Why Most Empirical Distributions Are Few-Modal

Julio Urenda[1,2], Olga Kosheleva[3],
and Vladik Kreinovich[2]

[1]Department of Mathematical Sciences
[2]Department of Computer Science
[3]Department of Teacher Education
University of Texas at El Paso
500 W. University
El Paso, TX 79968, USA
jcurenda@utep.edu, olgak@utep.edu
vladik@utep.edu

1. Empirical Distributions: We Expect Them to Be Multi-Modal

- Continuous distributions are characterized by their probability density functions $\rho(x)$.
- In principle, a probability density function can be any non-negative function.
- The only condition is that the overall probability should be equal to 1, i.e., that $\int \rho(x)\,dx = 1$.
- In such situations, it is natural to expect that:
  - in general, we will observe generic functions with this property,
  - e.g., functions which are random with respect to some reasonable measure on the set of all functions.

2. Empirical Distributions (cont-d)

- The first such measure was Wiener measure, corresponding to random walk.
- Later, many other random measures have been proposed.
- In most of these random measures, almost all functions are truly random, similar to random walk.
- They are very "wiggly", they have infinitely many local maxima and minima.
- In probabilistic terms, we expect the empirical probability density functions to be multi-modal.

3. Empirical Distributions Are Mostly Few-Modal

- In reality, empirical distributions are mostly either unimodal, or bimodal, or – in rare cases – trimodal.
- In other words, they are usually few-modal.
- Why?
- In science and engineering, the few-modality is often easy to explain.
- E.g., the distributions are normal or Gamma, or, in general, follow some theoretically justified law.
- But few-modal distributions are ubiquitous also:
  - in situations where we do not have exact equations,
  - such as econometrics.
- Why?
- In this talk, we provide a theoretical explanation for the few-modality of empirical distributions.

4. Main Idea

- Of course, the space of all possible probability density functions is infinite-dimensional.
- So to exactly describe each such function, we need to describe the values of infinitely many parameters.
- In practice, at each moment of time, we can only use finitely many parameters.
- So, we need to look into appropriate finite-dimensional families of probability density functions.
- And we need explain why functions from this appropriate family are few-modal.
- To answer this question, let us describe natural properties of such families F of distributions $\rho(c_1, \ldots, c_n.x)$.
- How do we gain the information about the distributions?

5. We Want Smoothness

- It is reasonable to require that:
  - small changes in the values of the parameters c_i and/or small changes in x
  - should lead to small changes in the probability density.
- In other words, we want the function $\rho(c_1, \ldots, c_n, x)$ to be smooth.

6. Combinining Pieces of Knowledge

- Suppose that:
  - one piece of evidence is described by a probability density function (pdf) $\rho_1(x)$, and
  - another – independent – piece of evidence – leads to pdf $\rho_2(x)$.
- If these were evidences about two different quantities x_1 and x_2, then:
  - due to independence, we would conclude that
  - the distribution of the pair (x_1, x_2) follows a product distribution $\rho_1(x_1) \cdot \rho_2(x_2)$.
- In our case, however, we know that this is the same quantity, i.e., that $x_1 = x_2$.
- Thus, to get the resulting distribution, we need to restrict the product distribution to the case $x_1 = x_2$.

7. Combinining Pieces of Knowledge (cont-d)

- In precise terms, we need to consider conditional distribution under the condition that $x_1 = x_2$.
- This means that we need to consider the distribution

$$\rho(x) = c \cdot \rho_1(x) \cdot \rho_2(x).$$

- Here c is a normalizing constant – which can be determined by the condition that $\int \rho(x)\,dx = 1$.
- Thus, it is reasonable to require that:
  - for every two distribution $\rho_1(x)$ and $\rho_2(x)$ from the desired family F,
  - their normalized product $c \cdot \rho_1(x) \cdot \rho_2(x)$ should also belongs to this family.

8. Knowledge Can Come In Parts

- Sometimes, we gain the knowledge right away.
- In many other cases, knowledge comes in small steps.
- Suppose that:
  - the resulting knowledge is described by a probability density function $\rho(x)$, and
  - it comes via several (n) independent similar pieces of knowledge,
  - each step characterized by some probability density function $\rho_1(x)$.
- Then, based on the previous subsection, we can conclude that $\rho(x) = c \cdot (\rho_1(x))^n$ for some constant c.
- So, $\rho_1(x) = c_1 \cdot (\rho(x))^{1/n}$ for an appropriate normalizing coefficient c_1.

9. Knowledge Can Come In Parts (cont-d)

- Thus, it is reasonable to require that:
  - for every distribution $\rho_1(x)$ from the desired family F and
  - for every natural number $n > 1$,
  - the normalized distribution $c_1 \cdot (\rho(x))^{1/n}$ should also belong to the family.

10. Scale- and Shift-Invariance

- The numerical value of a quantity depends:
  - on the starting point for measuring this quantity
  - and on the measuring unit.
- When we change numerical values, the expression for the probability distribution also changes.
- It is reasonable to require that:
  - if we simply change the starting point and/or the measuring unit in a distribution from the family F,
  - then we should still get a distribution from the same family.
- What if we change the starting point, i.e.,
  - we replace the original starting point
  - with a new one which is a units larger.

11. Scale- and Shift-Invariance (cont-d)

- Then in the new units $y = x - a$, the distribution:
  - described by pdf $\rho(x)$
  - will now be described by $\rho_1(y) = \rho(y + a)$.
- Similarly, we can change the measuring unit, i.e.:
  - replace the original measuring unit
  - with a new one which is λ times larger.
- Then in the new units $y = x/\lambda$, the distribution
  - described by the pdf $\rho(x)$
  - will now be described by $\rho_1(y) = \lambda \cdot \rho(\lambda \cdot y)$.

12. Definitions

- Let n be a natural number.
- By an *n-parametric family of distributions*, we mean a family $F = \{f(c_1, \ldots, c_n, x)\}_{c_1,\ldots,c_n}$ of pdfs, where:
  - the values $(c_1, \ldots, c_n)$ go over some set U, and
  - the function $f(c_1, \ldots, c_n, x)$ is continuously differentiable over the closure of this set.
- We say that a family F *allows combining knowledge* if:
  - for every two functions $\rho_1(x), \rho_2(x) \in F$,
  - there exists a real number $c > 0$ for which the product $c \cdot \rho_1(x) \cdot \rho_2(x)$ also belongs to F.

13. Definitions (cont-d)

- We say that a family F *allows partial knowledge* if:
  - for every function $\rho(x)$ from this family and for every natural number n,
  - there exists a real number $c > 0$ for which the function $c \cdot (\rho(x))^{1/n}$ also belongs to F.
- We say that a family F is *shift-invariant* if:
  - for every function $\rho(x)$ from this family and for every real number a,
  - the function $\rho(x + a)$ also belongs to F.
- We say that a family F is *scale-invariant* if:
  - for every function $\rho(x)$ from this family and for every real number $\lambda > 0$,
  - the function $\lambda \cdot \rho(\lambda \cdot x)$ also belongs to F.

14. Main Result

- **Proposition.**
  - *Let F be a shift- and scale-invariant n-parametric family that allows combining and partial knowledge.*
  - *Then, every function $\rho \in F$ has the form $\rho(x) = \exp(P(x))$ for some polynomial of degree $\leq n$.*
- **Corollary.**
  - *Let F be a shift- and scale-invariant n-parametric family that allows combining and partial knowledge.*
  - *Then, every function $\rho \in F$ has no more than $n-1$ local maxima and local minima.*
- This explain why empirical distributions are few-modal.

15. Proof of the Corollary

- Indeed, at local maxima and minima, the derivative $\rho'(x) = \exp(P(x)) \cdot P'(x)$ is equal to 0.
- This is equivalent to $P'(x) = 0$.
- The derivative $P'(x)$ is a polynomial of degree $\le n-1$.
- Such polynomials can have no more than $n - 1$ zeros.

16. Proof of the Main Result

- Let F be a family that satisfies all the given properties.
- To simplify the problem, let's consider a family G of all the functions $c \cdot \rho(x)$, where $c > 0$ and $\rho(x) \in F$.
- By definition, every function from the family F is also an element of G.
- To show this, it is sufficient to take $c = 1$.
- We will prove the desired form for all the function from the class G.
- This will automatically imply that all the functions from the family F also have this property.
- What is the dimension of the family G?
- I.e., how many parameters do we need to specify each function from this family?

17. Proof (cont-d)

- To describe a function from G, we need to specify:
  - the value c (1 parameter), and
  - the function $\rho(x) \in F$ – which requires n parameters.
- Thus, $n+1$ parameters are sufficient, and the dimension of the family G is $\leq n+1$.
- For the family G, allowing combining knowledge leads to a simpler property: that
  - for every two functions $f_1(x), f_2(x) \in G$
  - their product $f_1(x) \cdot f_2(x)$ also belong to G.
- Indeed, $f_i(x) \in G$ means that $f_i(x) = c_i \cdot \rho_i(x)$ for some $c_i > 0$ and $\rho_i(x) \in F$.
- Thus, the product $f(x) = f_1(x) \cdot f_2(x)$ of these functions has the from $f(x) = c_1 \cdot c_2 \cdot \rho_1(x) \cdot \rho_2(x)$.

18. Proof (cont-d)

- By the property of allowing combining knowledge, for some $c > 0$, we have $\rho_0(x) = c \cdot \rho_1(x) \cdot \rho_2(x) \in F$.
- Thus, $f(x) = \dfrac{c_1 \cdot c_2}{c} \cdot (c \cdot \rho_1(x) \cdot \rho_2(x)) = c_0 \cdot \rho_0(x)$, where $c_0 \stackrel{\text{def}}{=} \dfrac{c_1 \cdot c_2}{c}$.
- So indeed, $f(x) \in G$.
- Similarly, from the other properties of the family F, we can make the following conclusions:
  - that for every function $f(x) \in G$ and for every natural number n, we have $(f(x))^{1/n} \in G$;
  - that for every function $f(x) \in G$ and for every real number a, we have $f(x + a) \in G$;
  - that for every function $f(x) \in G$ and for every real number $\lambda > 0$, we have $f(\lambda \cdot x) \in G$.

19. Proof (cont-d)

- We can simplify the problem even more if:
  - instead of the family G,
  - we consider the family g of all the functions of the type $F(x) = \ln(f(x))$, where $f(x) \in G$.
- To such functions, we also add the limit functions.
- Adding limit cases does not increase the dimension, so the dimension of the family g is still $\le n+1$.
- In terms of this new family, we need to prove that all the functions from g are polynomials of order $\le n$.
- The fact that the family G is closed under multiplication means that the family g is closed under addition.

20. Proof (cont-d)

- The fact that the family G is closed under taking the n-th root means that:
  - the family g is closed
  - under multiplication by $1/n$ for each natural number n.
- Together with closing under addition, this means that:
  - for every two natural numbers m and n,
  - the function $\dfrac{m}{n}\cdot F(x) = \dfrac{1}{n}\cdot F(x) + \ldots + \dfrac{1}{n}\cdot F(x)$ (m times) also belongs to the family g.
- In other words, for every $F(x) \in g$ and for every rational number r, we have $r \cdot F(x) \in g$.
- Every real number is a limit of rational numbers.
- E.g., it is a limit of numbers obtained if we only keep the first N digits in the decimal or binary expansion.

21. Proof (cont-d)

- Since we added all limit cases, we can conclude that $r \cdot F(x) \in g$ for all non-negative real numbers r as well.
- One can easily show that shift- and scale-invariance properties are also satisfied for the new family:
  - that for every function $F(x) \in g$ and for every real number a, we have $F(x+a) \in g$;
  - that for every function $F(x) \in g$ and for every real number $\lambda > 0$, we have $F(\lambda \cdot x) \in g$.
- As a final simplification, we consider the family h of all the differences $d(x) = F_1(x) - F_2(x)$ between $F_i(x) \in g$.
- To describe each of the functions $F_1(x)$ and $F_2(x)$, we need $n+1$ parameters.
- So the dimension of the new family does not exceed $2 \cdot (n+1)$.

22. Proof (cont-d)

- For every function $F(x) \in g$, the function $2F(x)$ also belongs to the family g.
- So, we can conclude that the difference $F(x) = (2F(x)) - F(x)$ also belongs to the family h. Thus, $g \subseteq h$.
- The family h is also closed under addition.
- Indeed, if $d_1(x) = F_{11}(x) - F_{12}(x)$ and $d_2(x) = F_{21}(x) - F_{22}(x)$ for some $F_{ij}(x) \in g$, then

$$d_1(x) + d_2(x) = (F_{11}(x) - F_{12}(x)) + (F_{21}(x) - F_{22}(x)) =$$

$$(F_{11}(x) + F_{21}(x)) - (F_{12}(x) + F_{22}(x)).$$

- Since g is closed under addition, the sums $F_{11}(x) + F_{21}(x)$ and $F_{12}(x) + F_{22}(x)$ also belong to g.
- Thus, indeed, the sum $d_1(x) + d_2(x)$ is a difference between two functions from g and is, thus, in h.

23. Proof (cont-d)

- We can also prove that the family h is closed under multiplication by any real number c.
- Indeed, let $d(x) = F_1(x) - F_2(x)$.
- If $c > 0$, then $c \cdot d(x) = (c \cdot F_1(x)) - (c \cdot F_2(x))$, where both $c \cdot F_1(x)$ and $c \cdot F_2(x)$ belong to the family g.
- If $c < 0$, then $c \cdot F(x) = |c| \cdot F_2(x) - |c| \cdot F_1(x)$, where also $|c| \cdot F_2(x)$ and $|c| \cdot F_1(x)$ belong to the family g.
- So, the family h is closed under addition and under multiplication by any real number.
- Thus, h is a linear space.
- Let $d \le 2n + 2$ denote the dimension of this linear space.
- Let us select a basis $e_1(x), \ldots, e_d(x)$.

24. Proof (cont-d)

- This means that all functions from the space g have the form $c_1 \cdot e_1(x) + \ldots + c_d \cdot e_d(x)$.
- We know that the family g is shift- and scale-invariant.
- Thus, we can conclude that the family h is also shift- and scale-invariant.
- Shift-invariance means that for each $d(x) \in h$ and for each real number a, we have $d(x+a) \in h$.
- In particular, this is true for the basis functions

$$e_1(x), \ldots, e_d(x).$$

- Thus, for each i and a, there exist coefficients $c_{ij}(a)$ depending on a for which

$$e_i(x+a) = c_{i1}(a) \cdot e_1(x) + \ldots + c_{id}(a) \cdot e_d(x).$$

25. Proof (cont-d)

- In particular, for each i, we can select d different values

$$x_1, \ldots, x_d.$$

- Then we get the following system of d linear equations for determining the coefficients $c_{ij}(a)$:

$$e_i(x_1 + a) = c_{i1}(a) \cdot e_1(x_1) + \ldots + c_{id}(a) \cdot e_d(x_1),$$

$$\ldots$$

$$e_i(x_d + a) = c_{i1}(a) \cdot e_1(x_d) + \ldots + c_{id}(a) \cdot e_d(x_d).$$

- Here, the coefficients $e_j(x_k)$ are constants.
- So the values $c_{ij}(a)$ are linear combinations of the right-hand sides $e_i(x_k + a)$.
- Since the functions $e_i(x)$ are differentiable, the values $c_{ij}(a)$ are also differentiable functions of a.

26. Proof (cont-d)

- So, both sides of the following equality are differentiable: $e_i(x+a) = c_{i1}(a) \cdot e_1(x) + \ldots + c_{id}(a) \cdot e_d(x)$.
- Thus, we can differentiate them with respect to a and then plug in $a = 0$.
- As a result, we get the following system of differential equations, where $C_{ij} \stackrel{\text{def}}{=} c'_{ij}(0)$:

$$e'_1(x) = C_{11} \cdot e_1(x) + \ldots + C_{1d} \cdot e_d(x),$$

$$\ldots$$

$$e'_d(x) = C_{d1} \cdot e_1(x) + \ldots + C_{dd} \cdot e_d(x),$$

- In other words, for $e_i(x)$, we get a system of linear differential equations with constant coefficients.

27. Proof (cont-d)

- It is known that each solution of such system is a linear coefficient of the functions $x^p \cdot \exp(\alpha \cdot x)$, where:
  - the value p is a natural number and
  - α is a – possible complex – eigenvalue of the matrix C_{ij}.
- Similarly, scale-invariance means that for each function $d(x) \in h$ and for each positive real number $\lambda > 0$, we have $d(\lambda \cdot x) \in h$.
- In particular, this is true for the basis functions $e_i(x)$.
- For an auxiliary variable $X \stackrel{\text{def}}{=} \ln(x)$:
  - replacing x with $\lambda \cdot x$ corresponds to
  - replacing X with $X + a$, where $a \stackrel{\text{def}}{=} \ln(\lambda)$.

28. Proof (cont-d)

- So, for the correspondingly re-scaled functions $E_i(X) \stackrel{\text{def}}{=} e_i(\exp(X))$, we conclude that:
  - for each such function and for each real number a,
  - the function $E_i(X+a)$ is a linear combination of functions $E_1(X), \ldots, E_d(X)$.
- We already know, from the previous parts of this proof, that this implies that:
  - each function $E_i(X)$
  - is a linear combination of the functions $X^p \cdot \exp(\alpha \cdot X)$.
- Thus, each function $e_i(x) = E_i(\ln(x))$ is a linear combination of expressions

$$(\ln(x))^p \cdot \exp(\alpha \cdot \ln(x)) = (\ln(x))^p \cdot x^\alpha.$$

29. Proof (cont-d)

- One can see that:
  - the only possibility for a function to be represented in both forms
  - is to avoid logarithms and exponential functions altogether.
- So, $e_i(x)$ is a linear combination of the terms x^p for natural p, i.e., a polynomial.
- Thus, each function from the class g is a polynomial, as a linear combination of d polynomials $e_i(x)$.
- Since $g \subseteq h$, all functions from the class g are also polynomials.
- What is the order of these polynomials?
- Let D be the order of a polynomial $F(x)$ from the g.

30. Proof (cont-d)

- For a polynomial of order D, in general, $F(x)$, $F(x+h)$, $F(x+2h)$, $\ldots$, $F(x+D\cdot h)$ are linearly independent.
- Indeed, for $h \to 0$, this is equivalent to linear independence of x^D, x^{D-1}, $\ldots$, 1.
- Thus, in the generic case, the corresponding determinant is different from 0.
- Since we have $D+1$ independent functions, thus, the family g has dimension $D+1$.
- But we know that the dimension of this family is $\leq n+1$.
- From $D+1 \leq n+1$, we conclude that $D \leq n$.
- Thus, all functions $F(x) = \ln(f(x))$ from the class g are polynomials of order $\leq n$.

31. Proof (cont-d)

- Thus, all functions $F(x) = \ln(f(x))$ from the class g are polynomials of order $\le n$.
- Hence, each function $f(x) = \exp(F(x))$ from the class F has the desired form.
- The proposition is proven.

32. Acknowledgments

This work was supported in part by the National Science Foundation grants:

- 1623190 (A Model of Change for Preparing a New Generation for Professional Practice in Computer Science),
- HRD-1242122 (Cyber-ShARE Center of Excellence).