Entropy **2010**, *12*, 262-274; doi:10.3390/e12020262

OPEN ACCESS

entropy

ISSN 1099-4300

www.mdpi.com/journal/entropy

Article

Entropy and Divergence Associated with Power Function and the Statistical Application

Shinto Eguchi * **and Shogo Kato**

The Institute of Statistical Mathematics, Tachikawa, Tokyo 190-8562, Japan

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

Received: 29 December 2009; in revised form: 20 February 2010 / Accepted: 23 February 2010 / Published: 25 February 2010

Abstract: In statistical physics, Boltzmann-Shannon entropy provides good understanding for the equilibrium states of a number of phenomena. In statistics, the entropy corresponds to the maximum likelihood method, in which Kullback-Leibler divergence connects Boltzmann-Shannon entropy and the expected log-likelihood function. The maximum likelihood estimation has been supported for the optimal performance, which is known to be easily broken down in the presence of a small degree of model uncertainty. To deal with this problem, a new statistical method, closely related to Tsallis entropy, is proposed and shown to be robust for outliers, and we discuss a local learning property associated with the method.

Keywords: Tsallis entropy; projective power divergence; robustness

1. Introduction

Consider a practical situation in which a data set $\{x_1, \cdots, x_n\}$ is randomly sampled from a probability density function of a statistical model $\{f_\theta(x) : \theta \in \Theta\}$, where θ is a parameter vector and Θ is the parameter space. A fundamental tool for the estimation of unknown parameter θ is the log-likelihood function defined by

$$\ell(\theta) = \frac{1}{n}\sum_{i=1}^{n} \log f_\theta(x_i) \tag{1}$$

which is commonly utilized by statistical researchers ranging over both frequentists and Bayesians. The maximum likelihood estimator (MLE) is defined by

$$\widehat{\theta} = \operatorname*{argmin}_{\theta \in \Theta} \ell(\theta) \tag{2}$$

The Fisher information matrix for θ is defined by

$$I_\theta = \int f_\theta(x) \frac{\partial}{\partial \theta} \log f_\theta(x) \frac{\partial}{\partial \theta^{\mathrm{T}}} \log f_\theta(x) dx \tag{3}$$

where θ^{T} denotes the transpose of θ. As the sample size n tends to infinity, the variance matrix of $\sqrt{n}(\hat{\theta} - \theta)$ converges to I_θ^{-1}. This inverse matrix exactly gives the lower bound in the class of asymptotically consistent estimators in the sense of matrix inequality, that is,

$$\mathrm{AV}_\theta(\tilde{\theta}) \geq I_\theta^{-1} \tag{4}$$

for any asymptotically consistent estimator $\tilde{\theta}$ of θ, where AV_θ denotes the limiting variance matrix under the distribution with the density $f_\theta(x)$.

On the other hand, the Boltzmann-Shannon entropy

$$H_0(p) = \int p(x) \log p(x) dx \tag{5}$$

plays a fundamental role in various fields, such as statistical physics, information science and so forth. This is directly related with the MLE. Let us consider an underlying distribution with the density function $p(x)$. The cross entropy is defined by

$$C_0(p, f_\theta) = -\int p(x) \log f_\theta(x) dx \tag{6}$$

We note that $C_0(p, f_\theta) = \mathrm{E}_p\{-\ell(\theta)\}$, where E_p denotes the expectation with respect to $p(x)$. Hence, the maximum likelihood principle is equal to the minimum cross entropy principle. The Kulback-Leibler (KL) divergence is defined by

$$D_0(p, q) = \int p(x) \log \frac{p(x)}{q(x)} dx \tag{7}$$

which gives a kind of information distance between p and q. Note that $D_0(p, f_\theta) = C_0(p, f_\theta) - C_0(p, p)$

An exponential (type) distribution model is defined by the density form

$$f_\theta(x) = \exp\{\theta^{\mathrm{T}} t(x) - \psi(\theta)\} \tag{8}$$

where $\psi(\theta)$ is the cumulant transform defined by $\log \int \exp\{\theta^{\mathrm{T}} t(x)\} dx$. Under the assumption of this family, the MLE has a number of convenient properties such as minimal sufficiency, unbiasedness, efficiency [1]. In particular, the MLE for the expectation parameter $\eta = \mathrm{E}_\theta\{t(X)\}$ is explicitly given by

$$\hat{\eta}_0 = \frac{1}{n} \sum_{i=1}^{n} t(x_i)$$

which is associated with a dualistic relation of the canonical parameter θ and the expectation parameter η [2,3]. Thus, the MLE satisfies such an excellent property, which is associated with logarithmic and exponential functions as in (2) and (8).

The MLE has been widely employed in statistics, in which the properties are supported in theoretical discussion, for example, as in [4]. However, the MLE has some inappropriate properties when the underlying distribution does not belong to the model $\{f_\theta(x) : \theta \in \Theta\}$. A statistical model is just simulation of the true distribution as Fisher pointed in [1]. The model, which is just used as a working model, is wrong in most practical cases. In such situations, the MLE does not show proper performance because of model uncertainty. In this paper we explore alternative estimation method than the MLE.

2. Power Divergence

The logarithmic transform for observed values is widely employed in data analysis. On the other hand, a power transformation defined by

$$t_\beta(x) = \frac{x^\beta - 1}{\beta}$$

often gives more flexible performance to get good approximation to normal distribution [5]. In analogy with this transform, the power cross entropy is defined by

$$C_\beta(p, q) = -\int p(x)\frac{q(x)^\beta - 1}{\beta}dx + \int \frac{q(x)^{\beta+1}}{\beta+1}dx$$

where β is a positive parameter. Thus, it is defined by the power transform of the density. If we take the limit of β to 0, then $C_\beta(p, q)$ converges to $C_0(p, q)$, which is given in (6). In fact, the power parameter β is not fixed, so that different β's give different behaviors of the power entropy. The diagonal power entropy is defined by

$$H_\beta(p) = \int \frac{(\beta+1)p(x) - p(x)^{\beta+1}}{\beta(\beta+1)}dx$$

which is given by taking C_β the diagonal. Actually, this is equivalent to Tsallis q-entropy with the relation $\beta = q - 1$.

Let $\{x_1, \cdots, x_n\}$ be a random sample from unknown density function $p(x)$. Then we define the empirical mean power likelihood by

$$\ell_\beta(\theta) = \frac{1}{n}\sum_{i=1}^{n}\frac{f_\theta(x_i)^\beta - 1}{\beta} - \kappa_\beta(\theta) \tag{9}$$

where $\kappa_\beta(\theta) = \int f_\theta(x)^{\beta+1}dx/(\beta+1)$. See [6–9] for statistical applications. Accordingly, the minus expectation value of $\ell_\beta(\theta)$ is equal to $C_\beta(p, f_\theta)$. In general, a relation of cross and diagonal entropy leads to the inequality $C_\beta(p, q) \leq C_\beta(p, p)$, from which we define the power divergence by

$$D_\beta(p, q) = C_\beta(p, q) - C_\beta(p, p)$$

We extend power entropy and divergence defined over the space of all density functions, which are not always assumed to have a total mass one. In particular, this extension is useful for proposing boosting methods [10–16].

This derivation can be extended by a generator function U. Assume that $U(t)$ is strictly increasing and convex. The Fencel duality discussion leads to a conjugate convex function of $U(t)$ defined by

$$U^*(s) = \max_{t\in\mathbb{R}}\{st - U(t)\} \tag{10}$$

which reduces to $U^*(s) = s\xi(s) - U(\xi(s))$, where $\xi(s)$ is the inverse function of the derivation $\dot{U}$ of U. Then, U-cross entropy is defined by

$$C_U(\mu,\nu) = -\int \mu(x)\xi(\nu(x))dx + \int U(\xi(\nu(x)))dx$$

Similarly U-divergence is defined by

$$D_U(\mu,\nu) = \int \{U^*(\mu(x)) + U(\xi(\nu(x))) - \mu(x)\xi(\nu(x))\}dx \tag{11}$$

We note that $D_U(\mu,\nu) = C_U(\mu,\nu) - C_U(\mu,\mu)$. By the definition of U^* in (10) we see that the integrand of the right-hand side of (11) is always nonnegative. The power divergence is one example of U-divergence by fixing

$$U_\beta(t) = \frac{1}{\beta+1}(1+\beta t)^{\frac{\beta+1}{\beta}}$$

The power divergence can be defined on $\mathcal{M}$ as

$$D_\beta(\mu,\nu) = \int \left\{\mu(x)\frac{\mu(x)^\beta - \nu(x)^\beta}{\beta} + \frac{\nu(x)^{\beta+1} - \mu(x)^{\beta+1}}{\beta+1}\right\}dx \tag{12}$$

for μ and ν of $\mathcal{M}$ [17]. Thus $U_\beta(t)$ is strictly increasing and convex, which implies that the integrand of the right-hand side of (12) is nonnegative.

To explore this, it seems sufficient to restrict the definition domain of D_β to $\mathcal{P}$. However, we observe that the restriction is not useful for statistical considerations. We discuss the restriction on the projective space as follows. Fix two functions μ and ν in $\mathcal{M}$. We say that μ and ν are projectively equivalent if there exists a positive scalar λ such that

$$\nu(x) = \lambda\mu(x) \quad \text{(a.e. } x\text{)}$$

Thus, we write $\nu \sim \mu$. Similarly, we call a divergence D defined on $\mathcal{M}$ projectively invariant if for all $\lambda > 0, \kappa > 0$

$$D(\lambda\mu, \kappa\nu) = D(\mu,\nu) \tag{13}$$

We can derive a variant of power divergence as

$$\Delta_\beta(\mu,\nu) = \frac{1}{\beta(\beta+1)}\log\int \mu(x)^{\beta+1}dx - \frac{1}{\beta}\log\int \mu(x)\nu(x)^\beta dx + \frac{1}{\beta+1}\log\int \nu(x)^{\beta+1}dx$$

See Appendix 1 for the derivation. Immediately, we observe Δ_β satisfies (13), or projective invariance. Hereafter, we call Δ_β the projective power divergence. In this way, for $p(x) = \mu(x)/\int \mu(x)dx$ and $q(x) = \nu(x)/\int \nu(x)dx$, it is obtained that

$$\Delta_\beta(p,q) = \Delta_\beta(\mu,\nu)$$

If we take a specific value of β, then

$$\Delta_{\beta=1}(\mu,\nu)=\frac{1}{2}\log\frac{\int\mu(x)^2dx\int\nu(x)^2dx}{\Big(\int\mu(x)\nu(x)dx\Big)^2}$$

and

$$\lim_{\beta\to0}\Delta_\beta(\mu,\nu)=D_0\Big(\frac{\mu}{\int\mu(x)dx},\frac{\nu}{\int\nu(x)dx}\Big)$$

where D_0 is nothing but the KL divergence (7). We observe that the projective power divergence satisfies information additivity. In fact, if we write p and q as $p(x_1,x_2)=p_1(x_1)p_2(x_2)$ and $q(x_1,x_2)=q_1(x_1)q_2(x_2)$, respectively, then

$$\Delta_\beta(p,q)=\Delta_\beta(p_1,q_1)+\Delta_\beta(p_2,q_2)$$

which means information additivity. We note that this property is not satisfied by the original power divergence D_β. Furthermore, we know that Δ_β associates with the Pythagorean identity in the following.
Proposition 1. Assume that there exist three different points p,q and r in $\mathcal{M}$ satisfying

$$\Delta_\beta(p,r)=\Delta_\beta(p,q)+\Delta_\beta(q,r)\tag{14}$$

Define a path $\{p_t\}_{0\le t\le1}$ connecting p with q and a path $\{r_s\}_{0\le s\le1}$ connecting r with q as

$$p_t(x)=(1-t)p(x)+tq(x),\quad\{r_s(x)\}^\beta=(1-s)\{r(x)\}^\beta+s\{q(x)\}^\beta$$

Then

$$\Delta_\beta(p_t,r_s)=\Delta_\beta(p_t,q)+\Delta_\beta(q,r_s)\tag{15}$$

holds for all t $(0<t<1)$ and all s $(0<s<1)$.

Proof is given in Appendix 2. This Pytahgorean-type identity is also satisfied with the D_β [16].

3. Minimum Power Divergence Method

In the previous section we introduce a statistical method defined by minimization of the projective power divergence discussed. By the definition of Δ_β the cross projective power entropy is led to

$$\Gamma_\beta(\mu,\nu)=-\frac{1}{\beta}\log\int\mu(x)\nu(x)^\beta dx+c_\beta(\theta)$$

where $c_\beta(\theta)=(\beta+1)^{-1}\log\{\int f_\theta(x)^{\beta+1}dx\}$. We see that $\Delta_\beta(\mu,\nu)=\Gamma_\beta(\mu,\nu)-\Gamma_\beta(\mu,\mu)$. Hence, this decomposition leads the empirical analogue based on a given data set $\{x_1,\cdots,x_n\}$ to

$$L_\beta(\theta)=\frac{1}{\beta}\log\Big(\frac{1}{n}\sum_{i=1}^n f_\theta(x_i)^\beta\Big)-c_\beta(\theta)\tag{16}$$

which we call the mean power likelihood with the index β. Thus, the minus expectation of $L_\beta(\theta)$ with respect to the unknown density function $p(x)$ equals to $\Gamma_\beta(p,f_\theta)$. The limit of β to 0 leads that $L_\beta(\theta)$

converges to $\ell_0(\theta)$. Assume that $\{x_1, \cdots, x_n\}$ is a random sample exactly from $f_\theta(x)$. Then the strong law of large number yields that

$$L_\beta(\theta') \longrightarrow -\Gamma(f_\theta, f_{\theta'})$$

as n increases to infinity. From the property associated with the projective power divergence it follows that $\Gamma(f_\theta, f_{\theta'}) \geq \Gamma(f_\theta, f_\theta)$, which implies that $\theta = \mathrm{argmin}_{\theta' \in \Theta} \Gamma(f_\theta, f_{\theta'})$. Consequently, we conclude that the estimator $\widehat{\theta}_\beta = \mathrm{argmin}_{\theta' \in \Theta} L_\beta(\theta')$ converges to θ almost surely. The proof is similar to that for the MLE in Wald [18]. In general any minimum divergence estimator satisfies the strong consistency in the asymptotical sense.

The estimator $\widehat{\theta}_\beta$ is associated with the estimating function,

$$s_\beta(x, \theta) = f_\theta(x)^\beta \Big\{ s(x, \theta) - \frac{\partial}{\partial \theta} c_\beta(\theta) \Big\} \tag{17}$$

where $s(x, \theta)$ is the score vector, $(\partial/\partial\theta) \log f_\theta(x)$. We observe that the estimating function is unbiased in the sense that $\mathrm{E}_\theta\{s_\beta(x, \theta)\} = 0$. This is because

$$\mathrm{E}_\theta\{s_\beta(x, \theta)\} = \int f_\theta(x)^{\beta+1} s(x, \theta) dx - \int f_\theta(x)^{\beta+1} dx \frac{\partial}{\partial \theta} c_\beta(\theta) = 0$$

Thus the estimating equation is given by

$$S_\beta(\theta) = \frac{1}{n} \sum_{i=1}^{n} s_\beta(x_i, \theta) = 0$$

We see that the gradient vector of $L_\beta(\theta)$ is proportional to $S_\beta(\theta)$ as

$$\frac{\partial}{\partial \theta} L_\beta(\theta) = \Big(\frac{1}{n} \sum_{i=1}^{n} f_\theta(x_i)^\beta \Big)^{-1} S_\beta(\theta)$$

Hence, the estimating function (17) exactly leads to the estimator $\widehat{\theta}_\beta$.

Accordingly, we obtain the following asymptotic normality

$$\sqrt{n}(\widehat{\theta}_\beta - \theta) \longrightarrow_{\mathcal{D}} N(0, \mathrm{AV}_\beta(\theta))$$

where $\longrightarrow_{\mathcal{D}}$ denotes convergence in law, and $N(\mu, V)$ denotes a normal distribution with mean vector μ and variance matrix V. Here, the limiting variance matrix is

$$\mathrm{AV}_\beta(\theta) = \Big\{ \mathrm{E} \frac{\partial s_\beta(x, \theta)}{\partial \theta} \Big\}^{-\mathrm{T}} \mathrm{var}(s_\beta(x, \theta)) \Big\{ \mathrm{E} \frac{\partial s_\beta(x, \theta)}{\partial \theta} \Big\}^{-1}$$

The inequality (4) implies $\mathrm{AV}_\beta(\theta) \geq I_\theta^{-1}$ for any β, which implies that any estimator $\widehat{\theta}_\beta$ is not asymptotically efficient, where I_θ denotes the Fisher information matrix defined in (3). In fact, the estimator $\widehat{\theta}_\beta$ becomes efficient only when $\beta = 0$, which is reduced to the MLE. Hence, there is no optimal estimator except for the MLE in the class $\{\widehat{\theta}_\beta\}_{\beta \geq 0}$ as far as the asymptotic efficiency is concerned.

3.1. Super Robustness

We would like to investigate the influence of the estimator $\widehat{\theta}_\beta$ against outliers. We consider outliers in a probabilistic manner. An observation x_o is called an outlier if $f_\theta(x_\mathrm{o})$ is very small. Let us carefully look at the estimating equation (17). Then we observe that the larger the value of β is, the smaller $\|s_\beta(x_o, \theta)\|$ for all outliers x_o. The estimator $\widehat{\theta}_\beta$ is solved as

$$\widehat{\theta}_\beta = \underset{\theta\in\Theta}{\mathrm{argsolve}}\{\sum_{i=1}^{n} s_\beta(x_i, \theta) = 0\}$$

which implies that, for a sufficiently large β, the estimating equation has little impact from outliers contaminated in the data set because the value of the integral $\int f_\theta^\beta$ is hardly influenced by the outliers. In this sense, $\widehat{\theta}_\beta$ is robust for such β [19]. From an empirical viewpoint, we know it is sufficient to fix $\beta \geq 0.1$. In a case that $f_\theta(x)$ is absolutely continuous in $\mathbb{R}^p$ we see that $\lim_{|x|\to\infty} |s_\beta(x, \theta)| = 0$, which is quite contrast with the optimal robust method (cf. [20]). Consider an ϵ-contamination model

$$f_{\theta\epsilon}(x) = (1-\epsilon)f_\theta(x) + \epsilon\delta(x)$$

In this context, $\delta(x)$ is the density for outliers, which departs from the assumed density $f_\theta(x)$ with a large degree. It seems reasonable to suppose that $\int f_\theta(x)\delta(x)dx \simeq 0$. Thus if the true density function $p(x)$ equals $f_{\theta,\epsilon}(x)$, then $\widehat{\theta}_\beta$ becomes a consistent estimator for θ for all $\epsilon, 0 \leq \epsilon < 1$. In this sense we say $\widehat{\theta}_\beta$ satisfies super robustness. On the other hand, the mean power likelihood function $\ell_\beta(\theta)$ as given in (9) associates with the estimating function

$$f_\theta(x)^\beta s(x, \theta) - \frac{\partial}{\partial\theta}\kappa_\beta(\theta)$$

which is unbiased, but the corresponding estimator does not satisfy such super robustness.

Let us consider a multivariate normal model $N(\mu, V)$ with mean vector μ and variance matrix V in which the minimum projective power divergence method by (16) is applicable for the estimation of μ and V as follows:

$$(\widehat{\mu}_\beta, \widehat{V}_\beta) = \underset{(\mu,V)\in\mathbb{R}^p\times\mathcal{S}}{\mathrm{argmax}}\ L_\beta(\mu, V)$$

where $\mathcal{S}$ denotes the space of all symmetric, positive definite matrices.

Noting the projective invariance, we obtain

$$L_\beta(\mu, V) = \frac{1}{\beta}\log\Big[\frac{1}{n}\sum_{i=1}^{n}\exp\{-\frac{\beta}{2}(x_i-\mu)^{\mathrm{T}}V^{-1}(x_i-\mu)\}\Big] - \frac{1}{\beta+1}\log\det\big(\frac{V}{\beta+1}\big) \tag{18}$$

from which the estimating equation gives the weighted mean and variance as

$$\mu = \frac{\sum_{i=1}^n w(x_i,\mu,V)^\beta x_i}{\sum_{i=1}^n w(x_i,\mu,V)^\beta}, \tag{19}$$

$$V = (\beta+1)\frac{\sum_{i=1}^n w(x_i,\mu,V)^\beta (x_i-\mu)(x_i-\mu)^{\mathrm{T}}}{\sum_{i=1}^n w(x_i,\mu,V)^\beta} \tag{20}$$

where $w(x, \mu, V)$ is the weight function defined by $\exp\{-\frac{1}{2}(x-\mu)^{\mathrm{T}}V^{-1}(x-\mu)\}$. Although we do not know the explicit solution, a natural iteration algorithm can be proposed that the left-hand sides of (19)

and (20), say (μ_{t+1}, V_{t+1}) are both updated by plugging (μ_t, V_t) in the right-hand sides of (19) and (20). Obviously, for the estimator $(\widehat{\mu}_\beta, \widehat{V}_\beta)$ with $\beta = 0$, or the MLE, we need no iteration step but the sample mean vector and sample variance matrix as the exact solution.

3.2. Local Learning

We discuss a statistical idea beyond robustness. Since the expression (16) is inconvenient to investigate the behavior of the mean expected power likelihood function, we focus on

$$I_\beta(\theta) = \frac{1}{\beta}\Big\{ \int f_\theta(x)^\beta p(x)dx - 1\Big\}$$

as a core term, where $p(x)$ is the true density function, that is, the underlying distribution generating the data set. We consider K mixture model, while $p(x)$ is modeled as ϵ-contaminated density function $f_{\theta,\epsilon}(x)$ in the previous section. Thus, $p(x)$ is written by K different density functions $p_k(x)$ as follows:

$$p(x) = \pi_1 p_1(x) + \cdots + \pi_K p_K(x) \tag{21}$$

where π_k denotes the mixing ratio. We note that there exists redundancy for this modeling unless $p_k(x)$s are specified. In fact, the case in which $\pi_1 = 1$ and $p_1(x)$ is arbitrarily means no restriction for $p(x)$. However, we discuss $I_\beta(\theta)$ on this redundant model and find that

$$I_\beta(\mu, V) = \frac{1}{\beta}\Big(\sum_{k=1}^{K} \pi_k \{(2\pi)^p \det(V)\}^{-\frac{\beta}{2}} \int \exp\{-\frac{\beta}{2}(x-\mu)^{\mathrm{T}} V^{-1}(x-\mu)\} p_k(x)dx - 1\Big) \tag{22}$$

We confirm

$$I_0(\mu, V) = -\frac{1}{2}\Big\{ \sum_{k=1}^{K} \pi_k \int (x-\mu)^{\mathrm{T}} V^{-1}(x-\mu) p_k(x)dx + \log\det(V)\Big\}$$

taking the limit of β to 0. It is noted that $I_0(\mu, V)$ has a global maximizer $(\widehat{\mu}, \widehat{V})$ that is a pair of the mean vector and variance matrix with respect to $p(x)$ since we can write

$$I_0(\mu, V) = -\frac{1}{2}\Big\{ (\mu - \widehat{\mu}) V^{-1} (\mu - \widehat{\mu}) + \mathrm{trace}(\widehat{V} V^{-1}) + \log\det(V)\Big\}$$

This suggests a limitation of the maximum likelihood method. The MLE cannot change $N(\widehat{\mu}, \widehat{V})$ as the estimative solution even if the true density function is arbitrarily in (21). On the other hand, if β becomes larger, then the graph of $I_\beta(\mu, V)$ is flexibly changed in accordance with $p(x)$ in (21). For example, we assume

$$p(x) = \pi_1 g(x, \mu_1, V_1) + \cdots + \pi_K g(x, \mu_K, V_K) \tag{23}$$

where $g(x, \mu_k, V_k)$ is a normal density function $N(\mu_k, V_k)$. Then,

$$I_\beta(\mu, V) = \frac{1}{\beta}\Big(\sum_{k=1}^{K} \pi_k \{\beta^{-p}(2\pi)^p \det(V)\}^{\frac{1-\beta}{2}} \int g\big(x, \mu, \beta^{-1}V\big) g(x, \mu_k, V_k)dx - 1\Big)$$

Here, we see a formula

$$\int g(x,\mu,V)g(x,\mu^*,V^*)dx = g(\mu,\mu^*,V+V^*) \tag{24}$$

as shown in Appendix 3, from which we get that

$$I_\beta(\mu,V) = \frac{1}{\beta}\Big(\beta^{-p}\{(2\pi)^p \det(V)\}^{\frac{1-\beta}{2}} \sum_{k=1}^{K} \pi_k g\big(\mu,\mu_k,\beta^{-1}V+V_k\big) - 1\Big)$$

In particular, when $\beta = 1$,

$$I_1(\mu,V) = \sum_{k=1}^{K} \pi_k g\big(\mu,\mu_k,V+V_k\big) - 1$$

which implies that $I_1(\mu,O) = p(\mu) - 1$, where O is a zero matrix and $p(\cdot)$ is defined as in (23). If the normal mixture model has K modes, $I_1(\mu,V)$ has the same K modes for sufficiently small $\det V$. Therefore, the expected $I_\beta(\mu,V)$ with a large β adaptively behaves according to the true density function. This suggests that the minimum projective power divergence method can improve the weak point of the MLE if the true density function has much degree of model uncertainty. For example, such an adaptive selection for β is discussed in principal component analysis (PCA), which enables us to providing explanatory analysis rather than the conventional PCA.

Consider a problem for extracting principal components in which the data distribution has a density function with multimodality as described in (21). Then we wish to search all the sets of the principal vectors for V_k with $k = 1,\cdots,K$. The minimum projective power divergence method can properly provide the PCA to search the principal vectors for V_k at the centers μ_k separately for $k = 1,\cdots,K$. First we determine the first starting point, say $(\mu^{(1)},V^{(1)})$ in which we employ the iteratively reweighted algorithm (19) and (20) starting from $(\mu^{(1)},V^{(1)})$, so that we get the first estimator $(\widehat{\mu}^{(1)},\widehat{V}^{(1)})$. Then the estimator $\widehat{V}^{(1)}$ gives the first PCA with the center $\widehat{\mu}^{(1)}$ by the standard method. Next, we updates the second starting point $(\mu^{(2)},V^{(2)})$ to keep away from the first estimator $(\widehat{\mu}^{(1)},\widehat{V}^{(1)})$ by a heuristic procedure based on the weight function $w(x,\mu,V)$ (see [22] for the detailed discussion). Starting from $(\mu^{(2)},V^{(2)})$, the same algorithm (19) and (20) leads to the second estimator $(\widehat{\mu}^{(2)},\widehat{V}^{(2)})$ with the second PCA with the center $\widehat{V}^{(2)}$. In this way, we can make this sequential procedure to explore the multimodal structure with an appropriately determined stopping rule.

4. Concluding Remarks

We focus on that the optimality property of the likelihood method is fragile under model uncertainty. Such weakness frequently appears in practice when we got a data set typically from an observational study rather than a purely randomized experimental study. However, the usefulness of likelihood method is supported as the most excellent method in statistics. We note that the minimum projective power divergence method reduces to the MLE by taking a limit of the index β to 0 since it has one degree of freedom of β as a choice of method. A data-adaptive selection of β is possible by cross validation method. However, an appropriate model selection criterion is requested for faster computation.

Recently novel methods for pattern recognition from machine learning paradigm have been proposed [23–25]. These approaches are directly concerned with the true distribution in a framework of probability

approximate correct (PAC) learning in the computational learning theory. We need to employ this theory for the minimum projective power divergence method. In statistical physics there are remarkable developments on Tsallis entropy with reference to disequilibrium state, chaos phenomena, scale free network and econophysics. We should explore these developments from the statistical point of view.

Acknowledgements

We thank to the anonymous referees for their useful comments and suggestions, in particular, on Proposition 1.

Appendix 1

We introduce the derivation of Δ_β as follows. Consider the minimization for scalar multiplicity as

$$\kappa(\mu,\nu) = \underset{\kappa>0}{\operatorname{argmin}} D_\beta(\mu, \kappa\nu)$$

The gradient is

$$\frac{\partial}{\partial\kappa} D_\beta(\mu,\kappa\nu) = -\kappa^{\beta-1}\int \nu(x)^\beta \mu(x)dx + \kappa^\beta \int \nu(x)^{\beta+1}dx$$

which leads to $\kappa(\mu,\nu) = \int \nu(x)^\beta\mu(x)dx / \int \nu(x)^{\beta+1}dx$. Hence

$$\min_{\kappa>0} D_\beta(\mu,\kappa\nu) = \frac{1}{\beta(\beta+1)}\Big\{\int \mu(x)^{\beta+1}dx - \frac{\Big(\int \mu(x)\nu(x)^\beta dx\Big)^{\beta+1}}{\Big(\int \nu(x)^{\beta+1}dx\Big)^\beta}\Big\}$$

Taking the ratio as

$$\begin{aligned}\Delta_\beta(\mu,\nu) &= \frac{1}{\beta(\beta+1)}\log\frac{\Big(\int \mu(x)^{\beta+1}dx\Big)\Big(\int \nu(x)^{\beta+1}dx\Big)^\beta}{\Big(\int \mu(x)\nu(x)^\beta dx\Big)^{\beta+1}} \\ &= \frac{1}{\beta(\beta+1)}\log\int \mu(x)^{\beta+1}dx - \frac{1}{\beta}\log\int \mu(x)\nu(x)^\beta dx + \frac{1}{\beta+1}\log\int \nu(x)^{\beta+1}dx\end{aligned}$$

concludes the derivation of Δ_β in (13).

Appendix 2

We give a proof of Proposition 1.

Proof. By definition we get that

$$\Delta_\beta(p,r) - \{\Delta_\beta(p,q) + \Delta_\beta(q,r)\} = \frac{1}{\beta}\log\frac{\int p(x)q(x)^\beta dx \int q(x)r(x)^\beta dx}{\int q(x)^{\beta+1}dx \int p(x)r(x)^\beta dx}$$

which implies

$$\frac{\int p(x)q(x)^\beta dx \int q(x)r(x)^\beta dx}{\int q(x)^{\beta+1}dx \int p(x)r(x)^\beta dx} = 1 \tag{25}$$

from (14). Similarly,

$$\Delta_\beta(p_t,r_s) - \{\Delta_\beta(p_t,q) + \Delta_\beta(q,r_s)\} = \frac{1}{\beta}\log\frac{\int p_t(x)q(x)^\beta dx \int q(x)r_s(x)^\beta dx}{\int q(x)^{\beta+1}dx \int p_t(x)r_s(x)^\beta dx}$$

which is written as

$$\frac{1}{\beta}\log\frac{(1-t)\dfrac{\int p(x)q(x)^{\beta}dx}{\int q(x)^{\beta+1}dx}+t}{(1-t)\dfrac{\int p(x)r_s(x)^{\beta}dx}{\int q(x)r_s(x)^{\beta}dx}+t} \tag{26}$$

Furthermore, (26) is rewritten as

$$\frac{1}{\beta}\log\frac{(1-t)\dfrac{\int p(x)q(x)^{\beta}dx}{\int q(x)^{\beta+1}dx}+t}{(1-t)\dfrac{\int\{(1-s)p(x)r(x)^{\beta}+sp(x)q(x)^{\beta}\}dx}{\int\{(1-s)q(x)r(x)^{\beta}+sq(x)^{\beta+1}\}dx}+t}$$

which is

$$\frac{1}{\beta}\log\frac{(1-t)\dfrac{\int p(x)q(x)^{\beta}dx}{\int q(x)^{\beta+1}dx}+t}{(1-t)\dfrac{\int p(x)r(x)^{\beta}dx}{\int q(x)r(x)^{\beta}dx}\dfrac{1-s+s\dfrac{\int p(x)q(x)^{\beta}dx}{\int p(x)r(x)^{\beta}dx}}{1-s+s\dfrac{\int q(x)^{\beta+1}dx}{\int q(x)r(x)^{\beta}dx}}+t}$$

¿From (25) we can write

$$\Xi=\frac{\int p(x)q(x)^{\beta}dx}{\int q(x)^{\beta+1}dx}=\frac{\int p(x)r(x)^{\beta}dx}{\int q(x)r(x)^{\beta}dx}$$

Then, we conclude that

$$\Delta_{\beta}(p_t,r_s)-\{\Delta_{\beta}(p_t,q)+\Delta_{\beta}(q,r_s)\}=\frac{1}{\beta}\log\frac{(1-t)\Xi+t}{(1-t)\Xi\dfrac{(1-s)+s\Xi}{(1-s)+s\Xi}+t}$$

which vanishes for any $s, 0<s<1$ and $t, 0<t<1$. This completes the proof. □

Appendix 3

By writing a p-variate normal density function by

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

we have the formula

$$\int g(x,\mu,V)g(x,\mu^*,V^*)dx=g(\mu,\mu^*,V+V^*) \tag{27}$$

The proof of this formula is immediate. In fact, the left-hand side of (27) is written by

$$(2\pi)^p\{\det(V)\det(V^*)\}^{-\frac{1}{2}}\exp\{-\frac{1}{2}\mu^{\mathrm{T}}V^{-1}\mu-\frac{1}{2}\mu^{*\mathrm{T}}V^{*-1}\mu^*\}$$
$$\times\int\exp\{-\frac{1}{2}(x-A^{-1}b)^{\mathrm{T}}A(x-A^{-1}b)\}dx$$

where

$$A = V^{-1} + V^{*-1}, \quad b = V^{-1}\mu + V^{*-1}\mu^*$$

Hence, we get

$$\{(2\pi)^p \det(V) \det(V^*) \det(V^{-1} + V^{*-1})\}^{-\frac{1}{2}} \exp\{\frac{1}{2} b^{\mathrm{T}} A^{-1} b - \frac{1}{2}\mu^{\mathrm{T}} V^{-1}\mu - \frac{1}{2}\mu^{*\mathrm{T}} V^{*-1}\mu^*\}$$

Noting that

$$\{(2\pi)^p \det(V) \det(V^*) \det(V^{-1} + V^{*-1})\}^{-\frac{1}{2}} = \{(2\pi)^p \det(V + V^*)\}^{-\frac{1}{2}} \tag{28}$$

and

$$\begin{aligned} & \exp\{\frac{1}{2} b^{\mathrm{T}} A^{-1} b - \frac{1}{2}\mu^{\mathrm{T}} V^{-1}\mu - \frac{1}{2}\mu^{*\mathrm{T}} V^{*-1}\mu^*\} \\ = \; & \exp\{\frac{1}{2}\mu^{\mathrm{T}} V^{-1}(V^{-1} + V^{*-1})^{-1}\{I - (V^{-1} + V^{*-1})V\}V^{-1}\mu \\ & + \frac{1}{2}\mu^{*\mathrm{T}} V^{*-1}(V^{-1} + V^{*-1})^{-1}\{I - (V^{-1} + V^{*-1})V^*\}V^{*-1}\mu^* \\ & - \frac{1}{2}\mu^{\mathrm{T}} V^{-1}(V^{-1} + V^{*-1})^{-1} V^{*-1}\mu^*\} \end{aligned}$$

it is obtained that

$$\exp\{-\frac{1}{2}(\mu - \mu^*)^{\mathrm{T}}(V + V^*)^{-1}(\mu - \mu^*)\} \tag{29}$$

because of $V^{-1}(V^{-1} + V^{*-1})^{-1}V^{*-1} = (V + V^*)^{-1}$. Therefore, (28) and (29) imply (24). □

References

1. Fisher, R.A. On the mathematical foundations of theoretical statistics. *Philos. Trans. Roy. Soc. London Ser. A* **1922**, *222*, 309–368.
2. Amari, S. Lecture Notes in Statistics. In *Differential-Geometrical Methods in Statistics*; Springer-Verlag: New York, NY, USA, 1985; Volume 28.
3. Amari, S.; Nagaoka, H. Translations of Mathematical Monographs. In *Methods of Information Geometry*; Oxford University Press: Oxford, UK, 2000; Volume 191.
4. Akahira, M; Takeuchi, K. Lecture Notes in Statistics. In *Asymptotic Efficiency of Statistical Estimators: Concepts and Higher Order Asymptotic Efficiency*; Springer-Verlag: New York, NY, USA, 1981; Volume 7.
5. Box, G.E.P.; Cox, D.R. An Analysis of Transformations. *J. R. Statist. Soc. B* **1964**, *26*, 211–252.
6. Fujisawa, H.; Eguchi, S. Robust estimation in the normal mixture model. *J. Stat. Plan Inference* **2006**, *136*, 3989–4011.
7. Minami, M.; Eguchi, S. Robust blind source separation by beta-divergence. *Neural Comput.* **2002**, *14*, 1859–1886.
8. Mollah, N.H.; Minami, M.; Eguchi, S. Exploring latent structure of mixture ICA models by the minimum beta-divergence method. *Neural Comput.* **2006**, *18*, 166–190.

9. Scott, D.W. Parametric statistical modeling by minimum integrated square error. *Technometrics* **2001**, *43*, 274–285.
10. Eguchi, S.; Copas, J.B. A class of logistic type discriminant functions. *Biometrika* **2002**, *89*, 1–22.
11. Kanamori, T.; Takenouchi, T.; Eguchi, S.; Murata, N. Robust loss functions for boosting. *Neural Comput.* **2007**, *19*, 2183–2244.
12. Lebanon, G.; Lafferty, J. Boosting and maximum likelihood for exponential models. In *Advances in Neural Information Processing Systems* **2002**, *14*, 447-454. MIT Press: New York, NY, USA.
13. Murata, N.; Takenouchi, T.; Kanamori, T.; Eguchi, S. Information geometry of U-Boost and Bregman divergence. *Neural Comput.* **2004**, *16*, 1437–1481.
14. Takenouchi, T.; Eguchi, S. Robustifying AdaBoost by adding the naive error rate. *Neural Comput.* **2004**, *16*, 767–787.
15. Takenouchi, T.; Eguchi, S.; Murata, N.; Kanamori, T. Robust boosting algorithm for multiclass problem by mislabelling model. *Neural Comput.* **2008**, *20*, 1596–1630.
16. Eguchi, S. Information geometry and statistical pattern recognition. *Sugaku Expo.* **2006**, *19*, 197–216.
17. Basu, A.; Harris, I.R.; Hjort, N.L.; Jones, M.C. Robust and efficient estimation by minimising a density power divergence. *Biometrika* **1998**, *85*, 549–559.
18. Wald, A. Note on the Consistency of the Maximum Likelihood Estimate. *Ann. Math. Statist.* **1949**, *20*, 595–601.
19. Fujisawa, H.; Eguchi, S. Robust parameter estimation with a small bias against heavy contamination. *J. Multivariate Anal.* **2008**, *99*, 2053–2081.
20. Hampel, F.R.; Ronchetti, E.M.; Rousseeuw, P.J.; Stahel, W.A. *Robust Statistics: The Approach Based on Influence Functions*; Wiley: New York, NY, USA, 2005.
21. Eguchi, S.; Copas, J.A. Class of local likelihood methods and near-parametric asymptotics. *J. R. Statist. Soc. B* **1998**, *60*, 709–724.
22. Mollah, N.H.; Sultana, N.; Minami, M.; Eguchi, S. Robust extraction of local structures by the minimum beta-divergence method. *Neural Netw.* **2010**, *23*, 226–238.
23. Friedman, J.H.; Hastie, T.; Tibshirani, R. Additive logistic regression: A statistical view of boosting. *Annals of Statistics* **2000**, *28*, 337–407.
24. Hastie, T.; Tibishirani, R.; Friedman, J. *The Elements of Statistical Learning*; Springer: New York, NY, USA, 2001.
25. Schapire, R.E.; Freund, Y.; Bartlett, P.; Lee, W.S. Boosting the margin: A new explanation for the effectiveness of voting methods. *Ann. Statist.* **1998**, *26*, 1651–1686.

© 2010 by the authors; licensee Molecular Diversity Preservation International, Basel, Switzerland. This article is an open-access article distributed under the terms and conditions of the Creative Commons Attribution license http://creativecommons.org/licenses/by/3.0/.