MATHEMATICAL ENGINEERING
TECHNICAL REPORTS

The Information Geometric Structure of Generalized Empirical Likelihood Estimators

Tomoaki NISHIMURA and Fumiyasu KOMAKI

METR 2006–20 March 2006

DEPARTMENT OF MATHEMATICAL INFORMATICS
GRADUATE SCHOOL OF INFORMATION SCIENCE AND TECHNOLOGY
THE UNIVERSITY OF TOKYO
BUNKYO-KU, TOKYO 113-8656, JAPAN

WWW page: http://www.i.u-tokyo.ac.jp/mi/mi-e.htm

The Information Geometric Structure of Generalized Empirical Likelihood Estimators

Tomoaki NISHIMURA and Fumiyasu KOMAKI
Graduate School of Information Science and Technology
University of Tokyo

March, 2006

Abstract

The generalized empirical likelihood (GEL) method produces a class of estimators of parameters defined via general estimating equations. This class includes several important estimators, such as empirical likelihood (EL), exponential tilting (ET), and continuous updating estimators (CUE). We examine the information geometric structure of GEL estimators. We introduce a class of estimators closely related to the class of minimum divergence (MD) estimators and show that there is a one-to-one correspondence between this class and the class GEL.

Keywords and phrases: minimum divergence estimators, minimum discrepancy estimators, f-divergence, implied probabilities, Lagrange duality.

1 Introduction

Let x_i, $i = 1, \ldots, n$ be i.i.d. observations on a data vector x with unknown probability distribution F_0. Also, let θ be a $p \times 1$ parameter vector of interest and let $g(x, \theta)$ be a $q \times 1$ estimating function of the data observation x and a parameter θ, where $q \geq p$. The model we consider assumes that there exists a unique true parameter θ_0 such that $\mathrm{E}[g(x, \theta_0)] = 0$, where $\mathrm{E}[\cdot]$ denotes expectation with respect to F_0.

In this paper we consider generalized empirical likelihood (GEL) estimators (Smith (1997)) as a class of estimators of θ and examine the information geometric structure of this class.

GEL estimators are defined as follows: let $\rho(v)$ be a concave function of a scalar v whose domain is an interval $\mathcal{V}$ containing 0 as an interior point and which satisfies the normalization conditions

$$\rho(0) = 0, \quad \rho'(0) \ = \ \rho''(0) = \ -1, \tag{1}$$

and then the GEL estimator is defined by

$$\tilde{\theta}_{GEL} = \arg\min_{\theta\in\Theta} \sup_{\lambda\in\Lambda_n(\theta)} \frac{1}{n}\sum_{i=1}^{n} \rho(\lambda^{\mathrm{T}} g(x_i,\theta)), \tag{2}$$

where Θ is a parameter space, and $\Lambda_n(\theta) := \{\, \lambda \,|\, \lambda^{\mathrm{T}} g(x_i,\theta) \in \mathcal{V}, i = 1,\ldots,n\}$. All GEL estimators share the same first order asymptotic properties and the asymptotic variance is equal to that of the efficient generalized method of moments estimators studied by Hansen (1982).

GEL estimators include several important estimators suggested so far in the statistics and econometrics literatures. As shown by Smith (1997), the empirical likelihood (EL) estimator (Owen (1988), Qin and Lawless (1994)) and the exponential tilting (ET) estimator (Kitamura and Stutzer (1997), Imbens, Spady, and Johnson (1998)) are members of GEL estimators with $\rho(v) = \log(1-v)$ and $\rho(v) = 1 - \mathrm{e}^v$, respectively. As shown by Newey and Smith (2004), the continuous updating estimator (CUE) (Hasen, Heaton, and Yaron (1996)) is also a GEL estimator with $\rho(v) = -v^2/2 - v$.

Some GEL estimators are also members of the class of minimum divergence (MD) estimators formulated by Corcoran (1998) [1]. MD estimators are defined by using Csiszár's f-divergence (Csiszár (1967)). Let $f(u)$ be a convex function of a scalar u satisfying $f(1) = 0$ such that the domain of $f(u)$ is an interval $\mathcal{U}$ containing 1 as an interior point. For probability distributions P and Q, the f-divergence from P to Q is defined by

$$D_f(P,Q) = \int p(x) f\left(\frac{q(x)}{p(x)}\right) \mathrm{d}\nu(x), \tag{3}$$

where $p(x)$ and $q(x)$ are the densities of P and Q, respectively, with respect to a measure $\nu(x)$.

MD estimators are defined as follows: for fixed $\theta \in \Theta$, let $d_f(\theta)$ be the optimal value of the convex optimization problem

$$\min_{p_1,\ldots,p_n} \ \frac{1}{n}\sum_{i=1}^{n} f(np_i) \quad \text{s.t.} \ \sum_{i=1}^{n} p_i g(x_i,\theta) = 0, \ \sum_{i=1}^{n} p_i = 1. \tag{4}$$

Then the MD estimator is given by $\tilde{\theta}_{MD} = \arg\min_{\theta\in\Theta} \ d_f(\theta)$. For this problem, the objective function $\sum_{i=1}^{n} f(np_i)/n$ is the f-divergence from the empirical likelihood to the multinominal distribution $\{p_i\}_{i=1}^{n}$ which has the same support as the empirical likelihood.

Several important estimators are both GEL estimators and MD estimators. For example, EL and ET estimators can be derived as solutions to the problem above, where $f(u)$ is $-\log u$ and $u\log u$, respectively. In addition, as shown by Newey and Smith (2004), MD estimators based on the α-divergence, which is a subclass of the f-divergence, can

[1] MD estimators are sometimes called " minimum discrepancy " estimators. However, since in this paper some properties of divergence play an important role for elucidating properties of GEL estimators, we use the term "divergence " instead of "discrepancy ".

be formulated as GEL estimators [2]. However, GEL estimators do not have to be MD estimators, and furthermore, MD estimators based on an f-divergence other than the α-divergence are not GEL estimators.

In this paper, in order to study the information geometric structure of GEL estimators, we introduce a new class of estimators which have a structure similar to MD estimators. We do this by utilizing divergences obtained by extending the f-divergence (3) to be defined on measures instead of probability distributions and we consider estimators which minimize the extended f-divergence from the empirical distribution to a discrete measure [3]. We call them minimum extended divergence (MED) estimators and show that there is a one-to-one correspondence between GEL and MED estimators.

We also show that MD and MED estimators coincide when we use the α-divergence and that this fact results from an intrinsic property of the α-divergence. By using this equivalence between MD and MED based on the α-divergence, we show that the class of MD estimators based on the α-divergence is a subclass of GEL in a different way from Newey and Smith (2004).

Finally, we investigate relations between implied probabilities associated with GEL and discrete measures obtained by minimizing the extended f-divergences in MED. For a given function $\rho(v)$, the implied probability is defined by

$$\tilde{\pi}_i = \frac{\rho'(\tilde{\lambda}^{\mathrm{T}} g(x_i, \tilde{\theta}_{GEL}))}{\sum_{i'=1}^{n} \rho'(\tilde{\lambda}^{\mathrm{T}} g(x_{i'}, \tilde{\theta}_{GEL}))} , \quad i = 1, \ldots, n, \tag{5}$$

where $\tilde{\theta}_{GEL}$ and $\tilde{\lambda}$ are the saddle point of Eq.(2). Since implied probabilities take account of the structure of the model, they are important for construction of more efficient empirical estimates (Back and Brown (1993), Qin and Lawless (1994), Imbens (1997), Brown and Newey (2002)). In general, $\tilde{\pi}_i$ may take negative values and these negative implied probabilities sometimes cause problems in applications, such as the bootstrap method (Brown and Newey (2002)). We show that the implied probability with a GEL estimator is obtained by normalizing the discrete measure derived from the corresponding MED to sum to 1, and we can prevent the implied probability from taking negative values by considering only nonnegative measures for the corresponding MED.

The organization of this paper is as follows. In Section 2 we introduce a new class of estimators, that is, MED estimators. In Section 3 we discuss the duality between GEL and MED. Section 4 shows the equivalence between MD and MED based on the α-divergence. Section 5 presents relations between implied probabilities and MED.

[2] In Newey and Smith (2004), Cressie-Read divergence (Cressie and Read (1984)) is used in place of the α-divergence (Amari (1982)). These two divergences are equivalent under an appropriate transformation of parameters indexing the divergences.

[3] In this paper, a " measure " means a finite signed measure and may take negative values

2 Minimum extended divergence estimators

In this section, we introduce a class of estimators similar to MD estimators, which we call minimum extended divergence (MED) estimators.

To define MED estimators, we use the extended f-divergence to measures, which is obtained by modifying the f-divergence for probability distributions (3) (e.g., see Zhang (2004)). Suppose that $f(u)$ is continuously differentiable at $u = 1$. For a positive measure P and a measure Q, the extended f-divergence from P to Q is defined by

$$\begin{aligned} \bar{D}_f(P, Q) &= \int p(x) \left\{ f\left(\frac{q(x)}{p(x)}\right) - f'(1)\left(\frac{q(x)}{p(x)} - 1\right)\right\} \mathrm{d}\nu(x) \\ &= \int p(x) \bar{f}\left(\frac{q(x)}{p(x)}\right) \mathrm{d}\nu(x), \end{aligned} \tag{6}$$

where $\bar{f}(u) = f(u) - f'(1)(u-1)$ and $\bar{f}(u)$ satisfies $\bar{f}(1) = \bar{f}'(1) = 0$. If P and Q are both probability distributions, the extended f-divergence $\bar{D}_f(P, Q)$ coincides with the original f-divergence $D_f(P, Q)$. Note that if the domain of $\bar{f}(u)$ contains negative real numbers, Q may not be a positive measure and $q(x)$ may take negative values. In such a case, since the domain of $f(u)$ is the same as the domain of $\bar{f}(u)$, the original f-divergence (3) is defined for a measure Q such that $\int q(x)\mathrm{d}\nu(x) = 1$ and $p_i, i = 1, \ldots, n$ in the problem (4) may take negative values.

We define MED estimators as follows: for fixed $\theta \in \Theta$, let $\bar{d}_f(\theta)$ be the optimal value of a convex optimization problem

$$\min_{w_1,\ldots,w_n} \quad \frac{1}{n}\sum_{i=1}^{n} \bar{f}(nw_i) \quad \text{s. t.} \quad \sum_{i=1}^{n} w_i g(x_i, \theta) = 0. \tag{7}$$

Then the MED estimator is defined by $\tilde{\theta}_{MED} = \arg\min_{\theta\in\Theta} \bar{d}_f(\theta)$. Note that since we do not impose the normalization constraint for $\{w_i\}_{i=1}^n$ and the domain of $\bar{f}(u)$ may contain negative real numbers, the discrete measure $\{w_i\}_{i=1}^n$ is not necessarily a probability distribution or even a positive measure.

3 Duality between GEL and MED

In this section, we show that Lagrange duality holds between GEL and MED and that there is a one-to-one correspondence between the two classes of estimators. In this paper, in order to simplify description, we assume that a convex function takes the value ∞ outside its domain, and similarly that a concave function takes the value $-\infty$ outside its domain.

First, we derive the Lagrange dual problem for the convex optimization problem (7) in MED. The Lagrangian for the problem (7) is

$$L(w, \lambda) = \frac{1}{n}\sum_{i=1}^{n} \bar{f}(nw_i) - \lambda^{\mathrm{T}} \sum_{i=1}^{n} w_i g(x_i, \theta),$$

where λ is the Lagrange multiplier for $\sum_{i=1}^n w_i g(x_i, \theta) = 0$. The objective function of the dual problem is

$$\begin{aligned}\inf_{w_1,\ldots,w_n} L(w, \lambda) &= -\frac{1}{n}\sum_{i=1}^n \sup_{w_i} \left\{ n w_i \lambda^{\mathrm{T}} g(x_i, \theta) - \bar{f}(n w_i) \right\} \\ &= -\frac{1}{n}\sum_{i=1}^n \bar{f}^*(\lambda^{\mathrm{T}} g(x_i, \theta)),\end{aligned}$$

where $\bar{f}^*(v)$ is the conjugate function of $\bar{f}(u)$ by the Fenchel-Legendre transformation, i.e. $\bar{f}^*(v) = \sup_u \{uv - \bar{f}(u)\}$ and $\bar{f}^*(v)$ is a convex function whose domain $\mathcal{V}$ consists of v for which the supremum is finite (e.g., see Rockafellar (1970)). Note that since $\bar{f}(u)$ takes ∞ outside its domain $\mathcal{U}$, $\bar{f}^*(v)$ depends on $\mathcal{U}$, which is not described explicitly. Letting $\rho(v) = -\bar{f}^*(v)$, the Lagrange dual problem for the problem (7) can be written as

$$\max_{\lambda} \ \frac{1}{n}\sum_{i=1}^n \rho(\lambda^{\mathrm{T}} g(x_i, \theta)). \tag{8}$$

Since $\rho(v)$ is concave, minimizing the optimal value of this convex optimization problem with respect to θ yields a GEL estimator if $\rho(v)$ satisfies the normalization conditions (1).

Next, we consider the normalization conditions on $\bar{f}(u)$ corresponding to the normalization conditions (1) on $\rho(v)$. We assume that $\bar{f}(u)$ is twice continuously differentiable in a neighborhood of 1 and that the inverse function of $\bar{f}'(u)$, $(\bar{f}')^{-1}(v)$, exists in the neighborhood. From the definition of the conjugate function, for $v \in \mathcal{V}$ such that $v = \bar{f}'(u)$ in the neighborhood, $\rho(v) = -\bar{f}^*(v)$ can be written as

$$\rho(v) = \bar{f}((\bar{f}')^{-1}(v)) - v(\bar{f}')^{-1}(v),$$

and then we obtain

$$\rho'(v) \ = \ -(\bar{f}')^{-1}(v)\ , \quad \rho''(v) \ = \ -\frac{1}{\bar{f}''((\bar{f}')^{-1}(v))}\ .$$

From these equations, the normalization conditions (1) on $\rho(v)$ can be written as

$$\bar{f}(1) \ = \ \bar{f}'(1) = \ 0, \quad \bar{f}''(1) = \ 1. \tag{9}$$

The first two normalization conditions above are required for the extended f-divergence (6) to be well-defined. As long as $\bar{f}(u)$ satisfies these two normalization conditions and $\bar{f}''(1) > 0$, the last normalization condition can always be imposed by replacing $\bar{f}(u)$ with $\bar{f}(u)/\bar{f}''(1)$, which does not affect the MED estimator. As shown by Newey and Smith (2004), under the normalization conditions (1) on $\rho(v)$ and several regularity conditions, then

$$\min_{\theta\in\Theta} \sup_{\lambda\in\Lambda_n(\theta)} 2\sum_{i=1}^n \rho(\lambda^{\mathrm{T}} g(x_i, \theta)) \ \xrightarrow{\mathrm{d}} \ \chi^2_{q-p} \tag{10}$$

holds and this statistic can be used for testing the overidentified model, $\mathrm{E}[g(x,\theta_0)] = 0$, whereas the last normalization condition (9) on $\bar{f}(u)$ can be interpreted as the normalization for the metric derived from the divergence. Eguchi (1983) provides a generic way of constructing a metric from an arbitrary divergence on statistical manifolds. It is well known that the f-divergence results in a metric proportional to Fisher's information, and that the metric coincides with Fisher's information when $\bar{f}''(1) = 1$. Fisher's information is an important metric in the filed of information geometry and is given by the Kullback-Leibler divergence, the α-divergence and other divergences used often in statistics (e.g., see Amari and Nagaoka (2000), Zhang (2004)).

In the rest of this section we show that there is a one-to-one correspondence between the classes GEL and MED and that in fact the corresponding GEL and MED estimators coincide. We suppose that the convex function $\bar{f}(u)$ is closed, i.e. that the epigraph of $\bar{f}(u)$, $\{(u,t) \,|\, \bar{f}(u) \leq t\}$, is a closed set. Then the conjugate function of the conjugate function of $\bar{f}(u)$ is the original function, i.e. $\bar{f}^{**}(u) = (-\rho)^*(u) = \bar{f}(u)$ (e.g., see Rockafellar (1970)). Therefore we have a one-to-one correspondence via the Fenchel-Legendre transformation between $\rho(v)$ and $\bar{f}(u)$ satisfying the following Assumptions 1 and 2, respectively, and thus we also have a one-to-one correspondence between GEL and MED estimators via the respective functions.

Assumption 1. *(i) $-\rho(v)$ is closed. (ii) $\rho(v)$ is concave. (iii) The domain of $\rho(v)$, $\mathcal{V}$, is an interval containing 0 as an interior point. (iv) $\rho(v)$ is twice continuously differentiable in a neighborhood of 0. (v) $\rho(v)$ satisfies the normalization conditions (1).*

Assumption 2. *(i) $\bar{f}(u)$ is closed. (ii) $\bar{f}(u)$ is convex. (iii) The domain of $\bar{f}(u)$, $\mathcal{U}$, is an interval containing 1 as an interior point. (iv) $\bar{f}(u)$ is twice continuously differentiable in a neighborhood of 1. (v) $\bar{f}(u)$ satisfies the normalization conditions (9).*

We now show that the corresponding GEL and MED estimators coincide.

Theorem 3. *Suppose that $\rho(v)$ and $\bar{f}(u)$ satisfy Assumption 1.(ii)-(v) and 2, respectively, and $\bar{f}(u) = (-\rho)^*(u)$. Then for any $\theta \in \Theta$ the optimal values of the convex optimization problems (7) and (8) are equal, i.e.*

$$\inf_{w_1,\ldots,w_n} \left\{ \frac{1}{n}\sum_{i=1}^{n} \bar{f}(nw_i) \,\middle|\, \sum_{i=1}^{n} w_i g(x_i,\theta) = 0,\ nw_i \in \mathcal{U}, i = 1,\ldots,n \right\}$$
$$= \sup_{\lambda} \left\{ \frac{1}{n}\sum_{i=1}^{n} \rho(\lambda^{\mathrm{T}} g(x_i,\theta)) \,\middle|\, \lambda^{\mathrm{T}} g(x_i,\theta) \in \mathcal{V}, i = 1,\ldots,n \right\}.$$

Here we assume that the left-hand side of the equation above takes the value ∞ if the convex optimization problem (7) is infeasible and that the right-hand side takes the value $-\infty$ if the convex optimization problem (8) is infeasible. Thus the GEL estimator $\tilde{\theta}_{GEL}$ coincides with the MED estimator $\tilde{\theta}_{MED}$ if both of them uniquely exist.

Proof. First we show that the problem (7) for MED can be formulated as the Lagrange dual problem of the problem (8) for the corresponding GEL. Introducing new variables v_i, $i = 1, \ldots, n$, let us reformulate the problem (8) as

$$\min_{\lambda, v_1, \ldots, v_n} \quad -\frac{1}{n}\sum_{i=1}^{n} \rho(v_i) \quad \text{s. t.} \quad v_i = \lambda^{\mathrm{T}} g(x_i, \theta). \tag{11}$$

The Lagrangian of this problem is

$$L(\lambda, v, w) = -\frac{1}{n}\sum_{i=1}^{n} \rho(v_i) - \sum_{i=1}^{n} w_i (v_i - \lambda^{\mathrm{T}} g(x_i, \theta)),$$

where w_i, $i = 1, \ldots, n$ are the Lagrange multipliers for $v_i = \lambda^{\mathrm{T}} g(x_i, \theta)$, $i = 1, \ldots, n$. Minimizing over λ we find that $\inf_\lambda L(\lambda, v, w) = -\infty$ unless $w_i g(x_i, \theta) = 0$, $i = 1, \ldots, n$, in which case we have

$$\inf_{\lambda, v_1, \ldots, v_n} L(\lambda, v, w) = -\frac{1}{n}\sum_{i=1}^{n} \sup_{v_i} \{ n w_i v_i - (-\rho(v_i)) \} = -\frac{1}{n}\sum_{i=1}^{n} \bar{f}(n w_i).$$

Hence we obtain the problem (7) as the Lagrange dual problem of the problem (8). Since $\lambda = 0, v_i = 0, i = 1, \ldots, n$ is an interior feasible solution of the convex optimization problem (11), strong duality holds (e.g., see Boyd and Vandenberghe (2004)). Therefore the optimal value of the problem (11), which is equivalent to (8), is equal to the optimal value of the problem (7). Thus we obtain the desired result. □

Note that in Theorem 3 $-\rho(v)$ is not assumed to be closed, in which case we do not have a one-to-one correspondence between GEL and MED. We obtain the equivalence between GEL and MED estimators without requiring a one-to-one correspondence.

Example 1 (Empirical likelihood)**.** EL is an MD estimator by taking $f(u) = -\log u$ ($\mathcal{U} = \{u | u > 0\}$) and a GEL estimator by taking $\rho(v) = \log(1 - v)$ ($\mathcal{V} = \{v | v < 1\}$). For these $\rho(v)$ and $f(u)$, $\bar{f}(u) = (-\rho)^*(u)$ and $\bar{f}(u) = f(u) - f'(1)(u - 1)$ coincide and both can be expressed as $\bar{f}(u) = -1 + u - \log u$ ($\mathcal{U} = \{u | u > 0\}$).

Example 2 (Exponential tilting)**.** ET is an MD estimator by taking $f(u) = u \log u$ ($\mathcal{U} = \{u | u > 0\}$) and a GEL estimator by taking $\rho(v) = 1 - \mathrm{e}^v$ ($\mathcal{V} = \mathbf{R}$). For these $\rho(v)$ and $f(u)$, $\bar{f}(u) = (-\rho)^*(u)$ and $\bar{f}(u) = f(u) - f'(1)(u - 1)$ coincide as for the preceding example, and both can be expressed as $\bar{f}(u) = -1 + u - \log u$ ($\mathcal{U} = \{u | u > 0\}$).

Example 3 (Continuous updating)**.** CUE can be written as a GEL estimator by taking $\rho(v) = -v^2/2 - v$ ($\mathcal{V} = \mathbf{R}$). For this $\rho(v)$, we have that $\bar{f}(u) = u^2/2 + u + 1/2$ ($\mathcal{U} = \mathbf{R}$) The corresponding $f(u)$ can be written as $f(u) = \bar{f}(u) + a(u - 1)$ ($\mathcal{U} = \mathbf{R}$) for $^\forall a \in \mathbf{R}$. If we set $a = 1$ particularly, the associated divergence is the α-divergence with $\alpha = 3$, as described in the following section.

4 MD and MED based on the α-divergence

In this section we consider relations between MD and MED based on the α-divergence and show that the MD and MED estimators coincide when we use the α-divergence

The α-divergence $D_{f^{(\alpha)}}(P,Q)$ is defined as a subclass of the f-divergence by using a convex function $f^{(\alpha)}(u)$ indexed by a scalar parameter α. Although Amari (1982) introduces the α-divergence for probability distributions P and Q, in this paper we also consider the case where a measure Q may not be a probability distribution and may take negative values and therefore we consider two types of convex function, $f_j^{(\alpha)}(u)$, $j=1,2$ for the convex function $f^{(\alpha)}(u)$. For $\alpha \in \mathbf{R}$, let

$$h^{(\alpha)}(u) = \begin{cases} \frac{4}{1-\alpha^2}\left\{1 - u^{(1+\alpha)/2}\right\} & (\alpha \neq \pm 1) \\ u \log u & (\alpha = 1) \\ -\log u & (\alpha = -1). \end{cases} \tag{12}$$

Also, let

$$\mathcal{U}_1^{(\alpha)} = \begin{cases} \{\, u \,|\, u > 0\} & (\alpha \leq -1) \\ \{\, u \,|\, u \geq 0\} & \text{otherwise} \end{cases} \tag{13}$$

and

$$\mathcal{U}_2^{(\alpha)} = \begin{cases} \{\, u \,|\, u > 0\} & (\alpha \leq -1) \\ \mathbf{R} & (\alpha = 4k-1,\ k = 1,2,\dots) \\ \{\, u \,|\, u \geq 0\} & \text{otherwise}\,. \end{cases} \tag{14}$$

We define $f_j^{(\alpha)}(u)$ as

$$f_j^{(\alpha)}(u) = \begin{cases} h^{(\alpha)}(u) & (u \in \mathcal{U}_j^{(\alpha)}) \\ \infty & \text{otherwise,} \end{cases} \qquad j = 1,2. \tag{15}$$

From this definition the two functions $f_j^{(\alpha)}(u)$, $j=1,2$ are convex and, where both are defined, differ in value only when $\alpha = 4k-1$, $k = 1,2,\dots$. Note that in that case $f_1^{(\alpha)}(u)$ is defined only for nonnegative reals, whereas $f_2^{(\alpha)}(u)$ is defined for the whole of the reals. Therefore the α-divergence based on $f_1^{(\alpha)}(u)$, $D_{f_1^{(\alpha)}}(P,Q)$, is defined for probability distributions P and Q, whereas when $\alpha = 4k-1$, $k = 1,2,\dots$, the α-divergence based on $f_2^{(\alpha)}(u)$, $D_{f_2^{(\alpha)}}(P,Q)$, is defined for a probability distribution P and a normalized measure Q whose density $q(x)$ may take negative values. Note that the two types of convex function yield different MD estimators because the problem (4) for the class MD depends on the domain of the convex function $f(u)$ and that when $\alpha = 4k-1$, $k = 1,2,\dots$, p_i, $i = 1,\dots,n$ may take negative values if we adopt $f_2^{(\alpha)}(u)$.

Similarly we consider two types of extended α-divergence $\bar{D}_{f^{(\alpha)}}(P,Q)$. For $\alpha \in \mathbf{R}$, let

$$\bar{h}^{(\alpha)}(u) = \begin{cases} \frac{4}{1-\alpha^2}\left\{\frac{1-\alpha}{2} + \frac{1+\alpha}{2}u - u^{(1+\alpha)/2}\right\} & (\alpha \neq \pm 1) \\ 1 - u + u \log u & (\alpha = 1) \\ -1 + u - \log u & (\alpha = -1). \end{cases} \tag{16}$$

For $f_j^{(\alpha)}(u)$, $j=1,2$, the convex functions $\bar{f}_j^{(\alpha)}(u) = f_j^{(\alpha)}(u) - (f_j^{(\alpha)})'(1)(u-1)$, $j=1,2$ can be written as

$$\bar{f}_j^{(\alpha)}(u) = \begin{cases} \bar{h}^{(\alpha)}(u) & (u \in \mathcal{U}_j^{(\alpha)}) \\ \infty & \text{otherwise,} \end{cases} \qquad j=1,2, \tag{17}$$

and both of them satisfy Assumption 2. When $\alpha = 4k-1$, $k=1,2,\ldots$, the two types of convex function give different MED (GEL) estimators as in the case of MD estimators and the two concave functions $\rho_j^{(\alpha)}(v) = -(f_j^{(\alpha)})^*(v)$, $j=1,2$ are different because a conjugate function depends on the domain of the original convex function. However, these two concave functions have the same form in a neighborhood of 0 from the definition of a conjugate function and this common function can be expressed as

$$\rho^{(\alpha)}(v) = \begin{cases} \frac{2}{1+\alpha}\left\{1 - \left(1 - \frac{1-\alpha}{2}v\right)^{-(1+\alpha)/(1-\alpha)}\right\} & (\alpha \neq \pm 1) \\ 1 - \mathrm{e}^v & (\alpha = 1) \\ \log(1-v) & (\alpha = -1). \end{cases} \tag{18}$$

By setting $\alpha = -1$ and 1, we obtain EL and ET, respectively. In addition, as described in Example 4, by setting $\alpha = 3$ and adopting $\bar{f}_2^{(\alpha)}(u)$ we obtain CUE.

Example 4 (Continuation of Example 3)**.** We consider the case where $\alpha = 3$. CUE is an MED estimator by taking $\bar{f}_2^{(3)}(u)$ and $\rho_2^{(3)}(v) = -(\bar{f}_2^{(3)})^*(v) = -v^2/2 - v$ for $^\forall v \in \mathbf{R}$. On the other hand, for $\bar{f}_1^{(3)}(u)$, $\rho_1^{(3)}(v) = -(\bar{f}_1^{(3)})^*(v)$ can be written as

$$\rho_1^{(3)}(v) = \begin{cases} -\frac{1}{2}v^2 - v & v \geq -1 \\ \frac{1}{2} & v < -1 \end{cases}$$

and $\rho_1^{(3)}(v)$ is different from $\rho_2^{(3)}(v)$ when $v < -1$. Therefore we find that utilizing $\bar{f}_1^{(3)}(u)$ results in a different estimator from CUE.

We show that minimization of the α-divergence for probability distributions is equivalent to minimization of the extended α-divergence for measures in general settings. The results given below do not depend on which convex functions we adopt, $f_1^{(\alpha)}(u)$ or $f_2^{(\alpha)}(u)$. Therefore we write $f^{(\alpha)}(u)$ without distinguishing $f_1^{(\alpha)}(u)$ and $f_2^{(\alpha)}(u)$, and write $\bar{f}^{(\alpha)}(u)$ without distinguishing $\bar{f}_1^{(\alpha)}(u)$ and $\bar{f}_2^{(\alpha)}(u)$. Let P be a probability distribution and $\{Q_\xi | \xi \in \Xi\}$ be a family of measures indexed by a parameter ξ such that each element Q_ξ satisfies $\int q_\xi(x)\mathrm{d}\nu(x) = 1$.

Theorem 4. *For arbitrarily fixed $\alpha \in \mathbf{R}$, suppose that $D_{f^{(\alpha)}}(P, Q_\xi)$ can be defined for any $\xi \in \Xi$. Then for arbitrarily fixed $c_p > 0$ and $c_q > 0$, $\xi \in \Xi$ that minimizes $D_{f^{(\alpha)}}(P, Q_\xi)$ coincides with $\xi \in \Xi$ that minimizes $\bar{D}_{f^{(\alpha)}}(c_p P, c_q Q_\xi)$ if both of them uniquely exist.*

Proof. From Eq.(15) and (17) we have

$$\begin{aligned} \bar{D}_{f^{(-1)}}(c_p P, c_q Q_\xi) &= \int \left\{ c_q q_\xi(x) - c_p p(x) + c_p p(x) \log \frac{c_p p(x)}{c_q q_\xi(x)} \right\} \mathrm{d}\nu(x) \\ &= c_q - c_p - c_p \log \frac{c_p}{c_q} + c_p D_{f^{(-1)}}(P, Q_\xi)\,, \end{aligned}$$

$$\bar{D}_{f^{(1)}}(c_pP, c_qQ_\xi) = \int \left\{ c_p p(x) - c_q q_\xi(x) + c_q q_\xi(x) \log \frac{c_q q_\xi(x)}{c_p p(x)} \right\} \mathrm{d}\nu(x)$$
$$= c_p - c_q + c_q \log \frac{c_q}{c_p} + c_q D_{f^{(1)}}(P, Q_\xi)\,,$$

and for $\alpha \neq \pm 1$

$$\bar{D}_{f^{(\alpha)}}(c_pP, c_qQ_\xi) = \frac{4}{1-\alpha^2} \int \left\{ \frac{1-\alpha}{2} c_p p(x) + \frac{1+\alpha}{2} c_q q_\xi(x) - (c_p p(x))^{\frac{1-\alpha}{2}} (c_q q_\xi(x))^{\frac{1+\alpha}{2}} \right\} \mathrm{d}\nu(x)$$
$$= \frac{4}{1-\alpha^2} \left(\frac{1-\alpha}{2} c_p + \frac{1+\alpha}{2} c_q - c_p^{\frac{1-\alpha}{2}} c_q^{\frac{1+\alpha}{2}} \int p(x)^{\frac{1-\alpha}{2}} q_\xi(x)^{\frac{1+\alpha}{2}} \mathrm{d}\nu(x) \right)$$
$$= \frac{4}{1-\alpha^2} \left(\frac{1-\alpha}{2} c_p + \frac{1+\alpha}{2} c_q - c_p^{\frac{1-\alpha}{2}} c_q^{\frac{1+\alpha}{2}} \right) + c_p^{\frac{1-\alpha}{2}} c_q^{\frac{1+\alpha}{2}} D_{f^{(\alpha)}}(P, Q_\xi).$$

Therefore for all $\alpha \in \mathbf{R}$, $\bar{D}_{f^{(\alpha)}}(c_pP, c_qQ_\xi)$ is monotone increasing with respect to $D_{f^{(\alpha)}}(P, Q_\xi)$ and we have equivalence between the minimizations of $\bar{D}_{f^{(\alpha)}}(c_pP, c_qQ_\xi)$ and $D_{f^{(\alpha)}}(P, Q_\xi)$. □

Theorem 5. *For arbitrarily fixed $\alpha \in \mathbf{R}$, suppose that $D_{f^{(\alpha)}}(P, Q_\xi)$ can be defined for any $\xi \in \Xi$. Then for arbitrarily fixed $c_p > 0$, $\xi \in \Xi$ that minimizes $D_{f^{(\alpha)}}(P, Q_\xi)$ coincides with $\xi \in \Xi$ that minimizes $\min_{c_q \in \mathbf{R}} \bar{D}_{f^{(\alpha)}}(c_pP, c_qQ_\xi)$ if both of them uniquely exist.*

Proof. In order to obtain $\tilde{c}_q(\xi)$ minimizing $\bar{D}_{f^{(\alpha)}}(c_pP, c_qQ_\xi)$ for fixed $\xi \in \Xi$, we differentiate $\bar{D}_{f^{(\alpha)}}(c_pP, c_qQ_\xi)$ with respect to c_q. From Eq.(17) we have

$$\frac{\partial}{\partial c_q} \bar{D}_{f^{(1)}}(c_pP, c_qQ_\xi) = \log \frac{c_q}{c_p} + D_{f^{(1)}}(P, Q_\xi)$$

and for $\alpha \neq 1$

$$\frac{\partial}{\partial c_q} \bar{D}_{f^{(\alpha)}}(c_pP, c_qQ_\xi) = \frac{2}{1-\alpha} \left\{ 1 - \left(\frac{c_p}{c_q} \right)^{\frac{1-\alpha}{2}} \int p(x) \left(\frac{q_\xi(x)}{p(x)} \right)^{\frac{1+\alpha}{2}} \mathrm{d}\nu(x) \right\}.$$

For $\alpha = 1$, solving $\partial \bar{D}_{f^{(1)}}(c_pP, c_qQ_\xi)/\partial c_q = 0$ gives $\tilde{c}_q(\xi) = c_p \exp(-D_{f^{(1)}}(P, Q_\xi))$ and we have $\tilde{c}_q(\xi) > 0$. For $\alpha \neq 1$, we have

$$\tilde{c}_q(\xi) = c_p \left\{ \int p(x) \left(\frac{q_\xi(x)}{p(x)} \right)^{\frac{1+\alpha}{2}} \mathrm{d}\nu(x) \right\}^{\frac{2}{1-\alpha}}.$$

Since by Jensen's inequality

$$\int p(x) \left(\frac{q_\xi(x)}{p(x)} \right)^{\frac{1+\alpha}{2}} \mathrm{d}\nu(x) \geq \left(\int q_\xi(x) \mathrm{d}\nu(x) \right)^{\frac{1+\alpha}{2}} = 1,$$

it is the case that $\tilde{c}_q(\xi) > 0$ when $\alpha \neq 1$ Therefore the desired result follows from Theorem 4. □

Applying Theorem 5 into the classes MD and MED, the following corollaries are immediately obtained.

Corollary 6. *The MD estimator $\tilde{\theta}_{MD}$ with $f^{(\alpha)}(u)$ coincides with the MED estimator $\tilde{\theta}_{MED}$ with $\bar{f}^{(\alpha)}(u)$ if both of them uniquely exist.*

Corollary 7. *Suppose that for a fixed $\theta \in \Theta$ there exist a unique optimal solution $\{\tilde{p}_i(\theta)\}_{i=1}^n$ of the convex optimization problem (4) with $f^{(\alpha)}(u)$ and a unique optimal solution $\{\tilde{w}_i(\theta)\}_{i=1}^n$ of the convex optimization problem (7) with $\bar{f}^{(\alpha)}(u)$. Then*

$$\tilde{w}_i(\theta) = \frac{\tilde{p}_i(\theta)}{\sum_{i'=1}^n \tilde{p}_{i'}(\theta)}\ , \quad i = 1, \ldots, n,$$

holds.

Newey and Smith (2004) show that the MD estimator with $f^{(\alpha)}(u)$ and the GEL estimator with $\rho^{(\alpha)}(v)$ coincide by comparing the first order conditions of the two estimators. This result also follows from Theorem 3 and Corollary 6.

5 Implied probabilities

In this section, we consider relations between implied probabilities associated with GEL and discrete measures derived from MED.

Firstly, we show that the solution of the problem (7) for MED can be obtained by solving the problem (8) for the corresponding GEL. Note that in the following theorem it is not assumed that $\bar{f}(u)$ is differentiable on its domain $\mathcal{U}$.

Theorem 8. *Suppose that $\rho(v)$ and $\bar{f}(u)$ satisfy Assumption 1 and 2 respectively, $\rho(v) = -\bar{f}^*(v)$ and that $\rho(v)$ is differentiable on $\mathcal{V}$. If for fixed $\theta \in \Theta$ there exists an optimal solution $\tilde{\lambda}(\theta)$ for the convex optimization problem (8), then*

$$\tilde{w}_i(\theta) = -\frac{1}{n}\rho'(\tilde{\lambda}^{\mathrm{T}}(\theta)g(x_i, \theta))\ , \quad i = 1, \ldots, n \tag{19}$$

is the optimal solution of the convex optimization problem (7).

Proof. We have $\sum_{i=1}^n \rho'(\tilde{\lambda}^{\mathrm{T}}(\theta)g(x_i,\theta))g(x_i,\theta) = 0$ from the first order condition of the problem (8) and we also have $n\tilde{w}_i(\theta) \in \mathcal{U}, i = 1, \ldots, n$ since $\bar{f}(u) = (\bar{f}^*)^*(u)$ and $\rho'(v) = -(\bar{f}^*)'(v)$. Therefore $\{\tilde{w}_i(\theta)\}_{i=1}^n$ in Eq.(19) is a feasible solution of the problem (7). Let $\{w_i(\theta)\}_{i=1}^n$ be any feasible solution of the problem (7) and $\partial\bar{f}(u)$ be a subdifferential of $\bar{f}(u)$ at a point u, i.e. $\partial\bar{f}(u) := \{v | \bar{f}(u') \geq \bar{f}(u) + v(u' - u), {}^\forall u' \in \mathcal{U}\}$. ${}^\forall v_i \in \partial\bar{f}(n\tilde{w}_i(\theta))$ we have

$$\sum_{i=1}^n \bar{f}(nw_i(\theta)) - \sum_{i=1}^n \bar{f}(n\tilde{w}_i(\theta)) \geq \sum_{i=1}^n v_i(nw_i(\theta) - n\tilde{w}_i(\theta)). \tag{20}$$

Now from Eq.(19) and $\rho(v) = -\bar{f}^*(v)$, we have $n\tilde{w}_i(\theta) = (\bar{f}^*)'(\tilde{\lambda}^{\mathrm{T}}(\theta)g(x_i,\theta))$ and $n\tilde{w}_i(\theta) \in \partial\bar{f}^*(\tilde{\lambda}^{\mathrm{T}}(\theta)g(x_i,\theta))$. Hence it follows from Theorem 23.5 in Rockafellar (1970) p.218 that $\tilde{\lambda}^{\mathrm{T}}(\theta)g(x_i,\theta) \in \partial\bar{f}(n\tilde{w}_i(\theta))$. Therefore setting $v_i = \tilde{\lambda}^{\mathrm{T}}(\theta)g(x_i,\theta)$ in Eq.(20) gives

$$\sum_{i=1}^{n} \bar{f}(nw_i(\theta)) - \sum_{i=1}^{n} \bar{f}(n\tilde{w}_i(\theta)) \geq \sum_{i=1}^{n} \tilde{\lambda}^{\mathrm{T}}(\theta)g(x_i,\theta)(nw_i(\theta) - n\tilde{w}_i(\theta)) = 0,$$

where the equality follows from the fact that $\{\tilde{w}_i(\theta)\}_{i=1}^n$ and $\{w_i(\theta)\}_{i=1}^n$ are feasible solutions. Thus we obtain the desired result. □

From Theorems 3 and 8, the implied probability (5) associated with GEL can be written as $\tilde{\pi}_i = \tilde{w}_i(\tilde{\theta}_{MED})/\sum_{i'=1}^n \tilde{w}_{i'}(\tilde{\theta}_{MED})$, $i = 1,\ldots,n$, where $\{\tilde{w}_i(\theta)\}_{i=1}^n$ and $\tilde{\theta}_{MED}$ are the solutions of the problem (7) and the corresponding MED estimator, respectively. From Corollaries 6 and 7, when we adopt $f_j^{(\alpha)}(u)$ in Eq.(15) in MD and $\rho_j^{(\alpha)}(v) = -(\bar{f}_j^{(\alpha)})^*(v)$ in GEL, we also have $\tilde{\pi}_i = \tilde{p}_i(\tilde{\theta}_{MD})$, $i = 1,\ldots,n$, where $\{\tilde{p}_i(\theta)\}_{i=1}^n$ and $\tilde{\theta}_{MD}$ are the solution of the problem (7) and the corresponding MD estimator, respectively.

From Theorem 8 we also find that the possibility of negative values for implied probabilities results from the fact that a discrete measure in MED may take negative values. Therefore we can remove this possibility by restricting the domain of $\bar{f}(u)$ to nonnegative real numbers. For example, since $\bar{f}_2^{(\alpha)}(u)$ in Eq.(17) is defined for all real numbers when $\alpha = 4k-1, k = 1,2,\ldots$, the implied probability with $\rho_2^{(\alpha)}(v)$ may be negative; however, since $\bar{f}_1^{(\alpha)}(u)$ is defined only for nonnegative real numbers, the implied probability with $\rho_1^{(\alpha)}(v)$ is always nonnegative. In addition, since the problem (7) in MED is a convex optimization problem, if the implied probability with $\rho_2^{(\alpha)}(v)$ is nonnegative, i.e. the optimal solution of the problem (7) in MED is positive, then the difference in domains of $\bar{f}_1^{(\alpha)}(u)$ and $\bar{f}_2^{(\alpha)}(u)$ does not affect the optimal solutions of the problem (7) and the implied probability with $\rho_1^{(\alpha)}(v)$ coincides with the implied probability with $\rho_2^{(\alpha)}(v)$.

Example 5 (Continuation of Example 3 and 4)**.** The implied probability associated with $\rho_2^{(3)}(v)$ may take negative values, whereas the implied probability associated with $\rho_1^{(3)}(v)$ is always nonnegative. Note that the domain of $\rho_1^{(3)}(v)$ is the entire set of real numbers and $\rho_1^{(3)}(v)$ is differentiable on this domain, although the domain of $\bar{f}_1^{(3)}(u)$ is the set of nonnegative real numbers and $\bar{f}_1^{(3)}(u)$ is not differentiable at $u = 0$. The derivative of $\rho_1^{(3)}(v)$ can be written as

$$(\rho_1^{(3)})'(v) = \begin{cases} -v-1 & v \geq -1 \\ 0 & v < -1. \end{cases}$$

We find that since $-(\rho_1^{(3)})'(v)$ is nonnegative $^\forall v \in \mathcal{V}$, the associated implied probability is always nonnegative from the definition of implied probabilities (5).

Acknowledgements

We are very grateful to Dr. Tomonari Sei for his valuable discussions and useful comments.

References

[1] Amari, S. (1982). Differential geometry of curved exponential families – curvature and information loss. *Annals of Statistics*, **10**, 357-385.

[2] Amari, S. and Nagaoka, H. (2000). *Methods of Information Geometry*. AMS monograph, Oxford University Press.

[3] Back, K. and Brown, D. P. (1993). Implied probabilities in GMM estimators. *Econometrica*, **61**, 971-975.

[4] Boyd, S. and Vandenberghe, L. (2004). *Convex Optimization*. Cambridge University Press.

[5] Brown, B. W. and Newey, W. K. (2002). Generalized method of moments, efficient bootstrapping and improved inference. *Journal of Business and Economic Statistics*, **20**, 507-517.

[6] Csiszár, I. (1967). On topical properties of f-divergence. *Studia Mathematicarum Hungarica*, **2**, 329-339.

[7] Corcoran, S. A. (1998). Bartlett Adjustment of Empirical Discrepancy Statistics. *Biometrika*, **85**, 967-972.

[8] Cressie, N. and Read, T. (1984). Multinominal goodness-of-fit tests. *Journal of the Royal Statistical Society, Series B*, **46**, 440-464.

[9] Eguchi, S. (1983). Second order efficiency of minimum contrast estimators in a curved exponential family. *Annals of Statistics*, **10**, 793-803.

[10] Hansen, L. P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica*, **75**, 967-972.

[11] Hansen, L. P., Heaton, J. and Yaron, A. (1996). Finite-sample properties of some alternative GMM estimators. *Journal of Business and Economic Statistics*, **14**, 262-280.

[12] Imbens, G. W. (1997). One-step estimators for over-identified generalized method of moments models. *Review of Economic Studies*, **64**, 359-383.

[13] Imbens, G. W., Spady, R. H. and Johnson, P. (1998). Information theoretic approaches to inference in moment condition models. *Econometrica*, **66**, 333-357.

[14] Kitamura, Y. and Stutzer, M. (1997). An information-theoretic alternative to generalized method of moments estimation. *Econometrica*, **65**, 861-874.

[15] Newey, W. K. and Smith, R. J. (2004). Higher order properties of GMM and generalized empirical likelihood. *Econometrica*, **72**, 219-255.

[16] Owen, A. B. (1988). Empirical likelihood ratio confidence intervals for a single functional. *Biometrika*, **75**, 237-249.

[17] Qin, J. and Lawless, J. (1994). Empirical likelihood and general estimating equations. *Annals of Statistics*, **22**, 300-325.

[18] Rockafellar, T. R. (1970). *Convex Analysis*. Princeton University Press.

[19] Smith, R. J. (1997). Alternative semiparametric likelihood approaches to generalized method of moments estimation. *Economics Journal*, **22**, 300-325.

[20] Zhang, J. (2004). Divergence function, duality and convex analysis. *Neural Computation*, **16**, 159-195.