Why Kappa Regression?

Julio C. Urenda[a,b] and **Orsolya Csiszár**[c,d] and **Gábor Csiszár**[e] and **József Dombi**[f] and **György Eigner**[d] and **Olga Kosheleva**[g] and **Vladik Kreinovich**[a]

[a]Computer Science, Univ. of Texas at El Paso, El Paso, TX 79968, USA, vladik@utep.edu

[a]Math. Sciences, Univ. of Texas at El Paso, El Paso, TX 79968, USA, jcurenda@utep.edu

[c]Basic Sciences, Univ. of Applied Sciences, Esslingen, Germany, orsolya.csiszar@nik.uni-obuda.hu

[d]Inst. of Applied Math., Óbuda Univ., Budapest, Hungary, eigner.gyorgy@nik.uni-obuda.hu

[e]Inst. of Materials Physics, Univ. of Stuttgart, Germany, gabor.csiszar@mp.imw.uni-stuttgart.de

[f]Inst. of Informatics, Univ. of Szeged, Szeged, Hungary, dombi@inf.u-szeged.hu

[g]Teacher Education, Univ. of Texas at El Paso, El Paso, TX 79968, USA, olgak@utep.edu

Abstract

A recent book provide examples that a new class of probability distributions and membership functions – called kappa-regression distributions and membership functions – leads to better data processing results than using previously known classes. In this paper, we provide a theoretical explanation for this empirical success – namely, we show that these distributions are the only ones that satisfy reasonable invariance requirements.

Keywords: Kappa-regression distributions, Kappa-regression membership functions, Invariance

1 Formulation of the Problem

Empirical facts. Recent results have shown (see [2] and references therein) that in many practical situations, probability distributions with the cumulative distribution of the type

$$F(x) = \text{Prob}(X \le x) = \frac{1}{1 + C \cdot \left(\dfrac{b-x}{x-a}\right)^{\lambda}}, \qquad (1)$$

known as *kappa-regression distributions*, provide the best description of the data and the best data processing results.

Similarly, fuzzy processing that uses the membership functions

$$\mu(x) = \frac{1}{1 + C \cdot \left(\dfrac{b-x}{x-a}\right)^{\lambda}} \qquad (2)$$

leads, in many practical situations, to very good results – better than other tested membership functions.

Challenge. How can we explain these empirical results?

In this paper, we provide an explanation for these results, an explanation based on first principles.

2 Let Us Start with the Known Limit Case

A known limit case. While the kappa-regression-distribution itself is new, in the limit, it coincides with several known probability distributions; see [2]. One of such limit distributions is the *logistic* distribution

$$F(x) = \frac{1}{1 + C \cdot \exp(-k \cdot x)}, \qquad (3)$$

which is also known to lead to many useful results.

So, before we explain why the general case of the kappa-regression-distribution is so successful, let us first explain why this limit case has been very successful.

Idea of symmetry. In order to explain why logistic distribution is successful in describing real-life phenomena, let us recall how real-life phenomena are described and explained in the first place, what are the fundamental ideas behind these explanations.

Modern science – especially physics – has been very successful, we can predict many events. But what is the general basis for all these predictions? We observe that the Sun goes up day after day, and we conclude that in the similar situations, the Sun will go up again. We observe, at different locations on the Earth, that if you drop a pen, it will fall with the acceleration of 9.81 m/sec^2, and we conclude that in similar situations, it will fall down with the same acceleration. We observe, in many cases, that mechanical bodies follow Newton's laws, and we conclude that in the similar situations, the same laws will be observed.

In all these cases, we conclude that when we change

a situation to a similar one – e.g., by moving to a different location on Earth or to a different day, etc. – the processes will remain similar. This idea that many physical properties do not change if we perform certain transformations is known as it symmetry. Symmetries are indeed one of the fundamental ideas of modern physics, to the extent that many new theories – starting with theory of quarks – are proposed not by writing down differential equations, but by describing the corresponding symmetries; see, e.g., [3, 8].

What are the simplest symmetries? Some symmetries – e.g., the ones used in quark theory – are rather complicated. Let us start with the simplest possible symmetries.

These symmetries are related to the fact that when we write equations, we operate with numerical values of the physical quantities, but to describe physical quantities by numbers, we need to select a measuring unit and a starting point. For example, we can measure time starting with Year 0 – as in the commonly used calendar – or with any other moment of time; after the French revolution, the new calendar started with the year of the revolution as the first year. We can also change a measuring unit – e.g., count days or months instead of years.

In general, if you replace the original measuring unit with a new unit which is c times smaller, then all numerical values are multiplied by c: $x \to c \cdot x$. For example, if when measuring lengths we replace meters with centimeters, all numerical values will be multiplied by 100: 2 m social distance will become $2 \cdot 100 = 200$ cm.

Definition 1. *By a* scaling*, we mean a transformation (function) $f(x) = c \cdot x$ for some $c > 0$.*

Similarly, if we replace the original starting point with the one which is x_0 units earlier, then all numerical values increase by x_0: $x \to x + x_0$.

Definition 2. *By a* shift*, we mean a transformation $f(x) = x + x_0$ for some x_0.*

In many physical situations, there is no preferred starting point, so we expect that the processes remain similar if we replace change the starting point, i.e., if we replace all numerical values x with shifted values $x + x_0$. Similarly, in many physical situations, there is no preferred measuring unit, so so we expect that the processes remain similar if we replace the measuring unit, i.e., if we replace all numerical values x with re-scaled values $c \cdot x$.

How can we apply these ideas to probability distributions? Of course, if we change the units of one of the quantities, then, to preserve the same equations, we need to accordingly change the units of related quantities. For example, if we start with the formula $d = v \cdot t$ that the distance is velocity times time, and we change the unit for time from hours to seconds, then, to preserve the formula, we need to corresponding change the units for velocity: e.g., from km/h to km/sec.

In probability theory, there is a natural way to change probabilities: the Bayes formula, according to which if we have a new observation E, then the previous probability $P_0(H)$ of a hypothesis H changes to the new value

$$P(H\,|\,E) = \frac{P(E\,|\,H)\cdot P_0(H)}{P(E\,|\,H)\cdot P_0(H) + P(E\,|\,\neg H)\cdot P_0(\neg H)} = \frac{P(E\,|\,H)\cdot P_0(H)}{P(E\,|\,H)\cdot P_0(H) + P(E\,|\,\neg H)\cdot (1 - P_0(H))} = \frac{P_0(H)}{P_0(H) + (1-r)\cdot(1 - P_0(H)}, \tag{4}$$

where we denoted

$$r \stackrel{\text{def}}{=} \frac{P(E\,|\,\neg H)}{P(E\,|\,H)}; \tag{5}$$

see, e.g., [4, 7].

So, a natural idea is to require that if we apply a reasonable transformation to x – e.g., change the starting point or change the measuring unit – then the probability distribution will change according to the Bayes formula (4).

Definition 3. *We say that cumulative distribution functions $F(x)$ and $G(x)$ are* equivalent *if for some real number r, we have:*

$$G(x) = \frac{F(x)}{F(x) + (1-r)\cdot(1 - F(x))}.$$

Definition 4. *Let $f(x)$ be a transformation. We say that a cumulative distribution function $F(x)$ is* invariant with respect to f *(or, f-*invariant*, for short) if the functions $F(f(x))$ and $F(x)$ are equivalent, i.e., if for some real number $r > 0$, we have*

$$F(f(x)) = \frac{F(x)}{F(x) + (1-r)\cdot(1 - F(x))}.$$

What probability distributions satisfy this symmetry requirement? Let us analyze what are the probability distributions that satisfy this requirement.

Proposition 1. *For each cumulative distribution function $F(x)$, the following two conditions are equivalent to each other:*

- *$F(x)$ is invariant with respect to all shifts;*

- *$F(x)$ is a logistic distribution, i.e., is described by the formula (3).*

Proof. It is straightforward to prove that every logistic distribution is shift-invariant. Let us prove that every shift-invariant probability distribution is logistic.

In our proof, we will use the fact that the Bayes formula becomes even simpler if instead of probabilities P, we consider the *odds*

$$O \stackrel{\text{def}}{=} \frac{P}{1-P}. \tag{6}$$

Indeed, from the above formula

$$P' = \frac{P}{P + r \cdot (1-P)},$$

we conclude that

$$1 - P' = \frac{r \cdot (1-P)}{P + r \cdot (1-P)},$$

and thus, that

$$O' = \frac{P'}{1-P'} = \frac{P}{r \cdot (1-P)} = \frac{1}{r} \cdot \frac{P}{1-P} = s \cdot O, \tag{7}$$

where we denoted $s \stackrel{\text{def}}{=} \frac{1}{r}$.

In these terms, the fact that the shift $x \to x + x_0$ should lead to a Bayes-type transformation of the cumulative distribution function $F(x)$ means that for the corresponding odds $O(x)$ and $O(x + x_0)$, we must have

$$O(x + x_0) = s(x_0) \cdot O(x), \tag{8}$$

for some constant s – which is, in general, different for different shifts.

Each cumulative distribution function $F(x)$ is monotonic and thus, measurable. Thus, the odds function is also measurable. It is known (see, e.g., [1]) that all measurable solutions of the functional equation (8) have the form

$$O(x) = c \cdot \exp(k \cdot x) \tag{9}$$

for some values c and k.

It is known how to go back from odds to probabilities: from

$$O = \frac{P}{1-P} = \frac{1}{\frac{1}{P} - 1},$$

we conclude that

$$\frac{1}{P} - 1 = \frac{1}{O},$$

hence

$$\frac{1}{P} = 1 + \frac{1}{O}$$

and

$$P = \frac{1}{1 + \frac{1}{O}}. \tag{10}$$

Thus, in our case, we have

$$P(x) = \frac{1}{1 + \frac{1}{O(x)}} = \frac{1}{1 + C \cdot \exp(-k \cdot x)}, \tag{11}$$

where $C \stackrel{\text{def}}{=} \frac{1}{c}$, i.e., exactly the logistic distribution.

The proposition is proven.

Conclusions of this section. In this section, we have shown that a simple symmetry – namely, invariance with respect to shift – leads to the logistic distribution, and thus, explains why this distribution has been so successful in practice – because it corresponds to the frequent requirement that the physical processes do not change if we simply change the starting point for measuring the corresponding physical quantity.

3 What About the Fuzzy Case?

What about the fuzzy case? According to [2], logistic expression works well not only for the probability distributions, but also for membership functions as well. For membership functions, the above explanation does not work – this explanation is based on the Bayes formula, and this formula is not applicable to membership functions. So, to explain the success of logistic membership functions, we need to provide another explanation.

Idea. To come up with such an explanation, let us recall that one of the possible ways to get membership degrees is to ask experts. If m out of n experts think that the given statement is true, we assign to it the degree of confidence m/n. For example, we can say that a person of a certain age is young to a degree 0.7 if 70% of the experts consider this person young.

Resulting transformations. For complex statements – statements that require true expertise – we want to ask top experts, of whose opinion we are most confident. Suppose that out of n top experts, m thought that the given statement is true; then we assign, to this statement, the degree of confidence $\mu = m/n$.

The problem is that in many practical situations, there are very few top experts: the number n is small. In this case, we have a very limited number of possible degrees. For example, when $n = 5$, we only have 6

possible values: 0, 1/5, 2/5, 3/5, 4/5, and 1. The only way to make more meaningful distinction is to use a larger value of n, i.e., to ask more experts.

However, in the presence of the top experts, other not-so-top experts may be either silent, or simply follow the opinion of their peers. If we ask n' more experts and the new experts are silent, then the new degree of confidence is $\mu' = m/(n+n')$. In terms of the original degree of confidence $\mu = m/n$, we have $m = \mu \cdot n$ and thus, $\mu' = c \cdot \mu$, where $c \stackrel{\text{def}}{=} n/(n+n')$.

If the new experts follow the majority of top experts – and if this majority confirms our statement – then the new degree of confidence is $\mu' = (m+n')/(n+n')$. In terms of the original degree of confidence μ, we have $\mu' = c \cdot \mu + a$, where $a \stackrel{\text{def}}{=} n'/(n+n')$.

In both cases, we have a linear transformation $\mu \to \mu'$. A similar linear transformation occurs if some of the new experts remain silent, and some follow the majority of top experts. So, linear transformations make sense for fuzzy degrees as well.

Beyond linear transformations. In principle, not all functions are linear – for example, the Bayes formula describes a non-linear transformation. So let us look for a general class of transformations, i.e., functions from real line to real line, with respect to which physical properties can be invariant.

Clearly, if the properties do not change when we apply a transformation $x' = f(x)$, and do not change if we then apply the transformation $x'' = g(x')$, this means that the whole transformation from x to $x'' = g(x') = g(f(x))$ – which is the composition of two original transformations – also does not change the properties. Thus, the class of possible transformations must be closed under composition.

Similarly, if the physical properties do not change when we go from x to $y = f(x)$, this means that the transition back, from y to $x = f^{-1}(y)$, where f^{-1} denotes the inverse function, also preserves all physical properties. So, the class of possible transformation must contain the inverse transformation.

In mathematical terms, this means that the class of all possible transformations much be a *group*. Also, we want this to be constructive, we want to be able to simulate such transformations on a computer. At any given moment of time, a computer can only store and use finitely many parameters. Thus, elements of the desired transformation group must be uniquely determined by the values of finitely many parameters. In mathematical terms, this means that the corresponding group must be *finite-dimensional*. It is known that under reasonable conditions, any finite-dimensional transformation group that contains all linear transformation contains only fractional-linear transformations, i.e., transformations of the type [5, 6, 9, 10], etc.

$$f(x) = \frac{A + B \cdot x}{C + D \cdot x}. \tag{11}$$

So, we will consider fractional-linear transformations.

Comment. In particular, for $D = 0$, we get linear transformations.

Definition 5. *By a* reasonable transformation*, we mean a fractional-linear transformation, i.e., a transformation of type (11).*

Which reasonable transformations preserve the interval $[0,1]$? Possible degrees of confidence form the interval $[0,1]$. It is therefore reasonable to look for transformations that preserve this intervals, i.e., that map $[0,1]$ exactly into $[0,1]$.

Definition 6. *Let $a < b$ be real numbers. We say that a transformation $f(x)$* preserves *the interval $[a,b]$ if the range $f([a,b]) = \{f(x) : x \in [a,b]\}$ of this transformation on the interval $[a,b]$ is equal to this same interval: $f([a,b]) = [a,b]$.*

Proposition 2. *If a reasonable transformation $f(x)$ preserves the interval $[0,1]$, then this transformation has the form*

$$f(x) = \frac{x}{x + r \cdot (1-x)}, \tag{12}$$

for some real number r.

Proof. The requirement that the interval $[0,1]$ is invariant under the transformation (12) implies that we should have $f(0) = 0$ and $f(1) = 1$. Substituting $x = 0$ into the formula (12), we get $A = 0$ and thus,

$$f(x) = \frac{B \cdot x}{C + D \cdot x}. \tag{13}$$

To simplify this expression, we can divide both the numerator and the denominator of this fraction by B and get

$$f(x) = \frac{x}{C_0 + D_0 \cdot x}, \tag{14}$$

where $C_0 \stackrel{\text{def}}{=} \dfrac{C}{B}$ and $D_0 \stackrel{\text{def}}{=} \dfrac{D}{B}$. Now, the condition that $f(1) = 1$ leads to $C_0 + D_0 = 1$, i.e., to $D_0 = 1 - C_0$ and

$$x \to f(x) = \frac{x}{x + C_0 \cdot (1-x)}. \tag{15}$$

The proposition is proven.

Now, we can formulate the same invariance ideas as for cumulative distribution functions.

Definition 7. *We say that the membership functions $\mu(x)$ and $\nu(x)$ are* equivalent *if for some real number r, we have:*

$$\nu(x) = \frac{\mu(x)}{\mu(x) + (1-r)\cdot(1-\mu(x))}.$$

Definition 8. *Let $f(x)$ be a transformation. We say that a membership function $\mu(x)$ is* invariant with respect to *f (or, f-*invariant*, for short) if the functions $\mu(f(x))$ and $\mu(x)$ are equivalent, i.e., if for some real number $r > 0$, we have*

$$\mu(f(x)) = \frac{\mu(x)}{\mu(x) + (1-r)\cdot(1-\mu(x))}.$$

Proposition 3. *For each membership function $\mu(x)$, the following two conditions are equivalent to each other:*

- *$\mu(x)$ is invariant with respect to all shifts;*
- *$\mu(x)$ is a logistic distribution, i.e., is described by the formula*

$$\mu(x) = \frac{1}{1 + C\cdot\exp(-k\cdot x)}. \tag{16}$$

Proof. From the mathematical viewpoint, this is exactly Proposition 1 which we have already proven.

4 Another Special Case

Idea. In the previous sections, we showed that invariance with respect to changing the starting point leads to the logistic distribution (and logistic membership function). A natural question is: what if instead, we require that the probability distribution be invariant with respect to changing the measuring unit, i.e., with respect to the scaling transformation $x \to c\cdot x$.

Proposition 4. *For each cumulative distribution function $F(x)$, the following two conditions are equivalent to each other:*

- *$F(x)$ is invariant with respect to all scalings;*
- *$F(x)$ is described by the formula*

$$\mu(x) = \frac{1}{1 + C\cdot x^{-k}}. \tag{17}$$

Proposition 5. *For each membership function $\mu(x)$, the following two conditions are equivalent to each other:*

- *$\mu(x)$ is invariant with respect to all scalings;*
- *$\mu(x)$ is described by the formula*

$$\mu(x) = \frac{1}{1 + C\cdot x^{-k}}. \tag{18}$$

Comment. The resulting formulas (17) and (18) form yet another limit case of the kappa-regression formulas (1) and (2).

Proof of Propositions 4 and 5. From the mathematical viewpoint, the probabilistic and fuzzy formulations are identical. so it is sufficient to prove this result in the probabilistic case. In this case, similar to the case of shift, we conclude that the original odds function $O(x)$ and the re-scaled function $O(c\cdot x)$ must be related by the Bayes formula

$$O(c\cdot x) = s(c)\cdot O(x). \tag{19}$$

The function $F(x)$ is monotonic hence measurable, thus the odds function is also measurable, and it is known (see, e.g., [1]) that all measurable solutions of the functional equation (12) have the form

$$O(x) = c\cdot x^k \tag{20}$$

for some values c and k. So, by using the formula (10), we can go from the odds to the probability distribution, and get

$$P(x) = \frac{1}{1 + \dfrac{1}{O(x)}} = \frac{1}{1 + C\cdot x^{-k}}, \tag{21}$$

where $C \stackrel{\text{def}}{=} \dfrac{1}{c}$.

The proposition is proven.

5 Towards the General Case

Analysis of the problem. The general kappa-regression-distribution is concentrated, with probability 1, on the interval (a,b). This means that in this case, we cannot apply shift-invariance – since there is a natural starting value a, and we cannot apply scale-invariance – since there is a natural measuring unit, e.g., the difference $b-a$. Since we cannot use the usual linear transformations $x \to x + x_0$ and $x \to c\cdot x$, if we want to use symmetries, we need to use some more general transformations.

What are more general transformations? We have already discussed the need to go beyond linear transformations in one of the previous sections, and we concluded that reasonable requirements lead to fractional-linear transformations – which we then called *reasonable*. Now, we are ready to formulate our main results.

Proposition 6. *Let $a < b$ For each cumulative distribution function $F(x)$, the following two conditions are equivalent to each other:*

- *$F(x)$ is invariant with respect to all reasonable transformations that preserve the interval $[a,b]$;*
- *$F(x)$ is a kappa-regression distribution, i.e., it is described by the formula (1).*

Proposition 7. *Let $a < b$. For each membership function $\mu(x)$, the following two conditions are equivalent to each other:*

- *$\mu(x)$ is invariant with respect to all reasonable transformations that preserve the interval $[a,b]$;*
- *$\mu(x)$ is a kappa-regression membership function, i.e., it is described by the formula (2).*

Proof. The general interval $[a,b]$ can be easily reduced to the interval $[0,1]$ by an appropriate linear transformation. Thus, in the following derivation, it is sufficient to consider the case when $a = 0$ and $b = 1$.

Similar to the previous cases, without losing generality, we can consider only the probabilistic case. In this case, the requirement is that the distribution $F(x)$ is equivalent to $F(f(x))$ for all reasonable transformations that preserve the interval $[0,1]$. We have shown that these transformations have the form (15).

Similar to the Bayes case, we can show that for the expression

$$T(x) \stackrel{\text{def}}{=} \frac{x}{1-x}, \tag{22}$$

which is similar to the expression for odds, the transformation (15) leads to $T(f(x)) = c \cdot T(x)$, for $c \stackrel{\text{def}}{=} \frac{1}{k}$. Thus, for the auxiliary function $G(z) \stackrel{\text{def}}{=} F(T^{-1}(z))$, we conclude that the distributions $G(z)$ and $G(c \cdot z)$ are equivalent to teach other for all $c > 0$. We already know, from Proposition 4, that in this case, the auxiliary function $G(z)$ is equal to

$$G(z) = \frac{1}{1 + C \cdot z^{-k}}.$$

Thus, for $F(x) = G(T(x))$, we get

$$F(x) = \frac{1}{1 + C \cdot T(x)^{-k}} = \frac{1}{1 + C \cdot \left(\frac{1-x}{x}\right)^k},$$

which is exactly the kappa-regression-expression for $a = 0$ and $b = 1$. A similar proof can be repeated for any $a < b$.

The proposition is proven.

Conclusion. We have explained the efficiency of kappa-regression distributions and kappa-regression membership functions – they are the only ones which satisfy the reasonable invariance conditions.

Acknowledgement

This work was supported in part by the grant TUDFO/47138-1/2019-ITM from the Ministry of Technology and Innovation, Hungary. It was also supported in part by the National Science Foundation grants 1623190 (A Model of Change for Preparing a New Generation for Professional Practice in Computer Science), and HRD-1834620 and HRD-2034030 (CAHSI Includes), and by the program of the development of the Scientific-Educational Mathematical Center of Volga Federal District No. 075-02-2020-1478.

References

[1] J. Aczél and J. Dhombres, Functional Equations in Several Variables, Cambridge University Press, 2008.

[2] J. Dombi and T. Jónás, Advances in the Theory of Probabilistic and Fuzzy Scientific Methods, with Applications, Springer, Cham, Switzerland, 2018.

[3] R. Feynman, R. Leighton, and M. Sands, The Feynman Lectures on Physics, Addison Wesley, Boston, Massachusetts, 2005.

[4] E. T. Jaynes and G. L. Bretthorst, Probability Theory: The Logic of Science, Cambridge University Press, Cambridge, UK, 2003.

[5] V. Kreinovich and C. Quintana. Neural networks: what non-linearity to choose?, Proc. 4th Univ. of New Brunswick AI Workshop, Fredericton, New Brunswick, Canada, 1991, pp. 627–637.

[6] H. T. Nguyen and V. Kreinovich, Applications of Continuous Mathematics to Computer Science, Kluwer, Dordrecht, Netherlands, 1997.

[7] D. J. Sheskin, Handbook of Parametric and Non-Parametric Statistical Procedures, Chapman & Hall/CRC, London, UK, 2011.

[8] K. S. Thorne and R. D. Blandford, Modern Classical Physics: Optics, Fluids, Plasmas, Elasticity, Relativity, and Statistical Physics, Princeton University Press, Princeton, New Jersey, 2017.

[9] J. C. Urenda, O. Csiszár, G. Csiszár, J. Dombi, O. Kosheleva, V. Kreinovich, and G. Eigner, Why

squashing functions in multi-layer neural networks, Proceedings of the 2020 IEEE International Conference on Systems, Man, and Cybernetics SMC'2020, Toronto, Canada, October 11–14, 2020, pp. 296–300.

[10] N. Wiener, Cybernetics: Or Control and Communication in the Animal and the Machine, MIT Press, Cambridge, Massachisetts, 1948.