

Empirical likelihood inference in the framework of parametric likelihood inference

Mi-Ok Kim*

Center for Epidemiology and Biostatistics, Cincinnati Children's Medical Center
Cincinnati, OH 45229-3039

Abstract

We use a method of parameterized sub-family of probability distributions to connect empirical likelihood (EL) with parametric likelihoods and discuss the EL inference in the framework of parametric likelihood inference. The EL inference benefits from theoretical developments in the parametric case. We illustrate the method with general estimating equations and consider M-type linear regression as an example of practical applications where the proposed method promotes conditional EL inference with parameter orthogonality in place of profile EL inference. Also EL inference with M-type linear regression is a new extension of EL inference where the parameters of interest are not necessarily smooth functionals of distributions.

1 Introduction

Empirical likelihood (EL) is arguably the most successful and general methodology that extends the likelihood inference to a nonparametric setting. It has many desirable statistical properties (see (12) for an overview, and (8; 15; 14) for recent developments with censored data). We use a method of parameterized sub-family of probability distributions to connect the EL with parametric likelihoods and discuss the EL inference in the framework of parametric likelihood inference. We illustrate the method with general estimating equations and consider M-type linear regression as an example of practical applications where the parameterized sub-family of probability distributions method promotes conditional EL inference with parameter orthogonality in place of profile EL inference.

The idea of connecting a nonparametric problem with a parametric one is dated back to (18) where a nonparametric problem is considered as a union of finite dimensional parametric subproblems by assuming that enough knowledge of the unknown state of nature is known to restrict it to a finite-dimensional set. Let θ denotes the parameters of interest. In (18) each of the finite dimensional parametric subproblems is asymptotically a most difficult parametric problem through respective points of θ and crude estimates are made. The inference proceeds as if the true parameter point lays on the path created by the crude estimates.

With general estimating equations, a p -dimensional parameter of interest θ is a functional of a d -variate distribution F via r ($\geq p$) numbers of estimating equations $m(z, \theta) = (m_1(z, \theta), \dots, m_r(z, \theta))^T$

*Corresponding author email: miok.kim@cchmc.org

such that the true value θ_0 is defined by a solution to equations $E_{F_0}(m(z, \theta_0)) = 0$ where F_0 denotes the true underlying distribution. From the point of view of (18), the formulation of estimating equations $m(z, \theta)$ reduces the infinite dimensional parameter space of a nonparametric problem on F_0 to a finite dimensional parameter space problem on θ . Via a Lagrange multipliers method for fixed θ the EL takes an explicit expression where an r -dimensional Lagrange multiplier λ is given as a solution to $g_\theta(\lambda) = 0$ (16). The equation $g_\theta(\lambda) = 0$ corresponds to the most difficult finite dimensional parametric subproblem that exists for the fixed θ in (18). Hence the parameter space of θ is a union of parameter spaces of $g_\theta(\lambda)$ defined at each θ and the empirical likelihood inference is concerned about the path formed by solutions to $g_\theta(\lambda) = 0$. This is the key observation of this paper and details will be presented in Section 2. We show that with $m(z, \theta)$ continuous in θ in the neighborhood of θ_0 , this connection is exact, while it holds asymptotically with a non-continuous $m(z, \theta)$. We consider M-type linear regression as an example of the latter case.

The connection between the EL and a parametric likelihood is also implied in (9) where a score function of the parameter of interest θ is considered as a finite dimensional martingale and a dual likelihood is introduced for inference. With an independent data, $m(z, \theta)$ is a martingale in the number of observations and a dual likelihood $L_\theta(\lambda)$ of a *dual parameter* λ can be constructed at each fixed θ such that $l_\theta(\lambda) = \log(L_\theta(\lambda))$ is a log likelihood in λ for fixed θ and $m(z, \theta) = \partial l_\theta(\lambda) / \partial \lambda |_{\lambda=0}$ (9). It is readily seen that the *dual parameter* λ corresponds with the Lagrange multiplier of (16) and $l_\theta(\lambda)$ coincides with the log EL. This correspondence was considered only as an interesting special case of the dual likelihood with independent data (9). We focus on this connection and bring the EL inference into the realm of parametric likelihood inference.

The proposed method for the EL was considered with random right censored data but the EL ratio statistic were mainly concerned (13). In this paper we apply the method to discuss the EL inference in generality in the frame work of parametric likelihood. Then, we benefit from theoretical developments in the parametric case: we can formerly define mathematical constructs such as score function, hessian matrix and Fisher information matrix with the EL. This contrasts with the implicit use of such constructs in previous works such as (16; 9) without an explanation for how they were obtained. These mathematical constructs facilitate the development of theoretical results such as the efficiency of the maximum empirical likelihood estimator (MELE), profile likelihood inference and parameter orthogonality. Also a heuristic argument can be made that Bartlett correctability is as natural with the EL as with the parametric likelihood. On the other hand in the presence of nuisance parameters, the method facilitates conditional EL inference in place of profile EL under parameter orthogonality.

The proposed method is believed applicable for the EL with random right censoring and/or a increasing dimensional θ . Due to technicalities related to censoring and increasing dimensionality, we discuss these cases separately in subsequent works and focus on the application of the method for the EL of a finite dimensional θ with uncensored independent data. Section 2 introduces the proposed method for the EL in a setting of estimating equations. Section 3 discusses parameter orthogonality with the EL. Section 4 discusses profile and conditional EL inference in presence of nuisance parameters. Finally Section 5 discusses EL inference in M-type regression to illustrate the method for θ being a non-continuous functionale. Simulation results are provided for M-type regression in Section 5.3 where inference by the conditional EL is compared with inference by the profile EL and via bootstrap methods. Throughout the paper we adopt the following notations: for a square matrix ν , ν^- denotes a generalized (Moore's) inverse of ν . $\nu_{(i,j)}$ and $\nu^{(i,j)}$ denote the (i, j) th block sub-matrix of ν and ν^- . Block sub-matrices of a block sub-matrix $\nu_{(i,j)}$ are denoted

by $\nu_{(i,j(k,l))}$. If ν is a vector, $\nu_{(j)}$ denotes the j th element. For a probability distribution F of a random variable Z , $dF(z) = P(Z = z)$. Finally we let $ch(A)$ denote the convex hull of a set A .

2 Method of parameterized sub-family of distribution

We first outline EL inference in a setting of general estimating equations considered in (16). Let z_1, \dots, z_n be i.i.d. observations from a d -variate distribution F_0 . The empirical likelihood for some distribution function F is

$$L_E(F) = \prod_{i=1}^n dF(z_i) = \prod_{i=1}^n p_i. \quad (1)$$

Only distributions with an atom of probability on each z_i have nonzero likelihood and (1) is maximized by the empirical distribution function $\hat{F}_n(z) = n^{-1} \sum_{i=1}^n I_{[z_i \leq z]}$.

Suppose that we are interested in estimating the p -dimensional parameter θ_0 previously defined with respect to F_0 via r -dimensional $m(z, \theta_0)$ ($r \geq p$) such that $E_{F_0}(m(z, \theta_0)) = 0$. A (profile) empirical likelihood function for $\theta \in R^p$ is defined as

$$L_E(\theta) = \max \left\{ \prod_{i=1}^n p_i \left| \sum_{i=1}^n p_i m(z_i, \theta) = 0, \sum_{i=1}^n p_i = 1, p_i \geq 0 \right. \right\}, \quad (2)$$

where $p_i = dF(z_i)$ for some d -variate distribution F . A unique value for the right-hand side of (2) exists under the following convex hull condition:

Condition 1 As $n \rightarrow \infty$, $P(0 \in ch(\{m(z_1, \theta_0), \dots, m(z_n, \theta_0)\})) \rightarrow 1$.

As in (16), we define the maximum empirical likelihood estimate (MELE) for θ and F_0 as follows: $\tilde{\theta} = \operatorname{argmax} L_E(\theta)$ and $\tilde{F}_n(z) = \sum_{i=1}^n \tilde{p}_i I_{[z_i \leq z]}$, where \tilde{p}_i is the maximizer set that defines $L_E(\tilde{\theta})$ such that $L_E(\tilde{\theta}) = \prod_{i=1}^n \tilde{p}_i$. The likelihood ratio is given by $R_E(\theta) = L_E(\theta)/L_E(\tilde{\theta})$.

Given a random sample of $\{z_i\}_{i=1}^n$, we consider following parametric family of probability distributions parameterized by θ and λ :

$$\mathfrak{F}_n^{\theta, \lambda} = \{F_n^{\theta, \lambda} | F_n^{\theta, \lambda}(z) = \sum_{i=1}^n dF_n^{\theta, \lambda}(z_i) 1_{[z_i \leq z]}\}$$

where

$$dF_n^{\theta, \lambda}(z_i) = 1/[n\{1 + \lambda^\top m(z_i, \theta)\}], \quad 0 < dF_n^{\theta, \lambda}(z_i), \quad \sum_{i=1}^n dF_n^{\theta, \lambda}(z_i) \leq 1.$$

The last condition is assumed without loss of generality as we can normalize $F_n^{\theta, \lambda}(z)$ by $dF_n^{\theta, \lambda}(z_i)/\sum_{i=1}^n dF_n^{\theta, \lambda}(z_i)$ if $\sum_{i=1}^n dF_n^{\theta, \lambda}(z_i) > 1$. Inference remains same with the normalized $F_n^{\theta, \lambda}(z)$ by the likelihood principle. Note that $\mathfrak{F}_n^{\theta, \lambda}$ is a family of distributions with an atom of probability on each z_i . Likelihood and log likelihood functions are given by

$$L(\theta, \lambda) = \prod_{i=1}^n dF_n^{\theta, \lambda}(z_i), \quad l(\theta, \lambda) = \sum \log dF_n^{\theta, \lambda}(z_i).$$

Suppose that $m(z, \theta)$ is differentiable in some neighborhood of the true value θ_0 . Score function and hessian matrix are defined as follows:

$$\begin{aligned} S_n(\theta, \lambda) &= (S_{n(1)}(\theta, \lambda), S_{n(2)}(\theta, \lambda)) = (\partial l(\theta, \lambda)/\partial \theta, \partial l(\theta, \lambda)/\partial \lambda) \\ H_n(\theta, \lambda) &= \begin{pmatrix} \partial^2 l(\theta, \lambda)/(\partial \theta \partial \theta^\top) & \partial^2 l(\theta, \lambda)/(\partial \lambda \partial \theta^\top) \\ \partial^2 l(\theta, \lambda)/(\partial \theta \partial \lambda^\top) & \partial^2 l(\theta, \lambda)/(\partial \lambda \partial \lambda^\top) \end{pmatrix}. \end{aligned}$$

Distributions in $\mathfrak{F}_n^{\theta, \lambda}$ satisfy usual regularity conditions and we define Fisher information matrix as follows:

$$I_n(\theta, \lambda) = -E_{F_n^{\theta, \lambda}}[H_n(\theta, \lambda)]. \quad (3)$$

In what follows we let $A(z)$ be a generic notation for integrable functions in this neighborhood of θ_0 .

Lemma 1 Consider a subset of $\mathfrak{F}_n^{\theta, \lambda}$ with $\sum dF_n^{\theta, \lambda}(z_i) = 1$. For fixed θ ,

$$S_{n(2)}(\theta, \lambda) = \frac{1}{n} \sum \frac{m(z_i, \theta)}{1 + \lambda^\top m(z_i, \theta)} = 0. \quad (4)$$

Assume that $E[m(z, \theta_0)m^\top(z, \theta_0)]$ is positive definite, $\partial m/\partial \theta$ is continuous in a neighborhood of the true value θ_0 , and $\|\partial m/\partial \theta\|$ and $\|m\|^3$ are bounded by some integrable function $A(z)$ in this neighborhood. Also assume Condition 1. Then, for each θ within the surface ball of $\|\theta - \theta_0\| \leq n^{-1/3}$, λ that satisfies (4) is uniquely defined.

Let $\lambda(\theta)$ denote the value of λ that satisfies (4) and let $\mathfrak{F}_n^{\theta, \lambda(\theta)}$ denote the sub-family that satisfies (4). Then, $\mathfrak{F}_n^{\theta, \lambda(\theta)}$ is a set of distributions with an atom of probability on each z_i and z_i 's only. Also for fixed θ , $L(\theta, \lambda)$ is a parametric likelihood problem in λ , of which the minimum is defined by $S_{n(2)}(\theta, \lambda) = 0$. Hence $\mathfrak{F}_n^{\theta, \lambda(\theta)}$ is a set of the minimizers at each θ .

Via a Lagrange multipliers method applied to (2) we have $S_{n(2)}(\theta, \lambda) = g_\theta(\lambda)$. It follows that

$$L_E(\theta) = L(\theta, \lambda(\theta)). \quad (5)$$

In the language of (18), for fixed θ $L(\theta, \lambda(\theta))$ corresponds to the most difficult parametric sub-problem at θ . Hence $L_E(\theta)$ is concerned with the parameterized sub-family of distributions $\mathfrak{F}_n^{\theta, \lambda(\theta)}$, each member of which corresponds with the minimum of $L(\theta, \lambda)$ at fixed θ . The correspondences are unique in the neighborhood where $\lambda(\theta)$ are unique. That is, where $\lambda(\theta)$ are unique, $F_n^{\theta, \lambda(\theta)}$ are uniquely defined and correspond with θ one-to-one.

Let $\tilde{\lambda} = \lambda(\tilde{\theta})$. Trivially $L_E(\tilde{\theta}) = L(\tilde{\theta}, \tilde{\lambda}) = \max_{F_n^{\theta, \lambda} \in \mathfrak{F}_n^{\theta, \lambda(\theta)}} L(\theta, \lambda)$ and $R_E(\theta) = L(\theta, \lambda(\theta))/L(\tilde{\theta}, \tilde{\lambda})$. Also $\tilde{F}_n = F_n^{\tilde{\theta}, \tilde{\lambda}}$ where $dF_n^{\tilde{\theta}, \tilde{\lambda}}(z_i) = 1/[n\{1 + \tilde{\lambda}^\top m(z_i, \tilde{\theta})\}]$. If $r = p$, $\hat{F}_n(z) = n^{-1} \sum 1_{[z_i \leq z]} \in \mathfrak{F}_n^{\theta, \lambda(\theta)}$, and $\tilde{\theta} = \hat{\theta}$ and $F_n^{\tilde{\theta}, \tilde{\lambda}} = \hat{F}_n(z)$, where $\hat{\theta}$ is a solution to the equation $n^{-1} \sum m(z_i, \theta) = 0$.

Lemma 2 Assume that $E_{F_0}[m(z, \theta_0)m^\top(z, \theta_0)]$ is positive definite, $\partial m/\partial \theta$ is continuous in a neighborhood of the true value θ_0 , $\|\partial m/\partial \theta\|$ and $\|m\|^3$ are bounded by some integrable function $A(z)$ in this neighborhood, and the rank of $E_{F_0}[\partial m/\partial \theta|_{\theta=\theta_0}]$ is p . Also assume Condition 1. Then, as $n \rightarrow \infty$, with probability 1, $L_E(\theta)$ attains its maximum value at some point $\tilde{\theta}$ in the interior ball $\|\tilde{\theta} - \theta_0\| \leq n^{-1/3}$ and $\tilde{\theta}$ and $\tilde{\lambda}$ satisfy $S_n(\tilde{\theta}, \tilde{\lambda}) = 0$.

Theorem 1 Assume the conditions of Lemma 2 and additionally that $\partial^2 m / \partial \theta \partial \theta^\top$ is continuous in θ and $\|\partial^2 m / \partial \theta \partial \theta^\top\|$ is bounded by some integrable function $A(z)$ in this neighborhood. Then,

$$(a) \quad \begin{aligned} \sqrt{n}(\tilde{\theta} - \theta_0) &\rightarrow N(0, V), & \sqrt{n}(\tilde{\lambda} - 0) &\rightarrow N(0, U) \\ \sqrt{n}(\tilde{F}_n(z) - F_0(z)) &\rightarrow N(0, W(z)) \text{ where} \end{aligned}$$

$$\begin{aligned} V &= \left[E_{F_0} \left(\frac{\partial m}{\partial \theta} \Big|_{\theta=\theta_0} \right)^\top \left(E_{F_0}(m(z, \theta_0)m^\top(z, \theta_0)) \right)^{-1} E_{F_0} \left(\frac{\partial m}{\partial \theta} \Big|_{\theta=\theta_0} \right) \right]^{-1}, \\ U &= [E_{F_0}(m(z, \theta_0)m^\top(z, \theta_0))]^{-1} \\ &\quad \left\{ I - E_{F_0} \left(\frac{\partial m}{\partial \theta} \Big|_{\theta=\theta_0} \right) V E_{F_0} \left(\frac{\partial m}{\partial \theta} \Big|_{\theta=\theta_0} \right)^\top [E_{F_0}(m(z, \theta_0)m^\top(z, \theta_0))]^{-1} \right\}, \\ W(z) &= F_0(z)(1 - F_0(z)) - B(z)UB^\top(z), \quad B(z) = E_{F_0}[m(z, \theta_0)I_{[z_i \leq z]}], \end{aligned}$$

and $\tilde{\theta}$ and $\tilde{\lambda}$ are asymptotically uncorrelated.

$$(b) \quad W_E(\theta_0) = -2 \log R_E(\theta_0) \rightarrow \chi_p^2.$$

(c) The MELE $\tilde{\theta}$ is asymptotically efficient.

Remark 1 Lemma 2 and Theorem 1 are essentially the same results obtained in Lemma 1 and Theorem 1-3 of (16). The method of the parameterized sub-family of distribution does not add any new results in this regard. However, the development of the results are straightforward. For example, (16) utilized elements of S_n and H_n in their derivation of the results without explicitly explaining what they are and how they were obtained. Also the asymptotic efficiency of $\tilde{\theta}$ can be shown straightforwardly by the equivalence of V to the Cramér-Rao lower bound of the estimator. The simplicity of the derivation contrasts with Theorem 3 of (16) which relies on more elaborate semi- and non-parametric results (1; 19).

3 Parameter orthogonality with empirical likelihood

Suppose that $\theta = (\theta_1, \theta_2)$ with $\theta_1 \in R^{p_1}$ and $\theta_2 \in R^{p_2}$ ($p_1 + p_2 = p$). We let $(I_n^{(11)}(\theta, \lambda))_{(ij)}$ denote the i, j th sub-block matrix of the $I_n^{(11)}(\theta, \lambda)$. We define orthogonality between θ_1 and θ_2 analogously to the parametric case as follows: θ_1 is orthogonal to θ_2 if

$$(I_n^{(11)}(\theta, \lambda))_{(12)} = o(1). \quad (6)$$

Similar to parametric case, θ_1 and θ_2 are orthogonal globally if (6) holds for all θ . If (6) holds only for some θ^* , then the orthogonality is local at $\theta = \theta^*$.

Consider dividing $m(z, \theta)$ into $m_1(z, \theta)$ and $m_2(z, \theta)$ where $m_1(z, \theta)$ and $m_2(z, \theta)$ are the estimating equations of dimensions $r_1 (\geq p_1)$ and $r - r_1 (\geq p_2)$ that concern about θ_1 and θ_2 such that θ_{10} and θ_{20} are given by solutions to $E_{F_0}[m_1(z, \theta_1, \theta_{20})] = 0$ and $E_{F_0}[m_2(z, \theta_{10}, \theta_2)] = 0$ respectively.

Lemma 3 Assume the conditions of Theorem 1. Suppose that $\hat{\theta}_n^*$ is a \sqrt{n} -consistent estimator of θ_0 . Then, (6) holds for $\hat{\theta}_n^*$ in probability if and only if $m_1(z, \hat{\theta}_n^*)$ and $m_2(z, \hat{\theta}_n^*)$ are asymptotically uncorrelated.

This is consistent with parametric case where parameter orthogonality means that the relevant components of the score statistic are uncorrelated and orthogonal projection is defined with respect to the expected Fisher information matrix. Hence a general procedure of orthogonalization such as (2) or reference therein is applicable. The utility of the orthogonality for inference in presence of nuisance parameter will be discussed in generality in Section 4.2. EL inference with M-type regression is an example of practical applications where a simple linear transformation exists to achieve the orthogonality between vectors of parameters of interest and nuisance parameters. Details will be presented in Section 5.2.

4 Empirical likelihood inference in presence of nuisance parameter(s)

Suppose that we are only interested in the inference about the p_2 -dimensional parameter θ_2 and θ_1 is a (vector of) nuisance parameter(s). This has been a well-studied problem in the parametric case. A standard approach is via profiling. Results are provided with the EL (18). We show that similar results are yield yet in more straightforward and simpler manner with the proposed method (Section 4.1).

A widely used alternative approach is via plugging-in: maximum likelihood estimates are plugged-in for the nuisance parameters in the likelihood function and the resulting profile likelihood is examined as a function of the parameters of interest. This avoids profiling but can give inconsistent or inefficient estimates for problems with large numbers of nuisance parameters, which suggests that it may not be close to optimal for a subset of nuisance parameters (2). The “plug-in” method was considered with the EL and asymptotic results for the “plug-in” profile EL ratio (4). With the proposed method we show that the “plug-in” profile EL ratio has the usual asymptotic χ^2 distribution under the parameter orthogonality.

An alternative approach is via conditioning under parameter orthogonality. In the parametric case, likelihood was conditioned on the maximum likelihood estimates of orthogonalized nuisance parameters and the conditional profile likelihood function was used for inference (2). We take a similar approach with the EL. A caveat to the results in the following section is that with the EL, the “plug-in” and conditioning method produce the same EL. Details are presented in Section 4.2.

4.1 Inference by Profile Empirical Likelihood

Via profiling we have the following EL and associated likelihood ratio statistic:

$$\bar{L}_E(\theta_2) = \max_{\theta_1} L_E(\theta_1, \theta_2), \quad \bar{R}_E(\theta_2) = \bar{L}_E(\theta_2) / \max_{\theta_2} \bar{L}_E(\theta_2). \quad (7)$$

Trivially $\tilde{\theta}_2 = \operatorname{argmax}_{\theta_2} \bar{L}_E(\theta_2)$ and the MELE of θ_2 remains same whether profiling the EL or not. It follows that $\bar{R}_E(\theta_2) = \bar{L}_E(\theta_2) / \bar{L}_E(\tilde{\theta}_2)$.

Let $\bar{\theta}_1(\theta_2) = \operatorname{argmax}_{\theta_1} L_E(\theta_1, \theta_2)$. Via the method of parameterized sub-family of distribution, the profile likelihood is concerned with a subset of $\mathfrak{F}_n^{\theta, \lambda}$ that are defined by a solution to (4) at each fixed θ_2 with $\theta_1 = \bar{\theta}_1(\theta_2)$. Let $\bar{\lambda}(\theta_2)$ denote a value of λ that satisfies (4) with $\theta_1 = \bar{\theta}_1(\theta_2)$. Then,

$\mathfrak{F}_n^{\bar{\theta}_1(\theta_2), \theta_2, \bar{\lambda}(\theta_2)}$ denotes the sub-family of the parameterized probability distributions that $\bar{L}_E(\theta_2)$ is concerned with. We have

$$\bar{L}_E(\theta_2) = L(\bar{\theta}_1(\theta_2), \theta_2, \bar{\lambda}(\theta_2)), \quad \bar{R}_E(\theta_2) = L(\bar{\theta}_1(\theta_2), \theta_2, \bar{\lambda}(\theta_2)) / L(\tilde{\theta}_1, \tilde{\theta}_2, \tilde{\lambda}),$$

which is useful proving the following theorem.

Theorem 2 *Assume the conditions of Theorem 1. For θ_2 within the surface ball of $\|\theta_2 - \theta_{20}\| \leq n^{-1/2}$,*

- (a) $S_n^{-\theta_2}(\bar{\theta}_1(\theta_2), \theta_2, \bar{\lambda}(\theta_2)) = 0$ for all n , where $S_n^{-\theta_2}(\theta, \lambda) = (\partial l(\theta, \lambda) / \partial \theta_1, \partial l(\theta, \lambda) / \partial \lambda)$.
- (b) $\|\bar{\theta}_1(\theta_2) - \theta_{10}\| = O_p(n^{-1/2})$.
- (c) $\bar{W}_E(\theta_{20}) = -2 \log(\bar{R}_E(\theta_{20})) \rightarrow \chi_{p_2}^2$ as $n \rightarrow \infty$.

Theorem 2 (c) is similar to Corollary 5 of (16) but again the derivation is much straightforward with the proposed method.

4.2 Conditional Empirical Likelihood

We first define conditional empirical likelihood. Conditional EL of θ_2 given $\theta_1 = \theta_1^*$ and associated likelihood ratio statistic are given as follows:

$$L_E(\theta_2 | \theta_1^*) = \max \left\{ \prod_{i=1}^n p_i \mid \sum p_i m(z_i, \theta_1^*, \theta_2), \sum p_i = 1, p_i \geq 0 \right\} \quad (8)$$

$$R_E(\theta_2 | \theta_1^*) = L_E(\theta_2 | \theta_1^*) / \max_{\theta_2} L_E(\theta_2 | \theta_1^*).$$

One may consider the following definition of conditional EL.

$$L_E(\theta_2 | \theta_1^*) = \frac{\max \{ \prod_{i=1}^n p_i \mid \sum p_i m(z_i, \theta_1^*, \theta_2), \sum p_i = 1, p_i \geq 0 \}}{\max_{\theta_2} \max \{ \prod_{i=1}^n p_i \mid \sum p_i m(z_i, \theta_1^*, \theta_2), \sum p_i = 1, p_i \geq 0 \}}.$$

By this definition, $R_E(\theta_2 | \theta_1^*) = L_E(\theta_2 | \theta_1^*)$. (8) avoids the redundancy.

Note that $L_E(\theta_2 | \theta_1^*) = L_E(\theta_1^*, \theta_2)$. That is, conditioning and plug-in method produce same EL. We call (8) conditional EL for the following heuristic reasons. Conventionally plug-in method considers “good” values to fill in for θ_1 , for example, values of consistent estimators. In this paper θ_1 is replaced with not necessarily “good” values. For example, $\bar{L}(\theta_2) = L_E(\bar{\theta}(\theta_2), \theta_2)$ where $\bar{\theta}_1(\theta_2)$ is not necessarily close to θ_{10} unless θ_2 is close to θ_{20} . Also the name provides some consistency with parametric case in Bayesian analysis. EL can be considered as an alternative to a parametric likelihood in Bayesian analysis as a way to reflect uncertainties in the likelihood specification and to achieve some robustness in the analysis (6). With the EL, the posterior distributions need to be computed via Markov chain Monte Carlo (MCMC) methods which use conditional likelihood in the parametric case.

Let us consider inference by conditional EL for θ_2 under parameter orthogonality. An obvious candidate for the conditioning value of θ_1 is $\tilde{\theta}_1$ and we have

$$L_E(\theta_2|\tilde{\theta}_1) = \max\left\{\prod p_i \mid \sum p_i m(z_i, \tilde{\theta}_1, \theta_2) = 0, \sum p_i = 1, p_i \geq 0\right\}. \quad (9)$$

Trivially $\tilde{\theta}_2 = \operatorname{argmax}_{\theta_2} L_E(\theta_2|\tilde{\theta}_1)$ and the MELE of θ_2 remains same. Via the method of parameterized sub-family of distribution, the conditional EL (9) is concerned with a subset of $\mathfrak{F}_n^{\theta, \lambda}$, of which each member is defined by a solution to $S_{n(2)}(\tilde{\theta}_1, \theta_2, \lambda) = 0$ for fixed θ_2 . We let $\tilde{\lambda}(\theta_2)$ denote the value of λ that satisfies $S_{n(2)}(\tilde{\theta}_1, \theta_2, \lambda) = 0$. Then, $\mathfrak{F}_n^{\tilde{\theta}_1, \theta_2, \tilde{\lambda}(\theta_2)}$ denotes the sub-family of the parameterized probability distributions that $L_E(\theta_2|\tilde{\theta}_1)$ is concerned with. We have $L_E(\theta_2|\tilde{\theta}_1) = L(\tilde{\theta}_1, \theta_2, \tilde{\lambda}(\theta_2))$, which helps deriving the following results.

Theorem 3 *Assume the conditions of Theorem 1. Let $W_E(\theta_2|\tilde{\theta}_1) = -2\log R_E(\theta_2|\tilde{\theta}_1)$. For θ_2 within the surface ball of $\|\theta_2 - \theta_{20}\| \leq n^{-1/2}$,*

(i) *if the orthogonality condition (6) holds,*

$$\|\tilde{\theta}_1(\theta_2) - \tilde{\theta}_1\| = O_p(n^{-1}) \quad , \quad |\log \bar{L}_E(\theta_2) - \log L_E(\theta_2|\tilde{\theta}_1)| = o_p(1) \quad ,$$

and $|\overline{W}_E(\theta_2) - W_E(\theta_2|\tilde{\theta}_1)| = o_p(1)$. Furthermore $W_E(\theta_{20}|\tilde{\theta}_1) \rightarrow \chi_{p_2}^2$ as $n \rightarrow \infty$.

(ii) *If (6) does not hold,*

$$\|\tilde{\theta}_1(\theta_2) - \tilde{\theta}_1\| = O_p(n^{-1/2}) \quad , \quad |\log \bar{L}_E(\theta_2) - \log L_E(\theta_2|\tilde{\theta}_1)| = O_p(1) \quad ,$$

and $|\overline{W}_E(\theta_2) - W_E(\theta_2|\tilde{\theta}_1)| = O_p(1)$. Furthermore $W_E(\theta_{20}|\tilde{\theta}_1) \rightarrow d_1\chi_{1,1}^2 + \dots + d_r\chi_{1,r}^2$ where the weights, d_i , are the eigenvalues of \tilde{D} (defined in (??) in the appendix) and $\chi_{1,i}^2$ are independent chi-square distributions with one degree of freedom.

When the asymptotic properties of $W_E(\theta_{20}|\tilde{\theta}_1)$ are concerned, similar results exist in the literature as shown in the remarks below. Theorem 3 is more detailed and extensive, as it additionally provides results about $W_E(\theta_2|\tilde{\theta}_1)$ for θ_2 in the neighborhood of the true value and conditions under which $W_E(\theta_2|\tilde{\theta}_1)$ is a good approximation of $\overline{W}_E(\theta_2)$. This is useful when confidence interval/region is concerned as inference by the profile conditional EL is computationally much more efficient than profiling EL.

Remark 2 *In a univariate case with mean μ and variance σ^2 the Corollary 1 of (10) showed that $-2\log R_E(\mu_n)$ with $\mu_n = \mu + \tau\sigma n^{-1/2}$ converges to a non-central chi-squared distribution with the non-centrality parameter τ^2 . A complementary result of that is $\mu_n = \mu + \tau\sigma n^{-\gamma}$ with $\gamma > 1/2$ will lead to χ_1^2 . Theorem 3 can be seen as a version of the Corollary 1 of (10) and its complementary result with estimating equations. Note that $L_E(\theta_2|\tilde{\theta}_1) = \bar{L}_E(\theta_2 + \epsilon_n(\tilde{\theta}_1 - \bar{\theta}_1(\theta_2)))$ with the perturbation $\epsilon_n(\tilde{\theta}_1 - \bar{\theta}_1(\theta_2))$ as a function of $\tilde{\theta}_1 - \bar{\theta}_1(\theta_2)$. The theorem provides a condition under which the perturbation is at the magnitude of $o_p(n^{-1/2})$ for θ_2 within the surface ball of $\|\theta_2 - \theta_{20}\| \leq n^{-1/2}$.*

Remark 3 Theorem 2.1 of (4) reported similar results concerning $W_E(\theta_{20}|\tilde{\theta}_1)$. Their results are general in the sense that the nuisance parameters can be of infinite dimension and any consistent estimates of θ_1 are allowed for the plugging-in. However, they are limited to a just determined case. That is, in the given setting, the results of (4) are fully applicable only when $m(z, \theta)$ are reduced to p_2 number of equations that concerns with θ_2 via transformation (see Theorem 4 and related discussion below). However, the main ideas are valid: the asymptotic distribution of the likelihood ratio statistic depends on the asymptotic variance of $n^{-1/2} \sum m(z_i, \tilde{\theta}_1, \theta_{20})$ and the limit of $n^{-1} \sum m(z_i, \tilde{\theta}_1, \theta_{20})m^\top(z_i, \tilde{\theta}_1, \theta_{20})$. If they coincide, the limiting distribution is the usual $\chi_{p_2}^2$, which is the case under the orthogonality. Without the orthogonality, the limiting distribution is a non-central $\chi_{p_2}^2$. Lemma A.3 verifies this in the appendix. However, the expression of \tilde{D} is not same as the product of the variance of $n^{-1/2} \sum m(z_i, \tilde{\theta}_1, \theta_{20})$ with the inverse of the limit of $n^{-1} \sum m(z_i, \tilde{\theta}_1, \theta_{20})m^\top(z_i, \tilde{\theta}_1, \theta_{20})$ as proposed in Theorem 2.1 of (4).

Remark 4 Note that $\bar{L}_E(\theta_{20}) = L_E(\bar{\theta}_1(\theta_{20}), \theta_{20})$. Hence Theorem 2 can be seen as a special case of Theorem 2.1 of (4) with $\bar{\theta}_1(\theta_{20})$ plugged in for θ_1 . Lemma A.4 in the appendix verifies that the asymptotic variance of $n^{-1/2} \sum m(z_i, \bar{\theta}_1(\theta_{20}), \theta_{20})$ and the limit of $n^{-1} \sum m(z_i, \bar{\theta}_1(\theta_{20}), \theta_{20})m^\top(z_i, \bar{\theta}_1(\theta_{20}), \theta_{20})$ coincide always, which confirms the results of Theorem 2 using the results of (4).

Remark 5 We show that Condition 1 is sufficient for the usual convex hull conditions for $\bar{L}_E(\theta_2)$ and $L_E(\theta_2|\tilde{\theta}_1)$. The convex hull conditions are respectively that as $n \rightarrow \infty$,

$$\begin{aligned} P(0 \in \text{ch}(\{m(z_1, \bar{\theta}_1(\theta_{20}), \theta_{20}), \dots, m(z_n, \bar{\theta}_1(\theta_{20}), \theta_{20})\})) &\rightarrow 1, \\ P(0 \in \text{ch}(\{m(z_1, \tilde{\theta}_1, \theta_{20}), \dots, m(z_n, \tilde{\theta}_1, \theta_{20})\})) &\rightarrow 1. \end{aligned}$$

This is satisfied under Condition 1 as $m(z, \bar{\theta}_1(\theta_{20}), \theta_{20}) \rightarrow m(z, \theta_0)$ and $m(z, \tilde{\theta}_1, \theta_{20}) \rightarrow m(z, \theta_0)$ uniformly as $n \rightarrow \infty$ by the consistency of $\bar{\theta}_1(\theta_{20})$ and $\tilde{\theta}_1$ and the continuity of $m(z, \theta)$ in the neighborhood of θ_0 .

We can easily see from Lemma 3 that $m_1(z_i, \tilde{\theta}_1, \theta_2)$ and $m_2(z_i, \tilde{\theta}_1, \theta_2)$ are asymptotically uncorrelated for θ_2 within the surface ball of $\|\theta_2 - \theta_{20}\| \leq n^{-1/2}$ under the orthogonality. We consider the following EL for the inference of θ_2 .

$$\tilde{L}_E(\theta_2) = \max\left\{\prod p_i \mid \sum p_i m_2(z_i, \tilde{\theta}_1, \theta_2) = 0, \sum p_i = 1, p_i \geq 0\right\}.$$

Similarly to the earlier sections, we consider the following family of distributions parameterized by θ_2 and $\gamma \in R^{r-1}$:

$$\mathfrak{F}_n^{\theta_2, \gamma} = \{F_n^{\theta_2, \gamma} \mid F_n^{\theta_2, \gamma}(z) = \sum_{i=1}^n dF_n^{\theta_2, \gamma}(z_i) 1_{[z_i \leq z]}\}, \quad (10)$$

where $dF_n^{\theta_2, \gamma}(z_i) = 1/[n\{1 + \gamma^\top m_2(z_i, \tilde{\theta}_1, \theta_2)\}]$, $0 < dF_n^{\theta_2, \gamma}(z_i)$, $\sum_{i=1}^n dF_n^{\theta_2, \gamma}(z_i) \leq 1$. Then, $\tilde{L}_E(\theta_2)$ is concerned with $\mathfrak{F}_n^{\theta_2, \gamma(\theta_2)}$, a subset of $\mathfrak{F}_n^{\theta_2, \gamma}$ with $\sum dF_n^{\theta_2, \gamma}(z_i) = 1$ where $\gamma(\theta_2)$ denotes a solution to

$$S_{n(2)}(\theta_2, \gamma) = \frac{1}{n} \sum \frac{m_2(z, \tilde{\theta}_1, \theta_2)}{1 + \gamma^\top m_2(z_i, \tilde{\theta}_1, \theta_2)} = 0. \quad (11)$$

When $L(\theta_2, \gamma)$ denote the associated likelihood, $\tilde{L}_E(\theta_2) = L(\theta_2, \gamma(\theta_2))$, by which the following results are derived straightforwardly.

Theorem 4 Assume the conditions of Theorem 1. If the orthogonality condition (6) holds,

$$\tilde{L}_E(\theta_2) = \bar{L}_E(\theta_2) + o_p(1)$$

for θ_2 within the surface ball of $\|\theta_2 - \theta_{20}\| \leq n^{-1/2}$. Moreover, $\tilde{W}_E(\theta_{20}) \rightarrow \chi_{p_2}^2$ as $n \rightarrow \infty$, where $\tilde{W}_E(\theta_2) = -2 \log[\tilde{L}_E(\theta_2) - \tilde{L}_E(\hat{\theta}_2)]$.

5 Empirical likelihood for M-type linear regression

We first overview two different types of linear models and EL formulations, namely *casewise* and *residual* EL, by adapting the discussion of (20) to M-type regression. Let $\rho(u)$ be a convex loss function with $\psi(u) = \rho'(u)$ that is not necessarily continuous at zero but monotonically non-decreasing function. Following (3), the main characteristics of two different types of linear models concepts can be summarized as follows.

Correlation Model: We observe i.i.d. random vectors (Y_i, X_i) , $i = 1, \dots, n$, from a common joint distribution F_0 where the one-dimensional responses Y_i and the p -dimensional covariates X_i are related by a linear regression $Y_i = X_i^\top \theta + e_i$ where the true value of the parameter θ_0 is defined as a solution to $E_{F_0}[\psi(Y_i - X_i^\top \theta) X_i] = 0$ if $E_{F_0}[X^\top X]$ is of full rank. An M-type estimator of θ_0 is given by a minimizer of $\sum \rho(Y_i - X_i^\top \theta)$. A correlation model is used if, for example, the goal is to estimate a regression plane on the basis of a simple random sample of multivariate observations.

Regression Model: The covariates x_i , $i = 1, \dots, n$, are fixed, constant, and observable p -dimensional (column) vectors, forming a matrix of full rank. At fixed x_i , we observe the responses Y_i that are independent random variables with distributions having location parameter $x_i^\top \beta$ such that $Y_i = x_i^\top \beta + e_i$ where e_i are i.i.d. mean zero random variables with a common distribution F_e . The true value of the parameter β_0 is defined as a solution to $E_{F_e}[\psi(Y_i - x_i^\top \beta) x_i] = 0$ for all i . An M-type estimator of β_0 is given by a minimizer of $\sum \rho(Y_i - x_i^\top \beta)$. In general, regression models are used if measurement error of the response is the main source of uncertainty.

A major difference is found in the assumptions on the errors e_i : in the regression model, the errors are assumed homoscedastic, whereas the conditional distribution of the error term, given X , in the correlation model is allowed to depend on X . That is, the errors in the correlation model can be heteroscedastic. If bootstrap methods are applied in a correlation model, the appropriate resampling approach is resampling the vectors, and not the residuals, while an appropriate bootstrap for the regression model would resample the centered residuals. The names of two different EL formulations, *casewise* and *residual* EL, come from this analogy to bootstrap methods.

Suppose that $\psi(u)$ is a continuous function at zero for expository purpose. In the correlation model the probability distribution under consideration is the $(p+1)$ dimensional joint distribution and EL for θ is constructed *casewise*:

$$L_E(\theta) = \max \left\{ \prod_{i=1}^n p_i \mid \sum p_i \psi(y_i - x_i^\top \theta) x_i = 0, \sum p_i = 1, p_i \geq 0 \right\} \quad (12)$$

where $p_i = dF((y_i, x_i)) = P((Y, X) = (y_i, x_i))$ for some $(p+1)$ dimensional distribution function F . On the other hand in the regression model the univariate error distribution F_e is of interest and

EL for β is constructed *residual-wise*: when $r_i(\beta) = y_i - x_i^\top \beta$,

$$L_E(\beta) = \max\left\{\prod_{i=1}^n p_i \mid \sum p_i \psi(r_i(\beta)) x_i = 0, \sum p_i = 1, p_i \geq 0\right\} \quad (13)$$

where $p_i = dF(r_i(\beta)) = P(e = r_i(\beta))$ for some univariate distribution function F .

When there is no censoring, these EL functions and their ratios have identical values under either regression or correlation model (11). Thus, for uncensored data, the two different models mainly pertain to two different sets of assumptions under which the EL version of Wilks theorem holds. They yield identical p-values or confidence intervals, though interpreted differently. Under censoring, however, the two concepts lead to different estimators and empirical likelihood ratios. We refer to (20) for more details of comparisons between two models in censored case. We defer the discussions of a method of parameterized sub-family of distributions with EL and conditional EL inference for censored case in a subsequent work.

The regression model in this paper is different from the regression model of (11) where e_i mean zero independent yet not necessarily identically distributed random variables. With the heteroscedastic errors regression model, an EL function and its ratio are identically defined as to (13). Also all the results in the later section will hold similarly with slightly modified conditions. However, the interpretation is different. In (13) the EL is essentially concerned with the common underlying error distribution F_e and $F(s) = \sum I_{[s \leq r_i(\beta)]} dF(r_i(\beta))$ are considered as a candidate or an estimate of F_e . With the heteroscedastic model, the distribution under consideration is the mixture distribution $\lim_{n \rightarrow \infty} n^{-1} \sum F_{x_i}(s)$ where $F_{x_i}(s) = P(e_i < s | x_i)$.

From (12) note that the correlation model can be seen as a case of general estimation equations problems where $Z = (Y, X^\top)^\top$, $z_i = (y_i, x_i^\top)^\top$ and $m(z_i, \theta) = \psi(y_i - x_i^\top \theta) x_i$. Then, the results in the previous sections hold for the *casewise* EL and its ratio if $\psi(u)$ is continuous at zero. Hence it suffices to show similar results for $\psi(u)$ that are not continuous at zero to complete the theories for the *casewise* EL. We will discuss the discontinuous $\psi(u)$ in the next section for the *residual* EL in the regression model, which requires different sets of conditions and need to be explicated. We recycle most of the notations from the previous sections.

5.1 M-type linear regression model and residual EL inference

We first formally define an M-type estimator $\hat{\beta}_n$ with estimating equations: $\hat{\beta}_n$ is any estimator satisfying

$$n^{-1} \sum_{i=1}^n \psi(r_i(\beta)) x_i = \epsilon_n, \quad (14)$$

where $\epsilon_n \in R^p$ with $\|\epsilon_n\| = o(n^{-1/2})$. The rate is given by Lemma A.2 of (17) and implies that $\hat{\beta}_n$ is unique in the order of $n^{-1/2}$ with $\psi(u)$ that are not continuous at 0. With $\psi(u)$ continuous at 0, $\epsilon_n = 0$. ϵ_n is ψ dependent but the relationship is suppressed in the notation. Mean regression corresponds to a case with $\rho(u) = u^2$ and $\psi(u) = u$. When $\rho(u) = \rho_\tau(u) = u(\tau - I_{[u < 0]})$ for some $\tau \in (0, 1)$, $\psi(u) = \psi_\tau(u) = \tau - I_{[u < 0]}$ and this corresponds to quantile regression (see (5) for an overview). When $\rho(u) = |u|$, then $\psi(u) = \text{sign}(u)$ and it corresponds to least absolute deviation regression or median regression, which is a special case of quantile regression with $\tau = 0.5$.

We define a *residual-wise* (profile) EL for β as follows:

$$L_E(\beta) = \max\left\{\prod_{i=1}^n p_i \mid \sum p_i \psi(r_i(\beta)) x_i = \epsilon_n, \sum p_i = 1, p_i \geq 0\right\}, \quad (15)$$

where $p_i = dF(r_i(\beta))$ for some univariate distribution F . From (11) the usual convex hull condition is as follows:

Condition 2 As $n \rightarrow \infty$, $ch(N^+) \cap ch(N^-) \neq \emptyset$ with probability tending to 1, where $N^+ = \{r_i | r_i(\beta_0) > 0\}$ and $N^- = \{r_i | r_i(\beta_0) < 0\}$.

We define a maximum empirical likelihood estimate (MELE) for β and F_e by $\tilde{\beta} = \operatorname{argmax} L_E(\beta)$ and $\tilde{F}_e(e) = \sum_{i=1}^n \tilde{p}_i I_{[r_i(\tilde{\beta}) \leq e]}$, where \tilde{p}_i is the maximizer set that defines $L_E(\tilde{\beta})$ such that $L_E(\tilde{\beta}) = \prod_{i=1}^n \tilde{p}_i$. The likelihood ratio is given by $R_E(\beta) = L_E(\beta)/L_E(\tilde{\beta})$.

We consider following parametric family of univariate probability distributions:

$$\mathfrak{F}_n^{\beta, \lambda} = \left\{ F_n^{\beta, \lambda} \mid F_n^{\beta, \lambda}(e) = \sum_{i=1}^n dF_n^{\beta, \lambda}(r_i(\beta)) 1_{[r_i(\beta) \leq e]} \right\}$$

where

$$dF_n^{\beta, \lambda}(r_i(\beta)) = 1/[n\{1 + \lambda^\top x_i \psi(r_i(\beta))\}], \quad 0 < dF_n^{\beta, \lambda}(r_i(\beta)), \quad \sum dF_n^{\beta, \lambda}(r_i(\beta)) \leq 1.$$

The last condition is similarly assumed without loss of generality. $\mathfrak{F}_n^{\beta, \lambda}$ is a set of distributions with an atom of probability on each $r_i(\beta)$. Likelihood and log likelihood functions are given by $L(\beta, \lambda) = \prod_{i=1}^n dF_n^{\beta, \lambda}(r_i(\beta))$ and $l(\beta, \lambda) = \sum \log dF_n^{\beta, \lambda}(r_i(\beta))$ respectively.

We assume following conditions.

Condition 3 $\psi(u)$ is a monotonically nondecreasing function with strict monotonicity about 0 such that $\psi(u^-) < \psi(u^+)$ at $u = 0$ and $E_{F_e}[\psi(e)] = 0$ and $0 < E_{F_e}[\psi^2(e)] < \infty$. For some $b_1 > 0$,

$$E_{F_e}[\psi(e + s)] = b_1 s + o(s) \text{ as } s \rightarrow 0. \quad (16)$$

In some cases, we assume a stronger version as follows:

$$E_{F_e}[\psi(e + s)] = b_1 s + O(s^2) \text{ as } s \rightarrow 0. \quad (17)$$

Condition 4 There exist positive constants, b_2, b_3 , and b_4 such that $E_{F_e}[\psi(e + s) - \psi(e)]^2 \leq b_2 |s|$ and $|\psi(u + s) - \psi(u)| \leq b_3$ for all $|s| \leq b_4$ and $u \in R$.

Condition 5 $f_e(0) > 0$ and $f_e(u)$ is Lipschitz in a neighborhood of zero, where f_e denote the density function of the error variables.

Condition 6 $n^{-1/2} \max |x_{ij}| = o(1)$ and $n^{-2} \sum ||x_i||^4 E_{F_e}[\psi^4(e)] \rightarrow 0$ as $n \rightarrow \infty$. There exists a positive definite matrix Σ_x of rank p such that $n^{-1} \sum x_i x_i^\top \rightarrow \Sigma_x$.

Condition 5 is added for quantile regression.

Define $S_n(\beta, \lambda)$ and $H_n(\beta, \lambda)$ as follows:

$$\begin{aligned} S_n(\beta, \lambda) &= \left(b_1 \sum \frac{\lambda^\top x_i x_i^\top}{1 + \lambda^\top x_i \psi(r_i(\beta))}, - \sum \frac{\psi(r_i(\beta)) x_i^\top}{1 + \lambda^\top x_i \psi(r_i(\beta))} \right) \\ H_{n(11)} &= b_1^2 \sum \frac{(x_i x_i^\top \lambda)(\lambda^\top x_i x_i^\top)}{[1 + \lambda^\top x_i \psi(r_i(\beta))]^2} \\ H_{n(12)} &= H_{n(21)}^\top = b_1 \sum \frac{x_i x_i^\top}{[1 + \lambda^\top x_i \psi(r_i(\beta))]} - b_1 \sum \frac{(\psi(r_i(\beta)) x_i)(\lambda^\top x_i x_i^\top)}{[1 + \lambda^\top x_i \psi(r_i(\beta))]^2} \\ H_{n(21)} &= \sum \frac{\psi^2(r_i(\beta)) x_i x_i^\top}{[1 + \lambda^\top x_i \psi(r_i(\beta))]^2}, \end{aligned}$$

Let $I_n(\beta, \lambda) = -E_{F_n^{\beta, \lambda}}[H_n(\beta, \lambda)]$.

Lemma 4 Consider a subset of $\mathfrak{F}_n^{\beta, \lambda}$ where $\sum_{i=1}^n dF_n^{\beta, \lambda}(r_i(\beta)) = 1$. For fixed β ,

$$S_{n(2)}(\beta, \lambda) = \frac{1}{n} \sum \frac{\psi(r_i(\beta)) x_i}{1 + \lambda^\top x_i \psi(r_i(\beta))} = 0. \quad (18)$$

Assume Condition 3 with (16) and Conditions 4-6. Then, for each β within the surface ball of $\|\beta - \beta_0\| \leq n^{-1/2}$, λ that satisfies (18) is uniquely defined.

We let $\lambda(\beta)$ denote the value of λ that satisfies (18) and let $\mathfrak{F}_n^{\beta, \lambda(\beta)}$ denote the subset with $\sum dF_n^{\beta, \lambda}(r_i(\beta)) = 1$. It is a set of distributions with an atom of probability on each $r_i(\beta)$ and $r_i(\beta)$'s only. Also for fixed β , $L(\beta, \lambda)$ is a likelihood in λ and the minimum is given by (18). Hence $\mathfrak{F}_n^{\beta, \lambda(\beta)}$ is a set of minimizers of $L(\beta, \lambda)$ at each β .

If ψ is continuous at 0, $L_E(\beta) = L(\beta, \lambda(\beta))$ and $\log(L_E(\beta)) = l(\beta, \lambda(\beta))$. If not,

$$\log L_E(\beta) = l(\beta, \lambda(\beta)) + o_p(1)$$

for β within the surface ball of $\|\beta - \beta_0\| \leq n^{-1/2}$ (Lemma A.5(iii) in the appendix). Heuristic arguments are as follows: via Lagrange multiplier method, for fixed β , the right-side of (15) is explicitly given by $\prod p_i(\beta)$ where $p_i(\beta) = [n\{1 + (\lambda^*(\beta))^\top (x_i \psi(r_i(\beta)) - \epsilon_n)\}]^{-1}$ with $\lambda^*(\beta)$ given as the solution to

$$\sum p_i(\beta) \psi(r_i(\beta)) x_i = \epsilon_n. \quad (19)$$

If $\psi(u)$ is a continuous function at zero, (19) is equivalent to (18), $\lambda^*(\beta) = \lambda(\beta)$ and $p_i(\beta) = dF_n^{\beta, \lambda}(r_i(\beta))$. If not, (19) is equivalent to (18) only in a limit sense. In the language of (18), $L_E(\beta)$ is concerned with $\mathfrak{F}_n^{\beta, \lambda(\beta)}$ in the limit sense that each member of $\mathfrak{F}_n^{\beta, \lambda(\beta)}$ is obtained as a solution to a p -dimensional subproblem that is equivalent to (19) in the limit at each β . Let $\tilde{\lambda} = \lambda(\tilde{\beta})$.

Theorem 5 Assume Condition 3 with (16) and Conditions 4-6. Then,

(a) $\|\tilde{\beta} - \beta_0\| = O_p(n^{-1/2})$. If $\psi(u)$ is continuous at 0, $\|S_n(\tilde{\theta}, \tilde{\lambda})\| = 0$. Otherwise

$$\|S_n(\tilde{\theta}, \tilde{\lambda})\| = o_p(n\|\epsilon_n\|). \quad (20)$$

- (b) $\sqrt{n}(\tilde{\beta} - \beta_0) \rightarrow N(0, (\sigma_e^2/b_1^2)\Sigma_x^-)$ and $\|\tilde{\lambda}\| = o_p(n^{-1/2})$.
- (c) $W_E(\beta_0) = -2\log R_E(\beta_0) \rightarrow \chi_p^2$.

5.2 Conditional residual EL inference with M-type linear regression model

We consider EL inference for a part of β_0 as follows: $Y_i = x_{1i}^\top\beta_{10} + x_{2i}^\top\beta_{20} + e_i$ where $\beta_{20} \in R^{p_2}$ is of inference interest. With a standard approach via profiling, we have $\bar{L}_E(\beta_2) = \max_{\beta_1} L_E(\beta_1, \beta_2)$ and $\tilde{\beta}_2 = \operatorname{argmax}_{\beta_2} \bar{L}_E(\beta_2)$, and the likelihood ratio is $\bar{R}_E(\beta_2) = \bar{L}_E(\beta_2)/\bar{L}_E(\tilde{\beta}_2)$. We have the following results.

Theorem 6 *Assume the conditions of Theorem 5.*

$$\bar{W}_E(\beta_{20}) = -2\log(\bar{R}_E(\beta_{20})) \rightarrow \chi_{p_2}^2 \text{ as } n \rightarrow \infty.$$

Alternatively we consider following conditional EL

$$L_E(\beta_2|\tilde{\beta}_1) = \max\{\prod p_i | \sum p_i \psi(r_i(\tilde{\beta}_1, \beta_2))x_i = \epsilon_n, \sum p_i = 1, p_i \geq 0\}.$$

With respect to the M-type regression, parameter orthogonality means $\Sigma_{x(12)} = 0$. Then, the following results show that inference by $L_E(\beta_2|\tilde{\beta}_1)$ is asymptotically equivalent to the inference by $\bar{L}_E(\beta_2)$.

Corollary 1 *Assume the conditions of Theorem 5 with (17) instead of (16). Let $W_E(\beta_2|\tilde{\beta}_1) = -2\log R_E(\beta_2|\tilde{\beta}_1)$. If $\Sigma_{x(12)} = 0$, for β_2 within the surface ball of $\|\beta_2 - \beta_{20}\| \leq n^{-1/2}$,*

$$|\log \bar{L}_E(\beta_2) - \log L_E(\beta_2|\tilde{\beta}_1)| = o_p(1) \quad \text{and} \quad |\bar{W}_E(\beta_2) - W_E(\beta_2|\tilde{\beta}_1)| = o_p(1).$$

Moreover $W_E(\beta_{20}|\tilde{\beta}_1) \rightarrow \chi_{p_2}^2$ as $n \rightarrow \infty$.

An example of the global orthogonality with $\Sigma_{x(12)} = 0$ is ANOVA type analysis: $y_i = x_{1i}^\top\beta_1 + x_{2i}^\top\beta_2 + e_i$ where x_{1i} is a group indicator for the placebo, and x_{2i} is a vector of group indicators for an active control and experimental treatment group. β_2 denote a vector of treatment effects of the active control and the experimental treatment. If inference is concerned about whether the effect of the experimental treatment has the same efficacy as the active control, allowing it to be different from the placebo effect, we are only concerned about β_2 . Normally we do not have $\Sigma_{x(12)} = 0$ other than this trivial case.

In M-type regression we have a linear transformation that achieves the orthogonality. Define

$$\begin{aligned} X_{n(1)} &= (x_{11}, \dots, x_{1n})^\top, & X_{n(2)} &= (x_{21}, \dots, x_{2n})^\top, & W_n &= \operatorname{diag}(dF_n^{\tilde{\beta}, \tilde{\lambda}}(r_i(\tilde{\beta}))) \\ x_{1i}^* &= x_{1i}, & x_{2i}^* &= x_{2i} - X_{n(2)}^\top W_n X_{n(1)} (X_{n(1)}^\top W_n X_{n(1)})^{-1} x_{1i} \\ \zeta_1 &= \beta_1 + (X_{n(1)}^\top W_n X_{n(1)})^{-1} X_{n(1)}^\top W_n X_{n(2)} \beta_2, & \zeta_2 &= \beta_2, \end{aligned}$$

where $(X_{n(1)}^\top W_n X_{n(1)})^{-1}$ denotes a generalized (Moore's) inverse matrix. With the transformation $Y_i = x_i^{*\top} \zeta + e_i$ and a *residual-wise* EL is given by

$$L_E(\zeta) = \max\{\prod p_i | \sum p_i \psi(r_i(\zeta))x_i^* = \epsilon_n, \sum p_i = 1, p_i \geq 0\},$$

where $p_i = dF(r_i(\zeta))$ for some univariate distribution F .

Lemma 5 Let $\Sigma_{x^*} = \lim_n n^{-1} \sum x_i^* x_i^{*\top}$. Then, $\Sigma_{x^*(12)} = 0$.

When $\tilde{\zeta}_1 = \tilde{\beta}_1 + (X_{n(1)}^\top W_n X_{n(1)})^{-1} X_{n(1)}^\top W_n X_{n(2)} \tilde{\beta}_2$ and $\tilde{\zeta}_2 = \tilde{\beta}_2$, we have $\tilde{\zeta} = \operatorname{argmax} L_E(\zeta)$ by the invariance of MELE. As $\zeta_2 = \beta_2$, we consider a conditional EL $L_E(\zeta_2 | \tilde{\zeta}_1)$ for the EL inference of β_{20} . We have the following result from Corollary 1 and Lemma 5.

Corollary 2 Assume the conditions of Theorem 5 with (17) instead of (16). For β_2 within the surface ball of $\|\beta_2 - \beta_{20}\| \leq n^{-1/2}$,

$$|\log \bar{L}_E(\beta_2) - \log L_E(\zeta_2 | \tilde{\zeta}_2)| = o_p(1) \quad \text{and} \quad |\bar{W}_E(\beta_2) - W_E(\zeta_2 | \tilde{\zeta}_2)| = o_p(1).$$

Also as $n \rightarrow \infty$, $W_E(\zeta_{20} | \tilde{\zeta}_2) \rightarrow \chi_{p_2}^2$.

Also consider

$$\tilde{L}_E(\zeta_2) = \max\left\{\prod p_i \mid \sum p_i \psi(r_i(\zeta) x_{2i}^* = \epsilon_n^*, \sum p_i = 1, p_i \geq 0\right\}, \quad (21)$$

where $\epsilon_n^* \in R^{p_2}$ with $\|\epsilon_n^*\| = o(n^{-1/2})$. This has a computational advantage over $L_E(\zeta_2 | \tilde{\zeta}_2)$ that algorithms such as R `emplik` package are available as $\psi(r_i(\zeta) x_{2i}^*)$ is of p_2 -dimension. We have the following result from Theorem 4.

Corollary 3 Assume the conditions of Theorem 5 with (17) instead of (16). For β_2 within the surface ball of $\|\beta_2 - \beta_{20}\| \leq n^{-1/2}$,

$$|\log \bar{L}_E(\beta_2) - \log \tilde{L}_E(\zeta_2)| = o_p(1) \quad \text{and} \quad |\log \bar{W}_E(\beta_2) - \log \tilde{W}_E(\zeta_2)| = o_p(1).$$

Also $\tilde{W}_E(\zeta_{20}) \rightarrow \chi_{p_2}^2$ as $n \rightarrow \infty$, where $\tilde{W}_E(\zeta_{20}) = -2 \log[\tilde{L}_E(\zeta_{20}) - \tilde{L}_E(\tilde{\zeta}_2)]$.

A EL ratio test with an asymptotic level α for any $0 < \alpha < 1$ rejects the null hypothesis if $\tilde{W}_E(\zeta_2) > q_{1-\alpha}$ where $q_{1-\alpha}$ is $100(1-\alpha)$ -th percentile of $\chi_{p_2}^2$. Conversely, an EL confidence region for β_{20} with an asymptotic coverage probability $1 - \alpha$ is given by $\{\zeta_2 | \tilde{L}_E(\zeta_2) < q_{1-\alpha}\}$.

5.3 Monte Carlo Studies

Figures and tables are presented in the appendix 6.3.

5.3.1 Study 1

We examine EL ratio test by conditional profile EL with two M-type regressions, least squares and median regression. Data was generated from the following model:

$$Y_i = x_i^\top \beta + e_i,$$

where $x_i = (x_{i1}, \dots, x_{i5})^\top$, $\beta = (3, 2, 1, 1, 1)^\top$, and e_i are i.i.d. random error variables. The independent variables x_i were sampled as follows: $x_{i1} \sim \chi_1$, $x_{i2} \sim N(-1, 1)$, $x_{i3} \sim U(0, 1)$, $x_{i4} \sim N(2, 1)$ and $x_{i5} \sim \chi_1 - 1$. Sample sizes $n = 100, 200$ and 400 were considered for each of three

error distributions, a standard normal ($N(0, 1)$), a t distribution with degree of freedom 3 (t_3) and exponential distributions, centered respectively to be mean zero for the least squares regression and median zero for the median regression. 3000 simulations were conducted for each error distribution with each sample size to test null hypotheses $H_0 : \beta_2 = 2$ and $H_0 : \beta_2 = (2, 1)^\top$.

For computational simplicity, we used $\widetilde{W}_E(\zeta_{20})$ in the simulations. Figures 1 and 2 show that the distributions of the observed $\widetilde{W}_E(\zeta_{20})$ agree well with respective χ^2 reference distributions. The estimated type I error rates are reported in Table 1. For comparison, we considered the following EL:

$$\mathbb{L}_E(\beta_2) = \max\left\{\prod p_i \mid \sum p_i \psi(e_i - x_{2i}^\top(\beta_2 - \beta_{20}))x_{2i} = \epsilon_n^*, \sum p_i = 1, p_i \geq 0\right\},$$

where $\epsilon_n^* \in R^{p_2}$ with $\|\epsilon_n^*\| = o(n^{-1/2})$. We compute $\mathbb{W}_E(\beta_2) = -2[\log \mathbb{L}_E(\beta_{20}) - \log \max \mathbb{L}_E(\beta_2)]$. As the true parameter values are plugged in for β_1 , a test by $\mathbb{W}_E(\beta_{20})$ should show an ‘‘optimal performance’’ for an EL test. The estimated type I error rates of the proposed conditional EL are close to the nominal level with differences of magnitudes that are comparable with the case of the test by $\mathbb{W}_E(\beta_{20})$.

5.3.2 Study 2

We examined the length and coverage probability of a confidence interval constructed by the proposed conditional EL. We used the same simulation model and computed confidence intervals for $\beta_{20} = 2$ for three error distributions with $n = 100, 200, 400$. Again for the stated computation simplicity we used $\widetilde{W}_E(\zeta_2)$ instead of $W_E(\zeta_2|\zeta_1)$. For comparison, we used pairwise bootstrap method and computed two confidence intervals based on 200 bootstrapped samples, one by using percentiles of bootstrapped estimate distribution (BP) and the other via normal approximation based on standard error estimates from the bootstrapped estimates (BN). Table 2 summarizes the simulation results.

When the least squares mean regression is concerned, all three methods tend to cover the true parameter value less frequently than the nominal levels claim, while the EL based method tends to undercover the parameter value and the bootstrap methods tend to overcover when the median regression is concerned. The undercoverage with the least squares mean regression is severer in small samples ($n = 100$) and with the EL based method, while the shortest average interval length of the EL based confidence intervals indicates a trade-off between the coverage and the interval length. With the median regression, the EL method undercovers the true value to a similar degree that the bootstrap methods overcover and both problems are not serious compared with the under coverage of the least squares. When the trade-off between the interval length and coverage is accounted for, all three methods performed comparably in terms of the coverage. An interesting question that remains to be answered is whether $\widetilde{W}_E(\zeta_2)$ is Bartlett correctable or not. If correctable, improvement on the coverage probability can be achieved.

However, when the ratios of % miss below over % miss above the confidence intervals (the parenthesized numbers in the table) are considered, the EL method shows more balanced performance with the exponential error distribution uniformly in all sample sizes and in both the least squares and median regression. The closer to 1 the ratio is, the more balanced the performance is in terms of the % miss below and % miss above. The EL has the ratios closest to 1 among the three methods. This shows that the EL method better adapts to the error distribution. With the symmetric error distributions, performances by all three methods were well balanced with no clear

discrepancies. Another point to note is a generic advantage of EL based method over bootstrap method: when β_{20} is in a higher dimension ($p_2 \geq 2$), construction of confidence regions is not clear via bootstrap methods, while it is natural with EL based method.

5.3.3 Study 3

We examined the relationship between $\overline{W}(\beta_2)$ and the proposed conditional EL ratio. Contours (for $\beta_2 = (2, 1)$) and curves (for $\beta_2 = 2$) were computed by two different EL ratio functions, $\overline{W}(\beta_2)$ and $\overline{W}_E(\zeta_2)$ based on the same data generated from

$$Y_i = 3 + x_{i1} + 2x_{i2} + e_i,$$

where $x_{i1} \sim N(-1, 1)$, $x_{i2} \sim \chi_1$ and e_i are i.i.d. random error variables from $N(0, 1)$. Figure 3 shows that the confidence intervals and regions by two EL functions agree well increasingly with growing n . However, computational efforts were significantly different by the two methods. Computation of $\overline{W}_E(\zeta_2)$ is much simpler.

6 Appendix

6.1 Proofs for the results in Sections 2, 3 and 4

Define

$$M_n(\theta) = n^{-1} \sum m(z_i, \theta), \quad \Sigma_n(\theta) = n^{-1} \sum m(z_i, \theta) m^\top(z_i, \theta),$$

$$\mathcal{H} = \begin{pmatrix} 0 & - \left(E_{F_0} \left(\frac{\partial m}{\partial \theta} \Big|_{\theta=\theta_0} \right) \right)^\top \\ - E_{F_0} \left(\frac{\partial m}{\partial \theta} \Big|_{\theta=\theta_0} \right) & E_{F_0} (m(z, \theta_0) m^\top(z, \theta_0)) \end{pmatrix}.$$

Let $H_{n(ij(kl))}(\theta, \lambda)$ and $H_{n(ij(k))}(\theta, \lambda)$ denote sub-block matrices of the ij th sub-block matrices of $H_n(\theta, \lambda)$ such that $H_{n(11(ij))}(\theta, \lambda) = \partial^2 l(\theta, \lambda) / \partial \theta_i \partial \theta_j^\top$ and $H_{n(21(i))}(\theta, \lambda) = H_{n(12(i))}^\top(\theta, \lambda) = \partial^2 l(\theta, \lambda) / \partial \lambda \partial \theta_i^\top$. Let $H_n^{-\theta_2}(\theta, \lambda)$ and $\mathcal{H}^{-\theta_2}$ denote respectively $H_n(\theta, \lambda)$ and \mathcal{H} less the components with respect to θ_2 similarly to $S_n^{-\theta_2}(\theta, \lambda)$. Unless necessary, we suppress the subscript n . Let $M_0 = M_n(\theta_0)$, $\Sigma_0 = \Sigma_n(\theta_0)$, $S_0 = S_n(\theta_0, 0)$ and $H_0 = H_n(\theta_0, 0)$. Let $\lambda_0 = \lambda(\theta_0)$. We let δ_n be a generic notation for the remainder in Taylor expansion.

Proof of Lemma 1 The equation (4) can be derived by algebraically rearranging $\sum_{i=1}^n dF_n^{\theta, \lambda}(z_i) - 1 = 0$. As $0 < dF_n^{\theta, \lambda}(z_i) < 1$, $D_\theta = \{\lambda | 1 + \lambda^\top m(z_i, \theta) > 1/n, i = 1, \dots, n\}$ denotes the set of feasible λ for fixed θ . D_θ is open, convex and bounded under Condition 1. Also $\partial^2 l(\theta, \lambda) / (\partial \lambda \partial \lambda^\top) = \sum m(z_i, \theta) m^\top(z_i, \theta) \{1 + \lambda^\top m(z_i, \theta)\}^{-2}$. Then, $l(\theta, \lambda)$ for fixed θ is a continuous function of λ defined over D_θ with a positive definite hessian matrix, if $n^{-1} \sum m(z_i, \theta) m^\top(z_i, \theta)$ is positive definite, which is true with probability 1 for θ within the surface ball of $\|\theta - \theta_0\| \leq n^{-1/3}$.

Lemma A.1 Assume that $m(z, \theta)$ is continuous in some neighborhood of θ_0 and $E(m(z, \theta_0) m^\top(z, \theta_0))$ and $E[\partial m / \partial \theta]_{\theta=\theta_0}$ exist. Then, $n^{-1} H_0 \rightarrow \mathcal{H}$.

Proof Note that $H_{0(11)} = 0$ with $\lambda = 0$. The rest of the proof follows from the strong law of large numbers.

Proof of Lemma 2 From the equation (4), $S_{(2)}(\theta, \lambda(\theta)) = 0$ and

$$\lambda(\theta) = \Sigma^{-1}(\theta)M(\theta) + o(n^{-1/3}) = O(n^{-1/3}) \quad (\text{a.s.}), \quad (\text{A.1})$$

uniformly about $\theta \in \{\theta \mid \|\theta - \theta_0\| \leq n^{-1/3}\}$. Let $\theta = \theta_0 + un^{-1/3}$ for $\theta \in \{\theta \mid \|\theta - \theta_0\| = n^{-1/3}\}$ where $\|u\| = 1$. By Taylor expansion,

$$\begin{aligned} & l(\theta, \lambda(\theta)) \\ &= -n \log n - \frac{n}{2} M(\theta) [\Sigma(\theta)]^{-1} M(\theta) + o(n^{1/3}) \quad (\text{a.s.}) \\ &= -n \log n - \frac{n}{2} \left[O(n^{-1/2}(\log \log n)^{1/2}) - \mathcal{H}_{(21)} un^{-1/3} \right]^\top \mathcal{H}_{(22)}^{-1} \\ & \quad \left[O(n^{-1/2}(\log \log n)^{1/2}) - \mathcal{H}_{(21)} un^{-1/3} \right] + o(n^{1/3}) \quad (\text{a.s.}) \\ &\leq -n \log n - (c - \epsilon) n^{1/3} \quad (\text{a.s.}), \end{aligned} \quad (\text{A.2})$$

where $c - \epsilon > 0$ and c is the smallest eigenvalue of $\mathcal{H}_{(12)} \mathcal{H}_{(22)}^{-1} \mathcal{H}_{(21)}$. Similarly,

$$\begin{aligned} l(\theta_0, \lambda_0) &= -n \log(n) - \frac{n}{2} M_0^\top \Sigma_0^{-1} M_0 + o(1) \quad (\text{a.s.}) \\ &= -n \log(n) - O(\log \log n) \quad (\text{a.s.}) \end{aligned}$$

Note that $\max l(\theta, \lambda(\theta)) = \max \log L_E(\theta) \leq -n \log n$ and $l(\theta, \lambda(\theta))$ is a continuous function about $\theta \in \{\theta \mid \|\theta - \theta_0\| \leq n^{-1/3}\}$. Therefore, $l(\theta, \lambda(\theta))$ has maximum value in the interior of the ball and $S(\tilde{\theta}, \tilde{\lambda}) = 0$.

Proof of Theorem 1 (a) From the definition of $(\tilde{\theta}, \tilde{\lambda})$ and by Taylor expansion, for all n ,

$$0 = n^{-1/2} S(\tilde{\theta}, \tilde{\lambda}) = n^{-1/2} S_0 + \sqrt{n} ((\tilde{\theta} - \theta_0)^\top, \tilde{\lambda}^\top) (n^{-1} H_0) + o_p(\delta_n), \quad (\text{A.3})$$

where $\delta_n = \sqrt{n} \|\tilde{\theta} - \theta_0\| + \sqrt{n} \|\tilde{\lambda}\|$. We have

$$\sqrt{n} \begin{pmatrix} \tilde{\theta} - \theta_0 \\ \tilde{\lambda} \end{pmatrix} = (n^{-1} H_0)^{-1} (-n^{-1/2} S_0 + o_p(\delta_n)).$$

As $n^{-1} H_0 \rightarrow \mathcal{H}$ by Lemma A.1 and $\|n^{-1/2} S_0\| = O_p(1)$, we know that $\delta_n = O_p(1)$. As $V = -\mathcal{H}^{(11)}$, we have

$$\begin{aligned} \sqrt{n}(\tilde{\theta} - \theta_0) &= V \mathcal{H}_{(12)} \mathcal{H}_{(22)}^{-1} \sqrt{n} M_0 + O_p(n^{-1/2}) \\ \sqrt{n} \tilde{\lambda} &= \mathcal{H}_{(22)}^{-1} \{I - \mathcal{H}_{(21)} V \mathcal{H}_{(12)} \mathcal{H}_{(22)}^{-1}\} \sqrt{n} M_0 + O_p(n^{-1/2}). \end{aligned} \quad (\text{A.4})$$

The rest of the proof follows from the asymptotic normality of $\sqrt{n} M_0$ by multivariate central limit theorem. The results with $F_n^{\tilde{\theta}, \tilde{\lambda}}(z)$ follow similarly.

(b) Note that $W_E(\theta_0) = -2[l(\theta_0, \lambda_0) - l(\tilde{\theta}, \tilde{\lambda})]$. As (θ_0, λ_0) satisfies the equation (4),

$$\lambda_0 = \mathcal{H}_{(22)}^{-1} M_0 + o_p(n^{-1/2}). \quad (\text{A.5})$$

By Taylor expansion similar to (A.2) and from (A.4) and (A.5),

$$\begin{aligned} l(\tilde{\theta}, \tilde{\lambda}) - l(\theta_0, 0) &= -\frac{n}{2} M_0 \mathcal{H}_{(22)}^{-1} \left\{ I - \mathcal{H}_{(21)} V \mathcal{H}_{(12)} \mathcal{H}_{(22)}^{-1} \right\} M_0 + o_p(1), \\ l(\theta_0, \lambda_0) - l(\theta_0, 0) &= -\frac{n}{2} M_0 \mathcal{H}_{(22)}^{-1} M_0 + o_p(1). \end{aligned}$$

We have

$$\begin{aligned} W_E(\theta_0, \lambda_0) &= (\mathcal{H}_{(22)}^{-1/2} \sqrt{n} M_0)^\top [\mathcal{H}_{(22)}^{-1/2} \mathcal{H}_{(21)} V \mathcal{H}_{(12)} \mathcal{H}_{(22)}^{-1/2}] \\ &\quad (\mathcal{H}_{(22)}^{-1/2} \sqrt{n} M_0) + o_p(1), \end{aligned} \quad (\text{A.6})$$

which is the same results as in Qin and Lawless (1994). As in (16), the desired result follows from the fact that $(\mathcal{H}_{(22)}^{-1/2} \sqrt{n} M_0)$ converges to a standard multivariate normal distribution and $[\mathcal{H}_{(22)}^{-1/2} \mathcal{H}_{(21)} V \mathcal{H}_{(12)} \mathcal{H}_{(22)}^{-1/2}]$ is symmetric and idempotent, with trace equal to p .

(c) As $\lim_{n \rightarrow \infty} \left\{ \int H(\tilde{\theta}, \tilde{\lambda}) dF_n^{\tilde{\theta}, \tilde{\lambda}} - n^{-1} H(\tilde{\theta}, \tilde{\lambda}) + n^{-1} H(\tilde{\theta}, \tilde{\lambda}) - n^{-1} H(\theta_0, 0) \right\} = 0$, we have $\lim_{n \rightarrow \infty} I_n(\tilde{\theta}, \tilde{\lambda}) = -\mathcal{H}$ from Lemma A.1. The Cramér-Rao lower bound for the estimator of θ_0 is given by $-(\mathcal{H}_{(11)} - \mathcal{H}_{(12)} \mathcal{H}_{(22)}^{-1} \mathcal{H}_{(21)})^{-1}$. As $\mathcal{H}_{(11)} = 0$ and $V = (\mathcal{H}_{(12)} \mathcal{H}_{(22)}^{-1} \mathcal{H}_{(21)})^{-1}$, the asymptotic variance of $\tilde{\theta}$ achieves the lower bound.

Proof of Lemma 3 Similar to $F_n^{\tilde{\theta}, \lambda(\tilde{\theta})}(z)$, it can be shown that $F_n^{\hat{\theta}_n^*, \lambda(\hat{\theta}_n^*)}(z)$ is a \sqrt{n} -consistent estimator of F_0 . Then,

$$\begin{aligned} \int H(\hat{\theta}_n^*, \lambda(\hat{\theta}_n^*)) dF_n^{\hat{\theta}_n^*, \lambda(\hat{\theta}_n^*)} - n^{-1} H(\hat{\theta}_n^*, \lambda(\hat{\theta}_n^*)) + n^{-1} H(\hat{\theta}_n^*, \lambda(\hat{\theta}_n^*)) \\ - n^{-1} H(\hat{\theta}_n^*, \lambda(\hat{\theta}_n^*)) + n^{-1} H(\hat{\theta}_n^*, \lambda(\hat{\theta}_n^*)) - n^{-1} H(\theta_0, 0) = o_p(1), \end{aligned}$$

and $\lim_{n \rightarrow \infty} I_n(\hat{\theta}_n^*, \lambda(\hat{\theta}_n^*)) = -\mathcal{H}$ from Lemma A.1. On the other hand (6) implies $-(\mathcal{H}^{(11)})_{(12)} = 0$, which is

$$-\mathcal{H}_{(12(1))} \mathcal{H}_{(22)}^{-1} \mathcal{H}_{(21(2))} = 0, \quad (\text{A.7})$$

where $\mathcal{H}_{(12)}$ is of rank p and $\mathcal{H}_{(22)}$ is positive definite under the conditions. Hence (A.7) holds if and only if

$$0 = E_{F_0} \left[\frac{\partial m_2}{\partial \theta_1} \Big|_{\theta=\theta_0} \right] = E_{F_0} \left[\frac{\partial m_1}{\partial \theta_2} \Big|_{\theta=\theta_0} \right] = E_{F_0} \left[m_1(z, \theta_0) m_2^\top(z, \theta_0) \right]. \quad (\text{A.8})$$

(A.8) holds if and only if $m_1(z, \hat{\theta}_n^*)$ and $m_2(z, \hat{\theta}_n^*)$ are asymptotically uncorrelated.

Proof of Theorem 2 (a) For a given θ_2 , note that $\bar{L}_E(\theta_2) = \max_{\theta_1} L(\theta_1, \theta_2, \lambda(\theta_1))$ where $(\theta_1, \lambda(\theta_1))$ satisfies the equations (4) for the fixed θ_2 . That is, for the fixed θ_2 , $\partial l(\theta_1, \theta_2, \lambda) / \partial \lambda = 0$ for $\lambda = \lambda(\theta_1)$. Also $l(\theta_1, \theta_2, \lambda)$ is a continuous function in θ . Hence by the definition of $\bar{\theta}_1(\theta_2)$ and $\bar{\lambda}(\theta_2)$, $S^{-\theta_2}(\bar{\theta}_1(\theta_2), \theta_2, \bar{\lambda}(\theta_2)) = 0$ for all n and all θ_2 in the neighborhood that $m(z_i, \theta)$ is differentiable.

(b) By Taylor expansion similar to (A.3),

$$\sqrt{n} \begin{pmatrix} \bar{\theta}_1(\theta_2) - \theta_{10} \\ \bar{\lambda}(\theta_2) \end{pmatrix} = (n^{-1}H^{-\theta_2}(\theta_{10}, \theta_2, 0))^{-1}(-\sqrt{n}S^{-\theta_2}(\theta_{10}, \theta_2, 0) + o_p(\delta_n)), \quad (\text{A.9})$$

where $\delta_n = \sqrt{n}|\bar{\theta}_1(\theta_2) - \theta_{10}| + \sqrt{n}|\bar{\lambda}(\theta_2)|$. As $\|\theta_2 - \theta_{20}\| \leq n^{-1/2}$, $n^{-1}H^{-\theta_2}(\theta_{10}, \theta_2, 0) \rightarrow \mathcal{H}^{-\theta_2}$ by Lemma A.1. Also $\sqrt{n}M(\theta_{10}, \theta_2, 0) = \sqrt{n}M_0 - \mathcal{H}_{(21(2))}\sqrt{n}(\theta_2 - \theta_{20}) + o_p(1)$. From this we have $\delta_n = O_p(1)$. When $\bar{V}_1 = (\mathcal{H}_{(12(1))}\mathcal{H}_{(22)}^{-1}\mathcal{H}_{(21(1))})^{-1}$,

$$\sqrt{n} \begin{pmatrix} \bar{\theta}_1(\theta_2) - \theta_{10} \\ \bar{\lambda}(\theta_2) \end{pmatrix} = \bar{V}_1(\sqrt{n}M_0 - \mathcal{H}_{(21(2))}\sqrt{n}(\theta_2 - \theta_{20})) + o_p(1), \quad (\text{A.10})$$

and the proof is complete.

(c) Let $\bar{\theta}_{10} = \bar{\theta}_1(\theta_{20})$ and $\bar{\lambda}_0 = \bar{\lambda}(\theta_{20})$. Then,

$$\bar{W}_E(\theta_{20}) = W_E(\theta_0) - \bar{w},$$

where $\bar{w} = -2[l(\theta_{10}, \theta_{20}, \lambda_0) - l(\bar{\theta}_{10}, \theta_{20}, \bar{\lambda}_0)]$. Then, it follows from (A.10) and by the definition of $\bar{\theta}_{10}$ that

$$\begin{aligned} \sqrt{n}(\bar{\theta}_{10} - \theta_{10}) &= \bar{V}_1\mathcal{H}_{(12(1))}\mathcal{H}_{(22)}^{-1}\sqrt{n}M_0 + o_p(1), \\ \sqrt{n}\bar{\lambda}_0 &= \mathcal{H}_{(22)}^{-1}\{I - \mathcal{H}_{(21(1))}\bar{V}_1\mathcal{H}_{(12(1))}\mathcal{H}_{(22)}^{-1}\}\sqrt{n}M_0 + o_p(1). \end{aligned} \quad (\text{A.11})$$

Similarly as in the proof of Theorem 1(b), by Taylor expansion and (A.11),

$$\bar{w} = (\mathcal{H}_{(22)}^{-1/2}\sqrt{n}M_0)^\top \mathcal{H}_{(22)}^{-1/2}\mathcal{H}_{(21(1))}\bar{V}_1\mathcal{H}_{(12(1))}\mathcal{H}_{(22)}^{-1/2}(\mathcal{H}_{(22)}^{-1/2}\sqrt{n}M_0) + o_p(1).$$

Then, $\bar{W}_E(\theta_{20}) = (\mathcal{H}_{(22)}^{-1/2}\sqrt{n}M_0)^\top \mathcal{H}_{(22)}^{-1/2}\bar{D}\mathcal{H}_{(22)}^{-1/2}(\mathcal{H}_{(22)}^{-1/2}\sqrt{n}M_0) + o_p(1)$, where

$$\bar{D} = \{\mathcal{H}_{(21)}V\mathcal{H}_{(12)} - \mathcal{H}_{(21(1))}\bar{V}_1\mathcal{H}_{(12(1))}\}. \quad (\text{A.12})$$

This is the same result obtained in Corollary 5 of (16). The rest of the proof follow the same arguments as in their proof of corollary 5.

Proof of Theorem 3 (i) By Theorem 1 (a) and 2 (b), we have $O_p(n^{-1})$ in (A.3) and (A.9) in place of $o_p(\delta_n)$. It follows from (A.4) and (A.10) that

$$\begin{aligned} \sqrt{n}(\bar{\theta}_1(\theta_2) - \theta_{10}) &= \bar{V}_1\mathcal{H}_{(12(1))}\mathcal{H}_{(22)}^{-1}(\sqrt{n}M_0 - \mathcal{H}_{(21(2))}\sqrt{n}(\theta_2 - \theta_{20})) + O_p(n^{-1/2}), \\ \sqrt{n}(\tilde{\theta}_1 - \theta_{10}) &= (V_{(11)}\mathcal{H}_{(12(1))} + V_{(12)}\mathcal{H}_{(12(2))})\mathcal{H}_{(22)}^{-1}\sqrt{n}M_0 + O_p(n^{-1/2}) \end{aligned}$$

Under the orthogonality $-\mathcal{H}_{(12(1))}\mathcal{H}_{(22)}^{-1}\mathcal{H}_{(21(2))} = V_{(12)} = 0$ from (A.7) and $\bar{V}_1 = V_{(11)}$ and we have $\|\bar{\theta}_1(\theta_2) - \tilde{\theta}_1\| = O_p(n^{-1})$.

On the other hand by Taylor expansion similar to (A.2), for θ_2 within the surface ball of $\|\theta_2 - \theta_{20}\| \leq n^{-1/2}$,

$$\begin{aligned} &l(\bar{\theta}_1(\theta_2), \theta_2, \bar{\lambda}(\theta_2)) - l(\tilde{\theta}_1, \theta_2, \tilde{\lambda}(\theta_2)) \\ &= -\frac{n}{2}[M(\theta_{10}, \theta_2) - \mathcal{H}_{(21(1))}(\bar{\theta}_1(\theta_2) - \theta_{10})]^\top \mathcal{H}_{(22)}^{-1}[M(\theta_{10}, \theta_2) - \mathcal{H}_{(21(1))}(\bar{\theta}_1(\theta_2) - \theta_{10})] \\ &\quad + \frac{n}{2}[M(\theta_{10}, \theta_2) - \mathcal{H}_{(21(1))}(\tilde{\theta}_1 - \theta_{10})]^\top \mathcal{H}_{(22)}^{-1}[M(\theta_{10}, \theta_2) - \mathcal{H}_{(21(1))}(\tilde{\theta}_1 - \theta_{10})] + o_p(1). \end{aligned}$$

As $\|\tilde{\theta}_1(\theta_2) - \tilde{\theta}_1\| = O_p(n^{-1})$, we have $|l(\tilde{\theta}_1(\theta_2), \theta_2, \tilde{\lambda}(\theta_2)) - l(\tilde{\theta}_1, \theta_2, \tilde{\lambda}(\theta_2))| = o_p(1)$, which suffices to prove the rest of the results.

(ii) The proof is very similar to (i) except the part about the asymptotic distribution of $W_E(\theta_{20}|\tilde{\theta}_1)$. We sketch the proof.

Let $\tilde{\lambda}_0 = \tilde{\lambda}(\theta_{20})$ and $\tilde{w} = 2[l(\tilde{\theta}_1, \theta_{20}, \tilde{\lambda}_0) - l(\theta_{10}, \theta_{20}, \lambda_0)]$. Note that

$$\begin{aligned}\sqrt{n}M(\tilde{\theta}_1, \theta_{20}) &= \sqrt{n}M_0 - \mathcal{H}_{(21(1))}\sqrt{n}(\tilde{\theta}_1 - \theta_{10}) + o_p(1) \\ \Sigma(\tilde{\theta}_1, \theta_{20}) &= \Sigma_0 + o_p(1) \rightarrow \mathcal{H}_{(22)}.\end{aligned}\tag{A.13}$$

As $(\tilde{\theta}_1, \tilde{\lambda}_0)$ satisfies the equation (4) for $\theta_2 = \theta_{20}$, we have

$$\tilde{\lambda}_0 = \mathcal{H}_{(22)}^{-1}(M_0 - \mathcal{H}_{(21(1))}(\tilde{\theta}_1 - \theta_{10})) + o_p(n^{-1/2}).\tag{A.14}$$

By Taylor expansion and substituting $\tilde{\lambda}_0$ and $(\tilde{\theta}_1 - \theta_{10})$, we have

$$\begin{aligned}\tilde{w} &= (\mathcal{H}_{(22)}^{-1/2}\sqrt{n}M_0)^\top \mathcal{H}_{(22)}^{-1/2}\{D^* + D^{*\top} - D^{*\top}\mathcal{H}_{(22)}^{-1}D^*\} \\ &\quad \mathcal{H}_{(22)}^{-1/2}(\mathcal{H}_{(22)}^{-1/2}\sqrt{n}M_0) + o_p(1).\end{aligned}\tag{A.15}$$

where $D^* = \mathcal{H}_{(21(1))}V_{(11)}\mathcal{H}_{(21(1))} + \mathcal{H}_{(21(1))}V_{(12)}\mathcal{H}_{(12(2))}$. On the other hand, $W_E(\theta_{20}|\tilde{\theta}_1) = W_E(\theta_0) - \tilde{w}$. It follows from (A.6) and (A.15) that

$$W_E(\theta_{20}|\tilde{\theta}_1) = (\mathcal{H}_{(22)}^{-1/2}\sqrt{n}M_0)^\top \mathcal{H}_{(22)}^{-1/2}\tilde{D}\mathcal{H}_{(22)}^{-1/2}(\mathcal{H}_{(22)}^{-1/2}\sqrt{n}M_0) + o_p(1).$$

where $\tilde{D} = \mathcal{H}_{(21)}V\mathcal{H}_{(12)} - D^* - D^{*\top} + D^{*\top}\mathcal{H}_{(22)}^{-1}D^*$. By Lemma A.2, \tilde{D} is a non-negative definite matrix and the desired results follow from Lemma 3 of (16). Also $\tilde{D} = \overline{D}$ under the orthogonality by Lemma A.2 and we independently confirm that $W_E(\theta_{20}|\tilde{\theta}_1) = \overline{W}_E(\theta_{20}) + o_p(1)$.

Lemma A.2 *Assume the conditions of Theorem 3. \tilde{D} is a non-negative definite matrix. Under the orthogonality,*

$$\tilde{D} = \mathcal{H}_{22}^{-1/2}\mathcal{H}_{21(2)}(\mathcal{H}_{21(2)}\mathcal{H}_{22}^{-1}\mathcal{H}_{12(2)})^{-1}\mathcal{H}_{12(2)}\mathcal{H}_{22}^{-1/2}.$$

Proof Define $V = \begin{pmatrix} \nu_{11} & \nu_{12} \\ \nu_{21} & \nu_{22} \end{pmatrix}^{-1}$ so that $\nu_{ij} = \mathcal{H}_{(21(i))}\mathcal{H}_{(22)}^{-1}\mathcal{H}_{(12(j))}$. Then,

$$\begin{aligned}D^* + D^{*\top} &= \mathcal{H}_{21} \begin{pmatrix} V_{(11)} & V_{(12)} \\ V_{(21)} & 0 \end{pmatrix} \mathcal{H}_{(12)} + \mathcal{H}_{(21(1))}V_{(11)}\mathcal{H}_{(21(1))} \\ D^{*\top}\mathcal{H}_{(22)}^{-1}D^* &= \mathcal{H}_{(21)} \begin{pmatrix} V_{(11)} \\ V_{(21)} \end{pmatrix} \nu_{11}(V_{(11)}, V_{(12)})\mathcal{H}_{(12)}.\end{aligned}$$

With some simple matrix algebra, $\tilde{D} = \mathcal{H}_{(21)}[\tilde{D}_1 + \tilde{D}_2]\mathcal{H}_{(12)}$, where

$$\begin{aligned}\tilde{D}_1 &= \begin{pmatrix} V_{(11)}\{\nu_{11} - V_{(11)}^{-1}\}V_{(11)} - V_{(11)} & V_{(11)}\{\nu_{11} - V_{(11)}^{-1}\}V_{(12)} \\ V_{(21)}\{\nu_{11} - V_{(11)}^{-1}\}V_{(11)} & V_{(21)}\{\nu_{11} - V_{(11)}^{-1}\}V_{(12)} \end{pmatrix}, \\ \tilde{D}_2 &= \begin{pmatrix} 0 & V_{(12)} \\ V_{(21)} & V_{(22)} \end{pmatrix}.\end{aligned}$$

\tilde{D}_1 is non-negative definite as $V_{(11)}^{-1} = \nu_{11} - \nu_{12}\nu_{22}^{-1}\nu_{21}$. \tilde{D}_2 is also non-negative definite as it is symmetric and $V_{(22)}$ is non-negative definite. The result under the orthogonality follows as $V_{12} = V_{21} = 0$ and $V_{11} = \nu_{11}^{-1}$ and $V_{22} = \nu_{22}^{-1}$.

Lemma A.3 *Assume the conditions of Theorem 1. Then, $\sqrt{n}M(\tilde{\theta}_1, \theta_{20}) \sim N(0, \tilde{S})$ where*

$$\tilde{S} = \begin{bmatrix} \mathcal{H}_{(22)}^{1/2} - \mathcal{H}_{(21(1))}(V_{(11)}\mathcal{H}_{(12(1))} + V_{(12)}\mathcal{H}_{(12(2))})\mathcal{H}_{(22)}^{-1/2} \\ \mathcal{H}_{(22)}^{1/2} - \mathcal{H}_{(22)}^{-1/2}(\mathcal{H}_{(21(1))}V_{(11)} + \mathcal{H}_{(21(2))}V_{(21)})\mathcal{H}_{(12(1))} \end{bmatrix}.$$

Also $\Sigma(\tilde{\theta}_1, \theta_{20}) \rightarrow \mathcal{H}_{(22)}$ and $\tilde{S}^{-1}\mathcal{H}_{(22)} = I$ under orthogonality.

Proof The asymptotic normality follows directly from (A.4) and (A.13). Under the orthogonality, $V_{(12)} = V_{(21)}^\top = 0$ from (A.7) and $V_{(11)} = (\mathcal{H}_{(21(1))})\mathcal{H}_{(22)}^{-1}\mathcal{H}_{(12(1))}^{-1}$. Hence $\tilde{S} = \mathcal{H}_{22} - \mathcal{H}_{21(1)}V_{11}\mathcal{H}_{12(1)}$ and $(\tilde{S})^{-1} = \mathcal{H}_{22}^{-1}$. On the other hand $\Sigma_n(\tilde{\theta}_1, \theta_{20}) = \Sigma_n(\theta_0) + o_p(1)$, where $\Sigma_n(\theta_0) = n^{-1}H(\theta_0)$ and $n^{-1}H(\theta_0) \rightarrow \mathcal{H}_{22}$ by Lemma A.1.

Lemma A.4 *Assume the conditions of Theorem 1. Then, $\sqrt{n}M_n(\bar{\theta}_{10}, \theta_{20}) \sim N(0, \bar{S})$ where $\bar{S}^{-1}\mathcal{H}_{22} = I$. Also $\Sigma_n(\bar{\theta}_{10}, \theta_{20}) \rightarrow \mathcal{H}_{22}$.*

Proof As $\sqrt{n}M_n(\bar{\theta}_{10}, \theta_{20}) = \sqrt{n}M_n(\theta_0) - \mathcal{H}_{21(1)}\sqrt{n}(\bar{\theta}_{10} - \theta_{10}) + o_p(n^{-1/2})$, it follows from (A.11) that

$$\sqrt{n}M_n(\bar{\theta}_{10}, \theta_{20}) = [I - \mathcal{H}_{21(1)}\bar{V}_1\mathcal{H}_{12(1)}\mathcal{H}_{22}^{-1}]\sqrt{n}M_n(\theta_0) + o_p(1).$$

The normality follows from the multivariate central limit theorem and

$$\bar{S} = \left[\mathcal{H}_{22}^{1/2} - \mathcal{H}_{21(1)}\bar{V}_1\mathcal{H}_{12(1)}\mathcal{H}_{22}^{-1/2} \right] \left[\mathcal{H}_{22}^{1/2} - \mathcal{H}_{22}^{-1/2}\mathcal{H}_{21(1)}\bar{V}_1\mathcal{H}_{12(1)} \right].$$

As $\left[\mathcal{H}_{22}^{1/2} - \mathcal{H}_{21(1)}\bar{V}_1\mathcal{H}_{12(1)}\mathcal{H}_{22}^{-1/2} \right]^{-1} = \left[\mathcal{H}_{22}^{1/2} - \mathcal{H}_{22}^{-1/2}\mathcal{H}_{21(1)}\bar{V}_1\mathcal{H}_{12(1)} \right]^{-1} = \mathcal{H}_{22}^{-1/2}$, $\bar{S}^{-1} = \mathcal{H}_{22}^{-1}$. On the other hand, $\Sigma_n(\bar{\theta}_{10}, \theta_{20}) = \Sigma_n(\theta_0) + o_p(1)$, where $\Sigma_n(\theta_0) = n^{-1}H(\theta_0)$ and $n^{-1}H(\theta_0) \rightarrow \mathcal{H}_{22}$ by Lemma A.1.

Proof of Theorem 4 Without loss of generality, let $m(z, \theta) = (m_1^\top(z, \theta), m_2^\top(z, \theta))^\top$ and let $\Gamma = \begin{pmatrix} 0 & I \end{pmatrix}$. Then, $m_2(z, \theta) = \Gamma m(z, \theta)$. From (11) and by Taylor expansion,

$$\begin{aligned} & \tilde{L}_E(\theta_2) - L_E(\theta_2|\tilde{\theta}_1) \\ &= -\frac{n}{2}[M_0 - \mathcal{H}_{(21(1))}(\tilde{\theta}_1 - \theta_{10}) - \mathcal{H}_{(21(2))}(\theta_2 - \theta_{20})]^\top [\Gamma^\top (\Gamma\mathcal{H}_{(22)}\Gamma^\top)^{-1}\Gamma] \\ & \quad [M_0 - \mathcal{H}_{(21(1))}(\tilde{\theta}_1 - \theta_{10}) - \mathcal{H}_{(21(2))}(\theta_2 - \theta_{20})] \\ & + \frac{n}{2}[M_0 - \mathcal{H}_{(21(1))}(\tilde{\theta}_1 - \theta_{10}) - \mathcal{H}_{(21(2))}(\theta_2 - \theta_{20})]^\top [\mathcal{H}_{(22)}^{-1}] \\ & \quad [M_0 - \mathcal{H}_{(21(1))}(\tilde{\theta}_1 - \theta_{10}) - \mathcal{H}_{(21(2))}(\theta_2 - \theta_{20})] + o_p(1). \end{aligned}$$

From (A.4), (A.7) and (A.8),

$$\begin{aligned}
& [I - \mathcal{H}_{(22)}^{-1} \mathcal{H}_{(21(1))} \bar{V}_1 \mathcal{H}_{(12(1))}] [\Gamma^\top (\Gamma \mathcal{H}_{(22)} \Gamma^\top)^{-1} \Gamma] [I - \mathcal{H}_{(21(1))} \bar{V}_1 \mathcal{H}_{(12(1))} \mathcal{H}_{(22)}^{-1}] M_0 \\
&= [I - \mathcal{H}_{(22)}^{-1} \mathcal{H}_{(21(1))} \bar{V}_1 \mathcal{H}_{(12(1))}] [\mathcal{H}_{(22)}^{-1}] [I - \mathcal{H}_{(21(1))} \bar{V}_1 \mathcal{H}_{(12(1))} \mathcal{H}_{(22)}^{-1}], \\
& [\Gamma^\top (\Gamma \mathcal{H}_{(22)} \Gamma^\top)^{-1} \Gamma] \mathcal{H}_{(21(2))} = [\mathcal{H}_{(22)}^{-1}] \mathcal{H}_{(21(2))}, \\
& \mathcal{H}_{(12(2))} [\Gamma^\top (\Gamma \mathcal{H}_{(22)} \Gamma^\top)^{-1} \Gamma] \mathcal{H}_{(21(2))} = \mathcal{H}_{(12(2))} [\mathcal{H}_{(22)}^{-1}] \mathcal{H}_{(21(2))},
\end{aligned}$$

by which $\tilde{L}_E(\theta_2) = L_E(\theta_2 | \tilde{\theta}_1) + o_p(1)$ after simplification. The proof is complete by Theorem 3 (i).

6.2 Proofs for the results in Section 5

Let $H_n^{-\beta_2}(\beta, \lambda)$ denote $H_n(\beta, \lambda)$ less the components with respect to β_2 similarly to $S_n^{-\beta_2}(\beta, \lambda)$. Define

$$\begin{aligned}
M_n(\beta) &= n^{-1} \sum \psi(r_i(\beta)) x_i, & \Sigma_n(\beta) &= n^{-1} \sum \psi^2(r_i(\beta)) x_i x_i^\top, \\
\sigma_e^2 &= E_{F_e}[\psi^2(e)], & \mathcal{H} &= \begin{pmatrix} 0 & b_1 \Sigma_x \\ b_1 \Sigma_x & \sigma_e^2 \Sigma_x \end{pmatrix}.
\end{aligned}$$

We recycle most of the notations from the previous sections and skip proofs if they follow similar lines of arguments as in the proofs of previous sections.

Lemma A.5 *Assume Conditions 3 with (16) and Conditions 4-6.*

(i) $n^{-1} H_n(\beta_0, 0) \rightarrow \mathcal{H}$

(ii) For any β within the surface ball of $\|\beta - \beta_0\| \leq n^{-1/2}$,

$$\|n^{-1/2} \sum \psi(r_i(\beta)) x_i - n^{-1/2} \sum \psi(r_i(\beta_0)) x_i - b_1 \Sigma_x \sqrt{n}(\beta - \beta_0)\| = o_p(1)$$

and $n^{-1} \sum \psi^2(r_i(\beta)) x_i x_i^\top \rightarrow \sigma_e^2 \Sigma_x$ as $n \rightarrow \infty$.

(iii) For any β within the surface ball of $\|\beta - \beta_0\| \leq n^{-1/2}$, $\log L_E(\beta) = l(\beta, \lambda(\beta)) + o_p(1)$.

Proof of Lemma A.5 (i) can be shown similarly as Lemma A.1. (ii) follows from equations (A.8) and (A.9) of the proof of Theorem 1 in (17). As for (iii), we first note that from (18) and (19),

$$\begin{aligned}
\lambda^*(\beta) &= \left[\frac{1}{n} \sum \psi(r_i(\beta)) x_i x_i^\top \right]^{-1} \left(\frac{1}{n} \sum \psi(r_i(\beta)) x_i \right) + o_p(\|\epsilon_n\|) \\
\lambda(\beta) &= \left[\frac{1}{n} \sum \psi(r_i(\beta)) x_i x_i^\top \right]^{-1} \left(\frac{1}{n} \sum \psi(r_i(\beta)) x_i \right) + o_p(n^{-1/2})
\end{aligned}$$

Also by Taylor expansion, the chain rule and (16), for (b, κ) and (β, λ) such that $\|b - \beta_0\| \leq n^{-1/2}$, $\|\beta - \beta_0\| \leq n^{-1/2}$ and $\|\kappa\| = O_p(n^{-1/2})$, and $\|\lambda\| = O_p(n^{-1/2})$, we have $l(b, \kappa) - l(\beta, \lambda) = S_n(\beta, \lambda) \begin{pmatrix} b - \beta \\ \kappa - \lambda \end{pmatrix} + \frac{1}{2}((b - \beta)^\top$

, $(\kappa - \lambda)^\top H_n(\beta, \lambda) \begin{pmatrix} b - \beta \\ \kappa - \lambda \end{pmatrix} + o_p(1)$. It follows that $l(\beta, \lambda(\beta)) - l(\beta, 0) = -\sum (\lambda(\beta))^\top x_i \psi(r_i(\beta)) + \frac{1}{2} \sum (\lambda(\beta))^\top x_i x_i^\top (\lambda(\beta)) \psi^2(r_i(\beta)) + o_p(1)$. On the other hand we have

$$\begin{aligned} \log(L_E(\beta)) &= -n \log n - \sum \log\{1 + (\lambda^*(\beta))^\top [x_i \psi(r_i(\beta)) - \epsilon_n]\} \\ &= -n \log n - \sum (\lambda^*(\beta))^\top x_i \psi(r_i(\beta)) \\ &\quad - \frac{1}{2} \sum (\lambda^*(\beta))^\top x_i x_i^\top (\lambda^*(\beta)) \psi^2(r_i(\beta)) + o_p(1). \end{aligned}$$

by Taylor expansion. As $\|\epsilon_n\| = o(n^{-1/2})$, we have $\lambda^*(\beta) = \lambda(\beta) + o_p(n^{-1/2})$, which completes the proof.

Proof of Theorem 5 (a) We only need to show (20) and the rest of the proof is straightforward by Lemma A.5(iii) and strict monotonicity of $\psi(u)$ around 0. First note that $\|S_{n(2)}(\tilde{\beta}, \tilde{\lambda})\| = o_p(n\|\epsilon_n\|)$ from (19). Also from (19)

$$\left\| \sum \psi(r_i(\beta)) x_i \{1 + (\lambda^*(\beta))^\top [x_i \psi(r_i(\beta)) - \epsilon_n]\}^{-1} \right\| = n\epsilon_n.$$

As $\lambda^*(\beta) = O_p(n^{-1/2})$ for all β within the surface ball of $\|\beta - \beta_0\| \leq n^{-1/2}$ and by (16) it follows that

$$\left\| b_1 \sum \frac{(\lambda^*(\beta))^\top x_i x_i}{1 + (\lambda^*(\beta))^\top [x_i \psi(r_i(\beta)) - \epsilon_n]} \right\| = O_p(n^{1/2}),$$

for all β within the surface ball of $\|\beta - \beta_0\| \leq n^{-1/2}$. By the definition of $(\tilde{\beta}, \tilde{\lambda})$ and the strict monotonicity of $\psi(u)$ about 0, we have

$$\left\| b_1 \sum \frac{(\lambda^*(\tilde{\beta}))^\top x_i x_i}{1 + (\lambda^*(\tilde{\beta}))^\top [x_i \psi(r_i(\tilde{\beta})) - \epsilon_n]} \right\| = o_p(n^{1/2}).$$

As $\lambda^*(\tilde{\beta}) = \lambda(\tilde{\beta}) + o_p(n^{-1/2})$, we have $\|S_{n(1)}(\tilde{\beta}, \tilde{\lambda})\| = o_p(n\|\epsilon_n\|)$ with some algebra.

Proof of Corollary 2 Let $\bar{\beta}_1(\theta_2) = \operatorname{argmax}_{\theta_1} L_E(\beta_1, \beta_2)$. With (17), (20) holds with $\|\epsilon_n\| = o(n^{-1})$. We follow similar lines of arguments as in the proof of Theorem 3(i) and we have $\|\bar{\beta}_1(\theta_2) - \tilde{\beta}_1\| = O_p(n^{-1})$. The rest of the proof is straightforward.

Proof of Lemma 5 Note that $F_n^{\tilde{\beta}, \tilde{\lambda}} = F_n^{\tilde{\zeta}, \tilde{\gamma}}$ where γ denotes the Lagrange multiplier defined with respect to ζ and x_i^* . Then, $\sum x_{1i}^* x_{2i}^{*\top} = dF_n^{\tilde{\zeta}, \tilde{\gamma}} = X_{1n}^{*\top} W_n X_{2n}^* = 0$ where $X_{1n}^* = X_{1n}$ and $X_{2n}^* = X_{n(2)} - X_{n(1)} (X_{n(1)}^\top)^{-1} X_{n(1)}$. As $\sum x_{1i}^* x_{2i}^{*\top} dF_n^{\tilde{\zeta}, \tilde{\gamma}} \rightarrow \Sigma_{x^*(12)}$, the proof is complete.

6.3 Figures and tables

Figure 1: Q-Q plots of $-2 \log$ conditional EL ratio $(\widetilde{W}_E(\zeta_{20}))$ and quantiles of χ_1^2 reference distribution from 3000 simulations with $n = 400$ each when $H_0 : \beta_2 = 2$ is under consideration

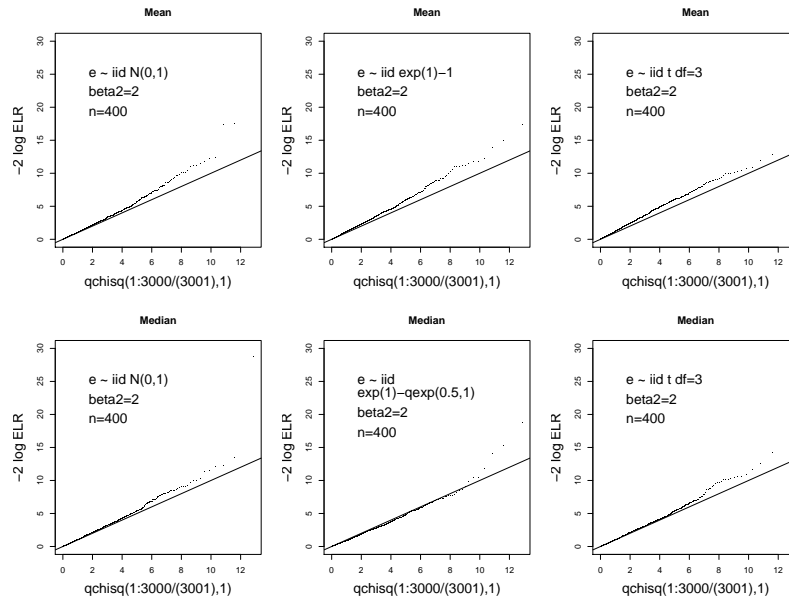


Figure 2: Q-Q plots of $-2 \log$ conditional EL ratio ($\widetilde{W}_E(\zeta_{20})$) and quantiles of χ^2_2 reference distribution from 3000 simulations with $n = 400$ each when $H_0 : \beta_2 = (2, 1)$ is under consideration

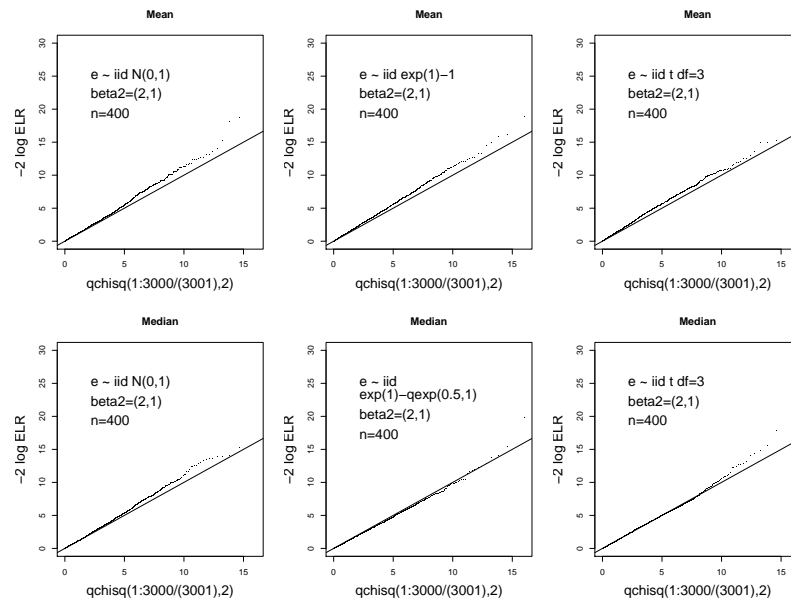
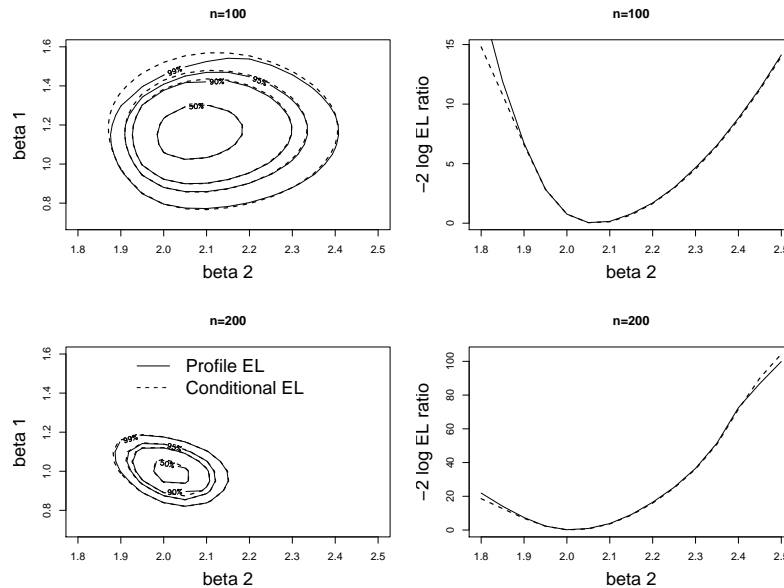


Table 1: Estimated type I-error rates for two EL ratio tests from 3000 simulations: ELR and ELR denote the tests by $\widetilde{W}_E(\zeta_2)$ and $\mathbb{W}_E(\zeta_2)$ respectively

			90%				95%			
			Mean		Median		Mean		Median	
			ELR	ELR	ELR	ELR	ELR	ELR	ELR	ELR
$\beta_2 = 2$	100	N(0,1)	.1417	.1280	.1300	.1117	.0830	.0743	.0703	.0587
		<i>exp</i>	.1683	.1557	.1103	.1157	.0997	.0917	.0597	.0630
		t_3	.1517	.1420	.1113	.1033	.0930	.0797	.0563	.0520
	200	N(0,1)	.1247	.1203	.1160	.1110	.0743	.0630	.0593	.0620
		<i>exp</i>	.1393	.1353	.0937	.0987	.0813	.0840	.0407	.0500
		t_3	.1327	.1393	.1113	.1073	.0757	.0767	.0603	.0610
	400	N(0,1)	.1147	.1107	.1070	.1073	.0607	.0593	.0560	.0533
		<i>exp</i>	.1257	.1207	.0883	.1013	.0707	.0680	.0430	.0513
		t_3	.1360	.1313	.1087	.1113	.0723	.0737	.0537	.0603
$\beta_2 = (2, 1)$	100	N(0,1)	.1407	.1317	.1243	.1150	.0810	.0763	.0673	.0590
		<i>exp</i>	.1717	.1643	.1023	.1123	.1040	.0993	.0513	.0593
		t_3	.1703	.1540	.1157	.1007	.1037	.0953	.0560	.0537
	200	N(0,1)	.1347	.1230	.1197	.1163	.0733	.0667	.0593	.0610
		<i>exp</i>	.1520	.1413	.0897	.0957	.0930	.0763	.0437	.0433
		t_3	.1473	.1390	.1087	.1053	.0853	.0893	.0570	.0550
	400	N(0,1)	.1163	.1163	.1170	.1147	.0660	.0653	.0610	.0580
		<i>exp</i>	.1237	.1243	.0897	.1113	.0677	.0693	.0457	.0587
		t_3	.1367	.1307	.0997	.1127	.0703	.0757	.0480	.0517

Figure 3: Plots of conditional ($\widetilde{W}_E(\zeta_2)$) and profile EL ratio functions ($\overline{W}_E(\beta_2)$)



References

- [1] BICKEL, P. J., KLAASSEN, C. A., RITOV, Y. and WELLNER, J. A. (1993). *Efficient and Adaptive Estimation for Semiparametric Models*. Johns Hopkins Univ. Press.
- [2] COX, D. R. and REID, N. (1987). Parameter orthogonality and approximate conditional inference. *J. R. Statist. Soc. B.* vol. 49. 1-39.
- [3] FREEDMAN, D. A. (1981). Bootstrapping Regression Models. *Annals of Statistics*. **9**. 1218-1228.
- [4] HJORT, N.L., MCKEAGUE, I.W. and KEILEGOM, I.V. (2006). Extending the scope of empirical likelihood. *Preprint. Submitted to Annals of Statistics*
- [5] KOENKER, R. (2005). *Quantile Regression*, Cambridge University Press.
- [6] LAZAR, N. A., (2003). Bayesian empirical likelihood *Biometrika*, vol. 90, 319-326
- [7] LAZAR, N. A. and MYKLAND, P. A. (1999). Empirical Likelihood in the Presence of Nuisance Parameters. *Biometrika*. vol 86. 203-211
- [8] LI, G. and WANG, Q.H. (2003). Empirical likelihood regression analysis for right censored data. *Statistica Sinica* **13** 51-68
- [9] MYKLAND, P. A., (1995). Dual Likelihood. *Annals of Statistics*. vol 23. 396-421

- [10] OWEN, A. B. (1988). Empirical likelihood ratio confidence intervals for a single functional. *Biometrika*, vol. 75, 237-249.
- [11] OWEN, A. (1991). Empirical Likelihood for Linear Models. *Annals of Statistics*, **19**, 4, 1725-1747.
- [12] OWEN, A. B., (2001). *Empirical likelihood*. Chapman & Hall/CRC, London
- [13] PAN, X.-R. and ZHOU, M. (1999). Using one-parameter sub-family of distributions in empirical likelihood ratio with censored data. *Journal of Statistical Planning and Inference*, **75**, 379-392.
- [14] PAN, X.-R. and ZHOU, M. (2002). Empirical likelihood ratio in terms of cumulative hazard function for censored data. *Journal of Multivariate Analysis*, **80**, 1, 166-188
- [15] QIN, G.S. and JING, B.Y. (2001). Empirical likelihood for censored linear regression. *Scand. J of Statist.*, **28**, 4, 661-673
- [16] QIN, J. and LAWLESS, J. (1994). Empirical likelihood and general estimating equations *Annals of Statistics*, **22**, 1, 300-325
- [17] RUPPERT, D. and CARROLL, R. J. (1980). Trimmed least squares estimation in the linear model *Journal of the American Statistical Association*, vol. 75, 828-838
- [18] STEIN, C. (1956). Efficient nonparametric testing and estimation. *Proc. 3rd Berkeley Symp. Math. Statist. Prob.*, **1**, Ed. Neyman, J., 166-188, University of California Press.
- [19] VAN DER VAART, A. W. (1988). *Statistical Estimation in Large Parameter Spaces*. Centrum voor Wiskunde en Informatica, Amsterdam.
- [20] ZHOU, M., BATHKE, A. and KIM, M. (2006). Empirical likelihood for Heteroscedastic AFT model. *Preprint. Submitted to J. Amer. Statist. Assoc.*

Table 2: Average length (Avl) and estimated coverage (Ecv) for three confidence interval methods, from 3000 simulations. The parenthesized numbers are the ratio of % miss below over % miss above the confidence intervals. The label ELR denotes the method by $\widetilde{W}_E(\zeta_2)$.

n	e_i		90%				95%				
			Mean		Median		Mean		Median		
			Avl	Ecv	Avl	Ecv	Avl	Ecv	Avl	Ecv	
100	N(0,1)	ELR	.23410	85.77	.30875	86.90	.28137	91.67	.37651	92.97	
		BP	.25217	87.97	.35241	94.30	.30344	92.77	.42448	97.17	
		BN	.25803	89.67	.36599	93.93	.30747	94.57	.43610	97.23	
	<i>exp</i>	ELR	.23141	83.00 (.619)	.31180	88.90	.28039	89.97 (.576)	.37487	94.00	
		BP	.24504	85.53 (.486)	.27816	94.10	.29501	91.80 (.398)	.33753	97.07	
		BN	.25199	88.37 (.460)	.29123	93.53	.30026	93.90 (.397)	.34702	96.90	
	t_3	ELR	.38602	84.80	.37155	88.83	.46951	90.70	.45491	94.37	
		BP	.40375	87.53	.41507	94.47	.48715	92.90	.50399	98.07	
		BN	.41565	91.23	.43384	95.13	.49528	95.93	.51695	97.73	
	200	N(0,1)	ELR	.16381	87.47	.21400	88.37	.19629	92.53	.25758	94.03
			BP	.16850	87.53	.23071	93.03	.20144	93.07	.27503	96.77
			BN	.17172	88.57	.23813	91.93	.20462	93.93	.28374	96.00
<i>exp</i>		ELR	.16280	86.07 (.526)	.22012	90.53	.19626	91.77 (.453)	.26387	95.87	
		BP	.16514	86.27 (.369)	.17949	93.40	.19685	91.97 (.324)	.21457	96.77	
		BN	.16854	88.10 (.347)	.18652	92.63	.20083	93.20 (.316)	.22225	96.57	
t_3		ELR	.27539	86.70	.24902	88.83	.33347	92.43	.30133	93.97	
		BP	.27463	88.30	.26483	93.07	.32888	93.30	.31800	96.87	
		BN	.28097	90.47	.27357	93.07	.33480	95.40	.32598	96.10	
400		N(0,1)	ELR	.11590	88.33	.14843	89.17	.13864	93.80	.17787	94.37
			BP	.11632	88.60	.15583	92.03	.13802	93.67	.18495	96.07
			BN	.11816	89.47	.16019	90.10	.14080	94.17	.19088	95.20
	<i>exp</i>	ELR	.11561	87.30 (.789)	.15575	91.10	.13894	92.87 (.740)	.18622	95.70	
		BP	.11489	86.97 (.602)	.12083	92.33	.13633	92.80 (.510)	.14358	96.30	
		BN	.11684	88.17 (.537)	.12480	91.17	.13922	94.23 (.352)	.14871	95.50	
	t_3	ELR	.19898	86.30	.16739	89.03	.24050	92.67	.20135	94.63	
		BP	.19396	87.00	.17393	91.67	.23044	92.37	.20737	95.80	
		BN	.19776	89.73	.17928	91.90	.23564	94.80	.21363	95.80	