Australian Government
Bureau of Meteorology

Point forecasts and forecast evaluation with generalised Huber loss

Robert Taggart

December 2020

Bureau Research Report - 050

Point forecasts and forecast evaluation with generalised Huber loss

Robert Taggart

Bureau Research Report No. 050

December 2020

National Library of Australia Cataloguing-in-Publication entry

Authors: Robert Taggart

Title: Point forecasts and forecast evaluation with generalised Huber loss

ISBN: 978-1-925738-25-4

Series: Bureau Research Report – BRR050

Enquiries should be addressed to:

Robert Taggart

Bureau of Meteorology
PO Box 413
Darlinghurst NSW 1300
Australia

Robert.Taggart@bom.gov.au

Copyright and Disclaimer

The Bureau of Meteorology advise that the information contained in this publication comprises general statements based on scientific research. The reader is advised and needs to be aware that such information may be incomplete or unable to be used in any specific situation. No reliance or actions must therefore be made on that information without seeking prior expert professional, scientific and technical advice. To the extent permitted by law and the Bureau of Meteorology (including each of its employees and consultants) excludes all liability to any person for any consequences, including but not limited to all losses, damages, costs, expenses and any other compensation, arising directly or indirectly from using this publication (in part or in whole) and any information or material contained in it.

Contents

List of Figures

List of Tables

Executive summary

In the practice of point forecasting, it is desirable that forecasters receive a directive in the form of a scoring function (such as the absolute error or squared error scoring functions) that will be used to evaluate forecasts, or of a statistical functional (such as the median or mean) of the predictive distribution that is being sought. In this report, we study the properties of a family of scoring functions that are intermediaries between the squared error or absolute error scoring functions, and an associated family of statistical functionals that are intermediaries between the median and mean, and demonstrate the value of these families for point forecasting contexts.

More precisely, we introduce a three-parameter family of functionals, called *Huber functionals*, which nest as limiting cases the family of quantiles (which includes the median) and the family of expectiles (which includes the mean). The Huber functional of a predictive distribution has the property that it is an optimal point forecast when scored with the generalised Huber loss scoring function. This scoring function loss applies a (possibly asymmetric) quadratic penalty to small errors and a (possibly asymmetric) linear penalty to large errors. A particular subfamily of Huber functionals are the *Huber means*, which are intermediaries between the median and mean functionals. The Huber mean of a distribution is the midpoint of the central interval of the distribution, and so is the natural functional to use when one wants a point summary of the location of the central bulk of a distribution while ignoring behaviour at its tails.

We give three important theoretical results about the Huber functional: a characterisation of its consistent scoring functions, that it is elicitable, and that its consistent scoring functions have a mixture representation in terms of elementary scoring functions. The elementary scoring functions of the Huber functional admit an economic interpretation in the context of investment problems with capped profits and losses, so that point forecasts targeting the Huber functional are used to construct optimal decision rules in such situations. Synthetic experiments illustrate the utility of the Huber loss scoring function as a robust scoring function for situations with contaminated observational data.

The Huber functional thus presents an attractive alternative to quantiles and expectiles in many point forecasting contexts.

1 Introduction

In many fields of human endeavour, it is desirable to make forecasts for an uncertain future. Hence, forecasts should be probabilistic, presented as probability distributions over possible future outcomes [Gneiting and Katzfuss, 2014]. Nonetheless, many practical situations require forecasters to issue single-valued point forecasts for reasons including ease of communication, reporting requirements and tradition. In this situation, a directive is required about the specific feature or functional of the predictive distribution that is being sought, or about the specific loss (or scoring) function that is to be minimised [Gneiting, 2011a, Ehm et al., 2016].

Well-known target functionals include the mean, median or a specific quantile. A lesser known family of functionals is the expectiles, which has recently attracted interest in risk management [Bellini and Di Bernardino, 2017]. Examples of scoring functions include the squared error scoring function $S(x, y) = (x - y)^2$ and absolute error scoring function

$S(x,y) = |x-y|$. More generally, a functional T is mapping from a class $\mathcal{F}$ of probability distributions to (subsets of) the real line $\mathbb{R}$, while a scoring function S is a non-negative function on a specified prediction–observation domain that assigns a loss or penalty $S(x,y)$ when the point forecast x is issued and the observation y realises [Gneiting, 2011a]. The predictive performance of a forecast system can be assessed by computing its mean score $\bar{S}$ over a finite number of forecast cases.

In this report we introduce a three-parameter family $\{\mathrm{H}^{\alpha}_{a,b} : 0 < \alpha < 1, a > 0, b > 0\}$ of functionals that act on probability distributions that are defined on the real line $\mathbb{R}$ or on subintervals of $\mathbb{R}$. This family nests, as limiting cases, the quantiles (including the median functional) and the expectiles (including the mean functional). The special case $\mathrm{H}^{1/2}_{a,a}$, where $a > 0$, has the property that it generates the set of optimal point forecasts for the scoring function

$$S(x,y) = \begin{cases} \frac{1}{2}(x-y)^2\,, & |x-y| \le a \\ a|x-y| - \frac{1}{2}a^2\,, & |x-y| > a\,, \end{cases} \tag{1.1}$$

in the sense that the expected score $\mathbb{E}_F S(x,Y)$ is minimised precisely whenever $x \in \mathrm{H}^{1/2}_{a,a}(F)$. (Here and throughout the notation $\mathbb{E}_F$ indicates that the expectation is taken with respect to $Y \sim F$ for a given distribution F.) The scoring function S of Equation (1.1) is essentially the classical Huber loss function [Huber, 1964], named for Peter Huber's pioneering work on robust parameter estimation, while the functional $\mathrm{H}^{1/2}_{a,a}$ also featured in [Huber, 1964] for finite discrete distributions. We therefore call $\mathrm{H}^{1/2}_{a,a}$ a *Huber mean*. Members $\mathrm{H}^{\alpha}_{a,b}$ of the larger family generate optimal point forecasts of a generalised version of the scoring function (1.1), and we name them *Huber functionals*.

The Huber mean has properties that make it a useful summary of the centre of a probability distribution F without being sensitive to the behaviour of F at its tails. Roughly speaking, the Huber mean $\mathrm{H}^{1/2}_{a,a}$ is the midpoint of the 'central interval' of F with length $2a$. It is an intermediary between the median (which it approaches as $a \downarrow 0$) and the mean (which it approaches as $a \to \infty$). This makes the Huber mean a useful target functional in situations where a point summary of the central 'bulk' of a distribution is desired, neglecting behaviour at the tails. Similarly, the Huber functional $\mathrm{H}^{\alpha}_{a,a}$ is an intermediary between the α-quantile (which it approaches as $a \downarrow 0$) and the α-expectile (which it approaches as $a \to \infty$).

This report has theoretical and applied aspects. The three main theoretical results are that the Huber functional is elicitable (Theorem 4.5), a characterisation of its consistent scoring functions (also Theorem 4.5), and that its consistent scoring functions have a mixture representation in terms of elementary scoring functions (Theorem 5.2).

A scoring function S is consistent for a functional T relative to some class $\mathcal{F}$ of distributions if every value of $\mathrm{T}(F)$ is a minimiser $\hat{x}$ of the expected score $\mathbb{E}_F S(x,Y)$, for every distribution F in $\mathcal{F}$. If, in addition, every minimiser $\hat{x}$ belongs to $\mathrm{T}(F)$ then S is strictly consistent. Under a strictly consistent scoring function, a forecaster will optimise their expected score by giving a truthful and accurate assessment of the functional $\mathrm{T}(F)$ [Gneiting, 2011a]. To give an example, the squared error scoring function is strictly consistent for the mean functional relative to the class of probability distributions with finite variance. But there are many other consistent scoring functions for this functional. Savage [Savage, 1971] showed that, under weak regularity conditions, S is consistent for the mean functional if and only if it is of the form

$$S(x,y) = \phi(y) - \phi(x) + (x-y)\phi'(x)\,, \tag{1.2}$$

where ϕ is a convex function with subgradient ϕ'; the squared error scoring function arises when $\phi(t) = t^2$. Analogous results exist for quantiles [Gneiting, 2011b] and expectiles [Gneiting, 2011a]. Our first main theoretical result is that, under weak regularity conditions, the consistent scoring functions S for the Huber functional can also be characterised as having a general form, where, similarly to Equation (1.2), S is parametrised by convex functions ϕ. Moreover, edge cases of this form recover the general form of the consistent scoring functions for quantiles and expectiles.

If point forecasts targeting a specified functional are to be evaluated against observations, it is critical that the functional has a strictly consistent scoring function relative to a suitable class $\mathcal{F}$ of probability distributions. If such a scoring function exists, the functional is said to be elicitable. In contrast, the quality of point forecasts targeting a nonelicitable functional cannot be adequately assessed using a scoring function [Gneiting and Katzfuss, 2014]. Our second main theoretical result is that, if I is an interval of $\mathbb{R}$ then the Huber functional is elicitable relative to the class of probability distributions on I, with the additional condition that the distributions have finite first moment in the case when $I = \mathbb{R}$.

Our third major theoretical result is that, subject to unimportant regularity conditions, every consistent scoring function S for a specific Huber functional admits a mixture representation of the form

$$S(x,y) = \int_{-\infty}^{\infty} S_\theta(x,y)\,\mathrm{d}M(\theta) \tag{1.3}$$

where M is a non-negative measure and $\{S_\theta : \theta \in \mathbb{R}\}$ is a one-parameter family of consistent scoring functions called elementary scoring functions. That is, every consistent scoring function can be written as a weighted average of elementary scoring functions. Analogous results for the consistent scoring functions of quantiles and expectiles were given by [Ehm et al., 2016], and these results are recoverable from the representation in the Huber functional case by taking appropriate limits. Mixture representations have several applications, such as determining whether one forecast system empirically dominates another (in the sense that the mean score of one system is lower than the other for every consistent scoring function), and aiding the selection of an appropriate consistent scoring function for forecasts disseminated to a heterogeneous user group.

From a practical standpoint, we highlight three use cases of the Huber functional or of Huber loss. First, as mentioned above, the Huber mean will be of interest when a point summary specifying the location the central interval of a distribution, unaffected by behaviour at the tails, is desired.

Second, the Huber functional arises in optimal decision rules for investment problems with fixed up-front costs and where profits and losses are both capped. The economic regret, relative to actions based on a perfect forecast, in such investment problems is proportional to one of the elementary scores $S_\theta(x,y)$ in the mixture representation (1.3). Thus the optimal decision rule specifies action based on point forecasts that target an appropriate Huber functional. Past performance of competing point forecasts that target the Huber functional can be evaluated using a Murphy diagram, which is a graph of mean elementary scores $\bar{S}_\theta(x, y)$ against θ [Ehm et al., 2016], so that each decision maker can choose to use the forecast system that would have historically minimised economic regret based on their particular decision rule.

Third, we demonstrate that Huber loss can be used as a robust scoring function for point forecasts targeting the mean functional in situations where forecasts are judged against observations that are contaminated, say, by faults in the observation measurement process. While the squared error scoring function is consistent for the mean, the presence of contaminated observations can grossly distort forecast rankings based on it. The Huber loss scoring function provides a palatable alternative.

The remainder of this report is organised as follows. In Section 2 we introduce notation and lay out the mathematical setting. Section 3 defines the Huber functional, states some of its basic properties, and compares it with quantiles and expectiles. Results about the elicitability of the Huber functional and the characterisation of its scoring functions are given in Section 4. Section 5 presents the mixture representation for consistent scoring functions of the Huber functional, and discusses a range of applications related to this result, including understanding the rankings of forecasts, choosing consistent scoring functions for heterogeneous user groups, and identifying the type of investment problems that naturally give rise to the Huber functional. The use of Huber loss for robust verification of forecasts targeting the mean functional is illustrated in Section 6 via a synthetic experiment. Conclusions are summarised in Section 7, and the proofs of the main results are given in the appendix.

2 Notation

We work in a setting where point forecasts x and observations y take values in some interval I of the real line $\mathbb{R}$, including the case when $I = \mathbb{R}$. A predictive distribution F can be issued for a future, as yet unknown, observation Y, which encodes the forecast $\mathbb{P}(Y \leq t) = F(t)$ whenever $t \in I$.

The family $\mathcal{F}$ of potential predictive distributions F that we consider is quite general. In practical settings, $\mathcal{F}$ includes those distributions having a probability density function (PDF), those having a discrete distribution, and those that are a mixture of the two. To be precise, let $\mathcal{F}(\mathbb{R})$ denote the class of probability measures on the Borel–Lebesgue sets of $\mathbb{R}$, and, for an interval $I \subset \mathbb{R}$, let $\mathcal{F}(I)$ denote the subset of probability measures on I. For simplicity, we do not distinguish between a measure F in $\mathcal{F}(\mathbb{R})$ and its associated cumulative density function (CDF) F. We follow standard conventions and assume that CDFs are right continuous. For F in $\mathcal{F}(I)$, write $Y \sim F$ to indicate that a random variable Y has distribution F; that is, $\mathbb{P}(Y \leq t) = F(t)$ whenever $t \in I$. Throughout this report, the notation $\mathbb{E}_F$ indicates that the expectation is taken with respect to $Y \sim F$.

Let $\mathbb{1}$ denote the indicator function, so that, for example, $\mathbb{1}_{\{x \geq y\}}$ equals 1 whenever $x \geq y$ and 0 otherwise. The power set of a set A will be denoted $\mathcal{P}(A)$. For a real-valued quantity X, we denote by X_+ the quantity $\max(0, X)$. The partial derivative with respect to the first argument of a function g is denoted $\partial_1 g$.

We will often use the 'capping function' κ. Whenever $a, b \in [0, \infty]$, define $\kappa_{a,b} : \mathbb{R} \to \mathbb{R}$ by

$$\kappa_{a,b}(x) = \max(\min(x, b), -a) \qquad \forall x \in \mathbb{R}\,.$$

That is, $\kappa_{a,b}(x)$ is x capped below by $-a$ and above by b. Note that $x_+ = \kappa_{0,\infty}(x)$.

3 Quantiles, expectiles and Huber functionals

Although a predictive distribution F of some unknown future quantity Y contains a wealth of information, in many contexts users or issuers of forecasts want a relevant point summary x of the predictive distribution. This can be generated by requesting a specific statistical *functional* of F (such as its mean, median or some specified quantile), or by specifying a scoring function S (such as the squared error scoring function) that will be used to evaluate forecasts. In this section we focus functionals, and in particular on three families of functionals: quantiles (which include the median), expectiles (which include the mean) and Huber functionals (which nests quantiles and expectiles as edge cases).

Given an interval $I \subseteq \mathbb{R}$ and some space $\mathcal{F}$ of probability distributions in $\mathcal{F}(I)$, a *statistical functional* (or simply a *functional*) T on $\mathcal{F}(I)$ is a mapping $\mathrm{T} : \mathcal{F}(I) \to \mathcal{P}(I)$ (e.g. [Horowitz and Manski, 2006], [Gneiting, 2011a]). Two important examples are quantiles and expectiles.

Example 3.1. Suppose that $I \subseteq \mathbb{R}$ and $\alpha \in (0, 1)$. The *α-quantile* functional $\mathrm{Q}^\alpha : \mathcal{F}(I) \to \mathcal{P}(I)$ is defined by

$$\mathrm{Q}^\alpha(F) = \{x \in I : \lim_{y \uparrow x} F(y) \leq \alpha \leq F(x)\}\,, \quad F \in \mathcal{F}(I)\,.$$

For any F, $\mathrm{Q}^\alpha(F)$ is a closed bounded interval of I. The two endpoints only differ when the level set $F^{-1}(\alpha)$ contains more than one point, so typically the functional is single valued. The median functional $\mathrm{Q}^{1/2}$ arises when $\alpha = 1/2$. If q is an α-quantile of F and F is continuous at q then $F(q)/(1 - F(q)) = \alpha/(1 - \alpha)$. Figure 1 illustrates the quantiles $\mathrm{Q}^{1/2}(F)$ (the median) and $\mathrm{Q}^{0.7}(F)$, where F is the exponential distribution. The aforementioned property is illustrated in the figure via the vertical dashed line segments, whose lengths are in the ratio $\alpha : (1 - \alpha)$.

Example 3.2. Given an interval $I \subseteq \mathbb{R}$, let $\mathcal{F}_1(I)$ denote the space of probability measures $\mathcal{F}(I)$ with finite first moment. The *α-expectile* functional $\mathrm{E}^\alpha : \mathcal{F}_1(I) \to \mathcal{P}(I)$ is defined by

$$\mathrm{E}^\alpha(F) = \left\{x \in I : \alpha \int_x^\infty (y - x)\,\mathrm{d}F(y) = (1 - \alpha) \int_{-\infty}^x (x - y)\,\mathrm{d}F(y)\right\}\,, \quad F \in \mathcal{F}_1(I)\,. \tag{3.1}$$

It can be shown there is a unique solution x to the defining equation, so expectiles are single valued. Expectiles were introduced by [Newey and Powell, 1987] in the context of least squares estimation and have recently attracted interest in financial risk management [Bellini and Di Bernardino, 2017]. Expectiles share properties of both expectations as well as quantiles, and nests the mean functional $\mathrm{E}^{1/2}$. Using integration by parts, one can show that $\{x\} = \mathrm{E}^\alpha(F)$ if and only if

$$\alpha \int_{[x,\infty) \cap I} (1 - F(t))\,\mathrm{d}t = (1 - \alpha) \int_{(-\infty,x] \cap I}^x F(t)\,\mathrm{d}t\,.$$

The latter equation gives a geometric interpretation of the α-expectile of F. It is the unique point x such that the $(1 - \alpha)$-weighted area of the region bounded by F and 0 on the interval $(-\infty, x] \cap I$ is equal to the α-weighted area of the region bounded by F and 1 on the interval $[x, \infty) \cap I$. Figure 1 illustrates this interpretation, via the areas of the shaded regions, for the expectiles $\mathrm{E}^{1/2}(F)$ (i.e. mean) and $\mathrm{E}^{0.7}(F)$, where F is the exponential distribution.

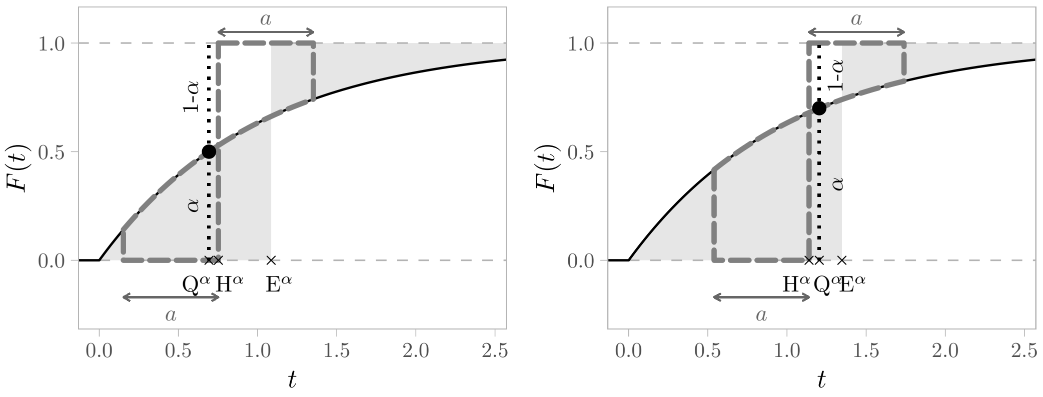

Figure 1: The quantile Q^α, expectile E^α and Huber quantile H^α (where $\mathrm{H}^\alpha = \mathrm{H}^\alpha_a(F)$, $a = 0.6$) when $\alpha = 0.5$ (left) and $\alpha = 0.7$ (right) for the exponential distribution $F(t) = 1 - \exp(-t)$, $t \geq 0$. The ratios of the areas of the two shaded regions, of the areas of the two regions bounded by thick dashed lines, and of the lengths of the two dotted line segments, are $\alpha : (1-\alpha)$.

Equation (3.1) can be re-written as

$$\mathrm{E}^\alpha(F) = \{x \in I : \alpha \mathbb{E}_F\, \kappa_{0,\infty}(Y - x) = (1-\alpha)\mathbb{E}_F\, \kappa_{0,\infty}(x - Y)\}\ . \tag{3.2}$$

By modifying the parameters of the capping function $\kappa_{0,\infty}$, we introduce a new functional.

Definition 3.3. Suppose that $a > 0$, $b > 0$, $\alpha \in (0,1)$ and that $I \subseteq \mathbb{R}$ is an interval. Then the *Huber functional* $\mathrm{H}^\alpha_{a,b} : \mathcal{F}(I) \to \mathcal{P}(I)$ is defined by

$$\mathrm{H}^\alpha_{a,b}(F) = \{x \in I : \alpha \mathbb{E}_F\, \kappa_{0,a}(Y - x) = (1-\alpha)\mathbb{E}_F\, \kappa_{0,b}(x - Y)\} \tag{3.3}$$

whenever $F \in \mathcal{F}(I)$. In the case when $a = b$, we simplify notation and write $\mathrm{H}^\alpha_a(F)$ for $H^\alpha_{a,a}(F)$. The special case $\mathrm{H}^{1/2}_a(F)$ is called a *Huber mean.*

The Huber functional is named after Peter Huber, whose loss function

$$h_a(u) = \begin{cases} \frac{1}{2}u^2\,, & |u| \leq a \\ a|u| - \frac{1}{2}a^2\,, & |u| > a \end{cases} \tag{3.4}$$

[Huber, 1964] now bears his name. The connection between the Huber functional and Huber loss will be made explicit in Section 4. Since the Huber functional is an example of a generalised quantile (see [Bellini et al., 2014], who follow [Breckling and Chambers, 1988]), $\mathrm{H}^\alpha_{a,b}(F)$ may also be called a *Huber quantile* of F. We note here that $x \in \mathrm{H}^\alpha_{a,b}(F)$ if and only if $\mathbb{E}_F V(x, Y) = 0$, where $V : I \times I \to \mathbb{R}$ is given by

$$V(x, y) = |\mathbb{1}_{\{x \geq y\}} - \alpha|\kappa_{a,b}(x - y)\,. \tag{3.5}$$

The function V is an *identification function* [Gneiting, 2011a, Section 2.4] for $H^\alpha_{a,b}$, and will be used to establish important properties of the Huber functional.

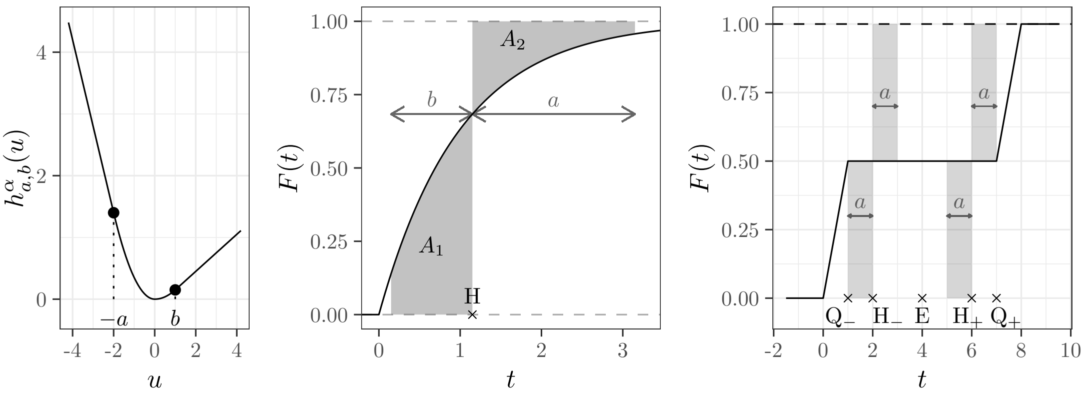

Figure 2: Left: Generalised Huber loss function $h^{\alpha}_{a,b}$ where $\alpha = 0.7$, $a = 2$ and $b = 1$. Centre: The Huber quantile $\mathrm{H} = \mathrm{H}^{\alpha}_{a,b}(F)$ where $\alpha = 0.7$, $a = 2$ and $b = 1$ for the exponential distribution $F(t) = 1 - \exp(-t)$, $t \geq 0$. The two shaded areas satisfy the equation $(1-\alpha)A_1 = \alpha A_2$. Right: A piecewise linear distribution F with endpoints H_- and H_+ of the interval $\mathrm{H}^{1/2}_a(F)$ where $a = 1$, endpoints Q_- and Q_+ of the median interval $\mathrm{Q}^{1/2}(F)$, and the mean value E. The area of each shaded rectangle is equal.

As with expectiles, a routine calculation using integration by parts shows that $x \in \mathrm{H}^{\alpha}_{a,b}(F)$ if and only if

$$\alpha \int_{[x,x+a]\cap I} (1 - F(t))\,\mathrm{d}t = (1-\alpha) \int_{[x-b,x]\cap I} F(t)\,\mathrm{d}t\,. \tag{3.6}$$

This gives a geometric interpretation of the Huber functional as the set of points x where the $(1-\alpha)$-weighted area of the region bounded by F and 0 on $[x-b,x]\cap I$ equals the α-weighted of the region bounded by F and 1 on $[x,x+a]\cap I$. In the case when $\alpha = 1/2$, the two areas are equal. This is illustrated for the exponential distribution in Figure 1 for $\mathrm{H}^{\alpha}_{0.6}(F)$ (when $\alpha = 1/2$ and $\alpha = 0.7$) and in Figure 2 for $\mathrm{H}^{\alpha}_{a,b}(F)$ (when $\alpha = 0.7$, $a = 2$ and $b = 1$).

In light of the corresponding geometric interpretations of quantiles and expectiles, and also the similarity between Equations (3.2) and (3.3), it should come as no surprise that quantiles and expectiles are nested as edge cases in the family H^{α}_{a} of Huber means. The following proposition makes this precise and lists several other basic properties of the Huber functional. In what follows, $\overline{F^{-1}(w)}$ denotes the closure in $\mathbb{R}$ of the level set $F^{-1}(w)$, and $R(F)$ denotes the smallest closed interval of $\mathbb{R}$ that contains the support of the measure F.

Proposition 3.4. *Suppose that $a > 0$, $b > 0$, $\alpha \in (0,1)$, $I \subseteq \mathbb{R}$ is an interval and $F \in \mathcal{F}(I)$.*

1. *Then $\mathrm{H}^{\alpha}_{a,b}(F)$ is a nonempty closed bounded subinterval of I contained in $R(F)$.*
2. *If $\mathrm{H}^{\alpha}_{a,b}(F) = [c,d]$ for some $c < d$, then there exists w in $(0,1)$ such that $\overline{F^{-1}(w)} = [c-b, d+a]$ and $\alpha = bw/(bw + a(1-w))$.*

3. *If there exists w in $(0,1)$ such that $\overline{F^{-1}(w)} = [c_0, d_0]$ for some c_0 and d_0 satisfying $d_0 - c_0 > a + b$, then $\mathrm{H}^\alpha_{a,b}(F) = [c_0 + b, d_0 - a]$ where $\alpha = bw/(bw + a(1-w))$.*

4. $\lim_{a\downarrow 0} \min(H^\alpha_a(F)) = \min(\mathrm{Q}^\alpha(F))$ *and* $\lim_{a\downarrow 0} \max(H^\alpha_a(F)) = \max(\mathrm{Q}^\alpha(F))$.

5. *If F has finite first moment then*

$$\lim_{a\to\infty} \min(H^\alpha_a(F)) = \lim_{a\to\infty} \max(H^\alpha_a(F)) = \mathrm{E}^\alpha(F)\,.$$

6. *If $\tilde{F} \in \mathcal{F}(I)$ and $F(t) = \tilde{F}(t)$ whenever $t \in [\min(\mathrm{H}^\alpha_{a,b}(F)) - b, \max(\mathrm{H}^\alpha_{a,b}(F)) + a]$, then $\mathrm{H}^\alpha_{a,b}(F) = \mathrm{H}^\alpha_{a,b}(\tilde{F})$.*

Part (1) is similar to [Bellini et al., 2014, Proposition 1(a)], whilst parts (4), (5) and (6) were noted, in the case of finite discrete distributions when $a = b$ and $\alpha = 1/2$, by [Huber, 1964]. The proof is given in the appendix.

Part (6) can be interpreted as saying that the Huber functional only depends on the values of the CDF F away from its tails. In situations where the tail of a predictive distribution is difficult to model, but a point summary describing its broad centre is desired, this property is useful. In particular, the Huber functional is invariant to the modification of F outside the interval $[\min(\mathrm{H}^\alpha_{a,b}(F)) - b, \max(\mathrm{H}^\alpha_{a,b}(F)) + a]$. In contrast, modification of the tails of F will generally change its mean and expectile values, whilst quantile values are invariant to modifications of F anywhere apart from at the quantile.

Parts (2) and (3) specify conditions on F for when $\mathrm{H}^\alpha_{a,b}(F)$ is multivalued. A corollary is that if each level set of F on $R(F)$ has length not exceeding $a+b$ then $\mathrm{H}^\alpha_{a,b}(F)$ is single valued for every α in $(0,1)$. Figure 2 illustrates a distribution F for which $\mathrm{H}^{1/2}_a(F)$ is multivalued whenever $0 < a < 3$. In this particular case, F has a symmetric bi-modal PDF, and also the property that $\mathrm{E}^{1/2}(F) \subset \mathrm{H}^{1/2}_a(F) \subset \mathrm{Q}^{1/2}(F)$ whenever $a > 0$.

Note that while $\mathrm{H}^\alpha_a(F)$ is in some sense an intermediary between $\mathrm{Q}^\alpha(F)$ and $\mathrm{E}^\alpha(F)$, the right-hand side of Figure 1 illustrates that the Huber quantile does not always lie between the corresponding quantile and expectile.

4 Scoring functions, consistency and elicitability

In this section we discuss scoring functions and their relationship to point forecasts and functionals. Two key concepts are those of consistency and elicitability. How these concepts relate to the Huber functional is the subject of Theorem 4.5, which is the first major theoretical result of this report.

4.1 Scoring functions and Bayes' rules

Definition 4.1. Suppose that $I \subseteq \mathbb{R}$. A function $S : I \times I \to \mathbb{R}^2$ is a called a *scoring function* if $S(x,y) \geq 0$ for all $(x,y) \in I \times I$ with $S(x,y) = 0$ whenever $x = y$. The scoring function S is said to be *regular* if (i) for each $x \in I$ the function $y \mapsto S(x,y)$ is measurable, and (ii) for each $y \in I$ the function $x \mapsto S(x,y)$ is continuous, with continuous derivative whenever $x \neq y$.

The score $S(x, y)$ can be interpreted as the loss or cost accrued when the point forecast x is issued and the observation y realises. Examples of scoring functions include the squared error scoring function $S(x, y) = (x - y)^2$, the absolute error scoring function $S(x, y) = |x - y|$ and the zero–one scoring function $S(x, y) = \mathbb{1}_{\{|x-y|\geq k\}}(x)$, for some positive k. Only first two of these are regular, whilst the zero–one scoring function fails to be regular on account of its discontinuity when $|x - y| = k$. The measurability condition (i) is a technical condition that is satisfied by most (if not all) scoring functions that arise in practice.

Huber loss (3.4) gives rise to the regular scoring function $S(x, y) = h_a(x - y)$. We introduce a more general version.

Definition 4.2. Suppose that $a > 0$, $b > 0$ and $\alpha \in (0, 1)$. The *generalised Huber loss function* $h^\alpha_{a,b} : \mathbb{R} \to \mathbb{R}$ is defined by

$$h^\alpha_{a,b}(u) = \begin{cases} |\mathbb{1}_{\{u\geq 0\}} - \alpha|\, \frac{1}{2}u^2\,, & -a \leq u \leq b \\ (1-\alpha)\, b(u - \frac{1}{2}b)\,, & u > b \\ -\alpha\, a(u + \frac{1}{2}a)\,, & u < -a\,. \end{cases}$$

The classical Huber loss function given by Equation (3.4) is $2h^{1/2}_{a,a}$. The same generalisation is used by [Zhao et al., 2019] for robust expectile regression. Figure 2 shows the graph of $h^{0.7}_{2,1}$. Note that $h^\alpha_{a,b}$ is differentiable on $\mathbb{R}$, with derivative

$$(h^\alpha_{a,b})'(u) = |\mathbb{1}_{\{u\geq 0\}} - \alpha|\kappa_{a,b}(u), \qquad u \in \mathbb{R}\,. \tag{4.1}$$

Generalised Huber loss gives rise to the regular scoring function $S(x, y) = h^\alpha_{a,b}(x - y)$.

Given a scoring function S, a forecast system that generates point forecasts can assessed by computing its mean score $\bar{S}$, where

$$\bar{S} = \frac{1}{n}\sum_{i=1}^{n} S(x_i, y_i)\,,$$

over a finite set of forecast cases $\{x_1, \ldots, x_n\}$ with corresponding observations $\{y_1, \ldots, y_n\}$. In this framework, if a number of competing forecast systems are being compared then the one with the lowest mean score is the best performer. Thus, given a scoring function S and predictive distribution F, an optimal point forecast is any $\hat{x}$ in I that minimises the expected score; that is,

$$\hat{x} = \arg\min_x \mathbb{E}_F S(x, Y)\,,$$

provided that the expectation exists. A point forecast that is optimal in this sense is also known as a *Bayes' rule* [Gneiting, 2011a, Ferguson, 1967].

It has long been known that the Bayes' rule under the squared error scoring function $S(x, y) = (x-y)^2$ is the mean of F, and under the absolute error scoring function $S(x, y) = |x - y|$ is any median of F. The Bayes' rule under the asymmetric piecewise linear scoring function

$$S(x, y) = |\mathbb{1}_{\{x\geq y\}} - \alpha||x - y| \tag{4.2}$$

is a quantile $\mathrm{Q}^\alpha(F)$ (e.g. [Ferguson, 1967]), whilst the Bayes' rule under the asymmetric quadratic scoring function

$$S(x, y) = |\mathbb{1}_{\{x\geq y\}} - \alpha|(x - y)^2 \tag{4.3}$$

is the expectile $\mathrm{E}^\alpha(F)$ [Newey and Powell, 1987, Gneiting, 2011a].

The Bayes' rule under the generalised Huber loss scoring function $S(x,y) = h^\alpha_{a,b}(x-y)$ can be found by looking for solutions x to the equation $\partial_1 \mathbb{E}_F S(x,Y) = 0$. If interchanging differentiation and integration can be justified then $\mathbb{E}_F \partial_1 S(x,Y) = 0$. Using Equation (4.1), one obtains $\mathbb{E}_F V(x,Y) = 0$, where V is the identification function given by (3.5). This implies that $x \in H^\alpha_{a,b}(F)$. That is, at least formally, the Bayes' rule under the generalised Huber loss scoring function is the corresponding Huber functional of F. A precise statement will be given in the next subsection.

4.2 Consistency and elicitability

Whenever a point forecast request specifies what functional of the predictive distribution is being sought, the scoring function used to evaluate the point forecast should be appropriate for that functional.

Definition 4.3. [Gneiting, 2011a, Murphy and Daan, 1985] Suppose that $I \subseteq \mathbb{R}$. A scoring function $S : I \times I \to \mathbb{R}$ is said to be *consistent* for the functional T relative to a class $\mathcal{F}$ of probability distributions on I if

$$\mathbb{E}_F S(t,Y) \leq \mathbb{E}_F S(x,Y) \tag{4.4}$$

for all probability distributions F in $\mathcal{F}$, all t in $\mathrm{T}(F)$ and all x in I. The functional T is said to be *strictly consistent* relative to the class $\mathcal{F}$ if it is consistent relative to the class $\mathcal{F}$ and if equality in (4.4) implies that $x \in \mathrm{T}(F)$.

Evaluating point forecasts with a strictly consistent scoring function rewards forecasters who give truthful point forecast quotes from carefully considered predictive distributions. This is because the requested functional of the predictive distribution coincides with the optimal point forecast (or Bayes' rule).

The families of consistent scoring functions for quantiles and expectiles each have a standard form. Subject to slight regularity conditions, a scoring function S is consistent for the quantile functional Q^α if and only if S is of the form

$$S(x,y) = |\mathbb{1}_{\{x \geq y\}} - \alpha||g(x) - g(y)|\,, \tag{4.5}$$

where g is a non-decreasing function [Gneiting, 2011b, Thomson, 1978, Saerens, 2000]. Moreover, if g is strictly increasing then S is strictly consistent. The standard asymmetric piecewise linear scoring function (4.2) for quantiles (which includes, up to a multiplicative constant, the absolute error scoring function for the median) is recovered from Equation (4.5) with the choice $g(x) = x$.

Subject to standard regularity conditions, a scoring function S is consistent for the expectile functional E^α if and only if S is of the form

$$S(x,y) = |\mathbb{1}_{\{x \geq y\}} - \alpha|\big(\phi(y) - \phi(x) + \phi'(x)(x-y)\big)\,, \tag{4.6}$$

where ϕ is a convex function with subgradient ϕ' [Gneiting, 2011a]. Moreover, if ϕ is strictly convex then S is strictly consistent. The standard asymmetric quadratic scoring function (4.3) for expectiles (including, up to a multiplicative constant, the squared error scoring function for the mean) is recovered from (4.6) by taking $\phi(x) = x^2$. When $\alpha = 1/2$, the function S of (4.6) is known as a *Bregman function.*

We will show that consistent scoring functions for the Huber functional also have a standard form. Before doing so, we introduce a critical concept related to the evaluation of point forecasts.

Definition 4.4. [Lambert et al., 2008] A statistical functional T is said to be *elicitable* relative to a class $\mathcal{F}$ of probability distributions if there exists a scoring function S that is strictly consistent for T relative to $\mathcal{F}$.

For example, quantiles are elicitable relative to the class $\mathcal{F}(\mathbb{R})$, while expectiles are elicitable relative to the class of distributions in $\mathcal{F}(\mathbb{R})$ with finite first moment [Gneiting, 2011a]. It is worth noting that some statistical functionals are not elicitable, including the sum of two distinct quantiles and *conditional value-at-risk*, a popular risk measure in finance [Gneiting, 2011a].

We turn now to the Huber functional. The main thrust (subject to appropriate regularity conditions) is that the Huber functional is elicitable, and that S is consistent for $\mathrm{H}^\alpha_{a,b}$ if and only if S is of the form

$$S(x,y) = |\mathbb{1}_{\{x\geq y\}} - \alpha|\big(\phi(y) - \phi(\kappa_{a,b}(x-y)+y) + \kappa_{a,b}(x-y)\phi'(x)\big)\,, \tag{4.7}$$

where ϕ is a convex function with subgradient ϕ'. Moreover, S is strictly consistent if ϕ is strictly convex. The generalised Huber loss scoring function $S(x,y) = h^\alpha_{a,b}(x-y)$ arises from Equation (4.7) with the choice $\phi(t) = t^2$. The following gives a precise statement.

Theorem 4.5. *Suppose that $I \subseteq \mathbb{R}$ is an interval and that $a > 0$, $b > 0$ and $\alpha \in (0,1)$.*

1. *The Huber functional $\mathrm{H}^\alpha_{a,b}$ is elicitable relative to the class of probability measures $\mathcal{F}(I)$ when I is bounded or semi-infinite, and elicitable relative to the class of probability measures $\mathcal{F}(I)$ with finite first moment when $I = \mathbb{R}$.*
2. *Suppose that $\phi : I \to \mathbb{R}$ is convex on I. Then the function $S : I \times I \to \mathbb{R}$, defined by Equation (4.7), is a consistent scoring function for the Huber functional $\mathrm{H}^\alpha_{a,b}$ relative to the class $\mathcal{F}(I)$ of probability measures F for which both $\mathbb{E}_F[\phi(Y) - \phi(Y-a)]$ and $\mathbb{E}_F[\phi(Y) - \phi(Y+b)]$ exist and are finite. If, additionally, ϕ is strictly convex then S is strictly consistent for $\mathrm{H}^\alpha_{a,b}$ relative to the same class of probability measures.*
3. *Suppose that the scoring function $S : I \times I \to \mathbb{R}$ is regular. If S is consistent for the Huber functional $\mathrm{H}^\alpha_{a,b}$ relative to the class of probability measures in $\mathcal{F}(I)$ with compact support, then S is of the form (4.7) for some convex function $\phi : I \to \mathbb{R}$. Moreover, if S is strictly consistent then ϕ is strictly convex.*

The proof is given in the appendix.

The general form (4.7) for the consistent scoring functions of the Huber functional yields, as edge cases, the general form for the consistent scoring functions of expectiles and quantiles. To be precise, let $S^{\mathrm{H},\phi}_a$ denote the scoring function S given by (4.7) when $a = b$, and let $S^{\mathrm{E},\phi}$ and $S^{\mathrm{Q},g}$ denote the consistent scoring functions of Equations (4.6) and (4.5) respectively. The relationship between $S^{\mathrm{H},\phi}_a$ and $S^{\mathrm{E},\phi}$ is straightforward: we have the pointwise limit

$$\lim_{a\to\infty} S^{\mathrm{H},\phi}_a(x,y) = S^{\mathrm{E},\phi}(x,y)\,. \tag{4.8}$$

For the other end of the spectrum we consider the rescaled consistent scoring function $S_a^{\mathrm{H},\phi}/a$, and obtain the pointwise limit

$$\lim_{a\downarrow 0} S_a^{\mathrm{H},\phi}(x,y)/a = S^{\mathrm{Q},\phi'}(x,y)\,, \tag{4.9}$$

where ϕ' is nondecreasing because ϕ is convex. Importantly, the relevant regularity conditions ensure that every non-decreasing function g in the representation (4.5) is the subderivative of some suitable convex ϕ.

The consistent scoring functions for the Huber functional thus show a mixture of the properties of the consistent scoring functions for quantiles and expectiles. Focusing on the functional $\mathrm{H}_a^{1/2}$ for positive a, the only consistent scoring function (up to a multiplicative constant) on $\mathbb{R}\times\mathbb{R}$ that only depends on the difference $x-y$ between the forecast and observation is the classical Huber loss scoring function $(x,y)\mapsto h_{a,a}^{1/2}(x-y)$. This is because the only Bregman function (up to a multiplicative constant) that has the same property for $\mathrm{E}^{1/2}$ is the squared error scoring function $(x,y)\mapsto(x-y)^2$ [Savage, 1971]. Hence, apart from multiples of classical Huber loss, other consistent scoring functions for $\mathrm{H}_a^{1/2}$ on $\mathbb{R}$ penalise under- and over-prediction asymmetrically. One such example is the exponential family

$$S_{\lambda;a}(x,y) = \begin{cases} \frac{1}{\lambda^2}\big(\exp(\lambda y)-\exp(\lambda x)\big) - \frac{1}{\lambda}\exp(\lambda x)(y-x)\,, & |x-y|\le a\\ \frac{1}{\lambda^2}\big(\exp(\lambda y)-\exp(\lambda(y+a))\big) + \frac{a}{\lambda}\exp(\lambda x)\,, & x-y>a\\ \frac{1}{\lambda^2}\big(\exp(\lambda y)-\exp(\lambda(y-a))\big) - \frac{a}{\lambda}\exp(\lambda x)\,, & x-y<-a\,, \end{cases} \tag{4.10}$$

parameterised by $\lambda\in\mathbb{R}$ and obtained from (4.7) via $\phi(t) = 2\exp(\lambda t)/\lambda^2$. These are analogous to the exponential family of Bregman functions considered by [Patton, 2020].

5 Mixture representations and Murphy diagrams

The main theoretical tool presented in this section is the mixture representation for consistent scoring functions of the Huber functional (Theorem 5.2). Mixture representations were introduced for quantiles and expectiles by [Ehm et al., 2016] and have several very useful applications, including providing insight into forecast rankings.

5.1 Ranking of forecasts

As mentioned in Section 4.1, point forecasts from two competing forecast systems A and B can be ranked by calculating their mean scores $\bar{S}_n^{\mathrm{A}}$ and $\bar{S}_n^{\mathrm{B}}$ over a finite number n of forecast cases for some scoring function S. If the forecast cases are independent, a statistical test for equal predictive performance can be based on the statistic t_n, where

$$t_n = \sqrt{n}\,\frac{\bar{S}_n^{\mathrm{A}}-\bar{S}_n^{\mathrm{B}}}{\hat{\sigma}_n} \quad\text{and}\quad \hat{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^{n}(S(x_i^{\mathrm{A}},y_i)-S(x_i^{\mathrm{B}},y_i))^2 \tag{5.1}$$

for forecasts $\{x_i^{\mathrm{A}}\}$ and $\{x_i^{\mathrm{B}}\}$ and corresponding realisations $\{y_i\}$. Corresponding p-values are computed and if the null hypothesis is rejected then A is preferred if $t_n<0$ and B is preferred otherwise [Gneiting and Katzfuss, 2014, Section 3.3]. Unfortunately, forecast rankings and the results of hypothesis tests can depend on the choice of consistent scoring function [Ehm et al., 2016, pp. 506, 515–516], as we now illustrate.

Example 5.1. Two forecast systems, BoM and OCF, produce point forecasts for the daily maximum temperature at Sydney Observatory Hill. The OCF system generates forecasts from a blend of bias-corrected numerical weather prediction forecasts. The BoM forecast is issued by meteorologists who have access to various information sources, including OCF. We consider forecasts for the period July 2018 to June 2020 with a lead time of one day. See Figure 3 for a sample time series of BoM and OCF forecasts with observations.

Suppose that these forecasts are targeting the Huber mean $\mathrm{H}_3^{1/2}$, and make the simplifying assumption that successive forecast cases are independent. If the consistent scoring function $S(x,y) = 2\,h_{3,3}^{1/2}(x-y)$ is used, then the mean score for BoM is lower than the mean score for OCF, and with a p-value of 6.52×10^{-4} the null hypothesis of equal predictive performance is rejected at the 5% significance level in favour of BoM forecasts. However, if the consistent scoring function $S_{2;3}$ defined by Equation (4.10) is used, then OCF has the lower mean score, albeit with a p-value of 0.333 that upholds the null hypothesis.

5.2 Mixture representations

In this subsection we state mixture representations for consistent scoring functions of the Huber functional. Practical applications of this theoretical result will follow in subsequent subsections.

In general, the choice of subgradient ϕ' in the representation (4.7) is not unique. To facilitate precise mathematical statements a special version of ϕ' will be chosen. Let $\mathcal{I}$ denote the class of all left-continuous non-decreasing functions on $\mathbb{R}$, and let $\mathcal{C}$ denote the class of all convex functions $\phi : \mathbb{R} \to \mathbb{R}$ with subgradient ϕ' in $\mathcal{I}$. This last condition will be satisfied if ϕ' is chosen to be the left-hand derivative of ϕ. Denote by $\mathcal{S}^{\mathrm{H}}_{\alpha,a,b}$ the class of scoring functions S of the form (4.7) such that $\phi \in \mathcal{C}$. For most practical purposes, $\mathcal{S}^{\mathrm{H}}_{\alpha,a,b}$ can be identified with the class consistent scoring functions for the Huber functional on $\mathbb{R}$.

Theorem 5.2. *Every member S of the class $\mathcal{S}^{\mathrm{H}}_{\alpha,a,b}$ has a representation of the form*

$$S(x,y) = \int_{-\infty}^{\infty} S^{\mathrm{H}}_{\alpha,a,b,\theta}(x,y)\,\mathrm{d}M(\theta), \qquad (x,y) \in \mathbb{R}^2\,, \tag{5.2}$$

where

$$S^{\mathrm{H}}_{\alpha,a,b,\theta}(x,y) = \begin{cases} (1-\alpha)\min(\theta - y, b) & \text{if } y \le \theta < x \\ \alpha \min(y-\theta, a) & \text{if } x \le \theta < y \\ 0 & \text{otherwise} \end{cases} \tag{5.3}$$

and M is a non-negative measure. The mixing measure is unique and satisfies $\mathrm{d}M(\theta) = \mathrm{d}\phi'(\theta)$ *whenever $\theta \in \mathbb{R}$, where ϕ' is the left-hand derivative of the convex function ϕ in the representation (4.7). Furthermore, $M(x) - M(y) = \partial_2 S(x,y)/(1-\alpha)$.*

The proof is given in the appendix and is a simple adaptation of the proof of the analogous results for expectiles and quantiles [Ehm et al., 2016, Theorem 1].

Each function $S^{\mathrm{H}}_{\alpha,a,b,\theta}$ of Theorem 5.2 is called an *elementary scoring function* for the Huber functional, and also belongs to $\mathcal{S}^{\mathrm{H}}_{\alpha,a,b}$ (use Equation (4.7) with the choice $\phi(t) = (t-\theta)_+$ and $\phi'(t) = \mathbb{1}_{\{\theta<t\}}$). So Theorem 5.2 essentially says that each consistent scoring function for the Huber functional can be expressed as a weighted average of elementary

scoring functions. The representation (5.2) holds pointwise via the same argument for expectiles [Ehm et al., 2016, p. 510]. Moreover, when $a = b$, the mixture representations for the consistent scoring functions of expectiles and quantiles emerge as edge cases of Theorem 5.2 by taking limits as $a \to \infty$ and as $a \downarrow 0$ and using the dominated convergence theorem. Details are given in Remark A.1.

5.3 Economic interpretation of elementary scoring functions

The elementary scoring functions $S^{\mathrm{H}}_{\alpha,a,b,\theta}$ admit an economic interpretation of the loss, relative to actions based on a perfect forecast, of an investment decision with fixed costs, differential tax rates for profits versus losses, and where profits and losses are capped. To illustrate, we give two examples. The first is an adaptation of the interpretation for the elementary scoring functions of expectiles [Ehm et al., 2016, p. 513]. The second illustrates how the Huber functional and its elementary scoring functions can arise in the context of investment decisions based on weather forecasts.

Example 5.3. Suppose that Alexandra considers investing a fixed amount θ in a start-up company in exchange for an unknown future amount y of the company's profits or losses. Additionally, Alexandra takes out an option to set a limit b on losses she could incur but which also imposes a limit a on the profits she could receive. Alexandra will make a profit if and only if $y > \theta$, and so adopts the decision rule to invest if and only if her point forecast x of y exceeds θ. Her pay-off structure is as follows:

1. If Alexandra refrains from the deal, her pay-off will be 0, independent of the outcome y.
2. If Alexandra invests and $y \leq \theta$ realises then her payout is negative at $-(1 - r_L)\min(\theta - y, b)$. Here $\min(\theta - y, b)$ is the monetary loss, bounded by b, and the factor $1 - r_L$ accounts for Alexandra's reduction in income tax with $r_L \in [0, 1)$ representing the deduction rate.
3. If Alexandra invests and $y > \theta$ realises then her pay-off is positive at $(1 - r_G)\min(y - \theta, a)$, where $r_G \in [0, 1)$ denotes the tax rate that applies to her profits.

The top matrix in Table 1 shows Alexandra's pay-off under her decision rule. The positively-oriented pay-off matrix can be reformulated as a negatively oriented regret matrix, by considering the difference between the pay-off for an (hypothetical) omniscient investor who has access to a perfect forecast and the pay-off for Alexandra. For example, if $x \leq \theta$ and $y > \theta$ realises, then the omniscient investor's pay-off is $(1 - r_G)\min(y - \theta, b)$ while Alexandra's pay-off is 0, and so Alexandra's regret is $(1 - r_G)\min(y - \theta, b)$. The bottom matrix of Table 1 is Alexandra's regret matrix, which up to a multiplication factor is the elementary score $S^{\mathrm{H}}_{\alpha,a,b,\theta}(x, y)$. So to minimise regret, Alexandra should invest if and only if $x > \theta$, where $x = \mathrm{H}^{\alpha}_{a,b}(F)$, F is Alexandra's predictive distribution of the future value of the investment and $\alpha = (1 - r_G)/(2 - r_L - r_G)$. The point forecast $x = H^{1/2}_{a}(F)$ arises if profits and losses are capped by the same value and if the rates r_G and r_L are equal.

Example 5.4. Hannah runs a business selling ice creams from a mobile cart at a sports stadium. Historically, there is an approximately linear relationship between the volume

Table 1: Overview of pay-off structure for Alexandra's decision rule to invest if and only if $x > \theta$.

	$y \leq \theta$	$y > \theta$
Monetary payoff		
$x \leq \theta$	0	0
$x > \theta$	$-(1-r_L)\min(\theta - y, b)$	$(1-r_G)\min(y-\theta, a)$
Score (regret)		
$x \leq \theta$	0	$(1-r_G)\min(y-\theta, a)$
$x > \theta$	$(1-r_L)\min(\theta - y, b)$	0

of ice cream sales on any given afternoon and the observed daily maximum temperature, so that the profit p from sales is modelled by $p = ky + c$, where y is the observed daily maximum temperature, $k > 0$ and $c \in \mathbb{R}$. Additionally, $0 \leq p \leq a$ for some positive a, since total sales are limited by cart capacity, while any unsold units can be sold at a later date. If Hannah chooses to sell ice creams on any given afternoon, she must also pay a fixed cost f (staff wages and stadium fees). If model assumptions are correct, Hannah will make a profit if and only if $ky + c > f$. So she adopts the decision rule to sell ice creams on any given afternoon if and only if her point forecast x of the maximum temperature exceeds the decision threshold θ, where $\theta = (f - c)/k$. Her pay-off structure is as follows.

1. If Hannah does not sell ice creams then her pay-off is 0.
2. If Hannah sells ice creams and $y > \theta$ then her profit after tax is $(1 - r_G)\min(ky + c - f, a - f)$, where $r_G \in [0, 1)$ denotes the tax rate. Her profit can be rewritten as $(1 - r_G)k\min(y - \theta, (a - f)/k)$.
3. If Hannah sells ice creams and $y < \theta$ then her loss after tax deductions is $(1 - r_L)\min(f - (ky + c), f)$, where $r_G \in [0, 1)$ denotes the deduction rate, and losses are capped by f since unsold ice creams go back into storage. Her loss can be rewritten as $(1 - r_L)k\min(\theta - y, f/k)$.

As with Example 5.3, these outcomes can be converted to a regret matrix, which up to a multiplication factor is the elementary score $S^{\mathrm{H}}_{\alpha,(a-f)/k,f/k,\theta}(x, y)$ where $\alpha = (1-r_G)/(2-r_L-r_G)$. Consequently, her optimal decision rule is to sell ice creams if and only if $x > \theta$, where $\theta = (f-c)/k$, $x \in \mathrm{H}^{\alpha}_{(a-f)/k,f/k}(F)$, F is her predictive distribution of the maximum temperature and $\alpha = (1 - r_G)/(2 - r_L - r_G)$.

The essential features of Example 5.4 also arise in the context of rainfall storage and water trading. Any profits made by selling harvested water are capped by storage capacity. The predicted volume v of water that is collected from any rainfall event can be modelled by $v = ky+c$, where y is the predicted rainfall at a representative point within the catchment, c is catchment initial loss and k is determined by catchment size and continuing loss.

5.4 Forecast dominance, Murphy diagrams and choice of consistent scoring function

We return to the problem of forecast rankings with the notion of *forecast dominance* [Ehm et al., 2016, Section 3.2]. We say that forecast system A *dominates* forecast system

B for point forecasts targeting a specific Huber functional if the expected score of point forecasts from A is not greater than the expected score of point forecasts from B, for every consistent scoring function. In practice this is impossible to check directly because the family of consistent scoring functions, parameterised by $\phi \in \mathcal{C}$, is very large. However, by the mixture representation of Theorem (5.2), one need only test for dominance over the family, parametrised by $\theta \in \mathbb{R}$, of elementary functions. In empirical situations, this is further reduced to checking forecast dominance for finitely many θ. In what follows, we consider tuples $(x_{i\mathrm{A}}, x_{i\mathrm{B}}, y_i)$ consisting of the ith point forecast from systems A and B along with the corresponding observation y_i.

Corollary 5.5. *Suppose that $\alpha \in (0,1)$, $a > 0$ and $b > 0$. The forecast system* A *empirically dominates* B *for predictions targeting* $\mathrm{H}^{\alpha}_{a,b}$ *if*

$$\frac{1}{n}\sum_{i=1}^{n} S^{\mathrm{H}}_{\alpha,a,b,\theta}(x_{i\mathrm{A}}, y_i) \le \frac{1}{n}\sum_{i=1}^{n} S^{\mathrm{H}}_{\alpha,a,b,\theta}(x_{i\mathrm{B}}, y_i)$$

whenever $\theta \in \bigcup\{x_{i\mathrm{A}}, x_{i\mathrm{B}}, y_i, y_i - a, y_i + b : 1 \le i \le n\}$ and in the left-hand limit as $\theta \uparrow \theta_0$, where $\theta_0 \in \bigcup\{x_{i\mathrm{A}}, x_{i\mathrm{B}}, : 1 \le i \le n\}$.

To see why, note that the score differential $\theta \mapsto d_i(\theta)$ for the ith forecast case is piecewise linear and right-continuous, and is zero unless θ lies between $x_{i\mathrm{A}}$ and $x_{i\mathrm{B}}$. The only possible discontinuities are at $x_{i\mathrm{A}}$ and $x_{i\mathrm{B}}$, and the only possible changes of slope are at y_i, $y_i - a$ and $y_i + b$.

An empirical check for forecast dominance is aided with the use of a *Murphy diagram* [Ehm et al., 2016, Section 3.3], which is a plot showing the graph of

$$\theta \mapsto \frac{1}{n}\sum_{i=1}^{n} S^{\mathrm{H}}_{\alpha,a,b,\theta}(x_i, y_i)$$

for each forecast source, computed at each of the points θ of Corollary 5.5. The top left of Figure 3 presents the Murphy diagram for three different forecasts targeting the Huber mean $\mathrm{H}^{1/2}_{3}$ of the daily maximum temperature at Sydney Observatory Hill (July 2018 to June 2020). The OCF and BoM forecasts were discussed in Example 5.1. For any given day, the Climate forecast is the mean of 46 observations, sampled from the previous 15 days and from a 31 day period this time last year centred on the day in question. A lower mean score is better.

The graph in the top right of Figure 3 represents forecast performance as a skill score with respect to two reference forecasts: the perfect forecast (skill score = 1) and the Climate forecast (skill score = 0). The difference in mean elementary scores between OCF and BoM forecasts is presented in the bottom left, with pointwise 95% confidence intervals. Neither of these forecasts dominates the other.

Returning to Example 5.4, if Hannah's decision rule is to sell ice creams if and only if the $\mathrm{H}^{1/2}_{3}$ point forecast x exceeds 30°C, then Hannah should base her decisions on the BoM forecast, since its mean elementary score, which is proportional to economic regret, is lowest (see the top left of Figure 3 where $\theta = 30$). But if her fixed investment costs f changed, then so would her decision threshold θ, and the Murphy diagram indicates which forecast system historically performed better at the new threshold.

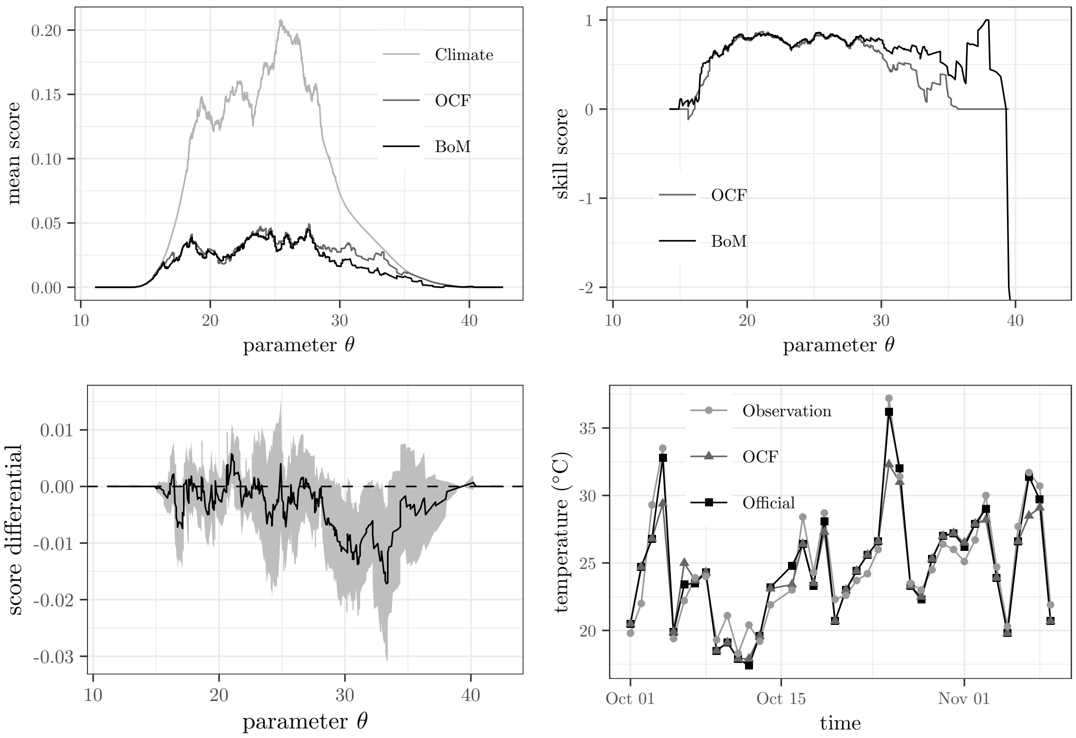

Figure 3: Competing forecast systems targeting the Huber mean $\mathrm{H}^{1/2}_{3,3}$ for the daily maximum temperature at Sydney Observatory Hill (July 2018 to June 2020). Top left: Murphy diagram of mean elementary scores. Top right: Murphy diagram of elementary skill scores. Bottom left: mean elementary score difference of OCF and BoM with pointwise 95% confidence intervals (less than 0 indicates that BoM is preferable). Bottom right: Sample of the forecast–observation time series.

The mixture representation and Murphy diagram also gives insight into why the two different scoring functions of Example 5.1 lead to different forecast rankings. The classical Huber loss scoring function $S(x,y) = 2\,h^{1/2}_{3,3}(x-y)$ is obtained from Equation (4.7) with the choice $\phi(t) = t^2$. The corresponding mixing measure is $\mathrm{d}M(\theta) = 2\,\mathrm{d}\theta$, implying that every elementary scoring function in the mixture representation (5.2) is weighted equally, and also that the area underneath each graph in the Murphy diagram (top left of Figure 3) is twice the mean Huber loss $\bar{S}$ for that forecast system. On the other hand, the exponential scoring function $S_{2;3}$ is obtained from Equation (4.7) with the choice $\phi(t) = \exp(2t)/2$. In this case $\mathrm{d}M(\theta) = 2\exp(2\theta)\,\mathrm{d}\theta$ and so mean elementary scores in the corresponding mixture representation are weighted heavily for higher values of θ. Hence when scored by $S_{2;3}$, a slight over-forecast of 40.4°C by BoM on 19 December 2019 (OCF forecast 35.4°C and the observation was 39.3°C) was penalised substantially more heavily than the OCF under-forecast, resulting in a higher mean score $\bar{S}_{2;3}$ for BoM than OCF.

Finally, we consider the choice of consistent scoring function for Huber quantile point predictions in the situation where the point forecast serves the needs of a diverse community of users. Classical Huber loss, obtained when $\phi(t) = t^2$, applies equal weight to all θ. For everyday use, this choice of ϕ may be justified by the desire to weight all decision thresholds θ equally. On the other hand, for weather forecasts there may be a desire, from a public risk perspective, to give greater weight to values of θ that lie in the hazardous climatological extremes, so that competing forecast system candidates are evaluated with that in mind. For maximum temperature forecasts, the mixing measure $\mathrm{d}M(\theta) = \phi''(\theta)\,\mathrm{d}\theta$, where

$$\phi''(\theta) = \begin{cases} (5-\theta)+1\,, & \theta \le 5 \\ 1\,, & 5 < \theta < 35 \\ (\theta - 35)+1\,, & \theta \ge 35\,, \end{cases}$$

puts increasing weight on decision thresholds below 5°C and above 35°C. This yields the convex function

$$\phi(\theta) = \begin{cases} \frac{1}{6}(5-\theta)^3 + \frac{1}{2}\theta^2\,, & \theta \le 5 \\ \frac{1}{2}\theta^2\,, & 5 < \theta < 35 \\ \frac{1}{6}(\theta - 35)^3 + \frac{1}{2}\theta^2\,, & \theta \ge 35\,. \end{cases}$$

and the corresponding consistent scoring function S can be computed from Equation (4.7). With this S, BoM maximum temperature forecasts for Sydney outperform those of OCF, and with a p-value of 1.48×10^{-3} the null hypothesis of equal predictive performance is rejected at the 5% significance level.

6 Robust verification of point forecasts for the mean

In parametric estimation, there is often a trade-off between robustness of and bias in the estimate, particularly in the presence of contaminated data [Huber, 1964]. These tensions also arise in the evaluation of point forecasts. The squared error scoring function, whilst consistent for point forecasts targeting the mean functional, is sensitive to large measured errors, which may be due to faulty measurements rather than poor forecasts. We therefore explore the use of Huber loss as a robust scoring function.

Three families of distributions will be used to generate synthetic data: the Normal distribution $\mathcal{N}(\mu, \sigma^2)$, the skew normal distribution $\mathcal{SN}(\xi, \omega, \nu)$ with location ξ, scale ω and shape ν, and the Beta distribution $\mathcal{B}(r, s)$ with shape parameters r and s. The skew normal distribution is right skewed if $\nu > 0$, normal if $\nu = 0$ and left skewed if $\nu < 0$. We also use the notion that a predictive distribution is *ideal* relative to an information set if it makes the best possible use of that information [Gneiting and Katzfuss, 2014, pp. 129–130].

Example 6.1. Suppose that the daily maximum temperature Y at a given location has conditional distribution $Y|(\xi, \omega, \nu) \sim \mathcal{SN}(\xi, \omega, \nu)$, where

$$\begin{aligned} &\xi \sim \mathcal{SN}(19, 6, 20) \\ &\omega \sim 1.4 + B_1 \max(20, \xi)/10\,, && B_1 \sim \mathcal{B}(2, 5) \\ &\nu \sim 40B_2 - 20\,, && B_2 \sim \mathcal{B}(1.5, 1.5\max(20, \xi)/20)\,. \end{aligned}$$

The PDF of Y is the uncontaminated distribution in the top left of Figure 4. This set-up has the following interpretation. On any given day, the maximum temperature has a skew normal distribution. Right and left skewed distributions are both possible, but if the temperature is likely to be high then odds favour a left skewed distribution. The

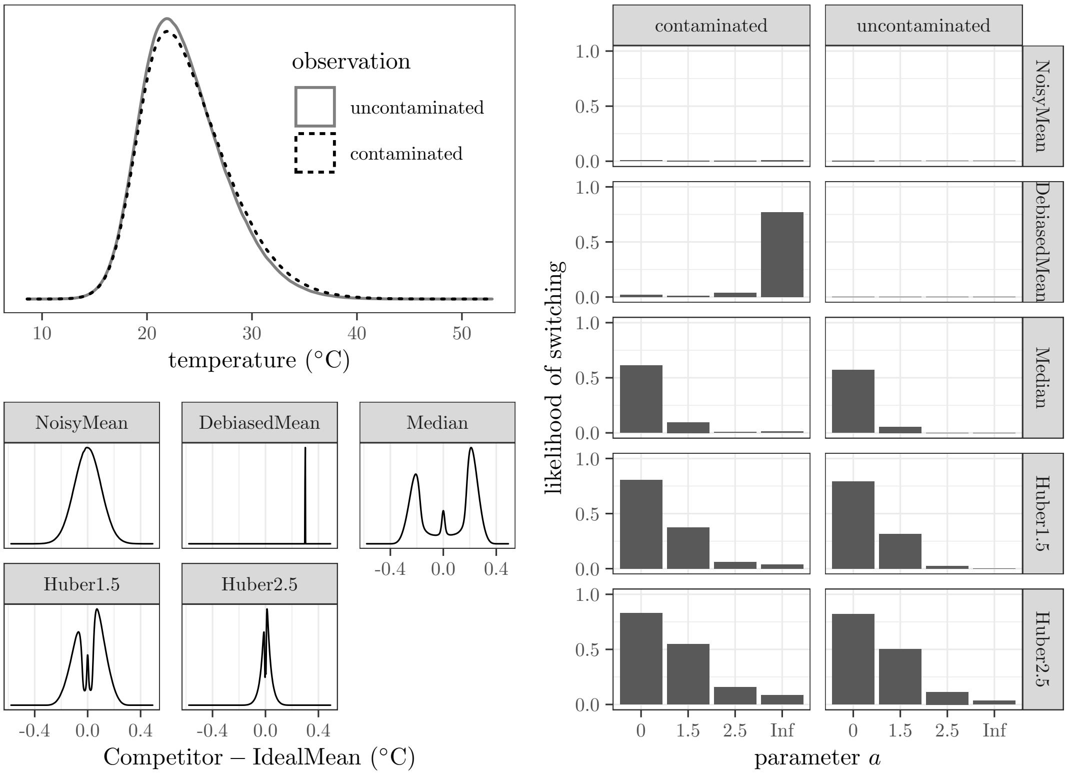

Figure 4: Top left: PDFs of uncontaminated observations Y and contaminated observations Y^{m}. Bottom left: PDFs of the difference between IdealMean and competitor forecasts. Right: The likelihood that Harry will switch from IdealMean to a competitor forecast, depending on the observation type and scoring function $S_a(x,y) = h^{1/2}_{a,a}(x-y)$ (squared loss when $a=\infty$ and absolute loss when $a=0$).

distribution on any given day also tends to be sharper for lower temperatures. This models the maximum temperature distribution for a site slightly inland of the New South Wales coast, where variability in the arrival of the sea breeze increases maximum temperature volatility during warmer months.

In addition to uncontaminated realisations Y we consider measured observations Y^{m}, of which 5% are contaminated by a measurement 'spike' of at least +5°C. To be precise, $Y^{\mathrm{m}} = Y + 5UV$, where $\mathbb{P}(U=1) = 0.05$ (the contamination rate), $\mathbb{P}(U=0) = 0.95$ and V is the exponential distribution with CDF $t \mapsto 1 - \exp(-0.8t)$, $t \geq 0$. The PDF of Y^{m} is also shown in Figure 4.

Harry pays Nick a handsome sum for accurate point forecasts of maximum temperature targeting the mean functional. Nick has access to the parameter set (ξ, ω, ν) and, for any given day, forecasts the ideal predictive distribution $F \sim \mathcal{SN}(\xi, \omega, \nu)$ relative to that information set and quotes its mean $\mathrm{E}^{1/2}(F)$. This is the *IdealMean* forecast. Five other forecasters have covertly obtained access to Nick's predictive distribution. One forecaster attempts to disguise blatant plagiarism by adding random noise to Nick's quote (this is the *NoisyMean* forecast $E^{1/2}(F) + \mathcal{N}(0, 0.5)$). Another forecaster compares Nick's point forecasts to a sample of (contaminated) observations, and issues the *DebiasedMean*

$E^{1/2}(F) + 0.3$ (which is close to the optimal correction 0.314). Three other forecasters suspect that Harry is aware that the observations are contaminated and may rank forecasts with a scoring function that is less sensitive to outliers. One quotes the *Median* $\mathrm{Q}^{1/2}(F)$, another the *Huber1.5* $\mathrm{H}^{1/2}_{1.5}(F)$ and the third the *Huber2.5* $\mathrm{H}^{1/2}_{2.5}(F)$. The PDFs of their difference from IdealMean is shown in the bottom left of Figure 4. Note that DebiasedMean is furthest, on average, from the ideal forecast.

Nick's contract with Harry is about to expire. Harry will switch from the IdealMean to a competitor if the null hypothesis that IdealMean is at least as good as the competitor is rejected at the 5% significance level, using a one-tailed test with the statistic of Equation (5.1) for two years of forecast data. To estimate the likelihood that the null hypothesis is rejected, we generate 4000 samples of 365×2 independent (Y, Y^{m}, F) tuples, and determine the test outcome for each sample. Results are obtained for both uncontaminated and contaminated observations, and for four different scoring functions $S_a(x, y) = 2\,h^{1/2}_{a,a}(x-y)$, where $a \in \{0, 1.5, 2.5, \infty\}$ and the cases $a = 0$ and $a = \infty$ are the absolute error and squared error scoring functions respectively. The likelihood of Harry rejecting the null hypothesis (i.e. switching) is shown in the right of Figure 4.

The results show that for uncontaminated observations, the likelihood of switching from IdealMean to a competitor is low when using the squared error scoring function ($a = \infty$), as one should expect since it is strictly consistent for the mean. However, the likelihood of switching to DebiasedMean (the competitor which is furthest from the ideal) is 77% when using the squared error scoring function with the contaminated data set. Of the four scoring functions considered, the one with parameter $a = 2.5$ performs best because the likelihood of switching from IdealMean is generally low for both observation sets across all competitors. Nonetheless, the chance of switching to Huber2.5 whilst using this scoring function with contaminated data is 16%. This risk may be tolerable since Huber2.5 is, on average, the competing forecast that is closest to the IdealMean.

We conclude this section with a general discussion of issues faced by a practitioner who wants to use generalised Huber loss for robust evaluation of point forecasts. In situations where forecasts targeting the mean are being evaluated with contaminated data, a good choice of tuning parameter a for the scoring function $S(x, y) = h^{1/2}_{a,a}(x - y)$ depends on the contamination rate ϵ, contamination distribution and the distribution of true errors of all competing forecasts. In practice, these are not precisely known, and if the estimated distribution of true errors is substantially asymmetric then using the classical Huber loss scoring is likely to be suboptimal. In the case where the distribution of true errors is approximately symmetric, empirical investigation of the author suggests that a be chosen such that (i) the bulk of (true) errors lies in $[-a, a]$ (say, a is larger than some robust measure of error spread), and (ii) so that a is somewhat less than the magnitude of typical contamination spikes (which may be estimated using knowledge of the measurement process).

We illustrate these recommendations using the simulation of Example 6.1. For (i), the absolute median deviation of forecast errors $x-y$ from a random sample of 2×365 forecast–observation cases (using contaminated data) ranged from 1.19 (for Median forecasts) to 1.33 (for IdealMean and DebiasedMean forecasts). Since absolute median deviation is a robust measure of spread, these results may be taken as proxies for the spread of the true errors. So in this example, it is recommended that a is at least 1.33. For (ii), suppose that some investigation of a sample of large errors suggest that observations spikes exceeding 6 occur sporadically. Then it is recommended that a is chosen to be somewhat less than 6.

We now turn attention to robust verification of forecasts targeting the α-expectile. For simplicity, suppose that contaminating measurement spikes are positive (as in Example 6.1), that $\alpha > 1/2$ and that the asymmetric quadratic scoring function (4.3) is used to assess forecast performance. In this case, apparent under-prediction forecast cases are penalised more heavily than apparent over-prediction forecast cases, and so forecasts systems that have a suitable over-prediction bias (relative to true realisations) will benefit from contaminated data. This leads to a more pronounced distortion of forecast rankings for higher values of α compared with case $\alpha = 1/2$. Unpublished empirical experiments of the author are consistent with this. However, this is not easily remedied. An alternative scoring function is the generalised Huber loss scoring function $S(x, y) = h^{\alpha}_{a,b}(x-y)$. Three parameters must be specified and it is not necessarily the case that the α parameter specifying of the target expectile should equal the α parameter of generalised Huber loss scoring function. Moreover, it is harder to distinguish statistically between the small proportion of true observations that correspond to mild under-prediction forecast cases and a similar size of contaminated observations that give the appearance of mild under-prediction.

Finally, we note that rankings of forecasts that target the Huber functional $\mathrm{H}^{\alpha}_{a,a}$ (when scored by $S(x, y) = h^{\alpha}_{a,b}(x - y)$) are less prone to distortion from contaminated data than forecasts targeting the expectile E^{α} (when scored by the asymmetric quadratic scoring function). The risk of distortion decreases as $a \downarrow 0$, and the situation with forecasts targeting Q^{α} is even more robust. Unpublished empirical experiments of the author are consistent with this.

Further light could be cast on the problem of tuning parameter selection and robust forecast evaluation by using insights from the theory of robust parameter estimation.

7 Conclusion

We have defined the Huber functional in such a way that it is the set of optimal point forecasts for minimising the expected score under the generalised Huber loss scoring function. The Huber functional is an intermediary between quantiles and expectiles, which it nests as edge cases. The Huber functional incorporates more information about a predictive distribution F than quantiles, yet unlike expectiles it is not sensitive to the behaviour of F at its tails. We have shown that the Huber functional is elicitable, given a characterisation of its consistent scoring functions and stated the mixture representation for those scoring functions. These theoretical results enable the use of the Huber functional and its associated consistent scoring functions within a theoretically sound framework for point forecasting and evaluation (see [Gneiting, 2011a], [Gneiting and Katzfuss, 2014], [Ehm et al., 2016] and the references therein). Moreover, the Huber functional is shown to arise naturally within decision theory for a broad class of investment problems, and within this context the mixture representation facilitates some justification for the choice of consistent scoring function when point forecasts targeting the Huber functional are utilised by a heterogeneous user group. Finally, we have shown that Huber loss can be used as a robust scoring function of forecasts targeting the mean in situations with contaminated observational data, noting that rankings of forecasts that target the Huber mean are more resilient to distortion in such situations.

Many organisations, including meteorological agencies, have traditionally issued point forecasts that are not well-defined, and that are consumed by a very broad user group. Where there is appetite to clarify forecast definitions, and where it is desirable that point forecasts target some ‘middle’ point of the predictive distribution, the Huber mean provides a good candidate functional, as it can be scored using a consistent scoring function that is more robust to contaminated data than the mean, and it incorporates more information from the predictive distribution than the median. The classical Huber loss scoring function $S(x, y) = h^{1/2}_{a,a}(x - y)$ is a natural choice for a consistent scoring function of the Huber mean, as it favours all user-decision thresholds equally (in the sense discussed in Section 5.4). Nonetheless, if it is desirable that forecast performance at some user-decision thresholds is more important than at others, the mixture representation provides a method for generating the appropriate scoring function.

Acknowledgements

The author would like to thank Jonas Brehmer and Professor Tilmann Gneiting for their interest in this work and for several very helpful comments. Deryn Griffiths, Robert Fawcett and Nicholas Loveday gave helpful feedback which improved the quality of this report. I also thank my family for their support as I wrote this manuscript. Finally, an expression of gratitude to Harry Jack, who introduced me to Huber loss, and to Michael Foley, who three years ago made the decision to start scoring Bureau of Meteorology temperature forecasts using Huber loss. I would not have otherwise embarked on this study.

A Proofs

Proof of Proposition 3.4. We first prove the proposition for the case when $I = \mathbb{R}$.

In light of the essential equivalence of Equations (3.3) and (3.6), define $G_{a,b} : \mathbb{R} \to \mathbb{R}$ by

$$G_{a,b}(u) = (1-\alpha)\int_{u-b}^{u} F(t)\,\mathrm{d}t - \alpha\int_{u}^{u+a}(1-F(t))\,\mathrm{d}t, \qquad u \in \mathbb{R}. \tag{A.1}$$

Where then is no confusion, we will drop the subscripts and simply be denote the function by G. Since the CDF F is nonnegative and nondecreasing, it follows that G is a nondecreasing function on $\mathbb{R}$. Moreover, G is also continuous on $\mathbb{R}$.

First we will show that the set of zeroes of G is nonempty and lies in $R(F)$, which will establish that $\mathrm{H}^{\alpha}_{a,b}(F)$ is a nonempty subset of $R(F)$. Since G is continuous and nondecreasing on $\mathbb{R}$, it suffices to show that that G takes at least one positive and one negative value in any neighbourhood of $R(F)$, which will also establish that the zero set is bounded.

Suppose that $\varepsilon > 0$. If $R(F)$ has finite left-endpoint r_0 then

$$\begin{aligned} G(r_0 - \varepsilon) &= -\alpha\int_{r_0-\varepsilon}^{r_0-\varepsilon+a}\mathrm{d}t + \alpha\int_{r_0}^{r_0-\varepsilon+a}F(t)\,\mathrm{d}t \\ &\le -\alpha a + \alpha(a+\varepsilon) \\ &< 0\,. \end{aligned}$$

Otherwise, let $\eta = \alpha a/((1-\alpha)b + \alpha a)$ and note that $0 < \eta < 1$. So there exists v in $R(F)$ such that $F(u) < \eta$ whenever $u \leq v$. So

$$G(v-a) < (1-\alpha)b\eta + \alpha a\eta - \alpha a = 0\,.$$

Similarly, if $R(F)$ has a finite right-endpoint r_1 then $G(r_1+\varepsilon) > 0$. Otherwise, there exists w in $R(F)$ such that $F(u) > \eta$ whenever $u \geq w$. So

$$G(w+b) > (1-\alpha)b\eta + \alpha a\eta - \alpha a = 0\,.$$

This shows that G has at least one zero, and since ε is arbitrary and G is nondecreasing and continuous, all the zeroes are contained in the interval $R(F)$, and the zero set is a closed bounded interval.

To prove parts (2) and (3) when $I = \mathbb{R}$, we note that the zero set of G is $[c,d]$ where $c < d$ only if there is a constant $w \in (0,1)$ such that $F(t) = w$ whenever $t \in \{\bigcup((u-b,u)\cup(u,u+a)) : u \in [c,d]\}$. The closure of this latter set is precisely $[c-b, d+a]$. Moreover, if u is any such zero of G, (A.1) implies that

$$0 = G(u) = (1-\alpha)bw + \alpha aw - \alpha w.$$

Rearranging gives $\alpha = bw/(bw + a(1-w))$ as required.

To prove part (4), fix α and F. Define q_0 and q_1 by

$$q_0 = \min(\mathrm{Q}^\alpha(F)) \qquad \text{and} \qquad q_1 = \max(\mathrm{Q}^\alpha(F))\,.$$

Suppose that $a > 0$. Denoting $\lim_{y\uparrow x} F(y)$ by $F(x^-)$, note that

$$F(q_i^-) \leq \alpha \leq F(q_i)\,, \quad F(q_0 - a) < \alpha \quad \text{and} \quad F(q_1 + a) > \alpha\,. \tag{A.2}$$

Therefore

$$\begin{aligned} G_{a,a}(q_0 - a) &\leq (1-\alpha)aF(q_0-a) + \alpha aF(q_0) - \alpha a \\ &< a\alpha(F(q_0) - F(q_0-a)) + a\alpha - \alpha a \\ &\leq 0\,, \end{aligned}$$

and similarly,

$$G_{a,a}(q_0 + a) \geq (1-\alpha)\alpha a + \alpha^2 a - \alpha a = 0\,.$$

This shows that $q_0 - a < \min(\mathrm{H}_a^\alpha(F)) \leq q_0 + a$, from which is obtained $\lim_{a\downarrow 0}\min(\mathrm{H}_a^\alpha(F)) = q_0$. Similarly, one can show that $G_{a,a}(q_1+a) > 0$ and $G_{a,a}(q_1 - a) \leq 0$, which are used to establish $\lim_{a\downarrow 0}\max(\mathrm{H}_a^\alpha(F)) = q_1$.

Part (5) follows from the definition of expectiles and the Huber functional.

To prove part (6), let $\tilde{G}_{a,b}$ denote the function of the form (A.1) defined using $\tilde{F}$ in the place of F, and suppose that $F = \tilde{F}$ on the interval $[\min(\mathrm{H}_{a,b}^\alpha(F)) - b, \max(\mathrm{H}_{a,b}^\alpha(F)) + a]$. If $x \in \mathrm{H}_{a,b}^\alpha(F)$ then $G_{a,b}(x) = 0$ and hence $\tilde{G}_{a,b}(x) = 0$, whence $\mathrm{H}_{a,b}^\alpha(F) \subseteq \mathrm{H}_{a,b}^\alpha(\tilde{F})$. The reverse inclusion is obtained similarly.

Finally, for each part, the case when $I \subset \mathbb{R}$ can be deduced from the case when $I = \mathbb{R}$ by considering the natural extension of $F \in \mathcal{F}(I)$ to $\mathcal{F}(\mathbb{R})$, and using the fact that $H_{a,b}^\alpha(F) \subseteq R(F) \subseteq I$. □

Proof of Theorem 4.5. To prove part (2), we take a similar approach to the proof presented in [Brehmer, 2017, pp. 38–39] (which follows [Gneiting, 2011a]) of the consistency theorem for expectiles. Suppose that $F \in \mathcal{F}(I)$ and that the expectations $\mathbb{E}_F[\phi(Y) - \phi(Y-a)]$ and $\mathbb{E}_F[\phi(Y) - \phi(Y+b)]$ both exist and are finite, which will guarantee the existence of the expectations that follow. Fix a and b in $(0,\infty)$ and α in $(0,1)$. For convenience, denote $\kappa_{a,b}$ by κ. Consider $t \in \mathrm{H}^{\alpha}_{a,b}(F)$ and let $x \in \mathbb{R}$. Suppose that ϕ is convex and let S be defined by (4.7). We need to show that

$$\mathbb{E}_F S(x,Y) - \mathbb{E}_F S(t,Y) \geq 0. \tag{A.3}$$

Define the function $g : I \times I \to \mathbb{R}$ by

$$g(u,v) = \phi(v) - \phi(u) - \phi'(u)(v-u)$$

and note that g is nonnegative by the convexity of ϕ, and strictly positive if ϕ is strictly convex. Define the function $f : I \times I \to \mathbb{R}$ by

$$f(u,v) = \phi'(u) - \phi'(v)\,,$$

and note that $f(u,v) \geq 0$ whenever $u \geq v$ by the convexity of ϕ, with $f(u,v) > 0$ whenever $u > v$ if ϕ is strictly convex.

To show (A.3), we break it up into two main cases (either $x < t$ or $t < x$) and then into several sub-cases. Consider first the case where $x < t$ with subcase $x - b < x < x + a \leq t - b < t < t + a$. Define the sets A_i, where $i \in \{1, 2, ..., 7\}$, by $A_1 = \{Y \in (\infty, x-b) \cap I\}$, $A_2 = \{Y \in [x-b, x] \cap I\}$, $A3 = \{Y \in (x, x+a] \cap I\}$, $A_4 = \{Y \in (x+a, t-b) \cap I\}$, $A_5 = \{Y \in [t-b, t] \cap I\}$, $A_6 = \{Y \in (t, t+a] \cap I\}$ and $A_7 = \{Y \in (t+a, \infty) \cap I\}$. Note that the sets A_i are disjoint and that their union is I. Hence

$$\mathbb{E}_F S(x,Y) - \mathbb{E}_F S(t,Y) = \mathbb{E}_F (S(x,Y) - S(t,Y)) \sum_{i=1}^{7} \mathbb{1}_{A_i}$$

We will calculate each term in the series and sum them together at the end. The calculations are:

$$\begin{aligned}
&\mathbb{E}_F(S(x,Y) - S(t,Y))\mathbb{1}_{A_1} \\
&\quad = (1-\alpha) f(x,t) \mathbb{E}_F \kappa(t-Y) \mathbb{1}_{A_1}\,, \\
&\mathbb{E}_F(S(x,Y) - S(t,Y))\mathbb{1}_{A_2} \\
&\quad = (1-\alpha)\mathbb{E}_F\big(g(x, Y+b) + f(x,t)\kappa(t-Y)\big)\mathbb{1}_{A_2}\,, \\
&\mathbb{E}_F(S(x,Y) - S(t,Y))\mathbb{1}_{A_3} \\
&\quad = \mathbb{E}_F\big(\alpha S(x,Y) - (1-\alpha)S(t,Y)\big)\mathbb{1}_{A_3} \\
&\quad = \alpha \mathbb{E}_F g(x,Y)\mathbb{1}_{A_3} + (1-\alpha)\mathbb{E}_F\big(g(Y, Y+b) + bf(Y,x) + f(x,t)\kappa(t-Y)\big)\mathbb{1}_{A_3}\,, \\
&\mathbb{E}_F(S(x,Y) - S(t,Y))\mathbb{1}_{A_4} \\
&\quad = \mathbb{E}_F\big(\alpha S(x,Y) - (1-\alpha)S(t,Y)\big)\mathbb{1}_{A_4} \\
&\quad = \alpha \mathbb{E}_F\big(g(Y-a, Y) + f(Y-a, x)\big)\mathbb{1}_{A_4} \\
&\qquad + (1-\alpha)\mathbb{E}_F\big(g(Y, Y+b) + bf(Y,x) + f(x,t)\kappa(t-Y)\big)\mathbb{1}_{A_4}\,, \\
&\mathbb{E}_F(S(x,Y) - S(t,Y))\mathbb{1}_{A_5}
\end{aligned}$$

$$
\begin{aligned}
&= \mathbb{E}_F\big(\alpha S(x,Y) - (1-\alpha)S(t,Y)\big)\mathbb{1}_{A_5} \\
&= \alpha\mathbb{E}_F\big(g(Y-a,Y) + f(Y-a,x)\big)\mathbb{1}_{A_5} \\
&\quad + (1-\alpha)\mathbb{E}_F\big(g(Y,t) + (t-Y)f(Y,x) + f(x,t)\kappa(t-Y)\big)\mathbb{1}_{A_5}\,, \\
&\mathbb{E}_F(S(x,Y) - S(t,Y))\mathbb{1}_{A_6} \\
&= \alpha\mathbb{E}_F\big(g(Y-a,t) + (t-Y+a)f(Y-a,x) + f(x,t)\kappa(t-Y)\big)\mathbb{1}_{A_6}\,, \\
&\mathbb{E}_F(S(x,Y) - S(t,Y))\mathbb{1}_{A_7} \\
&= \alpha f(x,t)\mathbb{E}_F\kappa(t-Y)\mathbb{1}_{A_7}\,.
\end{aligned}
$$

Now when summing these terms together, note that since $t \in \mathrm{H}^\alpha_{a,b}(F)$, Equation (3.3) implies that

$$
(1-\alpha)\mathbb{E}_F\kappa(t-Y)\mathbb{1}_{A_1\cup A_2\cup A_3\cup A_4\cup A_5} + \alpha\mathbb{E}_F\kappa(t-Y)\mathbb{1}_{A_6\cup A_7} = 0 \tag{A.4}
$$

and thus the all terms containing $f(x,t)$ vanish. The remaining terms are all nonnegative by the properties of f and g, which establishes (A.3) in this particular subcase and hence that S is consistent for $\mathrm{H}^\alpha_{a,b}$.

To prove strict consistency in this subcase, suppose that ϕ is strictly convex and that equality holds in (A.3). So we must have

$$
\begin{aligned}
0 &= \mathbb{E}_F(S(x,Y) - S(t,Y))\sum_{i=1}^{7}\mathbb{1}_{A_i} \\
&= (1-\alpha)\mathbb{E}_F g(x,Y+b)\mathbb{1}_{A_2} + (1-\alpha)\mathbb{E}_F g(Y,Y+b)\mathbb{1}_{A_3} + \alpha\mathbb{E}_F g(Y-a,Y)\mathbb{1}_{A_4} \\
&\quad + \alpha\mathbb{E}_F g(Y-a,Y)\mathbb{1}_{A_5} + \alpha\mathbb{E}_F g(Y-a,t)\mathbb{1}_{A_6} + K\,,
\end{aligned}
$$

where K can be written as a sum of nonnegative terms, having applied (A.4). Each of the terms in the final expression is nonnegative, so for equality to hold they must all equal 0. Now the terms involving A_3, A_4 and A_5 are all strictly positive unless $\mathbb{P}(Y \in A_i) = 0$ for $i = 3,4,5$. Similarly, the terms involving A_2 and A_6 are positive unless $\mathbb{P}(Y \in A_2\backslash\{x-b\}) = \mathbb{P}(Y \in A_6\backslash\{t+a\}) = 0$. Together, this implies that $\mathbb{P}(Y \in (x-b,t+a)\cap I) = 0$, or equivalently that F is constant on $(x-b,t+a)\cap I$. Combining this with the fact that $t \in \mathrm{H}^\alpha_{a,b}(F)$ if and only if (3.6) holds, it is easy to see that $x \in \mathrm{H}^\alpha_{a,b}(F)$. This establishes strict consistency.

For the main case $x < t$, there are four further subcases:

$$
\begin{aligned}
&x-b < x \le t-b < x+a \le t < t+a \\
&x-b < t-b < x \le x+a < t < t+a \\
&x-b < t-b < x < t < x+a < t+a \\
&x-b < x \le t-b < t \le x+a < t+a\,.
\end{aligned}
$$

The proof of consistency for each subcase proceeds in the same way as the first subcase, and if proceeding in this order most of the calculations in subcases that have already been proved can be used to prove subsequent subcases. The proof of strict consistency also proceeds similarly for the first case, by showing that F is constant on $(x-b,t+a)\cap I$. Details are left to the reader.

The case when $t < x$ is proved the same way, but calculations are quicker by exploiting symmetry and anti-symmetry. For example, the subcase

$$t - b < t < t + a \le x - b < x < x + a$$

proceeds by switching the roles of t and x in the definitions of A_i, and then making the switches $-a \leftrightarrow b$ and $A_i \leftrightarrow A_{8-i}$ in the calculations for each term. For example, in the case when $t > x$ we have $A_6 = \{Y \in (t, t+a] \cap I\}$ and

$$\begin{aligned}&\mathbb{E}_F(S(x,Y) - S(t,Y))\mathbb{1}_{A_6}\\&\quad= \alpha\mathbb{E}_F\big(g(Y-a,t) + (t-Y+a)f(Y-a,x) + f(x,t)\kappa(t-Y)\big)\mathbb{1}_{A_6}\,,\end{aligned}$$

while in the case when $x < t$, after making switches, we have $A_2 = \{Y \in (t-b, t] \cap I\}$ and

$$\begin{aligned}&\mathbb{E}_F(S(x,Y) - S(t,Y))\mathbb{1}_{A_2}\\&\quad= \alpha\mathbb{E}_F\big(g(Y+b,t) + (t-Y-b)f(Y+b,x) + f(x,t)\kappa(t-Y)\big)\mathbb{1}_{A_2}\,.\end{aligned}$$

All the terms are nonnegative apart from those involving $f(x,t)$, which will vanish when all the terms are summed together. Details are left to the reader.

The case when $t < x$ is proved the same way, but calculations are quicker by exploiting symmetry and anti-symmetry with previously proved subcases. Details are left to the reader. This completes the proof of part (2).

To prove part (1) for the cases when I is bounded or semi-finite, use the result of part (2) with the bounded (on I) strictly convex function $\phi(t) = e^{-t}$ (or $\phi(t) = e^t$ if I is the of the form $(-\infty, c)$). When $I = \mathbb{R}$, use the same approach with $\phi(t) = t^2$ and note that $\mathbb{E}_F[\phi(Y) - \phi(Y-a)]$ and $\mathbb{E}_F[\phi(Y) - \phi(Y+b)]$ exists and is finite if $\mathbb{E}_F Y$ exists and is finite.

To prove part (3), we apply Osband's principle with the identification function V of Equation (3.5). An argument similar to [Gneiting, 2011a, p. 753, 759] shows that

$$\partial_1 S(x,y) = h(x)V(x,y)$$

for $x, y \in I$ and some function $h : I \to I$, where ∂_1 denotes partial differentiation with respect to the first variable of the function. Integration by parts yields the representation (4.7), where the function ϕ is defined by

$$\phi(x) = \int_{x_0}^{x}\int_{x_0}^{v} h(u)\,\mathrm{d}u\,\mathrm{d}v$$

for some x_0 in I. Now since $S(x,y) \ge 0$ for all $x, y \in I$, it follows from (4.7) that $(x-y)\phi'(x) + \phi(y) - \phi(x) \ge 0$ whenever $-a \le x - y \le b$, which in turn implies that ϕ is convex on I. If S is strictly consistent, then $S(x,y) > 0$ for all non-identical x and y in I, whence a similar argument shows that ϕ is strictly convex. □

Proof of Theorem 5.2. Suppose that $a > 0$, $b > 0$ and $\phi \in \mathcal{C}$. Define the function $\Phi : \mathbb{R}^2 \to \mathbb{R}$ by

$$\Phi(x,y) = \phi(y) - \phi(\kappa_{a,b}(x-y) + y) + \kappa_{a,b}(x-y)\phi'(x)\,, \qquad x, y \in \mathbb{R}\,.$$

We will show that

$$\Phi(x,y) = 2\int_{-\infty}^{\infty} S^{\mathrm{H}}_{1/2,a,b,\theta}(x,y)\,\mathrm{d}\phi'(\theta)\,, \tag{A.5}$$

from whence follows the mixture representation (5.2), the fact that $\mathrm{d}M(\theta) = \mathrm{d}\phi'(\theta)$ and the relationship $M(x) - M(y) = \partial_2 S(x,y)/(1-\alpha)$ whenever $x > y$.

To show (A.5), we break into five cases. For the case $x - y < -a$,

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

The case $x - y > b$ is handled analogously. The case $-a \le x - y < 0$ is essentially the same as the proof of the case $x < y$ for expectiles [Ehm et al., 2016, p. 529], and the case $0 < x - y \le b$ is analogous. The final case $x = y$ is trivial.

Finally, note that the increments of M are determined by S and so the mixing measure is unique. □

Remark A.1. We show how the mixture representations for the consistent scoring functions of quantiles and expectiles [Ehm et al., 2016, Theorem 1] emerge as limiting cases of Theorem 5.2. Let $S^{\mathrm{E}}_{\alpha,\theta}$ denote the elementary scoring function for E^{α} with parameter θ, and let $S^{\mathrm{Q}}_{\alpha,\theta}$ denote the elementary scoring function for Q^{α} with parameter θ [Ehm et al., 2016]. Consider the case for expectiles first. For fixed x, y and θ we have

$$S^{\mathrm{E}}_{\alpha,\theta}(x,y) = \lim_{a\to\infty} S^{\mathrm{H}}_{\alpha,a,a,\theta}(x,y)\,.$$

Using the notation and limits following the statement of Theorem 4.5,

$$\begin{aligned}
S^{\mathrm{E},\phi}_{\alpha}(x,y) &= \lim_{a\to\infty} S^{\mathrm{H},\phi}_{\alpha,a}(x,y) \\
&= \lim_{a\to\infty} \int_{-\infty}^{\infty} S^{\mathrm{H}}_{\alpha,a,a,\theta}(x,y)\,\mathrm{d}M(\theta) \\
&= \int_{-\infty}^{\infty} S^{\mathrm{E}}_{\alpha,\theta}(x,y)\,\mathrm{d}M(\theta)\,,
\end{aligned}$$

where the interchange of limits and integration in the final equality is justified by the dominated convergence theorem and where $\mathrm{d}M(\theta) = \mathrm{d}\phi'(\theta)$. This recovers the mixture representation for expectiles. Turning now to quantiles, for fixed x, y and θ we have

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

and so $\lim_{a\downarrow 0}\frac{1}{a}S^{\mathrm{H}}_{\alpha,a,a,\theta}(x,y) = S^{\mathrm{Q}}_{\alpha,\theta}(x,y)$ for almost every θ (differing only when $\theta = y$). Hence, using the notation and limits following Theorem 4.5 and the dominated convergence theorem,

$$\begin{aligned} S^{\mathrm{Q},\phi'}_{\alpha}(x,y) &= \lim_{a\downarrow 0}\tfrac{1}{a}S^{\mathrm{H},\phi}_{\alpha,a}(x,y) \\ &= \lim_{a\downarrow 0}\int_{-\infty}^{\infty}\tfrac{1}{a}S^{\mathrm{H}}_{\alpha,a,a,\theta}(x,y)\,\mathrm{d}M(\theta) \\ &= \int_{-\infty}^{\infty}S^{\mathrm{Q}}_{\alpha,\theta}(x,y)\,\mathrm{d}M(\theta)\,, \end{aligned}$$

where $\mathrm{d}M(\theta) = \mathrm{d}\phi'(\theta)$. This recovers the mixture representation for quantiles.

References

[Bellini and Di Bernardino, 2017] Bellini, F. and Di Bernardino, E. (2017). Risk management with expectiles. *The European Journal of Finance*, 23(6):487–506.

[Bellini et al., 2014] Bellini, F., Klar, B., Müller, A., and Gianin, E. R. (2014). Generalized quantiles as risk measures. *Insurance: Mathematics and Economics*, 54:41–48.

[Breckling and Chambers, 1988] Breckling, J. and Chambers, R. (1988). M-quantiles. *Biometrika*, 75(4):761–771.

[Brehmer, 2017] Brehmer, J. (2017). Elicitability and its application in risk management. *arXiv preprint arXiv:1707.09604*.

[Ehm et al., 2016] Ehm, W., Gneiting, T., Jordan, A., and Krüger, F. (2016). Of quantiles and expectiles: consistent scoring functions, choquet representations and forecast rankings. *J. R. Statist. Soc. B*, 78:505–562.

[Ferguson, 1967] Ferguson, T. S. (1967). Probability and mathematical statistics.

[Gneiting, 2011a] Gneiting, T. (2011a). Making and evaluating point forecasts. *Journal of the American Statistical Association*, 106(494):746–762.

[Gneiting, 2011b] Gneiting, T. (2011b). Quantiles as optimal point forecasts. *International Journal of forecasting*, 27(2):197–207.

[Gneiting and Katzfuss, 2014] Gneiting, T. and Katzfuss, M. (2014). Probabilistic forecasting. *Annual Review of Statistics and Its Application*, 1:125–151.

[Horowitz and Manski, 2006] Horowitz, J. L. and Manski, C. F. (2006). Identification and estimation of statistical functionals using incomplete data. *Journal of Econometrics*, 132(2):445–459.

[Huber, 1964] Huber, P. J. (1964). Robust estimation of a location parameter. *Annals of Mathematical Statistics*, 35:73–101.

[Lambert et al., 2008] Lambert, N. S., Pennock, D. M., and Shoham, Y. (2008). Eliciting properties of probability distributions. In *Proceedings of the 9th ACM Conference on Electronic Commerce*, pages 129–138.

[Murphy and Daan, 1985] Murphy, A. H. and Daan, H. (1985). Forecast evaluation. In Murphy, A. H. and Katz, R. W., editors, *Probability, Statistics, and Decision Making in the Atmospheric Sciences*, pages 379–437. Westview Press, Boulder, CO.

[Newey and Powell, 1987] Newey, W. K. and Powell, J. L. (1987). Asymmetric least squares estimation and testing. *Econometrica: Journal of the Econometric Society*, pages 819–847.

[Patton, 2020] Patton, A. J. (2020). Comparing possibly misspecified forecasts. *Journal of Business & Economic Statistics*, 38(4):796–809.

[Saerens, 2000] Saerens, M. (2000). Building cost functions minimizing to some summary statistics. *IEEE Transactions on neural networks*, 11(6):1263–1271.

[Savage, 1971] Savage, L. J. (1971). Elicitation of personal probabilities and expectations. *Journal of the American Statistical Association*, 66(336):783–801.

[Thomson, 1978] Thomson, W. (1978). Eliciting production possibilities from a well-informed manager.

[Zhao et al., 2019] Zhao, J., Yan, G., and Zhang, Y. (2019). Robust estimation and shrinkage in ultrahigh dimensional expectile regression with heavy tails and variance heterogeneity. *arXiv preprint arXiv:1909.09302.*