計量經濟學報, 제18권 제2호(2007. 6. 30.), 85-113

Minimum Empirical ϕ-Divergence Estimation and Inference

Young Su Cho*

Utilizing the concept of unbiased estimating functions combined with the concept of ϕ-divergence of Csiszár (1963), we introduce the method of minimum empirical ϕ-divergence estimation and inference, which can be thought of as a generalization of empirical likelihood method of Owen (1988) and Qin and Lawless (1994). Efficiency results for estimators are obtained. Given the parallels with the conventional likelihood, classical-type tests based on the empirical ϕ-divergence for a simple parametric hypothesis and the moment conditions are constructed.

Keywords : Empirical Likelihood, Empirical ϕ-Divergence, Estimating Equations

JEL Classifications : C12, C13, C14

* Korea Rural Economic Institute, Seoul, Korea, 130-710, Tel : +82-2-3299-4337, Fax : +82-2-968-7340, E-mail : yscho@krei.re.kr

투고일: 2006. 06. 16 심사일: 2006. 06. 16 최종심사완료일: 2007. 06. 15

I. Introduction

Empirical likelihood, introduced and studied by Owen (1988, 1990), is a nonparametric method of statistical inference based on a likelihood function driven by the observations, i.e., the empirical likelihood inference is likelihood based, but does not require the assumption of a parametric family for the observations. The idea of empirical likelihood consists in maximization of the profile (or concentrated) likelihood function supported by the observations, under some model constraints. Owen and many followers have shown that empirical likelihood possesses properties similar to those of the conventional parametric likelihood methods. In particular, the empirical likelihood ratio statistics have asymptotic chi-square distributions in certain situations as in Wilks's theorem (see, e.g., Owen (1988, 1990, 1991) and Qin and Lawless (1994)), and the resulting confidence regions are Bartlett correctable. That is, by an explicit correction of confidence regions, the coverage error can be reduced to the order of $O(n^{-2})$ under some regularity assumptions (see, e.g., DiCiccio et al (1991)). Recently, empirical likelihood has been studied extensively because of its generality and effectiveness, and its applications can be found in a wide rage of areas. Examples are smooth function model (Hall and La Scala (1990) and DiCiccio et al (1988, 1991)), regression models (Owen (1991) and Chen (1993, 1994a, 1994b)), quantiles (Chen and Hall (1993), generalized linear models (Kolaczyk (1994)), and general estimating equations (Qin and Lawless (1994)).

As Owen (1991) noted, the empirical likelihood ratio statistic can be viewed as a measure of the distance of a multinomial distribution with support on the observations from the empirical distribution, its nonparametric maximum likelihood estimator. As a general distance statistic Baggely (1998) introduced the Cressie-Read power-divergence statistic (see Read and Cressie (1998)) and showed that all members of the empirical Cressie-Read family have an asymptotic chi-square calibration, but empirical likelihood is the only Bartlett correctable member of the family. The latter fact was also shown by Cocoran (1998) in the empirical discrepancy context.

In this paper, we propose an alternative nonparametric likelihood method to empirical likelihood, which utilizes the concept of more general ϕ-divergence introduced by Csiszár (1963). As Qin and Lawless (1994), we consider unbiased estimating functions to provide a flexible way to describe parameters and the corresponding statistics,

and our approach is based on the minimization of the empirical ϕ-divergence, defined as the ϕ-divergence of a multinomial distributions with support on the observations and its nonparametric maximum likelihood estimator, under the moments conditions (i.e., under the unbiasedness condition of the estimating functions). It thus can be thought of as a generalization of empirical likelihood of Owen (1988) and Qin and Lawless (1994). Our methodology also parallels the approaches based on the Cressie-Read power divergence statistic (Baggerly (1998)), empirical discrepancy (Cocoran (1998)), and generalized empirical likelihood (Smith (1997, 2001)), which provide general unifying frameworks for studying different nonparametric likelihood (or divergence) statistics.

The layout of this paper is the following. In part 2, the profile (or concentrated) empirical ϕ-divergence function is introduced for unbiased estimating functions with independent identically distributed observations, and some computational issues are discussed. In particular, there is shown that a dual problem allows us to reformulate the constrained minimization problem as one of an unconstrained maximization of a concave function, simplifying the computation of the profile empirical ϕ-divergence function. We then in part 3 describe the estimating procedure based on the empirical ϕ-divergence, and the asymptotic properties of the estimators are investigated. Part 4 gives some alternative statistics in the empirical ϕ-divergence context for testing parametric hypothesis and the validity of the moment conditions, of which the structure is analogous to that of conventional likelihood based test statistics. Part 5 concludes. The proofs of the main theorems are postponed to the appendix.

Ⅱ. Empirical ϕ-Divergence and Estimating Functions

In this part, we first recall some notions on ϕ-divergence (Csiszár (1963)). We show then how to link the (unbiased) estimating functions or equations and the concept of ϕ-divergence for statistical inference. Some computational issues are also discussed.

1. ϕ-divergence

Let P and Q be two probability distributions on a measurable space (Ω, F), and let μ be an arbitrary dominating measure of P and Q. Let f and g be the Radon-Nikodým-densities of P and Q with respect to μ, respectively.

As a measure of difference of the distributions P and Q, Csiszár (1963) introduced the ϕ-divergence defined by

$$D_\phi(P,Q) = \int_\Omega g\phi\left(\frac{f}{g}\right)d\mu \tag{1}$$

for any continuous convex function $\phi : [0,\infty) \to R^+ \cup \{\infty\}$ with $\phi(1) = 0$ and $\phi^{(2)}(1) > 0$, where $\phi^{(s)}(1)$ denotes the s-th derivative of $\phi(u)$ evaluated at $u = 1$. If P and Q are discrete on the same support, say $P = p_1, \cdots, p_m$ and $Q = (q_1, \cdots, q_m)$, $m \geq 2$, the count measure can be taken for μ, and the Φ-divergence (1) reduces to

$$D_\phi(P,Q) = \sum_{i=1}^{m} q_i\phi\left(\frac{p_i}{q_i}\right). \tag{2}$$

The family of ϕ-divergences includes several well-known measures of deviation between two distributions. Some examples of such measures are :

(1) Kullback-Leibler divergence (Kullback and Leibler (1951)), Pearson's χ^2-divergence (Pearson (1900)), and Neyman's modified χ^2-divergence (Neyman (1949)), which are obtained for $\phi(u) = u\log u$, $\phi(u) = (u-1)^2$, and $\phi(u) = (u-1)^2/u$, respectively.

(2) The family ϕ_α introduced by Liese and Vajda (1987) (also known as the power-divergence introduced by Read and Cressie (1988))

$$\phi_\alpha(u) = \begin{cases} u - 1 - \ln u, & \alpha = 0, \\ \dfrac{\alpha u + 1 - \alpha - u^\alpha}{\alpha(1-u)}, & \alpha \in R \setminus \{0, 1\}, \\ 1 - u + u\ln u, & \alpha = 1. \end{cases}$$

This family contains also many measures as special cases corresponding to

particular values of the parameter α.

Now, we impose the following regularity conditions on the function ϕ that will be maintained throughout the paper.

Assumptions 1 (a) The function $\phi:[0,\infty)\to R^+\cup\{\infty\}$ is convex on $(0,\infty)$. It is finite and twice continuously differentiable on $(0,\infty)$, and the second derivative $\phi^{(2)}(u)$ is positive for all $u\in(0,\infty)$; (b) It holds that $\phi(1)=\phi^{(1)}(1)=0$ and $\phi^{(2)}(1)=1$; (c) The domain U of the inverse function $\phi^{(1)-1}$ of $\phi^{(1)}:(0,\infty)\to R$ is an open interval containing zero.

The Assumption 1 (b) is simply normalization properties. Notice that every ϕ satisfying Assumption 1 (a) is strictly convex on $(0,\infty)$. Hence it follows under Assumptions 1 (a)-(b) that $D_\alpha(P,Q)\geq\phi(1)=0$ with equality only for $P=Q$. Since $\phi^{(1)}$ is continuous and monotonically increasing on $(0,\infty)$, there exists a continuous inverse function $\phi^{(1)-1}:U\to R^+$. In case of the ϕ_α family, for example, $U=(-2,\infty)$, $(-\infty,\infty)$, and $(-\infty,1)$ for $\alpha=2$, 1, and 0, respectively.

2. Empirical ϕ-divergence for General Estimating Equations

Let $X_1, X_2,\cdots$ be R^d-valued independent random vectors, for $d\geq 1$, with a common unknown distribution function F. Generically, X_i will be denoted by X. Given $X_1,\cdots,X_n$, let $P=(p_1,\cdots,p_n)$ be an arbitrary multinomial-type distribution that assigns the probability mass p_i to the observed value of X_i, and let $\hat{P}=(1/n,\cdots,1/n)$ be the empirical distribution, i.e., the nonparametric maximum likelihood estimator (NMLE) of P based on $X_1,\cdots,X_n$.

As a measure of deviation between P and the NMLE $\hat{P}$ that will underlie the concept of our approach to estimation and inference, we introduce the empirical ϕ-divergence defined as the ϕ-divergence of P and $\hat{P}$, i.e.,

$$D_\phi(P,\hat{P}) = \frac{1}{n}\sum_{i=1}^{n}\phi(p_i n). \tag{3}$$

Let $\theta \in \Theta \subset R^p$ be a parameter vector of inferential interest, which is associated with F. Assume[1] that the information about θ and F is available in the form of $r \geq p$ functionally independent[2] estimating functions $g_j : R^d \times \Theta \to R$, $j = 1, \cdots, r$, such that $E\{g_j(X,\theta)\} = 0$. Writing in vector form, we have

$$g(X,\theta) = (g_1(X,\theta), \cdots, g_r(X,\theta)),$$

where

$$E\{g(X,\theta)\} = 0. \tag{4}$$

By $r \geq p$, we are allowed to deal with the combination of pieces of information, and the information about θ and F is expressed by the moment conditions (i.e., by the unbiasedness of the estimating functions).

Examples (Qin and Lawless, 1994)

1. Let $(X_1, Y_1), \cdots, (X_n, Y_n)$ be bivariate i.i.d. observations with $E(X_i) = E(Y_i) = \theta$. In this case we can take $g((X_i, Y_i), \theta) = (X_i - \theta, Y_i - \theta)$.
2. Let $X_1, \cdots, X_n$ be i.i.d. univariate observations with mean θ, and suppose that it is known that $E(X^2) = m(\theta)$, where $m(\cdot)$ is a known function. The information about F can be expressed in the form given above by taking $g((X, \theta) = (X - \theta, X^2 - m(\theta))'$.

We will show in this paper how to use such information for estimating θ and testing hypotheses about θ, in conjunction with the empirical ϕ-divergence. The idea behind the empirical ϕ-divergence approach is as follows. Without the information about θ and F, the empirical distribution function is an optimal estimator for the distribution function F. But the empirical distribution function, in

---

1) The following notions and arguments are due to Qin and Lawless (1944).

2) The functions $g_j(X,\theta)$, $j = 1,\ldots,r$, are functionally independent if the Jacobian matrix $\partial(g_1,\ldots,g_r)/\partial(\theta_1,\ldots,\theta_p)$ has its rank p at some point of Θ.

general, does not contain such information of the moment conditions as that given by (4). We therefore choose the distribution function that is "closest" to the empirical distribution function, measured by the ϕ-divergence, among those distribution functions that satisfy the moment conditions, and use the minimized value of the divergence for further inference. Formally, we define the profile empirical ϕ-divergence function for θ as follows.

$$E_\phi D(\theta) = \min_P \left\{ \frac{1}{n} \sum_{i=1}^{n} \phi(p_i n) \mid p_i \geq 0, \sum_{i=1}^{n} p_i = 1, \sum_{i=1}^{n} p_i g(X_i, \theta) = 0 \right\}. \tag{5}$$

Remark : For $\phi(u) = u - 1 - \log u$, the profile empirical ϕ-divergence function, multiplied by n, corresponds to the profile empirical log-likelihood function introduced and studied by Owen (1988, 1990) and Qin and Lawless (1994), i.e.,

$$nE_\phi D(\theta) = \max_P \left\{ \sum_{i=1}^{n} \log p_i n \mid p_i \geq 0, \sum_{i=1}^{n} p_i = 1, \sum_{i=1}^{n} p_i g(X_i, \theta) = 0 \right\}, \tag{6}$$

where the profile empirical ϕ-divergence function assigns the most "favorable" probability masses possible to a given θ from the family of multinomial distributions with support on $g(X_1, \theta)$, $\cdots$, $g(X_n, \theta)$ that satisfies the moment conditions (see Mittelhammer et al (2000, Ch. 12)).

3. Computation and duality

Assuming the regular case that the minimum in (5) is attained at $p_i > 0$, $i = 1, \cdots, n$, the optimization problem can be solved by considering a Lagrangian as

$$L = \frac{1}{n} \sum_{i=1}^{n} \phi(p_i n) - \lambda_1 \left(\sum_{i=1}^{n} p_i - 1 \right) - \lambda_2' \sum_{i=1}^{n} p_i g(X_i, \theta), \tag{7}$$

where the multipliers λ_1 and λ_2 are scalar and a vector of dimension $r \times 1$, respectively.[3)]

---

3) Clearly, by the assumption made, the constraint $p_i \geq 0$ does not bind. Note that, by the functional independence of estimating functions, the rank condition is satisfied. Hence there exists λ satisfying the first-order conditions by the Lagrange Theorem.

From the first-order conditions, we have $p_i = \frac{1}{n}\phi^{(1)-1}(\lambda' G_i(\theta))$, where $\lambda = (\lambda_1, \lambda_2')'$ and $G_i = (1, g(X_i,\theta)')'$. Thus the profile empirical ϕ-divergence function is rewritten as

$$E_\phi D(\theta) = \frac{1}{n}\sum_{i=1}^{n} \phi \circ \phi^{(1)-1}(\lambda' G_i(\theta)), \tag{8}$$

where the optimal λ satisfies the $r+1$ equations

$$\frac{1}{n}\sum_{i=1}^{n} \phi^{(1)-1}(\lambda' G_i(\theta)) G_i(\theta) = \begin{bmatrix} 1 \\ 0 \end{bmatrix}. \tag{9}$$

The above optimization problem has dimension increasing with n. It, however, is possible to reformulate the problem as a much smaller dimensional one of maximization of a concave function (see Owen (1990) for some computational issues in the empirical likelihood context). For this, substituting the expression for p_i given above into (7), we define

$$D_{E\phi}(\lambda,\theta) = \frac{1}{n}\sum_{i=1}^{n} \phi \circ \phi^{(1)-1}(\lambda' G_i(\theta)) - \lambda' \left\{ \frac{1}{n}\sum_{i=1}^{n} \phi^{(1)-1}(\lambda' G_i(\theta)) G_i(\theta) - \begin{bmatrix} 1 \\ 0 \end{bmatrix} \right\}. \tag{10}$$

We now seek a maximum of $D_{E\phi}(\lambda,\theta)$ on $\Lambda(\theta) = \{\lambda \mid \lambda' G_i(\theta) \in U, i = 1, \cdots, n\}$, a random convex set depending on $X_1, \cdots, X_n$. By Assumption 1 (c) it contains a full neighborhood in R^{r+1} of $\lambda = 0$. We can show that setting the gradient of $D_{E\phi}(\lambda,\theta)$ with respect to λ to zero amounts to solving the equation (9). Moreover, the Hessian of $D_{E\phi}(\lambda,\theta)$ with respect to λ is

$$-\frac{1}{n}\sum_{i=1}^{n} \frac{G_i(\theta) G_i(\theta)'}{\phi^{(2)} \circ \phi^{(1)-1}(\lambda' G_i(\theta))}. \tag{11}$$

This is a negative definite function of λ if $n^{-1}\sum_{i=1}^{n} G_i(\theta) G_i(\theta)'$ is positive definite, as $\phi^{(2)} \circ \phi^{(1)-1}(\lambda' G_i(\theta)) > 0$ by the strict convexity of ϕ on $(0,\infty)$. Thus the solution to the first-order conditions defines a unique maximum of $D_{E\phi}(\lambda,\theta)$ on $\Lambda(\theta)$ and is a continuously differentiable function of θ by the implicit function theorem.

Now, we have the following dual problem: minimizing $n^{-1}\sum_{i=1}^{n}\phi(p_i n)$ over n variables p_i subject to $r+1$ constraints is equivalent to seeking an (interior) maximum of $D_{E\phi}(\lambda,\theta)$ on $\Lambda(\theta)=\{\lambda \mid \lambda' G_i(\theta)\in U, i=1,\ldots,n\}$. The latter is a $r+1$ dimensional maximization problem without constraints except some restrictions imposed on λ by $\lambda\in\Lambda(\theta)$. To summarize this dual problem, define $\hat{\lambda}_n(\theta)=(\hat{\lambda}_{1,n}(\theta),\hat{\lambda}_{2,n}(\theta)')'$ by

$$D_{E\phi}(\hat{\lambda}_n(\theta),\theta)=\max_{\lambda\in\Lambda(\theta)} D_{E\phi}(\lambda,\theta). \tag{12}$$

If $\hat{\lambda}_n(\theta)$ is an interior point of $\Lambda(\theta)$, then $\hat{\lambda}_n(\theta)$ satisfies the first-order conditions

$$\begin{bmatrix}1\\0\end{bmatrix}-\frac{1}{n}\sum_{i=1}^{n}\phi^{(1)-1}(\hat{\lambda}_n(\theta)'G_i(\theta))G_i(\theta)=0, \tag{13}$$

and, consequently, we have

$$E_\phi D(\theta)=D_{E\phi}(\hat{\lambda}_n(\theta),\theta)=\max_{\lambda\in\Lambda(\theta)} D_{E\phi}(\lambda,\theta). \tag{14}$$

Ⅲ. Estimation and hypothesis testing

1. Minimum empirical ϕ-divergence estimation

Now, suppose that we are interested in estimating the true parameter vector of θ_0. By the arguments in the previous section and the duality, we may take an estimator $\hat{\theta}_n$ of θ_0 as a solution to the problem $\min_{\theta\in\Theta} D_{E\phi}(\hat{\lambda}_n(\theta),\theta)=\min_{\theta\in\Theta}\max_{\lambda\in\Lambda(\theta)} D_{E\phi}(\lambda,\theta)$, which is called the minimum empirical ϕ-divergence estimator (MEϕDE) of θ_0, i.e.,

$$\hat{\theta}_n=\arg\min_{\theta\in\Theta} D_{E\phi}(\hat{\lambda}_n(\theta),\theta)=\arg\min_{\theta\in\Theta}\max_{\lambda\in\Lambda(\theta)} D_{E\phi}(\lambda,\theta). \tag{15}$$

In general, $\hat{\theta}_n$ need not be unique. Thus, $\hat{\theta}_n$ should be interpreted as a measurable selection from the pointwise solution set. The solutions $\hat{\lambda}_n(\theta)$ and $\hat{\theta}_n$ then define

a saddle point of $D_{E\phi}(\lambda,\theta)$ and satisfy the first-order conditions (13) and

$$\frac{1}{n}\sum_{i=1}^{n}\phi^{(1)-1}(\hat{\lambda}_n(\hat{\theta}_n)'G_i(\hat{\theta}_n))\left(\frac{\partial G_i(\hat{\theta}_n)}{\partial\theta'}\right)'\hat{\lambda}_n(\hat{\theta}_n)=0, \tag{16}$$

provided $\hat{\lambda}_n(\theta)$ and $\hat{\theta}_n$ lie interior of $\Lambda(\hat{\theta}_n)$ and Θ, respectively.

To study asymptotic properties of $\hat{\lambda}_n(\theta)$ and $\hat{\theta}_n$, we introduce some conditions that are in part quite standard in the literature. Throughout this paper, N_0 denotes an open neighborhood of θ_0, and $\|U\|$ denotes the Euclidean norm $\sqrt{trace(U'U)}$ of a column vector or matrix U. Also, C denotes a generic positive constant that may be different in different uses, and positive semi-definite will be abbreviated as p.s.d.

Assumptions 2 (a) The parameter space $\Theta\subset R^p$ is compact; (b) $\theta_0\in\Theta$ is the unique solution to $E\{g(X,\theta)\}=0$; (c) $g(\cdot,\theta)$ is Borel measurable for each $\theta\in\Theta$ and $g(x,\cdot)$ is continuous on Θ for each $x\in R^d$; (d) $E\left\{\sup_{\theta\in\Theta}\|g(X,\theta)\|^{\alpha}\right\}<\infty$ for some $\alpha>2$; (e) $S=\mathrm{var}(g(X,\theta_0))$ is positive definite; (f) θ_0 is an interior point of Θ; (g) $g(x,\cdot)$ is twice continuously differentiable on an open neighborhood N_0 for each $x\in R^d$, $E\left\{\sup_{\theta\in N_0}\|\partial g(X,\theta)/\partial\theta'\|\right\}<\infty$ and $E\left\{\sup_{\theta\in N_0}\|\partial^2 g_j(X,\theta)/\partial\theta\partial\theta'\|\right\}<\infty$, where $g_j(x,\theta)$ denotes the j-th element of $g(x,\theta)$, $j=1,\ldots,r$; (h) $D=E\{\partial g(X,\theta_0)/\partial\theta'\}$ is of full column rank p.

Assumptions 2 (a)-(e) are standard regularity conditions employed for the consistency of the estimators in the empirical likelihood and the generalized empirical likelihood context (see, e.g., Qin and Lawless (1994) and Newey and Smith (2001)). Note that the existence of higher than second moments of g is also required in the empirical ϕ-divergence approach.

The additional Assumptions 2 (f)-(h) will later be shown to be sufficient for asymptotic normality results for $\hat{\lambda}_n(\hat{\theta}_n)$ and $\hat{\theta}_n$.

Now, Assumptions 2 (a)-(e) allow us to state the following consistency results for the MEϕDE $\hat{\theta}_n$. In addition, we obtain a convergence rate for $\hat{\lambda}_n(\hat{\theta}_n)$ and $\bar{g}(\hat{\theta}_n)$, where $\bar{g}(\hat{\theta}_n) = n^{-1}\sum_{i=1}^{n} g(X_i, \hat{\theta}_n)$. Note that the second result of the theorem implies the weak consistency of $\hat{\lambda}_n(\hat{\theta}_n)$.

Theorem 1 If Assumptions 1 and 2 (a)-(e) are satisfied, then

$$\hat{\theta}_n \xrightarrow{p} \theta_0, \quad \| \hat{\lambda}_n(\hat{\theta}_n) \| = O_p(n^{-1/2}) \text{ and } \| \bar{g}(\hat{\theta}_n) \| = O_p(n^{-1/2}).$$

Under the additional Assumptions 2 (f)-(h), we have the following results on the asymptotic distributions of $\hat{\lambda}_n(\hat{\theta}_n)$ and $\hat{\theta}_n$

Theorem 2 If Assumptions 1 and 2 are satisfied, then

$$\sqrt{n}\hat{\lambda}_n(\hat{\theta}_n) \xrightarrow{d} N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 & 0 \\ 0 & \Sigma_\lambda \end{bmatrix}\right)^{4)} \text{ and } \sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N(0, \Sigma_\theta),$$

where $\Sigma_\lambda = S^{-1}(I_r - D\Sigma_\theta D'S^{-1})$ and $\Sigma_\theta = (D'S^{-1}D)^{-1}$, and $\hat{\lambda}_n(\hat{\theta}_n)$ and $\hat{\theta}_n$ are asymptotically uncorrelated.

In view of $p_i = \frac{1}{n}\phi^{(1)-1}(\lambda' G_i(\theta))$, $\hat{\lambda}_n(\hat{\theta}_n)$ and the MEϕDE $\hat{\theta}_n$ yield an estimator of the probability distribution

$$\hat{p}_i = \frac{1}{n}\phi^{(1)-1}(\hat{\lambda}_n(\hat{\theta}_n)' G_i(\hat{\theta}_n)), \quad i = 1, \ldots, n. \tag{17}$$

Thus, the $\hat{p}_i$'s are the empirical measure counterparts to the expectation operator in $E\{g(X, \theta_0)\} = 0$ of (4), which ensure that the moment conditions are satisfied in the observations (cf. (13)). Then the information contained in the moment conditions may be exploited using the $\hat{p}_i$'s to provide a more efficient estimator of the distribution

---

4) Note that, surprisingly, $\sqrt{n}\hat{\lambda}_{1,n}(\hat{\theta}_n) \xrightarrow{p} 0$. See the proof for details.

function F than the empirical distribution function (EDF) $F_n(x)=n^{-1}\sum_{i=1}^{n}I(X_i\le x)$, where $I(\cdot)$ denotes the indicator function. Define the empirical ϕ-divergence cumulative distribution function estimator (EϕDCDFE) of the distribution function F, based on the $\hat{p}_i$'s, as

$$\hat{F}_n(x)=\frac{1}{n}\sum_{i=1}^{n}\phi^{(1)-1}(\hat{\lambda}_n(\hat{\theta}_n)'G_i(\hat{\theta}_n))I(X_i\le x). \tag{18}$$

Then we have the following result for the asymptotic distribution of the EϕDCDFE $\hat{F}_n$.

Theorem 3 If Assumptions 1 and 2 are satisfied, then

$$\sqrt{n}(\hat{F}_n(x)-F(x))\xrightarrow{d}N(0,\Sigma_F(x)),$$

where $\Sigma_F(x)=F(x)(1-F(x))-b(x)'\Sigma_\lambda b(x)$ and $b(x)=E\{g(X,\theta_0)I(X\le x)\}$.

It is straightforward to show that the EDF F_n has an asymptotic distribution given by

$$\sqrt{n}(F_n(x)-F(x))\xrightarrow{d}N(0,F(x)(1-F(x))).$$

Since $F(x)(1-F(x)-\Sigma_F(x)=b(x)'\Sigma_\lambda b(x)\ge 0$ in the p.s.d. sense, the EϕDCDFE $\hat{F}_n$ is more efficient than the EDF F_n, a result which is clear from the definition of $\hat{F}_n$ as it incorporates the information contained in the moment conditions (4).

2. Hypothesis testing

By exploiting the asymptotic normality of $\hat{\lambda}_n(\hat{\theta}_n)$ and $\hat{\theta}_n$, we can construct some statistics based on the empirical ϕ-divergence for testing the hypothesis $H_0:\theta=\theta_0$ and the validity of the over-identifying $(r>p)$ moment conditions of $E\{g(X,\theta_0)\}=0$, of which the structure is analogous to that of classical likelihood based test statistics.

Consider the test of the hypothesis $H_0 : \theta = \theta_0$. From the classical viewpoint, a test of the hypothesis $H_0 : \theta = \theta_0$ could be conducted based on the difference of $D_{E\phi}(\hat{\lambda}_n(\theta_0), \theta_0)$ and $D_{E\phi}(\hat{\lambda}_n(\hat{\theta}_n), \hat{\theta}_n)$.

Theorem 4 If Assumptions 1 and 2 are satisfied, then under $H_0 : \theta = \theta_0$, the $E\phi D$ distance statistic

$$\overline{D}(\theta_0, \hat{\theta}_n) = 2n\left(D_{E\phi}(\hat{\lambda}_n(\theta_0), \theta_0) - D_{E\phi}(\hat{\lambda}_n(\hat{\theta}_n), \hat{\theta}_n)\right)$$

converges in distribution to $\chi^2_{(p)}$.

An asymptotic level α test of H_0 can be defined by the rule

$$\text{reject } H_0 : \theta = \theta_0 \text{ if } \overline{D}(\theta_0, \hat{\theta}_n) \geq \chi^2_{(p)}(1-\alpha),$$

where $\chi^2_{(p)}(1-\alpha)$ denotes the $100(1-\alpha)\%$ quantile of the central chi-square distribution with p degrees of freedom. An asymptotic $100(1-\alpha)\%$ confidence region $R_{(1-\alpha)}$ for θ_0 is then constructed in the usual way as

$$R_{(1-\alpha)} = \{\theta \mid \overline{D}(\theta, \hat{\theta}_n) \leq \chi^2_{(p)}(1-\alpha)\}.$$

If the dimension of θ and g are equal (i.e., $r = p$), then we can easily see that the profile empirical ϕ-divergence function $E_\phi D(\hat{\theta}_n)$ is attained at $p_i = n^{-1}$, $i = 1, \cdots, n$, where $\hat{\theta}_n$ is now defined as the solution to the estimating equations $\sum_{i=1}^n g(X_i, \theta) = 0$. This implies by the duality that $\hat{\lambda}_n(\hat{\theta}_n) = 0$ (cf. (2.13)). Thus the $E\phi D$ distance statistic $\overline{D}(\theta_0, \hat{\theta}_n)$ reduces to $2nD_{E\phi}(\hat{\lambda}_n(\theta_0), \theta_0)$.

Corollary 5 Suppose that Assumptions 1 and 2 (a)-(e) hold. If the estimating equations are just-identified (i.e., $r = p$), then under $H_0 : \theta = \theta_0$,

$$\overline{D}(\theta_0, \hat{\theta}_n) = 2nD_{E\phi}(\hat{\lambda}_n(\theta_0), \theta_0) \xrightarrow{d} \chi^2_{(p)}.$$

In the following, we introduce two statistics for testing the validity of the over-identifying ($r>p$) moment conditions $E\{g(X, \theta_0\}=0$. As will be shown below, these test statistics are based on the duality between the moment conditions $E\{g(X, \theta_0\}=0$ and the parametric restriction $\lambda=0$.

Firstly, suppose that the moment conditions are removed from the optimization problem in (5). Then the empirical ϕ-divergence $n^{-1}\Sigma_{i=1}^{n}\phi(p_i n)$ is minimized by $p_i=n^{-1}$, $i=1,\cdots,n$, which corresponds to the imposition of the parametric restriction $\lambda=0$. The value of the unrestricted empirical ϕ-divergence function is then $E_\phi D(\theta)=D_{E\phi}(0,\theta)=0$. Also recall that $E_\phi D(\hat{\theta}_n)=D_{E\phi}(\hat{\lambda}_n(\hat{\theta}_n),\hat{\theta}_n)$ is the minimum empirical ϕ-divergence possible subject to the moment conditions, where $\hat{\lambda}_n(\hat{\theta}_n)$ ensures that the moment conditions are satisfied in the observations (cf. (13)). Hence, for testing the moment conditions $E\{g(X,\theta_0)\}=0$ or, equivalently, for testing $\lambda=0$, the difference of $D_{E\phi}(\hat{\lambda}_n(\hat{\theta}_n), \hat{\theta}_n)$ and $D_{E\phi}(0, \hat{\theta}_n)$ could be used.

Theorem 6 If Assumptions 1 and 2 are satisfied, then under $H_0: E\{g(X,\theta_0)\}=0$ or $\lambda=0$, the $E\phi D$ distance statistic

$$\overline{D}(\hat{\lambda}_n(\hat{\theta}_n), 0)=2n\left(D_{E\phi}(\hat{\lambda}_n(\hat{\theta}_n), \hat{\theta}_n)-D_{E\phi}(0, \hat{\theta}_n)\right)$$

converges in distribution to $\chi^2_{(r-p)}$.

Secondly, because the parametric restriction $\lambda=0$ may be regarded as the dual of the moment conditions $E\{g(X,\theta_0)\}=0$, we can also consider a Lagrange multiplier (LM)-type test statistic given the asymptotic normality of $\hat{\lambda}_n(\hat{\theta}_n)$ by Theorem 2, which is asymptotically equivalent to the $E\phi D$ distance statistic $\overline{D}(\hat{\lambda}_n(\hat{\theta}_n),0)$ (i.e., they differ by $o_p(1)$).

Theorem 7 If Assumptions 1 and 2 are satisfied, then under $H_0: E\{g(X,\theta_0)\}=0$ or $\lambda=0$, the $E\phi D$ LM statistic

$$LM_{E\phi}=n\hat{\lambda}_n(\hat{\theta}_n)'\begin{bmatrix}0 & 0\\ 0 & \hat{\Sigma}_\lambda\end{bmatrix}^{+}\hat{\lambda}_n(\hat{\theta}_n)=n\hat{\lambda}_{2,n}(\hat{\theta}_n)'\hat{\Sigma}_\lambda^{+}\hat{\lambda}_{2,n}(\hat{\theta}_n)$$

is asymptotically equivalent to the $E\phi D$ distance statistic $\overline{D}(\hat{\lambda}_n(\hat{\theta}_n),0)$, where $\hat{\Sigma}_\lambda^+$ denotes the Moore-Penrose inverse of $\hat{\Sigma}_\lambda$, a consistent estimator of Σ_λ.

Remark : By following arguments like those in the proof of Theorem 2, it can be shown that $\hat{S} = \sum_{i=1}^n \hat{p}_i g(X_i,\hat{\theta}_n) g(X_i,\hat{\theta}_n)' \xrightarrow{p} S$ and $\hat{D} = \sum_{i=1}^n \hat{p}_i \partial g(X_i,\hat{\theta}_n)/\partial\theta' \xrightarrow{p} D$. Hence, the asymptotic covariance matrix Σ_λ can be consistently estimated by $\hat{\Sigma}_\lambda = \hat{S}^{-1}(I_r - \hat{D}\hat{\Sigma}_\theta\hat{D}'\hat{S}^{-1})$, where $\hat{\Sigma}_\theta = (\hat{D}'\hat{S}^{-1}\hat{D})^{-1}$. An alternative consistent estimator can also be defined by using the same expression with all of the $\hat{p}_i$'s replaced by n^{-1}, which defines the familiar maximum-likelihood-type estimator of the covariance matrix. However, the former estimator would be expected to be more efficient in finite samples, because it incorporates the additional information contained in the moment conditions via $\hat{p}_i$'s.

Ⅳ. Concluding Remarks

Utilizing the concept of unbiased estimating functions combined with the concept of ϕ-divergence of Csiszár (1963), we have introduced the method of minimum empirical ϕ-divergence estimation and inference. The resultant MEϕDE admits a number of special estimators that have been the focus of recent attention in the statistics and econometrics literature including the maximum empirical likelihood estimator of Qin and Lawless (1994), the exponential tilting estimator of Kitamura and Schutz (1997) and Imbens, Spady and Jonson (1998). Efficiency results for estimators are obtained. Given the parallels with the conventional likelihood, classical-type tests based on the empirical ϕ-divergence for a simple parametric hypothesis and the moment conditions are constructed.

The finite sample behavior of our estimator and EϕD test statistics, and choices of ϕ-function remain to be investigated.

References

Baggerly, K. A.(1998), "Empirical likelihood as a goodness-of-fit measure," *Biometrika,* 85, 535-547.

Chen, S. X.(1993), "On the accuracy of empirical likelihood confidence regions for linear regression model," *Annals of the Institute of Statistical Mathematics,* 45, 621-637.

Chen, S. X.(1994a), "Comparing empirical likelihood and bootstrap hypothesis tests," *Journal of Multivariate Analysis,* 51, 277-293.

Chen, S. X.(1994b), "Empirical likelihood confidence intervals for linear regression coefficients," *Journal of Multivariate Analysis,* 49, 24-40.

Chen, S. X. and P. Hall(1993), "Smoothed empirical likelihood confidence intervals for quantiles," *The Annals of Statistics,* 21, 1166-1181.

Corcoran, S. A.(1998). "Bartlett adjustment of empirical discrepancy statistics," *Biometrika,* 85, 967-972.

Csiszár, I.(1963), "Eine Informationstheoretische Ungleichung und ihre Anwendung auf den Beweis der Ergodizität von Markoffschen Ketten," *Publications of the mathematical Institute of Hungarian Academy of Sciences, Series A,* 8, 84-108.

DiCiccio, T. J., P. Hall, and J. Romano(1988), *Bartlett adjustment for empirical likelihood,* Technical Report, No. 298, Stanford University, Department of Statistics.

DiCiccio, T. J., P. Hall, and J. Romano(1991), "Empirical likelihood is Bartlett-correctable," *The Annals of Statistics,* 19, 1053-1061.

Hall, P. and La B. Scala(1990), "Methodology and algorithms of empirical likelihood," *International Statistical Review,* 58, 109-127.

Imbens, G. W., R. H. Spady, and P. Jonson(1998), "Information theoretic approaches to inference in moment condition models," *Econometrica,* 66, 2, 333-357.

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

Kolaczyk, E. D.(1994), "Empirical likelihood for generalized linear models," *Statistica Sinica,* 4, 199-218.

Kullback, S. and R. Leibler(1951). "On information and sufficiency," *Annals of*

Mathematical Statistics, 22, 79-86.

Liese, F. and I. Vajda(1987), *Convex Statistical Distances*, Leipzig: Teubner.

Mittelhammer, R. C., Judge, G. G., and D. J. Miller(2000), *Econometric Foundations*, Cambridge: Cambridge University Press.

Newey, W. K. and R. J. Smith(2001), *Higher order properties of GMM and generalized empirical likelihood estimators*, Working Paper, M.I.T. Department of Economics.

Neyman, J.(1949), "Contributions to the theory of the χ^2 test," In *Proceedings of the First Berkley Symposium on the Mathematical Statistics and Probability.*

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

Owen, A. B.(1990), "Empirical likelihood ratio confidence regions," *The Annals of Statistics,* 18, 90-120.

Owen, A. B.(1991), "Empirical likelihood for linear models," *The Annals of Statistics,* 19, 1725-1747.

Owen, A. B.(2001), *Empirical Likelihood,* New York: Chapman & Hall/CRC.

Pearson, K.(1900), "On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling," *Philos. Magazine Ser.* 50, 157-172.

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

Read, T. R. C. and N. A. C. Cressie(1988), *Goodness-of-fit Statistics for Discrete Multivariate Data,* New York: Springer-Verlag.

Smith, R. J.(1997), "Alternative semi-parametric likelihood approaches to generalized method of moments estimation," *Economic Journal,* 107, 503-519.

Smith, R. J.(2001), *GEL criteria for moment condition models*, Working Paper, University of Bristol, Department of Economics.

Appendix

Preceding the proofs of the theorems, we first give some preliminary lemmas. The following lemma shows that under Assumption (d), the largest value of $\| g(X_i,\theta)\|$ in a sample of size n is at most of order $n^{1/\alpha}$ in probability for all $\theta\in\Theta$.

Lemma A.1 If Assumption 2 (d) is satisfied, then

$$\max_{1\le i\le n}\sup_{\theta\in\Theta}\| g(X_i,\theta)\| = O_p(n^{1/\alpha}).$$

Proof. Let $b_i = \sup_{\theta\in\Theta}\| g(X_i,\theta)\|$. Then by the Markov inequality and Assumption 2 (d), we have $n^{-1}\sum_{i=1}^{n} b_i^{\alpha} = O_p(1)$. It thus follows that

$$\max_{1\le i\le n} b_i = \left(\max_{1\le i\le n} b_i^{\alpha}\right)^{1/\alpha} \le n^{1/\alpha}\left(\frac{1}{n}\sum_{i=1}^{n} b_i^{\alpha}\right)^{1/\alpha} = O_p(n^{1/\alpha}). \quad \square$$

The next lemma gives the stochastic order of $\| \hat{\lambda}_n(\theta_0)\|$ and an upper bound for $D_{E\phi}(\hat{\lambda}_n(\theta_0),\theta_0)$.

Lemma A.2 If Assumptions 1 and 2 (a)-(e) are satisfied, then $\| \hat{\lambda}_n(\theta_0)\| = O_p(n^{-1/2})$, and there exists some constant $C>0$ such that with probability approaching 1 (w.p.a.1),[5]

$$0 \le D_{E\phi}(\hat{\lambda}_n(\theta_0),\theta_0) \le \frac{1}{2C}\| \bar{g}(\theta_0)\|^2.$$

Proof. Define $\Lambda_n = \{\lambda : \| \lambda\| \le Dn^{-1/\zeta}\}$ for some $D>0$ with $2<\zeta<\alpha$. Then it follows by Lemma A.1 that $\max_{\lambda\in\Lambda_n}\max_{1\le i\le n}\sup_{\theta\in\Theta}|\lambda' G_i(\theta)| \le Dn^{-1/\zeta}\max_{1\le i\le n}\sup_{\theta\in\Theta}\| G_i(\theta)\| = O_p(n^{-1/\zeta+1/\alpha}) \xrightarrow{p} 0$,

5) An assertion will be said to hold true with probability approaching 1 if for every $\varepsilon>0$ there exists some $n_0(\varepsilon)$ such that for each $n\ge n_0(\varepsilon)$ the assertion is true on some event A_n with $P(A_n)\ge 1-\varepsilon$

since $n^{-1/\zeta+1/\alpha} \to 0$ as $n \to \infty$ by $\zeta < \alpha$. Hence, w.p.a.1, $\Lambda_n \subseteq \Lambda(\theta)$ uniformly for all $\theta \in \Theta$, in particular $\Lambda_n \subseteq \Lambda(\theta_0)$. By this and the twice continuous differentiability of ϕ on $(0,\infty)$ (Assumption 1 (a)), it follows that w.p.a.1, $D_{E\phi}(\lambda,\theta_0)$ is twice continuously differentiable on Λ_n, and that w.p.a.1, $\tilde{\lambda}_n = \arg\max_{\lambda \in \Lambda_n} D_{E\phi}(\lambda,\theta_0)$ exists.

Following arguments like those in Newey and Smith (2001), we first show that $\tilde{\lambda}_n = \hat{\lambda}_n(\theta_0) = \arg\max_{\lambda \in \Lambda_n} D_{E\phi}(\lambda,\theta_0)$. By a second-order Taylor expansion of $D_{E\phi}(\tilde{\lambda}_n,\theta_0)$ around $\lambda = 0$, we have

$$0 \le D_{E\phi}(\tilde{\lambda}_n,\theta_0) = \tilde{\lambda}_n' \left\{ \begin{bmatrix} 1 \\ 0 \end{bmatrix} - \frac{1}{n}\sum_{i=1}^{n} G_i(\theta_0) \right\} - \frac{1}{2}\tilde{\lambda}_n' \frac{1}{n}\sum_{i=1}^{n} \frac{G_i(\theta_0)G_i(\theta_0)'}{\phi^{(2)} \circ \phi^{(1)-1}(\bar{\lambda}_n' G_i(\theta_0))} \tilde{\lambda}_n$$

for some (measurable) $\bar{\lambda}_n$ lying between 0 and $\tilde{\lambda}_n$. Since $\bar{\lambda}_n \in \Lambda_n$, the above result implies $\max_{1 \le i \le n} |\bar{\lambda}_n' G_i(\theta_0)| \xrightarrow{p} 0$, so that $\max_{1 \le i \le n} |[\phi^{(2)} \circ \phi^{(1)-1}(\bar{\lambda}_n' G_i(\theta_0))]^{-1} - 1| \xrightarrow{p} 0$. Also, the Markov inequality and Assumption (d) ensure that $n^{-1}\sum_{i=1}^{n} |G_{ik}(\theta_0)G_{il}(\theta_0)| = O_p(1)$, where $G_{ik}(\theta)$ and $G_{il}(\theta)$ denote the k-th and l-th elements of $G_i(\theta)$, $k, l = 1, \cdots, r+1$, respectively. Thus

$$n^{-1}\sum_{i=1}^{n} \{[\phi^{(2)} \circ \phi^{(1)-1}(\bar{\lambda}_n' G_i(\theta_0))]^{-1} - 1\} G_{ik}(\theta_0)G_{il}(\theta_0) \le \max_{1 \le i \le n} |[\phi^{(2)} \circ \phi^{(1)-1}(\bar{\lambda}_n' G_i(\theta_0))]^{-1} - 1| \times$$

$n^{-1}\sum_{i=1}^{n} |G_{ik}(\theta_0)G_{il}(\theta_0)| \xrightarrow{p} 0$, $k, l = 1, \cdots, r+1$. Then it follows by the law of large numbers that

$$\frac{1}{n}\sum_{i=1}^{n} \frac{G_i(\theta_0)G_i(\theta_0)'}{\phi^{(2)} \circ \phi^{(1)-1}(\bar{\lambda}_n' G_i(\theta_0))} = \frac{1}{n}\sum_{i=1}^{n} G_i(\theta_0)G_i(\theta_0)' + o_p(1) \xrightarrow{p} diag\{1, S\}.$$

Since $diag\{1,S\}$ is a positive definite matrix by Assumption (e), there exists some constant $C > 0$ such that w.p.a.1,

$$\frac{1}{n}\sum_{i=1}^{n} \frac{G_i(\theta_0)G_i(\theta_0)'}{\phi^{(2)} \circ \phi^{(1)-1}(\bar{\lambda}_n' G_i(\theta_0))} \ge C I_{r+1}$$

in the p.s.d. sense, and we thus have w.p.a.1,

$$0 \le D_{E\phi}(\tilde{\lambda}_n, \theta_0) \le \tilde{\lambda}_n' \begin{bmatrix} 0 \\ -\bar{g}(\theta_0) \end{bmatrix} - \frac{1}{2} C \tilde{\lambda}_n' \tilde{\lambda}_n \le \| \tilde{\lambda}_n \| \cdot \| \bar{g}(\theta_0 \| - \frac{1}{2} C \| \tilde{\lambda}_n \|^2 .$$

Solving the above inequality for $\| \tilde{\lambda}_n \|$, it follows by the central limit theorem and $\zeta > 2$ that

$$\| \tilde{\lambda}_n \| \le \frac{2}{C} \| \bar{g}(\theta_0 \| = O_p(n^{-1/2}) = o_p(n^{-1/\zeta}) .$$

Hence, w.p.a.1, $\tilde{\lambda}_n$ lies interior to $\Lambda_n \subseteq \Lambda(\theta_0)$ and satisfies the first-order conditions for an interior maximum. Note that $D_{E\phi}(\lambda, \theta_0)$ is concave in λ and $\Lambda(\theta_0)$ is a convex set. It follows that $\tilde{\lambda}_n = \hat{\lambda}_n(\theta_0) = \arg \max_{\lambda \in \Lambda(\theta_0)} D_{E\phi}(\lambda, \theta_0)$ and $\| \hat{\lambda}_n(\theta_0) \| = O_p(n^{-1/2})$. Thus, by the above, w.p.a.1,

$$\begin{aligned} 0 \le D_{E\phi}(\hat{\lambda}_n(\theta_0), \theta_0) &\le \hat{\lambda}_n(\theta_0)' \begin{bmatrix} 0 \\ -\bar{g}(\theta_0) \end{bmatrix} - \frac{1}{2} C \hat{\lambda}_n(\theta_0)' \hat{\lambda}_n(\theta_0) \\ &\le \sup_{\lambda} \left(\lambda' \begin{bmatrix} 0 \\ -\bar{g}(\theta_0) \end{bmatrix} - \frac{1}{2} C \lambda' \lambda \right) \\ &= \frac{1}{2C} \| \bar{g}(\theta_0) \|^2 . \quad \square \end{aligned}$$

Proof of Theorem 1. First, we give the stochastic order of $\| \bar{g}(\hat{\theta}_n) \|$. Let $\tilde{\lambda}_n = [1, -\bar{g}(\hat{\theta}_n)' / \| \bar{g}(\hat{\theta}_n) \|]' \delta_n$, where $\delta_n = n^{-1/\zeta}$. Then Lemma 2 and $\zeta < \alpha$ imply that $\max_{1 \le i \le n} | \tilde{\lambda}_n' G_i(\hat{\theta}_n) | \le \sqrt{2} n^{-1/\zeta} \max_{1 \le i \le n} \| G_i(\hat{\theta}_n) \| = O_p(n^{-1/\zeta + 1/\alpha}) \xrightarrow{p} 0$. Thus $\tilde{\lambda}_n \in \Lambda(\hat{\theta}_n)$ w.p.a.1. By a second-order Taylor expansion of $D_{E\phi}(\tilde{\lambda}_n, \hat{\theta}_n)$ around $\lambda = 0$ gives

$$D_{E\phi}(\tilde{\lambda}_n, \hat{\theta}_n) = \tilde{\lambda}_n' \left\{ \begin{bmatrix} 1 \\ 0 \end{bmatrix} - \frac{1}{n} \sum_{i=1}^{n} G_i(\hat{\theta}_n) \right\} - \frac{1}{2} \tilde{\lambda}_n' \frac{1}{n} \sum_{i=1}^{n} \frac{G_i(\hat{\theta}_n) G_i(\hat{\theta}_n)'}{\phi^{(2)} \circ \phi^{(1)-1}(\bar{\lambda}_n' G_i(\hat{\theta}_n))} \tilde{\lambda}_n$$

for some $\bar{\lambda}_n$ lying between 0 and $\tilde{\lambda}_n$. Since $\| \bar{\lambda}_n \| \le \| \tilde{\lambda}_n \|$, it follows from the above results that $\max_{1 \le i \le n} | \bar{\lambda}_n' G_i(\hat{\theta}_n) | \xrightarrow{p} 0$, and thus $\max_{1 \le i \le n} | [\phi^{(2)} \circ \phi^{(1)-1}(\bar{\lambda}_n' G_i(\hat{\theta}_n))]^{-1} - 1 | \xrightarrow{p} 0$. Further,

the Markov inequality and Assumption (d) imply that $n^{-1}\sum_{i=1}^{n}|G_{ik}(\hat{\theta}_n)G_{il}(\hat{\theta}_n)| = O_p(1)$, $k, l = 1, \cdots, r+1$. Thus $n^{-1}\sum_{i=1}^{n}\{[\phi^{(2)} \circ \phi^{(1)-1}(\bar{\lambda}_n' G_i(\hat{\theta}_n))]^{-1} - 1\}G_{ik}(\hat{\theta}_n)G_{il}(\hat{\theta}_n) \leq \max_{1\leq i\leq n}|[\phi^{(2)} \circ \phi^{(1)-1}(\bar{\lambda}_n' G_i(\hat{\theta}_n))]^{-1} - 1| \, n^{-1}\sum_{i=1}^{n}|G_{ik}(\hat{\theta}_n)G_{il}(\hat{\theta}_n)| \xrightarrow{p} 0$, $k, l = 1, \cdots, r+1$. Then, w.p.a.1, for C sufficiently large,

$$\frac{1}{n}\sum_{i=1}^{n}\frac{G_i(\hat{\theta}_n)G_i(\hat{\theta}_n)'}{\phi^{(2)} \circ \phi^{(1)-1}(\bar{\lambda}_n' G_i(\hat{\theta}_n))} = \frac{1}{n}\sum_{i=1}^{n} G_i(\hat{\theta}_n)G_i(\hat{\theta}_n)' + o_p(1) \leq CI_{r+1},$$

and thus

$$D_{E\phi}(\tilde{\lambda}_n, \hat{\theta}_n) \geq \tilde{\lambda}_n' \begin{bmatrix} 0 \\ -\bar{g}(\hat{\theta}_n) \end{bmatrix} - \frac{1}{2}C\tilde{\lambda}_n'\tilde{\lambda}_n = \delta_n \| \bar{g}(\hat{\theta}_n) \| - C\delta_n^2. \tag{a.1}$$

Noting that, by $\hat{\lambda}_n(\hat{\theta}_n)$ and $\hat{\theta}_n$ being saddle point of $D_{E\phi}(\lambda, \theta)$,

$$D_{E\phi}(\tilde{\lambda}_n, \hat{\theta}_n) \leq D_{E\phi}(\hat{\lambda}_n(\hat{\theta}_n), \hat{\theta}_n) \leq D_{E\phi}(\hat{\lambda}_n(\hat{\theta}_n), \theta_0) \leq D_{E\phi}(\hat{\lambda}_n(\theta_0), \theta_0). \tag{a.2}$$

It follows by Lemma 2 and (a.1) that, w.p.a.1, $\delta_n \| \bar{g}(\hat{\theta}_n) \| - C\delta_n^2 \leq (1/2C) \| \bar{g}(\theta_0) \|^2$. Solving for $\| \bar{g}(\hat{\theta}_n) \|$ then gives

$$\| \bar{g}(\hat{\theta}_n) \| \leq C\delta_n + \frac{1}{2C\delta_n} \| \bar{g}(\theta_0) \|^2 = C\delta_n + o_p(1)/C = O_p(\delta_n). \tag{a.3}$$

Now, for any $a_n \to 0$, re-define $\tilde{\lambda}_n = [\| \bar{g}(\hat{\theta}_n) \|, -\bar{g}(\hat{\theta}_n)']' a_n$. Note that $\tilde{\lambda}_n = o_p(\delta_n)$ by (a.3), so that $\tilde{\lambda}_n \in \Lambda(\hat{\theta}_n)$ w.p.a.1. Then, following arguments similar to those above, we have, w.p.a.1, $D_{E\phi}(\tilde{\lambda}_n, \hat{\theta}_n) \geq a_n \| \bar{g}(\hat{\theta}_n) \|^2 (1 - Ca_n)$. It then follows by Lemma 2 and (a.2) that $a_n \| \bar{g}(\hat{\theta}_n) \|^2 (1 - Ca_n) \leq (1/2C) \| \bar{g}(\theta_0) \|^2 = O_p(n^{-1})$. Since, for any finite C, and for all n large enough, $1 - Ca_n$ is bounded away from zero, it follows that $a_n \| \bar{g}(\hat{\theta}_n) \|^2 = O_p(n^{-1})$, implying $\| \bar{g}(\hat{\theta}_n) \| = O_p(n^{-1/2})$.

Next, we show the weak consistency of the MEϕDE $\hat{\theta}_n$ using the arguments of Newley and Smith (2001). Note that, by the uniform law of large numbers, it follows

under Assumptions (a), (c) and (d) that $\sup_{\theta\in\Theta}\| \bar{g}(\theta) - E\{g(X,\theta)\}\| \xrightarrow{p} 0$ and $E\{g(X,\theta)\}$ is continuous on Θ. The triangle inequality and $\| \bar{g}(\hat{\theta}_n)\| = O_p(n^{-1/2})$ then gives

$$\| E\{g(X,\hat{\theta}_n)\}\| \le \| \bar{g}(\hat{\theta}_n) - E\{g(X,\hat{\theta}_n)\}\| + \| \bar{g}(\hat{\theta}_n)\| \xrightarrow{p} 0.$$

Since, by Assumption (b), $E\{g(X,\theta)\} = 0$ has a unique solution at θ_0, $\| E\{g(X,\theta)\}\|$ must be bounded away from zero outside any neighborhood of θ_0. Hence, $\hat{\theta}_n$ must be inside any neighborhood of θ_0 for sufficiently large n. Since the above arguments hold true w.p.a.1, $\hat{\theta}_n \xrightarrow{p} \theta_0$.

Finally, using arguments similar to those in the proof of Lemma 2, it follows by the consistency of $\hat{\theta}_n$ and $\| \bar{g}(\hat{\theta}_n)\| = O_p(n^{-1/2})$ that, w.p.a.1, $\hat{\lambda}_n(\hat{\theta}_n) = \arg\max_{\lambda\in\Lambda(\hat{\theta}_n)} D_{E\phi}(\lambda,\hat{\theta}_n)$ exists and $\| \hat{\lambda}_n(\hat{\theta}_n)\| = O_p(n^{-1/2})$. □

Proof of Theorem 2. For convenience, we introduce some notations. Define

$$\Delta_\lambda(\lambda,\theta) = \frac{\partial D_{E\phi}}{\partial\lambda} = \begin{bmatrix}1\\0\end{bmatrix} - \frac{1}{n}\sum_{i=1}^{n}\phi^{(1)-1}(\lambda' G_i(\theta))G_i(\theta) \quad \text{and}$$

$$\Delta_\theta(\lambda,\theta) = \frac{\partial D_{E\phi}}{\partial\theta} = -\frac{1}{n}\sum_{i=1}^{n}\phi^{(1)-1}(\lambda' G_i(\theta))\left(\frac{\partial G_i(\theta)}{\partial\theta'}\right)'\lambda.$$

Also define

$$\Delta_{\lambda\lambda}(\lambda,\theta) = \partial\Delta_\lambda(\lambda,\theta)/\partial\lambda', \quad \Delta_{\lambda\theta}(\lambda,\theta) = \partial\Delta_\lambda(\lambda,\theta)/\partial\theta',$$
$$\Delta_{\theta\lambda}(\lambda,\theta) = \partial\Delta_\theta(\lambda,\theta)/\partial\lambda', \quad \text{and} \quad \Delta_{\theta\theta}(\lambda,\theta) = \partial\Delta_\theta(\lambda,\theta)/\partial\theta'.$$

First, note that the consistency results given by Theorem 2 and Assumptions 2 (f)-(g) imply that $\hat{\lambda}_n(\hat{\theta}_n)$ and $\hat{\theta}_n$ satisfy the following first-order conditions w.p.a.1 :

$$\Delta_\lambda(\hat{\lambda}_n(\hat{\theta}_n),\hat{\theta}_n) = \begin{bmatrix}1\\0\end{bmatrix} - \frac{1}{n}\sum_{i=1}^{n}\phi^{(1)-1}(\hat{\lambda}_n(\hat{\theta}_n)' G_i(\hat{\theta}_n))G_i(\hat{\theta}_n) = 0,$$

$$\Delta_\theta(\hat{\lambda}_n(\hat{\theta}_n),\hat{\theta}_n) = -\frac{1}{n}\sum_{i=1}^{n}\phi^{(1)-1}(\hat{\lambda}_n(\hat{\theta}_n)'G_i(\hat{\theta}_n))\left(\frac{\partial G_i(\hat{\theta}_n)}{\partial\theta'}\right)'\hat{\lambda}_n(\hat{\theta}_n) = 0.$$

Taking a first-order Taylor expansion of these equations around $(0,\theta_0)$ and rearranging, we have

$$0 = -\begin{bmatrix}0\\ \sqrt{n}\bar{g}(\theta_0)\\ 0\end{bmatrix} + \begin{bmatrix}\Delta_{\lambda\lambda}(\bar{\lambda}_n,\bar{\theta}_n) & \Delta_{\lambda\theta}(\bar{\lambda}_n,\bar{\theta}_n)\\ \Delta_{\theta\lambda}(\bar{\lambda}_n,\bar{\theta}_n) & \Delta_{\theta\theta}(\bar{\lambda}_n,\bar{\theta}_n)\end{bmatrix}\begin{bmatrix}\sqrt{n}\hat{\lambda}_n(\hat{\theta}_n)\\ \sqrt{n}(\hat{\theta}_n-\theta_0)\end{bmatrix} \tag{a.4}$$

for some (measurable) $(\bar{\lambda}_n,\bar{\theta}_n)$ lying on the line segment joining $(0,\theta_0)$ and $(\hat{\lambda}_n(\hat{\theta}_n),\hat{\theta}_n)$ that is actually different from row to row of the matrix of second derivatives (thus, $(\bar{\lambda}_n,\bar{\theta}_n)\xrightarrow{p}(0,\theta_0)$ by the consistency), where

$$\Delta_{\lambda\lambda}(\bar{\lambda}_n,\bar{\theta}_n) = -\frac{1}{n}\sum_{i=1}^{n}\frac{G_i(\bar{\theta}_n)G_i(\bar{\theta}_n)'}{\phi^{(2)}\circ\phi^{(1)-1}(\bar{\lambda}_n'G_i(\bar{\theta}_n))},$$

$$\Delta_{\lambda\theta}(\bar{\lambda}_n,\bar{\theta}_n) = -\frac{1}{n}\sum_{i=1}^{n}\left(\frac{G_i(\bar{\theta}_n)\bar{\lambda}_n'}{\phi^{(2)}\circ\phi^{(1)-1}(\bar{\lambda}_n'G_i(\bar{\theta}_n))} + \phi^{(1)-1}(\bar{\lambda}_n'G_i(\bar{\theta}_n))I_{r+1}\right)\frac{\partial G_i(\bar{\theta}_n)}{\partial\theta'},$$

$$\Delta_{\theta\lambda}(\bar{\lambda}_n,\bar{\theta}_n) = \Delta_{\lambda\theta}(\bar{\lambda}_n,\bar{\theta}_n)',$$

$$\Delta_{\theta\theta}(\bar{\lambda}_n,\bar{\theta}_n) = -\frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{\phi^{(2)}\circ\phi^{(1)-1}(\bar{\lambda}_n'G_i(\bar{\theta}_n))}\left(\frac{\partial G_i(\bar{\theta}_n)}{\partial\theta'}\right)'\bar{\lambda}_n\bar{\lambda}_n'\frac{\partial G_i(\bar{\theta}_n)}{\partial\theta'}\right.$$

$$\left. + \phi^{(1)-1}(\bar{\lambda}_n'G_i(\bar{\theta}_n))\sum_{j=1}^{r+1}\frac{\partial^2 G_{ij}(\bar{\theta}_n)}{\partial\theta\partial\theta'}\bar{\lambda}_{nj}\right),$$

and $G_{ij}(\bar{\theta}_n)$ and $\bar{\lambda}_{nj}$ denote the j-th element of $G_i(\bar{\theta}_n)$ and $\bar{\lambda}_n$, respectively. Note that $\|\bar{\lambda}_n\| \le \|\hat{\lambda}_n(\hat{\theta}_n)\| = O_p(n^{-1/2})$ by Theorem 1. By Lemma 1 and $\alpha > 2$, it then follows that $\max_{1\le i\le n}|\bar{\lambda}_n'G_i(\bar{\theta}_n)| = O_p(n^{-1/2+1/\alpha})\xrightarrow{p}0$, so that $\max_{1\le i\le n}|[\phi^{(1)-1}(\bar{\lambda}_n'G_i(\bar{\theta}_n))]^{-1}-1|\xrightarrow{p}0$ and $\max_{1\le i\le n}|[\phi^{(2)}\circ\phi^{(1)-1}(\bar{\lambda}_n'G_i(\bar{\theta}_n))]^{-1}-1|\xrightarrow{p}0$. Further, Assumptions 2 (a), (c)-(d) ensure that, by the uniform law of large numbers, $n^{-1}\sum_{i=1}^{n}g(X_i,\theta)g(X_i,\theta)'\xrightarrow{p}E\{g(X,\theta)g(X,\theta)'\}$ uniformly

in $\theta \in \Theta$. Similarly, Assumptions 2 (a), (f)-(g) ensure that $n^{-1}\sum_{i=1}^{n} \partial g(X_i,\theta)/\partial\theta' \xrightarrow{p}$ $E\{\partial g(X,\theta)/\partial\theta'\}$ uniformly in $\theta \in N_0$. Then, for any sequence of measurable functions $\{\theta_n : \Omega \to \Theta\}$ such that $\theta_n \xrightarrow{p} \theta_0$, it follows that $n^{-1}\sum_{i=1}^{n} g(X_i,\theta_n)g(X_i,\theta_n)' \xrightarrow{p} S$ and $n^{-1}\sum_{i=1}^{n} \partial g(X_i,\theta_n)/\partial\theta' \xrightarrow{p} D$ by a standard argument (see, e.g., White (1994, Corollary 3.8)). Thus $n^{-1}\sum_{i=1}^{n} g(X_i,\bar{\theta}_n)g(X_i,\bar{\theta}_n)' \xrightarrow{p} S$ and $n^{-1}\sum_{i=1}^{n} \partial g(X_i,\bar{\theta}_n)/\partial\theta' \xrightarrow{p} D$. Using these results, it then follows

$$\Delta_{\lambda\lambda}(\bar{\lambda}_n,\bar{\theta}_n) = -\frac{1}{n}\sum_{i=1}^{n} G_i(\bar{\theta}_n)G_i(\bar{\theta}_n)' + o_p(1) \xrightarrow{p} -diag\{1, S\},$$

$$\Delta_{\lambda\theta}(\bar{\lambda}_n,\bar{\theta}_n) = -\frac{1}{n}\sum_{i=1}^{n} \frac{\partial G_i(\bar{\theta}_n)}{\partial\theta'} + o_p(1) \xrightarrow{p} -\begin{bmatrix} 0 \\ D \end{bmatrix},$$

$$\Delta_{\theta\lambda}(\bar{\lambda}_n,\bar{\theta}_n) \xrightarrow{p} -[0 \quad D'] \text{ and } \Delta_{\theta\theta}(\bar{\lambda}_n,\bar{\theta}_n) = o_p(1) \xrightarrow{p} 0.$$

Now, inverting and solving in the equation (a.4), we have by the above results

$$\begin{bmatrix} \sqrt{n}\hat{\lambda}_n(\hat{\theta}_n) \\ \sqrt{n}(\hat{\theta}_n - \theta_0) \end{bmatrix} = -\begin{bmatrix} 0 \\ S^{-1}(I_r - D(D'S^{-1}D)^{-1}D'S^{-1}) \\ (D'S^{-1}D)^{-1}D'S^{-1} \end{bmatrix} \sqrt{n}\bar{g}(\theta_0) + o_p(1) \qquad \text{(a.5)}$$

The results of the theorem follow from this equation, since $\sqrt{n}\bar{g}(\theta_0) \xrightarrow{d} N(0, S)$ by the central limit theorem. □

Proof of Theorem 3. A first-order Taylor expansion of $\sqrt{n}(\hat{F}_n(x) - F(x))$ around $\lambda = 0$ gives

$$\sqrt{n}(\hat{F}_n(x) - F(x)) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(I(X_i \le x) - F(x)) + \frac{1}{n}\sum_{i=1}^{n} \frac{G_i(\hat{\theta}_n)'I(X_i \le x)}{\phi^{(2)} \circ \phi^{(1)-1}(\bar{\lambda}_n'G_i(\hat{\theta}_n))} \sqrt{n}\hat{\lambda}_n(\hat{\theta}_n)$$

for some $\bar{\lambda}_n$ lying between 0 and $\hat{\lambda}_n(\hat{\theta}_n)$. By $\| \bar{\lambda}_n \| \le \| \hat{\lambda}_n(\hat{\theta}_n) \| = O_p(n^{-1/2})$ and following arguments like those in the proof of Theorem 2, it can be shown that the second term on the right hand side can be approximated by $n^{-1}\sum_{i=1}^{n} G_i(\hat{\theta}_n)'I(X_i \le x)$

$\sqrt{n}\hat{\lambda}_n(\hat{\theta}_n)+o_p(1)$. Further, the uniform law of large numbers and consistency imply $n^{-1}\sum_{i=1}^{n} G_i(\hat{\theta}_n)'I(X_i \le x) \xrightarrow{p} E\{G(\theta_0)'I(X_i \le x)\}$. Then, using these results and substituting $\sqrt{n}\hat{\lambda}_n(\hat{\theta}_n) = -[0, \Sigma_\lambda]'\sqrt{n}\bar{g}(\theta_0)+o_p(1)$ given by (a.5) into the above equation, we have

$$\begin{aligned}
&\sqrt{n}(\hat{F}_n(x)-F(x)) \\
&= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(I(X_i \le x)-F(x)) + E\{G(\theta_0)'I(X_i \le x)\}\sqrt{n}\hat{\lambda}_n(\hat{\theta}_n)+o_p(1) \\
&= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(I(X_i \le x)-F(x)) - E\{g(X,\theta_0)'I(X \le x)\}\Sigma_\lambda\sqrt{n}\bar{g}(\theta_0)+o_p(1) \\
&= [1, -E\{g(X,\theta_0)'I(X \le x)\}\Sigma_\lambda]\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\begin{bmatrix} I(X_i \le x)-F(x) \\ g(X_i,\theta_0) \end{bmatrix}+o_p(1).
\end{aligned}$$

The result of the theorem follows from this equation, since, by the central limit theorem

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\begin{bmatrix} I(X_i \le x)-F(x) \\ g(X_i,\theta_0) \end{bmatrix} \xrightarrow{d} N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} F(x)(1-F(x)) & b(x)' \\ b(x) & S \end{bmatrix}\right). \quad \square$$

Proof of Theorem 4. We first show that $D_{E\phi}(\hat{\lambda}_n(\theta_0),\theta_0)$ and $D_{E\phi}(\hat{\lambda}_n(\hat{\theta}_n),\hat{\theta}_n)$ can be approximated by simple quadratic forms, which will be useful in the proofs of Theorem 4 and of the subsequent theorems. By Lemma 2, $\hat{\lambda}_n(\theta_0)$ satisfies the first-order conditions

$$\frac{1}{n}\sum_{i=1}^{n}\phi^{(1)-1}(\hat{\lambda}_n(\theta_0)'G_i(\theta_0))G_i(\theta_0)-\begin{bmatrix} 1 \\ 0 \end{bmatrix}=0$$

w.p.a.1. A first-order Taylor expansion around $\lambda=0$ gives

$$0=\frac{1}{n}\sum_{i=1}^{n}G_i(\theta_0)-\begin{bmatrix} 1 \\ 0 \end{bmatrix}+\frac{1}{n}\sum_{i=1}^{n}\frac{G_i(\theta_0)G_i(\theta_0)'}{\phi^{(2)}\circ\phi^{(1)-1}(\bar{\lambda}_n'G_i(\theta_0))}\hat{\lambda}_n(\theta_0)$$

for some (measurable) $\bar{\lambda}_n$ lying between 0 and $\hat{\lambda}_n(\theta_0)$. Then, by the law of large numbers and Assumption (e), we have

$$\hat{\lambda}_n(\theta_0) = -\left(\frac{1}{n}\sum_{i=1}^{n} G_i(\theta_0)G_i(\theta_0)'\right)^{-1}\begin{bmatrix} 0 \\ \bar{g}(\theta_0) \end{bmatrix} + o_p(n^{-1/2}), \tag{a.6}$$

since $\| \bar{g}(\theta_0) \| = O_p(n^{-1/2})$ by the central limit theorem. Note that we have made use of the approximation

$$\frac{1}{n}\sum_{i=1}^{n} \frac{G_i(\theta_0)G_i(\theta_0)'}{\phi^{(2)} \circ \phi^{(1)-1}(\bar{\lambda}_n' G_i(\theta_0))} = \frac{1}{n}\sum_{i=1}^{n} G_i(\theta_0)G_i(\theta_0)' + o_p(1).$$

Also, taking a second-order Taylor expansion of $D_{E\phi}(\hat{\lambda}_n(\theta_0), \theta_0)$ around $\lambda = 0$ and using approximation similar to that above, we have

$$D_{E\phi}(\hat{\lambda}_n(\theta_0), \theta_0) = -\hat{\lambda}_n(\theta_0)'\begin{bmatrix} 0 \\ \bar{g}(\theta_0) \end{bmatrix} - \frac{1}{2}\hat{\lambda}_n(\theta_0)'\frac{1}{n}\sum_{i=1}^{n} G_i(\theta_0)G_i(\theta_0)'\hat{\lambda}_n(\theta_0) + o_p(n^{-1}). \tag{a.7}$$

Inserting the expression for $\hat{\lambda}_n(\theta_0)$ given by (a.5) into (a.6) and rearranging, it follows that

$$2nD_{E\phi}(\hat{\lambda}_n(\theta_0), \theta_0) = n\bar{g}(\theta_0)' S_n(\theta_0)^{-1} \bar{g}(\theta_0) + o_p(1), \tag{a.8}$$

where $S_n(\theta) = n^{-1}\sum_{i=1}^{n} (g(X_i, \theta) - \bar{g}(\theta))(g(X_i, \theta) - \bar{g}(\theta))'$. Similarly, using the results of Theorem 1, it can be shown that

$$2nD_{E\phi}(\hat{\lambda}_n(\hat{\theta}_n), \hat{\theta}_n) = n\bar{g}(\hat{\theta}_n)' S_n(\hat{\theta}_n)^{-1} \bar{g}(\hat{\theta}_n) + o_p(1). \tag{a.9}$$

Now, taking a first-order Taylor expansion of $\bar{g}(\hat{\theta}_n)$ around θ_0, we have

$$\bar{g}(\hat{\theta}_n) = \bar{g}(\theta_0) + \frac{\partial \bar{g}(\bar{\theta}_n)}{\partial \theta'}(\hat{\theta}_n - \theta_0)$$

for some $\bar{\theta}_n$ lying between θ_0 and $\hat{\theta}_n$. Since the uniform law of large numbers and consistency imply $\partial \bar{g}(\bar{\theta}_n)/\partial \theta' \xrightarrow{p} D$, and $(\hat{\theta}_n - \theta_0) = -\Sigma_\theta D' S^{-1} \bar{g}(\theta_0) + o_p(n^{-1/2})$ by (a.5), we may write $\bar{g}(\hat{\theta}_n) = (I_r - D\Sigma_\theta D' S^{-1})\bar{g}(\theta_0) + o_p(n^{-1/2}) = S\Sigma_\lambda \bar{g}(\theta_0) + o_p(n^{-1/2})$. Substituting this expression into (a.9) and simplifying then give

$$2nD_{E\phi}(\hat{\lambda}_n(\hat{\theta}_n),\hat{\theta}_n) = n\bar{g}(\theta_0)'\Sigma_\lambda S S_n(\hat{\theta}_n)^{-1} S\Sigma_\lambda \bar{g}(\theta_0) + o_p(1). \quad \text{(a.10)}$$

Then, by (a.8) and (a.10), we have

$$\begin{aligned}\overline{D}(\theta_0,\hat{\theta}_n) &= n\bar{g}(\theta_0)'(S_n(\theta_0)^{-1} - \Sigma_\lambda S S_n(\hat{\theta}_n)^{-1} S\Sigma_\lambda)\bar{g}(\theta_0) + o_p(1) \\ &= (S^{-1/2}\sqrt{n}\bar{g}(\theta_0))'S^{1/2}(S_n(\theta_0)^{-1} - \Sigma_\lambda S S_n(\hat{\theta}_n)^{-1} S\Sigma_\lambda)S^{1/2}(S^{-1/2}\sqrt{n}\bar{g}(\theta_0)) + o_p(1).\end{aligned}$$

Note that, by the law of large numbers,

$$S_n(\theta_0) = \frac{1}{n}\sum_{i=1}^{n} g(X_i,\theta_0)g(X_i,\theta_0)' - \bar{g}(\theta_0)\bar{g}(\theta_0)' \xrightarrow{p} S. \quad \text{(a.11)}$$

and, by the uniform law of large numbers and consistency,

$$S_n(\hat{\theta}_n) = \frac{1}{n}\sum_{i=1}^{n} g(X_i,\hat{\theta}_n)g(X_i,\hat{\theta}_n)' - \bar{g}(\hat{\theta}_n)\bar{g}(\hat{\theta}_n)' \xrightarrow{p} S. \quad \text{(a.12)}$$

It then follows by the continuous mapping theorem that

$$S^{1/2}(S_n(\theta_0)^{-1} - \Sigma_\lambda S S_n(\hat{\theta}_n)^{-1} S\Sigma_\lambda)S^{1/2} \xrightarrow{p} I_r - S^{1/2}\Sigma_\lambda S\Sigma_\lambda S^{1/2}.$$

Since $I_r - S^{1/2}\Sigma_\lambda S\Sigma_\lambda S^{1/2} = I_r - S^{1/2}\Sigma_\lambda S^{1/2}$ is symmetric and idempotent with trace equal to p and $S^{-1/2}\sqrt{n}\bar{g}(\theta_0) \xrightarrow{d} N(0,I_r)$ by the central limit theorem, the result of the theorem follows. □

Proof of Corollary 5. By the quadratic approximation for $2nD_{E\phi}(\hat{\lambda}_n(\theta_0),\theta_0)$ given by (a.8) and (a.11), the proof is straightforward.

Proof of Theorem 6. Observe that $\overline{D}(\hat{\lambda}_n(\hat{\theta}_n),0) = 2nD_{E\phi}(\hat{\lambda}_n(\hat{\theta}_n),\hat{\theta}_n)$ as $D_{E\phi}(0,\hat{\theta}_n) = 0$. By (a.10), (a.12), and the continuous mapping theorem, we have

$$\begin{aligned}2nD_{E\phi}(\hat{\lambda}_n(\hat{\theta}_n),\hat{\theta}_n) &= n\bar{g}(\theta_0)'\Sigma_\lambda S S_n(\hat{\theta}_n)^{-1} S\Sigma_\lambda \bar{g}(\theta_0) + o_p(1) \\ &= n\bar{g}(\theta_0)'\Sigma_\lambda S\Sigma_\lambda \bar{g}(\theta_0) + o_p(1) \\ &= (S^{-1/2}\sqrt{n}\bar{g}(\theta_0))'S^{1/2}\Sigma_\lambda S\Sigma_\lambda S^{1/2}(S^{-1/2}\sqrt{n}\bar{g}(\theta_0)) + o_p(1),\end{aligned}$$

where $S^{1/2}\Sigma_\lambda S\Sigma_\lambda S^{1/2} = S^{1/2}\Sigma_\lambda S^{1/2}$ is a symmetric and idempotent matrix with trace equal to $r-p$. Since $S^{-1/2}\sqrt{n}\bar{g}(\theta_0) \xrightarrow{d} N(0, I_r)$ by the central limit theorem, the result of the theorem follows. □

Proof of Theorem 7. Note first that, by the arguments in the proof of Theorem 6, we have $\bar{D}(\hat{\lambda}_n(\hat{\theta}_n),0) = 2nD_{E\phi}(\hat{\lambda}_n(\hat{\theta}_n),\hat{\theta}_n) = n\bar{g}(\theta_0)'\Sigma_\lambda\bar{g}(\theta_0) + o_p(1)$. Then, substituting the expression $\hat{\lambda}_{2,n}(\hat{\theta}_n) = -\Sigma_\lambda\sqrt{n}\bar{g}(\theta_0) + o_p(1)$ given by (a.5) into $LM_{E\phi}$, it follows by the consistency that

$$\begin{aligned} LM_{E\phi} &= n\hat{\lambda}_{2,n}(\hat{\theta}_n)'\hat{\Sigma}_\lambda^+\hat{\lambda}_{2,n}(\hat{\theta}_n) \\ &= n\bar{g}(\theta_0)'\Sigma_\lambda\hat{\Sigma}_\lambda^+\Sigma_\lambda\bar{g}(\theta_0) + o_p(1) \\ &= n\bar{g}(\theta_0)'\Sigma_\lambda\bar{g}(\theta_0) + o_p(1) \\ &= \bar{D}(\hat{\lambda}_n(\hat{\theta}_n),0) + o_p(1). \quad \square \end{aligned}$$

〔Abstract〕

경험적 ϕ-발산을 이용한 추정 및 추론

조 영 수

본 논문은 Owen(1988)의 경험적 우도방법론에서 사용되는 로그-우도비통계량을 특수한 경우로써 포함하는 보다 일반적인 두 확률분포간의 거리측도로서의 Csiszár(1963)의 ϕ-발산을 이용한 대안적인 비모수적 통계적 추론방법을 도입하였다. 특히 불편추정함수 또는 추정방정식 개념을 본 방법론에 연결시킴으로써 연구자에게 주어지는 여러 가지 종류의 모수 또는 확률분포에 관한 추가적인 정보들을 추론과정에 쉽게 도입할 수 있게끔 하였다.

먼저, 이 방법론이 표준적인 정칙조건하에서 모수와 확률분포에 대한 효율적인 추정을 가능하게 함을 보였다. 또한 단순한 모수적 가설과 모멘트조건에 대한 검정통계량을 도입하고 이 통계량들이 고전적 우도방법론에 근거한 검정통계량과 유사한 점근적 특성을 가지고 있음을 보였다.

핵심용어 : 경험적 우도, 경험적 ϕ-발산, 추정방정식