Robustness as a Criterion for Selecting a Probability Distribution Under Uncertainty

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

[1]Faculty of Economics, Chiang Mai University
Chiang Mai, Thailand, songsakecon@gmail.com
[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, El Paso, TX 79968, USA
olgak@utep.edu, vladik@utep.edu

1. Outline

- Often, we only have partial knowledge about a probability distribution.
- We would like to select a single probability distribution $\rho(x)$ out of all possible ones.
- *Idea:* take into account that the values x of the corresponding quantity are only known with some accuracy.
- It is therefore desirable to select a distribution which is the most robust – in the sense that:
  - x-inaccuracy
  - leads to the smallest possible inaccuracy in the resulting probabilities.
- We describe the most robust probability distributions, and we show that they make computations faster.

2. 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 to use these predictions to select the decision with the best results.
- When we have the full knowledge of the situations, we get an 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.

3. 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.
- In practice, we often have only partial knowledge about the probabilities.
- In such situations, several different probability distributions are consistent with the available data.
- To apply the existing techniques, we need:
  - to select a *single* probability distribution, and
  - to use this distribution in decision making.

4. Our Idea

- We have partial knowledge about *probabilities* of different values of the corresponding quantity x.
- We also have imprecise knowledge about the actual *values* of these quantities.
- Indeed, the knowledge about these values comes from measurements.
- There is always a difference between the measurement result $\widetilde{x}$ and the actual value x.
- 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.

5. Robustness: From an Informal Idea to a Precise Description

- 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.
- In mathematical terms, we are interested 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).

6. Robustness for Intervals

- 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 $[\underline{x}, \overline{x}]$.
- So, given an interval $[\underline{x}, \overline{x}]$, we want to compute:

$$P = \text{Prob}(X \in [\underline{x}, \overline{x}]) = \int_{\underline{x}}^{\overline{x}} \rho(x)\,dx.$$

7. Local Robustness

- Let us consider 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 P changes to:

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

$$P - \int_{\underline{x}}^{\underline{x}+\Delta x} \rho(x)\,dx.$$

- When Δx is small, $\int_{\underline{x}}^{\underline{x}+\Delta x} \rho(x)\,dx \approx \rho(\underline{x}) \cdot \Delta x$, so for $\Delta P \stackrel{\text{def}}{=} P' - P$, we have $|\Delta P| \approx \rho(\underline{x}) \cdot |\Delta x|$.
- Similarly, the effect of the uncertainty Δx in $\overline{x}$ on the change in probability P is $|\Delta P| \approx \rho(\overline{x}) \cdot |\Delta x|$.
- In both cases, the value $\rho(x)$ serves as a measure of local robustness at the point x.

8. 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.
- 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 means minimizing the largest possible approximation error $\sim \max\limits_x \rho(x)$.
- The mean squared case means minimizing the mean value of the squared error $\sim \rho^2(x)$.

9. Towards Global Robustness (cont-d)

- In the mean case, we have two possible choices:
  - we can interpret mean as the average $\int (\rho(x))^2\,dx$ over all possible values x, or
  - we can interpret mean as averaging over the random variable with pdf $\rho(x)$:

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

- In this talk, we consider all three options of selecting the distribution $\rho(x)$:
  - $\rho(x)$ for which $\max\limits_x \rho(x)$ is the smallest;
  - $\rho(x)$ for which $\int (\rho(x))^2\,dx$ is the smallest, and
  - $\rho(x)$ for which $\int (\rho(x))^3\,dx$ is the smallest.

10. 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 $S \stackrel{\text{def}}{=} -\int \rho(x) \cdot \ln(\rho(x))\, dx$ is the largest.
- It is known that:
  - if we assume that the criterion for selecting a probability distribution is scale-invariant,
  - then this criterion is equivalent to optimizing either entropy, or generalized entropy

$$\int \ln(\rho(x))\, dx \text{ or } \int \rho^\alpha(x)\, dx, \text{ for some } \alpha > 0.$$

11. Relation to Maximum Entropy Approach (cont-d)

- We have either entropy, or generalized entropy $\int \ln(\rho(x))\,dx$ or $\int \rho^\alpha(x)\,dx$ for some $\alpha > 0$.
- The generalized entropy corresponding to $\alpha = 2$ and $\alpha = 3$ describes mean-squared robustness.
- The worst-case criterion is a limit:

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

- Thus, minimizing $\max_x \rho(x)$ is, for large enough p:
  - equivalent to minimizing $\left(\int (\rho(x))^p\,dx\right)^{1/p}$ and hence,
  - equivalent to minimizing $\int (\rho(x))^p\,dx$.

12. Multi-D Case

- In the multi-D case, the probability density function $\rho(x)$ depends on several variables $x = (x_1, \ldots, x_m)$.
- We can also consider general connected sets S.
- If we add, to S, a small neighborhood of x, of volume ΔV, then the change in P is $\Delta P = \rho(x) \cdot \Delta V$.
- Vice versa, if 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.
- We can apply the usual techniques for combining local robustness measures into a single global one.
- So, we get one of three above criteria.

13. What We Do Next

- 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 simple examples, we explain what exactly probability distribution will be selected.
- *Observation:* a similar idea of selecting the most robust description is actively used in fuzzy logic:
  - we can select the most robust membership functions;
  - we can select the most robust "and"- and "or"-operations.
- While our problem is different, several related formulas are similar.
- This similarity helped us with our results.

14. Types of Partial Knowledge About the Probability Distribution

- What type of partial knowledge do we have about a random variable?
- We can have lower and upper bounds on the measurement error; we may also 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 asymmetry,
  - the excess (i.e., equivalently, the fourth moment) that describes heavy tails.
- In general, we will therefore consider the cases when we know the bounds and some moments (maybe none).

15. Minimizing $\int (\rho(x))^2\,dx$: 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 x.
- In this case, we 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.
- Lagrange multiplier method reduces this problem to an easier-to-solve unconstrained optimization problem:

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

- Differentiating with respect to $\rho(x)$ and equating derivative to 0, we get $2\rho(x) + \lambda = 0$.
- So, $\rho(x) = c$ – a uniform distribution.
- *Comment:* in this case, maximum entropy method leads to the exact same uniform distribution.

16. What If We Also Know the Mean?

- What if we know the bounds $\underline{a}$ and $\overline{a}$ and mean μ?
- Then, we minimize $\int (\rho(x))^2\,dx$ under the constraints $\int \rho(x)\,dx = 1$, $\int x\cdot\rho(x)\,dx = \mu$, and $\rho(x) \ge 0$.
- Lagrange multiplier method leads to

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

$$\lambda_1\cdot\left(\int_{\underline{a}}^{\overline{a}} x\cdot\rho(x)\,dx - \mu\right) \to \min.$$

- When $\rho(x) > 0$, the derivative of J relative to $\rho(x)$ is 0 hence $\rho(x) = p_0 + q\cdot x$.
- Thus, $\rho(x) = \max(0, p_0 + q\cdot x)$.
- When $\rho(x) > 0$ for all $x \in [\underline{a}, \overline{a}]$, then $p_0 = 1/(\overline{a} - \underline{a})$ (same as for uniform distribution), $q = \dfrac{12\cdot(\mu - \widetilde{a})}{(\overline{a} - \underline{a})^3}$.

17. What If We Also Know the Mean (cont-d)

- The condition $\rho(x) \geq 0$ holds when $|\mu - \widetilde{a}| \leq \dfrac{1}{6} \cdot (\overline{a} - \underline{a})$, i.e., when

$$\widetilde{a} - \frac{1}{6} \cdot (\overline{a} - \underline{a}) \leq \mu \leq \widetilde{a} + \frac{1}{6} \cdot (\overline{a} - \underline{a}).$$

- When μ is smaller, $\rho(x)$ is concentrated on a narrower interval containing $x = \overline{a}$.
- When μ is larger, $\rho(x)$ is concentrated on a narrower interval containing $x = \overline{a}$.
- *Comment:* in this case, maximum entropy leads to

$$-\ln(\rho(x)) - 1 + \lambda + \lambda_1 \cdot x = 0, \text{ so}$$

$$\rho(x) = \exp(a + b \cdot x).$$

18. What If We Know the First Two Moments?

- Sometimes, in addition to the bounds $\underline{a}$ and $\overline{a}$ and the mean μ, we also know the second moment M_2.
- Knowing M_2 is equivalent to knowing the variance

$$V = \sigma^2 = V - \mu^2.$$

- In this case, similar arguments lead to

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

- *Comment:* in this case, maximum entropy leads to Gaussian distribution.
- When $r < 0$, we get a bell-shaped distribution – i.e., somewhat similar in shape to the truncated Gaussian.

Need to Make . . .
Traditional Decision . . .
Our Idea
Robustness: From an . . .
From Local . . .
Types of Partial . . .
Minimizing . . .
A Distribution that . . .
A Distribution for . . .
Home Page
Title Page
◀◀ ▶▶
◀ ▶
Page 19 of 40
Go Back
Full Screen
Close
Quit

19. Advantages of the New Approach

- First, the new distribution is more robust – it is actually the most robust.
- Second, the new probability distribution function $\rho(x)$ is continuous on the entire real line.
- In contrast, the truncated Gaussian distribution is discontinuous at the endpoints.
- Third, the new distribution is computationally easier:
  - computation with polynomials (e.g., computing probabilities over intervals or 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*.

20. Bimodal Distributions (cont-d)

- Bimodal distributions are common in measuring instruments, for two reasons.
- The first is the effect of the sinusoid signal $\sin(\omega \cdot t)$ (at a random time t) in the electric grid;
- 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.
- Many independent factors contribute to this original measurement error.
- Thus, due to the Central Limit Theorem, we the overall effect of these factors is approximately normal.

21. Bimodal Distributions (cont-d)

- 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 extreme values $-\Delta$ and Δ get high values of pdf $\rho(x)$.
- So, we get a distribution which is either bimodal or even tri-modal (with a smaller original peak).

22. Advantages of Our Approach: Case of Bimodal Distributions

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

23. 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.
- If we know the first m moments, then we get

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

- This polynomial expression is easy to process.
- In contrast, the usual approaches like skew normal are not so computationally easy.

24. Multi-D Case: Good News

- What if 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 $\rho(x_1, \ldots, x_d) = \max(0, P(x_1, \ldots, x_d))$ 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.
- This is similar 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.

25. Multi-D Case (cont-d)

- If we also know $E[(x_i)^2]$ and $E[x_i \cdot x_j]$ (equivalently, the covariance matrix), $P(x_1, \ldots, x_d)$ is quadratic.
- 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 the most robust.
- 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.

26. Multi-D Case: Remaining Challenge

- Suppose that we only know the marginal distributions $\rho_1(x_1)$ and $\rho_2(x_2)$.
- We need to reconstruct the original 2-D distribution $\rho(x_1, x_2)$.
- In the maximum entropy approach, the corresponding optimization leads to the independence-related formula

$$\rho(x_1, x_2) = \rho_1(x_1) \cdot \rho_2(x_2).$$

- 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.

27. Multi-D Challenge (cont-d)

- We want to find $\rho(x_1, x_2) \geq 0$ on the box $A = [\underline{a}_1, \overline{a}_1] \times [\underline{a}_2, \overline{a}_2]$ for which $\int_A \rho(x_1, x_2)\, dx_1\, dx_2 = 1$,
  - $\int_{\underline{a}_2}^{\overline{a}_2} \rho(x_1, x_2)\, dx_2 = \rho_1(x_1)$ for all x_1, and
  - $\int_{\underline{a}_1}^{\overline{a}_1} \rho(x_1, x_2)\, dx_1 = \rho_2(x_2)$ for all x_2.
- Here, the first condition follows from the second one if we integrate over x_1.
- For this constraint optimization problem, the Lagrange multiplier technique means minimizing the functional

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

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

- When $\rho(x_1, x_2) > 0$, equating derivative to 0 leads to $\rho(x_1, x_2) = a_1(x_1) + a_2(x_2)$, where $a_i(x_i) = -\lambda_i(x_i)/2$.

28. Multi-D Challenge (cont-d)

- Integrating over x_2, we get $\rho_1(x_1) = (\overline{a}_2 - \underline{a}_2) \cdot a_1(x_1) + C_1$, hence $a_1(x_1) = \dfrac{1}{\overline{a}_2 - \underline{a}_2} \cdot \rho_1(x_1) + C_1$ and

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

- When both x_1 and x_2 are uniformly distributed, the result is the uniform distribution on the box.
- This is similar to the maximum entropy approach.
- However, in general, this is *not* independence, it is a *mixture* of the two distributions.
- It is not very clear what is the intuitive meaning of this mixture.
- In general, we get $\rho(x_1, x_2) = \max(a_1(x_1) + a_2(x_2), 0)$.

29. A Distribution that $\int (\rho(x))^3\,dx \to \min$

- If we only know the bounds, we minimize $\int_{\underline{a}}^{\overline{a}} (\rho(x))^3\,dx$ under the constraints $\int_{\underline{a}}^{\overline{a}} \rho(x)\,dx = 1$ and $\forall x\,\rho(x) \ge 0$.
- Lagrange multiplier methods leads to:

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

- The solution is a uniform distribution.
- If we also know moments, we minimize $\int_{\underline{a}}^{\overline{a}} (\rho(x))^3\,dx$ under the constraints $\int_{\underline{a}}^{\overline{a}} \rho(x)\,dx = 1$, and

$$\int_{\underline{a}}^{\overline{a}} x^k \cdot \rho(x)\,dx = M_k \text{ for } k = 1, \ldots, m.$$

30. A Distribution that $\int (\rho(x))^3\,dx \to \min$ (cont-d)

$$\text{So } \int_{\underline{a}}^{\overline{a}} (\rho(x))^3\,dx + \lambda \cdot \left(\int_{\underline{a}}^{\overline{a}} \rho(x)\,dx - 1\right) +$$

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

- When $\rho(x) > 0$, differentiation leads to

$$(\rho(x))^2 = p_0 + \sum_{k=1}^{m} q_k \cdot x^k.$$

- So, $\rho(x) = \sqrt{\max\left(0, p_0 + \sum\limits_{k=1}^{m} q_k \cdot x^k\right)}.$
- Similarly, in the multi-D case, for an appropriate polynomial $P(x_1, \ldots, x_d)$, we get:

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

31. Limitations of This Approach

- From the computational viewpoint:
  - integrating polynomials is easy, but
  - integrating square roots of polynomials is not easy.
- From this viewpoint, the first robustness criterion is more computationally advantageous.
- Square roots also lead to less accurate approximations.
- Let us illustrate it on the example of approximating a Gaussian $N(0,1)$:

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

32. Limitations (cont-d)

- If we approximate this expression in the vicinity of 0, then:
  - the best quadratic approximation takes the first two terms in the Taylor expansion

$$f_1(x) = 1 - \frac{x^2}{2};$$

  - the approximation accuracy $\delta_1 = |f_1(x) - f(x)|$ is determined by the first ignored term: $\delta_1 \approx \dfrac{x^4}{8}$.
- What if we use square roots of quadratic polynomials?
- The problem is symmetric with respect to $x \to -x$.
- So, we have to use symmetric quadratic polynomials

$$a \cdot (1 + b \cdot x^2).$$

33. Limitations (cont-d)

- For this polynomial $a \cdot (1 + b \cdot x^2)$, we have:

$$\sqrt{a \cdot (1 + b \cdot x^2)} = \sqrt{a} \cdot \left(1 + \frac{b \cdot x^2}{2}\right) + \ldots$$

- For these terms to coincide with the first two terms in the Taylor expansion of $f(x)$, we must take

$$a = 1 \text{ and } b = -1.$$

- For the resulting approximating function $f_2(x) = \sqrt{1 - x^2}$, we have

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

- Here, the approximation accuracy is equal to

$$f_2(x) - f(x) = -\frac{x^4}{8} + \ldots$$

34. Limitations (final)

- So asymptotically:
  - if we use the criterion $\int (\rho(x))^3\,dx \to \min$,
  - then the approximation error has the form

$$\delta_2 = |f_2(x) - f(x)| \approx \frac{x^4}{4}.$$

- This is twice larger than when we use the first robustness criterion $\int (\rho(x))^2\,dx \to \min$.

35. A Distribution for which $\max\limits_x \rho(x) \to \min$

- 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 so
  - satisfying 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.

36. Case of $\max\limits_x \rho(x) \to \min$ (cont-d)

- This argument can be formally confirmed.
- Indeed, the worst-case criterion is a limit, when $p \to \infty$, of $\int_{\underline{a}}^{\overline{a}} (\rho(x))^p\,dx \to \min$.
- Let's minimize $\int_{\underline{a}}^{\overline{a}} (\rho(x))^p\,dx$ under the constraints $\int_{\underline{a}}^{\overline{a}} \rho(x)\,dx = 1$, and

$$\int_{\underline{a}}^{\overline{a}} x^k \cdot \rho(x)\,dx = M_k \text{ for } k = 1, \ldots, m.$$

- Lagrange multiplier method leads to:

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

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

37. Case of $\max\limits_x \rho(x) \to \min$ (final)

- When $\rho(x) > 0$, equating derivative to 0 leads to

$$(\rho(x))^{p-1} = p_0 + \sum_{k=1}^{m} q_k \cdot x^k.$$

- So $\rho(x) = \text{const} \cdot (P(x))^{1/(p-1)}$ 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: $\rho(x) = 0$ when $P(x) \le 0$, $\rho(x) = c$ when $P(x) > 0$.

38. 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.

39. Acknowledgments

- We acknowledge the support of the Center of Excellence in Econometrics, Chiang Mai University.
- 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 from the Prudential Foundation.