Non-Gaussianity as a data analysis artifact

Edoardo Milotti*

Dipartimento di Fisica, Università di Trieste, and I. N. F. N.-Sezione di Trieste, Via Valerio 2, I-34127 Trieste, Italy

(Received 29 December 2010; revised manuscript received 22 March 2011; published 27 April 2011)

Non-Gaussian effects are important features in many fields of physics, and the search for non-Gaussianity motivates several new experiments. Here I show that an insidious form of non-Gaussianity can easily arise as a finite-size effect in a data analysis tool that is guaranteed to be asymptotically Gaussian. This means that experimental searches for non-Gaussianity should also include an extremely careful scrutiny of the statistical tools used to analyze data.

DOI: 10.1103/PhysRevE.83.042103 PACS number(s): 02.50.Tt, 02.60.Ed, 07.05.Kf

Non-Gaussian fluctuations arise in several fields of physics such as, for instance, in cosmology because of primordial inflation [1,2], in glassy materials [3], in nonequilibrium thermodynamics [4], in the interplanetary magnetic field [5,6], and as a consequence of nonextensive phenomena in thermodynamics [7]. The non-Gaussianity usually means that distributions that are nearly Gaussian close to their mean value display long tails with a power-law behavior, which is associated to rare events. These rare events may actually spoil the statistics of gravitational wave detectors, and at least one experiment is actively studying them [8].

Its frequent association with the tails of distributions makes non-Gaussianity a very delicate effect, which requires a careful handling of data. Consider a common data-handling tool, the least squares method (LS): in this method, we minimize the chi-square

$$\chi^2 = (\mathbf{y} - A\boldsymbol{\theta})^T V^{-1} (\mathbf{y} - A\boldsymbol{\theta}), \tag{1}$$

where V is the covariance matrix of the measured data $\mathbf{y}$, and where we assume that the mean value $\boldsymbol{\mu}$ of each measurement $\mathbf{y}$ is a function of the independent input variable $\mathbf{x}$:

$$\mu_i(\mathbf{x}_i,\theta) = \sum_{j=1}^{m} a_j(\mathbf{x}_i)\theta_j = \sum_{j=1}^{m} A_{ij}\theta_j, \tag{2}$$

in which $A_{ij} = a_j(\mathbf{x}_i)$, and the a's are arbitrary functions of $\mathbf{x}$ (see, e.g., [9] or [10]). Minimizing the χ^2 in Eq. (1) to obtain the parameters $\boldsymbol{\theta}$ of model (2), we find the parameter estimate

$$\hat{\boldsymbol{\theta}} = (A^T V^{-1} A)^{-1} A^T V^{-1} \mathbf{y} = B\mathbf{y}. \tag{3}$$

Thus, the estimated parameter values are linear combinations of the data values y_i's, and if these data have a Gaussian distribution, the parameter estimates are Gaussian as well. Unfortunately, this is not always the case, and data often have different distributions. In most cases, we can still rely on the central limit theorem, which guarantees that eventually the fit parameters are normally distributed for a large number n of data, although the question remains as to how large a data set should be before we can safely assume a Gaussian distribution for the parameter estimates. A partial answer comes from the Berry-Esseen theorem and its variants [11], which guarantee that, under rather mild conditions, the difference between the actual cumulative distribution function of a linear combination of Gaussian variates and a true Gaussian cumulative distribution function is less than $C'/\sqrt{n}$, where C' is a constant; more precisely, in its basic form, the Berry-Esseen theorem states that, in the case of n independent and identically distributed (i.i.d.) variates ξ_k with mean μ and variance σ^2,

$$\sup_x \left| P(S_n^* < x) - \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-y^2/2} dy \right| \leqslant \frac{C_{\mathrm{BE}}\alpha}{\sqrt{n}\sigma^3}, \tag{4}$$

where $S_n^* = \sum_{k=1}^{n} (\xi_k - \mu)/\sigma$, $\alpha = \mathbf{E}|\xi_k - \mu|^3$, and C_{BE} is a constant ($C_{\mathrm{BE}} < 0.7915$) [11]. The variants of this theorem extend its validity to random variables with different variances, etc.; however, there are no results on the specific shape of the tails of the distributions of linear combinations of variates, so we turn now to Monte Carlo simulation. To be more specific, we study the LS parameter estimates with data distributed according to a gamma distribution. The rationale of this choice is that the gamma distribution emerges naturally in sums of exponentially distributed variates (see, e.g., [10]), it turns up in estimates of power spectra [12], it is the general case of the χ^2 distribution, it can be related to the Rayleigh and the Weibull distributions [13], and, more generally, it often appears wherever there are physical variables that are bounded from below. We take a linear LS fit $S_k = af_k + b$ of the spectral estimate S_k at frequency f_k, obtained from a discrete Fourier transform (DFT) analysis of a Gaussian white noise. Each value of the DFT estimate has a known distribution function; if there is no spectral averaging, then the shape parameter of the gamma distribution is equal to 1, and we find an exponential probability density function (PDF) for the noise spectral estimate

$$p_1(S_k) = \frac{n^2}{\sigma^2} \exp\left(-\frac{n^2}{\sigma^2} S_k\right), \tag{5}$$

where σ^2 is the total mean square fluctuation of the Gaussian white noise in the measurement frequency band, and n is the total number of signal samples [12]. Otherwise, if the spectral estimate is the result of m averages, we find the gamma PDF with shape parameter m and scale parameter σ^2/n^2, i.e.,

$$p_m(S_k) = \left(\frac{n^2}{\sigma^2}\right)^m \frac{S_k^{m-1}}{(m-1)!} \exp\left(-\frac{n^2}{\sigma^2} S_k\right). \tag{6}$$

The slope a is computed repeatedly in a Monte Carlo (MC) simulation where the estimated spectral density (PSD) has no averaging and is fitted with the linear model $S_k = af_k + b$ over 10, 20, and 30 (gamma-distributed) data points. In the simulated data sets, the frequencies f_k are evenly spaced, as

*Edoardo.Milotti@ts.infn.it

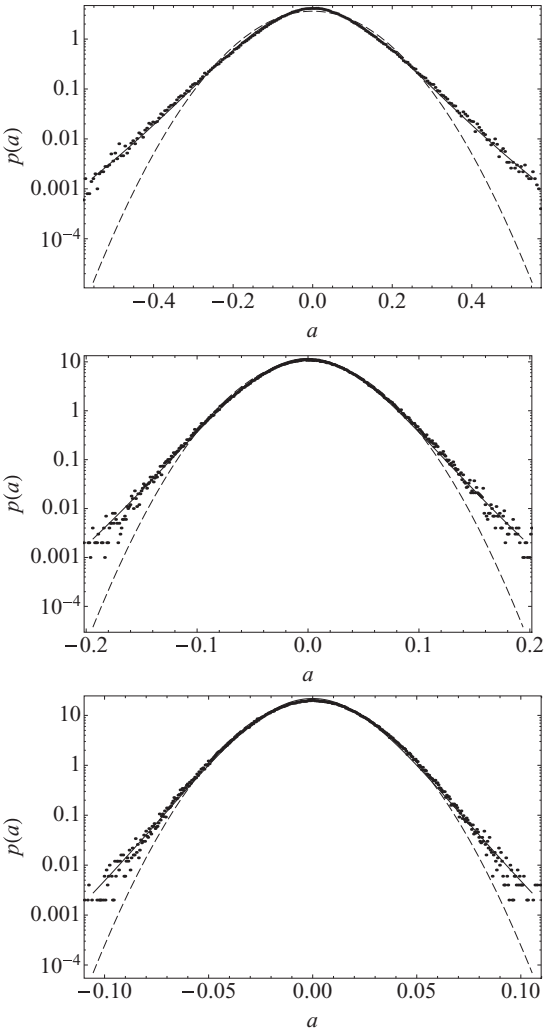

FIG. 1. Empirical PDF of the slope estimator $\hat{a}$ in the MC simulation described in the text, obtained from 10^6 MC simulated data sets for different numbers of data points. The dashed line is the normal PDF with the same mean and standard deviation as the empirical PDF. The solid line is a q-Gaussian fit. Upper panel: 10 data points; middle panel: 20 data points; lower panel: 30 data points. The higher the number of data points, the smaller the width of the distribution, and this means that, for a proper representation, the range of the horizontal axis must be reduced as the number of points gets larger. As the number of points increases, the empirical PDF gets closer and closer to a true Gaussian PDF [and the q value of the q-Gaussian fit approaches 1 (see Table I)].

is customary in DFT analyses. Figure 1 shows the PDF of the estimate of the slope for a large number of MC iterations (10^6) and different numbers of data points.

While the empirical PDF of the slope estimator a has an excellent Gaussian behavior close to the peak, it is also clear that it has non-Gaussian tails. It is interesting to remark that the PDF, including these tails, can be well approximated by a q-Gaussian PDF [14]

$$p_q(a) = C_q\left[1-(1-q)\left(\frac{a}{a_0}\right)^2\right]^{1/(1-q)}, \tag{7}$$

TABLE I. Numerical values for the q-Gaussian fits of Fig. 1. The q value decreases significantly and approaches 1 as the number of data points increases.

Number of points	q	a_0
10	1.185 ± 0.004	0.132 ± 0.002
20	1.148 ± 0.004	0.0469 ± 0.0005
30	1.128 ± 0.004	0.0257 ± 0.0003

where C_q is a normalization constant, a_0 is a width parameter, and q specifies the deviation from normality (if $q \to 1$, one recovers the usual Gaussian PDF). It has been stated that q-Gaussians are ubiquitous and that their observations strengthen the case for nonextensive statistical mechanics [15,16]. However, it has also been argued that, although q-Gaussians exhibit many interesting properties, there is no support for the idea that they do play a special role as limit distributions of correlated sums [17,18], and it is therefore important to examine carefully those features of statistical tools that may lead to wrong experimental conclusions. The q-Gaussian fits are shown in each panel of Fig. 1 as well, and the corresponding q values are listed in Table I. The estimator of b is less interesting, it follows the data points, and, thus, rather unsurprisingly, it has a skew distribution.

This is an artificial example; in this case, we know the exact asymptotic statistics for the estimator $\hat{a}$ (we know that it is Gaussian) and the appearance of the q-Gaussian is clearly a finite-size effect. However, Gaussianity may not be so obvious in more complex cases, and this simple example is an interesting lesson. Figure 1 shows very clearly that, as the number of data points in the data set increases, the deviation from normality in the PDF of the slope estimator becomes less evident, as shown numerically by the q values in Table I. The upper panel of Fig. 2 shows that, for an even larger number of data points, the deviation from normality is barely perceptible (notice also that the same happens if we fix the number of data points, but average over many data sets).

q-Gaussian PDFs have a power-law behavior for very large deviations; however, these regions are unexplored in the plots of Fig. 1, and the tails of the PDFs in Fig. 1 display interpolating regions with exponential behavior. It turns out that, in the simple example of the linear LS fit, it is possible to understand analytically how this behavior arises. Indeed, in the case of the linear model $y_k = ax_k + b$, we find the LS estimate

$$\hat{a} = \frac{S_0 S_{xy} - S_x S_y}{S_{xx} S_0 - S_x^2}, \tag{8a}$$

$$\hat{b} = \frac{S_{xx} S_y - S_x S_{xy}}{S_{xx} S_0 - S_x^2}, \tag{8b}$$

where

$$S_0 = \sum_k \frac{1}{\sigma_k^2}, \quad S_x = \sum_k \frac{x_k}{\sigma_k^2}, \quad S_y = \sum_k \frac{y_k}{\sigma_k^2},$$

$$S_{xx} = \sum_k \frac{x_k^2}{\sigma_k^2}, \quad S_{xy} = \sum_k \frac{x_k y_k}{\sigma_k^2},$$

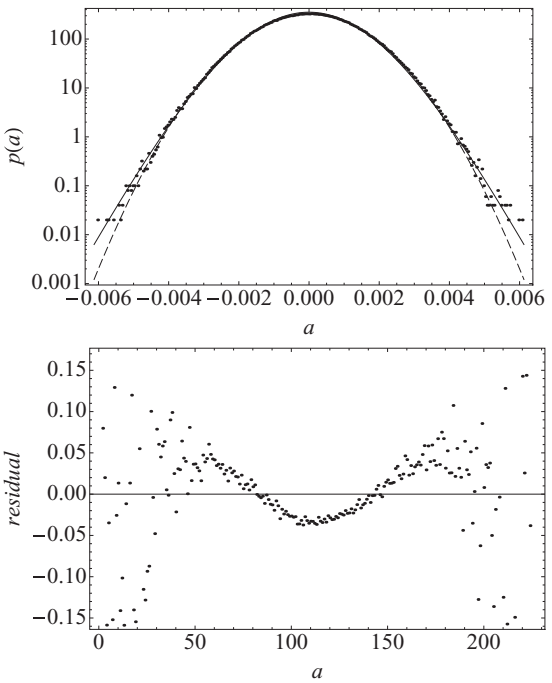

FIG. 2. Upper panel: Empirical PDF of the slope estimator $\hat{a}$ for a larger number of data points (200) obtained from 10^6 MC simulated data sets. The dashed line is the normal PDF with the same mean and standard deviation as the empirical PDF. The solid line is a q-Gaussian fit to the empirical PDF. A nearly identical empirical distribution is obtained if we keep 20 data points, but assume 10 averages (and thus a gamma PDF with shape parameter 10). Lower panel: Residuals of the q-Gaussian fit. Although the fit looks good, the residuals show that the q-Gaussian PDF deviates significantly from the numerical simulation data.

and the estimate of $\hat{a}$ [Eq. (8a)] can also be written in the form

$$\hat{a} = \frac{1}{S_{xx}S_0 - S_x^2} \sum_k \frac{S_0 x_k - S_x}{\sigma_k^2} y_k. \tag{9}$$

If the y_k's are i.i.d., exponential variates such as the white noise spectral data in the example discussed above, with PDF $p(y) = \tau^{-1} e^{-y/\tau}$, and we rewrite the sum (9) in the shorthand form

$$\hat{a} = \sum_k c_k y_k, \tag{10}$$

we see that the characteristic function (CF)[1] for the PDF of $\hat{a}$ is

$$f_{\hat{a}}(z) = \prod_k f_k(c_k z) = \prod_k \frac{1}{1 - i c_k z \tau}, \tag{11}$$

where $f_k(z) = f(z) = (1 - iz\tau)^{-1}$ is the CF of $p(y)$. There are some subtleties in the evaluation of this sum and general discussions can be found in [19–21]. For simplicity, here we assume that all the c_k's are different and positive, then the product can be expanded in partial fractions:

$$f_{\hat{a}}(z) = \prod_k \frac{1}{1 - i c_k z \tau} = \sum_k \frac{A_k}{1 - i c_k z \tau} \tag{12}$$

and the values of the A_j's can be found by multiplying both expressions times $1 - i c_k z \tau$, and taking $z = i/(c_j \tau)$, i.e.,

$$A_j = \prod_{k \neq j} \frac{c_j}{c_j - c_k}. \tag{13}$$

Mathematically, the CF is just a Fourier transform, and now the inversion of the CF is straightforward, and the PDF is a weighted distribution of exponentials

$$p(\hat{a}) = \sum_k \frac{A_k}{c_k \tau} e^{-\hat{a}/(c_k \tau)}. \tag{14}$$

It is important to remark that this result can be extended to negative c_k's, as in the present case: indeed, the differences $S_0 x_k - S_x$ in Eq. (9) are both positive and negative. To appreciate the meaning of Eq. (14), we take a very simple example, with two i.i.d. exponential variates with $\tau = 10$ and $c_1 = 1$, $c_2 = 1.1$. Then, $A_1 = -10$, $A_2 = 11$, so that the PDF of the sum $s = c_1 y_1 + c_2 y_2$ is $p(s) = -e^{-s/10} + e^{-s/11}$, i.e., it has the generic functional shape $-e^{\lambda_1 s} + e^{-\lambda_2 s}$ with $\lambda_1 > \lambda_2$. We can rearrange this expression as follows for $(\lambda_1 - \lambda_2)s \ll 1$:

$$\begin{aligned} p(s) &= -e^{-\lambda_1 s} + e^{-\lambda_2 s} = e^{-\lambda_2 s}(1 - e^{-(\lambda_1 - \lambda_2)s}) \\ &\approx e^{-\lambda_2 s}(\lambda_1 - \lambda_2)s, \end{aligned} \tag{15}$$

so that

$$\ln p(s) \approx -\lambda_2 s + \ln s + \ln(\lambda_1 - \lambda_2), \tag{16}$$

and we see that the tail of the distribution has a region that is nearly linear in a logarithmic plot and reproduces the

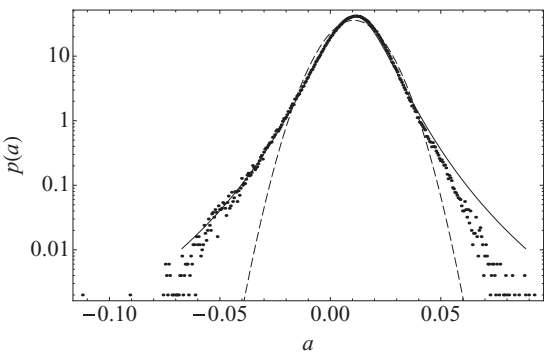

FIG. 3. Empirical PDF of the estimated slope a for a single data set of 20 points, with different measurement errors, and unevenly spaced values of the independent variable. The empirical distribution has been obtained with 10^6 bootstrap resamplings. The dashed line is the normal PDF with the same mean and standard deviation as the empirical PDF. The solid line is a q-Gaussian PDF adapted to the empirical PDF. Now, in addition to the long tails of the distribution, there is also some asymmetry, however, the q-Gaussian still provides a better fit. It is interesting to notice that this non-Gaussianity disappears when the independent variable has evenly spaced values.

[1]The characteristic function $f(z)$ for a PDF $p(x)$ is defined by $f(z) = \mathbf{E}(e^{izx}) = \int_{-\infty}^{+\infty} p(x) e^{izx} dx$.

intermediate region with the exponential decay displayed in Figs. 1 and 2.

Non-Gaussianity can also arise in other subtle ways; even with Gaussian data, the LS method can yield non-Gaussian tails if we somehow "forget" the original Gaussian distribution. This happens, e.g., in analyses that use the statistical bootstrap [22]. Indeed, the usual MC bootstrap procedure resamples data points from a single data set, and the resampling procedure does not retain any memory of the original distribution of data points. An example is shown in Fig. 3, which shows the empirical PDF of the slope a in a straight line fit problem similar to the one discussed above.

The q-Gaussians obtained in the fits of Figs. 1, 2, and 3 are all artifacts due to finite-size effects in the data analysis. A careful scrutiny also shows that the q-Gaussians slightly deviate from simulation data, as shown in the lower panel of Fig. 2. The similarity that q-Gaussians bear to several experimental results thus calls for great caution, and finite-size effects such as those observed in this paper may be an additional explanation of the frequent occurrence of good q-Gaussian fits [18,23].

[1] N. Bartolo, E. Komatsu, S. Matarrese, and A. Riotto, Phys. Rep. **402**, 103 (2004).

[2] Y. Wiaux, P. Vielva, R. B. Barreiro, E. Martínez-González, and P. Vandergheynst, Mon. Not. R. Astron. Soc. **385**, 939 (2008).

[3] A. J. Moreno, I. Saika-Voivod, E. Zaccarelli, E. L. Nave, S. V. Buldyrev, P. Tartaglia, and F. Sciortino, J. Chem. Phys. **124**, 204509 (2006).

[4] F. Ritort, C. R. Phys. **8**, 528 (2007).

[5] E. Marsch and C. Y. Tu, Ann. Geophys. **12**, 1127 (1994).

[6] D. Koga, A. C.-L. Chian, R. A. Miranda, and E. L. Rempel, Phys. Rev. E **75**, 046401 (2007).

[7] S. Umarov, C. Tsallis, and S. Steinberg, Milan J. Math. **76**, 307 (2008).

[8] M. Saraceni, M. Bonaldi, L. Castellani, L. Conti, A. B. Gounda, S. Longo, and M. Pegoraro, Rev. Sci. Instrum. **81**, 035115 (2010).

[9] R. J. Barlow, *Statistics: A Guide to the Use of Statistical Methods in the Physical Sciences (Manchester Physics Series)* (Wiley, New York, 1989).

[10] G. Cowan, *Statistical Data Analysis* (Oxford University Press, New York, 1998).

[11] V. Zolotarev, J. Math. Sci. **92**, 4112 (1998).

[12] M. Priestley, *Spectral Analysis and Time Series (Probability and Mathematical Statistics)*, Vols. I and II (Academic, New York, 1983).

[13] J. A. Gubner, *Probability and Random Processes for Electrical and Computer Engineers* (Cambridge University Press, New York, 2006)

[14] C. Tsallis, S. Levy, A. Souza, and R. Maynard, Phys. Rev. Lett. **75**, 3589 (1995).

[15] C. Tsallis and E. Brigatti, Continuum Mech. Thermodyn. **16**, 223 (2004).

[16] C. Tsallis and U. Tirnakli, J. Phys.: Conf. Ser. **201**, 012001 (2010).

[17] H. J. Hilhorst and G. Schehr, J. Stat. Mech.: Theory Exp. (2007) P06003.

[18] T. Dauxois, J. Stat. Mech.: Theory Exp. (2007) N08001.

[19] A. Mathai, Commun. Stat.: Theory Methods **12**, 625 (1983).

[20] P. G. Moschopoulos, Ann. Inst. Stat. Math. **37**, 541 (1985).

[21] H. V. Khuong and H.-Y. Kong, Commun. Lett., IEEE **10**, 159 (2006).

[22] B. Efron, Ann. Stat. **7**, 1 (1979).

[23] C. Vignat and A. Plastino, Phys. A (Amsterdam) **388**, 601 (2009).