Mean Empirical Likelihood

Wei Liang,

School of Mathematical Sciences, Xiamen University, Xiamen, China.

Hongsheng Dai,

Department of Mathematical Sciences, University of Essex, Colchester, UK.

Shuyuan He,

School of Mathematical Sciences, Capital Normal University, Beijing, China.

Abstract

Empirical likelihood methods are widely used in different settings to construct the confidence regions for parameters which satisfy the moment constraints. However, the empirical likelihood ratio confidence regions may have poor accuracy, especially for small sample sizes and multi-dimensional situations. In this paper, we propose a novel Mean Empirical Likelihood (MEL) method. This new method constructs a new pseudo dataset using the means of observation values to define the empirical likelihood ratio and we prove that this MEL ratio satisfies the Wilks' theorem. Simulations with different examples are given to assess its finite sample performance, which shows that the confidence regions constructed by Mean Empirical Likelihood is much more accurate than that of the other Empirical Likelihood methods.

Keywords: Confidence interval; Empirical likelihood; Exponentially tilted likelihood; Two sample comparison

1 Introduction

Empirical likelihood (EL) method proposed by Owen (1990) is a very powerful tool in non-parametric and semi-parametric statistics Qin and Lawless (1994); Newey and

Smith (2004). In particular, the confidence regions based on EL method are more appealing than those constructed based on asymptotic normality; not requiring the calculation of variance estimates, providing natural shape for the confidence regions and so on.

Suppose that we have independent and identically distributed random vectors $\boldsymbol{X}_1$, $\boldsymbol{X}_2, \cdots, \boldsymbol{X}_n$, with an unknown distribution function $F(\boldsymbol{x})$. We are interested in the estimation problem for a d-dimensional parameter $\boldsymbol{\theta} = \boldsymbol{\theta}(F)$. The true parameter value $\boldsymbol{\theta}_0$ is a unique solution of a system of equations $\mathbf{E}\,\boldsymbol{g}(\boldsymbol{X}, \boldsymbol{\theta}) = \mathbf{0}$ for some d-dimensional function $\boldsymbol{g}$. The original Empirical Likelihood (OEL) is defined as

$$\mathcal{R}^O(\boldsymbol{\theta}) = \sup \left\{ \prod_{i=1}^n np_i \Big| \sum_{i=1}^n p_i \boldsymbol{g}(\boldsymbol{X}_i, \boldsymbol{\theta}) = 0, \sum_{i=1}^n p_i = 1,\, p_i \geq 0,\, i = 1,\, 2,\, \cdots,\, n \right\}. \quad (1.1)$$

Assume $\boldsymbol{g}(\boldsymbol{X}, \boldsymbol{\theta}_0)$ has a finite covariance matrix of rank $d > 0$, Owen (1990) proved that

$$\mathcal{L}^O(\boldsymbol{\theta}_0) = -2 \log \mathcal{R}^O(\boldsymbol{\theta}_0) \to \chi^2(d), \text{ in dist.} \quad (1.2)$$

Therefore, the $(1-\alpha)$ confidence region can be constructed as $I^O = \{\boldsymbol{\theta} : \mathcal{L}^O(\boldsymbol{\theta}) < \chi^2_\alpha(d)\}$, where $\chi^2_\alpha(d)$ is such that $P(\chi^2(d) \leq \chi^2_\alpha(d)) = 1 - \alpha$.

Although EL method has found its application in many statistical areas, its finite sample properties may not work well because of low precision of χ^2 approximation. Hall and La Scala (1990), DiCiccio et al. (1991) and Tsao (2004) showed that empirical likelihood ratio confidence regions could have poor accuracy, especially in small sample and multi-dimensional situations. Many methods have been proposed to improve the performance of the EL approach in the literature. For parameters defined by standard estimating equations, the Bartlett correction Empirical Likelihood (BEL) (DiCiccio et al., 1991) achieves the second order accuracy, which is substantially more accurate than the original EL approach. An alternative method is to add a pseudo-observation to the sample. This leads to the adjusted Empirical Likelihood (AEL) (Chen et al., 2008) and it also achieves the second order accuracy. Recently, Tsao and Wu (2013) developed an extended Empirical Likelihood method (EEL), attaining the second order accuracy as well. However, all the above-mentioned methods require the calculation of the Bartlett correction constant, which has no analytical formula since it depends on the moments of

$\boldsymbol{g}(\boldsymbol{X}, \boldsymbol{\theta})$. In practice, using a $\sqrt{n}$-consistent estimator for Bartlett correction constant is feasible, but it may be difficult to calculate the estimator in certain practical scenarios (Liu and Chen, 2010). Apart from the practical estimation challenge for Bartlett correction constant, Jing (1996) proved that exponentially tilted likelihood is actually not Bartlett correctable and all existing methods only have the first order accuracy. Therefore, all the above practical and theoretical challenges motivate us to search new EL approaches.

In this paper, we will present a new method, named Mean Empirical Likelihood (MEL). It constructs an empirical likelihood function based on a set of pseudo data and it is easy to compute (not requiring the calculation of the Bartlett correction constant). The large sample properties of MEL are presented. This new MEL is particularly more important for exponentially tilted likelihood, where Bartlett correction is not available. In the simulation studies, we find that the confidence intervals constructed by MEL is much more accurate than those found by the other Empirical Likelihood methods with second order accuracy, such as BEL and AEL. In particular, MEL outperforms all other methods for heavy-tail or highly-skewed distributions and for exponentially tilted likelihood.

This paper is constructed as follows. In Section 2, we present the MEL methodologies in different settings: standard estimating equation framework, two sample mean comparison problem, exponentially tilted likelihood and generalised empirical likelihood framework, with all theoretical proofs provided in Appendix. Simulation studies are presented in Section 3 and they demonstrate that MEL outperforms all other existing methods. Section 4 provides a real data analysis and the paper concludes with a discussion in Section 5.

2 Methodology

2.1 Mean Empirical Likelihood for standard Estimating Equations

We here follow notations in the previous section and for simplicity we denote $\boldsymbol{V}_i(\boldsymbol{\theta}) = \boldsymbol{g}(\boldsymbol{X}_i, \boldsymbol{\theta})$, $i = 1$, 2, ..., n and further denote the pairwise-mean dataset as follows,

$$\mathcal{W} = \left\{ \frac{\boldsymbol{V}_i(\boldsymbol{\theta}) + \boldsymbol{V}_j(\boldsymbol{\theta})}{2} : 1 \le i \le j \le n \right\}, \tag{2.1}$$

which can also be written as $\mathcal{W} = \{\boldsymbol{W}_1(\boldsymbol{\theta}), \boldsymbol{W}_2(\boldsymbol{\theta}), \cdots, \boldsymbol{W}_N(\boldsymbol{\theta})\}$ with $N = n(n+1)/2$. Based on this pairwise-mean dataset, the empirical likelihood ratio for $\boldsymbol{\theta}$ is defined as

$$\mathcal{R}^M(\boldsymbol{\theta}) = \sup \left\{ \prod_{k=1}^N Np_k \middle| \sum_{k=1}^N p_k \boldsymbol{W}_k(\boldsymbol{\theta}) = \mathbf{0}, \sum_{k=1}^N p_k = 1, p_k \ge 0, k = 1, 2, \cdots, N \right\},$$

which is named as *mean empirical likelihood* ratio. It follows that

$$\mathcal{R}^M(\boldsymbol{\theta}) = \prod_{k=1}^N Np_k = \prod_{k=1}^N \frac{1}{1 + \boldsymbol{\lambda}' \boldsymbol{W}_k(\boldsymbol{\theta})},$$

where $\boldsymbol{\lambda}$ satisfying

$$\frac{1}{N} \sum_{k=1}^N \frac{\boldsymbol{W}_k(\boldsymbol{\theta})}{1 + \boldsymbol{\lambda}' \boldsymbol{W}_k(\boldsymbol{\theta})} = \mathbf{0},$$

and

$$p_k = \frac{1}{N(1 + \boldsymbol{\lambda}' \boldsymbol{W}_k(\boldsymbol{\theta}))}.$$

Denote $\boldsymbol{\theta}_0$ as the unknown true parameter value. Then the mean empirical log-likelihood ratio is given by

$$\mathcal{L}^M(\boldsymbol{\theta}_0) = -2 \log \mathcal{R}^M(\boldsymbol{\theta}_0)/(n+1).$$

Now we have the following main theorem.

Theorem 2.1. *Under the conditions that* $\texttt{Cov}(\boldsymbol{V}_i(\boldsymbol{\theta}_0)) = \boldsymbol{\Sigma}$ *exists and* $\texttt{rank}(\boldsymbol{\Sigma}) = d$*, we have*

$$\mathcal{L}^M(\boldsymbol{\theta}_0) \to \chi^2(d), \quad \textit{in dist.}$$

Proof. See Appendix A. □

Following Theorem 2.1, a confidence region for the parameter $\boldsymbol{\theta}$ with asymptotic coverage probability $1-\alpha$ can be defined as

$$I^M = \{\boldsymbol{\theta} : \mathcal{L}^M(\boldsymbol{\theta}) \le \chi^2_\alpha(d)\}.$$

2.2 Mean Empirical Likelihood for Two-Sample Comparison

In this section, we consider applying the Mean Empirical Likelihood idea to a two-sample problem. To be specific, let $\boldsymbol{U}_1, \boldsymbol{U}_2, \cdots, \boldsymbol{U}_{n_1}$ be a d-dimensional i.i.d. random sample from distribution F, and $\boldsymbol{V}_1, \boldsymbol{V}_2, \cdots, \boldsymbol{V}_{n_2}$ be a d-dimensional i.i.d. random sample from distribution G. We want to construct confidence regions for $\boldsymbol{\theta} \in \mathbb{R}^d$,

$$\boldsymbol{\theta} = \int \boldsymbol{g}(\boldsymbol{u})\,\mathrm{d}F(\boldsymbol{u}) - \int \boldsymbol{g}(\boldsymbol{v})\,\mathrm{d}G(\boldsymbol{v}),$$

where $\boldsymbol{g}(\cdot)$ is a known d-dimensional function. For instance, $\boldsymbol{g}(\boldsymbol{x}) = \boldsymbol{x}$, then $\boldsymbol{\theta}$ stands for the difference between two distributions.

Let $\boldsymbol{\theta}_0$ be the true value, $\boldsymbol{X}_i = \boldsymbol{g}(\boldsymbol{U}_i)$ and $\boldsymbol{Y}_i = \boldsymbol{g}(\boldsymbol{V}_i) - \boldsymbol{\theta}_0$. Denote the elements in the dataset $\{(\boldsymbol{X}_i + \boldsymbol{X}_j)/2,\ 1 \le i \le j \le n_1\}$ by $\boldsymbol{W}^X_1, \boldsymbol{W}^X_2, \cdots, \boldsymbol{W}^X_{N_1}$, where $N_1 = n_1(n_1+1)/2$, and the elements in the dataset $\{(\boldsymbol{Y}_i + \boldsymbol{Y}_j)/2,\ 1 \le i \le j \le n_2\}$ by $\boldsymbol{W}^Y_1, \boldsymbol{W}^Y_2, \cdots, \boldsymbol{W}^Y_{N_2}$, where $N_2 = n_2(n_2+1)/2$. Then the mean empirical likelihood for $\boldsymbol{\theta}$, evaluated at $\boldsymbol{\theta}_0$, is defined as

$$\begin{aligned} l^M_2(\boldsymbol{\theta}_0) = \sup\Big\{ \prod_{s=1}^{N_1} p_s \prod_{t=1}^{N_2} q_t \ \ \Big| \ \ & \sum_{s=1}^{N_1} p_s \left(\boldsymbol{W}^X_s - \boldsymbol{\mu}\right) = \boldsymbol{0}, \sum_{t=1}^{N_2} q_t \left(\boldsymbol{W}^Y_t - \boldsymbol{\mu}\right) = \boldsymbol{0}, \\ & \sum_{s=1}^{N_1} p_s = 1,\ \sum_{t=1}^{N_2} q_t = 1,\ p_s \ge 0,\ q_t \ge 0 \Big\}, \end{aligned}$$

and the mean empirical likelihood ratio for $\boldsymbol{\theta}_0$ is

$$\mathcal{R}^M_2(\boldsymbol{\theta}_0) = \frac{l^M_2(\boldsymbol{\theta}_0)}{l^M_2(\hat{\boldsymbol{\theta}})},$$

where $l^M_2(\hat{\boldsymbol{\theta}}) = N_1^{-N_1} N_2^{-N_2}$.

Let $N = N_1 + N_2$, $\delta = N_1/N$. The Lagrange multiplier method leads to

$$p_s = \frac{1}{N\delta}\frac{1}{1+\delta^{-1}\boldsymbol{\lambda}_1'(\boldsymbol{W}^X_s - \boldsymbol{\mu})}, \qquad q_t = \frac{1}{N(1-\delta)}\frac{1}{1+(1-\delta)^{-1}\boldsymbol{\lambda}_2'(\boldsymbol{W}^Y_t - \boldsymbol{\mu})},$$

and the maximum log-likelihood ratio

$$\log(\mathcal{R}_2^M(\boldsymbol{\theta}_0)) = -\left[\sum_{s=1}^{N_1} \ln\left(1+\delta^{-1}\boldsymbol{\lambda}_1'(\boldsymbol{W}_s^X-\boldsymbol{\mu})\right) + \sum_{t=1}^{N_2} \ln\left(1+(1-\delta)^{-1}\boldsymbol{\lambda}_2'(\boldsymbol{W}_t^Y-\boldsymbol{\mu})\right)\right],$$

where $(\boldsymbol{\lambda}_1,\ \boldsymbol{\lambda}_2,\ \boldsymbol{\mu})$ satisfies the following equations

$$\begin{aligned}
&\frac{1}{N\delta}\sum_{s=1}^{N_1}\frac{\boldsymbol{W}_s^X-\boldsymbol{\mu}}{1+\delta^{-1}\boldsymbol{\lambda}_1'(\boldsymbol{W}_s^X-\boldsymbol{\mu})} = \mathbf{0},\\
&\frac{1}{N(1-\delta)}\sum_{t=1}^{N_2}\frac{\boldsymbol{W}_t^Y-\boldsymbol{\mu}}{1+(1-\delta)^{-1}\boldsymbol{\lambda}_2'(\boldsymbol{W}_t^Y-\boldsymbol{\mu})} = \mathbf{0},\\
&\boldsymbol{\lambda}_1+\boldsymbol{\lambda}_2 = \mathbf{0}.
\end{aligned}$$

Thus the corresponding mean empirical log-likelihood ratio is defined as

$$\mathcal{L}_2^M(\boldsymbol{\theta}_0) = -2\log(\mathcal{R}_2^M(\boldsymbol{\theta}_0))/n.$$

Theorem 2.2. *Let $n = n_1 + n_2$, $\Delta = n_1/n$. Assume that $\Delta \to \Delta_0 \in (0,\ 1)$ as $n \to \infty$, $\boldsymbol{Cov}(X) = \boldsymbol{\Sigma}_X$ and $\boldsymbol{Cov}(Y) = \boldsymbol{\Sigma}_Y$ exist, and $\boldsymbol{rank}(\boldsymbol{\Sigma}_X) = \boldsymbol{rank}(\boldsymbol{\Sigma}_Y) = d$. Then, mean empirical log-likelihood ratio $\mathcal{L}_2^M(\boldsymbol{\theta}_0)$ converges in distribution to a weighted sum of independent standard chi-square random variables, each with one degree of freedom and weight r_k. That is*

$$\mathcal{L}_2^M(\boldsymbol{\theta}_0) \to \sum_{k=1}^{d} r_k\eta_k, \quad \textit{in dist.}, \qquad \eta_k \sim \chi^2(1)$$

where r_k are the eigenvalues of $(\boldsymbol{R}^M)^{-1}\boldsymbol{R}$,

$$\boldsymbol{R} = \frac{1}{\Delta_0}\boldsymbol{\Sigma}_X + \frac{1}{1-\Delta_0}\boldsymbol{\Sigma}_Y, \qquad \boldsymbol{R}^M = \frac{1}{\Delta_0^2}\boldsymbol{\Sigma}_X + \frac{1}{(1-\Delta_0)^2}\boldsymbol{\Sigma}_Y.$$

Proof. See Appendix B. □

Theorem 2.2 is the MEL version of Wilks' theorem in the two-sample problem. For computational simplicity, we can consider the following adjusted statistic in practice. Let $r = d/\mathrm{tr}((\boldsymbol{R}^M)^{-1}\boldsymbol{R})$ with $\mathrm{tr}(\cdot)$ denoting the trace operator. Then, following Rao and Scott (1981) and Xue and Wang (2012), the distribution of $r\sum_{k=1}^{d} r_k\eta_k$ can be approximated by a standard chi-square distribution with d degrees of freedom $\chi^2(d)$.

Motivated by this approach, we now define an adjusted mean empirical log-likelihood $\hat{\mathcal{L}}_2^M(\boldsymbol{\theta}_0)$ whose asymptotic distribution is approximately $\chi^2(d)$,

$$\hat{\mathcal{L}}_2^M(\boldsymbol{\theta}_0) = \hat{r}\,\mathcal{L}_2^M(\boldsymbol{\theta}_0), \tag{2.2}$$

where $\hat{r} = d/\mathrm{tr}((\hat{\boldsymbol{R}}^M)^{-1}\hat{\boldsymbol{R}})$, $\hat{\boldsymbol{R}}^M$ and $\hat{\boldsymbol{R}}$ are the estimators of $\boldsymbol{R}^M$ and $\boldsymbol{R}$, respectively. Hence, a simple approach to construct an α-level confidence region for $\boldsymbol{\theta}$, based on (2.2), is

$$I_2^M = \{\boldsymbol{\theta} : \hat{\mathcal{L}}_2^M(\boldsymbol{\theta}) \le \chi_\alpha^2(d)\}.$$

2.3 Mean Empirical Likelihood for Exponentially Tilted Likelihood

Exponentially tilted (ET) likelihood is a useful nonparametric approach to evaluate estimates and confidence regions of parameters $\boldsymbol{\theta}$. In this section, we develop a Mean ET likelihood procedure for parameter estimation, and prove this likelihood ratio statistic is asymptotically distributed as the chi-squared distribution (the Wilks' theorem holds), which can be used to construct confidence regions of parameters of interest.

Suppose that $\boldsymbol{X}_1, \boldsymbol{X}_2, \cdots, \boldsymbol{X}_n$ are i.i.d. random vectors from an unknown distribution $F(\boldsymbol{x})$. Using the same notations as Section 2.1, i.e. $\boldsymbol{V}_i(\boldsymbol{\theta}) = \boldsymbol{g}(\boldsymbol{X}_i, \boldsymbol{\theta}), i = 1, 2, \cdots, n$, the ET likelihood for $\boldsymbol{\theta} \in \mathbb{R}^d$ can be defined as

$$\mathcal{H}(\boldsymbol{\theta}) = \sup\left\{-\prod_{i=1}^n w_i \log(w_i) : \sum_{i=1}^n w_i \boldsymbol{V}_i(\boldsymbol{\theta}) = \boldsymbol{0}, \sum_{i=1}^n w_i = 1,\, w_i \ge 0,\, i = 1,\, 2,\, \cdots,\, n\right\},$$

which is maximized at

$$w_i = \frac{\exp\{\boldsymbol{\lambda}'\boldsymbol{V}_i(\boldsymbol{\theta})\}}{\sum_{j=1}^n \exp\{\boldsymbol{\lambda}'\boldsymbol{V}_j(\boldsymbol{\theta})\}}.$$

Note that, the function $-\prod_{i=1}^n w_i \log(w_i)$ attains its maximum value $\log(n)$ at $w_i = n^{-1}$. Therefore existing methods use the empirical entropy difference

$$\Delta\mathcal{H}(\boldsymbol{\theta}) = \mathcal{H}(\boldsymbol{\theta}) - \log(n)$$

to derive the empirical confidence regions. Newey and Smith (2004) and Jaynes (1982) proposed two adjusted empirical entropy differences, respectively, the adjusted Newey-Smith empirical entropy difference:

$$\mathcal{T}_1(\boldsymbol{\theta}) \;=\; -2n\,\{\exp(\Delta\mathcal{H}(\boldsymbol{\theta})) - 1\},$$

and the adjusted Jaynes empirical entropy difference:

$$\mathcal{T}_2(\boldsymbol{\theta}) \;=\; -2n\,\Delta\mathcal{H}(\boldsymbol{\theta}).$$

Both $\mathcal{T}_1(\boldsymbol{\theta}_0)$ and $\mathcal{T}_2(\boldsymbol{\theta}_0)$ converge in distribution to a $\chi^2(d)$ distribution.

We can also extend the above empirical entropy difference idea by using the mean exponentially tilted likelihood for $\boldsymbol{\theta}$, i.e.

$$\mathcal{H}^M(\boldsymbol{\theta}) = \sup\left\{-\prod_{i\le j} w_{ij}\log(w_{ij}) : \sum_{i\le j} w_{ij}\left(\boldsymbol{V}_i(\boldsymbol{\theta}) + \boldsymbol{V}_j(\boldsymbol{\theta})\right) = \mathbf{0},\, \sum_{i\le j} w_{ij} = 1,\, w_{ij}\ge 0\right\}.$$

Under some regularity conditions, it is easily to show that the mean empirical entropy difference can be expressed as

$$\Delta\mathcal{H}^M(\boldsymbol{\theta}) = \mathcal{H}^M(\boldsymbol{\theta}) - \log(N) = \log\left\{\frac{1}{N}\sum_{i\le j}\exp\left\{-\boldsymbol{\lambda}'\left(\boldsymbol{V}_i(\boldsymbol{\theta}) + \boldsymbol{V}_j(\boldsymbol{\theta})\right)\right\}\right\},$$

where $N = n(n+1)/2$ and $\boldsymbol{\lambda}$ satisfies

$$\frac{1}{N}\sum_{i\le j}\left(\boldsymbol{V}_i(\boldsymbol{\theta}) + \boldsymbol{V}_j(\boldsymbol{\theta})\right)\exp\{-\boldsymbol{\lambda}'\left(\boldsymbol{V}_i(\boldsymbol{\theta}) + \boldsymbol{V}_j(\boldsymbol{\theta})\right)\} = \mathbf{0}.$$

By defining two adjusted mean empirical entropy differences

$$\begin{aligned}
\mathcal{T}_1^M(\boldsymbol{\theta}) &= -2N\{\exp\left(\Delta\mathcal{H}^M(\boldsymbol{\theta})\right) - 1\}/(n+1),\\
\mathcal{T}_2^M(\boldsymbol{\theta}) &= -2N\,\Delta\mathcal{H}^M(\boldsymbol{\theta})/(n+1),
\end{aligned}$$

we get the following theorem.

Theorem 2.3. *Assume $\boldsymbol{Cov}(\boldsymbol{V}_1(\boldsymbol{\theta}_0)) = \boldsymbol{\Sigma}$ exists, $\boldsymbol{rank}(\boldsymbol{\Sigma}) = d$. Then both mean empirical entropy differences $\mathcal{T}_1^M(\boldsymbol{\theta}_0)$ and $\mathcal{T}_2^M(\boldsymbol{\theta}_0)$ are asymptotically a chi-square random variable, that is*

$$\mathcal{T}_1^M(\boldsymbol{\theta}_0) \to \chi^2(d), \quad \textit{in dist.} \qquad \mathcal{T}_2^M(\boldsymbol{\theta}_0) \to \chi^2(d), \quad \textit{in dist.}$$

Proof. The proof is similar to the proof of Theorem 2.1 and therefore it is omitted. □

Based on Theorem 2.3, the α-level confidence region for $\boldsymbol{\theta}$ can be constructed by

$$\begin{aligned}
I_{4,1}^M &= \left\{\boldsymbol{\theta} :\, -n\left\{\exp\left(\Delta\mathcal{H}^M(\boldsymbol{\theta})\right) - 1\right\} \le \chi^2_\alpha(d)\right\},\\
I_{4,2}^M &= \left\{\boldsymbol{\theta} :\, -n\,\Delta\mathcal{H}^M(\boldsymbol{\theta}) \le \chi^2_\alpha(d)\right\}.
\end{aligned}$$

2.4 Generalized Mean Empirical Likelihood

In this subsection, we make some further extensions to generalized empirical likelihood(GEL) inference. Suppose that we have d-dimensional independent and identically distributed random vectors $\boldsymbol{X}_1,\ \boldsymbol{X}_2,\ \cdots,\boldsymbol{X}_n$ with an unknown distribution function $F(\boldsymbol{x})$. We are still interested in the estimation problem for a d-dimensional parameter $\boldsymbol{\theta} = \boldsymbol{\theta}(F)$. The estimating equations for $\boldsymbol{\theta}$ is $\mathbf{E}\,\boldsymbol{g}(\boldsymbol{X}, \boldsymbol{\theta}) = \mathbf{0}$, where $\boldsymbol{g}$ is an m-dimensional function with $m \geq d$.

If we denote $h(p)$ as a convex function of a scalar p, the minimum discrepancy(MD) estimators, first discussed by Corcoran (1998), can be calculated as

$$\hat{\boldsymbol{\theta}}_{\text{MD}} = \arg_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \min \sum_{i=1}^{n} h(p_i),$$

subject to

$$\sum_{i=1}^{n} p_i\, \boldsymbol{g}(\boldsymbol{X}_i,\, \boldsymbol{\theta}) = \mathbf{0}, \quad \sum_{i=1}^{n} p_i = 1.$$

Newey and Smith (2004) explained that for each MD estimators there is a dual GEL estimator when $h(p)$ is a member of Cressie and Read family of discrepancies in which

$$h(p) = \frac{(np)^{\gamma+1} - 1}{n\gamma(\gamma+1)}.$$

With these notations, we can introduce the GEL estimators as follows. Let

$$\rho(v) = -\frac{(1+\gamma v)^{(\gamma+1)/\gamma}}{\gamma+1}$$

be a function of a scalar v that is concave on its domain $\mathcal{V}$, an open interval containing 0. Let $\hat{\boldsymbol{\Lambda}}_n(\boldsymbol{\theta}) = \{\boldsymbol{\lambda} :\ \boldsymbol{\lambda}^T \boldsymbol{g}(\boldsymbol{X}_i,\, \boldsymbol{\theta}) \in \mathcal{V},\ i = 1,\, 2,\, \cdots,\, n\}$, then

$$\hat{\boldsymbol{\theta}}_{\text{GEL}} = \arg \min_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \sup_{\boldsymbol{\lambda}\in\hat{\boldsymbol{\Lambda}}_n(\boldsymbol{\theta})} \sum_{i=1}^{n} \rho(\boldsymbol{\lambda}^T \boldsymbol{g}(\boldsymbol{X}_i,\, \boldsymbol{\theta})).$$

From Theorem 2.2 in Newey and Smith (2004), $\hat{\boldsymbol{\lambda}}_{\text{GEL}} = \arg\sup_{\boldsymbol{\lambda}\in\hat{\boldsymbol{\Lambda}}_n(\boldsymbol{\theta})} \sum_{i=1}^{n} \rho(\boldsymbol{\lambda}^T \boldsymbol{g}(\boldsymbol{X}_i,\, \boldsymbol{\theta}))$ exists and $\hat{\boldsymbol{\theta}}_{\text{MD}} = \hat{\boldsymbol{\theta}}_{\text{GEL}}$.

EL is a special case with $h(p) = -\ln p,\ \rho(v) = \ln(1-v)$, and ET is another special case when $h(p) = p\ln p,\ \rho(v) = -e^v$.

Denote $\boldsymbol{V}_i(\boldsymbol{\theta}) = \boldsymbol{g}(\boldsymbol{X}_i,\, \boldsymbol{\theta}), i = 1, 2, \cdots, n$, and denote the pairwise-mean equation as $\boldsymbol{W}_k(\boldsymbol{\theta})$ with $k = 1,\, 2,\, \cdots,\, N = n(n+1)/2$. It is easy to see that

$$\hat{\boldsymbol{\Lambda}}_N(\boldsymbol{\theta}) = \{\boldsymbol{\lambda} :\ \boldsymbol{\lambda}^T \boldsymbol{W}_k(\boldsymbol{\theta}) \in \mathcal{V},\ k = 1,\, 2,\, \cdots,\, N\} = \hat{\boldsymbol{\Lambda}}_n(\boldsymbol{\theta}).$$

Then the Generalized Mean Empirical Likelihood (GMEL) estimator can be defined as

$$\hat{\boldsymbol{\theta}}_{\text{GMEL}} = \arg\min_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \sup_{\boldsymbol{\lambda}\in\hat{\boldsymbol{\Lambda}}_n(\boldsymbol{\theta})} \sum_{k=1}^{N} \rho(\boldsymbol{\lambda}^T \boldsymbol{W}_k(\boldsymbol{\theta})).$$

For simplicity, we denote $\hat{\boldsymbol{\theta}}_{\text{GMEL}}$ as $\hat{\boldsymbol{\theta}}$, and the corresponding $\hat{\boldsymbol{\lambda}}_{\text{GMEL}}$ as $\hat{\boldsymbol{\lambda}}$. Assume that $\boldsymbol{\theta}_0$ is the true value of the parameter $\boldsymbol{\theta}$, and

$$\boldsymbol{\Sigma}_0 = \mathbf{E}\left(\boldsymbol{g}(\boldsymbol{X},\, \boldsymbol{\theta}_0)\, \boldsymbol{g}^T(\boldsymbol{X},\, \boldsymbol{\theta}_0)\right), \qquad \boldsymbol{\Gamma}_0 = \mathbf{E}\left(\frac{\partial \boldsymbol{g}(\boldsymbol{X},\, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}\right).$$

Under the following assumptions we have the main result Theorem 2.4.

A1. $\boldsymbol{\theta}_0 \in \boldsymbol{\Theta}$ is the unique solution to $\mathbf{E}\,\boldsymbol{g}(\boldsymbol{X},\, \boldsymbol{\theta}) = \mathbf{0}$.

A2. $\boldsymbol{\Theta}$ is compact.

A3. $\boldsymbol{g}(\boldsymbol{x},\, \boldsymbol{\theta})$ is continuous at each $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, continuously differentiable in a neighborhood $\mathcal{N}_{\boldsymbol{\theta}_0}$ of $\boldsymbol{\theta}_0$.

A4. For some $\alpha > 2$,

$$\mathbf{E}\left(\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} ||\boldsymbol{g}(\boldsymbol{X},\, \boldsymbol{\theta})||^{\alpha}\right) < \infty, \qquad \mathbf{E}\left(\sup_{\boldsymbol{\theta}\in\mathcal{N}_{\boldsymbol{\theta}_0}} ||\partial \boldsymbol{g}(\boldsymbol{X},\, \boldsymbol{\theta})/\partial \boldsymbol{\theta}^T||\right) < \infty.$$

A5. $\boldsymbol{\Sigma}_0$ is nonsingular and $\text{rank}(\boldsymbol{\Gamma}_0) = d$ (definitions see (C.2)).

Theorem 2.4. *Under the assumptions A1-A5, let $\rho_0 = \rho(0)$, we have*

$$n\left(\frac{1}{N}\sum_{k=1}^{N} \rho(\hat{\boldsymbol{\lambda}}^T \boldsymbol{W}_k(\hat{\boldsymbol{\theta}})) - \rho_0\right) \to \chi^2_{(m-d)}, \quad \textit{in dist.}$$

Proof. See Appendix C. □

If we consider $\rho(v) = \ln(1-v)$, then Theorem 2.4 is an MEL version of Corollary 4 in Qin and Lawless (1994).

3 Simulation

We present four simulation studies in this section, which correspond to the finite sample performance of different MEL methods in Section 2. We will compare the MEL confidence regions with AEL and BEL results. For one dimensional estimating equation $g(X, \theta)$, we use the theoretical Bartlett correction b and $a = b/2$ for constructing the BEL and AEL confidence regions, respectively. In Section 3.1 and 3.2, we will also compare our method with EEL confidence regions, which are built with the first order of EEL expansion factor Tsao and Wu (2013).

3.1 Example I: Single Parameter Model

Suppose that $g(x,\, \theta) = x - \theta$ is the estimating equation for θ. We aim to compare the different confidence intervals derived from MEL(I^M), OEL(I^O), BEL(I^B), AEL(I^A) and EEL(I^E), for a given sample size n. We consider different scenarios by generate observations $X_1, X_2, \cdots, X_n$ from Norm$(0, 1)$, t(5), $\chi^2(1)$ and LogNorm$(0, 1)$, respectively. Based on $10,000$ replicates, the coverage proportions were calculated. The simulation results are summarized in Table 1.

We noticed that MEL confidence intervals are easy to calculate, just as the first order method of OEL. However, the second order methods BEL and AEL need to evaluate the theoretical Bartlett correction factors, which is difficult to estimate. The method EEL, the most accurate method in Tsao and Wu (2013), needs to solve an equation to obtain the extended parameter. Therefore, from the aspects of computational efficiency, MEL method is recommended.

(1) Comparison for different sample sizes. For small sample size n, MEL is much better than OEL, and it is a little better than EEL. All of the coverage probabilities are close to the nominal levels when the sample size increase.

(2) Comparison for different methods. In fact, for most cases, MEL has the similar coverage probabilities as EEL. However, BEL and AEL use the theoretical Bartlett correction factors, which is not available in practice. In practice, we have to estimate the Bartlett correction factor, therefore we cannot achieve such good performance for BEL and AEL.

Table 1: Coverage probabilities for the mean parameter.

		OEL	MEL	BEL	AEL	EEL	OEL	MEL	BEL	AEL	EEL
n	$1-\alpha$	Norm(0, 1)					t (5)				
20	0.90	0.8805	0.9039	0.8935	0.8942	**0.9018**	0.8621	0.8883	**0.9041**	0.9109	0.8891
	0.95	0.9327	**0.9536**	0.9430	0.9434	0.9538	0.9221	0.9419	**0.9487**	0.9556	0.9447
	0.99	0.9804	**0.9904**	0.9834	0.9841	0.9912	0.9763	0.9881	0.9876	0.9953	**0.9908**
40	0.90	0.8908	0.9052	**0.8979**	0.8979	0.9025	0.8801	0.8957	**0.9005**	0.9005	0.8932
	0.95	0.9433	0.9538	0.9476	**0.9477**	0.9526	0.9357	0.9472	**0.9474**	0.9474	0.9458
	0.99	0.9858	**0.9917**	0.9872	0.9871	0.9917	0.9832	**0.9895**	0.9873	0.9873	0.9895
100	0.90	0.8969	0.9021	**0.8998**	0.8998	0.9017	0.8924	0.8975	**0.9000**	0.9000	0.8968
	0.95	0.9481	0.9527	**0.9501**	0.9501	0.9524	0.9449	**0.9501**	0.9503	0.9503	0.9499
	0.99	**0.9900**	0.9919	0.9906	0.9906	0.9915	0.9886	0.9918	**0.9904**	0.9904	0.9910
n	$1-\alpha$	$\chi^2(5)$					$\text{LogNorm}(0,1)$				
20	0.90	0.8273	0.8538	0.8745	**0.8840**	0.8511	0.8064	**0.8359**	1.0000	1.0000	0.8323
	0.95	0.8858	0.9156	0.9256	**0.9359**	0.9138	0.8680	**0.8957**	1.0000	1.0000	0.8941
	0.99	0.9512	0.9692	0.9667	**0.9841**	0.9671	0.9373	**0.9618**	1.0000	1.0000	0.9599
40	0.90	0.8643	0.8828	0.8874	**0.8890**	0.8753	0.8489	**0.8671**	1.0000	1.0000	0.8601
	0.95	0.9228	**0.9397**	0.9370	0.9383	0.9333	0.9027	**0.9195**	1.0000	1.0000	0.9153
	0.99	0.9745	**0.9852**	0.9808	0.9818	0.9815	0.9615	**0.9759**	1.0000	1.0000	0.9712
100	0.90	0.8919	0.8988	0.8994	**0.8995**	0.8957	0.8695	**0.8795**	0.9413	0.9627	0.8755
	0.95	0.9428	**0.9502**	0.9486	0.9491	0.9473	0.9231	**0.9335**	0.9728	0.9923	0.9292
	0.99	0.9861	0.9913	0.9887	0.9889	**0.9892**	0.9784	**0.9854**	0.9948	1.0000	0.9819

The boldface results are the most accurate coverage probabilities among EL methods.

Table 2: Numerical characteristics of different EL methods.

		t(5)					LogNorm$(0, 1)$				
	n	OEL	MEL	BEL	AEL	EEL	OEL	MEL	BEL	AEL	EEL
mean	20	1.2379	1.0670	0.7635	0.6381	1.0597	2.4854	1.8078	-1.8998	0.0062	1.5666
	40	1.1074	1.0193	0.8952	0.8607	1.0288	1.5087	1.3013	0.1777	0.2345	1.3590
	100	1.0577	1.0190	0.9766	0.9734	1.0264	1.2132	1.1301	0.7851	0.7008	1.1721
median	20	0.5410	0.5089	0.3337	0.3459	0.5269	0.6528	0.6078	-0.4990	0.0057	0.6310
	40	0.4958	0.4803	0.4008	0.4036	0.4898	0.6135	0.5911	0.0723	0.1668	0.6046
	100	0.4613	0.4553	0.4259	0.4262	0.4592	0.5334	0.5261	0.3452	0.3562	0.5307
variance	20	3.9691	2.4035	1.5098	0.5643	2.0427	314.4747	83.2156	187.7386	0.0000	6.6908
	40	2.6470	1.9846	1.7298	1.3644	1.9754	7.0534	3.8520	0.0979	0.0492	4.3292
	100	2.3290	2.0207	1.9856	1.9498	2.0743	3.2536	2.4522	1.3625	0.7679	2.8476

(3) Comparison for different distributions. Under Norm, t and χ^2 distributions, about a third of the cases, MEL performs best among these five EL methods. Under logNorm distribution, AEL and BEL give coverage probabilities larger than nominal, but MEL and EEL perform much better. Note that for heavy tail distributions, the large kurtosis leads to a large negative theoretical Bartlett correction b and further leads to smaller values for $\mathcal{L}^E(\theta_0)$, thus the coverage probabilities of BEL is much larger than nominal levels. This explains, for log-normal distribution, why MEL is much better than other methods.

Table 2 shows the basic numerical characteristics of different log-EL ratio statistics, such as expectations, variances and medians. Since all these asymptotically equivalent statistics converges in distribution to $\chi^2(1)$, the true mean and variance should be 1 and 2. From this table, we can see that

(1) Compared with other EL methods, OEL log-likelihood ratios are much larger. It leads to the lower coverage proportions showed in Table 1.

(2) The results of BEL and AEL are similar. BEL has some negative values while AEL has many 0s.

(3) MEL and EEL are similar. Meanwhile the mean and variance of MEL are closer to true values than those of EEL.

Table 3: The comparison between MEL and other EL methods.

		Norm(0, 1)			t(5)			LogNorm(0, 1)		
	n	proportion	mean	var	proportion	mean	var	proportion	mean	var
OEL−MEL	20	1	0.1680	0.1916	1	0.1798	0.2098	1	0.6728	8.0499
	40	1	0.0825	0.0703	1	0.0902	0.0645	1	0.2084	0.5278
	100	1	0.0312	0.0106	1	0.0382	0.1768	1	0.0901	0.1362
BEL−MEL	20	0.3173	0.0783	0.0940	0.0225	-0.0997	0.0202	0	-4.1773	836.7197
	40	0.3243	0.0429	0.0460	0.0416	-0.0314	0.0155	0	-1.2584	3.4900
	100	0.3210	0.0160	0.0075	0.0405	-0.0077	0.0085	0	-0.3949	0.2585
EEL−MEL	20	0.9146	0.0516	0.0209	0.8770	-0.0073	0.0379	0.8236	-0.2445	7.1008
	40	0.9409	0.0117	0.0037	0.9186	0.0121	0.0060	0.8559	0.0635	0.0820
	100	0.9065	0.0035	0.0016	0.9079	0.0082	0.0045	0.8657	0.0467	0.0550

The ”proportion” shows the proportion of positive difference. OEL is $\mathcal{L}^O(\theta_0)$, BEL is $\mathcal{L}^B(\theta_0)$, MEL is $\mathcal{L}^M(\theta_0)$ and EEL is $\mathcal{L}^E(\theta_0)$

For further study of the difference of these EL method, we get the following Table 3, which shows the difference of log-empirical likelihood ratio between MEL and other EL methods. From this table, we can further understand the reason why MEL performs similar as EEL. The difference, $\mathcal{L}^E(\theta_0) - \mathcal{L}^M(\theta_0)$, is very small (see the last row and the column ‘mean’ for all three distributions; the mean differences are very small). For the heavy tail distribution LogNorm(0, 1), we clearly have (see the column ‘proportion - proportion of positive difference’)

$$\mathcal{L}^B(\theta_0) < \mathcal{L}^M(\theta_0) < \mathcal{L}^O(\theta_0).$$

It implies that for large kurtosis, MEL will correct the empirical likelihood automatically, but BEL makes the empirical likelihood too small due to a large Bartlett correction factor b.

3.2 Example II: Regression Models

In this section, we consider the linear regression model $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, where ε_is are independent random variables with mean zero and finite variance and the parameters of interest are (β_0, β_1). Therefore the estimating equations are

$$g_1(x, Y, \beta_0, \beta_1) = Y - \beta_0 - \beta_1 x, \qquad g_2(x, Y, \beta_0, \beta_1) = (Y - \beta_0 - \beta_1 x)x.$$

Table 4: Coverage probabilities for regression parameters with homoscedastic error ε.

		$\varepsilon \sim$ Norm(0, 1)					$\varepsilon \sim$ t(5)				
n	$1-\alpha$	OEL	MEL	BEL	AEL	EEL	OEL	MEL	BEL	AEL	EEL
20	0.90	0.7992	0.8446	0.8490	0.8766	**0.8528**	0.7826	0.8314	0.9482	1.0000	**0.8456**
	0.95	0.7296	0.9100	0.9074	0.9396	**0.9186**	0.8506	0.8906	0.9700	1.0000	**0.9132**
30	0.90	0.8324	0.8662	0.8618	0.8686	**0.8656**	0.8176	0.8536	0.9192	1.0000	**0.8584**
	0.95	0.8894	0.9204	0.9124	0.9190	**0.9212**	0.8832	0.9134	0.9578	1.0000	**0.9198**
50	0.90	0.8702	**0.8970**	0.8896	0.8908	0.8920	0.8574	0.8822	**0.9154**	0.9366	0.8794
	0.95	0.9234	**0.9454**	0.9384	0.9406	0.9430	0.9174	0.9394	**0.9558**	0.9812	0.9368
100	0.90	0.8818	**0.8948**	0.8890	0.8888	0.8912	0.8712	0.8868	**0.8986**	0.9028	0.8820
	0.95	0.9378	**0.9496**	0.9440	0.9440	0.9470	0.9254	0.9366	**0.9428**	0.9468	0.9346

Two different scenarios are considered here: the homoscedastic case and the heteroscedastic case.

SCENARIO 1: Homoscedasticity

The true parameter values are $\beta_0 = 1$, $\beta_1 = 2$ and x is generated from standard uniform distribution. We consider two different types of errors ε_i, one is drawn from Norm(0, 1) and the other is a heavy tail distribution t(5). Table 4 shows the coverage probabilities of confidence intervals derived from MEL, OEL, BEL , AEL and EEL under different sample sizes n = 20, 30, 50. Each entry in the Table 4 is based on 5000 random replicates of size n.

SCENARIO 2: Heteroscedasticity

In this scenario, we explore the performance of MEL ratio statistics under heteroscedasticity. We choose the true parameter values $\beta_0 = 3, \beta_1 = 2$, and generate x from Norm$(2, 3)$ distribution. We set $\varepsilon_i = x_i^2 * \xi_i$, and generate ξ_i from Norm$(0, 1)$ and heavy tail distribution t(5), respectively. Similarly as Scenario 1, we still use the theoretical Bartlett correction constant b to construct confidence regions. In this simulation, we choose n = 50, 200, 500 respectively and all results are based on 5000 Monte Carlo replications. The simulation results are summarized in Table 5.

Apart from the similar observations as Example I, we can also get the following observations based on Table 4 and 5,

(1) Homo case: The AEL statistic suffers from a boundedness problem which may lead to 100% confidence regions. For the heavy tail distribution t(5), the BEL

Table 5: Coverage probabilities for regression parameters with heteroscedastic error $\varepsilon = x^2\xi$.

		$\xi \sim$ Norm(0, 1)					$\xi \sim$ t(5)				
n	$1-\alpha$	OEL	MEL	BEL	AEL	EEL	OEL	MEL	BEL	AEL	EEL
50	0.90	0.7450	0.7864	**0.8986**	0.9874	0.7766	0.7234	**0.7618**	0.9978	1.0000	0.7594
	0.95	0.8212	0.8644	**0.9422**	0.9972	0.8562	0.8046	**0.8498**	0.9998	1.0000	0.8496
200	0.90	0.8324	0.8600	**0.8950**	0.9238	0.8392	0.8218	**0.8472**	0.9702	0.9978	0.8284
	0.95	0.9016	0.9258	**0.9434**	0.9678	0.9082	0.8890	**0.9100**	0.9868	1.0000	0.8928
500	0.90	0.8730	0.8952	**0.8994**	0.9064	0.8752	0.8598	**0.8798**	0.9412	0.9748	0.8618
	0.95	0.9288	0.9460	**0.9504**	0.9568	0.9308	0.9230	**0.9370**	0.9746	0.9944	0.9246

and AEL coverage probabilities are much larger than nominal levels when sample size is small. MEL has the similar performance as EEL, but the coverage errors of MEL is smaller than EEL when sample size increases to $n = 50$.

(2) Hetero case: All of the coverage probabilities slowly converge to the nominal levels. When sample size $n = 500$, the coverage probabilities of OEL are still smaller than nominal levels. Under Norm distribution, theoretical BEL performs best; but for the same reason as Example I, under t distribution, both BEL and AEL are much larger. In this case, MEL performs better than EEL uniformly, especially for t distribution.

3.3 Example III: Two Sample Comparison

In this section, we report a simulation study designed to evaluate the performance of the MEL confidence regions proposed in Section 2.2. For simplicity, we consider $d = 1$ and $g(x) = x$. Simulated data sets of various n_1 and n_2 are generated under the following two scenarios: (1) X drawn from the standard exponential distribution and Y from $\chi^2(3)$ distribution; (2) X from t(5) distribution and Y from Log-Norm$(0, 1)$ distribution. The nominal coverage levels under consideration are $\alpha = 0.90$, 0.95. Table 6 shows the coverage percentage comparisons for constructing the confidence intervals, where all results are based on 5000 Monte Carlo replicates. In all cases, the MEL works uniformly better than OEL, and it is much more accurate when sample sizes are small.

Table 6: Coverage probabilities for two sample comparison.

$X \sim \text{Exp}(1),\ Y \sim \chi^2(3)$					$X \sim \text{t}(5),\ Y \sim \text{LogN}(0,\ 1)$				
	0.90		0.95			0.90		0.95	
$(n_1,\, n_2)$	OEL	MEL	OEL	MEL	$(n_1,\, n_2)$	OEL	MEL	OEL	MEL
$(10, 10)$	0.8459	**0.8836**	0.8993	**0.9356**	$(20, 20)$	0.8586	**0.8793**	0.9238	**0.9376**
$(10, 20)$	0.8659	**0.8835**	0.9177	**0.9357**	$(20, 50)$	0.8670	**0.9123**	0.9270	**0.9582**
$(10, 30)$	0.8791	**0.9074**	0.9308	**0.9580**	$(20, 100)$	0.8799	**0.9253**	0.9219	**0.9627**
$(20, 10)$	0.8376	**0.8903**	0.8973	**0.9440**	$(50, 20)$	0.8417	**0.8806**	0.8988	**0.9312**
$(20, 20)$	0.8752	**0.8976**	0.9321	**0.9491**	$(50, 50)$	0.8490	**0.8800**	0.9120	**0.9350**
$(20, 30)$	0.8823	**0.8936**	0.9307	**0.9429**	$(50, 100)$	0.8810	**0.8878**	0.9330	**0.9389**
$(30, 10)$	0.8378	**0.8921**	0.8976	**0.9432**	$(100, 20)$	0.8298	**0.8718**	0.8859	**0.9354**
$(30, 20)$	0.8737	**0.9023**	0.9249	**0.9454**	$(100, 50)$	0.8430	**0.8758**	0.9050	**0.9343**
$(30, 30)$	0.8782	**0.8942**	0.9344	**0.9466**	$(100, 100)$	0.8770	**0.8840**	0.9310	**0.9350**
$(50, 20)$	0.8573	**0.8931**	0.9195	**0.9458**	$(200, 50)$	0.8450	**0.8756**	0.9050	**0.9455**
$(50, 30)$	0.8810	**0.9030**	0.9332	**0.9513**	$(200, 100)$	0.8860	**0.9013**	0.9470	**0.9587**
$(50, 50)$	0.8878	**0.8987**	0.9388	**0.9479**	$(200, 200)$	0.8970	**0.8990**	0.9490	**0.9510**

3.4 Example IV: Exponentially Tilted Empirical Likelihood

To assess the performance of MEL based on Theorem 2.3, we first carry out the following simulation study. We fixed the nominal level $\alpha = 0.95$ and sample size $n = 100$. The following distributions are considered, Normal$(0, 1)$, Weibull$(2, 1)$, Generalized Pareto$(1/4, 1, 0)$ and LNi for LogNorm$(0,\ i^2/4)$, $i = 1,\ 2,\ 3$. The skewnesses of these distributions are: 0, 2, 7.07, 1.75, 6.18 and 33.47. Under all these distributions, we implement the following six different methods, (1) OEL, (2) ETL-1: the Newey-Smith empirical entropy difference $\mathcal{T}_1(\theta)$, (3) ETL-2: Jaynes empirical entropy difference $\mathcal{T}_2(\theta)$, (4) MEL, (5) METL-1: the Newey-Smith mean empirical entropy difference $\mathcal{T}_1^M(\theta)$ and (6) METL-2: the Jaynes mean empirical entropy difference $\mathcal{T}_2^M(\theta)$.

The results of coverage probabilities based on 5000 Monte Carlo replicates are shown in Figure 3.4 (details can be found in the supplementary file). For all cases, the performance of MEL is the best. Also note that as the skewness increases, MEL becomes much better than other methods. Since Exponential Tilted EL is not Bartlett correctable, see Jing (1996). MEL provides a good way to improve the accuracy of coverage probabilities.

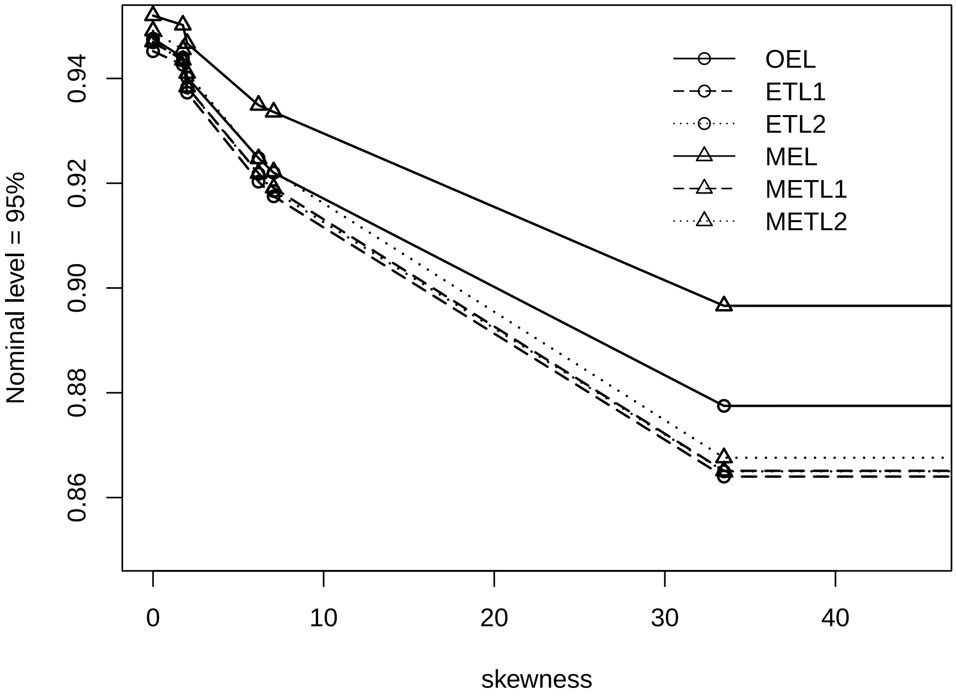

Figure 1: Coverage probabilities of mean under different distributions when sample size is 100. Notation ∘ stands for EL-type coverage probabilities and △ for MEL-type coverage probabilities.

Table 7: Boston Housing Study: confidence intervals for the mean parameter

OEL	ETL-1	ETL-2	MEL	METL-1	METL-2
$(2.9611, 4.4970)$	$(2.9211, 4.4188)$	$(2.9204, 4.4197)$	$(2.9599, 4.5343)$	$(2.9205, 4.4198)$	$(2.9193, 4.4215)$

4 Real Data Analysis

In this section, we compare our proposed methods with existing methods using a real dataset. This example is taken from the Boston Housing Study, to illustrate our proposed MEL for Exponentially Tilted Likelihood method in Section 2.3. The distribution of per capita crime rate by town (CRIM) in the dataset follows an unknown heavy-tailed distribution. The data set, which has even been analyzed by Harrison and Rubinfeld (1978), consists of 506 observations. We are interested in the mean of CRIM, θ. A comparison of OEL, ETL-1, ETL-2, MEL, METL-1 and METL-2 was carried out. The 95% confidence intervals are list in Table 7., which show similar performance of our proposed methods and other existing methods.

5 Conclusion

This paper developed a mean empirical likelihood approach, which gives much more accurate confidence region estimates and coverage probabilities. We presented the method and proved its large sample properties under different application problems, regression models, two-sample comparisons and exponentially tilted likelihood. This new approach outperforms existing methods, in particular for heavy-tail or highly-skewed distributions.

The new method gains its advantage by using the pairwise-mean data and is equivalent to using each data point more than once. A well-known example is the Hodges-Lehmann sign-based estimator, see Hodges and Lehmann (1963), using a similar mean-pair data idea, which provides much more reliable nonparametric estimators than standard median estimator. Such mean-pair data approach will not be very useful for standard estimation approaches where observations have the same weights, however, it brings new insights to the area of empirical likelihood which assigns different weights for observations. We are currently working on a more general approach of constructing such pseudo dataset, which can determine the percentage of pairwise mean data and the percentage of multiple mean data in the pseudo dataset and provide an optimum solution.

Note that Wood et al. (1996) proposed a novel sequential linearization method for empirical likelihood with nonlinear constraints, which can be applied to solve the U-statistics problem in OEL. But from the practical viewpoint, this method typically requires iterations to get satisfactory results. The Jackknife Empirical Likelihood(JEL) method (Jing et al., 2009) can improve the computational efficiency for the sequential linearization method in some degree, but it still needs bootstrap calibrations to improve the performance on coverage probabilities. Our MEL approach uses U-statistics as well, however, it is actually different from the sequential linearization method. In our MEL approach, each pair-wise mean data point $W_k = (V_i + V_j)/2$ corresponds to a weight $p_k, k = 1, \cdots, N$. On the contrary, in the sequential linearization method, each pair-wise mean data point $(V_i + V_j)/2$ corresponds to the product $p_i p_j, i, j = 1, \cdots, n$. This is why the sequential linearization method has non-linear constraints and involves a heavier computational cost. It is interesting to study how well MEL performs for EL with non-linear constraints and we left this to future research.

A Proof of Theorem 2.1

Denote $\boldsymbol{V}_i = \boldsymbol{V}_i(\boldsymbol{\theta}_0)$ and $\boldsymbol{W}_i = \boldsymbol{W}_i(\boldsymbol{\theta}_0)$. We shall introduce the following lemma, which is a key for the proof of Theorem 2.1.

Lemma A.1. *Under the condition $\boldsymbol{Cov}(\boldsymbol{V}_1) = \boldsymbol{\Sigma}$ exists, $\boldsymbol{rank}(\boldsymbol{\Sigma}) = d$, we have*

$$(i) \qquad \max_{1 \le k \le N} \|\boldsymbol{W}_k\| = o_p(n^{1/2}),$$

$$(ii) \qquad \boldsymbol{e}' \left(\frac{1}{N} \sum_{k=1}^{N} \boldsymbol{W}_k \right) = O_p(n^{-1/2}), \quad \textit{where } \boldsymbol{e} \textit{ is any unit vector in } \mathbb{R}^d,$$

$$(iii) \qquad \frac{1}{N} \sum_{k=1}^{N} \boldsymbol{W}_k \boldsymbol{W}_k' = \frac{1}{2} \boldsymbol{\Sigma} + o_p(1),$$

$$(iv) \qquad \frac{1}{N} \sum_{k=1}^{N} \boldsymbol{W}_k' \boldsymbol{W}_k = O_p(1).$$

Proof. (i) Since $\texttt{Cov}(\boldsymbol{V}_i) = \boldsymbol{\Sigma}$ exists, we immediately have $\max_i \|\boldsymbol{V}_i\| = o_p(n^{1/2})$ and

$$\max_{1\leq k\leq N} \|\boldsymbol{W}_k\| = \max_{i\leq j} \left\|\frac{\boldsymbol{V}_i + \boldsymbol{V}_j}{2}\right\| \leq \frac{1}{2}\left(\max_i \|\boldsymbol{V}_i\| + \max_j \|\boldsymbol{V}_j\|\right) = o_p(n^{1/2}).$$

(ii) Noticing

$$\frac{1}{N}\sum_{k=1}^{N} \boldsymbol{W}_k = \frac{1}{2N}\left(\sum_{i=1}^{n}\sum_{j=1}^{n} \frac{\boldsymbol{V}_i + \boldsymbol{V}_j}{2} + \sum_{i=1}^{n} \boldsymbol{V}_i\right) = \frac{n+1}{2N}\sum_{i=1}^{n} \boldsymbol{V}_i = \bar{\boldsymbol{V}}_n,$$

and following the assumption $\texttt{Cov}(\boldsymbol{V}_1) = \boldsymbol{\Sigma}$, we can obtain $\boldsymbol{e}'\bar{\boldsymbol{V}}_n = \boldsymbol{e}'(\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{V}_i) = O_p(n^{-1/2})$. Therefore

$$\boldsymbol{e}'\left(\frac{1}{N}\sum_{k=1}^{N} \boldsymbol{W}_k\right) = O_p(n^{-1/2}).$$

(iii)

$$\begin{aligned}\frac{1}{N}\sum_{k=1}^{N} \boldsymbol{W}_k\boldsymbol{W}_k' &= \frac{1}{2N}\left(\sum_{i=1}^{n}\sum_{j=1}^{n}\left(\frac{\boldsymbol{V}_i + \boldsymbol{V}_j}{2}\right)\left(\frac{\boldsymbol{V}_i + \boldsymbol{V}_j}{2}\right)' + \sum_{i=1}^{n} \boldsymbol{V}_i\boldsymbol{V}_i'\right)\\ &= \frac{1}{2(n+1)}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} \boldsymbol{V}_i\right)\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} \boldsymbol{V}_i'\right) + \frac{n+2}{2(n+1)}\left(\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{V}_i\boldsymbol{V}_i'\right)\\ &= \frac{1}{2}\boldsymbol{\Sigma} + o_p(1).\end{aligned}$$

(iv)

$$\frac{1}{N}\sum_{k=1}^{N} \|\boldsymbol{W}_k\|^2 = \frac{n+2}{2(n+1)}\left(\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{V}_i'\boldsymbol{V}_i\right) + \frac{1}{2(n+1)}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} \boldsymbol{V}_i'\right)\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} \boldsymbol{V}_i\right) = O_p(1).$$

□

Now we can prove Theorem 2.1.

Proof. According to Lemma 11.1 in Owen (2001), with probability tending to 1, $\mathbf{0}$ is inside the convex hull of $\boldsymbol{W}_k$, $k = 1, 2, \cdots, n$. By using the Lagrange multiplier, we have

$$p_k = \frac{1}{N(1 + \boldsymbol{\lambda}'\boldsymbol{W}_k)} > 0,$$

where $\boldsymbol{\lambda}$ satisfies the equation $\sum p_k \boldsymbol{W}_k = 1$. Then applying Lemma A.1, we have

$$\|\boldsymbol{\lambda}\| = O_p(n^{-1/2}). \tag{A.1}$$

On the other hand, with following equation

$$\mathbf{0} = \frac{1}{N}\sum_{k=1}^{N}\frac{\boldsymbol{W}_k}{1+\boldsymbol{\lambda}'\boldsymbol{W}_k} = \frac{1}{N}\sum_{k=1}^{N}\boldsymbol{W}_k - \frac{1}{N}\left(\sum_{k=1}^{N}\boldsymbol{W}_k\boldsymbol{W}_k'\right)\boldsymbol{\lambda} + \frac{1}{N}\sum_{k=1}^{N}\frac{\boldsymbol{W}_k(\boldsymbol{\lambda}'\boldsymbol{W}_k)^2}{1+\boldsymbol{\lambda}'\boldsymbol{W}_k},$$

equation((A.1)) and Lemma A.1, we get

$$\boldsymbol{\lambda} = \left(\sum_{k=1}^{N}\boldsymbol{W}_k\boldsymbol{W}_k'\right)^{-1}\left(\sum_{k=1}^{N}\boldsymbol{W}_k\right) + o_p(n^{-1/2}).$$

Using Taylor's expansion, we can write

$$\mathcal{L}^M(\boldsymbol{\theta}_0) = \frac{2}{n+1}\sum_{k=1}^{N}\log(1+\boldsymbol{\lambda}'\boldsymbol{W}_k) = \frac{2}{n+1}\sum_{k=1}^{N}\left(\boldsymbol{\lambda}'\boldsymbol{W}_k - \frac{1}{2}(\boldsymbol{\lambda}'\boldsymbol{W}_k)^2\right) + \frac{r_N}{n+1}, \quad \text{(A.2)}$$

where

$$\|r_N\| \le C\|\boldsymbol{\lambda}\|^3 \max_{1\le k\le N}\|\boldsymbol{W}_k\|\sum_{k=1}^{N}\|\boldsymbol{W}_k\|^2 = O_p(n^{-3/2})o_p(n^{1/2})O_p(n^2) = o_p(n).$$

Substitute $\boldsymbol{\lambda}$ into ((A.2)), we obtain

$$\begin{aligned} \mathcal{L}^M(\boldsymbol{\theta}_0) &= \frac{1}{n+1}\left(\sum_{k=1}^{N}\boldsymbol{W}_k\right)\left(\sum_{k=1}^{N}\boldsymbol{W}_k\boldsymbol{W}_k'\right)^{-1}\left(\sum_{k=1}^{N}\boldsymbol{W}_k\right) + o_p(1) \\ &= \frac{2N}{n+1}\bar{\boldsymbol{V}}_n'\boldsymbol{\Sigma}^{-1}\bar{\boldsymbol{V}}_n + o_p(1) = n\bar{\boldsymbol{V}}_n'\boldsymbol{\Sigma}^{-1}\bar{\boldsymbol{V}}_n + o_p(1) \to \chi^2(d), \quad \text{in dist.} \end{aligned}$$

□

B Proof of Theorem 2.2

The notations follow that in Section 2.2. First we shall introduce the following Lemma giving the relation of $\boldsymbol{\lambda}_1$, $\boldsymbol{\lambda}_2$ and $\boldsymbol{\mu}$.

Lemma B.1. *Assume $\boldsymbol{\Sigma}_X := \mathtt{Cov}(\boldsymbol{X})$ and $\boldsymbol{\Sigma}_Y := \mathtt{Cov}(\boldsymbol{Y})$ exist and $\mathtt{rank}(\boldsymbol{\Sigma}_X) = \mathtt{rank}(\boldsymbol{\Sigma}_Y) = d$. Denote $\boldsymbol{\mu}_0$ as the mean of $\boldsymbol{X}$ and*

$$\boldsymbol{V}_X = \frac{1}{2\delta}\boldsymbol{\Sigma}_X, \quad \boldsymbol{V}_Y = \frac{1}{2(1-\delta)}\boldsymbol{\Sigma}_Y, \quad \boldsymbol{C}_X = \frac{1}{N\delta}\sum_{s=1}^{N_1}(\boldsymbol{W}_s^X - \boldsymbol{\mu}_0), \quad \boldsymbol{C}_Y = \frac{1}{N(1-\delta)}\sum_{t=1}^{N_2}(\boldsymbol{W}_t^Y - \boldsymbol{\mu}_0).$$

Then we have

$$(i) \qquad \boldsymbol{\lambda}_1 = (\boldsymbol{V}_X)^{-1}(\boldsymbol{C}_X - \boldsymbol{\mu}) + o_p(n^{-1/2}), \qquad \boldsymbol{\lambda}_2 = (\boldsymbol{V}_Y)^{-1}(\boldsymbol{C}_Y - \boldsymbol{\mu}) + o_p(n^{-1/2}),$$

$$\boldsymbol{\mu} = \boldsymbol{V}_X \left(\boldsymbol{V}_X + \boldsymbol{V}_Y\right)^{-1} \boldsymbol{V}_Y \left(\boldsymbol{V}_X^{-1}\boldsymbol{C}_X + \boldsymbol{V}_Y^{-1}\boldsymbol{C}_Y\right) = \boldsymbol{\mu}_0 + O_p(n^{-1/2}),$$

$$(ii) \qquad \boldsymbol{S}_X = \frac{1}{N_1}\sum_{s=1}^{N_1}(\boldsymbol{W}_s^X - \mu)(\boldsymbol{W}_s^X - \mu)^T = \frac{1}{2}\boldsymbol{\Sigma}_X + o_p(1),$$

$$\boldsymbol{S}_Y = \frac{1}{N_2}\sum_{t=1}^{N_2}(\boldsymbol{W}_t^Y - \boldsymbol{\mu})(\boldsymbol{W}_t^Y - \boldsymbol{\mu})^T = \frac{1}{2}\boldsymbol{\Sigma}_Y + o_p(1).$$

The proof of Lemma B.1 is similar to Liu et al. (2008) and is omitted here. Now we provide the proof of Theorem 2.2.

Proof. Using Taylor's expansion, we can write

$$\mathcal{L}_2^M(\boldsymbol{\theta}_0) = \frac{2}{n}\sum_{s=1}^{N_1}\log\left[1 + \delta^{-1}\boldsymbol{\lambda}_1'(\boldsymbol{W}_s^X - \boldsymbol{\mu})\right] + \frac{2}{n}\sum_{t=1}^{N_2}\log\left[1 + (1-\delta)^{-1}\boldsymbol{\lambda}_2'(\boldsymbol{W}_t^Y - \boldsymbol{\mu})\right] + \frac{r_N}{n},$$

With a similar argument as that in Theorem 2.1, we know that r_N is of order $o_p(n^{-1})$. Therefore we have

$$\begin{aligned}
\mathcal{L}_2^M(\boldsymbol{\theta}_0) &= \frac{2}{n}\sum_{s=1}^{N_1}\delta^{-1}\boldsymbol{\lambda}_1'(\boldsymbol{W}_s^X - \boldsymbol{\mu}) - \frac{1}{n}\sum_{s=1}^{N_1}\left[\delta^{-1}\boldsymbol{\lambda}_1'(\boldsymbol{W}_s^X - \boldsymbol{\mu})\right]^2 \\
&\quad + \frac{2}{n}\sum_{t=1}^{N_2}(1-\delta)^{-1}\boldsymbol{\lambda}_2'(\boldsymbol{W}_t^Y - \boldsymbol{\mu}) - \frac{1}{n}\sum_{t=1}^{N_2}\left[(1-\delta)^{-1}\boldsymbol{\lambda}_2'(\boldsymbol{W}_t^Y - \boldsymbol{\mu})\right]^2 + o_p(1) \\
&= \frac{2N_1}{n}\delta^{-1}\boldsymbol{\lambda}_1'(\boldsymbol{C}_X + \boldsymbol{\mu}_0 - \boldsymbol{\mu}) - \frac{N_1}{n}\delta^{-1}\boldsymbol{\lambda}_1'\boldsymbol{S}_X\delta^{-1}\boldsymbol{\lambda}_1 \\
&\quad + \frac{2N_2}{n}(1-\delta)^{-1}\boldsymbol{\lambda}_2'(\boldsymbol{C}_Y + \boldsymbol{\mu}_0 - \boldsymbol{\mu}) - \frac{N_2}{n}(1-\delta)^{-1}\boldsymbol{\lambda}_2'\boldsymbol{S}_Y(1-\delta)^{-1}\boldsymbol{\lambda}_2 + o_p(1).
\end{aligned}$$

Substituting $(\boldsymbol{\lambda}_1,\, \boldsymbol{\lambda}_2)$ into $\mathcal{L}_2^M(\boldsymbol{\theta}_0)$ and with some simple calculations, we further obtain

$$\begin{aligned}
\mathcal{L}_2^M(\boldsymbol{\theta}_0) &= 2n^{-1}N\delta(\boldsymbol{C}_X - \boldsymbol{\mu})'\boldsymbol{\Sigma}_X^{-1}(\boldsymbol{C}_X - \boldsymbol{\mu}) + 2n^{-1}N(1-\delta)(\boldsymbol{C}_Y - \boldsymbol{\mu})'\boldsymbol{\Sigma}_Y^{-1}(\boldsymbol{C}_Y - \boldsymbol{\mu}) \\
&\quad + n^{-1}N\left[(\boldsymbol{C}_X - \boldsymbol{\mu})'\boldsymbol{V}_X^{-1} + (\boldsymbol{C}_Y - \boldsymbol{\mu})'\boldsymbol{V}_Y^{-1}\right]\boldsymbol{\mu}_0 + o_p(1).
\end{aligned}$$

Since $(\boldsymbol{C}_X - \boldsymbol{\mu})'\boldsymbol{V}_X^{-1} + (\boldsymbol{C}_Y - \boldsymbol{\mu})'\boldsymbol{V}_Y^{-1} = 0$, we can rewrite the above $\mathcal{L}_2^M(\boldsymbol{\theta}_0)$ as

$$\begin{aligned}
\mathcal{L}_2^M(\boldsymbol{\theta}_0) &= 2n^{-1}N\delta(\boldsymbol{C}_X - \boldsymbol{\mu})'\boldsymbol{\Sigma}_X^{-1}(\boldsymbol{C}_X - \boldsymbol{\mu}) + 2n^{-1}N(1-\delta)(\boldsymbol{C}_Y - \boldsymbol{\mu})'\boldsymbol{\Sigma}_Y^{-1}(\boldsymbol{C}_Y - \boldsymbol{\mu}) + o_p(1) \\
&= n^{-1}N(\boldsymbol{C}_X - \boldsymbol{C}_Y)'(\boldsymbol{V}_X + \boldsymbol{V}_Y)^{-1}(\boldsymbol{C}_X - \boldsymbol{C}_Y) + o_p(1) \\
&= n(\boldsymbol{C}_X - \boldsymbol{C}_Y)'\left(\frac{1}{\Delta^2}\boldsymbol{\Sigma}_X + \frac{1}{(1-\Delta)^2}\boldsymbol{\Sigma}_Y\right)^{-1}(\boldsymbol{C}_X - \boldsymbol{C}_Y) + o_p(1). \qquad \text{(B.1)}
\end{aligned}$$

Noting that $\boldsymbol{C}_X - \boldsymbol{C}_Y = \bar{\boldsymbol{X}} - \bar{\boldsymbol{Y}}$, we then apply the Central Limit Theorem and have

$$(\boldsymbol{C}_X - \boldsymbol{C}_Y)'\left(\frac{1}{n_1}\boldsymbol{\Sigma}_X + \frac{1}{n_2}\boldsymbol{\Sigma}_Y\right)^{-1}(\boldsymbol{C}_X - \boldsymbol{C}_Y)$$

$$= \quad n(\boldsymbol{C}_X - \boldsymbol{C}_Y)' \left(\frac{1}{\Delta} \boldsymbol{\Sigma}_X + \frac{1}{1-\Delta} \boldsymbol{\Sigma}_Y \right)^{-1} (\boldsymbol{C}_X - \boldsymbol{C}_Y) \to \chi^2(d). \tag{B.2}$$

The theorem is then proved by the fact that equations ((B.1)) and ((B.2)) together imply

$$n^{-1} \mathcal{L}_2^M(\boldsymbol{\theta}_0) \to \sum_{k=1}^{d} r_k \, \chi_k^2(1), \quad \text{in dist.}$$

where $\chi_k^2(1)$ are standard χ^2 distribution, and r_k are the eigenvalues of $(\boldsymbol{R}^M)^{-1}\boldsymbol{R}$,

$$\boldsymbol{R} = \frac{1}{\Delta_0} \boldsymbol{\Sigma}_X + \frac{1}{1-\Delta_0} \boldsymbol{\Sigma}_Y, \qquad \boldsymbol{R}^M = \frac{1}{\Delta_0^2} \boldsymbol{\Sigma}_X + \frac{1}{(1-\Delta_0)^2} \boldsymbol{\Sigma}_Y.$$

□

C Proof of Theorem 2.4

Proof. Denote $\rho_1 = \mathrm{d}\,\rho(v)/\mathrm{d}\,v$, $\rho_2 = \mathrm{d}^2\,\rho(v)/\mathrm{d}\,v^2$, and $\boldsymbol{\Gamma}_k(\boldsymbol{\theta}) = \partial \boldsymbol{W}_k(\boldsymbol{\theta})/\partial \boldsymbol{\theta}$.

Under the assumptions A1-A5, similarly as proof of Theorem 2.2 and 3.1 in Newey and Smith (2004), GMEL estimator $\hat{\boldsymbol{\theta}}$ and the corresponding $\hat{\boldsymbol{\lambda}} = \hat{\boldsymbol{\lambda}}(\hat{\boldsymbol{\theta}})$ satisfy

$$\sum_{k=1}^{N} \rho_1(\hat{\boldsymbol{\lambda}}^T \boldsymbol{W}_k(\hat{\boldsymbol{\theta}})) \left(\frac{\partial \boldsymbol{W}_k(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} \right)^T \hat{\boldsymbol{\lambda}} = \boldsymbol{0},$$

$$\sum_{k=1}^{N} \rho_1(\hat{\boldsymbol{\lambda}}^T \boldsymbol{W}_k(\hat{\boldsymbol{\theta}})) \boldsymbol{W}_k(\hat{\boldsymbol{\theta}}) = \boldsymbol{0},$$

and

$$\hat{\boldsymbol{\theta}} \to \boldsymbol{\theta}_0, \qquad \hat{\boldsymbol{\lambda}} = O_p(n^{-1/2}).$$

By Taylor expansion for $(\hat{\boldsymbol{\theta}}^T, \hat{\boldsymbol{\lambda}}^T)^T$ at $(\boldsymbol{\theta}_0^T, \boldsymbol{0})^T$,

$$\begin{pmatrix} \boldsymbol{0} \\ \boldsymbol{0} \end{pmatrix} = \begin{pmatrix} \boldsymbol{0} \\ -\frac{1}{N}\sum_{k=1}^{N} \boldsymbol{W}_k(\boldsymbol{\theta}_0) \end{pmatrix} + \boldsymbol{\Psi}_N \begin{pmatrix} \hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0 \\ \hat{\boldsymbol{\lambda}} - \boldsymbol{0} \end{pmatrix}, \tag{C.1}$$

where $\boldsymbol{\Psi}_N$ is a $(d+m)*(d+m)$ matrix,

$$\boldsymbol{\Psi}_N = \frac{1}{N} \begin{pmatrix} \boldsymbol{0}, & \sum_{k=1}^{N} \rho_1(\tilde{\boldsymbol{\lambda}}^T \boldsymbol{W}_k(\tilde{\boldsymbol{\theta}})) \boldsymbol{\Gamma}_k^T(\tilde{\boldsymbol{\theta}}) \\ \sum_{k=1}^{N} \rho_1(\tilde{\boldsymbol{\lambda}}^T \boldsymbol{W}_k(\tilde{\boldsymbol{\theta}})) \boldsymbol{\Gamma}_k(\tilde{\boldsymbol{\theta}}) & \sum_{k=1}^{N} \rho_2(\tilde{\boldsymbol{\lambda}}^T \boldsymbol{W}_k(\tilde{\boldsymbol{\theta}})) \boldsymbol{W}_k^T(\tilde{\boldsymbol{\theta}}) \boldsymbol{W}_k(\tilde{\boldsymbol{\theta}}) \end{pmatrix},$$

$\tilde{\boldsymbol{\theta}}$ and $\tilde{\boldsymbol{\lambda}}$ are vectors between $(\hat{\boldsymbol{\theta}}^T, \hat{\boldsymbol{\lambda}}^T)^T$ and $(\boldsymbol{\theta}_0^T, \mathbf{0})^T$. Since $\rho_1(0) = -1$, $\rho_2(0) = -1$, together with Lemma A1 in Newey and Smith (2004), we get

$$\begin{aligned} &\max_k |\rho_1(\tilde{\boldsymbol{\lambda}}^T \boldsymbol{W}_k(\tilde{\boldsymbol{\theta}}) + 1| \to 0, \qquad \text{in pr.} \\ &\max_k |\rho_2(\tilde{\boldsymbol{\lambda}}^T \boldsymbol{W}_k(\tilde{\boldsymbol{\theta}}) + 1| \to 0, \qquad \text{in pr.} \end{aligned}$$

Further, using the similar proof as Lemma A.1 in Appendix A, we have

$$\begin{aligned} &\frac{1}{N}\sum_{k=1}^{N} \rho_1(\tilde{\boldsymbol{\lambda}}^T \boldsymbol{W}_k(\tilde{\boldsymbol{\theta}}))\boldsymbol{\Gamma}_k(\tilde{\boldsymbol{\theta}}) = -\boldsymbol{\Gamma}_0 + o_p(1), \\ &\frac{1}{N}\sum_{k=1}^{N} \rho_2(\tilde{\boldsymbol{\lambda}}^T W_k(\tilde{\boldsymbol{\theta}}))\boldsymbol{W}_k^T(\tilde{\boldsymbol{\theta}})\boldsymbol{W}_k(\tilde{\boldsymbol{\theta}}) = -\frac{1}{2}\boldsymbol{\Sigma}_0 + o_p(1). \end{aligned} \tag{C.2}$$

Hence

$$\boldsymbol{\Psi}_N \to \boldsymbol{\Psi} = \begin{pmatrix} \mathbf{0}, & -\boldsymbol{\Gamma}_0^T \\ -\boldsymbol{\Gamma}_0 & -\frac{1}{2}\boldsymbol{\Sigma}_0 \end{pmatrix}, \qquad \boldsymbol{\Psi}^{-1} = \begin{pmatrix} \frac{1}{2}\boldsymbol{K}, & -\boldsymbol{L} \\ -\boldsymbol{L}^T & -2\boldsymbol{H} \end{pmatrix},$$

where

$$\boldsymbol{K} = (\boldsymbol{\Gamma}_0^T \boldsymbol{\Sigma}_0^{-1} \boldsymbol{\Gamma}_0)^{-1}, \quad \boldsymbol{L} = \boldsymbol{K}\boldsymbol{\Gamma}_0^T \boldsymbol{\Sigma}_0^{-1}, \quad \boldsymbol{H} = \boldsymbol{\Sigma}_0^{-1} - \boldsymbol{\Sigma}_0^{-1}\boldsymbol{\Gamma}_0 \boldsymbol{K}\boldsymbol{\Gamma}_0^T \boldsymbol{\Sigma}_0^{-1}.$$

Denote $\bar{\boldsymbol{W}}(\boldsymbol{\theta}) = N^{-1}\sum_{k=1}^{N} \boldsymbol{W}_k(\boldsymbol{\theta})$. Since $\bar{\boldsymbol{W}}(\boldsymbol{\theta}_0) = n^{-1}\sum_{i=1}^{n} \boldsymbol{V}_i(\boldsymbol{\theta}_0) = O_p(n^{-1/2})$, after solving equation (C.1), we get

$$\begin{aligned} \sqrt{n}\begin{pmatrix} \hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0 \\ \hat{\boldsymbol{\lambda}} - \mathbf{0} \end{pmatrix} &= -\boldsymbol{\Psi}_N^{-1}\sqrt{n}\begin{pmatrix} \mathbf{0} \\ -\bar{\boldsymbol{W}}(\boldsymbol{\theta}_0) \end{pmatrix} = \Psi^{-1}\begin{pmatrix} \mathbf{0} \\ \sqrt{n}\bar{\boldsymbol{W}}(\boldsymbol{\theta}_0) \end{pmatrix} + o_p(1) \\ &= \begin{pmatrix} -\boldsymbol{L}\sqrt{n}\bar{\boldsymbol{W}}(\boldsymbol{\theta}_0) \\ -2\boldsymbol{H}\sqrt{n}\bar{\boldsymbol{W}}(\boldsymbol{\theta}_0) \end{pmatrix} + o_p(1). \end{aligned}$$

Therefore

$$\begin{aligned} \left(\frac{1}{N}\sum_{k=1}^{N} \boldsymbol{W}_k(\hat{\boldsymbol{\theta}})\right) &= \left(\frac{1}{N}\sum_{k=1}^{N} \boldsymbol{W}_k(\boldsymbol{\theta}_0)\right) - \left(\frac{1}{N}\sum_{k=1}^{N} \boldsymbol{\Gamma}_k(\tilde{\boldsymbol{\theta}})\right)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \\ &= (\boldsymbol{I} - \boldsymbol{\Gamma}_0 \boldsymbol{L})\left(\frac{1}{N}\sum_{k=1}^{N} \boldsymbol{W}_k(\boldsymbol{\theta}_0)\right) = -\frac{1}{2}\boldsymbol{\Sigma}_0\,\hat{\boldsymbol{\lambda}} + o_p(n^{-1/2}). \end{aligned} \tag{C.3}$$

Using Taylor expansion,

$$\frac{1}{N}\sum_{k=1}^{N} \rho(\hat{\boldsymbol{\lambda}}^T \boldsymbol{W}_k(\hat{\boldsymbol{\theta}})) = \rho_0 - \hat{\boldsymbol{\lambda}}^T\left(\frac{1}{N}\sum_{k=1}^{N} \boldsymbol{W}_k(\hat{\boldsymbol{\theta}})\right) + \frac{1}{2}\hat{\boldsymbol{\lambda}}^T\left(\frac{1}{N}\sum_{k=1}^{N} \rho_2(\tilde{\boldsymbol{\lambda}}^T \boldsymbol{W}_k(\hat{\boldsymbol{\theta}}))\boldsymbol{W}_k(\hat{\boldsymbol{\theta}})\boldsymbol{W}_k^T(\hat{\boldsymbol{\theta}})\right)\hat{\boldsymbol{\lambda}}$$

where $\tilde{\boldsymbol{\lambda}}$ is between $\hat{\boldsymbol{\lambda}}$ and $\mathbf{0}$, and (C.2), we have

$$\begin{aligned}\frac{1}{N}\sum_{k=1}^{N}\rho(\hat{\boldsymbol{\lambda}}^T\boldsymbol{W}_k(\hat{\boldsymbol{\theta}})) &= \rho_0 - \hat{\boldsymbol{\lambda}}^T\bar{\boldsymbol{W}}(\hat{\boldsymbol{\theta}}) - \frac{1}{4}\hat{\boldsymbol{\lambda}}^T\boldsymbol{\Sigma}_0\hat{\boldsymbol{\lambda}} + o_p(n^{-1}) \\ &= \rho_0 + \bar{\boldsymbol{W}}^T(\hat{\boldsymbol{\theta}})\boldsymbol{\Sigma}_0^{-1}\bar{\boldsymbol{W}}(\hat{\boldsymbol{\theta}}) + o_p(n^{-1}) \\ &= \rho_0 + \left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{V}_i(\hat{\boldsymbol{\theta}})\right)^T\boldsymbol{\Sigma}_0^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{V}_i(\hat{\boldsymbol{\theta}})\right) + o_p(n^{-1}).\end{aligned}$$

It follows Newey and Smith (2004) that

$$n\left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{V}_i(\hat{\boldsymbol{\theta}})\right)^T\boldsymbol{\Sigma}_0^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{V}_i(\hat{\boldsymbol{\theta}})\right) \to \chi^2_{(m-d)}, \qquad \text{in dist.}$$

Hence

$$n\left(\frac{1}{N}\sum_{k=1}^{N}\rho(\hat{\boldsymbol{\lambda}}^T\boldsymbol{W}_k(\hat{\boldsymbol{\theta}})) - \rho_0\right) \to \chi^2_{(m-d)}, \qquad \text{in dist.}$$

□

References

CHEN, S. (1993). On the Accuracy of Empirical Likelihood Confidence Regions for Linear Regression Model. *Annals of the Institute of Statistical Mathematics.* **45**, 621–637.

CHEN, J., VARIYATH, A.M. & ABRAHAM, B. (2008). Adjusted Empirical Likelihood and its Properties. *Journal of Computational Graphical Statistics.* **17**, 426–443.

CORCORAN, S. A. (1998). Bartlett Adjustment of Empirical Discrepancy Statistics. *Biometrika.* **85(4)**, 967–972.

DICICCIO, T., HALL, P. & ROMANO, J. (1991). Empirical Likelihood is Bartlett-correctable. *Annals of Statistics.* **19**, 1053–1061.

HALL, P. & LA SCALA, B. (1990). Methodology and Algorithm of Empirical Likelihood. *International Statistical Review.* **58**, 109–127.

HARRISON, D. & RUBINFELD, D.L. (1978). Hedonic Prices and the Demand for Clean Air. *Journal of Environmental Economics and Management.* **5**, 81–102.

Hodges, Jr. J.L. & Lehmann, E.L. (1963). Estimates of Location Based on Rank Tests. *The Annals of Mathematical Statistics.* **34**, 598–611.

Jaynes, E.T. (1982). On the Rational of Maximum Entropy Methods. *Proceedings of IEEE International Conference.* **70**, 939–952.

Jing B.Y.(1996). Exponential Empirical Likelihood is Not Bartlett Correctable. *Annals of Statistics* **24**, 365–369.

Jing B.Y. Yuan J. & Zhou W. (2009). Jackknife Empirical Likelihood. *Journal of the American Statistical Association.* **104**, 1224–1232.

Liu, Y. & Chen, J. (2010). Adjusted Empirical Likelihood with High-order Precision. *Annals of Statistics.* **38**, 1341–1362.

Liu, Y., Zou, C. & Zhang, R. (2008). Empirical Likelihood for the Two-sample Mean Problem. *Statistics and Probability Letters.* **78**, 548–556.

Newey, W.K. & Smith, R.J. (2004). Higher Order Properties of GMM Generalized Empirical Likelihood Estimators. *Econometrica,* **72**, 219–255.

Owen, A. B. (1990). Empirical Likelihood Ratio Confidence Regions. *Annals of Statistics.* **18**, 90–120.

Owen, A. B. (2001). *Empirical Likelihood.* Chapman and Hall, London.

Qin, J. & Lawless, L. (1994). Empirical Likelihood and General Estimating Equations. *Annals of Statistics.* **22**, 300–325.

Rao J.N. K. & Scott A. J. (1981). The Analysis of Categorical Data from Complex Sample Surveys: Chi-Squared Tests for Goodness of Fit and Independence in Two-Way Tables. *Journal of the American Statistical Association.* **76**, 221–230.

Tsao, M. (2004). Bounds on Coverage Probabilities of the Empirical Likelihood Ratio Confidence Regions. *Annals of Statistics.* **32**, 1215–1221.

Tsao, M. & Wu, F. (2013). Empirical Likelihood on the Full Parameter Space. *Annals of Statistics.* **41**, 2176–2196.

Wood, A.T., Do, K.A. & Broom, B.M. (1996). Sequential linearization of empirical likelihood constraints with application to U-statistics. *Journal of Computational and Graphical Statistics.* **54**, 365–85.

Xue L.G. & Wang Q.H.(2012). Empirical likelihood for single-index varying-coefficient models. *Bernoulli.* **18**, 836–856.