Need for Most Accurate Discrete Approximations Explains Effectiveness of Statistical Methods Based on Heavy-Tailed Distributions

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

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

Abstract. In many practical situations, it is effective to use statistical methods based on Gaussian distributions, and, more generally, distribution for which tails are *light* – in the sense that as the value increases, the corresponding probability density tends to 0 very fast. There are many theoretical explanations for this effectiveness. On the other hand, in many other cases, it is effective to use statistical methods based on *heavy-tailed* distributions, in which the probability density is asymptotically described, e.g., by a power law. In contrast to the light-tailed distributions, there is no convincing theoretical explanation for the effectiveness of the heavy-tail-based statistical methods. In this paper, we provide such a theoretical explanation. This explanation is based on the fact that in many applications, we approximate a continuous distribution by a discrete one. From this viewpoint, it is desirable, among all possible distributions which are consistent with our knowledge, to select a distribution for which such an approximation is the most accurate. It turns out that under reasonable conditions, this requirement (of allowing the most accurate discrete approximation) indeed leads to the statistical methods based on the power-law heavy-tailed distributions.

1 Formulation of the Problem

Many effective statistical methods are based on light-tailed probability distributions. In many practical situations, we encounter Gaussian distributions. Their ubiquity can be explained by the fact that in many of these situations, the corresponding random quantity is the result of joint effect of many different factors, and according to the Central Limit Theorem, the distribution of the sum of many independent small random variables is close to Gaussian; see, e.g., [24].

This argument explains why many actual distributions are close to Gaussian. For example, an empirical study of measuring instruments shows that for about 60% of them, the probability distribution of the measurement error is close to Gaussian; see, e.g., [19, 20]. In such situations – when we have enough data about the probability distribution to know that it is close to Gaussian – it is reasonable to use statistical methods based on Gaussian distributions.

Interestingly, however, statistical methods based on Gaussian distributions have been effectively used in much more general situations, when we only have a *partial* information about the probability distribution – e.g., when we only know the first two moments of the distribution; see, e.g., [21]. The effectiveness of Gaussian methods in such situations comes from the use of the Maximum Entropy approach [9]: if we have only partial information about the probability distribution – i.e., if several different probability density functions $\rho(x)$ are consistent with our knowledge – then it is reasonable to select the distributions with the largest uncertainty, i.e., with the largest value of the entropy $S = -\int \rho(x) \cdot \ln(\rho(x))\, dx$. Often, the only information that we have about the probability distribution is its first two moments E and M_2, i.e., we only know that $\int \rho(x)\, dx = 1$, $\int x \cdot \rho(x)\, dx = E$, and $\int x^2 \cdot \rho(x)\, dx = M_2$. If we apply the Lagrange multiplier method to maximize entropy under these three constraints, we thus reduce the original constraint optimization problem to the unconstrained problem of maximizing the following functional

$$-\int \rho(x) \cdot \ln(\rho(x))\, dx + \lambda_0 \cdot \left(\int \rho(x)\, dx - 1 \right) + \lambda_1 \cdot \left(\int x \cdot \rho(x)\, dx - E \right) +$$

$$\lambda_2 \cdot \left(\int x^2 \cdot \rho(x)\, dx - M_2 \right).$$

Differentiating this expression with respect to each unknown $\rho(x)$ and equating the resulting derivative to 0, we conclude that

$$-\ln(\rho(x)) - 1 + \lambda_0 + \lambda_1 \cdot x + \lambda_2 \cdot x^2 = 0,$$

hence $\rho(x) = \exp((\lambda_0 - 1) + \lambda_1 \cdot x + \lambda_2 \cdot x^2)$. One can check that for the constraints to be satisfied, this function has to have the usual Gaussian form

$$\rho(x) = \frac{1}{\sqrt{2\pi} \cdot \sigma} \cdot \exp\left(-\frac{(x-a)^2}{2\sigma^2} \right),$$

where $a = E$ and $\sigma = \sqrt{M_2 - E^2}$.

For the Gaussian distribution, the probability density has *light tails* in the sense that as x increases, the probability density tends to 0 fast: so fast that for every k, we have a finite moment $\int x^k \cdot \rho(x)\, dx$.

Similarly, many other effective statistical methods are based on light-tailed distributions.

Some effective statistical methods are based on heavy-tailed distributions. On the other hand, in many empirical situations, it turns out to be

efficient to use statistical methods based on *heavy-tailed* distributions, i.e., distributions for which the probability density $\rho(x)$ tends to 0 much slower – so that some moments become infinite; see, e.g., [3–6, 8, 13–17, 23, 25, 26]. A typical case is when asymptotically, the probability density function has a power law distribution $\rho(x) = C \cdot x^{-\alpha}$ for some $\alpha > 0$.

Natural question. If usual arguments lead to the effectiveness of statistical methods based on light-tailed distributions, how can we explain the effectiveness of statistical methods based on the heavy-tailed distributions?

What we do in this paper. In this paper, we provide a possible answer to the above question: namely, we show that the need for most accurate discrete approximations naturally leads to statistical methods based on heavy-tailed distributions.

2 Need for Most Accurate Discrete Approximations: The Main Idea

Need for discrete approximations: an example. Many real-life situations are well-described by *random walk* models; such models are especially ubiquitous in econometrics; see, e.g., [12] and references therein.

In the general random walk, each component $x(t+1)$ of the state at the next moment of time $t+1$ is obtained from the state $x(t)$ at the previous moment of time t by adding a random step $r(t)$: $x(t+1) = x(t) + r(t)$. In many cases, the random steps $r(t)$ and $r(t')$ corresponding to different moments of time $t \neq t'$ are independent.

In some applications, the empirical data is well-described by the simplest type of random walk, in which, for some constant $r_0 > 0$, the random step is equal either to r_0 or to $-r_0$, with probability 0.5 of each of these two values. In other applications, this simple model is not sufficient, so we need to use a more complex model, in which the step takes three, four, or more values with different probabilities. In the limit, when the number n of values tends to infinity, we get a continuous description, in which a step $r(t)$ is distributed according to some probability density function $\rho(x)$.

Resulting need for the most accurate discrete approximations. The more accurately we can approximate the actual continuous distribution with an n-point one, the more accurate are the discrete-approximation models (e.g., the corresponding random walk models).

Since random walk models are widely used, it makes sense to use the existence of the most accurate discrete approximations as an alternative criterion for selecting a probability distribution in situations when we only have partial information about the probabilities – i.e., in which several different probability distributions are consistent with our knowledge.

This is our main idea. This is our main idea, an idea that, as we show in this paper, will lead to an explanation for power-law distributions. To come up with this explanation, we first need to formulate our idea in precise terms.

3 Towards Formalizing the Main Idea

Approximating a continuous distribution by discrete ones: analysis of the problem. When we approximate a continuous probability distribution by a discrete one, we thus:

- approximate the actual random variable r which can take, in principle, any value from the real line,
- by a discrete variable that can only take a finite number of values

$$r_1, \ldots, r_n.$$

The approximating value r_i is, in general, different from the actual value r, so there is the approximation inaccuracy $\delta \stackrel{\text{def}}{=} |r_i - r|$. To describe continuous distributions which allow the most accurate discrete approximations, we need to formalize what it means, for a given probability distribution and for a given n, to select the most accurate discrete approximation.

Rational decision making: reminder. It is known that, in general, decision of a rational decision maker can be described as maximizing the expected value of an appropriate objective function $u(x)$ called *utility* [7, 11, 18, 22]. This function is determined modulo a general linear transformation $u \to a \cdot u + b$ for some $a > 0$ and b.

Maximizing utility is equivalent to mininizing *disutility* $U \stackrel{\text{def}}{=} -u$. Disutility is also defined modulo a general linear transformation $U \to a \cdot U + b$.

Let us apply the general ideas of rational decision making to our situation. To apply the general ideas of rational decision making to our case, we need to describe the disutility $U(\delta)$ caused by inaccuracy δ.

To describe this disutility, we can take into account that in most case, the numerical value r of the corresponding quantity depends on the choice of a measuring unit.

- For a geometric random walk, the value r represents a distance, whose numerical value depends on our choice of a distance unit – e.g., meters or feet.
- For a financial random walk, e.g., for the financial random walk describing the stock market index, the value r represents the price, and its numerical value depends on the monetary unit – e.g., dollars or Euros.

If we choose a new measuring unit which is k times smaller than the original one, then the numerical value of the corresponding quantity increases by a factor of k: $r \to k \cdot r$. For example, if we use centimeters instead of meters, then in centimeters, the distance of $r = 2$ m takes the value $k \cdot r = 100 \cdot 2$ cm.

The choice of a measuring unit is usually rather arbitrary. It is therefore reasonable to require that the disutility function $U(\delta)$ not depend on the choice of a measuring unit. Of course, we cannot simply require that the disutility function does not change at all, i.e., that $U(k \cdot \delta) = U(\delta)$ for all k and δ, since that would imply that $U(\delta) \equiv$ const. However, we can take into account the

fact that the disutility function is defined modulo some linear transformation. If we fix $U(0) = 0$, this still leaves us with a transformation $U \to a \cdot U$. We can therefore require that when we re-scale the unit for measuring the original quantity, the disutility function remains the same modulo an appropriate linear transformation, i.e., that for every $k > 0$, there exist a value $a(k)$ for which, for every $\delta \geq 0$, we have $U(k \cdot \delta) = a(k) \cdot U(\delta)$.

Small changes in accuracy should lead to small changes in utility. In mathematical terms, this means that the disutiliuty function $U(\delta)$ must be continuous. It is known that for a continuous function, the above functional equation implies that $U(\delta) = C \cdot \delta^\beta$ for some values $C > 0$ and β; see, e.g., [1]. The larger the inaccuracy δ, the larger the disutility. So, we must have $\beta > 0$.

Towards formalizing the problem. Let us use this disutility function to describe the expected disutility of approximating the original continuous variable with a discrete one. Since the disutility increase with inaccuracy, once the points $r_1 < \ldots < r_n$ are selected, we should assign, to each value r, the point r_i for which the inaccuracy is the smallest possible, i.e., the value which is the closest to r. Thus, for every i, the point r_i is assigned to all the values r from the interval $\left[\frac{r_{i-1} + r_i}{2}, \frac{r_i + r_{i+1}}{2}\right]$.

For large n, we can describe the selection of the points $r_1 < r_2 < \ldots < r_n$ by describing the frequency $\rho_0(r)$, i.e., the number of points per unit length. The overall number of points is n, so we have $\int \rho_0(r)\, dr = n$. In this case, the length of each intervals $[r_i, r_{i+1}]$ is approximately equal to $\dfrac{1}{\rho_0(r)}$, and similarly, the length of the interval $I_i \stackrel{\text{def}}{=} \left[\dfrac{r_{i-1} + r_i}{2}, \dfrac{r_i + r_{i+1}}{2}\right]$, which is composed of the halves of $[r_{i-1}, r_i]$ and of $[r_i, r_{i+1}]$, is also approximately equal to $\dfrac{1}{\rho_0(r)}$. On this interval, the average value of inaccuracy is proportional to the interval width and thus, the average value of the disutility is proportional to the β-th power of this width, i.e., to $U_i \approx \dfrac{1}{(\rho_0(r_i))^\beta}$.

The overall expected value E_U of the disutility – the one that we need to minimize – is equal to $E_U = \sum\limits_{i=1}^{n} p_i \cdot U_i$, where p_i is the probability that the original random variable r occurs in the interval I_i. Here, $p_i = \int_{I_i} \rho(r)\, dr$, where $\rho(r)$ is the probability density of the original random variable. Thus,

$$E_U = \sum_{i=1}^{n} \frac{1}{(\rho_0(r_i))^\beta} \cdot \int_{I_i} \rho(r)\, dr.$$

By moving the term $\dfrac{1}{(\rho_0(r_i))^\beta}$ inside the intervals, we get

$$E_U = \sum_{i=1}^{n} \int_{I_i} \frac{1}{(\rho_0(r_i))^\beta} \cdot \rho(r)\, dr.$$

For large n and narrow intervals, we have $r_i \approx r$ and thus,

$$E_U \approx \sum_{i=1}^{n} \int_{I_i} \frac{1}{(\rho_0(r))^\beta} \cdot \rho(r)\, dr.$$

The intervals I_i cover the whole real line. Thus, the sum of integrals of the same function over all intervals I_i is simply the integral over the whole real line:

$$E_U \approx \int \frac{1}{(\rho_0(r))^\beta} \cdot \rho(r)\, dr.$$

So, we arrive at the following precise reformulation of the best discrete optimization problem:

- we know the probability density $\rho(r)$;
- we want to find the density $\rho_0(r)$ of the distribution of the discrete points as the one that minimizes the integral $\int \frac{1}{(\rho_0(r))^\beta} \cdot \rho(r)\, dr$ under the constraint $\int \rho_0(r)\, dr = n$.

After that, we will need to select, in each class of probability distributions, a distribution for which this best-case expected disutility has the smallest possible value.

4 Solving the Resulting Optimization Problem

Solving the resulting optimization problem: first step. Let us first find the optimal function $\rho_0(r)$ corresponding to a given probability density function $\rho(r)$.

To find this function $\rho_0(r)$, we must solve the above constraint optimization problem. For this problem, the Lagrange multiplier method leads to minimizing the following objective function:

$$\int \frac{1}{(\rho_0(r))^\beta} \cdot \rho(r)\, dr + \lambda \cdot \left(\int \rho_0(r)\, dr - n \right).$$

Differentiating this expression with respect to each unknown $\rho_0(r)$ and equating the derivative to 0, we conclude that

$$-\beta \cdot \frac{\rho(r)}{(\rho_0(r))^{\beta+1}} + \lambda = 0,$$

hence $(\rho_0(r))^{\beta+1} = \text{const} \cdot \rho(r)$, with $\text{const} = \frac{\lambda}{\beta}$. Thus,

$$\rho_0(r) = c \cdot (\rho(r))^{1/(\beta+1)},$$

for some constant c. This constant can be determined from the condition that $\int \rho_0(r)\, dr = n$. Substituting the above expression into this condition, we conclude that

$$\rho_0(r) = n \cdot \frac{(\rho(r))^{1/(\beta+1)}}{\int (\rho(s))^{1/(\beta+1)}\, ds}.$$

Solving the resulting optimization problem: second step. Now that we know which point density $\rho_0(r)$ is optimal for the given probability density function $\rho(r)$, we need to find the probability density function $\rho(r)$ for which the corresponding best-case disutility function attains the smallest possible value.

To find such $\rho(r)$, let us first use our result of solving the first-step optimization problem to come up with an explicit (and thus, easier-to-minimize) expression for the best-case average disutility.

Substituting the above expression for $\rho_0(r)$ into the formula for E_U, we get

$$E_U = \int \frac{\rho(r)}{(\rho_0(r))^{1/(\beta+1)}}\,dr = \frac{\int (\rho(r))^{1/(\beta+1)}\,dr}{\left(\int (\rho(r))^{1/(\beta+1)}\,dr\right)^{\beta}} = \frac{1}{\left(\int (\rho(r))^{1/(\beta+1)}\,dr\right)^{\beta-1}}.$$

Thus, depending on whether $\beta < 1$ or $\beta > 1$, minimizing the best-case expected disutility is equivalent to either minimizing or maximizing the integral

$$J \stackrel{\text{def}}{=} \int (\rho(r))^{1/(\beta+1)}\,dr.$$

Comment. It is worth mentioning that the resulting objective function for selecting a probability distribution is one of the scale-invariant objective functions described in [10]. This is not surprising since we based our derivation on the ideas of invariance with respect to selecting a measuring unit – which, in mathematical terms, is exactly scale invariance.

Solving the resulting optimization problem: final step. Now, in a class of probability density functions which are consistent with our knowledge, we need to find the ones for which the above expression J is optimal. In particular, if our knowledge consists of a moment-related constraint $\int |r|^k \cdot \rho(r)\,dr = M_k$, then we need to optimize the expression J under this constraint and the additional constraint $\int \rho(r)\,dr = 1$. For this constraint optimization problem, the Lagrange multiplier method leads to the need for optimizing the following expression:

$$\int (\rho(r))^{1/(\beta+1)}\,dr + \lambda_0 \cdot \left(\int \rho(r)\,dr - 1\right) + \lambda_k \cdot \left(\int |r|^k \cdot \rho(r)\,dr - M_k\right).$$

Differentiating this expression with respect to each unknown $\rho(r)$ and equating the derivative to 0, we conclude that

$$\frac{1}{\beta+1} \cdot (\rho(r))^{-\beta/(\beta+1)} = -\lambda_0 + \lambda_k \cdot |r|^k,$$

hence

$$\rho(r) = \frac{1}{(C_0 + C_1 \cdot |r|^k)^{1+1/\beta}},$$

for some constants C_0 and C_1.

Asymptotically, when $|r|$ increases, we have

$$\rho(r) \sim |r|^{-\alpha}$$

for $\alpha = k \cdot \left(1 + \dfrac{1}{\beta}\right)$. Thus, *we indeed have an explanation for the effectiveness of statistical methods based on heavy-tailed distributions.*

Comment. In this paper, we concentrated on the use of heavy-tail-based statistical methods in econometric (and similar) applications. However, such methods are also effectively used in many other application areas, e.g., in the analysis of graphs and networks; see, e.g., [2] and references therein. It would be interesting to analyze to what extent similar ideas can explain the effectiveness of these methods in network studies and in other application areas.

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.

The authors are thankful to the anonymous referees for valuable suggestions.

References

1. J. Aczél, *Lectures on Functional Equations and Their Applications*, Dover, New York, 2006.
2. A.-L. Barabasi, *Network Science*, Cambridge University Press, Cambridge, Massachusetts, 2016.
3. J. Beirlant, Y. Goegevuer, J. Teugels, and J. Segers, *Statistics of Extremes: Theory and Applications*, Wiley, Chichester, 2004.
4. B. K. Chakrabarti, A. Chakraborti, and A. Chatterjee, *Econophysics and Sociophysics: Trends and Perspectives*, Wiley-VCH, Berlin, 2006.
5. A. Chatterjee, S. Yarlagadda, and B. K. Chakrabarti, *Econophysics of Wealth Distributions*, Springer-Verlag Italia, Milan, 2005.
6. J. D. Farmer and T. Lux (eds.), *Applications of statistical physics in economics and finance*, a special issue of the *Journal of Economic Dynamics and Control*, 2008, Vol. 32, No. 1, pp. 1–320.
7. P. C. Fishburn, *Utility Theory for Decision Making*, John Wiley & Sons Inc., New York, 1969.
8. C. P. Gomez and D. B. Shmoys, "Approximations and Randomization to Boost CSP Techniques", *Annals of Operations Research*, 2004, Vol. 130, pp. 117–141.
9. E. T. Jaynes and G. L. Bretthorst, *Probability Theory: The Logic of Science*, Cambridge University Press, Cambridge, UK, 2003.

10. 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.
11. R. D. Luce and R. Raiffa, *Games and Decisions: Introduction and Critical Survey*, Dover, New York, 1989.
12. B. G. Malkiel, *A Random Walk Down Wall Street: The Time-Tested Strategy for Successful Investing*, W. W. Norton & Company, New York, 2007.
13. B. Mandelbrot, "The variation of certain speculative prices", *J. Business*, 1963, Vol. 36, pp. 394–419.
14. B. Mandelbrot, *The Fractal Geometry of Nature*, Freeman, San Francisco, California, 1983.
15. B. Mandelbrot and R. L. Hudson, *The (Mis)behavior of Markets: A Fractal View of Financial Turbulence*, Basic Books, 2006.
16. N. Markovich (ed.), *Nonparametric Analysis of Univariate Heavy-Tailed Data: Research and Practice*, Wiley, Chichester, 2007.
17. J. McCauley, *Dynamics of Markets, Econophysics and Finance*, Cambridge University Press, Cambridge, Massachusetts, 2004.
18. 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.
19. P. V. Novitskii and I. A. Zograph, *Estimating the Measurement Errors*, Energoatomizdat, Leningrad, 1991 (in Russian).
20. A. I. Orlov, "How often are the observations normal?", *Industrial Laboratory*, 1991, Vol. 57, No. 7, pp. 770–772.
21. S. G. Rabinovich, *Measurement Errors and Uncertainty: Theory and Practice*, Springer Verlag, Berlin, 2005.
22. H. Raiffa, *Decision Analysis*, Addison-Wesley, Reading, Massachusetts, 1970.
23. S. I. Resnick, *Heavy-Tail Phenomena: Probabilistic and Statistical Modeling*, Springer-Varlag, New York, 2007.
24. D. J. Sheskin, *Handbook of Parametric and Nonparametric Statistical Procedures*, Chapman and Hall/CRC, Boca Raton, Florida, 2011.
25. S. V. Stoyanov, B. Racheva-Iotova, S. T. Rachev, and F. J. Fabozzi, "Stochastic models for risk estimation in volatile markets: a survey", *Annals of Operations Research*, 2010, Vol. 176, pp. 293–309.
26. P. Vasiliki and H. E. Stanley, "Stock return distributions: tests of scaling and universality from three distinct stock markets", *Physical Review E: Statistical, Nonlinear, and Soft Matter Physics*, 2008, Vol. 77, No. 3, Pt. 2, Publ. 037101.