ELSEVIER

Physica D 132 (1999) 111–132

PHYSICA D

www.elsevier.com/locate/physd

On probability distribution functions in turbulence. Part 1. A regularisation method to improve the estimate of a PDF from an experimental histogram

B. Andreotti *, S. Douady

Laboratoire de Physique Statistique de l'Ecole Normale Supérieure, Laboratoire associé au CNRS at Universités Paris VI et Paris VII, 24 rue Lhomond, 75231 Paris cedex 05, France

Received 11 February 1998; received in revised form 22 February 1999; accepted 22 February 1999
Communicated by U. Frisch

Abstract

The most common method to estimate a probability distribution function (PDF) from experimental data is to compute a normalised histogram. This approximation implicitly assumes that the PDF is smooth at the scale of one histogram bin. Usually, the normalised histogram is ill defined for the rarer events since the points are very scattered in that region. In order to increase the quality of the PDF estimate, the assumption that the PDF is smooth can be used explicitly. A specially designed regularisation method is constructed and tested on both synthetic and real turbulence signals. Using this procedure, the estimated PDFs are now smooth and well-defined up to the unique rarest event (the last histogram point). Among its direct applications, the method allows to get a better estimate of high order PDF moments and of PDFs convolution products.

Keywords: Statistics; Probability distribution function; Turbulence

1. Introduction

The aim of this paper is to construct a procedure to improve the estimate of a probability distribution function (PDF) from an experimental or numerical signal. This study has been motivated by turbulence problems. As in many other physical systems, turbulence appears to be highly disorganised, quite unpredictable and presents a very large range of physical scales. However, for given experimental conditions, the statistical properties of any signal measured in turbulence are almost reproducible. These basic observations have induced physicists to look for a probabilistic description of turbulence. For this purpose, PDFs of various physical quantities (velocity, velocity increments, pressure, passive scalar, etc.) have been measured both experimentally and numerically. The question of the intermittency of these signals has induced a strong interest for the rare (but large) events which correspond to

* Corresponding author.

0167-2789/99/$ – see front matter
PII: S0167-2789(99)00040-8

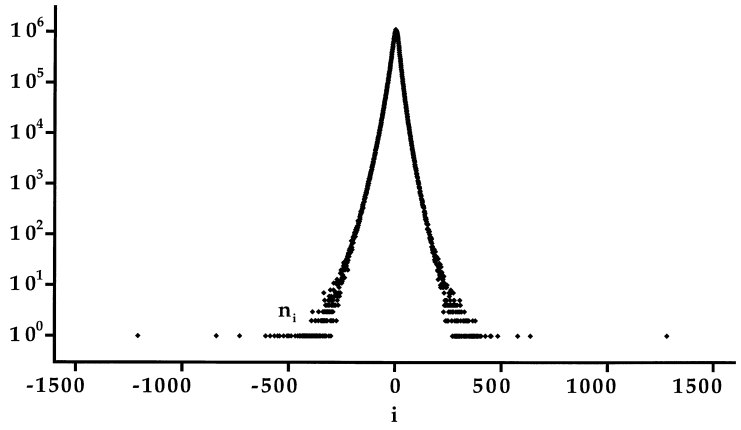

Fig. 1. The example histogram $\{n_i\}$ (velocity gradient in a particular turbulent flow [2]), as a function of the bin number i and computed for $N \approx 3.7 \times 10^7$ points.

the tails of the corresponding PDFs. To investigate the influence of these rare events, for instance by the measurement of PDFs high order moments, it is necessary to record and process very long data signals, however limited at any rate by numerical and experimental constraints. A recent state of the art can be found in [1] and in references therein.

We want, thus, to develop a regularisation method to increase, for a given data set, the accuracy of the PDF estimate, and particularly for the rarest events. In a second article, we will present the results obtained on a real turbulent velocity signal using this procedure. We will here present the method and discuss its validity. It is based on the a priori constraint that the PDF is smooth (Section 2). It consists in the minimisation of the sum of two functionals, one which measures the PDF smoothness (Section 3) and the other which characterises the likelihood of the PDF to the histogram (Section 4). An adequate test is used to adjust the best compromise between these two requirements (Section 5). The applications of the method to experimental histograms are discussed in Sections 6 and 7, before showing the benefits of the method on the example of a real turbulence signal (Section 8).

2. Basic principles

2.1. Histogram, cumulated histogram and rank ordering

For a start, let us consider a particular velocity gradient signal $\gamma(t)$ measured as a function of time in one point of a turbulent flow (see [2] for details). This signal $\gamma(t)$ is sampled at evenly spaced time intervals. Let τ denote the time interval between consecutive samples. The signal is now a sequence of N sampled values: $\gamma(n\tau)$ with $n = 1, 2, \ldots, N$ ($N \approx 3.7 \times 10^7$ for the example chosen). In order to characterise the statistics of γ, the temporal information has no importance. A first method is then to compute the histogram of the velocity gradient γ. For this, the sampled values are binned into channels of width Δ and centred around $\gamma_i = (i - 1/2)\Delta$. The number of events observed in the ith bin is denoted by n_i. The resulting histogram $\{n_i\}$ is shown in Fig. 1 in the linear-logarithmic representation in order to highlight the rarest events. An alternative is to consider the cumulated histogram, i.e., to compute the number of events smaller than γ_i and that, for each grid point i. The histogram is the discrete difference of the cumulated histogram and thus contains strictly the same information: they are two ways of presenting the same discrete data, the cumulated histogram being however smoother by construction. A striking consequence is that the moments of order p estimated from one or the other are strictly equal whereas the quantity to be integrated to obtain the moment is much smoother with the cumulated histogram than with the histogram. There is, however,

a possibility to construct a cumulative distribution function without binning, and thus keeping more information than a binned histogram.

When the histogram (or the cumulated histogram) is computed, a part of the information contained in the signal is lost: the whole list of the N values $\gamma(n\tau)$ is no longer known exactly but is defined with an imprecision $\pm\Delta/2$ on each sampled value. From the probabilistic point of view, this error is negligible for the rarest events (Fig. 1) since the latter could have appear with the same probability several bins away from their real position. The direct consequence is a large reduction of data without losing much information. For instance, instead of the $N \approx 3.7\times10^7$ sampled values, the histogram of Fig. 1, obtained with $\Delta = 2.16\times10^{-3}\gamma_{\rm rms}$, is composed by 3×10^3 histogram bins (from first to last empty bin). As a conclusion, the loss in information between the whole signal and the histogram is not that important while the data compression between the two is massive. An alternative is the rank ordering [3,4] which consists in keeping a part of the sampled values together with their rank.

2.2. On the normalised histogram as an (bad) approximation of a PDF

Before discussing the estimate of the velocity gradient PDF from this histogram, let us briefly recall its definition. Roughly speaking, $p(\gamma)\mathrm{d}\gamma$ is the probability of observing γ at an arbitrary time between γ and $\gamma+\mathrm{d}\gamma$. To define it more precisely, we have to consider first the cumulative distribution function $P(g)$ which is the probability to observe γ smaller than g at an arbitrary time. The cumulative distribution function increases from 0, for the smallest accessible velocity gradient, to 1, for the largest one. Its derivative $p(\gamma)$ is positive and is by definition the PDF of γ. The PDF $p(\gamma)$ can evidently be either a function or a distribution (for instance if γ only takes discrete values).

Usually, the PDF $p(\gamma)$ is directly deduced from the experimental histogram $\{n_i\}$ via a normalisation by the factor:

$$p(\gamma_i) \approx \frac{n_i}{N\Delta} \tag{1}$$

This estimate of the PDF can be justified by two successive approximations:

- Firstly, n_i is an estimate of its average $\tilde{\pi}_i$ obtained with an infinite number of realisations of the experiment with the same number N of sampled values ($\tilde{\pi} \equiv \langle n_i\rangle$). To say it in another way, $\tilde{\pi}_i/N$ is the probability that a point falls in the ith bin ($\tilde{\pi}_i \equiv N(P(\gamma_i+\Delta/2) - P(\gamma_i-\Delta/2))$).
- Secondly, $\tilde{\pi}_i/N\Delta$ is itself a finite difference approximation of the PDF:

$$p(\gamma_i) \approx \frac{P(\gamma_i+\Delta/2) - P(\gamma_i-\Delta/2)}{\Delta} \approx \frac{\tilde{\pi}_i}{N\Delta} \tag{2}$$

The histogram points (Fig. 1) form a well-defined curve in the large probability region. On the contrary, in the tails, the histogram is composed of bins containing only one point ($n_i = 1$) separated by empty bins ($n_i = 0$). Besides, it is interesting to note that the linear-logarithmic representation of the histogram tails is problematic since the bins for which $n_i = 0$ ($\ln(n_i) = -\infty$) cannot be shown (Fig. 1). To investigate the link between an histogram and its corresponding PDF in the tails, let us consider an analytical PDF which exhibits large algebraic tails such as for instance:

$$\tilde{\pi}_i \propto \frac{N}{(1+(ai)^2)^2} \tag{3}$$

The curve $\tilde{\pi}_i$ is shown in Fig. 2 (solid line) together with a histogram example randomly generated using the multinomial distribution and having $N = 4\times10^7$ points (see Section 4). If the histogram is dispersed around the PDF for large probability bins ($\tilde{\pi}_i \geq 1$), it is as previously composed by scattered points in the tails (for $\tilde{\pi}_i \leq 1$). $\tilde{\pi}_i$ can be interpreted as the mean local density of points. If there is on the average less than one point per bin, this density is of the order of the inverse of the distance between neighbouring points. For instance, $\tilde{\pi}_i = 10^{-3}$

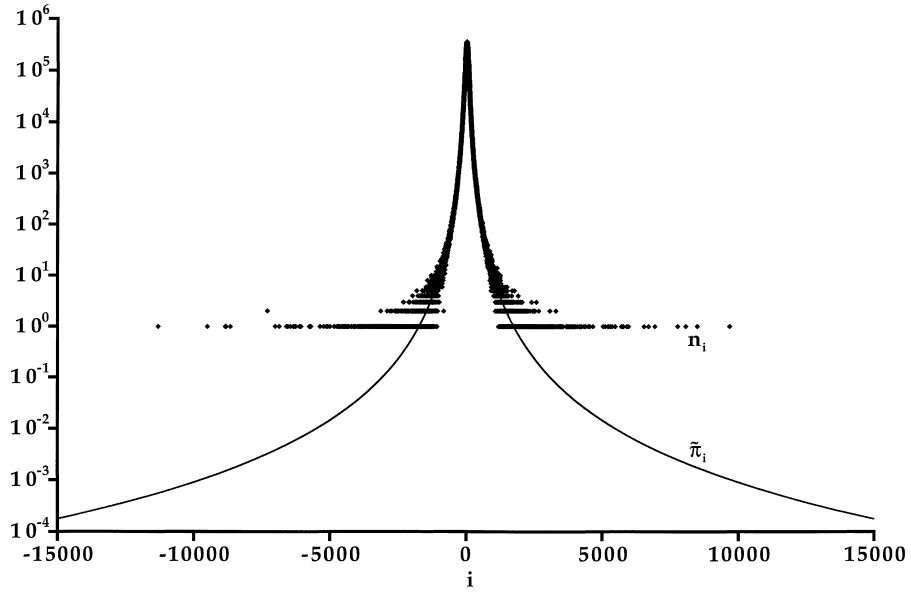

Fig. 2. The PDF $\{\tilde{\pi}_i\}$ (solid line) corresponding to Eq. (3) compared to an histogram $\{n_i\}$ generated from it and containing $N = 4 \times 10^7$ random trials. The bins which do not contain any event are not shown. Note that in the tails, the distance between non empty bins is of the order of $1/\tilde{\pi}_i$.

corresponds, on the average, to one bin containing one event ($n_i = 1$) surrounded by 1000 empty bins (Fig. 2). As a conclusion, the last histogram points still capture some information, even if it is not directly accessible. It is this information we want to use and reveal.

The PDF directly estimated by normalisation of the histogram (Eq. (1)) depends on the bin width Δ. How does its quality evolve when Δ is changed? Firstly, Δ should be sufficiently small to insure the validity of the second approximation (Eq. (2)). It is noteworthy that Eq. (2) requires the PDF be smooth at the scale Δ, i.e., its radius of curvature be everywhere much larger than Δ. On the other hand, the relative error made in the first approximation is of the order of $\tilde{\pi}_i^{-1/2}$ and thus decreases when the mean number of points in a given channel increases. This means that the second approximation requires a value of Δ as large as possible. A compromise on Δ has to be made. A usual good choice for Δ is around 1/100 of the PDF typical radius of curvature (which is usually close to the root mean square value) [4].

The conclusion that, at least for small values of Δ, the precision increases with Δ (for fixed N) is somehow paradoxical. Indeed, it is clear that the initial signal contains the whole available information (see Section 2.1). The best estimate of the PDF, which uses the whole information, should, thus, correspond to the limit when Δ tends to zero:

$$p(\gamma) = \sum_{n=1}^{N} \frac{1}{N} \delta(\gamma - \gamma(n\tau)) \tag{4}$$

This is in fact the best estimate of the PDF which *only* uses the signal information. However, the usual estimate (Eq. (1)) is better because the smoothness of the PDF $p(\gamma)$ over a scale Δ is implicitly used. To say this in another way, the smoothness of a PDF $p(\gamma)$ is an essential prior information which must be *added* to the information contained in the signal. On the other hand, the problem of the PDF tails, i.e., of the rare events fully remains since it corresponds to a region of the curve $p(\gamma)$ where at least one of the approximations made is a bad one. This induces a second paradoxical remark: to compute a PDF from a histogram using Eq. (1), the hypothesis that the PDF is smooth is implicitly used but the resulting estimate of the PDF is not smooth at all.

2.3. Regularisation methods

The aim of this article is, thus, to construct an alternative procedure to estimate the PDf by using *explicitly* its a priori smoothness. Our basic strategy is to fix a small bin width Δ in order to use safely the finite difference approximation of the derivative (Eq. (2)) and to concentrate on the estimate $\{\pi_i\}$ of the mean histogram $\tilde{\pi}_i$ from the real realisation $\{n_i\}$. To fix the notations used below, $\{n_i\}$ is the experimental histogram, $\{\pi_i\}$ is the real PDF (forgetting the factor $N\Delta$) and we want to compute an estimate $\{\pi_i\}$ of this PDF. We consider non-normalised histograms (in fact normalised to N but not divided by $N\Delta$, neither rescaled by the standard deviation) both because here the discrete nature of the histogram is of fundamental importance and because arrays of integers are easier to manipulate than small real numbers (and thus save computer time). After the smoothing procedure, the result $\{\pi_i\}$ can of course be normalised and rescaled.

In the particular example chosen here (Fig. 1), the PDF seems to be smooth in linear-logarithmic representation, whereas it presents a sharper maximum in linear–linear representation (not shown). This choice of considering the PDF logarithm is of course arbitrary, but related to our interest in the rarest events. We will, thus, consider in parallel to the PDF $\{\pi_i\}$, its logarithm $\{\alpha_i\}$:

$$\pi_i = \exp(\alpha_i) \tag{5}$$

Using $\{\alpha_i\}$ instead of $\{\pi_i\}$ has the strong advantage of imposing the positive sign of π_i so that this constraint has not to be specified explicitly. It will also simplify the requirement that the PDF logarithm $\{\alpha_i\}$ be smooth.

The natural methods to use the a priori information that $\{\alpha_i\}$ is smooth are the so-called regularisation methods (see [5] and references therein). The central idea of these methods is the minimisation of a functional $\varphi_T[\alpha_i]$ with respect to a set of unknowns $\{\alpha_i\}$. This functional $\varphi_T[\alpha_i]$ quantify the compromise between two extreme requirements:

- the PDF should be perfectly smooth
- the PDF should pass through all the data.

φ_T is thus defined as the sum of two positive functionals which have to be made explicit ($\varphi_T = \varphi_L + \lambda\varphi_S$). One, $\varphi_L[\alpha_i]$, measures the agreement between the data (here, the histogram $\{n_i\}$) and the model (here, the PDF $\{\exp(\alpha_i)\}$). We will call it the likelihood, which does not mean only that $\{\pi_i\}$ is close to the histogram $\{n_i\}$ but more precisely that $\{n_i\}$ can have occurred assuming that $\{\pi_i\}$ is the real PDF. The other, $\varphi_S[\alpha_i]$, reflects the smoothness of the solution. λ is the relative weight of one requirement with respect to the other and can thus be interpreted as a Lagrange multiplier. Finding the best solution corresponds to choosing a parameter λ which defines the 'best' compromise between *smoothness* and *likelihood*.

In our case, the PDF is submitted to an additional constraint of normalisation:

$$\sum_i \exp(\alpha_i) = \sum_i \pi_i = N \tag{6}$$

Using the method of Lagrange multipliers, the new functional φ_T to minimise is defined as

$$\varphi_T = \varphi_L + \lambda\varphi_S + \mu\varphi_N \tag{7}$$

where

$$\varphi_N[\alpha_i] \equiv \sum_i \exp(\alpha_i) \tag{8}$$

and where μ is the Lagrange multiplier which has to be adjusted to get $\varphi_N[\alpha_i] = N$.

We will now construct explicitly this regularisation method. The smoothing functional $\varphi_S[\alpha_i]$ will be defined in Section 3, the likelihood functional $\varphi_L[\alpha_i]$ in Section 4 and the criterion to adjust the Lagrange parameter λ in Section 5. This construction will be made on the basis of the experimental histogram shown in Fig. 1.

3. Construction of the smoothing functional

The role of the smoothing functional $\varphi_S[\alpha_i]$ is to specify the a priori belief that the PDF logarithm is smooth. All the candidates $\{\alpha_i\}$ to be this PDF logarithm are not equivalent: $\varphi_S[\alpha_i]$ must characterise their roughness and thus their *prior* probability to be the right one. One of the simplest possibility is to use a kind of bending energy of the curve α_i which has to be minimal. The local contribution to this bending energy depends quadratically on the curvature of the curve. Using the finite difference approximation of the second derivative, the smoothing functional reads

$$\varphi_S[\alpha_i] \equiv \frac{1}{2}\int \left(\frac{\mathrm{d}^2\alpha}{\mathrm{d}i^2}\right)^2 \mathrm{d}i \approx \frac{1}{2}\sum_i (\alpha_{i+1} - 2\alpha_i + \alpha_{i-1})^2 \tag{9}$$

This smoothing functional is local in the sense that it links one bin i only to its two neighbours. In fact, it can be generalised (and probably improved) by introducing a bending energy computed from the square modulus of a wavelet transform of the curve α_i. The wavelet can then be slowly decaying (and defined on a large number of bins), creating thus, a non-local smoothing constraint. Here, we use the basic discrete curvature (Eq. (9)) both for the sake of simplicity and because a local constraint allows to save much computing time. It should be kept in mind that the results can be straightforwardly extended to other forms of smoothing functionals.

In the mechanical analogy, the curve α_i to be determined is an elastic line attracted by the histogram points. We have now, in the likelihood functional, to specify the force field created by these points.

4. Construction of the likelihood functional

4.1. Construction

By analogy with the statistical negentropy, we consider for the likelihood functional $\varphi_L[\alpha_i]$ the opposite of the logarithm of the conditional probability that the experimental histogram $\{n_i\}$ would be observed given that $\{\pi_i\}$ is the true PDF:

$$\varphi_L = -\ln(\mathrm{Prob}_{\{\pi_i\}}[n_i]) \tag{10}$$

The problem is to model the real distribution $\mathrm{Prob}_{\{\pi_i\}}[n_i]$ in the case of a histogram. Let us suppose for the moment that the N values set apart from the signal correspond to N independent trials. As introduced above, if n_i is the number of events observed in the interval $[\gamma_i - \Delta/2, \gamma_i + \Delta/2]$ and $\tilde{\pi}_i$ its average over an infinite number of realisations, the probability that a point falls in the ith bin is $\tilde{\pi}_i/N$. If we suppose that $\{\pi_i\}$ is the true PDF $\{\tilde{\pi}_i\}$, the probability to obtain a particular realisation $\{n_i\}$ is given by the multinomial distribution [6]:

$$\mathrm{Prob}_{\{\pi_i\}}[n_i] \equiv \frac{N!}{N^N}\prod_i \frac{\pi^{n_i}}{n_i!} = \frac{N!}{N^N}\prod_i \frac{\exp(\alpha_i n_i)}{n_i!} \tag{11}$$

The negative of its logarithm can be decomposed in two parts, one which does not depend explicitly on $\{\pi_i\}$, and the other being the interesting part (the likelihood functional):

$$\varphi_L[\alpha_i] \equiv -\sum_i n_i\alpha_i \tag{12}$$

With the simple form taken by $\varphi_L[\alpha_i]$ in Eq. (12), the Lagrange multiplier μ can be determined explicitly. Indeed, the set $\{\alpha_i\}$ which minimises the functional φ_T verifies:

$$\sum_i \frac{\partial \varphi_T}{\partial \alpha_i} = \sum_i (\mu \exp(\alpha_i) - n_i) = N(\mu - 1) = 0 \tag{13}$$

As a consequence, μ is equal to 1 and the functional φ_T simplifies in:

$$\varphi_T[\alpha_i] \equiv \sum_i \left[\exp(\alpha_i) - n_i \alpha_i + \frac{\lambda}{2} (\alpha_{i+1} - 2\alpha_i + \alpha_{i-1})^2 \right] \tag{14}$$

It is interesting to note that the normalisation of the PDF $\{\pi_i\}$ to N is a nonlocal constraint. Contrarily to other smoothing filters which can be applied to the histogram, the regularisation method proposed here (see Section 3) makes this constraint local in the sense that the Hessian of the functional φ is a band diagonal matrix with bandwidth 5. Thus, using the Levenberg–Marquardt algorithm to perform the minimisation of the estimator φ_T, the locality of the potential to minimise allows to easily store and quickly invert its Hessian matrix.

4.2. Comparison with the 'chi-square' method

It should be noted that for quite high probability events ($N \gg \pi_i \gg 1$), the multinomial distribution of Eq. (11) can be normally approximated [6]:

$$\mathrm{Prob}_{\{\pi_i\}}[n_i] \propto \exp\left(-\sum_i \frac{(n_i - \pi_i)^2}{2\pi_i} \right) \tag{15}$$

The negative of its logarithm, i.e., the corresponding negentropy is then simply

$$-\ln(\mathrm{Prob}_{\{\pi_i\}}[n_i]) \approx \sum_i \frac{(n_i - \pi_i)^2}{2\pi_i} \tag{16}$$

which is Pearson's 'chi-square' formula [7]. However, for small probability events, ($\pi_i \ll 1$) the multinomial distribution exhibits a strong deviation from Gaussianity. For the purpose of estimating the PDF in the tails, it is very important to have captured this non-Gaussian feature which makes the difference with the usual 'chi-square' method [5]. Another empirical generalisation was also proposed by Holy [8].

4.3. Results

For a given Lagrange multiplier λ the solution of the minimisation verifies:

$$\frac{\partial \varphi_T}{\partial \alpha_i} = \lambda(\exp(\alpha_i) - n_i) + (\alpha_{i+2} - 4\alpha_{i+1} + 6\alpha_i - 4\alpha_{i-1} + \alpha_{i-2}) = 0 \tag{17}$$

To avoid any confusion, this solution will be denoted by $\alpha_i(\lambda)$ and the corresponding PDF by $\pi_i(\lambda)$. Fig. 3 shows the result of the minimisation on the example of the velocity gradient histogram introduced above (Fig. 1). The curve $\pi_i(\lambda)$ as a function of i is plotted for three values of the Lagrange multiplier λ. In the limit where λ tends to 0 (when the smoothing is not imposed) the estimated PDF $\pi_i(\lambda)$ remains close to the experimental histogram $\{n_i\}$. For a finite (but small) value of this parameter ($\lambda = 10^4$), the curve $\alpha_i(\lambda)$ is smoother but still feels the histogram points individually. However, the curve is now continuous. For a larger value ($\lambda = 10^8$), the peaks due to the separated points are completely smoothed. The large probability region of the histogram appears to be more resistant to the smoothing than the tails. These tails could be seen as rather arbitrary since they are below the scattered points corresponding to large values. This can be understood by looking again at Fig. 2. For larger values of λ (e.g., $\lambda = 10^{12}$) the histogram becomes even smoother but on a scale much larger than the curvature around 0 so that it

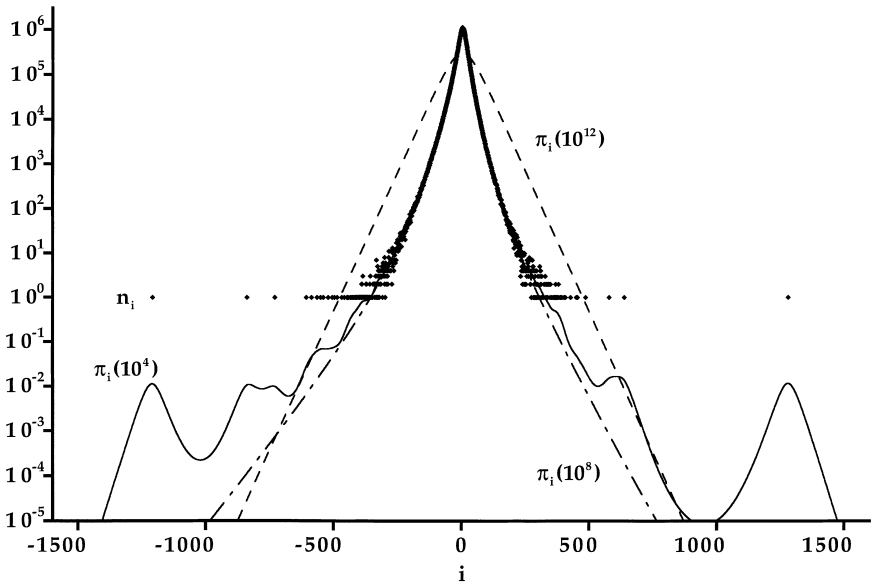

Fig. 3. The PDFs $\pi_i(\lambda)$ estimated from the regularisation of the initial histogram $\{n_i\}$ for three values of the Lagrange multiplier: $\lambda = 10^4$ (solid line), $\lambda = 10^8$ (dotted dashed line) and $\lambda = 10^{12}$ (dashed line).

deviates strongly from the real histogram $\{n_i\}$. The curve $\pi_i(\lambda)$ quickly tends towards exponential tails which are the natural asymptotes selected by the functional φ_T.

It can be seen from these observations that the regularisation procedure acts as a smoothing filter. By construction, $\exp(-\lambda\varphi_S[\alpha_i])$ can be interpreted here as the prior probability that $\{\alpha_i\}$ be the true PDF logarithm. This probability is Gaussian, with a variance equal to $1/\lambda$ the inverse of the Lagrange multiplier. The regularisation method can thus be seen as a low pass filter in curvature with a typical curvature cut-off of order $1/\sqrt{\lambda}$. However, it is better than a Fourier transform filtering for at least two reasons. Firstly, it automatically conserves the PDF normalisation. Secondly, the noise in the histogram tails cannot be removed by Fourier filtering since they are formed of localised peaks with a large band signature on the Fourier transform. It is worth noting that it induces drastic problems to compute PDFs convolution products using directly Fourier transforms. On the contrary, the method proposed here solves this problem by taking into account the specific non-Gaussian statistic in the PDF tails.

4.4. Adding further constraints

It interesting to note that further constraints on the PDF can be easily introduced. For instance, the average of a longitudinal velocity derivative should be zero (it is not the case on a finite signal, i.e., for a particular realisation). This can be easily introduced by adding to the total functional φ_T a new functional φ_A which specifies the constraint,

$$\varphi_A[\alpha_i] \equiv \int \gamma p(\gamma)\,\mathrm{d}\gamma \approx \frac{\Delta}{N}\sum_i \rho_i \gamma_i \tag{18}$$

and thus, a new Lagrange multiplier κ,

$$\varphi_T = \varphi_L + \varphi_N + \lambda\varphi_S + \kappa\varphi_A \tag{19}$$

As previously, the Lagrange multiplier κ has to be adjusted in order to verify the constraint $\varphi_A = 0$. Following this example, any other constraint can (and should) be added in the regularisation.

5. Construction of a criterion of selection for the lagrange multiplier

5.1. Construction

It can be seen in Fig. 3 that there exist between the smoothest curve ($\lambda = 10^{12}$) and the roughest one ($\lambda = 10^4$), a range of values of the Lagrange multiplier λ which gives a correct estimate of the PDF $\pi_i(\lambda)$. We have now to construct an objective criterion to select the 'best' value of the parameter λ. Basically, we want to get the smoothest estimate $\pi_i(\lambda)$ which does not contradict the hypothesis that $\{n_i\}$ 'can' have been generated if $\pi_i(\lambda)$ is the real PDF. In other words, we want to test if the dispersion of the histogram $\{n_i\}$ around the curve $\pi_i(\lambda)$ approximately corresponds to the dispersion predicted by the multinomial distribution. We chose among the possible criteria to use the statistical negentropy defined by Eqs. (10) and (11). Using the Stirling formula, it can be approximated by:

$$-\ln(\mathrm{Prob}_{\{\pi_i\}}[n_i]) \approx \sum_i n_i \ln\left(\frac{n_i}{\pi_i}\right) \tag{20}$$

The test consists in requiring this negentropy to be equal to the canonical negentropy. If we suppose that $\pi_i(\lambda)$ is the real PDF, the canonical negentropy can be defined as the average of the negentropy (Eq. (20)) over the possible realisations $\{n_i\}$. It can be written in the form:

$$\langle -\ln(\mathrm{Prob}_{\{\pi_i\}}[n_i])\rangle \equiv -\sum_{\{n_i\}} \mathrm{Prob}_{\{\pi_i\}}[n_i]\ln(\mathrm{Prob}_{\{\pi_i\}}[n_i]) = \sum_i T(\pi_i, N) \tag{21}$$

where $T(\rho, N)$ is the canonical negentropy considering only one bin of probability ρ/N and N independent trials, as previously. Since the number of points η in this bin is distributed according to the binomial distribution, $T(\rho, N)$ reads:

$$T(\rho, N) \equiv \left\langle \eta \ln\left(\frac{\eta}{\rho}\right)\right\rangle = \frac{N!}{N^N}\sum_{n=0}^{N}\frac{\rho^n(N-\rho)^{N-n}}{\eta!(N-\eta)!}\eta\ln\left(\frac{\eta}{\rho}\right) \tag{22}$$

We chose to define the test quantity χ_T as the ratio between the two negentropies defined by Eqs. (20) and (21):

$$\chi_T[n_i, \pi_i] = \frac{\sum_i n_i \ln(n_i/\pi_i)}{\sum_i T(\pi_i, N)} \tag{23}$$

The test is to require from χ_T to be equal to 1. Let us call $\chi_T(\lambda)$, the value taken by this test functional for the PDF $\pi_i(\lambda)$ and for the experimental histogram $\{n_i\}$ ($\chi_T(\lambda) = \chi_T[n_i, \pi_i(\lambda)]$). When λ tends to infinity, the estimated PDF $\pi_i(\lambda)$ tends towards the histogram $\{n_i\}$. By construction, the test quantity χ_T then tends to 0 ($\chi_T(+\infty) = 0$). The Lagrange parameter should thus be decreased from infinity down to the value λ_T for which the test quantity is equal to 1 ($\chi_T(\lambda_T) = 1$). The PDF $\pi_i(\lambda_T)$ is, from this point of view, the 'best' compromise between smoothness and likelihood.

5.2. Interpretation of the selection criterion

There are many ways of constructing a test function which, as χ_T, quantifies the likelihood of a PDF $\{\pi_i\}$ to the histogram $\{n_i\}$. Any selection criterion thus contains a part of arbitrariness. However the choice of the negentropy, which is at least a natural quantity in the probability field, can be justified by several interesting properties.

It should first be noted that for quite high probability events, the multinomial distribution can be normally approximated (Eq. (16)). In this limit, the χ_T test can be approximated by the Pearson 'chi-square' [2,3]:

$$\chi^2 = \frac{\sum_i ((n_i - \pi_i)^2 / 2\pi_i)}{\sum_i 1/2} \tag{24}$$

The χ_T test appears, thus, as the natural adaptation for the multinomial case of the 'chi-square' test. This adaptation is necessary as, in our case, the numerator of the 'chi-square' converge (n_i quickly becomes zero in the tails), while the denominator, which is simply the number of bins, tends to infinity. The 'chi-square' is thus here always zero ($\chi^2 = 0$). This is due to the strong non-Gaussianity of the multinomial distribution for rare events.

$T(\rho, N)$ is plotted in Fig. 4 (a) as a function of ρ for $N = 10^5$. It can be simply understood as the typical weight in χ_T of a bin which would have a probability ρ/N. In the case of the 'chi-square' this weight is independent of the probability and is equal to $1/2$ (dashed line). As predicted, for a large mean number of events ($1 \ll \rho \ll N$) the curve $T(\rho, N)$ appears to be almost constant and equal to 1/2. It has a maximum around $\rho = 1$ and tends to zero both when ρ tends to zero (as $-\rho \ln(\rho)$) and when ρ becomes of the order of the total number of point N (as $(1 - \rho/N)/2$). The negentropy (Eq. (21)) can, thus, be interpreted as an effective number of useful bins, the real number of bins being recovered if π_i is larger than 1.

The test quantity χ_T can be interpreted as a measure of the dispersion of the data around the candidate $\{\pi_i\}$. Assuming that the realisations are multinomially dispersed around the PDF $\{\pi_i\}$, we can compute for each bin the typical dispersion of any virtual realisation (Eq. (21)). Ideally, we should compare it to the dispersion for each bin over many real realisations of n_i. But in our case we only have one realisation of the experiment, the histogram $\{n_i\}$. In fact, the statistical negentropy (Eq. (17)) measures the gap between the candidate $\{\pi_i\}$ and our particular realisation $\{n_i\}$ by averaging over all the bins. The χ_T test can, thus, be interpreted as the condition that the multinomial dispersion around the candidate $\{\pi_i\}$ correspond to the real dispersion of $\{n_i\}$ averaged over the bins.

Whatever the real PDF $\{\tilde{\pi}_i\}$ can be, from a probabilistic point of view, any histogram $\{n_i\}$ containing N points can occur. The χ_T test can be understood as a reduction of these probabilities to only two values: all the histograms $\{n_i\}$ which verify $\chi_T[n_i, \pi_i] \leq 1$ are equiprobable, and all the others cannot be realisations of $\{\tilde{\pi}_i\}$. Reciprocally, the condition $\chi_T[n_i, \pi_i] \leq 1$ defines, from a known realisation $\{n_i\}$, a border for the possible $\{\pi_i\}$ that can have generated it. In the regularisation method, we start from a PDF ($\pi_i = n_i$ for $\lambda = 0$) which is automatically inside the region of possible PDFs ($\chi_T[n_i, n_i] = 0$). Smoother and smoother PDFs are computed until the border $\chi_T = 1$ be crossed.

We can check that this negentropy test ($\chi_T = 1$) is a good criterion by computing χ_T for many (real) realisations $\{n_i\}$ of a known PDF $\{\tilde{\pi}\}$. We then estimate the corresponding PDF of χ_T. In the case of the analytical PDF $\{\tilde{\pi}_i\}$ (Eq. (3)) shown in Fig. 2, the PDFs of χ_T are computed numerically for $N = 10^2$, $N = 10^5$ and $N = 10^8$ points and presented in Fig. 4(b). They are effectively centred on 1 and have, even for few points, standard deviations small compared to 1 (0.079 for $N = 10^2$, 0.031 for $N = 10^5$ and 0.014 for $N = 10^8$). This indicates that all the realisations of a known PDF $\{\tilde{\pi}_i\}$ nearly verify the criterion $\chi_T = 1$. Thus, a value χ_T in the vicinity of 1 is at least a necessary condition for a histogram to be a realisation of the PDF $\{\pi_i\}$. As an indirect but interesting consequence, it is then possible (and recommended) to use it to quantify the quality of any histogram fit. If the value of χ_T is far from 1 this means that, however nice looking it may be, the fit is not statistically realistic.

5.3. Results

For the example of the velocity gradient PDF (Fig. 1), $\chi_T(\lambda)$ is plotted in Fig. 5. It is a increasing function which tends to 0 when λ tends to 0. $\chi_T(\lambda)$ crosses 1 for $\lambda_T \approx 6.64 \times 10^6$. The corresponding wavelength cut-off is of

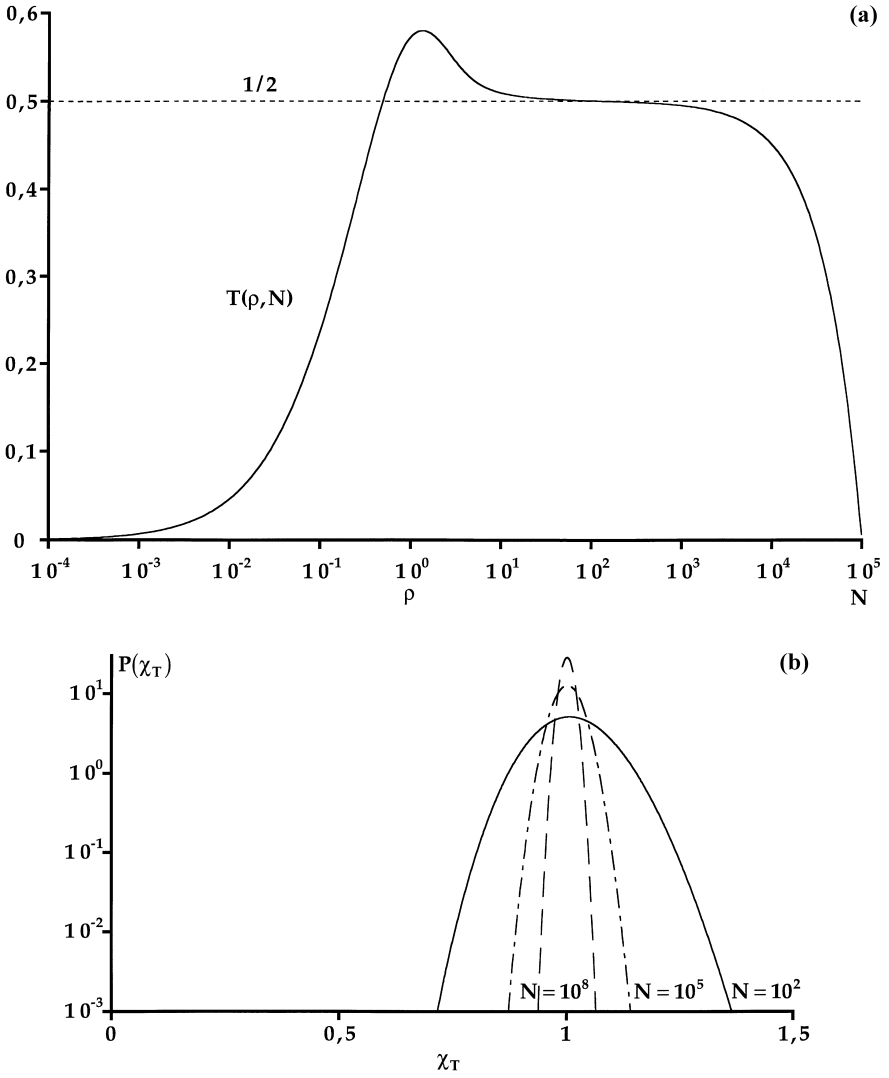

Fig. 4. (a) The canonical negentropy $T(\rho, N)$ as a function of ρ and computed for $N = 10^5$ independent trials. It corresponds to the mean weight in the χ_T test quantity of one bin which would have a probability ρ/N to occur. For a Gaussian statistic, the canonical negentropy (the weight in the 'chi-square') is always 1/2 (dashed line). (b) PDF of the value taken by the test quantity χ_T for histograms randomly generated from the PDF shown in Fig. 2 for $N = 10^2$ points (solid line), for $N = 10^5$ points (dotted-dashed line) and for $N = 10^8$ points (dashed line). Even with few points, the standard deviation is small compared to 1 (0.079 for $N = 10^2$, 0.031 for $N = 10^5$ and 0.014 for $N = 10^8$).

the order of $1.4\,\gamma_{\text{rms}}$, For λ larger than λ_T, the increase in λ is very rapid and becomes slower for smaller λ. The selected value λ_T approximately corresponds to the inflection point of the curve. This means that it does not 'cost' much in likelihood to filter all the histogram peaks in the PDF. On the other hand, increasing too much the Lagrange parameter λ increases drastically χ_T since it moves the whole curve $\pi_i(\lambda)$ away from the real PDF (see Fig. 3). The shape of the curve $\chi_T(\lambda)$ also indicates that the selection value (initially 1 (dashed line)) can be slightly increased without serious consequences.

We investigated for realisations $\{n_i\}$ of a given PDF $\{\tilde{\pi}_i\}$, the mean value of λ_T as a function of Δ: it appears to scale approximately as Δ^{-2} for small values of this parameter. This first confirms that $\Delta\lambda_T^{-1/2}$ is directly related

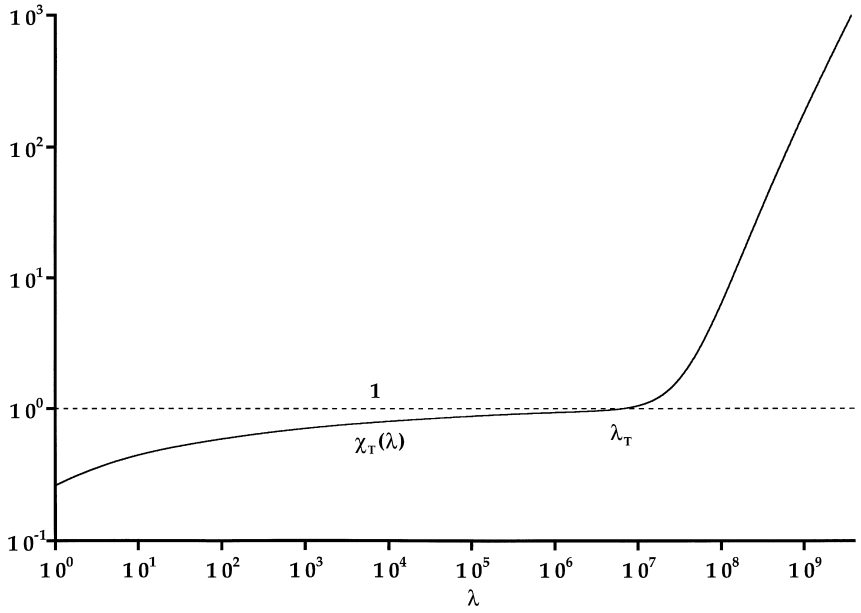

Fig. 5. The evolution of the test quantity $\chi_T(\lambda)$ as a function of the Lagrange multiplier λ for the histogram shown in Fig. 1. The optimal value of the test, 1 (dashed line), is obtained for $\lambda_T \approx 6.64 \times 10^6$.

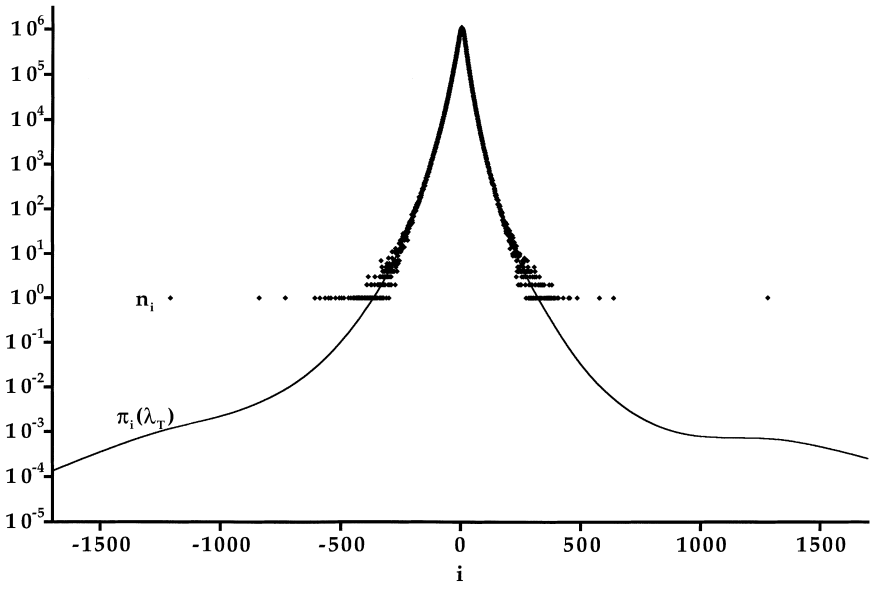

Fig. 6. The velocity gradient histogram $\{n_i\}$ and the corresponding optimal PDF estimate $\pi_i(\lambda_T)$ as functions of the bin number i.

to a typical curvature of the PDF logarithm. Moreover, the results are roughly independent of Δ provided that this parameter be small enough (see Sections 2.1 and 2.2).

The 'best' estimate of the PDF $\pi_i(\lambda_T)$ is shown in Fig. 6 together with the initial histogram $\{n_i\}$. It is, as required, a smooth function which interpolates, at least by eye, the histogram.

6. On the possible deviations to the multinomial model

It is in fact surprising that the PDF $\pi_i(\lambda_T)$ shown in Fig. 6 'can have generated' the histogram $\{n_i\}$ according to the multinomial distribution. There are in fact many reasons for which an experimental histogram could be over-dispersed, by comparison to the ideal multinomial case. It should first be noted that the noise in the measurement of the signal essentially convolves the PDF and thus has a tendency to smooth the PDF rather than to make it rougher. The noise is thus a source of error but not a source of dispersion of the histogram.

A second experimental problem which can be encountered is a drift of the experiment control parameters in particular if the total record time $N\tau$ is too large. There can be in this case a slow change of the real PDF during the experiment, and thus strictly speaking no global probability distribution $\{\tilde{\pi}_i\}$. The natural model for this drift is the so called 'Poisson case' for N independent trials. For a given bin i, the dispersion $\langle(n_i - \tilde{\pi}_i)^2\rangle$ is affected by an additive term of order $\delta\tilde{\pi}_i^2/N$, where $\delta\tilde{\pi}_i^2$ is the variance of the $\tilde{\pi}_i$ drift. This corresponds to a relative over-dispersion of order $\delta\tilde{\pi}_i^2/(N\tilde{\pi}_i)$ which should not affect strongly the regularisation method in most of the cases.

The third problem encountered regularising experimental histograms is even more serious: it is linked to the physical system studied and to the data sampling. Indeed, the sampling time τ is in general chosen much smaller than the typical correlation time, because one wants both to access to the small time scale features and to get a number N of sampled values as large as possible. For the example of the velocity gradient presented in Fig. 1, the signal is smooth and thus strongly correlated at the sampling time scale. This means that there is a redundancy in the data. This is problematic since the multinomial distribution requires that the data samples be completely independent and uncorrelated.

There is no way to escape from this over-sampling problem. However, it has different consequences depending on the bin width Δ. This effect is illustrated schematically in Fig. 7 where we consider a smooth signal. On the left, the bin width Δ is large so that when one event falls in the ith bin, there is a strong probability that the next event also falls in the ith bin. On the right, Δ is smaller so that the probability that the next event falls *exactly* in the ith bin is not much increased. In fact, when one event falls in the ith bin there is a strong probability that the next one falls in a bin *close* to it. Only looking at one given bin, each trial can be considered as independent from

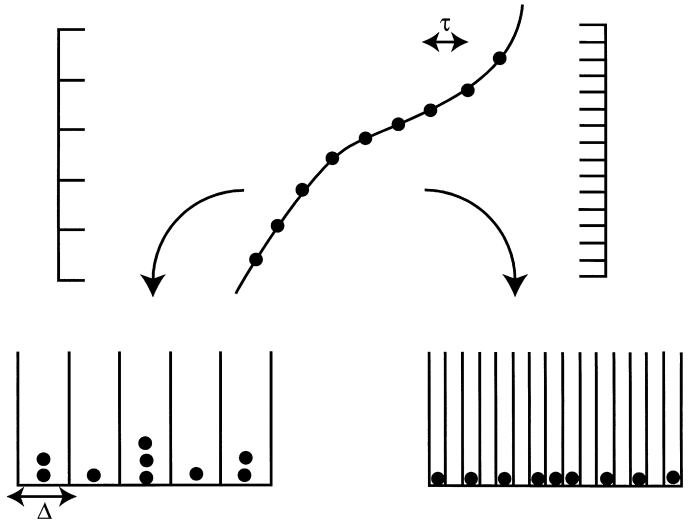

Fig. 7. A schematic of the over-sampling effect. If the signal is correlated at the scale of the sampling time interval τ, there is a strong probability that two consecutive points fall in *neighbouring* bins (left histogram). It is better in this case to chose a small bin width Δ so that two correlated points will rarely fall in the *same* bin (right histogram).

the previous one: the number of points in this bin is reasonably given by the binomial distribution. However, there should be a drastic effect on the covariance $\langle (n_i - \tilde{\pi}_i)(n_j - \tilde{\pi}_j)\rangle$. For the multinomial distribution it is a negative quantity (equal to $-\tilde{\pi}_i\tilde{\pi}_j/N$): the negative sign corresponds to the fact that if an event falls in the ith bin, it cannot also be in the jth one. On the other hand, for an over-sampled histogram, this quantity should be positive if i and j are close to each other: if there are too much points ($n_i > \tilde{\pi}_i$) in the ith bin, there is a strong probability that it will be the same in the neighbouring bins. This will have some consequences on the PDF moments estimate (see Section 4.3). As a conclusion, there are two simple solutions which can be tried if the histogram appears to be really over dispersed: decreasing the bin width Δ and increasing the sampling time τ.

The fourth problem we want to discuss can also explain an over-dispersion. It may appear when an analog-digital converter is used to acquire the signal. If the number of quantified levels by bin is too small, the quantification of the signal can appear in the histogram. In this case, the apparent width Δ of a given bin can quantitatively differ from the real width which takes into account the signal quantification. This problem can be corrected either by increasing the bin width Δ or paradoxically by adding a random noise to the sampled signal: adding a random real number between -0.5 and 0.5 to all the integer values allows in general to overcome the problem.

7. On the tails of the estimated probability distribution function

7.1. Test on a synthetic PDF exhibiting algebraic tails

If the validity of the regularisation procedure is quite clear in the large probability region of the PDF (Fig. 6), the tails could at first sight be seen as rather arbitrary. In particular, the smoothing potential selects exponential asymptotes. This arbitrariness is intrinsically due to the principle of the method which adds a priori some information on the nature of a PDF. We will thus investigate the estimated tails on the synthetic example of Fig. 2 for which the real PDF $\tilde{\pi}_i$ is known.

The PDF estimate $\pi_i(\lambda_T)$ is shown on Fig. 8 (a). It is a continuous curve with, however, some remaining oscillations *around* the real PDF $\tilde{\pi}_i$. This demanding example clearly shows that the estimated PDF tails are not at all arbitrary extrapolations but still reflect the local point density. Moreover, it can be seen that the exponential asymptotes only start after the last histogram points.

We also observe on Fig. 8(a) the main limit of the smoothing functional. To give a visual interpretation of the parameter λ_T, an arc of parabola of same curvature is drawn above the histogram. It appears to be clearly too large compared to the curvature in the central part of the histogram. However, it corresponds well to the slight oscillations remaining in the tails. This means that the central part of the histogram is too much constrained. In the central region, the curve has the highest resistance to smoothing : the curve $\pi_i(\lambda)$ cannot be moved away from the large histogram points. As a consequence, increasing λ, the test function becomes larger than 1 because of the discrepancy in this central region. As a consequence, the estimated PDF remains rough in the tails where the curvature is lower. The regularly spaced bins together with the basic smoothing functional thus lead to some problems as the curvature is not constant along the histogram (Fig. 8(a)). This suggests to make a change of variable and more generally to adapt the smoothing functional in order to obtain a more regularly dispatched curvature and then to regularise completely the tails.

The positive counterpart of this problem is that a ‘corrupted’ prior information cannot be forced (contrarily to a fit) since the estimated PDF has to be plausible (selection criterion). The example (Fig. 8(a)) shows that the histogram cannot be over-smoothed and that its effect is easy to diagnose. This observation emphasises the fundamental difference between this regularisation method and a fit, which always gives by construction a result corresponding to the attempt. This again shows the interest of the criterion $\chi_T = 1$ (and more generally of chi-square like quantities) to check the likelihood of a fit to the data.

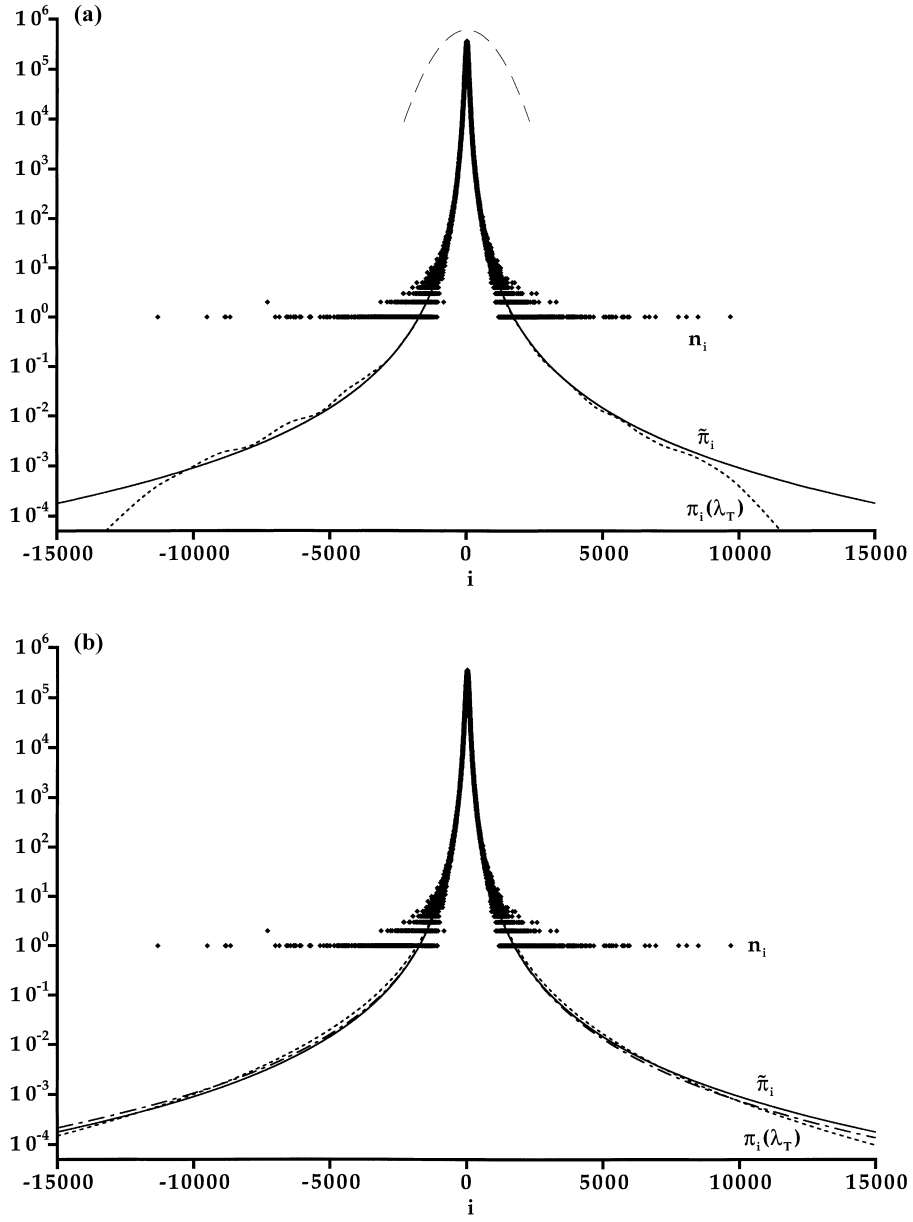

Fig. 8. The PDF and the synthetic histogram of Fig. 2. (a) The optimal PDF estimate $\pi_i(\lambda_T)$ (dashed line) using the variable i for the smoothing potential. The parabola corresponds to the curvature cut-off of the filter. (b) The PDF estimates $\pi_i(\lambda_T)$ for two changes of variable: one with a power law asymptote (dashed line) and the other with a logarithmic asymptote (dotted dashed line).

7.2. Change of variable and generalisation of the smoothing functional

A first possible modification is to change the variable used to compute the smoothing functional (Eq. (9)). Even if the natural coordinate is that chosen to compute the histogram, another coordinate $x(i)$ may be easily introduced and φ_S modifies in:

$$\varphi_S[\alpha_i] \equiv \frac{1}{2}\int \left(\frac{\mathrm{d}^2\alpha}{\mathrm{d}x^2}\right)^2 \mathrm{d}x \approx \frac{1}{2}\sum_i \frac{1}{x'(i)^3}\left[(\alpha_{i+1} - 2\alpha_i + \alpha_{i-1}) - \frac{x''(i)}{2x'(i)}(\alpha_{i+1} - \alpha_{i-1})\right]^2 \tag{25}$$

Compared to the previous definition (Eq. (9)), this modified smoothing functional has two interests. It allows to weight the terms in the sum and thus to selectively smooth the different parts of the histogram. Following the mechanical analogy, this weight corresponds to a distribution of elasticity along the curve. For the two examples discussed above (Figs. 6 and 8(a)), the tails have to be stiffened: this corresponds to a derivative $x'(i)$ which should decrease from the centre to the tails. The other additive term can be interpreted as a reference curvature. The null bending energy no longer corresponds to an exponential in i but to an exponential in $x(i)$. In other terms, the elastic line is no more straight at rest (when no external force field is applied to it) but already has a nontrivial shape specified by $x(i)$. These observations suggest a generalisation of the smoothing potential which keeps the two positive aspects of the change of variable:

- in order to improve the PDF tails, they have generally to be more smoothed than the high probability regions;
- the curvature should be compared to a reference which roughly describes the prior shape of the PDF.

Introducing a smoothness weighting w_i and a curvature reference c_i, the smoothing functional can then be written in the general form:

$$\varphi_S[\alpha_i] = \frac{1}{2}\sum_i w_i[(\alpha_{i+1} - 2\alpha_i + \alpha_{i-1}) - c_i(\alpha_{i+1} - \alpha_{i-1})]^2 \tag{26}$$

The difference with the simple change of variables is just the independence of the weight and the curvature reference. Some criteria are now needed to determine objectively w_i and c_i.

7.3. Guidelines for the generalised smoothing functional selection

A first guideline originates in the difference highlighted on Fig. 8(a) between the regularisation method (or a series of local fits) and a global fit. Namely, if the smoothing procedure is not adapted to the histogram considered, then the result is not good. This means that the result does not correspond to the prior assumptions made. In Fig. 8(a), this leads to remaining oscillations in the tails together with a sudden change of behaviour around the last points. This suggests an intuitive self-consistency criterion: if the result does not exhibit the properties assumed, this means that these assumptions are not good. On the contrary, we can expect that if the hypotheses made are finally verified by the estimate, they are reasonable.

A second test to check the validity of the prior assumptions is the stability of the method. Once a PDF is estimated, it can be used to produce several realisations (Monte Carlo method see [5]), which can in turn leads to estimated PDFs, and so on and so forth. The stability of the procedure can be investigated by looking at the difference between the successive estimated PDFs and the first one.

The next step in this work is to use these principles (self-consistency and stability) to build a self-adaptive procedure to determine for each histogram the 'best' smoothing functional (the optimal weight w_i and curvature reference c_i), without adding by hand any information. In particular, it would allow to get rid of the prior choice of variable. We can imagine an iterative procedure which starts from the basic smoothing functional ($c_i = 0$ and $w_i = 1$) to obtain a first PDF estimate $\{\pi_i^1\}$. The next step is to determine the 'best' parameters w_i and c_i if this PDF $\{\pi_i^1\}$ is the real one. As explained above, $\{\pi_i^1\}$ allows to generate 'false' realisations, each leading to other estimates $\{\pi_i'^1\}_k$. We can for instance ask the average of these estimates to be equal to the first one $\{\pi_i^1\}$. We can also try to minimise the standard deviation of these estimates around $\{\pi_i^1\}$ with respect to the set of parameters w_i

and c_i. If this is achieved, a second estimate $\{\pi_i^2\}$ can be computed from the initial histogram and the algorithm can be iterated.

However, there are still some difficulties (due for instance to the nontrivial selection of the Lagrange multiplier λ, to the attraction towards the stable estimate $\pi_i = 0$ if $n_i = 0 \ldots$), and the work has still to be done. Before this general method, we can however use the criteria defined above to determine semi-empirically a working procedure, for instance in the particular case of turbulent-like signals. After having shown the effect of a simple change of variables (Section 7.4), a working smoothing functional will be presented (Section 7.5).

7.4. Test of the change of variable

The PDF shown in Fig. 8(a) (Eq. (24)) exhibits algebraic tails: the best choice for the change of variable should thus have a logarithmic asymptote. It should also be linear around the PDF maximum. Using a change of variable which verifies these two constraints, the regularisation method gives a PDF estimate $\pi_i(\lambda_T)$ (Fig. 8(b), dotted dashed line) which is smooth everywhere and which nearly collapses the real PDF $\tilde{\pi}_i$. By construction of the change of variable, the smoothed histogram $\pi_i(\lambda_T)$ exhibits far algebraic tails.

In order to investigate the effect of this asymptote, we tested a change of variable which is also linear around the position of the PDF maximum but which exhibits a 0.5 power law asymptote ($x(i) \propto i^{0.5}$). The corresponding PDF estimate $\pi_i(\lambda_T)$ is shown by the dashed line in Fig. 8(b). It is again a smooth curve which is also very close to the exact PDF (solid line in Fig. 8(b)). It exhibits by construction a stretched exponential asymptote which starts only *after* the last histogram points. Thus, there is only a slight dependence on the change of variable: the two results with algebraic and stretched exponential asymptotes are close to each other and to the original PDF up to the last point. So we could say that even with a selected asymptote in stretched exponential, an algebraic behaviour is recovered with the smoothing procedure. This was already obtained in Fig. 8(a), which shows that even with exponential asymptotes the algebraic tails are reasonably well approximated. A fortiori, distributions close to a Gaussian or an exponential, which are particular stretched exponentials, can be well fitted up to the last point.

As a conclusion, if the selected asymptote fits very badly the histogram tails, there remains some oscillations in the PDF tails. These tails consequently exhibit statistically larger deviations to the real PDF. Reciprocally *any* reasonable change of variable (and it appeared to be a loose condition) allows to estimate the tails up to the last histogram points.

7.5. Working regularisation method

An interest of the change of variable is, as noted above, to assign naturally a weight to each bin in the smoothing potential. This weight is directly related to the change of variable $x(i)$, which is itself determined by the PDF curvature in the tails. If $\{\alpha_i\}$ is concave, $x'(i)$ will decrease from the centre and the tails will be more smoothed than with the simple method. If $\{\alpha_i\}$ is convex, the tails will be less smoothed than with the variable i. Checking this property on various synthetic PDFs, we observed that the tails also require to be more smoothed in this case. The simple change of variable is thus only a partial improvement of the tails estimate, and the generalised smoothing function should be used.

We introduce here a parametrised curvature reference. The basic regularisation method (Eqs. (14) and (23)) is applied to get a first idea of the PDF $\{\pi_i^1\}$. We determine from this initial estimate $\{\pi_i^1\}$, the curvature reference under the form $c_i = \theta_\pm/(i - i_0)$ where i_0 is the position of the PDF maximum and θ_+ (respectively θ_-) is obtained by minimisation of the curvature (Eq. (26) with $w_i = 1$) in the positive tail (respectively the negative one). An algebraic tail correspond to $\theta_\pm = -0.05$. The other values of $\theta_\pm$ corresponds to stretched (or compressed) exponential tails ($\theta_\pm = 0$ for the exponential and $\theta_\pm = 0.5$ for the Gaussian).

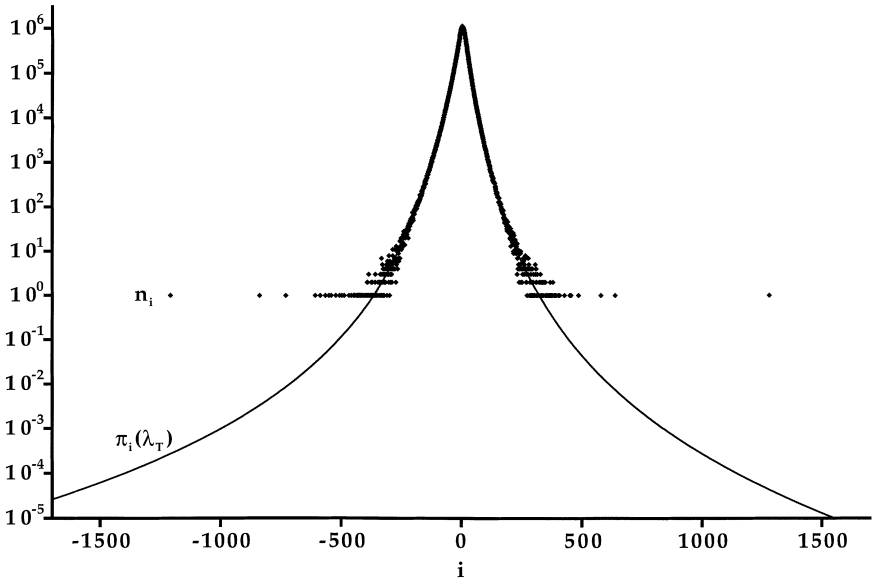

Fig. 9. The velocity gradient histogram $\{n_i\}$ and the corresponding PDF estimate $\pi_i(\lambda_T)$ using the modified smoothing functional.

The weight w_i has to be adjusted to prevent two extreme effects. If a region has locally a too low weight (a large flexibility), some oscillations remain (Figs. 6 and 8(a)). On the contrary, if the weight is too large (a very rigid curve), the exponential behaviour (in i or $x(i)$) is forced. If this exponential asymptote is locally a good fit of the PDF, the estimate becomes very good (Fig. 8(b)). But if it is not the case, the curve $\{\alpha_i\}$ 'breaks' into portions of lines joined in nearly singular places around isolated points, which concentrate the essential of the bending energy. There is also in this case a strong dependence of the resulting curve on the particular starting histogram $\{n_i\}$ (breaking of the stability criterion). The best compromise we could achieve uses a second time the basic regularised PDF $\{\pi_i^1\}$. We construct a weight which is almost constant in the large probability region ($\alpha_i > 0$) and increases in the tails as α_i^2:

$$w_i = \left(\ln\left(\frac{1+2\pi_i}{\pi_i}\right)\right)^2 = (\ln(2+\exp(-\alpha_i)))^2 \tag{27}$$

It should be noted that the weight is computed once for all with the initial PDF estimate and must not be minimised in φ_T. The final velocity gradient PDF, estimated with this complete regularisation method (including the null average (see Section 4.4)), is plotted in Fig. 9. This new estimate collapses with the previous one (Fig. 6) in the central region but the oscillations due to the two largest points have been smoothed. Finally, the results become self-consistent and reasonably stable, and the test χ_T, which is strictly verified for the whole histogram (by construction), is also approximately verified locally.

8. Benefits of the regularisation method

8.1. Improvement of the PDF estimate

We will now test the efficiency of the complete regularisation method constructed in the previous parts. For this purpose, synthetic histograms were generated from a known PDF . We chose to use the velocity gradient PDF shown in Fig. 9 as this real PDF (Monte Carlo method, see Section 7.3 and [5]). One thousand histograms were generated

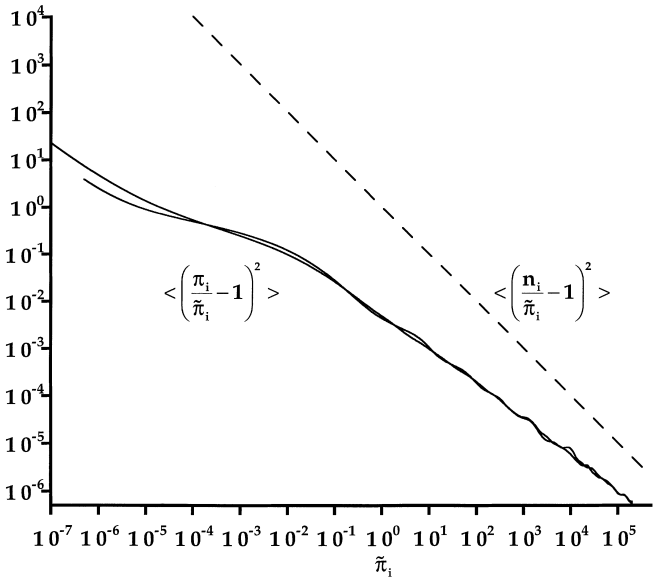

Fig. 10. The relative error statistically made on the PDF estimated by direct normalisation of the histogram ($\langle(n_i/\tilde{\pi}_i - 1)^2\rangle$, dashed line) and using the regularisation method ($\langle(\pi_i/\tilde{\pi}_i - 1)^2\rangle$, solid line). The two curves correspond to the two sides of the PDF.

using the multinomial distribution with the same number of points $N \approx 3.7 \times 10^7$ as initially. The regularisation method was applied to each of these histograms and gives thus 1000 estimated PDF $\pi_i(\lambda_T)$ (noted π_i below for simplicity). We computed the average over the 1000 trials of the quantities $(n_i/\tilde{\pi} - 1)^2$ and $(\pi_i/\tilde{\pi}_i - 1)^2$ for each bin. $\langle(n_i/\tilde{\pi}_i - 1)^2\rangle$ and $\langle(\pi_i/\tilde{\pi}_i - 1)^2\rangle$ are, respectively, the relative statistical error on $\{\tilde{\pi}_i\}$ estimated by direct normalisation of the histogram and using the regularisation method. The resulting functions are both plotted in Fig. 10 as a function of π_i (see also Fig. 9 for the correspondence between i and π_i). The two sides of the PDF correspond to the two curves (in solid line) shown in Fig. 10.

The precision of the PDF obtained using the regularisation method is much better than the direct normalisation of the histogram $\{n_i\}$, and that even in the tails (for $\tilde{\pi}_i < 1$). For the example chosen, a 10% uncertainty on $\tilde{\pi}_i$ ($\langle(\pi_i/\tilde{\pi}_i - 1)^2\rangle = 10^{-2}$) is obtained for $\pi_i \approx 0.4$ with the regularisation method and for $\pi_i = 100$ with the direct histogram normalisation. Similarly, the limit of 100% error ($\langle(\pi_i/\tilde{\pi}_i - 1)^2\rangle = 1$) is reached for $\pi_i \approx 10^{-5}$ with $\pi_i(\lambda_T)$ and for $\pi_i = 1$ with n_i. The change of slope of the statistical error $\langle(\pi_i/\tilde{\pi}_i - 1)^2\rangle$ corresponds to the region where the last histogram points are statistically observed. As a conclusion, the PDF $\{\pi_i(\lambda_T)\}$ is a very good estimate of the real PDF $\{\tilde{\pi}_i\}$ up to the last histogram point (the rarer event).

8.2. Improvement of the PDF moments estimate

If it is clear that the estimate of the PDF is improved by the regularisation method, it is less evident that there can be a gain on the computation of PDF moments. Indeed, a moment is a quantity which uses all the histogram bins and which is thus less sensitive to the dispersion of the data than the PDF itself.

If we consider an estimate $\{\rho_i\}$ of the PDF (this can be, for instance, directly the histogram $\{n_i\}$ or the regularised PDF $\{\pi_i(\lambda_T))\}$, the PDF moment of order q, $M[q, \tilde{\pi}_i]$, can be approximated, if the bin width Δ is sufficiently small, by:

$$M[q, \rho_i] \equiv \int \gamma^q p(\gamma)\, \mathrm{d}\gamma \approx \frac{1}{N}\sum_i \rho_i (\Delta\gamma_i)^q \tag{28}$$

As in the previous part, the statistical error $\sigma^2[q, \rho_i]$ on these moments, defined as

$$\sigma^2[q, \rho_i] \equiv \left\langle \left(\frac{M_q[\rho_i]}{M_q[\tilde{\pi}_i]} - 1 \right)^2 \right\rangle \tag{29}$$

will be used to quantify the improvement. As in the previous part, the simplest method is to generate a large number of 'false' realisations of the PDF estimate (Fig. 9) and to compute these errors by averaging over this synthetic set. At this step, it is important to note that the statistical error $\sigma^2[q, \rho_i]$ is directly related to the $\{\pi_i\}$ covariant matrix:

$$\sigma^2[q, \rho_i] = \frac{\left[\sum_{i,j} (\gamma_i \gamma_j)^q \langle (\rho_i - \tilde{\pi}_i)(\rho_j - \tilde{\pi}_j) \rangle \right]}{\left[\sum_i \gamma_i^q \tilde{\pi}_i \right]^2} \tag{30}$$

This expression brings us back to the discussion of Section 4.2. Indeed, $\sigma^2[q, n_i]$ (for the basic estimate $\{\rho_i\} = \{n_i\}$) directly depends on the dispersion of the histogram and thus on the possible redundancy in the sampling. There is thus potentially a difference between the 'multinomial' variance (denoted $\sigma^2_{\text{mul}}[q, \rho_i]$ below) computed from synthetic histograms and the real experimental one, $\sigma^2_{\text{exp}}[q, \rho_i]$.

A possible trick consists of computing from the signal m histograms which contain N/m points, and this for several values of m. $m = 1$ corresponds to the initial histogram. We used for commodity powers of two for both N and m ($m = 1, 2, \ldots, 128$). For each subdivision, the sum of the m histograms is strictly equal to the initial N point histogram. We can thus compute m estimates of the PDF moments based on N/m points histograms, and this, for each value of m. $M[q, \rho_i, m, j]$ denotes the jth estimate (over m) of the moment of order q (Eq. (28)). The variance on this moment, considering realisations of N/m points is denoted by $S^2[q, \rho_i, m]$ (with the interesting relation $S^2[q, \rho_i, 1] = \sigma^2_{\text{exp}}[q, \rho_i]$). This variance can be approached by an averaging over the m realisations considered:

$$S^2[q, \rho_i, m] \approx \frac{1}{m-1} \sum_{j=1}^{m} \left(\frac{m M_q[\rho_i, m, j]}{\sum_{k=1}^{m} M_q[\rho_i, m, k] - 1} \right)^2 \tag{31}$$

We obtain finally the variance on the PDF moment of order q, as a function of the number of points used to compute it. This variance is plotted in Fig. 11(a) for the fourth order moment, as a function of m, for both $\{n_i\}$ (black diamonds) and $\{\pi_i\}$ (black squares). $S^2[4, n_i, m]$ is everywhere larger than $S^2[4, \pi_i, m]$. In order to turn to the interesting quantity $\sigma^2_{\text{exp}}[q, \rho_i]$ we have to extrapolate the curve $S^2[q, \rho_i, m]$ to $m = 1$. As for multinomial trials, $S^2[q, \rho_i, m]$ appears to be proportional to m for both $\{n_i\}$ and $\{\pi_i\}$. This allows to fit objectively the experimental points (dashed lines) and finally to compute the statistical errors $\sigma^2_{\text{exp}}[q, \rho_i]$.

The four curves $\sigma^2_{\text{exp}}[q, n_i]$, $\sigma^2_{\text{exp}}[q, \pi_i]$, $\sigma^2_{\text{mul}}[q, n_i]$ and $\sigma^2_{\text{mul}}[q, \pi_i]$ are shown on Fig. 11(b). The diamonds correspond to the basic estimate $\{\rho_i\} = \{n_i\}$ and the circles to the regularisation method $\{\rho_i\} = \{\pi_i\}$. The black points are those measured experimentally by subdivision of the histogram and the white ones are computed by averaging over the synthetic multinomial trials. In both cases, there is an improvement of the PDF moments in using the regularisation method which allows to compute higher moments than what is directly possible. For instance, the limit of 75% error is reached for $p \approx 5$ for $\{n_i\}$ and $p \approx 7$ for $\{\pi_i\}$. For orders q lower than 3 (in the high probability region of the PDF), the curves collapse two by two: the two methods are there strictly equivalent. In the same zone, we observe that the experimental error is much larger than in the strict multinomial case. This is the confirmation that there is a quite high redundancy in the sampled data (see Section 6). However, for higher orders the regularisation method seems to be insensitive to this effect. This also means that the variance of the high order moments estimated from the regularised PDF may be precisely computed by the Monte Carlo method developed above.

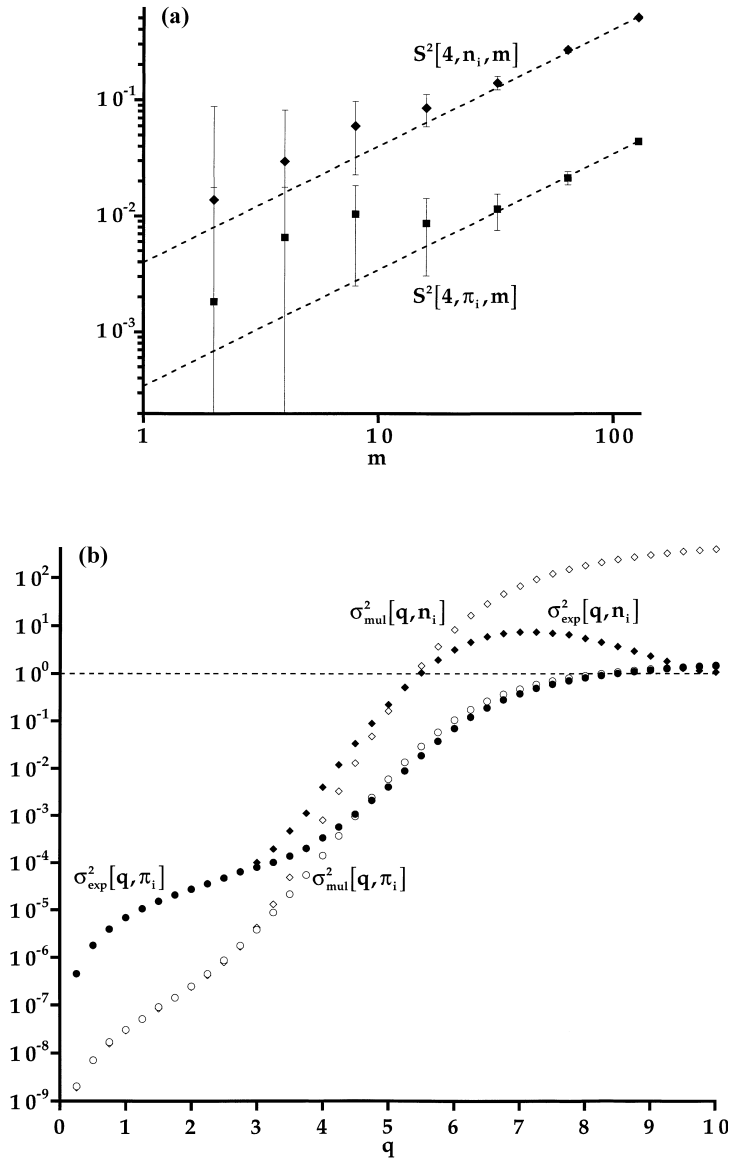

Fig. 11. (a) The relative error statistically made on the PDF fourth moment by direct normalisation of the histogram ($S^2[4, n_i, m]$, black diamonds) and using the regularisation method ($S^2[4, \pi_i, m]$, black squares) for sub-histograms containing N/m points. The shift between the two curves indicates an improvement of the moment estimate by the regularisation method. The fit by a linear function of m allows to measure the statistical error for the complete histogram (which contains the whole N points). (b) The relative statistical error on the q PDF moments as a function of the order q using the direct normalisation (diamonds) and the regularisation method (squares) on both the experimental histogram (black points) and synthetic multinomial trials (white points).

9. Concluding remarks

The estimate of a PDF $\{\pi_i\}$ from an experimental histogram $\{n_i\}$ was investigated. A specifically designed regularisation method was constructed to take advantage of the PDF smoothness. The PDFs estimated this way are precisely defined up to the last point of the histogram tails (the rarest event). Using this method, the statistical

errors approximately correspond to what was obtained classically with an experimental sampling time a hundred times longer [1]. This clearly demonstrates the importance of using prior information on the nature of the PDF (smoothness, null moments, inequalities between PDFs, etc.). The regularisation procedure presented here takes into account the smoothness of the PDF logarithm and can be easily extended to any other constraint.

Our method assumes that the dispersion of histograms around a PDF follows the multinomial distribution. This allowed to discuss some experimental problems (over-sampling, discretisation due to analog-digital converters. . .) which induce an over-dispersion of histograms. The regularisation method is thus also a very useful tool to 'inspect' experimental histograms. The essential point is the χ_T test (Eq. (23)) developed to characterise a histogram dispersion. It is for instance an objective criterion to test some PDF models (fitted using $\varphi_L + \varphi_N$ as defined above instead of the usual chi-square).

It is noteworthy that this smoothing procedure does not correspond to a fit. Although the first simple form (Eq. (9)) is improved using a parametrised change of variable (Eq. (25)) and more generally a change in the smoothing functional Eq. (26), this smoothing procedure still fundamentally differs from a simple fit. It can be understood as a series of local fits, which local form imposes a loose enough constraint so that a different dependence can be globally recovered. Another evidence of this is that if the guess used is really wrong, it is directly shown by the result which is then not consistent with the guess. This self-consistency criterion is very useful to choose the form of the smoothing functional. More work is still needed to take profit from stability and self-consistency criteria to construct a general regularisation method powerful in any cases.

Finally, the precision in the computation of the PDFs moments is improved using the regularisation method and the statistical errors can be correctly estimated. However, we want to emphasise that models directly on PDFs shapes would be easier to test objectively (with χ_T test for instance) than models on the moments. In a forthcoming article, we will present the results obtained on experimental and numerical turbulence signals using this procedure.

Acknowledgements

We would like to thank Y. Couder for his suggestions and P. Tabeling, F. Belin and H. Willaime for the useful discussions we had. Finally, we thank D. Sornette for his suggestions on further improvements of the method.

References

[1] A. Arneodo et al., Structure functions in turbulence, in various flow configurations, at Reynolds number between 30 and 5000 using extended self similarity, Europhys. Lett. 34 (1996) 411–416 .
[2] B. Andreotti, J. Maurer, Y. Couder, S. Douady, Experimental investigation of turbulence near a large scale vortex, Eur. J. Mech. B 17 (1998).
[3] D. Sornette, L. Knopoff, Y.Y. Kagan, C. Vanneste, Rank-ordering statistics of extreme events: application to the distribution of large earthquakes, J. Geophys. Res. 101 (1996) 13883–13893.
[4] J.S. Bendat, A.G. Piersol, Random Data: Analysis and Measurement Procedures, 2nd ed., Wiley, New York, 1986.
[5] W.H. Press, S.A. Teukolsky, W.T. Vettering, B.P. Flannery, Numerical Recipes, Cambridge University Press, Cambridge, 1992, pp. 656–706, 804–826.
[6] W. Feller, An Introduction to Probability Theory and its Aapplications, 3rd ed., vol. 1/2, Wiley, New York, 1968.
[7] K. Pearson, On a criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling, Philos. Mag. 50 (1900) 157–175.
[8] T.E. Holy, Analysis of data from continuous probability distributions, Phys. Rev. Lett. 79 (1997) 3545–3548.