Preprint Accepted for publication in *Inform. Sciences* (2013) **235**: 214–223

Measures of statistical dispersion based on Shannon and Fisher information concepts

Lubomir Kostal,* Petr Lansky, and Ondrej Pokora
Institute of Physiology of the Czech Academy of Sciences, Videnska 1083, 14220 Prague 4, Czech Republic

We propose and discuss two information-based measures of statistical dispersion of positive continuous random variables: the entropy-based dispersion and Fisher information-based dispersion. Although standard deviation is the most frequently employed dispersion measure, we show, that it is not well suited to quantify some aspects that are often expected intuitively, such as the degree of randomness. The proposed dispersion measures are not entirely independent, though each describes the quality of probability distribution from a different point of view. We discuss relationships between the measures, describe their extremal values and illustrate their properties on the Pareto, the lognormal and the lognormal mixture distributions. Application possibilities are also mentioned.

Keywords: Statistical dispersion, Entropy, Fisher information, Positive random variable

1. INTRODUCTION

In recent years, information-based measures of randomness (or "regularity") of signals have gained popularity in various branches of science [1–4]. In this paper we construct measures of statistical dispersion based on Shannon and Fisher information concepts and we describe their properties and mutual relationships. The effort was initiated in [5], where the entropy-based dispersion was employed to quantify certain aspects of neuronal timing precision. Here we extend the previous effort by taking into account the concept of Fisher information (FI), which was employed in different contexts [2, 6–9]. In particular, FI about the location parameter has been employed in the analysis of EEG [8, 10], of the atomic shell structure [11] (together with Shannon entropy) or in the description of variations among the two-electron correlated wavefunctions [12].

The goal of this paper is to propose different dispersion measures and to justify their usefulness. Although the *standard deviation* is used ubiquitously for characterization of *variability*, it is not well suited to quantify certain "intuitively intelligible" properties of the underlying probability distribution. For example highly variable data might not be random at all if it only consists of "very small" and "very large" values. Although the probability density function (or histogram of data) provides a complete view, one needs quantitative methods in order to make a comparison between different experimental scenarios.

The methodology investigated here does not adhere to any specific field of applications. We believe, that the general results are of interest to a wide group of researchers who deal with positive continuous random variables, in theory or in experiments.

* E-mail: kostal@biomed.cas.cz

2. MEASURES OF DISPERSION

2.1. Generic case: standard deviation

We consider a continuous positive random variable (r.v.) T with a probability density function (p.d.f.) $f(t)$ and finite first two moments. Generally, statistical dispersion is a measure of "variability" or "spread" of the distribution of r.v. T, and such a measure has the same physical units as T. There are different dispersion measures described in the literature and employed in different contexts, e.g., standard deviation, interquartile range, mean difference or the L_V coefficient [13–16].

By far, the most common measure of dispersion is the standard deviation, σ, defined as

$$\sigma = \sqrt{\mathbb{E}\left([T - \mathbb{E}(T)]^2\right)}. \tag{1}$$

The corresponding *relative* dispersion measure is obtained by dividing σ with $\mathbb{E}(T)$. The resulting quantity is denoted as the coefficient of variation, C_V,

$$C_V = \frac{\sigma}{\mathbb{E}(T)}. \tag{2}$$

The main advantage of C_V over σ is, that C_V is dimensionless and thus probability distributions with different means can be compared meaningfully.

From Eq. (1) follows, that σ (or C_V) essentially measures how off-centered (with respect to $E(T)$) is the distribution of T. Furthermore, since the difference $(T - \mathbb{E}(T))$ is squared in Eq. (1), it follows that σ is sensitive to outlying values [15]. On the other hand, σ does not quantify how random, or unpredictable, are the outcomes of r.v. T. Namely, high value of σ (high variability) does not indicate that the possible values of T are distributed evenly [5].

2.2. Dispersion measure based on Shannon entropy

For continuous r.v.'s the association between entropy and randomness is less straightforward than for discrete r.v.'s. The (differential) entropy $h(T)$ of r.v. T is defined as [17]

$$h(T) = -\int_0^\infty f(t)\,\ln f(t)\,dt. \tag{3}$$

The value of $h(T)$ may be positive or negative, therefore $h(T)$ is not directly usable as a measure of statistical dispersion [5]. In order to obtain a properly behaving quantity, the entropy-based dispersion, σ_h, is defined as

$$\sigma_h = \exp[h(T)]. \tag{4}$$

The interpretation of σ_h relies on the asymptotic equipartition property theorem [17]. Informally, the theorem states that almost any sequence of n realizations of the random variable T comes from a rather small subset (the typical set) in the n-dimensional space of all possible values. The volume of this subset is approximately $\sigma_h^n = \exp[nh(T)]$, and the volume is bigger for those random variables, which generate more diverse (or unpredictable) realizations. Further connection between σ_h and σ follows from the analogue to the entropy power concept [17]: σ_h/e is equal to the standard deviation of an exponential distribution with entropy equal to $h(T)$.

Analogously to Eq. (2), we define the relative entropy-based dispersion coefficient, C_h, as

$$C_h = \frac{\sigma_h}{\mathbb{E}\,(T)}. \tag{5}$$

The values of σ_h and C_h quantify how "evenly" is the probability distributed over the entire support. From this point of view, σ_h is more appropriate than σ when discussing randomness of data generated by r.v. T.

2.3. Dispersion measure based on Fisher information

The FI plays a key role in the theory of statistical estimation of continuously varying parameters [18]. Let $X \sim p(x;\theta)$ be a family of r.v.'s defined for all values of parameter $\theta \in \Theta$, where Θ is an open subset of the real line. Let $\hat\theta(X)$ be an unbiased estimator of parameter θ, i.e., $\mathbb{E}\left(\hat\theta(X) - \theta\right) = 0$. If for all $\theta \in \Theta$ and both $\varphi(x) \equiv 1$ and $\varphi(x) \equiv \hat\theta(x)$ the following equation is satisfied ([19, p.169] or [18, p.31]),

$$\frac{\partial}{\partial\theta}\int_X \varphi(x)p(x;\theta)\,dx = \int_X \varphi(x)\frac{\partial p(x;\theta)}{\partial\theta}\,dx, \tag{6}$$

then the variance of the estimator $\hat\theta(X)$ satisfies the Cramer-Rao bound,

$$\mathrm{Var}(\hat\theta(X)) \geq \frac{1}{J(\theta|X)}, \tag{7}$$

where

$$J(\theta|X) = \int_X \left[\frac{\partial \ln p(x;\theta)}{\partial\theta}\right]^2 p(x;\theta)\,dx, \tag{8}$$

is the FI about parameter θ contained in a single observation of r.v. X.

Exact conditions (the regularity conditions) under which Eq. (7) holds are stated slightly differently by different authors. In particular, it is sometimes required that the set $\{x : p(x;\theta) > 0\}$ does not depend on θ, which is an unnecessarily strict assumption [18]. For any given p.d.f. $f(t)$ one may conveniently "generate" a simple parametric family by introducing a location parameter. The appropriate regularity conditions for this case are stated below.

The family of location parameter densities $p(x;\theta)$ satisfies

$$p(x;\theta) = p_0(x-\theta), \tag{9}$$

where we consider Θ to be the whole real line and $p_0(x)$ is the p.d.f. of the "generating" r.v. X_0. Let the location family $p(x;\theta)$ be generated by the r.v. $T \sim f(t)$, thus $p(x;\theta) = f(x-\theta)$ and Eq. (8) can be written as

$$\begin{aligned} J(\theta|X) &= \int_\theta^\infty \left[\frac{\partial \ln f(x-\theta)}{\partial\theta}\right]^2 f(x-\theta)\,dx = \\ &= \int_0^\infty \left[\frac{\partial \ln f(t)}{\partial t}\right]^2 f(t)\,dt \equiv J(T), \end{aligned} \tag{10}$$

where the last equality follows from the fact that the derivatives of $f(x-\theta)$ with respect to θ or x are equal up to a sign and due to the location-invariance of the integral (thus justifying the notation as $J(T)$). Since the value of $J(T)$ depends only on the "shape" of the p.d.f. $f(t)$, it is sometimes denoted as the FI *about the random variable* T [17].

To interpret $J(T)$ according to the Cramer-Rao bound in Eq. (7), the required regularity conditions on $f(t)$ are: $f(t)$ must be continuously differentiable *for all* $t > 0$ and $f(0) = f'(0) = 0$. The integral (10) may exist and be finite even if $f(t)$ does not satisfy these conditions, e.g., if $f(t)$ is differentiable almost everywhere or $f(0) \neq 0$. However, in such a case the value of $J(T)$ does not provide any information about the efficiency of the location parameter estimation [18].

The units of $J(T)$ correspond to the inverse of the squared units of T, therefore we propose the FI based dispersion measure, σ_J, as

$$\sigma_J = \frac{1}{\sqrt{J(T)}}. \tag{11}$$

Heuristically, σ_J quantifies the change in the p.d.f. $f(t)$ subject to an infinitesimally small shift $\delta\theta$ in t, i.e, it quantifies the difference between $f(t)$ and $f(t-\delta\theta)$. Any peak, or generally "non-smoothness" in the shape of $f(t)$ decreases σ_J. Analogously to Eqns. (2) and (5) we define the relative dispersion coefficient C_J as

$$C_J = \frac{\sigma_J}{\mathbb{E}\,(T)}. \tag{12}$$

In this paper we do not introduce different symbols for C_J in dependence on whether the Cramer-Rao bound holds or not. We evaluate C_J whenever the integral in Eq. (10) exists and we comment on the regularity conditions in the text.

3. RESULTS

3.1. Extrema of variability

Generally, the value C_V can be any non-negative real number, $0 \leq C_V < \infty$. The lower bound, $C_V \to 0$, is approached by a p.d.f. highly peaked at the mean value, in the limit corresponding to the Dirac's delta function, $f(t) = \delta(t - \mathbb{E}(T))$. There is, however, no unique upper bound distribution for which $C_V \to \infty$. For example, the p.d.f. examples analyzed in the next section allow arbitrarily high values of C_V and yet their shapes are different.

3.2. Extrema of entropy and its relation to variability

The relation between C_V and entropy was investigated in a series of papers [5, 20, 21]. The results can be re-stated in terms of C_h as follows. From the definition of C_h by Eq. (5) and from the properties of entropy [17] follows, that $0 < C_h < e$. The lower bound, $C_h \to 0$, is not realized by any unique distribution. Highly-peaked (possibly multimodal) densities approach the bound and in the limit any discrete-valued distribution achieves it. From this fact follows, that the relationship between C_V and C_h is not unique, small C_V implies small C_h but not vice versa.

The maximum value of C_h is connected with the problem of maximum entropy (ME), which is well known in the literature, see e.g., [17, 22]. The goal is to find such a p.d.f., that maximizes the functional (3) subject to n constraints of the form $\mathbb{E}(\alpha_i(T)) = \xi_i$, where $\alpha_i(t)$ and ξ_i are known and $i = 1, \ldots, n$. The ME p.d.f. satisfying these constraints can be written in the form [17]

$$f(t) = \frac{1}{Z(\lambda_1, \ldots, \lambda_n)} \exp\left[\sum_{i=1}^{n} \lambda_i \alpha_i(t)\right], \tag{13}$$

where the "partition function" $Z(\lambda_1, \ldots, \lambda_n)$ is the normalization factor, $Z(\lambda_1, \ldots, \lambda_n) = \int_0^\infty \exp\left[\sum_{i=1}^n \lambda_i \alpha_i(t)\right] dt$. The introduced Lagrange multipliers, λ_i, are related to the averages ξ_i as [22]

$$-\frac{\partial}{\partial \lambda_i} \ln Z(\lambda_1, \ldots, \lambda_n) = \xi_i. \tag{14}$$

It is well known [17], that the distribution maximizing the entropy on $[0, \infty)$ for given $\mathbb{E}(T)$ is the exponential distribution,

$$f(t) = \frac{1}{\mathbb{E}(T)} \exp\left[-\frac{t}{\mathbb{E}(T)}\right], \tag{15}$$

and entropy $h(T) = 1 + \ln \mathbb{E}(T)$. Thus the upper bound, $C_h = e$, is unique: it is achieved *only if* $f(t)$ is exponential. For the exponential distribution holds $C_V = 1$, however, non-exponential distributions may have $C_V = 1$ too (see the next section). In other words, the maximum of C_h does not correspond to any exclusive value of C_V. This fact highlights the main difference between these two measures: the variability (described by C_V) and randomness (described by C_h) are not interchangeable notions. High variability (overdispersion), $C_V > 1$, results in decreased randomness for many common distributions, see Fig. 2, although there are exceptions, e.g., the Pareto distribution discussed later.

In order to find the ME distribution on $[0, \infty)$ given both $\mathbb{E}(T)$ and C_V, we first realize that the problem is equivalent to finding the ME distribution given $\mathbb{E}(T)$ and $\mathbb{E}\left(T^2\right)$. Applying the Lagrange formalism results in a p.d.f. based on the Gaussian, with the probability of all negative values aliased onto the positive half-line,

$$f(t) = \frac{1}{Z} \exp\left[-\frac{(t-\alpha)^2}{2\beta^2}\right], \tag{16}$$

where

$$Z = \beta\sqrt{\frac{\pi}{2}}\left[1 + \operatorname{erf}\left(\frac{\alpha}{\sqrt{2}\beta}\right)\right]. \tag{17}$$

The density in Eq. (16) is also known as the density of the folded normal r.v. [23]. The parameters α, $\beta > 0$, and $\mathbb{E}(T)$, C_V are related as

$$\mathbb{E}(T) = \beta + \frac{\beta^2}{Z} \exp\left(-\frac{\alpha^2}{2\beta^2}\right), \tag{18}$$

$$C_V = \beta\sqrt{\exp\left(\frac{\alpha^2}{\beta^2}\right) - \frac{\alpha}{Z}\exp\left(\frac{\alpha^2}{2\beta^2}\right) - \frac{\beta^2}{Z^2}} \times \left[\alpha \exp\left(\frac{\alpha^2}{2\beta^2}\right) + \frac{\beta^2}{Z}\right]^{-1}. \tag{19}$$

The entropy and FI can be calculated for Eq. (16) to be

$$h(T) = \frac{1}{2} - \frac{\alpha}{2Z}\exp\left(-\frac{\alpha^2}{2\beta^2}\right) + \ln Z, \tag{20}$$

$$J(T) = \frac{1}{\beta^2}\left[1 - \frac{\alpha}{Z}\exp\left(-\frac{\alpha^2}{2\beta^2}\right)\right]. \tag{21}$$

Note, that C_V in Eq. (19) is limited to $C_V \in (0, 1)$, and therefore the p.d.f. in Eq. (16) provides a solution to the ME problem only in this range. The density of the ME distribution given by Eq. (16) is shown for different values of C_V in Fig. 1. Although it is not possible to express α, β in terms of $\mathbb{E}(T), C_V$ from Eqns. (18) and (19), we obtain all distinct shapes (neglecting the scale) of the folded normal density by fixing, e.g., $\beta = 1$ and varying $\alpha \in (-\infty, \infty)$, since $\lim_{\alpha\to-\infty} C_V(\alpha) = 1$, $\lim_{\alpha\to\infty} C_V(\alpha) = 0$ and noting that $C_V(\alpha)$ is monotonously decreasing. In the limit $C_V = 1$ the density in Eq. (16) becomes exponential, and for $C_V > 1$ there is no unique ME distribution. However, we can always construct a p.d.f. with $C_V > 1$, which is arbitrarily close to the exponential p.d.f., e.g., almost-exponential with a small peak located at some large value of t. Therefore, the maximum value of entropy is $1 + \ln \mathbb{E}(T) - \varepsilon$ for $C_V > 1$, where $\varepsilon > 0$ can be arbitrarily small. The corresponding C_h is shown in Fig. 2.

3.3. Extrema of Fisher information and its relation to entropy

From Eqns. (10) and (12) follows $C_J > 0$. Similarly to C_h, the lower bound is not achieved by a unique distribution, since any continuous, highly peaked density (possibly multimodal) approaches it. Determination of the maximum value of C_J is, however, more difficult. In the following we solve the problem of C_J maximization (FI minimization) subject to $\xi = \mathbb{E}(T)$, both when the regularity conditions hold and when they do not, see Fig. 1.

It is convenient [1, 2] to rewrite the FI functional by employing the real probability amplitude $u(t) = \sqrt{f(t)}$, so that Eq. (10) becomes

$$J(T) = 4 \int_T u'(t)^2 \, dt, \tag{22}$$

where $u'(t) = du(t)/dt$. The extrema of FI satisfies the Euler-Lagrange equation

$$\frac{\partial L}{\partial u} - \frac{d}{dt}\frac{\partial L}{\partial u'} = 0, \tag{23}$$

where the Lagrangian L is

$$L = \int_T u'(t)^2 \, dt + \lambda_1 \left[\int_T u(t)^2 \, dt - 1 \right] + \\ + \lambda_2 \left[\int_T t u(t)^2 \, dt - \xi \right], \tag{24}$$

and the multiplicative constants resulting from the substitution $f \to u$ are contained in Lagrange multipliers λ_1, λ_2. Substituting from Eq. (24) into Eq. (23) results in the differential equation

$$u''(t) - u(t)[\lambda_1 - \lambda_2 t] = 0. \tag{25}$$

The solution to this equation can be written as [24]

$$u(t) = C_1 \mathrm{Ai}\left(\frac{\lambda_1 + \lambda_2 t}{\lambda_2^{2/3}}\right) + C_2 \mathrm{Bi}\left(\frac{\lambda_1 + \lambda_2 t}{\lambda_2^{2/3}}\right), \tag{26}$$

where C_1, C_2 are constants and $\mathrm{Ai}(\cdot), \mathrm{Bi}(\cdot)$ are the Airy functions. Since the integrability of the solution is required, it must be $C_2 = 0$. The remaining parameters $\lambda_1, \lambda_2, C_1$ are determined by requiring that $\int_0^\infty f(t)\, dt = 1$, that the mean equals ξ and from the regularity conditions ($f(0) = f'(0) = 0$). Due to the presence of the Airy function, these parameters must be determined by numerical means. The resulting p.d.f. can be written as

$$f(t) = \frac{1}{Z_1} \mathrm{Ai}^2 \left(a_1 + \frac{b_1 t}{\mathbb{E}(T)}\right), \tag{27}$$

where Z_1 is the normalizing constant, and $a_1 \doteq -2.3381$, $b_1 \doteq 1.5587$. The expression for FI of this p.d.f. can be obtained by integrating Eq. (22) and by combining Eq. (25) with the constraint values,

$$J(T) = -4 \left(\frac{b_1^3 + a_1 b_1^2}{\mathbb{E}(T)^2}\right) \doteq \frac{7.5744}{\mathbb{E}(T)^2}, \tag{28}$$

thus the maximum value of C_J is $C_J \doteq 0.363$. The density from Eq. (27) is shown in Fig. 1. Due to convexity of the FI functional (similarly to the concavity of the entropy functional) in $f(t)$, the maximum of C_J is global. For p.d.f. (27) also holds $C_V \doteq 0.447$, $C_h \doteq 1.77$.

If the regularity conditions are relaxed, we arrive by similar means to the p.d.f.

$$f(t) = \frac{1}{Z_2} \mathrm{Ai}^2 \left(a_2 + \frac{b_2 t}{\mathbb{E}(T)}\right), \tag{29}$$

where $a_2 \doteq -1.0188$, $b_2 \doteq 0.6792$. It holds $C_J \doteq 1.263$, $C_V \doteq 0.79$ and $C_h \doteq 2.63$, the density is shown in Fig. 1. The resulting p.d.f. differs from the exponential shape in both cases, showing that C_h and C_J are two different measures. On the other hand, the p.d.f. which achieves maximum C_h can be fitted to approximate the extremal density of C_J rather well (shown in Fig. 1), further demonstrating the complex relationships between C_V, C_h and C_J. Particularly, even though the shape of C_J-maximizing (C-R not valid) density differs from the exponential density, it holds $C_h \doteq 2.63$, which is close to the maximum value $C_h = e \doteq 2.72$ (corresponding to the exponential density).

The main properties of the dispersion coefficients C_V, C_h and C_J are summarized in Table 1. The evenness of the p.d.f. (described by C_h) is related to the “smoothness” of the density (described by C_J). However, more detailed analysis of C_J shows, that C_h and C_J are not interchangeable, and that the requirement on the differentiability of $f(t)$ plays an important role. Namely, C_J is sensitive to the modes of the density, while C_h is sensitive to the overall spread of the density. Since multimodal densities can be more evenly spread than unimodal ones, it is obvious that the behavior of C_h cannot be deduced from C_J (and vice versa).

Another relationship between C_h and C_J follows from the de-Bruijn’s identity [17, p.672]

$$\left.\frac{\partial}{\partial \varepsilon} h(T + \sqrt{\varepsilon} Z)\right|_{\varepsilon=0} = \frac{1}{2} J(T), \tag{30}$$

where r.v. $Z \sim \mathcal{N}(0,1)$ is standard normal; and from the entropy power inequality [17, p.674]

$$e^{2h(X+Y)} \geq e^{2h(X)} + e^{2h(Y)}, \tag{31}$$

for independent r.v.’s X and Y. The entropy of r.v. $\sqrt{\varepsilon} Z$ in Eq. (30) is $h(\sqrt{\varepsilon} Z) = \frac{1}{2} \ln 2\pi e \varepsilon$, thus from Eq. (31) we have

$$h(T + \sqrt{\varepsilon} Z) \geq \frac{1}{2} \ln \left[e^{2h(T)} + 2\pi e \varepsilon \right]. \tag{32}$$

Taking the derivative with respect to ε and evaluating it at $\varepsilon = 0$ leads to

$$\pi e^{1-2h(T)} \leq J(T), \tag{33}$$

with equality if and only if T is Gaussian (the inequality is thus always strict in the context of this paper). In terms of the relative dispersion coefficients C_h, given by Eq. (5), and C_J, given by Eq. (12), we have from Eq. (33)

$$C_h C_J \leq \frac{1}{\sqrt{\pi e}}. \tag{34}$$

	C_V	C_h	C_J
Interpretation	Distribution of the probability mass with respect to $\mathbb{E}(T)$	Predictability of the outcomes of T	Smoothness of $f(t)$
Sensitive to	Probability of the values away from $\mathbb{E}(T)$	Concentration of the probability mass	Changes in $f'(t)$, modes
Assumptions	$\mathrm{Var}(T)$ exists	No assumptions	$f(t)$ continuously differentiable for $t > 0$ and$^{(*)}$ $f(0) = f'(0) = 0$
Minimum	0	0	0
Minimizing density	$\delta(t - \mathbb{E}(T))$	Not unique	Not unique
Maximum	∞	e	1.263 $0.363^{(*)}$
Maximizing density	Not unique	$\exp[-t/\mathbb{E}(T)]/\mathbb{E}(T)$	$\mathrm{Ai}^2[a + bt/\mathbb{E}(T)]/Z$
Peaked unimodal $f(t) \to \delta(t - \mathbb{E}(T))$	$\to 0$	$\to 0$	$\to 0$
Peaked multimodal $f(t) \to \sum_i \delta(t - \tau_i)$	> 0	$\to 0$	$\to 0$
Extreme variance of T	$\to \infty$	≥ 0	≥ 0
$f(t)$ exponential	implies $C_V = 1$	equal to $C_h = e$	implies $C_J = 1$

Table 1. Summary of properties of the discussed statistical dispersion coefficients of positive continuous random variable T with probability density function $f(t)$ and finite mean value $\mathbb{E}(T)$. The starred$^{(*)}$ entries are valid if the Cramer-Rao bound holds.

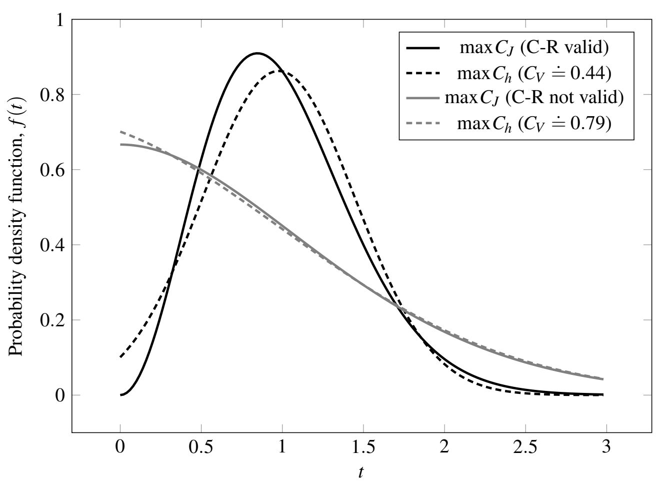

Figure 1. Comparison of probability density functions maximizing the relative dispersion coefficients, C_h and C_J, for different values of coefficient of variation, C_V, and $\mathbb{E}(T) = 1$.

4. APPLICATIONS

4.1. Lognormal and Pareto distributions

Both lognormal and Pareto distributions appear in a broad range of scientific applications [25]. The lognormal distribution is found in the description of, e.g., concentration of elements in the Earth's crust, distribution of organisms in environment or in human medicine, see [26] for a review. The Pareto distribution is often described as an alternative model in situations similar as in the lognormal case, e.g, the sizes of human settlements, sizes of particle or allocation of wealth among individuals [27, 28]. Another common aspect of lognormal and Pareto distributions is, that both can be derived from exponential transforms of common distributions: normal and exponential.

The lognormal p.d.f., parametrized by the mean value and coefficient of variation, is

$$f_{\ln}(t) = \frac{1}{t\sqrt{2\pi \ln(1 + C_V^2)}} \times \times \exp\left\{ -\frac{1}{8} \frac{\left[\ln(1 + C_V^2) + 2\ln(t/\mathbb{E}(T))\right]^2}{\ln(1 + C_V^2)} \right\}. \quad (35)$$

The coefficients C_h and C_J of the lognormal distribution can be calculated to be,

$$C_h = \sqrt{2\pi e}\sqrt{\frac{\ln(1 + C_V^2)}{1 + C_V^2}}, \quad (36)$$

$$C_J = \sqrt{\frac{\ln(1 + C_V^2)}{[1 + C_V^2]^3[1 + \ln(1 + C_V^2)]}}. \quad (37)$$

The dependencies of C_h and C_J on C_V are shown in Fig. 2a, b. We see, that both C_h and C_J as functions of C_V show a "∩" shape with maximum for $C_V = \sqrt{e - 1} \doteq 1.31$ (for C_h) and around $C_V \doteq 0.55$ (for C_J), confirming that each of the proposed dispersion coefficients provides a different point of view. The max C_h p.d.f., Eq. (16), exists only for $C_V \leq 1$, for $C_V > 1$ the upper bound $C_h = 1$ is shown in Fig. 2a. Note, that the max C_h distribution does generally not satisfy the regularity conditions, since $f(0) \neq 0$.

The dependence of C_J on C_h is shown in Fig. 2c. We observe, that C_h and C_J indeed do not describe the same qualities of the distribution, since for the lognormal distribution a single C_h value does not correspond to a single C_J value (and vice versa). In the lognormal case, the dependence between C_h and C_J forms a closed loop, where $C_h = C_J = 0$ for both $C_V \to 0$ and $C_V \to \infty$. In other words, both C_h and C_J fail to distinguish between very different p.d.f. shapes ($C_V \to 0$ or $C_V \to \infty$).

The p.d.f. $f_P(t)$ of the Pareto distribution is

$$f_P(t) = \begin{cases} 0, & t \in (0, b) \\ ab^a t^{-a-1}, & t \in [b, \infty) \end{cases} \tag{38}$$

with parameters $a > 0$ and $b > 0$ (the expression in terms of $\mathbb{E}(T), C_V$ is cumbersome). Note, that $\mathbb{E}(T)$ exists only if $a > 1$ and $\mathrm{Var}(T)$ only if $a > 2$, thus we restrict ourselves to the case $a > 1$ if only C_h and C_J are to be evaluated, and additionally to $a > 2$ if C_V is required. From Eq. (38) follows that $f_P(t)$ is not differentiable at $t = b$, thus $J(T)$ cannot be interpreted in terms of the Cramer-Rao bound, although $J(T)$ is finite for all $a > 0$.

The parameters a and b are related to $\mathbb{E}(T)$ and C_V by

$$a = 1 + \frac{\sqrt{1 + C_V^2}}{C_V}, \tag{39}$$

$$b = \mathbb{E}(T)\left[1 + C_V^2 - C_V\sqrt{1 + C_V^2}\right]. \tag{40}$$

The coefficients C_h and C_J of the Pareto distribution can be expressed in terms of parameter a as

$$C_h = \frac{a-1}{a^2}\exp\left(1 + \frac{1}{a}\right), \tag{41}$$

$$C_J = \frac{(a-1)\sqrt{2+a}}{\sqrt{a^3}(1+a)}. \tag{42}$$

Both C_h and C_J have non-zero limit as $C_V \to \infty$, namely $C_h = \sqrt{e^3}/4 \doteq 1.12$ and $C_J = 1/(3\sqrt{2}) \doteq 0.2357$. However, while C_h as a function of C_V is monotonously increasing, C_J attains its maximum value, $\max C_J \doteq 0.2361$, for $C_V \doteq 2.3591$ (Fig. 2). The monotonous shape of C_h versus the non-monotonous shape of C_J in dependence on C_V is a significant qualitative difference in the behavior of C_h and C_J, although the effect is numerically very small. The shape of the dependence between C_h and C_J forms a closed loop if both $a > 2$ and $2 \geq a > 1$ regions are added together, since $C_h = C_J = 0$ occurs for both $C_V \to 0$ and $a \to 1$ (C_V does not exist).

4.2. Example: lognormal mixture

Finally, we analyze a more complex example, a mixture of two distributions of the same type. The mixture models are met in diverse situations, e.g., in modeling of populations composed of subpopulations, in neuronal coding of odorant mixtures [29] or in the description spiking activity of bursting neurons [30, 31].

Recently, Bhumbra and Dyball [32] have successfully employed a mixture of two lognormal distributions to describe the neuronal firing in supraoptic nucleus.

The p.d.f. of the lognormal mixture model is

$$f_m(t) = pf_{\ln}(t; \mu_1, C_{V1}) + (1-p)f_{\ln}(t; \mu_2, C_{V2}), \tag{43}$$

where $0 < p < 1$ gives the weight of mixture components, and $f_{\ln}(t; \mu, C_V)$ is the lognormal density parametrized by the mean μ and C_V given by Eq. (35). The lognormal mixture does not allow to express C_h or C_J in a closed form. Numerical evaluation of the involved integrals is more convenient in terms of a logarithmically transformed r.v. X, $X = \ln T$, since X is described by a mixture of two normals. Let the density of the r.v. X be denoted as $g_m(x)$, then

$$g_m(x) = p\phi(x, m_1, s_1) + (1-p)\phi(x, m_2, s_2), \tag{44}$$

where $\phi(x, m, s) = \exp[-(x-m)^2/(2s^2)]/\sqrt{2\pi s^2}$ is the density of the normal distribution with mean m and variance s^2. The mean value, $\mu = \mathbb{E}(T)$, and C_V of the random variable T can be expressed as

$$\mu = p\exp\left(m_1 + \frac{s_1^2}{2}\right) + (1-p)\exp\left(m_2 + \frac{s_2^2}{2}\right), \tag{45}$$

$$\begin{aligned} C_V = &\frac{1}{\mu}\Big[p\exp\left(2m_1 + 2s_1^2\right) + \\ &+(1-p)\exp\left(2m_2 + 2s_2^2\right) - \mu^2\Big]^{1/2}. \end{aligned} \tag{46}$$

Since it holds $f_m(t) = g_m(\ln t)/t$ and $dx = dt/t$, the entropy $h(T)$, given by Eq. (3), can be expressed by employing $g_m(x)$ as

$$h(T) = h(X) + \mathbb{E}(X) \tag{47}$$

where $h(X)$ is the entropy of r.v. X and $\mathbb{E}(X) = pm_1 + (1-p)m_2$ is the mean value of r.v. X.

Similarly, we employ r.v. X for the evaluation of C_J, since it holds

$$\frac{d}{dt}g_m(\ln t) = e^{-x}\frac{d}{dx}g_m(x), \tag{48}$$

thus Eq. (10) can be written in terms of $g_m(x)$ as

$$J(T) = \int_{-\infty}^{\infty}\left[\frac{1}{g_m(x)}\frac{dg_m(x)}{dx} - 1\right]^2 e^{-2x}g_m(x)\,dx. \tag{49}$$

The parametric space of the lognormal mixture model is large, in the following we illustrate the behavior of this model just in two different situations.

First, we vary the weight p while keeping the other parameters fixed, Fig. 3 (top row). While the mean value, $\mathbb{E}(T)$, decreases with p monotonically, C_V reaches its maximum $C_V \doteq 1.4$ for $p \doteq 0.5$. The shapes of C_h and C_J in dependence on C_V are radically different: C_h initially increases while C_J decreases. This difference in behavior is explained by the basic properties of C_h and C_J, namely, that C_h is lowest when the p.d.f. is most concentrated (smallest C_V), however, the shape of the density is "smoother" for higher values of C_V. Obviously, this behavior is distribution-dependent. Furthermore, to each C_h corresponds a unique C_V (the reverse statement is not true), while the relationship between C_J and C_V is non-unique both ways. The relationship between C_J and C_h is unique only in the sense that to each C_J corresponds a unique C_h (the reverse statement is not true).

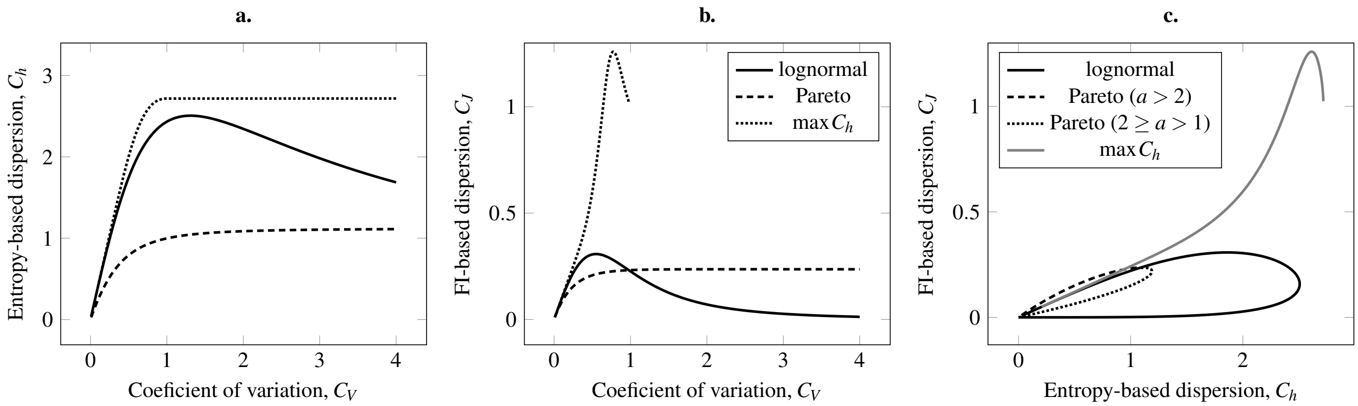

Figure 2. Relationships between C_V, C_h and C_J for the lognormal, Pareto and C_h-maximizing distribution. The max C_h density is unique for $C_V \leq 1$, for $C_V > 1$ only the upper bound can be given. The dependence of C_h on C_V for the lognormal has a global maximum, while for the Pareto distribution C_h grows monotonously. For all distributions holds $C_h \to 0$ as $C_V \to 0$. For the lognormal distribution the dependence of C_J on C_V resembles a scaled version of the C_h-C_V dependence. For the Pareto distribution the C_J-C_V dependence shows a global maximum at $C_V \doteq 2.36$, contrary to the monotonicity of the C_h-C_V dependence. This confirms that "smoothness" and "evenness" of the distribution are different notions, although, e.g., $C_J = 0$ for $C_V \to 0$ for all distributions. The Pareto distribution with parameter $1 < a \leq 2$ is added to the C_h-C_J dependence plot, since for this case both C_h and C_J can be calculated but C_V is undefined.

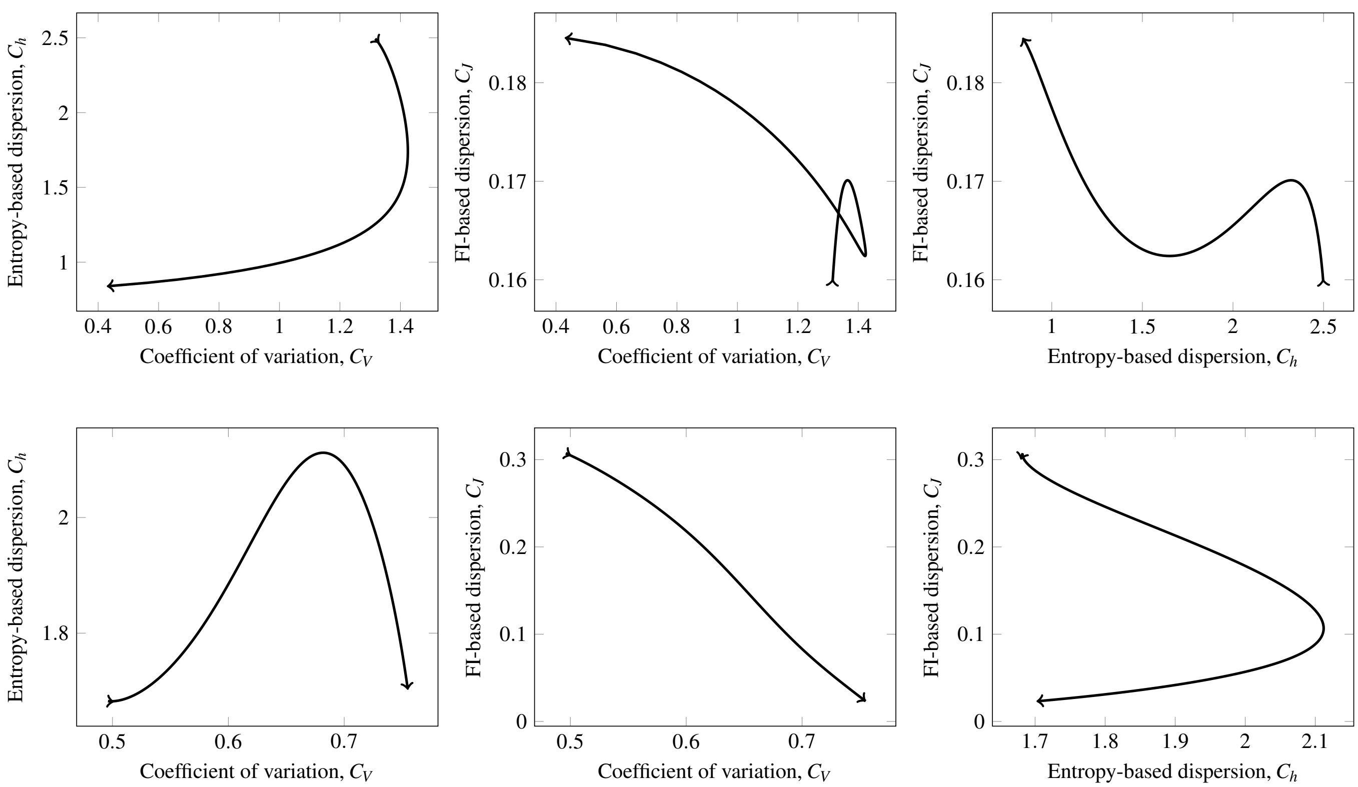

Figure 3. **Top row**: lognormal mixture with variable weight of the components, $p \in [0, 1]$, in the direction of arrows ($m_1 = -1$, $m_2 = -0.5$, $s_1 = 0.2$ and $s_2 = 1$). Although $\mathbb{E}(T)$ decreases with p, C_V exhibits maximum at $p \doteq 0.6$. While the relationship between C_V and C_J is non-unique, C_h describes C_V uniquely (although the reverse statement is not true). **Bottom row**: lognormal mixture with increasing separation between the mean values of the logarithmically transformed components, $m_2 \in [-1, 2]$, in the direction of arrows ($p = 0.2$, $m_1 = -1$, $s_1 = 0.2$ and $s_2 = 0.5$). Although the mean value $\mathbb{E}(T)$ and C_V increase and C_J decreases monotonically, the shape of C_h is unimodal with maximum at $C_V \doteq 0.69$. The example shows the specific sensitivity of C_J to modes (C_J decreases as the modes become more apparent), while C_h is sensitive to the overall spread (at $C_V \doteq 0.69$ the bimodal distribution is more evenly distributed than for any other value of C_V).

In the second example, shown in Fig. 3 (bottom row), we vary the parameter m_2. Both $\mathbb{E}(T)$ and C_V increase monotonically with increasing m_2. While C_J decreases monotonically, C_h shows a unimodal behavior with maxima around $C_V \doteq 0.69$. Thus, although the shape of $f_m(t)$ becomes increasingly bimodal with growing m_2 (i.e., C_J decreases), at the same time the distribution becomes more spread (or equiprobable, thus C_h increases) up to a point $C_V \doteq 0.69$. From that point on, the bimodality becomes too strong and decreases the evenness (or equiprobability) of the distribution and both C_h and C_J decrease. The increasing tendency of the density to become multimodal (decreasing C_J) may result in more unpredictable outcomes of the random variable T (increasing C_h). Both these examples show, that C_h and C_J describe different aspects of the p.d.f. shape.

5. DISCUSSION AND CONCLUSIONS

We propose and discuss two measures of statistical dispersion for continuous positive random variables: the entropy-based dispersion (C_h) and the Fisher information-based dispersion (C_J). Both C_h and C_J describe the overall spread of the distribution differently than the coefficient of variation. While C_h is most sensitive to the concentration of the probability mass (the predictability of random variable outcomes), C_J is sensitive to the modes of the p.d.f. or any non-smothness in the p.d.f. shape in general. The difference between C_h and C_J is further demonstrated by the fact, that the distributions maximizing their values are not the same. On the other hand, we do not claim that C_h (or C_J) is "more informative" than C_V due to taking into account, e.g., higher moments of the distribution. For example, one can find different distributions with equal C_V's but differing C_h's, and vice-versa, distributions with equal C_h's but differing C_V's, see Fig 2a.

It is also important to emphasize what is the benefit of employing the proposed measures once the full distribution function (and therefore a complete description of the situation) is known. The answer is, that it is often required to compare (or "categorize") individual distributions according some specific property, i.e., to assign a number to each function. The advantage of employing the newly proposed measures lies in the possibility to describe p.d.f. qualities from different points of view, that might be of interest in various applications, see e.g., [1, 2, 9]. The parametrical estimates of the proposed coefficients (for both simulated and experimental data from olfactory neurons) were treated in detail in [33]. However, it is natural to ask for the non-parametric versions, which are arguably more valuable in practice. Lansky and Ditlevsen [34] discussed the disadvantages of the "classical" C_V estimator based on sample mean and deviation, proposing solutions especially for the problem of biasedness. Non-parametric reliable estimates of the entropy (and thus of C_h) are well known [35, 36]. Recently, Kostal and Pokora [37] employed the *maximum penalized likelihood* method of Good and Gaskins [38] to jointly estimate C_h and C_J from simulated data.

Acknowledgements. This work was supported by the Institute of Physiology RVO:67985823, the Centre for Neuroscience P304/12/G069 and by the Grant Agency of the Czech Republic projects P103/11/0282 and P103/12/P558.

[1] J. F. Bercher and C. Vignat, "On minimum Fisher information distributions with restricted support and fixed variance," Inform. Sciences **179**, 3832–3842 (2009).

[2] B. R. Frieden, *Physics from Fisher information: a unification* (Cambridge University Press, New York, 1998).

[3] A. L. Berger, V. J. Della Pietra, and S. A. Della Pietra, "A maximum entropy approach to natural language processing," Comput. Linguist. **22**, 39–71 (1996).

[4] S. A. Della Pietra, V. J. Della Pietra, and J. Lafferty, "Inducing features of random fields," IEEE Trans. on Pattern Anal. and Machine Int. **19**, 380–393 (1997).

[5] L. Kostal, P. Lansky, and J-P. Rospars, "Review: neuronal coding and spiking randomness," Eur. J. Neurosci. **26**, 2693–2701 (2007).

[6] L. Bonnasse-Gahot and J-P. Nadal, "Perception of categories: From coding efficiency to reaction times," Brain Res. **1434**, 47–61 (2012).

[7] L. Telesca, V. Lapenna, and M. Lovallo, "Fisher information measure of geoelectrical signals," Physica A **351**, 637–644 (2005).

[8] C. Vignat and J. F. Bercher, "Analysis of signals in the Fisher–Shannon information plane," Phys. Lett. A **312**, 27–33 (2003).

[9] V. Zivojnovic, "A robust accuracy improvement method for blind identification usinghigher order statistics," in *IEEE Internat. Conf. Acous. Speech. Signal Proc., 1993. ICASSP-93.*, Vol. 4 (1993) pp. 516–519.

[10] M. T. Martin, F. Pennini, and A. Plastino, "Fisher's information and the analysis of complex signals," Phys. Lett. A **256**, 173–180 (1999).

[11] J. B. Szabó, K. D. Sen, and Á. Nagy, "The Fisher-Shannon information plane for atoms," Phys. Lett. A **372**, 2428–2430 (2008).

[12] I. A. Howard, A. Borgoo, P. Geerlings, and K. D. Sen, "Comparative characterization of two-electron wavefunctions using information-theory measures," Phys. Lett. A **373**, 3277–3280 (2009).

[13] S. R. Chakravarty, *Ethical social index numbers* (Springer-Verlag, New York, 1990).

[14] H. Cramér, *Mathematical methods of statistics* (Princeton University Press, Princeton, 1946).

[15] M. Kendall, A. Stuart, and J. K. Ord, *The advanced theory of statistics. Vol. 1: Distribution theory* (Charles Griffin, London, 1977).

[16] S. Shinomoto, K. Shima, and J. Tanji, "Differences in spiking patterns among cortical neurons," Neural Comput. **15**, 2823–2842 (2003).

[17] T. M. Cover and J. A. Thomas, *Elements of Information Theory* (John Wiley and Sons, Inc., New York, 1991).

[18] E. J. G. Pitman, *Some basic theory for statistical inference* (John Wiley and Sons, Inc., New York, 1979).

[19] J Shao, *Mathematical statistics* (Springer, New York, 2003).

[20] L. Kostal and P. Lansky, "Classification of stationary neuronal activity according to its information rate," Netw. Comput. Neural Syst. **17**, 193–210 (2006).

[21] L. Kostal, P. Lansky, and C. Zucca, "Randomness and variability of the neuronal activity described by the Ornstein-Uhlenbeck model," Netw. Comput. Neural Syst. **18**, 63–75 (2007).

[22] E. T. Jaynes and G. L. Bretthorst, *Probability Theory: The Logic of Science* (Cambridge University Press, Cambridge, 2003).

[23] F. C. Leone, L. S. Nelson, and R. B. Nottingham, "The folded normal distribution," Technometrics **3**, 543–550 (1961).

[24] M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions, With Formulas, Graphs, and Mathematical Tables* (Dover, New York, 1965).

[25] N. Johnson, S. Kotz, and N. Balakrishnan, *Continuous Univariate Distributions, Vol. 1* (John Wiley & Sons, New York, 1994).

[26] E. Limpert, W. A. Stahel, and M. Abbt, "Log-normal distributions across the sciences: keys and clues," BioScience **51**, 341–352 (2001).

[27] H. A. Simon and C. P. Bonini, "The size distribution of business firms," Am. Economic Rev. , 607–617 (1958).

[28] W. J. Reed, "The Pareto, Zipf and other power laws," Economics Lett. **74**, 15–19 (2001).

[29] J-P. Rospars, P. Lansky, M. Chaput, and P. Viret, "Competitive and noncompetitive odorant interaction in the early neural coding of odorant mixtures," J. Neurosci. **28**, 2659–2666 (2008).

[30] B. C. DeBusk, E. J. DeBruyn, R. K. Snider, J. F. Kabara, and A. B. Bonds, "Stimulus-dependent modulation of spike burst length in cat striate cortical cells," J. Neurophysiol. **78**, 199–213 (1997).

[31] H. C. Tuckwell, *Introduction to Theoretical Neurobiology, Vol. 2* (Cambridge University Press, New York, 1988).

[32] G. S. Bhumbra, A. N. Inyushkin, and R. E. J. Dyball, "Assessment of spike activity in the supraoptic nucleus," J. Neuroendocrinol. **16**, 390–397 (2004).

[33] L. Kostal, P. Lansky, and O. Pokora, "Variability measures of positive random variables," PLoS ONE **6**, e21998 (2011).

[34] S. Ditlevsen and P. Lansky, "Firing variability is higher than deduced from the empirical coefficient of variation," Neural Comput. **23**, 1944–1966 (2011).

[35] O. Vasicek, "A test for normality based on sample entropy," J. Roy. Stat. Soc. B **38**, 54–59 (1976).

[36] A. B. Tsybakov and E. C. van der Meulen, "Root-n consistent estimators of entropy for densities with unbounded support," Scand. J. Statist. **23**, 75–83 (1994).

[37] L. Kostal and O. Pokora, "Nonparametric estimation of information-based measures of statistical dispersion," Entropy **14**, 1221–1233 (2012).

[38] I. J. Good and R. A. Gaskins, "Nonparametric roughness penalties for probability densities," Biometrika **58**, 255–277 (1971).