From Normal Distribution to What? How to Best Describe Distributions with Known Skewness

Olga Kosheleva and Vladik Kreinovich

Abstract In many practical situations, we only have partial information about the probability distribution – e.g., all we know is its few moments. In such situations, it is desirable to select one of the possible probability distributions. A natural way to select a distribution from a given class of distributions is the maximum entropy approach. For the case when we know the first two moments, this approach selects the normal distribution. However, when we also know the third central moment – corresponding to skewness – a direct application of this approach does not work. Instead, practitioners use several heuristic techniques, techniques for which there is no convincing justification. In this paper, we show that while we cannot directly apply the maximum entropy approach to the skewness situation, we can apply it approximately – with any approximation accuracy we want – and get a meaningful answer to the above selection problem.

1 Formulation of the Problem

Need to select a single probability distribution from a class of possible distributions. Many statistical methods assume that we know the corresponding probability distribution. In many practical situations, however, we only have partial information about the probability distribution, i.e., all we know about the actual distribution is that it belongs to some class C of distributions.

A frequent example is when all we know about a distribution are the values of the first several moments.

Olga Kosheleva
Department of Teacher Education, University of Texas at El Paso
500 W. University, El Paso, Texas 79968, USA, e-mail: olgak@utep.edu

Vladik Kreinovich
Department of Computer Science, University of Texas at El Paso
500 W. University, El Paso, Texas 79968, USA, e-mail: vladik@utep.edu

In general, to apply the corresponding statistical method, we need to select one of the distributions from the class C.

Maximum entropy approach. This problem is ubiquitous is many applications of statistics, and there is a known way to solve this problem: out of all possible distributions, we should select the distribution with the largest possible entropy

$$S \stackrel{\text{def}}{=} -\int f(x) \cdot \ln(f(x))\, dx, \tag{1}$$

where $f(x)$ denotes the probability density function (pdf); see, e.g., [2].

Often, this approach works very well. This method works well for many classes of distributions. For example, if all we know are the first two moments – i.e., equivalently, the mean

$$\mu = \int x \cdot f(x)\, dx \tag{2}$$

and the second central moment (the variance)

$$V = \int (x-\mu)^2 \cdot f(x)\, dx, \tag{3}$$

then the maximum entropy approach select the normal (Gaussian) distribution

Indeed, in this case, the maximum entropy approach means selecting:

- among all distributions $f(x)$ that satisfy the usual condition

$$\int f(x)\, dx = 1 \tag{4}$$

  and the additional conditions (2) and (3) for given μ and V,
- the distribution with the largest value of the entropy S.

By using the Lagrange multiplier method, we can reduce this constraint optimization problem to the unconstrained problem of maximizing a function

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

$$\lambda_2 \cdot \left(\int (x-\mu)^2 \cdot f(x)\, dx - V\right) \tag{5}$$

for some Lagrange multipliers λ_i. Differentiating the expression (5) with respect to each unknown $f(x)$, we conclude that

$$-\ln(f(x)) - 1 + \lambda_0 + \lambda_1 \cdot x + \lambda_2 \cdot (x-\mu)^2 = 0, \tag{6}$$

i.e., $\ln(f(x)) = a_0 + a_1 \cdot x + a_2 \cdot x^2$ for some a_i, and thus

$$f(x) = \exp(a_0 + a_1 \cdot x + a_2 \cdot x^2). \tag{7}$$

For $a_2 < 0$, we indeed get the normal distribution.

Selecting the normal distribution is a very good choice, since this distribution is indeed ubiquitous; see, e.g., [3].

The problem: the maximum entropy approach does not work for skewness (and higher moments). Sometimes, in addition to the mean and variance, we also know higher central moments: the third central moment

$$\mu_3 = \int (x - \mu)^3 \cdot f(x)\, dx$$

describing the distribution's asymmetry (skewness), and higher-order central moments

$$\mu_n = \int (x - \mu)^n \cdot f(x)\, dx,$$

for $n > 3$.

Indeed, for the case of skewness, the Lagrange multiplier method – similar to the one used before – leads to the expression

$$f(x) = \exp(a_0 + a_1 \cdot x + a_2 \cdot x^2 + a_3 \cdot x^3). \tag{8}$$

Unfortunately, it is not possible to select the coefficient a_i for which the integral of this expression is equal to 1; indeed:

- when $a_3 > 0$, this expression tends to infinity for $x \to \infty$ – so its integral is infinite, and
- when $a_3 < 0$, this expression tends to infinity for $x \to -\infty$ – so its integral is also infinite.

Comment. The same problem appears not only for the third moment, but also for any case in which the highest known central moment is of odd order.

What people do in this case. Since we cannot directly use the maximum entropy approach, researchers has proposed several heuristic methods. For the case of skewness, the most widely known approach is to use a special family known as |it skew-normal distributions; see, e.g., [1].

The problem is that this approach is heuristic, there is no convincing explanation of why this particular family has been selected.

What we do in this paper. In this paper, we show that while we cannot directly apply the maximum entropy approach, we can apply it approximately – with any approximation accuracy we want – and get a meaningful answer to the selection problem. Our idea works not only for skewness, it works for moments of arbitrary order.

2 Our Idea: Motivation, Description, and Computational Aspects

Motivations for the idea. If we do not have any information about the third moment, we get a normal distribution (7). The normal distribution is symmetric with respect to the mean μ. So, for this distribution, the third central moment if equal to 0.

When we add information about the third central moment, we get a more general expression (8). This expression can be written as

$$f(x) = \exp(a_0 + a_1 \cdot x + a_2 \cdot x_2) \cdot \exp(a_3 \cdot x^3). \tag{9}$$

When the third central moment is 0, we have $a_3 = 0$. Thus, when the third central moment is small, the value a_3 is close to 0 and is, thus, also small. So, we can safely expand the corresponding expression $\exp(a_3 \cdot x^3)$ in Taylor series and keep only the first few terms in this expansion. For example, in the first approximation, we can take $\exp(a_3 \cdot x^3) \approx 1 + a_3 \cdot x^3$.

Issues related to this idea. There are two issues related to the above formulation.

The first issue is that we are describing the probability density function which is applicable for all possible values x. When a_3 is small, for reasonable values x, the product $a_3 \cdot x^3$ is also small. However, for large x, the product $a_3 \cdot x^3$ become large, and the linear expression $1 + a_3 \cdot x^3$ is no longer a good approximation to the exponential terms $\exp(a_3 \cdot x^3)$. This is true, but good news is that this term is multiplied by the Gaussian expression (7) which, for large x, is indistinguishable from 0. So, whatever we multiply this practically-zero by, we still get practically zero.

The second issue is more serious: that for some x, the approximating expression $1 + a_3 \cdot x^3$ – and thus, its product with (7) which is supposed to be a probability density function – becomes negative:

- when $a_3 > 0$, this expression becomes negative when $x \to -\infty$, and
- when $a_3 < 0$, this expression becomes negative when $x \to +\infty$.

There are two ways to deal with this issue:

- as we have mentioned, for x close to ∞ or to $-\infty$, the value (7) is practically 0, so its product with our expression is practically zero too; thus, we can simply ignore the fact that this practically-zero product is negative;
- alternatively, we can force this expression to be always non-negative, by taking the maximum of this expression and 0.

With this in mind, we arrive at the following family of distributions.

Description of the proposed family of distributions. We select the value $k \geq 1$. For each k, we propose to use one of the following two 3-parametric families of distributions:

$$f(x) = \exp(a_0 + a_1 \cdot x + a_2 \cdot x^2) \cdot \left(1 + a_3 \cdot x^3 + \ldots + \frac{(a_3 \cdot x^3)^k}{k!}\right) \tag{10}$$

or

$$f(x) = \exp(a_0 + a_1 \cdot x + a_2 \cdot x^2) \cdot \max\left(0, 1 + a_3 \cdot x^3 + \ldots + \frac{(a_3 \cdot x^3)^k}{k!}\right). \quad (11)$$

Comments.

- In particular, for the simplest case $k = 1$, we have the families

$$f(x) = \exp(a_0 + a_1 \cdot x + a_2 \cdot x^2) \cdot (1 + a_3 \cdot x^3) \quad (12)$$

and

$$f(x) = \exp(a_0 + a_1 \cdot x + a_2 \cdot x^2) \cdot \max(0, 1 + a_3 \cdot x^3). \quad (13)$$

- As we mentioned earlier, each expression (10)-(11) is not exactly what the maximum likelihood approach suggests, but it is an approximation to this suggestion – and the larger k, the more accurate is this approximation. Thus, whatever accuracy we want, if we select a sufficiently large k, we will get the approximation with the desired accuracy.
- A similar idea can be used for the cases when we know higher moments: in this case, we use a Taylor expansion of the expression

$$\exp(a_3 \cdot x^2 + \ldots + a_n \cdot x^n).$$

- Similarly, in the multi-dimensional case, we consider the product of the Gaussian terms

$$\exp\left(a_0 + \sum_i a_i \cdot x_i + \sum_{i,j} a_{i,j} \cdot x_i \cdot x_j\right) \quad (14)$$

and the sum of several first terms in the Taylor expression of the additional factor

$$\exp\left(\sum_{i,j,k} a_{i,j,k} \cdot x_i \cdot x_j \cdot x_k\right). \quad (15)$$

How difficult is it to perform computations with this new family of distributions: computing moments. The main purpose of the corresponding family is to select the parameters a_i that match the given moments, and to make predictions based on the selected values of the parameters.

To fit the given values of the moments, we need to be able to compute the integrals $\int f(x) \cdot x^d \, dx$ for $d = 0, 1, 2, 3, \ldots$ The expression (10) is a linear combination of the products of Gaussian pdf and a power x^p for some p. Thus, to find the corresponding moments, it is sufficient to be able to compute the products of the Gaussian pdf and the power x^{d+p} – i.e., the moments of the Gaussian distribution, and there are known formulas for these moments; see, e.g., [3].

Comments. For normal distribution with center at μ, moments of odd order ar 0s.

For moments of even order, these formulas can be contained from the known expression for the full probability formula for the normal distribution

$$\int \frac{1}{\sqrt{2\pi}\cdot\sigma}\cdot\exp\left(-\frac{x^2}{2\sigma^2}\right)\,dx = 1, \tag{15}$$

if we differentiate both sides by σ. If we differentiate once, we get an expression proportional to the integral describing the second moment. If we differentiate once again, we get the expression for the 4th moment, etc.

How difficult is it to perform computations with this new family of distributions: computing cdf. Similarly, we can come up with an explicit expression for the cumulative distribution function (cdf) for the new distribution, We start with the formula

$$\int_{-\infty}^{x_0} \frac{1}{\sqrt{2\pi}\cdot\sigma}\cdot\exp\left(-\frac{x^2}{2\sigma^2}\right)\,dx = \Phi\left(\frac{x_0}{\sigma}\right), \tag{16}$$

where $\Phi(x)$ denotes the cdf of the standard normal distribution, with 0 mean and standard deviation 1. If we differentiate both sides with respect to σ, we get the interval of x^2 times Gaussian from $-\infty$ to x_0. If we differentiate twoce, we get a similar integral for x^4 times Gaussian, etc., all even order expressions can be thus attained.

For odd number expressions, we can introduce a new variable $t = x^2$, then we get

$$\int x^{2k+1}\cdot\exp(-x^2)\,dx = \frac{1}{2}\cdot\int \exp(-t)\cdot t^k\,dt. \tag{17}$$

Here, by using the integration by part formula $\int u\cdot v'\,dt = u\cdot v - \int v\cdot u'\,dt$, with $v = -\exp(-t)$ and $u = t^k$, we conclude that

$$\int \exp(-t)\cdot t^k = -\exp(-t)\cdot t^k + k\cdot\int \exp(-t)\cdot t^{k-1}\,dt. \tag{18}$$

So, the computation of the integral (17) for any integer k is reduced to the case of $k-1$, etc., until we reach the case $k = 0$ for which $\int \exp(-t)\,dt = -\exp(-t) + C$, where C is the integration constant.

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-1834620 and HRD-2034030 (CAHSI Includes), EAR-2225395 (Center for Collective Impact in Earthquake Science C-CIES), and by the AT&T Fellowship in Information Technology.

It was also supported by the program of the development of the Scientific-Educational Mathematical Center of Volga Federal District No. 075-02-2020-1478, and by a grant from the Hungarian National Research, Development and Innovation Office (NRDI).

The authors are thankful to all the participants of the 7th International Conference on Financial Econometrics ECONVN'2024 (Ho Chi Minh City, Vietnam, January 9–11, 2024), especially to Tonghui "Tony" Wang, for valuable discussions.

References

1. A. Azzalini and A. Capitanio, *The Skew-Normal and Related Families*, Cambridge University Press, Cambridge, Massachusetts, 2013.
2. E. T. Jaynes and G. L. Bretthorst, *Probability Theory: The Logic of Science*, Cambridge University Press, Cambridge, UK, 2003.
3. D. J. Sheskin, *Handbook of Parametric and Nonparametric Statistical Procedures*, Chapman and Hall/CRC, Boca Raton, Florida, 2011.