An Easy Empirical Likelihood Approach To Efficient Estimation In Models With Side Information

Fei Tan * Hanxiang Peng *

Abstract

In this article, we construct semiparametrically efficient estimates in a nonparametric model with side information which can be described by expectation equations of some known functions. These estimates are given by easy maximum empirical likelihood estimates (MELEs) of the model that can found as solutions to certain estimating equations, which extend those of Qin and Lawless (1994) for MELEs from smooth estimating functions to discontinuous ones. In comparison with the usual MELEs, these are computationally fast and mathematically tractable. We calculate their computational complexities and derive easy MELEs for some problems of which the usual MELEs are difficult or even unable to be obtained. We give asymptotic normality and efficiency for MELEs in general M-estimation models which allow for discontinuous criterion and estimating functions. We also derive the MELEs for differentiable statistical functionals, von-Mises functionals, L-estimators, and U-statistics. Besides, we present several examples about side information.

Key Words: maximum empirical likelihood estimator, semiparametric efficiency, side information, statistical functional, U-statistics

1. Introduction

The empirical likelihood approach was introduced by Owen (1990, 2001) to construct confidence intervals in a nonparametric setting. Soon it was realized that it can also be used to construct point estimators. Qin and Lawless (1994) linked empirical likelihood with generalized estimating equations and investigated maximum empirical likelihood estimators (MELEs). They established consistency and asymptotic normality of MELEs under the usual regularity conditions, and demonstrated that the variance of a MELE will not increase when the number of estimating equations is increased. Furthermore, they showed that MELEs are fully semiparametrically efficient in the sense of least dispersed regular estimators (Bickel, *et al.* (1993); van der Vaart (2000)). Peng and Schick (2013) explored MELEs in the case of constraint functions that may be discontinuous and/or depend on additional parameters. The later is the case in applications to semiparametric models where the constraint functions may depend on the nuisance parameter.

Let $(\mathscr{X}, \mathscr{S})$ be a measurable space, $\mathscr{Q}$ be a family of probability measures on $\mathscr{S}$, and κ be a functional from $\mathscr{Q}$ onto an open subset Θ of $\mathbb{R}^k$. Let $Z_1, \dots, Z_n$ be independent and identically distributed (i.i.d.) copies of Z taking value in $\mathscr{X}$ with an unknown distribution Q belonging to the model $\mathscr{Q}$. We are interested in statistical inference about the characteristic $\theta = \kappa(Q)$ when side information is available. Let us now introduce the following condition (K1). Write $A^{\otimes 2} = AA^\top$ for a matrix A.

(K1a) There is a measurable function $u : \mathscr{X} \to \mathbb{R}^m$ such that $\int u\, dQ = 0$ and $W_u = \int u^{\otimes 2}\, dQ$ is positive definite.

*Indiana University Purdue University Indianapolis, Department of Mathematical Sciences, Indianapolis, IN 46202-3216, USA.

(K1b) There is a measurable function $v : \mathscr{Z} \times \Theta \to \mathbb{R}^d$, with $d \geq k$, such that for every R in $\mathscr{Q}$, $\int v(z, \kappa(R))\, dR(z) = 0$, and $W_v(R) = \int v(z, \kappa(R))^{\otimes 2}\, dR(z)$ is positive definite.

(K1c) For every R in $\mathscr{Q}$, the matrix $W(R) = \int w(z, \kappa(R))^{\otimes 2}\, dR(z)$ is positive definite, where $w = (v^\top, u^\top)^\top$.

Let u, v satisfy (K1). To construct a confidence set in this setup, Owen confronted the maximization problem:

$$\mathscr{R}_n(\vartheta) = \sup\Big\{\prod_{j=1}^n n\pi_j : \pi \in \mathscr{P}_n, \sum_{j=1}^n \pi_j w(Z_j, \vartheta) = 0\Big\}, \quad \vartheta \in \Theta,$$

where $\mathscr{P}_n$ denotes the closed probability simplex in dimension n, i.e.,

$$\mathscr{P}_n = \Big\{\pi = (\pi_1, \dots, \pi_n)^\top \in [0,1]^n : \sum_{i=1}^n \pi_i = 1\Big\}.$$

Qin and Lawless (1994) tackled point estimation for $\theta = \kappa(Q)$ and studied the maximum empirical likelihood estimator of θ:

$$\hat{\theta}_n = \arg\max_{\vartheta \in \Theta} \mathscr{R}_n(\vartheta). \tag{1.1}$$

Arguably, it could be difficult to establish desired properties for a possibly irregular $\mathscr{R}_n(\vartheta)$ even if strong assumptions are imposed on the estimating (or constraint) function $w(z, \vartheta)$. In fact, only in a few cases does $\mathscr{R}_n(\vartheta)$ have explicit formulas, see e.g. Remark 6 of Peng and Schick (2013). As another example, consider the case that $w(z, \vartheta)$ is convex in ϑ. It would be nontrivial to prove or disprove that $\mathscr{R}_n(\vartheta)$ is convex, even for the simple case $w(z, \vartheta) = z - \vartheta$.

Let us now consider the M-estimator (or Z-estimator) $\tilde{\vartheta}_n$ of θ, which is defined to be a solution to the sample equation

$$\bar{v}_n(\vartheta) := \frac{1}{n}\sum_{j=1}^n v(Z_j, \vartheta) = 0. \tag{1.2}$$

When side information is available given by expectation equation $E(u(Z)) = 0$, we are naturally motivated to look at the empirical likelihood

$$\mathscr{R}_n = \sup\Big\{\prod_{j=1}^n n\pi_j : \pi \in \mathscr{P}_n, \sum_{j=1}^n \pi_j u(Z_j) = 0\Big\}.$$

For example, we want to estimate the p-th quantile ϑ when there is the side information that the median is zero. In this case, we may take $v(z, \vartheta) = 1[z \leq \vartheta] - p, p \in (0,1)$ and $u(z) = 1[z \leq \theta] - 1/2$. Following Owen, one uses Lagrange multipliers to derive

$$\pi_{nj} = \frac{1}{n}\frac{1}{1 + \zeta_n^\top u(Z_j)}, \quad j = 1, \dots, n, \tag{1.3}$$

where ζ_n satisfies the equation

$$\frac{1}{n}\sum_{j=1}^n \frac{u(Z_j)}{1 + \zeta^\top u(Z_j)} = 0. \tag{1.4}$$

These π_{nj}'s carry the side information. To utilize this information it is very natural for us to employ π_{nj}'s as carriers to pass on the information to the estimator of parameter through equation (1.2). This consideration leads us to estimating the parameter θ by $\hat{\vartheta}_n$ as a solution to the equation

$$\bar{\boldsymbol{v}}_n(\vartheta) := \sum_{j=1}^{n} \pi_{nj} v(Z_j, \vartheta) = \frac{1}{n} \sum_{j=1}^{n} \frac{v(Z_j, \vartheta)}{1 + \zeta_n^\top u(Z_j)} = 0. \quad (1.5)$$

We are not the first to use this idea. It was used by Zhang (1995, 1997) in M-estimation and quantile processes in the presence of auxiliary (side) information. He established consistency and asymptotic normality, and proved that the asymptotic variances of the resulting estimators are smaller than those of the usual sample M-estimators and sample quantiles. Hellerstein and Imbens (1999) utilized this idea for the least squares estimators in a linear regression model, while Bravo (2010) introduced a class of M-estimators based on generalized empirical likelihood estimation with side information with empirical likelihood as a special case. More discussions about the latter two can be found in Example 2.2. Relatively recently, Yuan *et al.* (2012) explored this idea in U-statistics with side information. Tang and Leng (2012) employed the idea to construct improved estimators of parameters in quantile regression. For convenience, we shall refer to (1.5) as empirical likelihood (EL-) weighted equation and the resulting estimators as EL-weighted estimators.

We show in this article that the M-estimator $\hat{\vartheta}_n$ defined in (1.5) is asymptotically equivalent to the MELE $\hat{\theta}_n$ defined in (1.1) in the sense that both have the identical asymptotic normal distribution, see Theorem 4.2 – Theorem 4.4 below. Consequently, the estimator $\hat{\vartheta}_n$ obtained from (1.5) is the MELE for θ in the model defined by estimating equations $E(u(Z)) = 0$ and $E(v(Z, \theta)) = 0$, hence is fully semiparametrically efficient for θ in the sense of least dispersed regular estimators as shown in Qin and Lawless (1994). It is noteworthy that if there is no side information available, then we simply take $u \equiv 0$ and the EL-weighted equation $\bar{\boldsymbol{v}}_n(\vartheta) = 0$ boils down to the usual sample equation $\bar{v}_n(\vartheta) = 0$, and the MELE $\hat{\vartheta}_n$ reduces to the usual sample M-estimator $\tilde{\vartheta}_n$. In other words, the usual sample M-estimators are the MELEs for the parameters.

Qin and Lawless (1994) derived the estimating equation for finding the MELE $\hat{\theta}_n$ defined in (1.1) under the regularity assumptions including, in particular, the differentiability of w w.r.t. parameter ϑ. In Section 3 we show that (1.4)-(1.5) are the estimating equations for the MELE's even when w is discontinuous, thus our results extend those of Qin and Lawless to irregular constraints in this case. We also demonstrate in Section 3 that $\hat{\vartheta}_n$ is computationally faster than the usual MELE $\hat{\theta}_n$ by comparing their computational complexities. We exhibit that $\hat{\vartheta}_n$ is mathematically more convenient than the usual MELE $\hat{\theta}_n$ and explain that the EL-weight method can be applied to extend the scope of the usual MELEs. We have obtained MELEs and their asymptotic normality results under general conditions and, in particular, allow for irregular estimating functions.

We give several examples about the side information in Section 2 to illustrate the applicability of the method. In Example 2.1, the side information is the knowledge about marginal distributions, which has a long history, see the discussions therein. That the side information in census data can be used to improve estimation is discussed in Example 2.2. That the independence of gene and environment can increase efficiency is considered in Example 2.3. That raters share similar rating behaviors in the interrater agreement study is investigated in Example 2.4. That the symmetry of random errors in a random-effects model can be used is described in Example 2.5.

The rest of this paper is organized as follows. Examples about side information are given in Section 2. In Section 3, we discuss computational complexity and mathematical simplicity of the easy MELE's. In Section 4, we first report the consistency result of EL-weighted M-estimators, followed by asymptotic normality of EL-weighted M-, GM- and AGM-estimators under general conditions, and ended up with maximum empirical likelihood estimators. In Section 5, we apply the obtained results to various situations.

2. Examples of side information

In this section, we give several examples about side information.

Example 2.1. Bickel, *et al.* (1991) studied efficient estimation of linear functional $\theta = E(h(X,Y)) = \int h\,dP$ of a probability measure P for known h when the marginal distributions P_X of X and P_Y of Y are known. One of their justifications of known marginals is as follows. One observes a random sample $(X_i, Y_i), i = 1, \dots, n$ from the joint P. In addition, one also observes random samples $X_{1i} : i = 1, \dots, n_1$ from P_X and $Y_{2i} : i = 1, \dots, n_2$ from P_Y. If n_1, n_2 are very large relative to n, we could act as if P_X, P_Y are known and equal to the empirical distributions of the n_1 auxiliary X's and n_2 auxiliary Y's. Such a real example is the census data which provide nearly exact information of moments of the marginal distributions of economic variables, see e.g. Imbens and Lancaster (1994). Peng and Schick (2005) investigated the same estimation problem but with equal but unknown marginals. Haberman (1984) considered minimum Kullback-Leibler divergence -type estimators for this problem involving a fixed number of side information. Vitale (1979) looked at a regression version of this problem. All these can be viewed as an extension of the work on estimation of cell probabilities on contingency tables with known marginals (Deming and Stephan (1940); Ireland and Kullback (1968)).

Bickel, *et al.* constructed efficient estimators based on minimizing modified chisquare statistic while Peng and Schick used the least squares criterion. We can use the EL- weight method to construct efficient estimators to the above problem. In fact, we can directly apply the results for a finite number of constraints while for an infinite number of constraints this is also doable which is pursued in elsewhere. Note that if a distribution is discrete and has finite a support then known marginals or equal marginals are equivalent to a finite number of constraints. Let us illustrate the EL- weight method with the case of known marginals. Suppose P_X and P_Y have finite supports $x_l : l = 1, \dots, L$ and $y_m : m = 1, \dots, M$ with known probability distributions p_l and q_m respectively. Then the side information is given by $E(u(X,Y)) = 0$ where $u(x,y) = (\mathbf{1}[x = x_l] - p_l, \mathbf{1}[y = y_m] - q_m : l = 1, \dots, L, m = 1, \dots, M)^\top$. By Theorem 4.4 below, an efficient estimator of $\theta = E(h(X,Y))$ is given by

$$\hat{\theta}_n = \frac{1}{n} \sum_{j=1}^{n} \frac{h(X_j, Y_j)}{1 + \zeta_n^\top u(X_j, Y_j)},$$

where ζ_n is the solution to (1.4) with the above u as side information.

Example 2.2. Let $Y_i, i = 1, \dots, n$ be independent with mean $E(Y_i) = \mu(\theta_i)$ and variance $\mathrm{Var}(Y_i) = V(\theta_i)$ for some real parameter θ_i. Let $X_1, \dots, X_n$ be i.i.d.

covariate vectors. A popular estimate of a regression parameter β is a solution to the generalized estimating equation (GEE)

$$\sum_{i=1}^{n} \frac{Y_i - \mu_i(\beta)}{V_i(\beta)} h_i'(\beta) X_i = 0, \tag{2.1}$$

where h is a usual inverse link function with $h_i(\beta) = h(X_i^\top \beta)$, $\mu_i(\beta) = \mu(h_i(\beta))$ and $V_i(\beta) = V(\mu_i(\beta))$. In GEE models only up to second order moments are used while higher order moments are replaced by what are called "working matrices", the resulting estimators are thus usually not efficient. One can use the EL-weight method to improve efficiency when side information is available. This was pursued in Hellerstein and Imbens (1999) and Bravo (2010), the former looked at the least squares estimators in linear models while the latter studied a general setting which takes the empirical likelihood as a special case. Side information can be moment equalities in census data. Examples of such data are the national employment rate, data on the frequency of unemployment spells of certain length, and aggregate expenditure for various goods, see e.g. Imbens and Lancaster (1994). By reweighing the least squares estimators in a linear regression model with the empirical likelihood weights where constraints are derived from side information, Hellerstein and Imbens (1999) gave efficient estimators of the regression coefficients and applied the results to analyze a real data. Relatively recently Bravo (2010) introduced a class of M-estimators based on generalized empirical likelihood estimation with side information and showed that the resulting class of estimators is efficient in the sense that it achieves the same asymptotic lower bound as that of the efficient GMM estimator with the same side information. The empirical likelihood is a special case of the M-estimators.

As in Hellerstein and Imbens (1999), the side information is $E(u(X,Y)) = 0$ with, for example, $u(X,Y) = Y - m_0$ for some known m_0 or $u(X,Y) = \mathbf{1}[(X,Y) \in C] - u_0$ where the researcher knows the probability (u_0) that (X,Y) is in a particular subset C of the sample space. Then an efficient estimate of β is any solution to the equation

$$\sum_{i=1}^{n} \frac{Y_i - \mu_i(\beta)}{(1 + \zeta_n^\top u(X_i, Y_i)) V_i(\beta)} h_i'(\beta) X_i = 0, \tag{2.2}$$

where ζ_n is the solution to the equation (1.4) with the above u as side information.

Example 2.3. Chatterjee and Carroll (2005) pointed out that a special feature of the gene-environment interaction problem is that it may often be reasonable to assume that a subject's genetic susceptibility (G), a factor which is determined from birth, is independent of his/her subsequent environmental exposure (E). Standard logistic regression analysis remains a valid option for analyzing case-control data. However, the method may not be efficient because it fails to exploit the gene-environment independence assumption. They gave semiparametric maximum likelihood estimation to make use of the independence assumption. Here we propose to use the EL- weight method applied on the standard logistic regression. Let $A_l : l = 1, \ldots, L$ be a partition of the range of G and $B_m : m = 1, \ldots, M$ a partition of the range of E, so that $A_l \times B_m : l = 1, \ldots, L, m = 1, \ldots, M$ form a partition of the range of (G, E). Independence implies

$$E(u_{lm}(G,E)) = 0, \quad l = 1, \ldots, L, m = 1, \ldots, M, \tag{2.3}$$

where $u_{lm}(G,E) = \mathbf{1}[(G,E) \in A_l \times B_m] - p_l q_m$ with $p_l = P(G \in A_l)$ and $q_m = P(E \in B_m)$. With a similar argument used in Example 2.1 let us assume p_l, q_m

are known, so that we can devote all our attention to the interaction of the genetic factor G and environmental risk factor E. Thus we take the side information as $E(u(G,E)) = 0$ where $u = (u_{lm} : l = 1, \ldots, L, m = 1, \ldots, M)^\top$.

Let D be the binary indicator of presence, D=1, or absence, D=0, of a disease. Suppose the prospective risk model for the disease given G and E is given by the logistic regression model $P(D = 1|G,E) = h(\beta_0 + \beta_1 G + \beta_2 E + \beta_3 G \times E)$, where $h(x) = (1+\exp(-x))^{-1}$ is the inverse logistic link. Suppose that n_0 controls and n_1 cases are sampled from the conditional distributions $P(G,E|D = 1)$ and $P(G,E|D = 0)$, respectively, and let (G_i, E_i) be $n_0 + n_1$ denote the corresponding covariate data of the n_0+n_1 study subjects. This is a special case of (2.1) and (2.2) with $Y = D$, $X = (1, G, E, G \times E)^\top$, the inverse logistic link h, and the above side information u.

Example 2.4. Consider the interrater agreement of two raters with two rating categories *yes* and *no*. The widely used Cohen's unweighted kappa is defined as

$$\kappa = \frac{\Pi_0 - \Pi_c}{1 - \Pi_c},$$

where $\Pi_0 = P(X = Y)$ denotes the measure of agreement and $\Pi_c = P(X = yes)P(Y = yes) + P(X = no)P(Y = no)$ the measure of agreement by chance. Clearly $-1 \le \kappa \le 1$. Sinha (2013) recently studied modifications of the kappa to cope with different issues. One of these is the case that it is apriori known that there is known agreement between the two raters, i.e., $\Pi_0 = P(X = Y) = \pi_0$ for some known $0 < \pi_0 < 1$. This happens in real-life situations that two raters may exhibit similar rating behavior with certain probability. Thus the side information can be expressed with the expectation equation $E(u(X,Y)) = 0$ where $u(x,y) = \mathbf{1}[x = y] - \pi_0$ and (K1) is met as $\int u^2\, dQ = \pi_0(1 - \pi_0) > 0$. The MELE for κ is given by

$$\hat{\kappa} = \frac{\pi_0 - \hat{\Pi}_c}{1 - \hat{\Pi}_c},$$

where $\hat{\Pi}_c$ is the MELE for Π_c (we omit the details). It is easy to verify the conditions of Theorem 4.3, hence $\hat{\kappa}$ is an efficient estimator for κ.

Example 2.5. In a balanced one-way random effects model, the response Y_{ij}, random effect u_i and random error ϵ_{ij} satisfy

$$Y_{ij} = \mu + u_i + \epsilon_{ij}, \quad i = 1, \ldots, n,\ j = 1, \ldots, J (J \ge 2), \tag{2.4}$$

where μ is the mean response, the ϵ_{ij}'s are i.i.d. with mean zero and variance $\sigma_\epsilon^2 = \mathrm{Var}(\epsilon_{ij})$, the u_i's are i.i.d. with mean zero and variance $\sigma_u^2 = \mathrm{Var}(u_j)$, and ϵ_{ij}'s and u_i's are independent and have finite fourth moments. Following Arvesen (1969), put

$$\mathbf{X}_i = \begin{pmatrix} Y_{i\cdot} \\ (J-1)^{-1} \sum_{j=1}^J (Y_{ij} - Y_{i\cdot})^2 \end{pmatrix}, \quad i = 1, \ldots, n, \tag{2.5}$$

where $A_{i\cdot} = J^{-1} \sum_{j=1}^J A_{ij}$ denotes the average of A_{ij} over j. Clearly $\mathbf{X}_1, \ldots, \mathbf{X}_n$ are i.i.d. Suppose there is available additional information about the model, for instance, ε as an i.i.d. copy of $\varepsilon_i = u_i + \epsilon_{i\cdot}$ is symmetric about zero. In this formulation, the model (2.4) becomes

$$Y_{i\cdot} = \mu + \varepsilon_i, \quad i = 1, \ldots, n.$$

This is the well known symmetric location model. For illustration, let us consider improved estimation of the variance σ_u^2 of the random effect. Let

$$h(\mathbf{X}_1, \mathbf{X}_2) = 2^{-1}\big((Y_{1\cdot} - Y_{2\cdot})^2 - J^{-1}(\kappa(\mathbf{X}_1) + \kappa(\mathbf{X}_2))\big),$$

where $\kappa(\mathbf{X}_i) = (J-1)^{-1}\sum_{j=1}^J (Y_{ij} - Y_{i\cdot})^2$. It is easy to see h is a symmetric kernel and satisfies $E(h(\mathbf{X}_1, \mathbf{X}_2)) = \sigma_u^2$. Then an unbiased estimator of σ_u^2 is the U-statistic

$$U_n(h) = T(F_n) = \binom{n}{2}^{-1} \sum_{i<j} h(\mathbf{X}_i, \mathbf{X}_j),$$

where F_n is the empirical distribution of $Y_{i\cdot} - \mu_0, i = 1, \dots, n$. We assume $\mu = \mu_0$ for some known μ_0. We are motivated by the fact that if J is big then we take $\mu_0 = J^{-1}\sum_{j=1}^J Y_{1j}$ and work with the observations $\mathbf{X}_2, \dots, \mathbf{X}_n$. Otherwise we take $\mu_0 = Y_{\cdot\cdot}$ the grand mean of all Y_{ij}. This does not fit our theory but one can expect that the results approximately hold in the sense that the asymptotic variance will be a bit bigger than when μ_0 is known. For $B \subset \mathbb{R}$, write $-B = \{-x : x \in B\}$. Let $\{B_k, -B_k : k = 1, \dots, K\}$ be a partition of the real line $\mathbb{R}$. Then symmetry implies

$$E(\mathbf{1}[Y_{i\cdot} - \mu_0 \in B_k] - \mathbf{1}[Y_{i\cdot} - \mu_0 \in -B_k]) = 0, \quad i = 1, \dots, n, k = 1, \dots, K.$$

These suggest us to take

$$u(\mathbf{X}_i) = (\mathbf{1}[Y_{i\cdot} - \mu_0 \in B_k] - \mathbf{1}[Y_{i\cdot} - \mu_0 \in -B_k],\, k = 1, \dots, K)^\top.$$

Assume $p_k = P(Y_{1\cdot} - \mu_0 \in B_k) \neq 0, k = 1, \dots, K$. Then $W_u = \int uu^\top \, dQ = 2\mathrm{diag}(p_1, \dots, p_K)$ is nonsingular. Thus the EL-weighted estimator of σ_u^2 is given by

$$T(\mathbb{F}_n) = \frac{1}{n^2}\sum_{i<j} \frac{h(\mathbf{X}_i, \mathbf{X}_j)}{(1 + \zeta_n^\top u(\mathbf{X}_i))(1 + \zeta_n^\top u(\mathbf{X}_j))}.$$

where ζ_n satisfies $n^{-1}\sum_{j=1}^n u(\mathbf{X}_i)/(1 + \zeta_n^\top u(\mathbf{X}_i)) = 0$.

3. Computational complexity and mathematical tractability

In this section, we explain why the estimating equations for the MELEs of Qin and Lawless (1994) still holds for discontinuous constraint functions, calculate the complexities and discuss the possibilities of breaking the limitations of the usual MELEs.

Under the regularity conditions on the constraint function $w(z, \vartheta)$ such as the differentiability of w w. r. t. ϑ, Qin and Lawless (1994) in their Lemma 1 derived that the MELE $\hat{\theta}_n$ defined in (1.1) can be found as the solution to the estimating equations for ϑ and t,

$$\frac{1}{n}\sum_{j=1}^n \frac{w(Z_j, \vartheta)}{1 + t^\top w(Z_j, \vartheta)} = 0, \quad \frac{1}{n}\sum_{j=1}^n \frac{\dot{w}^\top(Z_j, \vartheta)t}{1 + t^\top w(Z_j, \vartheta)} = 0. \tag{3.1}$$

Under the current setting, $\dot{w} = (\dot{v}^\top, 0^\top)^\top$. Substituting this in the second equation of (3.1) and partitioning $t = (t_1^\top, t_2^\top)^\top$, one finds $t_1 = 0$. Now substituting this in the first equation, one derives exactly the estimating equations (1.4) and (1.5). However, when w is discontinuous, equations (3.1) can't be applied. Thus our result that the solution $\hat{\vartheta}_n$ of (1.4)-(1.5) is the MELE extends equations (3.1) to discontinuous constraint functions. Apparently what happens here is that the solution $t_1 = 0$ leads $\dot{v}(z, \vartheta)t_1^\top = 0$, which holds even if v hence w is not differentiable (at least heuristically and we justify this in Section 4). Moreover, equations (1.4)-(1.5) reduces the computational burden of equations (3.1) as we explained below.

The preceding equations can numerically be solved by the iterative newton method such as the commonly used elm algorithm. Let us now look at the time complexity. The execution time of a program depends on the number of floating-point operations (FLOPs) and the processor speed of the computer used which is defined by FLOPs/sec. Specifically,

Time required (sec)=Number of FLOPs / Processor speed (FLOPs/sec).

It is well known that the number of FLOPs using the newton method to solve the two equations in (3.1) is $O(nL_1(d+m)^3) + O(nL_2d^3)$, where L_1, L_2 are the desired numbers of loops of the first and second equation in (3.1) and $d+m$, d are the dimensions of w, ϑ. Let $L = \max(L_1, L_2)$. Then the number of FLOPs for solving equations (3.1) is $FL_0 = O(nL(d+m)^3)$. Likewise, the number of FLOPs for solving equations (1.4) and (1.5) is $O(nLm^3) + O(nLd^3) = O(nL\max(d,m)^3)$. Thus the FLOPs are reduced from $FL_0 = O(nL(d+m)^3)$ to $FL_1 = O(nL\max(d,m)^3)$.

The results about the MELE's of statistical functionals in Section 5 break the limitations of the usual empirical likelihood. For example, the usual empirical likelihood for U-statistics involve the quadratic products $\pi_i\pi_j$ of the probability weights. This causes difficulty to obtain explicit formulas for π_i as we have for the usual case when the π_i's are linear in the empirical likelihood ratio. Our results beat this difficulty and show that the plug-in estimates of the statistical functionals with the MELE of the distribution are also MELE's. Besides, it might worth to mention that the EL-weighted estimating equation (1.5) is generally not more difficult to deal with than the usual sample estimating equation (1.2), but is a lot easier to deal with than the maximization problem (1.1).

4. Asymptotic Efficiency

Recall in the Introduction $(\mathscr{Z}, \mathscr{S})$ denotes a measurable space, $\mathscr{Q}$ is a family of probability measures on $\mathscr{S}$, and κ is a function from $\mathscr{Q}$ onto an open subset Θ of $\mathbb{R}^k$. We are interested in statistical inference about $\theta = \kappa(Q)$ when side information is available.

Throughout this article, we assume that u satisfies (K1a). Let v be a function from $\mathscr{Z} \times \Theta$ to $\mathbb{R}^d$ such that condition (K1b) is met, where Θ is a nonempty subspace of $\mathbb{R}^k$. In the usual M-estimation, typically the number of parameters is equal to the number of estimating equations, i.e. $d = k$. From now on, we shall assume this for v unless otherwise indicated.

The usual sample M-estimator of θ is defined as a solution to the sample estimating equation (1.2), where the identical probability weight n^{-1} is assigned to every observation Z_j. When side information is available via equality $E(u(Z)) = 0$ about the underlying distribution, this assignment does not make use of the side information. Based on the empirical likelihood theory, each distinct observation Z_j must be assigned a distinct EL-probability weight, which is the π_{nj} given in (1.3). Then find an estimator of θ as a solution to (1.5). We shall refer this to as the principle of maximum empirical likelihood.

Following the definitions in Chapter 7 of Bickel, Klaassen, Ritov and Wellner (1993), a solution $\hat{\vartheta}_n$ of (1.5) is called an EL-weighted generalized M-estimator (EL-weighted GM-estimator for short). If $\hat{\vartheta}_n$ satisfies

$$\bar{\boldsymbol{v}}_n(\hat{\vartheta}_n) = \frac{1}{n}\sum_{j=1}^{n} \frac{v(Z_j,\hat{\vartheta}_n)}{1+\zeta_n^\top u(Z_j)} = o_p(n^{-1/2}),$$

then it is called an EL-weighted asymptotic generalized M-estimator (EL- weighted AGM-estimator for short).

Let us first address consistency. There are numerous results in the literature. One can use these results to derive the consistency of the EL-weighted estimators. As an illustration, we present the following consistency result based on Theorem 5.9 of van der Vaart (2000), and omit the proof due to the page limit. Here as usual, a well separated zero point of the criterion function (i.e. (4.3)) plays the key role. Denote by $\|a\|$ the euclidean norm of a vector a.

Theorem 4.1. *Let v be a measurable function from $\mathscr{X} \times \Theta$ to $\mathbb{R}^k$ such that*

$$\sup_{\vartheta\in\Theta} \frac{1}{n}\sum_{j=1}^{n} \|v(Z_j,\vartheta)\|^2 < \infty, \quad \textit{and} \tag{4.1}$$

$$\sup_{\vartheta\in\Theta} \Big\|\frac{1}{n}\sum_{j=1}^{n} \big(v(Z_j,\vartheta) - E(v(Z,\vartheta))\big)\Big\| = o_p(1). \tag{4.2}$$

If for every $\epsilon > 0$,

$$\inf_{\vartheta:\|\vartheta-\theta\|\ge\epsilon} \|E(v(Z,\vartheta))\| > 0 = \|E(v(Z,\theta))\|, \tag{4.3}$$

then any sequence of estimators $\hat{\vartheta}_n$ such that $\bar{\boldsymbol{v}}_n(\hat{\vartheta}_n) = o_p(1)$ converges in probability to the true value θ of parameter.

We now study asymptotic normality. As in the case of consistency, there are numerous results about asymptotic normality of M-estimators in the literature. We can use these results and apply our method to obtain asymptotic normality results for the EL-weighted estimators. Here for illustration, we are satisfied with the aforementioned EL-weighted AGM-estimators which cover irregular constraint functions such as indicator functions. Our results are formulated in the framework of the master theorem for asymptotic generalized estimates, see e.g. Theorem 1 (page 514) of Bickel, *et al.* (1993).

Recall $\bar{v}_n(\vartheta)$ and $\bar{\boldsymbol{v}}_n(\vartheta)$ defined in the Introduction. From now on, we shall write $\bar{u}_n = n^{-1}\sum_{j=1}^n u(Z_j)$ and $C(v) = E(v(Z,\theta)\otimes u^\top(Z))$, where $\otimes$ denotes the Kronecker product. Recall the W_u given in (K1a), the $W_v(R)$ in (K1b) and set $W_v = W_v(Q) = \mathrm{Var}(v(Z,\theta))$ and $M = W_v - C(v)W_u^{-1}C(v)^\top$. Note that the positive definiteness of $W = W(Q)$ given in (K1c) implies M is also positive definite. We omit the proof due to the page limit.

Theorem 4.2. *Suppose (K1) holds with $d = k$. Suppose also $E(v(Z,\vartheta))$ is differentiable with respect to ϑ for every $\vartheta \in \Theta$ such that its negative gradient A at $\vartheta = \theta = \kappa(Q)$ with $Q \in \mathscr{Q}$ is nonsingular. Assume as $\varepsilon_n \downarrow 0$,*

$$\sup_{\|\vartheta-\theta\|\le\varepsilon_n} \left\{ \frac{\sqrt{n}\|\bar{v}_n(\vartheta) - \bar{v}_n(\theta) - E(v(Z,\vartheta) - v(Z,\theta))\|}{1+\sqrt{n}\|\vartheta-\theta\|} \right\} = o_p(1), \tag{4.4}$$

$$\sup_{\|\vartheta-\theta\|\le\varepsilon_n}\left\{\frac{n^{-1}\sum_{j=1}^n\|v(Z_j,\vartheta)-v(Z_j,\theta)\|^2}{1+n\|\vartheta-\theta\|^2}\right\}=o_p(1). \tag{4.5}$$

Let $\hat\vartheta_n$ be an EL-weighted AGM-estimator of θ on $\mathscr{Q}$. If $\hat\vartheta_n$ is consistent, then it satisfies the stochastic expansion,

$$\hat\vartheta_n=\theta+A^{-1}\frac{1}{n}\sum_{j=1}^n\big(v(Z_j,\theta)-C(v)W_u^{-1}u(Z_j)\big)+o_p(n^{-1/2}). \tag{4.6}$$

Accordingly, $\sqrt{n}(\hat\vartheta_n-\theta)\implies\mathscr{N}(0,A^{-1}MA^{-\top})$.

It may worth to note that the uniform convergence in (4.2), (4.4) and (4.5) is commonly used in the literature and can be proved using such as exponential inequalities.

Qin and Lawless (1994) derived the asymptotic results for the MELE $\hat\theta_n$ of parameter θ defined in (1.1) for constraint functions which possess continuous second order differentiability. Here we give the asymptotic results of the MELE for constraint functions which may be discontinuous. The results can be obtained as a special case of Theorem 1.2 in Peng and Schick (2013), which is stated below. Notice here that $d\ge k$ is allowed. We need the following conditions.

(K2) For every finite constant C,

$$D_n(C)=\sup_{\|t\|\le C}\frac{1}{n}\sum_{j=1}^n\|v(Z_j,\theta+n^{-1/2}t)-v(Z_j,\theta)\|^2=o_p(1).$$

(K3) There is a $d\times k$ matrix A of full rank k such that

$$\sup_{\|t\|\le C}\|\frac{1}{\sqrt{n}}\sum_{j=1}^n[v(Z_j,\theta+n^{-1/2}t)-v(Z_j,\theta)]+At\|^2=o_p(1)$$

for every constant C.

Typically, $A=-E[\dot v(Z,\theta)]$, where $\dot v(z,\vartheta)$ denotes the derivative of $v(z,\vartheta)$ with respect to parameter ϑ for $z\in\mathscr{Z}$. These assumptions are modifications of (K0) – (K2) in Peng and Schick (2013). Under (K3), a consistent estimate of A is given by $\hat A=(\hat A_{rs})$ with

$$\hat A_{rs}=-\frac{1}{c_s\sqrt{n}}\sum_{j=1}^n[v_r(Z_j,\hat\theta_n+c_sn^{-1/2}e_s)-v_r(Z_j,\hat\theta_n)], \tag{4.7}$$

where c_s is a constant and e_s denotes the unit vector with the s-th component being one.

Theorem 4.3. *Suppose (K1) – (K3) hold. Assume that the random function $\mathscr{R}_n(\vartheta)$ is upper semicontinuous on Θ. Then there is a local maximizer $\hat\theta_n$ such that*

$$\hat\theta_n=\theta+J^{-1}A^\top M^{-1}\frac{1}{n}\sum_{j=1}^n\big(v(Z_j,\theta)-C(v)W_u^{-1}u(Z_j)\big)+o_p(n^{-1/2}), \tag{4.8}$$

where $J=A^\top M^{-1}A$. Accordingly $\sqrt{n}(\hat\theta_n-\theta)\implies\mathscr{N}(0,J^{-1})$.

Remark 4.1. *Under the conditions of Theorem 4.3, a consistent estimate of J is given by $\hat{J} = \hat{A}^\top \hat{M}^{-1}\hat{A}$, where $\hat{A}$ is given in (4.7) and $\hat{M} = \hat{W}_v - \hat{C}\hat{W}_u^{-1}\hat{C}^\top$ with $\hat{W}_u = n^{-1}\sum_{j=1}^n u(Z_j)^{\otimes 2}$, $\hat{W}_v = n^{-1}\sum_{j=1}^n v(Z_j,\hat{\theta}_n)^{\otimes 2}$ and $\hat{C} = n^{-1}\sum_{j=1}^n v(Z_j,\hat{\theta}_n)u(Z_j)^\top$.*

Note that the MELE $\hat{\theta}_n$ given in Theorem 4.3 possesses the identical asymptotic variance-covariance matrix as that of the EL-weighted estimator $\hat{\vartheta}_n$ in Theorem 4.2 when A is nonsingular. Consequently, we obtain our main result below.

Theorem 4.4. *Suppose $u : \mathscr{Z} \mapsto \mathbb{R}^m$ and $v : \mathscr{Z} \times \Theta \mapsto \mathbb{R}^k$ with $\Theta \subset \mathbb{R}^k$ satisfy all the assumptions in Theorem 4.2 and Theorem 4.3. Then $\hat{\vartheta}_n$ is the MELE for θ in the model defined by $E(u(Z)) = 0$ and $E(v(Z,\theta)) = 0$. In particular, the usual sample M-estimator is the MELE for θ if no additional information is available.*

Remark 4.2. *Zhang (1995) investigated the M-estimation with auxiliary information and his estimators are the same as ours. Our results show that his estimators are in fact the MELEs for the parameters.*

5. MELEs for statistical functionals

5.1 The MELEs for linear functionals of a probability measure

Consider efficient estimation of the expectation $\theta = E(\psi(Z)) = \int \psi\, dQ$ for some known measurable function ψ from $\mathscr{Z}$ to $\mathbb{R}^k$ such that the variance-covariance matrix $\mathrm{Var}(\psi(Z))$ is (componentwise) finite and positive definite, when side information is available expressed via $E(u(Z)) = 0$. We take $v(z,\vartheta) = \vartheta - \psi(z), z \in \mathscr{Z}, \vartheta \in \Theta$ for some compact subset Θ of $\mathbb{R}^k$. The EL-weighted M-estimator $\hat{\theta}_n$ of θ is then the solution to the equation

$$\bar{\boldsymbol{v}}_n(\vartheta) = \sum_{j=1}^n \pi_{nj} v(Z_j,\vartheta) = \sum_{j=1}^n \pi_{nj}\,(\vartheta - \psi(Z_j)) = 0, \quad \vartheta \in \Theta,$$

where π_{nj} are the EL-weights given in (1.3). This has an explicit solution, yielding the following natural estimator of θ,

$$\hat{\theta}_n = \sum_{j=1}^n \pi_{nj}\psi(Z_j). \tag{5.1}$$

We have proved the following result with the proof omitted.

Theorem 5.1. *Suppose ψ is a measurable function from $\mathscr{Z}$ to $\mathbb{R}^d$ such that $W(\psi)$ is positive definite. Then $\hat{\theta}_n$ given in (5.1) is the MELE for $\theta = E(\psi(Z))$ and satisfies the stochastic expansion,*

$$\hat{\theta}_n = \bar{\psi}_n - C(\psi)W_u^{-1}\bar{u}_n + o_p(n^{-1/2}),$$

where $\bar{\psi}_n = n^{-1}\sum_{j=1}^n \psi(Z_j)$. Thus $\sqrt{n}(\hat{\theta}_n - \theta) \implies \mathscr{N}(0,\Sigma(\psi))$.

5.2 The MELE for a distribution function

As an application of Theorem 5.1, take $\psi(Z) = \mathbf{1}[Z \le z], Z \in \mathbb{R}^r$ for a fixed z. The resulting expected value is the cdf $F(z) = P(Z \le z)$. When side information is available via $E(u(Z)) = 0$, the MELE $\mathbb{F}_n(z)$ for $F(z)$ is given by

$$\mathbb{F}_n(z) = \sum_{j=1}^n \pi_{nj}\mathbf{1}[Z_j \le z] = \frac{1}{n}\sum_{j=1}^n \frac{\mathbf{1}[Z_j \le z]}{1+\zeta_n^\top u(Z_j)}, \quad z \in \mathbb{R}^r,$$

where π_{nj} are the EL-weights given in (1.3).

Let us now consider the case of $r = 1$ and view a cdf F as a member of the usual Skorohod space $D[-\infty, \infty]$, the set of functions h which are right continuous with left limits, equipped with the uniform norm $\| \|_\infty$, i.e. $\|h\|_\infty = \sup_{z \in [-\infty,\infty]} |h(z)|$. According to van der Vaart (2000), an estimator T_n of a parameter $\psi(Q)$ in $D[-\infty, \infty]$ is *asymptotically efficient* at Q if $\sqrt{n}(T_n - \psi(Q))$ converges under Q in distribution to a tight limit and d^*T_n is asymptotically efficient for estimating $d^*\psi(Q)$ for every d^* in the dual space. It is not difficult to establish that uniformly in $z \in \bar{\mathbb{R}}$ (the extended real line),

$$\mathbb{F}_n(z) = F_n(z) - \frac{1}{n}\sum_{j=1}^n \mathbf{1}[Z_j \le z]u(Z_j)^\top \zeta_n + o_p(n^{-1/2}). \tag{5.2}$$

By the Dvoretsky-Kiefer-Wolfowits inequality,

$$\sup_n P(\|\sqrt{n}(F_n - F)\|_\infty > M) \le 2\exp(-2M^2) \to 0, \quad M \to \infty.$$

This shows that the empirical process (EP) $\sqrt{n}(F_n - F)$ is asymptotically tight (bounded in probability). Noticing $\zeta_n = O_p(n^{-1/2})$, it can be seen that the second term on the right side of (5.2) hence $\sqrt{n}(\mathbb{F}_n - F)$ is also asymptotically tight. Clearly the cdf $F(z)$ as a functional of the probability P is differentiable (as a statistical parameter on the model) with derivative $\mathbf{1}[Z \le z] - F(z)$. Also, the EL-weighted estimator $\mathbb{F}_n(z)$ is asymptotically efficient at Q for estimating the coordinatewise projection $F(z)$ of F. Thus an application of Lemma 25.49 in van der Vaart (2000) yields that $\sqrt{n}(\mathbb{F}_n - F)$ is asymptotically efficient at Q for estimating F. As a consequence, by the continuous mapping theorem, $\sqrt{n}\|\mathbb{F}_n - F\|_\infty$ converges in distribution. Summarizing our preceding discussion, we have proved the following theorem. For a fixed $z \in \mathbb{R}^r$, let $W_{\mathbf{1}}(z)$ be the matrix when $\psi(Z) = \mathbf{1}[Z \le z]$ and set

$$C_{\mathbf{1}}(z) = E(\mathbf{1}[Z \le z] \otimes u(Z)^\top), \quad \sigma_F^2(z) = F(z)(1 - F(z)) - C_{\mathbf{1}}(z)W_u^{-1}C_{\mathbf{1}}(z)^\top.$$

Theorem 5.2. *If $W_{\mathbf{1}}(z)$ is positive definite for fixed $z \in \mathbb{R}^r$, then the EL-weighted $\mathbb{F}_n(z)$ is the MELE for the cdf $F(z)$ and satisfies the expansion,*

$$\mathbb{F}_n(z) = F_n(z) - C_{\mathbf{1}}(z)W_u^{-1}\bar{u}_n + o_p(n^{-1/2}), \quad z \in \mathbb{R}^d.$$

Hence $\sqrt{n}(\mathbb{F}_n(z) - F(z)) \implies \mathscr{N}(0, \sigma_F^2(z))$. Furthermore, the EL-weighted EP $\sqrt{n}(\mathbb{F}_n - F)$, viewed as a process over $D[-\infty, \infty]$, is asymptotically efficient at Q for estimating F.

5.3 The MELEs for differentiable statistical functionals

Suppose that T is a functional $T : \mathscr{F} \to \mathbb{R}$, where $\mathscr{F}$ is a convex set of probability distribution functions on $\mathbb{R}^r$ including all point masses and F. Suppose also that T is Gâteaux differentiable at F with derivative L_F representable as an integral, i.e.,

$$L_F(G - F) = \frac{\partial}{\partial t}T(F + t(G - F))|_{t=0} = \int \psi_F(z)\, dG(z), \quad G \in \mathscr{F}, \tag{5.3}$$

where necessarily $\int \psi_F(z)\, dF(z) = 0$. Suppose we have available side information expressed by $E(u(Z)) = 0$ for some known function u. We are interested in efficient estimation of $T(F)$. Naturally, we estimate the cdf F by the EL-weighted cdf $\mathbb{F}_n$ and wish that the plug-in estimator $T(\mathbb{F}_n)$ of $T(F)$ is efficient under suitable conditions. This is pursued next. To this end, we take $G = \mathbb{F}_n$, and obtain the following stochastic expansion,

$$T(\mathbb{F}_n) - T(F) = \int \psi_F(z)\, d\mathbb{F}_n(z) = \sum_{j=1}^{n} \pi_{nj}\psi_F(Z_j) + R_n,$$

where R_n is the reminder such that $R_n = o_p(L_F(\mathbb{F}_n - F))$. Suppose the matrix $W(\psi_F)$ when $\psi = \psi_F$ is (componentwise) finite and positive definite, so that (K1) is met and $\sigma_F^2 := \mathrm{Var}(\psi_F(Z)) - C(\psi_F)W_u^{-1}C(\psi_F)^\top$ is positive definite. Thus one can derive

$$\sum_{j=1}^{n} \pi_{nj}\psi_F(Z_j) = \frac{1}{n}\sum_{j=1}^{n}\Big(\psi_F(Z_j) - C(\psi_F)W_u^{-1}u(Z_j)\Big) + o_p(n^{-1/2}), \tag{5.4}$$

where $C(\psi_F) = E(\psi_F(Z) \otimes u^\top(Z))$. Thus,

$$\sqrt{n}(T(\mathbb{F}_n) - T(F)) \implies \mathscr{N}(0, \sigma_F^2), \tag{5.5}$$

if R_n satisfies

$$R_n = o_p(n^{-1/2}). \tag{5.6}$$

Sufficient conditions for these to hold can be found e.g. in Serfling (1980) (page 293). Summarizing the above discussion, we have the following result about plug-in estimators with the proof omitted.

Theorem 5.3. *Suppose that T is a functional from $\mathscr{F}$ to $\mathbb{R}$ which is Gâteaux differentiable such that (5.3) holds. Suppose $W(\psi_F)$ is positive definite. If the remainder R_n satisfies (5.6), then $T(\mathbb{F}_n)$ satisfies (5.5). Furthermore, if $T(G)$ is Hadamard differentiable at F in $D[-\infty, \infty]$, then $T(\mathbb{F}_n)$ is asymptotically efficient at F for estimating $T(F)$.*

VON-MISES FUNCTIONALS. We shall not pursue general considerations here but focus on the simplest von Mises functional:

$$T(G) = \iint \omega(x, y)\, dG(x)dG(y), \quad G \in \mathscr{F}, \tag{5.7}$$

where $w(x, y) = w(y, x), x, y \in \mathbb{R}$ and T is such that $T(G)$ is well defined. The usual empirical estimator of $T(F)$ is the plug-in estimator given by

$$T(F_n) = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} \omega(Z_i, Z_j).$$

This is related to the U-statistic of order 2 defined by

$$U_n(\omega) = \binom{n}{2}^{-1} \sum_{i<j} \omega(Z_i, Z_j)$$

by the relationship

$$U_n(\omega) = \frac{n}{n-1}T(F_n) - \frac{1}{n(n-1)}\sum_{i=1}^{n} \omega(Z_i, Z_j).$$

Thus their asymptotic behaviors are equivalent under suitable conditions. Naturally, we would like to replace F_n with the EL-weighted $\mathbb{F}_n$ and estimate $T(F)$ by $T(\mathbb{F}_n)$ given by

$$T(\mathbb{F}_n) = \sum_{i,j=1}^{n} \pi_{ni}\pi_{nj}\omega(Z_i, Z_j) = \frac{1}{n^2}\sum_{i,j=1}^{n} \frac{\omega(Z_i, Z_j)}{(1+\zeta_n^\top u(Z_i))(1+\zeta_n^\top u(Z_j))}.$$

Suppose ω satisfies

$$\int \omega^2(x,x)\, dF(x)dF(x) + \int \omega^2(x,y)\, dF(x)dF(y) < \infty. \tag{5.8}$$

Then $T(G)$ is Gâteaux differentiable such that (5.3) and (5.6) are met with

$$\varphi_F(x) = 2\left\{\int \omega(x,y)\, dF(y) - T(F)\right\}.$$

It is well known that if (5.8) holds then

$$\sqrt{n}(T(F_n) - T(F)) \implies \mathscr{N}(0, v_F^2),$$

where $v_F^2 = 4\left\{\int\left\{\int \omega(x,y)\, dF(y)\right\}^2 dF(x) - T^2(F)\right\}$. Suppose the matrix $W(\varphi_F)$ when $\psi = \varphi_F$ is finite and positive definite. In a similar fashion to the above discussion, we derive

$$\sqrt{n}(T(\mathbb{F}_n) - T(F)) \implies \mathscr{N}(0, \sigma_F^2), \tag{5.9}$$

where $\sigma_F^2 = v_F^2 - C(\varphi_F)W_u^{-1}C(\varphi_F)^\top$.

Theorem 5.4. *Let T be the functional defined in (5.7). Suppose that (5.8) holds such that $W(\varphi_F)$ is positive definite. Then $T(\mathbb{F}_n)$ satisfies (5.9). If, furthermore,*

$$\sup_{G\in\mathscr{F}} \left|\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \omega(x,y)\, dG(x)dG(y)\right| < \infty, \tag{5.10}$$

then $T(\mathbb{F}_n)$ is asymptotically efficient at F for estimating $T(F)$.

Obviously, if ω is bounded then (5.10) is satisfied.

L-, M-, R-ESTIMATORS AND RANK STATISTICS. Let F be a cdf on $\mathbb{R}$ and J be a function on $[0,1]$. An L-functional is defined as

$$T(G) = \int zJ(G(z))\, dG(z), \quad G \in \mathscr{F}. \tag{5.11}$$

The usual L-estimator of $T(F)$ is $T(F_n)$. Naturally, we would replace the ecdf F_n by the EL-weighted cdf $\mathbb{F}_n$ and estimate $T(F)$ by the plug-in estimator $T(\mathbb{F}_n)$. For an L-estimator T, one derives (details can be found on page 297 in Serfling (1980))

$$T(G) - T(F) = \int \phi_F(z)\, d(G-F)(z) + R(G,F),$$

where

$$\phi_F(x) = -\int (\delta_x - F)(z)J(F(z))\, dz,$$

$$R(G,F) = -\int W_G(z)(G(z) - F(z))\, dz,$$

and

$$W_G(z) = \begin{cases} (G(z) - F(z))^{-1} \int_{F(z)}^{G(z)} J(t)\, dt - J(F(z)) & G(z) \neq F(z) \\ 0 & G(z) = F(z). \end{cases}$$

Thus by Theorem 5.3, Theorem 5.6 in Serfling (1980) and in view of the fact T is also Hadamard differentiable, we obtain the following result. Let $W(\phi_F)$ denote the matrix when $\psi = \phi_F$ and set $\sigma_F^2 = \mathrm{Var}(\phi_F(Z)) - C(\phi_F)W_u^{-1}C(\phi_F)^\top$.

Theorem 5.5. *Let T be an L-functional defined in (5.11). Suppose that J is bounded, $J(t) = 0$ when $t \in [0, \alpha] \cup [\beta, 1]$ for some $\alpha < \beta$, and that the set $\{z : J \text{ is discontinuous at } F(z)\}$ has Lebesgue measure zero. Assume $W(\phi_F)$ is positive definite. Then $\sqrt{n}(T(\mathbb{F}_n) - T(F)) \implies \mathscr{N}(0, \sigma_F^2)$, and $T(\mathbb{F}_n)$ is asymptotically efficient at F for estimating $T(F)$.*

REFERENCES

Bickel, P.J., Klaassen, C.A.J., Ritov, Y. and Wellner, J.A. (1993), "Efficient and Adaptive Estimation in Semiparametric Models," Baltimore, USA: Johns Hopkins Univ. Press.

Bickel, P. J., Ritov, Y. and Wellner, J.A. (1991), "Efficient estimation of linear functionals of a probability measure P with known marginal distributions," *Ann. Statist.*, 19, 1316–1346.

Bravo, F. (2010), "Efficient M-estimators with auxiliary information," *J. Statist. Plan. Infer.*, 140, 3326–3342.

Chatterjee, N. and Carroll R.J. (2005), "Semiparametric maximum likelihood estimation exploiting gene-environment independence in case-control studies," *Biometrika*, 92, 399–418.

Deming, W.E. and Stephan, F.F. (1940), "On a least squares adjustment of a sampled frequency table when the expected marginal totals are known," *Ann. Math. Statist.*, 11, 423–444.

Haberman, S. J. (1984), "Adjustment by minimum discriminant information," *Ann. Statist.*, 12, 971-988.

Hellerstein, J. and Imbens, G. W. (1999), "Imposing moment restrictions from auxiliary data by weighting," *Review of Economics and Statistics*, 81, 1–14.

Imbens, G. W. and Lancaster, T. (1994). Combining Micro and Macro Data in Microeconometric Models. Review of Economic Studies 61, 655–680.

Ireland, C. T. and Kullback, S. (1968), "Contingency tables with given marginals," *Biometrika*, 55, 179–188.

Koltchinskii, V.I. (1997), "M-estimation, convexity and quantiles," *Ann. Statist.*, 25, 435–477.

Owen, A. (1990), "Empirical likelihood ratio confidence regions," *Ann. Statist.*, 18, 90–120.

Owen, A. (2001), *Empirical Likelihood*, London, UK: Chapman & Hall/CRC.

Peng, H. and Schick, A. (2013), "Maximum empirical likelihood estimation and related topics," The *Manuscript* can be found at the URL: http://www.math.binghamton.edu/anton/preprint.html.

Peng, H. and Schick, A. (2012), "An empirical likelihood approach of goodness of fit testing," *Bernoulli*, 19, 954–981

Peng, H. and Schick, A. (2005), "Efficient estimation of linear functionals of a bivariate distribution with equal, but unknown marginals: the least-squares approach," *J. Multiv. Anal.*, 95, 385–409.

Qin, J. and Lawless, J. (1994), "Empirical likelihood and general estimating equations," *Ann. Statist.*, 22, 30–325.

Serfling, R.J. (1980), *Approximation Theorems of Mathematical Statistics*, John Wiley & Sons, Inc.

Tang, C. Y. and Leng, C. (2012), "An empirical likelihood approach to quantile regression with auxiliary information," *Statist. & Probabil. Lett.* 82, 29–36.

Vaart, van der A.W. (2000), *Asymptotic Statistics*, Cambridge, UK: Cambridge University Press.

Vitale, R.A. (1979), "Regression with given marginals," *Ann. Statist.*, 7, 653 – 658.

Yuan, A., He, W., Wang, B. and Qin, G. (2012), "U-statistics with side information," *J. Multiv. Analy.*, 111, 20 – 38.

Zhang, B. (1995), "M-estimation and quantile estimation in the presence of auxiliary information," *J. Statist. Plan. Infer.*, 44, 77 – 94.

Zhang, B. (1997), "Quantile processes in the presence of auxiliary information," *Ann. Inst. Statist. Math.*, 49, 35 – 55.