On Solutions to Multivariate Maximum α-entropy Problems

J. Costa[†], A. Hero[†] and C. Vignat[*]

[†]University of Michigan, Ann Arbor, MI 48109-2122, USA
[†]Université Marne la Vallée, France

Abstract. Entropy has been widely employed as an optimization function for problems in computer vision and pattern recognition. To gain insight into such methods it is important to characterize the behavior of the maximum-entropy probability distributions that result from the entropy optimization. The aim of this paper is to establish properties of multivariate distributions maximizing entropy for a general class of entropy functions, called Renyi's α-entropy, under a covariance constraint. First we show that these entropy-maximizing distributions exhibit interesting properties, such as spherical invariance, and have a stochastic Gaussian-Gamma mixture representation. We then turn to the question of stability of the class of entropy-maximizing distributions under addition.

1 Introduction

Entropy has been widely employed as an optimization function for problems in computer vision, image reconstruction, communications, clustering, and pattern recognition, see [2, 3, 6, 5, 8] for representative examples. Entropy has also played a role in statistical physics where physical probability laws are deduced from maximum entropy principles [6]. Among many other applications, maximum entropy models have been successful in describing the distribution of the interior solar plasma [1], the behavior of dissipative, low dimensional chaotic systems [2] and self-gravitating systems [3]. Studying the properties of entropy maximizing distributions is important for understanding the advantages and limitations of entropy maximization methods. In this paper we give an overview of properties of multivariate distributions maximizing entropy for a general class of entropy functions, called Renyi's α-entropy, under a covariance constraint.

The Renyi α-entropy [7] is a generalization of the Shannon entropy and is defined as follows:

$$S_\alpha(f) = \frac{1}{1-\alpha} \log \int_{\mathbb{R}^n} f^\alpha(\mathbf{x})\, d\mathbf{x} \tag{1}$$

where f is the n-variate probability distribution of the n-dimensional random variable X, and α is a real positive parameter. It can be easily shown that as $\alpha \to 1$ the α-entropy S_α converges to the well known Shannon entropy

$$S_1(f) \stackrel{\Delta}{=} \lim_{\alpha \to 1} S_\alpha(f) = -\int_{\mathbb{R}^n} f(\mathbf{x}) \log f(\mathbf{x})\, d\mathbf{x}. \tag{2}$$

It is well-known that among all multivariate continuous distributions, the classical Gaussian distribution maximizes the Shannon entropy under a covariance (power) constraint. The question addressed in this paper is the characterization of the maximizing distribution of Renyi entropy under the same covariance constraint: this distribution will be called Renyi distribution in the following. We show below that the multivariate Student-t ($\alpha < 1$) and Student-r ($\alpha > 1$) densities are the maximum entropy distributions under a covariance constraint for different ranges of the parameter α. In addition they satisfy the following properties: 1) the multivariate α-entropy maximizing distributions maximize the Shannon entropy under a logarithmic constraint; 2) these distributions satisfy can be represented as Gaussian scale mixtures; 3) these distributions are spherically invariant.

2 The multivariate α-entropy maximizing distribution

Renyi-entropy maximizing distributions have been studied for the restricted case of $\alpha > 1$, by Moriguti in the scalar case [4] and by Kapur [5] in the multivariate case. The case of $\alpha \in [0, 1]$ is of special interest since in this region the Renyi-entropy generalizes easily to Renyi-divergence via measure transformation [4].

Throughout, $\mathbf{x}$ will denote an $n-$dimensional real random vector with covariance matrix

$$\begin{aligned}\mathbf{K} &= \mathbf{E}\left(\mathbf{x}-\mu_X\right)\left(\mathbf{x}-\mu_X\right)^T \\ &= \int_{\mathbb{R}^n}\left(\mathbf{x}-\mu_X\right)\left(\mathbf{x}-\mu_X\right)^T f_X\left(\mathbf{x}\right)d\mathbf{x}\end{aligned}$$

In what follows, we consider, without loss of generality, the centered case $\mu_X = 0$. Define next the following constants:

$$m = \begin{cases} n + \frac{2}{\alpha - 1} \text{ if } \alpha > 1 \\ \frac{2}{1-\alpha} - n \text{ if } \alpha < 1 \end{cases}$$

$$\mathbf{C}_\alpha = \begin{cases} (m+2)\,\mathbf{K} \text{ if } \alpha > 1 \\ (m-2)\,\mathbf{K} \text{ if } \alpha < 1 \end{cases}$$

and

$$A_\alpha = \begin{cases} \frac{1}{|\pi \mathbf{C}_\alpha|^{\frac{1}{2}}} \frac{\Gamma\left(\frac{m}{2}+1\right)}{\Gamma\left(\frac{m-n}{2}+1\right)} \text{ if } \alpha > 1 \\ \frac{1}{|\pi \mathbf{C}_\alpha|^{\frac{1}{2}}} \frac{\Gamma\left(\frac{m+n}{2}\right)}{\Gamma\left(\frac{m}{2}\right)} \quad \text{if } \frac{n}{n+2} < \alpha < 1 \end{cases}$$

and the following sets

$$\Omega_\alpha = \begin{cases} \left\{\mathbf{x} \in \mathbb{R}^{\mathbf{n}} \,:\, \mathbf{x}^T \mathbf{C}_\alpha^{-1} \mathbf{x} \leq 1\right\} \text{ if } \alpha > 1 \\ \mathbb{R}^n \qquad\qquad\qquad\qquad \text{if } \frac{n}{n+2} < \alpha < 1 \end{cases}$$

Define the $n-$variate probability density f_α as follows:

– if $\alpha > 1$

$$f_\alpha(\mathbf{x}) = \begin{cases} A_\alpha \left(1 - \mathbf{x}^T \mathbf{C}_\alpha^{-1} \mathbf{x}\right)^{\frac{1}{\alpha-1}} & \text{if } \mathbf{x} \in \Omega_\alpha \\ 0 & \text{else} \end{cases} \tag{3}$$

– if $\frac{n}{n+2} < \alpha < 1$

$$f_\alpha(\mathbf{x}) = A_\alpha \left(1 + \mathbf{x}^T \mathbf{C}_\alpha^{-1} \mathbf{x}\right)^{\frac{1}{\alpha-1}} \quad \forall \mathbf{x} \in \mathbb{R}^n \tag{4}$$

The following theorem provides a general description of the $\alpha-$entropy maximizing densities.

Theorem 1. *the functions f_α defined by equations (3) and (4) are the unique solutions of the following problem*

$$f_\alpha = \arg \max_{f : \int_{\Omega_\alpha} \mathbf{x}\mathbf{x}^T f(\mathbf{x}) d\mathbf{x} = \mathbf{K}} S_\alpha(f) \tag{5}$$

We prove this theorem by introducing a new divergence measure and adopting an information theoretic approach similar to that used by [7, p.234 theorem 6.9.5]) to prove that the Gaussian distribution maximizes Shannon entropy.

Consider the following non-symmetric directed divergence measure

$$D_\alpha(f||g) = \operatorname{sign}(\alpha - 1) \int_{\Omega_\alpha} \frac{f^\alpha}{\alpha} + \frac{\alpha - 1}{\alpha} g^\alpha - f g^{\alpha-1} \tag{6}$$

The general theory of dirrected divergence measures is discussed in [12] and [13]. Convexity of D_α gives the following positivity property: for any two probability densities f and g, we have

$$D_\alpha(f||g) \geq 0$$

with equality if and only if

$$f = g \text{ a.e.}$$

Proposition 1. *For any $n-$variate probability density f with covariance matrix* $\mathbf{K}$,

$$\int_{\Omega_\alpha} f_\alpha^{\alpha-1} (f - f_\alpha) = 0 \tag{7}$$

Proof. suppose for example $\alpha > 1$ then

$$\int_{\Omega_\alpha} f(\mathbf{x}) f_\alpha^{\alpha-1}(\mathbf{x}) d\mathbf{x} = \int_{\Omega_\alpha} f(\mathbf{x}) A_\alpha^{\alpha-1} \left(1 - \mathbf{x}^T \mathbf{C}_\alpha^{-1} \mathbf{x}\right) d\mathbf{x}$$

But f and f_α have the same covariance matrix, so that

$$\int_{\Omega_\alpha} \mathbf{x}^T \mathbf{C}_\alpha^{-1} \mathbf{x} f(\mathbf{x}) d\mathbf{x} = \int_{\Omega_\alpha} \mathbf{x}^T \mathbf{C}_\alpha^{-1} \mathbf{x} f_\alpha(\mathbf{x}) d\mathbf{x}$$

and finally

$$\int_{\Omega_\alpha} f(\mathbf{x}) f_\alpha^{\alpha-1}(\mathbf{x})\, d\mathbf{x} = \int_{\Omega_\alpha} f_\alpha A_\alpha^{\alpha-1}\left(1-\mathbf{x}^T\mathbf{C}_\alpha^{-1}\mathbf{x}\right) d\mathbf{x}$$
$$= \int_{\Omega_\alpha} f_\alpha^\alpha$$

The proof is similar in the case $\alpha < 1$ and will be omitted.

Observe that for $\alpha = 1$, the orthogonality property (7) implies

$$\int f \log f_1 = \int f_1 \log f_1$$

Now we can deduce the extremal property of the density f_α.

Theorem 2. *For any probability density f with covariance matrix $\mathbf{K}$ and any $\alpha > \frac{n}{n+2}$*

$$S_\alpha(f) \leq S_\alpha(f_\alpha)$$

with equality if and only if $f = f_\alpha$ almost everywhere.

Proof. suppose for example $\alpha > 1$ then

$$D_\alpha(f||f_\alpha) = \int_{\Omega_\alpha} \left(\frac{f^\alpha}{\alpha} + \frac{\alpha-1}{\alpha} f_\alpha^\alpha - f_\alpha^{\alpha-1} f\right)$$

but, from the orthogonality property:

$$\int_{\Omega_\alpha} f_\alpha^\alpha = \int_{\Omega_\alpha} f_\alpha^{\alpha-1} f$$

and thus

$$0 \leq D_\alpha(f||f_\alpha) = \int_{\Omega_\alpha} \left(\frac{f^\alpha}{\alpha} + \frac{\alpha-1}{\alpha} f_\alpha^\alpha - f_\alpha^\alpha\right)$$
$$= \frac{1}{\alpha}\int_{\Omega_\alpha} (f^\alpha - f_\alpha^\alpha) = \frac{\alpha-1}{\alpha}\left(S_\alpha(f_\alpha) - S_\alpha(f)\right)$$

The proof is similar in the case $\alpha < 1$.

Note that theorem 2 implies that the entropy $S_\alpha(f)$ has a unique maximizer f_α.

Definition 1. *a distribution is called elliptically invariant if it has the form*

$$p_{\mathbf{X}}(\mathbf{x}) = \phi_{\mathbf{X}}\left(\mathbf{x}^T\mathbf{C}^{-1}\mathbf{x}\right) \tag{8}$$

for some function $\phi_{\mathbf{X}} : \mathbb{R}^+ \to \mathbb{R}^+$ and some positive definite matrix $\mathbf{C}$, called the characteristic matrix.

It is easily seen that f_α, defined by equations (3) and (4), is an elliptically invariant density. A consequence of this elliptical invariance property is that if X is a random vector with density f_α having $\alpha < 1$, then it can be represented as a Gaussian scale mixture [10]):

$$\mathbf{X} = \mathbf{A}\mathbf{N} \tag{9}$$

where $\mathbf{A}$ is a scalar Chi-square random variable with $m = \frac{2}{1-\alpha} - n$ degrees of freedom and $\mathbf{N}$ is an n-variate Gaussian random vector independent of $\mathbf{A}$ and with covariance matrix $\mathbf{C}_\alpha$. For more details see [1]. Rewriting equivalently $\mathbf{X}$ as

$$\mathbf{X} = \frac{\mathbf{C}_\alpha^{\frac{1}{2}}\mathbf{N}_0}{\sqrt{\sum_{i=1}^{m} N_i^2}} \tag{10}$$

where $\mathbf{N}_0$ is a zero mean Gaussian random vector with covariance matrix $\mathbf{I}_n$, it is evident that, by Slutzky's theorem, as $m \to +\infty$

$$\frac{\mathbf{C}_\alpha^{\frac{1}{2}}}{\sqrt{\sum_{i=1}^{m} N_i^2}} = \frac{\mathbf{K}^{\frac{1}{2}}}{\sqrt{\frac{1}{m-2}\sum_{i=1}^{m} N_i^2}} \tag{11}$$

converges a.s. to the constant matrix $\mathbf{K}^{\frac{1}{2}}$ and $\mathbf{X}$ converges to a Gaussian random vector.

Although the Gaussian scale mixture representation does not hold in the case $\alpha > 1$, we can extend the stochastic representation based on the existence of a natural bijection between the cases $\alpha < 1$ and $\alpha > 1$ gives the following theorem.

Proposition 2. *If* $\mathbf{X}$ *is an* $n-$*variate random vector distributed according to* f_α *with* $\alpha > 1$, *and if* m *defined as*

$$\alpha = \frac{m+n}{m+n-2} \tag{12}$$

is an integer not equal to zero, then $\mathbf{X}$ *has the representation*

$$\mathbf{X}_\alpha = \mathbf{C}_\alpha^{\frac{1}{2}} \frac{\mathbf{N}}{\sqrt{\|\mathbf{N}\|_2^2 + N_1^2 + \cdots + N_m^2}} \tag{13}$$

where $\{N_i\}_{1 \le i \le m}$ *are Gaussian* $\mathcal{N}(O,1)$ *mutually independent, and independent of* $\mathbf{N}$ *which is Gaussian* $\mathcal{N}(O, \mathbf{I}_n)$.

We remark here that the denominator in (13) is a chi random variable with $m+n$ degrees of freedom which, contrarily to the case $\alpha < 1$, is not independent of the numerator. Using these stochastic representations, random samples from f_α with integer degrees of freedom can be easily implemented with a Gaussian random number generator and a squarer.

Characteristic function The characteristic function φ_α of f_α can be deduced from the following formula [9]:

$$\varphi_\alpha(\mathbf{u}) = \mathcal{L}\left[w^{-2} f_W\left(w^{-1}\right)\right]_{s=\mathbf{u}^T \mathbf{C}_\alpha \mathbf{u}}$$

where $\mathcal{L}$ denotes the Laplace transform.

a- Case $\alpha < 1$ As, in the case $\alpha < 1$ ([17]),

$$\mathcal{L}\left[w^{-2} f_W\left(w^{-1}\right)\right] = \frac{2^{1-\frac{m}{2}}}{\Gamma\left(\frac{m}{2}\right)} s^{\frac{m}{2}} K_{\frac{m}{2}}(s)$$

the characteristic function of the Renyi distribution writes

$$\varphi_\alpha(\mathbf{u}) = \frac{2^{1-\frac{m}{2}}}{\Gamma\left(\frac{m}{2}\right)} \left(\mathbf{u}^T \mathbf{C}_\alpha \mathbf{u}\right)^{\frac{m}{2}} K_{\frac{m}{2}}\left(\mathbf{u}^T \mathbf{C}_\alpha \mathbf{u}\right) \tag{14}$$

where $K_{\frac{m}{2}}$ denotes the modified Bessel function of the second kind.

b- Case $\alpha > 1$ Although the preceding technique does not apply in the case $\alpha > 1$, a direct computation yields the characteristic function in this case as

$$\varphi_\alpha(\mathbf{u}) = 2^{\frac{m}{2}} \Gamma\left(\frac{m}{2} + 1\right) \left(\mathbf{u}^T \mathbf{C}_\alpha \mathbf{u}\right)^{-\frac{m}{2}} J_{\frac{m}{2}}\left(\mathbf{u}^T \mathbf{C}_\alpha \mathbf{u}\right) \tag{15}$$

where $J_{\frac{m}{2}}$ denotes the Bessel function of the first kind.

We remark that both families of characteristic functions (14) and (15) are normalized in such a way that

$$\varphi_\alpha(\mathbf{u}) = 1 + O\left(\left(\mathbf{u}^T \mathbf{C}_\alpha \mathbf{u}\right)^2\right)$$

Moreover, it can be checked that, as $\alpha \to 1$, these functions converge to the classical Gaussian characteristic function.

2.1 An alternative entropic characterization

The Renyi-entropy maximizing distribution can be characterized as a Shannon entropy maximizer under a logarithmic constraint: this property was first derived by Kapur in his seminal paper [5]. It was remarked also by Zografos [14] in the multivariate case, but not connected to the Renyi entropy. We state here an extension of Kapur's main result to the correlated case. This result can be proven using the stochastic representation (see [1] for details).

Theorem 3. *f_α with $\alpha < 1$ (resp. $\alpha > 1$) and characteristic matrix $\mathbf{C}_\alpha$ is the solution of the following optimization problem*

$$f_\alpha = \arg\max_f S_1(f)$$

under constraint

$$\int_{\Omega_\alpha} \log\left(1+\mathbf{x}^T\mathbf{C}_\alpha^{-1}\mathbf{x}\right) f(\mathbf{x})\, d\mathbf{x} = \psi\left(\frac{m+n}{2}\right) - \psi\left(\frac{m}{2}\right) \tag{16}$$

(resp.

$$\int_{\Omega_\alpha} \log\left(1-\mathbf{x}^T\mathbf{C}_\alpha^{-1}\mathbf{x}\right) f(\mathbf{x})\, d\mathbf{x} = \psi\left(\frac{m}{2}\right) - \psi\left(\frac{m+n}{2}\right))$$

where ψ is the digamma function.

We make the following observations. Firstly, the constraint in this multivariate optimization problem is real-valued, and its value is independent of the characteristic matrix $\mathbf{C}_\alpha$. Secondly, as the logarithmic moment $E\log\left(1+\mathbf{X}^T\mathbf{C}_\alpha^{-1}\mathbf{X}\right)$ exists whenever $\alpha > 0$, the distributions f_α as defined by (4) are solutions of the logarithmically constrained maximum Shannon entropy problem even in the case $\alpha < \frac{n}{n+2}$. However, in this case the covariance matrix does not exist and therefore the matrix $\mathbf{C}_\alpha$ can not be interpreted as a covariance matrix.

3 Convolution of entropy maximizing distributions

We first discuss the issue of renormalization as presented by Mendes et al. [15]. Then we address the issue of stability by addition.

3.1 Renormalizability of f_α

Mendes and Tsallis ([15]) have shown that Renyi distributions have the important property of "renormalizability," but contrarily to the Gaussian case, they are not "factorizable." f_α has the renormalizability property when

$$\int_{-\infty}^{+\infty} f_\alpha(x_1, x_2)\, dx_2 = f_{\alpha'}(x_1)$$

for some α'. In statistical terms, this expresses the fact that the $2-$dimensional distributions remain of the same type after marginalization. Using the elliptical invariance property, we provide here a much more general result, as stated by the following theorem.

Theorem 4. *If $\mathbf{X}$ is an $n-$variate Renyi random variable with index α and characteristic matrix $\mathbf{C}$, if $\mathbf{X}_\alpha^T = \left[\mathbf{X}_1^T, \mathbf{X}_2^T\right]$ $(\dim \mathbf{X}_i = n_i, n_1+n_2 = n)$ and $\mathbf{C}$ is partitioned accordingly as $\mathbf{C} = [\mathbf{C}_{11}, \mathbf{C}_{12;}\mathbf{C}_{21}, \mathbf{C}_{22}]$ $(\dim \mathbf{C}_{ij} = n_i \times n_j)$, then the marginal distribution of vector $\mathbf{X}_i$ $(i = 1, 2)$ is nothing but a Renyi distribution with index α_i such that*

$$\frac{1}{1-\alpha_i} = \frac{1}{1-\alpha} - \frac{n_i}{2}$$

and characteristic matrix $\mathbf{C}_{ii}$

Proof. suppose first $\alpha < 1$ and consider the stochastic representation

$$\mathbf{X} = \mathbf{C}^{\frac{1}{2}} \frac{\left[\mathbf{N}_1^T, \mathbf{N}_2^T\right]^T}{\chi_m}$$

where $\left[\mathbf{N}_1^T, \mathbf{N}_2^T\right]$ is a Gaussian vector with identity covariance and partitioned according to $\mathbf{X}$. Then the stochastic representation of $\mathbf{X}_i$ is

$$\mathbf{X}_i = \frac{\tilde{\mathbf{N}}_i}{\chi_m}$$

for some n_i-variate Gaussian vector $\tilde{\mathbf{N}}_i$ so that the indices α and α_i are characterized by

$$\alpha = \frac{m+n-2}{m+n}, \quad \alpha_i = \frac{m+n_i-2}{m+n_i}$$

thus

$$\frac{1}{1-\alpha_i} = \frac{1}{1-\alpha} - \frac{n_i}{2}$$

The characteristic matrix of $\mathbf{X}_i$ can be deduced by remarking that $\mathbf{X}_i$ can be expressed as

$$\mathbf{X}_i = \mathbf{H}\mathbf{X}$$

where $\mathbf{H}$ is a $n_i \times n$ matrix whose $i-th$ block is the $n_i \times n_i$ identity matrix so that the characteristic matrix of $\mathbf{X}_i$ writes (see [9, corollary 3.2])

$$\mathbf{H}\mathbf{C}\mathbf{H}^T = \mathbf{C}_{ii}$$

The case $\alpha > 1$ writes accordingly.

Thus the renormalization property as observed in [15] is nothing but a consequence of the elliptical invariance property, which is itself induced by the orthogonal invariance of both the Renyi entropy and the covariance constraint.

3.2 Stability of Renyi distributions

It is well known that the Gaussian distributions are stable in the sense that the sum of two Gaussian random vectors is also Gaussian, although with possibly different means and variances. An interesting question is the stability of the class of Renyi-entropy maximizing distributions defined as the set of all densities f_α of the form (3)-(4) for some $\alpha \in (0, 1]$ and some positive definite characteristic matrix $\mathbf{C}_\alpha$. In the following, we characterize the conditions under which stability of the Renyi-entropy maximizing distributions is ensured, and link this feature with their elliptical invariance property, distinguishing between three important cases: the Renyi mutually dependent case, the mutually independent case and the special case of odd degrees of freedom. For proofs of these results see the referenced article or [1].

Mutually dependent case

Theorem 5 ([9]). *If* $\mathbf{X}_1$ *and* $\mathbf{X}_2$ *are* n_1 *and* n_2*-variate vectors mutually distributed according to a Renyi-entropy maximizing density with index* α *and characteristic matrix* $\mathbf{C}_\alpha$*, and if* $\mathbf{H}$ *is a* $n' \times n$ *matrix with* $n = n_1 + n_2$*, then the* n'*-variate vector*

$$\mathbf{Z} = \mathbf{H} \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{bmatrix}$$

is distributed according to a Renyi-entropy maximizing density f_α *with index* α' *and characteristic matrix* $\mathbf{C}_{\alpha'}$ *such that*

$$\mathbf{C}_{\alpha'} = \mathbf{H}\mathbf{C}_\alpha\mathbf{H}^T$$
$$\frac{1}{1-\alpha'} = \frac{1}{1-\alpha} + \frac{n'-n}{2}$$

Independent Renyi-entropy maximizing random variables

Theorem 6 ([18]). *If* X *and* Y *are two scalar Renyi random variables with index* α *then* $Z = X + Y$ *is* ***nearly*** *Renyi with index* α' *such that*

$$\alpha' = 2 - (2-\alpha)\left(1 - 4\frac{\alpha(\alpha-1)}{(3\alpha-5)(\alpha+3)}\right) \tag{17}$$

The relative mean square error of this approximation is numerically bounded by 10^{-5}.

The relation (17) was obtained in [18] by evaluating all derivatives up to order 5 at point 0 of the distribution of $X+Y$ and showing that they are nearly identical (up to numerical precision of the simulations) to those of a Renyi-entropy maximizing distribution with parameter α'. In the case where m is an odd integer stronger results can be established. For the sake of clarity, we denote in the following by $f^{(m)}$ a Renyi-entropy maximizing distribution with m degrees of freedom [1].

The first original result we state now is an extension to the multivariate case of the classical one-dimensional result, for which a rich literature already exists (see for example [19],[20]).

Theorem 7. *Suppose that* $\mathbf{X}$ *and* $\mathbf{Y}$ *are two independent* n*-variate* ($\alpha < 1$) *random vectors from* $f^{(m)}$ *with respective characteristic matrices* $\mathbf{C}_\mathbf{X} = \mathbf{C}_\mathbf{Y} = \mathbf{I}_n$ *and* ***odd*** *degrees of freedom* $m_\mathbf{X}$ *and* $m_\mathbf{Y}$*. Then if* $0 \leq \beta \leq 1$*, the distribution of* $\mathbf{Z} = \beta\mathbf{X} + (1-\beta)\mathbf{Y}$ *is*

$$p_\mathbf{Z}(\mathbf{z}) = \sum_{k=0}^{k_\mathbf{Z}} \alpha_k f^{(2k+1)}(\mathbf{z}) \tag{18}$$

where $k_\mathbf{Z} \leq \frac{m_\mathbf{X}+m_\mathbf{Y}}{2} - 1$.

[1] thus $f^{(m)}(\mathbf{x}) = f_q(\mathbf{x})$ with $q = \frac{m+n-2}{m+n}$

Proof. denote $k_{\mathbf{X}} \in \mathbb{N}$ *such that, by hypothesis,* $m_{\mathbf{X}} = 2k_{\mathbf{X}}+1$, *and* $k_{\mathbf{Y}}$ *accordingly. The characteristic function of* $\mathbf{X}$ *in this special case writes*

$$\phi_{\mathbf{X}}(\mathbf{u}) = e^{-\|\mathbf{u}\|} Q_{k_{\mathbf{X}}}(\|\mathbf{u}\|)$$

where $\|\mathbf{u}\| = \sqrt{\mathbf{u}^T \mathbf{u}}$ *and* $Q_{k_{\mathbf{X}}}$ *is a polynomial of degree* $d(Q_{k_{\mathbf{X}}}) = k_{\mathbf{X}}$. *By the independence assumption, the characteristic function of* $\mathbf{Z}$ *writes*

$$\begin{aligned}\phi_{\mathbf{Z}}(\mathbf{u}) &= \phi_{\mathbf{X}}(\beta\mathbf{u})\,\phi_{\mathbf{Y}}((1-\beta)\mathbf{u}) \\ &= e^{-|\beta|\|\mathbf{u}\|} Q_{k_{\mathbf{X}}}(\beta\|\mathbf{u}\|)\, e^{-|1-\beta|\|\mathbf{u}\|} Q_{k_{\mathbf{Y}}}((1-\beta)\|\mathbf{u}\|) \\ &= e^{-\|\mathbf{u}\|} Q_{k_{\mathbf{X}}}(\beta\|\mathbf{u}\|)\, Q_{k_{\mathbf{Y}}}((1-\beta)\|\mathbf{u}\|)\end{aligned}$$

As each polynomial Q_k *has exactly degree* k, *the set of polynomials* $\{Q_l\}_{0\le k\le k_{\mathbf{Z}}}$ *is a basis of the linear space of polynomials with degree lower or equal to* $k_X + k_Y$: *thus,* $Q_{k_{\mathbf{X}}}(\beta\|\mathbf{u}\|)\, Q_{k_{\mathbf{Y}}}((1-\beta)\|\mathbf{u}\|)$, *itself a polynomial of degree* $k_{\mathbf{Z}} \le k_{\mathbf{X}} + k_{\mathbf{Y}} = \frac{m_X+m_Y}{2} - 1$, *can be expressed in a unique way in this basis and consequently, there exists a unique set* $\{\alpha_k\}_{0\le k\le k_{\mathbf{Z}}}$ *of real numbers such that*

$$Q_{k_{\mathbf{X}}}(\beta\|\mathbf{u}\|)\, Q_{k_{\mathbf{Y}}}((1-\beta)\|\mathbf{u}\|) = \sum_{k=0}^{k_{\mathbf{Z}}} Q_k(\|\mathbf{u}\|)$$

so that

$$\phi_{\mathbf{Z}}(\mathbf{u}) = e^{-\|\mathbf{u}\|} \sum_{k=0}^{k_{\mathbf{Z}}} \alpha_k Q_k(\|\mathbf{u}\|)$$

and the result (18) is obtained by inverse Fourier transform. Note that coefficients $\{\alpha_k\}$ *depend on* β.

This result can be restated as follows: the distribution of a convex linear combination of independent Renyi-entropy maximizing random variables with odd degrees of freedom is distributed according to a **discrete scale mixture** of Renyi-entropy maximizing distributions with odd degrees of freedom. However, although the fact that

$$\sum_{k=0}^{k_{\mathbf{Z}}} \alpha_k = 1$$

holds trivially by integrating relation (18) over $\mathbb{R}$, the positiveness of coefficients α_k was, to our best knowledge, never proved in the literature. We are currently working on this conjecture, for which numerical simulations show however that it holds with high probability.

A second result: an information projection property The second result that we propose in this context allows to characterize the Renyi distribution that is the closest to a convolution of $f^{(m)}$' s with odd degrees of freedom.

Theorem 8. *Consider* $\mathbf{X}$ *and* $\mathbf{Y}$ *two independent* n*-variate random vectors following density* f_α *with* $\alpha < 1$*, characteristic matrices* $\mathbf{C_X} = \mathbf{C_Y} = \mathbf{I}_n$ *and* ***odd*** *degrees of freedom* $m_\mathbf{X}$ *and* $m_\mathbf{Y}$*, and* $\mathbf{Z} = \frac{1}{2}(\mathbf{X} + \mathbf{Y})$*. Then the Renyi distribution which is the closest to the distribution of* $\mathbf{Z}$ *in the sense of the Kullback-Leibler divergence has degrees of freedom* m' *such that*

$$w_n(m') = E w_n[M] \tag{19}$$

where

- *function* w_n *is defined as*

$$w_n(m) = \psi\left(\frac{m+n}{2}\right) - \psi\left(\frac{m}{2}\right)$$

- *the random variable* M *is distributed according to*

$$\Pr\{M = 2k+1\} = \alpha_k \tag{20}$$

  where coefficients α_k *are defined by (18) for* $\beta = \frac{1}{2}$*.*

Moreover, condition (19) is equivalent to

$$E_{f^{(m')}} \log\left(1 + \mathbf{x}^T\mathbf{x}\right) = E_{f_Z} \log\left(1 + \mathbf{x}^T\mathbf{x}\right)$$

Proof. the Kullback-Leibler distance between the distribution p_Z *of* Z *and a Renyi distribution* $f^{(m')}$ *with parameter* m' *writes*

$$\begin{aligned} D\left(p_\mathbf{Z} || f^{(m')}\right) &= \int p_\mathbf{Z} \log \frac{p_\mathbf{Z}}{f^{(m')}} \\ &= -H_1(p_\mathbf{Z}) - \int p_\mathbf{Z} \log f^{(m')} \end{aligned}$$

where H_1 *denotes the Shannon entropy. The distribution* $p_\mathbf{Z}$ *takes the form*

$$p_\mathbf{Z}(\mathbf{z}) = \sum_{k=0}^{k_\mathbf{Z}} \alpha_k f^{(2k+1)}(\mathbf{z})$$

with $k_\mathbf{Z} = \frac{m_X + m_Y}{2} - 1$*. Finding the optimal value of* m' *is thus equivalent to maximizing the integral* $\int p_\mathbf{Z} \log f^{(m')}$ *that can be explicitly computed using a result obtained by Zografos [14]: if* $\mathbf{X} \sim f_m$ *then* [2]

$$E \log\left(1 + \mathbf{X}^T\mathbf{X}\right) = w_n(m) \stackrel{\Delta}{=} \mathrm{Psi}\left(\frac{m+n}{2}\right) - \mathrm{Psi}\left(\frac{m}{2}\right)$$

[2] function $w_n(m)$ is denoted as $w_2(m,n)$ in [14]

Thus

$$\begin{aligned}
\int p_{\mathbf{Z}} \log f^{(m')} &= \int \sum_{k=0}^{m_Z} \alpha_k f^{(2k+1)}(\mathbf{z}) \log f^{(m')}(\mathbf{z})\, d\mathbf{z} \\
&= \sum_{k=0}^{m_Z} \alpha_k \int f^{(2k+1)} \log A_{\alpha'} \left(1+\mathbf{z}^T\mathbf{z}\right)^{-\frac{m'+n}{2}} d\mathbf{z} \\
&= \sum_{k=0}^{m_Z} \alpha_k \log A_{\alpha'} - \frac{m'+n}{2} \sum_{k=0}^{m_Z} \alpha_k E_{f^{(2k+1)}}\left(1+\mathbf{Z}^T\mathbf{Z}\right) \\
&= \log \frac{\Gamma\left(\frac{m'+n}{2}\right)}{\Gamma\left(\frac{1}{2}\right)\Gamma\left(\frac{m'}{2}\right)} - \frac{m'+n}{2} \sum_{k=0}^{m_Z} \alpha_k w_n(2k+1)
\end{aligned}$$

Taking the derivative and equating to zero yields

$$w_n(m') = E w_n(M)$$

where M is distributed according to (20). The fact that m' corresponds to a maximum of the considered integral (and thus to a minimum of the Kullback-Leibler distance) is a direct consequence of the negativity of the second derivative of the ψ function, together with

$$\psi'(\frac{m'+n}{2}) - \psi'(\frac{m'}{2}) = \frac{\partial^2}{\partial m'^2} \int p_Z \log f^{(m')}$$

Finally, computing

$$\begin{aligned}
E_{f^{(m')}} \log\left(1+\mathbf{Z}^T\mathbf{Z}\right) &= w_n(m') \\
&= \sum_{k=0}^{m_Z} \alpha_k w_n(2k+1) \\
&= \sum_{k=0}^{m_Z} \alpha_k E_{f^{(2k+1)}} \log\left(1+\mathbf{Z}^T\mathbf{Z}\right) \\
&= E_{f_Z} \log\left(1+\mathbf{Z}^T\mathbf{Z}\right)
\end{aligned}$$

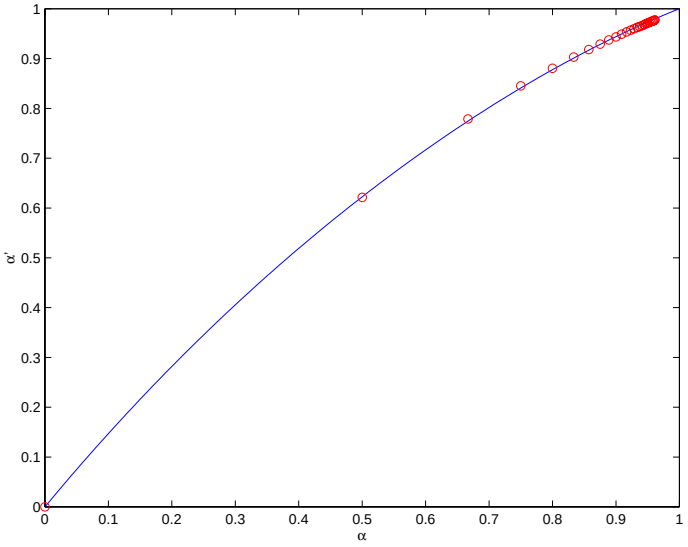
yields the final result.

The equation (19) defining variable m' in terms of dimension n and degrees of freedom m does not seem to have any closed-form solution. However, it can be solved numerically[3]: the following figure represents the resulting values of α' as a function of α, when m takes all odd values from 1 to 51 (red circles); moreover, the superimposed solid line curve shows α' as a function of α as defined by (17) in the approach by Oliveira et al. This curve shows a very accurate agreement between our results and Oliveira's results.

[3] note that in the case $m = 1$, the solution is obviously $m' = 1$ since the Cauchy distributions are stable

	$m=3$	5	9	11	21	51
$n=1$	4.2646	8.0962	16.026	20.017	40.004	100.0
$n=2$	4.2857	8.1116	16.047	20.021	40.005	100.0
$n=5$	4.318	8.1406	16.032	20.031	40.008	100.0

Table 1. m' as a function of m



Oliveira's results and the solutions of equation 19

Moreover, by inspecting the numerical solutions m' of equation(19) for different values of m and n, as depicted in the table below, we propose an approximation rule called the "$m' = 2m - 2$" rule.

Proposition 3. *Given m and n, the solution m' of (19) can be approximated, for m large enough, as follows:*

$$m' \simeq 2m - 2$$

or equivalently, as $\alpha = \frac{m+n-2}{m+n}$,

$$\alpha' \simeq \frac{(4+n)\,\alpha - n}{(2+n)\,\alpha - (n-2)}$$

We note that this approximation is all the more accurate when α is near 1, and it is in agreement with the approximation provided by Oliveira et al.

A third result: almost additivity Unfortunately, a closed form expression for the distance between p_Z and $f^{(m')}$ is difficult to derive. The following theorem, however, allows us to derive an upper bound on this distance.

Theorem 9. *The distribution of the form $f^{(m')}$ closest to $p_{\mathbf{Z}}$ satisfies the orthogonality property*

$$D\left(f^{(m')}||p_{\mathbf{Z}}\right) = H_1\left(f^{(m')}\right) - H_1\left(p_{\mathbf{Z}}\right) \tag{21}$$

Moreover, the corresponding minimum Kullback-Leibler distance can be bounded as follows:

$$D\left(f^{(m')}||p_{\mathbf{Z}}\right) \leq H_1\left(f^{(m')}\right) - H_1\left(f^{(m)}\right) + \frac{1}{2}\log 2 \tag{22}$$

Proof. Remarking that

$$\begin{aligned}\int p_{\mathbf{Z}} \log f^{(m')} &= \log A_{\alpha'} - \frac{m'+n}{2}\sum_{k=0}^{m_Z} \alpha_k w_n(2k+1) \\ &= \log A_{\alpha'} - \frac{m'+n}{2} w_n(m')\end{aligned}$$

which is exactly the Shannon entropy $H_1\left(f^{(m')}\right)$, we deduce

$$\begin{aligned}D\left(p_{\mathbf{Z}}||f^{(m')}\right) &= -H_1(p_Z) - \int p_{\mathbf{Z}} \log f^{(m')} \\ &= H_1\left(f^{(m')}\right) - H_1(p_Z)\end{aligned}$$

Let us now consider

$$H_1(p_{\mathbf{Z}}) = H_1\left(p_{\frac{\mathbf{X}+\mathbf{Y}}{2}}\right) = H_1(p_{\mathbf{X}+\mathbf{Y}}) - \log 2$$

A classical inequality on the Shannon entropy of the sum of independent random variables is

$$H_1(p_{\mathbf{X}+\mathbf{Y}}) \geq H_1\left(p_{\tilde{\mathbf{X}}+\tilde{\mathbf{Y}}}\right) \tag{23}$$

where $\tilde{\mathbf{X}}$ and $\tilde{\mathbf{Y}}$ are independent Gaussian random variables such that

$$H_1\left(p_{\tilde{\mathbf{X}}}\right) = H_1(p_{\mathbf{X}}) \text{ and } H_1\left(p_{\tilde{\mathbf{Y}}}\right) = H_1(p_{\mathbf{Y}})$$

These constraints are equivalently

$$\sigma_{\tilde{\mathbf{X}}} = \sigma_{\tilde{\mathbf{Y}}} = \frac{\exp\left(\frac{m+n}{2} w_n(m)\right)}{A_\alpha\sqrt{2\pi e}}$$

so that

$$\begin{aligned}H_1\left(p_{\tilde{\mathbf{X}}+\tilde{\mathbf{Y}}}\right) &= \frac{1}{2}\log\left(2\pi e 2\sigma_{\tilde{\mathbf{X}}}\right) \\ &= H_1\left(f^{(m)}\right) + \frac{1}{2}\log 2\end{aligned}$$

	$m=3$	5	7	9	11	13	15	21	25	31
$D_{rel}\left(f^{(m')}\|\|p_Z\right)\times 10^4$	9.176	5.931	3.501	148.7	1.875	1.407	0.516	0.028	0.042	0.031
bound $(25)\times 10^4$	660	480	476	783	1718	2 75	1 25	33.18	18.75	9.82

Table 2. relative Kullback-Leibler distance, upper bound and numerical estimation

Let us remark that, as m grows, the Shannon inequality (23) and the bound expressed by (22) become tighter.

For the sake of comparison, it is more convenient to consider a **relative** Kullback-Leibler distance defined as

$$D_{rel}\left(f^{(m')}||p_Z\right)=\left|\frac{H_1\left(f^{(m')}\right)-H_1\left(p_Z\right)}{H_1\left(f^{(m')}\right)}\right| \tag{24}$$

so that the computed upper bound is defined now by

$$D_{rel}\left(f^{(m')}||p_Z\right)\leq\left|\frac{H_1\left(f^{(m')}\right)-H_1\left(f^{(m)}\right)+\frac{1}{2}\log 2}{H_1\left(f^{(m')}\right)}\right| \tag{25}$$

In the following table, we present, for $n=1$ and several values of m, the values of the relative upper bound as defined by the right hand side of (25). Moreover, we give an approximated numerical value of the true relative distance as defined by (24).

Inspection of the numerical values of $D_{rel}\left(f^{(m')}||p_Z\right)$ as a function of m shows that the approximation of p_Z by $f^{(m')}$ holds up to a relative error bounded by 0.1%, which is decreasing function of m for $m\geq 11$. The bound (25) is weaker but has the advantage of being in closed form.

4 Conclusion

In this paper, we have provided a complete characterizations of the $\alpha-$entropy maximizers under covariance constraints in the multivariate context. Elliptical invariance and a Gaussian mixture representation where established and the issue of stability of the entropy-maximizing densities was addressed. Applications of these results to pattern recognition, inverse problems, communications, and indpendent components analysis are currently being pursued.

References

1. J. Costa, A. O. Hero, and B. Ma, "Asymptotic convergence of random graphs and entropy estimation," Technical Report 315, Comm. and Sig. Proc. Lab. (CSPL), Dept. EECS, University of Michigan, Ann Arbor, Dec, 2002.

2. D. Geman and B. Jedynak, "An active testing model for tracking roads in satellite images," *IEEE Trans. on Pattern Anal. and Machine Intell.*, vol. 1, pp. 10–17, 1996.
3. G. T. Gullberg and B. M. W. Tsui, "Maximum entropy reconstruction with constraints: Iterative algorithms for solving the primal and dual programs," in *Information Processing in Medical Imaging*, C. N. de Graaf and M. A. Viergever, editors, chapter 23, Plenum Press, New York and London, 1988.
4. A. O. Hero, B. Ma, O. Michel, and J. D. Gorman, "Alpha-divergence for classification, indexing and retrieval," Technical Report 328, Comm. and Sig. Proc. Lab. (CSPL), Dept. EECS, University of Michigan, Ann Arbor, May, 2001. `http://www.eecs.umich.edu/~hero/det_est.html`.
5. R. J. McEliece, E. R. Rodemich, and L. Swanson, "An entropy maximization problem related to optical communication," *IEEE Trans. on Inform. Theory*, vol. 32, pp. 322–325, March, 1986.
6. M. I. Miller and D. L. Snyder, "The role of likelihood and entropy in incomplete-data problems: applications to estimating point-process intensities and Toeplitz constrained covariances," *IEEE Proceedings*, vol. 75, no. 7, pp. 892–907, July 1987.
7. A. Rényi, "On measures of entropy and information," in *Proc. 4th Berkeley Symp. Math. Stat. and Prob.*, volume 1, pp. 547–561, 1961.
8. B. D. Ripley, *Pattern recognition and neural networks*, Cambridge U. Press, 1996.

[1] G. Kaniadakis, A. Lavagno and P. Quarati, *Generalized statistics and solar neutrinos*, Physics Letters B, 369, 3-4, (1996), 308-312

[2] C. Tsallis, A. R. Plastino and W. -M. Zheng, *Power-law Sensitivity to Initial Conditions–New Entropic Representation*, Chaos, Solitons & Fractals, Volume 8, Issue 6, (1997), 885-891

[3] A. R. Plastino, A. Plastino, *Stellar polytropes and Tsallis' entropy*, Phys. Lett. A 174 (1993), 5-6, 384–386

[4] S. Moriguti, *A lower bound for a probability moment of any absolutely continuous distribution with finite variance* Ann. Math. Statistics 23, (1952). 286–289.

[5] J. N. Kapur, *Generalised Cauchy and Student's distributions as maximum-entropy distributions*, Proc. Nat. Acad. Sci. India Sect. A 58 (1988), 2, 235–246

[6] A.M.C. de Souza, C. Tsallis, *Student's t- and r- distributions: unified derivation from an entropic variational principle*, Physica A, 236 (1997), 52-57

[7] Thomas M. Cover, Joy A. Thomas, *Elements of Information Theory, (1991),* Wiley-Interscience

[8] K. T. Fang, S. Kotz, K. W. Ng, *Symmetric multivariate and related distributions*, Monographs on Statistics and Applied Probability, (1990), 36, Chapman and Hall, Ltd., London

[9] K. C. Chu, *Estimation and decision for linear systems with elliptical random processes*, IEEE Trans. on Automatic Control, 18 (1973), 499-505

[10] W. Feller, *An introduction to probability theory and its applications*, Vol. I. (1968), Third edition, John Wiley & Sons, Inc.

[11] M. Rahman, A.K. Md. Ehsanes Saleh, *Explicit form of the distribution of the Behrens-Fisher d-statistic*, J. Roy. Statist. Soc. Ser. B 36 (1974), 54-60

[12] I. Csiszár, *Information-type measures of difference of probability distributions and indirect observations*, Studia Sci. Math. Hungar. 2 (1967), 299– 318

[13] S. M. Ali, S. D. Silvey, *A general class of coefficients of divergence of one distribution from another*, J. Roy. Statist. Soc. Ser. B 28 (1966), 131– 142

[14] K. Zografos, *On maximum entropy characterization of Pearson's type II and VII multivariate distributions*, Journal of Multivariate Analysis, 71 (1999), 1, 67–75

[15] R. S. Mendes, C. Tsallis, *Renormalization group approach to nonextensive statistical mechanics*, Phys. Lett. A 285 (2001), no. 5-6, 273–278

[16] Q. A. Wang, M. Pezeril, L. Nivanen and A. Le Méhauté, *Nonextensive distribution and factorization of the joint probability*, Chaos Solitons & Fractals, Volume 13, Issue 1, January 2002, 131-137

[17] M. Abramowitz and I. Stegun, *Handbook of mathematical functions with formulas, graphs, and mathematical tables*, Applied mathematics series; 55, U.S. Govt. Print. Off., 1964

[18] F.A. Oliveira, B.A. Mello and I.M. Xavier Jr, *Scaling transformation of random walk distributions in a lattice*, Physical Review. E, Volume 61, Issue 6, Part B, June 2000, 7200-7203

[19] G.A. Walker, J.G. Saw, *The distribution of linear combinations of t-variables*, J. Amer. Statist. Assoc. 73 (1978), no. 364, 876–878

[20] V. Witkovsky, *On the exact computation of the density and of the quantiles of linear combinations of t and F random variables*, J. Statist. Plann. Inference 94 (2001), no. 1, 1–13

[21] J. F. C. Kingman, *Random walks with spherical symmetry*, Acta Math 109, 11-53, 1953

[22] K. Urbanik, *Generalized Convolutions I to V*, Studia Mathematica, 23, 45, 80 -1, 83 -2 and 91-2 resp., 1963, 73, 84, 86 and 88 resp., pp. 217-45, 57-70, 167-189, 57-95, 153-178 resp.