5-2020

Why Most Empirical Distributions Are Few-Modal

Julio Urenda
The University of Texas at El Paso, jcurenda@utep.edu

Olga Kosheleva
The University of Texas at El Paso, olgak@utep.edu

Vladik Kreinovich
The University of Texas at El Paso, vladik@utep.edu

Comments:
Technical Report: UTEP-CS-20-50

Why Most Empirical Distributions Are Few-Modal

Julio C. Urenda, Olga Kosheleva, and Vladik Kreinovich
University of Texas at El Paso, 500 W. University
El Paso, TX 79968, USA
jcurenda@utep.edu, olgak@utep.edu, vladik@utep.edu

Abstract

In principle, any non-negative function can serve as a probability density function – provided that it adds up to 1. All kinds of processes are possible, so it seems reasonable to expect that observed probability density functions are random with respect to some appropriate probability measure on the set of all such functions – and for all such measures, similarly to the simplest case of random walk, almost all functions have infinitely many local maxima and minima. However, in practice, most empirical distributions have only a few local maxima and minima – often one (unimodal distribution), sometimes two (bimodal), and, in general, they are few-modal. From this viewpoint, econometrics is no exception: empirical distributions of economics-related quantities are also usually few-modal. In this paper, we provide a theoretical explanation for this empirical fact.

1 Formulation of the Problem

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. 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 – in the sense that 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.

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; see, e.g., [1]. Why?

This is especially puzzling in econometrics. 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?

What we do in this paper. In this paper, we provide a theoretical explanation for the few-modality of empirical distributions.

2 Analysis of the Problem

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 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 $\rho(c_1, \ldots, c_n.x)$. To come up with these properties, let us recall how we gain the information about the corresponding distributions.

We want smoothness. 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.

We can combine different pieces of knowledge. Suppose that:

- one piece of evidence leads us to conclude that the distribution of the corresponding quantity is described by a probability density function $\rho_1(x)$, and
- another – independent – piece of evidence – leads to a slightly different probability density function $\rho_2(x)$.

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 when $x_1 = x_2$, i.e., 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 $\rho(x) = c \cdot \rho_1(x) \cdot \rho_2(x)$, where 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.

Knowledge can come in parts. Sometimes, we gain the knowledge right away. In many other cases, knowledge comes in small steps. If the resulting knowledge is described by a probability density function $\rho(x)$, and it comes via several (n) independent similar pieces of knowledge, each 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, i.e., that $\rho_1(x) = c_1 \cdot (\rho(x))^{1/n}$ for an appropriate normalizing coefficient c_1.

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.

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.

If we change the starting point, i.e., if we replace the original starting point with a new one which is a units larger, then in the new units $y = x - a$, the distribution described by the original probability density function $\rho(x)$ will now be described by the new function $\rho_1(y) = \rho(y + a)$.

Similarly, if we change the measuring unit, i.e., if we 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 original probability density function $\rho(x)$ will now be described by the new function $\rho_1(y) = \lambda \cdot \rho(\lambda \cdot y)$.

Now, we are ready. Now, we are ready to formulate our main result.

3 Definitions and the Main Result

Definition 1. *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 probability density functions, 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 very two functions $\rho_1(x)$ and $\rho_2(x)$ from this family, there exists a real number $c > 0$ for which the product $c \cdot \rho_1(x) \cdot \rho_2(x)$ also belongs to F.*

- *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.*

Proposition 1. *Every function from a shift- and scale-invariant n-parametric family of distributions that allows combining knowledge and partial knowledge has the form $\rho(x) = \exp(P(x))$ for some polynomial of degree $\le n$.*

Corollary. *Every function from a shift- and scale-invariant n-parametric family of distributions that allows combining knowledge and partial knowledge has no more than $n - 1$ local maxima and local minima.*

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

Discussion. This explain why empirical distribution are few-modal.

Proof of the main result.

1°. Let F be a family that satisfies all the given properties. To somewhat simplify the problem, let us consider a family G of all the functions of the type $c \cdot \rho(x)$, 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? To describe a function from the family 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 $\le n + 1$.

2°. For the family G, the first property of the family F – allowing combining knowledge – leads to a simpler property: that for every two functions $f_1(x)$ and $f_2(x)$ from the family G their produc $f_1(x) \cdot f_2(x)$ also belong to G.

Indeed, the fact that each function $f_i(x)$ belongs to G means that it has the form $c_i \cdot \rho_i(x)$ for some $c_i > 0$ and for some function $\rho_i(x)$ from the family F. Thus, the product $f(x) = f_1(x) \cdot f_2(x)$ of these functions has the from $f(x) = c_1 \cdot c_2 \cdot \rho_1(x) \cdot \rho_2(x)$. By the property of allowing combining knowledge, for some $c > 0$, the function $\rho_0(x) = c \cdot \rho_1(x) \cdot \rho_2(x)$ also belongs to the family F. Thus, we have

$$f(x) = \frac{c_1 \cdot c_2}{c} \cdot (c \cdot \rho_1(x) \cdot \rho_2(x)) = c_0 \cdot \rho_0(x),$$

where we denoted $c_0 \stackrel{\text{def}}{=} \dfrac{c_1 \cdot c_2}{c}$. So indeed, $f(x) \in G$.

3°. Similarly, from the other properties of the family F, we can make the following conclusions:

- that for every function $f(x)$ from the family G and for every natural number n, the function $(f(x))^{1/n}$ also belongs to G;
- that for every function $f(x)$ from the family G and for every real number a, the function $f(x + a)$ also belongs to G; and
- that for every function $f(x)$ from the family G and for every real number $\lambda > 0$, the function $f(\lambda \cdot x)$ also belongs to G.

6°. 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 this family 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. 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

$$\frac{m}{n} \cdot F(x) = \frac{1}{n} \cdot F(x) + \ldots + \frac{1}{n} \cdot F(x) \ (m \text{ times})$$

also belongs to the family g. In other words, for every function $F(x) \in g$ and for every rational number r, the product $r \cdot F(x)$ also belongs to g. Since every real number is a limit of rational numbers – e.g., of numbers obtained if we only keep the first N digits in the decimal or binary expansion – and 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 $F(x)$ from the family g and for every real number a, the function $F(x+a)$ also belongs to g; and
- that for every function $F(x)$ from the family G and for every real number $\lambda > 0$, the function $F(\lambda \cdot x)$ also belongs to g.

7°. As a final simplification, we consider the family h of all the differences $d(x) = F_1(x) - F_2(x)$ between functions from the class 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 $2 \cdot (n+1)$.

Since for every function $F(x) \in g$, the function $2F(x)$ also belongs to the family g, 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

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

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

where, 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, an element of the family h.

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 g is closed under addition and under multiplication by any real number – and is, thus, a linear space. Let $d \leq 2n+2$ denote the dimension of this linear space, and let us select a basis $e_1(x), \ldots, e_d(x)$. This means that all functions from the space g have the form $c_1 \cdot e_1(x) + \ldots + c_d \cdot e_d(x)$.

From the fact that the family g is shift- and scale-invariant, we can conclude that the family h is also shift- and scale-invariant.

8°. Shift-invariance means that for each function $d(x)$ from the family h and for each real number a, the function $d(x+a)$ also belongs to 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

$$e_i(x+a) = c_{i1}(a) \cdot e_1(x) + \ldots + c_{id}(a) \cdot e_d(x). \quad (1)$$

In particular, for each i, if we 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, we conclude that the values $c_{ij}(a)$ are also differentiable functions of a.

So, both sides of the equality (1) are differentiable. 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:

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

$$\ldots$$

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

where $C_{ij} \stackrel{\text{def}}{=} c_{ij}'(0)$.

In other words, for the functions $e_1(x)$, $\ldots$, $e_d(x)$, we get a system of linear differential equations with constant coefficients. It is known that each solution of such system is a linear coefficient of the functions

$$x^p \cdot \exp(\alpha \cdot x), \tag{2}$$

where p is a natural number and α is a – possible complex – eigenvalue of the matrix C_{ij}.

9°. Similarly, scale-invariance means that for each function $d(x)$ from the family h and for each positive real number $\lambda > 0$, the function $d(\lambda \cdot x)$ also belongs to h. In particular, this is true for the basis functions $e_i(x)$.

If we introduce an auxiliary variable $X \stackrel{\text{def}}{=} \ln(x)$, then replacing x with $\lambda \cdot x$ corresponds to replacing X with $X + a$, where $a \stackrel{\text{def}}{=} \ln(\lambda)$. 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 Part 8 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

$$(\ln(x))^p \cdot \exp(\alpha \cdot \ln(x)) = (\ln(x))^p \cdot x^\alpha. \tag{3}$$

One can see that the only possibility for a function to be represented both in forms (2) and (3) is to avoid logarithms and exponential functions altogether,

i.e., to have $e_i(x)$ equal to a linear combination of the terms x^p for natural p, i.e., to have all functions $e_i(x)$ polynomials. 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.

10°. What is the order of these polynomials? Let D be the order of a polynomial $F(x)$ from the class g. For each polynomial of order D, in general, the functions $F(x)$, $F(x+h)$, $F(x+2h)$, $\ldots$, $F(x+D\cdot h)$ are linearly independent: indeed, for $h \to 0$, this is equivalent to linear independence of x^D, x^{D-1}, $\ldots$, 1, and 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 $\le n+1$. From $D+1 \le n+1$, we conclude that $D \le n$. 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.

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) and HRD-1242122 (Cyber-ShARE Center of Excellence).

References

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