65 (2020) APPLICATIONS OF MATHEMATICS No. 3, 311–330

NOTION OF INFORMATION AND INDEPENDENT COMPONENT ANALYSIS

Una Radojičić, Klaus Nordhausen, Vienna, Hannu Oja, Turku,

Received November 29, 2019. Published online May 25, 2020.

Abstract. Partial orderings and measures of information for continuous univariate random variables with special roles of Gaussian and uniform distributions are discussed. The information measures and measures of non-Gaussianity including the third and fourth cumulants are generally used as projection indices in the projection pursuit approach for the independent component analysis. The connections between information, non-Gaussianity and statistical independence in the context of independent component analysis is discussed in detail.

Keywords: dispersion; entropy; kurtosis; partial ordering

MSC 2020: 62B10, 94A17, 62H99

1. Introduction

In the engineering literature, independent component analysis (ICA) [12], [23] is often described as a search for the uncorrelated linear combinations of the original variables that maximize non-Gaussianity. The estimation procedure then usually has two steps. First, the vector of principal components is found and the components are standardized to have zero means and unit variances, and second, the vector is further rotated so that the new components maximize a selected measure of non-Gaussianity. It is then argued that the components obtained in this way are made as independent as possible or that they display the components with maximal information. In [12], for example, a heuristic argument is given that, according to the central limit theorem, weighted sums of independent non-Gaussian random variables are closer to Gaussian than the original ones. In this paper, we discuss and clarify the

The work of K. Nordhausen has been supported by the Austrian Science Fund (FWF) Grant number P31881-N32.

DOI: 10.21136/AM.2020.0326-19

somewhat vague connections between non-Gaussianity, independence and notions of information in the context of the independent component analysis.

In Section 2 we first introduce descriptive measures for location, dispersion, skewness and kurtosis of univariate random variables with some discussion of corresponding partial orderings. In this part of the paper we assume that the considered univariate random variable x has a finite mean $E(x)$ and variance $\mathrm{Var}(x)$, cumulative distribution function F and continuously differentiable probability density function f. Skewness, kurtosis and other cumulants of the standardized variable $(x - E(x))/\sqrt{\mathrm{Var}(x)}$ are often used to measure non-Gaussianity of the distribution of x. The most popular measures of statistical information are the differential entropy $H(f) = -\int f(x)\log(f(x))\,\mathrm{d}x$ and the Fisher information in the location model, that is, $J(f) = \int f(x)[f'(x)/f(x)]^2\,\mathrm{d}x$. These and other information measures with related partial orderings and their use as measures of non-Gaussianity are discussed in the latter part of Section 2.

The multivariate independent components model is discussed in Section 3. It is then assumed that, for a p-variate random vector $\mathbf{x}$, there is a linear operator $\mathbf{A} \in \mathbb{R}^{p\times p}$ such that $\mathbf{Ax}$ has independent components. Under certain assumptions, the projection pursuit approach can be used to find the rows of $\mathbf{A}$ one-by-one while various information measures as well as cumulants are used as projection indices. In Section 3 the connections between non-Gaussianity, independence and information in this context are discussed in detail. The paper ends with some final remarks in Section 4.

2. Some characteristics of a univariate distribution

2.1. Location, dispersion, skewness and kurtosis. We consider a continuous random variable x with the finite mean $E(x)$, finite variance $\mathrm{Var}(x)$, density function f and cumulative density function F. Location, dispersion, skewness and kurtosis are often considered by defining the corresponding measures or functionals for these properties. Location and dispersion measures, write $T(x)$ and $S(x)$, are functions of the distribution of x and defined as follows.

Definition 2.1.

(1) $T(x) \in \mathbb{R}$ is a location measure if $T(ax+b) = aT(x) + b$ for all $a, b \in \mathbb{R}$.
(2) $S(x) \in \mathbb{R}_+$ is a dispersion measure if $S(ax+b) = |a|S(x)$ for all $a, b \in \mathbb{R}$.

Clearly, if T is a location measure and x is symmetric around μ, then $T(x) = \mu$ for all location measures. For squared dispersion measures S^2, Huber [10] considered the concepts of additivity, subadditivity and superadditivity. These concepts appear to be crucial in developing tools for the independent component analysis and are defined as follows.

Definition 2.2. Let S^2 be a squared dispersion measure.

(1) S^2 is *additive* if $S^2(x+y) = S^2(x) + S^2(y)$ for all independent x and y.
(2) S^2 is *subadditive* if $S^2(x+y) \leqslant S^2(x) + S^2(y)$ for all independent x and y.
(3) S^2 is *superadditive* if $S^2(x+y) \geqslant S^2(x) + S^2(y)$ for all independent x and y.

The mean $E(x)$ and the variance $\mathrm{Var}(x)$ are important and most popular location and squared dispersion measures. It is well known that $\mathrm{Var}(x+y) = \mathrm{Var}(x) + \mathrm{Var}(y)$ for independent x and y, and $E(x+y) = E(x) + E(y)$ is true even for dependent x and y. These additivity properties are highly important in certain applications and in fact characterize the mean and variance among continuous measures as follows.

Theorem 2.1.

(1) *Let a location measure T be additive and weakly continuous at $N(0,1)$, that is, $z_n \to_d z \sim N(0,1)$ implies that $T(z_n) \to T(z) = 0$. Then $T(x) = E(x)$ for all x with finite second moments.*
(2) *Let a squared dispersion measure S^2 be additive and continuous at $N(0,1)$, that is, $z_n \to_d z \sim N(0,1)$ implies that $S^2(z_n) \to S^2(z) > 0$. Then $S^2(x) = S^2(z)\,\mathrm{Var}(x)$ for all x with finite second moments.*

The weak continuity of a statistical functional is a popular and crucial assumption for finding a consistent estimate of the population value, for example. Let $x_1, \dots, x_n$ be a random sample of size n from the distribution F and $F_n(x) = n^{-1} \sum_{i=1}^{n} 1_{x_i \leqslant x}$ be the sample cumulative distribution function. If T is weakly continuous at F then T at F_n is a consistent estimate of T at F as F_n converges in distribution to F. In Theorem 2.1 we in fact need the weak continuity only at the Gaussian distribution, see the proof in Section 5.

The comparison of different location measures T_1 and T_2, and dispersion measures S_1 and S_2 provides measures of skewness and kurtosis as

$$\mathrm{Sk}(x) = \frac{T_1(x) - T_2(x)}{S(x)} \quad \text{and} \quad \mathrm{Ku}(x) = \frac{S_1^2(x)}{S_2^2(x)}.$$

Classical measures of skewness and kurtosis proposed in the literature can be written in this way. Note that both the measures are affine invariant in the sense that

$$\mathrm{Sk}(ax+b) = \mathrm{sgn}(a)\mathrm{Sk}(x) \quad \text{and} \quad \mathrm{Ku}(ax+b) = \mathrm{Ku}(x).$$

If x has a symmetric distribution, then $\mathrm{Sk}(x) = 0$. In the literature, kurtosis measures are aimed to measure the peakedness and/or the heaviness of the tails of the density of x but, as we will see in Section 2.3, $\mathrm{Ku}(x)$ as defined here may be a global measure of deviation from the normality and has also been used as an affine invariant information measure for some special choices of the dispersion measures S_1 and S_2.

Moment and cumulant generating functions defined as

$$E[\mathrm{e}^{tx}] = \sum_{k=0}^{\infty} \mu_k t^k/k! \quad \text{and} \quad \log E[\mathrm{e}^{tx}] = \sum_{k=0}^{\infty} \kappa_k t^k/k!,$$

respectively, generate classical measures, i.e., moments $E(x) = \mu_1(x)$ and $\mathrm{Var}(x) = \mu_2(x - \mu_1(x))$, and cumulants $\kappa_3(x^{\mathrm{st}})$ and $\kappa_4(x^{\mathrm{st}})$ where $x^{\mathrm{st}} = (x - E(x))/\sqrt{\mathrm{Var}(x)}$. The cumulants κ_k for all $k = 1, 2, \ldots$ are additive as $\log E[\mathrm{e}^{tx}]$ is additive and $\kappa_k^{2/k}(x - E(x))$, $k = 2, 3, \ldots$, are subadditive squared dispersion measures which follows from the Minkowski inequality, see [10]. Another class of measures is given by the quantiles $q_u = F^{-1}(u)$, $0 < u < 1$, with the corresponding measures such as

$$q_{1/2},\ q_{1-u} - q_u,\ \frac{q_u + q_{1-u} - 2q_{1/2}}{q_{1-u} - q_u}, \quad \text{and} \quad \frac{q_{1-u} - q_u}{q_{1-v} - q_v}, \quad 0 < u < v < \frac{1}{2}.$$

These quantile based measures provide robust alternatives to moment based measures. To our knowledge, the quantile based location and squared dispersion measures, however, lack the additivity, subadditivity or superadditivity properties as defined in Definition 2.2. As proven later in Theorem 3.2, the squared dispersion measures, which are either subadditive or superadditive, can be used to find independent components. The use of quantile based measures in independent component analysis is therefore questionable.

An alternative approach when discussing and comparing certain distribution properties, such as location, dispersion, skewness and kurtosis, is to define the corresponding *partial orderings* of such measures. For continuous x and y with the cumulative distribution functions F and G, write $\Delta(x) = G^{-1}(F(x)) - x$. The function $\Delta(x)$ is called a shift function of x as x when shifted by $\Delta(x)$ has the distribution of y. The transformation $x \mapsto x + \Delta(x)$ is also known as the (univariate) Monge-Kantorovich optimal transport map. Using the function Δ, we can naturally define the following partial orderings [3], [4], [32], [25].

(1) Location ordering: Δ is positive.
(2) Dispersion ordering: Δ is increasing.
(3) Skewness ordering: Δ is convex.
(4) Kurtosis ordering: Δ is concave-convex.

Papers [3], [4], [25] then stated that, in addition to the affine equivariance and invariance properties, the measures of location, dispersion, skewness and kurtosis should be monotone with respect to the corresponding orderings. For example, if one needs to order the random variables x and y in the location sense, one can proceed by checking whether the corresponding Δ function is positive. For finding monotone measures in the dispersion case, for example, Δ is increasing if and only if

$$E[C(x - E(x))] \leqslant E[C(y - E(y))] \quad \text{for all convex } C,$$

which is also called the dilation order. It implies, for example, that the measures $(E[|x - E(x)|^k])^{1/k}$, $k > 1$, are monotone dispersion measures. Now that we have defined various partial orderings of random variables with respect to location, dispersion, kurtosis and skewness, we consider ordering random variables with respect to the amount of information they carry.

2.2. Information and discrete distributions. Consider a discrete random variable with k possible values ('alphabets') with probabilities listed in $p = (p_1, \ldots, p_k)$. Write $p_{(1)} \leqslant \ldots \leqslant p_{(k)}$ for the ordered probabilities. It is sometimes presumed that a distribution p is informative if it can provide 'surprises' with very small p_i's. On the other hand, people often claim that p is informative if the result of the experiment is known with a high probability, that is, if only one or few values have high p_i's. These somewhat naive characterizations suggest the following well-known partial ordering for discrete distributions [19].

Definition 2.3. *Majorization*: $p \prec q$ if $\sum_{i=1}^{j} p_{(i)} \geqslant \sum_{i=1}^{j} q_{(i)}$, $j = 1, \ldots, k$. Then p is said to be majorized by q.

Majorization is nothing but a dispersion ordering (and a dilation order) for the discrete distributions with k equiprobable values $p_1, \ldots, p_k$ in $[0, 1]$ with the mean $1/k$. Then, according to [26],

$$\begin{aligned} p \prec q &\Leftrightarrow p = q\mathbf{L} \text{ with some doubly stochastic matrix } \mathbf{L} \\ &\Leftrightarrow \sum_{i=1}^{k} C(p_i) \leqslant \sum_{i=1}^{k} C(q_i) \quad \text{for all continuous convex } C. \end{aligned}$$

The doubly stochastic matrix $\mathbf{L}$ is a matrix with non-negative elements such that all row sums and all column sums equal one. The doubly stochastic operator $\mathbf{L}$ is then in fact a convex combination of permutations; p is obtained from q by this 'smoothing' and is therefore less informative. Further, for all p,

$$(1/k, \ldots, 1/k) \prec p \prec (0, \ldots, 0, 1)$$

and, for simple mixtures, $p \prec q \Rightarrow p \prec \lambda p + (1 - \lambda)q \prec q$, $0 \leqslant \lambda \leqslant 1$.

We can now give the following definition.

Definition 2.4. Let $p = (p_1, \dots, p_k)$ list the probabilities of k possible values of a discrete random variable, that is, $p_1, \dots, p_k \in [0, 1]$, $\sum_{i=1}^{k} p_i = 1$. A measure $M(p)$ is an information measure if it is monotone with respect to majorization.

Note that, as $(p_1, \dots, p_k) \prec (p_{(1)}, \dots, p_{(k)}) \prec (p_1, \dots, p_k)$, the definition implies that the information measures are invariant under permutations of the probabilities in $(p_1, \dots, p_k)$. The equivalent conditions for majorization then suggest quantities such as

$$H(p) = -\sum_{i=1}^{k} \log(p_i)p_i, \quad H^*(p) = \sum_{i=1}^{k} p_i^2 \quad \text{and} \quad H^{**}(p) = p_{(k)},$$

and $-H$, H^* and H^{**} are monotone information measures that easily extend to continuous and multivariate cases. *Shannon's entropy* [29], $-\sum_{i=1}^{k} \log_2(p_i)p_i$, is often seen as a measure of ability to compress the data (e.g. lower bound for the expected number of bits to store the data).

2.3. Some information measures for continuous distributions. Consider next a continuous random variable x with the continuously differentiable probability density function f and finite variance $\mathrm{Var}(x)$. The three measures from the discrete case straightforwardly extend in the continuous case to

$$\begin{aligned}
H(x) &= -E[\log f(x)] = -\int_{-\infty}^{\infty} f(x) \log f(x)\, \mathrm{d}x, \\
H^*(x) &= E[f(x)] = \int_{-\infty}^{\infty} f^2(x)\, \mathrm{d}x, \quad \text{and} \\
H^{**}(x) &= \sup_x f(x) = f(x_{\mathrm{mode}}), \ \text{if the mode } x_{\mathrm{mode}} \text{ exists.}
\end{aligned}$$

The Fisher information in the location model $f(\cdot - \mu)$ at $\mu = 0$ given by

$$J(x) = \int_{-\infty}^{\infty} f(x) \Big(\frac{f'(x)}{f(x)}\Big)^2 \mathrm{d}x$$

is also often used as an information measure [16].

The measure $H(x)$ is popular in the literature and known as the *differential entropy.* Under certain restrictions, the measure has the following maximizers [7]. For the distributions on $\mathbb{R}$ with a fixed variance, $H(x)$ is maximized if x has a normal distribution. For the distributions on $\mathbb{R}_+$ with a fixed mean, $H(x)$ is maximized at the exponential distribution. For the distributions on a finite interval, $H(x)$ is maximized at the uniform distribution on that interval. Note that, in the Bayesian analysis, these three distributions are often used as priors that reflect 'total ignorance'. An interesting connection between the Fisher information and the differential entropy can be found in [1].

We next show that the three straightforward extensions H, H^*, and H^{**} as well as the Fisher information J provide squared dispersion measures as in Definition 2.1 but with an interesting additional invariance property. First note that the measures are invariant under the location shift of the distribution but not under the rescaling of the variable. Recall that information as stated for discrete distributions is invariant under the permutations of the probabilities in $(p_1, \dots, p_k)$. All permutations consist of successive pairwise exchanges of two probabilities. In the continuous case, similar elemental probability density transformations may be constructed as follows. For all $a < a + \Delta < b < b + \Delta$ and density function f, write

$$f_{a,b,\Delta}(x) = \begin{cases} f(x), & x \in \mathbb{R} - [a, a+\Delta] - [b, b+\Delta], \\ f(b + (x - a)), & x \in [a, a+\Delta], \\ f(a + (x - b)), & x \in [b, b+\Delta]. \end{cases}$$

The transformation allows the manipulation of the properties of the distribution in many ways. The transformation can, for example, be used to move some probability mass from the centre of distribution to the tails and in this way to manipulate the variance and the kurtosis of the distribution. As far as we know, this transformation has not been discussed in the literature yet. It is surprising that the information measures H, H^*, H^{**}, and J provide dispersion measures which are invariant under these transformations.

Theorem 2.2. *The entropy power* $\mathrm{e}^{2H(x)}$ *and measures* $[H^*(x)]^{-2}$, $[H^{**}(x)]^{-2}$ *and* $[J(x)]^{-1}$ *are squared dispersion measures that are invariant under the transformations* $f \to f_{a,b,\Delta}$. *The measures* $\mathrm{e}^{2H(x)}$ *and* $[J(x)]^{-1}$ *are superadditive.*

2.4. Affine invariant information measures. We now further discuss the properties of the dispersion measures from Theorem 2.2 and, to find affine invariant information measures, consider the ratios of the variance to these squared dispersion measures. The ratio of the variance to the entropy power, that is, $\mathrm{Var}(x)\mathrm{e}^{-2H(x)}$, is minimized at the normal distribution [7]. In a neighbourhood of a normal distribution, the negative entropy $-H(x)$ possesses an interesting approximation using third and fourth cumulants. Paper [13] showed that the negative differential entropy for the density $f(x) = \varphi(x)(1 + \varepsilon(x))$, where φ is the density of $N(0,1)$ and ε is a well-behaved "small" function that satisfies $E[\varepsilon(z)z^k] = 0$, $z \sim N(0,1)$, $k = 0, 1, 2$, can be approximated by $\frac{1}{2}\int \varphi(x)\varepsilon^2(x)\,\mathrm{d}x \approx \frac{1}{12}(\kappa_3^2(x) + \frac{1}{4}\kappa_4^2(x))$.

Next, $[H^*(x)]^{-2}$ is a (squared) dispersion measure, and therefore $[H^*(x)]^2 \operatorname{Var}(x)$ provides an affine invariant information measure. For symmetric distributions, it preserves the concave-convex kurtosis ordering of van Zwet and $12[H^*(x)]^2 \operatorname{Var}(x)$ is in fact the efficiency of the Wilcoxon rank test with respect to the t-test. Also, for symmetric distributions, $4[H^{**}(x)]^2 \operatorname{Var}(x)$ is a kurtosis measure in the van Zwet sense and simultaneously the efficiency of the sign test with respect to the t-test. We also mention that, if $Q(x) = E[f(F^{-1}(u))/\varphi(\Phi^{-1}(u))]$ with $u \sim U(0,1)$, then $Q^{-2}(x)$ is a squared dispersion measure and $Q^2(x) \operatorname{Var}(x)$ is the efficiency of the van der Waerden test with respect to the t-test in the symmetric case. By the Chernoff-Savage theorem, it attains its minimum 1 at the normal distribution. See [6], [9].

Finally, the information measure $\operatorname{Var}(x)J(x) \geqslant 1$ is minimized at the normal distribution. In the location estimation problem in the symmetric case, $\operatorname{Var}(x)J(x)$ is also the asymptotic relative efficiency of the maximum likelihood estimate of the symmetry centre with respect to the sample mean [28].

2.5. Information orders for continuous distributions. We next outline how to construct partial orderings for information in the univariate continuous case as an extension to the discrete case. Let first x be a continuous random variable with density f on $(0,1)$. If $m(y) = \mu\{u\colon f(u) > y\}$ where μ is the Lebesgue measure, then the function $f_\downarrow(u) = \sup\{y\colon m(y) > u\}$, $u \in (0,1)$, provides the *decreasing rearrangement* of f. Note that any density function on $(0,1)$ can be approximated by a simple density function $f(x) = \sum\limits_{i=1}^{k} \alpha_i \chi_{A_j}(x)$, where $\alpha_1 < \alpha_2 < \ldots < \alpha_k$, $A_1, \ldots, A_k$ are disjoint Lebesque-measurable sets on $(0,1)$ and χ_A is the characteristic function of set A. Then

$$m(y) = \sum_{i=1}^{k} \beta_i \chi_{B_i}(y) \quad \text{and} \quad f_\downarrow(u) = \sum_{i=1}^{k} \alpha_i \chi_{[\beta_{i-1},\beta_i)}(u),$$

where $\beta_i = \sum\limits_{j=1}^{i} \mu(A_j)$, $B_i = [\alpha_{i+1}, \alpha_i)$ for $i = 1, 2, \ldots, k$, and $\beta_0 = \alpha_{k+1} = 0$. For a better insight on how the decreasing rearrangement $f_\downarrow$ is constructed, see Figure 1. For more details and examples, see e.g. [15].

Using the decreasing rearrangement we can give the following definitions.

Definition 2.5. Let f and g be density functions on the interval $(0,1)$. Then g has more information than f, write $f \prec g$, if

$$\int_0^u f_\downarrow(v)\,\mathrm{d}v \leqslant \int_0^u g_\downarrow(v)\,\mathrm{d}v \quad \forall\, u \in (0,1).$$

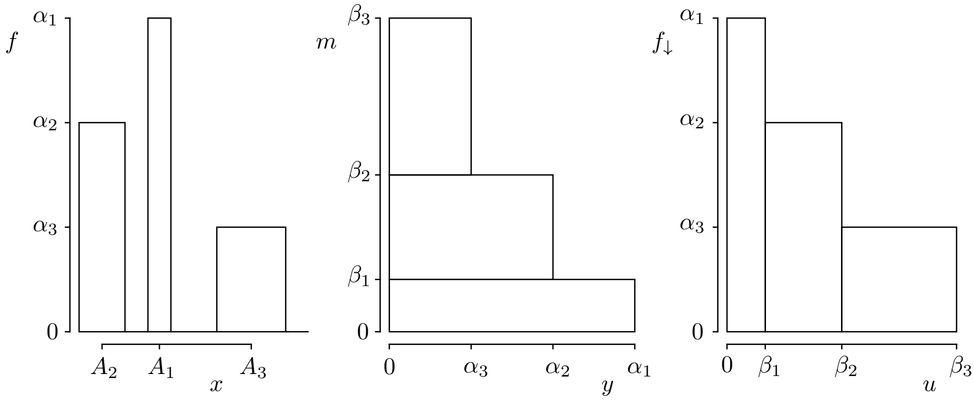


Figure 1. A simple function f (left), corresponding function m (middle) and decreasing rearrangement $f_{\downarrow}$ (right).

Definition 2.6. Let $\mathcal{F}_{(0,1)}$ be the set of density functions f on the interval $(0,1)$. Then $M_{(0,1)}\colon \mathcal{F}_{(0,1)} \to \mathbb{R}$ is an information measure if it is monotone with respect to the partial ordering from Definition 2.5

The distribution with minimum information is the uniform distribution on $(0,1)$. Information measures are easily found, see [27], as $f \prec g$ if and only if

$$\int_0^1 C(f(u))\,\mathrm{d}u \leqslant \int_0^1 C(g(u))\,\mathrm{d}u \quad \text{for all continuous convex functions } C.$$

Ryff [27] also discusses how to construct linear operators L for which $f = Lg \prec g$ when $f \prec g$.

Consider next a continuous random variable x on $\mathbb{R}$ with a pdf f. To find a location and a scale-free version of the density, Staudte [30] proposed the transformation

$$f(x),\ \mathbb{R} \ni x \mapsto f^*(u) = \frac{f(F^{-1}(u))}{H^*(x)}, \quad u \in (0,1).$$

Then f^*, called *the probability density quantile* (*pdQ*), is a probability density function on $(0,1)$ which is invariant under linear transformations of the original variable x [30]. It is also true that, for a given f^*, the original f is known up the location and scale. Using this density transformation, the definition of an invariant information measure for densities on $\mathbb{R}$ can be given as follows.

Definition 2.7. Let $\mathcal{F}_{\mathbb{R}}$ be a set of density functions f on $\mathbb{R}$ and let $M_{(0,1)}\colon \mathcal{F}_{(0,1)} \to \mathbb{R}$ be an information measure for distributions on $(0,1)$. Then $M_{\mathbb{R}}\colon f \to M_{(0,1)}(f^*)$ is an information measure in the set $\mathcal{F}_{\mathbb{R}}$.

Note that $M_{\mathbb{R}}$ is not an extension of $M_{(0,1)}$ meaning that $f \in \mathcal{F}_{(0,1)}$ does not imply that $M_{\mathbb{R}}(f) = M_{(0,1)}(f)$. $M_{\mathbb{R}}$ is invariant under rescaling of f while $M_{(0,1)}$ is not.

Applying Definition 2.7 and choosing convex $C(u) = -\log(u)$ and $C(u) = \log(u)u$, we get location and scale invariant information measures for f such as

$$\exp\left\{-2\int \log(f^*(u))\,\mathrm{d}u\right\} = \mathrm{e}^{2H(x)}[H^*(x)]^2$$

and

$$\exp\left\{4\int \log(f^*(u))f^*(u)\,\mathrm{d}u\right\} = \mathrm{e}^{-2H(f^2/H^*(x))}[H^*(x)]^{-2},$$

which attain their minimum at the uniform distribution and are invariant under the transformations $f \mapsto f_{a,b,\Delta}$. For more details, see e.g. [31].

To replace the transformation $f \mapsto f^*$ by a transformation to densities on $(0,1)$ for which the minimum information is attained at any density g, one can use the following adjustment.

Theorem 2.3. *Let x and y be random variables on $\mathbb{R}$ with the probability density functions f and g and cumulative distribution functions F and G, respectively. Then*

$$f : g(u) = \frac{f(G^{-1}(u))}{g(G^{-1}(u))}$$

is a density function on $(0,1)$ and its differential entropy $-H(f : g) \geqslant 0$ is the Kullback-Leibler (KL) divergence between the distributions of x and y.

Let again x have a density f and let φ and Φ be the pdf and the cdf of a normal distribution with the mean $E(x)$ and variance $\mathrm{Var}(x)$. Then one can show, using similar arguments as in [30], that

$$f : \varphi(u) = \frac{f(\Phi^{-1}(u))}{\varphi(\Phi^{-1}(u))}, \quad u \in (0,1),$$

is a location and scale-free density and information measures from Definition 2.6 applied to the set of densities $\tilde{f} = f : \varphi$ attain their minima when f has a normal distribution. A collection of information measures is given by $\int C(\tilde{f}(u))\,\mathrm{d}u$ with continuous and convex functions C and then we get, for example, again

$$\exp\left\{2\int \log(\tilde{f}(u))\tilde{f}(u)\,\mathrm{d}u\right\} = (2\pi\mathrm{e})\mathrm{e}^{-2H(x)}\,\mathrm{Var}(x).$$

We next provide examples of the probability density functions f, f^*, and $\tilde{f}$ when f is the density of the Gaussian, Laplace, lognormal and uniform distributions. Also a mixture of two Gaussian distributions denoted by $\mathrm{GMM}(\mu_1, \mu_2, \sigma_1, \sigma_2, w)$ is considered with the densities $w\varphi_{\mu_1,\sigma_1}(x) + (1-w)\varphi_{\mu_2,\sigma_2}(x)$, $0 \leqslant w \leqslant 1$. Figure 2 then shows the impact of the transformations $f \to f^*$ and $f \to \tilde{f}$ in these cases.

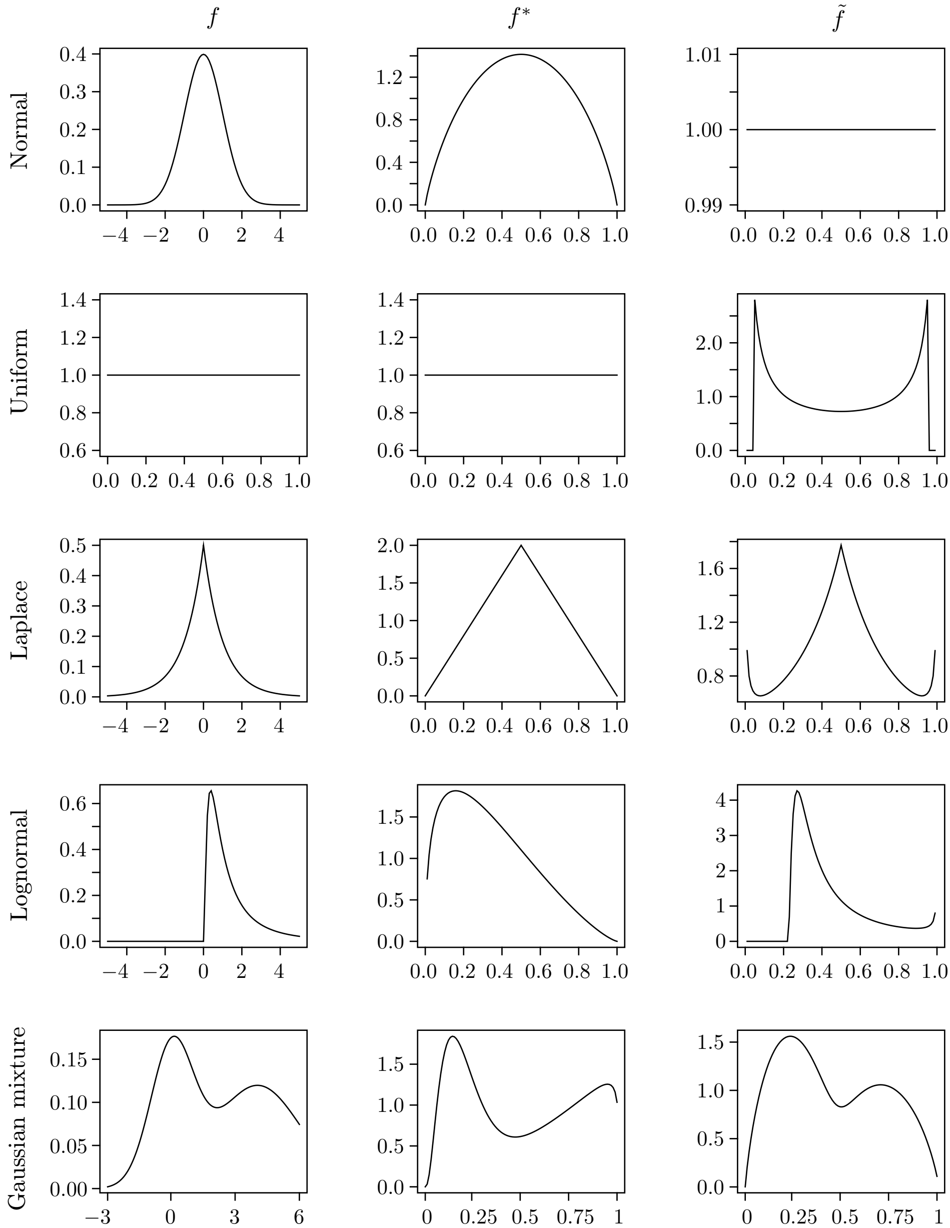


Figure 2. Comparison of f, f^*, and $\tilde{f}$ for five distributions.

Table 1 provides for the same distributions the power entropies $e^{2H(\cdot)}$ and $H^*(\cdot)^{-2}$ for f, f^*, and $\tilde{f}$. Note that the information measures applied to f are not invariant under rescaling of x as opposed to f^* and $\tilde{f}$. For example, for the settings we use in Table 1, the normal and lognormal densities have the same power entropy just by accident and the equality is not generally true.

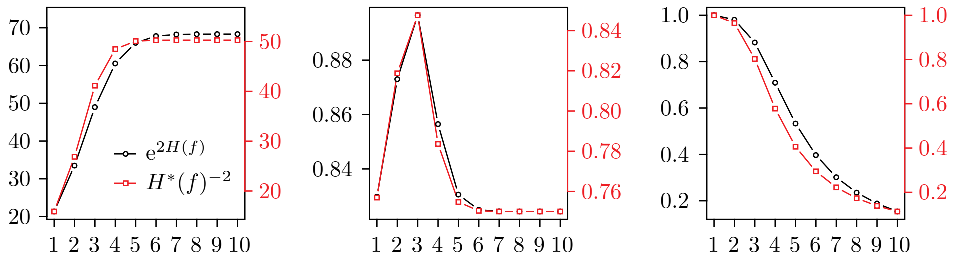
Distribution	$e^{2H(f)}$	$e^{2H(f^*)}$	$e^{2H(\tilde{f})}$	$H^*(f)^{-2}$	$H^*(f^*)^{-2}$	$H^*(\tilde{f})^{-2}$
$N(0,1)$	17.079	0.824	1.000	12.566	0.750	1.000
Laplace(1)	29.556	0.680	0.887	16.000	0.719	0.783
Lognormal(0,1)	17.079	0.642	0.308	7.622	0.537	0.186
$U(0,1)$	1.000	1.000	0.703	1.000	1.000	0.567
GMM$(0,4,1,2,0.4)$	100.000	0.862	0.855	78.000	0.792	0.756

Table 1. The power entropy and the $[H^*]^{-2}$ measure for some continuous distributions and their transformations.

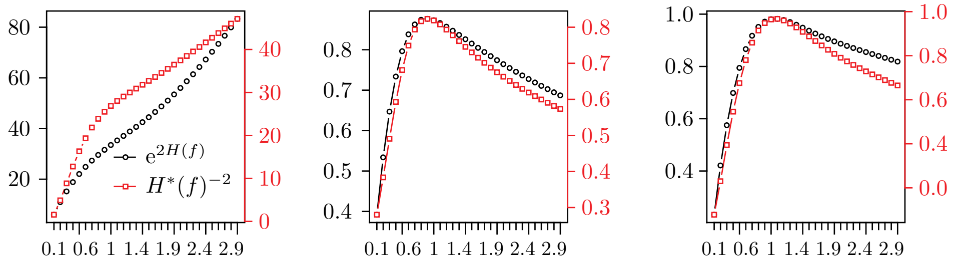
For better understanding of the measures, we illustrate the behavior of $e^{H(\cdot)}$ and $H^*(\cdot)^{-2}$ in the GMM model with four fixed parameters and one parameter that varies in turn. In Figure 3 both information measure curves are plotted in the corresponding graphs to compare the shapes of the curves as well as the occurrences of the extreme values. The curves for $\tilde{f}$ with varying location and scale seem natural since minimum information is attained as GMM gets "closer" to the normal distribution. Results for f^* and varying location seem strange in a sense. One would expect a decreasing behaviour of both measures when the distance in means increases, as is the case for $\tilde{f}$, while the result for f in all three cases could simply be explained by the decrease in information as a result of the increase in overall variance of the mixture. The functions $e^{H(\cdot)}$ and $H^*(\cdot)^{-2}$ seem to behave almost proportionally in all cases. In cases of f^* and $\tilde{f}$, where the majorization is well defined, such behaviour is indeed expected, as the reciprocals of both $e^{H(\cdot)}$ and $H^*(\cdot)^{-2}$ are information measures for both f^* and $\tilde{f}$. However, further investigations into this matter will be conducted in the future.

3. Independent component analysis

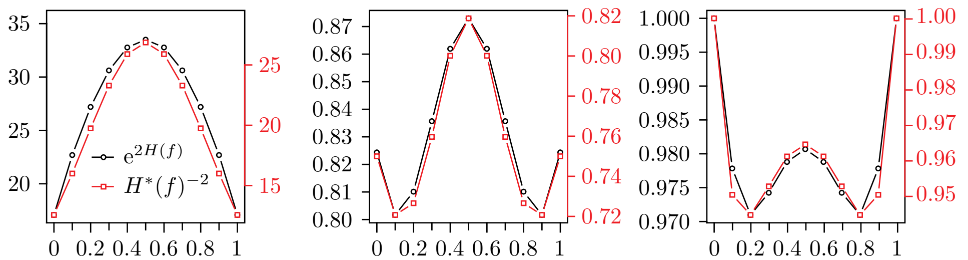
In this section we discuss independent component analysis from a projection pursuit point of view, and motivate the application of the above considered information and dispersion measures as projection indices when estimating the independent components.



(a) GMM$(0, \mu_2, 1, 1, 0.5)$ where μ_2 varies.



(b) GMM$(0, 2, 1, \sigma_2, 0.5)$ where σ_2 varies.



(c) GMM$(0, 2, 1, 1, w)$ where w varies.

Figure 3. Power entropy and $[H^*]^{-2}$ for different GMMs when always one parameter varies. The left vertical axis corresponds to power entropy and the right axis to $[H^*]^{-2}$. The left panel gives the measures for f, the middle one for f^* and the right one for $\tilde{f}$. Horizontal axis: (a) μ_2, (b) σ_2 and (c) w.

3.1. Some preliminaries. In this section we consider multivariate random variables. For a p-variate random vector $\mathbf{x}$ with finite second moments, the mean vector and covariance matrix are $E(\mathbf{x}) \in \mathbb{R}^p$ and $\mathrm{Cov}(\mathbf{x}) \in \mathbb{R}^{p\times p}$, respectively. Let $\mathrm{Cov}(\mathbf{x}) = \mathbf{UDU}'$ be the eigenvector-eigenvalue decomposition of the covariance matrix. Then $\mathrm{Cov}(\mathbf{x})^{-1/2} := \mathbf{UD}^{-1/2}\mathbf{U}'$ and $\mathbf{x}^{\mathrm{st}} = \mathrm{Cov}(\mathbf{x})^{-1/2}(\mathbf{x} - E(\mathbf{x}))$ standardizes $\mathbf{x}$, that is, $E(\mathbf{x}^{\mathrm{st}}) = \mathbf{0}$ and $\mathrm{Cov}(\mathbf{x}^{\mathrm{st}}) = \mathbf{I}_p$. The set of $p \times r$, $r \leqslant p$, matrices with orthonormal columns is denoted by $\mathcal{O}^{p\times r}$. Thus $\mathbf{U} \in \mathcal{O}^{p\times r}$ implies $\mathbf{U}'\mathbf{U} = \mathbf{I}_r$. The set of $p \times p$ diagonal matrices with positive diagonal elements is denoted by $\mathcal{D}^{p\times p}$. If $\mathbf{U}\in\mathcal{O}^{p\times p}$ and $\mathbf{D}\in\mathcal{D}^{p\times p}$ then $\mathbf{x}\to\mathbf{U}\mathbf{x}$ and $\mathbf{x}\to\mathbf{D}\mathbf{x}$, $\mathbf{x}\in\mathbb{R}^p$, are a *rotation* operator and a *componentwise rescaling* operator, respectively. Let $\mathbf{A}\in\mathbb{R}^{p\times q}$ be a matrix with rank $r\leqslant\min\{p,q\}$. Then the linear operator $\mathbf{A}$ may be written as (singular value decomposition, SVD) $\mathbf{A}=\mathbf{U}\mathbf{D}\mathbf{V}'=\sum_{i=1}^{r}d_i\mathbf{u}_i\mathbf{v}_i'$, where $\mathbf{U}=(\mathbf{u}_1,\dots,\mathbf{u}_r)\in\mathcal{O}^{p\times r}$, $\mathbf{V}=(\mathbf{v}_1,\dots,\mathbf{v}_r)\in\mathcal{O}^{q\times r}$, and $\mathbf{D}\in\mathcal{D}^{r\times r}$.

3.2. Elliptical model and independent components model. Let $\mathbf{x}$ be a p-variate vector with the full-rank covariance matrix $\operatorname{Cov}(\mathbf{x})$. We say that $\mathbf{x}$ has a *spherical distribution* if there exists $\boldsymbol{\mu}$ such that $(\mathbf{x}-\boldsymbol{\mu})\sim\mathbf{U}(\mathbf{x}-\boldsymbol{\mu})$ for all orthogonal $\mathbf{U}$. In the following we first define the elliptical and independent components distributions (see, for example, [23], [24] for more details).

Definition 3.1. Let $\mathbf{x}\in\mathbb{R}^p$ be a p-variate random vector.

(1) $\mathbf{x}$ has an *elliptical distribution* if there exists a nonsingular $\mathbf{A}\in\mathbb{R}^{p\times p}$ such that $\mathbf{A}\mathbf{x}$ has a spherical distribution.
(2) $\mathbf{x}$ has an *independent components distribution* if there exists a nonsingular $\mathbf{A}\in\mathbb{R}^{p\times p}$ such that $\mathbf{A}\mathbf{x}$ has independent components.

We next provide some results on how the matrix $\mathbf{A}$ can be found in different cases.

Theorem 3.1. *Let $\mathbf{x}$ be a p-variate random vector with a full-rank covariance matrix $\operatorname{Cov}(\mathbf{x})=\mathbf{U}\mathbf{D}\mathbf{U}'$. Then we have the following:*

(1) $[\mathbf{V}\mathbf{D}^{-1/2}\mathbf{U}']\mathbf{x}$ *has uncorrelated components for all orthogonal* $\mathbf{V}$.
(2) *If* $\mathbf{x}$ *has an elliptical distribution,* $[\mathbf{V}\mathbf{D}^{-1/2}\mathbf{U}']\mathbf{x}$ *has a spherical distribution for all orthogonal* $\mathbf{V}$.
(3) *If* $\mathbf{x}$ *has an independent components distribution,* $[\mathbf{V}\mathbf{D}^{-1/2}\mathbf{U}']\mathbf{x}$ *has independent components for some choice(s) of orthogonal* $\mathbf{V}$.
(4) *If* $\mathbf{x}$ *has both an elliptical distribution and an independent components distribution then* $[\mathbf{V}\mathbf{D}^{-1/2}\mathbf{U}']\mathbf{x}$ *has independent Gaussian components for all orthogonal* $\mathbf{V}$, *that is,* $\mathbf{x}$ *has a multivariate Gaussian distribution.*

3.3. Projection pursuit and independent component analysis. Let $\mathbf{x}$ have an independent components distribution such that $\mathbf{z}=\mathbf{A}\mathbf{x}+\mathbf{b}$ is standardized ($E(\mathbf{z})=\mathbf{0}$ and $\operatorname{Cov}(\mathbf{z})=\mathbf{I}_p$) and has independent components. Theorem 3.1 then implies that $\mathbf{A}=\mathbf{V}'\operatorname{Cov}(\mathbf{x})^{-1/2}$ where the rotation matrix $\mathbf{V}$ can be chosen as $\mathbf{V}=(\mathbf{V}_1,\mathbf{V}_2)$ separating non-Gaussian independent components in $\mathbf{V}_1'\operatorname{Cov}(\mathbf{x})^{-1/2}\mathbf{x}$ and Gaussian independent components in $\mathbf{V}_2'\operatorname{Cov}(\mathbf{x})^{-1/2}\mathbf{x}$. Note that $\mathbf{V}_2$ is only unique up to right multiplication by an orthogonal matrix. A generally accepted strategy is to find $\mathbf{V}_1 = (\mathbf{v}_1, \ldots, \mathbf{v}_q) \in \mathcal{O}^{p\times q}$ such that the components of $\mathbf{V}_1'\mathbf{x}^{\mathrm{st}}$ are 'as non-Gaussian as possible'. The Gaussian part $\mathbf{V}_2' \operatorname{Cov}(\mathbf{x})^{-1/2}\mathbf{x}$ is thought to be just the noise part and, for other components, it is argued that the sum of independent random variables is 'more Gaussian' than the original variables. The noise interpretation of the Gaussian part may be motivated by the following. A random vector has a multivariate normal distribution if and only if all linear combinations of the marginal variables have univariate normal distributions, that is, there are no 'interesting' directions. The normal distribution is the only distribution for which all third and higher cumulants are zero. As seen before, a Gaussian distribution is the distribution with the poorest information among distributions with the same variance (highest entropy, smallest Fisher information). For a thorough discussion of Gaussian distributions, see [14].

Let $D(x)$ then be the *projection index*, i.e., the functional that is used to measure non-Gaussianity. In the one-by-one projection pursuit approach the first direction $\mathbf{v}_1$ ($\mathbf{v}_1'\mathbf{v}_1 = 1$) maximizes $D(\mathbf{v}_1'\mathbf{x}^{\mathrm{st}})$, the second direction $\mathbf{v}_2$ is orthogonal to $\mathbf{v}_1$ ($\mathbf{v}_2'\mathbf{v}_2 = 1, \mathbf{v}_2'\mathbf{v}_1 = 0$) and maximizes $D(\mathbf{v}_2'\mathbf{x}^{\mathrm{st}})$ and so on. After finding $\mathbf{v}_1, \ldots, \mathbf{v}_{j-1}$, we optimize the Lagrangian function

$$L(\mathbf{v}; \lambda_{j1}, \ldots, \lambda_{jj}) = D(\mathbf{v}'\mathbf{x}^{\mathrm{st}}) - \lambda_{jj}(\mathbf{v}'\mathbf{v} - 1) - \sum_{i=1}^{j-1} \lambda_{ji}\mathbf{v}'\mathbf{v}_i.$$

Then $\mathbf{v}_j$ solves the (estimation) equation $\left(\mathbf{I}_p - \sum_{i=1}^{j-1} \mathbf{v}_i\mathbf{v}_i'\right)\mathbf{T}(\mathbf{v}) = (\mathbf{T}(\mathbf{v})'\mathbf{v})\mathbf{v}$, where $\mathbf{T}(\mathbf{v}) = \partial D(\mathbf{v}'\mathbf{x}^{\mathrm{st}})/\partial\mathbf{v}$. From the computational point of view, this suggests a *fixed-point algorithm*. The estimation equation also provides a way to find the limiting distribution of the estimate, since the estimate is obtained when the theoretical multivariate distribution is replaced by the empirical one. See, for example, [20], [21], [22] and references therein for more details.

The following questions naturally arise. How should one choose the projection index $D(x)$ to find the independent components? Are the independent components provided by the most informative directions as has been often stated in the literature? These questions are partially answered by the following.

Theorem 3.2. *Let $\mathbf{z} = \mathbf{A}\mathbf{x} + \mathbf{b} = (z_1, \ldots, z_p)'$ be the vector of standardized independent components.*

(1) *Let $D(x)$ be a subadditive squared dispersion measure. Then $D(\mathbf{v}'\mathbf{x}^{\mathrm{st}}) \leqslant \max_j D(z_j)$.*
(2) *Let $D(x)$ be a superadditive squared dispersion measure. Then $D(\mathbf{v}'\mathbf{x}^{\mathrm{st}}) \geqslant \min_j D(z_j)$.*

Based on Theorem 3.2 and the discussion above we can now end the paper with the following conclusions. If $D(x)$ is subadditive then it can be used as a projection index. For example the cumulants $\kappa_{2k+1}^{2/(2k+1)}(x)$ and $\kappa_{2k+2}^{2/(2k+2)}(x)$, $k = 1, 2, \ldots$, when calculated for standardized distributions, provide squared dispersion measures that are subadditive. Therefore they can be used as projection indices. For superadditive $D(x)$, the functional $(D(x))^{-1}$ is a valid projection index as $(D(\mathbf{v}'\mathbf{x}^{\mathrm{st}}))^{-1} \leqslant \max_j (D(z_j))^{-1}$. As seen before, the entropy power $\mathrm{e}^{H(x)}$ and the inverse of the Fisher information $J^{-1}(x)$ are superadditive squared dispersion measures. Note that in both cases $D(\mathbf{v}'\mathbf{x}^{\mathrm{st}})$ is in fact a ratio of two squared dispersion functions, and the projection index measures the deviation from Gaussianity using a skewness, kurtosis or information measure. As mentioned in Section 3.3, $\frac{1}{12}(\kappa_3^2(x) + \frac{1}{4}\kappa_4^2(x))$ provides an approximation of negative differential entropy in a neighborhood of the Gaussian distribution and is a valid projection index as well. For further discussion, see [10]. Note also that one of the most popular ICA procedures in the engineering community, the so called *fast ICA*, uses a projection index of the form $D(x) = |E[C(x)]|$, where C is such a function that if $z \sim N(0,1)$ then $E[C(z)] = 0$. Examples of valid choices of C are $C(z) = z^3$ and $C(z) = z^4 - 3$ providing again the third and fourth cumulants, respectively. Many other indices have been used; for example [35] discuss how to choose C when the marginal distributions are Gaussian mixtures. Entropy related suggestions for projection pursuit indices are given by [11], [8], [17], [2] just to name a few, while the usage of Fisher information has been discussed in [18], [33], for example.

4. Final remarks

The usage of various information criteria is popular in independent component analysis. The connections between notions of information and statistical independence, and the special role of the Gaussian distribution have been discussed in detail in the paper. We also introduced new ideas and partial orderings for information which utilize transformed location and scale-free probability density functions. In independent component analysis with unknown marginal densities, the estimation of the value of the adapted information measure in a given direction is highly challenging and it has to be done again and again when applying the fixed-point algorithm for the correct direction. Substantial research is therefore still needed for these tools to be of practical value.

5. Appendix: The Proofs

P r o o f of Theorem 2.1. Let $x_1, \dots, x_n$ be a random sample from a distribution of x with the mean value $E(x)$ and variance $\mathrm{Var}(x)$. By the central limit theorem,

$$z_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{x_i - E(x)}{\sqrt{\mathrm{Var}(x)}} \to_d z \sim N(0,1).$$

Therefore, by additivity and affine equivariance,

$$T(z_n) = \sqrt{\frac{n}{\mathrm{Var}(x)}}(T(x) - E(x)) \to 0 \quad \text{and} \quad S^2(z_n) = \frac{S^2(x)}{\mathrm{Var}(x)} \to S^2(z)$$

and the result follows. For similar results in the multivariate case, see [34]. □

P r o o f of Theorem 2.2. The invariances of the measures $H(x)$, $H^*(x)$, $H^{**}(x)$ and $J(x)$ under location shifts $f(x) \to f(x+b)$ and sign change $f(x) \to f(-x)$ follow easily from their definitions and from the definition of the Riemann integral. To show that the measures $H(x)$, $H^*(x)$, $H^{**}(x)$, and $J(x)$ are invariant under the transformation $f \mapsto f_{a,b,\Delta}$ observe that

$$f|_{\mathbb{R}-[a,a+\Delta]-[b,b+\Delta]} = f_{a,b,\Delta}|_{\mathbb{R}-[a,a+\Delta]-[b,b+\Delta]}, \quad f|_{[a,a+\Delta]} = f_{a,b,\Delta}|_{[b,b+\Delta]},$$
$$f|_{[b,b+\Delta]} = f_{a,b,\Delta}|_{[a,a+\Delta]}.$$

Since the measures $H(x)$, $H^*(x)$, $H^{**}(x)$, and $J(x)$ are defined as transformations of $\int g(f(x))\,\mathrm{d}x$ for some function g, where f is the pdf of x, and as the Riemann integral is a functional which is additive over an area of integration, the *entropy power* $\mathrm{e}^{2H(x)}$ and measures $[H^*(x)]^{-2}$, $[H^{**}(x)]^{-2}$, and $[J(x)]^{-1}$ are invariant under the transformations $f \to f_{a,b,\Delta}$.

We therefore have only to consider the rescaling $f(x) \to (1/a)f(x/a)$ with $a > 0$. Then $H(ax) = -\int (1/a)f(x/a)\log((1/a)f(x/a))\,\mathrm{d}x = -\int f(x)\log((1/a)f(x))\,\mathrm{d}x = H(x) + \log(a)$ and therefore $\mathrm{e}^{2H(ax)} = a^2\mathrm{e}^{H(x)}$. In a similar way one can show that $[H^*(ax)]^{-2} = a^2[H^*(x)]^{-2}$ and also easily $[H^{**}(ax)]^{-2} = a^2[H^{**}(x)]^{-2}$. As $f'(x) \to (1/a^2)f'(x/a)$ one further shows that $[J(ax)]^{-1} = a^2[J(x)]^{-1}$. Thus all four measures are scale equivariant and therefore squared dispersion measures. □

P r o o f of Theorem 2.3. The ratio $f : g$ is indeed a density function, since it is trivially nonnegative and

$$\int_0^1 (f:g)(u)\,\mathrm{d}u = \int_0^1 \frac{f(G^{-1}(u))}{g(G^{-1}(u))}\,\mathrm{d}u = \int_{-\infty}^{\infty} f(x)\,\mathrm{d}x = 1$$

with the substitution $x=G^{-1}(u)$. Similarly,

$$-H(f:g)=\int_0^1 (f:g)(u)\log((f:g)(u))\,\mathrm{d}u=\int_{-\infty}^{\infty} f(x)\log\frac{f(x)}{g(x)}\,\mathrm{d}x=D(f||g).$$

$\square$

Proof of Theorem 3.1. (1) Let $\mathbf{V}$ be orthogonal. As $\operatorname{Cov}([\mathbf{V}\mathbf{D}^{-1/2}\mathbf{U}']\mathbf{x}) = \mathbf{V}\mathbf{D}^{-1/2}\mathbf{U}'\operatorname{Cov}(\mathbf{x})\mathbf{U}\mathbf{D}^{-1/2}\mathbf{V}' = \mathbf{V}\mathbf{V}' = \mathbf{I}_p$, the components of $[\mathbf{V}\mathbf{D}^{-1/2}\mathbf{U}']\mathbf{x}$ are uncorrelated.

(2) Assume that $\mathbf{A}\mathbf{x}$ is spherical with $\mathbf{A}=\mathbf{V}\mathbf{C}\mathbf{W}'$ rescaled so that $\operatorname{Cov}(\mathbf{A}\mathbf{x})=\mathbf{I}_p$. As $\mathbf{A}\operatorname{Cov}(\mathbf{x})\mathbf{A}'=\mathbf{I}_p$, then $\operatorname{Cov}(\mathbf{x})=(\mathbf{A}'\mathbf{A})^{-1}$ and $\mathbf{W}\mathbf{C}^{-2}\mathbf{W}'=\mathbf{U}\mathbf{D}\mathbf{U}'$. Therefore, $\mathbf{W}=\mathbf{U}$ and $\mathbf{C}=\mathbf{D}^{-1/2}$ and we can conclude that $[\mathbf{V}\mathbf{D}^{-1/2}\mathbf{U}']\mathbf{x}$ is spherical for any orthogonal $\mathbf{V}$. (If $\mathbf{x}$ is spherical then $\mathbf{V}\mathbf{x}$ is spherical for all orthogonal $\mathbf{V}$.)

(3) Let $\mathbf{A}\mathbf{x}$ with $\mathbf{A}=\mathbf{V}\mathbf{C}\mathbf{W}'$ have independent and standardized components so that $\operatorname{Cov}(\mathbf{A}\mathbf{x})=\mathbf{I}_p$. As in (2), $\mathbf{A}$ must be $\mathbf{V}\mathbf{D}^{-1/2}\mathbf{U}'$ but now for some $\mathbf{V}$ only. (It is not true that if $\mathbf{x}$ has independent standardized components then $\mathbf{V}\mathbf{x}$ has independent components for any choice of $\mathbf{V}$.)

(4) Based on (2) and (3), there exists an $\mathbf{A}=\mathbf{V}\mathbf{D}^{-1/2}\mathbf{U}'$ such that $\mathbf{A}\mathbf{x}$ has a spherical distribution with independent components. Then by the Maxwell-Hershell theorem, $\mathbf{A}\mathbf{x}$ has a multivariate normal distribution. For the proof of the Maxwell-Hershell theorem, see e.g. Proposition 4.11 in [5]. $\square$

Proof of Theorem 3.2. Let $\mathbf{z}=\mathbf{A}\mathbf{x}+P\mathbf{b}=(z_1,\ldots,z_p)'$ be a vector of standardized independent components. By Theorem 3.1, $\mathbf{z}=\mathbf{V}\mathbf{x}^{\mathrm{st}}$ with some orthogonal $\mathbf{V}$. If $\mathbf{u}'\mathbf{u}=1$ then also $(\mathbf{V}\mathbf{u})'(\mathbf{V}\mathbf{u})=1$ and therefore $D(\mathbf{u}'\mathbf{x}^{\mathrm{st}})=D(\mathbf{u}'\mathbf{V}\mathbf{z})\leqslant \sum(\mathbf{V}'\mathbf{u})_i^2 D(z_i)\leqslant \max_j D(z_j)$ for a subadditive squared dispersion measure D and $D(\mathbf{u}'\mathbf{x}^{\mathrm{st}})=D(\mathbf{u}'\mathbf{V}\mathbf{z})\geqslant \sum(\mathbf{V}'\mathbf{u})_i^2 D(z_i)\geqslant \min_j D(z_j)$ for a superadditive squared dispersion measure D. $\square$

Acknowledgement. The authors wish to express their gratitude to the anonymous referee, whose insightful comments greatly helped improving the quality of the manuscript.

References

[1] *A. R. Barron*: Entropy and the central limit theorem. Ann. Probab. *14* (1986), 336–342. zbl MR doi

[2] *A. J. Bell, T. J. Sejnowski*: An information-maximization approach to blind separation and blind deconvolution. Neural Comput. *7* (1995), 1129–1159. doi

[3] *P. J. Bickel, E. L. Lehmann*: Descriptive statistics for nonparametric models. II: Location. Ann. Stat. *3* (1975), 1045–1069. zbl MR doi

[4] *P. J. Bickel, E. L. Lehmann*: Descriptive statistics for nonparametric models. III: Dispersion. Ann. Stat. *4* (1976), 1139–1158. zbl MR doi

[5] *M. Bilodeau, D. Brenner*: Theory of Multivariate Statistics. Springer Texts in Statistics, Springer, New York, 1999. zbl MR doi
[6] *H. Chernoff, I. R. Savage*: Asymptotic normality and efficiency of certain nonparametric test statistics. Ann. Math. Stat. *29* (1958), 972–994. zbl MR doi
[7] *T. M. Cover, J. A. Thomas*: Elements of Information Theory. Wiley Series in Telecommunications, John Wiley & Sons, New York, 1991. zbl MR doi
[8] *L. Faivishevsky, J. Goldberger*: ICA based on a smooth estimation of the differential entropy. NIPS'08: Proceedings of the 21st International Conference on Neural Information Processing Systems. Curran Associates, New York, 2008, pp. 433–440.
[9] *J. L. Hodges,* Jr., *E. L. Lehmann*: The efficiency of some nonparametric competitors of the t-test. Ann. Math. Stat. *27* (1956), 324–335. zbl MR doi
[10] *P. J. Huber*: Projection pursuit. Ann. Stat. *13* (1985), 435–475. zbl MR doi
[11] *A. Hyvärinen*: New approximations of differential entropy for independent component analysis and projection pursuit. NIPS '97: Proceedings of the 1997 Conference on Advances in Neural Information Processing Systems 10. MIT Press, Cambridge, 1998, pp. 273–279.
[12] *A. Hyvärinen, J. Karhunen, E. Oja*: Independent Component Analysis. John Wiley & Sons, New York, 2001. doi
[13] *M. C. Jones, R. Sibson*: What is projection pursuit? J. R. Stat. Soc., Ser. A *150* (1987), 1–36. zbl MR doi
[14] *K. Kim, G. Shevlyakov*: Why Gaussianity? IEEE Signal Processing Magazine *25* (2008), 102–113. doi
[15] *E. Kristiansson*: Decreasing Rearrangement and Lorentz $L(p,q)$ Spaces. Master Thesis. Department of Mathematics, Lulea University of Technology, Lulea, 2002; Available at `http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.111.1244&rep=rep1&type=pdf`.
[16] *S. Kullback*: Information Theory and Statistics. Wiley Publication in Mathematical Statistics, John Wiley & Sons, New York, 1959. zbl MR
[17] *E. G. Learned-Miller, J. W. Fisher III*: ICA using spacings estimates of entropy. J. Mach. Learn. Res. *4* (2004), 1271–1295. zbl MR doi
[18] *B. G. Lindsay, W. Yao*: Fisher information matrix: A tool for dimension reduction, projection pursuit, independent component analysis, and more. Can. J. Stat. *40* (2012), 712–730. zbl MR doi
[19] *A. W. Marshall, I. Olkin*: Inequalities: Theory of Majorization and Its Applications. Mathematics in Science and Engineering 143, Academic Press, New York, 1979. zbl MR doi
[20] *J. Miettinen, K. Nordhausen, H. Oja, S. Taskinen*: Deflation-based FastICA with adaptive choices of nonlinearities. IEEE Trans. Signal Process *62* (2014), 5716–5724. zbl MR doi
[21] *J. Miettinen, K. Nordhausen, H. Oja, S. Taskinen, J. Virta*: The squared symmetric FastICA estimator. Signal Process. *131* (2017), 402–411. doi
[22] *J. Miettinen, S. Taskinen, K. Nordhausen, H. Oja*: Fourth moments and independent component analysis. Stat. Sci. *30* (2015), 372–390. zbl MR doi
[23] *K. Nordhausen, H. Oja*: Independent component analysis: A statistical perspective. WIREs Comput. Stat. *10* (2018), Article ID e1440, 23 pages. MR doi
[24] *K. Nordhausen, H. Oja*: Robust nonparametric inference. Annu. Rev. Stat. Appl. *5* (2018), 473–500. MR doi
[25] *H. Oja*: On location, scale, skewness and kurtosis of univariate distributions. Scand. J. Stat., Theory Appl. *8* (1981), 154–168. zbl MR
[26] *J. E. Pečarić, F. Proschan, Y. L. Tong*: Convex Functions, Partial Orderings, and Statistical Applications. Mathematics in Science and Engineering 187, Academic Press, Boston, 1992. zbl MR doi

[27] *J. V. Ryff*: On the representation of doubly stochastic operators. Pac. J. Math. *13* (1963), 1379–1386. zbl MR doi

[28] *R. Serfling*: Asymptotic relative efficiency in estimation. International Encyclopedia of Statistical Science. Springer, Berlin, 2011, pp. 68–72. doi

[29] *C. E. Shannon*: A mathematical theory of communication. Bell Syst. Tech. J. *27* (1948), 379–423. zbl MR doi

[30] *R. G. Staudte*: The shapes of things to come: Probability density quantiles. Statistics *51* (2017), 782–800. zbl MR doi

[31] *R. G. Staudte, A. Xia*: Divergence from, and convergence to, uniformity of probability density quantiles. Entropy *20* (2018), Article ID 317, 10 pages. MR doi

[32] *W. R. van Zwet*: Convex Transformations of Random Variables. Mathematical Centre Tracts 7, Mathematisch Centrum, Amsterdam, 1964. zbl MR

[33] *V. Vigneron, C. Jutten*: Fisher information in source separation problems. International Conference on Independent Component Analysis and Signal Separation. Lecture Notes in Computer Science 3195, Springer, Berlin, 2004, pp. 168–176. doi

[34] *J. Virta*: On characterizations of the covariance matrix. Available at `https://arxiv.org/abs/1810.01147` (2018), 11 pages.

[35] *J. Virta, K. Nordhausen*: On the optimal non-linearities for Gaussian mixtures in FastICA. Latent Variable Analysis and Signal Separation. Theoretical Computer Science and General Issues, Springer, Berlin, 2017, pp. 427–437. doi

Authors' addresses: *Una Radojičić* (corresponding author), *Klaus Nordhausen*, CSTAT-Computational Statistics, Institute of Statistics & Mathematical Methods in Economics, Vienna University of Technology, Wiedner Hauptstraße 7, 1040 Vienna, Austria, e-mail: `una.radojicic@tuwien.ac.at`, `klaus.nordhausen@tuwien.ac.at`; *Hannu Oja*, Department of Mathematics and Statistics, University of Turku, 20014 Turku, Finland, e-mail: `hannu.oja@utu.fi`.