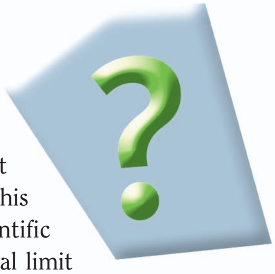
Kiseon Kim and Georgy Shevlyakov

WHY GAUSSIANITY?

"Physicists believe that the Gaussian law has been proved
in mathematics while mathematicians think that
it was experimentally established in physics."

—Henri Poincaré

An attempt to explain this phenomenon



This witty remark of a great mathematician [1] reflects the fact of the ubiquitous use and success of the Gaussian distribution law and at the same moment gives both a humorous and serious hint to explain this phenomenon. The majority of members of the scientific community shares the common belief that it is due to the central limit theorem (CLT). We will show that the CLT is not only a unique reason but perhaps it is even not the main reason.

In this article, we try to answer the question: "Why the ubiquitous use and success of the Gaussian distribution law?" The history of the Gaussian or normal distribution is rather long, having existed for nearly 300 years since it was discovered by de Moivre in 1733, and the related literature is immense. An extended and thorough treatment of the topic and a

Digital Object Identifier 10.1109/MSP.2007.913700

1053-5888/08/$25.00©2008IEEE

survey of the works in the related area are given in the posthumously edited book of E.T. Jaynes [2], and we partially follow this source, in particular while considering the history of the posed question. The important aspects of the general history of noise, especially of Brownian motion, are given in [3]. Our main contribution to the topic is concerned with highlighting the role of Gaussian models in signal processing based on the optimal property of the Gaussian distribution minimizing Fisher information over the class of distributions with a bounded variance.

In what follows, we deal only with the univariate Gaussian distribution, omitting the properties of multivariate Gaussian distribution. First of all, we present the ideas of classical derivations of the Gaussian law. Then we consider its properties and characterizations including the CLT and minimization of the distribution entropy and Fisher information. Finally, we dwell on the connections between Gaussianity and robustness in signal processing.

HISTORICAL PRELIMINARIES

The Gaussian or normal distribution density is defined as

$$\mathcal{N}(x;\ \mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right], \quad -\infty < x < \infty, \tag{1}$$

where μ and σ are the parameters of location (mean) and scale (standard deviation), respectively. Its standard form is commonly denoted by $\phi(x) = \mathcal{N}(x;\ 0,1)$. (See Figure 1.)

Using Stirling's approximations for factorials, it can be shown that the Gaussian distribution is a limiting form of the binomial distribution [4]

$$P_{n,k}(p) = C_n^k p^k q^{n-k}, \qquad k = 0, 1, \dots, n,$$

$$0 < p < 1, \qquad q = 1-p, \qquad C_n^k = \frac{n!}{k!(n-k)!},$$

$$P_{n,k}(p) \to \frac{1}{\sqrt{npq}}\,\phi\left(\frac{k-np}{\sqrt{npq}}\right) \quad \text{as} \quad n,k \to \infty$$

with $(k-np)/\sqrt{npq}$ finite.

In the particular case $p = q = 1/2$, the Gaussian distribution had been found by de Moivre [5] who did not recognize its significance. In the general case $0 < p < 1$, Laplace [6] had derived its main properties and suggested that it should be tabulated due to its importance. Gauss [7] considered another derivation of this distribution (not as a limiting form of the binomial distribution); it became popularized by his work and thus his name was attached to it. The fundamental Boltzmann distribution of statistical mechanics [8], exponential in energies, is the Gaussian in velocities [9].

It seems likely that the term "normal" is associated with a linear regression model $y = \Phi\beta + e$, where the vector y and the matrix Φ are known, the vector of parameters β and the noise vector e unknown; to solve this linear regression problem, Gauss [10] suggested the least-squares (LS) method and called the system of equations $\Phi'\Phi\beta = \Phi' y$, which give the least square parameter estimates $\widehat{\beta}$, the *normal equations*.

One more name *central distribution* originating from the term CLT was suggested by Pólya [11] and then it was actively backed by Jaynes [2].

A well-known historian of statistics, Stigler [12] formulates an universal law of eponymy that "no discovery is named for its original discoverer." Jaynes [2] truly notices that "the history of this terminology excellently confirms this law, since the fundamental nature of this distribution and its main properties were derived by Laplace when Gauss was six years old; and the distribution itself had been found by de Moivre before Laplace was born."

DERIVATIONS OF THE GAUSSIAN DISTRIBUTION

DERIVATION OF GAUSS (1809)

Consider a sample of $n+1$ independent observations $x_0, x1, \dots, x_n$ taken from the distribution with density $f(x;\theta)$, where θ is a parameter of location. Its maximum likelihood estimate $\widehat{\theta}$ must satisfy

$$\frac{\partial}{\partial\theta}\log\left\{\prod_{i=0}^{n} f(x_i;\theta)\right\} = \sum_{i=0}^{n}\frac{\partial}{\partial\theta}\log f(x_i;\theta) = 0.$$

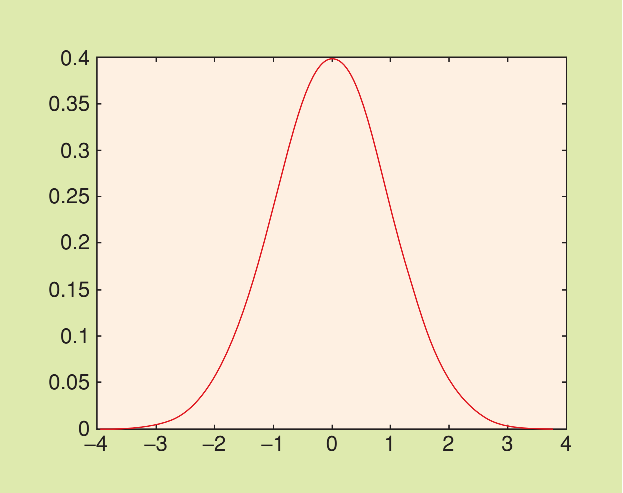
Assuming differentiability of $f(x;\theta)$ and denoting

$$\log f(x;\theta) = g(x-\theta),$$

we have that the maximum likelihood estimate will be the solution of

$$\sum_{i=0}^{n} g'(x_i - \widehat{\theta}) = 0. \tag{2}$$

Gauss [7] asked the following question: "What would be a distribution density $f(x;\theta)$ for which the maximum likelihood estimate $\widehat{\theta}$ is the sample mean



[FIG1] **Standard Gaussian distribution density.**

$$\widehat{\theta} = \overline{x} = \frac{1}{n+1}\sum_{i=0}^{n} x_i \quad ? \quad ”$$

Note that here we use the modern terminology adopted by the scientific community more than a century later (the method of maximum likelihood was proposed by Fisher in 1921 [13]).

To answer the posed question, let us apply the following method of functional equations [14]. First, set $x_0 = x_1 = \cdots = x_n = 0$, then check that $\overline{x} = 0$,write out (2) and get that

$$g'(0) = 0. \tag{3}$$

Second, set $x_0 = u,\ x_1 = -u,\ x_2 = 0,\ \ldots,\ x_n = 0$, check that $\overline{x} = 0$ and then from (2) and (3) it follows that

$$g'(-u) = -g'(u). \tag{4}$$

Third, set $x_0 = -u(n+1),\ x_1 = 0,\ \ldots,\ x_n = 0$, get that $\overline{x} = -u$ and then from (2) and (4) it follows that

$$g'(-nu) + \sum_{i=1}^{n} g'(u) = 0$$

and the function $g(u)$ must satisfy the functional equation

$$g'(nu) = n\,g'(u), \qquad n = 1, 2, 3, \ldots. \tag{5}$$

From (5) it follows that

$$g'(1) = n\,g'\left(\frac{1}{n}\right) \qquad n = 1, 2, 3, \ldots,$$
$$g'\left(\frac{m}{n}\right) = m\,g'\left(\frac{1}{n}\right) \qquad m, n = 1, 2, 3, \ldots,$$

and we get the linear equation that holds for all rational numbers

$$g'\left(\frac{m}{n}\right) = a\frac{m}{n} \qquad m, n = 1, 2, 3, \ldots, \tag{6}$$

where $a = g'(1)$. Since any real number can be arbitrarily accurately approximated by rational numbers, linear equation (6) holds for real u

$$g'(u) = au$$

with the corresponding quadratic form of $g\ (u)$

$$g(u) = \frac{1}{2}au^2 + b,$$

and the Gaussian density

$$f(x;\theta) = \mathcal{N}(x;\ \theta, 1/\sqrt{\alpha}), \qquad \alpha = -a > 0.$$

Look at this derivation from another point of view: Gauss assumed the sample mean (the estimate of the LS method, the honor of inventing that he shares with Legendre [15]) due to its computational convenience and derived the Gaussian law. This line of reasoning is quite the opposite to the modern exposition in textbooks on statistics and signal processing where the LS method is derived from the assumed Gaussianity.

[IN THIS ARTICLE, WE TRY TO ANSWER THE QUESTION: “WHY THE UBIQUITOUS USE AND SUCCESS OF THE GAUSSIAN DISTRIBUTION LAW?”]

DERIVATION OF HERSCHEL (1850) AND MAXWELL (1860)

The astronomer John Herschel [16] considered the two-dimensional probability distribution for errors in measuring the position of a star and, ten years later, James Clerk Maxwell [9] gave a three-dimensional version of the same derivation for the probability distribution density for the velocities of molecules in a gas, which has become well-known to physicists as the *Maxwellian velocity distribution law* fundamental in kinetic theory and statistical mechanics.

Here we consider the two-dimensional case. Let x be the error in the east-west direction and y the error in the north-south direction, and $f(x, y)$ be the joint probability distribution density. First, assume the independence and identity of coordinate error distributions

$$f(x, y)\,dx\,dy = f(x)\,dx \times f(y)\,dy. \tag{A1}$$

Second, require that this distribution should be invariant to the rotation of the coordinate axes

$$f(x, y) = g(x^2 + y^2). \tag{A 2}$$

From assumptions (A1) and (A2) it immediately follows that

$$g(x^2) = f(x)f(0), \qquad g(y^2) = f(y)f(0),$$

yielding the functional equation

$$g(x^2 + y^2) \propto g(x^2)\,g(y^2),$$

with the exponential solution [14]

$$g(u^2) = \exp(\lambda u^2), \qquad \lambda < 0,$$

and the Gaussian law for the coordinate error distribution

$$f(x) = \mathcal{N}(x;\ 0, 1/\sqrt{-2\lambda}).$$

DERIVATION OF LANDON (1941)

Vernon D. Landon [17], an electrical engineer, considered the distribution density $p(x;\ \sigma^2)$ of the electrical noise voltage $x(t)$ observed in a circuit at time t, where σ is the standard deviation of the noise voltage. He suggested that this distribution is so universal that it must be determined theoretically: namely, that there exists a hierarchy of distributions $p(x;\ \sigma^2)$ of the same functional form characterized only by σ. Moreover, all the different levels of σ at which it occurs correspond to different noise environments, such as temperatures, amplifications, impedance levels, and even to different kinds of sources—natural or man-made industrial, resulting only in a new value of σ and preserving the functional shape. Landon's original derivation concerned the particular case of a sinusoidal noise amplitude; in what follows, we use the generalization of Landon's approach proposed by Jaynes [2].

Suppose the noise amplitude x has the distribution density $p(x;\ \sigma^2)$. Let it be incremented by a small extra contribution Δx so that $x' = x + \Delta x$, where Δx is small compared with σ, and let Δx have a distribution density $q(\Delta x)$ independent of $p(x;\ \sigma^2)$. Then, given a specific Δx, the probability for the new noise amplitude to have the value x' would be the previous probability that x should have the value $(x' - \Delta x)$. Next, by the product and sum rules of probability theory, the new distribution density is the convolution

$$f(x') = \int p(x' - \Delta x;\ \sigma^2)\, q(\Delta x)\, d(\Delta x). \tag{7}$$

Expanding (7) in powers of the small quantity Δx and dropping the prime, we get

$$\begin{aligned} f(x) =& p(x;\sigma^2) - \frac{\partial p(x;\sigma^2)}{\partial x}\int \Delta x\, q(\Delta x)\, d(\Delta x) \\ &+ \frac{1}{2}\frac{\partial^2 p(x;\sigma^2)}{\partial x^2}\int (\Delta x)^2\, q(\Delta x)\, d(\Delta x) + \cdots, \end{aligned}$$

or

$$\begin{aligned} f(x) = p(x;\ \sigma^2) - \overline{\Delta x}\,\frac{\partial p(x;\ \sigma^2)}{\partial x} \\ + \frac{1}{2}\,\overline{\Delta x^2}\,\frac{\partial^2 p(x;\ \sigma^2)}{\partial x^2} + \cdots, \end{aligned} \tag{8}$$

where $\overline{\Delta x}$ and $\overline{\Delta x^2}$ stand for the expectation and second moment of the increment Δx, respectively.

Since the increment is as likely to be positive as negative, assume that $\overline{\Delta x} = 0$. Moreover, assume also that the moments of order higher than two can be neglected, that is, $\overline{\Delta x^k} = o\left(\overline{\Delta x^2}\right)$ for all $k > 2$. Then (8) can be rewritten as follows

$$f(x) = p(x;\ \sigma^2) + \frac{1}{2}\overline{\Delta x^2}\,\frac{\partial^2 p(x;\ \sigma^2)}{\partial x^2} + o\left(\overline{\Delta x^2}\right). \tag{9}$$

Further, the variance of x is increased to $\sigma^2 + \mathrm{Var}[\Delta x]$, and Landon's invariancy property requires that $f(x)$ should be equal to

$$f(x) = p(x;\ \sigma^2 + \mathrm{Var}[\Delta x]). \tag{10}$$

Expanding (10) with respect to small $\mathrm{Var}[\Delta x]$, we get

$$f(x) = p(x;\ \sigma^2) + \mathrm{Var}[\Delta x]\,\frac{\partial p(x;\ \sigma^2)}{\partial(\sigma^2)} + o\left(\overline{\Delta x^2}\right). \tag{11}$$

Equating the main parts of (9) and (11), we obtain the following condition for this invariance:

$$\frac{\partial p(x;\ \sigma^2)}{\partial(\sigma^2)} = \frac{1}{2}\,\frac{\partial^2 p(x;\ \sigma^2)}{\partial x^2}$$

that is the "diffusion equation" [3], whose solution with the initial condition $p(x;\ \sigma^2 = 0) = \delta(x)$ is given by the Gaussian distribution

$$p(x;\sigma^2) = \mathcal{N}(x;\ 0, \sigma).$$

The two crucial points of this derivation are, first, to guarantee the expansions (8) and (11) hold, we should consider smooth distributions $p(x;\sigma^2)$ and $q(\Delta x)$; second, to neglect the moments of Δx of order higher than two, we should at least assume their existence. Thus, discontinuous and heavy-tailed distributions, such as Laplace and Cauchy, are excluded. Here we conclude quoting Jaynes (see [2, p. 206]): "… This is, in spirit, an incremental version of the CLT; instead of adding up all the small contributions at once, it takes them into account one at a time, requiring that at each step the new probability distribution has the same functional form (to second order in Δx). … this is just the process by which noise is produced in Nature—by addition of many small increments, one at a time (for example, collisions of individual electrons with atoms, each collision radiating another tiny impulse of electromagnetic waves, whose sum is the observed noise). Once a Gaussian form is attained, it is preserved; this process can be stopped at any point, and the resulting final distribution still has the Gaussian form."

PROPERTIES OF THE GAUSSIAN DISTRIBUTION

Here we enlist several properties of the Gaussian distribution:

- the convolution of two Gaussian functions is another Gaussian function
- the Fourier transform of a Gaussian function is another Gaussian function.
- the CLT
- maximizing entropy
- minimizing Fisher information.

Apparently, the CLT, and based on it Gaussian approximations of the sums of random variables, can be regarded as one of the main reasons for the ubiquitous use of a Gaussian distribution. Nevertheless, we begin from the other ones, which also relate to the CLT and explain why a Gaussian form, once attained, is further preserved; the remaining properties play each its own role deserving a separate consideration.

Henceforth, a function $f(x)$ is said to be Gaussian or of a Gaussian form if it is equal to the Gaussian distribution density with accuracy up to the norming constant: $f(x) \propto \mathcal{N}(x;\ \mu, \sigma)$.

CONVOLUTION OF GAUSSIANS

The operation of convolution arises in computing the distribution density $f_Y(y)$ of the sum $Y = X_1 + X_2$ of two independent random variables X_1 and X_2 with densities $f_1(x_1)$ and $f_2(x_2)$, respectively, and it is given by the following relations

$$\begin{aligned} f_Y(y) &= \int_{-\infty}^{\infty} f_1(x_1) f_2(y - x_1)\, dx_1 \\ &= \int_{-\infty}^{\infty} f_1(y - x_2) f_2(x_2)\, dx_2. \end{aligned} \tag{12}$$

Let the independent random variables X_1 and X_2 be Gaussian with densities $\mathcal{N}(x_1; \mu_1, \sigma_1)$ and $\mathcal{N}(x_2; \mu_2, \sigma_2)$. Substitute these densities into (12) and get

$$\begin{aligned} f_Y(y) =& \frac{1}{2\pi\sigma_1\sigma_2} \\ & \int_{-\infty}^{\infty} \exp\left\{-\frac{1}{2}\left[\left(\frac{x-\mu_1}{\sigma_1}\right)^2 + \left(\frac{y-x-\mu_2}{\sigma_2}\right)^2\right]\right\} dx \\ =& \frac{1}{2\pi\sigma_1\sigma_2} \int_{-\infty}^{\infty} \exp\left\{-\frac{1}{2}\left(ax^2 + bx + c\right)\right\} dx, \end{aligned}$$

where

$$a = \frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2}, \quad b = \frac{\mu_2 - y}{\sigma_2^2} - \frac{\mu_1}{\sigma_1^2}, \quad c = \frac{\mu_1^2}{\sigma_1^2} + \frac{(y-\mu_2)^2}{\sigma_2^2}.$$

Further use the following formula [18]

$$\int_{-\infty}^{\infty} \exp\left\{-\frac{1}{2}\left(ax^2 + bx + c\right)\right\} dx = \sqrt{\frac{2\pi}{a}} \exp\left\{\frac{b^2 - ac}{2a}\right\}, \quad a > 0,$$

and obtain that the sum of independent Gaussian random variables is distributed according to the Gaussian law

$$f_{X_1+X_2}(y) = \mathcal{N}\left(y;\ \mu_1 + \mu_2, \sqrt{\sigma_1^2 + \sigma_2^2}\right).$$

FOURIER TRANSFORM OF A GAUSSIAN

The Fourier transform of the Gaussian distribution density is defined as

$$\phi_X(t) = \int_{-\infty}^{\infty} e^{itx} \mathcal{N}(x; \mu, \sigma)\, dx \tag{13}$$

and it is well-known as the characteristic function of the Gaussian random variable X.

Setting $z = x - \mu$, we can rewrite (13) as

$$\begin{aligned} \phi_X(t) =& \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-z^2/2\sigma^2} e^{it(\mu+z)}\, dx \\ =& e^{it\mu} \frac{1}{\sigma\sqrt{2\pi}} \\ & \times \left(\int_{-\infty}^{\infty} e^{-z^2/2\sigma^2} \cos tz\, dz + i \int_{-\infty}^{\infty} e^{-z^2/2\sigma^2} \sin tz\, dz\right). \end{aligned}$$

The second integral is zero as the integral of an odd function over a symmetric interval. For computing the first integral, we use the Laplace integral [18]

$$\int_0^{\infty} e^{-\alpha z^2} \cos \beta z\, dz = \frac{1}{2}\sqrt{\frac{\pi}{\alpha}} \exp\left(-\frac{\beta^2}{4\alpha}\right)$$

and get that

$$\phi_X(t) = \exp\left(it\mu - \frac{\sigma^2 t^2}{2}\right).$$

For the standard Gaussian random variable when $\mu = 0$ and $\sigma = 1$, we have that

$$\phi_X(t) = e^{-t^2/2}.$$

THE CLT

The history of the CLT is long. It begins with the results of de Moivre [5] and Laplace [6], who obtained the limit shape of a binomial distribution. It was then followed the work of Lyapunov [19], who invented the method of characteristic functions in probability theory and used it to essentially generalize the de Moivre-Laplace results. Lindeberg [20], Lévy [21], [22], Khintchine [24], and Feller [23] formulated general necessary and sufficient conditions of asymptotic normality.

Basing on a simple sufficient condition in the case of identical distributions, we formulate and prove the Lindeberg-Lévy CLT [20], [21].

Let $X_1, X_2, \ldots, X_n, \ldots$ be independent identically distributed (i.i.d.) random variables with finite mean μ and variance σ^2. Then the distribution function of the centered and standardized random variable

$$Y_n = \frac{1}{\sigma\sqrt{n}} \sum_{k=1}^{n} (X_k - \mu) \tag{14}$$

tends to the Gaussian distribution function

$$\begin{aligned} F_{Y_n}(x) = P\{Y_n \le x\} \quad &\to \quad \Phi(x) \\ = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2}\, dt \quad &\text{as} \quad n \to \infty \end{aligned}$$

for every fixed x.

The proof is based on the asymptotical expansion of the characteristic function of the sum of random variables, so in the sequel, we use some properties of the characteristic functions (Fourier transforms).

Consider the centered and standardized random variables

$$X_k' = \frac{X_k - \mu}{\sigma}, \qquad k = 1, 2, \ldots, n,$$

which are i.i.d.; hence they have the same characteristic function $\phi_{X'}(t)$. Next return to formula (14) $Y_n = \sum_{k=1}^{n}(X_k'/\sqrt{n})$ and write out its characteristic function. Since the characteristic function of the random variable $X_k'/\sqrt{n}$ is given by $\phi_{X'}(t/\sqrt{n})$, the characteristic function for Y_n is the product of $\phi_{X_k'}(t/\sqrt{n})$

$$\phi_{Y_n}(t) = \phi^n_{X'}(t/\sqrt{n}).$$

Now expand $\phi_{X'}(t)$ into the Taylor series about the point $t = 0$ with the remainder in the Peano form

$$\phi_{X'}(t) = \phi_{X'}(0) + \phi'_{X'}(0)\,t + \left[\phi''_{X'}(0)/2 + \alpha(t)\right]\,t^2.$$

The remainder is $\alpha(t)\,t^2$ where $\alpha(t) \to 0$ as $t \to 0$.

Further use the properties of the characteristic functions: $\phi_{X'}(0) = 1$ and $\phi^{(k)}_{X'}(0) = i^k E[(X')^k]$, $k = 1, 2, \ldots, n$. Since $E[X'] = 0$ and $E[X'^2] = 1$, $\phi'_{X'}(0) = 0$ and $\phi\,''X\,'(0) = -1$. Hence,

$$\phi_{X'}(t) = 1 - \frac{t^2}{2} + \alpha(t)\,t^2,$$
$$\phi_{X'}\left(\frac{t}{\sqrt{n}}\right) = 1 - \frac{t^2}{2n} + \alpha\left(\frac{t}{\sqrt{n}}\right)\frac{t^2}{n},$$
$$\phi_{Y_n}(t) = \left[1 - \frac{t^2}{2n} + \alpha\left(\frac{t}{\sqrt{n}}\right)\frac{t^2}{n}\right]^n.$$

Taking the logarithm of the both parts of the last equation and passing to the limit, we get

$$\lim_{n\to\infty} \log\phi_{Y_n}(t) = \lim_{n\to\infty} n\,\log\left[1 - \frac{t^2}{2n} + \alpha\left(\frac{t}{\sqrt{n}}\right)\frac{t^2}{n}\right].$$

Using the relation of equivalency $\log(1 + x) \sim x$ as $x \to 0$, we obtain

$$\lim_{n\to\infty} \log\phi_{Y_n}(t) = \lim_{n\to\infty}\left[n\left(-\frac{t^2}{2n} + \alpha\left(\frac{t}{\sqrt{n}}\right)\frac{t^2}{n}\right)\right]$$
$$= \lim_{n\to\infty}\left(-\frac{t^2}{2} + \alpha\left(\frac{t}{\sqrt{n}}\right)t^2\right) = -\frac{t^2}{2}.$$

Thus,

$$\lim_{n\to\infty} \log\phi_{Y_n}(t) = e^{-t^2/2}.$$

Since the convergence of the characteristic functions to a certain limit implies the convergence of the distribution functions to the corresponding limit [25], the limit law of the random variables Y_n is the standard Gaussian with the parameters $\mu = 0$ and $\sigma = 1$.

If we assume the existence of the third absolute moment of each X_k about its mean $\nu_{3k} = E[|X_k - \mu_k|^3] < \infty$, then the requirement of distributions identity can be dropped. Asymptotic normality, precisely the Lyapunov CLT [19], holds if the Xs have different distributions with finite means μ_k and variances σ_k^2, and if $\lim_{n\to\infty} \nu_3/B_n^3 = 0$, where $\nu_3 = \sum_1^n \nu_{3k}$ and $B_n^2 = \sum_1^n \sigma_k^2$. Then the random variable $Y_n = \sum_1^n (X_k - \mu_k)/B_n$ has the limit distribution $\Phi(x)$.

Asymptotic normality may be established under conditions that do not require the existence of third moments. Actually, it is a necessary and sufficient condition that

$$\lim_{n\to\infty} \frac{1}{B_n^2}\sum_{k=1}^{n}\int_{|x-\mu_k|>\varepsilon B_n}(x - \mu_k)^2 dF_k(x) = 0, \qquad (15)$$

where ε is an arbitrary positive number and F_k is the distribution function of X_k, $k = 1, 2, \ldots, n$.

This condition, due to Lindeberg [20] who proved its sufficiency and Feller [23] who proved its necessity, implies that the total variance B_n^2 tends to infinity and that every σ_k^2/B_n^2 tends to zero, in fact that no random variable dominates the others. The theorem may fail to hold for random variables that do not possess a second moment; for instance, the mean of n variables each distributed according to the Cauchy law

$$dF(x) = \frac{dx}{\pi(1 + x^2)}, \qquad -\infty < x < \infty,$$

is distributed in precisely the same form. This can be easily seen from the characteristic function $\phi(t) = e^{-|t|}$ for the Cauchy distribution [25].

The practical applications of the CLT are based on the corresponding Gaussian approximations of the sums of random variables and their accuracy significantly depends on convergence rate estimates in the CLT.

Return again to the Lindeberg-Lévy formulation of the CLT (14) for the sums of i.i.d. random variables $\{X_k\}_{k=1}^n$. In this case, the classical Berry-Esseen convergence rate estimate in the uniform metric is given by

$$\rho(F_{Y_n}, \Phi) = \sup_x |F_{Y_n}(x) - \Phi(x)| \le C\,\frac{\nu_3}{\sigma^3\sqrt{n}}, \qquad (16)$$

where σ^2 and ν_3 are respectively the variance and the absolute third moment about mean of the parent distribution F_X, and C is an absolute constant (this means that there exists such an F_X for which the upper boundary in inequality (16) is attained) [26], [27]. The latest improvement of the value of the constant C is given by 0.7655 [28].

It's noteworthy that the Berry-Esseen boundary with the convergence rate of $1/\sqrt{n}$ is, although pessimistic, fundamentally related to the Gram-Charlier and Edgeworth series [29]. The practical implications of the aforementioned results on the CLT probability approximations are usually justified by the relative error of probability approximation of the real-life problem, in many cases, to evaluate a certain tail probability.

The classical Lindeberg-Lévy, Lyapunov, and Lindeberg-Feller versions of the CLT state the convergence of the distribution $F_{Y_n}(x)$ of the standardized sums Y_n to the Gaussian distribution $\Phi(x)$. Evidently, the conditions under which the distribution density of Y_n converges to the Gaussian density should be stricter than for the classical versions of the CLT, since the convergence $F_{Y_n}(x) \to \Phi(x)$ does not imply the convergence $F'_{Y_n}(x) \to \Phi\,'(x)$. In the case of continuous i.i.d. random variables X_k, the sufficient condition for the uniform convergence of distribution densities is just the existence of mean and variance [30].

Concluding our remarks on the CLT, we note that the distributions with finite third moments are of a special interest in theory not only because of that Lyapunov's sufficient condition $\nu_3 < \infty$ is evidently simpler to verify than Lindeberg-Feller condition (15), but due to the following results. The finiteness of the moment $\nu_{2+\delta}$, $0 < \delta < 1$, guarantees the decrease rate $n^{-\delta/2}$ for $\rho(F_{Y_n}, \Phi)$ as $n \to \infty$ [19]; for $\delta \geq 1$, $\rho(F_{Y_n}, \Phi) = O(n^{-1/2})$, that is, just the Berry-Esseen convergence rate, and this rate is not improvable [31].

We have mentioned several classical results on the limit theorems in probability theory dealing with the sums of independent random variables. Further extensions and generalizations of the CLT are concerned i) with the study of different schemes of dependency between the summands: homogeneous Markov chains with the finite number of states [32], m-dependent random variables [33], martingales [34]; ii) with the random number of summands, mostly the Poisson and generalized Poisson models being considered [35], [36], and iii) with the properties of various metrics and measures characterizing convergence rates in the CLT [37]. This topic still remains rather popular among mathematicians: in [37], a comprehensive study of the former and recent results in this area is given, focusing on the classical versions of the CLT as well as on the CLT analogs in the classes of non-Gaussian infinitely divisible and stable distribution laws.

MAXIMIZING ENTROPY

Consider the variational problem of maximizing entropy

$$H(f) = -\int_{-\infty}^{\infty} f(x) \log f(x)\, dx$$

in the class of symmetric distributions with a bounded variance

$$\begin{gathered} f^*(x) = \arg\max_{f(x)} H(f), \qquad (17) \\ f(x) \geq 0, \qquad f(-x) = f(x), \\ \int_{-\infty}^{\infty} f(x)\, dx = 1, \quad \sigma^2(f) = \int_{-\infty}^{\infty} x^2 f(x)\, dx \leq \overline{\sigma}^2. \end{gathered}$$

Its solution is given by the Gaussian distribution density $f^*(x) = \mathcal{N}(x; 0, \overline{\sigma})$ [38].

To show this, first note that the entropy of any distribution increases with increasing of its variance, say, for the Gaussian as $\log \sigma$. Thus, it suffices to solve problem (17) under given variance $\sigma^2(f) = d^2$ assuming $d^2 \leq \overline{\sigma}^2$. Second, consider two random variables X and Y with zero mean and variance σ^2 such that $f_X(x)$ is th probability density function (pdf) for X and Y is a Gaussian with pdf f_Y, and use the IT-inequality for entropies

$$\begin{aligned} H(f_Y) &= -\int f_Y(y) \log f_Y(y)\, dy \\ &= \int f_Y(y) \left[\log(\sigma\sqrt{2\pi}) + \frac{y^2}{2\sigma^2} \right] dy \\ &= \log(\sigma\sqrt{2\pi}) + \frac{1}{2\sigma^2} E[Y^2] \\ &= \log(\sigma\sqrt{2\pi}) + \frac{1}{2\sigma^2} E[X^2] \\ &= -\int f_X(y) \log f_Y(x)\, dx \geq \\ &\quad -\int f_X(y) \log f_X(x)\, dx = H(f_X). \end{aligned}$$

So, we arrive at the inequality $H(f_Y) \geq H(f_X)$ with equality if and only if $f_X = f_Y$. In other words, the Gaussian distribution has higher entropy than any other with the same variance: thus, any operation on a distribution, which discards information and preserves variance bounded, leads us to a Gaussian. The best example of this is given by the CLT as, evidently, the summation discards information and the appropriate standardizing even conserves variance.

MINIMIZING FISHER INFORMATION

The notion of the Fisher information arises in the Cramér-Rao inequality [29], one of the principal results of the mathematical statistics, which gives the lower boundary upon an parameter estimator's variance

$$\operatorname{Var} \widehat{\theta}_n \geq \frac{1}{n I(f)}, \qquad (18)$$

where $\widehat{\theta}_n$ is an unbiased estimator of a parameter θ of the distribution density $f(x, \theta)$ from a sample $x_1, \ldots, x_n$ and $I(f)$ is the functional of the Fisher information given by

$$I(f) = \int_{-\infty}^{\infty} \left(\frac{\partial \log f(x, \theta)}{\partial \theta} \right)^2 f(x, \theta)\, dx. \qquad (19)$$

In the case of estimation of a location parameter, say, the mean, when the distribution density depends on θ as $f(x - \theta)$, it can be easily seen that formula (19) takes the following form

$$I(f) = \int_{-\infty}^{\infty} \left(\frac{f(x)}{f(x)} \right)^2 f(x)\, dx. \qquad (20)$$

Now we show that the solution to the variational problem of minimization of Fisher information for location (20) for the distributions with a bounded variance is achieved at the Gaussian [39], precisely that

$$\mathcal{N}(x; 0, \overline{\sigma}) = f^*(x) = \arg\min_{f(x)} I(f)$$

subject to

$$f(x) \geq 0, \qquad f(-x) = f(x),$$

$$\int_{-\infty}^{\infty} f(x)\,dx = 1, \qquad \sigma^2(f) = \int_{-\infty}^{\infty} x^2 f(x)\,dx \leq \overline{\sigma}^2.$$

Similar to the aforementioned derivation for entropy, it suffices to consider the case of a given variance $\sigma^2(f) = d^2 \leq \overline{\sigma}^2$. Next we use the following version of the Cauchy-Bunyakovskiy inequality:

$$\left(\int \phi(x)\psi(x)f(x)dx\right)^2 \leq \int \phi^2(x)f(x)dx \int \psi^2(x)f(x)\,dx, \tag{21}$$

THE ARGUMENTS PRO GAUSSIANITY CAN BE CLASSIFIED IN THE FOLLOWING TWO GROUPS: THE ARGUMENTS FOR THE GRAVITY AND STABILITY OF A GAUSSIAN SHAPE AND THE ARGUMENTS FOR THE OPTIMALITY OF A GAUSSIAN SHAPE.

where the functions $\phi(x)$ and $\psi(x)$ should only provide the existence of the integrals in (21) and remain arbitrary in all other aspects.

Now choose $\phi(x) = x$ and $\psi(x) = -f'(x)/f(x)$. The integrals in the right-hand part of (21) are the distribution variance $\sigma^2(f) = d^2$ and the Fisher information (20), respectively. Using symmetry and integrating by parts, we compute the integral in the left-hand part of (21)

$$\begin{aligned} -\int_{-\infty}^{\infty} xf'(x)\,dx &= -2\int_{0}^{\infty} xf'(x)\,dx \\ &= -2\left[xf(x)|_0^{\infty} - \int_0^{\infty} f(x)\,dx\right] \\ &= 1, \end{aligned}$$

assuming that the distribution tails satisfy $\lim_{x\to\infty} xf(x) = 0$. Collecting the obtained results and substituting them into (21), we get the lower boundary upon Fisher information

$$I(f) \geq \frac{1}{d^2}.$$

As this lower boundary just gives the Fisher information value for the Gaussian distribution density

$$\int_{-\infty}^{\infty}\left(\frac{\mathcal{N}'(x;0,d)}{\mathcal{N}(x;0,d)}\right)^2 \mathcal{N}(x;0,d)\,dx = \frac{1}{d^2}$$

and the minimization problem in the class of distributions with a bounded variance allows for the following two-step decomposition:

$$f^* = \arg\min_{f:\sigma^2(f)\leq\overline{\sigma}^2} I(f) = \arg\min_{d^2\leq\overline{\sigma}^2}\left\{\min_{f:\sigma^2(f)=d^2} I(f)\right\},$$

we arrive at the required relation $f^*(x) = \mathcal{N}(x;0,\overline{\sigma})$.

This important result that the Gaussian distribution is the least favorable distribution in the class of distributions with a bounded variance gives another reason for the ubiquitous use of the Gaussian distribution in signal processing and, moreover, links Huber's results in robustness.

ROBUSTNESS VERSUS GAUSSIANITY

In this section, we show that the Gaussian distribution being least favorable and, therefore, the LS method being robust in Huber's sense naturally arise in robustness, despite the conventional emphasis on the departures from Gaussianity.

The field of mathematical statistics called robust statistics appeared due to the pioneer works of Tukey [40], Huber [41], and Hampel [42], respectively; it has been intensively developed since 1960 and is rather definitely formed by present. The term "robust" (strong, sturdy, rough) as applied to statistical procedures was proposed by Box [43].

Robustness deals with the consequences of possible deviations from the assumed statistical model and suggests the methods providing stability of statistical procedures against such deviations.

Using the model of ε-contaminated normal distributions, Tukey [40] showed that the LS estimators are not stable under small deviations from Gaussianity, furthermore, that the LS estimators are catastrophically bad in the presence of outliers in the Gaussian data. The simplest way to see this is to consider the Cauchy distribution contamination of the Gaussian underlying distribution

$$f(x;\theta) = (1-\varepsilon)\mathcal{N}(x;\theta,\sigma) + \varepsilon C(x;\theta), \qquad 0 \leq \varepsilon < 1,$$
$$C(x;\theta) = \frac{1}{\pi[1+(x-\theta)^2]}.$$

It is easy to see that for any $\varepsilon > 0$, the sample mean, the optimal LS estimator of location for the Gaussian distribution, is not even consistent in this case. Since such negligible deviations from Gaussianity in the tail area cannot be detected by any statistical procedure, it seems that the aforementioned phenomenon seriously undermines the belief in the ubiquitous applicability of Gaussian models.

Next, we are going to show that, nevertheless, Gaussian models also successfully work in robust procedures within Huber's minimax approach [41], [44].

HUBER'S MINIMAX APPROACH

We now briefly recall the basic stages of Huber's minimax approach to robust estimation of location. In general, the minimax principle aims at the least favorable situation for which it suggests the best solution. Thus, in some sense, this approach provides a guaranteed result, possibly too pessimistic. Huber's minimax approach in robustness represents a good example of application of the minimax principle.

Let $x_1, \ldots , x_n$ be i.i.d. random variables with common density $f(x-\theta)$ in a convex class $\mathcal{F}$. Then the M-estimator $\widehat{\theta}_n$ of a location parameter θ is defined as

$$\widehat{\theta}_n = \arg \min_{\theta} \sum_{i=1}^{n} \rho(x_i - \theta)$$

or

$$\sum_{i=1}^{n} \psi(x_i - \widehat{\theta}_n) = 0,$$

where $\rho(x)$ is a loss function and $\psi(x) = \rho\,'(x)$ is a score function [41].

The minimax approach implies the determination of the least favorable density f^* minimizing Fisher information $I(f) = \int (f\,'/f)^2 f\,dx$ over the class $\mathcal{F}$:

$$f^* = \arg\min_{f\in\mathcal{F}} I(f), \qquad (22)$$

followed by designing the maximum-likelihood estimator with the loss function $\rho^* = -\log f^*$ and the score function $\psi^* = -f^{*\,\prime}/f^*$. The necessary and sufficient condition for f^* to minimize Fisher information $I(f)$ is given by the condition

$$\int \left(2\psi^{*} - \psi^{*2}\right)(f - f^*)\,dx \ge 0 \qquad (23)$$

that must hold for any density $f \in \mathcal{F}$.The required convexity of class $\mathcal{F}$ guarantees that the variations of the optimal density f^* retain densities f in this class [41], [44].

Under rather general conditions of regularity [41], $\sqrt{n}\,(\widehat{\theta}_n - \theta)$ is asymptotically normal with variance

$$V(\psi, f) = \frac{\int \psi^2 f\,dx}{\left[\int \psi' f\,dx\right]^2}$$

satisfying the minimax property

$$V(\psi^*, f) \le V(\psi^*, f^*) \le V(\psi, f^*).$$

The both sides of this saddle-point inequality have sense: the right-hand side is just the Cramér-Rao inequality (18) whereas the left-hand side provides the guaranteed accuracy of estimation

$$\mathrm{Var}\,\widehat{\theta}_n = \frac{V(\psi^*, f)}{n} \le \frac{V(\psi^*, f^*)}{n} = \frac{1}{n\,I(f^*)} \quad \text{for all } f \in \mathcal{F}.$$

Concluding, we may say that Huber proposed to use the supremum of the asymptotic variance $V(\psi^*, f^*) = \sup_{f\in\mathcal{F}} V(\psi^* f)$ as a measure of robustness of the optimal M-estimator: the less the range of the optimal estimator variance $V(\psi^*, f)$ over the class $\mathcal{F}$, the more robust is this estimator in this class, and vice versa.

LEAST FAVORABLE DISTRIBUTIONS

The shape of the least favorable density f^* and the corresponding score function ψ^* is wholly determined by the structure of class $\mathcal{F}$. We now describe how to obtain a least favorable distribution and enlist several examples. The symmetry and unimodality of distribution densities are assumed.

Consider the restrictions defining the classes of distribution densities $\mathcal{F}$. In general, these restrictions are of the following forms:

$$\int_{-\infty}^{\infty} s_k(x) f(x)\,dx \le \alpha_k, \quad k = 1, \ldots , m, \qquad (24)$$

$$f(x) \ge \varphi(x), \qquad (25)$$

where α_k, $k = 1, \ldots , m$, and $\varphi(x)$ are given constraints.

In particular, the normalization condition $\int f(x)\,dx = 1$ $(s(x) = 1)$ and the restriction on the variance $\int x^2 f(x)\,dx \le \overline{\sigma}^2$ $(s(x) = x^2)$ are referred to (24); the condition of non-negativeness $f(x) \ge 0$ is described by (25), etc.

The variational problem of minimization of Fisher information under conditions (24) and (25) is nonstandard, and at present, there are no general methods of its solution.

Nevertheless, using heuristic and plausible considerations (in the Polya sense [45]), it is possible to find a candidate for the optimal solution and then check its validity. Certainly, such a reasoning must ground on the classical results of the calculus of variations. In general, it may be described as follows: first, use the restrictions of form (24); second, solve the Euler-Lagrange equation and determine the family of extremals; third, try to satisfy the restrictions of form (25) by gluing the pieces of free extremals with the constraints $\varphi(x)$; and finally, verify the obtained solution checking condition (23).

Now we describe a procedure of searching for a candidate for the solution of problem (22) under conditions (24). In this case, the Lagrange functional is composed as

$$L(f, \lambda_1, \ldots , \lambda_m) = I(f) + \sum_{k=1}^{m} \lambda_k \left(\int_{-\infty}^{\infty} s_k(x) f(x)\,dx - \alpha_k \right),$$

and by (20) it can be rewritten as

$$L(f, \lambda_1, \ldots , \lambda_m) = \int_{-\infty}^{\infty} \left[\frac{\left(f\,'(x)\right)^2}{f(x)} + \sum_{k=1}^{m} \lambda_k s_k(x)\, f(x) \right] dx - \sum_{k=1}^{m} \lambda_k \alpha_k, \qquad (26)$$

where $\lambda_1, \ldots , \lambda_m$ are the Lagrange multipliers. Denoting by $G(x, f(x), f\,'(x))$ the integrand, we get

$$L(f, \lambda_1, \ldots , \lambda_m) = \int_{-\infty}^{\infty} G\left(x, f(x), f\,'(x)\right) dx - \sum_{k=1}^{m} \lambda_k \alpha_k.$$

Noting the necessary condition of minimum of $L(f, \lambda_1, \ldots , \lambda_m)$, namely the Euler-Lagrange equation

$$\frac{d}{dx}\frac{\partial G}{\partial f'} - \frac{\partial G}{\partial f} = 0,$$

we obtain

$$2\left(\frac{f'(x)}{f(x)}\right)' + \left(\frac{f'(x)}{f(x)}\right)^2 - \sum_{k=1}^{m}\lambda_k s_k(x) = 0. \tag{27}$$

Equation (27), as a rule, cannot be solved in a closed form. Hence, one should use numerical methods (for details, see [46] and [47]). In what follows, we consider some classes $\mathcal{F}$ with analytical solutions for the least favorable density.

THE EXPONENTIAL EXTREMALS OF THE BASIC VARIATIONAL PROBLEM

Consider the problem of minimization of Fisher information with the only side normalization condition

$$\text{minimize} \quad I(f) = \int_{-\infty}^{\infty}\left(\frac{f'(x)}{f(x)}\right)^2 f(x)\,dx$$
$$\text{subject to} \quad \int_{-\infty}^{\infty} f(x)\,dx = 1.$$

Then, from (27), it follows that the Euler-Lagrange equation has the form

$$2\left(\frac{f(x)}{f(x)}\right)' + \left(\frac{f'(x)}{f(x)}\right)^2 - \lambda = 0. \tag{28}$$

Changing the variable $f(x) = g^2(x) \geq 0$, we can rewrite (28) as follows

$$4g''(x) - \lambda g(x) = 0. \tag{29}$$

For the positive λ, the system of the fundamental solutions for (29) is given by

$$g_1(x) = e^{-kx}, \quad g_2(x) = e^{kx}$$

with the corresponding exponential extremals

$$f_1(x) = e^{-2kx}, \quad f_2(x) = e^{2kx}, \tag{30}$$

where $k = \sqrt{\lambda}/2$.

CONTAMINATED GAUSSIAN DISTRIBUTIONS

Though it is not the simplest example of a least favorable distribution, we begin with historically the first Huber's solution [41] for the class of ε-contaminated Gaussian distributions

$$\mathcal{F}_H = \left\{f\colon f(x) = (1-\varepsilon)\mathcal{N}(x;\ 0,\sigma) + \varepsilon h(x)\right\},$$

where $h(x)$ is an arbitrary density and ε $(0 \leq \varepsilon < 1)$ is a contamination parameter.

Using condition (25) with $\varphi(x) = (1-\varepsilon)\mathcal{N}(x;\ 0,\sigma)$ for defining this class

$$\mathcal{F}_H = \left\{f\colon f(x) \geq (1-\varepsilon)\mathcal{N}(x;\ 0,\sigma)\right\},$$

we can foresee the qualitative structure of the least favorable density: there should be the exponential extremals of form (30) smoothly sewed with the constraint $\varphi(x) = (1-\varepsilon)\mathcal{N}(x;\ 0,\sigma)$. Its exact form is given by

$$f_H^*(x) = \begin{cases} (1-\varepsilon)\mathcal{N}(x;0,\sigma), & \text{for } |x| \leq k\sigma, \\ \frac{1-\varepsilon}{\sqrt{2\pi}\sigma}\exp\left(-k\sigma|x| + \frac{k^2\sigma^2}{2}\right), & \text{for } |x| > k\sigma, \end{cases}$$

where the dependence $k = k(\varepsilon)$ is tabulated in [44]. The optimality of f_H^* is established by checking inequality (23): here it is just $f(x) \geq (1-\varepsilon)\mathcal{N}(x;\ 0,\sigma)$ taking the form of the characterization condition of class $\mathcal{F}_H$. The optimal score function has the following limited linear form

$$\psi_H^*(x) = \begin{cases} x/\sigma^2, & \text{for} \quad |x| \leq k\sigma, \\ k\ \mathrm{sgn}(x)/\sigma, & \text{for} \quad |x| > k\sigma \end{cases}$$

with the Winsorized mean as the minimax M-estimator of location. The qualitatively similar solution also holds for the class of approximately Gaussian distributions in which the ε-neighborhood of a Gaussian distribution is defined by the Kolmogorov distance as $\sup_x|F(x) - \Phi(x)| \leq \varepsilon$ [41]. These both results exhibit the direct way how Gaussian models can be used in robust settings.

NONDEGENERATE DISTRIBUTIONS

In the class $\mathcal{F}_1$ of nondegenerate distribution densities (with a bounded density value at the center of symmetry)

$$\mathcal{F}_1 = \left\{f\colon\ f(0) \geq \frac{1}{2a} > 0\right\},$$

the least favorable density is known to be the Laplace [49], [50]

$$f_1{}^*(x) = \mathcal{L}(x;0,a) = \frac{1}{2a}\exp\left(-\frac{|x|}{a}\right),$$

here the scale parameter a characterizes the distribution dispersion about the center of symmetry. In this case, we also observe the two exponential extremals of form (30) sewed at the center of symmetry and satisfying the constraint $f(0) = 1/(2a)$ of class $\mathcal{F}_1$.

The score function is of the sign form $\psi_1^*(x) = \mathrm{sgn}(x)/a$ with the conventional robust estimator of location, the sample median as the optimal L_1-norm estimator [44]. Class $\mathcal{F}_1$ is one of the most wide classes: any unimodal distribution density with a nonzero value at the center of symmetry belongs to it. The condition of belonging to this class is very close to the complete lack of information about an underlying distribution.

DISTRIBUTIONS WITH A BOUNDED VARIANCE

As it was shown before, in the class $\mathcal{F}_2$ of distributions with a bounded variance

$$\mathcal{F}_2 = \left\{ f\colon \sigma^2(f) = \int_{-\infty}^{\infty} x^2 f(x)\, dx \le \overline{\sigma}^2 \right\},$$

the least favorable density is the Gaussian

$$f_2{}^*(x) = \mathcal{N}(x; 0, \overline{\sigma}) = \frac{1}{\overline{\sigma}\sqrt{2\pi}} \exp\left(-\frac{x^2}{2\overline{\sigma}^2}\right).$$

The optimal score function is linear $\psi_2^*(x) = x/\overline{\sigma}^2$ and the minimax estimator of location is the sample mean $\overline{x}_n$.

Taking into account the role of a least favorable distribution in Huber's minimax approach, we have arrived to a rather strange result: the sample mean $\overline{x}_n$ is robust in the Huber sense in the class of distributions with a bounded variance!

ALL THE ARGUMENTS CONTRA GAUSSIANITY ARISE WHEN THE AFOREMENTIONED CONDITIONS OF SMOOTHNESS ARE VIOLATED.

Let us dwell on this phenomenon in more detail. Since the Fisher information for the least favorable Gaussian distribution attains its minimum value at $I(f_2^*) = 1/\overline{\sigma}^2$, the sample mean is an estimator of guaranteed accuracy in $\mathcal{F}_2$, that is

$$\mathrm{Var}\, \overline{x}_n \le \overline{\sigma}^2/n \quad \text{for} \quad \text{all} \quad f \in \mathcal{F}_2.$$

Thus, if the bound on variance $\overline{\sigma}^2$ is small, then the minimax approach yields a reasonable result and the LS method can be successfully used with relatively short-tailed distributions in estimation and detection of signals, e.g., see [48].

On the contrary, if we deal with really heavy-tailed distributions (gross errors, impulse noise) when $\overline{\sigma}^2$ is large or even infinity like for the Cauchy-type distributions, then the minimax solution in class $\mathcal{F}_2$ is still trivially correct as $\mathrm{Var}\widehat{\theta}_n \le \infty$ but practically senseless. In this case, we must use robust versions of the LS method such as Huber's M-estimators optimal for the class of ε-contaminated Gaussian distributions.

We also may say that the minimax principle gives an unrealistic result in this case. However, this disadvantage becomes a significant advantage of the LS estimator if to consider the class of nondegenerate distributions with a bounded variance, in other words, the intersection of the classes $\mathcal{F}_1$ and $\mathcal{F}_2$

$$\mathcal{F}_{12} = \left\{ f\colon \quad f(0) \ge \frac{1}{2a} > 0, \quad \sigma^2(f) \le \overline{\sigma}^2 \right\}.$$

This class comprises qualitatively different densities, for example, the Gaussian, the heavy-tailed ε-contaminated Gaussian, Laplace, Cauchy-type (with $\overline{\sigma}^2 = \infty$), and short-tailed densities.

For this class, the least favorable density simultaneously depends on the two parameters a and $\overline{\sigma}$ through their ratio $\overline{\sigma}/a$ having the Gaussian and Laplace densities as the particular cases (for details, see [47]).

In this case, the corresponding minimax estimator of location can be described as follows: 1) with relatively small variances when $\overline{\sigma}^2/a^2 < 2/\pi$ or with relatively short tails, it is the sample mean or the L_2-norm estimator; 2) with relatively large variances when $\overline{\sigma}^2/a^2 > 2$ or with relatively heavy tails, it is the sample median or the L_1-norm estimator; 3) and with relatively moderate variances when $2/\pi \le \overline{\sigma}^2/a^2 \le 2$, it is a compromise between the L_1-norm and the L_2-norm estimators. This solution is robust and close to Huber's solution for the class $\mathcal{F}_H$ of heavy-tailed distributions due to the presence of the Laplace branch and more efficient than Huber's for short-tailed distributions due to the presence of the Gaussian branch [47], [48]. In other words, the additional information on the relative weight of distribution tails given by the ratio $\overline{\sigma}^2/a^2$ may significantly improve the quality of estimation and detection.

CONCLUSIONS

Now we return to the initial question posed at the beginning: "Why the ubiquitous use and success of Gaussian distributions?"

All the arguments pro Gaussianity can be classified in the following two groups: 1) the arguments for the gravity and stability of a Gaussian shape: the statistical gravity (the CLT, the Landon derivation), the stability (the convolution property), and geometrical invariancy (the Herschel-Maxwell derivation) and 2) the arguments for the optimality of a Gaussian shape (the Gauss derivation, the maximization of entropy, and minimization of Fisher information). In this list, we skipped various characterization properties of a multivariate Gaussian, especially of a bivariate one, represented in [39] and the stability aspects related to the Gaussian infinite divisibility analyzed in [37]; some additional reasons pro Gaussianity can be found in [2] and [29].

On the whole, we may repeat after Jaynes that, "in Nature, all smooth processes with increasing entropy lead to Gaussianity and once it is reached, it is then preserved" [2]. The fact that a Gaussian is the least favorable distribution minimizing Fisher information is significantly important in signal and data processing.

All the arguments contra Gaussianity arise when the aforementioned conditions of smoothness are violated: this refers to the presence of gross errors and outliers in the data, impulse noises in observed signals, etc. Moreover, we may add that most of the formulated properties of a Gaussian, say, the CLT, are of an asymptotic nature, so on finite samples, they hold only approximately. For instance, we never know the tails of distributions in real-life data. On the whole, these reasons lead to robust methods and algorithms of signal processing and what is important that a Gaussian again naturally emerges in robustness either in the form of various Gaussian ε-neighborhoods or as the least favorable distribution.

ACKNOWLEDGMENT

We are grateful to the reviewers whose comments helped much to improve the contents of this article.

AUTHORS

Kiseon Kim (kskim@gist.ac.kr) received the B.Eng. and M.Eng. degrees, in electronics engineering, from Seoul National University, Korea, in 1978 and 1980, and the Ph.D. degree in electrical engineering-systems from University of Southern California, Los Angeles, in 1987. From 1988 to 1991, he was with Schlumberger, Houston, Texas. From 1991 to 1994, he was with the Superconducting Super Collider Lab, Texas. He joined Gwangju Institute of Science and Technology (GIST), Korea, in 1994, where he is currently a professor. His current interests include wideband digital communications system design, sensor network design, analysis and implementation both at the physical layer and at the resource management layer.

Georgy Shevlyakov (shev@gist.ac.kr) received the M.S. degree in control theory and the Ph. D. degree in cybernetics from Leningrad Polytechnic Institute, U.S.S.R., in 1973 and 1976, respectively, and the Dr. Sc. degree in applied statistics from St. Petersburg Polytechnic University, St. Petersburg, U.S.S.R., in 1991. From 1976 to 1979, he was with Vavilov Research Institute, Leningrad. From 1979 to 1986, he was with the Department of Mechanics and Control Processes, Leningrad Polytechnic Institute. From 1986 to 1992, he was with Department of Mathematics, St. Petersburg Polytechnic University. He is currently a visiting professor at the Department of Information and Communications, Gwangju Institute of Science and Technology (GIST), Korea. His research interests include mathematical methods of robust statistics and data analysis with their applications to signal processing.

REFERENCES

[1] H. Poincaré, *Science et Hypothesis*, 1904; English translation. New York: Dover, 1952.

[2] E.T. Jaynes, *Probability Theory. The Logic of Science*. Cambridge, U.K.: Cambridge Univ. Press, 2003.

[3] L. Cohen, "The history of noise," *IEEE Signal Process. Mag*., vol. 22, no. 6, pp. 20–45, Nov. 2005.

[4] W. Feller, *An Introduction to Probability Theory and its Applications*, vol. 1. New York: Wiley, 1950.

[5] Photographic reproduction in Archibald, R.C., "A rare pamphlet of Moivre and some of his discoveries," Isis 8, pp. 671–683, 1926.

[6] P.S. Laplace, "Mémoire sur les probabilités*," Mem. Acad. Roy*., 1781, Paris, France; reprinted in Laplace (1878–1912), vol. 9, pp. 384–485.

[7] K.F. Gauss, *Theoria Motus Corporum Celestium, Perthes, Hamburg, 1809; English translation, Theory of the Motion of the Heavenly Bodies Moving About the Sun in Conic Sections*. New York: Dover, 1963.

[8] L.W. Boltzmann, "Über das Wärmegleichgewicht zwischen mehratomigen Gasmolekülen," *Wiener Berichte*, 1871, vol. 63, pp. 397–418, 679–711, 712–732.

[9] J.C. Maxwell, "Illustration of the dynamical theory of gases. Part I. On the motion and collision of perfectly elastic spheres," *Phil. Mag.*, vol. 56, 1860.

[10] K.F. Gauss, *Theoria Combinationis Observationum Erroribus Minimis Obnoxiae*. Göttingen, Germany, 1823; Suppl., 1826.

[11] G. Pólya, *Collected Papers*, 4 vols, G-C. Rota, Ed. Cambridge, MA: MIT Press, 1984.

[12] S.M. Stigler, "Stigler's law of eponymy," *Trans. NY Acad. Sci.*, vol. 39, series 2, pp. 147–159, 1980.

[13] R.A. Fisher, "On the mathematical foundations of theoretical statistics," *Phil. Trans. Roy. Soc.*, A, vol. 222, 1921; reproduced in R.A. Fisher, *Contributions to Mathematical Statistics*. New York: Wiley, 1950.

[14] J. Aczel, *Functional Equations: History, Applications and Theory*. Norwell, MA: Kluwer, 2002.

[15] A.M. Legendre, *Nouvelles méthods pour la détermination des orbits des cométes.* Paris, France: Didot, 1806.

[16] J. Herschel, "Quetelet on probabilities," *Edinburgh Rev.*, vol. 92, no. 14, 1850.

[17] V.D. Landon, "The distribution of amplitude with time in fluctuation noise," *Proc. IRE*, vol. 29, no. 1, pp. 50–54, 1941.

[18] M. Abramowitz and I. Stegun, *Handbook of Mathematical Functions*. New York: Dover, 1972.

[19] A. Liapounoff, "Nouvelle forme du théoréme sur la limite de probabilité," *Mém. Acad. Sci. St. Pétersbourg*, vol. 12, no. 5, pp. 1-24, 1901.

[20] J.W. Lindeberg, "Eine neue Herleitung des Exponentialgesetzes in der Wahrscheinlichkeitsrechnung," *Math. Zeitschr.*, vol. 15, pp. 211–225, 1922.

[21] P. Lévy, *Calcul des probabilités*, Paris, France, 1925.

[22] P. Lévy, *Théorie de l'addition des variables aléatoires*, Paris, France, 1935.

[23] W. Feller, "Über den zentralen Grenzwertsatz der Wahrscheinlichkeitsrechnung," *Math. Zeitschr.*, vol. 40, pp. 521-559, 1936.

[24] A. Khintchine, "Sul dominio di attrazione della legge di Gauss," *Giorn. Ist. Italiano d. Attuari*, vol. 6, pp. 378-393, 1935.

[25] W. Feller, *An Introduction to Probability Theory and its Applications*, vol. 2. New York: Wiley, 1971.

[26] A.C. Berry, "The accuracy of the Gaussian approximation to the sum of independent variables," *Trans. Amer. Math. Soc.*, vol. 49, no. 1, pp. 122–126, 1941.

[27] C.G. Esseen, "On the Lyapunov limit error in the theory of probability," *Ark. Mat. Astr. Fys.*, vol. 28A, no. 9, pp. 1–19, 1942.

[28] I.S. Shiganov, "Refinement of the upper bound on the constant in the central limit theorem," *J. Soviet Math.*, vol. 35, pp. 2545–2551, 1986.

[29] H. Cramér, *Mathematical Methods of Statistics*. Princeton, NJ: Princeton Univ. Press, 1974.

[30] B.V. Gnedenko and A.N. Kolmogorov, *Limit Distributions for Sums of Independent Random Variables*. Reading, MA: Addison–Wesley, 1954.

[31] I.A. Ibragimov, "On the accuracy of approximation to the distribution functions of the sums of independent variables by normal distribution," *Theory Probab. Appl.*, vol. 11, no. 4, pp. 632–655, 1966.

[32] A. Rényi, "On the theory of order statistics," *Acta Math. Hung.*,vol. 4, pp. 191–231, 1953.

[33] J. Davidson, *Stochastic Limit Theory—An Introduction for Econometricians*. Oxford, U.K.: Oxford Univ. Press, 1994.

[34] P. Hall and C.C. Heyde, *Martingale Limit Theory and Its Applications*. New York: Academic, 1980.

[35] B.V. Gnedenko and V.Yu. Korolev, *Random Summation: Limit Theorems and Applications*. Boca Raton, FL: CRC Press, 1996.

[36] V.E. Bening and V.Yu. Korolev, *Generalized Poisson Models*. Utrecht, The Neterlands: VSP, 2002.

[37] V.M. Zolotarev, *Modern Theory of Summation of Random Variables*. Utrecht, The Neterlands: VSP, 1997.

[38] C.E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, pp. 329–423, 623–656, 1948.

[39] A.M. Kagan, Yu.V. Linnik, and S.R. Rao, *Characterization Problems in Mathematical Statistics*. New York: Wiley, 1973.

[40] J.W. Tukey, "A survey of sampling from contaminated distributions," in *Contributions to Probability and Statistics,* I. Olkin, Ed. Stanford, CA: Stanford Univ. Press, 1960, pp. 448–485.

[41] P.J. Huber, "Robust estimation of a location parameter," *Ann. Math. Statist.*, vol. 36, no. 1, pp. 1–72, 1964.

[42] F.R. Hampel, *Contributions to the theory of robust estimation*, Ph.D. dissertation, Univ. California, Berkeley, 1968.

[43] G.E.P. Box, "Non-normality and test on variances," *Biometrika*, vol. 40, no. 3, pp. 318–335, 1953.

[44] P.J. Huber, *Robust Statistics*. New York: Wiley, 1981.

[45] G. Polya, *How to Solve It*, Princeton University Press, 1957.

[46] Ya.Z. Tsypkin, *The Informational Identification Theory*. Moscow, State Publishing House of Sciences, 1995 (in Russian).

[47] G.L. Shevlyakov and N.O. Vilchevski, *Robustness in Data Analysis: Criteria and Methods*. Utrecht: VSP, 2002.

[48] G.L. Shevlyakov and K. Kim, "Robust minimax detection of a weak signal in noise with a bounded variance and density value at the center of symmetry," *IEEE Trans. Inf. Theory*, vol. 52, no. 3, pp. 1206–1211, Mar. 2006.

[49] B.T. Polyak and Ya.Z. Tsypkin, "Robust identification," in *Identif. Syst. Parameter Estim., Part 1, Proc. 4th IFAC Symp.*, Tbilisi, 1976, pp. 203–224.

[50] H. Delic, P. Papantoni-Kazakos, and D. Kazakos, "Fundamental structures and asymptotic performance criteria in decentralized binary hypothesis testing," *IEEE Trans. Commun.*, vol. 43, no. 1, pp. 32–43, Jan. 1995.