On Solutions to Multivariate Maximum α-entropy Problems

J. Costa[1], A. Hero[1], and C. Vignat[2]

[1] Department of Electrical Engineering and Computer Science,
University of Michigan, Ann Arbor, MI 48109-2122, USA
jcosta@umich.edu, hero@eecs.umich.edu
[2] Laboratoire Systèmes de Communications, Université Marne la Vallée, France
vignat@univ-mlv.fr

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 Rényi'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, communications, clustering, and pattern recognition; see [1–6] for representative examples. In particular, entropy maximization/minimization methods have found natural application in areas where an entropy or information divergence can be used as a discriminant of the data. These include: texture classification, feature clustering, image indexing or image registration, which are all core problems in areas such as geographical information systems, medical information processing, multi-sensor fusion and image content based retrieval. For example, the mutual information method of image registration (see [5] and references therein) searches through a set of coordinate transformations to find the one that minimizes the α-entropy of the joint feature distribution of the two images. In a similar way, a statistical image retrieval algorithm ([5]) searches trough a database of images to choose the image whose feature distribution is the closest to the query image in a minimum information divergence sense. Thus, studying the entropy maximizing distributions is important for understanding the advantages and limitations of such entropy maximization methods.

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

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

where f is the multivariate probability density of the n-dimensional random variable $\mathbf{X}$, 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:

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

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 the Rényi entropy under the same covariance constraint.

The remainder of this paper is organized as follows. In Section 2, we show 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. We then show that these distributions are elliptically invariant, which will allow a representation in terms of Gaussian scale mixtures. In addition, we give an alternative characterization for the maximum entropy distributions in terms of the Shannon entropy and a logarithmic constraint. In Section 3, we address the question of stability of the class of entropy-maximizing distributions under addition.

2 The Multivariate α-entropy Maximizing Distribution

Rényi-entropy maximizing distributions have been studied for the restricted case of $\alpha > 1$, by Moriguti in the scalar case [8] and by Kapur [9] in the multivariate case. The case of $\alpha \in [0,1]$ is of special interest since, in this region, the Rényi-entropy generalizes easily to Rényi-divergence via measure transformation [5].

Throughout, $\mathbf{X}$ will denote an $n-$dimensional real random vector with covariance matrix $\mathbf{K} = E(\mathbf{X}-\mu_{\mathbf{X}})(\mathbf{X}-\mu_{\mathbf{X}})^T$. In what follows, we consider, without loss of generality, the centered case $\mu_{\mathbf{X}} = E\mathbf{X} = 0$. Define next the following constants:

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

and

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

and the following sets

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

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

– if $\alpha > 1$

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

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

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

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

Note that Theorem 1 implies that the entropy $S_\alpha(f)$ has a unique maximizer f_α. We also point out that, when $0 < \alpha \leq \frac{n}{n+2}$, f_α has infinite covariance and so the covariance constraint cannot be met.

We prove this theorem by introducing a new divergence measure and adopting an information theoretic approach similar to that used by [10, Theorem 6.9.5] to prove that the Gaussian distribution maximizes Shannon entropy.

Consider the following non-symmetric directed divergence measure

$$D_\alpha(f||g) = \operatorname{sign}(\alpha - 1) \int_{\mathbb{R}_n} \left(\frac{f^\alpha}{\alpha} + \frac{\alpha - 1}{\alpha} g^\alpha - f g^{\alpha-1} \right) \tag{5}$$

The general theory of directed divergence measures is discussed in [11] and [12]. 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 $f = g$ *a.e.*

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

$$\int_{\mathbb{R}_n} f f_\alpha^{\alpha-1} \geq \int_{\mathbb{R}_n} f_\alpha \ , \tag{6}$$

with equality iff supp$(f) \subseteq \Omega_\alpha$.

Proof. Suppose for example $\alpha > 1$. Then

$$\begin{aligned} \int_{\mathbb{R}_n} f f_\alpha^{\alpha-1} &= \int_{\Omega_\alpha} f(\mathbf{x}) A_\alpha^{\alpha-1} \left(1 - \mathbf{x}^T \mathbf{C}_\alpha^{-1} \mathbf{x}\right) \mathrm{d}\mathbf{x} \\ &\geq \int_{\mathbb{R}_n} f(\mathbf{x}) A_\alpha^{\alpha-1} \left(1 - \mathbf{x}^T \mathbf{C}_\alpha^{-1} \mathbf{x}\right) \mathrm{d}\mathbf{x} \ , \end{aligned}$$

with equality iff $\mathrm{supp}(f) \subseteq \Omega_\alpha$. But, as f and f_α have the same covariance matrix,

$$\int_{\mathbb{R}_n} \mathbf{x}^T \mathbf{C}_\alpha^{-1} \mathbf{x} f(\mathbf{x})\, \mathrm{d}\mathbf{x} = \int_{\mathbb{R}_n} \mathbf{x}^T \mathbf{C}_\alpha^{-1} \mathbf{x} f_\alpha(\mathbf{x})\, \mathrm{d}\mathbf{x} ,$$

which implies

$$\int_{\mathbb{R}_n} f f_\alpha^{\alpha-1} \geq \int_{\mathbb{R}_n} f_\alpha A_\alpha^{\alpha-1} \left(1 - \mathbf{x}^T \mathbf{C}_\alpha^{-1} \mathbf{x}\right) \mathrm{d}\mathbf{x} = \int_{\mathbb{R}_n} f_\alpha^\alpha .$$

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

We can now deduce the extremal property of the density f_α.

Proof (of Theorem 1). Suppose, for example, $\alpha > 1$. Then, by Lemma 1 and positivity of D_α,

$$0 \leq D_\alpha(f||f_\alpha) \leq \int_{\mathbb{R}_n} \left(\frac{f^\alpha}{\alpha} + \frac{\alpha-1}{\alpha} f_\alpha^\alpha - f_\alpha^\alpha \right) = \frac{1}{\alpha} \int_{\mathbb{R}_n} \left(f^\alpha - f_\alpha^\alpha\right) .$$

Theorem 1 now follows. The proof is similar for $\alpha < 1$. □

Although the case $\alpha = 1$ was not explicitly addressed above, it can easily be shown that f_α converges pointwise to the density of $\mathcal{N}(0, \mathbf{K})$ when $\alpha \to 1$. Likewise, the corresponding entropies also converge to the Shannon entropy, thus extending, by continuity, Theorem 1 to the well known case of $\alpha = 1$.

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

$$p_{\mathbf{X}}(\mathbf{x}) = \phi_{\mathbf{X}} \left(\mathbf{x}^T \mathbf{C}^{-1} \mathbf{x}\right) \tag{7}$$

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 $\mathbf{X}$ is a random vector with density f_α, $\alpha < 1$, then it can be represented as a Gaussian scale mixture [13]: $\mathbf{X} = A\mathbf{N}$, where A is a Gamma random variable with shape parameter $\frac{m}{2} = \frac{1}{1-\alpha} - \frac{n}{2}$ and scale parameter 2, i.e., $A \sim \Gamma(\frac{m}{2}, 2)$. When $m = \frac{2}{1-\alpha} - n$ is a positive integer, A can be represented as a Chi-square random variable with m degrees of freedom. $\mathbf{N}$ is a n-variate Gaussian random vector, independent of A, with covariance matrix $\mathbf{C}_\alpha$. For more details see [14]. Equivalently, $\mathbf{X}$ can be rewritten as

$$\mathbf{X} = \frac{\mathbf{C}_\alpha^{\frac{1}{2}} \mathbf{N}_0}{\sqrt{\sum_{i=1}^m N_i^2}} , \tag{8}$$

where $\mathbf{N}_0$ is a zero mean Gaussian random vector with identity covariance matrix $\mathbf{I}_n$. As

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

converges a.s. to the constant matrix $\mathbf{K}^{\frac{1}{2}}$ when $m \to +\infty$ (i.e. $\alpha \to 1$), it is evident that, by Slutzky's theorem, $\mathbf{X}$ converges in distribution 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 $\alpha > 1$. This gives the following proposition:

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

$$\alpha = \frac{m+n}{m+n-2} \ , \tag{9}$$

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

$$\mathbf{X}_\alpha = \mathbf{C}_\alpha^{\frac{1}{2}} \frac{\mathbf{N}}{\sqrt{\|\mathbf{N}\|_2^2 + N_1^2 + \cdots + N_m^2}} \ , \tag{10}$$

where $\{N_i\}_{1 \leq i \leq m}$ are Gaussian $\mathcal{N}(0,1)$ mutually independent, and independent of $\mathbf{N}$ which is Gaussian $\mathcal{N}(0, \mathbf{I}_n)$.

We remark here that the denominator in (10) 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 [15]:

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

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

(a) – Case $\alpha < 1$. From [16],

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

The characteristic function of the Rényi distribution can then be written as

$$\varphi_\alpha(\mathbf{u}) = \frac{2^{1-\frac{m}{2}}}{\Gamma\left(\frac{m}{2}\right)} \left(\mathbf{u}^T \mathbf{C}_\alpha \mathbf{u}\right)^{\frac{m}{2}} K_{\frac{m}{2}}\left(\mathbf{u}^T \mathbf{C}_\alpha \mathbf{u}\right) \ , \tag{11}$$

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

$$\varphi_\alpha(\mathbf{u}) = 2^{\frac{m}{2}} \Gamma\left(\frac{m}{2}+1\right) \left(\mathbf{u}^T \mathbf{C}_\alpha \mathbf{u}\right)^{-\frac{m}{2}} J_{\frac{m}{2}}\left(\mathbf{u}^T \mathbf{C}_\alpha \mathbf{u}\right) \ , \tag{12}$$

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

We remark that both families of characteristic functions (11) and (12) 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 pointwise to the classical Gaussian characteristic function.

2.1 An Alternative Entropic Characterization

The Rényi-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 [9]. It was remarked also by Zografos [17] in the multivariate case, but not connected to the Rényi 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 [14] for details).

Theorem 2. *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

$$\int \log\left(1 + \mathbf{x}^T \mathbf{C}_\alpha^{-1} \mathbf{x}\right) f(\mathbf{x}) \, \mathrm{d}\mathbf{x} = \psi\left(\frac{m+n}{2}\right) - \psi\left(\frac{m}{2}\right) \tag{13}$$

(resp. $\int \log\left(1 - \mathbf{x}^T \mathbf{C}_\alpha^{-1} \mathbf{x}\right) f(\mathbf{x}) \, \mathrm{d}\mathbf{x} = \psi\left(\frac{m}{2}\right) - \psi\left(\frac{m+n}{2}\right)$*), where* $\psi(m) = \frac{\Gamma'(m)}{\Gamma(m)}$ *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 for all $\alpha > 0$, the distributions f_α as defined by (4) are solutions of the logarithmic 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. [18]. Then we address the issue of stability under the addition operation.

3.1 Renormalizability of f_α

Mendes and Tsallis ([18]) have shown that Rényi distributions have the important property of "renormalizability", but contrarily to the Gaussian case, they are not "factorizable". f_α has the renormalizability property when

$$\int_{-\infty}^{+\infty} f_\alpha\left(x_1, x_2\right) \mathrm{d}x_2 = f_{\alpha'}\left(x_1\right)$$

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 3. *Let* $\mathbf{X}^T = \left[\mathbf{X}_1^T, \mathbf{X}_2^T\right]$ $(\dim \mathbf{X}_i = n_i, n_1 + n_2 = n)$ *be a random vector distributed according to* f_α *with characteristic matrix* $\mathbf{C} = [\mathbf{C}_{11}, \mathbf{C}_{12}; \mathbf{C}_{21}, \mathbf{C}_{22}]$ $(\dim \mathbf{C}_{ij} = n_i \times n_j)$. *Then the marginal density of vector* $\mathbf{X}_i$ $(i = 1, 2)$ *is* f_{α_i}, *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 similarly 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

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

Hence

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

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 [15, corollary 3.2])

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

The case $\alpha > 1$ follows accordingly. □

Thus the renormalization property, as observed in [18], is nothing but a consequence of the elliptical invariance property, which is itself induced by the orthogonal invariance of both the Rényi entropy and the covariance constraint.

3.2 Stability of Rényi 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 Rényi-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 Rényi-entropy maximizing distributions is ensured, and link this feature with their elliptical invariance property, distinguishing between three important cases: the Rényi mutually dependent case, the mutually independent case and the special case of odd degrees of freedom. For proofs of these results see the referenced articles or [14].

Mutually Dependent Case

Theorem 4 ([15]). *If $\mathbf{X}_1$ and $\mathbf{X}_2$ are n_1 and n_2-variate vectors mutually distributed according to a Rényi-entropy maximizing density f_α 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 Rényi-entropy maximizing density $f_{\alpha'}$ 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 Rényi-entropy Maximizing Random Variables

Theorem 5 ([19]). *If X and Y are two scalar i.i.d. random variables with density f_α, then $Z = X + Y$ has a density which is **nearly** equal to $f_{\alpha'}$, with index α' such that*

$$\alpha' = 2 - (2-\alpha)\left(1 - 4\frac{\alpha\,(\alpha-1)}{(3\alpha-5)\,(\alpha+3)}\right) \ . \tag{14}$$

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

Relation (14) was obtained in [19] 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 Rényi-entropy maximizing distribution $f_{\alpha'}$ with the given parameter α'. In the case where m is an odd integer stronger results can be established. For economy of notation, we define, for m a positive integer,

$$f^{(m)} = f_\alpha \ , \quad \alpha = \frac{m+n-2}{m+n} \ . \tag{15}$$

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 [20],[21]).

Theorem 6. *Suppose that* $\mathbf{X}$ *and* $\mathbf{Y}$ *are two independent* n*-variate random vectors with densities* $f^{(m_\mathbf{X})}$ *and* $f^{(m_\mathbf{Y})}$*, respectively, and characteristic matrices* $\mathbf{C_X} = \mathbf{C_Y} = \mathbf{I}_n$*, with* ***odd*** *degrees of freedom* $m_\mathbf{X}$ *and* $m_\mathbf{Y}$*. Then, for* $0 \leq \beta \leq 1$*, the distribution of* $\mathbf{Z} = \beta\mathbf{X} + (1-\beta)\,\mathbf{Y}$ *is*

$$p_\mathbf{Z}(\mathbf{z}) = \sum_{k=0}^{k_\mathbf{Z}} \alpha_k f^{(2k+1)}(\mathbf{z}) \ , \tag{16}$$

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

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 $\{Q_l\}_{0\leq k\leq k_\mathbf{Z}}$ 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 $k_\mathbf{Z} \leq k_\mathbf{X} + k_\mathbf{Y} = \frac{m_X+m_Y}{2} - 1$, can be expressed in a unique way in this basis. Consequently, there exists a unique set $\{\alpha_k\}_{0\leq k\leq k_\mathbf{Z}}$ of real numbers such that

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

and

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

Result (16) now follows 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 Rényi-entropy maximizing random variables with odd degrees of freedom is distributed according to a **discrete scale mixture** of Rényi-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 (16) over $\mathbb{R}^n$, the positiveness of the coefficients α_k has, to our best knowledge, never proved in the literature. We are currently working on this conjecture, for which numerical simulations have confirmed the positivity of α_k's for a large number of special cases.

A Second Result: an Information Projection Property

The second result that we propose in this context allows us to characterize the projection of the Rényi entropy maximizing distribution onto a convolution of $f^{(m')}$'s with odd degrees of freedom.

Theorem 7. *Consider* $\mathbf{X}$ *and* $\mathbf{Y}$ *two independent* n*-variate random vectors following densities* $f^{(m_{\mathbf{X}})}$ *and* $f^{(m_{\mathbf{Y}})}$*, respectively, with characteristic matrices* $\mathbf{C_X} = \mathbf{C_Y} = \mathbf{I}_n$ *and* ***odd*** *degrees of freedom* $m_{\mathbf{X}}$ *and* $m_{\mathbf{Y}}$*. Let* $\mathbf{Z} = \frac{1}{2}(\mathbf{X} + \mathbf{Y})$*. Then, the Rényi distribution which is the closest to the distribution of* $\mathbf{Z}$ *in the sense of the Kullback-Leibler divergence has* m' *degrees of freedom such that*

$$w_n(m') = E w_n[M] \ , \tag{17}$$

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*

$$\Pr\{M = 2k+1\} = \alpha_k \ , \tag{18}$$

  where coefficients α_k *are defined by (16) for* $\beta = \frac{1}{2}$*.*

  Moreover, condition (17) 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 Rényi distribution $f^{(m')}$ with parameter m' is given by

$$\begin{aligned} D\left(p_{\mathbf{Z}} || f^{(m')}\right) &= \int p_{\mathbf{Z}} \log \frac{p_{\mathbf{Z}}}{f^{(m')}} \\ &= -S_1\left(p_{\mathbf{Z}}\right) - \int p_{\mathbf{Z}} \log f^{(m')} \ . \end{aligned}$$

Distribution $p_{\mathbf{Z}}$ takes the form

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

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 [17]: if $\mathbf{X} \sim f_m$ then [3]

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

Thus

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

Taking the derivative and equating to zero yields

$$w_n\left(m'\right) = E w_n\left(M\right) \ ,$$

where M is distributed according to (18). 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 ψ, together with

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

[3] Function $w_n\left(m\right)$ is denoted as $w_2\left(m, n\right)$ in [17].

Finally, computing

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

yields the final result. □

Equation (17) 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[4]: Fig. 1 represents the resulting values of α' as a function of α, when m takes all odd values from 1 to 51 (circles); moreover, the superimposed solid line curve shows α' as a function of α as defined by (14) in the approach by Oliveira et al [19]. This curve shows a very accurate agreement between our results and Oliveira's results.

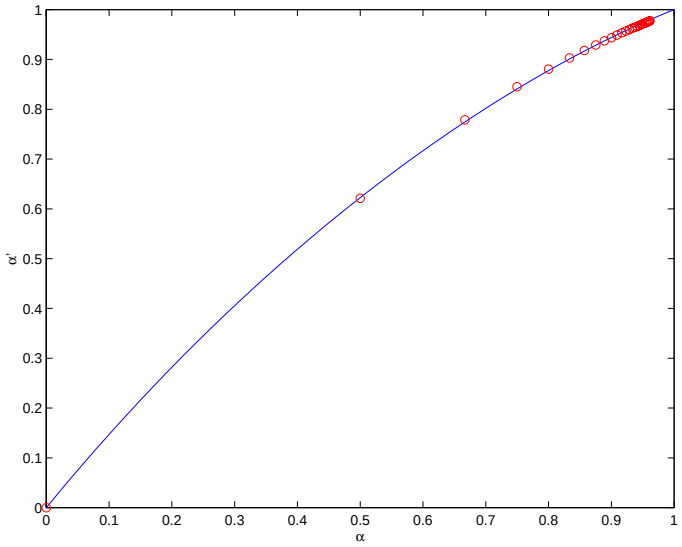

Fig. 1. Equation (14) (*solid line*) and the solutions of equation (17) (*circles*). See text for explanation

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

Moreover, by inspecting the numerical solutions m' of equation (17) for different values of m and n, as depicted in Table 1, we propose an approximation rule called the "$m' = 2m - 2$" rule.

Table 1. m' as a function of m

$m =$	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

Proposition 2. *Given m and n, the solution m' of (17) can be approximated, for m sufficiently large, as:*

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

or, equivalently $\left(\text{as } \alpha = \frac{m+n-2}{m+n}\right)$

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

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

$$D\left(f^{(m')}||p_{\mathbf{Z}}\right) = S_1\left(f^{(m')}\right) - S_1\left(p_{\mathbf{Z}}\right) \ . \tag{19}$$

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

$$D\left(f^{(m')}||p_{\mathbf{Z}}\right) \leq S_1\left(f^{(m')}\right) - S_1\left(f^{(m)}\right) + \frac{1}{2}\log 2 \ . \tag{20}$$

Proof. Remarking that

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

we deduce

$$D\left(p_{\mathbf{Z}}||f^{(m')}\right) = -S_1\left(p_Z\right) - \int p_{\mathbf{Z}} \log f^{(m')} \\ = S_1\left(f^{(m')}\right) - S_1\left(p_Z\right) \ .$$

Let us now consider

$$S_1\left(p_{\mathbf{Z}}\right) = S_1\left(p_{\frac{\mathbf{X}+\mathbf{Y}}{2}}\right) = S_1\left(p_{\mathbf{X}+\mathbf{Y}}\right) - \log 2 \ .$$

A classical inequality on the Shannon entropy of the sum of independent random variables is the so called entropy power inequality [10]:

$$S_1\left(p_{\mathbf{X}+\mathbf{Y}}\right) \geq S_1\left(p_{\tilde{\mathbf{X}}+\tilde{\mathbf{Y}}}\right) \ , \tag{21}$$

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

$$S_1\left(p_{\tilde{\mathbf{X}}}\right) = S_1\left(p_{\mathbf{X}}\right) \text{ and } S_1\left(p_{\tilde{\mathbf{Y}}}\right) = S_1\left(p_{\mathbf{Y}}\right) \ .$$

These constraints are equivalent to

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

so that

$$S_1\left(p_{\tilde{\mathbf{X}}+\tilde{\mathbf{Y}}}\right) = \frac{1}{2}\log\left(2\pi e 2\sigma_{\tilde{\mathbf{X}}}\right) \\ = S_1\left(f^{(m)}\right) + \frac{1}{2}\log 2 \ .$$

□

Let us remark that, as m grows, the Shannon inequality (21) and the bound expressed by (20) become tighter.

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

$$D_{rel}\left(f^{(m')}||p_Z\right) = \left|\frac{S_1\left(f^{(m')}\right) - S_1\left(p_Z\right)}{S_1\left(f^{(m')}\right)}\right| \ , \tag{22}$$

so that the computed upper bound is now defined by

$$D_{rel}\left(f^{(m')}||p_Z\right) \leq \left|\frac{S_1\left(f^{(m')}\right) - S_1\left(f^{(m)}\right) + \frac{1}{2}\log 2}{S_1\left(f^{(m')}\right)}\right| \ . \tag{23}$$

In Table 2, 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 (23). Moreover, we

Table 2. Relative Kullback-Leibler distance, upper bound and numerical approximation

$m =$	3	5	7	9	11	13	15	21	25	31
$D_{rel}(f^{(m')}\|\|p_Z)\times 10^4$	9.176	5.931	3.501	148.7	1.875	1.407	0.516	0.028	0.042	0.031
bound $(23)\times 10^4$	660	480	476	783	1718	2 75	1 25	33.18	18.75	9.82

give an approximated numerical value of the true relative distance as defined by (22).

Inspection of the numerical values of $D_{rel}(f^{(m')}||p_Z)$ 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 a function of m, for $m \geq 11$. The bound (23) is weaker but has the advantage of being in closed form.

4 Conclusion

In this paper, we have provided a complete characterization of the $\alpha-$entropy maximizers under covariance constraints for multivariate densities. 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 independent components analysis are currently being pursued.

Acknowledgments

This work was partially supported by Fundação para a Ciência e Tecnologia under the project SFRH/BD/2778/2000, a Dept. EECS UM Fellowship, and DARPA-MURI Grant Number DAAD19-02-1-0262.

References

1. Geman, D., Jedynak, B.: An active testing model for tracking roads in satellite images. IEEE Trans. on Pattern Anal. and Machine Intell. **1** (1996) 10–17
2. Gullberg, G.T., Tsui, B.M.W.: Maximum entropy reconstruction with constraints: Iterative algorithms for solving the primal and dual programs. In de Graaf, C.N., Viergever, M.A., eds.: Information Processing in Medical Imaging. Plenum Press, New York and London (1988)
3. Miller, M.I., Snyder, D.L.: The role of likelihood and entropy in incomplete-data problems: applications to estimating point-process intensities and Toeplitz constrained covariances. IEEE Proceedings **75** (1987) 892–907
4. McEliece, R.J., Rodemich, E.R., Swanson, L.: An entropy maximization problem related to optical communication. IEEE Trans. on Inform. Theory **32** (1986) 322–325

5. Hero, A., Ma, B., Michel, O., Gorman, J.: Applications of entropic spanning graphs. IEEE Signal Processing Magazine **19** (2002) 85–95
6. Heemuchwala, H., Hero, A.O., Carson, P.: Image registration using entropy measures and entropic graphs. To appear in European Journal of Signal Processing, Special Issue on Content-based Visual Information Retrieval (2003)
7. Rényi, A.: On measures of entropy and information. In: Proc. 4th Berkeley Symp. Math. Stat. and Prob. Volume 1. (1961) 547–561
8. Moriguti, S.: A lower bound for a probability moment of any absolutely continuous distribution with finite variance. Ann. Math. Statistics **23** (1952) 286–289
9. Kapur, J.N.: Generalised Cauchy and Student's distributions as maximum-entropy distributions. Proc. Nat. Acad. Sci. India Sect. A **58** (1988) 235–246
10. Cover, T., Thomas, J.: Elements of Information Theory. Wiley, New York (1987)
11. Csiszár, I.: Information-type measures of difference of probability distributions and indirect observations. Studia Sci. Math. Hungar. **2** (1967) 299–318
12. S. M. Ali, S.D.S.: A general class of coefficients of divergence of one distribution from another. J. Roy. Statist. Soc. Ser. B **28** (1966) 131–142
13. Feller, W.: An introduction to probability theory and its applications. Third edn. Volume I. John Wiley & Sons, Inc. (1968)
14. Vignat, C., Costa, J., Hero, A.: Characterization of the multivariate distributions maximazing Tsallis entropy under covariance constraint. Technical report (2003)
15. Chu, K.C.: Estimation and decision for linear systems with elliptical random processes. IEEE Trans. on Automatic Control **18** (1973) 499–505
16. Abramowitz, M., Stegun, I.: Handbook of mathematical functions with formulas, graphs, and mathematical tables. Volume 55 of Applied Mathematics Series. U.S. Govt. Print. Off. (1964)
17. Zografos, K.: On maximum entropy characterization of Pearson's type II and VII multivariate distributions. Journal of Multivariate Analysis **71** (1999) 67–75
18. Mendes, R.S., Tsallis, C.: Renormalization group approach to nonextensive statistical mechanics. Phys. Lett. A **285** (2001) 273–278
19. Oliveira, F., Mello, B., Jr., I.X.: Scaling transformation of random walk distributions in a lattice. Physical Review. E **61** (2000) 7200–7203
20. Walker, G., Saw, J.: The distribution of linear combinations of t-variables. J. Amer. Statist. Assoc. **73** (1978) 876–878
21. Witkovsky, V.: 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) 1–13