Why Some Families of Probability Distributions Are Practically Efficient: A Symmetry-Based Explanation

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

[1]University of Texas at El Paso,
El Paso, TX 79968, USA
vladik@utep.edu, olgak@utep.edu
[2]Department of Mathematical Sciences
New Mexico State University
Las Cruces, NM 88003, USA, hunguyen@nmsu.edu
[3]Faculty of Economics, Chiang Mai University
Chiang Mai, Thailand,
songsakecon@gmail.com

1. Formulation of the Problem

- Theoretically, we can have infinite many different families of probability distributions.
- In practice, only a few families have been empirically successful.
- For some of these families, there is a good theoretical explanation for their success.
- For example, the Central Limit theorem explains the ubiquity of normal distributions.
- However, for many other families, there is no theoretical explanation for their empirical success.
- In this talk, we provide a theoretical explanation of their success.

2. Our Main Idea

- We are looking for a family which is the best among all the families that satisfy appropriate constraints.
- So, we need to select:
  - objective functions and
  - constraints.
- The numerical value of each quantity x depends:
  - on the starting point for measurement and
  - on the choice of the measuring unit.
- If we change the starting point to the one x_0 units smaller, then all the values shift by x_0: $x \to x + x_0$.
- Similarly, if we change the original measuring unit by a one λ times smaller, then $x \to \lambda \cdot x$: 2 m = 200 cm.
- Shifts and scaling do not change the physical quantities – just change the numbers.

3. Main Idea (cont-d)

- Shifts and scaling do not change the physical quantities – just change the numbers.
- It is therefore reasonable to require that objective functions and constraints are shift- and scale-invariant.
- We look for distributions which are optimal w.r.t. invariant objective functions under invariant constraints.
- It turns out that the resulting optimal families indeed include many empirically successful families of distributions.
- Thus, our approach explains the empirical success of many such families.
- This approach is in good accordance with modern physics, where symmetries are ubiquitous.

4. Which Objective Functions Are Invariant?

- According to decision theory, decisions of a rational agent are equivalent to maximizing *utility*.
- It is reasonable to require that:
  - if have two distribution which differ only in some local region,
  - and the first distribution is better, then
  - if we replace a common distribution outside this region by another common distribution,
  - the first distribution will still be better.
- It is known that each utility function with this property is either a sum or a product of functions $A(\rho(x), x)$.
- Maximizing the product is equivalent to maximizing its logarithm: the sum of logarithms.

5. Invariant Objective Functions (cont-d)

- Thus, the general expression of an objective function with the above "localness" property is $\int A(\rho(x), x)\,dx$.
- Shift-invariance implies no explicit dependence on x:

$$u = \int A(\rho(x))\,dx.$$

- Scaling $y = \lambda \cdot x$ changes $\rho(x)$ to $\lambda^{-1} \cdot \rho(\lambda^{-1} \cdot y)$.
- We require that if $\int A(\rho(x))\,dx = \int A(\rho'(x))\,dx$, then this equality remains after re-scaling.
- This requirement leads to:
  - entropy $S = -\int \rho(x) \cdot \ln(\rho(x))\,dx$ and
  - generalized entropy: $\int \ln(\rho(x))\,dx$ and $\int (\rho(x))^{\alpha}\,dx$.

6. Which Constraints Are Invariant: Definitions

- Decision making is based on expected values of utility.
- So, we consider constraints of the type

$$\int f_i(x) \cdot \rho(x)\, dx = c_i.$$

- We says that constraints corr. to $f_i(x)$ are *shift-invariant* if:
  - the values of the corr. quantities $\int f_i(x) \cdot \rho(x)\, dx$
  - uniquely determine the values of these quantities for a shifted distribution.
- We says that constraints corr. to $f_i(x)$ are *scale-invariant* if:
  - the values of the corr. quantities $\int f_i(x) \cdot \rho(x)\, dx$
  - uniquely determine the values of these quantities for a scaled distribution.

7. Which Constraints Are Invariant: Results

- Functions $f_i(x)$ corresponding to shift-invariant constraints are linear combinations of the functions

$$x^k \cdot \exp(a \cdot x) \cdot \sin(\omega \cdot x + \varphi), \quad k = 0, 1, 2, \ldots$$

- Functions $f_i(x)$ corresponding to scale-invariant constraints are linear combinations of the functions

$$(\ln(x - x_0))^k \cdot (x - x_0)^a \cdot \sin(\omega \cdot \ln(x - x_0) + \varphi).$$

- Only functions which are both shift- and scale-invariant are polynomials.
- We optimize:
  - an invariant objective function $J(\rho)$
  - under the constraint $\int \rho(x)\,dx = 1$ and
  - under the constraints $\int f_i(x) \cdot \rho(x)\,dx = c_i$ for invariant functions $f_i(x)$.

8. Optimal Distributions: General Formula

- The Lagrange multiplier methods means optimizing

$$J(\rho)+\lambda\cdot\left(\int \rho(x)\,dx-1\right)+\sum_i\left(\int f_i(x)\cdot\rho(x)\,dx-c\right).$$

- Differentiating this expression with respect to $\rho(x)$ and equating the resulting derivative to 0, we get:

$\ln(\rho(x)) = -1+\lambda+\sum\limits_i \lambda_i\cdot f_i(x)$ for the usual entropy;

$-(\rho(x))^{-1} = \lambda+\sum\limits_i \lambda_i\cdot f_i(x)$ for $J(\rho)=\int \ln(\rho(x))\,dx;$ and

$(-\alpha)\cdot(\rho(x))^{\alpha-1} = \lambda+\sum\limits_i \lambda_i\cdot f_i(x)$ for $J(\rho)=\int (\rho(x))^{\alpha}\,dx.$

- This is how we will explain all empirically successful distributions.

9. All Constraints Are Both Shift- and Scale-Invariant, Objective Function is Entropy

- In this case, $f_i(x)$ are polynomials $P_i(x)$, and equation is
$$\ln(\rho(x)) = -1 + \lambda + \sum_i \lambda_i \cdot P_i(x).$$
- The right-hand side of this formula is a polynomial $P(x)$, so $\rho(x) = \exp(P(x))$.
- The most widely used distribution, the normal distribution, is exactly of this type:
$$\rho(x) = \frac{1}{\sqrt{2\pi}} \cdot \exp\left(-\frac{(x-\mu)^2}{\sigma^2}\right).$$
- It is a known fact that it has the largest entropy among all the distributions with given mean and variance.

10. Constraints Are Shift- and Scale-Invariant, Objective Function: Generalized Entropy

- General formulas: $(\rho(x))^{-1} = \lambda + \sum\limits_i \lambda_i \cdot P_i(x) = 0$ or $-\alpha \cdot (\rho(x))^{\alpha-1} = \lambda + \sum\limits_i \lambda_i \cdot P_i(x) = 0.$
- In both cases, $\rho(x) = (P(x))^\beta$ for some polynomial $P(x)$.
- *Example:* Cauchy distribution ($\beta = 1$):
$$\rho(x) = \frac{\Delta}{\pi} \cdot \frac{1}{1 + \dfrac{(x-\mu)^2}{\Delta^2}}.$$
- This distribution is actively used:
  - in physics, to describe resonance energy distribution and the corr. widening of spectral lines; and
  - to estimate the uncertainty of the results of data processing.

11. Constraints Scale-Invariant With Same Value x_0, Objective Function: Entropy

- *Generalized Gamma distribution:*
  - for $f_1(x) = \ln(x)$, $f_2(x) = x^\alpha$, and
  - for a scale-invariant constraint corresponding to $x \geq 0$,
  - we get $\ln(\rho(x)) = \lambda + \lambda_1 \cdot \ln(x) + \lambda_2 \cdot x^\alpha$ and

$$\rho(x) = \text{const} \cdot x^{\lambda_1} \cdot \exp(\lambda_2 \cdot x^\alpha).$$

- This distribution is efficiently used in survival analysis in social sciences.
- Several probability distributions are particular cases of this general formula.
- χ^2*:* when λ_1 is a natural number and $\alpha = 2$.
- This distribution is used to check how well the model fits the data.

12. Generalized Gamma Distribution (cont-d)

- As *Nakagami distribution*, χ^2 is used to model attenuation of wireless signals traversing multiple paths.
- *Inverse Gamma distribution:* $\alpha = -1$, used as a prior in Bayesian analysis, e.g., as a prior for variance.
- In particular, when $2\lambda_1$ is a negative integer, we get the *scaled-inverse chi-square distribution.*
- *Exponential distribution:* $\lambda_1 = 0$, $\alpha = 1$,

$$\rho(x) = \text{const} \cdot \exp(-k \cdot x).$$

- This distribution describe the time between consecutive events (queuing theory, radioactive decay, etc.).
- *Gamma distribution:* $\alpha = 1$, used as a prior distribution in Bayesian analysis.

13. Generalized Gamma Distribution (cont-d)

- In particular, when $\lambda_1 = k$ is a natural number, we get the *Erlang distribution*.
- It describes the time during which k consecutive events occur.
- *Fréchet distribution:* $\lambda_1 = 0$, describes the frequency of extreme events, such as:
  - the yearly maximum and minimum stock prices in economics,
  - yearly maximum rainfalls in hydrology, etc.
- *Half-normal distribution:* When $\lambda_1 = 0$ and $\alpha = 2$.
- *Rayleigh distribution:* $\lambda_1 = 1$ and $\alpha = 2$.
- It is used to describe the length of random vectors – e.g., the distribution of wind speed in meteorology.

14. Constraints Scale-Invariant With Same Value x_0, Objective: Entropy (cont-d)

- *Type-2 Gumbel (Weibull) distribution:* $\lambda_1 = \alpha - 1$.
- It is used to describe the frequency of extreme events and time to failure.
- *3-parametric Gamma distribution:* $f_1(x) = \ln(x - \mu)$, $f_2(x) = (x - \mu)^\alpha$, and $x \geq \mu$:

$$\rho(x) = \text{const} \cdot (x - \mu)^{\lambda_1} \cdot \exp(\lambda_2 \cdot (x - \mu)^\alpha).$$

- It is efficiently used in hydrology.
- *Inverse Gaussian (Wald) distribution:* $f_1(x) = \ln(x)$, $f_2(x) = x$, $f_3(x) = x^{-1}$, and$x > 0$:

$$\rho(x) = \text{const} \cdot x^{\lambda_1} \cdot \exp(\lambda_2 \cdot x + \lambda_3 \cdot x^{-1}).$$

- For $\lambda_1 = -1.5$, we get the *inverse Gaussian (Wald) distribution.*

15. Constraints Scale-Invariant With Same Value x_0, Objective: Entropy (cont-d)

- Wald's $\rho(x)$ describes the time a Brownian Motion with positive drift takes to reach a fixed positive level.
- *Laplace distribution:* $f_1(x) = |x - \mu|$, so

$$\rho(x) = \text{const} \cdot \exp(\lambda_1 \cdot |x - \mu|).$$

- It is used, e.g.:
  - in *speech recognition*, as a prior distribution for the Fourier coefficients;
  - in *databases*, where, to preserve privacy, each record is modified by adding a Laplace-generated noise.
- *Lévy (van der Waals) distribution:* $f_1(x) = \ln(x - \mu)$, $f_2(x) = (x - \mu)^{-1}$, $x - \mu > 0$, so

$$\rho(x) = \text{const} \cdot (x - \mu)^{\lambda_1} \cdot \exp(\lambda_2 \cdot (x - \mu)^{-1}).$$

16. Constraints Scale-Invariant With Same Value x_0, Objective: Entropy (end)

- We have $\rho(x) = \text{const} \cdot (x-\mu)^{\lambda_1} \cdot \exp(\lambda_2 \cdot (x-\mu)^{-1})$.
- For $\lambda_1 = -1.5$, we get the *Lévy (van der Waals) distribution* describing spectra.
- *Log-normal distribution:* $f_1(x) = \ln(x)$, $f_2(x) = (\ln(x))^2$, $x > 0$.
- This distribution describes the product of several independent random factors.
- It is used in econometrics to describe:
  - the compound return of a sequence of multiple trades,
  - a long-term discount factor, etc.

17. All Constraints Are Shift-Invariant, Objective Function Is Entropy

- *Gumbel distribution:* $f_1(x) = \exp(k \cdot x)$, so

$$\rho(x) = \text{const} \cdot \exp(\lambda_1 \cdot \exp(k \cdot x)).$$

- It is used to describe the frequency of extreme events.
- *Type I Gumbel distribution:*
  - $f_1(x) = x$,
  - $f_2(x) = \exp(k \cdot x)$, so
  - $\rho(x) = \text{const} \cdot \exp(\lambda_1 \cdot x + \lambda_2 \cdot \exp(k \cdot x))$.
- For $\lambda_1 = k$, we get *type I Gumbel distribution* which is used to describe frequencies of extreme values.

18. All Constraints Are Shift-Invariant, Objective Function Is Generalized Entropy

- Objective function $\int \ln(\rho(x))\,dx$.
- Shift-invariant constraint
$$f_1(x) = \exp(k \cdot x) + \exp(-k \cdot x).$$
- Result: $(\rho(x))^{-1} = -\lambda - \lambda_1 \cdot (\exp(k \cdot x) + \exp(-k \cdot x)) = -\lambda + c \cdot \cosh(k \cdot x)$.
- So $\rho(x) = \text{const} \cdot \dfrac{1}{-\lambda + c \cdot \cosh(k \cdot x)}$.
- The requirement that $\int \rho(x)\,dx = 1$ leads to $\lambda = 0$, so we get a *hyperbolic secant distribution*.
- This distribution is similar to the normal one, but it has a more acute peak and heavier tails.
- It is used when we have a distribution which is close to normal but has heavier tails.

19. Different Constraints Have Different Symmetries, Objective Function Is Entropy

- Sometimes, to get the desired distribution, we need constraints with symmetries of different type.
- *Uniform distribution:* constraints leading to $x \geq a$ and $x \leq b$ lead to uniform distribution on the interval $[a, b]$.
- *Comment:* the same result holds if we use generalized entropy.
- *Beta distribution:*
  - constraints $\int \ln(x) \cdot \rho(x)\,dx$ is scale-invariant relative to $x_0 = 0$,
  - constraint $\int \ln(a - x) \cdot \rho(x)\,dx$ is scale-invariant relative to $x_0 = a$,
  - we add $0 \leq x \leq a$,
  - result: Beta $\rho(x) = A \cdot x^\alpha \cdot (a - x)^\beta$.

20. Different Constraints Have Different Symmetries, Objective Function: Entropy (cont-d)

- B is used in agriculture, epidemiology, geosciences, meteorology, population genetics, project management.
- For $a = 1$ and $\alpha = \beta = 0.5$, we get the *arcsine distribution* $\rho(x) = \dfrac{1}{\pi \cdot \sqrt{x \cdot (1 - x)}}$.
- It describes, e.g., the measurement error caused by an external sinusoidal signal.
- *Beta prime (F-) distribution:* $f_1(x) = \ln(x)$, $f_2(x) = \ln(x + a)$, $x > 0$, lead to $\rho(x) = \text{const} \cdot x^{\lambda_1} \cdot (x + a)^{\lambda_2}$.
- *Log distribution:* $f_1(x) = x$, $f_2(x) = \ln(x)$, $x \geq a$, and $x \leq b$ lead to $\rho(x) = \text{const} \cdot \exp(\lambda_1 \cdot x) \cdot x^{\lambda_2}$.
- For $\lambda_1 = -1$, we get the *log distribution.*

21. Different Constraints Have Different Symmetries, Objective Function: Entropy (cont-d)

- *Generalized Pareto distribution:* $f_1(x) = \ln(x + x_0)$, $x > x_m$, lead to $\rho(x) = \text{const} \cdot (x + x_0)^{\lambda_1}$.
- It describes the frequency of large deviations in economics, in geophysics, etc.
- The case $x_0 = 0$ is known as the *Pareto distribution.*
- *Comment.* The Generalized Pareto distribution can also be derived by using generalized entropy.
- *Gompertz distribution:* $f_1(x) = \exp(b \cdot x)$, $f_2(x) = x$, and $x > 0$ lead to

$$\rho(x) = \text{const} \cdot \exp(\lambda_1 \cdot x) \cdot \exp(\lambda_2 \cdot \exp(b \cdot x)).$$

- It describes aging and life expectancy.
- It is also used in software engineering, to describe the "life expectancy" of software.

22. Reciprocal and U-Quadratic Distribution

- *Reciprocal and U-quadratic distribution:* $f_1(x) = \ln(x-\beta)$, $x \geq a$, and $x \leq b$, lead to

$$\rho(x) = A \cdot x^{\alpha} \text{ for } x \in [a,b].$$

- For $\alpha = -1$ and $\beta = 0$, we get the *reciprocal distribution* $\rho(x) = \text{const} \cdot x^{-1}$.
- It is used in *computer arithmetic*, to describe the frequency with which different numbers occur.
- for $\alpha = 2$, we get the *U-quadratic distribution*

$$\rho(x) = \text{const} \cdot (x-\beta)^2.$$

- It is used to describe quantities with a bimodal distribution.
- *Comment.* Both distributions can also be obtained by using generalized entropy.

23. Different Constraints Have Different Symmetries, Objective Function: Entropy (end)

- *Truncated normal distribution:* constrains on mean and variance, $x \geq a$, and $x \leq b$.
- It is used in econometrics, to model quantities about which we only know lower and upper bounds.
- *von Mises distribution:*
  - $f_1(x) = \cos(x - \mu)$,
  - $x \geq -\pi$, and
  - $x \leq \pi$
  - lead to $\rho(x) = \text{const} \cdot \exp(\lambda_1 \cdot \cos(x - \mu))$.
- It is used to describe random angles $x \in [-\pi, \pi]$.

24. Constraints With Different Symmetries, Objective Function: Generalized Entropy

- *Raised cosine distribution:*
  - objective function $\int (\rho(x))^2\,dx$,
  - shift-invariant constraint $f_1(x) = \cos(\omega \cdot x + \varphi)$,
  - scale-invariant constraints corresponding to $x \geq a$ and $x \leq b$,
  - we get $\rho(x) = c_1 + c_2 \cdot \cos(\omega \cdot x + \varphi)$.
- *Uniform distribution:* constraints $x \geq a$ and $x \leq b$ lead to uniform distribution on $[a, b]$.
- Similarly, we can get exponential, Erlang, reciprocal, and U-quadratic distributions.

25. Conclusion

- We have listed numerous families of distributions which are optimal if we:
  - optimize symmetry-based utility functions
  - under symmetry-based constraints.
- One can see that:
  - this list includes many empirically successful families of distributions, and
  - most empirically successful families of distributions are on this list.
- Thus, we indeed provide a symmetry-based explanation for the empirical success of these families.

26. Acknowledgments

- We acknowledge the partial support of:
  - the Center of Excellence in Econometrics, Faculty of Economics, Chiang Mai University, Thailand,
  - the National Science Foundation grants:
    * HRD-0734825 and HRD-1242122 (Cyber-ShARE Center of Excellence) and
    * DUE-0926721.

27. Proof for Scale-Invariant Objective Functions

- Scale-invariance means, in particular, that:
  - if we add a small deviation $\delta\rho(x)$ to the original distribution
  - in such a way that the value of the objective function does not change,
  - then the value of the re-scaled objective function should not change either.
- The fact that we still get a pdf means $\int \delta\rho(x)\,dx = 0$.
- For small deviations, $A(\rho(x) + \delta\rho) = A(\rho(x)) + A'(\rho(x)) \cdot \delta\rho(x)$.
- Thus, the fact that the value of the re-scaled objective function does not change means that

$$\int A'(\rho(x)) \cdot \delta\rho(x)\,dx = 0.$$

28. Proof (cont-d)

- Similarly, the fact that the value of the original objective function does not change means that

$$\int A'(\mu \cdot \rho(x)) \cdot \delta\rho(x)\, dx = 0.$$

- So, we arrive at the following requirement:
  - for every function $\delta\rho(x)$ for which $\int \delta\rho(x)\, dx = 0$ and $\int A'(\rho(x)) \cdot \delta\rho(x)\, dx = 0$,
  - we should have $\int A'(\mu \cdot \rho(x)) \cdot \delta\rho(x)\, dx = 0$.
- In Hilbert space terms: if $\delta\rho \perp 1$ and $\delta\rho \perp A'(\rho(x))$, then $\delta\rho \perp A'(\mu \cdot \rho(x))$.
- Thus, $A'(\mu \cdot \rho(x))$ should belong to the linear space spanned by 1 and $A'(\rho(x))$:

$$A'(\mu \cdot \rho(x)) = a(\mu, \rho) + b(\mu, \rho) \cdot A'(\rho(x)).$$

29. Proof (cont-d)

- We got $A'(\mu \cdot \rho(x)) = a(\mu, \rho) + b(\mu, \rho) \cdot A'(\rho(x))$.
- For $x_1 \neq x_2$, we get:

$$A'(\mu \cdot \rho(x_1)) = a(\mu, \rho) + b(\mu, \rho) \cdot A'(\rho(x_1));$$

$$A'(\mu \cdot \rho(x_2)) = a(\mu, \rho) + b(\mu, \rho) \cdot A'(\rho(x_2)).$$

- Hence $b(\mu, \rho) = \dfrac{A'(\mu \cdot \rho(x_2)) - A'(\mu \cdot \rho(x_1))}{A'(\rho(x_2)) - A'(\rho(x_1))}$ and $a(\mu, \rho) - b(\mu, \rho) \cdot A'(\rho(x_1))$.
- Thus, $a(\mu, \rho)$ and $b(\mu, \rho)$ depend only on $\rho(x_1)$ and $\rho(x_2)$ and do not depend on $\rho(x)$ for $x \neq x_1, x_2$.
- We can start with $x'_1, x'_2 \neq x_1, x_2$.
- Then, we conclude that $a(\mu, \rho)$ and $b(\mu, \rho)$ do not depend on the values $\rho(x_1)$ and $\rho(x_2)$ either.

30. Proof (cont-d)

- So, a and b don't depend on $\rho(x)$ at all: $a(\mu, \rho) = a(\mu)$, $b(\mu, \rho) = b(\mu)$, $A'(\mu \cdot \rho(x)) = a(\mu) + b(\mu) \cdot A'(\rho(x))$.
- Differentiating both side with respect to μ and taking $\mu = 1$, we get $\rho \cdot \dfrac{dA'}{d\rho} = a'(1) + b'(1) \cdot A'(\rho)$.
- We can separate A and ρ: $\dfrac{dA'}{a'(1) + b'(1) \cdot A'} = \dfrac{d\rho}{\rho}$.
- When $b'(1) = 0$, we get $A' = a'(1) \cdot \ln(\rho) + \text{const}$ and so, $A(\rho) = a'(1) \cdot \rho \cdot \ln(\rho) + c_1 \cdot \rho + c_2$.
- For the term $c_1 \cdot \rho$, the integral is always constant

$$\int (c_1 \cdot \rho(x))\, dx = c_1 \cdot \int \rho(x)\, dx = c_1.$$

- Thus, optimizing the expression $\int A(\rho(x))\, dx$ is equivalent to optimizing the entropy $-\int \rho(x) \cdot \ln(\rho(x))\, dx$.

31. Proof (end)

- When $b'(1) \neq 0$, then for $B \stackrel{\text{def}}{=} A' + \dfrac{a'(1)}{b'(1)}$, we get $\dfrac{dB}{b'(1) \cdot B} = \dfrac{d\rho}{\rho}$.
- So, integration leads to $\ln(B) = b'(1) \cdot \ln(\rho) + \text{const}$ and $B = C \cdot \rho^\beta$ for $\beta \stackrel{\text{def}}{=} b'(1)$.
- Hence, $A'(\rho) = B - \text{const} = C \cdot \rho^\beta + \text{const}$, and

$$A(\rho) = \text{const} \cdot \rho^\alpha + c_1 \cdot \rho + c_2, \text{ where } \alpha \stackrel{\text{def}}{=} \beta + 1.$$

- Similarly to the above case, optimizing $\int A(\rho(x))\,dx$ is equivalent to optimizing $\int (\rho(x))^\alpha\,dx$.
- When $\beta = -1$, integration leads to

$$A(\rho) = \text{const} \cdot \ln(\rho) + c_1 \cdot \rho + c_2.$$

- So optimizing $\int A(\rho(x))\,dx$ is equivalent to optimizing generalized entropy $\int \ln(\rho(x))\,dx$.