Skewed distributions as limits of a formal evolutionary process

F Sattin
Consorzio RFX, Corso Stati Uniti 4, I-35127 Padova, Italy
E-mail: fabio.sattin@igi.cnr.it

Received 9 January 2018, revised 8 May 2018
Accepted for publication 12 June 2018
Published 27 June 2018

Abstract
Time series of observables measured from different turbulent systems do often exhibit non-normal statistics, their probability distribution functions (PDFs) are not Gaussian and often skewed, with roughly exponential tails, skewness and kurtosis. Departure from gaussianity is related to the intermittent development of large-scale coherent structures. The existence of these structures is rooted into the nonlinear dynamical equations obeyed by each system, therefore it is expected that some prior knowledge or guessing of these equations is needed if one wishes to infer the corresponding PDF; conversely, the empirical knowledge of the PDF does provide information about the underlying dynamics. In this work we advance the suggestion that this is not always necessary. We show that, under some assumptions, a formal evolution equation for the PDF $p(x)$ can be written down, corresponding to the progressive accumulation of measurements of the generic observable $x$. The limiting solution to this equation is computed analytically, and shown to interpolate between some of the most common distributions, Gamma, Beta and Gaussian PDFs. The control parameter is just the ratio between the root mean square of the fluctuations and the range of allowed values. Thus, no information about the dynamics is required.

Keywords: statistical distributions, data analysis, replicator equation, evolutionary dynamics

(Some figures may appear in colour only in the online journal)

1. Introduction
Signals produced by turbulent, random or chaotic systems do fluctuate around their mean value. While the normal (Gaussian) curve is the paradigmatic distribution for these fluctuations, there exist as well several counterexamples of signals produced from turbulent systems which exhibit non-normal statistical properties: their probability distribution functions (PDF’s) are not Gaussian and often skewed, with roughly exponential tails, skewness and kurtosis. Typical examples are Gamma, log-normal, Weibull distributions [1]. The fields where they are encountered span several scientific disciplines (plasma physics, meteorology, financial data, oceanography, biology, ...). Departure from gaussianity in those cases is commonly related to the intermittent development of large-scale coherent structures that break the independence between measurements [2]. Ultimately, the existence of these structures is argued on the basis of the equations governing the dynamics of the system, and in particular by their nonlinearities. In some cases the true equations are not known and replaced by phenomenological models. Thus, a prerequisite for inferring the possible shape of the PDF should be some knowledge or guessing about the system’s dynamics [2–9].

Conversely, one might expect to employ the empirical knowledge about the PDF in order to infer some information about the unknown underlying dynamics of the system studied. The same kind of distribution may appear in totally different fields. On the other hand, different experiments measuring the same quantity may yield different PDFs: for instance, laboratory plasma physics where particle density measurements are mostly interpolated by Gamma distributions, but other PDF’s are suggested to be suitable candidates as well [10–15]. This might imply that the governing equations must admit several possible classes of solutions.

In this work we investigate the possibility that the analytical form of the PDF might be fixed by just few gross constraints which are extracted by the measured data, without
invoking any detailed knowledge of the underlying dynamics, even at the statistical level. To make pictorial this claim we resort to the following paradigmatic case: let us imagine to be measuring some positive-definite quantity, i.e. a (number, mass, ...) density. Suppose as well that the measured quantity is fluctuating and the typical amplitude of its fluctuations (the rms) is comparable to the average value. It is obvious that negative (below the average) fluctuations are penalized, since no less-than-zero values are permitted, whereas no similar constraint holds for the positive (above the average) ones. Hence, one must expect a fortiori the PDF to be skewed and non-Gaussian: a log-normal PDF, e.g., might be appropriate in this instance [16]. Of course, dynamical equations do contain the constraint of positiveness but our point is that, in this instance, it is not necessary to invoke them. Another intriguing aspect is that it is not necessary that these constraints do reflect an intrinsic property of the system; rather, they may be imposed by the measuring apparatus as well. Hence—in principle—the same system observed by different observers may produce different PDFs.

2. Evolutionary processes

We describe the measurement process as an evolutionary process. The evolution does not occur in real time, but in a fictitious time whose flow parametrizes any increase of information about the system, acquired, e.g., via measurements, within a Bayesian viewpoint, and is therefore a fairly generic feature. The evolving quantity is the PDF, obtained by binning the measured signal, and the quantities of interest are the stable solutions of this equation. Our approach is thus closely affine to that pursued by Choi et al [17] (see also [18, 19]) and Smerlak and Youssef [20].

Let us introduce the model. Hereafter, we will suppose that the output of our experiment is a real-valued scalar signal \( z \). Any such signal is bounded on an interval \((a, b)\). The two extrema are either fixed by physical constraints, or are empirically determined by experiments. For the forthcoming analytical manipulations it is convenient to deal with a PDF whose support is semicompact: either \( a \to -\infty \) or \( b \to +\infty \).

If both of them are finite, then it is necessary to transform the measured variable, say switch to \( x = \log[(z - a)/(b - a)] \) which converts the initial interval \((a, b)\) into \((-\infty, 0)\). In most scenarios choosing the opposite range \((0, +\infty)\) would be a more convenient choice, since measured signals are ordinarily defined as positive numbers. However, the difference is immaterial since amounts to just a change of sign of the observable involved. The present choice has the advantage of being consistent with the existing literature on evolutionary dynamics.

Let thus \( x \) be the measured quantity, and \( p_0(x) \) its prior density, which may have been obtained via some sets of measurements. Any further investigation provides new information about \( x \) and, correspondingly, acts to modify \( p_0 \): \( p_0 \mapsto p \). Formally we can represent this through a state-transition operator:

\[
p(x) = \int W(x, y)p_0(y)dy. \tag{1}
\]

After subtracting \( p_0 \) from both members we arrive to

\[
\Delta p \equiv p(x) - p_0(x) = \int (W(x, y) - \delta(x - y))p_0(y)dy \\
\equiv \int \Theta(x, y)p_0(y)dy. \tag{2}
\]

It is implicit that all the integrals are performed over the whole domain of the variable \( y \). Equation (2) is an instance of a Master equation [21]; a differential (or, like in this case, finite-differences) equation for the evolution of the probability density of a state \( p(x) \) in terms of transition rates between different states. The function \( \Theta \) accounts for the flow of probability from \( y \) to \( x \). Generally, it is expected to be system-dependent. However, we will look for universal traits that restrict its possible analytical forms. There are actually two constraints; the first one comes by integrating (2) over \( x \):

\[
\int dx \Delta p = 0 \\
\rightarrow \int dx \int dy \Theta(x, y)p_0(y) = 0 \quad \forall p_0(y). \tag{3}
\]

Secondly, the diagonal part of the operator, i.e. \( \Theta(x, x) \), accounts for a flow of density \( p_0 \) into itself. Since this is already accounted by \( p_0 \) in the lhs of (2), this term must be zero: \( \Theta(x, x) = 0 \).

These two constraints allow to guess a plausible form for \( \Theta \). First, we factorize it as: 

\[
\Theta(x, y) \equiv T(x - y)\Omega(x), \tag{4}
\]

where \( T \) quantifies the probability flux along distance \( |x - y| \). We must account for inhomogeneity as well, and this is brought by \( \Omega \), that quantifies the propensity for each point \( x \) to act as final target. Let us start guessing an analytical expression for \( \Omega \). The propensity may be estimated a posteriori on the basis of the past history of the system, which is quantified by \( p_0 \) itself.

Thus, we may guess that \( \Omega(x) = \Phi(p_0(x)) \), and we choose the simplest proxy: \( \Omega(x) = \Omega_0 \times p_0(x) \) (incidentally, this implies that the Master equation in equation (2) is actually a nonlinear one [22]; transition rates contain the probability density itself). As far as the function \( T \) is concerned, the simplest choice compliant with the requirement \( \Theta(x, x) = 0 \) is

\[
T(x - y) = T_0 \times (x - y). \tag{5}
\]

In conclusion, equation (2) rewrites

\[
\Delta p = \Delta t (x - \mu)p_0(x), \quad \Delta t = T_0\Omega_0, \quad \mu = \langle x \rangle. \tag{6}
\]

Notice that \( \Delta t \) does not depend from \( x \) or \( p_0 \); it is a factor proportional to the difference between the two successive estimates \( p_0 \). By assuming that each set of measurements increases only marginally the information about \( p_0 \), i.e. \( \Delta p \to 0 \), we require therefore \( \Delta t \to 0 \), and \( \Delta t \) may thus be attributed the meaning of an infinitesimal increment of a variable \( t \) that parametrizes the state of information about \( x \). Equation (5) takes formally the expression of a finite-difference equation,

\[
\Delta p = \Delta t (x - \mu)p_0(x), \quad \Delta t = T_0\Omega_0, \quad \mu = \langle x \rangle. \tag{7}
\]

Notice that \( \Delta t \) does not depend from \( x \) or \( p_0 \); it is a factor proportional to the difference between the two successive estimates \( p_0 \). By assuming that each set of measurements increases only marginally the information about \( p_0 \), i.e. \( \Delta p \to 0 \), we require therefore \( \Delta t \to 0 \), and \( \Delta t \) may thus be attributed the meaning of an infinitesimal increment of a variable \( t \) that parametrizes the state of information about \( x \). Equation (5) takes formally the expression of a finite-difference equation,
which may further be turned into a differential equation:
\[
\frac{\Delta p}{\Delta t} \approx \frac{\partial p}{\partial t} = (x - \mu)p(x, t). \tag{6}
\]

Equation (6) is a nonlinear integro-differential equation, often encountered in biosciences and game theory, where is known as ‘replicator equation’ [23]. In those contexts, \( t \) is the true time whereas in our case it is a fictitious time, parameterizing the flow of information. Several different choices for \( \Theta \) other than \( T, \Omega \) above are obviously possible in principle. However, they are penalized in terms of complexity criteria: equation (6) is not the only possible evolution equation but is the simplest one.

3. Solutions of the replicator equation

Let us discuss some features of (6). First of all, the support of \( p \) does not vary throughout time: if \( p(x) = 0 \) at some time for some \( x \) then \( p(x) \) will be zero at all times. It is possible to generalize equation (6) by including the possibility for \( p \) to spread over larger and larger intervals of \( x \) (see [24]) but for the moment we will not discuss this option and will return back to it only in the final part of this work. Second, equation (6) admits a Dirac-delta stationary solution: \( p(x) = \delta(x - x_0) \). Within our picture, this solution corresponds to the case where enough information has been collected to assign a univocal value to the observable \( x: t \to \infty \). Physically, this limit does not make sense and is impossible for a number of reasons, among others, quantum-mechanical effects; only finite \( t \) is meaningful.

Smerlak and Youssef [24] worked out analytically the solution to equation (6) for large times. We do not replicate here their study, rather we follow the other way round by plugging their result into equation (6) and showing that it is a valid solution at all times. The solution takes the form of a flipped Gamma function:
\[
p(x, t) = \frac{C(t)^{x+1}}{\Gamma(\gamma + 1)} \exp[C(t)x](-x)\gamma. \tag{7}
\]

The parameter \( C \) is conveniently expressed in terms of the average value \( \mu \):
\[
C(t) = -\frac{\gamma + 1}{\mu(t)}. \tag{8}
\]

In equation (8) we have anticipated that \( C \) (and \( \mu \)) is a function of time. By taking the time derivative of the logarithm of (7) we get
\[
\frac{\partial \log p}{\partial t} = \frac{(1 + \gamma)(x - \mu)}{\mu^2} \mu'(t) \tag{9}
\]
which reduces to (6) when
\[
\frac{(1 + \gamma)}{\mu^2} \mu'(t) = 1 \to \nu(t) \to \frac{1}{1 + \gamma - t\mu(0)}, \quad \mu(0) \equiv \mu(t = 0). \tag{10}
\]

It is easy to check that:

\begin{itemize}
  \item[(i)] \( \mu(t) \to 0 \) as \( t \to \infty \), hence \( p \) converges towards \( p_0 = \delta(x) \) and \( p \) is self-similar: it does not change shape as time grows;
  \item[(ii)] the rate of convergence slows down and is very small as time increases: \( \mu'(t) \propto t^{-2}, t \to \infty \). On the other hand, \( \mu \) varies rapidly in the first stages, which suggests that, regardless of the starting point \( p_0P \) approaches quickly its limiting value (7). The numerical integration of equation (6) confirms the validity of this guess (see figure 1).
\end{itemize}

Notice that \( p \) of equation (7) is a stable solution: if one starts from a different curve, the solution evolves eventually approaching equation (7).

Ultimately, we have to revert to the physical variable \( z \) by replacing into (7) \( x \) with: \( x \to \log[(z - a)/(b - a)] \). The result is
\[
p(z) \propto [-\log(y)]^{y-1}, \quad y \equiv \frac{z - a}{b - a}. \tag{11}
\]

This expression is not particularly illuminating, but its limiting forms are revealing. The relevant figures here are the ratios \( \sigma/a, b \), where \( \sigma \) is a measure of the typical amplitude of the signal: say the average value, or the mean fluctuation around the average. We will consider both \( a, b \) to be non-zero in order not to bother with sub-cases.

\begin{itemize}
  \item[(i)] Let us keep \( \sigma/b \) finite and take \( |\sigma/a| \ll 1 \). This obviously corresponds practically to reverting back to the original variable \( x \): indeed, \( x = \log[(z - a)/(b - a)] \approx \log(1 - z/a) \approx z/(z-a) \). Hence, \( p(z) \) turns into a Gamma function like equation (7).
  \item[(ii)] A second limiting case involves both \( a, b \) to be very large: \( |\sigma/a|, |\sigma/b| \ll 1 \). The function \( p(z) \) has its maximum in \( z_M : \log((z_M - a)/(b - a)) = \gamma/(C - 1) \). If we expand \( \log p \) around this point we get
\[
\log p/p(z_M) \approx -\left(\frac{z - z_M}{b - a}\right)^2 + O(\frac{z - z_M}{b - a})^3. \tag{12}
\]
\end{itemize}
Thus now, \( \sigma = \sqrt{\langle (z-z_m)^2 \rangle} \). As long as \( \sigma \) remains much smaller than \( b - a \), hence, the lowest-order term dominates and \( p \) reduces to a Gaussian PDF.

(iii) Finally, let us consider the case when both \( \sigma/a \) and \( \sigma/b \) are finite. A fairly versatile family of functions, often employed to model distributions within finite domains, is the Beta distribution [1], with possibly its generalizations [25]:

\[
\begin{align*}
B_{p,q}(y) &= \frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)} y^{p-1}(1-y)^{q-1},
\end{align*}
\]

where \( \Gamma \) is the usual gamma function. In figure 2 we show that (11) indeed does fit remarkably well equation (13). This is not unexpected, since, for \( y \approx 1 \), we can Taylor expand:

\[
-\ln(y) \approx (1-y) + O(1-y^2)
\]

(see e.g. [26]) and there is an explicit correspondence between equations (11) and (13): the Beta PDF can be regarded as the lowest-order series expansion of (11). In conclusion, we may claim that even this family of distributions is accounted within our model.

4. The skewness–kurtosis characteristic

Rather than inspecting the whole PDF, an often employed strategy when dealing with experimental data is to study the mutual relations between the lowest-order statistical moments, in particular the third (skewness \( S \)) and fourth (kurtosis \( K \)) one. Low-order statistical moments are robust quantities to compute from raw data. Furthermore, several PDFs feature characteristic correlations between them. For the Gamma distribution (7), in particular, it is easy to show that

\[
K = \frac{3}{2} S^2 + 3.
\]

(The Gaussian PDF is a special case of this law, with \( S \equiv 0, K \equiv 3 \)). It is a well established empirical result that data from very diverse fields do obey a law like (15): \( K = AS^2 + B \), with \( A, B \) close to the values 3/2, 3, respectively. This scaling has been studied extensively in laboratory plasma physics [12, 13, 15, 27–29], but also in oceanography [30], meteorology [31], seismology and financial time series [32, 33]. In figure 3 we plot the \( (S, K) \) curve from equation (15) together with a large sample of \( (S, K) \) couples computed from (11) by varying \( \gamma \) and \( C \). Expression (11) does not allows for a unique relation between \( S \) and \( K \), rather a cloud of points is generated, which is bounded from above by the Gamma limiting solution (15). Visual comparison of figure 3 with the similar ones from experiments shows that there is a good overlap between the two datasets, but for one aspect: equation (11) does not appear to include solutions featuring \( K > 3/2S^2 + 3 \), whereas part of the experimental data do lie in this region. In terms of PDFs this implies that datasets do exist whose histograms are not interpolated by none of the distributions studied above. This is, at this stage, a shortcoming of the model but we do not think it is a critical one. One possible way of addressing this issue is by acknowledging that we have so far dealt with immutable ranges \( (a, b) \). This is justified as long \( a, b \) are fixed \( a \text{ fortiiori} \) by the physics of the system or by the measuring apparatus. In several situations, on the other hand, \( a, b \) are only defined \( a \text{ posteriori} \); as the extrema of the signal measured so far, hence they are susceptible of varying just like \( p \) does. Smerlak and Youssef [24] argue that this effect can be accounted by adding an effective diffusive term in equation (6)—converting it into a nonlinear Fokker–Planck equation with reaction terms. Their numerical studies of this modified equation show that—transiently—its solutions do move in the \( (S, K) \) plane above the curve \( 3/2S^2 + 3 \).

5. Conclusions

We are speculating that some of the statistical distributions encountered in the analysis of experimental data might be
justified as enforced by few natural constraints. Admittedly, our results cannot be viewed as conclusive nor exhaustive: a number of important PDFs are excluded from our analysis, including, e.g., power-law, skew-t and skew-normal PDFs [26]. Our whole discussion relies basically on just the three ansatzs defining the functions $\Theta, \Omega, T$, which have been justified on the basis of mathematical simplicity considerations. There are some cases where it is known a fortiori that these ansatz alone do not exhaust the relevant physics; for example, the PDF of particle flux measured at the edge of fusion devices has been analytically computed as the convolution between two correlated Gaussian PDFs [34], which does not reduce to any of the above mentioned PDFs. The missing element unaccounted in our model is the correlation between the two ingredients: particle density and velocity. The appealing consequence of our hypotheses, on the other hand, is that we are able to reduce ourselves to just one basic solution, equation (11). Several of the best known and most common distributions do arise just as several different cases of this solution. The relevant parameters for interpolating between the one and the other limits are, roughly, the ratios between the typical measured values (say, the rms) and the maximum measurable ones, with the Gaussian PDF appearing when both ratios go to zero. It is intriguing to notice that we are thus able to justify the universality of PDFs across several different systems, while at the same time providing a rationale for one kind of signals being modelled by different curves in different experiments.

Acknowledgments

The author wishes to thank S Cappello, D Escande, N Vianello, M Agostini, M Zuin and I Predebon for useful discussions.

ORCID iDs

F Sattin https://orcid.org/0000-0002-4308-350X

References

[1] Johnson N L, Kotz S and Balakrishnan N 1995 Continuous Univariate Distributions 2nd edn (New York: Wiley)
[2] Portelli B, Holdsworth P C W and Pinton J-F 2003 Phys. Rev. Lett. 90 104501
[3] Sattin F and Vianello N 2005 Phys. Rev. E 72 016407
[4] Garcia O E 2012 Phys. Rev. Lett. 108 265001
[5] Garcia O E, Kube R, Theordosen A and Pécseti H L 2016 Phys. Plasmas 23 052308
[6] Guszejnov D, Lazányi N, Bencze A and Zoletnik S 2013 Phys. Plasmas 20 112305
[7] Krommes J A 2008 Phys. Plasmas 15 030703
[8] Carbone V et al 2002 Europhys. Lett. 58 349
[9] Sandberg I, Benkadda S, Garbet X, Ropokis G, Hizadinis K and del-Castillo-Negrete D 2009 Phys. Rev. Lett. 103 165001
[10] Sattin F, Vianello N and Valisa M 2004 Phys. Plasmas 11 5032
[11] Graves J P, Horacek J, Pitts R A and Hopcraft K I 2005 Plasma Phys. Control. Fusion 47 L1
[12] Labit B, Furno I, Fasoli A, Diallo A, Müller S H, Pulyushchev G, Podestá M and Poli F M 2007 Phys. Rev. Lett. 98 255002
[13] Sattin F, Agostini M, Cavazzana R, Scarin P and Terry J L 2009 Plasma Phys. Control. Fusion 51 055013
[14] Garcia O E, Fritzner S M, Kube R, Cziegler I, LaBombard B and Terry J L 2013 Phys. Plasmas 20 055901
[15] Banerjee S et al 2014 Phys. Plasmas 21 072311
[16] Sattin F, Vianello N, Valisa M, Antoni V and Serianni G 2005 J. Phys.: Conf. Ser. 7 247
[17] Choi M Y, Choi H, Fortin J-Y and Choi J 2009 Europhys. Lett. 85 30006
[18] Kaupuzs J, Mahnke R and Weber H 2010 Europhys. Lett. 91 30004
[19] Gadomski A 2010 Europhys. Lett. 89 40002
[20] Smerlak M and Youssef A 2017 Europhys. Lett. 117 50003
[21] Gardiner C W 2009 Handbook of Stochastic Methods (Berlin: Springer)
[22] Nicolis G and Prigogine I 1977 Self-Organization in Nonequilibrium Systems (New York: Wiley)
[23] Taylor P D and Jonker L B 1978 Math. Biosci. 40 145
[24] Smerlak M and Youssef A 2017 J. Theor. Biol. 416 68
[25] Pham-Gia T and Duong Q P 1989 Math. Comput. Modelling 12 1613
[26] Contreras-Reyes J E 2014 Physica A, 385 200
[27] Sattin F, Agostini M, Cavazzana R, Serianni G, Scarin P and Vianello N 2009 Phys. Scr. 79 045006
[28] Yan N et al 2013 Plasma Phys. Control. Fusion 55 115007
[29] Kube R, Theordosen A, Garcia O E, LaBombard B and Terry J L 2016 Plasma Phys. Control. Fusion 58 054001
[30] Sura P and Sardeshmukh P D 2008 J. Phys. Oceanogr. 38 638
[31] Schlopfchler T P and Sullivan P J 2005 Bound.-Layer Meteorol. 115 341
[32] Cristelli M, Zaccaria A and Pietronero L 2012 Phys. Rev. E 85 066108
[33] Adcock C, Eling M and Loperfido N 2015 Eur. J. Finance 21 1253
[34] Carreras B A et al 1996 Phys. Plasmas 3 2664