University of British Columbia
Department of Statistics
Technical Report #262
December 2010

"An invariant loss function for quantile approximation, estimation and summarizing data"

by

Reza Hosseini
University of British Columbia
reza1317@gmail.com

Department of Statistics, University of British Columbia

ABSTRACT

This paper develops a loss function to assess the goodness of approximation or estimation of quantiles of a distribution (or data). We propose one that is invariant under monotonic transformations and we show by examples why this property is desirable in applications, in particular for making decisions that are invariant under (non-linear) change of scale of data. We show that the sample version of this loss function tends uniformly to the distributional version. This loss function can also be used to find optimal ways to summarize data (specially massive data), find equivariant ways to estimate quantiles and to define a measure of distance among random variables. We also show the usefulness of this loss function in interpreting results for quantiles. For example we show that even if the quantiles are not equivariant under increasing transformations and in fact the transformed quantile can be arbitrarily far from the quantile of the transformed random variable in terms of typical losses (such as absolute value), the distance is zero using this loss. It is also discussed how this loss function can be extended to multi-dimensions and statistics of data.

Keywords: Loss function; quantiles; invariant; large datasets; estimation; approximation

1 Introduction

This paper develops a "loss function" to assess the goodness of an approximation or an estimator of quantiles of a distribution (or a data vector). Suppose a quantile of a very large data vector, q is approximated by $\hat{q}$. Several classic losses can be considered. For example: absolute error $L(q, \hat{q}) = |q - \hat{q}|$ or squared error $L(q, \hat{q}) = (q - \hat{q})^2$ which was proposed by Gauss. Quoting from [6]: "Gauss proposed the square of the error as a measure of loss or inaccuracy. Should someone object to this specification as arbitrary, he writes, he is in complete agreement. He defends his choice by an appeal to mathematical simplicity and convenience." An obvious problem with this loss is its lack of invariance under (possibly non-linear) re-scaling of data. We propose a loss function that is invariant under strictly monotonic transformations. We also show that the sample version of this loss function tends uniformly to the distributional version. This loss function can be used also to find optimal ways to summarize a data vector and to define a measure of distance among random variables as discussed here.

We define the loss of estimating/approximating q by $\hat{q}$ to be the probability that the random variable falls in between the two values. A limited version of this concept only for data vectors can be found in computer science literature, where ϵ-approximations are used to approximate quantiles of large datasets (see for example [7]). However, this concept has not been introduced as a measure of loss and the definition is limited to data vectors rather than arbitrary distributions.

Since we use quantiles in this paper, we give a definition (slightly different from the customary one) and a lemma that gives the elementary properties. The traditional definition of quantiles for a random variable X with distribution function F,

$$lq_X(p) = \inf\{x | F(x) \geq p\},$$

appears in classic works as [8]. We call this the "left quantile function". In some books (e.g. [9]) the quantile is defined as

$$rq_X(p) = \sup\{x | F(x) \leq p\},$$

this is what we call the "right quantile function". Also in robustness literature people talk about the upper and lower medians which are a very specific case of these definitions. [2] considers both definitions, explore their relation and shows that considering both has several advantages. He also proves the following lemma regarding the properties of the quantiles.

Lemma 1.1 *(Quantile Properties Lemma) Suppose X is a random variable on the probability space (Ω, Σ, P) with distribution function F:*

a) $F(lq_F(p)) \geq p$.

b) $lq_F(p) \leq rq_F(p)$.

c) $p_1 < p_2 \Rightarrow rq_F(p_1) \leq lq_F(p_2)$.

d) $rq_F(p) = \sup\{x | F(x) \leq p\}$.

e) $P(lq_F(p) < X < rq_F(p)) = 0$. *i.e.* F *is flat in the interval* $(lq_F(p), rq_F(p))$.

f) $P(X < rq_F(p)) \leq p$.

g) If $lq_F(p) < rq_F(p)$ *then* $F(lq_F(p)) = p$ *and hence* $P(X \geq rq_F(p)) = 1 - p$.

h) $lq_F(1) > -\infty, rq_F(0) < \infty$ *and* $P(rq_F(0) \leq X \leq lq_F(1)) = 1$.

i) $lq_F(p)$ *and* $rq_F(p)$ *are non-decreasing functions of* p.

j) If $P(X = x) > 0$ *then* $lq_F(F(x)) = x$.

k) $x < lq_F(p) \Rightarrow F(x) < p$ *and* $x > rq_F(p) \Rightarrow F(x) > p$.

For continuous variable [2] shows:

Lemma 1.2 *(Continuous distributions inverse) If* F *is continuous* $F(x) = p \Leftrightarrow x \in [lq_X(p), rq_X(p)]$.

Section 2 introduces the probability loss for data vectors by showing the motivation of the definition in terms of the "degree of separation" between data points in a sorted vector. Section 3 extends the definition to distribution functions and shows the elementary properties of this loss function under strictly increasing or decreasing transformations. This section also contains examples to show the usefulness of this loss specially in taking decisions that are invariant of the scale of data. Section 4 shows that the sample version of probability loss tends to the distribution version when the sample size goes to infinity which is an easy consequence of Glivenko-Cantelli Theorem. Section 5 shows the desirable properties of the probability loss when the underlying distribution function is continuous. For example in that case the probability loss satisfies the triangular inequality. Section 6 interprets many results about the quantiles and sample quantiles using the probability loss. For example even though the sample quantiles are not almost surely convergent to the distribution quantiles, they converge to the distribution quantile in terms of the probability loss. Also it is shown that even though quantiles of a distribution are not equivariant under strictly monotonic transformations, the probability loss of the transformed quantile and the quantile of the transformed random variable is zero. Section 7 studies how large the probability loss between two quantiles or a

group of consequent quantiles can become when we change the underlying distribution function or data vector. This is useful when one wants to create quantile data summaries and so on using the probability loss as shown in [2] (Chapter 8). Section 8 introduces the "penalized" probability loss which is non-zero whenever the two values differ. It studies its properties and shows it satisfies the triangular inequality for appropriate penalties and also shows the choice of the penalty is not very influential in most cases. Section 9 discusses possible extensions of probability loss to multi-dimensions and statistics. Section 10 shows the applications of the probability loss in handling large data sets, summarizing them and inferring about their "exact quantiles" in the presence of missing values or contaminated data. It also shows how the probability loss can be used to approximate quantiles of large data sets specially when sorting the whole data set is not possible and the sorting can be done for smaller partitions. Section 11 shows how the probability loss can be used in estimating parameters (quantiles) of distributions in a framework similar to Wald decision theoretic approach with the desirable property that the estimators are equivariant under changes of scale of data (even non-linear). Finally Section 12 discusses other applications of the probability loss, for example in defining a measure of distance among random variables.

2 Probability loss for data vectors

Our purpose is to find good approximations to the median and other quantiles. We need a method to asses such approximations. We contend that such a method should not depend on the scale of the data. In other words it should be invariant under monotonic transformations. We define a function δ that measures a natural "degree of separation" between data points of a data vector x. For the sake of illustration, consider the example $sort(x) = (1, 2, 3, 3, 4, 4, 4, 5, 6, 6, 7)$. Now suppose, we want to define the degree of separation of 3,4 and 7 in this example. Since 4 comes right after 3, we consider their degree of separation to be zero. There are 3 elements between 4 and 7 so it is appealing to measure their degree of separation as 3 but since the degree of separation should be relative, we also divide by $n = 11$, the length of the vector, and get: $\delta(4, 7) = 3/11$. We can generalize this idea to get a definition for all pairs in $\mathbb{R}$. With the same example, suppose we want to compute the degree of separation between 2.5 and 4.5 that are not members of the data vector. Then since there are 5 elements of the data vector between these two values, we define their degree of separation as 5/11. More formally, we give the following definition.

Definition 2.1 *Suppose $x = (x_1, \cdots, x_n)$, a data vector and $z < z'$ let $\Delta_x(z, z') = \{i | z < x_i < z', i = 1, \cdots, n\}$. Then we define*

$$\delta_x(z, z') = \frac{|\Delta_x(z, z')|}{n},$$

and $\delta_x(z,z) = 0$, *where* $|\Delta_x(z,z')|$ *is the cardinality of* $\Delta_x(z,z')$. *We call* δ_x *the "degree of separation" (DOS) or the "probability loss" associated with* x.

We then have the following lemma about the properties of δ.

Lemma 2.1 *The degree of separation* δ_x *has the following properties:*

a) $\delta_x \geq 0$.

b) $y < y' < y'' \Rightarrow \delta_x(y,y'') \geq \delta_x(y,y')$.

c) $\delta_{\phi(x)}(\phi(z),\phi(z')) = \delta_x(z,z')$ *if* ϕ *is a strictly monotonic transformation.*

d) $y = sort(x)$ *and* $y_i < y_j \Rightarrow \delta_x(y_i,y_j) \leq (j-i-1)/n$.

Proof Both a) and b) are straightforward. To show (c), suppose $z < z'$ and ϕ is strictly decreasing. (The strictly increasing case is similar.) Then $\phi(z') < \phi(z)$ and hence

$$\Delta_{\phi(x)}(\phi(z),\phi(z')) = \{i|\phi(z') < \phi(x_i) < \phi(z)\} = \{i|z < x_i < z'\} = \Delta_x(z,z').$$

Finally d) is true because $|\Delta_x(y_i,y_j)| = |\{l|y_i < x_l < y_j, l = 1,\cdots,n\}| \leq j-i-1$. ■

Remark. The definition and results above can be applied to random vectors $S = (X_1,\cdots,X_n)$ as well. In that $\delta_S(z,z')$ is random. To develop our theory, we need to study the asymptotic behavior of these statistics. We do so in later sections.

3 Probability loss for distributions

We define a degree of separation for distributions which corresponds to the notion of probability defined for data vectors to measure separation between data points.

Definition 3.1 *Suppose* X *has a distribution function* F. *Let*

$$\delta_F(z',z) = \delta_F(z,z') = \lim_{u \to z^-} F(u) - F(z') = P(z' < X < z), \quad z > z',$$

and $\delta_F(z,z) = 0$, $z \in \mathbb{R}$. *We also denote this by* δ_X *whenever a random variable* X *with distribution* F *is specified. We call* δ_X *the "degree of separation" or the "probability loss" associated with* X.

The following lemma is a straightforward consequence of the definition.

Lemma 3.1 *Suppose* $x = (x_1,\cdots,x_n)$ *is a data vector with the empirical distribution* F_n. *Then*

$$\delta_{F_n}(z,z') = \delta_x(z,z'), \ z,z' \in \mathbb{R}.$$

This lemma implies that to prove a result about the degree of separation of data vectors, it suffices to show the result for the degree of separation of random variables.

Theorem 3.1 *Let X, Y be random variables and F_X, F_Y, their corresponding distribution functions.*
a) Assume $Y = \phi(X)$, for a strictly increasing or decreasing function $\phi : \mathbb{R} \to \mathbb{R}$. Then $\delta_{F_X}(z, z') = \delta_{F_Y}(\phi(z), \phi(z'))$, $z < z' \in \mathbb{R}$.
b) $\delta_{F_X}(z, z') \leq \delta_{F_X}(z, z'')$, $z \leq z' \leq z''$.
c) $\delta_{F_X}(z_1, z_3) \leq \delta_{F_X}(z_1, z_2) + \delta_{F_X}(z_2, z_3) + P(X = z_2)$.
d) Suppose, $p \in [0, 1]$. Then $\delta_{F_X}(lq_{F_X}(p), rq_{F_X}(p)) = 0$.
e) Suppose, $p_1 < p_2 \in [0, 1]$. Then $\delta_{F_X}(lq_{F_X}(p_1), rq_{F_X}(p_2)) \leq p_2 - p_1$.

Remark. We may restate Part (c), for data vectors: Suppose x has length n and z_2 is of multiplicity m, (which can be zero). Then the inequality in (c) is equivalent to $\delta_x(z_1, z_3) \leq \delta_x(z_1, z_2) + \delta_x(z_2, z_3) + m/n$.

Proof
a) Note that for a strictly increasing function ϕ, we have

$$P(z < X < z') = P(\phi(z) < \phi(X) < \phi(z')).$$

Now suppose ϕ is strictly decreasing. Then $z < z' \Rightarrow \phi(z') < \phi(z)$. Let $Y = \phi(X)$. Then

$$\delta_X(z, z') = P(z < X < z') = P(\phi(z') < \phi(X) < \phi(z)) = \delta_Y(\phi(z), \phi(z')).$$

b) This is trivial.
c) Consider the case $z_1 < z_2 < z_3$. (The other cases are easier to show.) Then

$$\delta_{F_X}(z_1, z_3) = P(z_1 < X < z_3) = P(z_1 < X < z_2) + P(X = z_2) + P(z_2 < X < z_3)$$

$$= \delta_{F_X}(z_1, z_2) + \delta_{F_X}(z_2, z_3) + P(X = z_2).$$

d) This result is a straightforward consequence of Lemma 1.1 b) and c).
e) This result follows from

$$\delta_{F_X}(lq(p_1), rq(p_2)) = P(lq(p_1) < X < rq(p_2))$$

$$= P(X < rq(p_2)) - P(X \leq lq(p_1)) \leq p_2 - p_1.$$

The last inequality being a result of Lemma 1.1 a) and d). ■

Remark: (e),(b) immediately imply

$$\delta_{F_X}(lq_{F_X}(p_1), lq_{F_X}(p_2)) \leq p_2 - p_1,$$

and

$$\delta_{F_X}(rq_{F_X}(p_1), lq_{F_X}(p_2)) \leq p_2 - p_1.$$

Remark. We call Part c) of the above theorem the pseudo-triangle inequality.

Here we give two examples about using the probability loss function and its interpretation.

Example 1:

We showed above that the triangle property does not hold for the probability loss function and that might lead to the criticism that this definition is not intuitively appealing. By an example, we now show why it makes sense that the triangle property should not hold for such a situation. Suppose a few mathematicians are standing in a line

Euclid, Khawarzmi, Khayyam, Gauss, Von Neumann.

If we were to ask Khwarzmi about his distance from Euclid, he would answer: "0, since I am right beside him." If we ask Khwarazmi again about his distance to Khayyam, he will say that "my distance is 0 since I am right beside him." However if we were to ask Euclid about his distance to Khayyam he would answer: "One unit (person) since Khwarzmi is in the middle." We observe that this distance does not satisfy the triangle property as well. In this example the people sitting in the middle are the relevant factors. If we deal with a vector of sorted observations, then observations in the middle are the relevant factors.

The following example shows the importance of invariance of loss (used to take a decision) under monotonic transformations.

Example 2:

A student is told that he will receive a scholarship if he ranks first in an exam in his class in either of the subjects mathematics and physics. The teacher of the courses differ and take a practice exam in each subject. They return the students back their marks out of 100. They also publish the lists of all the marks after removing the names, to give the students a feeling of how they did in the class. Table 1 shows the marks in mathematics and physics.

Mathematics	Physics	Physics before re-scaling
80	90	81.0
65	89	79.2
63	86	74.0
61	85	72.2
54	83	68.9
54	82	67.2
53	79	62.4
50	79	62.4
49	76	57.8
48	75	56.2
47	72	51.8
47	72	51.8
46	69	47.6
44	68	46.2
30	55	30.2

Table 1: A class marks in mathematics and physics. The third column are the raw physics marks before the physics teacher scaled them.

Sarina got 63 in math and 75 in physics. He decided to focus on just one subject that gives him a better chance in order to win the scholarship. He compared his mark in math with the best student in math: 63 against 80. So he needed

$$|\text{best mark} - \text{Sarina's mark}| = 80 - 63 = 17$$

more marks to be as good as the best student. Then he compared his physics mark to the best student in physics. He found he needs 90-75=15 marks to be as good as him. So he thought it's better to focus on physics. But then he realized that different teachers use different exam and scoring methods. He had heard that the physics teacher scales the marks upward by the formula

$$\text{new mark} = \sqrt{100 \times \text{old mark}}.$$

So the student calculated the untransformed values and put the result in the third column. Now he noticed that his new mark is 56.2 while the best mark is 91. The difference this time is 24.8 which is a larger difference than before. According to his "decision-making tool", the absolute difference, he should focus on math since the absolute difference for math was only 17. But what if the mathematics teacher had used another transformation to re-scale the marks without him knowing it? This made him see a disadvantage to using the absolute value difference. Instead he realized, he can use the number of the students between himself and the best student as a measure of the difficulty of getting the best mark. He noticed his decision in this case will be independent of how the teachers re-scaled the marks. In the math case there is only one and for physics there are 8 students between him and the best student. Hence he decided that he should focus on math.

This example shows in order to avoid contrary decisions when the scale changes, we need invariance of the loss under such transformations.

This example was under the assumption that other students do not change their study habits or do not have access to the marks. If the other students had access to their marks or were ready to change their study focus, we need to take into account other possible actions of the other students and the problem will become game-theoretical in nature, a very interesting problem on its own right. The solution for that problem we conjecture to be the same.

4 Limit theory for probability loss function

Suppose a random variable X with a distribution function F is given and $S = (X_1, \cdots, X_n)$ as an *i.i.d* sample from X and let F_n be the empirical distribution of the sample. We defined a distribution loss associated with F, δ_F a deterministic function and the loss associated to the sample δ_{F_n}, a random variable. The following theorem shows the sample loss tends to the distribution loss almost surely.

Theorem 4.1 *Suppose $X_1, X_2, \cdots$, is a sequence of i.i.d random variables with distribution function F. Then as $n \to \infty$,*

$$\delta_{F_n}(z, z') \to \delta_F(z, z'), \quad a.s.,$$

uniformly in $z, z' \in \mathbb{R}$. In other words

$$\sup_{z > z' \in \mathbb{R}} |\delta_{F_n}(z, z') - \delta_F(z, z')| \to 0, \quad a.s..$$

Proof If $z = z'$, the result is trivial. Suppose $z > z'$. We need to show that

$$\lim_{u \to z^-} F_n(u) - F_n(z') \underset{a.s.}{\to} \lim_{u \to z^-} F(u) - F(z'), \tag{1}$$

as $n \to \infty$, uniformly in $z > z' \in \mathbb{R}$. Suppose $\epsilon > 0$ is given. By Glivenko-Cantelli Theorem there exist $N \in \mathbb{N}$ such that for every $n > N$:

$$|F_n(u) - F(u)| < \frac{\epsilon}{2}, \quad a.s., \quad \forall u \in \mathbb{R}.$$

Now for $n > N$,

$$|(\lim_{u \to z^-} F_n(u) - F_n(z')) - (\lim_{u \to z^-} F(u) - F(z'))| \leq$$

$$|\lim_{u \to z^-} (F_n(u) - F(u))| + |F_n(z') - F(z')| = \lim_{u \to z^-} |F_n(u) - F(u)| + |F_n(z') - F(z')|.$$

But since $|F_n(u) - F(u)| < \frac{\epsilon}{2}$, $\lim_{u \to z^-} |F_n(u) - F(u)| \leq \frac{\epsilon}{2}$. Also $|F_n(z') - F(z')| < \frac{\epsilon}{2}$. Hence

$$|(\lim_{u \to z^-} F_n(u) - F_n(z')) - (\lim_{u \to z^-} F(u) - F(z'))| < \epsilon.$$

■

5 Probability loss for continuous distributions

This section studies the probability loss when the distribution function is continuous. The results are given in the following lemmas, which show some of its desirable properties in the continuous case.

Lemma 5.1 *(Probability loss for continuous distributions) Suppose X is a random variable with distribution function F_X. Then $\delta_X(lq_X(p_1), rq_X(p_2)) = p_2 - p_1,\ p_2 > p_1,\ \forall p_1, p_2 \in [0,1]$ if and only if F_X is continuous.*

Proof If F_X is continuous then for $p_1 < p_2$ and by Lemma 1.2,

$$\delta(lq_X(p_1), rq_X(p_2)) = P(lq_X(p_1) < X < rq_X(p_2)) =$$

$$P(X < rq_X(p_2)) - P(X \leq lq_X(p)) = F(rq_X(p_2)) - F(lq_X(p_2)) = p_2 - p_1.$$

If F is not continuous then there exists an x_0 such that $a = P_X(X = x_0) > 0$. Let $p_1 = P(X < x_0) + a/3$ and $p_2 = P(X < x_0) + a/2$. Clearly $lq_X(p_1) = x_0$ and $rq_X(p_2) = x_0$. Hence

$$\delta(lq_X(p_1), rq_X(p_2)) = 0 \neq p_2 - p_1.$$

■

Lemma 5.2 *Suppose $\delta(lq_X(p_1), rq_X(p_2)) = \delta(rq_X(p_1), lq_X(p_2)) = a,\ p_1 < p_2$. Then also*

$$\begin{aligned} a &= \delta(lq_X(p_1), lq_X(p_2)) \\ &= \delta(rq_X(p_1), lq_X(p_2)) \\ &= \delta(rq_X(p_1), rq_X(p_2)). \end{aligned}$$

Moreover, if X is continuous, all the above are equal to $p_2 - p_1$.

Proof The result follows immediately from the fact that all the three quantities are greater than or equal to $\delta(rq_X(p_1), lq_X(p_2)) = a$ and smaller than or equal to $\delta(lq_X(p_1), rq_X(p_2)) = a$. The second part is straightforward using the previous lemma. ■

6 Interpreting results about quantiles using probability loss

This section shows the usefulness of probability loss for interpreting results about quantiles that do not appear intuitive under typical losses such as absolute or square error.

Closeness of left and right quantiles

By definition left and right quantiles. The left and right quantile at a point p can disagree. In fact their absolute difference can become arbitrarily large. For example for $k > 0$ define $P(X = 0) = P(X = k) = 1/2$. Then $lq_X(1/2) = 0, rq_X(1/2) = k$ and hence $|lq_X(1/2) - rq_X(1/2)| = k$, where k can be taken arbitrarily large. However, it is easy to see $P(lq_X(p) < X < rq_X(p)) = 0$ (Lemma 1.1, Part (e)) and we conclude

$$\delta_X(lq_X(p), rq_X(p)) = 0.$$

Convergence of sample quantiles

We can consider the left and right quantiles of the empirical distribution function F_n of a sample $X_1, \cdots, X_n$ of independent random variables identically distributed with distribution function F. Then one would hope that $lq_{F_n} \to lq_F(p)$ and $rq_{F_n}(p) \to rq_F(p)$. This is actually true if $lq_F(p) = rq_F(p)$. In fact one can show by an example that this is not true if $lq_F(p) = rq_F(p)$. Moreover [3] showed if $lq_F(p) \neq rq_F(p)$ then the sample quantiles diverge almost surely. More precisely he showed

Theorem 6.1 *(Quantile Convergence/Divergence Theorem)*

a) Suppose $rq_F(p) = lq_F(p)$ *then*

$$rq_{F_n}(p) \to rq_F(p), \quad a.s.,$$

and

$$lq_{F_n}(p) \to lq_F(p), \quad a.s..$$

b) When $lq_F(p) < rq_F(p)$ *then both* $rq_{F_n}(p), lq_{F_n}(p)$ *diverge almost surely.*

Proof See [3]. ■

Unfortunately based on the above theorem the sample quantiles do not converge in general to the distribution version. In fact [3] shows that when $lq_F(p) < rq_F(p)$ the liminf of the sample quantile is $lq_F(p)$ and the limsup is $rq_F(p)$. Moreover, [2] shows in the following theorem the nice property that in general the sample quantiles converge to distribution quantiles in the probability loss sense uniformly. In the following proof for a random variable X we define $F_X^c(x) = P(X \leq x)$ and $F_X^o(x) = P(X < x)$.

Theorem 6.2 *Let* $X_1, X_2, \cdots$ *be an i.i.d. random sample drawn from an arbitrary distribution function* F*. Then*

$$(a) \quad \sup_{p \in (0,1)} \delta_F(lq_{F_n}(p), lq_F(p)) \to 0., \ a.s.,$$

and

$$(b) \quad \int_{p\in(0,1)} \delta_F(lq_{F_n}(p), lq_F(p)) \to 0., \ a.s..$$

Proof We only need to prove (a) since (b) is a straightforward consequence of (a). Clearly $lq_{F_n}(p) = X_{i:n}$ for $p \in ((i-1)/n, i/n], i = 1, 2, \cdots, n$. Also $F_n^c(X_{i:n}) \geq i/n$ and $F_n^o(X_{i:n}) \leq (i-1)/n$. Pick an N large enough in the Glivenko-Cantelli Theorem such that

$$n > N \Rightarrow |F_n(x) - F(x)| < \epsilon, \text{ and } |F_n^o(x) - F^o(x)| < \epsilon,$$

uniformly in x. Consider two cases:
Case I: $X_{i:n} < lq_F(p)$. Then

$$\begin{aligned} \delta_F(lq_{F_n}(p), lq_F(p)) &= \delta_F(X_{i:n}, lq_F(p)) = \\ F^o(lq_F(p)) - F^c(X_{i:n}) &\leq F^o(lq_F(p)) - F_n^c(X_{i:n}) + \epsilon \\ &\leq p - i/n + \epsilon \leq \epsilon. \end{aligned}$$

Case II: $X_{i:n} > lq_F(p)$. Then

$$\begin{aligned} \delta_F(lq_{F_n}(p), lq_F(p)) &= \delta_F(X_{i:n}, lq_F(p)) = \\ F^o(X_{i:n}) - F^c(lq_F(p)) &\leq F_n^o(X_{i:n}) + \epsilon - p \\ &\leq (i-1)/n + \epsilon - p \leq \epsilon. \end{aligned}$$

Since this holds for $i = 1, 2, \cdots, n$ and $(0,1) = \cup_{i=1,2,\cdots,n}(\frac{i-1}{n}, \frac{i}{n}]$, the supremum is also less than ϵ. ■

Equivariance of quantiles

It is claimed that the classic quantile function, i.e. the left quantile function, is equivariant under strictly increasing transformations ([5] and [1]). However, [4] showed that continuity is a necessary (and sufficient) condition for this to hold. A counterexample for the claim is given below.
Counterexample: Suppose X is distributed uniformly on [0,1]. Then $lq_X(1/2) = 1/2$. Now consider the following strictly increasing transformation

$$\phi(x) = \begin{cases} x & -\infty < x < 1/2 \\ x + 5 & x \geq 1/2 \end{cases}.$$

Let $T = \phi(X)$ then the distribution of T is given by

$$P(T \leq t) = \begin{cases} 0 & t \leq 0 \\ t & 0 < t \leq 1/2 \\ 1/2 & 1/2 < t \leq 5 + 1/2 \\ t - 5 & 5 + 1/2 < t \leq 5 + 1 \\ 1 & t > 5 + 1 \end{cases}.$$

It is clear form above that $lq_T(1/2) = 1/2 \neq \phi(lq_X(1/2)) = \phi(1/2) = 5 + 1/2$.

In fact for any increasing but not left continuous function we can build an example as above. Moreover the example can be built in a way that the transformed quantile and the quantile of the transformation are arbitrarily far in terms of the absolute difference (replace 5 by k in above example). In the following theorem [4] showed that with continuity this problem is resolved.

Theorem 6.3 *(Quantile Equivariance Theorem) Suppose $\phi : \mathbb{R} \to \mathbb{R}$ is non-decreasing.*

a) If ϕ is left continuous then

$$lq_{\phi(X)}(p) = \phi(lq_X(p)).$$

b) If ϕ is right continuous then

$$rq_{\phi(X)}(p) = \phi(rq_X(p)).$$

Proof See [4]. ■

It is unappealing that the equivariance property does not hold for arbitrary increasing transformations. However, [4] showed that using the probability loss a version of equivariance can be shown for such functions.

Lemma 6.1 *(Equivariance under non-decreasing transformations) Suppose X is a random variable with distribution function F and $\phi : \mathbb{R} \to \mathbb{R}$ a non-decreasing transformation on $\mathbb{R}$. Also let $Y = \phi(X)$. Then*
a) $\phi(lq_X(p)) \in [lq_Y(p), rq_Y(p)]$
b) $\phi(rq_X(p)) \in [lq_Y(p), rq_Y(p)]$.

Remark. We conclude

$$\delta_Y(\phi(lq_X(p)), lq_Y(p)) = 0,$$

and

$$\delta_Y(\phi(rq_X(p)), rq_Y(p)) = 0.$$

7 The supremum of probability loss

This section investigates how large the probability loss can become under various scenarios. The results are given in the following lemmas.

Lemma 7.1 *Let $\mathcal{F}$ be the set of all univariate distribution functions. Then*

$$\sup_{F\in\mathcal{F}} \delta_F(lq_F(p_1), lq_F(p_2)) = p_2 - p_1, \;\; p_2 > p_1, \; p_1, p_2 \in (0,1).$$

Proof This follows from the fact that $\delta_F(lq_F(p_1), lq_F(p_2)) \leq p_2 - p_1$ in general, as shown in Lemma 3.1 and $\delta_F(lq_F(p_1), lq_F(p_2)) = p_2 - p_1$ for continuous variables. ■

The same is true for data vectors as shown in the following lemma.

Lemma 7.2 *Suppose the supremum in the following is taken over all data vectors, then*

$$\sup_{x} \delta_x(lq_x(p_1), lq_x(p_2)) = p_2 - p_1, \;\; p_2 > p_1, \; p_1, p_2 \in (0,1).$$

Proof We know that $\delta_x(lq_x(p_1), lq_x(p_2)) \leq p_2 - p_1$. To show that the supremum attains the upper bound, let $x^n = (1, \cdots, n)$. Then $lq_{x^n}(p_1) = [np_1]$ or $[np_1] + 1$. Also $lq_{x^n}(p_2) = [np_2]$ or $[np_2] + 1$. Then Δ, the number of elements of x between $lq_{x^n}(p_1)$ and $lq_{x^n}(p_2)$ satisfies:

$$[np_2] - [np_1] - 1 \leq \Delta \leq [np_2] - [np_1] + 1 \Rightarrow$$

$$np_2 - 1 - np_1 - 1 - 1 \leq \Delta \leq np_2 - np_1 + 1 \Rightarrow$$

$$-3/n \leq \delta_{x^n}(p_1, p_2) - (p_2 - p_1) \leq 1/n.$$

This shows that $\delta_{x^n}(p_1, p_2)$ tends to $p_2 - p_1$ uniformly for all $p_1 < p_2 \in [0,1]$. ■

Lemma 7.3 *Suppose $p_1, p_2, \cdots, p_m \in [0,1]$ and $m = 2k$. Then*

$$\sup_{x} \max\{\delta_x(lq_x(p_1), lq_x(p_2)), \delta_x(lq_x(p_3), lq_x(p_4)), \cdots, \delta_x(lq_x(p_{m-1}), lq_x(p_m))\}$$

$$= \max\{|p_2 - p_1|, \cdots, |p_m - p_{m-1}|\}.$$

Proof The supremum is less than or equal to the left hand side by Lemma 1.1. Let $x^n = (1, 2, \cdots, n)$. Without loss of generality suppose $p_1 < p_2, p_3 < p_4, \cdots, p_{2k-1} < p_{2k}$. By the properties of quantiles of data vectors:
$lq_{x^n}(p_i) = x_{[np_i]} = [np_i]$ or $lq_{x^n}(p_i) = x_{[np_i]+1} = [np_i] + 1$.
Also, $lq_{x^n}(p_{i+1}) = x_{[np_{i+1}]} = [np_{i+1}]$ or $lq_{x^n}(p_{i+1}) = x_{[np_{i+1}]+1} = [np_{i+1}] + 1$.
Then, $\delta_{x^n}(lq_{x^n}(p_i), lq_{x^n}(p_{i+1})) \geq \frac{1}{n}([np_{i+1}] - [np_i] - 1) \geq \frac{1}{n}(np_{i+1} - np_i - 2) = (p_{i+1} - p_i) - \frac{2}{n}$. Hence

$$\delta_{x^n}(lq_{x^n}(p_i), lq_{x^n}(p_{i+1})) > |p_{i+1} - p_i| - \frac{2}{n}, \quad i = 1, \cdots, m-1.$$

The inequality shows the supremum is greater than

$$= \max\{|p_2 - p_1| - \frac{2}{n}, \cdots, |p_m - p_{m-1}| - \frac{2}{n}\},$$

for all $n \in \mathbb{N}$. Now let $n \to +\infty$ to get the conclusion.

■

Lemma 7.4 *Suppose* $p_1, p_2, \cdots, p_m \in [0, 1]$ *and* $a_1, a_1, \cdots, a_{2m} \in [0, 1]$. *Then*

$$\sup_x [\int_{a_1}^{a_2} \delta_x(lq_x(p_1), lq_x(p))dp + \int_{a_3}^{a_4} \delta_x(lq_x(p_2), lq_x(p))dp +$$

$$\cdots + \int_{a_{2m-1}}^{a_2 m} \delta_x(lq_x(p_m), lq_x(p))dp]$$

$$= \int_{a_1}^{a_2} |p - p_1|dp + \int_{a_3}^{a_4} |p - p_2|dp + \cdots + \int_{a_{2m-1}}^{a_2 m} |p - p_m|dp.$$

Proof The proof is similar to the previous lemmas and we skip the details. ■

8 penalized probability loss

This section introduces a family of loss functions that are very similar to the probability loss function but might be more useful in some contexts, particularly when the distribution function is not continuous. A defect of the probability loss function is: it can be equal to zero even if $a \neq b, a, b \in \mathbb{R}$. Also we noted that even though it resembles a metric it is not one. For example the triangular inequality does not hold. We introduce the "c-probability loss" to solve these problems.

Definition 8.1 *Suppose X is a random variable, δ_X its associated probability loss function and $c \geq 0$. Then let*

$$\delta_X^c(a,b) = \delta_X(a,b) + c(1 - 1_{\{0\}}(a-b)),$$

where $1_{\{0\}}$ is the indicator function at zero.

Note that the c-probability loss is the sum of two losses. The first, $\delta_X(a,b)$, is the probability of being between the two values (a and b), the second, $c(1-1_{\{0\}}(a-b))$, is the penalty for a and b not being equal. One question is what value of c should be chosen as the "penalty" of not being equal to the true value. It turns out that the value of c is not very important for many purposes as shown in the following lemma.

Lemma 8.1 *(Properties of the c-probability loss functions)*
a) $\delta_X^c(a,b) = c \Leftrightarrow a \neq b$ and $\delta_X(a,b) = 0$.
b) $\delta_X^c(a,b) = 0$ or $\delta_X^c(a,b) \geq c$.
c) δ_X^c is invariant under strictly monotonic transformations.
d) Let $d = \sup_{x_0 \in \mathbb{R}} P(X = x_0)$. Then if $c \geq d$, δ^c satisfies the triangle inequality.
e) $\delta_X^c(lq_X(p), rq_X(p)) \leq c$. (It is either zero or c.)
f) Suppose δ_X^c is given for any $c > 0$. Then we can obtain any other δ_X^d for $d \geq 0$.

Proof a) and b) are trivial.
c) Both δ_X and $c(1 - 1_{\{0\}}(a-b))$ are invariant under monotonic transformations.
d) We use the pseudo-triangle inequality for the probability loss function. Take $z_1, z_2, z_3 \in \mathbb{R}$. We need to show $\delta_X^c(z_1,z_3) \leq \delta_X^c(z_1,z_2) + \delta_X^c(z_2,z_3)$. If $z_1 = z_3$, the result is trivial. Otherwise $c(1 - 1_{\{0\}}(z_1 - z_3)) = c$ and

$$\delta_X^c(z_1,z_3) = \delta_X(z_1,z_3) + c \leq \delta_X(z_1,z_2) + \delta_X(z_2,z_3) + P(X = z_2) + c$$

$$\leq \delta_X(z_1,z_2) + \delta_X(z_2,z_3) + c(1 - 1_{\{0\}}(z_1 - z_2)) + c(1 - 1_{\{0\}}(z_2 - z_3)) =$$

$$\delta_X^c(z_1,z_2) + \delta_X^c(z_2,z_3).$$

e) Trivial by properties of lq, rq and δ_X as shown in Lemma 1.1.
f) Suppose δ_X^c is given. If $\delta_X^c(a,b) = 0$ then $a = b$ and hence $\delta_X^d(a,b) = 0$. If $a \neq b$ then $\delta_X^c(a,b) = \delta_X(a,b) + c$. From this we can obtain $\delta_X(a,b) = \delta_X^c(a,b) - c$ and hence $\delta_X^d(a,b) = \delta_X^c(a,b) - c + d$. ■

$\delta_X(X_1,X_2)$ (or $\delta_X^c(X_1,X_2)$), if $X_1, X_2 \overset{i.i.d}{\sim} X$ can be considered as a measure of disparity of the common distribution. The following lemma shows that the expectation of this quantity is constant for all continuous random variables. It is also easy to show in the non-continuous case the expectation is smaller than the one given below.

Lemma 8.2 *Suppose X is a continuous random variable, then*

$$E(\delta_X(X_1, X_2)) = 2/3,$$

where $X_1, X_2 \overset{i.i.d}{\sim} X$. Also

$$E(\delta_X^c(X_1, X_2)) = 2/3 + c.$$

Proof We know that $F_X(X_1)$ and $F_X(X_2)$ are both uniformly distributed on (0,1) and independent. Hence

$$\begin{aligned} E(\delta_X(X_1, X_2)) &= E(|F(X_1) - F(X_2)|) = \\ \int_0^1 \int_0^1 |p_1 - p_2| dp_1 dp_2 &= 2\int_0^1 \int_{p_2}^1 (p_1 - p_2) dp_1 dp_2 = \\ 2\int_0^1 (1 - 2p_2 + p_2^2) dp_2 &= 2/3. \end{aligned}$$

$E(\delta_X^c(X_1, X_2)) = 2/3 + c$ is obtained by noting that $P(X_1 = X_2) = 0$ for continuous random variables. ■

It is interesting to note that $\delta_F(X_1, X_2)$ in this case is a special case of the coverage probabilities discussed by [10].

9 Extensions to statistics and multi-dimensional data

This section shows how probability loss function can be extended to more than one dimension and also to measure the distance between statistics. However, we do not study this case in details and the applications are left to future research.

Suppose $X_1, \cdots, X_n$ is a random sample and consider two statistics

$$T_1(X_1, \cdots, X_n) \quad \text{and} \quad T_2(X_1, \cdots, X_n).$$

For example these statistics might be estimators of $lq_X(p)$ e.g. $X_{i:n}$ (ith order statistics) for some i. Then we can consider the random loss

$$\delta_X(T_1(X_1, \cdots, X_n), T_2(X_1, \cdots, X_n)).$$

In order to find optimal estimators we need a deterministic measure and one can settle for the expected probability loss (EPL)

$$E(\delta_X(T_1(X_1, \cdots, X_n), T_2(X_1, \cdots, X_n))).$$

In the following definition we offer a new idea to measure this loss.

Definition 9.1 *Suppose $X_1, X_2, \cdots$ a random sample drawn from distribution function F_X and consider another draw X_F which is independent of the random sample (can think of this as a "future value"). Then for two statistics*

$$T_1(X_1, X_2, \cdots), T_1(X_1, X_2, \cdots),$$

define the future value probability loss function (FPL) as follows

$$\gamma_X(T_1, T_2) = P(T_1 < X_F < T_2) + P(T_2 < X_F < T_1).$$

Remark. Note that for constant numbers a, b we have $\gamma_X(a, b) = \delta_X(a, b)$. It can easily be shown that if S, T are equivariant under strictly monotonic transformations of the random sample then so is the FPL.

Lemma 9.1 *Suppose $X_1, X_2, \cdots$ a random sample from F_X and X_F is another draw independent from the random sample and consider a strictly monotonic transformation of the reals $\phi : \mathbb{R} \to \mathbb{R}$. If*

$$T_1(X_1, X_2, \cdots), T_2(X_1, X_2, \cdots),$$

are two equivariant statistics under ϕ i.e.

$$\phi(T_j(X_1, X_2, \cdots)) = T_j(\phi(X_1), \phi(X_2), \cdots),\ j = 1, 2$$

then we have

$$\gamma_{\phi(X)}(T_1(\phi(X_1), \cdots), T_2(\phi(X_2), \cdots)) = \gamma_X(T_1(X_1, \cdots), T_2(X_1, \cdots))$$

Remark. Order statistics clearly satisfy the above property.
Remark. As a less trivial example consider a sample of size $n = n_1 + n_2$:

$$X_1, \cdots, X_{n_1}, Y_1, \cdots, Y_{n_2}$$

and consider $\min(X_{i:n_1}, Y_{j:n_2})$. As yet a more interesting example let

$$Z_1 = X_{i_1:n_1}, \cdots, Z_k = X_{i_k:n_1}, Z_{k+1} = Y_{j_1:n_2}, \cdots, Z_{k+l} = Y_{j_l:n_2}$$

and consider $Z_{s:(k+l)}$. This example has important applications in dealing with massive data sets. Suppose n_1, n_2 are very large and hence loading the two sample in the same time on the computer memory is not possible, then one can 1-load them individually, 2-save a subset of their order statistics of size k and l and 3-use these summaries to infer about order statistics of the full sample or quantiles of the underlying distributions. It is also interesting to find optimal ways to choose these subsets.

As an application, we can estimate a quantile $lq_X(p)$ using a statistics T. Then the loss can be assessed by either EPL or FPL. In fact in Section 11, we use these losses to find optimal estimators.

Extensions to multi-dimensional data is possible in several ways. We introduce two ways here. In the first one, we extend the point loss definition defined on pairs of points in $\mathbb{R}$ to multi-dimensional pairs of points in $\mathbb{R}^k$. Suppose $X = (X^1, \cdots, X^k)$ a random vector in $\mathbb{R}^k$ and $a = (a^1, \cdots, a^k), b = (b^1, \cdots, b^k) \in \mathbb{R}^k$ then we can define

$$\delta_X(a, b) = P(X^i \in (c^i, d^i),\ i = 1, \cdots, k),$$

where $c^i = \min(a^i, b^i)$ and $d^i = \max(a^i, b^i)$. Obviously this matches with the definition previously given for $k = 1$. We can again show the invariance property of this loss under componentwise strictly monotonic transformations

$$\phi = (\phi^1, \cdots, \phi^k) : \mathbb{R}^k \to \mathbb{R}^k,$$

meaning each $\phi^i : \mathbb{R} \to \mathbb{R}$ is strictly monotonic.

Another way to extend the definition is to consider a class "parallel" hyper surfaces in $\mathbb{R}^k$ (every two hyper surface in the class are identical or disjoint) and for two hyper surfaces S_1 and S_2 define the loss as the probability the random variable falls in between the two hyper surfaces. Such a class will be transformed to another class of parallel hyper surfaces using a componentwise strictly monotonic transformations. We leave the study of the properties of these extensions and applications to multi-variate data to future research.

10 Applications in approximating quantiles in large or imperfect datasets

First we prove two lemmas. These lemmas show what happens to the quantiles if we throw away a small portion of the data vector or add some more data to it. The first lemma is for a situation that we have thrown away or ignored a small part of the data. The second lemma is for a situation that a small part of the data are contaminated or includes outliers. In both cases, we show how the quantiles computed in the "imperfect" vectors correspond to the quantiles of the original vector. In both case x stands for the imperfect vector and w is the complete/clean data.

Lemma 10.1 *(Missing data quantile approximation lemma)*
Suppose $x = (x_1, \cdots, x_n)$, $sort(x) = (y_1, \cdots, y_n)$ *and* $y' = lq_x(p), p \in [0, 1]$. *Consider a vector* $x^\star$ *of length* $n^\star$ *and let* $w = stack(x, x^\star)$. *Then* $y' = lq_w(p')$, *where* $p' \in [p - \epsilon, p + \epsilon]$ *and* $\epsilon = \frac{n^\star}{n+n^\star}$. *In other words* $\delta_w(y, y') \leq \epsilon$.

Similarly if $y' = rq_x(p)$ *and* $p \in [0, 1]$, $y' = rq_w(p')$, *where* $p' \in [p - \epsilon, p + \epsilon]$ *and* $\epsilon = \frac{n^\star}{n+n^\star}$. $\delta_w(y, y') \leq \epsilon$.

Remark. Note that no error guarantee can be given using typical loss functions such as absolute value.

Proof We prove the result for lq_x only and a similar argument works for rq_x.
Let $z = sort(w)$ then $lq_z = lq_w$. For $p = 1$ the result is easy to see. Otherwise, $\frac{i}{n} \leq p < \frac{i+1}{n}$ for some $i = 0, \cdots, n-1$. But then $y' = lq_x(p) = y_i$. In the new vector z since we have added $n^\star$ elements $y' = z_j$ for some j, $i \leq j < i + n^\star$. Hence $y' = lq_z(\frac{j}{n+n^\star})$. From $np - 1 < i \leq np$, we conclude

$$\frac{np-1}{n+n^\star} < \frac{i}{n+n^\star} \leq \frac{j}{n+n^\star} < \frac{i+n^\star}{n+n^\star} \leq \frac{np+n^\star}{n+n^\star}.$$

Hence,

$$\frac{n^\star(1-p)-1}{n+n^\star} < \frac{j}{n+n^\star} - p < \frac{n^\star(1-p)}{n+n^\star} \Rightarrow$$

$$|\frac{j}{n+n^\star} - p| < \max\{|\frac{n^\star(1-p)-1}{n+n^\star}|, |\frac{n^\star(1-p)}{n+n^\star}|\}.$$

But $|\frac{n^\star(1-p)}{n+n^\star}| \leq \frac{n^\star}{n+n^\star}$ and $|\frac{n^\star(1-p)-1}{n+n^\star}| \leq \max\{\frac{n^\star-1}{n+n^\star}, \frac{1}{n+n^\star}\}$ since p ranges in $[0,1]$. We conclude that that

$$|\frac{j}{n+n^\star} - p| < \frac{n^\star}{n+n^\star}.$$

■

Lemma 10.2 *(Contaminated data quantile approximation lemma)*
Suppose $x = (x_1, \cdots, x_n)$, $sort(x) = (y_1, \cdots, y_n)$ *and* $y' = lq_x(p), p \in [0,1]$. *Consider the vector* $w = (x_1, x_2, \cdots, x_{n-n^\star})$ *then* $y' = lq_w(p')$, *where* $p' \in [p - \epsilon, p + \epsilon]$ *and* $\epsilon = \frac{n^\star}{n-n^\star}$. $\delta_w(y, y') \leq \epsilon$.

Similarly if $y' = rq_x(p)$ *and* $p \in [0,1]$, $y' = rq_w(p')$, *where* $p' \in [p - \epsilon, p + \epsilon]$ *and* $\epsilon = \frac{n^\star}{n-n^\star}$. $\delta_w(y, y') \leq \epsilon$.

Proof We only show the case for lq_x and a similar argument works for rq_x.
Let $z = sort(w)$. Then $lq_z = lq_w$. If $p = 1$ the result is easy to see. Otherwise, $\frac{i}{n} \leq p < \frac{i+1}{n}$ for some $i = 0, \cdots, n-1$. But then $y' = lq_x(p) = y_i$. In the new vector z since we have removed $n^\star$ elements $y' = z_j$ for some j, $i - n^\star \leq j \leq i$. Hence $y' = lq_z(\frac{j}{n-n^\star})$. From $np - 1 < i \leq np$, we conclude $np - 1 - n^\star < j \leq np \Rightarrow np - n^\star \leq j \leq np$. Hence

$$\frac{-n^\star + n^\star p}{n-n^\star} \leq \frac{j}{n-n^\star} - p \leq \frac{n^\star p}{n-n^\star} \Rightarrow$$

$$|\frac{j}{n-n^\star} - p| \leq \frac{n^\star}{n-n^\star}.$$

■

[2] introduced an algorithm to approximate quantiles of very large datasets. The idea of the algorithm is to read partitions of a very large data vectors sequentially and save a summary of the partitions and then refer about the quantiles of the original vector using the summaries. The algorithm allows for non-equal partition sizes. The accuracy of the approximation can be given deterministically using the probability loss function as stated in the following theorem.

Theorem 10.1 *Suppose x is of length $n = \sum_{i=1}^{m} l_i$, $m \geq 2$ and $l_i = c_i d$. Let $C = \sum_{i=1}^{m} c_i$. Apply the "coarsening algorithm" ([2]) to x and find μ to approximate $rq_x(p)$ (or $lq_x(p)$). Then μ is a (left and right) quantile in the interval*

$$[p - \epsilon, p + \epsilon],$$

where $\epsilon = \frac{m+1}{C-m}$. In other words $\delta_x(\mu, rq_x(p)) \leq \epsilon$ and $\delta_x(\mu, lq_x(p)) \leq \epsilon$. When $l_i = cd,\ i = 1, \cdots, m$, $\epsilon = \frac{m+1}{m-1}\frac{1}{c-1} \leq \frac{3}{c-1}$.

Remark. Note that again no error guarantee can be given using typical loss functions such as absolute value.

11 Estimation

This section shows how the probability loss idea can be used to estimate parameters of a distributions, in particular quantiles.

Here we discuss estimating a quantile of a random variable only. However, the method can be used in estimating parameters of any family of random variables that are specified by their quantiles at specific points. For example the normal family, $N(\mu, \sigma^2)$, can be specified by $lq(0) = \mu, lq(1) = \mu + \sigma^2$ and we call such families quantile-specified families. It can be seen that most of typical families of distributions are characterized by their values on specific quantiles. The idea is much the same as Wald's decision theoretic approach ([6]) except for the loss is defined differently using expected probability loss δ. In fact we can avoid using expectation by using future probability loss. [2] showed that this estimation method is equivariant under changes of scale of data (even non-linear), a property that does not hold for typical loss functions such as the square error. Here we only focus on estimating a specific quantile and for simplicity suppose the distribution function is continuous and strictly increasing (Hence $lq(p) = rq(p) = q(p)$). However, by some modifications the results can re-stated for the general case.

Suppose a random sample $X_1, \cdots, X_n$ and class of estimators are given $\mathcal{D}$ to estimate a quantile $q(p)$. Then we propose to minimize either of the losses:

a. Expected probability loss:

$$argmin_{D \in \mathcal{D}} EPL(D, q(p)) = argmin_{D \in \mathcal{D}} E(\delta_F(D, q(p))).$$

b. Future probability loss:

$$argmin_{D\in\mathcal{D}} FPL(D, q(p)) = argmin_{D\in\mathcal{D}} \gamma_F(D, q(p)).$$

[2] shows such estimators are equivariant (if $\mathcal{D}$ is closed under strictly monotonic transformations). It is natural to consider $\mathcal{D}$ to include the order statistics $X_{1:n}, \cdots, X_{n:n}$ (from the smallest to largest) with distribution functions $F_{1:n}, \cdots, F_{n:n}$. Also define $G_{i:n}(y) = F_{i:n}(q(y)) = \sum_{j=i}^{n} \binom{n}{j} y^j (1-y)^{(n-j)}$. Then for $D = X_{i:n}$:

a. We have

$$E(\delta_F(X_{i:n}, q(p))) = E|F(X_{i:n}) - p| = \int_{-\infty}^{\infty} |p - F(x)| d_{F_{i:n}}$$

$$= \int_0^1 |p - y| d_{G_{i:n}} = \int_0^p (p - y) d_{G_{i:n}} + \int_p^1 (y - p) d_{G_{i:n}}.$$

b. We have

$$\gamma_F(D, q(p)) = P(X_{i:n} < X_F < q(p)) + P(q(p) < X_F < X_{i:n})$$

$$= \int_{-\infty}^{+\infty} P(x < X_F < q(p)) d_{F_{i:n}} + \int_{-\infty}^{+\infty} P(q(p) < X_F < x) d_{F_{i:n}}$$

$$= \int_{-\infty}^{q(p)} (p - F(x)) d_{F_{i:n}} + \int_{q(p)}^{+\infty} (F(x) - p) d_{F_{i:n}} = \int_0^p (p - y) d_{G_{i:n}} + \int_p^1 (y - p) d_{G_{i:n}}$$

Note that the estimator does not depend on the distribution as $G_{i:n}$ is the same for all continuous random variables and hence invariant under (possibly non-linear) changes of scale of data. It can also be shown that in the general (possibly non-continuous) these losses are equal to or smaller than the above. Interestingly in this case the two methods give rise to the same answer. Note that the solution does need to be unique. For example if $p = 1/2$ and $n = 2$ as it should be the case both $X_{1:2}$ and $X_{2:2}$ are equally eligible. These equations can be solved numerically (or by simulations) for any given n and i, however a theoretical solution is desirable and is left for future research.

12 Other applications

We saw before that the probability loss function is invariant under re-scaling of data. This is of great importance since the results obtained by using this loss function do not depend on the scale of the data as they should not in most applications. We also showed how this loss can be used to assess quantile approximations in large or imperfect datasets. We found bounds on the error of quantile approximations in

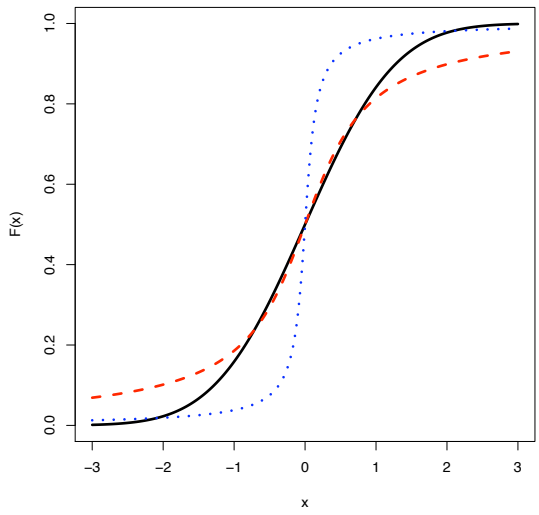

Figure 1: Comparing the standard normal distribution (solid) with optimal Cauchy picked by quantile distance (dashed) and the optimal Cauchy picked by tail quantile distance minimization (dotted).

such datasets in terms of the probability loss. It can be easily seen that no such bound can be found using other classic measures of loss such as the absolute difference or the square of the difference (Since they are not invariant under monotonic transformations). [2] uses this loss function in many other applications and we point out some of them here.

- Suppose a large data vector of length n is given. We want to find $m << n$ elements of this vector in such a way we can optimally refer to the original vector. [2] solved this problem using results of Section 7, by looking for m (left) quantiles of the original vector $p_1, \cdots, p_m$ which minimize the probability loss when we use these quantiles to refer about the other unspecified quantiles. The solution is

$$p_1 = \frac{1}{2m}, p_2 - p_1 = 1/m, p_3 - p_2 = 1/m, \cdots, p_{m-1} = 1/m, p_m = 1 - \frac{1}{2m},$$

  which is different from the naive conjecture that $p_i = i/(n+1)$. See Chapter 8 of [2] for more details.

- Suppose a random sample $X_1, \cdots, X_n$ is given with the order statistics

$$X_{1:n}, \cdots, X_{n:n}$$

  then in order to make a quantile-quantile plot the order statistics must be assigned to the theoretical distribution quantiles $lq(p_1), \cdots, lq(p_n)$. Hosseini

used the probability loss function to find the p_i 's by minimizing the expected probability loss

$$E(\delta_F(X_{i:n}, lq_F(p_i))).$$

See Chapter 8 of [2] for more details.

- [2] used the probability loss function to define distance measures among distribution functions. For example one can consider the *sup quantile distance* and *integral quantile distance* respectively

$$SQD(X,Y) = \sup_{p\in E}[\delta_X(lq_X(p), lq_Y(p)) + \delta_Y(lq_X(p), lq_Y(p))],$$

$$IQD(X,Y) = \int_{p\in E}[\delta_X(lq_X(p), lq_Y(p)) + \delta_Y(lq_X(p), lq_Y(p))]dp,$$

where E is a fixed measurable subset of $[0,1]$ for example $E = [0,1]$ or $E = (0, 0.025) \cup (0.0975, 1)$, where the later is more appropriate for studying the distance of the random variables X and Y on the tails. [2] showed the invariance of such measures under strictly monotonic transformations. Figure 1 shows that the closest Cauchy to the standard normal on the tails differs significantly from the closest Cauchy overall using the integral quantile distance.

[2] [Chapter 9] also used the quantile distance minimization to estimate parameters of distributions. He also compare the results to maximum likelihood estimates.

In summary this paper shows that the probability loss is useful to find results regarding quantiles that does not depend on the distribution function or scale of data. We showed this loss function is quite useful when referring about large data vectors using smaller subsets or in the presence contaminated data or when some data are missing. It also provide a decision-theoretic framework for inference that is invariant under changes of scale of data.

Acknowledgements: I would like to thank my PhD supervisors, Prof. Jim Zidek and Prof. Nhu Le for insightful comments. This work was partially supported by NSERC.

References

[1] L. Hao and D. Q. Naiman. *Quantile Regression*. Quantitative Applications in the Social Sciences Series. SAGE publications, 2007.

[2] R. Hosseini. *Statistical Models for Agroclimate Risk Analysis*. PhD thesis, Department of Statistics, UBC, 2009.

[3] R. Hosseini. Divergence of sample quantiles. *eprint arXiv:1005.2781*, 2010.

[4] R. Hosseini. Quantiles equivariance. *eprint arXiv:1004.0533*, 2010.

[5] R. Koenker. *Quantile Regression*. Cambridge university press, 2005.

[6] E. L. Lehmann and G. Casella. *The theory of point estimation*. Springer-Verlag, 1998.

[7] G. S. Manku, S. Rajagopalan, and B. G. Lindsay. Approximate medians and other quantiles in one pass and with limited memory. In *ACM SIGMOD*, volume 27, pages 426–435, 1998.

[8] E. Parzen. Nonparametric statistical data modeling. *Journal of the American Statistical Association*, 74:105–121, 1979.

[9] T. Rychlik. *Projecting statistical functionals*. Springer, 2001.

[10] S. S. Wilks. Order statistics. *Bulletin of the American Mathematical Society*, 5:6–50, 1948.