Entropy **2011**, *13*, 1746-1764; doi:10.3390/e13101746

OPEN ACCESS

entropy

ISSN 1099-4300

www.mdpi.com/journal/entropy

Article

Projective Power Entropy and Maximum Tsallis Entropy Distributions

Shinto Eguchi *, Osamu Komori and Shogo Kato

The Institute of Statistical Mathematics, Tachikawa, Tokyo 190-8562, Japan;
E-Mails: komori@ism.ac.jp (O.K.); skato@ism.ac.jp (S.K.)

* Author to whom correspondence should be addressed; E-Mail: eguchi@ism.ac.jp.

Received: 26 July 2011; in revised form: 20 September 2011 / Accepted: 20 September 2011 / Published: 26 September 2011

Abstract: We discuss a one-parameter family of generalized cross entropy between two distributions with the power index, called the projective power entropy. The cross entropy is essentially reduced to the Tsallis entropy if two distributions are taken to be equal. Statistical and probabilistic properties associated with the projective power entropy are extensively investigated including a characterization problem of which conditions uniquely determine the projective power entropy up to the power index. A close relation of the entropy with the Lebesgue space L_p and the dual L_q is explored, in which the escort distribution associates with an interesting property. When we consider maximum Tsallis entropy distributions under the constraints of the mean vector and variance matrix, the model becomes a multivariate q-Gaussian model with elliptical contours, including a Gaussian and t-distribution model. We discuss the statistical estimation by minimization of the empirical loss associated with the projective power entropy. It is shown that the minimum loss estimator for the mean vector and variance matrix under the maximum entropy model are the sample mean vector and the sample variance matrix. The escort distribution of the maximum entropy distribution plays the key role for the derivation.

Keywords: elliptical contoured distribution; escort distribution; L_p space; maximum entropy distribution; statistical distribution; Tsallis entropy

1. Introduction

In the classical statistical physics and the information theory the close relation with Boltzmann-Shannon entropy has been well established to offer elementary and clear understandings. The Kullback-Leibler divergence is directly connected with maximum likelihood, which is one of the most basic ideas in statistics. Tsallis opened new perspectives for the power entropy to elucidate non-equilibrium states in statistical physics, and these give the strong influence on the research for non-extensive and chaotic phenomenon, *cf.* [1,2]. There are proposed several generalized versions of entropy and divergence, *cf.* [3–7]. We consider generalized entropy and divergence defined on the space of density functions with finite mass,

$$\mathcal{F} = \Big\{ f : \int f(x)dx < \infty, f(x) \geq 0 \text{ for almost everywhere } x \Big\}$$

in a framework of information geometry originated by Amari, *cf.* [8,9].

A functional $D : \mathcal{F} \times \mathcal{F} \mapsto [0, \infty)$ is called a divergence if $D(g, f) \geq 0$ with equality if and only if $g = f$. It is shown in [10,11] that any divergence associates with a Riemannian metric and a pair of conjugate connections in a manifold modeled in $\mathcal{F}$ under mild conditions.

We begin with the original form of power cross entropy [12] with the index β of $\mathbb{R}$ defined by

$$C_\beta^{(\mathrm{o})}(g, f) = -\frac{1}{\beta} \int g(x)\{f(x)^\beta - 1\}dx + \frac{1}{1+\beta} \int f(x)^{1+\beta} dx$$

for all g and f in $\mathcal{F}$, and so the power (diagonal) entropy

$$H_\beta^{(\mathrm{o})}(f) = C_\beta^{(\mathrm{o})}(f, f) = -\frac{1}{\beta(\beta+1)} \int f(x)^{1+\beta} dx + \frac{1}{\beta} \int f(x)dx$$

See [13,14] for the information geometry and statistical applications for the independent component analysis and pattern recognition. Note that this is defined in the continuous case for probability density functions, but can be reduced to a discrete case, see Tsallis [2] for the extensive discussion on statistical physics. In fact, the Tsallis entropy

$$S_q(f) = \frac{1}{q-1} \Big\{ 1 - \int f(x)^q dx \Big\}$$

for a probability density function $f(x)$ is proportional to the power entropy to a constant with $qH_\beta^{(\mathrm{o})}(g) - 1$, where $q = 1 + \beta$. The power divergence is given by

$$D_\beta^{(\mathrm{o})}(g, f) = C_\beta^{(\mathrm{o})}(g, f) - H_\beta^{(\mathrm{o})}(g)$$

as, in general, defined by the difference of the cross entropy and the diagonal entropy.

In this paper we focus on the projective power cross entropy defined by

$$C_\gamma(g, f) = -\frac{1}{\gamma(1+\gamma)} \frac{\int g(x) f(x)^\gamma dx}{\Big\{ \int f(x)^{1+\gamma} dx \Big\}^{\frac{\gamma}{1+\gamma}}} \tag{1}$$

and so the projective power entropy is

$$H_\gamma(f) = -\frac{1}{\gamma(1+\gamma)} \Big\{ \int f(x)^{1+\gamma} dx \Big\}^{\frac{1}{1+\gamma}} \tag{2}$$

The log expression for $C_\gamma(g,f)$ is defined by

$$C_\gamma^{\log}(g,f) = -\frac{1}{\gamma}\log\{-\gamma(1+\gamma)C_\gamma(g,f)\}$$

See [15,16] for the derivation of $C_\gamma^{\log}$, and detailed discussion on the relation between $C_\beta^{(\mathrm{o})}(g,f)$ and $C_\gamma(g,f)$. The projective power cross entropy $C_\gamma(g,f)$ satisfies the linearity with respect to g and the projective invariance, that is $C_\gamma(g,\lambda f) = C_\gamma(g,f)$ for any constant $\lambda > 0$. Note that $H_\gamma(f)$ has a one-to-one correspondence with $S_q(f)$ as given by

$$H_\gamma(f) = -\frac{1}{q(q-1)}\{1-(q-1)S_q(f)\}^{\frac{1}{q}}$$

where $q = 1+\gamma$. The projective power divergence is

$$D_\gamma(g,f) = C_\gamma(g,f) - H_\gamma(g) \tag{3}$$

which will be discussed on a close relation with the Hölder's inequality. The divergence defined by $C_\gamma(g,f)$ satisfies

$$D_\gamma^{\log}(g,f) = C_\gamma^{\log}(g,f) - C_\gamma^{\log}(g,g) \geq 0$$

for all γ of $\mathbb{R}$ if there exist integrals in $D_\gamma^{\log}(g,f)$. The nonnegativity leads to

$$D_\gamma(g,f) \geq 0 \tag{4}$$

We remark that the existence range of the power index γ for $C_\gamma(g,f)$ and $H_\gamma(f)$ depends on the sample space on which f and g are defined. If the sample space is compact, both $C_\gamma(g,f)$ and H_γ are well-defined for all $\gamma \in \mathbb{R}$. If the sample space is not compact, $C_\gamma(g,f)$ is defined for $\gamma \geq 0$ and $H_\gamma(f)$ is for $\gamma > -1$. More precisely we will explore the case that the sample space is $\mathbb{R}^d$ in a subsequent discussion together with moment conditions. Typically we observe that

$$\lim_{\gamma\to 0} D_\gamma(g,f) = D_0(g,f) \tag{5}$$

where $D_0(g,f)$ denotes the Kullback-Leibler divergence,

$$D_0(g,f) = \int g(x)\log\frac{g(x)}{f(x)}dx \tag{6}$$

See Appendix 1 for the derivation of (5).

Let $\{x_1,\cdots,x_n\}$ be a random sample from a distribution with the probability density function $g(x)$. A statistical model $\{f(x,\theta) : \theta \in \Theta\}$ with parameter θ is assumed to sufficiently approximate the underlying density function $g(x)$, where Θ is a parameter space. Then the loss function associated with the projective power entropy $C_\gamma(g, f(\cdot,\theta))$ based on the sample is given by

$$L_\gamma(\theta) = -\frac{1}{\gamma(1+\gamma)}\frac{1}{n}\sum_{i=1}^{n} k_\gamma(\theta)f(x_i,\theta)^\gamma$$

in which we call

$$\hat{\theta}_\gamma \equiv \underset{\theta\in\Theta}{\operatorname{argmin}}\, L_\gamma(\theta) \tag{7}$$

the γ-estimator, where

$$k_\gamma(\theta) = \Big\{ \int f(x,\theta)^{1+\gamma} dx \Big\}^{-\frac{\gamma}{1+\gamma}}$$

We note that

$$\mathbb{E}_g\{L_\gamma(\theta)\} = C_\gamma(g, f(\cdot,\theta))$$

where $\mathbb{E}_g$ denotes the statistical expectation with respect to g. It is observed that the 0-estimator is nothing but the maximum likelihood estimator (MLE) since the loss $L_\gamma(\theta)$ converges to the minus log-likelihood function,

$$L_0(\theta) = -\frac{1}{n}\sum_{i=1}^{n} \log f(x_i,\theta)$$

in the sense that

$$\lim_{\gamma\to 0}\Big\{ L_\gamma(\theta) + \frac{1}{\gamma(1+\gamma)} \Big\} = L_0(\theta)$$

If the underlying density function $g(x)$ belongs to a Gaussian model with mean μ and variance σ^2, then the MLEs for μ and σ^2 are the sample mean and sample variance. The reverse statement is shown in [17,18]. We will extend this theory to a case of the γ-estimator under γ-model.

In Section 2 we discuss characterization of the projective power entropy. In Section 3 the maximum entropy distribution with the Tsallis entropy S_q with $q = 1+\gamma$ under the constraints of mean vector μ and variance matrix Σ is considered. We discuss the model of maximum entropy distributions, called the γ-model, in which 0-model and 2-model equal Gaussian and Wigner models, respectively. Then we show that the γ-estimators for μ and Σ under the γ-model are the sample mean and sample variance. Section 4 gives concluding remarks and further comments.

2. Projective Invariance

Let us look at a close relation of $\mathcal{F}$ with Lebesgue's space

$$L_p = \Big\{ f(x) : \int |f(x)|^p dx < \infty \Big\}$$

where $p \geq 1$ and the L_p-norm $\| \ \|_p$ is defined by

$$\|f\|_p = \Big\{ \int |f(x)|^p dx \Big\}^{\frac{1}{p}}$$

Let q be the conjugate index of p satisfying $1/p+1/q = 1$, in which p and q can be expressed as functions of the parameter $\gamma > 0$ such that $p = 1+\gamma^{-1}$ and $q = 1+\gamma$. We note that this q is equal to the index q

in Tsallis entropy S_q in the relation $q = 1 + \gamma$. For any probability density function $f(x)$ we define the escort distribution with the probability density function,

$$e_q(f(x)) = \frac{f(x)^q}{\int f(y)^q dy}$$

cf. [2] for extensive discussion. We discuss an interesting relation of the projective cross entropy (1) with the escort distribution. By the definition of the escort distribution,

$$C_\gamma(g, f) = -\frac{1}{\gamma(1+\gamma)} \int \{e_q(f(x))\}^{\frac{1}{p}} g(x) dx \tag{8}$$

We note that $e_q(f)^{\frac{1}{p}}$ is in the unit sphere of L_p in the representation. The projective power diagonal entropy (2) is proportional to the L_q-norm, that is,

$$H_\gamma(f) = -\frac{1}{\gamma(1+\gamma)} \|f\|_q$$

from which the Hölder's inequality

$$\int g(x) f(x)^\gamma dx \leq \|g\|_q \, \|f^\gamma\|_p \tag{9}$$

claims that $C_\gamma(g, f) \geq H_\gamma(g)$, or equivalently

$$D_\gamma(g, f) \geq 0 \tag{10}$$

for all f and g in $\mathcal{F}$, which is also led by $C_\gamma^{(\mathrm{o})}(g, f) \geq H_\gamma^{(\mathrm{o})}(g)$. The equality in (10) holds if and only if $f(x) = \lambda g(x)$ for almost everywhere x, where λ is a positive constant. The power transform suggests an interplay between the space L_p and L_q by the relation,

$$\|f^\gamma\|_p = \|f\|_q{}^\gamma$$

Taking the limit of γ to 0 in the Hölder's inequality (9) yields that

$$\int g(x) \log f(x) dx \leq \int g(x) \log g(x) dx$$

since

$$\lim_{\gamma \to 0} \int g(x) \frac{f(x)^\gamma - 1}{\gamma} dx = \int g(x) \log f(x) dx$$

and

$$\lim_{\gamma \to 0} \frac{\|f^\gamma\|_p \|g\|_q - 1}{\gamma} = \int g(x) \log g(x) dx \tag{11}$$

This limit regarding p associates with another space rather than the L_∞ space, which is nothing but the space of all density functions with finite Boltzmann-Shannon entropy, say $L_{\log}$. The power index γ reparameterizes the Lebesgue space L_p and the dual space L_q with the relation $p = 1 + \gamma^{-1}$, however, to take the power transform $f(x)^\gamma$ is totally different from the ordinary discussion of the Lebesgue

space, so that the duality converges to $(L_{\log}, L_1)$ as observed in (11). In information geometry the pair $(L_{\log}, L_1)$ corresponds to that of mixture and exponential connections, *cf.* [9]. See also another one-parameterization of L_p space [19].

We now discuss a problem of the uniqueness for $C_\gamma(g, f)$ as given in the following theorem. A general discussion on the characterization is given in [16], however, the derivation is rather complicated. Here we assume a key condition that a cross entropy $\Gamma(g, f)$ is linear in g to give an elementary proof. The Riesz representation theorem suggests

$$\Gamma(g, f) = c(f) \int g(x)\psi(f(x))dx$$

where $c(f)$ is a constant that depends on f. Thus we observe the following theorem when we make a specific form for $c(f)$ to guarantee the scale invariance.

Theorem 1. Define a functional $\Gamma : \mathcal{F} \times \mathcal{F} \mapsto \mathbb{R}$ by

$$\Gamma(g, f) = \varphi\Big(\int \rho(f(x))dx \Big) \int g(x)\psi(f(x))dx \tag{12}$$

where φ, ρ and ψ are differentiable and monotonic functions. Assume that

(i). $\Gamma(g, g) = \min_{f\in\mathcal{F}} \Gamma(g, f)$ for all $g \in \mathcal{F}$,

and that

(ii). $\Gamma(g, f) = \Gamma(g, \lambda f)$ for all $\lambda > 0$ and all $g, f \in \mathcal{F}$.

Then there exists γ such that $\Gamma(g, f) = C_\gamma(g, f)$ up to a constant factor, where $C_\gamma(g, f)$ is the projective power cross entropy defined by (1).

Proof. The requirement (ii) means that

$$\frac{\partial}{\partial\lambda}\Big\{\varphi\Big(\int \rho(\lambda f(x))dx \Big) \int \psi(\lambda f(x))g(x)dx\Big\} = 0$$

which implies that, if f is absolutely continuous and g is the Dirac measure at x_0, then

$$\frac{\dot\psi(\lambda f(x_0))}{\psi(\lambda f(x_0))}\lambda f(x_0) = c(\lambda)$$

where

$$c(\lambda) = -\frac{\lambda\dot\varphi\Big(\int \rho(\lambda f(x))dx \Big) \int \dot\rho(\lambda f(x))f(x)dx}{\varphi\Big(\int \rho(\lambda f(x))dx \Big)}$$

Since we can take an arbitrary value $f(x_0)$ for any fixed λ,

$$\frac{\dot\psi(t)}{\psi(t)} = c(\lambda)t^{-1}$$

which is uniquely solved as $\psi(t) = t^\gamma$ where $\gamma = c(\lambda)$. Next let us consider a case of a finite discrete space, $\{x_i : 1 \le i \le m\}$. Then, since $\psi(f) = f^\gamma$, we can write

$$\Gamma(g, f) = \varphi\Big(\sum_{i=1}^m \rho(f_i)\Big) \sum_{i=1}^m g_i f_i^\gamma$$

where $f_i = f(x_i)$ and $g_i = g(x_i)$. The requirement (i) leads that $(\partial/\partial f_j)\Gamma(g, f)|_{f=g} = 0$ for all $j, 1 \le j \le m$, which implies that

$$\dot\rho(g_j) = -\gamma c(g_1, \cdots, g_m) g_j^\gamma \tag{13}$$

where

$$c(g_1, \cdots, g_m) = \frac{\varphi\Big(\sum_{i=1}^m \rho(g_i)\Big)}{\sum_{i=1}^m g_i^{1+\gamma} \dot\varphi\Big(\sum_{i=1}^m \rho(g_i)\Big)} \tag{14}$$

It follows from (13) that $c(g_1, \cdots, g_m)$ must be a constant in $g_1, \cdots, g_m$, say C, so that we solve (13) as $\rho(g_j) = -\gamma C g_j^{1+\gamma}/(1+\gamma)$. Therefore, Equation (14) is written by

$$\frac{\dot\varphi(t)}{\varphi(t)} = -\frac{\gamma}{1+\gamma} t^{-1}$$

which leads to $\varphi(t) = t^{-\frac{\gamma}{1+\gamma}}$. We conclude that $\Gamma(g, f) \propto C_\gamma(g, f)$, which completes the proof.□

Remark 1. The proof above is essentially applicable for the case that the integral (11) is given by the summation just for a binary distributions. In this sense the statement of Theorem 1 is not tight, however, statistical inference is discussed in a unified manner such that the distribution is either continuous or discrete. In a subsequent discussion we focus on the case for continuous distributions defined on $\mathbb{R}^d$.

Remark 2. We see the multiplicative decomposition for $C_\gamma(g, f)$ for statistical independence. In fact, if f and g are decomposed as $f = f_1 \otimes f_2, g = g_1 \otimes g_2$ in the same partition, then

$$C_\gamma(g, f) = C_\gamma(g_1, f_1) C_\gamma(g_2, f_2)$$

This property is also elemental, but we do not assume this decomposability as the requirement in Theorem 1.

3. Model of Maximum Entropy Distributions

We will elucidate a dualistic structure between the maximum entropy model on H_γ, defined in (2) and the minimum cross entropy estimator on C_γ, defined in (1). Before the discussion we overview the classical case in which the maximum likelihood estimation nicely makes good performance under the maximum entropy model on the Boltzmann-Shannon entropy, that is, a Gaussian model if we consider the mean and variance constraint. We will use conventional notations that X denotes random variable with value x. Let $\{x_1, \cdots, x_n\}$ be a random sample from a Gaussian distribution with the density function

$$f_0(x, \mu, \Sigma) = \det(2\pi\Sigma)^{-\frac{1}{2}} \exp\{-\frac{1}{2}(x-\mu)^{\mathrm{T}} \Sigma^{-1} (x-\mu)\}$$

The Gaussian density function is written by a canonical form

$$(2\pi)^{-\frac{d}{2}} \exp\Big\{ -\frac{1}{2}(x-\mu)^{\mathrm{T}}\Xi(x-\mu) + \frac{1}{2}\log\det(\Xi)\Big\} \tag{15}$$

where Ξ is called the canonical parameter defined by Σ^{-1}. The differentiation of (15) on μ and Ξ yields

$$\mathbb{E}_{f_0(\cdot,\mu,\Sigma)}(X) = \mu \quad \text{and} \quad \mathbb{V}_{f_0(\cdot,\mu,\Sigma)}(X) = \Sigma$$

where $\mathbb{E}_f$ and $\mathbb{V}_f$ denote the expectation vector and variance matrix with respect to a probability density function $f(x)$, respectively.

The maximum likelihood estimator is given by

$$(\hat{\mu}_0, \widehat{\Sigma}_0) = (\bar{x}, S) \tag{16}$$

where $\bar{x}$ and S are the sample mean vector and the sample variance matrix,

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \quad S = \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})(x_i-\bar{x})^{\mathrm{T}} \tag{17}$$

This is because the minus log-likelihood function is

$$L_0(\mu,\Sigma) = -\frac{1}{n}\sum_{i=1}^{n}\log f_0(x_i,\mu,\Sigma)$$

which is written by

$$\frac{1}{2}\mathrm{trace}(S(\mu)\Xi) - \frac{1}{2}\log\det(\Xi)$$

apart from a constant, where

$$S(\mu) = \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)(x_i-\mu)^{\mathrm{T}} \tag{18}$$

Hence the estimating equation system is

$$\begin{bmatrix} \dfrac{\partial}{\partial\mu}L_0(\mu,\Sigma) \\ \dfrac{\partial}{\partial\Xi}L_0(\mu,\Sigma) \end{bmatrix} = \begin{bmatrix} \Xi(\bar{x}-\mu) \\ \dfrac{1}{2}\{S(\mu)-\Sigma\} \end{bmatrix} = \begin{bmatrix} 0 \\ \mathrm{O} \end{bmatrix}$$

which concludes the Expression (16) of the MLE since $S(\mu) = S + (\bar{x}-\mu)(\bar{x}-\mu)^{\mathrm{T}}$. Alternatively, we have another route to show (16) as follows. The Kullback-Leibler divergence defined in (6) is given by

$$\begin{aligned} & D_0(f_0(\cdot,\mu,\Sigma), f_0(\cdot,\mu_1,\Sigma_1)) \\ = \; & \frac{1}{2}(\mu-\mu_1)^{\mathrm{T}}\Sigma_1^{-1}(\mu-\mu_1) + \frac{1}{2}\mathrm{trace}\{(\Sigma-\Sigma_1)\Sigma_1^{-1}\} - \frac{1}{2}\log\det(\Sigma\Sigma_1^{-1}) \end{aligned}$$

Thus, we observe that

$$L_0(\mu,\Sigma) - L_0(\bar{x},S) = D_0(f_0(\cdot,\bar{x},S), f_0(\cdot,\mu,\Sigma)) \tag{19}$$

which is nonnegative with equality if and only if $(\mu, \Sigma) = (\bar{x}, S)$. This implies (16).

Under mild regularity conditions the reverse statement holds, that is, the MLE for a location and scatter model satisfies (16) if and only if the model is Gaussian, *cf.* [17,18]. However, even if we do not assume anything for the underling distribution $g(x)$, the statistics $\bar{x}$ and S are asymptotically consistent for

$$\mu_g = \mathbb{E}_g(X) \quad \text{and} \quad \Sigma_g = \mathbb{V}_g(X)$$

This is a direct result from the strong law of large numbers, and the central limit theorem leads to the asymptotic normality for these two statistics. In this sense, $(\bar{x}, S)$ is also a nonparametric estimator for (μ_g, Σ_g).

We explore a close relation of the statistical model and the estimation method. We consider a maximum entropy distribution with the γ-entropy H_γ over the space of d-dimensional distributions with a common mean and variance,

$$\mathcal{F}_{(\mu,\Sigma)} = \{f \in \mathcal{F} : \int f(x)dx = 1, \mathbb{E}_f(x) = \mu, \mathbb{V}_f(x) = \Sigma\} \tag{20}$$

Then we define a distribution with a probability density function written by

$$f_\gamma(x, \mu, \Sigma) = \frac{c_\gamma}{\det(2\pi\Sigma)^{\frac{1}{2}}} \Big\{1 - \frac{\gamma}{2 + d\gamma + 2\gamma}(x-\mu)^{\mathrm{T}}\Sigma^{-1}(x-\mu)\Big\}_+^{\frac{1}{\gamma}} \tag{21}$$

where $(\ )_+$ denotes a positive part and c_γ is the normalizing factor,

$$c_\gamma = \begin{cases} \big(\frac{2\gamma}{2+d\gamma+2\gamma}\big)^{\frac{d}{2}} \frac{\Gamma(1+\frac{d}{2}+\frac{1}{\gamma})}{\Gamma(1+\frac{1}{\gamma})} & \text{if } \gamma > 0 \\ \big(-\frac{2\gamma}{2+d\gamma+2\gamma}\big)^{\frac{d}{2}} \frac{\Gamma(-\frac{1}{\gamma})}{\Gamma(-\frac{1}{\gamma}-\frac{d}{2})} & \text{if } -\frac{2}{d+2} < \gamma < 0 \end{cases} \tag{22}$$

See the derivation for c_γ in Appendix 2. If the dimension d equals 1, then $f_\gamma(x, \mu, \Sigma)$ is a q-Gaussian distribution with $q = \gamma + 1$. We remark that

$$\lim_{\gamma \uparrow 0} c_\gamma = \lim_{\gamma \downarrow 0} c_\gamma = 1$$

in which $f_\gamma(x, \mu, \Sigma)$ is reduced to a d-variate Gaussian density when $\gamma = 0$. The support of $f_\gamma(\cdot, \mu, \Sigma)$ becomes an ellipsoid defined as

$$\Big\{x \in \mathbb{R}^d : (x-\mu)^{\mathrm{T}}\Sigma^{-1}(x-\mu) \le \frac{2 + d\gamma + 2\gamma}{\gamma}\Big\}$$

if $\gamma > 0$. On the other hand, if $-\frac{2}{d+2} < \gamma < 0$, the density function (21) is written as

$$f_\gamma(x, \mu, \Sigma) = \det(\pi\tau\Sigma)^{-\frac{1}{2}} \frac{\Gamma(-\frac{1}{\gamma})}{\Gamma(-\frac{1}{\gamma} - \frac{d}{2})} \Big\{1 + \frac{1}{\tau}(x-\mu)^{\mathrm{T}}\Sigma^{-1}(x-\mu)\Big\}^{\frac{1}{\gamma}} \tag{23}$$

where

$$\tau = -\frac{2 + (d+2)\gamma}{\gamma}$$

The d-variate t-distribution is defined by

$$g_\nu(x,\mu,P)=\det(\pi\nu P)^{-\frac{1}{2}}\frac{\Gamma(\frac{\nu+d}{2})}{\Gamma(\frac{\nu}{2})}\Big\{1+\frac{1}{\nu}(x-\mu)^{\mathrm{T}}P^{-1}(x-\mu)\Big\}^{-\frac{\nu+d}{2}} \tag{24}$$

cf. [20] for the extensive discussion. Assume that

$$\frac{\nu+d}{2}=-\frac{1}{\gamma}\quad\text{and}\quad \nu P=\tau\Sigma$$

Then we observe from (23) and (24) that

$$f_\gamma(x,\mu,\Sigma)=g_\nu(x,\mu,P)$$

Accordingly, the density function $f_\gamma(x,\mu,\Sigma)$ with $-\frac{2}{d+2}<\gamma<0$ is a t-distribution. The distribution has elliptical contours on the Euclidean space $\mathbb{R}^d$ for any $\gamma>-\frac{2}{d+2}$, as shown in Figure 1 for typical cases of γ.

Figure 1. t-distribution $(\gamma=-0.4)$, Gaussian $(\gamma=0)$ and Wigner $(\gamma=2)$ distributions.

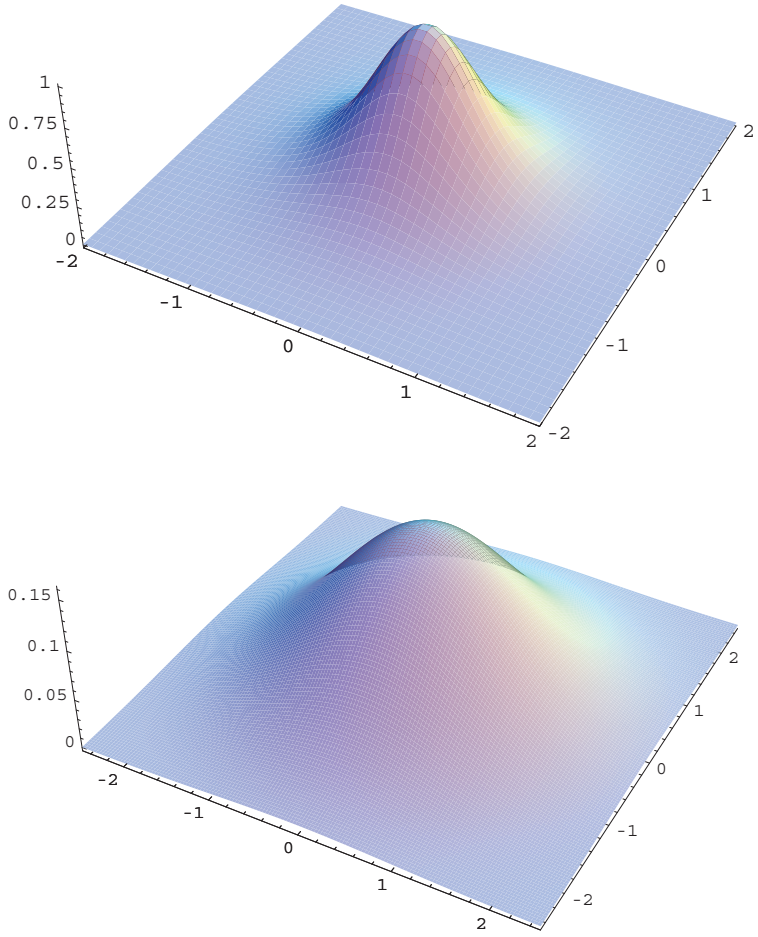
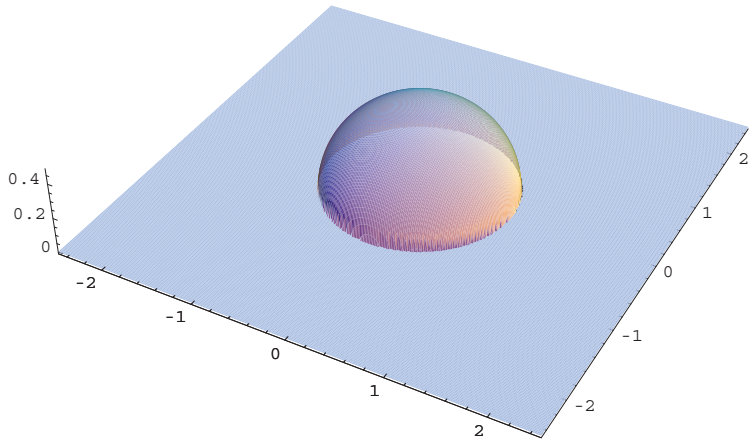


Figure 1. *Cont.*



Let

$$M_\gamma = \left\{ f_\gamma(x, \mu, \Sigma) : \mu \in \mathbb{R}^d, \Sigma \in \mathcal{S}_d \right\} \tag{25}$$

which we call γ-model, where $\mathcal{S}_d$ denotes the space of all symmetric, positive-definite matrices of order d. We confirm the mean and variance of the γ-model as follows.

Lemma. Under the model M_γ defined in (25) with the index $\gamma > -\frac{2}{d+2}$,

$$\mathbb{E}_{f_\gamma(\cdot,\mu,\Sigma)}(X) = \mu \quad \text{and} \quad \mathbb{V}_{f_\gamma(\cdot,\mu,\Sigma)}(X) = \Sigma$$

Proof. We need to consider a family of escort distributions. In the model M_γ we can define the escort distribution as

$$e_q(f_\gamma(x, \mu, \Sigma)) = \frac{c_\gamma^*}{\det(\Sigma)^{\frac{1}{2}}} \left\{ 1 - \frac{\gamma}{2 + d\gamma + 2\gamma} (x - \mu)^{\mathrm{T}} \Sigma^{-1} (x - \mu) \right\}_+^{\frac{1+\gamma}{\gamma}} \tag{26}$$

where $q = 1 + \gamma$ and c_γ^* is the normalizing factor. Hence,

$$e_q(f_\gamma(x, \mu, \Sigma)) = c_\gamma^* \left\{ \det(\Sigma)^{-\frac{1}{2}\frac{\gamma}{1+\gamma}} - \frac{\gamma}{2 + d\gamma + 2\gamma} (x - \mu)^{\mathrm{T}} \{ \det(\Sigma)^{-\frac{1}{2}\frac{\gamma}{1+\gamma}} \Sigma^{-1} \} (x - \mu) \right\}_+^{\frac{1+\gamma}{\gamma}} \tag{27}$$

Here we define alternative parameter Ξ_γ to the original parameter Σ by the transform

$$\Xi_\gamma = \det(\Sigma)^{-\frac{1}{2}\frac{\gamma}{1+\gamma}} \Sigma^{-1} \tag{28}$$

and so that the inverse transform is given by

$$\Sigma = \det(\Xi_\gamma)^{\frac{\gamma}{d\gamma+2\gamma+2}} \Xi_\gamma^{-1} \tag{29}$$

noting that $\det(\Xi_\gamma) = \det(\Sigma)^{-\frac{1}{2}\frac{d\gamma+2\gamma+2}{1+\gamma}}$. Thus, we get a canonical form of (26) as

$$e_q(f_\gamma(x, \mu, \Sigma)) = c_\gamma^* \left\{ \det(\Xi_\gamma)^{\frac{\gamma}{2+d\gamma+2\gamma}} - \frac{\gamma}{2 + d\gamma + 2\gamma} (x - \mu)^{\mathrm{T}} \Xi_\gamma (x - \mu) \right\}_+^{\frac{1+\gamma}{\gamma}} \tag{30}$$

By analogy of the discussion for an exponential family we have the following expression for the braced term in (30) as

$$-\frac{2\gamma}{2+d\gamma+2\gamma}\Big\{\frac{1}{2}\mathrm{trace}(xx^{\mathrm{T}}\Xi_\gamma)-\mu^{\mathrm{T}}\Xi_\gamma x+\frac{1}{2}\mu^{\mathrm{T}}\Xi_\gamma\mu-\frac{2+d\gamma+2\gamma}{2\gamma}\det(\Xi_\gamma)^{\frac{\gamma}{d\gamma+2\gamma+2}}\Big\} \tag{31}$$

A property of the escort distribution suggests moment formulae for the distribution (25) as follows: We have an identity

$$\frac{\partial}{\partial\mu}\int c_\gamma^*\Big\{\det(\Xi_\gamma)^{\frac{\gamma}{d\gamma+2\gamma+2}}-\frac{\gamma}{2+d\gamma+2\gamma}(x-\mu)^{\mathrm{T}}\Xi_\gamma(x-\mu)\Big\}_+^{\frac{1+\gamma}{\gamma}}dx=0$$

which implies that

$$\int\Big\{\det(\Xi_\gamma)^{\frac{\gamma}{d\gamma+2\gamma+2}}-\frac{\gamma}{2+d\gamma+2\gamma}(x-\mu)^{\mathrm{T}}\Xi_\gamma(x-\mu)\Big\}_+^{\frac{1}{\gamma}}\Xi_\gamma(x-\mu)dx=0$$

which concludes that

$$\mathbb{E}_{f_\gamma(\cdot,\mu,\Sigma)}(X)=\mu$$

Similarly,

$$\frac{\partial}{\partial\Xi_\gamma}\int c_\gamma^*\Big\{\det(\Xi_\gamma)^{\frac{\gamma}{d\gamma+2\gamma+2}}-\frac{\gamma}{2+d\gamma+2\gamma}(x-\mu)^{\mathrm{T}}\Xi_\gamma(x-\mu)\Big\}_+^{\frac{1+\gamma}{\gamma}}dx=0$$

which is

$$\begin{aligned}&\int c_\gamma^*\Big\{\det(\Xi_\gamma)^{\frac{\gamma}{d\gamma+2\gamma+2}}-\frac{\gamma}{2+d\gamma+2\gamma}(x-\mu)^{\mathrm{T}}\Xi_\gamma(x-\mu)\Big\}_+^{\frac{1}{\gamma}}\\ \times\ &\Big\{\frac{\gamma}{d\gamma+2\gamma+2}\det(\Xi_\gamma)^{\frac{\gamma}{d\gamma+2\gamma+2}}\Xi_\gamma^{-1}-\frac{\gamma}{2+d\gamma+2\gamma}(x-\mu)(x-\mu)^{\mathrm{T}}\Big\}dx=0\end{aligned} \tag{32}$$

which concludes that

$$\mathbb{V}_{f_\gamma(\cdot,\mu,\Sigma)}(X)=\Sigma$$

because of the relation of Ξ_γ and Σ as observed in (29). The proof is complete. □

Remark 3. The canonical form (30) of the escort distribution (26) plays an important role on the proof of Lemma. Basically we can write the canonical form of (21), however it is not known any link to distributional properties like a case of exponential family.

Remark 4. In Equation (31) the function

$$\varphi(\Xi)=\frac{1}{2\omega}\det(\Xi)^{\omega} \tag{33}$$

is viewed as a potential function in the Fenchel convex duality, where

$$\omega=\frac{\gamma}{2+d\gamma+2\gamma}$$

cf. [21,22] for the covariance structure model.

From Lemma we observe that $f_\gamma(\cdot,\mu,\Sigma)\in\mathcal{F}(\mu,\Sigma)$. Next we show that the distribution with density $f_\gamma(\cdot,\mu,\Sigma)$ maximizes the γ-entropy H_γ over the space $\mathcal{F}(\mu,\Sigma)$, where H_γ is defined in (2).

Theorem 2.
(i). If $-\frac{2}{d+2}<\gamma\le 0$, then

$$f_\gamma(\cdot,\mu,\Sigma)=\operatorname*{argmax}_{f\in\mathcal{F}_{(\mu,\Sigma)}} H_\gamma(f) \tag{34}$$

where $\mathcal{F}_{(\mu,\Sigma)}$ is defined in (20).
(ii). If $\gamma>0$, then

$$f_\gamma(\cdot,\mu,\Sigma)=\operatorname*{argmax}_{f\in\mathcal{F}_{(\mu,\Sigma)}{}^{(\gamma)}} H_\gamma(f) \tag{35}$$

where

$$\mathcal{F}_{(\mu,\Sigma)}{}^{(\gamma)}=\{f\in\mathcal{F}_{(\mu,\Sigma)}: f(x)=0 \text{ for almost everywhere } x\in B(\mu,\Sigma)\}$$

with $B(\mu,\Sigma)$ being $\{x\in\mathbb{R}^d:(x-\mu)^{\mathrm{T}}\Sigma^{-1}(x-\mu)>\frac{2+d\gamma+2\gamma}{\gamma}\}$.

Proof. By the definition of $\mathcal{F}_{(\mu,\Sigma)}$, we see from Lemma that $f_\gamma(\cdot,\mu,\Sigma)\in\mathcal{F}_{(\mu,\Sigma)}$ for any $\gamma\in(-\frac{2}{d+2},0)$. This leads to

$$\mathbb{E}_{f_\gamma(\cdot,\mu,\Sigma)}\{f_\gamma(X,\mu,\Sigma)^\gamma\}=\mathbb{E}_f\{f_\gamma(X,\mu,\Sigma)^\gamma\}$$

for any f in $\mathcal{F}_{(\mu,\Sigma)}$, which implies that

$$H_\gamma(f_\gamma(\cdot,\mu,\Sigma))=C_\gamma(f,f_\gamma(\cdot,\mu,\Sigma))$$

Hence

$$H_\gamma(f_\gamma(\cdot,\mu,\Sigma))-H_\gamma(f)=D_\gamma(f,f_\gamma(\cdot,\mu,\Sigma)) \tag{36}$$

which is nonnegative as discussed in (4). This concludes (34). Similarly, we observe that (36) holds for any $\gamma>0$ and any f in $\mathcal{F}_{(\mu,\Sigma)}{}^{(\gamma)}$ since the support of f includes that of $f(\cdot,\mu,\Sigma)$. This concludes (35). □

We would like to elucidate a similar structure for the statistical inference by the minimum projective cross entropy in which the data set $\{x_1,\cdots,x_n\}$ is assumed to follow the model M_γ. We recall (8) from the relation of the projective cross entropy with the escort distribution

$$C_\gamma(g,f)=-\frac{1}{\gamma(1+\gamma)}\int e_q(f(x))^{\frac{\gamma}{1+\gamma}}g(x)dx$$

When we have got data $\{x_1,\cdots,x_n\}$ to be fitted to the model M_γ, the loss function is

$$L_\gamma(\mu,\Sigma)=-\frac{1}{\gamma(1+\gamma)}\frac{1}{n}\sum_{i=1}^{n}e_q(f_\gamma(x_i,\mu,\Sigma))^{\frac{\gamma}{1+\gamma}}$$

where $f_\gamma(x,\mu,\Sigma)$ defined in (21). The γ-estimator is defined by

$$(\hat{\mu}_\gamma, \widehat{\Sigma}_\gamma) = \underset{(\mu,\Sigma)}{\operatorname{argmin}}\, L_\gamma(\mu,\Sigma)$$

see the general definition (7). It follows from the canonical form defined in (30) with the canonical parameter Ξ_γ defined in (28) that

$$L_\gamma(\mu,\Sigma) = -\frac{1}{\gamma(1+\gamma)}(c_\gamma^*)^{\frac{\gamma}{\gamma+1}}\left[\det(\Xi_\gamma)^\omega - \omega\{\mathrm{trace}(\Xi_\gamma S) + (\mu-\bar{x})^{\mathrm{T}}\Xi_\gamma(\mu-\bar{x})\}\right] \tag{37}$$

where $(\bar{x}, S)$ and ω are defined in (17) and (33), and c_γ^* is the normalizing factor defined in (27). Here we note that if $\gamma > 0$, then the parameter (μ,Σ) must be assumed to be in Θ_n, where

$$\Theta_n = \{(\mu,\Sigma) \in \mathbb{R}^d \times \mathcal{S}_d \ :\ (x_i-\mu)^{\mathrm{T}}\Sigma^{-1}(x_i-\mu) < \omega^{-1}\ \ (\forall i = 1,\cdots,n)\ \} \tag{38}$$

We note that $L_\gamma(\mu,\Sigma) = C_\gamma(f(\cdot,\bar{x},S), f(\cdot,\mu,\Sigma))$ and $L_\gamma(\bar{x},S) = H_\gamma(f(\cdot,\bar{x},S))$ since

$$\mathbb{E}_{f(\cdot,\bar{x},S)}(X) = \bar{x}, \quad \text{and} \quad \mathbb{V}_{f(\cdot,\bar{x},S)}(X) = S$$

Accordingly, we observe the argument similar to (19) for the MLE. The projective divergence D_γ defined in (3) equals the difference of the γ-loss functions as

$$L_\gamma(\mu,\Sigma) - L_\gamma(\bar{x},S) = D_\gamma(f_\gamma(\cdot,\bar{x},S), f_\gamma(\cdot,\mu,\Sigma)), \tag{39}$$

which is nonnegative with equality if and only if $(\mu,\Sigma) = (\bar{x},S)$. See the discussion after equation (10). In this way, we can summarize the above discussion as follows:

Theorem 3. Let $\{x_1,\cdots,x_n\}$ be a random sample from a γ-model defined in (21). Then the γ-estimator defined in (7) for (μ,Σ) is $(\bar{x},S)$, where $(\bar{x},S)$ is defined in (17).

Proof. Let us give another proof. The estimating equation system is given by

$$\begin{bmatrix} \dfrac{\partial}{\partial\mu}L_\gamma(\mu,\Sigma) \\ \dfrac{\partial}{\partial\Xi_\gamma}L_\gamma(\mu,\Sigma) \end{bmatrix} = \begin{bmatrix} \Xi_\gamma(\bar{x}-\mu) \\ \omega\{\det(\Xi_\gamma)^\omega\Xi_\gamma^{-1} - S(\mu)\} \end{bmatrix} = \begin{bmatrix} 0 \\ \mathrm{O} \end{bmatrix} \tag{40}$$

which is equivalent to

$$\begin{bmatrix} \mu-\bar{x} \\ \Sigma - S(\mu) \end{bmatrix} = \begin{bmatrix} 0 \\ \mathrm{O} \end{bmatrix}$$

because of the relation of Ξ_γ into Σ as given in (29). Thus, we also attain the conclusion $(\hat{\mu}_\gamma, \widehat{\Sigma}_\gamma) = \big(\bar{x}, S\big)$. In this way, we obtain the solution of the equation system defined by (40) via the parameter Ξ_γ using the relation of the escort distribution with the loss function (37). □

Remark 5. Consider the location model $\{f_\gamma(\cdot,\mu,\Sigma)\}$ with the location parameter μ, where Σ is known in Theorem 3. Then we easily see that the γ-estimator for μ is $\bar{x}$. What about the reverse statement? We observe that if the γ-estimator for μ is $\bar{x}$ with the sample size $n \geq 3$, then the model is the γ-model, $\{f_\gamma(\cdot,\mu,\Sigma)$ with the known Σ. The proof is parallel to that of Theorem 2 given in [17]. In fact, we conclude that the model density function $f(x)$ satisfies that

$$\{f(x-\mu)\}^\gamma = a + b(x-\mu)^{\mathrm{T}}\Sigma^{-1}(x-\mu)$$

where a and b are constants.

Remark 6. If we look at jointly Theorem 2 and 3, then

$$\min_{(\mu,\Sigma)\in\mathbb{R}^d\times\mathbb{S}_d} L_\gamma(\mu,\Sigma) = \max_{f\in\mathcal{F}(\bar{x},S)} H_\gamma(f) \tag{41}$$

since $L_\gamma(\bar{x},S) = H_\gamma(f_\gamma(\cdot,\bar{x},S))$. Both sides of (41) associate with inequalities (39) and (36) on γ-divergence in separate discussion.

Remark 7. The derivation of the γ-estimator in Theorem 3 is provided by the canonical parameter Ξ_γ of the escort distribution as given in (28). Here we directly calculate the gradient of the loss with respect to Σ as follows:

$$\begin{aligned}
\frac{\partial}{\partial\Sigma}L_\gamma(\mu,\Sigma) &= -\frac{1}{2(1+\gamma)^2}\det(\Sigma)^{-\frac{1}{2}\frac{\gamma}{1+\gamma}}(1-\omega\mathrm{trace}\{S(\mu)\Sigma^{-1}\})\Sigma^{-1} \\
&\quad +\frac{\gamma}{(1+\gamma)}\omega\det(\Sigma)^{-\frac{1}{2}\frac{\gamma}{1+\gamma}}\,\Sigma^{-1}S(\mu)\Sigma^{-1} \\
&= -\frac{1}{2(1+\gamma)^2}\det(\Sigma)^{-\frac{1}{2}\frac{\gamma}{1+\gamma}} \\
&\quad \times\Big[(1-\omega\mathrm{trace}\{S(\mu)\Sigma^{-1}\})\Sigma^{-1} - \frac{1+\gamma}{1+\frac{1}{2}d\gamma+\gamma}\Sigma^{-1}S(\mu)\Sigma^{-1}\Big]
\end{aligned}$$

Therefore we observe that if we put $\mu=\bar{x}$ and $\Sigma=\alpha S(\bar{x})$, then

$$\begin{aligned}
&\frac{\partial}{\partial\Sigma}L_\gamma(\bar{x},\alpha S(\bar{x})) = -\frac{1}{2}\frac{\gamma}{1+\gamma}\det(\alpha S(\bar{x}))^{-\frac{1}{2}\frac{\gamma}{1+\gamma}}(\alpha S(\bar{x}))^{-1} \\
&\times\Big[(1-\omega\mathrm{trace}\{S(\bar{x})(\alpha S(\bar{x}))^{-1}\})\alpha S(\bar{x}) - \frac{1+\gamma}{1+\frac{1}{2}d\gamma+\gamma}S(\bar{x})\Big](\alpha S(\bar{x}))^{-1}
\end{aligned} \tag{42}$$

The bracketed term of (42) is given by

$$\begin{aligned}
&\Big[\alpha(1-\omega\mathrm{trace}\{S(\bar{x})(\alpha S(\bar{x}))^{-1}\}) - \frac{1+\gamma}{1+\frac{1}{2}d\gamma+\gamma}\Big]S(\bar{x}) \\
= &\Big(\alpha - \frac{d\gamma}{2+d\gamma+2\gamma} - \frac{1+\gamma}{1+\frac{1}{2}d\gamma+\gamma}\Big)S(\bar{x})
\end{aligned}$$

which concludes that if $\alpha=1$, then $(\partial/\partial\Sigma)L_\gamma(\bar{x},\alpha S(\bar{x}))=0$. This is a direct proof for Theorem 3, but it would accompany with a heuristic discussion for the substitution of (μ,Σ) into $(\bar{x},\alpha S(\bar{x}))$.

4. Concluding Remarks

We explored the elegant property (39), the empirical Pythagoras relation between the γ-model and γ-estimator, in the sense that (39) directly shows Theorem 3 without any differential calculus. Another elegant expression is in the minimax game between Nature and a decision maker, see [23]. Consider the space $\mathcal{F}(\mu, \Sigma)$ defined in (20). The intersection of the γ-model (21) and $\mathcal{F}(\mu, \Sigma)$ is a singleton $\{f_\gamma(\cdot, \mu, \Sigma)\}$, which is the minimax solution of

$$\max_{g \in \mathcal{F}(\mu,\Sigma)} \min_{f \in \mathcal{F}} C_\gamma(g, f) = C_\gamma(f_\gamma(\cdot, \mu, \Sigma), f_\gamma(\cdot, \mu, \Sigma))$$

Consider different indices γ and γ^* which specify the γ-model and γ^*-estimator, respectively. Basically the γ^*-estimator is consistent under the γ-model for any choice of γ and γ^*. If we specifically fix $\gamma = 0$ for the model, that is, a Gaussian model, then the γ^*-estimator is shown to be qualitatively robust for any $\gamma^* > 0$, see [16]. The degree of robustness is proportional to the value of γ^* with a trade for the efficiency. The γ^*-estimator for (μ, Σ) of the Gaussian model is given by the solution of

$$\begin{aligned} \mu &= \frac{\sum_{i=1}^n f_0(x_i, \mu, \Sigma)^{\gamma^*} x_i}{\sum_{i=1}^n f_0(x_i, \mu, \Sigma)^{\gamma^*}} \\ \Sigma &= (1+\gamma^*) \frac{\sum_{i=1}^n f_0(x_i, \mu, \Sigma)^{\gamma^*} (x_i - \mu)(x_i - \mu)^{\mathrm{T}}}{\sum_{i=1}^n f_0(x_i, \mu, \Sigma)^{\gamma^*}} \end{aligned}$$

The weight function $f_0(x_i, \mu, \Sigma)^{\gamma^*}$ for the i-th observation x_i becomes almost 0 when x_i is an outlier. Alternatively, the classical robust method employs $\gamma^* = 0$, that is, the MLE for the misspecified model $\gamma < 0$ or t-distribution model, see [24,25]. Thus, the different indices γ and γ^* work robust statistics in a dualistic manner.

This property is an extension of that associated between the exponential model and MLE, however, it is fragile in the sense that (19) does not hold if the indices in the γ-model and γ^*-estimator are slightly different. In practice, we find some difficulties for a numerical task for solving the MLE under the γ-model with $\gamma > 0$ because the support of the density depends on the parameter and the index γ. We discussed statistical and probabilistic properties on the model and estimation associated with the specific cross entropy. A part of properties discussed still holds for any cross entropy in a much wider class, which is investigated from the point of the Fenchel duality in [13,26].

Acknowledgements

We would like to express our thanks to two referees for their helpful comments and constructive suggestions.

References

1. Tsallis, C. Possible generalization of Boltzmann-Gibbs statistics. *J. Statist. Physics.* **1988**, *52*, 479–487.
2. Tsallis, C. *Introduction to Nonextensive Statistical Mechanics: Approaching a Complex World*; Springer-Verlag: New York, NY, USA, 2009.

3. Cichocki, A.; Cruces, S.; Amari, S. Families of alpha- beta- and gamma- divergences: Flexible and robust measures of similarities. *Entropy* **2010**, *12*, 1532–1568.
4. Cichocki, A.; Cruces, S.; Amari, S. Generalized alpha-beta divergences and their application to robust nonnegative matrix factorization. *Entropy* **2011**, *13*, 134–170.
5. Csiszàr, I. Information-type measures of difference of probability distributions and indirect observation. *Studia Scientiarum Mathematicarum Hungarica* **1967**, *2*, 229–318.
6. Rèny I, A. On measures of entropy and information. *Proc. Fourth Berkeley Symp. Math. Statist. Prob.* **1961**, *1*, 547–561.
7. Tøpsoe, F. Some inequalities for information divergence and related measures of discrimination. *IEEE Trans. Inform. Theor.* **2000**, *46*, 1602–1609.
8. Amari, S. Differential-geometrical methods in statistics. In *Lecture Notes in Statistics*; Springer-Verlag: New York, NY, USA, 1985; Volume 28.
9. Amari, S.; Nagaoka, H. Methods of information geometry. In *Translations of Mathematical Monographs*; American Mathematical Society: Providence, RI, USA, 2000; Volume 191.
10. Eguchi, S. Second order efficiency of minimum contrast estimators in a curved exponential family. *Ann. Statist.* **1983**, *11*, 793–803.
11. Eguchi, S. Geometry of minimum contrast. *Hiroshima Math. J.* **1992**, *22*, 631–647.
12. Basu, A.; Harris, I.R.; Hjort, N.L.; Jones, M.C. Robust and efficient estimation by minimizing a density power divergence. *Biometrika* **1988**, *85*, 549–559.
13. Eguchi, S. Information divergence geometry and the application to statistical machine learning. In *Information Theory and Statistical Learning*; Emmert-Streib, F., Dehmer, M., Eds.; Springer: New York, NY, USA, 2008; 309–332.
14. Minami, M.; Eguchi, S. Robust blind source separation by beta-divergence. *Neural Comput.* **2002**, *14*, 1859–1886.
15. Eguchi, S.; Kato, S. Entropy and divergence associated with power function and the statistical application. *Entropy* **2010**, *12*, 262–274.
16. Fujisawa, H.; Eguchi, S. Robust parameter estimation with a small bias against heavy contamination. *J. Multivariate Anal.* **2008**, *99*, 2053–2081.
17. Azzalini, A.; Genton, M.G. On Gauss's characterization of the normal distribution. *Bernoulli* **2007**, *13*, 169–174.
18. Teicher, H. Maximum likelihood characterization of distributions. *Ann. Math. Statist.* **1961**, *32*, 1214–1222.
19. Amari, S.; Ohara, A. Geometry of q-exponential family of probability distributions. *Entropy* **2011**, *13*, 1170–1185.
20. Kotz, S.; Nadarajah, S. *Multivariate T Distributions and Their Applications*; Cambridge University Press: Cambridge, UK, 2004.
21. Eguchi, S. A differential geometric approach to statistical inference on the basis of contrast functionals. *Hiroshima Math. J.* **1985**, *15*, 341–391.
22. Wakaki, H.; Eguchi, S.; Fujikoshi, Y. A class of tests for general covariance structure. *J. Multivariate Anal.* **1990**, *32*, 313–325.

23. Grünwald, P.D.; Dawid, A.P. Game theory, maximum entropy, minimum discrepancy, and robust Bayesian decision theory. *Ann. Statist.* **2004**, *32*, 1367–1433.
24. Kent, J.T.; Tyler, D.E. Redescending M-estimates of multivariate location and scatter. *Ann. Statist.* **1991**, *19*, 2102–2119.
25. Marrona, R.A. Robust M-estimators of multivariate location and scatter. *Ann. Statist.* **1976**, *4*, 51–67.
26. Eguchi, S. Information geometry and statistical pattern recognition. *Sugaku Exposition* **2006**, *19*, 197–216.

Appendix 1

We show (5). It follows from l'Hôpital's rule that

$$\lim_{\gamma\to 0} D_\gamma(g,f) = \left(\frac{\partial}{\partial\gamma}\left[\left\{ \int g(x)^{1+\gamma}dx \right\}^{\frac{1}{1+\gamma}} - \frac{\int g(x)f(x)^\gamma dx}{\left\{ \int f(x)^{1+\gamma}dx \right\}^{\frac{\gamma}{1+\gamma}}} \right] \right)_{\gamma=0}$$

which is written as

$$\left(\frac{1}{1+\gamma}\left\{ \int g(x)^{1+\gamma}dx \right\}^{\frac{-\gamma}{1+\gamma}} \int g(x)^{1+\gamma}\log g(x)dx - \frac{\int g(x)f(x)^\gamma \log f(x)dx}{\left\{ \int f(x)^{1+\gamma}dx \right\}^{\frac{\gamma}{1+\gamma}}} + \frac{\gamma}{1+\gamma}\frac{\int g(x)f(x)^\gamma dx}{\left\{ \int f(x)^{1+\gamma}dx \right\}^{\frac{1+2\gamma}{1+\gamma}}} \int f(x)^{1+\gamma}\log f(x)dx \right)_{\gamma=0}$$

which is reduced to

$$\int g(x)\log g(x)dx - \int g(x)\log f(x)dx$$

This completes the proof of (5). □

Appendix 2

First, we give the formula for c_γ in (22) when $\gamma > 0$. Let

$$I = \frac{1}{\det(2\pi\Sigma)^{\frac{1}{2}}} \int \left\{ 1 - \omega(x-\mu)^{\mathrm{T}}\Sigma^{-1}(x-\mu) \right\}_+^{\frac{1}{\gamma}} dx$$

where $\omega = \frac{\gamma}{2+d\gamma+2\gamma}$. The integral is rewritten as

$$I = (2\pi\omega)^{-\frac{d}{2}} \int (1-y^{\mathrm{T}}y)_+^{\frac{1}{\gamma}} dy$$

where $y = (\omega)^{\frac{1}{2}}\Sigma^{-1/2}(x-\mu)$. It is expressed in polar coordinates as

$$I = (2\pi\omega)^{-\frac{d}{2}} S^{d-1} \int_0^1 (1-r^2)^{\frac{1}{\gamma}} r^{d-1} dr \tag{43}$$

where S^{d-1} is the surface area of the unit sphere of $d-1$ dimension, that is,

$$S^{d-1} = \frac{2\pi^{\frac{d}{2}}}{\Gamma(\frac{d}{2})}$$

Since the integral in (43) is expressed by a beta function, we have

$$c_\gamma = I^{-1} = (2\omega)^{\frac{d}{2}} \frac{\Gamma(1+\frac{d}{2}+\frac{1}{\gamma})}{\Gamma(1+\frac{1}{\gamma})}$$

Second, we give the formula when $-\frac{2}{d+2} < \gamma < 0$. The argument similar to the above

$$I = (-2\pi\omega)^{-\frac{d}{2}} \int (1+y^{\mathrm{T}}y)^{\frac{1}{\gamma}} dy$$

where $y = (-2\pi\omega)^{1/2}\Sigma^{-1/2}(x-\mu)$. It is expressed in polar coordinates as

$$I = (-2\pi\omega)^{-\frac{d}{2}} S^{d-1} \int_0^\infty (1+r^2)^{\frac{1}{\gamma}} r^{d-1} dr$$

which leads that

$$c_\gamma = (-2\omega)^{\frac{d}{2}} \frac{\Gamma(-\frac{1}{\gamma})}{\Gamma(-\frac{1}{\gamma}-\frac{d}{2})}$$