The Verification of Probabilistic Forecasts in Decision and Risk Analysis

by

Victor Richmond R. Jose

Department of Business Administration
Duke University

Date: _____

Approved:

_____
Robert L. Winkler, Supervisor

_____
Robert F. Nau

_____
Luca Rigotti

_____
Mark L. Huber

A dissertation submitted in partial fulfillment of the
requirements for the degree of Doctor of Philosophy
in the Department of Business Administration
in the Graduate School of
Duke University

2009

ABSTRACT

The Verification of Probabilistic Forecasts in Decision and Risk Analysis

by

Victor Richmond R. Jose

Department of Business Administration
Duke University

Date: _____

Approved:

_____
Robert L. Winkler, Supervisor

_____
Robert F. Nau

_____
Luca Rigotti

_____
Mark L. Huber

An abstract of a dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Business Administration in the Graduate School of Duke University

2009

Abstract

Probability forecasts play an important role in many decision and risk analysis applications. Research and practice over the years have shown that the shift towards distributional forecasts provides a more accurate and appropriate means of capturing risk in models for these applications. This means that mathematical tools for analyzing the quality of these forecasts, may it come from experts, models or data, become important to the decision maker. In this regard, strictly proper scoring rules have been widely studied because of their ability to encourage assessors to provide truthful reports. This dissertation contributes to the scoring rule literature in two main areas of assessment - probability forecasts and quantile assessments.

In the area of probability assessment, scoring rules typically studied in the literature, and commonly used in practice, evaluate probability assessments relative to a default uniform measure. In many applications, the uniform baseline used to represent some notion of ignorance is inappropriate. In this dissertation, we generalize the power and pseudospherical family of scoring rules, two large parametric families of commonly-used scoring rules, by incorporating the notion of a non-uniform baseline distribution for both the discrete and continuous cases. With an appropriate normalization and choice of parameters, we show that these new families of scoring rules relate to various well-known divergence measures from information theory and to well-founded decision models when framed in an expected utility maximization context.

In applications where the probability space considered has an ordinal ranking between states, an important property often considered is sensitivity to distance. Scoring rules with this property provide higher scores to assessments that allocate higher probability mass to events "closer" to that which occurs based on

some notion of distance. In this setting, we provide an approach that allows us to generate new sensitive to distance strictly proper scoring rules from well-known strictly proper binary scoring rules. Through the use of the weighted scoring rules, we also show that these new scores can incorporate a specified baseline distribution, in addition to being strictly proper and sensitive to distance.

In the inverse problem of quantile assessment, scoring rules have not yet been well-studied and well-developed. We examine the differences between scoring rules for probability and quantile assessments, and demonstrate why the tools that have been developed for probability assessments no longer encourage truthful reporting when used for quantile assessments. In addition, we shed light on new properties and characterizations for some of these rules that could guide decision makers trying to choosing an appropriate scoring rule.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Mathematical models of uncertainty abound in many areas of research from the traditional fields of business such as finance, operations and marketing to the natural and social sciences. In all of these areas, a decision maker often needs to quantify certain relevant uncertainties through the use of probabilities. In many cases, decision makers do this either by expressing their own beliefs or by eliciting probabilities from experts whom they feel have better information.

When little care is given to this assessment procedure, poor probability judgments often lead the decision maker to the wrong conclusions. Research in the fields of decision and risk analysis has sought to characterize good probability assessments and develop methods one could use to elicit and evaluate these judgments from experts. Scoring rules have been developed for this purpose.

Scoring rules are functions that assign a 'reward' for a probability assessment given an outcome of an uncertain event. From an ex post point of view, scoring rules are useful because they can provide a means by which assessors can be evaluated with respect to their predictive abilities. Also, with an appropriate choice of scoring rule, decision makers can generate payment schemes ex ante such that probability assessors would have the incentive to truthfully report their beliefs. In these contexts, scoring rules have been useful tools in the areas of probability elicitation and verification.

This dissertation looks at the issues of probability elicitation and verification by suggesting new families of rules with useful properties. In particular, we focus on the two main types of elicitation used in practice - probability forecasts and quantile assessments.

In the area of probability assessment, we investigate the notion of generating families of scoring rules with the ability to incorporate a specified baseline distribution that could be used as a reference point to measure the quality of information provided

by an assessor. Scoring rules typically studied in the literature and commonly used in practice evaluate probability assessments relative to a default uniform distribution. However, in many applications, the uniform baseline used to represent some notion of ignorance is often inappropriate. As an alternative, we suggest the use of weighted versions of the power and pseudospherical families of scoring rules, which are two large classes of scoring rules that can incorporate a non-uniform baseline distribution.

We also show that these new parametric families of scoring rules are connected to well-known information measures. In particular, the expected scores for these rules correspond to two well-known parametric families of generalized divergence functions in the information theory literature. In addition, we show that these scoring rules are related to some generic expected utility optimization models that provide a decision-theoretic connection between these new rules and the notion of information value in an economic setting.

In many applications, the probability space considered may have an ordinal ranking between states. In these settings, an important property often considered is sensitivity to distance. Scoring rules with this property provide higher scores to assessments that allocate higher probability mass to events "closer" to that which occurs based on some notion of distance. In this setting, we provide an approach that allows us to generate new sensitive to distance strictly proper scoring rules from well-known strictly proper binary scoring rules. Through the use of the weighted scoring rules, we also show that these new scores can incorporate a specified baseline distribution, in addition to being strictly proper and sensitive to distance. We illustrate this in detail using the weighted power and pseudospherical scoring rules we have developed earlier.

Aside from probability elicitation, another common method of eliciting probabilistic forecasts is to ask experts to provide quantile assessments for an unknown distribution. For example, it is common in decision and risk analysis to assess the 0.10, 0.50 and 0.90, or the 0.05, 0.50 and 0.95 quantiles of a distribution. We then ask how forecast evaluation tools can be applied to this area. The notion of scoring rules has been hardly studied in the area of quantile assessment. We begin by examining the differ-

ences between scoring rules for probability and quantile assessments, and demonstrate why the tools that have been developed for probability assessments no longer apply. We then shed light on new properties and characterizations for some of the new scoring rules for quantiles that could later guide decision makers in choosing an appropriate scoring rule to use.

This dissertation proposal proceeds in the following manner. Chapter 2 describes basic definitions and concepts that are used in this dissertation. Chapter 3 discusses the motivation and development behind the new families of scoring rules. Next, connections to information theory and utility maximization are presented. Chapter 4 describes an extension to accommodate for the notion of sensitivity to distance in ordered spaces. Chapter 5 discusses the distinction between the scoring rules under probability assessment and under quantile assessment, and develops new families of rules suited for quantile assessment. Properties of these scoring rules and some of their generalizations are then highlighted. Chapter 6 summarizes.

Chapter 2

Foundations

In this chapter, we introduce the basic definitions, properties and concepts that we will be using throughout this dissertation. In particular, we discuss two main topics - scoring rules and information measures. For those interested in reading further, more comprehensive reviews of scoring rules are provided by Winkler (1996), and Gneiting and Raftery (2007), while Verdu & McLaughlin (1999) is a good reference for information measures.

2.1 Scoring Rules

2.1.1 Definition and Set-Up

Suppose that we want to make probability statements about a random variable $\mathbf{X}$ defined on some measurable space $(\Omega, \mathscr{F})$, where Ω is the sample space and $\mathscr{F}$ the σ-field of Borel subsets B of Ω. If we let $\mathscr{P}$ be a convex class of probability measures on the space $(\Omega, \mathscr{F})$ then any probability measure belonging to $\mathscr{P}$ is a probability assessment about $\mathbf{X}$. In this dissertation, we only consider probability distributions P that are absolutely continuous with respect to a σ-finite measure μ on the measurable space $(\Omega, \mathscr{F})$. Moreover, we only consider two cases: (i) μ is the Lebesgue measure and (ii) μ is a counting measure. This implies that the Radon-Nikodym derivative of the distribution P with respect to μ can be written in the following manner (assuming it exists)

$$\frac{dP}{d\mu} = \begin{cases} p(x) & \text{if } \mu \text{ is the Lebesgue measure} \\ \mathbb{P}(\mathbf{X} = x) = p_x & \text{if } \mu \text{ is a counting measure} \end{cases} \tag{2.1}$$

for $x \in supp(\mathbf{X})$, the support of $\mathbf{X}$. We refer to the Radon-Nikodym derivative as the *probability density function* when μ is the Lebesgue measure and the *probability*

mass function when μ is the counting measure. This covers most useful and commonly encountered applications in practice where typically the focus is on Borel subsets of $\mathbb{R}^n$ and the elicited distributions are expressed as either discrete or continuous probability distributions.

Given this set-up, we can now formally define a scoring rule.

Definition 1. *An extended real-valued function $S : \mathscr{P} \times \Omega \to \bar{\mathbb{R}}$ is a scoring rule if for all $P \in \mathscr{P}$,*

(i) $S(P, \cdot)$ is measurable (with respect to $\mathscr{F}$), and

(ii) $S(P, \cdot)$ is quasi-integrable, i.e. $\int S^+(P, \omega) dP(\omega) := \int [S(P, \omega) \vee 0] dP(\omega) < \infty$ or $\int S^-(P, \omega) dP(\omega) := -\int [S(P, \omega) \wedge 0] dP(\omega) < \infty$.

In addition, we consider the following regularity condition similar to Gneiting and Raftery (2007): for all $P, Q \in \mathscr{P}$, $\mathbb{E}_Q S(P) \in \mathbb{R}$ except possibly that $S(P, Q) = -\infty$ if $P \neq Q$. This is a technical requirement which will later play a role when we try to characterize rules which encourage truthful reporting.

Under this scheme, if $x = X(\omega)$ materializes then the score assigned to a forecast P is given by $S(P, \omega)$. In an *ex post* sense, the score received by the assessor can be used to monitor her performance over a series of assessments. Here, we say that the scoring rule is useful in the context of forecast verification.

From an ex ante perspective, scoring rules can also be useful. Prior to the observance of ω, an analysis based on expectations is possible. More specifically, these scores could be used to generate a reward scheme such that the assessor who wants to maximize her expected score would report truthfully.

If P represents the assessor's belief about the random variable $\mathbf{X}$ and she chooses

to report R, then his expected score is given by

$$\begin{aligned}\mathbb{E}_P S(R) &= S(R,P) \\ &= \int_\Omega \frac{dP}{d\mu}(\omega) S(R,\omega) d\mu(\omega) \\ &= \begin{cases} \int_\Omega p(\omega) S(r(\omega),\omega) d\omega & \text{if } \mu \text{ is the Lebesgue measure} \\ \sum_\Omega p_\omega S(r(\omega),\omega) & \text{if } \mu \text{ is a counting measure} \end{cases}.\end{aligned}$$

We slightly abuse notation by using $d\omega$ instead of $d\mu(\omega)$ when there is no confusion about the setting where are in.

As an illustration, consider the state space $\Omega = \{0,1\}$ with two possible outcomes and the set $\mathscr{P} = \Delta^2 = \{(p, 1-p) : p \in [0,1]\}$. Next, let the *quadratic score* be defined as the function which assigns the following score to an assessment $\mathbf{r} = (r, 1-r)$ if event 1 happens:

$$S(\mathbf{r}, 1) = 2r - r^2 - (1-r)^2 \tag{2.2}$$

and

$$S(\mathbf{r}, 0) = 2(1-r) - r^2 - (1-r)^2 \tag{2.3}$$

if the other event happens. For simplicity, we denote the actual score for a discrete case as is conventionally done in the literature as $S_j(\mathbf{r})$ given that event j occurs. Hence, under this scheme, if a probability assessment of $(0.7, 0.3)$ is provided and event 1 happens, then the assessor receives a score of 0.82 if state 1 happens but only 0.02 if other state occurs. Note that to get the highest possible score of one, the assessor has to provide an assessment of 1 for the event that occurs. This represents a categorical forecast. Unfortunately, if a categorical forecast is given to an event that does not occur then the assessor would get the lowest possible score, -1.

If the assessor's true belief is $\mathbf{p}$ then her expected score is given by

$$\begin{aligned}\mathbb{E}_{\mathbf{p}} S(\mathbf{r}) &= pS_1(\mathbf{r}) + (1-p)S_2(\mathbf{r}) \\ &= 1 + (1-2r)(1-2p) - r^2 - (1-r)^2.\end{aligned}$$

To maximize this expected score among all assessments $\mathbf{r} \in \mathscr{P}$, $\mathbf{r}$ has to be exactly $\mathbf{p}$. This implies that a person wanting to maximize this expected score should report

truthfully. Note that not all scoring rules satisfy this property. If the *linear rule*

$$S_1(\mathbf{r}) = S_2(\mathbf{1} - \mathbf{r}) = r \tag{2.4}$$

is used, then the expected score is maximized by providing a categorical forecast to the event with higher probability.

On the other hand, if $\Omega = \mathbb{R}$ then the quadratic score changes as follows: if x is the realization of $\mathbf{X}$ and a probability density r is reported, then the score for reporting r would be

$$S(R, x) = 2r(x) - \int_{-\infty}^{\infty} r(y)^2 dy. \tag{2.5}$$

while the expected score is

$$\mathbb{E}_P S(R) \quad = \quad \int_{-\infty}^{\infty} 2p(x)r(x)dx - \int_{-\infty}^{\infty} r(y)^2 dy. \tag{2.6}$$

The integral in (2.5) is the natural extension of the sum of squares for the probabilities of each state in the discrete case.

2.1.2 Some Well-Known Examples of Scoring Rules

Numerous scoring rules have been proposed in the literature. Here, we describe some of the most commonly used and well-studied rules. We provide some background and references but will defer discussing specific properties of these rules until the next section.

Quadratic Scoring Rule. The first scoring rule introduced in literature is the quadratic scoring rule, also known as the Brier score, named after G. W. Brier (1950) who suggested it in the context of weather forecasting. His score was similar to the forms (2.2) and (2.3), with the exception that the orientation was negative, i.e. lower scores imply a better performance. This perhaps comes from the long tradition of using least squares in many statistical applications.

Later on, de Finetti (1962) provided the generalization to the general n-state discrete case. In this form, the score when event j happens is given by

$$S_j(\mathbf{r}) = 2r_j - \sum_{k=1}^{n} r_k^2. \tag{2.7}$$

The continuous version of this rule (and many others) was introduced by Matheson and Winkler (1976) and is exactly the same form given in (2.5). We note that the expected score when the true belief is reported is $\int_\Omega p(\omega)^2 d\omega$. The quadratic rule is the most commonly used scoring rule.

Logarithmic Scoring Rule. Another score introduced early in the literature is the logarithmic scoring rule. Good (1952) suggested this score in the context of betting. In the discrete case, this score has the form

$$S_j(\mathbf{r}) = \log r_j, \tag{2.8}$$

while in the continuous setting,

$$S(R, \omega) = \log r(\omega). \tag{2.9}$$

An interesting property of this rule is that the expected score under truth-telling is related to a popular measure from information theory. In particular, it is the negative of Shannon's entropy:

$$\mathbb{E}_P S(P) = \int p(\omega) \log p(\omega) d\omega. \tag{2.10}$$

Spherical Scoring Rule. This rule was first introduced by Roby (1965) in the context of psychological testing, where respondents had to provide probability assessments for events described in an experiment. When the state space is discrete, this scoring rule has the form

$$S_j(\mathbf{r}) = \frac{r_j}{||\mathbf{r}||_2}, \tag{2.11}$$

where $||\mathbf{r}||_k$ is the k-norm given by

$$||\mathbf{r}||_k = \Big(\sum_{j=1}^{n} r_j^k \Big)^{1/k}.$$

Analogously, the continuous version of this rule is given by

$$S(R, \omega) = \frac{r(\omega)}{\Big(\int_\Omega r(\omega)^2 d\omega \Big)^{1/2}}. \tag{2.12}$$

We can denote the integral in the denominator as $||R||_2$. The expected score for truthful reporting, $\mathbb{E}_R S(R) = ||R||_2$, is the square root of the expected score for the quadratic scoring rule.

Pseudospherical Scoring Rule. One of the first parametric generalizations for a scoring rule was provided in a short comment by Good (1971). He suggested a measure that contains two well-known scoring rules and has a strong connection to a statistical measure for surprise used in inferential experiments. The scoring rule, which Good called pseudospherical, is given by:

$$S(R,\omega) = \begin{cases} \frac{1}{\beta-1}\Big[\Big(\frac{r_\omega}{||\mathbf{r}||_{\beta-1}}\Big)^\beta - 1\Big] & \text{if } \mu \text{ is a counting measure} \\ \\ \frac{1}{\beta-1}\Big[\Big(\frac{R(\omega)}{||R||_{\beta-1}}\Big)^\beta - 1\Big] & \text{if } \mu \text{ is the Lebesgue measure} \end{cases} \tag{2.13}$$

for $\beta \in \mathbb{R}$. When $\beta = 2$, the score is equivalent to the spherical rule. In the limit as $\beta \to 1$, this scoring rule converges to the logarithmic rule.

Power Scoring Rule. Similar to the pseudospherical scoring rule, a generalization of the quadratic rule exists. Selten (1998) introduced the power family of scoring rules, which has the form

$$S(R,\omega) = \begin{cases} \frac{\beta r_\omega^{\beta-1} - 1 - (\beta-1)||\mathbf{r}||_\beta^\beta}{\beta(\beta-1)} & \text{if } \mu \text{ is a counting measure} \\ \\ \frac{\beta r(\omega)^{\beta-1} - 1 - (\beta-1)||R||_\beta^\beta}{\beta(\beta-1)} & \text{if } \mu \text{ is the Lebesgue measure} \end{cases} \tag{2.14}$$

for $\beta \in \mathbb{R}$. When $\beta = 2$, we get the quadratic rule, and as $\beta \to 1$, we get the logarithmic rule. Selten's original proposed score is not in the same form as (2.14), which uses the normalization suggested in Jose, Nau & Winkler (2008) since this has nice properties for all β.

2.1.3 Properties

Given the wide array of possible scoring rules, many authors have investigated a collection of properties which could be desirable in different decision contexts. Knowledge of these properties could be used in the selection of an appropriate scoring rule. We introduce some of these properties and discuss their implications in practice.

Strict Propriety. The most commonly studied property of scoring rules is strict propriety. Though the notion was first coined by Winkler and Murphy (1968), the idea was already present in the earlier works of Brier (1950), Good (1952) and de Finetti (1962).

Definition 2. *A scoring rule is proper (relative to $\mathscr{P}$) if*

$$\mathbb{E}_P S(P) \geq \mathbb{E}_P S(R) \tag{2.15}$$

for all $P, R \in \mathscr{P}$. It is said to be strictly proper if the equality in (2.15) holds only when $P = R$.

The implication of this definition is that ex ante assessors who would like to maximize their expected score have to report truthfully. From the standpoint of economics and game theory, this principle is equivalent to the interim incentive compatibility condition in mechanism design, which encourages agents to truthfully report their beliefs. This is perhaps why among all properties, this is the only one which almost all studies consider desirable. Note that not all rules satisfy this property. As discussed earlier, linear rule is not strictly proper and encourages the reporting of extreme probabilities.

All the scoring rules presented in Section 2.1.2 are strictly proper. In addition, positive affine transformations of strictly proper scoring rules are also strictly proper, i.e.

Remark 1. *If $S(R, \omega)$ is (strictly) proper with respect to $\mathscr{P}$ then so is $\check{S} := aS(R, \omega) + b$ for $a \in \mathbb{R}^+$ and $b \in \mathbb{R}$. [Winkler & Murphy 1968, Gneiting & Raftery 2007]*

Affine transformations however are not the only way to generate new scores. We can generate infinitely many scoring rules by taking strictly convex functions and applying a result by Hendrickson and Buehler (1971), which generalizes an earlier result of McCarthy (1956) and Savage (1971). To state this result, we first briefly recall the notion of a subtangent from convex analysis.

Definition 3. *A function $G^*(P,\cdot) : \Omega \to \bar{\mathbb{R}}$ is a subtangent of G at $P \in \mathscr{P}$ if the following are satisfied:*

(i) G^* is integrable with respect to P

(ii) G^* is quasi-integrable with respect to all $R \in \mathscr{P}$, and

(iii) [Subtangent Inequality] For all $R \in \mathscr{P}$,

$$G(R) \geq G(P) + \int G^*(P,\omega)d(R-P)(\omega). \tag{2.16}$$

In the discrete case, the notion of a subtangent G^* is simply that of the familiar subgradient G' with (2.16) being equivalent to

$$G(R) \geq G(P) + \langle G'(P), R-P\rangle, \tag{2.17}$$

where $\langle\cdot,\cdot\rangle$ is the standard scalar product operator. Using this notion, we have the following theorem:

Remark 2 (McCarthy-Savage-Hendrickson-Buehler Characterization Theorem)**.** *A scoring rule S is (strictly) proper relative to a class $\mathscr{P}$ if and only if there exists a (strictly) convex function G on $\mathscr{P}$ such that*

$$S(P,\omega) = G(P) - \int G^*(P,\omega)dP(\omega) + G^*(P,\omega) \tag{2.18}$$

for $P \in \mathscr{P}$, where $G^*(P,\cdot) : \Omega \to \bar{\mathbb{R}}$ is a subtangent of G at $P \in \mathscr{P}$. If Ω is finite, (2.18) is equivalent to

$$S_i(\mathbf{p}) = G(\mathbf{p}) - \langle G'(\mathbf{p}), \mathbf{p}\rangle + G_i'(\mathbf{p}) \tag{2.19}$$

for every $i \in \Omega = \{1, \ldots, n\}$ *and* $\mathbf{p} \in \mathscr{P}$ *with* G *being a (strictly) convex function and* $G'(\mathbf{p})$ *a subgradient of* G *at* $\mathbf{p}$*. [McCarthy 1956, Savage 1971, Hendrickson & Buehler 1971]*

This theorem is equivalent to McCarthy's (1956) original result which states that (strict) propriety is equivalent to requiring that the expected score $\mathbb{E}_P S(P)$ is (strictly) convex with $S(P, \omega)$ being a subtangent of $\mathbb{E}_P S(P)$ at P for all measures in the space $\mathscr{P}$. Thus, given any convex function, we can use (2.18) to generate a strictly proper scoring rule. For example, when Ω is discrete and $G(\mathbf{p}) = ||\mathbf{p}||_2$, (2.18) generates the spherical scoring rule. The reason that we obtain the spherical scoring rule is that it turns out that under this constructive theorem $G(P)$ is simply $\mathbb{E}_P S(P)$.

As an illustration, consider the case where the state space has only two elements. Equation (2.19) can be expressed as

$$S_1(\mathbf{p}) = G(\mathbf{p}) + (1-p)G'(\mathbf{p}) \tag{2.20}$$

$$S_2(\mathbf{p}) = G(\mathbf{p}) - pG'(\mathbf{p}). \tag{2.21}$$

This is commonly referred to as the Schervish representation (Dawid 1986, Schervish 1989) for binary strictly proper scoring rules. Figure 2.1 provides a graphical illustration for the spherical rule when $n = 2$. Given a report $\mathbf{r} = (0.6, 0.4)$, the score when states 1 (state 2) occurs is the point at which the tangent line to G at $p = 0.6$ intersects the lines $p = 0$ ($p = 1$). The expected score under truth-telling is the functional value evaluated at the report, while the expected score of report $\mathbf{r}$ under some other distribution $\hat{\mathbf{p}} = (\hat{p}, 1-\hat{p})$ is the point where the tangent line intersects the line $p = \hat{p}$.

Symmetry. In cases where the ordering of state space does not matter, one property that has been considered by many authors is symmetry.

Definition 4. *Let* Ω *be discrete and* $\mathbf{r}, \vartheta(\mathbf{r}) \in \mathscr{P}$ *such that the elements of* $\vartheta(\mathbf{r})$ *are mapped from the elements of* $\mathbf{r}$ *under the permutation operator* ϑ*. A scoring rule is symmetric if for all* $\omega \in \Omega$*,*

$$S_\omega(\mathbf{r}) = S_{\vartheta(\omega)}(\vartheta(\mathbf{r})). \tag{2.22}$$

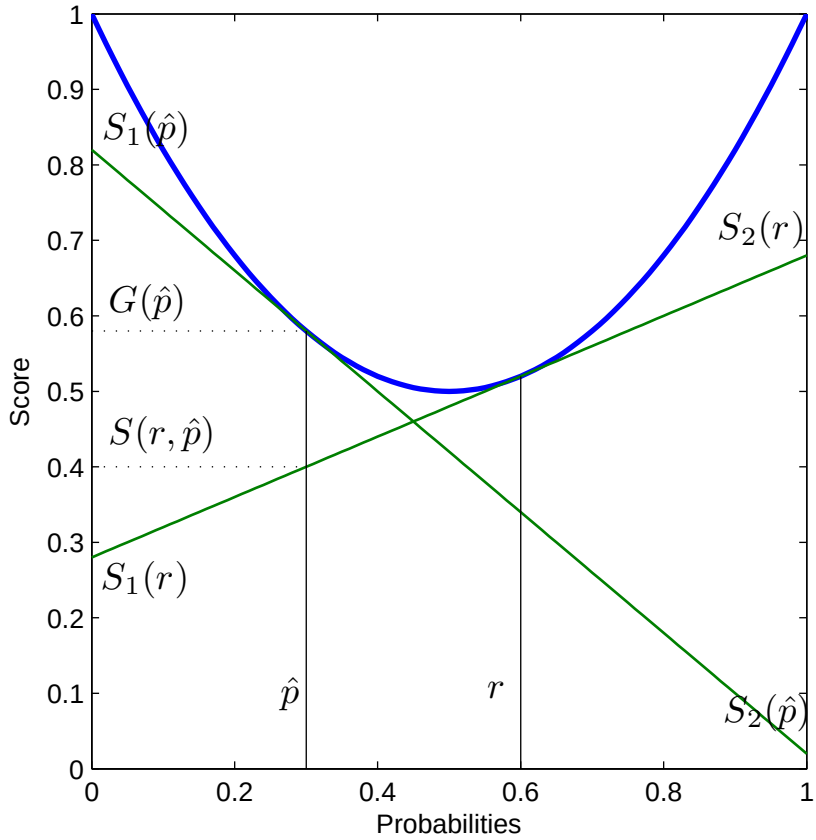

Figure 2.1: Graphical explanation for the McCarthy-Savage-Hendrickson-Buehler characterization theorem

When $n = 2$, (2.22) is equivalent to $S_1(r, 1-r) = S_2(1-r, r)$. This property has been hardly mentioned in the continuous setting but we can extend this by saying that a continuous scoring rule is symmetric if the definition above holds for any countable subset of Ω.

Most scoring rules present in the literature, including all those mentioned in Section 2.1.2 are symmetric, but there are a few notable exceptions. One of them is Winkler's (1994) asymmetric score $\breve{S}(\mathbf{r}, \omega)$ for the binary case (generated by a symmetric scoring rule S). Consider a symmetric scoring rule for binary events and a baseline parameter q, then

$$\begin{aligned} \breve{S}_1(r, 1-r||q, 1-q) &= \frac{S_1(r, 1-r) - S_1(r, 1-r)}{T(q, r)} \\ \breve{S}_2(r, 1-r||q, 1-q) &= \frac{S_2(r, 1-r) - S_2(q, 1-q)}{T(q, r)} \end{aligned} \tag{2.23}$$

is a strictly proper scoring rule, where

$$T(q, r) = \begin{cases} S_2(0, 1) - S_2(c, 1-c) & \text{if } p < q \\ S_1(1, 0) - S_1(c, 1-c) & \text{if } p \geq q \end{cases}.$$

The implication of the symmetry property is that the ordering of events does not matter and labeling of events could interchangeably be done. This is why some have also referred to this property as order invariance (e.g. Jose 2008).

Locality. One key observation in the computation of many scoring rules is that the score is dependent on the assessments provided for all states, including the states which did not occur. For example, note that the quadratic score is dependent on the term $||\mathbf{r}||_2^2$, which may vary depending on the number of states and the manner by which probabilities are assigned to the events that did not occur.

The logarithmic score however is not affected by this issue in the sense that the only input necessary to compute the score is the assessment made for the observed outcome ω. This property is referred to as locality.

Definition 5. *A scoring rule satisfies the locality property if* $\exists f$ *such that for all* ω, $S(R, \omega) = f(r(\omega))$.

The main implication of locality is that the score is solely dependent on the assessment made for the event that occurs. This makes local scoring rules consistent with the likelihood principle, which states that the information critical to an observed event is fully contained solely in the likelihood of that event, while probabilities for events that did not transpire should be treated as irrelevant.

When $n = 2$ all strictly proper scoring rules are local since the assessment on the second state is fixed upon the choice of the first due to coherence, i.e. $\int_\Omega p(\omega)d\mu(\omega) = 1$. When $n \geq 3$, this is no longer true.

Remark 3. *For any Ω, the only strictly proper scoring rule that satisfies locality is the logarithmic scoring rule. [Shuford, Albert & Massengill 1966]*

This characterization implies that if you want a rule to be strictly proper and local, then it has to be the logarithmic rule.

Effectiveness. The notion of effectiveness of scoring rules was first introduced by Friedman (1983).

Definition 6. *Given a metric (or distance function) ρ measurable with respect to $\mathscr{P}$, we say that a scoring rule is effective with respect to this metric if for every $P, Q, R \in \mathscr{P}$,*

$$\mathbb{E}_P S(Q) > \mathbb{E}_P S(R) \qquad \Leftrightarrow \qquad \rho(P, Q) < \rho(P, R). \tag{2.24}$$

Moreover, if there exists a scoring rule S where (2.24) holds then we say that the metric satisfying (2.24) is co-effective with respect to S. The desired result of this property is to guarantee that the expected score is a decreasing function of the distance between the reported distribution and the belief of the assessor under some metric, hence imposing a monotonicity condition on the expected score.

An example of an effective scoring rule is the spherical rule given by (2.11) which satisfies (2.24), under the re-normalized L^2 metric given by

$$\rho(P, R) = \left|\left| \frac{p(\omega)}{||P||_2} - \frac{r(\omega)}{||R||_2} \right|\right|.$$

Not all scoring rules are effective. For example, the logarithmic scoring rule is not effective with respect to any metric. Though some authors have suggested the theoretical desirability of effectiveness, Nau (1985) showed that most metrics do not have an associated co-effective scoring rule and that an implication of effectiveness is transitivity, which has been difficult to defend behaviorally.

Properties of the Loss Function. We can fully characterize a scoring rule by its expected score under truthful reporting using (2.18). Instead of expected scores, we can alternatively talk about the scoring rule in terms of the expected loss that one could get by not truthfully reporting probabilities. The expected loss in score from reporting R given the belief P is

$$L(R, P) = S(P, P) - S(R, P) = \mathbb{E}_P S(P) - \mathbb{E}_P S(R). \tag{2.25}$$

Since S is strictly proper, then $L(R, P) = 0$ if and only if $P = R$, implying that the maximization of one's expected score yields the same result as the minimization of one's expected loss.

Using this notation, conditions on the expected loss yield new properties that may be of interest in practice. One of these properties is the notion of neutrality.

Definition 7. *A scoring rule satisfies the neutrality property if for any $P, R \in \mathscr{P}, \mathscr{R}$, the expected loss of P given R is equal to the expected loss of R given P, i.e. $L(P, R) = L(R, P)$.*

Savage (1971) points out that the neutrality property can be replaced by the condition that there exists a function H such that $L(P, R) = H(P(\omega) - R(\omega))$, which implies that the differences between the assessor's beliefs and her reported probabilities are sufficient statistics for the expected loss. The imposition of the neutrality condition is pretty strong in the sense that it restricts the possible class of scoring rules to a single family.

Remark 4. *A scoring rule S satisfies symmetry, strict propriety and neutrality if and*

only if it is a positive affine transformation of the quadratic scoring rule. [Savage 1971, Selten 1998]

Selten (1998) defends this property by saying that

> *It is only fair to require that this [loss function] is 'neutral' in the sense that it treats both theories equally. If p is wrong and q is right then p should be considered to be as far from the truth as q in the opposite case that q is wrong and p is right.*
>
> *A scoring rule should not be prejudiced in favor of one of both theories in contest between p and q... Therefore, the neutrality axiom is a natural requirement to be imposed on a reasonable scoring rule. (Selten 1998, p. 54)*

In some cases, such an argument might hold. But sometimes the treatment of the two theories may not be equal. Consider the case where there are 3 possible outcomes and an assessor has two assessments in mind, namely, $(1, 0, 0)$ and $(1/3, 1/3, 1/3)$. It seems plausible that for some instances a greater loss should be given to one compared to the other depending on what he knows. Knowing that event 1 is going to occur and choosing to report an 'ignorant' assessment is very different than having an 'ignorant' prior and choosing to provide the categorical forecast. The argument that the two losses should be equal in the two cases is not compelling.

Jose (2009) introduced the notion of proportionality, which leads to some interesting geometric properties.

Definition 8. *A scoring rule satisfies the proportionality property if for any $P \neq R \in \mathscr{P}, \mathscr{R}$, the relationship between the expected loss of P given R and the expected loss of R given P satisfies:*

$$\frac{L(P,R)}{L(R,P)} = \frac{\mathbb{E}_R S(R)}{\mathbb{E}_P S(P)}.$$

This property leads to the spherical scoring rule.

Remark 5. *A scoring rule S satisfies symmetry, strict propriety and proportionality if and only if it is a positive affine transformation of the spherical scoring rule. [Jose 2008]*

This proportionality axiom provides an alternative representation of (expected) losses in the sense that the loss one incurs is dependent on the quality of information one has. Choosing to lie when one has "precise or valuable information" incurs a greater expected loss compared to when one has "less precise information".

The neutrality and proportionality axioms are closely related, differing only in terms of a constant determined by the assessments $\mathbf{p}$ and $\mathbf{r}$. If $||\mathbf{p}||_2 = ||\mathbf{r}||_2$ then the proportionality axiom becomes equivalent to the neutrality axiom. But for the spherical rule, this could only happen when the two predictions are equally sharp, meaning that the two theories generate the same expected gain when the assessor decides to report truthfully. This reinforces the idea that under the spherical rule expected losses are only equal in relative terms.

Decomposability. In probability assessments, many decision makers often consider several notions of goodness. For instance, a decision maker may want his assessor to provide a sharp forecast in the sense that it provides larger probabilities to some events. On the other hand, some may want to see the probabilities assigned to match in relative frequency with the actual observations.

This latter property is referred to as calibration or reliability, which relates to "the degree to which an appraiser's probabilities correspond in a relative frequency sense to what eventually occurs." (Winkler 1986) In an ex post sense, if an assessor asked to provide probabilities for the event that the Dow Jones Industrial Average will close lower than its opening value provides a forecast of 0.20 for twenty different days in a month, perfect or good calibration suggests that the Dow Jones average should have closed lower than its opening value approximately 4 days (20% of those 20 days) in that set.

To one extent, calibration may not seem to be an unreasonable expectation. As

Dawid (1982) notes, a coherent Bayesian assessor should expect to be well-calibrated. Using scoring rules, we can study the notion of calibration. Recalling (2.25), we have

$$\mathbb{E}_P S(R) = \mathbb{E}_P S(P) + L(R, P).$$

If we set $F \in \mathscr{P}$ to be the unknown distribution for the random variable $\mathbf{X}$, then we have

$$\begin{aligned} \mathbb{E}_F[\mathbb{E}_P S(R)] &= \mathbb{E}_F[\mathbb{E}_P S(P)] + \mathbb{E}_F L(R, P) \\ &= \int_\Omega \mathbb{E}_P S(P) dF(\omega) + \int_\Omega L(R, P) dF(\omega). \end{aligned}$$

For example, if we use the quadratic scoring rule S for a repeated experiment, then the average score measured by $\mathbb{E}_F \mathbb{E}_P S$ is given by

$$\mathbb{E}_F[\mathbb{E}_P S] = -\int_\Omega P(\omega)(1 - P(\omega)) dF(\omega) - \int_\omega (R(\omega) - P(\omega))^2 dF(\omega). \tag{2.26}$$

The first integral in the RHS of (2.26) refers to the sharpness of the forecaster. We say this refers to sharpness because it measures the over-all expected score of the assessor, which is maximized when the assessor provides forecasts much further from 1/2. This implies that truth-telling is encouraged and truthful assessments which are "more categorical" are given better expected scores. Moreover, the sharpness measure is independent of the assessment R because it assumes truth-telling.

The second integral in (2.26) refers to a calibration measure. To better illustrate this, consider the discrete case of (2.26) whose average score $\bar{S}$ can be written as:

$$\bar{S} = -\sum_{j=1}^{n} \frac{n_i}{n} p_i(1 - pi) - \sum_{j=1}^{n} \frac{n_i}{n} (r_i - p_i)^2,$$

where n_i is the number of actual observations for event i. Note that a well-calibrated person will have a calibration score of 0. This happens only when $P = R$, meaning the percentage of times that the event occurred when the assessor reported r_i is p_i.

The idea of separating the score into components has been a long standing tradition stemming from the statistical tradition of dividing the analysis of variance statistic into its components. Sanders (1963) was the first to provide such an analysis with a focus

on the quadratic rule (see also Blattenberger & Lad 1985, Murphy & Winkler 1987). The primary reasons for this are (1) the quadratic rule is the most commonly used rule, especially in the meteorological community where there is an active amount of research in this area, and (2) the availability of statistical tools that could be used stemming from the tradition of analyzing least squares. Similar decompositions exist for other scores with different functional forms for the sharpness and calibration measures.

Decompositions allow a decision maker to separate an assessor's skill via his ability to provide sharp forecasts and his ability to attain calibration. However, a decision maker should not focus solely on calibration as a measure of performance, ignoring skill. Foster and Vohra (1998) shows that the use of calibration scores instead of scoring rules could lead an assessor to game the system. This means that any assessor who would like to improve his score based on calibration can do so if he is given a sufficient number of tries.

Beyond Risk Neutrality. The idea that strict propriety implies that all assessor who likes to maximize their expected score (or minimize their expected loss) has to report truthfully is based on an implicit assumption that the assessor's utility is linear in the score. In some instances this may not be the case. Winkler and Murphy (1970) showed that the property of truth-telling may be lost when the assessor has a nonlinear utility for the score. For example, suppose that the assessor has exponential utility (with risk tolerance parameter $\tau > 0$) for the score:

$$u(S) = 1 - \exp^{-S/\tau}.$$

If $|\Omega| = 2$ and a quadratic scoring rule is used, then the expected utility of an assessor with belief $\mathbf{p}$ and report $\mathbf{r}$ is

$$\mathbb{E}_{\mathbf{p}} u(S(\mathbf{r})) = 1 - \mathbb{E}[\exp(S(\mathbf{r})/R)].$$

In this situation, truthful reporting is optimal only when $p \in \{0, \frac{1}{2}, 1\}$. If $p \in (1/2, 1)$, the optimal report is a value less than the true belief p, while if $p \in (0, 1/2)$ we

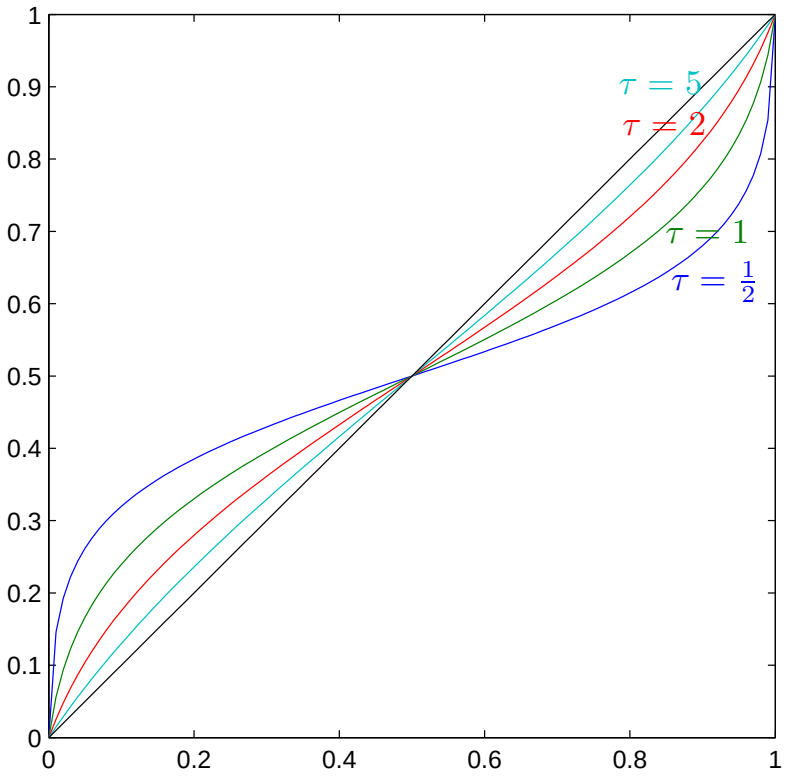

Figure 2.2: The effect of risk aversion

expect the reports to be greater than the true belief. This is not surprising since risk aversion suggests that the assessor would like to minimize her risk of assessing a large probability for an event which turns out not to occur. Figure 2.2 illustrates this point for various levels of τ. Note that as the individual becomes more and more risk averse, her forecasts tend towards 1/2 showing conservatism in the sense that she prefers not to commit too much probability to either state. On the other hand, as his risk tolerance increases, he becomes more and more risk neutral and his optimal strategy becomes closer to the truth-telling scenario.

The decision maker can correct for this if he knows the utility function by using the scoring rule $u^{-1}(S(\cdot,\cdot))$ so that the expected utility is simply the expected score. However, such knowledge of the assessor's utility function is a very strong assumption. Bickel (2007) suggests the logarithmic score as opposed to the quadratic and spherical scoring rules because it is not affected as much by distortions in the utility function.

Another possible complication is that the assessor may be placed in a competitive environment, competing with other experts. In this case, the assessor may care not only about her score but also about her rank among her peers. Lichtendahl and Winkler (2007) studied this environment and showed that overconfidence and extreme forecasts may result from this additional pressure on the assessor. Here, we ignore this possibility and note that this scenario can be avoided by making the environment feel less competitive to avoid the game-theoretic repercussions of competition among forecasters.

2.2 Information Measures

2.2.1 Entropy

One area of research which could provide insight when dealing with the notion of informativeness of probability assessments is information theory. This strand of literature originally started with the question of how engineers could efficiently design a com-

munication channel to reliably compress and transmit data, typically coming from a sequence of random events.

Shannon (1948) was the first to provide a quantitative model to measure information passing through an information channel. He showed that under the most efficient scheme, the average number of bits (the fundamental unit of information) required to report the occurrence of an event with relative frequency p is proportional to $\ln(1/p) = -\ln p$. Thus, using expectation under the distribution $\mathbf{p}$, the average number of bits required to report an event is proportional to

$$H(\mathbf{p}) = -\sum_{i=1}^{n} p_i \ln p_i. \tag{2.27}$$

In a more general setting, this measure which is called (Shannon's) entropy can be expressed in the form:

$$H(P) = -\int_{\Omega} \frac{dP}{d\mu}(\omega) \ln \frac{dP}{d\mu}(\omega) d\mu(\omega) \tag{2.28}$$

where $P \in \mathscr{P}$. This entropy measure satisfies several appealing properties.

Remark 6. *Shannon's entropy satisfies the following properties*

1. *Non-negativity.* $H(P) \geq 0$ *for all* $P \in \mathscr{P}$.

2. *Concavity.* $H(P)$ *is concave in* P.

3. *Continuity.* $H(P)$ *is continuous in* P. *This means that small changes in the probability should not lead to large jumps in the function.*

4. *Symmetry and Maximality. Based on the principle of insufficient evidence, if no event is more likely than another, i.e. the event space is equiprobable then* $H(p)$ *should be maximized at this point. Moreover, if events are reordered then the entropy score should be the same, i.e.* H *must be symmetric in its arguments.*

5. *Additivity. Consider two statistically independent partitions of the state space which generate distributions* $P_1, P_2 \in \mathscr{P}$. *If we measure the entropy of the distribution attained by consecutively imposing the partitions into the state space,*

additivity implies that the entropy measure should be the same regardless of which partition is used first, i.e.

$$H(P_1 \times P_2) = H(P_1) + H(P_2). \tag{2.29}$$

6. *Recursivity. This implies that*

$$\begin{aligned} H(p_1, p_2, p_3, \ldots, p_n) &= H(p_1 + p_2, p_3 \ldots, p_n) \\ &\quad +(p_1 + p_2)H(p_1, p_2). \end{aligned} \tag{2.30}$$

for all $P \in \mathscr{P}$.

It turns out that if one wishes to satisfy continuity, maximality and additivity, then the information measure has to be (2.27). We note that additivity can be replaced by recursivity.

Based on the maximality principle, Shannon's entropy, in effect, measures informativeness relative to the uniform distribution. If instead, we want to measure informativeness relative to some other distribution, an extension to Shannon's entropy that does this is known as Kullback-Leibler (1959) divergence.

Definition 9. *Let* $P, Q \in \mathscr{P}$. *The Kullback-Leibler divergence of* P *relative to* Q *is given by*

$$D_{KL}(P||Q) = \int_\Omega p(\omega) \ln \frac{p(\omega)}{q(\omega)} d\omega = \mathbb{E}_P[\ln(P/Q)]. \tag{2.31}$$

Note that if Q is the uniform measure then we obtain Shannon's entropy, showing that entropy is based on measuring the gain in moving from the uniform distribution to the probability distribution P.

Remark 7. *The Kullback-Leibler Divergence satisfies the following properties:*

1. *Non-negativity.* $D_{KL}(P||Q) \geq 0$ *for all* $P, Q \in \mathscr{P}$.

2. *Convexity.* $H(P)$ *is convex in* P.

3. *Continuity.* $H(P)$ *is continuous in* P. *This means that small changes in the probability should not lead to large jumps in the function.*

4. *Additivity. Suppose that* A *and* B *are statistically independent partitions of the state space whose prior distributions are* $\mathbf{q}_A$ *and* $\mathbf{q}_B$, *so that their prior joint distribution is* $\mathbf{q}_A \times \mathbf{q}_B$. *Now suppose that independent experiments are performed, which result in the updating of* $\mathbf{q}_A$ *and* $\mathbf{q}_B$ *to* $\mathbf{p}_A$ *and* $\mathbf{p}_B$, *respectively, so that the posterior joint distribution is* $\mathbf{p}_A \times \mathbf{p}_B$. *Then the total information gain of the two experiments is the sum of their separate Kullback-Leibler divergences:*

$$D_{KL}(\mathbf{p}_A \times \mathbf{p}_B \| \mathbf{q}_A \times \mathbf{q}_B) = D^{KL}(\mathbf{p}_A \| \mathbf{q}_A) + D^{KL}(\mathbf{p}_B \| \mathbf{q}_B). \tag{2.32}$$

5. *Recursivity. This implies that*

$$\begin{aligned} D_{KL}(\mathbf{p}\|\mathbf{q}) &= D_{KL}(p_1 + p_2, p_3, ..., p_n \| q_1 + q_2, q_3, ..., q_n) \\ &\quad + (p_1 + p_2) D_{KL}\left(\frac{p_1}{p_1 + p_2}, \frac{p_2}{p_1 + p_2} \| \frac{q_1}{q_1 + q_2}, \frac{q_2}{q_1 + q_2}\right). \end{aligned} \tag{2.33}$$

The recursivity property is a strong requirement but it seems very appealing. It suggests that when a refinement is provided after the transmission of a message which clarifies the distinction between two states which are combined, then the entropy measurement should be the same if only a "refined" message was sent.

Similar to Shannon's entropy, Kullback-Leibler divergence satisfies both additivity and recursivity. Weakening additivity and/or recursivity leads to more general measures of information.

2.2.2 ϕ-Divergence and Other Generalizations

Again, consider the σ-finite measure space $(\Omega, \mathscr{F}, \mu)$ and let Φ be the space of all continuous, convex functions ϕ defined on $[0, \infty]$ which satisfy the following properties:

$$\begin{aligned} &\text{(i) } \phi \text{ is finite on } (0,\infty) \text{ and strictly convex on some point } x \in (0,\infty), \\ &\text{(ii) } \phi(1) = 0, \\ &\text{(iii) } 0 \cdot \phi(0/0) = 0, \\ &\text{(iv) } 0 \cdot \phi(y/0) = \lim_{x\to\infty} \tfrac{\phi(x)}{x}. \end{aligned} \tag{2.34}$$

Divergence Measure	$\phi(x)$
Harmonic Mean	$\frac{1-x}{2} - \left(\frac{1+x^{-r}}{2}\right)^{-1/r}$
J-Divergence	$(x-1)\ln x$
Kullback-Leibler	$x \ln x$
Matsusita	$\lvert 1 - x^a \rvert^{1/a}$
Minimum Discrimination Information	$-\ln x + x - 1$
Pearson / Chi-Square	$\frac{1}{2}(x-1)^2$

Table 2.1: Some well-known ϕ-divergence families

Divergence Measure	$h(x)$	$\phi(x)$
Bhattacharya	$-\ln(1-x)$	$-\sqrt{x} + \frac{1+x}{2}$
Renyi	$\frac{1}{\beta(\beta-1)} \ln(\beta(\beta-1)x + 1)$	$\frac{x^\beta - r(\beta-1) - 1}{\beta(\beta-1)}$
Sharma-Mittal	$\frac{1}{\alpha-1}\left((1+\beta(\beta-1)x)^{\frac{\alpha-1}{\beta-1}} - 1\right)$	$\frac{x^\beta - r(\beta-1) - 1}{\beta(\beta-1)}$

Table 2.2: Some well-known (h, ϕ)-divergence families

Under this framework, if $P, Q \in \mathscr{P}$ we can define the following generalization of Kullback-Leibler divergence.

Definition 10. *Let $\phi \in \Phi$. The ϕ-divergence of P relative to Q is given by*

$$D_\phi(P||Q) = \int_\Omega \frac{dQ}{d\mu}(\omega)\phi\Big(\frac{dP/d\mu(\omega)}{dQ/d\mu(\omega)}\Big)d\mu(\omega) = \mathbb{E}_Q[\phi(P/Q)] \tag{2.35}$$

This measure is also known as an f-divergence, which was first introduced by Csiszár (1963). When $\phi(x) = x \ln x$, we obtain $D_{KL}(P||Q)$. Table 2.1 gives ϕ-functions that generate some of the well-known ϕ-divergences (Pardo 2006).

Remark 8. *$D_\phi(P||Q)$ satisfies the following properties:*

1. *Theorem of Range.* $0 \leq D_\phi(P||Q) \leq \phi(0) + \lim_{x\uparrow\infty} \dfrac{\phi(x)}{x}$.
2. *Theorem of Uniqueness. If $\phi \in \Phi$ then $D_\phi(P||Q) = D_\eta(P||Q)$ for all P, Q if and only if $\phi \cong \eta$.*

[Vajda 1989]

Other parametric families of divergences can be generated using monotonic transformations. In particular, we have:

Definition 11. *Let $\phi \in \Phi$ and let h be a differentiable increasing function defined from $\left[0, \phi(0) + \lim_{x\uparrow\infty} \frac{\phi(x)}{x}\right]$ onto $[0, \infty)$. The (h, ϕ)-divergence of P relative to Q is given by*

$$D_\phi^h(P||Q) = h(D_\phi(P||Q)). \tag{2.36}$$

The domain of the function h comes from the Theorem of Range. Table 2.2 provides some examples of the more well-known and well-studied families of (h, ϕ)-divergences. Note that when $h(x) = x$, $D_\phi^h(P||Q) = D_\phi(P||Q)$.

Chapter 3

Scoring Probability Assessments

3.1 Scoring Rules with Baselines

In many instances we are interested in evaluating probabilities relative to some *reference* or *baseline* distribution. For example, in a location where precipitation occurs on average 5% of the time it might be appropriate to measure the quality of a probability forecast of rain relative to this baseline.

A typical measure used in the meteorological community to test the performance of an assessor relative to some baseline distribution is a skill score (SS):

$$\mathrm{SS}(R,\omega) = \frac{S(R,\omega) - S(Q,\omega)}{S(T,\omega) - S(Q,\omega)} \tag{3.1}$$

where S is a strictly proper scoring rule, Q a baseline distribution and T a perfect forecast (i.e. a Dirac measure on the state that occurs). SS measures the improvement in score from the baseline Q to R divided by the improvement from Q to a perfect forecast. The denominator of this ratio is always positive by the monotonicity of S but the numerator may either be positive or negative, thus providing a measure of the intensity and direction of an assessor's forecasting skill.

However, the skill score does not encourage truth telling. Murphy (1973) shows that since the skill score is no longer strictly proper, hedging could pose a problem for individuals who solely want to maximize their expected score.

The first method suggested to incorporate the notion of a baseline which avoids the problem of hedging is the asymmetric rule of Winkler (1994) given in (2.23). Using a symmetric binary scoring rule such as the quadratic score to generate $\breve{S}$, the expected score function attains a maximum when the forecasts are categorical. Also since maximization is in terms of $\mathbf{r}$, truth telling ($\mathbf{r} = \mathbf{p}$) turns out to be still optimal. The asymmetric rule is scaled such that the lowest expected score under honest

reporting, zero, is achieved when $\mathbf{p} = \mathbf{q}$. The incorporation of a baseline suggests that the expected score should attain the lowest possible expected score for truth-telling if a person's state of information or belief is exactly the same as that of the baseline distribution.

For the other scoring rules mentioned earlier, particularly those which are symmetric, the expected score is measured relative to the uniform measure, which gives an equal weight to all the states in Ω.

Proposition 3.1.1. *Suppose Ω is a finite set with n elements. For a symmetric, strictly proper scoring rule S, the expected score under honest reporting is uniquely minimized when $\mathbf{r} = \mathbf{p} = (\frac{1}{n}, \frac{1}{n}, \ldots, \frac{1}{n}, \frac{1}{n})$.*

Proof. Using the McCarthy-Savage-Hendrickson-Buehler Characterization Theorem, we know that by the strict propriety of S, there exists a strictly convex function G for which $S_i(\mathbf{p}) = G(\mathbf{p}) - \langle G'(\mathbf{p}), \mathbf{p}\rangle + G_i'(\mathbf{p})$, where $G'(\mathbf{p})$ is a subgradient of G at $\mathbf{p}$. In particular we know that $G(\mathbf{p}) = \mathbb{E}_{\mathbf{p}} S(\mathbf{p})$. Now, since S is symmetric, then at $\tilde{\mathbf{p}} = (\frac{1}{n}, \frac{1}{n}, \ldots, \frac{1}{n}, \frac{1}{n})$, $S_i(\tilde{\mathbf{p}}) = S_j(\tilde{\mathbf{p}})$ and $G_i'(\tilde{\mathbf{p}}) = G_j'(\tilde{\mathbf{p}})$ for all $i, j \in \{1, \ldots, n\}$. Hence, using this result and the definition of a subgradient, we have for any $\mathbf{g}$,

$$\begin{aligned}
\mathbb{E}_{\mathbf{g}} S(\mathbf{g}) &\geq \mathbb{E}_{\tilde{\mathbf{p}}} S(\tilde{\mathbf{p}}) + \langle G'(\tilde{\mathbf{p}}), \mathbf{g} - \tilde{\mathbf{p}}\rangle \\
&= \mathbb{E}_{\tilde{\mathbf{p}}} S(\tilde{\mathbf{p}}) + \sum_{j=1}^{n} g_j G_j'(\tilde{\mathbf{p}}) - \sum_{j=1}^{n} \frac{1}{n} G_j'(\tilde{\mathbf{p}}) \\
&= \mathbb{E}_{\tilde{\mathbf{p}}} S(\tilde{\mathbf{p}}) + G_1'(\tilde{\mathbf{p}}) \Big(\sum_{j=1}^{n} g_j - \sum_{j=1}^{n} \frac{1}{n} \Big) \\
&= \mathbb{E}_{\tilde{\mathbf{p}}} S(\tilde{\mathbf{p}}),
\end{aligned}$$

which proves the result. □

One possible justification for this is Laplace's principle of insufficient reason, which states that without additional information the default prior should be the uniform distribution since we do not know which state is more likely to occur. In certain instances a uniform baseline distribution may not be an unreasonable assumption, e.g. in an assessment of the probability that a coin will land heads. But in many other

applications, this may not be reasonable.

Consider an insurance company asked to specify a price for a policy related to a catastrophic hurricane similar to Hurricane Andrew or Hurricane Katrina. In this case, a decision maker interested in the probability of a hurricane at a particular location, say Florida or Louisiana, would not view $(\frac{1}{2}, \frac{1}{2})$ as a reasonable baseline from which to measure how useful a probability is. Similarly, someone participating in a betting market for a particular sporting event might view the probabilities implied by the current market odds in betting sites such as Tradesports or VegasLine as reasonable benchmarks, since these reflect the decisions (hence the information) of a large number of presumably informed bettors.

By incorporating this notion of a baseline distribution, our objective is to try and better capture the skill of an assessor while at the same time reducing the "arbitrariness" of a scoring rule by allowing it to be tailored to the situation at hand instead of being held at an equal-probabilities default.

Are there other scoring rules that can incorporate this notion and be extended easily to more general cases (e.g. when Ω has more than 2 elements or if it were a subset of $\mathbb{R}^n$)? We will answer this through a generalization of the power and pseudospherical scoring rules referred to in (2.13) and (2.14).

Consider a baseline Q which sufficiently covers the set $\mathscr{P}$; i.e. for every ω, there exists no $R \in \mathscr{P}$ such that $r(\omega) > 0$ whenever $q(\omega) = 0$. In the discrete case, this is equivalent to having $q_j > 0$ for all states, while in the continuous case, this implies that Q does not assign zero measure on a set for which some P does in the set of allowable measures. This assumption may be somewhat strong but it coincides with the spirit of a baseline. If a state is believed by everyone to be an impossible event then no one should assign any probability mass on such a state.

The generalization then works by using this baseline as the ratio R/Q in place of R in the regular setting, such that the new "weighted" power and pseudospherical scores

can now be expressed as:

$$S^{P(\beta)}(R,\omega||Q) := \frac{(r(\omega)/q(\omega))^{\beta-1}-1}{\beta-1} - \frac{\mathbb{E}_Q[(R/Q)^{\beta}]-1}{\beta} \tag{3.2}$$

$$= \frac{(r(\omega)/q(\omega))^{\beta-1}-1}{\beta-1} - \frac{\mathbb{E}_R[(R/Q)^{\beta-1}]-1}{\beta} \tag{3.3}$$

$$\tag{3.4}$$

and

$$S^{S(\beta)}(R,\omega||Q) := \frac{1}{\beta-1}\left[\left(\frac{r(\omega)/q(\omega)}{(\mathbb{E}_Q[(R/Q)^{\beta}])^{1/\beta}}\right)^{\beta-1}-1\right] \tag{3.5}$$

$$= \frac{1}{\beta-1}\left[\left(\frac{r(\omega)/q(\omega)}{(\mathbb{E}_R[(R/Q)^{\beta-1}])^{1/\beta}}\right)^{\beta-1}-1\right] \tag{3.6}$$

for $\beta \in \mathbb{R}$. Note here that unlike the original construction of the power and pseudospherical scoring rules by Good (1971) and Selten (1998), these hold for all real β instead of simply $\beta > 1$. As a special case, when μ is a counting measure,

$$S_i^{P(\beta)}(\mathbf{r}||\mathbf{q}) := \frac{(r_i/q_i)^{\beta-1}-1}{\beta-1} - \frac{\mathbb{E}_{\mathbf{q}}[(\mathbf{r}/\mathbf{q})^{\beta}]-1}{\beta} \tag{3.7}$$

$$= \frac{(r_i/q_i)^{\beta-1}-1}{\beta-1} - \frac{\mathbb{E}_{\mathbf{r}}[(\mathbf{r}/\mathbf{q})^{\beta-1}]-1}{\beta} \tag{3.8}$$

$$\tag{3.9}$$

and

$$S_i^{S(\beta)}(\mathbf{r}||\mathbf{q}) := \frac{1}{\beta-1}\left[\left(\frac{r_i/q_i}{(\mathbb{E}_{\mathbf{q}}[(\mathbf{r}/\mathbf{q})^{\beta}])^{1/\beta}}\right)^{\beta-1}-1\right] \tag{3.10}$$

$$= \frac{1}{\beta-1}\left[\left(\frac{r_i/q_i}{(\mathbb{E}_{\mathbf{r}}[(\mathbf{r}/\mathbf{q})^{\beta-1}])^{1/\beta}}\right)^{\beta-1}-1\right] \tag{3.11}$$

for $i \in \Omega = \{1,\ldots,n\}$. Table 3.1 provides some special cases of these rules.

Note that for any fixed values of $\mathbf{r}$, $\mathbf{q}$, and β, the pseudospherical score vector $(S_1^{S(\beta)}(\mathbf{r}\|\mathbf{q}),\ldots,S_n^{S(\beta)}(\mathbf{r}\|\mathbf{q}))$ is a positive affine transformation of the power score vector $(S_1^{P(\beta)}(\mathbf{r}\|\mathbf{q}),\ldots,S_n^{P(\beta)}(\mathbf{r}\|\mathbf{q}))$, since both vectors are affine transformations of $(\mathbf{r}/\mathbf{q})^{\beta-1}$, although the origins and scale factors of the transformations vary with $\mathbf{r}$, $\mathbf{q}$, and β. Thus, although the two rules lead to different expected payoffs under truthful

	$S_i^{P(\beta)}(\mathbf{r}\|\|\mathbf{q})$	$S_i^{S(\beta)}(\mathbf{r}\|\|\mathbf{q})$
$\beta=-1$	$-\frac{1}{2}(1+(q_i/r_i)^2)+\mathbb{E}_{\mathbf{q}}[\mathbf{q}/\mathbf{r})]$	$\frac{1}{2}(1-((q_i/r_i)/\mathbb{E}_{\mathbf{q}}[\mathbf{q}/\mathbf{r})])^2)$
$\beta=0$	$1-(q_i/r_i)+\mathbb{E}_{\mathbf{q}}[\ln(\mathbf{q}/\mathbf{r})]$	$1-(q_i/r_i)\exp\left(-\mathbb{E}_{\mathbf{q}}[\ln(\mathbf{q}/\mathbf{r})]\right)$
$\beta=\frac{1}{2}$	$2\left(2-\sqrt{q_i/r_i}-\mathbb{E}_{\mathbf{r}}[\sqrt{\mathbf{q}/\mathbf{r}}]\right)$	$2\left(1-\sqrt{q_i/r_i}\mathbb{E}_{\mathbf{r}}[\sqrt{\mathbf{q}/\mathbf{r}}]\right)$
$\beta=1$	$\ln(r_i/q_i)$	$\ln(r_i/q_i)$
$\beta=2$	$((r_i/q_i)-1)-\frac{1}{2}\left(\mathbb{E}_{\mathbf{r}}[\mathbf{r}/\mathbf{q}]-1\right)$	$((r_i/q_i)/\sqrt{\mathbb{E}_{\mathbf{r}}[\mathbf{r}/\mathbf{q}]})-1$

Table 3.1: Special cases of the power and pseudospherical scoring rules.

reporting as a function of $\mathbf{p}$ (for the same $\mathbf{q}$ and β), and they create different incentives for information-gathering and different penalties for dishonest reporting, they nevertheless present the same relative risk profile to a truthful forecaster whose $\mathbf{r}$ is already fixed. At $\beta=1$ both rules converge to the weighted logarithmic score $\ln(r_i/q_i)$. At $\beta=2$, weighted forms of the quadratic and spherical scoring rules are obtained. The cases $\beta=0$ and $\beta=\frac{1}{2}$ have not received much (if any) attention in the antecedent literature, but it will be shown later that these yield interesting connections to some well-known concepts in information theory.

Note that the scaling and centering used here makes the rules (3.4)-(3.6) strictly proper for all values of β as will be later shown. In the earlier versions of these scores, they are strictly proper only for certain values of β. For example if we consider the pseudospherical scoring rule

$$S_i(\mathbf{r})=\frac{r_i^{\beta-1}}{||\mathbf{r}||_\beta^{\beta-1}}$$

with $\beta>1$, the expected score is

$$\begin{aligned}\mathbb{E}S_{\mathbf{p}}(\mathbf{r})=\frac{\sum_{j\in\Omega}p_j r_j^{\beta-1}}{||\mathbf{r}||_\beta^{\beta-1}} &\leq \frac{\left(\sum_{j\in\Omega}p_j^\beta\right)^{\frac{1}{\beta}}\left(\sum_{j\in\Omega}r_j^\beta\right)^{\frac{\beta-1}{\beta}}}{||\mathbf{r}||_\beta^{\beta-1}}\\ &= \frac{||\mathbf{p}||_\beta\cdot||\mathbf{r}||_\beta^{\beta-1}}{||\mathbf{r}||_\beta^{\beta-1}}\\ &= ||\mathbf{p}||_\beta\\ &= \mathbb{E}S_{\mathbf{p}}(\mathbf{p}),\end{aligned}$$

by Hölder's inequality. Then by coherence it would follow that the equality holds only when $\mathbf{p} = \mathbf{r}$; but for $\beta \leq 1$, this is no longer true.

For the weighted power and pseudospherical rules, the score received by an assessor is scaled such that when $P = Q$ she receives an expected score of 0. In the symmetric case, the highest score is achieved when a categorical forecast is given to the state that occurs. For these new scores it is achieved when the least likely event under the baseline (i.e. the state with the lowest q_i) occurs and a categorical forecast is given to that event. In the case where Q is the uniform measure, all events are equiprobable and we simply recover the regular symmetric scores.

What is achieved under this new asymmetric score is a correction in the way by which we measure the information value of knowing the distribution P instead of some other distribution Q as seen from the decision maker's perspective.

As an illustration of how the asymmetry affects the score, consider the case $\Omega = \mathbb{R}^2$, Figure 3.1 shows how the score of an assessor who provides an assessment of

$$P = N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \right)$$

changes as various realizations of ω are observed. For the regular "unweighted" version of the logarithmic rule, the plot clearly shows that the maximum score is achieved when the point $(0, 0)$ is observed since it is the point where the density function is highest. But when we introduce a baseline

$$Q = N\left(\begin{bmatrix} 1 \\ 1 \end{bmatrix}, \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix} \right),$$

the maximum value is attained at a different point. In fact is happens at the point $(-1, -1)$ because among all the values in $\mathbb{R}^2$ this yields the highest ratio $p(\omega)/q(\omega)$.

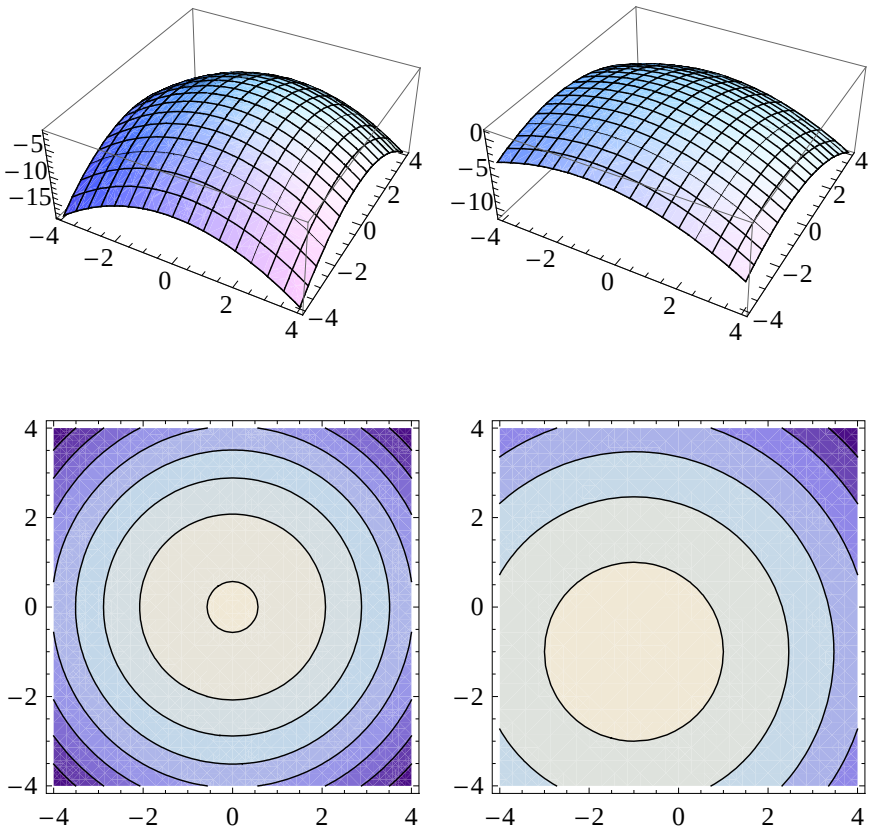

Figure 3.1: A comparison of the regular [left] and weighted [right] logarithmic scoring rule under various outcomes ω

This can be easily verified:

$$\begin{aligned} r(x,y) &= \frac{p((x,y))}{q((x,y))} \\ &= \exp\Big[\frac{1}{2}\Big(\frac{1}{2}(x-1)^2+\frac{1}{2}(y-1)^2\Big)+\frac{1}{2}(-x^2-y^2)\Big]. \\ \frac{\partial r}{\partial x} &= \Big[\frac{1}{2}(x-1)-x\Big]r(x,y)=0 \quad \Rightarrow \quad \frac{1}{2}(x-1)-x=0 \\ &\qquad\qquad\qquad\qquad\qquad\quad \Rightarrow \quad -1+x=2x \\ &\qquad\qquad\qquad\qquad\qquad\quad \Rightarrow \quad x^*=-1 \end{aligned}$$

which yields the maximal point $(-1,-1)$ after a similar computation for y^*, and verification of the second order condition.

With respect to expected scores, the power and pseudospherical scores with baseline Q yield "similar" forms given by:

$$\begin{aligned} \mathbb{E}S^{P(\beta)}(R||Q) &= \frac{E_P[(R/Q)^{\beta-1}]-1}{\beta(\beta-1)}, \\ \mathbb{E}S^{S(\beta)}(R||Q) &= \frac{\big(E_P[(R/Q)^{\beta-1}]\big)^{1/\beta}-1}{\beta-1}, \end{aligned}$$

which are close in form to the expected scores of (2.14) and (2.13) with the exception that now we use the ratio R/Q instead of simply R. We will show later that the expected scores are maximized when $P=R$, yielding the following expected scores under truth-telling:

$$\mathbb{E}S^{P(\beta)}(P||Q) = \frac{E_Q[(P/Q)^{\beta}]-1}{\beta(\beta-1)} = \frac{E_P[(P/Q)^{\beta-1}]-1}{\beta(\beta-1)}, \tag{3.12}$$

$$\mathbb{E}S^{S(\beta)}(R||Q) = \frac{\big(E_Q[(P/Q)^{\beta}]\big)^{1/\beta}-1}{\beta-1} = \frac{\big(E_P[(P/Q)^{\beta-1}]\big)^{1/\beta}-1}{\beta-1}. \tag{3.13}$$

To illustrate how the new expected score compares to the uniform baseline case, we consider the case when $|\Omega|=3$. Figure 3.2 shows the contours of the expected score in the simplex for four different scoring rules. These are the quadratic and spherical scores with baselines $(1/3,1/3,1/3)$ and $(7/10,2/10,1/10)$. Under the uniform baseline, these scores correspond to (2.7) and (2.11), while under the other baseline, we apply formulas (3.9) and (3.11).

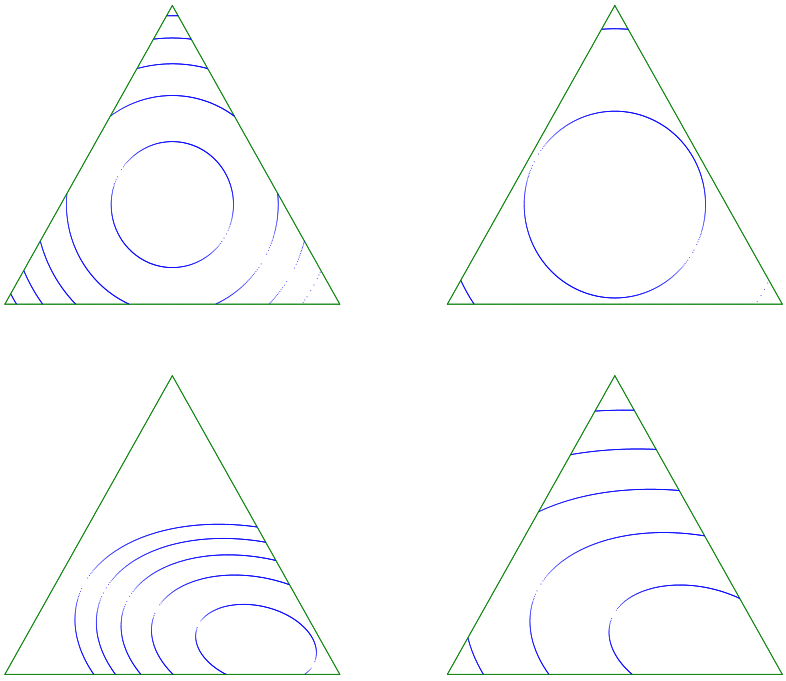

Figure 3.2: Expected score contours for the regular quadratic score [top left], the spherical score [top right], the power score with $\beta = 2$ [bottom left] and the pseudospherical score with $\beta = 2$ [bottom right] with baseline (7/10,2/10,1/10) for the latter two

For the symmetric quadratic and spherical rules the expected score under truthful reporting attains a minimum at the center of the simplex and behaves symmetrically as we move from the center to any corner of the simplex; for the lower plots in Figure 3.2 the "center" of the contours now shifts to the new baseline. Moving from left to right, we notice that the contours of the spherical rule are more dispersed in the sense that the range of values for the spherical score is much smaller. This is not too surprising since

$$Var[S^{S(\beta)}] = \frac{1}{(\beta-1)^2}\mathbb{E}_P[(P/Q)^{\beta-1}]^{2\left(\frac{1}{\beta}-1\right)}Var\Big[\Big(\frac{P}{Q}\Big)^{\beta-1}\Big]$$

and

$$Var[S^{P(\beta)}] = \frac{1}{(\beta-1)^2}Var\Big[\Big(\frac{P}{Q}\Big)^{\beta-1}\Big],$$

which indicates that the variance of the spherical score is a "dampened" version of the variance of the power score through a proportionality constant of $\mathbb{E}_P[(P/Q)^{\beta-1}]^{2\left(\frac{1}{\beta}-1\right)}$, which is typically less than 1.

In addition, the contours generate a convex function.

Proposition 3.1.2. *The expected scores for the power and pseudospherical scores given by (3.12) and (3.13) are strictly convex.*

Proof. We begin by showing that the pseudospherical score is convex. Consider two distributions $\mathbf{f}$ and $\mathbf{g}$, both in $\mathscr{P}$. For any $\lambda \in [0,1]$, if we denote $\mathbf{h} \equiv \lambda\mathbf{f} + (1-\lambda)\mathbf{g}$,

then we have:

$$\begin{aligned}
\mathbb{E}_{\mathbf{h}} S^{S(\beta)}(\mathbf{h}||\mathbf{q}) &= \frac{1}{\beta-1}\left[\left(\mathbb{E}_{\mathbf{h}}[(\mathbf{h}/\mathbf{q})^{\beta-1}]\right)^{1/\beta} - 1\right] \\
&= \frac{1}{\beta-1}\left\{\left[\left(\sum_{i\in\Omega}\frac{\lambda f_i + (1-\lambda)g_i}{q_i^{\beta-1}}\right)^{\beta}\right]^{1/\beta} - 1\right\} \\
&\leq \frac{1}{\beta-1}\left\{\left[\left(\sum_{i\in\Omega}\frac{\lambda f_i}{q_i^{\beta-1}}\right)^{\beta}\right]^{1/\beta}\right] + \left[\left(\sum_{i\in\Omega}\frac{(1-\lambda)g_i}{q_i^{\beta-1}}\right)^{\beta}\right]^{1/\beta}\right] - 1\right\} \\
&\quad \text{(via a generalized Minköwski's inequality (ref. Hardy et al. 1934))} \\
&= \frac{1}{\beta-1}\left\{\lambda\left[\left(\sum_{i\in\Omega}\frac{f_i}{q_i^{\beta-1}}\right)^{\beta}\right]^{1/\beta}\right] - \lambda \\
&\qquad + \left[(1-\lambda)\left(\sum_{i\in\Omega}\frac{g_i}{q_i^{\beta-1}}\right)^{\beta}\right]^{1/\beta} - (1-\lambda)\right]\right\} \\
&= \lambda\mathbb{E}_{\mathbf{f}} S^{S(\beta)}(\mathbf{f}||\mathbf{q}) + (1-\lambda)\mathbb{E}_{\mathbf{g}} S^{S(\beta)}(\mathbf{g}||\mathbf{q}),
\end{aligned}$$

which proves convexity. Strict convexity comes from the fact that the inequality holds only when all the values are equal to 0 or are all equal to 1, which can not happen in our context. Next consider the function

$$h(x) = \frac{[(\beta-1)x+1]^{\beta} - 1}{\beta(\beta-1)}$$

which has a derivative

$$h'(x) = [(\beta-1)y+1]^{\beta-1}.$$

When $\beta > 1$, $h'(x)$ is clearly positive. When $\beta < 1$, $h'(x) > 0$ if and only if $y \leq -\frac{1}{\beta-1}$. We then note that $\mathbb{E}_{\mathbf{p}} S^{S(\beta)}(\mathbf{p}||\mathbf{q}) \geq 0$ and attains a maximum when a categorical forecast is given to some state say i, i.e. the state for which q_j is lowest for all $j = 1, \ldots, n$. Therefore, for $\beta < 1$,

$$\mathbb{E}_{\mathbf{p}} S^{S(\beta)}(\mathbf{p}||\mathbf{q}) \leq -\frac{1}{\beta-1} - \frac{q_i^{-1+\frac{1}{\beta}}}{1-\beta} < -\frac{1}{\beta-1}.$$

We use this fact to show that for all possible values x of $\mathbb{E}_{\mathbf{p}} S^{S(\beta)}(\mathbf{p}||\mathbf{q})$, the function $h(x)$ is increasing. Hence, the expected score of the power rule $\mathbb{E}_{\mathbf{p}} S^{P(\beta)}(\mathbf{p}||\mathbf{q})$ must also be strictly convex since it is generated by an increasing function, namely p. For the continuous case, the proof follows exactly the same procedure using the generalized

	$S^{P(\beta)}(P,\omega\|\|Q)$	$S^{S(\beta)}(P,\omega\|\|Q)$
$\beta=-1$	$-\frac{1}{2}(1+(q(\omega)/p(\omega))^2)+E_Q[Q/P]$	$\frac{1}{2}(1-((q(\omega)/p(\omega))/E_Q[Q/P])^2)$
$\beta\to 0$	$1-(q(\omega)/p(\omega))+E_Q[\ln(Q/P)]$	$1-(q(\omega)/p(\omega))\exp(-E_Q[\ln(Q/P)])$
$\beta=\frac{1}{2}$	$4-2\sqrt{q(\omega)/p(\omega)}-2E_P[\sqrt{Q/P}]$	$2-2\sqrt{q(\omega)/p(\omega)}E_P[\sqrt{Q/P}]$
$\beta\to 1$	$\ln(p(\omega)/q(\omega))$	$\ln(p(\omega)/q(\omega))$
$\beta=2$	$(p(\omega)/q(\omega))-\frac{1}{2}(E_P[P/Q]+1)$	$((p(\omega)/q(\omega))/\sqrt{\mathbb{E}_P[P/Q]})-1$

Table 3.2: The expected scores under truth-telling with the weighted power and pseudospherical scoring rules for special cases of β

Minköwski inequality for integrals and the fact that the maximum score is attained with a Dirac measure on the least likely state with non-zero density. □

This property of the expected score is important since it will guarantee that the scoring rule encourages truth-telling.

Proposition 3.1.3. *The power and pseudospherical scoring rules given by (3.4) and (3.6) are strictly proper.*

Proof. Since the expected scores are strictly convex, the McCarthy-Savage-Hendrickson-Buehler Theorem implies that the scores are strictly proper for all β. □

Table 3.2 shows the functional forms of these rules for certain special or limiting values of β. In particular, for $\beta=2$ we have weighted versions of the quadratic and spherical rules. For $\beta\to 1$, we simply get the logarithmic rule for both families. Since the weighted logarithmic score is an affine transformation of the unweighted logarithmic score, it retains its locality property. In contrast, some other properties of rules from (2.14) and (2.13) for specific values of β do not carry over to (3.4) and (3.6). For example the weighted quadratic and spherical rules no longer retain their neutrality or proportionality properties.

In addition, these rules can also be easily decomposed into refinement and calibration terms. We provide the decomposition for the average score $\bar{S}$ in the discrete case

and note that the continuous versions follow analogously:

WEIGHTED POWER SCORE

$$
\begin{aligned}
\bar{S} &= \frac{1}{\beta-1}\sum_{i=1}^{n} p_i\Big(\Big(\frac{r_i}{q_i}\Big)^{\beta-1}-1\Big) - \frac{1}{\beta}\Big(\mathbb{E}_{\mathbf{r}}\Big[\Big(\frac{\mathbf{r}}{\mathbf{q}}\Big)^{\beta-1}\Big]-1\Big) \\
&= \frac{1}{\beta-1}\sum_{i=1}^{n} p_i\Big(\Big(\frac{r_i}{q_i}\Big)^{\beta-1}-1\Big) - \frac{1}{\beta}\big(\mathbb{E}_{\mathbf{r}}[(\mathbf{r}/\mathbf{q})^{\beta-1}]-1\big)-1\big) \\
&\qquad\qquad -\frac{\mathbb{E}_{\mathbf{p}}[(\mathbf{p}/\mathbf{q})^{\beta-1}]-1}{\beta(\beta-1)} + \frac{\mathbb{E}_{\mathbf{p}}[(\mathbf{p}/\mathbf{q})^{\beta-1}]-1}{\beta(\beta-1)} \\
&= \frac{1}{\beta(\beta-1)}\Big[\beta\sum_{i=1}^{n}\frac{p_i r_i^{\beta-1}}{q_i^{\beta-1}} - \beta - (\beta-1)\sum_{i=1}^{n}\frac{r_i^{\beta}}{q_i^{\beta-1}} + (\beta-1) - \sum_{i=1}^{n}\frac{p_i^{\beta}}{q_i^{\beta-1}} + 1\Big] \\
&\qquad\qquad +\frac{\mathbb{E}_{\mathbf{p}}[(\mathbf{p}/\mathbf{q})^{\beta-1}]-1}{\beta(\beta-1)} \\
&= \frac{1}{\beta(\beta-1)}\Big[\sum_{i=1}^{n}\Big(\frac{\beta p_i r_i^{\beta-1} - (\beta-1)r_i^{\beta} - p_i^{\beta}}{q_i^{\beta-1}}\Big)\Big] + \frac{\mathbb{E}_{\mathbf{p}}[(\mathbf{p}/\mathbf{q})^{\beta-1}]-1}{\beta(\beta-1)}
\end{aligned}
$$

The first term refers to a measure of calibration. It is equal to zero when $\mathbf{r} = \mathbf{p}$. Since we hold $\mathbf{r}$ to be fixed, the FOC w.r.t. $\mathbf{p}$ is $\beta r_i^{\beta-1} = \beta p_i^{\beta-1} = 0$. Note that the second order condition is satisfied since $\frac{\partial^2 \text{Calibration}}{\partial p^2} = -p_i^{\beta-2}$, which is negative when we have a sufficiently large sample. The second term refers to sharpness. This term attains a minimum as expected when $\mathbf{p} = \mathbf{q}$. Moreover, it attains a maximum when $p_j = 1$ (and $p_i = 0$ for $i \neq j$), where $j := \arg\min\{j : q_j\}$.

WEIGHTED PSEUDOSPHERICAL SCORE

Similarly, we have the following:

$$
\begin{aligned}
\bar{S} &= \frac{1}{\beta-1}\sum_{i=1}^{n} p_i \frac{(r_i/q_i)^{\beta-1}}{||\mathbf{r}/\mathbf{q}||_\beta^{\beta-1}} - \frac{1}{\beta-1} \\
&= \frac{1}{\beta-1}\sum_{i=1}^{n} p_i \frac{(r_i/q_i)^{\beta-1}}{||\mathbf{r}/\mathbf{q}||_\beta^{\beta-1}} - \frac{1}{\beta-1} - \frac{1}{\beta-1}\mathbb{E}_p[(\mathbf{p}/\mathbf{q})^{\beta-1}]^{1/\beta} \\
&\qquad + \frac{1}{\beta-1}\mathbb{E}_p[(\mathbf{p}/\mathbf{q})^{\beta-1}]^{1/\beta} \\
&= \frac{1}{\beta-1}\sum_{i=1}^{n} p_i \frac{(r_i/q_i)^{\beta-1}}{||\mathbf{r}/\mathbf{q}||_\beta^{\beta-1}} - \frac{1}{\beta-1}\mathbb{E}_p[(\mathbf{p}/\mathbf{q})^{\beta-1}]^{1/\beta} + \frac{1}{\beta-1}\Big[\mathbb{E}_p[(\mathbf{p}/\mathbf{q})^{\beta-1}]^{1/\beta} - 1\Big] \\
&= \frac{1}{\beta-1}\sum_{i=1}^{n} p_i \frac{(r_i/q_i)^{\beta-1}}{||\mathbf{r}/\mathbf{q}||_\beta^{\beta-1}} - \frac{1}{\beta-1}\mathbb{E}_p[(\mathbf{p}/\mathbf{q})^{\beta-1}]^{1/\beta}\frac{||\mathbf{p}/\mathbf{q}||_\beta^{\beta-1}}{||\mathbf{p}/\mathbf{q}||_\beta^{\beta-1}} \\
&\qquad + \frac{1}{\beta-1}\Big[\mathbb{E}_p[(\mathbf{p}/\mathbf{q})^{\beta-1}]^{1/\beta} - 1\Big] \\
&= \frac{1}{\beta-1}\Big[\frac{1}{||\mathbf{r}/\mathbf{q}||_\beta^{\beta-1}||\mathbf{p}/\mathbf{q}||_\beta^{\beta-1}}\sum_{i=1}^{n} p_i\Big((r_i/q_i)^{\beta-1}||\mathbf{p}/\mathbf{q}||_\beta^{\beta-1} - (p_i/q_i)^{\beta-1}||\mathbf{r}/\mathbf{q}||_\beta^{\beta-1}\Big)\Big] \\
&\qquad + \frac{1}{\beta-1}\Big[\mathbb{E}_p[(\mathbf{p}/\mathbf{q})^{\beta-1}]^{1/\beta} - 1\Big]
\end{aligned}
$$

Note that the the right hand term refers to a sharpness term and again the left hand term refers to a calibration term which is minimized when $\mathbf{p} = \mathbf{r}$. By quick inspection, this is 0 when the assessor is perfectly-calibrated. It can be shown that this generates the minimum score for the calibration term.

3.2 Connections and Motivations

Next, we show that these new weighted scoring rules have interesting connections with some measures that have been proposed in the information theory literature. After that, we present some generic expected utility optimization models that provide a decision theoretic connection between these scoring rules and the notion of information value in an economic setting.

3.2.1 Information Theoretic Connections

One of the key properties which made the logarithmic scoring rule popular with some researchers is the connection it has to information theory. Recall that the expected score for truthful reporting under the logarithmic score (2.9) is equal to the negative of Shannon's entropy. For the new families we have proposed, the logarithmic score is attained as $\beta \to 1$. With respect to the other values of β, there exist connections to two parametric families of divergence measures suggested in the literature.

We first show that the power score is related to a well-studied ϕ-divergence, while the pseudospherical score has an expected score which is an (h, ϕ)-divergence.

Consider the function

$$\phi_\beta(x) = \frac{x^\beta - \beta(x-1) - 1}{\beta(\beta-1)}, \qquad x \neq 0, 1 \tag{3.14}$$

This function satisfies (2.34), and thus $\phi \in \Phi$ for all $\beta \in \mathbb{R}\backslash\{0, 1\}$. Moreover, in the limit, as $\beta \to 0$ and 1 respectively, we have:

$$\begin{aligned} \phi_0(x) &= x \ln x - x + 1 \\ \phi_1(x) &= -\ln x + x - 1 \end{aligned}$$

which also satisfies the requirements of (2.34). This generates the divergence family

$$D_{\phi_\beta}(P||Q) = \frac{\mathbb{E}_P[(P/Q)^{\beta-1}] - 1}{\beta(\beta-1)} \tag{3.15}$$

for $\beta \in \mathbb{R}$. This family which only differs by a scaling constant was originally proposed by Havrda and Chavrát (1967), who called it the *directed divergence of order β between* $\mathbf{p}$ *and* $\mathbf{q}$. The notation used was for the discrete case since the original paper did not consider the continuous case. The scaling used in (3.15) was independently derived by Rathie and Kannappan (1972). Later on, Cressie and Read (1982) and Hausler and Opper (1997) independently came up with the same functional form in a different context.

Interestingly, $D_{\phi_\beta}(P||Q)$ is the expected score for honest reporting under the weighted power scoring rule. Hence, we shall refer to this as the *power divergence*

following Cressie and Read, who used this family in the context of statistical inference.

Similar to the weighted power scoring rule, this divergence family has some interesting special cases. In particular, at $\beta = 1$, the power divergence between P and Q is equal to the KL divergence, while at $\beta = \frac{1}{2}$, the power divergence is

$$D_{\phi_{1/2}}(P||Q) = 4\left(1 - \int \sqrt{p(\omega)q(\omega)}d\omega\right), \tag{3.16}$$

which is proportional to the squared Hellinger distance between P and Q, as noted by Haussler and Opper (1997). The Hellinger distance $D_H(P||Q)$ widely used in statistics is defined by

$$D_H(P||Q) \equiv \left(\int \left(\sqrt{p(\omega)} - \sqrt{q(\omega)}\right)^2 d\omega\right)^{1/2}, \tag{3.17}$$

which gives us the following relationship:

$$D_{\phi_{1/2}}(P||Q) = 2D_H(P||Q)^2. \tag{3.18}$$

At $\beta = 2$ the power divergence reduces to (a multiple of) another well-known divergence, the χ^2-divergence (Pearson 1900):

$$D_2(P||Q) = \frac{1}{2}(\mathbb{E}_P[P/Q] - 1) = \frac{1}{2}\chi^2(P||Q). \tag{3.19}$$

while at $\beta = -1$ it is the reverse Chi-square divergence $\frac{1}{2}\chi^2(Q||P)$.

On the other hand, the pseudospherical score is not a ϕ-divergence. Instead, it is a monotonic transformation of a ϕ-divergence. In particular, it is an h-transformation of the power divergence, with

$$h(x) = \frac{1}{\beta - 1}\left[\left(\beta(\beta-1)x + 1\right)^{\frac{1}{\beta}} - 1\right]. \tag{3.20}$$

The (h, ϕ)-divergence generated by (3.14) and (3.20) yields

$$D^h_{\phi_\beta}(P||Q) = \frac{\left(\mathbb{E}_P[(P/Q)^{\beta-1}]\right)^{\frac{1}{\beta}} - 1}{\beta - 1}, \tag{3.21}$$

which is the expected score of the pseudospherical scoring rule. For this reason, we shall call this the *pseudospherical divergence*. The roots of this divergence measure came much earlier in a paper written by Arimoto (1971) and further elaborated by Sharma

	$\mathbb{E}S^{P(\beta)}(P\|\|Q) = D_{\phi_\beta}(P\|\|Q)$	$\mathbb{E}S^{S(\beta)}(P\|\|Q) = D^h_{\phi_\beta}(P\|\|Q)$
$\beta = -1$	$\frac{1}{2}\chi^2(Q\|\|P)$	$\frac{1}{2}\left(1 - (\chi^2(Q\|\|P) + 1)^{-1}\right)$
$\beta \to 0$	$D_{KL}(Q\|\|P)$	$1 - \exp(-D_{KL}(Q\|\|P))$
$\beta = \frac{1}{2}$	$2D_H(P\|\|Q)^2$	$2\left(1 - \left(1 - \frac{1}{2}D_H(P\|\|Q)^2\right)^2\right)$
$\beta \to 1$	$D_{KL}(P\|\|Q)$	$D_{KL}(P\|\|Q)$
$\beta = 2$	$\frac{1}{2}\chi^2(P\|\|Q)$	$\sqrt{\chi^2(P\|\|Q) + 1} - 1$

Table 3.3: Weighted expected scores and corresponding generalized divergences

and Mittal (1975), Boekee and Van der Lubbe (1980) and Lavenda and Dunning-Davies (2003). Arimoto's original form is different from our form by a scaling constant and is expressed only for the discrete case, but the connections can be easily drawn from these forms to arrive at (3.21). Note that because of the transformation the special cases for the pseudospherical divergence will simply be transformations of the well-known measures in the power case. Table 3.3 provides the forms of these special cases for $\beta = -1, 0, \frac{1}{2}, 1$ and 2.

We now discuss some properties of these divergences. First, we note that these measures are always non-negative. For the power divergence, this follows immediately from the Theorem of Range, while for the pseudospherical divergence, it is true since h is increasing and $h(0) = 0$. This fact allows us to prove the following proposition easily.

Proposition 3.2.1. *The weighted power and pseudospherical scores attain a minimum expected score under truth-telling when $P = Q$.*

Proof. Since $0 \leq \mathbb{E}_P S^{P(\beta)}(P\|\|Q)$ and $\mathbb{E}_Q S^{P(\beta)}(Q\|\|Q) = 0$, the minimum is attained at the baseline distribution. The same is true for $S^{S(\beta)}$ since h is a monotonic transformation. □

In the discrete case when $\beta = 1$, we know that the power and pseudospherical divergences satisfy additivity and recursivity, but in general the power and pseudospherical divergences are neither additive nor recursive.

Proposition 3.2.2. *Let A and B be statistically independent partitions of Ω whose prior distributions are $\mathbf{q}_A$ and $\mathbf{q}_B$, so that their prior joint distribution is $\mathbf{q}_A \times \mathbf{q}_B$. Then suppose that independent experiments are performed, which result in the updating of $\mathbf{q}_A$ and $\mathbf{q}_B$ to $\mathbf{p}_A$ and $\mathbf{p}_B$, respectively, so that the posterior joint distribution is $\mathbf{p}_A \times \mathbf{p}_B$. Then the power and pseudospherical divergences satisfy the following pseudoadditivity conditions:*

$$\begin{aligned} D_{\phi_\beta}(\mathbf{p}_A \times \mathbf{p}_B || \mathbf{q}_A \times \mathbf{q}_B) &= D_{\phi_\beta}(\mathbf{p}_A||\mathbf{q}_A) + D_{\phi_\beta}(\mathbf{p}_B||\mathbf{q}_B) \\ &\quad +\beta(\beta-1)D_{\phi_\beta}(\mathbf{p}_A||\mathbf{q}_A)D_{\phi_\beta}(\mathbf{p}_B||\mathbf{q}_B) \\ D^h_{\phi_\beta}(\mathbf{p}_A \times \mathbf{p}_B || \mathbf{q}_A \times \mathbf{q}_B) &= D^h_{\phi_\beta}(\mathbf{p}_A||\mathbf{q}_A) + D^h_{\phi_\beta}(\mathbf{p}_B||\mathbf{q}_B) \\ &\quad +(\beta-1)D^h_{\phi_\beta}(\mathbf{p}_A||\mathbf{q}_A)D^h_{\phi_\beta}(\mathbf{p}_B||\mathbf{q}_B) \end{aligned}$$

Proof. We prove this by showing that

$$G^\alpha_\beta(\mathbf{p}_A \times \mathbf{p}_B || \mathbf{q}_A \times \mathbf{q}_B) = G^\alpha_\beta(\mathbf{p}_A||\mathbf{q}_A) + G^\alpha_\beta(\mathbf{p}_B||\mathbf{q}_B) + \alpha\beta(\beta-1)G^\alpha_\beta(\mathbf{p}_A||\mathbf{q}_A)G^\alpha_\beta(\mathbf{p}_B||\mathbf{q}_B), \tag{3.22}$$

where

$$G^\alpha_\beta(P||Q) = \frac{(\mathbb{E}_P[(P/Q)^{\beta-1}])^\alpha - 1}{\alpha\beta(\beta-1)}$$

with $\alpha = 1$ for the power case and $\alpha = \frac{1}{\beta}$ for the pseudospherical case. The RHS of (3.22) can now be expressed as:

$$\begin{aligned} RHS &= G^\alpha_\beta(\mathbf{p}_A||\mathbf{q}_A) + G^\alpha_\beta(\mathbf{p}_B||\mathbf{q}_B) + \alpha\beta(\beta-1)G^\alpha_\beta(\mathbf{p}_A||\mathbf{q}_A)G^\alpha_\beta(\mathbf{p}_B||\mathbf{q}_B) \\ &= \frac{1}{\alpha(\beta)(\beta-1)}\Big[\big(\mathbb{E}_{\mathbf{p_A}}[(\mathbf{p}_A/\mathbf{q}_A)^{\beta-1}]\big)^\alpha - 1 + \big(\mathbb{E}_{\mathbf{p_B}}[(\mathbf{p}_B/\mathbf{q}_B)^{\beta-1}]\big)^\alpha - 1 - 1 \\ &\quad +\Big(\big(\mathbb{E}_{\mathbf{p_A}}[(\mathbf{p}_A/\mathbf{q}_A)^{\beta-1}]\big)^\alpha - 1\Big)\cdot\Big(\big(\mathbb{E}_{\mathbf{p_B}}[(\mathbf{p}_B/\mathbf{q}_B)^{\beta-1}]\big)^\alpha - 1\Big)\Big] \\ &= \frac{1}{\alpha\beta(\beta-1)}\Big[\big(\mathbb{E}_{\mathbf{p_A}}[(\mathbf{p}_A/\mathbf{q}_A)^{\beta-1}]\big)^\alpha\big(\mathbb{E}_{\mathbf{p_B}}[(\mathbf{p}_B/\mathbf{q}_B)^{\beta-1}]\big)^\alpha - 1\Big] \\ &= \frac{1}{\alpha\beta(\beta-1)}\Big[\Big(\sum_i p_{A_i}\Big(\frac{p_{A_i}}{q_{A_i}}\Big)^{\beta-1}\sum_j p_{B_j}\Big(\frac{p_{B_j}}{q_{B_j}}\Big)^{\beta-1}\Big)^\alpha - 1\Big] \\ &= \frac{1}{\alpha\beta(\beta-1)}\Big[\Big(\sum_i\sum_j p_{A_i}p_{B_j}\Big(\frac{p_{A_i}p_{B_j}}{q_{A_i}q_{B_j}}\Big)^{\beta-1}\Big)^\alpha - 1\Big] \\ &= \frac{1}{\alpha\beta(\beta-1)}\Big[\big(\mathbb{E}_{\mathbf{p}_A\times\mathbf{p}_B}[((\mathbf{p}_A\times\mathbf{p}_B)/(\mathbf{q}_A\times\mathbf{q}_B))^{\beta-1}]\big)^\alpha - 1\Big] \\ &= G^\alpha_\beta(\mathbf{p}_A \times \mathbf{p}_B||\mathbf{q}_A \times \mathbf{q}_B). \qquad \square \end{aligned}$$

Under this new notion of additivity, the divergence acts differently in various regions. For $\beta \in (0,1)$, the power divergence is subadditive, i.e. $D_{\phi_\beta}(\mathbf{p}_A \times \mathbf{p}_B || \mathbf{q}_A \times \mathbf{q}_B) \leq D_{\phi_\beta}(\mathbf{p}_A||\mathbf{q}_A) + D_{\phi_\beta}(\mathbf{p}_B||\mathbf{q}_B)$, while for $\beta \notin [0,1]$, it is superadditive, i.e. $D_{\phi_\beta}(\mathbf{p}_A \times \mathbf{p}_B || \mathbf{q}_A \times \mathbf{q}_B) \geq D_{\phi_\beta}(\mathbf{p}_A||\mathbf{q}_A) + D_{\phi_\beta}(\mathbf{p}_B||\mathbf{q}_B)$. On the other hand, subadditivity (superadditivity) holds for the pseudospherical divergence only when $\beta < 1$ ($\beta > 1$), since the coefficient of the cross-term is simply $\beta - 1$ in this case instead of $\beta(\beta-1)$. When $\beta = 1$, pseudoadditivity reduces to additivity, but for the power divergence this is also true when $\beta = 0$, which is the special case of the inverse Kullback-Leibler divergence.

Neither rule satisfies recursivity. The power score, however, satisfies a recursive-like property.

Proposition 3.2.3. *The power divergence function satisfies the following pseudorecursivity property:*

$$\begin{aligned} D_{\phi_\beta}(\mathbf{p}||\mathbf{q}) &= D_{\phi_\beta}(p_1+p_2,\ldots,p_n||q_1+q_2,\ldots,q_n) \\ &\quad +(p_1+p_2)\Big(\frac{p_1+p_2}{q_1+q_2}\Big)^{\beta-1} D_{\phi_\beta}\Big(\frac{p_1}{p_1+p_2},\frac{p_2}{p_1+p_2}\Big|\Big|\frac{q_1}{q_1+q_2},\frac{q_2}{q_1+q_2}\Big). \end{aligned} \tag{3.23}$$

Proof. For ease of notation, let

$$\eta = (p_1+p_2)\big(\frac{p_1+p_2}{q_1+q_2}\big)^{\beta-1}.$$

The RHS of (3.24) can be expressed as

$$\begin{aligned} RHS &= \frac{1}{\beta(\beta-1)}\Big[\Big(\eta + \sum_{j=3}^{\infty} p_j\Big(\frac{p_j}{q_j}\Big)^{\beta-1} - 1\Big) + \eta\Big(\frac{p_1}{p_1+p_2}\Big(\frac{p_1(q_1+q_2)}{q_1(p_1+p_2)}\Big)^{\beta-1} \\ &\qquad + \frac{p_2}{p_1+p_2}\Big(\frac{p_2(q_1+q_2)}{q_2(p_1+p_2)}\Big)^{\beta-1} - 1\Big)\Big] \\ &= \frac{1}{\beta(\beta-1)}\Big[\Big(\eta + \sum_{j=3}^{\infty} p_j\Big(\frac{p_j}{q_j}\Big)^{\beta-1} - 1\Big) + \eta\Big(\eta^{-1}p_1\Big(\frac{p_1}{q_1}\Big)^{\beta-1} \\ &\qquad + \eta^{-1}p_2\Big(\frac{p_2}{q_2}\Big)^{\beta-1} - 1\Big)\Big] \\ &= \frac{1}{\beta(\beta-1)}\Big[p_1\Big(\frac{p_1}{q_1}\Big)^{\beta-1} + p_2\Big(\frac{p_2}{q_2}\Big)^{\beta-1} + \sum_{j=3}^{\infty} p_j\Big(\frac{p_j}{q_j}\Big)^{\beta-1} - 1\Big] \\ &= D_{\phi_\beta}(\mathbf{p}||\mathbf{q}). \qquad \square \end{aligned}$$

Notice that when $\beta = 1$, this reduces to the original notion of recursivity. For other values of β, the extra factor $\left(\frac{p_1+p_2}{q_1+q_2}\right)^{\beta-1}$ may either increase or decrease the sum. For example, when $\beta > 1$ the entropy of the "refined" distribution (i.e. the one which needs only one transmission) is smaller (larger) than the "unrefined" distribution if $p_1 + p_2 < (>) q_1 + q_2$. On the other hand, the pseudospherical score does not satisfy any obvious recursive property.

3.2.2 A Utility Maximization Framework

We begin by considering two generic optimization problems whose solution provides an appropriate measure of the information value contained in an assessor's probabilistic forecast. In these models we consider a risk averse decision maker with probability distribution P who bets against a non-strategic risk neutral opponent with distribution Q. Alternatively, we can represent this problem as one where the same decision maker invests in a complete market for contingent claims whose supporting risk neutral distribution is described by Q.

Now, let $\omega \in \Omega \subset \mathbb{R}^n$ denote a generic state, let $X \subseteq \mathbb{R}$ be the set of outcomes for the decision maker (gains or losses in units of money), let $a : \Omega \mapsto X$ denote an act (or decision) whose outcome in state ω is $a(\omega) \in X$, let $\mathscr{A}$ denote the set of available acts, and let p and q denote the probability density functions of the forecaster and opponent, respectively (i.e., the forecaster's probability density evaluated at state ω is $p(\omega)$).

In the first model (**"S"**), we consider a single period consumption model wherein a decision maker is endowed with a single-attribute von Neumann-Morgenstern utility function $u(x)$. Her objective is to find the payoff vector $\mathbf{x}$ that maximizes her subjective expected utility subject to the self-financing constraint $\mathbb{E}_{\mathbf{q}}[\mathbf{x}] \leq 0$. It then follows that this decision maker's optimal expected utility, denoted $U^{\mathbf{S}}(P||Q)$, is determined

by:

$$\textbf{Model S}: \qquad U^{\mathbf{S}}(P||Q) \equiv \max_{a \in \mathscr{A}} \mathbb{E}_P[u(a)] \text{ s.t. } \mathbb{E}_Q[a] \leq 0. \tag{3.24}$$

In the second problem (**"P"**), we consider a two-period scenario in which consumption occurs and the decision maker with probability distribution P has a quasilinear utility function $u(a,b) = a + u(b)$ where a is money consumed at time-0 and b is money consumed at time-1. The decision maker's objective is to maximize the overall consumption in both periods by purchasing at market prices a vector of time-1 payoffs using time-0 funds. Then the cost at time-0 for the act a or the cost of achieving the vector $\mathbf{x}$ is $\mathbb{E}_Q[a]$. Therefore, the decision maker's optimal expected utility, denoted $U^{\mathbf{P}}(P||Q)$ can be obtained from:

$$\textbf{Model P}: \qquad U^{\mathbf{P}}(P||Q) \equiv \max_{a \in \mathscr{A}} \mathbb{E}_P[u(a)] - \mathbb{E}_Q[a] \tag{3.25}$$

Note that the assumption of a risk neutral opponent is without loss of generality. For the models we are considering it is sufficient to require at least one player to be risk averse and the forecaster to be more risk averse than the opponent is risk seeking. Alternatively, we can simply require that the forecaster is less risk seeking than the opponent is risk averse, which is guaranteed when the sum of their Arrow-Pratt absolute risk aversion measures is positive throughout any transaction. Hence, if we denote $u(x) \equiv \tilde{u}(-v^{-1}(x))$ and $\tilde{v}(z) \equiv -v(-\tilde{u}^{-1}(-z))$, then the following problems are equivalent:

$$\begin{aligned}
\textbf{Model S}: &\qquad U^{\mathbf{S}}(P||Q) \equiv \max_{a \in \mathscr{A}} \mathbb{E}_P[u(a)] \text{ s.t. } \mathbb{E}_Q[a] \leq 0. \\
\textbf{Model S1}: &\qquad U^{\mathbf{S}}(P||Q) \equiv \max_{b \in -v(\mathscr{A})} \mathbb{E}_P[\tilde{u}(b)] \text{ s.t. } \mathbb{E}_Q[v(b)] \leq 0. \\
\textbf{Model S2}: &\qquad U^{\mathbf{S}}(P||Q) \equiv \max_{c \in \tilde{u}(\mathscr{A})} \mathbb{E}_P[c] \text{ s.t. } \mathbb{E}_Q[\tilde{v}(c)] \leq 0.
\end{aligned}$$

Analogously, for Model P, we have

$$\begin{aligned}
\textbf{Model P}: &\qquad U^{\mathbf{P}}(P||Q) \equiv \max_{a \in \mathscr{A}} \mathbb{E}_P[u(a)] - \mathbb{E}_Q[a] \\
\textbf{Model P1}: &\qquad U^{\mathbf{P}}(P||Q) \equiv \max_{b \in -v(\mathscr{A})} \mathbb{E}_P[\tilde{u}(b)] - \mathbb{E}_Q[v(b)] \\
\textbf{Model P2}: &\qquad U^{\mathbf{P}}(P||Q) \equiv \max_{c \in \tilde{u}(\mathscr{A})} \mathbb{E}_P[c] - \mathbb{E}_Q[\tilde{v}].
\end{aligned}$$

	Name	$u_\beta(x)$
$\beta = -1$	Quadratic utility	$-\frac{1}{2}((1-x)^2 - 1)$
$\beta \to 0$	Exponential utility	$1 - \exp(-x)$
$\beta = \frac{1}{2}$	Reciprocal utility	$2\left(1 - \frac{1}{1+x/2}\right)$
$\beta \to 1$	Logarithmic utility	$\ln(1+x)$
$\beta = 2$	Square-root utility	$\sqrt{1+2x} - 1$

Table 3.4: Examples of normalized linear-risk-tolerance utility functions

Next, we focus our attention on the utility function $u(x)$. We consider a normalized risk averse utility function whose risk tolerance (the reciprocal of the Arrow-Pratt risk aversion measure, $-u''(x)/u'(x)$) is a linear function of wealth, parameterized as:

$$u_\beta(x) = \frac{1}{\beta - 1}[(1 + \beta x)^{(\beta-1)/\beta} - 1] \tag{3.26}$$

for $\beta \in \mathbb{R}$. Under this parametrization, this family contains some well-known special (or limiting) cases such as the exponential and logarithmic utility functions. The function in (3.26) with the slope of $u_\beta(x)$ at the origin equalling 1 is normalized in the sense that it is centered at the origin, i.e., $u_\beta(0) = 0$ for all $\beta \in \mathbb{R}$. Hence, for Model **P** the marginal rates of substitution between time-0 consumption and time-1 consumption are the same. Table 3.4 provides the functional form of some of these special cases, while Figure 3.3 shows how these plots compare. Since the risk tolerance is linear in wealth, this family is known as the hyperbolic absolute risk aversion (HARA) family of utility functions.

HARA utility functions are all continuously differentiable functions in some finite neighborhood around the origin. They exhibit a convenient symmetry around $\beta = \frac{1}{2}$, meaning that $u_{1-\beta}(x) = -u_\beta(-x)$, or equivalently, $u_\beta(-u_{1-\beta}(-x)) = x$. Graphically, this means that the plot of $u_{1-\beta}(x)$ is obtained from the graph of $u_\beta(x)$ by reflecting it around the line $y = -x$. The power (exponent) in u_β is the term $(\beta - 1)/\beta$, which has the property that $((\beta - 1)/\beta)^{-1} = ((1-\beta) - 1)/(1 - \beta)$, so that swapping β for $1 - \beta$ results in another power utility function whose power is the reciprocal of the

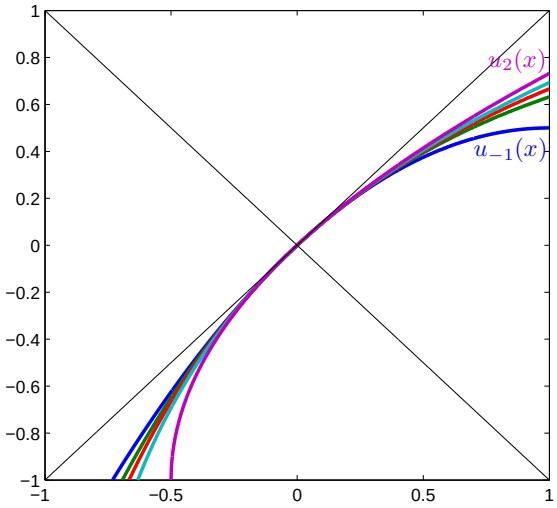

Figure 3.3: Some commonly used HARA utility functions

original. Thus, under this parametrization, the reciprocal utility function ($\beta = \frac{1}{2}$) is its own reflection around the line $y = -x$, the exponential and log utility functions ($\beta = 0$ and $\beta = 1$) are reflections of each other, and the power utility function with exponent δ is the reflection of the power utility function with exponent $1/\delta$ for any positive or negative δ other than 0 or 1.

Descriptively, we say that the set of functions u_β is a family of functions which can be described by its skewness with respect to a reference function. At the center of this family (at $\beta = 0$) is the exponential utility function. For values less than 0, we have functions whose left tail is said to be more linear than the exponential, namely the power functions with positive exponents. On the other hand, for $\beta > 0$ we have the opposite: the right tail is more linear than the exponential function. This side contains the logarithmic function and the power family with negative exponents. Finally, as β approaches positive or negative infinity, the utility function approaches a linear function, implying risk neutrality. Though all HARA utility functions exhibit risk aversion, those with non-decreasing risk tolerance (i.e. $\beta \geq 0$) are the ones which are generally compatible with economic intuition.

Putting all of these together, the main result we have is that these models yield a connection to the scoring rules developed in this chapter. Thus, there exists a decision theoretic motivation for these new parametric families of scoring rules using utility functions that are commonly encountered in practice. What will come out of all these models is that the optimal expected utility that we get coincides with the expected scores of these new families of scoring rules, which imply that the information value that is measured in these decision models coincides with that of the scoring rule.

Proposition 3.2.4. *The optimal expected utility for Model* $\boldsymbol{P}$ *given by (3.25) with a utility function parameterized by* β *in (3.26) is the expected score of the pseudospherical score, i.e.* $U^{\mathbf{P}}(P||Q) = \mathbb{E}_P S^{P(\beta)}(P||Q)$. *The same result holds for Model* $\boldsymbol{S}$ *and the expected score of the power scoring rule.*

Proof. We begin the proof for Model $\mathbf{P}$ in the continuous setting. Writing out the maximization problem, we have

$$\max \qquad \mathbb{E}_P[u_\beta(a)] + \mathbb{E}_Q[-a] = \int [p(\omega)u_\beta(a(\omega)) - q(\omega)a(\omega)]d\omega.$$

Applying the Euler-Lagrange Theorem, we have

$$\begin{aligned} 0 &= \frac{\partial}{\partial a}\Big[p(\omega)\Big(\frac{1}{\beta-1}\Big(1+\beta a(\omega)\Big)^{\frac{\beta-1}{\beta}} - 1\Big) - q(\omega)\Big] - 0 \\ &= p(\omega)\Big(1+\beta a(\omega)\Big)^{-\frac{1}{\beta}} - q(\omega). \end{aligned}$$

Simplifying, this leads to

$$\begin{aligned} \Big(1+\beta a(\omega)\Big)^{\frac{1}{\beta}} &= \frac{p(\omega)}{q(\omega)} \\ 1+\beta a(\omega) &= \Big(\frac{p(\omega)}{q(\omega)}\Big)^{\beta} \\ a(\omega) &= \frac{1}{\beta}\Big[\Big(\frac{p(\omega)}{q(\omega)}\Big)^{\beta} - 1\Big]. \end{aligned}$$

Plugging this into the objective function leads to an optimal utility payoff in state ω

of

$$\begin{aligned} g_\beta(\mathbf{d}) + \mathbb{E}_{\mathbf{q}}[-(\mathbf{d})] &= \frac{1}{\beta-1}\left\{\left[\left(\frac{p(\omega)}{q(\omega)}\right)^\beta\right]^{\frac{\beta-1}{\beta}} - 1\right\} - \left[\int \frac{p(\omega)^\beta}{\beta q(\omega)^{\beta-1}} d\omega - \int \frac{q(\omega)}{\beta} d\omega\right] \\ &= \frac{1}{\beta-1}\left[\left(\frac{p(\omega)}{q(\omega)}\right)^{\beta-1} - 1\right] - \frac{1}{\beta}\left[\int q(\omega)\left(\frac{p(\omega)}{q(\omega)}\right)^\beta d\omega - 1\right] \\ &= \frac{1}{\beta-1}\left[\left(\frac{p(\omega)}{q(\omega)}\right)^{\beta-1} - 1\right] - \frac{1}{\beta}[\mathbb{E}_Q[(P/Q)^\beta] - 1] \\ &= \mathbb{E}_P S^{P(\beta)}(P||Q), \end{aligned}$$

which is the desired result. For Model **S**

$$\begin{aligned} \max \quad & \mathbb{E}_P[u_\beta(a)] = \int u_\beta(a(\omega)) dP(\omega) \\ \text{s.t.} \quad & \mathbb{E}_Q[-a] = -\int a(\omega) dQ(\omega) \geq 0, \end{aligned}$$

we begin by introducing a Lagrange multiplier λ and proceed in a similar fashion with Model **P**. In particular, we have the following problem:

$$\max \quad \int [p(\omega) g_\beta(a(\omega)) - \lambda q(\omega) a(\omega)] d\omega$$

with the Euler-Lagrange condition (based on what we have done earlier)

$$p(\omega)\Big(1 + \beta a(\omega)\Big)^{-\frac{1}{\beta}} - \lambda q(\omega) = 0.$$

This simplifies to

$$\begin{aligned} 1 + \beta a(\omega) &= \left(\frac{p(\omega)}{\lambda q(\omega)}\right)^\beta \\ a(\omega) &= \frac{1}{\beta}\left[\left(\frac{p(\omega)}{\lambda q(\omega)}\right)^\beta - 1\right]. \end{aligned}$$

Since $u_\beta(a(\omega))$ is a utility function, we know that it is strictly increasing in $a(\omega)$. Also, since $q(\omega)$ is always non-negative then we will attain a maximum for the objective

function when this single constraint is binding. This gives us

$$\begin{aligned}
\int q(\omega)a(\omega)d\omega = \int q(\omega)\Big[\frac{1}{\beta}\Big[\Big(\frac{p(\omega)}{\lambda q(\omega)}\Big)^{\beta} - 1\Big]\Big]d\omega &= 0 \\
\frac{1}{\beta}\Big[\int q(\omega)\Big(\frac{p(\omega)}{\lambda q(\omega)}\Big)^{\beta}d\omega - 1\Big] &= 0 \\
\int q(\omega)\Big(\frac{p(\omega)}{\lambda q(\omega)}\Big)^{\beta}d\omega &= 1 \\
\int q(\omega)\Big(\frac{p(\omega)}{q(\omega)}\Big)^{\beta}d\omega &= \lambda^{\beta} \\
\lambda^{*} &= \mathbb{E}_q[(p/q)^{\beta}]^{1/\beta}.
\end{aligned}$$

Plugging this into $a(\omega)$, we then have

$$\begin{aligned}
a^{*}(\omega) &= \frac{1}{\beta}\Big[\Big(\frac{p(\omega)}{\mathbb{E}_Q[(P/Q)^{\beta}]^{1/\beta}q(\omega)}\Big)^{\beta} - 1\Big] \\
&= \mathbb{E}_P S^{S(\beta)}(P||Q).
\end{aligned}$$

The proof for the discrete case is similar (cf. Jose, Nau & Winkler 2008). □

Proposition 3.2.5. *For all $P, Q \in \mathscr{P}$ and $\beta \in \mathbb{R}$, $\mathbb{E}_P S^{P(\beta)}(P||Q) \geq \mathbb{E}_P S^{S(\beta)}(P||Q)$.*

Proof. This follows from the fact that both expected scores are solutions to the optimization problem: $\max_x \mathbb{E}_P[u_\beta(x)] + \lambda\mathbb{E}_Q[-x]$, with the power score fixing λ at 1. On the other hand, the pseudospherical score allows λ to change in a way such that it minimizes the maximum expected utility. Hence, the result holds. □

As always, since $\beta = 1$ is a common point for the two families of functions, we know that the equality holds for all measures $P, Q \in \mathscr{P}$. Figure 3.4 clearly shows this for two examples. The left plot shows how the expected score changes as we move β when $\mathbf{p} = (1/3, 1/3, 1/3)$ and the baseline is $\mathbf{q} = (7/10, 2/10, 1/10)$, as given in an earlier example; the right one illustrates the expected scores for a continuous assessment $P = N(0, 1)$ and baseline $Q = N(1, 1)$.

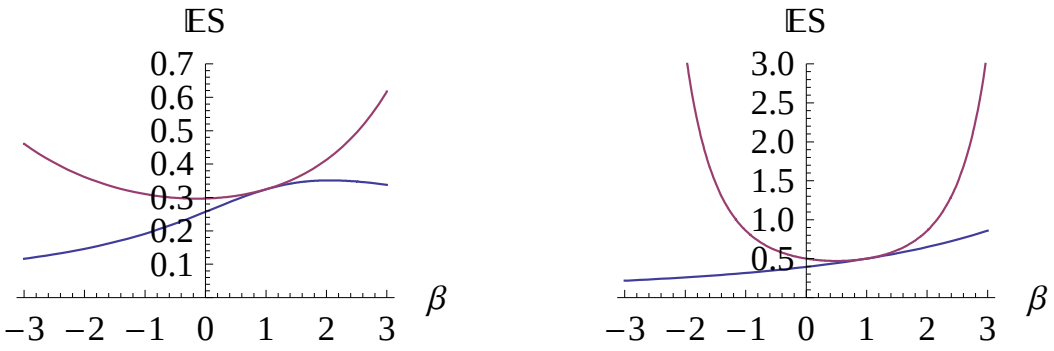

Figure 3.4: Expected score for various values of β

3.3 An Application in Finance

In many instances, it is difficult to single out the risk neutral supporting distribution Q that supports the market prices. For example, consider prices posted in a prediction where the bid-ask spread is wide. In this instance, the most that we can say is that the risk neutral supporting distribution lies in this interval. Hence, instead of a singleton Q, we may have to deal with some set $\mathscr{Q}$ that contains the risk-neutral distribution. This set is generally broad and the only typical constraint for this set is that it is non-empty and convex. The first is a natural requirement while the latter follows from the no arbitrage principle. Recently, several papers have studied the connection between maximizing expected utility and minimizing divergence in this context. (e.g. Fritelli 2000, Goll and Ruschendorf 2001, Delbaen et al 2002) Most of these however have focused on the exponential utility for which the dual problem is associated with the inverse Kullback-Leibler divergence.

Here we will show that the duality holds in this setting as well for the two parametric families that we have introduced. For ease of discussion, we focus on the discrete case. We note that the continuous version also holds. A more detailed discussion of this topic and further references can be found in Nau et al. (2007, 2009).

Again, we consider a risk-averse individual with utility function u_β whose belief about the market is represented by the distribution $\mathbf{p}$ and who invests in a contin-

gent claims market whose set of risk neutral distribution lies in the set $\mathscr{Q}$. If we let $\mathbf{x} \in \mathscr{A} = \mathbb{R}^n$ be the vector of monetary payoffs, then the decision maker has a utility vector $u_\beta(\mathbf{x}) = (u_\beta(x_1), \ldots, u_\beta(x_n))$ which contains the utility she will receive for each possible outcome. Next, we note that we can characterize the set $\mathscr{Q}$ by some $m \times n$ matrix $\mathbf{Q} = \{q_{ij}\}$. The rows of this matrix contains the extremal risk neutral distributions of the set $\mathbf{Q}$, while the q_{ij} entry refers to the probability that the jth state occurs under the ith risk neutral distribution. This characterization is possible since any point in this convex set can be represented as a convex combination of the extremal points of the set.

For Model **S**, we consider again a single time period in which the decision maker wants to find the vector of payoffs $\mathbf{x}$ that maximizes her subjective expected utility subject to a self-financing constraint. In the complete market case, we represent this constraint by $\mathbb{E}_\mathbf{q}[\mathbf{x}] \leq 0$. In this setting, we write this as $\mathbf{Qx} \leq 0$, which means that market's expected payoff vector under all extremal risk neutral distribution is non-positive. This implies that the incentive compatibility constraint holds for every possible distribution in the set $\mathscr{Q}$. Mathematically, we can write this as:

$$\textbf{Model S:} \quad U_\beta^S(\mathbf{p}||\mathbf{q}) \equiv \max_{x \in \mathbb{R}^n} \mathbb{E}_\mathbf{p}[u_\beta(\mathbf{x})] \quad \text{subject to} \quad \mathbf{Qx} \leq 0. \tag{3.27}$$

Next for Model **P**, we can consider this as a two-period consumption problem where the decision maker chooses a vector $\mathbf{x}$ of future payoffs that she can purchase using funds during the current period. This can be written as

$$\textbf{Model P:} \quad U_\beta^P(\mathbf{p}||\mathbf{q}) \equiv \max_{y \in \mathbb{R}^n, x \in \mathbb{R}^n} \mathbb{E}_\mathbf{p}[u_\beta(\mathbf{x})] - y \quad \text{subject to} \quad \mathbf{Qx} \leq y\mathbf{1}. \tag{3.28}$$

Under this setting, the constraint implies that the market will never give the decision maker an expected payoff larger than what she expects to gain from the prevailing market rate. Later on, we will see that the optimal solution for this problem would yield $y^* = \mathbf{E}_\mathbf{q}[\mathbf{x}^*]$, which is the cost of buying $\mathbf{x}^*$ at time-0 prices.

Having set-up these problems, our next proposition tells us that the duality holds not only for exponential utility but for the entire spectrum of β for these two families of divergence measures.

Proposition 3.3.1. *(a) In an incomplete, single-period market, the dual problem for Model* ***S*** *is the minimization of the pseudospherical divergence, i.e.*

$$D_\beta^{\mathbf{S}}(\mathbf{p}\|\mathbf{Q}) \equiv \min_{\mathbf{z}\in\triangle^k} D_\beta^{\mathbf{S}}(\mathbf{p}\|\mathbf{z}^T\mathbf{Q}). \tag{3.29}$$

(b) In an incomplete, two-period market, the dual problem for Model ***P*** *is the minimization of the power divergence, i.e.*

$$D_\beta^{\mathbf{P}}(\mathbf{p}\|\mathbf{Q}) \equiv \min_{\mathbf{z}\in\triangle^k} D_\beta^{\mathbf{P}}(\mathbf{p}\|\mathbf{z}^T\mathbf{Q}). \tag{3.30}$$

Proof. Denote the solutions to Models **S** and **P** by $\mathbf{x}_\beta^{\mathbf{S}}(\mathbf{p}||\mathbf{q})$ and $\mathbf{x}_\beta^{\mathbf{P}}(\mathbf{p}||\mathbf{q})$ respectively, with the i^{th} elements $\mathbf{x}_{\beta,i}^{\mathbf{S}}(\mathbf{p}||\mathbf{q})$ and $\mathbf{x}_{\beta,i}^{\mathbf{P}}(\mathbf{p}||\mathbf{q})$. Also, let$\mathbf{z} \in \triangle^k$ denote a vector of weights, so that $\mathbf{z}^T\mathbf{Q}$ is a mixture of the rows of $\mathbf{Q}$, which is an element of the convex polytope $\mathscr{Q}$ of risk neutral distributions.

For part (a), Lagrangian relaxation is applicable because the primal problem has a strictly concave, continuously differentiable objective function and linear constraints. Let λ denote the vector of Lagrange multipliers associated with the constraints $\mathbf{Qx} \leq \mathbf{0}$. The Lagrangian relaxation of Model **S** is then $\min_{\lambda\in\mathbb{R}^{k+}} L(\lambda)$ where

$$L(\lambda) = \max_{\mathbf{x}\in\mathbb{R}^n} E_{\mathbf{p}}[u_\beta(\mathbf{x})] - \lambda^T\mathbf{Q}\mathbf{x}. \tag{3.31}$$

The Lagrangian $L(\lambda)$ is an unconstrained maximum of a continuously differentiable concave function, so it can be solved for $\mathbf{x}$ in terms of λ by setting $\nabla(E_{\mathbf{p}}[u_\beta(x)] - \lambda^T\mathbf{Q}x) = \mathbf{0}$, which yields

$$\mathbf{x} = \frac{1}{\beta}\left(\left(\frac{\mathbf{p}}{\lambda^T\mathbf{Q}}\right)^\beta - 1\right), \tag{3.32}$$

whence

$$\begin{aligned} L(\lambda) &= E_{\mathbf{p}}\left[\frac{1}{\beta-1}\left[\left(\frac{\mathbf{p}}{\lambda^T\mathbf{Q}}\right)^{\beta-1} - 1\right]\right] - \lambda^T\mathbf{Q}\left[\frac{1}{\beta}\left[\left(\frac{\mathbf{p}}{\lambda^T\mathbf{Q}}\right)^{\beta} - 1\right]\right] \\ &= \frac{1}{\beta-1}\left[E_{\mathbf{p}}\left[\left(\frac{\mathbf{p}}{\lambda^T\mathbf{Q}}\right)^{\beta-1}\right] - 1\right] - \frac{1}{\beta}\left[E_{\mathbf{p}}\left[\left(\frac{\mathbf{p}}{\lambda^T\mathbf{Q}}\right)^{\beta-1}\right] - \mathbf{1}^T(\lambda^T\mathbf{Q})\right] \end{aligned} \tag{3.33}$$

In the optimal solution λ^*, where the constraints are satisfied, the second term will be zero, which implies

$$\mathbf{1}^T(\lambda^{*T}\mathbf{Q}) = E_{\mathbf{p}}\left[\left(\frac{\mathbf{p}}{\lambda^{*T}\mathbf{Q}}\right)^{\beta-1}\right] \tag{3.34}$$

and consequently

$$L(\lambda^*) = \frac{1}{\beta - 1}\left(E_{\mathbf{p}}\left[\left(\frac{\mathbf{p}}{\lambda^{*T}\mathbf{Q}}\right)^{\beta-1}\right] - 1\right). \tag{3.35}$$

Now let $\mathbf{z}^* = \lambda^*/\mathbf{1}^T\lambda^*$ be the probability distribution that is obtained by normalization of the optimal Lagrange multipliers λ^*. Then it follows from (21) that:

$$\mathbf{z}^{*T}\mathbf{Q} = \frac{\lambda^{*T}\mathbf{Q}}{E_{\mathbf{p}}[(\mathbf{p}/\lambda^{*T}\mathbf{Q})^{\beta-1}]}. \tag{3.36}$$

The pseudospherical divergence between $\mathbf{p}$ and $\mathbf{z}^{*T}\mathbf{Q}$ can therefore be expressed in terms of λ^* as:

$$\begin{aligned}
D_\beta^{\mathbf{S}}(\mathbf{p}\|\mathbf{z}^{*T}\mathbf{Q}) &= \frac{(E_{\mathbf{p}}[(\mathbf{p}/\mathbf{z}^{*T}\mathbf{Q})^{\beta-1}])^{1/\beta} - 1}{\beta - 1} \\
&= \frac{(E_{\mathbf{p}}[(E_{\mathbf{p}}[(\mathbf{p}/\lambda^{*T}\mathbf{Q})^{\beta-1}](\mathbf{p}/\lambda^{*T}\mathbf{Q}))^{\beta-1}])^{1/\beta} - 1}{\beta - 1} \\
&= \frac{(E_{\mathbf{p}}[(\mathbf{p}/\lambda^{*T}\mathbf{Q})^{\beta-1}])^{(\beta-1)/\beta}(E_{\mathbf{p}}[(\mathbf{p}/\lambda^{*T}\mathbf{Q})^{\beta-1}])^{1/\beta} - 1}{\beta - 1} \\
&= \frac{1}{\beta - 1}\left(E_{\mathbf{p}}\left[\left(\frac{\mathbf{p}}{\lambda^{*T}\mathbf{Q}}\right)^{\beta-1}\right] - 1\right) \\
&= L(\lambda^*),
\end{aligned} \tag{3.37}$$

which is the optimal objective value of the primal problem. Furthermore $\mathbf{z}^* = \lambda^*/\mathbf{1}^T\lambda^*$ must also minimize $D_\beta^{\mathbf{S}}(\mathbf{p}\|\mathbf{z}^T\mathbf{Q})$ over all $\mathbf{z} \in \triangle^k$, because if there were some other $\mathbf{z}^{**} \in \triangle^k$ such that $D_\beta^{\mathbf{S}}(\mathbf{p}\|\mathbf{z}^{**T}\mathbf{Q}) < D_\beta^{\mathbf{S}}(\mathbf{p}\|\mathbf{z}^{*T}\mathbf{Q})$, then it would be possible to find some $\lambda^{**} \in \mathbb{R}^{k+}$ proportional to $\mathbf{z}^{**}$ such that $\mathbf{z}^{**T}\mathbf{Q} = \lambda^{**T}\mathbf{Q}/(\mathbf{E}_{\mathbf{p}}[(\mathbf{p}/(\lambda^{**T}\mathbf{Q}))^{\beta-\mathbf{1}}])$. By construction this λ^{**} would satisfy $E_{\mathbf{p}}[(\mathbf{p}/\lambda^{**T}\mathbf{Q})^{\beta-1}] - \mathbf{1}^T(\lambda^{**T}\mathbf{Q}) = 0$, implying $L(\lambda^{**}) = D_\beta^{\mathbf{S}}(\mathbf{p}\|\mathbf{z}^{**T}\mathbf{Q})$, and it would follow that $L(\lambda^{**}) < L(\lambda^*)$, contradicting the assumption that λ^* was optimal.

For part (b), the problem of finding the feasible risk neutral distribution that minimizes the power divergence of order β:

$$\min_{\mathbf{z}\in\triangle^k} D_\beta^{\mathbf{P}}(\mathbf{p}\|\mathbf{z}^T\mathbf{Q}), \tag{3.38}$$

is equivalent to the Lagrangian problem $\min_{\lambda\in\triangle^k} L(\lambda)$, where $L(\lambda) = \max_{\mathbf{x}\in\mathbb{R}^n} E_{\mathbf{p}}[u_\beta(x)] - \lambda^T\mathbf{Q}x$ is the same Lagrangian that was used in the proof of part (a) to minimize the

pseudospherical divergence, except that here λ is constrained to be in the simplex, not just the non-negative orthant ($\lambda \in \triangle^k$ rather than $\lambda \in \mathbb{R}^{k+}$), which requires a Lagrange multiplier for the constraint $\mathbf{1}^T\mathbf{q} = 1$ in addition to the m Lagrange multipliers for the constraints $\mathbf{Aq} \geq \mathbf{0}$. The latter divided by the former are equal to the optimal values of the decision variables in Primal Problem $\mathbf{P}$ multiplied by $-\beta$. The power divergence is minimized by the same risk neutral distribution $\mathbf{q}^* = \mathbf{z}^{*T}\mathbf{Q}$ that minimizes the pseudospherical divergence (for the same $\mathbf{p}$, β and $\mathbf{Q}$), because they are both monotonic functions of $E_{\mathbf{p}}[(\mathbf{p}/\mathbf{q})^{\beta-1}]$. The optimal value of λ is a unit vector selecting the largest element of $\mathbf{Qx}$. Let z denote this largest element. Then $\min_{\lambda \in \triangle^k} \max_{\mathbf{x} \in \mathbb{R}^n} E_{\mathbf{p}}[u_\beta(x)] - \lambda^T\mathbf{Q}x$ is equivalent to $\max_{\mathbf{x} \in \mathbb{R}^n} E_{\mathbf{p}}[u_\beta(\mathbf{x})] - z$ subject to $\mathbf{Qx} \leq z\mathbf{1}$. □

This result shows that the duality between maximizing expected utility and minimization of entropy holds for a wider class of functions. In particular, the reliance on exponential utility in many models can be relaxed to cover a wide class of commonly-used utility functions or family of divergence measures. A simple illustration in finance of how different entropy functions affect a decision problem can be seen in Nau et al. (2009). A geometric illustration is also provided.

Chapter 4

Incorporating Sensitivity to Distance in Probabilistic Forecast Evaluation

4.1 Motivation for Sensitivity to Distance

In certain instances, the decision situation requires probability assessments for ordered (or ranked) events. Consider a betting situation wherein probabilities are assessed for three possible outcomes in a soccer match (win, tie, loss). We say that the outcomes are ordered in the sense that a tie is considered worse than a win but better than a loss. In cases like these, scoring rules have been developed to take into consideration a notion called *sensitivity to distance*. This implies that probability assessments that are "closer" to the event that occurs based on some definition of distance are given higher scores. In the soccer example, if the team actually loses (third category) an assessment of $(0.1, 0.4, 0.5)$ would be given a better score than $(0.3, 0.2, 0.5)$ because the first one differs from the second one by having more weight on the event closer (tie) to the actual outcome.

Most scoring rules studied in the literature and used in practice do not take into account the ordering of events and this notion of being sensitive to distance. A notable exception to this is the ranked probability score (RPS) introduced by Epstein (1969) in the context of weather forecasting. It is strictly proper (thereby encouraging honest reporting) and gives higher scores for assessments that are "closer" to the event that occurs based on a particular concept of distance (thereby taking the order of events into consideration).

Similar to the standard scoring rules, there also exists a default baseline distribution for the RPS from which assessments are measured. Unlike the standard symmetric scoring rules which assign a uniform baseline distribution, scoring rules for ranked

categories provide the lowest expected score when a probability of 1/2 is provided for the two most extreme states with zero probabilities for all intermediate states. In certain instances this is not a reasonable baseline. In a soccer game wherein a top ranked team plays an amateur team, it may not seem reasonable to use a 50-50 reference distribution on winning and losing.

The focus of this chapter is threefold. First, we develop other strictly proper scoring rules that take into account the notion of sensitivity to distance. Second, we discuss how the notion of a baseline distribution could be incorporated in the context of sensitive to distance scoring rules. Finally, we address the issue of how these scores could be extended to the continuous case.

4.2 Generating Sensitive to Distance Rules

Consider a countable state space Ω with elements that have a predetermined ordering. When $|\Omega| \geq 3$, it may be of interest to study forecasts which are considered to be "more distant" than others after the uncertainty has been resolved and the actual state is known. First, we define what makes one forecast "more distant" from a particular state than another forecast.

Definition 12. *A forecast* $\mathbf{r}'$ *is said to be more distant from event* j *than another forecast* $\mathbf{r}$ *if*

$$\begin{aligned} R'_i &\geq R_i \quad \text{for } i = 1, \ldots, j-1, \\ R'_i &\leq R_i \quad \text{for } i = j, \ldots, n-1 \end{aligned} \tag{4.1}$$

where $R_i = \sum_{j \leq i} r_i$ *and* $R'_i = \sum_{j \leq i} r'_i$.

To illustrate this, consider forecasts provided for six possible states of the world, say the change in the price of a Microsoft stock falls in one of the following categories: below -10%, -10% to -5%, -5% to 0%, 0% to 5%, 5% to 10%, above 10%. If we observe that the actual change is 3% (i.e., event 4) then we might feel that forecasts that put higher probabilities on event 4 and events close to it should have higher scores.

For example, if we consider forecast $A = (0.1, 0.3, 0.2, 0.1, 0.1, 0.2)$ in Figure 4.1, then we note that forecast $B = (0.05, 0.15, 0.33, 0.17, 0.1, 0.2)$ should have better score than A since A is more distant than B with respect to event 4. Here we see that the probability assigned to events 1 and 2 have decreased from A to B, and this probability has been transferred to events 3 and 4. Similarly, forecast $C = (0.1, 0.3, 0.2, 0.15, 0.2, 0.05)$ is closer to event 4 than A since we have probability mass being moved toward the actual event from states on the right hand side of that event.

Shifts of probability masses from both sides are also allowed. Forecast $D = (0.05, 0.15, 0.33, 0.22, 0.2, 0.05)$, which results from performing the shifts done in forecast B and then in forecast C, is also less distant from event 4 when compared to forecast A. Caution though has to be exercised when changes of probability masses occur on both sides of the actual event. Notice that forecast $E = (0.05, 0.3, 0.2, 0.1, 0.15, 0.2)$ shows a shift of mass from the most distant state to an adjacent state of the actual event, but forecast E fails to satisfy the definition because the shift is to a state on the other side of the actual event.

When there are only two states of interest, the distance between forecasts becomes trivial and the use of any strictly proper scoring rule would give a higher score to forecasts with greater probability on the actual event. Coherence takes away the notion of "sensitivity to distance" since probability assessments are limited to a single degree of freedom. But when $n \geq 3$, scores may differ depending on how probabilities are assigned to the different states that did not occur. Hence, we can now define what it means for a scoring rule to respect this notion of forecasts being less distant than others to the actual event and therefore to be sensitive to distance.

Definition 13. *A scoring rule S is said to be sensitive to distance if $S_j(\mathbf{r}) > S_j(\mathbf{r}')$ whenever $\mathbf{r}'$ is more distant than $\mathbf{r}$ for $j = 1, \ldots, |\Omega|$.*

A characterization for strictly proper scoring rules that are sensitive to distance has been provided by Nau (2007).

The earliest scoring rule in the literature with this property is the ranked proba-

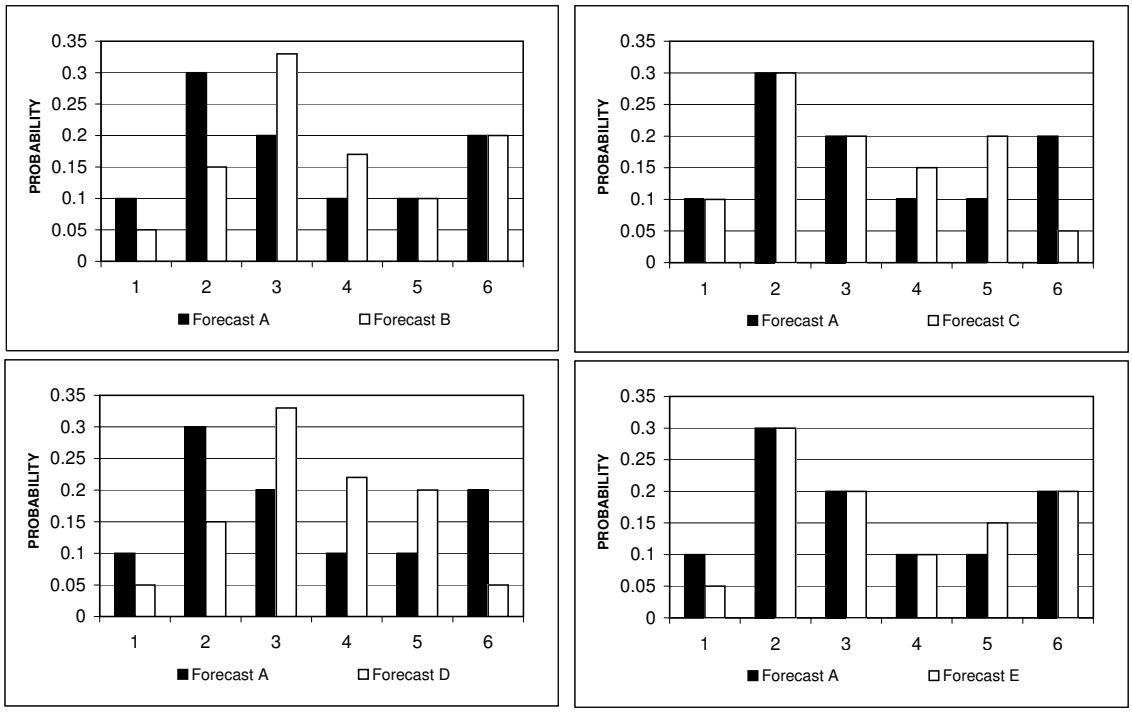

Figure 4.1: Examples of forecasts more distant from an event

bility score (RPS), which was introduced by Epstein (1969):

$$RPS_j(\mathbf{r}) = \frac{3}{2} - \frac{1}{n-1}\sum_{i=1}^{n-1}\left[R_i^2 + (1-R_i)^2\right] - \frac{1}{n-1}\sum_{i=1}^{n}|i-j|r_i. \tag{4.2}$$

Murphy (1971) points out that (4.2) can be simplified to

$$RPS_j(\mathbf{r}) = -\sum_{i=1}^{j-1} R_i^2 - \sum_{i=j}^{n-1}(1-R_i)^2. \tag{4.3}$$

The motivation for (4.3) is that for each state, we can divide the state space with $n-1$ dividers between adjoining states and score each of the $n-1$ resulting dichotomies. The score attained when these scores are added is strictly proper and now sensitive to distance.

There is nothing particularly special about the quadratic score, however. If S is any strictly proper binary scoring rule, then we can generate a new score $\tilde{S}$ defined by

$$\tilde{S}_j(\mathbf{r}) = \sum_{i=1}^{j-1} S_2(R_i, 1-R_i) + \sum_{i=j}^{n-1} S_1(R_i, 1-R_i). \tag{4.4}$$

Proposition 4.2.1. *The scoring rule $\tilde{S}$ given by (4.4) is strictly proper.*

Proof. The expected score is

$$\begin{aligned}\mathbb{E}_{\mathbf{p}}\widetilde{S}(\mathbf{r}) &= \sum_{j=1}^{n} p_j \Big[\sum_{i=1}^{j-1} S_2(R_i, 1-R_i) + \sum_{i=j}^{n-1} S_1(R_i, 1-R_i) \Big] \\ &= \sum_{i=1}^{n} P_i S_1(R_i, 1-R_i) + (1-P_i) S_2(R_i, 1-R_i). \end{aligned} \tag{4.5}$$

For each i in (4.5), the maximizing solution can be obtained from the following system of equations:

$$\sum_{k \le i} p_k = \sum_{k \le i} r_k \quad \text{and} \quad \sum_{k > i} p_k = \sum_{k > i} r_k.$$

Iteratively, we get $p_i = r_i$ for all i. Hence, $\mathbf{p} = \mathbf{r}$ is a maximizing solution for the aggregate problem. To show that it is unique, let $\mathbf{r}'$ be a maximizing solution distinct from $\mathbf{p}$. Let m be the first state for which $\mathbf{r}$ differs from $\mathbf{p}$. Then $P_m S_1(R'_m, 1-R'_m) + (1-P_m) S_2(R'_m, 1-R'_m)$ is strictly lower than $P_m S_1(P_m, 1-P_m) + (1-P_m) S_2(P_m, 1-P_m)$ since S is strictly proper, while the other terms in the summation can do no better. Hence, $\mathbf{r}'$ can not be a maximizing solution. □

To illustrate the form this new score generated by (4.4) would take, consider the logarithmic score. When event j occurs,

$$\widetilde{S}_j(\mathbf{r}) = \sum_{i=1}^{j-1} \log(1-R_i) + \sum_{i=j}^{n-1} \log R_i.$$

When $n = 4$, for example, the expected score is:

$$\mathbb{E}\widetilde{S} = \begin{bmatrix} p_1 \\ p_2 \\ p_3 \\ p_4 \end{bmatrix}^T \begin{bmatrix} \log R_1 & \log R_2 & \log R_3 \\ \log(1-R_1) & \log R_2 & \log R_3 \\ \log(1-R_1) & \log(1-R_2) & \log R_3 \\ \log(1-R_1) & \log(1-R_2) & \log(1-R_3) \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}.$$

Similar to the RPS, this scoring rule is sensitive to distance. Consider two distributions $\mathbf{r}$ and $\mathbf{r}'$ which only differ in two states. For example, let $a < b \le j$, with

$r'_a = r_a + \epsilon$ and $r'_b = r_b - \epsilon$, where $\epsilon > 0$. This implies that r' is more distant than r if state j occurs. Now, if we let $H_j(\mathbf{r}, \mathbf{r}') = \tilde{S}_j(\mathbf{r}) - \tilde{S}_j(\mathbf{r}')$, the improvement in score in moving from $\mathbf{r}'$ to $\mathbf{r}$, we then have

$$H_j(\mathbf{r}, \mathbf{r}') = \sum_{i=a}^{b-1} \log(1 - R_i) - \sum_{i=a}^{b-1} \log(1 - R_i - \epsilon) = \sum_{i=a}^{b-1} \log \frac{(1 - R_i)}{(1 - R_i - \epsilon)}.$$

Note that $(1 - R_i)/(1 - R_i - \epsilon) > 1$, so that $\log[(1 - R_i)/(1 - R_i - \epsilon)] > 0$ for each of the terms in the summation. Therefore, $H_j(\mathbf{r}, \mathbf{r}') > 0$. A similar argument can be made when $a > b \geq j$. Thus, as shown more generally in Proposition 4.2.2, the rule is sensitive to distance.

Proposition 4.2.2. *Let $\tilde{S}$ be a ranked scoring rule generated from (4.4) for a strictly proper binary scoring rule S. Then $\tilde{S}$ is sensitive to distance.*

Proof. Using the notation from the log example, we have the following form for $H_j(\mathbf{r}, \mathbf{r}')$ when $a < b \leq j$, with $r'_a = r_a + \epsilon$ and $r'_b = r_b - \epsilon$:

$$H_j(\mathbf{r}, \mathbf{r}') = \sum_{i=a}^{b-1} S_2(R_i, 1 - R_i) - S_2(R_i + \epsilon, 1 - R_i - \epsilon).$$

When $a > b \geq j$, with $r'_a = r_a + \epsilon$ and $r'_b = r_b - \epsilon$,

$$H_j(\mathbf{r}, \mathbf{r}') = \sum_{i=b}^{a-1} S_1(R_i, 1 - R_i) - S_1(R_i - \epsilon, 1 - R_i + \epsilon).$$

Because S is strictly proper, $S_1(r, 1-r)$ must be increasing in its first argument, while $S_2(r, 1 - r)$ must be increasing in the second. Hence, H_j must be positive in both cases. Finally, any distribution which is more distant with respect to an event j can be created through a step-by-step transfer of probability mass further away from the event of concern. Hence, $\tilde{S}$ is sensitive to distance. □

The construction given by (4.4) allows us to have greater flexibility in the type of scoring rule that we can use in problems with ranked categories. It provides an extensive set of strictly proper scoring rules $\tilde{S}$ that are sensitive to distance.

In addition, Propositions 4.2.1 and 4.2.2 can be extended to allow for the use of a different scoring rule for every dichotomy i, thus increasing the generality of the results. Hence, if $\{S^{(i)}\}_{i=1}^{i=n-1}$ is any set of strictly proper binary scoring rules, then

$$\tilde{S}_j(\mathbf{r}) = -\sum_{i=1}^{j-1} S_2^{(i)}(R_i, 1-R_i) - \sum_{i=j}^{n-1} S_1^{(i)}(R_i, 1-R_i). \tag{4.6}$$

will also be a strictly proper scoring rule that is sensitive to distance. These results can be proven in same way as the previous propositions, since the fact that identical scoring rules were used for each i was never used.

A possible application of (4.6) is when $S^{(i)} = a_i S + b_i$. Under this scheme, we could choose appropriate parameters $\{a_i, b_i\}_{i=1}^{n-1}$ such that greater (or less) weight could be placed on more extreme events. We note though that the choice of same S for every i provides sufficient flexibility in most cases.

The rules generated by (4.4) take into consideration the ordering of events and are sensitive to distance, but their sensitivity to distance is highly dependent on the way by which we define the notion of "more distant." For example, Murphy (1970) suggested a notion of distance based on symmetric sums, i.e. we say that $\mathbf{r}'$ is more distant than another forecast $\mathbf{r}$ from event j if

$$C_i' := \sum_{k=j-i}^{j+i} r_i' \leq \sum_{k=j-i}^{j+i} r_i =: C_i \tag{4.7}$$

for all i. Sensitivity to this notion of difference would imply that if (4.7) is satisfied for every j then $S_j(\mathbf{r}') < S_j(\mathbf{r})$. No scoring rule has been developed for this notion of distance.

4.3 Incorporating a Baseline Distribution

In the previous chapter, we discussed the notion of incorporating a non-uniform baseline distribution and how it could be of use in many applications. In the case of a standard (non-ranked) symmetric strictly proper scoring rule such as the quadratic,

Proposition 3.1.1 tells us that a uniform distribution provides the lowest possible expected score. In contrast, the RPS attains the lowest expected score when probabilities of one-half are assigned to the two extreme events in the ordered space (with zero on all other events). An interpretation for this is that these distributions can be viewed as the "least informative" beliefs that one could have in the two situations, where the difference between the two is that with the RPS, we care about sensitivity to distance. Proposition 4.3.1 shows that the baseline distribution for the RPS holds for a wide class of strictly proper sensitive-to-distance scoring rules.

Proposition 4.3.1. *For a scoring rule $\tilde{S}$ generated from (4.4) using a symmetric, strictly proper scoring rule S, the expected score under honest reporting is minimized when $\mathbf{r} = \mathbf{p} = (0.5, 0, \ldots, 0, 0.5)$.*

Proof. From (4.5), the expected score $\mathbb{E}_{\mathbf{p}} S(\mathbf{r})$ is the sum of expected scores using S for $n-1$ dichotomies. Because S is symmetric, the expected score when $R_i = P_i$ is minimized at $P_i = 0.50$. Setting $\mathbf{p} = (0.5, 0, \ldots, 0, 0.5)$ yields $P_i = 0.5$ for $i = 1, \ldots, n-1$; so each of the $n-1$ expected scores is minimized individually. Therefore, the overall expected score is minimized. □

When there is a reference distribution, or baseline distribution, that represents the notion of "least informative" in a given situation, it is useful to be able to scale a scoring rule such that it attains a minimum expected score for honest reporting at this reference distribution. One way to address this issue is by making strictly proper rules "asymmetric" by relating them to a baseline distribution while retaining the property of strict propriety, as done in the previous chapter for unordered spaces.

Proposition 4.3.2. *For a scoring rule $\tilde{S}$ generated from (4.4) using an asymmetric strictly proper binary scoring rule S with baseline $(b, 1-b)$, $b \notin \{0, 1\}$, the expected score under honest reporting is minimized when $\mathbf{r} = \mathbf{p} = (b, 0, \ldots, 0, 1-b)$.*

Proof. When $\mathbf{r} = \mathbf{p} = (b, 0, \ldots, 0, 1-b)$, $P_i = b$ for $i = 1, \ldots, n-1$, which minimizes the expected score in the same manner as the proof in Proposition 4.3.1. □

This is slightly less restrictive in the sense that the extreme states need not have a value of 0.5. Unfortunately in many applications, having a baseline which is always 0 in the middle states is quite unrealistic. However, using (4.6), we can create similar scores for ordered state spaces which allow a wider set of possible baseline distributions. In particular, by using the families of power and pseudospherical scores, we can create new scoring rules using (4.6) that are strictly proper, are sensitive to distance, and have a baseline distribution that is not necessarily uniform.

Let $\mathbf{q} = (q_1, q_2, \ldots, q_n)$ be a baseline distribution that is strictly positive (i.e., $q_i > 0$ for $i = 1, ..., n$). Note that if there exists a state for which the baseline is 0, we can easily combine this state with an adjacent state. If the baseline is 0, then every assessment for this state must be 0 as well; otherwise the spirit of a baseline distribution is violated.

We can write the ranked power and pseudospherical scores with baseline $\mathbf{q}$ as follows, where $S^{P(\beta)}$ and $S^{S(\beta)}$ are the weighted power and pseudospherical scores in (3.9) and (3.11) for a dichotomy, $\mathbf{R}_j = (R_j, 1 - R_j)$, and $\mathbf{Q}_j = (Q_j, 1 - Q_j)$:

$$\tilde{S}^{P(\beta)}(\mathbf{r}||\mathbf{q}) = \sum_{i=1}^{j-1} S_2^{P(\beta)}(\mathbf{R}_i||\mathbf{Q}_i) + \sum_{i=j}^{n-1} S_1^{P(\beta)}(\mathbf{R}_i||\mathbf{Q}_i) \tag{4.8}$$

and

$$\tilde{S}^{S(\beta)}(\mathbf{r}||\mathbf{q}) = \sum_{i=1}^{j-1} S_2^{S(\beta)}(\mathbf{R}_i||\mathbf{Q}_i) + \sum_{i=j}^{n-1} S_1^{S(\beta)}(\mathbf{R}_i||\mathbf{Q}_i). \tag{4.9}$$

Figure 4.2 shows plots for some special cases of (4.8) and (4.9) overlayed with the cases which do not have a baseline distribution nor are sensitive to distance. We consider the quadratic, spherical, and logarithmic scores as a function of r_2 for a three-event example with $\mathbf{r} = (0.4, r_2, 0.6 - r_2)$ and $j = 1$. In each case, four scores are considered: the standard score from (2.14) or (2.13), the ranked score (4.4) using the generators (2.14) or (**??**), the standard score from (3.9) or (3.11) with baseline $\mathbf{q} = (0.3, 0.6, 0.1)$, and the ranked score from (4.8) or (4.9) with the same baseline. These plots illustrate how scores can differ when a ranked score is used as opposed to a standard (non-ranked) score and when a baseline different from the default baseline is

considered, and they also demonstrate that the scores generated by different values of β have different properties. For example, the standard quadratic and spherical scores are maximized when $r_2 = 0.3$, the value for which $r_2 = r_3$. The corresponding scores with the baseline are maximized near $r_2 = 0.5$. The logarithmic scores with and without the baseline are constant in r_2 since they depend on $\mathbf{r}$ only through r_1, although they differ from each other because the score with the baseline also depends on q_1. All of the ranked scores, with or without the baseline, are increasing in r_2; with $j = 1$ and r_1 fixed, the ranked scores are higher as more probability is shifted to event 2, the event closest to event 1.

Note that the shapes of the curves in Figure 4.2 and the relationships among these curves differ not only depending on whether the score is ranked or whether a non-default baseline is used, but also for the different families of scoring rules (different values of β). Of course, this is a single example, and the curves will differ as the details change. It is clear, however, that considering sensitivity to distance and making comparisons with a non-default baseline can significantly change the nature of the scores.

Proposition 4.3.3. *The scoring rule $\tilde{S}$ generated in (4.8) or (4.9) by the weighted power or pseudospherical scoring rule attains a minimum possible expected score for honest reporting if and only if $\mathbf{p} = \mathbf{q}$.*

Proof. Given a fixed $\mathbf{q}$, the expected scores for honest reporting can be written as

$$\begin{aligned}
\mathbb{E}\tilde{S}^{P(\beta)}(\mathbf{p}||\mathbf{q}) &= \sum_{j=1}^{n} p_j \Big[\sum_{i=1}^{j-1} S_2^{P(\beta)}(\mathbf{P}_i||\mathbf{Q}_i) + \sum_{i=j}^{n-1} S_1^{P(\beta)}(\mathbf{P}_i||\mathbf{Q}_i) \Big] \\
&= \sum_{j=1}^{n} \frac{\mathbb{E}_{\mathbf{P}_j}[(\mathbf{P}_j/\mathbf{Q}_j)^{\beta}] - 1}{\beta(\beta-1)}
\end{aligned}$$

and

$$\begin{aligned}
\mathbb{E}\tilde{S}^{S(\beta)}(\mathbf{p}||\mathbf{q}) &= \sum_{j=1}^{n} p_j \Big[\sum_{i=1}^{j-1} S_2^{S(\beta)}(\mathbf{P}_i||\mathbf{Q}_i) + \sum_{i=j}^{n-1} S_1^{S(\beta)}(\mathbf{P}_i||\mathbf{Q}_i) \Big] \\
&= \sum_{j=1}^{n} \frac{\mathbb{E}_{\mathbf{P}_j}[(\mathbf{P}_j/\mathbf{Q}_j)^{\beta}]^{1/\beta} - 1}{\beta - 1},
\end{aligned}$$

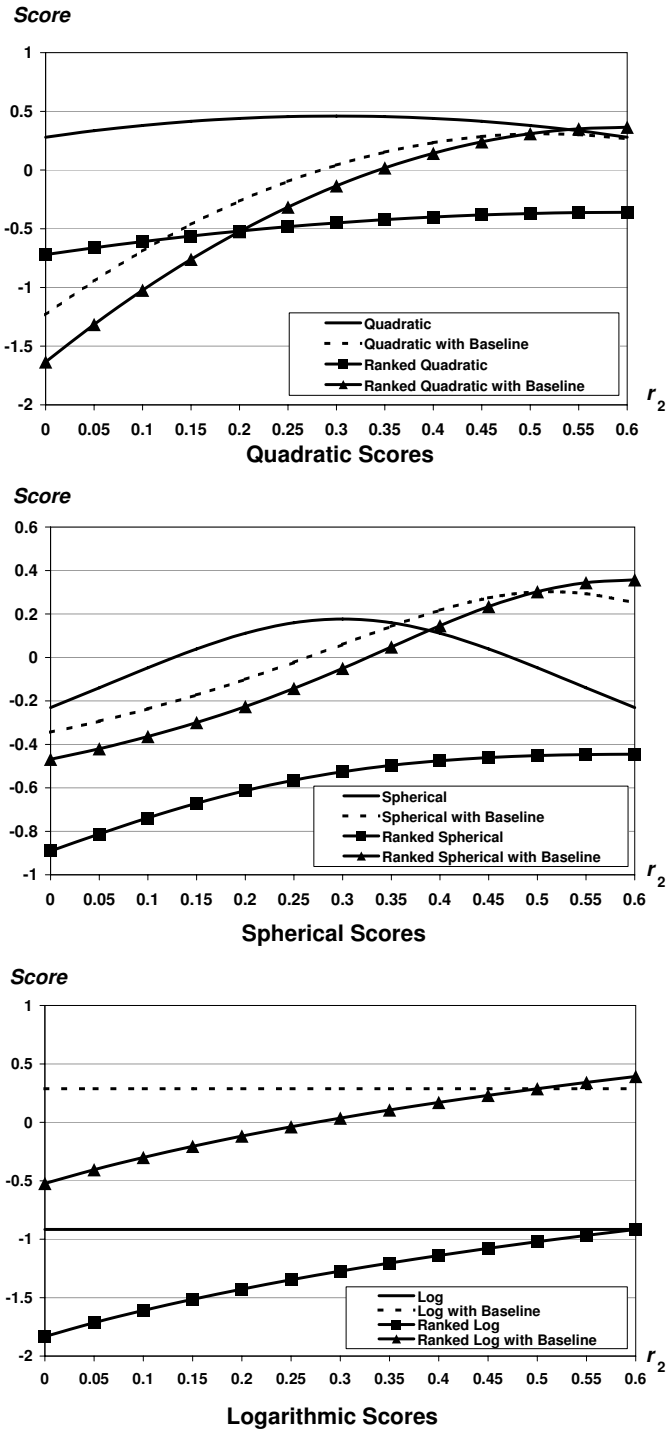

Figure 4.2: Scores as a function of r_2 for different scoring rules when $\mathbf{r} = (0.4, r_2, 0.6 - r_2)$ and $j = 1$

where $\mathbf{P}_j = (P_j, 1 - P_j)$ and $\mathbf{Q}_j = (Q_j, 1 - Q_j)$. For every j, the terms in these summations are non-negative since the power and pseudospherical divergences are non-negative. Hence, $\mathbb{E}\tilde{S} \geq 0$. Also, when $\mathbf{p} = \mathbf{q}$, these terms are 0, so they attains a minimum when $\mathbf{p} = \mathbf{q}$. The uniqueness follows from the fact that the divergences attain a minimum only when $\mathbf{P}_i = \mathbf{Q}_i$ for every i. □

A good way to illustrate this is by looking at the shape of the contours of these new ranked scores in a simplex. Figure 4.3 shows the different expected score contours when $n = 3$ for the quadratic scoring rule, the RPS, and the power scoring rule with $\beta = 2$ and baselines $(1/3, 1/3, 1/3)$ and $(7/10, 2/10, 1/10)$. The quadratic score attains a minimum expected score at the point $(1/3, 1/3, 1/3)$, similar to the ranked power scoring rule centered at the same point. Unlike the quadratic score, however, the latter score is only symmetric along the middle state since this score is sensitive to distance. For the RPS, Figure 4.3 shows that a minimum is attained at $(1/2, 0, 1/2)$. The two plots at the bottom of Figure 4.3 indicate how the expected score changes when a baseline other than $(1/2, 0, 1/2)$ is selected. The minimum expected score for the ranked power scoring rule is attained at the baseline distribution, as shown in Proposition 4.3.3. Though the baseline is restricted to be strictly positive, we can approximate the RPS by choosing a baseline $(1/2 - \epsilon/2, \epsilon, 1/2 - \epsilon/2)$ and making $\epsilon > 0$ arbitrarily small. Though the ranked power and pseudospherical families provide sufficient richness for all practical purpose, one may ask whether there exists a much larger set of ranked scoring rules which can be assigned a pre-specified baseline. The response to this is yes. Let f be any function on the unit square such that: $(1)|f(x; y)| < \infty$ for all $(x; y) \in [0, 1]^2$, $(2) f(x; y^*)$ is strictly convex in x for any fixed $y^* \in (0, 1)$ and attains a minimum at $x = y^*$; and (3) given any fixed $y^* \in (0, 1)$ and for every $x \in [0, 1]$, f has a subgradient f' at x. Using the Schervish representation given in (2.20) and (2.21), we can generate a strictly proper binary scoring rule for a

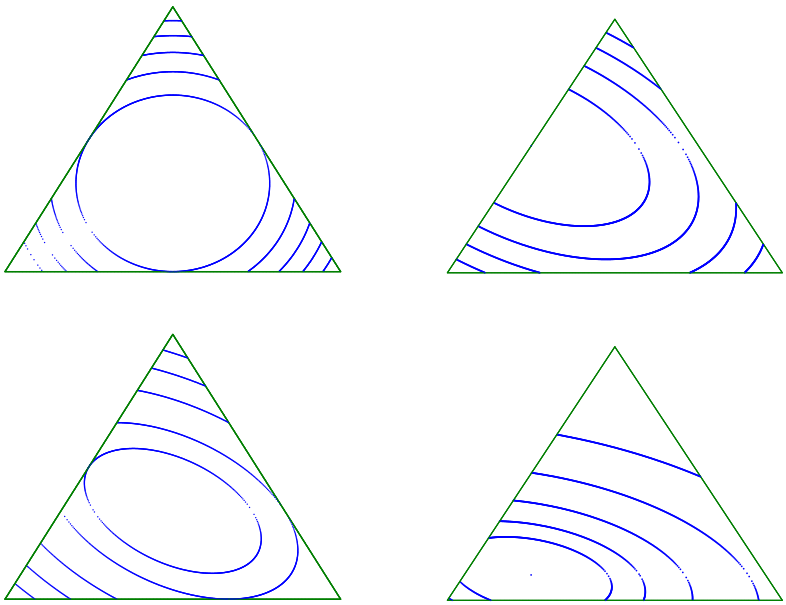

Figure 4.3: Expected score contours for the regular quadratic score [top left], the RPS [top right] and the ranked power score with $\beta = 2$ and baselines (1/3,1/3,1/3) [bottom left] and (7/10,2/10,1/10) [bottom right]

fixed baseline $(q, 1-q)$ as follows:

$$S_1^f((r,1-r)||(q,1-q)) = f(r,q) + (1-p)f'(r,q) \tag{4.10}$$

$$S_2^f((r,1-r)||(q,1-q)) = f(r,q) - pf'(r,q). \tag{4.11}$$

Proposition 4.3.4. *The scoring rule*

$$\tilde{S}^f(\mathbf{r}, \mathbf{e}_j||\mathbf{q}) = \sum_{i=1}^{j-1} S_2^f(\mathbf{R}_i||\mathbf{Q}_i) + \sum_{i=j}^{n-1} S_1^f(\mathbf{R}_i||\mathbf{Q}_i) \tag{4.12}$$

generated using the binary scoring rule S^f from (4.10) and (4.11) is strictly proper, sensitive to distance and attains a minimum possible expected score for honest reporting if and only if $\mathbf{p} = \mathbf{q}$.

Proof. The proof for strict propriety and sensitivity to distance is analogous to the proofs for Propositions (4.2.1) and (4.2.2). To show the last part, we write the expected scores for honest reporting given a fixed $\mathbf{q}$ as follows, where $\mathbf{P}_j = (P_j, 1-P_j)$ and $\mathbf{Q}_j = (Q_j, 1-Q_j)$:

$$\begin{aligned}
\mathbb{E}_{\mathbf{p}}\tilde{S}^f(\mathbf{p}||\mathbf{q}) &= \sum_{j=1}^{n} p_j \Big[\sum_{i=1}^{j-1} S_2^f(\mathbf{P}_i||\mathbf{Q}_i) + \sum_{i=j}^{n-1} S_1^f(\mathbf{P}_i||\mathbf{Q}_i) \Big] \\
&= \sum_{j=1}^{n} \Big[P_j S_1^f(\mathbf{P}_j||\mathbf{Q}_j) + (1-P_j) S_2^f(\mathbf{P}_j||\mathbf{Q}_j) \Big] \\
&= \sum_{j=1}^{n-1} f(P_j, Q_j).
\end{aligned}$$

Since f is minimized when $\mathbf{P}_j = Q_j$, $\mathbb{E}_{\mathbf{p}}\tilde{S}^f(\mathbf{p}||\mathbf{q})$ attains a minimum when $\mathbf{p} = \mathbf{q}$. Uniqueness follows from the strict convexity of f. □

Under Proposition 4.3.4, the ranked power and pseudospherical scores become special cases. For example, if f is the power divergence given by (3.12) then the scoring rule given by (4.12) is the ranked power rule. Similarly, if the pseudospherical divergence given by (3.13) is used then the score becomes the ranked pseudospherical scoring rule in (4.9). Though (4.12) provides a plethora of possibilities, for most practical purposes the ranked power and pseudospherical scoring rules offer sufficient richness. In

addition, the scores used to generate these rules have several interesting properties as mentioned in the previous chapter.

4.4 Continuous Analogues

Suppose that the state space $\Omega \subseteq \mathbb{R}$ is no longer countable, and the appropriate measure μ is Lebesgue. Moreover, suppose forecasts now are taken from $\mathscr{P}$, a non-empty set of Borel probability measure on Ω. Continuous versions of the previous scores can now be created. For the ranked probability score, Matheson and Winkler (1976) suggested a continuous analogue, which they referred to as the cumulative ranked probability score (CRPS). Suppose that an assessor provides a forecast $F \in \mathscr{P}$ and x is the outcome that is observed then this new score is given by

$$\begin{aligned} CRPS(F,x) &= -\int_{-\infty}^{\infty} (F(\omega) - \mathbf{1}\{\omega \geq x\})^2 d\omega \\ &= -\int_{-\infty}^{x} F(\omega)^2 dy - \int_{x}^{\infty} (1 - F(\omega))^2 dy \end{aligned} \tag{4.13}$$

where $\mathbf{1}$ is the indicator function. The intuition of (4.14) is the same as the old score where the space is divided into two regions - the ones which are ordered below the event observed and those above it. In this case, the CRPS hinges on the use of the quadratic score, as does the RPS, but there is nothing particularly special about it, leading us to the following generalization.

Let S be a strictly proper binary scoring rule. For simplicity, we consider only those functions S which are continuously differentiable and integrable with respect to $\mathscr{P}$. These may limit the set of possible functions but this is wide enough to allow for great flexibility in most applications. A new continuous ranked scoring (CRS) rule for the continuous case can then be generated by

$$CRS(F,x) = \int_{-\infty}^{x} S_2(F(\omega), 1 - F(\omega)) d\omega + \int_{x}^{\infty} S_1(F(\omega), 1 - F(\omega)) d\omega. \tag{4.14}$$

Similar to the CRPS, the CRS can be viewed, as Hersbach (2000) puts it, as a ranked score similar to (4.4) with "an infinite number of classes, each of zero width." Moreover

we can see that this extension of the discrete case is also strictly proper. To show this, suppose R is the reported distribution and P is the measure by which we compute the expected score. Then, using Fubini's Theorem, we can write the expected score as

$$\begin{aligned}
\mathbb{E}[CRS] &= \int_{-\infty}^{\infty} \Big[\int_{-\infty}^{x} S_2(R(\omega), 1-R(\omega)) d\omega + \int_{x}^{\infty} S_1(R(\omega), 1-R(\omega)) d\omega \Big] dP(\omega) \\
&= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \Big(S_2(R(\omega), 1-R(\omega)) \mathbf{1}\{\omega \leq x\} \\
&\qquad + S_1(R(\omega), 1-R(\omega)) \mathbf{1}\{\omega \geq x\} \Big) d\omega dP(x) \\
&= \int_{-\infty}^{\infty} P(z) S_2(R(z), 1-R(z)) + (1-P(z)) S_1(R(z), 1-R(z)) dz \\
&= \int_{-\infty}^{\infty} \mathbb{E}_P[S(R(z), 1-R(z))] dz.
\end{aligned}$$

Since S is strictly proper, the expectation is maximized when $P(u) = R(u)$ for every u, thereby encouraging truth-telling.

As an illustration, consider the power scoring rule with parameter β. The continuous ranked power score $CRS^{P(\beta)}S$ for a report R under a baseline Q is given by

$$\begin{aligned}
CRS^{P(\beta)}S &= \frac{1}{\beta(\beta-1)} \Big\{ \int_{-\infty}^{x} \Big[\frac{\beta(1-R(\omega))}{1-Q(\omega)} - (\beta-1) \int_{-\infty}^{\infty} \frac{(1-R(x))^{\beta}}{(1-Q(x))^{\beta-1}} dx - 1 \Big] d\omega \\
&\qquad + \int_{x}^{\infty} \Big[\frac{\beta R(\omega)}{Q(\omega)} - (\beta-1) \int_{-\infty}^{\infty} \frac{R(x)^{\beta}}{Q(x)^{\beta-1}} dx - 1 \Big] d\omega \Big\},
\end{aligned}$$

and its expected score is given by:

$$\mathbb{E}[CRS^{P(\beta)}S] = \int_{-\infty}^{\infty} \mathbb{E}_P[S^{P(\beta)}(R(y), 1-R(y) || Q(y), 1-Q(y))] dy.$$

This implies that $\mathbb{E}[CRS^{P(\beta)}S]$ is maximized only when $P(u) = R(u)$ for all u and moreover, among the set of all possible $P \in \mathscr{P}$, the expected score $\mathbb{E}[CRS^{P(\beta)}S]$ attains a lowest possible score when $P = Q$. Hence, the use of the power (or pseudospherical) scoring rule leads to an asymmetric version for this new score, which attains a minimum expected score for truthful reporting only when the assessor's belief coincides with the baseline.

Similarly, the new rules that we had generated also satisfies some notion of sensitivity to distance. The following provides the extension of the definition of closeness and distance to the continuous setting.

Definition 14. *A distribution F is said to be more distant from an outcome x than another forecast G (with $F \neq G$) if*

$$\begin{array}{ll} F(y) \geq G(y) & \text{if } y < x \\ F(y) \leq G(y) & \text{if } y \geq x \end{array}. \tag{4.15}$$

Under this definition, the scoring rule defined by (4.14) is sensitive to distance: If F is more distant from an outcome x than another forecast G then $CRS(F,x) \leq CRS(G,x)$ for all x. The argument for this result hinges on the discretization of the state space into small intervals and using the proof of Proposition 4.2.2. Once, x is fixed and a partition is selected, we can consider two distributions which only differ in area in two intervals on the same side of x. Since S is strictly proper then the difference H should be strictly positive, implying an improvement. The sensitivity to distance is then guaranteed by the fact that the result holds for every x and for every partition of Ω, which implies that under a continuity condition, $CRS(F,x) \leq CRS(G,x)$ for all x.

Under this definition, the distributions may coincide as many times as they want but once a distribution crosses the other, it can coincide with the other distribution but never cross it again. This is called the *single upcrossing condition*.

An issue that has been raised with respect to the CRPS which may be applicable to these new scores is the lack of analytic closed-form solutions. (Gneiting and Raftery 2007) For example, the CRPS for a simple normal probability assessment with mean μ and variance σ^2 is given by

$$CRPS_{N(\mu,\sigma^2)}(x) = \frac{\sigma}{\sqrt{\pi}} - (x-\mu)(2\Phi(z)-1) - 2\sigma\phi(z),$$

where $z = (x-\mu)/\sigma$ is the standardized value of x, ϕ the density function of the normal and Φ its distribution function. For more complicated functions S and many distribution functions F, the forms may lead to some difficulties in evaluating the score. In recent years, this problem has become less important with the growth of computing power. Traditionally, numerical analysis tools such as quadrature techniques (Staël von Holstein 1977, Unger 1985) have been suggested as a solution. Now, Monte Carlo techniques could be applied easily to provide a quick and accurate estimate for these

complicated integrals.

In the meteorological community, the CRPS has gained some popularity in practice (e.g. Hersbach 2000; Candille and Talagrand 2005; Gneiting et al. 2005). In particular, they cite that an advantage of this score is that it retains both strict propriety and the ability of the RPS to take order into consideration without having to specify a set of predefined classes, whose choice could greatly affect the results.

Chapter 5

Scoring Quantile Assessment

In this chapter, we investigate another common method in assessing probabilistic information. In many instances, an assessor may simply be asked to provide specific values for a distribution instead of an entire distribution. For example, a common practice in decision and risk analysis is to ask an assessor to provide assessments for the 5-50-95 or 10-50-90 percentiles (or equivalently, their 0.05-0.5-0.95 and 0.1-0.5-0.9 quantile assessments) rather than the entire predictive distribution. In this case, we would like to use scoring rules to better assess the quality of quantile assessments that are provided by an expert.

5.1 The Quantile Setting

5.1.1 Why New Tools are Needed

Again consider the state space $\Omega \subset \bar{\mathbb{R}}$ associated with the random variable $\mathbf{X}$ of interest. Moreover, let $\mathscr{P}$ be a convex class of Borel probability measures defined on Ω and have finite moments of all order. Without loss of generality, we also assume that all elements of $\mathscr{P}$ are strictly increasing in Ω. The results can be easily extended to the non-decreasing case with some slight change of phrasing, but we focus on the strictly increasing case for clarity of exposition.

In principle, quantile assessment can be viewed as the inverse of probability forecasting. Suppose an assessor believes that X can be represented by some $F \in \mathscr{P}$. In the traditional probability assessment setting, the expert can report a cumulative probability p corresponding to a fixed value ξ of $\mathbf{X}$. But in the quantile assessment protocol, the expert provides a quantile ξ corresponding to a fixed cumulative probability p. At the very start, one may be tempted to simply use the machinery that

we have for probability assessment because of this connection. Unfortunately, our old tools are inadequate in this set-up.

Proposition 5.1.1. *A scoring rule that is strictly proper for probability assessment does not encourage truth-telling in the quantile assessment setting for any* $p \in (0,1)$.

Proof. The expected score for this assessor can be expressed as a weighted average of the two possible scores, i.e. $\mathbb{E}_F S = F(\xi)S_1(p,1-p) + (1-F(\xi))S_2(p,1-p)$. Since p is fixed then so are $S_1(p,1-p)$ and $S_2(p,1-p)$. Therefore, the expert can maximize her score by providing as much weight as she can on the larger of the two scores. If $S_1(p,1-p) < S_2(p,1-p)$, any ξ such that $F(\xi) = 0$ is optimal (i.e. any ξ equal or below the infimum of the support of $\mathbf{X}$). Similarly, any ξ such that $F(\xi) = 1$ is optimal when $S_1(p,1-p) > S_2(p,1-p)$. Thus, truth-telling is not encouraged. □

As we transition from fixing ξ to fixing p, then the old incentive for truthful reporting no longer exists. In this new set-up, a hedging strategy of providing extreme forecasts can be obtained since the score does not provide any incentive for how close you are to the event that occurs. It simply provides a score depending on whether your assessment is above or below the event that is observed.

Practitioners have focused on a certain measure of "goodness" called calibration. In quantile assessments with a fixed probability p, good calibration implies that the proportion of random variables that are observed to have values below the assessed quantiles should be as close as possible to p. For example, if an assessor is continually asked for his 0.25 quantile assessment for various items, good calibration implies that the observed value should be below her reported quantile approximately 25% of the time. If she is asked for several quantiles then each one should match appropriately.

Given a series of quantile assessments (say $p_1 < p_2 < \cdots < p_n$) for N different uncertainties, some measures that have been proposed for use are the following: (1) Chi-Squared Measure or Mean-Squared Error: [Cooke 1991]

$$\chi^2 = \sum_{i=1}^{n} \frac{(r_j - Np_j)^2}{Np_j} \tag{5.1}$$

and (2) Kolmogorov-Smirnov Statistic: [Wiper et. al. 1994]

$$KS = \max\{|r_j/N - p_j| : j = 1, 2, \ldots, m\} \tag{5.2}$$

where r_j is the number of realizations that fall within the quantiles provided for p_{j-1} and p_j. Wiper, et. al. (1994) provide the following justification for these scores:

> *Both the KS and the χ^2-scores can be used to evaluate experts' calibration. Both scores, when multiplied by a suitable information score, act as (average) strictly proper scoring rules. (Wiper, et. al. 1994, p. 236)*

If only calibration is of interest the expert can game the system and lose the flavor of a strictly proper scoring rule as the following proposition shows.

Proposition 5.1.2. *Let $\mathbf{p} = (p_1, \ldots, p_n)$ with $p_i \in (0, 1)$ for $i = 1, 2, \ldots, n$. For every $\varepsilon > 0$, given a sufficient number of trials a forecaster providing quantile assessments (associated with $\mathbf{p}$) without any knowledge of the unknown quantity can always perform well in terms of calibration (i.e., she can always be within ε units of the perfect score for the calibration measures (5.1) and (5.2).*

Proof. We begin by analyzing the case where p_i is a rational number. In this case, $\exists$ integers s_i, t_i such that $p_i = s_i/t_i$.

Now, fix an $\varepsilon > 0$ and consider the algorithm which provides an assessment of $+\infty$ for s_i consecutive times and then an assessment of $-\infty$ for $t_i - s_i$ consecutive times. Then, the function $|r_i/n - p_i|$ will attain a perfect calibration score for all multiples of t_i. Moreover, for $n \in \{(w-1)\cdot t_i + 1, \ldots, w \cdot t_i\}$, the maximum occurs at $(w-1)\cdot t_i + s_i$ for any $w \in \{n > 1 | n \in \mathbb{Z}\}$. Hence, there exists a w_i such that

$$\left|\frac{(w_i - 1)t_i + s_i}{w_i t_i}\right| \leq \varepsilon,$$

for all $n > N_i := w_i t_i$. To show the theorem holds, simply choose $N := \max\{N_1, \ldots, N_n\}$.

Next we note that since ε was chosen arbitrarily, then $KS \to 0$. But it is a well known fact that $KS \to 0$ is equivalent to $\chi^2 \to 0$. This implies that χ^2 would also approach the perfect score using this algorithm. Hence, there exists a sufficiently large

n such that the other calibration score would always be within ε units of the perfect score.

For the case of irrational numbers, simply choose a $\hat{p}_i \in (0,1)$ that is $\frac{\varepsilon}{4}$ units from p_i and use the previous result to find N such that the proportion will always be in an $\frac{\varepsilon}{2}$-neighborhood around $\hat{p}_i$, which implies that the score is within ϵ units of perfect calibration for p_i. □

In general, the proposition applies to many other goodness-of-fit statistic such as a Bayes Factor or a Cressie-Read power statistic.

The main implication of Proposition 5.1.2 is that someone who knows very little about $\mathbf{X}$ can achieve a good calibration score in a sequence of quantile assessments. The proposition suggests a strategy that will work regardless of being able to observe the events that occur. However, if the expert is allowed to monitor the results, she can always adjust assessments sequentially to correct for miscalibration and obtain a better score more quickly. For example, if $p = 0.25$ and the expert realizes that only 20% of the observed values have been below her 0.25 quantiles, she can give extremely high values for some assessed 0.25 quantiles to increase that percentage and bring it closer to 25%. Calibration statistics provide no penalty for doing so, just as they provide no penalty for the extreme reporting shown in Proposition 5.1.1. For probability assessment, in contrast, attempts to improve calibration in this manner result in reductions in expected score, although an assessor simply focusing on calibration can "game the system" in a similar fashion as shown by Foster and Vohra (1998).

These results show that new measures are needed under this setting to correctly incentivize probability assessors to provide truthful reports. Scoring rules developed for probability assessments do not provide appropriate incentives.

5.1.2 Scoring Rules for Quantile Assessments

We begin by defining a scoring rule for the quantile setting.

Definition 15. *A scoring rule for a p-quantile is an extended real value function*

$S_p : \Omega \times \Omega \to \bar{\mathbb{R}}$ *such that* $S_p(\Omega, \cdot)$ *is measurable with respect to* $\mathscr{F}$ *and quasi-integrable with respect to* $\mathscr{P}$.

Similar to the conventional scoring rule, this function provides a score in the ex post sense when an observation $x = X(\omega)$ materializes. In this instance, we say that the scoring rule provides a score $S(\xi, x)$ when x is observed and ξ is reported. We drop the subscript p when there is no confusion about which value of p is being used. In an ex ante sense, we can use these functions to compute the expected score of an assessor with belief $F \in \mathscr{P}$ as:

$$\mathbb{E}_F S(\xi) = \int_\Omega S(\xi, x) dF(x). \tag{5.3}$$

Now, define the set $\Pi(\xi) \subset \mathscr{P}$ to be the set of all probability measures $F \in \mathscr{P}$ which have the property $F(\xi) = p$. Using this we can define what it means to be strictly proper in the quantile setting.

Definition 16. *A scoring rule for quantiles* S *is strictly proper relative to* $\mathscr{P}$ *if*

$$\mathbb{E}_P[S(\xi)] \geq \mathbb{E}_P[S(\xi')] \tag{5.4}$$

for all $P \in \Pi(\xi)$.

The first explicit mention of a strictly proper scoring rule was only in Cervera and Muñoz (1997). They suggested the following rule:

$$S_p(\xi, x) = p\xi + (x - \xi)\mathbf{1}\{x \leq \xi\} + g(x), \tag{5.5}$$

where g is any arbitrary function solely dependent on x. Note that when $g(x) = -px + h(x)$, then the scoring rule can be written as

$$\begin{aligned} S(\xi, x) &= p\xi + (x - \xi)\mathbf{1}\{x \leq \xi\} - px + g(x) && (5.6) \\ &= -p(x - \xi) + (x - \xi)\mathbf{1}\{x - \xi \leq 0\} + h(x) \\ &= \begin{cases} -p(x - \xi) + h(x) & \text{if } x - \xi > 0 \\ (1 - p)(x - \xi) + h(x) & \text{if } x - \xi \leq 0 \end{cases} \\ &= -p(x - \xi)^+ - (1 - p)(\xi - x)^+ + h(x) && (5.7) \end{aligned}$$

where $z^+ := \max\{z, 0\}$. This decomposition generally shows the behavior of the score in the two major types of event that could occur - namely, the observed value being either above or below the reported assessment. The maximization of this score

$$\max_{\xi} \mathbb{E}[-p(x-\xi)^+ - (1-p)(\xi - x)^+ + h(x)]$$

is equivalent to the solution of the so-called *newsvendor problem* from the operations management literature:

$$\min_{\xi} p\mathbb{E}[(x-\xi)^+] + (1-p)\mathbb{E}[(\xi - x)^+] - \mathbb{E}[h(x)]$$

where p and $1-p$ are the underestimation and overestimation costs, respectively. The optimal solution to the newsvendor problem is given by $\xi^* = F^{-1}(p)$, which is the Bayes act under this piecewise linear loss function.

Remark 9. *The scoring rule given by (5.7) is strictly proper in the quantile setting. [Cervera & Munoz 1997]*

The solution to the newsvendor problem shows that the value that maximizes $\mathbb{E}_F S(q)$ is $F^{-1}(p)$, verifying that the score is strictly proper. We note that an equivalent version of (5.7) called the check (or tick) function has been used in the econometric literature to verify the quality of quantile regressions (e.g. Koenker & Bassett 1978, Giacomini & Komunjer 2005).

Interestingly, a linear scoring rule is strictly proper for quantile assessment, whereas a quadratic scoring rule (involving squared deviations between q and x) is not proper. In a probability-assessment setting, on the other hand, a quadratic scoring rule is strictly proper, but a linear scoring rule is not, because it leads to extreme assessed probabilities of zero or one. This relates to the point made earlier that evaluating assessments in terms of probabilities is different than evaluating them in terms of quantiles, and it demonstrates the importance of finding evaluation measures that are suitable for the particular assessment process that is used.

5.1.3 Properties

We will now consider some properties of the strictly proper scoring rule $S(\xi, x)$ for quantile assessments.

Proposition 5.1.3. *The scoring rule $S(\xi, x)$ given by (5.6) satisfies the following properties:*

i. Positive affine transformations of S as functions of x are still proper, i.e. $aS(\xi, x)+b$ is strictly proper whenever $a \in \mathbb{R}^+$. Moreover, b can be a constant or any function of x.

ii. If h is bounded then S is bounded above. Moreover, if $\mathbf{X}$ has bounded support, i.e. $supp(\mathbf{X}) = [\underline{v}, \bar{v}]$ then it is bounded below as well. In fact, the following holds:

$$h(x) - \varrho(\bar{v} - \underline{v}) \leq S(\xi, x) \leq h(x),$$

where $\varrho = \max\{p, 1-p\}$.

Proof. (i) By (5.4), we have for all $\pi \in \Pi(\xi)$, $\mathbb{E}_\pi[S(\xi, x)] \geq \mathbb{E}_\pi[S(\xi', x)] \Leftrightarrow a\mathbb{E}_\pi[S(\xi, x)]+\mathbb{E}_\pi[b(x)] \geq a\mathbb{E}_\pi[S(\xi', x)] + \mathbb{E}_\pi[b(x)] \Leftrightarrow \mathbb{E}_\pi[aS(\xi) + b(x)] \geq \mathbb{E}_\pi[aS(q') + b(x)]$, (ii) The upper bound follows from the fact that $0 = supS^*(\xi, x)$, while the lower bound follows from the decomposition shown earlier in (5.6) $\square$

Property (i) is similar to Remark 1 which provides a wider family of strictly proper scoring rules through affine transformations. It implies that strict propriety is retained in the quantile setting for positive affine transformations. Property (ii), on the other hand, provides bounds for the score whenever the support is bounded. We note that the bounds are independent of q and are bounded on both sides when h is integrable with respect to $\mathscr{P}$.

The properties in Proposition 5.1.3 are related to the actual score. From an ex ante perspective, we might be interested in properties related to the expected score rather than the actual score. Our first result for the expected score provides a partial ordering

for certain classes of probability distributions. To avoid complications brought about by the extra term $h(x)$, we use the following notation:

$$S_p(\xi, x) = S^*(\xi, x) + h(x) \tag{5.8}$$

where $S^*(\xi, x) = -p(x-\xi)^+ - (1-p)(\xi - x)^+$.

Consider a distribution F and randomize each outcome x further so that the value of $y = x + z$, where z has distribution function $H(z)$ with mean zero. This generates a function G over y. In this case, we say that a distribution G obtained in this manner is a mean preserving spread of F through a randomization distribution H. This is mean preserving since H has mean 0.

Proposition 5.1.4. *Let $F, G \in \mathscr{P}$ and h be a concave function. If G is a mean-preserving spread of F then the expected scores under truthful reporting of the p-quantile for these beliefs imply that $\mathbb{E}_F S \geq \mathbb{E}_G S$.*

Proof. Note that the expected score under truthful reporting can be expressed as:

$$\begin{aligned}
\mathbb{E}_F S^* &= -(1-p)\int_{-\infty}^{F^{-1}(p)} (F^{-1}(p) - x) dF(x) - p\int_{F^{-1}(p)}^{\infty} (x - F^{-1}(p)) dF(x) \\
&= -(1-p)F^{-1}(p)F(F^{-1}(p)) + (1-p)\int_{-\infty}^{F^{-1}(p)} x dF(x) \\
&\qquad -p\Big[\mathbb{E}_F[X] - \int_{-\infty}^{F^{-1}(p)} x dF(x) - F^{-1}(p)(1 - F(F^{-1}(p)))\Big] \\
&= -p\mathbb{E}_F[X] + \int_{-\infty}^{F^{-1}(p)} x dF(x)
\end{aligned} \tag{5.9}$$

Now, since G is a mean-preserving spread of F, $\mathbb{E}_F[X] = \mathbb{E}_G[X]$, and the following property holds: (see Mas-Colell, et. al. 1995)

$$\int_{-\infty}^{\infty} u(x) dF(x) \geq \int_{-\infty}^{\infty} u(x) dG(x)$$

for all non-decreasing concave functions u. Set $u(x) = x\mathbf{1}\{x \leq r\} + r\mathbf{1}\{x > r\}$. And we get that $\int_{-\infty}^{F^{-1}(p)} x dF(x) \geq \int_{-\infty}^{G^{-1}(p)} x dG(x)$, which shows that $\mathbb{E}_F S^* \geq \mathbb{E}_G S^*$. If h is concave then $\mathbb{E}_F[h(x)] \geq \mathbb{E}_G[h(x)]$, which means that $\mathbb{E}_F S \geq \mathbb{E}_G S$. $\square$

Proposition 5.1.4 shows that given two distribution with the same mean if one second order stochastically dominates the other then the dominant one will have a better expected score. In the newsvendor framework, the intuition is that since there is less dispersion with one distribution, large errors resulting from being too far from the actual value are less likely in one setting compared to another.

This requirement of a mean-preserving spread however is somewhat restrictive. We are able to get this ordering condition under several other settings. An example is the Karlin-Novikoff(1963) condition. Suppose that two assessors report truthfully and it turns out that they have the same report, a. If $F(x) \leq G(x)$ for all $x < a$ and $F(x) \geq G(x)$ for all $x > a$, then we still get the same result: $\mathbb{E}_F S^* \geq \mathbb{E}_G S^*$. This follows because, on $[-\infty, r]$, F first order stochastically dominates G. Hence,

$$\int u(x)dF(x) \geq \int u(x)dG(x)$$

for any increasing function. Now, if we set $u(x) = (1-p)(x-r)\mathbf{1}\{x < r\}$, we have

$$-(1-p)\int_{-\infty}^{r}(r-x)dF(x) \geq -(1-p)\int_{-\infty}^{r}(r-x)dG(x).$$

Similarly, on the interval (r, ∞), G first order stochastically dominates F. Hence, when $u(x) = p(x-r)\mathbf{1}\{x > r\}$ is chosen. We have

$$\begin{aligned} p\int_{r}^{\infty}(x-r)dG(x) &\geq p\int_{r}^{\infty}(x-r)dG(x) \\ -p\int_{r}^{\infty}(x-r)dG(x) &\leq -p\int_{r}^{\infty}(x-r)dG(x), \end{aligned}$$

which yields the result.

To provide a more general result, we consider a dilation ordering.

Definition 17. *Let X and Y be two random variables. We say that Y is more dispersed than X in the dilation (or dilatation) sense, denoted by $X \leqslant_{dil} Y$, if*

$$\mathbb{E}[\phi(X - \mu_X)] \leq \mathbb{E}[\phi(Y - \mu_Y)]$$

for all convex functions ϕ, provided that these expectations exist. [Shaked & Shantikumar 2007]

This provides a generalization of comparing variances as means of studying the dispersion of distributions. Note that if $\phi(x) = x^2$ then the condition is equivalent to a comparison of variances. Unlike the traditional convex order (a.k.a. "increasing riskiness" condition) which was popularized in economics by Rothschild and Stiglitz (1970), the dilation order is location invariant, which makes it suitable for comparing distributions in our setting. For those who are familiar with income economics, the dilation order is a generalization of the well-known Lorenz order which is used in studying income inequality.

Our result can then be written as:

Proposition 5.1.5. *Let F and G be two distribution functions with finite-means, and let S be defined as in (5.6). For all $p \in (0,1)$,*

$$\mathbb{E}_F[S(\xi_F^*)] \geq \mathbb{E}_G[S(\xi_G^*)] \quad \Leftrightarrow \quad F \leq_{dil} G. \tag{5.10}$$

where ξ_F^* and ξ_G^* is the expected maximizing solution under distributions F and G respectively.

Proof. Note that from (5.9), we have:

$$\mathbb{E}_F S^* = -p\mathbb{E}_F[X] + \int_{-\infty}^{F^{-1}(p)} x dF(x).$$

Next, we note that $\mathbb{E}_F[S(\xi_F^*)] \geq \mathbb{E}_G[S(\xi_G^*)]$ is equivalent to

$$\begin{aligned} -p\mathbb{E}_F[X] + \int_{-\infty}^{F^{-1}(p)} x dF(x) &\geq -p\mathbb{E}_G[X] + \int_{-\infty}^{G^{-1}(p)} x dG(x) \\ \int_{-\infty}^{F^{-1}(p)} x dF(x) - \int_{-\infty}^{G^{-1}(p)} x dG(x) &\geq p(\mathbb{E}_F[X] - \mathbb{E}_G[X]) \\ \int_0^p (F^{-1}(u) - G^{-1}(u)) du &\geq p(\mathbb{E}_F[X] - \mathbb{E}_G[X]). \end{aligned}$$

But the last statement is a necessary and sufficient condition for dilation ordering (see Corollary 2.1 of Fagiuoli, et al. 1999). □

The main implication of this result is that an assessor who has a less informative prior distribution on the variable of interest should expect a lower score than a person

who is more informed. Hence, information gain should be measured in this context not by changes in the mean or variance but rather improvements in the distribution with respect to dilation ordering.

Proposition 5.1.6. *Given $F \in \mathscr{P}$, the expected score under truthful reporting satisfies the following:*

$$\lim_{p\to 0} \mathbb{E}_F S_p(F^-(p)) = 0$$
$$\lim_{p\to 1} \mathbb{E}_F S_p(F^-(p)) = 0$$

Proof. Let A and B be the lower and upper support (on $\bar{\mathbb{R}}$) of the distribution F

$$\begin{aligned}
\lim_{p\to 0} \mathbb{E}S^* &= \lim_{p\to 0} -p\mathbb{E}[X] + \int_A^{F^{-1}(p)} x dF(x) = 0 + \int_A^A x dF(x) = 0 \\
\lim_{p\to 1} \mathbb{E}S^* &= \lim_{p\to 1} -p\mathbb{E}[X] + \int_A^{F^{-1}(p)} x dF(x) = -\mathbb{E}[X] + \int_A^B x dF(x) = 0 \qquad \square
\end{aligned}$$

Proposition 5.1.6 shows that as we move towards more extreme forecasts the expected score that an assessor gets for truthful reporting diminishes. This may provide an incentive for her to be negligent with respect to the effort she puts into the assessment of more extreme values of p. To respond to this, we might scale the score up for values of p closer to 0 and 1 so that $\mathbb{E}S$ maintains a reasonable range of values to motivate the decision maker.

To illustrate Proposition 5.1.6, Figure 5.1 provides two plots for the expected score under truthful reporting. The left-hand side refers to plots for the exponential distribution (with $\lambda = 0.25, 0.5, 1, 2, 4$), while the right plot refers to a uniform distribution on the interval $(-\theta, \theta)$ (with $\theta = \{0.5, 1, 2, 5, 10\}$. Note that as the variance of the distribution becomes smaller, the expected scores tend to be closer to 0, and the discrepancy between expected scores for closer values of p become smaller as p approaches either 0 or 1.

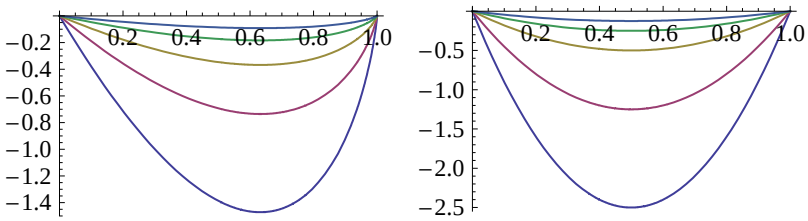

Figure 5.1: The expected report for truthful reporting in the quantile setting: an exponential and uniform example.

5.1.4 Multiple Quantiles

There are situations where a single quantile is sufficient for decision-making purposes. An example is the decision about how many papers to order in the newsvendor problem. Even in that situation, however, further information about F is needed if the newsvendor is considering getting additional information about the demand for papers. More generally, we are typically interested in more information about $\mathbf{X}$ and want to assess multiple quantiles. Suppose that we want to assess m quantiles $\xi_1, \ldots, \xi_m$ for the random variable $\mathbf{X}$. For the assessment of $\xi_i, i = 1, \ldots, m$, we use the linear scoring rule in (5.7) with coefficients p_i and $1 - p_i$, where $0 < p_1 < \cdots < p_m < 1$. That yields ($\boldsymbol{\xi}$,$\mathbf{p}$), where $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_m)$ and $\mathbf{p} = (p_1, \ldots, p_m)$. Adding the m scores, as Cervera and Muñoz (1996), and Gneiting and Raftery (2007) do, encourages truthful reporting $[\xi_i = F^{-1}(p_i), i = 1, \ldots, m]$ for an expert who wants to maximize her expected score. In our exposition, we assume $h(x)$ is a constant, i.e. $h(x) := h$, but note that general results for the $h(x)$ case can be attained with some extra steps.

Proposition 5.1.7. *The scoring rule*

$$S(\boldsymbol{\xi}, x) = \sum_{i=1}^{m} S(\xi_i, x) = mh - \sum_{i=1}^{m} p_i(x - \xi_i)^+ - \sum_{i=1}^{m}(1 - p_i)(\xi_i - x)^+ \tag{5.11}$$

is strictly proper, with the assessed quantiles ordered as follows: $\xi_1 < \xi_2 < \cdots < \xi_m$.

Proof. The additive scoring rule implies separability of the assessments. Therefore, $S(\boldsymbol{\xi}, x)$ is strictly proper since each $S(\xi_i, x)$ is strictly proper, and the ordering of the

quantiles follows from the ordering of $p_1, \ldots, p_m$ and the assumption that F is strictly increasing. □

From (5.11), the score attains a maximum value of mh when $\xi_1 = \xi_2 = \cdots = \xi_m = x$, i.e., when all of the assessed quantiles equal the observed value x. Next, we consider some properties of $S(\boldsymbol{\xi}, x)$. For convenience, we define $\xi_0 = -\infty$ and $\xi_{m+1} = \infty$.

Proposition 5.1.8. *$S(\boldsymbol{\xi}, x)$ is piecewise linear and concave in x, with*

$$S(\boldsymbol{\xi}, x) = b_j x + mh + \sum_{i=1}^{j} p_i \xi_i - \sum_{i=j+1}^{m} (1 - p_i)\xi_i \tag{5.12}$$

if $\xi_j \le x \le \xi_{j+1}$, $j = 0, \ldots, m$, where

$$b_0 = \sum_{i=1}^{m} (1 - p_i)$$

and

$$b_j = \sum_{i=j+1}^{m} (1 - p_i) - \sum_{i=1}^{j} p_i = m - j - \sum_{i=1}^{m} p_i = b_{j-1} - 1$$

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

Proof. If $\xi_j \le x \le \xi_{j+1}$,

$$\begin{aligned} \sum_{i=1}^{m} p_i (x - \xi_i)^+ + \sum_{i=1}^{m} (1 - p_i)(\xi_i - x)^+ &= \sum_{i=1}^{j} p_i (x - \xi_i)^+ + \sum_{i=j+1}^{m} (1 - p_i)(\xi_i - x)^+ \\ &= -b_j x - \sum_{i=1}^{j} p_i \xi_i + \sum_{i=j+1}^{m} (1 - p_i)\xi_i. \end{aligned}$$

Thus, (5.12) follows from (5.11). $S(\boldsymbol{\xi}, x)$ is piecewise linear in x from (5.12) and concave because the slopes b_j are decreasing in j. □

Proposition 5.1.9. *Let k be the smallest integer greater than or equal to b_0. Then $S(\boldsymbol{\xi}, x)$ is maximized at $x = \xi_k$, uniquely so if $b_0 = \sum_{i=1}^{m}(1 - p_i)$ is not an integer, and*

$$S(\boldsymbol{\xi}, \xi_k) = mh - w(\boldsymbol{\xi}), \tag{5.13}$$

where

$$w(\boldsymbol{\xi}) = \sum_{i=1}^{k-1} \Big[\sum_{l=1}^{i} p_l \Big] (\xi_{i+1} - \xi_i) + \sum_{i=k}^{m-1} \Big[\sum_{l=i+1}^{m} (1 - p_l) \Big] (\xi_i - \xi_{i-1}). \tag{5.14}$$

Proof. From Proposition 5.1.8, $b_0 > 0$ and $b_k = b_{k-1} - 1 = b_0 - k$. Thus, $b_i > (\leq) 0$ for $i < (\geq) k$, and $S(\boldsymbol{\xi}, x)$ is maximized at $x = \xi_k$. If $b_0 = \sum_{i=1}^{m}(1 - p_i)$ is (is not) an integer, $b_k = (>) 0$ and ξ_k is not (is) a unique maximum. Letting $x = \xi_k$ in (5.12) and rearranging terms yield (5.13) with

$$w(\boldsymbol{\xi}) = \sum_{i=1}^{k-1} p_i(\xi_k - \xi_i) - \sum_{i=k+1}^{m} (1 - p_i)(\xi_i - \xi_k),$$

from which expanding $\xi_k - \xi_i$ and $\xi_i - \xi_k$ into terms $(\xi_{i+1} - \xi_i)$ and $(\xi_i - \xi_{i-1})$ and rearranging again gives us (5.14). □

From Propositions 5.1.8 and 5.1.9, we see that the slopes of the piecewise linear, concave $S(\boldsymbol{\xi}, x)$ can be expressed in terms of sums of left-hand and right-hand cumulative probabilities p_i and $1 - p_i$. Slopes of successive pieces always decrease by one, which means that $\arg\max S(\boldsymbol{\xi}, x)$ can be determined by the initial slope b_0, which is the sum of the m right-hand cumulative probabilities. The maximum score for a given $\boldsymbol{\xi}$ is equal to the upper bound of mg minus $w(\boldsymbol{\xi}) \geq 0$, which represents a penalty for spreading out the quantiles. The contribution to this penalty of each interval between adjoining quantiles is the product of the width of the interval and a weight that becomes smaller as we move from intervals near the highest score toward intervals in the lower or upper tail.

Proposition 5.1.10. *If k is the smallest integer greater than equal to b_0, as in Proposition 5.1.9, and $\xi_j \leq x \leq \xi_{j+1}$, then we can decompose the score and expected score as follows:*

$$S(\boldsymbol{\xi}, x) = mh(x) - w(\boldsymbol{\xi}) - v(\boldsymbol{\xi}, x) \tag{5.15}$$

and

$$\mathbb{E}_F[S(\boldsymbol{\xi})] = mh - w(\boldsymbol{\xi}) - \mathbb{E}_F[v(\boldsymbol{\xi}, x)], \tag{5.16}$$

where

$$v(\boldsymbol{\xi}, x) = \begin{cases} b_{k-1}(\xi_k - x) + \sum_{i=j+1}^{k-1} (\xi_i - x) = b_j(\xi_{j+1} - x) + \sum_{i=j+1}^{k-1} b_i(\xi_{i+1} - \xi_i) & \text{if } j < k \\ -b_k(x - \xi_k) + \sum_{i=k+1}^{j} (x - \xi_i) = -b_j(x - \xi_j) - \sum_{i=k}^{j-1} b_i(\xi_{i+1} - \xi_i) & \text{if } j \geq k \end{cases} \tag{5.17}$$

Proof. $S(\boldsymbol{\xi}, x) = S(\boldsymbol{\xi}, \xi_k) - [S(\boldsymbol{\xi}, \xi_k) - S(\boldsymbol{\xi}, x)] = mh - w(\boldsymbol{\xi}) - [S(\boldsymbol{\xi}, \xi_k) - S(\boldsymbol{\xi}, x)]$. From (5.12), $S(\boldsymbol{\xi}, \xi_k) - S(\boldsymbol{\xi}, x)$ simplifies to (5.17). This gives us (5.15) and (5.16) follows directly. □

If $x = \xi_k$ or if $\xi_k \leq x \leq \xi_{k+1}$ and the slopes are integers so that $b_k = 0$, then $v(\boldsymbol{\xi}, x) = 0$. In this case, the only reduction from the maximum score $mh(x)$ is the penalty $w(\boldsymbol{\xi}) \geq 0$ for spreading out the quartiles. Otherwise, the score does not reach its maximum $mh - w(\boldsymbol{\xi})$ for the assessed $\boldsymbol{\xi}$, but is penalized further by an amount $v(\boldsymbol{\xi}, x) > 0$. This additional penalty is a function of how far x is from ξ_k, with a different weight for each interval between quantiles from x to ξ_k (including the partial interval from x to the nearest quantile in the direction of ξ_k). The contribution to the penalty of each interval is the product of the width of the interval and a weight that becomes larger as the intervals are further from ξ_k and the corresponding slopes get larger.

Consider an expert who is asked to assess her 0.05, 0.25, 0.50, 0.75, and 0.95 quantiles for $\mathbf{X}$, using the scoring rule in (5.11) with $h(x) = 200$. Here $m = 5$, the slopes $b_0, \ldots, b_5$ of the segments of $S(\boldsymbol{\xi}, x)$ are 2.5, 1.5, 0.5, -0.5, -1.5, and -2.5, and $S(\boldsymbol{\xi}, x)$ is maximized at $x = \xi_3$, the median of the expert's distribution of $\mathbf{X}$. Suppose that distribution is normal with mean 500 and standard deviation 100, i.e. $\mathbf{X} \sim N(500, 100^2)$. The score $S(\boldsymbol{\xi}, x)$ is shown in Figure 5.2 as a function of x assuming that the expert reports quantiles based on $N(500, 100^2)$ and assuming that she reports quantiles based on $N(500, 50^2)$. From (5.14), the lower standard deviation yields a lower value of $w(\boldsymbol{\xi})$ because the quantiles are less spread out, leading to narrower intervals between suc-

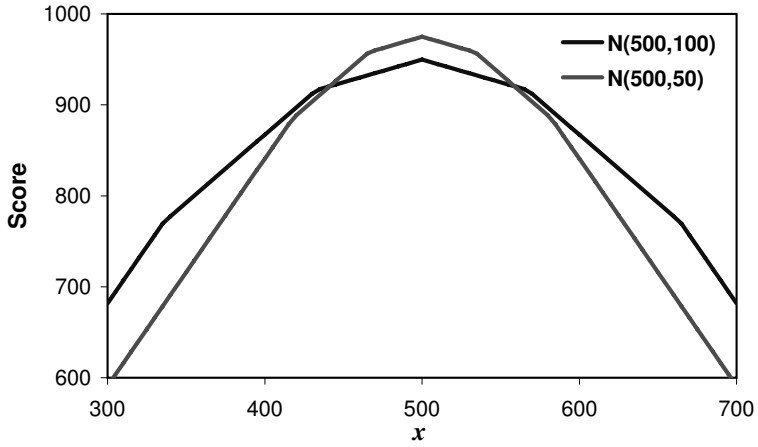

Figure 5.2: $S(\boldsymbol{\xi}, x)$ as a function of x for quantiles based on $N(500, 100^2)$ and $N(500, 50^2)$.

cessive quantiles. That is why the maximum score at $x = \xi_3 = 500$, which equals $1000 - w(\boldsymbol{\xi})$, is higher for the standard deviation of 50. However, with the less-spread-out quantiles the slopes get steeper more quickly as we move away from $x = 500$, and the score is higher for the standard deviation of 100 if $x < 441$ or $x > 559$.

The scoring rule in (5.11) is strictly proper, so not reporting quantiles truthfully leads to a decrease in the expert's expected score. What happens if the expert's distribution of $\mathbf{X}$ is $N(500, 100^2)$ but she reports quantiles based on a normal distribution with a different mean or standard deviation? Figure 5.3 shows the expected score as a function of μ for reports based on $N(\mu, 100^2)$ and Figure 5.4 shows the expected score as a function of σ for reports based on $N(500, \sigma^2)$. The expected score appears to be more sensitive to changes in the mean: basing quantiles on a mean of 450 or 550 instead of 500 results in about the same reduction in expected score as basing quartiles on a standard deviation of 50 or 175 instead of 100. This provides an indication of reductions in expected score due to not reporting truthfully.

If the expert does report her quantiles truthfully to maximize her expected score, as per Proposition 5.1.7, then how will her expected score vary with a change in the mean or standard deviation of her distribution? A change in the mean shifts all of

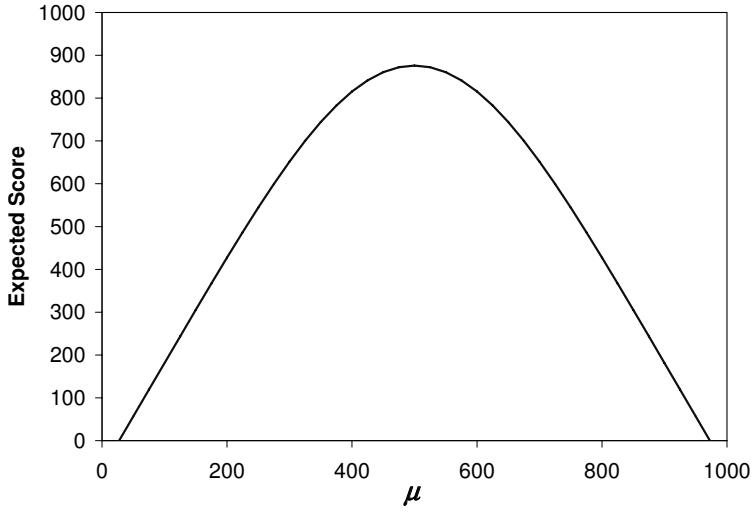

Figure 5.3: Expected score as a function of μ when the expert's distribution is $N(500, 100^2)$ and reports are based on $N(\mu, 100^2)$.

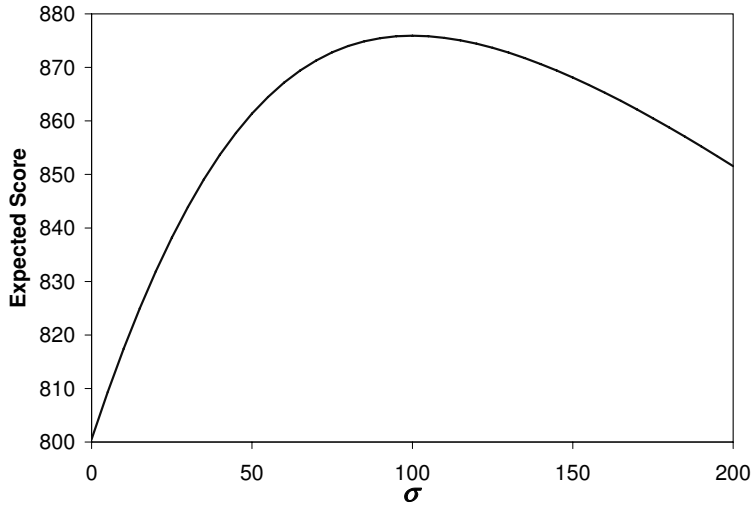

Figure 5.4: Expected score as a function of σ when the expert's distribution is $N(500, 100^2)$ and reports are based on $N(500, \sigma^2)$.

the quantiles by the same amount, and the width of the intervals between quantiles remains unchanged, as does the expected score. As for dispersion, we would expect the expected score to be lower if her standard deviation increases. As discussed above, $w(\boldsymbol{\xi})$ decreases as the dispersion of the expert's distribution decreases. Similarly, (5.17) implies that $\mathbb{E}_F[v(\boldsymbol{\xi}, x)]$ will decrease as the dispersion of F decreases because the quantiles are less spread out and x itself is expected to be less distant from $\arg\max S(\boldsymbol{\xi}, x)$. For the normal example, the expected score for an expert who truthfully reports quantiles based on her $N(500, \sigma^2)$ distribution for $\mathbf{X}$ is linear as a function of σ with a negative slope, -1.24. Thus, as σ increases, the expected score is reduced.

Not all of the details of the normal example generalize to other distributions. For instance, the linearity of the relationship between $\mathbb{E}_F[S(\boldsymbol{\xi})]$ and σ does not hold for all distributions. The dilation property for a single quantile can however be used in this setting. For this, we again set $h(x)$ to be a constant say h.

Proposition 5.1.11. *If F and G are cdfs of $\mathbf{X}$ with finite means, $\boldsymbol{\xi}^F$ and $\boldsymbol{\xi}^G$ are quantiles based on F and G for any fixed $\mathbf{p}$, and S is the scoring rule from (5.11), then if and only if $\mathbb{E}_F[S(\boldsymbol{\xi}^F)] \geq \mathbb{E}_G[S(\boldsymbol{\xi}^G)]$ if and only if F is less dispersed than G in the dilation ordering sense.*

Proof. The expected score under F can be expressed as

$$\mathbb{E}_F[S(\boldsymbol{\xi}^F)] = mh - \sum_{i=1}^{m} p_i \mathbb{E}_F[X] + \sum_{i=1}^{m} \int_{-\infty}^{F^{-1}(p_i)} x dF(x),$$

and similarly for G. Thus, $\mathbb{E}_F[S(\boldsymbol{\xi}^F)] \geq \mathbb{E}_G[S(\boldsymbol{\xi}^G)]$ is equivalent to

$$\begin{aligned} \sum_{i=1}^{m} \Big(\int_{-\infty}^{F^{-1}(p_i)} x dF(x) - \int_{-\infty}^{G^{-1}(p_i)} x dG(x) \Big) &\geq \sum_{i=1}^{m} p_i (\mathbb{E}_F[X] - \mathbb{E}_G[X]) \\ \sum_{i=1}^{m} \Big(\int_{-\infty}^{p_i} [F^{-1}(u) - G^{-1}(u)] du \Big) &\geq \sum_{i=1}^{m} p_i (\mathbb{E}_F[X] - \mathbb{E}_G[X]). \end{aligned}$$

This is a sufficient and necessary condition for F being less dispersed than G in the dilation ordering sense (Fagiuoli et al. 1999). □

Proposition 5.1.11 shows that our scoring rule for quantiles rewards sharpness in the sense that an expert whose distribution has less dispersion than the distribution

of another expert will have a higher expected score. In other words, a more informed expert can expect a higher score. This provides an incentive for an expert to obtain additional information or to do further analysis with existing information in order to reduce the dispersion of her distribution.

A special case of the use of multiple quantiles in practice is interval forecasts. They are widely encountered, in part because of their frequent use, along with interval estimates, in statistics. This means that experts are familiar with the notion of interval forecasts, so a possible assessment method is to assess an interval. If the relevant probabilities are fixed in advance, as in the typical case of a 90% interval that is symmetric in probability (i.e., has equal tail probabilities of 5%), then this is equivalent to assessing two quantiles (the 0.05 and 0.95 quantiles). Alternatively, the two quantiles can be assessed separately without mention of an interval, perhaps with additional assessments (e.g., $0.05 - 0.50 - 0.95$ quantiles). Of course, it is also possible to start with a fixed interval and assess the probability of the interval, which is probability assessment instead of quantile assessment, but it is more common to fix the probability associated with the interval and assess the quantiles that are the end points.

An interval forecast is an important special case of the assessment of multiple quantiles. For an interval forecast using the linear scoring rule in (5.11), $m = 2$ and the probability associated with the interval is $1 - \alpha = p_2 - p_1$. The slopes of $S(\boldsymbol{\xi}, x)$ below, within, and above the interval are $b_0 = 2 - (p_1 + p_2)$, $b_1 = 1 - (p_1 + p_2)$, and $b_2 = -(p_1 + p_2)$. Here $k = 1 + \mathbf{1}\{p_1 + p_2 \geq 1\}$. Dunsmore (1968) and Winkler (1972) consider linear loss functions for interval estimation related to the linear scoring rule in (5.11).

As noted above, intervals encountered in practice are typically symmetric in probability, which implies that $0 < p_1 < 0.5$, $p_2 = 1 - p_1$, $\alpha = 2p_1$, $(b_0, b_1, b_2) = (1, 0, -1)$, and $k = 1$. The scoring rule (5.11) is of the form

$$S(\boldsymbol{\xi}, x) = 2h - \frac{\alpha}{2}(\xi_2 - \xi_1) - (\xi_1 - x)^+ - (x - \xi_2)^+. \tag{5.18}$$

The penalty associated with the width of the interval is $w(\boldsymbol{\xi}) = (\alpha/2)(\xi_2 - \xi_1)$. The

penalty $v(\boldsymbol{\xi}, x)$ for deviations of x from ξ_k is $\xi_1 - x$ if $x < \xi_1$ (x is below the interval), zero if $\xi_1 \leq x \leq \xi_2$ (x is in the interval), and $x - \xi_2$ if $x > q_2$ (x is above the interval). Thus, the strictly proper nature of S results from a tradeoff between the penalty associated with the width of the interval and the penalty associated with having x fall outside the interval. The rule is flat when x falls within the interval because $v(\boldsymbol{\xi}, x)$ comes into play only if x is outside the interval. The slopes of the piecewise linear scoring rule are integers, and $\arg\max S(\boldsymbol{\xi}, x)$ is not unique; any $x \in [q_1, q_2]$ maximizes $S(\boldsymbol{\xi}, x)$.

From (5.18), the expected score under truthful reporting for a symmetric interval with $\alpha = 2p$ can be simplified to

$$\begin{aligned} \mathbb{E}_F[S(\boldsymbol{\xi})] &= 2h - \Big(\int_{\xi_2}^{\infty} x dF(x) - \int_{-\infty}^{\xi_1} x dF(x) \Big) \\ &= 2h - (\mathbb{E}_{\xi_2}^{\infty}[X] - \mathbb{E}_{-\infty}^{\xi_1}[X]). \end{aligned}$$

Intuitively, the partial expectations in the tails should get more extreme (further apart) as the dispersion increases, leading to a decrease in expected score. If F is $N(\mu, \sigma^2)$, then

$$\begin{aligned} w(\boldsymbol{\xi}) &= 2p\sigma|z|, \text{ and} \\ \mathbb{E}_F[v(\boldsymbol{\xi}, \mathbf{X})] &= 2[\sigma(2\pi)^{-1/2} \exp(-z^2/2) - p\sigma|z|], \end{aligned}$$

where $z = (F^{-1}(p) - \mu)/\sigma$. Therefore,

$$\mathbb{E}_F[S(\boldsymbol{\xi})] = 2h - \frac{\sqrt{2}\sigma}{\sqrt{\pi}} exp(-z^2/2);$$

the expected score decreases linearly in σ. If F is uniformly distributed on $[a, b]$, then

$$\begin{aligned} w(\boldsymbol{\xi}) &= p(1-2p)(b-a), \text{ and} \\ \mathbb{E}_F[v(\boldsymbol{\xi}, \mathbf{X})] &= p^2(b-a), \end{aligned}$$

so that

$$\mathbb{E}_F[S(\boldsymbol{\xi})] = 2h - p(1-p)(b-a),$$

which also decreases linearly in the standard deviation, $(b-a)/\sqrt{12}$. For both distributions, the expected score increases as $p \to 0$ for a given standard deviation, which

might be thought of as a reflection of an increased difficulty in assessing more extreme quantiles. The same general idea operates separately above and below the interval in the case of an interval that is asymmetric in terms of probability.

5.2 Monotonic Transformations

One of the ways that the scoring rule in (5.7) can be generalized is through a transformation on x and ξ. In particular, Gneiting and Raftery (2007) suggest the following generalization:

$$S(\xi, x) = ps(\xi) + (s(x) - s(\xi))\mathbf{1}\{x \leq \xi\} + h(x), \tag{5.19}$$

where s is any non-decreasing function. To avoid complications involving non-uniqueness of the optimal report, we focus on functions s which are strictly increasing on Ω.

The introduction of the function s retains the strict propriety of the score but changes the behavior of the scoring rule. When $s(\xi) = \xi$, the rule in (5.19) reduces to the linear rule.

Remark 10. *The scoring rule S given by (5.19) is strictly proper. [Gneiting & Raftery 2007]*

As an example, consider the power function defined for $z \geq 0$,

$$s(z) = \frac{1}{\beta}(z^{\beta} - 1). \tag{5.20}$$

When $\beta = 1$, we have the linear rule, while as $\beta \to 0$, we get the log rule, i.e. $s(z) = \ln z$. For all β, s is strictly increasing. Moreover, for $\beta > 1$ it is strictly concave, while for $\beta < 1$ it is strictly convex. Hence, the use of the function s enables us to generate a wider variety of rules. We shall refer to this function s as the *generator* of S and the resulting score as the s-transformed (or simply s) scoring rule.

We can think of the generator s simply as a transformation we apply to both our assessment and the actual observation. For example, if we are providing an assessment for a project's rate of return, s might represent a profit function for the project. This

provides an assessor the opportunity to step into the shoes of the decision maker and her score could then be interpreted as her share of the company's profit. This echoes the general idea that Savage (1971) had about scoring rules as functions that allow assessors to have a "share of the business." A stringent constraint however here is that s has to be a non-decreasing function, which may not always be realistic.

We can characterize this new generalized scoring rule by looking at the behavior of the scores for two assessments ξ and ξ'.

Proposition 5.2.1. *Let $s \in C^1$ be a strictly increasing function, ξ, ξ' be two reports such that $\xi < \xi'$. Report ξ receives a higher score than report ξ' if and only if the observed value is less than a unique critical value $\alpha \in [\xi, \xi']$, which is a weighted mean of the two reports.*

Proof. Consider WLOG $\xi < \xi'$. If $x < \xi$, then we note that $S(\xi, x) \geq S(\xi', x)$ iff

$$\begin{aligned} -(1-p)(s(\xi) - s(x)) &\geq -(1-p)(s(\xi') - s(x)) \\ s(\xi) &\leq s(\xi'). \end{aligned}$$

This is always true since s is increasing. Hence, ξ always gets a better score. Similarly, if $x > \xi'$ then ξ' always gets a better score. Finally, if $x \in [\xi, \xi']$ then ξ attains a better score than ξ' iff

$$\begin{aligned} p(s(\xi) - s(x)) &\geq (1-p)(s(x) - s(\xi')) \\ ps(\xi) + (1-p)s(\xi') &\geq s(x) \\ x &\leq s^{[-1]}(ps(\xi) + (1-p)s(\xi')). \end{aligned}$$

Since s is continuous and strictly increasing then this inverse is unique and moreover it lies in the interval $[\xi, \xi']$. □

This threshold value α is a quasi-arithmetic mean. For the power example in (5.20), the threshold value can be computed as follows:

$$\alpha = s^{[-1]}(ps(\xi) + (1-p)s(\xi')) = (p\xi^c + (1-p)\xi'^c)^{1/c} := M_c(\xi, \xi'; p),$$

which simply is the weighted mean of degree c with weight $\mathbf{p} = (p, 1-p)$. Therefore, under this transformation, the threshold value satisfies the following properties for $\xi \leq \xi'$:

$$\begin{aligned}
&\lim_{c \to -\infty} M_c(\xi, \xi'; p) = \xi \\
&\lim_{c \to \infty} M_c(\xi, \xi'; p) = \xi' \\
&M_c(\xi, \xi'; p) \leq M_{c'}(\xi, \xi'; p) \quad \text{for } c < c'.
\end{aligned}$$

with the last equality holding only when $\xi = \xi'$.

Proposition 5.2.1 paves the way for a characterization for the s-transformed scoring rule. This means that if the scoring rule gives a better score for a forecast which is based on the quasi-arithmetic mean threshold generated by s, then it is the s-transformed scoring rule. For the p-quantile linear rule, the threshold divides the interval $[\xi, \xi']$ according to the ratio $\frac{1-p}{p}$.

The next proposition demonstrates other properties that the threshold value may have under different generator functions leading to the notion of comparability between generators.

Proposition 5.2.2. *Under the conditions of the previous proposition, the following holds:*

i. If $a > 0$ is any constant then the critical values for the scoring rules with generators s and $s' = as + b$ are the same.

ii. Given the generator s, if $\xi \leq \tau$ and $\xi' \leq \tau'$ then $\alpha_{[\xi,\xi']} \leq \alpha_{[\tau,\tau']}$ where $\alpha_{[i,j]}$ is the threshold associated with reports i, j.

iii. Given the generator s, for every fixed ξ, ξ' and $a \in [\xi, \xi']$, there exists a unique p such that $\alpha_{[\xi,\xi']} = a$.

iv. We say that a function f is (strictly) convex with respect to a monotonic function g if $f(g^{-1})$ is (strictly) convex. Then given two scoring rules with generators s and s', $\alpha_{s,[\xi,\xi']} \leq \alpha_{s',[\xi,\xi']}$ for all $\xi, \xi' \in \mathbb{R}$ if and only if s is convex with respect to s'.

Proof. (i) The critical value is $x' = s'^{[-1]}(ps'(\xi) + (1-p)s'(\xi')) = s'^{[-1]}(a(ps(\xi) + (1-p)s(\xi')) + b) = x$, (ii) $(\xi \leq \tau, \xi' \leq \tau') \Rightarrow (ps(\xi) + (1-p)s(\xi') \leq ps(\tau) + (1-p)s(\tau'))$. Then the conclusion follows from the strict monotonicity of s, (iii) This immediately follows from the fact that for fixed values of ξ and ξ', $a : (\xi, \xi', p)$ is a continuous function from a compact set of the weights p to the closed interval $[\xi, \xi']$, (iv) By the convexity of s w.r.t. s' and Jensen's Inequality, we have:

$$\phi(ps(\xi) + (1-p)s(\xi')) \leq p\phi(s(\xi)) + (1-p)\phi(s(\xi')),$$

where $\phi := s(s'^{-1})$. Then the conclusion follows. □

These results generally follow from the fact that α is a quasi-arithmetic mean (cf. Bullen, et. al. 1988). Proposition 5.2.2 mentions the following properties: (i) invariance to affine transformations in s, (ii) monotonicity, (iii) uniqueness in p and (iv) comparability. The first two represent notions which are to be reasonably expected. In fact, using affine transformations imply that the scoring rules are essentially equivalent. The third property related to the uniqueness in p simply comes from the fact that s is restricted to be a strictly positive function.

The last property talks about the mathematical notion of comparability (in the sense of Hardy, et. al. 1934). We say that functions of two variables are comparable when given two values, the order between the functions never changes. For example, we say that the harmonic, geometric and arithmetic means denoted by $H(x,y), G(x,y)$ and $A(x,y)$ are comparable in the positive real line since for all x, y, $H(x,y) \leq G(x,y) \leq A(x,y)$. Hence, in terms of ordering, the direction of the inequalities does not reverse for any pair of x, y. Using the requirement of Proposition 5.2.2, it turns out that a wide class of functions are comparable to the linear generator. Thus, the choice of a generator asides from a linear generator must have a reasonable motivation and should consider other factors asides from simple ordering.

Next, we consider another property called sufficiency that may be useful in choosing a generator. A statistic T is *sufficient* for a measurable function $S(\mathbf{\Theta})$ if the conditional distribution of $S(\mathbf{\Theta})$ given $T = t$ is independent of $\mathbf{\Theta}$. For the quantile scoring rule

$S(\xi, x)$, this means that given the sufficient statistic we can compute the value of S regardless of whether we know ξ and x.

Proposition 5.2.3. *If $s(z) = az$ (or $s(z) = a \log z$) then the signed deviation (ratio) ξ between the assessment and the observed value is sufficient for S^*. In general, the signed s-deviation (ratio) ξ_s together with x is sufficient for S, with x being necessarily sufficient only through $h(x)$, i.e. ξ_s is sufficient for S whenever h is not dependent on x (i.e. it is a constant).*

Proof. We can write $S^*(\xi, x) = K_1(\xi - x)$ where K_1 can take on two possible constant values depending on the sign of $\xi - x$ for a linear generator. Similarly, $S^*(\xi, x) = K_2 \log(\xi/x)$ for a logarithmic generator, with K_2 being dependent on whether the ratio ξ/x is greater than or equal to 1. □

A nice feature of the linear scoring rule is that when $\tilde{h}(x)$ is a constant then the score is completely determined by the deviation of the assessment from the observed value. This is one reason why the solution obtained by papers looking only at the errors $\xi - x$ lead to the linear rule (e.g. Giacomini and Komunjer 2005). Moreover, we extend this result by looking at the ratio ξ/x, which relates to the logarithmic scoring rule. If we restrict our attention to the space of all s-transformed scoring rules, we have have the following characterization result.

Proposition 5.2.4. *The only continuous generator for which the scoring rule S^* is dependent on the deviations $\xi_i - x$, is $s(z) = az + b$ with $a > 0$. Also, the only generator in $\mathbb{R}^+$ for which the scoring rule is dependent on the ratios ξ_i/x is $s(z) = a \log z + b$ with $a > 0$. If h is independent of x, then the result also holds for S.*

Proof. Sufficiency in terms of the deviations $\xi - x$ implies that the term $s(\xi) - s(x)$ can be expressed as some function $\pi(\xi - x)$. First consider the case, $\xi \leq x$, we note that this leads to a modified Pexider functional equation: $\pi(\xi - x) = w(\xi) - w(x)$, with $w(z) = ps(z)$. This implies that $\pi(\xi - x)$ is exactly the same as $w(\xi - x)$ but translated by some constant $(-w(0))$. Hence, if we set $g(x) = w(x) - w(0)$, then we have

$g(x-y)+g(y) = g(x)$. Therefore under the assumption of continuity and monotonically increasing for s, the general solution for this is $g(x) = ax = w(x) + \text{constant} \Rightarrow s(x) = \alpha x + \beta$. A similar argument can be made for the case when $\xi > x$. Similarly, the case for the ratio follows using Cauchy's logarithmic functional equation. $\square$

Note that the sufficiency of ξ leads us to several things. First, it may provide us with insights on an appropriate choice for the generator s. In particular, if we want the sufficient statistic to be determined solely by the deviations from the observed value then $s(z)$ can be a linear function (i.e. $s(z) = az$ for $a > 0$). In a similar fashion, if our decision variables are restricted to the positive half-line, we may want the sufficient statistic to be determined solely by the ratio of the assessed values to the observed value. In this case, the function $s(z) = a \log z$, $a > 0$ satisfies this requirement (since the unique solution in $C^1(\mathbb{R}^+)$ to the functional equation $f(x/y) = f(x) - f(y)$ is this function). This may be useful when comparing values over a series of experiments or periods where the random quantity is non-stationary. For example, the score for an assessment of 100 and an outcome of 200 is the same as it is for an assessment of 1900 and the outcome of 2000 when a linear rule is used, but not when a logarithmic score is used.

Another property of potential interest is boundedness. If a scoring rule is used as an incentive tool for reporting, then it may be reasonable for a decision maker to place a cap such that the highest and lowest payoffs are restricted to some range. In this case, it may be desirable to use a transformation which is strictly increasing but also bounded.

For the linear and logarithmic rule, this is not satisfied since these functions are unbounded below. However, other rules are not bounded in this fashion. For example, some of the well-known functions in the area of population dynamics could be used such as:

$$\text{Logistic Curve} : s(x) = \alpha + \frac{\gamma}{(1 + \tau \exp(-\beta(x - \mu)))^{1/\tau}}$$
$$\text{Gompertz Curve} : s(x) = \alpha + (\gamma + \alpha) \exp(\zeta \exp(\chi x))$$

where $\zeta, \chi < 0$. The range for both functions is $[\alpha, \gamma + \alpha]$. Note that if the support is bounded, then all functions discussed earlier would apply, e.g. a linear function defined on this region (or truncated outside this region) would work.

Chapter 6

Summary and Future Research

6.1 Summary of Results

In this dissertation, we examine new methods for evaluating probabilistic information from expert opinion using scoring rules. We contribute to the literature and practice of expert elicitation and verification through (i) the development of two parametric families of scoring rules that can incorporate non-uniform baseline distributions, (ii) the creation of a new class of sensitive to distance strictly proper scoring rules and (iii) the analysis of scoring rules for quantile assessments and their properties.

First, we extend the power and pseudospherical families of scoring rules by allowing them to incorporate a baseline distribution from which the value of a forecast is measured. A nice feature of these families is that they provide a continuum of possibilities and contain as special cases several well-known scoring rules in the literature. They also have strong connections to two other large areas of research, namely information theory and utility maximization. We showed that the expected scores under truthful reporting for these scores correspond to some well-known generalized divergence measures and are the solutions to two utility maximization problems. These connections provide an information-theoretic justification and a decision-theoretic motivation for these new rules.

Next, in cases where the state space has a natural ordering, we noted that sensitivity to distance may play an important role in many decision problems. The ranked probability score is presently the only one considered in almost all of these applications. We enrich this class of scoring rules by providing a new algorithm that generates strictly proper scoring rules that are sensitive to distance. By an appropriate adjustment, we show that the power and pseudospherical families of scoring rules can be extended to take into account a baseline distribution while maintaining the property

of sensitivity to distance. These strictly proper scoring rules are the first to incorporate both baseline distributions and sensitivity to distance.

Finally, we provide a discussion on how forecasts can be evaluated in a quantile setting, which is another common method of eliciting probabilistic information from experts. By showing that the scoring rules developed in the probability assessment setting lose their incentive compatibility property in a quantile setting, we motivate the need to consider new scoring rules for quantile assessment. In particular, we study the (piecewise) linear scoring rule and its generalization through positive monotonic transformations. We also study some of its properties and provide new characterizations for some of its special and limiting cases. We argue that the knowledge of these properties can be useful in the selection of an appropriate scoring rule to be used in practice.

6.2 Future Research

This dissertation has focused on one interesting aspect of probability elicitation and verification. In particular, its contributions are primarily focused on the topic of scoring rules with some applications and connections outside the scoring rules literature such as finance and information theory also present.

Though the development and investigation of scoring rules is quite extensive in this dissertation, several interesting questions still remain. Some of these are old problems that have not yet been resolved, while others are questions that have risen from this dissertation.

In the development of weighted scoring rules, we have shown that there is a general connection between scoring rules, information theory and utility maximization. One interesting area of investigation would be the application of these concepts in other areas. For example, we have shown that there is room for such tools to be used in the context of financial markets. It would be interesting to see how these could be applied to areas such as marketing or operations where the need to measure information from

moving from one distribution to another also exists.

In addition, it would be interesting to see how the different rules compare against each other in various contexts. A comparative analysis of weighted scoring rules similar to what Bickel (2007) has done would provide interesting insights on instances where one scoring rule might be preferred over another. In addition, the discovery of new properties of these scoring rules would be useful when rules are chosen in practice.

In cases where the state space has an ordering, we have argued that sensitivity to distance may be a useful property for scoring rules. An open problem in this area has involved the existence of sensitive to distance scoring rules for other definitions of distance. For example, Murphy (1970) asks whether a sensitive to distance scoring rule exists for a different notion of distance such as one based on symmetric sums of probabilities for events centered at the event that is observed. If one exists, questions such as (1) "Does a complete characterization along the lines of Nau (2007) exist for these rules?", (2) "What natural decision-theoretic story can be given to these rules?", and (3) "For what applications is this new definition useful?", and (4) "What other definitions of distance might also be of interest?"

This is even more complicated when one moves from the univariate case to the multivariate case. Since it is feasible to discuss the notion of distance between probability distributions in higher dimensions, the challenge is to find ways to give the notion of sensitivity to distance a natural interpretation in this context. Winkler and Jose (2008) mention that this could be very challenging, especially when the different dimensions are measured using incomparable scales, since value tradeoffs would be needed to develop an appropriate distance measure.

In terms of quantile assessment, the literature on scoring rules for quantiles is still very young. An open problem suggested by Cervera and Muñoz (1996) is the existence of a characterization theorem for scoring rules in this context. The existence of such a result would have an impact not only on the scoring rules literature but on a wider class of problems such as quantile regression and estimation theory. With respect to the research that we have started, it would be interesting to find a way to incorporate

the notion of a baseline distribution into scoring rules for quantiles.

On a much broader level, one important question that needs further study involves the practical application of scoring rules in probability assessment. Though the mathematical theory has a solid foundation, a question that many practitioners might ask is how different scoring rules affect the behavior of individuals whose probabilities are being elicited. Though some experimental work (e.g. Staël von Holstein 1970) has been done, there is still no general consensus on the practical impact of scoring rules in elicitation contexts. Are they really useful guides for making probability assessments or are they simply suited for ex post verification? This is also true for quantile assessment. Further experimental work using scoring rules for these types of elicitation would be useful.

For both probability and quantile assessments, one interesting research question is the applicability of scoring rules in the context of sequential assessments. Given that information on the uncertainty and performance is revealed over time, it would be interesting to investigate how information gains should be appropriately accounted for using scoring rules.

Bibliography

Arimoto, S. 1971. Information-theoretical considerations on estimation problems. *Information and Control*, **19**:181–194.

Bickel, E. 2007. Some comparisons among quadratic, spherical and logarithmic scoring rules. *Decision Analysis*, **4**(2):49–65.

Blattenberger, G., F. Lad. 1985. Separating the Brier score into calibration and refinement components: a graphical exposition. *American Statistician*, **39**:26–32.

Boekee, D. E., J. C. A. van der Lubbe. 1980. The R-norm information measure. *Information and Control*, **45**:136-155.

Brier, G. W. 1950. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, **78**(1):1–3.

Bullen, P. S., D. S. Mitrinović, P. M. Vasić. 1988. Means and their inequalities. *Mathematics and its Applications: East European Series*. D. Reidel Publishing Company, Dodrecht.

Candille, G., O. Talagrand. 2005. Evaluation of probabilistic prediction systems for a scalar variable. *Quarterly Journal of the Royal Meteorological Society*, **131**:2131–2150.

Cervera, J.L., J. Muñoz. 1996. Proper scoring rules for fractiles. In *Bayesian Statistics 5*, J.M. Bernardo, J.O. Berger, A.P. Dawid, A.F.M. Smith, eds. Oxford University Press, Oxford, England, 513–519.

Cooke, R. M. 1991. Experts in uncertainty: opinion and subjective probability in science. Oxford University Press, Oxford, England.

Cressie, N., T. R. C. Read. 1984. Multinomial goodness of fit. *Journal of the Royal Statistical Society Series B*, **46**(3):440–464.

Csiszár, I. 1963. Eine informationstheoretische ungleichung und ihre andwendung auf den bewis der ergodizität on markhoffschen ketten. *Publications of the Mathematical Institute of the Hungarian Academy of Sciences*, **8**:84–108.

Dawid, A. P. 1982. The well-calibrated Bayesian. *Journal of the American Statistical Association*, **77**:605–613.

Dawid, A. P. 1986. Probability forecasting. In *Encyclopedia of Statistical Sciences*, Vol. 7, S. Kotz, N. L. Johnson, C. B. Read, eds. Wiley Publishers, New York, 210-218.

De Finetti, B. 1962. Does it make sense to speak of 'good probability appraisers'? In *The Scientist Speculates: An Anthology of Half-baked Ideas*, I. J. Good ed. Heinemann, London, 357–364.

Delbaen, F., P. Grandits, T. Rheinländer, D. Samperi, M. Schweizer, C. Stricker. 2002. Exponential hedging and entropy penalties. *Mathematical Finance*, **12:**99–123.

Dunsmore, I. R. 1968. A Bayesian approach to calibration. *Journal of the Royal Statistical Society Series B*, **30**(2): 396–405.

Epstein, E. S. 1969. A scoring system for probability forecasts of ranked categories. *Journal of Applied Meteorology*, **8**(6):985–987.

Fagiuoli, E., F. Pellerey, M. Shaked. 1999. A characterization of dilation order and its applications. *Statistical Papers*, **40**:393–406.

Foster, D. P., R. Vohra. 1998. Asymptotic calibration. *Biometrika*, **85**:379–390.

Friedman, D. 1983. Scoring rules for probabilistic forecasts. *Management Science*, **29**(4):447–454.

Frittelli, M. 2000. The minimal entropy martingale measure and the valuation problem in incomplete markets. *Mathematical Finance*, **10:**39–52.

Giacomini, R., I. Komunjer. 2005. Evaluation and combination of conditional quantile forecasts. *Journal of Business & Economic Statistics*, **23**(4):416–431.

Gneiting, T., A. E. Raftery, A. H. Westveld, T. Goldman. 2005. Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. *Monthly Weather Review*, **33**:1098–1118.

Gneiting, T., A. Raftery. 2007. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, **102**:359–378.

Good, I. J. 1952. Rational decisions. *Journal of the Royal Statistical Society Series B*, **14**(1):107–114.

Good, I. J. 1971. Comment on Buehler. In *Foundations of Statistical Inference*, Godambe and Sprott, eds., Holt, Reinhart, & Winston, Toronto, 337–339.

Goll, T., Rüschendorf, L. 2001. Minimmax and minimal distance martingale measures and their relationship to portfolio optimization. *Finance and Stochastics*, **5**:557–581.

Hardy, G. H., J. E. Littlewood, G. Polya. 1934. Inequalities. Cambridge University Press, England.

Haussler, D., M. Opper. 1997. Mutual information, metric entropy, and cumulative relative entropy risk. *The Annals of Statistics*, **25**:2451–2492.

Havrda, J., F. Chavrát. 1967. Quantification method of classification processes: the concept of structural α-entropy. *Kybernetika*, **3**:30–35.

Hendrickson, A. D., R. J. Buehler 1971. Proper scores for probability forecasters. *The Annals of Mathematical Statistics*, **42**:1916–1921.

Hersbach, H. 2000. Decomposition of the continuous ranked probability score for ensemble prediction systems. *Weather and Forecasting*, **15**:559–570.

Jose, V. R. R. 2009. A characterization for the spherical scoring rule. *Theory and Decision*, **66**(3):263–281.

Jose, V. R. R., R. F. Nau, R. L. Winkler. 2008. Scoring rules, generalized entropy, and utility maximization. *Operations Research*, **56**(5):1146–1157.

Jose, V. R. R., R. F. Nau, R. L. Winkler. 2009. Sensitivity to distance and baseline distributions in forecast evaluation. *Management Science*, forthcoming.

Jose, V. R. R., R. L. Winkler. 2009. Evaluating quantile assessments. *Operations Research*, forthcoming.

Karlin, S., A. Novikoff. 1963. Generalized convex inequalities. *Pacific Journal of Mathematics*, **13**:1251–1279.

Koenker, R., G. Bassett. 1978. Regression quantiles. *Econometrica*, **46**(1):33–50.

Kullback, S. 1959. Information theory and statistics. John Wiley & Sons, New York.

Lavenda, B. H., J. Dunning-Davies. 2003. Qualms concerning Tsallis's condition of pseudo-additivity as a definition of non-extensivity. http://arxiv.org/abs/cond-mat/0312132.

Lichtendahl, K. C., R. L. Winkler. 2007. Probability elicitation, scoring rules and competition among forecasters. *Management Science*, **53**(11):1745–1755.

Mas-Colell, A., M. Whinston, J. R. Green. 1995. Microeconomic theory. Oxford University Press, England.

MATHESON, J., R. L. WINKLER. 1976. Scoring rules for continuous probability distributions. *Management Science*, **22**(10):1087–1096.

MCCARTHY, J. 1956. Measures of the value of information. *Proceedings of the National Academy of Sciences of the United States of America*, **42**:654–655.

MURPHY, A. H. 1970. The ranked probability score and the probability score: a comparison. *Monthly Weather Review*, **98**(12):917–924.

MURPHY, A. H. 1971. A note on the ranked probability score. *Journal of Applied Meteorology*, **10**(1):155–156.

MURPHY, A. H. 1973. Hedging and skill scores for probability forecasts. *Journal of Applied Meteorology*. **12**(1):215–223.

MURPHY, A., R. L. WINKLER. 1987. A general framework for forecast verification. *Monthly Weather Review*, **115**(7):1330-1338.

NAU, R. F. 1985. Should scoring rules be 'effective'? *Management Science*, **31**(5):527–535.

NAU, R. F. 2007. Scoring rules that are sensitive to distance and dominance. Working Paper. The Fuqua School of Business, Duke University, Durham, NC.

NAU, R. F., V. R. R. JOSE, R. L. WINKLER. 2007. Scoring rules, entropy and imprecise probabilities. Proceedings of the 5th International Symposium on Imprecise Probabilities and their Applications (ISIPTA), Prague, Czech Republic: Action M Agency, 307-315.

NAU, R. F., V. R. R. JOSE, R. L. WINKLER. 2009. Duality between maximization of expected utility and minimization of relative entropy when probabilities are imprecise. Working Paper. The Fuqua School of Business, Duke University, Durham, NC.

PARDO, L. 2006. Statistical inference based on divergence measure. Taylor & Francis, Boca Raton, Florida.

PEARSON, K. 1900. On a criterion that a given system of deviations from the probable in the case of correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *The London, Edinburgh and Dublin Philosophical Magazine and Journal of Science Series 5*, **50**:157–175.

RATHIE, P. N., P. KANNAPPAN. 1972. A directed-divergence function of type β. *Information and Control*, **20**:38–45.

Roby, T. B. 1965. Belief states and the use of evidence. *Behavioral Science*, **10**:255–270.

Sanders, F. 1963. On subjective probability forecasting. *Journal of Applied Meteorology*, **2**(2):191–201.

Rothschild, M., J. Stiglitz. 1970. Increasing risk I: a definition. *Journal of Economic Theory*, **2**:225-243.

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. *The Annals of Statistics*, **17**:1856-1879.

Selten, R. 1998. Axiomatic characterization of the quadratic scoring rule. *Experimental Economics*, **1**:43–62.

Shaked, M., J. G. Shanthikumar. 2007. Stochastic Orders. Springer Series in Statistics. Springer, New York.

Shannon, C. E. 1948. A mathematical theory of communication. *Bell System Technical Journal*, **27**:379–423.

Sharma, B. D., D. P. Mittal. 1975. New non-additive measures of entropy for discrete probability distributions. *Journal of Mathematical Science*, **10**:28–40.

Shufford, E. H., A. Albert, H. E. Massengill. Admissible probability measurements procedures. *Psychometrika*, **31**:125–145.

Staël von Holstein, C. A. S. 1970. *Assessment and Evaluation of Subjective Probability Distributions*. The Economic Research Institute at the Stockholm School of Economics, Stockholm.

Staël von Holstein, C. A. S. 1977. The continuous ranked probability score in practice. In *Decision Making and Change in Human Affairs*, Jungermann, H., G. de Zeeuw, eds., D. Reidel, Dordrecht, 263–273.

Unger, D. A. 1985. A method to estimate the continuous ranked probability score. In *Preprints of the 9th Conference on Probability and Statistics in Atmospheric Sciences*, American Meteorological Society, Virginia Beach, Virginia, 206–213.

Vajda, I. 1989. Theory of statistical inference and information. *Theory and Decision Library Series B: Mathematical and Statistical Methods*. Kluwer Academic Press, Do-

drecht.

Verdu, S., S. W. McLaughlin. 1999. Information theory: 50 years of discovery. Wiley-IEEE Press, New York.

Winkler, R. L. 1972. A decision-theoretic approach to interval estimation. *Journal of the American Statistical Association*, **67**:187–191.

Winkler, R. L. 1986. On "good probability appraisers." In *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti*, Goel, P., A. Zellner eds. North-Holland, Amsterdam, 265–278.

Winkler, R. L. 1994. Evaluating probabilities: asymmetric scoring rules. *Management Science*, **40**(11):1395–1405.

Winkler, R. L. 1996. Scoring rules and the evaluation of probabilities (with discussions). *Test*, **5**(1):1–60.

Winkler, R. L., A. H. Murphy. 1968. 'Good' probability assessors. *Journal of Applied Meteorology*, **7**(5):751–758.

Winkler, R. L., A. H. Murphy. 1970. Nonlinear utility and the probability score. *Journal of Applied Meteorology*, **9**(1):143–148.

Winkler, R. L., V. R. R. Jose. 2008. Comment: Assessing probabilistic forecasts of multivariate quantities, with an application to ensemble predictions of surface winds, *Test*, **17**(2), 251–255.

Wiper, M. P., S. French, R. Cooke. 1994. Hypothesis-based calibration scores. *The Statistician*, **43**(2):231–236.

Biography

YEAR AND PLACE OF BIRTH

- 1981, Metro Manila, Philippines

EDUCATION

- Ph.D. in Business Administration, The Fuqua School of Business, Duke University, 2009
- B.S. in Mathematics, School of Science & Engineering, Ateneo de Manila University, 2003
- B.S. in Management Engineering, John Gokongwei School of Management, Ateneo de Manila University, 2002

RECENT PUBLICATIONS

1. Jose, V. R. R., R. F. Nau, R. L. Winkler. 2009. Sensitivity to Distance and Baseline Distributions in Forecast Evaluation. *Management Science*, forthcoming.
2. Jose, V. R. R., R. L. Winkler. 2009. Evaluating Quantile Assessments. *Operations Research*, forthcoming.
3. Jose, V. R. R. 2009. A Characterization for the Spherical Scoring Rule. *Theory and Decision*, **66**(3):263–281.
4. Jose, V. R. R., R. F. Nau, R. L. Winkler. 2008. Scoring Rules, Generalized Entropy, and Utility Maximization. *Operations Research*, **56**(5): 1146-1157.
5. Winkler, R. L., V. R. R. Jose. 2008. Comment on "Assessing Probabilistic Forecasts of Multivariate Quantities, with an Application to Ensemble Prediction of Surface Winds," *Test*(*A Journal of the Spanish Society of Statistics*), **17**(2):251-255.
6. Jose, V. R. R., R. L. Winkler. 2008. Simple Robust Average of Forecasts: Some Empirical Results. *International Journal of Forecasting*, **24**(1):163-169.
7. Jose, V. R. R., K. C. Lichtendahl, R. F. Nau, R. L. Winkler. 2007. Comment on "Objective Priors for the Multivariate Normal Distribution," in *Bayesian Statistics 8*, J. Bernardo, et. al. eds. Oxford: University Press, 557-558.
8. Nau, R. F., V. R. R. Jose, R. L. Winkler. 2007. Scoring Rules, Entropy and Imprecise Probabilities. Proceedings of the 5th International Symposium on Imprecise Probabilities and their Applications (ISIPTA), Prague, Czech Republic: Action M Agency, 307-315.

SELECTED HONORS AND SCHOLARSHIPS

1. Finalist, Decision Analysis Student Paper Competition, 2008
2. Graduate Fellowship, The Fuqua School of Business, Duke University, 2004-2009
3. First Place, Dean's Award for Undergraduate Research in Science and Engineering, Ateneo de Manila University, 2003
4. Departmental Award for Management Engineering and Mathematics, Ateneo de Manila University, 2002 and 2003
5. Undergraduate Scholarship, Ateneo de Manila Scholarship Foundation, 1998-2002