times, they may have an influence on the decision of "choose a forecaster" against the option "wait for more data", but their main contribution to forecast evaluation lies in their interpretation as indicators of a forecast ranking's stability with respect to the collection of additional data. In this context, we illustrate that the paradigm shift to probabilistic forecasts rekindles the interest in one of the most basic statistical tests, the sign test. Figure 7.2 recapitulates some of the results from Figure 5.3, and illustrates that the performance measure can play an important role in significance testing. Under certain conditions or assumptions, the sign test can be used to test for the null hypothesis of equal predictive performance and may exhibit superior power properties. An intriguing question is whether the accommodation of an autocorrelation estimator in the sign test allows to omit the practice of multiple testing for forecast horizons beyond one time step ahead.

Bibliography

Abdi, H. (2007). The Bonferroni and Šidàk corrections for multiple comparisons. In *Encyclopedia of Measurement and Statistics* (N. Salkind, ed.). Thousand Oaks (CA): Sage, 103–107.

Apagodu, M. and Zeilberger, D. (2006). Multi-variable Zeilberger and Almkvist–Zeilberger algorithms and the sharpening of Wilf–Zeilberger theory. *Advances in Applied Mathematics*, 37, 139–152.

Baran, S. and Lerch, S. (2015). Log-normal distribution based ensemble model output statistics models for probabilistic wind speed forecasting. *Quarterly Journal of the Royal Meteorological Society*, 141, 2289–2299.

Barone Adesi, G. (2016). VaR and CVaR implied in option prices. *Journal of Risk and Financial Management*, 9, 2.

Bellini, F. and Bignozzi, V. (2015). On elicitable risk measures. *Quantitative Finance*, 15, 725–733.

Bellini, F. and Di Bernardino, E. (2015). Risk management with expectiles. *The European Journal of Finance*, to be published, doi: 10.1080/1351847X.2015.1052150.

Bellini, F., Klar, B., Müller, A. and Rosazza Gianin, E. (2014). Generalized quantiles as risk measures. *Insurance: Mathematics and Economics*, 54, 41–48.

Bentzien, S. and Friederichs, P. (2014). Decomposition and graphical portrayal of the quantile score. *Quarterly Journal of the Royal Meteorological Society*, 140, 1924–1934.

Bernardo, J. M. (1979). Expected information as expected utility. *Annals of Statistics*, 7, 686–690.

Berrocal, V. J., Raftery, A. E., Gneiting, T. and Steed, R. C. (2010). Probabilistic weather forecasting for winter road maintenance. *Journal of the American Statistical Association*, 105, 522–537.

Bradley, A. A. and Schwartz, S. S. (2011). Summary verification measures and their interpretation for ensemble forecasts. *Monthly Weather Review*, 139, 3075–3089.

Breckling, J. and Chambers, R. (1988). M-quantiles. *Biometrika*, 75, 761–771.

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78, 1–3.

Bröcker, J. (2009). Reliability, sufficiency, and the decomposition of proper scores. *Quarterly Journal of the Royal Meteorological Society*, 135, 1512–1519.

Bronshtein, E. M. (1978). Extremal convex functions. *Siberian Mathematical Journal*, 19, 6–12.

Brown, B. M. (1972). Formulae for absolute moments. *Journal of the Australian Mathematical Society*, 13, 104–106.

Buja, A., Stuetzle, W. and Shen, Y. (2005). Loss functions for binary class probability estimation and classification: Structure and applications. *Working paper*. Wharton School, University of Pennsylvania, Philadelphia. Available from `http://www-stat.wharton-upenn.edu/~buja/PAPERS/paper-proper-scoring.pdf`.

Busetti, F. and Marcucci, J. (2013). Comparing forecast accuracy: A Monte Carlo investigation. *International Journal of Forecasting*, 29, 13–27.

Clark, T. E. and McCracken, M. W. (2001). Tests of equal forecast accuracy and encompassing for nested models. *Journal of Econometrics*, 105, 85–110.

Clark, T. E. and McCracken, M. W. (2013). Advances in forecast evaluation. In *Handbook of Economic Forecasting* (G. Elliot and A. Timmerman, eds.), vol. 2. Amsterdam: Elsevier, 1107–1201.

Conover, W. J. (1999). *Practical Nonparametric Statistics* (3rd edition). New York: Wiley.

Cox, D. D. and Lee, J. S. (2008). Pointwise testing with functional data using the Westfall–Young randomization method. *Biometrika*, 95, 621–634.

Dawid, A. P. (2007). The geometry of proper scoring rules. *Annals of the Institute of Statistical Mathematics*, 59, 77–93.

DeGroot, M. H. and Fienberg, S. E. (1983). The comparison and evaluation of forecasters. *The Statistician*, 32, 12–22.

Delbaen, F. (2012). *Monetary Utility Functions*. Osaka: Osaka University Press.

Delbaen, F., Bellini, F., Bignozzi, V. and Ziegel, J. F. (2016). Risk measures with the CxLS property. *Finance and Stochastics*, 20, 433–453.

Dell'Aquila, R. and Ronchetti, E. (2004). Robust tests of predictive accuracy. *Metron*, 62, 161–184.

Diebold, F. X. (2015). Comparing predictive accuracy, twenty years later: A personal perspective on the use and abuse of Diebold-Mariano tests. *Journal of Business & Economic Statistics*, 33, 1–8.

Diebold, F. X. and Mariano, R. S. (1995). Comparing predictive accuracy. *Journal of Business and Economic Statistics*, 13, 253–263.

Eaton, M. L. (1982). A method for evaluating improper prior distributions. In *Statistical Decision Theory and Related Topics III* (S. S. Gupta and J. O. Berger, eds.). New York: Academic Press, 329–352.

Efron, B. (1991). Regression percentiles using asymmetric squared error loss. *Statistica Sinica*, 1, 93–125.

Ehm, W., Gneiting, T., Jordan, A. and Krüger, F. (2016). Of quantiles and expectiles: Consistent scoring functions, Choquet representations, and forecast rankings. *Journal of the Royal Statistical Society: Series B*, 78, 505–562.

Elliott, G., Ghanem, D. and Krüger, F. (2016). Forecasting conditional probabilities of binary outcomes under misspecification. *Review of Economics and Statistics*, 98, 742–755.

Embrechts, P., Puccetti, G., Rüschendorf, L., Wang, R. and Beleraj, A. (2014). An academic response to Basel 3.5. *Risks*, 2, 25–48.

Engelberg, J., Manski, C. F. and Williams, J. (2009). Comparing the point predictions and subjective probability distributions of professional forecasters. *Journal of Business & Economic Statistics*, 27, 30–41.

Epstein, E. S. (1969). A scoring system for probability forecasts of ranked categories. *Journal of Applied Meteorology*, 8, 985–987.

Fasciati, F., Jordan, A., Krüger, F. and Ziegel, J. F. (2016). Murphy diagrams for evaluating forecasts of value-at-risk and expected shortfall. *Working paper.*

Faust, J. and Wright, J. H. (2013). Forecasting inflation. In *Handbook of Economic Forecasting* (G. Elliott and A. Timmermann, eds.), vol. 2A. Amsterdam: Elsevier, 2–56.

Feuerverger, A. and Rahman, S. (1992). Some aspects of probability forecasting. *Communications in Statistics – Theory and Methods*, 21, 1615–1632.

Fisher, R. A. (1935). *The Design of Experiments.* London: Oliver and Boyd.

Fissler, T. and Ziegel, J. F. (2016). Higher order elicitability and Osband's principle. *The Annals of Statistics*, 44, 1680–1707.

Fissler, T., Ziegel, J. F. and Gneiting, T. (2016). Expected shortfall is jointly elicitable with value at risk–implications for backtesting. *Risk*, January, 55–61.

Fraley, C., Raftery, A. E., Sloughter, J. M. and Gneiting, T. (2016). *ensembleBMA: Probabilistic Forecasting using Ensembles and Bayesian Model Averaging.* R package version 5.1.3, URL `https://CRAN.R-project.org/package=ensembleBMA`.

Friederichs, P. and Thorarinsdottir, T. L. (2012). Forecast verification for extreme value distributions with an application to probabilistic peak wind prediction. *Environmetrics*, 23, 579–594.

Gneiting, T. (2011). Making and evaluating point forecasts. *Journal of the American Statistical Association*, 106, 746–762.

Gneiting, T., Balabdaoui, F. and Raftery, A. E. (2007). Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society: Series B*, 69, 243–268.

Gneiting, T., Larson, K., Westrick, K., Genton, M. G. and Aldrich, E. (2006). Calibrated probabilistic forecasting at the stateline wind energy center: The regime-switching space–time method. *Journal of the American Statistical Association*, 101, 968–979.

Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102, 359–378.

Gneiting, T., Raftery, A. E., Westveld III, A. H. and Goldman, T. (2005). Calibrated probabilistic forecasting using ensemble model output statistics and minimum crps estimation. *Monthly Weather Review*, 133, 1098–1118.

Gneiting, T. and Ranjan, R. (2011). Comparing density forecasts using threshold and quantile weighted scoring rules. *Journal of Business and Economic Statistics*, 29, 411–422.

Gneiting, T. and Ranjan, R. (2013). Combining predictive distributions. *Electronic Journal of Statistics*, 7, 1747–1782.

Gneiting, T. and Thorarinsdottir, T. L. (2010). Predicting inflation: Professional experts versus no-change forecasts. *arXiv preprint, arXiv:1010.2318.*

Gradshteyn, I. S. and Ryzhik, I. M. (2007). *Table of Integrals, Series, and Products* (7th edition). Amsterdam: Elsevier.

Grimit, E. P., Gneiting, T., Berrocal, V. J. and Johnson, N. A. (2006). The continuous ranked probability score for circular variables and its application to mesoscale forecast ensemble verification. *Quarterly Journal of the Royal Meteorological Society*, 132, 2925–2942.

Habiger, J. D. and Peña, E. A. (2011). Randomised p-values and nonparametric procedures in multiple testing. *Journal of Nonparametric Statistics*, 23, 583–604.

Hand, D. J. and Vinciotti, V. (2003). Local versus global models for classification problems: fitting models where it matters. *The American Statistician*, 57, 124–131.

Harvey, D., Leybourne, S. and Newbold, P. (1997). Testing the equality of prediction mean squared errors. *International Journal of Forecasting*, 13, 281–291.

Harvey, D., Leybourne, S. and Newbold, P. (1998). Tests of forecast encompassing. *Journal of Business and Economic Statistics*, 16, 281–291.

Hersbach, H. (2000). Decomposition of the continuous ranked probability score for ensemble prediction systems. *Weather and Forecasting*, 15, 559–570.

Holzmann, H. and Eulert, M. (2014). The role of the information set for forecasting–with applications to risk management. *The Annals of Applied Statistics*, 8, 595–621.

Holzmann, H. and Klar, B. (2016). Comment on: Of quantiles and expectiles: Consistent scoring functions, Choquet representations and forecast rankings by Ehm, Gneiting, Jordan and Krüger. *Journal of the Royal Statistical Society: Series B*, 78, 545–546.

Hong, T., Pinson, P. and Fan, S. (2014). Global energy forecasting competition 2012. *International Journal of Forecasting*, 30, 357–363.

Hsu, P. L. (1951). Absolute moments and characteristic functions. *Journal of the Chinese Mathematical Society*, 1, 259–280.

Huber, P. J. (1964). Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 35, 73–101.

Johansen, S. (1974). The extremal convex functions. *Mathematica Scandinavica*, 34, 61–68.

Jordan, A. and Krüger, F. (2016). *murphydiagram: Murphy Diagrams for Forecast Comparisons.* R package version 0.11, URL https://CRAN.R-project.org/package=murphydiagram.

Jordan, A., Krüger, F. and Lerch, S. (2016). *scoringRules: Scoring Rules for Parametric and Simulated Distribution Forecasts.* R package version 0.9.2, URL https://CRAN.R-project.org/package=scoringRules.

Koenker, R. (2005). *Quantile Regression.* Cambridge: Cambridge University Press.

Koenker, R. and Bassett, G. (1978). Regression quantiles. *Econometrica*, 46, 33–50.

Koltchinskii, V. I. (1997). M-estimation, convexity and quantiles. *The Annals of Statistics*, 25, 435–477.

Krämer, W. (2005). On the ordering of probability forecasts. *Sankhyā*, 67, 662–669.

Krüger, F., Lerch, S., Thorarinsdottir, T. L. and Gneiting, T. (2016). Probabilistic forecasting and comparative model assessment based on Markov chain Monte Carlo output. *arXiv preprint, arXiv:1608.06802*.

Laio, F. and Tamea, S. (2007). Verification tools for probabilistic forecasts of continuous hydrological variables. *Hydrology and Earth System Sciences*, 11, 1267–1277.

Lambert, N. S. (2013). Elicitation and evaluation of statistical forecasts. *Preprint*. Stanford University, Stanford. Available from `http://web.stanford.edu/~nlambert/papers/elicitation.pdf`.

Lavoie, J. L. (1987). Some summation formulas for the series ${}_3F_2(1)$. *Mathematics of Computation*, 49, 269–274.

Lieli, R. P. and Springborn, M. (2013). Closing the gap between risk estimation and decision making: Efficient management of trade-related invasive species risk. *Review of Economics and Statistics*, 95, 632–645.

Matheson, J. E. and Winkler, R. L. (1976). Scoring rules for continuous probability distributions. *Management Science*, 22, 1087–1096.

McNeil, A. J., Frey, R. and Embrechts, P. (2015). *Quantitative Risk Management*. Princeton: Princeton University Press.

Merkle, E. C. and Steyvers, M. (2013). Choosing a strictly proper scoring rule. *Decision Analysis*, 10, 292–304.

Messner, J. W., Mayr, G. J. and Zeileis, A. (2016). Heteroscedastic censored and truncated regression with crch. *The R-Journal*, 8/1, 173–181. URL `https://journal.r-project.org/archive/2016-1/messner-mayr-zeileis.pdf`.

Mitra, S. (2015). The relationship between conditional Value at Risk and option prices with a closed-form solution. *The European Journal of Finance*, 21, 400–425.

Murphy, A. H. (1970). The ranked probability score and the probability score: A comparison. *Monthly Weather Review*, 98, 917–924.

Murphy, A. H. (1977). The value of climatological, categorical and probabilistic forecasts in the cost-loss ratio situation. *Monthly Weather Review*, 105, 803–816.

Murphy, A. H. and Winkler, R. L. (1987). A general framework for forecast verification. *Monthly Weather Review*, 115, 1330–1338.

Mylne, K. R. (2002). Decision-making from probability forecasts based on forecast value. *Meteorological Applications*, 9, 307–315.

National Center for Atmospheric Research (2015). *verification: Weather Forecast Verification Utilities*. R package version 1.42, URL `https://CRAN.R-project.org/package=verification`.

Nau, R. F. (1985). Should scoring rules be effective? *Management Science*, 31, 527–535.

Newey, W. K. and Powell, J. L. (1987). Asymmetric least squares estimation and testing. *Econometrica*, 55, 819–847.

Newey, W. K. and West, K. D. (1987). A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica*, 55, 703–708.

Neyman, J. and Pearson, E. S. (1933). On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 231, 289–337.

Olver, F. W. J., Olde Daalhuis, A. B., Lozier, D. W., Schneider, B. I., Boisvert, R. F., Clark, C. W., R, M. B. and Saunders, B. V. (eds.) (2016). *NIST Digital Library of Mathematical Functions* (Release 1.0.13). URL `http://dlmf.nist.gov/`.

Patton, A. J. (2011). Volatility forecast comparison using imperfect volatility proxies. *Journal of Econometrics*, 160, 246–256.

Patton, A. J. (2015). Evaluating and comparing possibly misspecified forecasts. *Working Paper*. Department of Economics, Duke University, Durham. Available from `http://public.econ.duke.edu/~ap172/Patton_bregman_comparison_27mar15.pdf`.

Phelps, R. R. (2001). *Lectures on Choquet's Theorem* (2nd edition). Heidelberg: Springer.

R Core Team (2016). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL `https://www.R-project.org`.

Richardson, D. S. (2000). Skill and relative economic value of the ECMWF ensemble prediction system. *Quarterly Journal of the Royal Meteorological Society*, 126, 649–668.

Richardson, D. S. (2012). Economic value and skill. In *Forecast Verfication: a Practitioner's Guide in Atmospheric Science* (2nd edition) (I. T. Jolliffe and D. B. Stephenson, eds.). Chichester: Wiley, 167–184.

Rudebusch, G. D. and Williams, J. C. (2009). Forecasting recessions: the puzzle of the enduring power of the yield curve. *Journal of Business & Economic Statistics*, 27, 492–503.

Savage, L. J. (1971). Elicitation of personal probabilities and expectations. *Journal of the American Statistical Association*, 66, 783–801.

Schervish, M. J. (1989). A general method for comparing probability assessors. *Annals of Statistics*, 17, 1856–1879.

Scheuerer, M. and Möller, D. (2015). Probabilistic wind speed forecasting on a grid based on ensemble model output statistics. *The Annals of Applied Statistics*, 9, 1328–1349.

Schulze Waltrup, L., Sobotka, F., Kneib, T. and Kauermann, G. (2015). Expectile and quantile regression–David and Goliath? *Statistical Modelling*, 15, 433–456.

Shuford, E. H., Albert, A. and Massengill, H. E. (1966). Admissible probability measurement procedures. *Psychometrika*, 31, 125–145.

Siegel, S. (1956). *Nonparametric Statistics for the Behavioral Sciences*. New-York: McGraw-Hill.

Siegert, S. (2015). *SpecsVerification: Forecast Verification Routines for the SPECS FP7 Project*. R package version 0.4-1, URL `https://CRAN.R-project.org/package=SpecsVerification`.

Smet, G., Termonia, P. and Deckmyn, A. (2012). Added economic value of limited area multi-EPS weather forecasting applications. *Tellus A*, 64, 18901.

Steinwart, I., Pasin, C., Williamson, R. C. and Zhang, S. (2014). Elicitation and identification of properties. *JMLR Workshop and Conference Proceedings*, 35, 1–45.

Strähl, C. and Ziegel, J. F. (2015). Cross-calibration of probabilistic forecasts. *arXiv preprint, arXiv:1505.05314*.

Székely, G. J. and Rizzo, M. L. (2013). Energy statistics: A class of statistics based on distances. *Journal of Statistical Planning and Inference*, 143, 1249–1272.

Taillardat, M., Mestre, O., Zamo, M. and Naveau, P. (2016). Calibrated ensemble forecasts using quantile regression forests and ensemble model output statistics. *Monthly Weather Review*, 144, 2375–2393.

Thompson, J. C. and Brier, G. W. (1955). The economic utility of weather forecasts. *Monthly Weather Review*, 83, 249–253.

Thomson, W. (1979). Eliciting production possibilities from a well-informed manager. *Journal of Economic Theory*, 20, 360–380.

Thorarinsdottir, T. L. and Gneiting, T. (2010). Probabilistic forecasts of wind speed: Ensemble model output statistics by using heteroscedastic censored regression. *Journal of the Royal Statistical Society: Series A*, 173, 371–388.

Tsyplakov, A. (2014). Theoretical guidelines for a partially informed forecast examiner. *Working Paper*. Available from `http://mpra.ub.uni-muenchen.de/55017/`.

Vardeman, S. and Meeden, G. (1983). Calibration, sufficiency, and domination considerations for Bayesian probability assessors. *Journal of the American Statistical Association*, 78, 808–816.

Varian, H. R. (1975). A Bayesian approach to real estate assessment. In *Studies in Bayesian econometrics and statistics in honor of Leonard J. Savage* (S. E. Fienberg and A. Zellner, eds.). Amsterdam: North Holland, 195–208.

Ventura, S. L. and Nugent, R. (2016). Comment on: Of quantiles and expectiles: Consistent scoring functions, Choquet representations and forecast rankings by Ehm, Gneiting, Jordan and Krüger. *Journal of the Royal Statistical Society: Series B*, 78, 555.

von Bahr, B. (1965). On the convergence of moments in the Central Limit Theorem. *The Annals of Mathematical Statistics*, 36, 808–818.

Wei, W. and Held, L. (2014). Calibration tests for count data. *TEST*, 23, 787–805.

West, K. D. (1996). Asymptotic inference about predictive ability. *Econometrica*, 64, 1067–1084.

West, K. D. (2006). Forecast evaluation. In *Handbook of Economic Forecasting* (G. Elliot, C. W. Granger and A. Timmerman, eds.), vol. 1. Amsterdam: Elsevier, 99–134.

Westfall, P. H. and Young, S. S. (1993). *Resampling-based Multiple Testing*. New York: Wiley.

White, H. (2000). A reality check for data snooping. *Econometrica*, 64, 1067–1084.

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics Bulletin*, 1, 80–83.

Wilks, D. S. (2001). A skill score based on economic value for probability forecasts. *Meteorological Applications*, 8, 209–219.

Wolfers, J. and Zitzewitz, E. (2008). Prediction markets in theory and practice. In *The New Palgrave Dictionary of Economics* (2nd edition) (S. N. Durlauf and L. E. Blume, eds.). London: Palgrave Macmillan.

Yuen, R. A., Gneiting, T., Thorarinsdottir, T. and Fraley, C. (2013). *ensembleMOS: Ensemble Model Output Statistics*. R package version 0.7, URL https://CRAN.R-project.org/package=ensembleMOS.

Ziegel, J. F. (2016). Coherence and elicitability. *Mathematical Finance*, 26, 901–918.