Robustness as a Criterion for Selecting a Probability Distribution Under Uncertainty

Songsak Sriboonchitta[1], Hung T. Nguyen[1,2],
Vladik Kreinovich[3], and Olga Kosheleva[3]

[1] Faculty of Economics, Chiang Mai University
Chiang Mai, Thailand,
songsak@econ.chiangmai.ac.th
[2]Department of Mathematical Sciences
New Mexico State University
Las Cruces, New Mexico 88003, USA
hunguyen@nmsu.edu
[3]University of Texas at El Paso
500 W. University
El Paso, TX 79968, USA
vladik@utep.edu, olgak@utep.edu

Abstract

Often, we only have partial knowledge about a probability distribution, and we would like to select a single probability distribution $\rho(x)$ out of all probability distributions which are consistent with the available knowledge. One way to make this selection is to take into account that usually, the values x of the corresponding quantity are also known only with some accuracy. It is therefore desirable to select a distribution which is the most robust – in the sense the x-inaccuracy leads to the smallest possible inaccuracy in the resulting probabilities. In this paper, we describe the corresponding most robust probability distributions, and we show that the use of resulting probability distributions has an additional advantage: it makes related computations easier and faster.

1 Formulation of the Problem

Need to make decisions under uncertainty. One of the main objectives of science is to understand the world, to predict the future state of the world under different possible decisions – and then, to use these predictions to select the decision for which the corresponding prediction is the most preferable.

When we have the full knowledge of the situations, the problem of selecting the best decision becomes a straightforward optimization problem. In practice, however, we rarely have the full knowledge. Usually, we have some uncertainty about the future situations. It is therefore important to make decisions under uncertainty.

Traditional decision making assumes that we know the probabilities. There exist many techniques for decision making under uncertainty. Most of these techniques assume that we know the probabilities of different outcomes – i.e., in precise terms, that we know the probability distribution on the set of all possible outcomes; see, e.g., [2, 6, 7, 14].

In practice, we often have only partial knowledge about the probabilities. In many real-life random phenomena, we only have partial knowledge about the corresponding probability distributions. In such situations, several different probability distributions are consistent with the available data.

The resulting need to select a single probability distribution. As we have mentioned, most decision making techniques use a single probability distribution. So, to be able to apply these techniques to the practical situations, when *several* different probability distributions are consistent with our knowledge, we need to be able to select a *single* probability distribution – and use it in decision making.

What we do in this paper. To select a probability distribution, we can take into account that, in addition to imprecise knowledge about *probabilities* of different values of the corresponding quantity x (or quantities), we also have imprecise knowledge about the actual *values* of these quantities.

Indeed, the knowledge about these values comes from measurements, and measurements are never absolutely accurate: there is always a difference between the measurement result and the actual value, the difference known as the *measurement error*; see, e.g., [13]. In other words, when the measurement result is $\widetilde{x}$, the actual value x can be (and usually is) slightly different. It is therefore reasonable to select a probability distribution which is the most *robust*, i.e., for which the change from $\widetilde{x}$ to x has the smallest possible effect on the resulting probabilities.

In this paper, we show that this robustness idea indeed enables us to select a single distribution.

2 Robustness: From an Informal General Idea to a Precise Description

1-D case: analysis of the problem. Let us start with a 1-D case, when we have a single quantity x. In this case, we are interested in the probability of different events related to this quantity, i.e., in mathematical terms, in the probabilities of different subsets of the real line.

In many cases, it makes sense to limit ourselves to connected sets. In the 1-D case, the only connected sets are intervals $[\underline{x}, \overline{x}]$ (finite or infinite).

This make practical sense: e.g., it corresponds to checking whether x is larger than or equal to a certain lower threshold $\underline{x}$ and/or checking whether x is smaller than or equal to a certain upper threshold $\overline{x}$, or to checking whether x belongs to the given tolerance interval $[\underline{x}, \overline{x}]$.

From this viewpoint, all we need is for different intervals $[\underline{x}, \overline{x}]$, to find the probability that the value x belongs to this interval.

A 1-D probability distribution can be naturally described in terms of the corresponding probability density function $\rho(x)$. In terms of this function, the desired probability is equal to the integral $P = \int_{\underline{x}}^{\overline{x}} \rho(x)\,dx$.

Local robustness. As we have mentioned earlier, all the values of the quantity – in particular, the threshold values – and known with uncertainty. Let us consider, for example, the effect of uncertainty in $\underline{x}$ on the resulting probability. If we replace the value $\underline{x}$ with a slightly different value $\underline{x}' = \underline{x} + \Delta x$, then the original probability P changes to the slightly different probability

$$P' = \int_{\underline{x}+\Delta x}^{\overline{x}} \rho(x)\,dx = \int_{\underline{x}}^{\overline{x}} \rho(x)\,dx - \int_{\underline{x}}^{\underline{x}+\Delta x} \rho(x)\,dx = P - \int_{\underline{x}}^{\underline{x}+\Delta x} \rho(x)\,dx. \quad (1)$$

When the value Δx is small, we can, in the first approximation, ignore the changes of the function $\rho(x)$ on the narrow interval $[\underline{x}, \underline{x} + \Delta x]$ and thus, get $\int_{\underline{x}}^{\underline{x}+\Delta x} \rho(x)\,dx \approx \rho(\underline{x}) \cdot \Delta x$. Then, the resulting change in probability $\Delta P \stackrel{\text{def}}{=} P' - P$ can be described as $\Delta P \approx -\rho(\underline{x}) \cdot \Delta x$, so $|\Delta P| \approx \rho(\underline{x}) \cdot |\Delta x|$.

Thus, the effect of the uncertainty Δx (with which we know $\underline{x}$) on the change in probability P is determined by the value $\rho(\underline{x})$. Similarly, the effect of the uncertainty Δx with which we know $\overline{x}$ on the change in probability P is determined by the value $\rho(\overline{x})$.

We can summarize both cases by saying that for any point x, the the effect of the uncertainty Δx (with which we know x) on the change in probability P is determined by the value $\rho(x)$. This value $\rho(x)$ thus serves as a measure of local robustness at the point x.

From local robustness to global robustness. For different values x, the local robustness degree is, in general, different. To select a distribution, we need to combine these values into a single criterion.

Local robustness values are proportional to approximation errors caused by uncertainty Δx. There are two natural ways to combine different approximation errors:

- we can consider the worst-case error, or
- we can consider the mean squared error.

The worst-case error corresponds to selecting the largest possible value of the approximation error, i.e., in our terms, the largest possible value $\max\limits_{x} \rho(x)$.

The mean squared error means considering the mean value of the squared error, i.e., equivalently, of the squared coefficient $\rho^2(x)$. In contrast to the worst-case approach, where the global criterion is uniquely determined, here, we have two possible choices:

- we can interpret mean as the average over all possible x, i.e., as a quantity proportional to the integral $\int (\rho(x))^2\,dx$;
- alternatively, we can interpret mean as averaging over the probability distribution characterized by the probability density $\rho(x)$; in this case, as a criterion of global robustness, we get the quantity

$$\int \rho(x) \cdot (\rho(x))^2\,dx = \int (\rho(x))^3\,dx. \tag{2}$$

Thus, we arrive at the following conclusion.

Resulting criteria of global robustness. We have three possible choices of selecting the most robust probability distribution:

- we can select a probability distribution $\rho(x)$ for which the maximum $\max\limits_x \rho(x)$ attains the smallest possible value;
- we can select a probability distribution $\rho(x)$ for which the integral $\int (\rho(x))^2\,dx$ attains the smallest possible value; and
- we can select a probability distribution $\rho(x)$ for which the integral $\int (\rho(x))^3\,dx$ attains the smallest possible value.

Relation to maximum entropy approach. Traditionally in probability theory, when we only have partial knowledge about the probability distribution, we select a distribution for which the entropy $-\int \rho(x) \cdot \ln(\rho(x))\,dx$ attains the largest possible value (see, e.g., [3]), or, equivalently, for which the integral $\int \rho(x) \cdot \ln(\rho(x))\,dx$ attains the smallest possible value.

It is worth mentioning that, in general, if we assume that the criterion for selecting a probability distribution is scale-invariant (in some reasonable sense), then this criterion is equivalent to optimizing either entropy, or generalized entropy $\int \ln(\rho(x))\,dx$ or $\int \rho^\alpha(x)\,dx$, for some $\alpha > 0$; see, e.g., [5]. Our analysis shows that the generalized entropy corresponding to $\alpha = 2$ and $\alpha = 3$ describes mean-squared robustness.

The worst-case criterion can also be thus interpreted. Indeed, it is known that for non-negative values $v_1, \ldots, v_n$, we have

$$\max(v_1, \ldots, v_n) = \lim_{p \to \infty} ((v_1)^p + \ldots + (v_n)^p)^{1/p} \tag{3}$$

and similarly,

$$\max_x \rho(x) = \lim_{p \to \infty} \left(\int (\rho(x))^p\,dx \right)^{1/p}. \tag{4}$$

Thus, minimizing $\max\limits_{x} \rho(x)$ is, for large enough p, equivalent to minimizing the expression $\left(\int (\rho(x))^p\,dx\right)^{1/p}$ and hence, equivalent to minimizing the corresponding generalized entropy $\int (\rho(x))^p\,dx$.

Multi-D case. In the multi-D case, when the probability density function $\rho(x)$ depends on several variables $x = (x_1, \ldots, x_m)$, we can also consider general connected sets S. Similarly to the 1-D case, if we add, to the set S, a small neighborhood of a point x, of volume ΔV, then the resulting change in probability is equal to $\Delta P = \rho(x) \cdot \Delta V$. Vice versa, if the set S contained the point x with some neighborhood, and we delete an x-neighborhood of volume ΔV from the set S, then we get $\Delta P = -\rho(x) \cdot \Delta V$.

In both cases, we have $|\Delta P| = \rho(x) \cdot \Delta V$. Thus, in the multi-D case too, the value $\rho(x)$ serves as a measure of local robustness at a point x. So, when we apply the usual techniques for combining local robustness measures into a single global one, we get one of three criteria described above.

What we do in the following sections. Now that we know that we have three possible ways of selecting the most robust probability distribution, let us consider these three ways one by one. For each way, on several simple examples, we explain what exactly probability distribution will be thus selected.

Comment. It is worth mentioning that a similar idea of selecting the most robust description is actively used in fuzzy logic [4, 11, 16]; namely, in [8, 9, 10, 11], it is shown how we can select the most robust membership functions and the most robust "and"- and "or"-operations.

While our problem is different, several related formulas are similar – and this similarity helped us with our results.

3 Selecting a Probability Distribution that Minimizes $\int (\rho(x))^2\,dx$

General idea. In this section, we will describe, for several reasonable types of partial knowledge, which probability distribution corresponds to the smallest possible values of the global robustness criterion $\int (\rho(x))^2\,dx$.

Types of partial knowledge about the probability distribution. What type of partial knowledge do we have about a random variable? For example, about a random measurement error?

First, we can have lower and upper bounds on the measurement error (and, more generally, on the possible values of the random variable).

Second, we may know:

- the mean value, i.e., the first moment of the corresponding random variable,
- the variance (i.e., equivalently, the second moment),

- sometimes the skewness (i.e., equivalently, the third moment) that characterizes the distribution's asymmetry, and
- the excess (i.e., equivalently, the fourth moment) that describes how heavy are the distribution's tails.

In general, we will therefore consider the cases when we know the bounds and some moments (maybe none).

Simplest case, when we only know the bounds. Let us start with the simplest case, when we only know the bounds $\underline{a}$ and $\overline{a}$ on the values of the corresponding random variable x, i.e., we know that always $\underline{a} \le x \le \overline{a}$ and thus, that $\rho(x) = 0$ for values x outside the interval $[\underline{a}, \overline{a}]$.

In this case, the problem of selecting the most robust distribution takes the following form: minimize $\int_{\underline{a}}^{\overline{a}} (\rho(x))^2\,dx$ under the constraints that $\int_{\underline{a}}^{\overline{a}} \rho(x)\,dx = 1$ and $\rho(x) \ge 0$ for all x. To solve this constrained optimization problem, we can apply the Lagrange multiplier methods to reduce it to an easier-to-solve unconstrained optimization problem

$$\int_{\underline{a}}^{\overline{a}} (\rho(x))^2\,dx + \lambda \cdot \left(\int_{\underline{a}}^{\overline{a}} \rho(x)\,dx - 1 \right) \to \min_{\rho(x)}, \tag{5}$$

under the condition that $\rho(x) \ge 0$ for all x.

According to calculus, for every x, when the value $\rho(x)$ corresponding to the optimum is inside the corresponding range $(0, \infty)$, the derivative of the above objective function with respect to $\rho(x)$ should be equal to 0. Differentiating the above expression and equating its derivative to 0, we get $2\rho(x) + \lambda = 0$, hence $\rho(x) = c$ for some constant c (equal to $-\lambda/2$; strictly speaking, we should be talking here about *variational* derivative, not a regular derivative).

So, for every x from the interval $[\underline{a}, \overline{a}]$, $\rho(x) > 0$ implies that $\rho(x) = c$. In other words, for every $x \in [\underline{a}, \overline{a}]$, we have either $\rho(x) = 0$ or $\rho(x) = c$.

Let S denote the set of all the points $x \in [\underline{a}, \overline{a}]$ for which $\rho(x) > 0$. Let L denote the total length (1-D Lebesgue measure) of this set. Then, the condition $\int_{\underline{a}}^{\overline{a}} \rho(x)\,dx = \int_S \rho(x)\,dx = 1$ implies that $c \cdot L = 1$, hence $c = \dfrac{1}{L}$. Thus, the value of the desired objective function takes the form

$$\int_{\underline{a}}^{\overline{a}} (\rho(x))^2\,dx = L \cdot \left(\frac{1}{L} \right)^2 = \frac{1}{L}. \tag{6}$$

One can easily see that this value is the smallest if and only if the length L is the largest.

The largest possible length of a set $S \subseteq [\underline{a}, \overline{a}]$ is attained when this subset coincide with the interval – and is equal to the length $\overline{a} - \underline{a}$ of this interval. In this case, $\rho(x) = \text{const}$ for all points $x \in [\underline{a}, \overline{a}]$.

Thus, in this case, the most robust distribution is the uniform distribution on the interval $[\underline{a}, \overline{a}]$.

Comment. It is worth mentioning that in this case, when we only know the bounds $\underline{a}$ and $\overline{a}$ on the values of the corresponding random variable x, maximum entropy method leads to the exact same uniform distribution.

What if we also know the mean? Let us now consider the next case, when, in addition to the bounds $\underline{a}$ and $\overline{a}$ on the values of the corresponding random variable x, we also know its mean μ.

In this case, we need to minimize the functional $\int (\rho(x))^2\,dx$ under the constraints $\int \rho(x)\,dx = 1$, $\int x \cdot \rho(x)\,dx = \mu$, and $\rho(x) \geq 0$. By using the Lagrange multiplier method, we can reduce this constraint optimization problem to the following unconstrained optimization problem:

$$\int_{\underline{a}}^{\overline{a}} (\rho(x))^2\,dx + \lambda \cdot \left(\int_{\underline{a}}^{\overline{a}} \rho(x)\,dx - 1 \right) + \lambda_1 \cdot \left(\int_{\underline{a}}^{\overline{a}} x \cdot \rho(x)\,dx - \mu \right) \to \min \quad (7)$$

under the constraint that $\rho(x) \geq 0$ for all $x \in [\underline{a}, \overline{a}]$.

Similarly to the previous case, for the points x for which $\rho(x) > 0$, the derivative of the above expression relative to $\rho(x)$ should be equal to 0, so we conclude that for some x, we have $\rho(x) = p_0 + q \cdot x$ for appropriate constants $p_0 = -\lambda/2$ and $q = -\lambda_1/2$. In other words, the probability density $\rho(x)$ is either determined by a linear expression or it is equal to 0.

One can check that, in general, the desired minimum is attained when

$$\rho(x) = \max(0, p_0 + q \cdot x). \quad (8)$$

In particular, in the case when $\rho(x) > 0$ for all $x \in [\underline{a}, \overline{a}]$, the probability density function $\rho(x)$ is linear for all $x \in [\underline{a}, \overline{a}]$: $\rho(x) = p_0 + q \cdot x$. We can get explicit expressions for p_0 and q if we reformulate this linear function in an equivalent form $\rho(x) = \rho_0 + q \cdot (x - \widetilde{a})$, where $\widetilde{a} \stackrel{\text{def}}{=} \dfrac{\underline{a} + \overline{a}}{2}$ is the interval's midpoint. In this case, the condition $\int_{\underline{a}}^{\overline{a}} \rho(x)\,dx = 1$ takes the form $\int_{-\Delta}^{\Delta} (\rho_0 + q \cdot t)\,dt = 1$, where $t \stackrel{\text{def}}{=} x - \widetilde{a}$ and $\Delta \stackrel{\text{def}}{=} \dfrac{\overline{a} - \underline{a}}{2}$ is the half-width (radius) of the interval $[\underline{a}, \overline{a}]$. The integral of an odd function t over a symmetric interval $[-\Delta, \Delta]$ is equal to 0, so we have $2\Delta \cdot \rho_0 = 1$ and thus,

$$\rho_0 = \frac{1}{2\Delta} = \frac{1}{\overline{a} - \underline{a}}, \quad (9)$$

exactly the value corresponding to the uniform distribution on the interval $[\underline{a}, \overline{a}]$.

The value q can be determined by the condition $\int_{\underline{a}}^{\overline{a}} x \cdot \rho(x)\,dx = \mu$. Since $\int_{\underline{a}}^{\overline{a}} \rho(x)\,dx = 1$, this condition is equivalent to $\int_{\underline{a}}^{\overline{a}} (x - \widetilde{a}) \cdot \rho(x)\,dx = \mu - \widetilde{a}$ and thus, to

$$\int_{-\Delta}^{\Delta} t \cdot (\mu_0 + q \cdot t)\,dt = \int_{-\Delta}^{\Delta} (t \cdot \mu_0 + q \cdot t^2)\,dt = \mu - \widetilde{a}. \quad (10)$$

Here similarly, the integral of t is equal to 0, and the integral of t^2 is equal to

$$\int_{-\Delta}^{\Delta} t^2\,dt = \left.\frac{t^3}{3}\right|_{-\Delta}^{\Delta} = \frac{2\Delta^3}{3}, \quad (11)$$

thus the above condition leads to $q \cdot \dfrac{2\Delta^3}{3} = \mu - \widetilde{a}$ and to

$$q = \frac{3(\mu - \widetilde{a})}{2\Delta^3}. \tag{12}$$

Substituting, into this formulas, the definition of half-width in terms of the bounds $\underline{a}$ and $\overline{a}$, we get an equivalent formula

$$q = \frac{12 \cdot (\mu - \widetilde{a})}{(\overline{a} - \underline{a})^3}. \tag{12a}$$

The resulting linear formula $\rho(x) = \rho_0 + q \cdot (x - \widetilde{a})$ only works when the resulting expression is non-negative for all x, i.e., when $\rho_0 + q \cdot t \geq 0$ for all $t \in [-\Delta, \Delta]$. This, in its turn, it equivalent to $\rho_0 \geq |q| \cdot \Delta$, i.e., to $\dfrac{1}{2\Delta} \geq \dfrac{3 \cdot |\mu - \widetilde{a}|}{2\Delta^2}$, and, equivalently, to $|\mu - \widetilde{a}| \leq \dfrac{1}{3} \cdot \Delta$.

When $|\mu - \widetilde{a}| > \dfrac{1}{3} \cdot \Delta$, we have to consider probability density functions $\rho(x)$ which are equal to 0 on some subinterval of the interval $[\underline{a}, \overline{a}]$. For a random variable $x \in [\underline{a}, \overline{a}]$, its means value μ also has to be within the same interval, so we must have $\mu \in [\underline{a}, \overline{a}]$ and $\mu - \widetilde{a} \in [-\Delta, \Delta]$.

- When $\mu - \widetilde{a} \to \Delta$, i.e., when $\mu \to \overline{a}$, the corresponding probability distribution get concentrated on a narrower and narrower interval containing the point $x = \overline{a}$.
- Similarly, when $\mu - \widetilde{a} \to -\Delta$, i.e., when $\mu \to \underline{a}$, the corresponding probability distribution get concentrated on a narrower and narrower interval containing the point $x = \underline{a}$.

Comment. If instead of our robustness criterion, we would look for the probability distribution with the largest entropy, then the corresponding derivative would take a form $-\ln(\rho(x)) - 1 + \lambda + \lambda_1 \cdot x = 0$, so $\ln(\rho(x)) = a + b \cdot x$, where $a = \lambda - 1$ and $b = \lambda_1$, and we would get an exponential distribution $\rho(x) = \exp(a + b \cdot x)$.

What if we also know the first two moments? Let us now consider the next case, when, in addition to the bounds $\underline{a}$ and $\overline{a}$ on the values of the corresponding random variable x, and the mean μ, we also know the second moment M_2 – or, equivalently, the variance $V = \sigma^2 = V - \mu^2$.

In this case, we need to minimize the functional $\int_{\underline{a}}^{\overline{a}} (\rho(x))^2\, dx$ under the constraints $\int_{\underline{a}}^{\overline{a}} \rho(x)\, dx = 1$, $\int_{\underline{a}}^{\overline{a}} x \cdot \rho(x)\, dx = \mu$, $\int_{\underline{a}}^{\overline{a}} x^2 \cdot \rho(x)\, dx = M_2$, and $\rho(x) \geq 0$. By using the Lagrange multiplier method, we can reduce this constraint optimization problem to the following unconstrained optimization problem:

$$\int_{\underline{a}}^{\overline{a}} (\rho(x))^2\, dx + \lambda \cdot \left(\int_{\underline{a}}^{\overline{a}} \rho(x)\, dx - 1 \right) +$$

$$\lambda_1 \cdot \left(\int_{\underline{a}}^{\overline{a}} x \cdot \rho(x)\, dx - \mu \right) + \lambda_2 \cdot \left(\int_{\underline{a}}^{\overline{a}} x^2 \cdot \rho(x)\, dx - M_2 \right) \to \min \tag{13}$$

under the constraint that $\rho(x) \geq 0$ for all $x \in [\underline{a}, \overline{a}]$.

Similarly to the previous case, for the points x for which $\rho(x) > 0$, the derivative of the above expression relative to $\rho(x)$ should be equal to 0, so we conclude that for some x, we have $\rho(x) = p_0 + q \cdot x + r \cdot x^2$, where $p_0 = -\lambda/2$, $q = -\lambda_1/2$, and $r = -\lambda_2/2$. In other words, the probability density $\rho(x)$ is either determined by a quadratic expression or it is equal to 0. One can check that, in general, the desired minimum is attained when

$$\rho(x) = \max(0, p_0 + q \cdot x + r \cdot x^2). \tag{14}$$

It should be mentioned that for the maximum entropy case, similar arguments lead to the Gaussian distribution $\rho_G(x) = \text{const} \cdot \exp\left(-\dfrac{(x-\mu_0)^2}{2\sigma_0^2} \right)$ truncated to some interval $[\underline{b}, \overline{b}] \subseteq [\underline{a}, \overline{a}]$ of the given interval $[\underline{a}, \overline{a}]$: $\rho(x) = \rho_G(x)$ for $x \in [\underline{b}, \overline{b}]$ and $\rho(x) = 0$ for all other x.

Let us consider particular cases. When $r < 0$, we get a bell-shaped distribution – i.e., somewhat similar in shape to the Gaussian distribution. However, the new distribution has several advantages over the Gaussian distribution:

- first, the new distribution is more robust – it is actually the most robust of all the distributions on the given interval with the given two moments (this is how we selected it);
- second, the new probability distribution function $\rho(x)$ is continuous on the entire real line – while, due to the fact that the probability density of ta Gaussian distribution is always positive, the pdf of the truncated Gaussian distribution is discontinuous at the endpoints $\underline{b}$ and $\overline{b}$ of the corresponding interval;
- third, the new distribution is computationally easier, since computation with polynomials (e.g., computing probability over different intervals or different moments) is much easier than computation with the Gaussian pdf.

When the variance is sufficiently high, we get $r > 0$, which corresponds to a *bimodal distribution*. Bimodal distributions are common in measuring instruments (see, e.g., [12]). There are two main reasons for the bimodal distribution. The first is the effect of the sinusoid signal in the electric grid. Electric grids are ubiquitous, and the electromagnetic field created by the electric plugs affects all electromagnetic devices. The resulting noise is proportional to $\sin(\omega \cdot t)$ at a random time t – and the resulting random variable indeed has a bimodal distribution.

The second reason is related to the very process of manufacturing the corresponding measuring instrument. Indeed, usually, we have a desired upper bound

on the measurement error. At first, the measurement error of the newly manufactured measuring instrument is normally distributed. This can be explained by the fact that there are many different independent factors that contribute to this original measurement error and thus, due to the Central Limit Theorem, we expect the overall effect of these factors to be approximately normally distributed; see, e.g., [13, 15]. However, the range of the corresponding errors Δx is usually much wider than the desired tolerance bounds. Thus, the manufacturers start tuning the instrument until it fits into the bounds; this tuning stops as soon as we get into the desired intervals $[-\Delta, \Delta]$. As a result:

- all the cases when originally, we had $\Delta x \le -\Delta$ are converted to $-\Delta$ and
- all the cases when originally, we had $\Delta x \ge \Delta$ are converted to Δ.

Hence, the vicinities of the two extreme values $-\Delta$ and Δ get a high probability — and thus, very high values of probability density $\rho(x)$. So, we get a distribution which is either bimodal or even tri-modal (with a smaller original peak).

In our robust approach, we cover bimodal distributions by using the same easy-to-process quadratic formulas as the more usual unimodal ones – a clear advantage over the more traditional approach, when bimodal distributions are modeled by using much more computationally complex expressions.

What if we also know higher moments? In many cases, we also know higher moments. For example, often, we know third and/or fourth moments, i.e., equivalently, skewness and excess. For such situations, traditionally, there are no easy-to-use expression. However, in our case, we do get such an expression.

Namely, let us now consider the case, when, in addition to the bounds $\underline{a}$ and $\overline{a}$ on the values of the corresponding random variable x, we also know the values of the first m moments $\int_{\underline{a}}^{\overline{a}} x^k \cdot \rho(x)\,dx = M_i$, $k = 1, 2, \ldots, m$.

In this case, we need to minimize the functional $\int_{\underline{a}}^{\overline{a}} (\rho(x))^2\,dx$ under the constraints $\int_{\underline{a}}^{\overline{a}} \rho(x)\,dx = 1$, $\int_{\underline{a}}^{\overline{a}} x^k \cdot \rho(x)\,dx = M_k$ for $k = 1, \ldots, m$, and $\rho(x) \ge 0$ for all x. By using the Lagrange multiplier method, we can reduce this constraint optimization problem to the following unconstrained optimization problem:

$$\int_{\underline{a}}^{\overline{a}} (\rho(x))^2\,dx + \lambda \cdot \left(\int_{\underline{a}}^{\overline{a}} \rho(x)\,dx - 1 \right) + \sum_{k=1}^{m} \lambda_k \cdot \left(\int_{\underline{a}}^{\overline{a}} x^k \cdot \rho(x)\,dx - M_k \right) \to \min \tag{15}$$

under the constraint that $\rho(x) \ge 0$ for all $x \in [\underline{a}, \overline{a}]$.

For the points x for which $\rho(x) > 0$, the derivative of the above expression relative to $\rho(x)$ should be equal to 0, so we conclude that for some x, we have $\rho(x) = p_0 + \sum\limits_{k=1}^{m} q_k \cdot x^k$, where $p_0 = -\lambda/2$ and $q_k = -\lambda_k/2$. In other words, the

probability density $\rho(x)$ is either determined by a polynomial expression or it is equal to 0. One can check that, in general, the desired minimum is attained when

$$\rho(x) = \max\left(0, p_0 + \sum_{k=1}^{m} q_k \cdot x^k\right). \tag{16}$$

This polynomial expression is easy to process, so we have a distribution whose processing is computationally easy – as opposed to the usual not-so-computationally easy approaches of dealing with, e.g., skew-normal distributions [1].

Multi-D case: good news. What is we want to analyze a joint distribution of several variables? Similarly to the 1-D case, if we know several moments, then the most robust pdf $\rho(x_1, \ldots, x_d)$ on a given box $[\underline{a}_1, \overline{a}_1] \times \ldots \times [\underline{a}_d, \overline{a}_d]$ is described by a polynomial, or, to be more precise, by an expression

$$\rho(x_1, \ldots, x_d) = \max(0, P(x_1, \ldots, x_d)) \tag{17}$$

for some polynomial $P(x_1, \ldots, x_d)$.

The degree of this polynomial depends on what moments we know:

- if we do not know any moments, then $P(x_1, \ldots, x_d)$ is a constant, and thus, we get a uniform distribution – similarly to what we get if we use the maximum entropy approach;
- if we only know the means $E[x_i]$, then $P(x_1, \ldots, x_d)$ is a linear function;
- if we also know second moments $E[(x_i)^2]$ and $E[x_i \cdot x_j]$ – i.e., equivalently, the covariance matrix – then $P(x_1, \ldots, x_d)$ is a quadratic function;
- if we also know third (and fourth) order moments, then $P(x_1, \ldots, x_d)$ is a cubic (quartic) polynomial, etc.

These polynomial pdf's are not only more robust, but they are also much easier to process than Gaussian or other usually used pdf's.

But maybe we are missing something? Not really, since, as it is well known, polynomials are *universal approximators* – in the sense that any arbitrary continuous function on a given box can be, with any desired accuracy, approximated by a polynomial.

Multi-D case: remaining challenges. While, as we have mentioned on several example, the robust approach to selecting a probability distribution has many advantages over the maximum entropy approach, there are situations in which the use of the robust approach faces some challenges.

One such situation is when we know the marginal distributions $\rho_1(x_1)$ and $\rho_2(x_2)$, and we need to reconstruct the original 2-D distribution $\rho(x_1, x_2)$. In the usual maximum entropy approach, the corresponding optimization problem leads to the independence-related formula $\rho(x_1, x_2) = \rho_1(x_1) \cdot \rho_2(x_2)$; see, e.g., [3]. This makes perfect sense: if we know nothing about the relation

between two random variables, it is reasonable to assume that they are independent.

For our robust approach, however, the situation is less intuitive. Specifically, we want to find a distribution $\rho(x_1, x_2) \geq 0$ on the box $A = [\underline{a}_1, \overline{a}_1] \times [\underline{a}_2, \overline{a}_2]$ for which the following conditions are satisfied:

- $\int_A \rho(x_1, x_2)\, dx_1\, dx_2 = 1$,
- $\int_{\underline{a}_2}^{\overline{a}_2} \rho(x_1, x_2)\, dx_2 = \rho_1(x_1)$ for all x_2, and
- $\int_{\underline{a}_1}^{\overline{a}_1} \rho(x_1, x_2)\, dx_1 = \rho_2(x_2)$ for all x_2.

(Strictly speaking, we do not need the first condition, since it automatically follows from, e.g., the second one if we integrate both sides over x_1.

For this constraint optimization problem, the Lagrange multiplier technique means minimizing the functional

$$\int_A (\rho(x_1, x_2))^2\, dx_1\, dx_2 + \int_{\underline{a}_1}^{\overline{a}_1} dx_1\, \lambda_1(x_1) \cdot \left(\int_{\underline{a}_2}^{\overline{a}_2} \rho(x_1, x_2)\, dx_2 \right) +$$

$$\int_{\underline{a}_2}^{\overline{a}_2} dx_2\, \lambda_2(x_2) \cdot \left(\int_{\underline{a}_1}^{\overline{a}_1} \rho(x_1, x_2)\, dx_1 \right)$$

for appropriate values $\lambda_i(x_i)$. When $\rho(x_1, x_2) > 0$, differentiating this objective function with respect to $\rho(x_1, x_2)$ leads to $\rho(x_1, x_2) = a_1(x_1) + a_2(x_2)$, where $a_1(x_1) = -\lambda_1(x_1)/2$ and $a_2(x_2) = -\lambda_2(x_2)/2$.

In general, we get

$$\rho(x_1, x_2) = \max(0, a_1(x_1) + a_2(x_2)). \tag{18}$$

In particular, when $\rho(x_1, x_2) > 0$ for all $x_1 \in [\underline{a}_1, \overline{a}_1]$ and $x_2 \in [\underline{a}_2, \overline{a}_2]$, then we get $\rho(x_1, x_2) = a_1(x_1) + a_2(x_2)$. Integrating over x_2, we conclude that $\rho_1(x_1) = (\overline{a}_2 - \underline{a}_2) \cdot a_1(x_1) + C_1$, where $C_1 \stackrel{\text{def}}{=} \int_{\underline{a}_1}^{\overline{a}_1} a_2(x_2)$. Thus, $a_1(x_1) = \dfrac{1}{\overline{a}_2 - \underline{a}_2} \cdot \rho_1(x_1) + C_1$, for some constant C_1.

Similarly, we get $a_2(x_2) = \dfrac{1}{\overline{a}_1 - \underline{a}_1} \cdot \rho_2(x_2) + C_2$, for some constant C_1. So,

$$\rho(x_1, x_2) = \frac{1}{\overline{a}_2 - \underline{a}_2} \cdot \rho_1(x_1) + \frac{1}{\overline{a}_1 - \underline{a}_1} \cdot \rho_2(x_2) + C, \tag{19}$$

where $C \stackrel{\text{def}}{=} C_1 + C_2$. We can find the constant C if we integrate both sides of this equality over all $x_1 \in [\underline{a}_1, \overline{a}_1]$ and all $x_2 \in [\underline{a}_2, \overline{a}_2]$; we then get

$$1 = 1 + 1 + C \cdot (\overline{a}_1 - \underline{a}_1) \cdot (\overline{a}_2 - \underline{a}_2). \tag{20}$$

Thus, $C = -\dfrac{1}{(\overline{a}_1 - \underline{a}_1) \cdot (\overline{a}_2 - \underline{a}_2)}$ and so,

$$\rho(x_1, x_2) = \frac{1}{\overline{a}_2 - \underline{a}_2} \cdot \rho_1(x_1) + \frac{1}{\overline{a}_1 - \underline{a}_1} \cdot \rho_2(x_2) - \frac{1}{(\overline{a}_1 - \underline{a}_1) \cdot (\overline{a}_2 - \underline{a}_2)}. \quad (21)$$

When both x_1 and x_2 are uniformly distributed, the result is the uniform distribution on the box in which the random variables x_1 and x_2 are independent – similarly to the maximum entropy approach. However, in general, this is *not* independence, it is a *mixture* of the two distributions – and it is not very clear what is the intuitive meaning of this mixture.

4 Selecting a Probability Distribution that Minimizes $\int (\rho(x))^3\, dx$

General idea. As we have mentioned earlier, one of the possible ways to describe robustness is to select the probability distribution corresponds to the smallest possible values of the global robustness criterion $\int (\rho(x))^3\, dx$.

Simplest case, when we only know the bounds. Let us start with the simplest case, when we only know the bounds $\underline{a}$ and $\overline{a}$ on the values of the corresponding random variable x, i.e., we know that always $\underline{a} \le x \le \overline{a}$ and thus, that $\rho(x) = 0$ for values x outside the interval $[\underline{a}, \overline{a}]$.

In this case, the problem of selecting the most robust distribution takes the following form: minimize $\int_{\underline{a}}^{\overline{a}} (\rho(x))^3\, dx$ under the constraints that $\int_{\underline{a}}^{\overline{a}} \rho(x)\, dx = 1$ and $\rho(x) \ge 0$ for all x. To solve this constrained optimization problem, we can apply the Lagrange multiplier methods to reduce it to an easier-to-solve unconstrained optimization problem

$$\int_{\underline{a}}^{\overline{a}} (\rho(x))^3\, dx + \lambda \cdot \left(\int_{\underline{a}}^{\overline{a}} \rho(x)\, dx - 1 \right) \to \min_{\rho(x)}. \quad (22)$$

When $\rho(x) > 0$, then differentiation over $\rho(x)$ leads to $3(\rho(x))^2 + \lambda = 0$, i.e., to $\rho(x) = c$, where $c = \sqrt{-\lambda}$.

Similarly to the case of the criterion $\int (\rho(x))^2\, dx$, we can conclude that the smallest value of the robustness criterion is attained when $\rho(x) > 0$ for all $x \in [\underline{a}, \overline{a}]$, i.e., when we have a uniform distribution on the given interval. In other words, in this simplest case, we have the same distribution as when we use the first robustness criterion or when we use the maximum entropy approach.

What if we also know several moments? Let us now consider the case, when, in addition to the bounds $\underline{a}$ and $\overline{a}$ on the values of the corresponding random variable x, we also know the values of the first m moments

$$\int_{\underline{a}}^{\overline{a}} x^k \cdot \rho(x)\, dx = M_i, \quad k = 1, 2, \ldots, m. \quad (23)$$

In this case, we need to minimize the functional $\int_{\underline{a}}^{\overline{a}} (\rho(x))^3\,dx$ under the constraints $\int_{\underline{a}}^{\overline{a}} \rho(x)\,dx = 1$, and $\int_{\underline{a}}^{\overline{a}} x^k \cdot \rho(x)\,dx = M_k$ for $k = 1, \ldots, m$. By using the Lagrange multiplier method, we can reduce this constraint optimization problem to the following unconstrained optimization problem:

$$\int_{\underline{a}}^{\overline{a}} (\rho(x))^3\,dx + \lambda \cdot \left(\int_{\underline{a}}^{\overline{a}} \rho(x)\,dx - 1 \right) + \sum_{k=1}^{m} \lambda_k \cdot \left(\int_{\underline{a}}^{\overline{a}} x^k \cdot \rho(x)\,dx - M_k \right) \to \min \tag{24}$$

under the constraint that $\rho(x) \geq 0$ for all $x \in [\underline{a}, \overline{a}]$.

For the points x for which $\rho(x) > 0$, the derivative of the above expression relative to $\rho(x)$ should be equal to 0, so we conclude that for some x, we have $(\rho(x))^2 = p_0 + \sum\limits_{k=1}^{m} q_k \cdot x^k$, where $p_0 = -\lambda/3$ and $q_k = -\lambda_k/3$. In other words, the probability density $\rho(x)$ is either determined by a square root of a polynomial expression or it is equal to 0. One can check that, in general, the desired minimum is attained when

$$\rho(x) = \sqrt{\max\left(0, p_0 + \sum_{k=1}^{m} q_k \cdot x^k\right)}. \tag{25}$$

Similarly, in the multi-D case, we get

$$\rho(x_1, \ldots, x_d) = \sqrt{\max(0, P(x_1, \ldots, x_d)}, \tag{26}$$

for an appropriate polynomial $P(x_1, \ldots, x_d)$.

The results of this approach are less desirable than the results of using the first robustness criterion. From the computational viewpoint, integrating polynomials is easy, but integrating square roots of polynomials is not easy. From this viewpoint, the first robustness criterion – that was analyzed in the previous section – is much more computationally advantageous that the second robustness criterion that we analyze in this section.

It turns out that square roots also lead to less accurate approximations. Let us illustrate it on the example of approximating a Gaussian distribution by a quadratic polynomial vs. by a square root of a quadratic polynomial. Without losing generality, we can restrict ourselves to a standard normal distribution with 0 mean and standard deviation 1, for which the probability density is proportional to

$$f(x) \stackrel{\text{def}}{=} \exp\left(-\frac{x^2}{2}\right) = 1 - \frac{x^2}{2} + \frac{x^4}{8} + \ldots \tag{27}$$

If we approximate this expression in the vicinity of 0, then the best quadratic approximation corresponds to taking the first two terms in the above Taylor

expansion $f_1(x) = 1 - \dfrac{x^2}{2}$, and the accuracy $\delta_1 = |f_1(x) - f(x)|$ of this approximation is largely determined by the first ignored terms: $\delta_1 \approx \dfrac{x^4}{8}$.

On the other hand, if we use square roots of quadratic polynomials, then, due to the symmetry of the problem with respect to the transformation $x \to -x$, we have to use symmetric quadratic polynomials $a \cdot (1 + b \cdot x^2)$. For this polynomial, we have $\sqrt{a \cdot (1 + b \cdot x^2)} = \sqrt{a} \cdot \left(1 + \dfrac{b \cdot x^2}{2}\right) + \ldots$ For these terms to coincide with the first two terms in the Taylor expansion of the function $f(x)$, we must therefore take $a = 1$ and $b = -1$. For the resulting approximating function $f_2(x) = \sqrt{1 - x^2}$, we have

$$f_2(x) = \sqrt{1 - x^2} = 1 - \frac{x^2}{2} - \frac{x^4}{8} + \ldots \tag{28}$$

Here, the approximation accuracy is equal to $f_2(x) - f(x) = -\dfrac{x^4}{8} + \ldots$ So asymptotically, the approximation error has the form $\delta_2 = |f_2(x) - f(x)| \approx \dfrac{x^4}{4}$ – which is twice larger than when we use the first robustness criterion

$$\int (\rho(x))^2 \, dx \to \min. \tag{29}$$

5 Selecting a Probability Distribution that Minimizes $\max\limits_x \rho(x)$

Reminder. If we use the worst-case description of robustness, then we should select a distribution $\rho(x)$ for which the value $\max\limits_x \rho(x)$ is the smallest possible.

Analysis of the problem. In this case, whatever moments conditions we impose, if there is a point x_0 in the vicinity of which $0 < \rho(x_0) < \max\limits_x \rho(x)$, then we can decrease all the value $\rho(x)$ for which $\rho(x) = \max$ by some small amount – compensating it with an appropriate increase in the vicinity of x_0, and satisfy the same criteria while decreasing the value $\max\limits_x \rho(x)$.

Thus, when the desired criterion $\max\limits_x \rho(x)$ is the smallest possible, then for every x, we either have $\rho(x) = 0$ or $\rho(x)$ is equal to this maximum.

This somewhat informal argument can be formally confirmed if we take into account that, that we have mentioned earlier, the worst-case criterion can be viewed as a limit, when $p \to \infty$, of the criteria $\int_{\underline{a}}^{\overline{a}} (\rho(x))^p \, dx \to \min$.

Let us thus consider the case, when, in addition to the bounds $\underline{a}$ and $\overline{a}$ on the values of the corresponding random variable x, we also know the values of the first m moments $\int_{\underline{a}}^{\overline{a}} x^k \cdot \rho(x) \, dx = M_i$, $k = 1, 2, \ldots, m$. In this case, we need to minimize the functional $\int_{\underline{a}}^{\overline{a}} (\rho(x))^p \, dx$ under the constraints $\int_{\underline{a}}^{\overline{a}} \rho(x) \, dx = 1$,

and $\int_{\underline{a}}^{\overline{a}} x^k \cdot \rho(x)\,dx = M_k$ for $k = 1, \ldots, m$. By using the Lagrange multiplier method, we can reduce this constraint optimization problem to the following unconstrained optimization problem:

$$\int_{\underline{a}}^{\overline{a}} (\rho(x))^p\,dx + \lambda \cdot \left(\int_{\underline{a}}^{\overline{a}} \rho(x)\,dx - 1 \right) +$$

$$\sum_{k=1}^{m} \lambda_k \cdot \left(\int_{\underline{a}}^{\overline{a}} x^k \cdot \rho(x)\,dx - M_k \right) \to \min \tag{30}$$

under the constraint that $\rho(x) \geq 0$ for all $x \in [\underline{a}, \overline{a}]$.

For the points x for which $\rho(x) > 0$, the derivative of the above expression relative to $\rho(x)$ should be equal to 0, so we conclude that for some x, we have $(\rho(x))^{p-1} = p_0 + \sum\limits_{k=1}^{m} q_k \cdot x^k$, where $p_0 = -\lambda/p$ and $q_k = -\lambda_k/p$. Thus, for such points x, we have $\rho(x) = \text{const} \cdot (P(x))^{1/(p-1)}$ for some polynomial $P(x)$. When $p \to \infty$, we have $1/(p-1) \to 0$, and the 0-th power of a positive number is always 1. Thus, we indeed have $\rho(x) = \text{const}$ whenever $\rho(x) > 0$. A similar conclusion can be made in the multi-D case. So, we arrive at the following conclusion.

Resulting formulas. For the worst-case robustness criterion, for the optimal distribution $\rho(x)$, the probability density is either equal to 0, or to some constant.

To be more precise, both in the 1-D case and in the multi-D case, the zone at which $\rho(x) > 0$ is determined by some polynomial $P(x)$, i.e., we have:

- $\rho(x) = 0$ when $P(x) \leq 0$ and
- $\rho(x) = c$ when $P(x) > 0$.

The value c is determined by the condition that the total probability should be equal to 1: $\int \rho(x)\,dx = 1$, hence $c = \dfrac{1}{A}$, where A is the Lebesque measure (length, areas, volume, etc., depending on the dimension d) of the set of all the points $x = (x_1, \ldots, x_d)$ for which $P(x_1, \ldots, x_d) > 0$.

Shall we recommend this approach? It depends on what we want:

- If the goal is to get a good approximation to the original cdf, then clearly no: in contrast to polynomials, these functions do not have a universal approximation property.
- On the other hand, in critical situations, when we want to minimize worst-case dependence on the input's uncertainty, these are the distributions that we should use.

Acknowledgments

We acknowledge the partial support of the Center of Excellence in Econometrics, Faculty of Economics, Chiang Mai University, Thailand. This work was also supported in part by the National Science Foundation grants HRD-0734825 and HRD-1242122 (Cyber-ShARE Center of Excellence) and DUE-0926721, and by an award "UTEP and Prudential Actuarial Science Academy and Pipeline Initiative" from Prudential Foundation.

References

[1] A. Azzalini and A. Capitanio, *The Skew-Normal and Related Families*, Cambridge University Press, Cambridge, Massachusetts, 2013.

[2] P. C. Fishburn, *Utility Theory for Decision Making*, John Wiley & Sons Inc., New York, 1969.

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

[4] G. Klir and B. Yuan, "Fuzzy Sets and Fuzzy Logic", Prentice Hall, Upper Saddle River, New Jersey, 1995.

[5] V. Kreinovich, O. Kosheleva, H. T. Nguyen, and S. Sriboonchitta, "Why some families of probability distributions are practically efficient: a symmetry-based explanation", In: V. N. Huynh, V. Kreinovich, and S. Sriboonchitta (eds.), *Causal Inference in Econometrics*, Springer Verlag, Cham, Switzerland, 2016, pp. 133–152.

[6] R. D. Luce and R. Raiffa, *Games and Decisions: Introduction and Critical Survey*, Dover, New York, 1989.

[7] H. T. Nguyen, O. Kosheleva, and V. Kreinovich, "Decision making beyond Arrow's 'impossibility theorem', with the analysis of effects of collusion and mutual attraction", *International Journal of Intelligent Systems*, 2009, Vol. 24, No. 1, pp. 27–47.

[8] H. T. Nguyen, V. Kreinovich, B. Lea, and D. Tolbert, "How to control if even experts are not sure: robust fuzzy control", *Proceedings of the Second International Workshop on Industrial Applications of Fuzzy Control and Intelligent Systems*, College Station, Texas, December 2–4, 1992, pp. 153–162.

[9] H. T. Nguyen, V. Kreinovich, and D. Tolbert, "On robustness of fuzzy logics", *Proceedings of the 1993 IEEE International Conference on Fuzzy Systems FUZZ-IEEE'93*, San Francisco, California, March 1993, Vol. 1, pp. 543–547.

[10] H. T. Nguyen, V. Kreinovich, and D. Tolbert, "A measure of average sensitivity for fuzzy logics", *International Journal on Uncertainty, Fuzziness, and Knowledge-Based Systems*, 1994, Vol. 2, No. 4, pp. 361–375.

[11] H. T. Nguyen and E. A. Walker, *A First Course in Fuzzy Logic*, Chapman and Hall/CRC, Boca Raton, Florida, 2006.

[12] P. V. Novitskii and I. A. Zograph, *Estimating the Measurement Errors*, Energoatomizdat, Leningrad, 1991 (in Russian).

[13] S. G. Rabinovich, *Measurement Errors and Uncertainty. Theory and Practice*, Springer Verlag, Berlin, 2005.

[14] H. Raiffa, *Decision Analysis*, Addison-Wesley, Reading, Massachusetts, 1970.

[15] D. J. Sheskin, *Handbook of Parametric and Nonparametric Statistical Procedures*, Chapman & Hall/CRC, Boca Raton, Florida, 2011.

[16] L. A. Zadeh, "Fuzzy sets", *Information and Control*, 1965, Vol. 8, pp. 338–353.