Semiparametric analysis in double-sampling designs via empirical likelihood

Wen Yu
Department of Statistics, School of Management, Fudan University, Shanghai 200433, PR China

A R T I C L E   I N F O

Article history:
Received 30 August 2010
Available online 6 May 2011

AMS subject classifications:
62G05
62G20

Keywords:
Doubling sampling
Empirical likelihood
Growing constraints
Missing completely at random
Semiparametric efficiency
Wilks theorem

A B S T R A C T

Double-sampling designs are commonly used in real applications when it is infeasible to collect exact measurements on all variables of interest. Two samples, a primary sample on proxy measures and a validation subsample on exact measures, are available in these designs. We assume that the validation sample is drawn from the primary sample by the Bernoulli sampling with equal selection probability. An empirical likelihood based approach is proposed to estimate the parameters of interest. By allowing the number of constraints to grow as the sample size goes to infinity, the resulting maximum empirical likelihood estimator is asymptotically normal and its limiting variance–covariance matrix reaches the semiparametric efficiency bound. Moreover, the Wilks-type result of convergence to chi-squared distribution for the empirical likelihood ratio based test is established. Some simulation studies are carried out to assess the finite sample performances of the new approach.

© 2011 Elsevier Inc. All rights reserved.

1. Introduction

During the process of data collection, sometimes it is prohibitive to collect exact measurements on all variables of interest for the subjects because of the limitation of resource. Under these circumstances, a two-stage or double-sampling design may be adopted. At first, a primary sample of poor (proxy) measurements is drawn from the target population. At the second stage, a validation subsample is drawn from the primary sample. The exact information is collected for each subject in the validation subsample in addition to the proxy data. As a result, two sets of data, the proxy data on all subjects, along with the validation data on the subsample, are available for statistical analysis.

For the double-sampling designs, some efforts have been taken to provide statistical inferential procedures, cf., Tenenbein [12] on binomial data, Breslow and Cain [1] on logistic regression, Pepe and Fleming [8] on nonparametric approach, etc. Since exact information for the whole sample is not observed completely, the two-stage sampling designs can be put into the framework of missing data analysis. Robins et al. [11] provided a general class of estimators for missing data under the missing at random (MAR) assumption. The estimators proposed by them include all possible regular asymptotic linear (RAL) estimators and can be extended to the double-sampling designs under suitable conditions. The semiparametric efficiency bound is attainable by choosing the optimal estimating function. However, the optimal one involves the knowledge about the underlying joint distribution of the two samples, and is therefore difficult to implement. More recently, Chen and Chen [3] proposed a simple estimation procedure when the validation subsample is randomly chosen from the primary sample with equal selection probability. This corresponds to the missing completely at random (MCAR) assumption. Their estimator belongs to the general class of Robins et al. [11]. Instead of constructing the optimal estimating function, they used a relatively simple function as the augmentation part.
The methods developed by Robins et al. [11] and Chen and Chen [3] require that the number of unknown parameters is equal to that of estimating functions. However, in many problems, people may have more estimating functions than unknown parameters. Qin and Lawless [10] introduced the empirical likelihood method to utilize all the estimating functions. Empirical likelihood was first introduced by Owen [5,6] for constructing generalized likelihood ratio test statistics and corresponding confidence regions. In Qin and Lawless’s [10] work, they used the method to do estimation and showed that asymptotically, the empirical likelihood function is able to make the optimal linear combination of the original estimation functions automatically. Therefore, empirical likelihood has been widely used in these so-called over-identified problems. Another important feature of empirical likelihood is that the approach may utilize auxiliary information in the data. This feature is attractive since in double-sampling designs, there usually exist auxiliary information in the primary sample, which could be utilized to improve the estimation efficiency.

In this paper, we develop a new empirical likelihood approach for the double-sampling designs when the validation sample is a random subsample of the primary data. The Bernoulli sampling with equal selection probability is used to draw the validation sample. That is, the sampling of the validation sample does not dependent on the primary data. The new approach is able to deal with over-identified situations flexibly, as well as utilize auxiliary information by introducing constraints from the primary data. Moreover, we allow the number of constraints grows with the sample size at certain rates. By doing so, the resulting estimator reaches the semiparametric efficiency bound asymptotically under suitable conditions. The growing constraints in empirical likelihood was discussed by Hjort et al. [4]. Wu and Ying [14] proposed an empirical likelihood approach with growing constraints to covariate adjustment in randomized clinical trials. Compared with [11], the proposed empirical likelihood approach automatically attains the semiparametric efficiency bound without the need of constructing optimal estimating equations. The empirical likelihood ratio tests are also available for making inferences.

The rest of the paper is organized as follows. In Section 2, we give out the model specification and briefly discuss the semiparametric efficiency bound. In Section 3, the new empirical likelihood approach with growing constraints is introduced and its asymptotic properties are explored. Some numerical results are presented in Section 4. Section 5 contains concluding remarks and some discussions. All technical details are summarized in the Appendix.

2. Notation and model specification

Assume that the primary or proxy data, \( \tilde{X}_i, i = 1, 2, \ldots, n \), are i.i.d. copies drawn from the proxy population represented by a \( d \)-dimensional random vector \( \tilde{X} \) at the first stage. At the second stage, a random subset of \( \{1, 2, \ldots, n\} \), denoted by \( V \), with size \( n_V (n_V < n) \) is obtained. For each subject in \( V \), the exact information \( X (X \in \mathbb{R}^d) \) is measured. Consequently, the two available data sets are the primary sample \( \{X_i, i = 1, 2, \ldots, n\} \) and the validation sample \( \{\tilde{X}_i, i = 1, 2, \ldots, n\} \). The distribution of \( X \), denoted by \( F \), is the target population. Suppose that we are interested in making inference about a \( p \)-dimensional parameter \( \theta \) associated with \( F \). The information about \( \theta \) is contained in a set of functionally independent unbiased estimating functions \( g(X; \theta) = (g_1(X; \theta), g_2(X; \theta), \ldots, g_p(X; \theta))^T \), where \( s \geq p \), that is, we have

\[
E \left[ g(X; \theta_0) \right] = 0,
\]

where \( \theta_0 \) denotes the true value of \( \theta \). Note that when \( s > p \), model [1] corresponds to the over-identified situations discussed by Qin and Lawless [10].

As we mentioned, the two-stage sampling design can be put into the framework of missing data analysis. Introduce a dichotomous variable \( \delta \). Set \( \delta = 1 \) if an observation belongs to the validation sample and \( \delta = 0 \) otherwise. Then for the individual with \( \delta_i = 1 \), the available data is \( (\tilde{X}_i, X_i) \), while for \( \delta_i = 0 \), one can only observe \( \tilde{X}_i \) and \( X_i \) is unobservable, or equivalently, is missing. Let \( \{(X_i, \tilde{X}_i, \delta_i), i = 1, 2, \ldots, n\} \) be the i.i.d. copies of \( (X, \tilde{X}, \delta) \). That the validation data is drawn by the Bernoulli sampling with equal selection probability means that \( \delta_i \) is independent of \( (\tilde{X}_i, X_i) \), which corresponds to the assumption that the missing data are MAR.

Assume that the selection probability, \( \rho = P(\delta = 1) \), is known. When \( s = p \), Robins et al. [11] proposed a general class of RAL estimators with influence functions

\[
-A^{-1} \left( \frac{\delta}{\rho} g(X; \theta_0) - \frac{\delta - \rho}{\rho} \kappa(\tilde{X}) \right),
\]

where \( A = E \left[ \frac{\partial g(X; \theta_0)}{\partial \theta^T} \right] \) and \( \kappa(\cdot) \) is arbitrary \( p \)-dimensional function of the primary data. The second term in the bracket of (2) is known as the augmentation part, which can be used to improve the efficiency of the estimation. Intuitively, different choice of \( \kappa(\cdot) \) would result in different influence function with different efficiency. Robins et al. [11] pointed out the optimal choice of \( \kappa(\cdot) \), which results in the semiparametric efficient RAL estimator, is given by

\[
\kappa^*(\tilde{X}) = E \left[ g(X; \theta_0) | \tilde{X} \right],
\]

and the corresponding influence function with \( \kappa^*(\cdot) \) is the efficient influence function. A review of the related theory may be found in [13].

For the over-identified case (i.e. \( s > p \)), Cattaneo [2] gave out the efficient influence function

\[
- \left( A^T \Sigma^{-1} A \right)^{-1} A \Sigma^{-1} \left( \frac{\delta}{\rho} g(X; \theta_0) - \frac{\delta - \rho}{\rho} E \left[ g(X; \theta_0) | \tilde{X} \right] \right),
\]

(3)
For a given θ, it is straightforward that performing a linear transformation on b

E

Moreover, the efficient influence function

E

series, arbitrary functions. Now let

b

with suitable constraints. It is well known that the linear combination of a series of basis functions may approximate arbitrary functions. Now let

\( b_n(x) = (b_1(x), b_2(x), \ldots, b_{n}(x))^T \) \( x \in \mathbb{R}^d \) be a series of basis functions, where the number of series, \( r_n \), will grow to infinity with the sample size \( n \) at a certain rate. Loosely speaking, it is expectable that under suitable conditions, the conditional expectation \( E(g(X; \theta_0)|\tilde{X}) \) can be approximated by some proper linear combination of \( b_{\theta}(\tilde{X}) \). Moreover, the efficient influence function (3) itself is a linear combination of \( \delta g(X; \theta_0)/\rho \) and \( (\delta - \rho)E(g(X; \theta_0)|\tilde{X})/\rho \). In light of these facts, we propose the following constraints

\[
\sum_{i=1}^{n} p_i \delta_i g(X_i; \theta) = 0; \quad \sum_{i=1}^{n} p_i (\delta_i - \rho) b_{\theta}(\tilde{X}_i) = 0
\]

as well as the standard unit total probability constraint

\[
\sum_{i=1}^{n} p_i = 1
\]

in addition to the empirical likelihood function (4). Note that the first set of constraints in (5) corresponds to the fact that \( E(\delta g(X; \theta_0)) = 0 \), while the second set corresponds to \( E[(\delta - \rho)b_{\theta}(\tilde{X})] = 0 \). The number of elements of \( b_{\theta}(\tilde{X}) \) grows to infinity as \( n \to \infty \), so does the number of constraints in (5).

For each fixed \( \theta \), maximizing (4) subject to (5) and (6) gives out the empirical likelihood function,

\[
\mathcal{L}_{n}(\theta) = \max_{p_i \geq 0 \left\{ \frac{1}{n} \sum_{i=1}^{n} p_i \delta_i g(X_i, \theta) = 0, \sum_{i=1}^{n} p_i (\delta_i - \rho) b_{\theta}(\tilde{X}_i) = 0, \sum_{i=1}^{n} p_i = 1 \right\} .
\]

It is straightforward that performing a linear transformation on \( b_{\theta}(\cdot) \) does not change the second set of constraints, that is, \( b_{\theta}(\tilde{X}_i) \) can be replaced by \( W_{n}b_{\theta}(\tilde{X}_i) \), where \( W_{n} \) is a non-random \( r_n \times r_n \) matrix. Let

\[
m_{i,n}(\theta) = \left( \frac{\delta_i}{\rho} g(X_i; \theta)^T, \frac{\delta_i - \rho}{\rho} (W_{n}b_{\theta}(\tilde{X}_i))^T \right)^T.
\]

For a given \( \theta \), a unique solution can be obtained through the Lagrange multipliers, provided that 0 is inside the convex hull of \( \{m_{i,n}(\theta), i = 1, 2, \ldots, n\} \), and is given by

\[
p_i(\theta) = \frac{1}{n} \left[ 1 + \eta_n(\theta)^T m_{i,n}(\theta) \right]^{-1}.
\]
where \( \eta_n(\theta) \) solves the following equation of \( \eta \)

\[
\frac{1}{n} \sum_{i=1}^{n} \frac{\mathbf{m}_{in}(\theta)}{1 + \eta_i^2 \mathbf{m}_{i,n}(\theta)} = 0.
\]  

(7)

Consequently, \( L_n(\theta) = \prod_{i=1}^{n} p_i(\theta) \). Since \( \prod_{i=1}^{n} p_i \) is maximized for \( p_i = n^{-1} \) in the absence of constraints (5), the minus log-empirical likelihood ratio is given by

\[
L_n(\theta) = -\log \left( \prod_{i=1}^{n} p_i(\theta) \right) = \sum_{i=1}^{n} \log \left( 1 + \eta_n(\theta)^T \mathbf{m}_{i,n}(\theta) \right).
\]

We may minimize \( L_n(\theta) \) to obtain an estimator, denoted by \( \hat{\theta}_{MELE} \), of \( \theta_0 \), called the maximum empirical likelihood estimator (MELE). If \( r_n \) is fixed, by Qin and Lawless’s [10] result, \( \hat{\theta}_{MELE} \) has the smallest asymptotic variance among all the \( p \)-dimensional linear combinations of \( \mathbf{m}_{i,n}(\theta) \). Now given \( r_n \) growing to infinity as \( n \to \infty \) at certain rate, we expect the MELE \( \hat{\theta}_{MELE} \) to reach the semiparametric efficiency bound under suitable conditions. Let

\[
\mathbf{m}_n(\theta) = \left( \frac{\delta}{\rho} \mathbf{g}(X; \theta)^T, \frac{\delta - \rho}{\rho} W_n \mathbf{b}_n(\bar{X})^T \right)^T
\]

and \( \Sigma_{n,m} = E(\mathbf{m}_n(\theta_0) \mathbf{m}_n(\theta_0)^T) \). Throughout, \( \| \cdot \| \) is used to denote the Euclidean norm. We assume the following conditions:

(C.1) \( A \) is of rank \( p \). There exists a neighborhood of \( \theta_0 \), denoted by \( \Theta \), and an integrable function \( M(X) \) such that \( \sup_{\theta \in \Theta} \| \mathbf{g}(X; \theta) \|^3 \leq M(X), \sup_{\theta \in \Theta} \| \partial \mathbf{g}(X; \theta)/\partial \theta \| \leq M(X) \) and \( \| \partial^2 \mathbf{g}(X; \theta)/\partial \theta \| \leq M(X) \); \( \partial \mathbf{g}(X; \theta)/\partial \theta \) and \( \partial^2 \mathbf{g}(X; \theta)/\partial \theta \) are continuous in \( \theta \).

(C.2) Components of \( W_n \mathbf{b}_n(\bar{X}) \) are uniformly bounded by a finite constant \( C > 0 \).

(C.3) Eigenvalues of \( \Sigma_{n,m} \) are bounded away from zero and infinity.

(C.4) There exists a non-random \( p \times r_n \) matrix \( B_n \) such that

\[
B_n W_n \mathbf{b}_n(\bar{X}) \to E \left( \mathbf{g}(X; \theta_0) | \bar{X} \right) \text{ in } L^2 \text{ as } n \to \infty.
\]

(C.5) The growth rate of \( r_n \) is limited to \( r_n = o(n^{1/3}) \).

(C.6) \( \Sigma^* \) is positive definite.

The asymptotic properties of the MELE \( \hat{\theta}_{MELE} \) are then to be explored. We prove in the Appendix the following theorem:

**Theorem 1.** Under conditions (C.1)–(C.6), \( \sqrt{n}(\hat{\theta}_{MELE} - \theta_0) \) converges in distribution to the \( N(0, (A^T \Sigma^*^{-1} A)^{-1}) \) distribution as \( n \to \infty \).

Theorem 1 claims that the proposed MELE is asymptotically semiparametric efficient under the listed conditions when the number of constraints grow to infinity at a certain rate. This is desirable since we do not have to construct the optimal estimating function and the MELE automatically reaches the semiparametric efficiency bound. Moreover, the fact that the MELE will not be affected by performing linear transformation of the constraints greatly facilitates the applicability of our approach since we can just put all the constraints we have without forming the proper combination of them. For instance, it is possible that \( E(\mathbf{m}_n \mathbf{m}_n^T) \) is ill conditioned but we can still use \( \mathbf{m}_n \) as the constraints as long as there exists a \( W_n \) such that \( E(\mathbf{m}_n \mathbf{m}_n^T) \) is better conditioned, where \( \mathbf{m}_n = W_n \mathbf{m}_n \).

Next we give some discussions about the conditions we assumed. Condition (C.1) are the regular conditions for the estimating functions \( \mathbf{g}(X; \theta) \), which can be found in [10]. Condition (C.5) gives the upper bound on the growth rate of the number of constraints at which a well-behaved MELE can be obtained. It is of course of theoretical interest to find out the optimal rate of the growth, but we will not take trouble to discuss here since this is not our main concern. Condition (C.6) requires a non-degenerate variance-covariance matrix for \( \delta \mathbf{g}(X; \theta_0)/\rho - (\delta - \rho)E(\mathbf{g}(X; \theta_0) | \bar{X})/\rho \) and is fairly mild. The key part lies in (C.2)–(C.4). To satisfy the conditions, some smoothness assumptions should be imposed on \( E(\mathbf{g}(X; \theta_0) | \bar{X}) \) and we also need to use certain basis functions with orthogonality and boundedness. For example, suppose that \( d = 1 \) and let \( \tilde{F} \) be the cumulative distribution function (CDF) of the primary population \( \bar{X} \). The basis functions \( \mathbf{b}_n(\bar{X}) \) can be chosen as

\[
(1, \sin(2\pi \tilde{F}(\bar{X})), \cos(2\pi \tilde{F}(\bar{X})), \sin(4\pi \tilde{F}(\bar{X})), \cos(4\pi \tilde{F}(\bar{X})), \ldots, \sin(2\pi r_n \tilde{F}(\bar{X})), \cos(2\pi r_n \tilde{F}(\bar{X})))
\]

which are the Fourier series. By applying the fact that these basis are orthogonal when their arguments are distributed as \( U[0, 1] \) and they are bounded, we can show that (C.2) and (C.3) hold. (C.4) can be verified by taking the expansion of
the conditional expectation as long as some smoothness conditions are satisfied. By making use of multivariate Fourier expansion, the arguments can be generalized to \(d \geq 2\) situations. Similarly, some orthogonal and bounded basis functions other than the Fourier can be used. For instance, we can use the Legendre polynomials \(1, x, (3x^2 - 1)/2, \ldots\), and the argument takes the form of \(2\hat{F}(\bar{X}) - 1\). Note that the Legendre polynomials are linear transformations of polynomial terms \(1, x, x^2, \ldots\).

Thus we can just use polynomial terms in the constraints due to linear transformation invariance of the proposed empirical likelihood. In applications, usually \(\hat{F}\) is unknown. One may use the empirical CDF \(\hat{F}_n(x) = n^{-1} \sum_{i=1}^n 1_{[x_i \leq x]}\) instead and the conclusion of Theorem 1 does not change.

Note that Cattaneo [2] also applied series function method to approximate \(E(g(X; \theta_0)|\bar{X})\). In his work, the conditional expectation is estimated by an explicit linear combination of certain series functions and the number of series goes to infinity at a well-chosen rate. In order to get the linear combination, one has to use proper approach to estimate the combination coefficients. By contrast, the proposed empirical likelihood approach automatically gives out the optimal linear combination of the constraints in asymptotic sense. There is no need to estimate the combination coefficients explicitly.

In order to make inferences about \(\theta_0\), one needs to estimate the semiparametric efficiency bound. Let \(\bar{m}_n(\theta) = n^{-1} \sum_{i=1}^n m_{i,n}(\theta)\) and \(S_n(\theta) = n^{-1} \sum_{i=1}^n m_{i,n}(\theta)m_{i,n}(\theta)^T\). The following theorem, proved in the Appendix, gives out a consistent estimator using the usual plug-in method.

**Theorem 2.** Under conditions (C.1)–(C.6),

\[
\left\| \left( \frac{\partial \bar{m}_n(\hat{\theta}_{EL})}{\partial \theta^T} \right)^T S_n(\hat{\theta}_{EL})^{-1} \left( \frac{\partial \bar{m}_n(\hat{\theta}_{EL})}{\partial \theta^T} \right) - \left( A^T \Sigma^* A \right)^{-1} \right\| \rightarrow 0
\]

in probability as \(n \rightarrow \infty\).

The results claimed by Theorems 1 and 2 enable us to construct the Wald-type confidence region for \(\theta_0\). Since the approach is based on empirical likelihood, an empirical likelihood ratio test is available. Similar to the case with fixed number of constraints discussed by Qin and Lawless [10], the log-empirical likelihood ratio test statistic is defined as

\[
T_{1n}(\theta) = 2I_n(\theta) - 2I_n(\hat{\theta}_{EL}).
\]

A Wilks-type theorem of convergence to the chi-squared distribution is given as follows.

**Theorem 3.** Under conditions (C.1)–(C.6), \(T_{1n}(\theta_0)\) converges in distribution to \(\chi^2_p\) as \(n \rightarrow \infty\).

Based on Theorem 3, an empirical likelihood based confidence region can be constructed. Moreover, when only certain subset of \(\theta_0\) is of interest, one may turn to the profile empirical likelihood approach. Intuitively, similar to the estimating problem, adding more constraints will result in more powerful test. When the number of constraints grows to infinity, the corresponding tests become asymptotically most powerful under contiguous alternatives.

### 4. Numerical studies

In this section we carry out some simulation studies to assess the finite sample performances of the proposed empirical likelihood based estimation procedure.

**Example 1.** Related mean and variance.

First we consider a variable with information relating the first and second moments. The exact measurement, \(X \in \mathbb{R}\) (i.e., \(d = 1\)), has the first and second moments \(E(X) = \theta_0\) and \(E(X^2) = m(\theta_0)\) where \(m(\cdot)\) is a known function. In this example, \(X\) is generated from the \(N(\theta_0, \theta_0^2 + 1)\), which implies \(m(\theta) = 2\theta^2 + 1\). \(\theta_0\) is the parameter of interest and consequently we have two estimating functions

\[g_1(X; \theta) = X - \theta, \quad g_2(X; \theta) = X^2 - 2\theta^2 - 1,\]

that is, \(p = 1\) and \(s = 2\). \(\theta_0\) is set to be 1. The primary population \(\bar{X} = X + \epsilon\), where \(\epsilon\) follows the \(N(0, 1)\) distribution and is independent of \(X\).

The primary sample size \(n\) is taken to be \(500\) and \(1000\). The validation percentage \(\rho\) is taken to be \(0.4\). We use the Fourier series

\[(1, \sin(2k \pi \bar{F}_n(x)), \cos(2k \pi \bar{F}_n(x)), \quad k = 1, 2, \ldots, K)\]

for \(b_n(\cdot)\) in the constraints (5) and gradually add the number of series \(K\). 1000 data sets are generated. For each replication, we calculate four MELEs with different number of series functions. “Fourier 2” corresponds to \(K = 1\), “Fourier 4” corresponds to \(K = 2\), “Fourier 6” corresponds to \(K = 3\), “Fourier 8” corresponds to \(K = 4\) and “Fourier 10” corresponds to \(K = 5\). We also calculate the MELE based on the validation data only (“vdo”) for comparison.
Theorem 2

Theorem 3

Common mean.

Table 2

Example 1

Growing the number of constraints, the efficiency of the proposed MELE gradually increases and approaches the semiparametric estimators. We also calculate the MELE based on the validation data only ("vdo") for comparison.

 respectivley. Again we gradually grow \( b \) and herewe grow \( K \) till 3. 1000 data sets are generated. “Fourier 2” corresponds to \( K = 1 \), “Fourier 4” corresponds to \( K = 2 \), “Fourier 6” corresponds to \( K = 3 \) and “Fourier 8” corresponds to \( K = 4 \). Besides the proposed estimators, we also calculate the MELE based on the validation data only (“vdo”) for comparison.

The notations have the same meaning as those in Table 1. Similar findings can be obtained from Table 2. Firstiy, the proposed MELEs based on the two samples are more efficient than the MELE based on the validation data only. Secondy, by growing the number of constraints, the efficiency of the proposed MELE gradually increase and approach the semiparametric

Example 2. Common mean.

Next we consider a bivariate vector with common mean. The exact measurement \( X = (X_1, X_2)^T \in \mathbb{R}^2 \) (i.e., \( d = 2 \)) satisfies \( E(X_1) = E(X_2) = \theta_0 \). In this example, \( X_1 \) is generated from the \( \text{Exp}(0.5) \), \( X_2 \) is generated from the \( \mathcal{U}(0, 4) \) distribution and \( X_1 \) and \( X_2 \) are independent. The parameter of interest is the common mean \( \theta_0 \) and consequently we have two estimating functions

\[
g_1(X; \theta) = X_1 - \theta, \quad g_2(X; \theta) = X_2 - \theta,
\]

that is, \( p = 1 \) and \( s = 2 \). It is easy to see that \( \theta_0 = 2 \). The primary population \( \tilde{X}_1 = X_1 + \varepsilon_1 \) and \( \tilde{X}_2 = X_2 + \varepsilon_2 \), where \( \varepsilon_1 \) and \( \varepsilon_2 \) follow the \( N(-0.5, 1) \) and the \( N(-0.5, 0.5) \) distribution independently.

The primary sample size \( n \) is taken to be 500 and 1000. The validation percentage \( \rho \) is taken to be 0.4. In this example the multivariate Fourier series

\[
(1, \sin(2k\pi \tilde{F}_{1n}(x_1)), \cos(2k\pi \tilde{F}_{1n}(x_1)), \sin(2k\pi \tilde{F}_{2n}(x_2)), \cos(2k\pi \tilde{F}_{2n}(x_2)), \quad k = 1, 2, \ldots, K)
\]

are applied for \( \hat{b}_n(\cdot) \) in the constraints (5), where \( \tilde{F}_{1n} \) and \( \tilde{F}_{2n} \) are the empirical CDF based on the samples of \( \tilde{X}_1 \) and \( \tilde{X}_2 \), respectively. Again we gradually grow \( K \) and here we grow \( K \) till 3. 1000 data sets are generated. “Fourier 2” corresponds to \( K = 1 \), “Fourier 4” corresponds to \( K = 2 \), “Fourier 6” corresponds to \( K = 3 \) and “Fourier 8” corresponds to \( K = 4 \). Besides the proposed estimators, we also calculate the MELE based on the validation data only (“vdo”) for comparison.

\begin{table}[ht]
\centering
\caption{Simulation results for Example 1.}
\begin{tabular}{|c|c|c|c|c|c|c|}
\hline
\( n(\rho) \) & Estimator & Bias & Var & Est. Var & Wald & ELR \\
\hline
500 (0.4) & vdo & 0.0009 & 0.0048 & 0.0049 & 0.896 & 0.946 \\
& Fourier 2 & -0.0006 & 0.0042 & 0.0041 & 0.895 & 0.943 \\
& Fourier 4 & -0.0002 & 0.0038 & 0.0039 & 0.895 & 0.946 \\
& Fourier 6 & -0.0001 & 0.0037 & 0.0037 & 0.892 & 0.946 \\
& Fourier 8 & -0.0001 & 0.0037 & 0.0036 & 0.891 & 0.939 \\
& Fourier 10 & 0.0007 & 0.0036 & 0.0036 & 0.888 & 0.939 \\
& Optimal & 0.0001 & 0.0033 & & & 0.891 & 0.945 \\
\hline
1000 (0.4) & vdo & 0.0004 & 0.0025 & 0.0025 & 0.897 & 0.947 \\
& Fourier 2 & 0.0014 & 0.0021 & 0.0021 & 0.897 & 0.949 \\
& Fourier 4 & 0.0014 & 0.0019 & 0.0019 & 0.901 & 0.949 \\
& Fourier 6 & 0.0016 & 0.0019 & 0.0019 & 0.890 & 0.937 \\
& Fourier 8 & 0.0015 & 0.0019 & 0.0018 & 0.892 & 0.939 \\
& Fourier 10 & 0.0011 & 0.0018 & 0.0018 & 0.892 & 0.939 \\
& Optimal & 0.0008 & 0.0017 & & & 0.897 & 0.939 \\
\hline
\end{tabular}
\end{table}
Table 2
Simulation results for Example 2.

<table>
<thead>
<tr>
<th>Estimator</th>
<th>Parameter</th>
<th>Bias</th>
<th>Var</th>
<th>Est. Var</th>
<th>Wald 0.90</th>
<th>Wald 0.95</th>
<th>ELR 0.90</th>
<th>ELR 0.95</th>
</tr>
</thead>
<tbody>
<tr>
<td>vdo</td>
<td></td>
<td>-0.0008</td>
<td>0.0048</td>
<td>0.0049</td>
<td>0.899</td>
<td>0.948</td>
<td>0.899</td>
<td>0.949</td>
</tr>
<tr>
<td>Fourier 2</td>
<td></td>
<td>-0.0004</td>
<td>0.0032</td>
<td>0.0034</td>
<td>0.910</td>
<td>0.954</td>
<td>0.913</td>
<td>0.959</td>
</tr>
<tr>
<td>Fourier 4</td>
<td></td>
<td>-0.0007</td>
<td>0.0029</td>
<td>0.0030</td>
<td>0.895</td>
<td>0.958</td>
<td>0.905</td>
<td>0.964</td>
</tr>
<tr>
<td>Fourier 6</td>
<td></td>
<td>-0.0005</td>
<td>0.0027</td>
<td>0.0028</td>
<td>0.907</td>
<td>0.961</td>
<td>0.911</td>
<td>0.963</td>
</tr>
<tr>
<td>Fourier 8</td>
<td></td>
<td>-0.0010</td>
<td>0.0026</td>
<td>0.0027</td>
<td>0.908</td>
<td>0.956</td>
<td>0.915</td>
<td>0.959</td>
</tr>
<tr>
<td>Optimal</td>
<td></td>
<td>-0.0003</td>
<td>0.0024</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 3
Simulation results for Example 3.

<table>
<thead>
<tr>
<th>Estimator</th>
<th>Parameter</th>
<th>Bias</th>
<th>Var</th>
<th>Est. Var</th>
<th>Wald(0.95)</th>
<th>Est. Var</th>
<th>Wald(0.95)</th>
</tr>
</thead>
<tbody>
<tr>
<td>MLE</td>
<td>$\beta_0$</td>
<td>-0.0341</td>
<td>0.0631</td>
<td>0.0620</td>
<td>0.953</td>
<td>-0.0157</td>
<td>0.0307</td>
</tr>
<tr>
<td></td>
<td>$\beta_1$</td>
<td>0.0571</td>
<td>0.1051</td>
<td>0.0951</td>
<td>0.947</td>
<td>0.0242</td>
<td>0.0462</td>
</tr>
<tr>
<td>C&amp;C</td>
<td>$\beta_0$</td>
<td>-0.0158</td>
<td>0.0536</td>
<td>0.0496</td>
<td>0.940</td>
<td>-0.0119</td>
<td>0.0253</td>
</tr>
<tr>
<td></td>
<td>$\beta_1$</td>
<td>0.0377</td>
<td>0.0868</td>
<td>0.0744</td>
<td>0.931</td>
<td>0.0225</td>
<td>0.0406</td>
</tr>
<tr>
<td>Fourier 2</td>
<td>$\beta_0$</td>
<td>-0.0314</td>
<td>0.0549</td>
<td>0.0519</td>
<td>0.947</td>
<td>-0.0160</td>
<td>0.0273</td>
</tr>
<tr>
<td></td>
<td>$\beta_1$</td>
<td>0.0518</td>
<td>0.0883</td>
<td>0.0762</td>
<td>0.930</td>
<td>0.0233</td>
<td>0.0407</td>
</tr>
<tr>
<td>Fourier 4</td>
<td>$\beta_0$</td>
<td>-0.0307</td>
<td>0.0528</td>
<td>0.0484</td>
<td>0.933</td>
<td>-0.0163</td>
<td>0.0257</td>
</tr>
<tr>
<td></td>
<td>$\beta_1$</td>
<td>0.0505</td>
<td>0.0842</td>
<td>0.0708</td>
<td>0.930</td>
<td>0.0230</td>
<td>0.0395</td>
</tr>
<tr>
<td>Fourier 6</td>
<td>$\beta_0$</td>
<td>-0.0313</td>
<td>0.0517</td>
<td>0.0464</td>
<td>0.933</td>
<td>-0.0165</td>
<td>0.0250</td>
</tr>
<tr>
<td></td>
<td>$\beta_1$</td>
<td>0.0530</td>
<td>0.0835</td>
<td>0.0684</td>
<td>0.922</td>
<td>0.0235</td>
<td>0.0393</td>
</tr>
</tbody>
</table>

efficiency bound. In this example, $K = 3$ seems to be an adequate choice from the view of variance reduction. Thirdly, the plug-in estimator for the asymptotic variance estimates the variance quite well. Finally, both the Wald-type and empirical likelihood based confidence intervals have adequate coverage probabilities.

Example 3. Errors-in-variables.

In the third example we consider the logistic regression model with errors in covariate. The exact measurement $X = (Y, Z)^T \in \mathbb{R}^2 \ (i.e., \ d = 2)$, where $Y$ is a binary 0–1 variable, satisfies the logistic regression $P(Y = 1|Z) = 1/(1 + \exp(-\beta_0 - \beta_1Z))$. $\beta_0 = 0, \beta_1 = 1$ and $Z$ is generated from a log-normal distribution with $\log(Z)$ following the $N(-0.5, 1)$ distribution. The parameters of interest are $\beta_0$ and $\beta_1$ and the corresponding estimating functions are

$$
g_1(X; \beta_0, \beta_1) = Y - \frac{\exp(\beta_0 + \beta_1Z)}{1 + \exp(\beta_0 + \beta_1Z)}, \quad g_2(X; \beta_0, \beta_1) = Z \left( Y - \frac{\exp(\beta_0 + \beta_1Z)}{1 + \exp(\beta_0 + \beta_1Z)} \right),
$$

that is, $p = s = 2$. The primary population $\tilde{X} = (Y, \tilde{Z})^T$, where $\log(\tilde{Z})$, given $Z$, follows the $N(\log(Z), 0.5)$ distribution. It means that there exist measurement error in the covariate $Z$.

The primary sample size $n$ is taken to be 500 and 1000. The validation percentage $\rho$ is taken to be 0.4. Again, the multivariate Fourier series are applied for $\mathbf{b}_n(\cdot)$ in the constraints (5). 1000 data sets are generated. “Fourier 2”, “Fourier 4” and “Fourier 6” refer to the proposed MELEs with growing number of constraints. Besides the proposed estimators, the usual maximum likelihood estimator based on the validation data only (“MLE”) is calculated for comparison. Moreover, we also included the estimator proposed by Chen and Chen [3] (“C&C”) which is applicable when $p = s$. The simulated bias, simulated variance, simulated mean of the estimated variance and Wald-type confidence interval with 95% confidence level (“Wald (0.95)”) are recorded. All the simulation results are summarized in Table 3.

From the table, we find out that both Chen and Chen’s [3] estimator and the proposed MELEs based on the two samples are more efficient than the MLE based on the validation data only. Moreover, by growing the number of constraints, the simulated variances for the proposed MELEs decrease and the proposed MELEs could possess smaller simulated variances than Chen and Chen’s [3] estimator. In general, the plug-in estimator for the asymptotic variance estimates the variance adequately and the Wald-type confidence intervals have reasonable coverage probabilities. Nevertheless, the results indicate that the variance estimates may slightly suffer from underestimating the true variance for larger number of constraints.

To conclude, the illustrated examples show that the proposed method has reasonable finite sample performances. The simulation results validate the theoretical findings presented in Section 3.
5. Conclusion and discussion

Two-stage sampling designs are often applied when collecting exact measurements on all variables of interest for the subjects is infeasible. If the validation sample is a random subsample with equal selection probability, the problem can be viewed as a missing data problem with MCAR. An empirical likelihood based approach is proposed to estimate the parameters of interest. With the growing constraints, the resulting MELE reaches the semiparametric efficiency bound asymptotically without constructing optimal estimating functions. The efficiency bound can be consistently estimated. The corresponding log-empirical likelihood ratio test statistic is shown to have usual standard chi-squared limiting distribution. The computation of the proposed inferential procedure is easy to implement by adopting some standard algorithms for empirical likelihood.

One may concern about the choice of basis functions in the constraints. Although we have provided two examples, it appears to us that there is no universal way to deal with this issue. A related issue is the number of basis functions to choose. One ad hoc way to choose is to consider variance reduction when additional constraints are added, as we do in the numerical studies. Usually if initial basis functions are properly chosen, only small number of constraints will be needed. Both the choice of optimal basis functions and the number of constraints are interesting challenges to be tackled in future.

Note that the proposed method requires a known validation percentage $\rho$. It is quite common to know $\rho$ when the second stage of sampling is designed by the researchers. In the cases when the selection percentage is unavailable, one may need to estimate it from the data, for example, estimate it by the sampling percentage $\hat{\rho}$. When $\rho$ is estimated from the data, the basis functions should be chosen carefully to keep the asymptotic optimality of the proposed method. One sufficient way is to choose basis function with mean zero. For example, one can use $\tilde{b}_i(X) = b_i(X) - \hat{E}[b_i(X)]$ as the basis functions and replace the unknown expectation by its sample analogy. Using arguments similar to those in [15], one can show that the resulting MELE still achieves the semiparametric information bound asymptotically.

Another key assumption of the proposed method is the MCAR, or missing by design. The case where the missing data are MAR is not covered here. In fact, Cattaneo [2] obtained the semiparametric efficiency bound under the MAR assumption. It will also be of interest to consider extending the proposed approach to the MAR situations in future work.

Acknowledgments

The author thanks two anonymous referees whose insightful comments have resulted in a much improved paper. The research was supported by the National Natural Science Foundation of China (10971033) and a research fund from Fudan University (EYH1019145).

Appendix

Here we prove the three theorems presented in Section 3. We introduce the following notations: $\mathbf{M}_n(\theta) = \max_{1 \leq i \leq n} \| \mathbf{m}_{n,i}(\theta) \|$, $\mathbf{M}^*(\theta) = \frac{\delta}{\rho} \mathbf{g}(X; \theta_0) - \frac{\delta - \rho}{\rho} \hat{E}[\mathbf{g}(X; \theta_0)|\tilde{X}]$, $\tilde{\mathbf{m}}_n(\theta) = \frac{\delta}{\rho} \mathbf{g}(X; \theta) - \frac{\delta - \rho}{\rho} \hat{B}_n W_n b_n(X)$, $\tilde{S}_n = E[\tilde{\mathbf{m}}_n(\theta_0) \tilde{\mathbf{m}}_n(\theta_0)^T]$ and $A_{n,m} = E(\partial \mathbf{m}_n(\theta_0)/\partial \theta^T)$. The following lemmas are necessary for the proof.

**Lemma A.1.** The probability that zero is inside the convex hull spanned by $\{\mathbf{m}_{n,i}(\theta), i = 1, 2, \ldots, n\}$ goes to one as $n \to \infty$ uniformly for $\theta$ such that $\|\theta - \theta_0\| \leq n^{-1/3}$.

**Proof.** This follows from Lemma 4.2 in [4] and discussions thereof. \qed

**Lemma A.2.** Under (C.1), (C.2), (C.3) and (C.5), the eigenvalues of $S_n(\theta_0)$ are bounded away from 0 and $\infty$ as $n \to \infty$.

**Proof.** This follows from Lemma 4.5 and proofs of condition D4 in [4]. \qed

**Lemma A.3.** Under (C.1), (C.2), (C.3) and (C.5), $\|\eta_n(\theta_0)\| = O_p(n^{-1/3})$, $\sup_{\|\theta - \theta_0\| \leq n^{-1/3}} 2 \|\eta_n(\theta)\| = O_p(n^{-2/3})$, $\|\eta_n(\theta) - S_n(\theta)^{-1}\tilde{\mathbf{m}}_n(\theta)\| = o_p(n^{-1/3})$.

**Proof.** Under (C.1), (C.2), (C.3) and (C.5), we can apply [9] to obtain

$$\|\tilde{\mathbf{m}}_n(\theta)\| = O_p(n^{-2/3} r_n^{1/2}).$$

(A.1)
Under (C.1) and (C.2),
\[
\sup_{n \geq 1} \max_{1 < i < n} \| \mathbf{g}(x_i; \theta) \| = o_p(n^{1/3}), \quad \text{and} \quad \max_{1 < i < n} \| \mathbf{b}_n(x_i) \| = O_p(r_n^{1/2}).
\]

Thus, we have
\[
\mathbf{M}_n(\theta) = o_p(n^{-1/3}) \quad \text{(A.2)}
\]
uniformly in \( \| \theta - \theta_0 \| \leq n^{-1/3} \). Write \( \eta_n(\theta) = \| \eta_n(\theta) \| \alpha \), where \( \| \alpha \| = 1 \). Then by the proof of Theorem 3.2 in [7], we can obtain
\[
\| \eta_n(\theta) \| (\alpha^T S_n(\theta) \alpha - \alpha^T \tilde{\mathbf{M}}_n(\theta) \mathbf{M}_n(\theta)) \leq \alpha^T \tilde{\mathbf{m}}_n(\theta) \quad \text{uniformly in \( \| \theta - \theta_0 \| \leq n^{-1/3} \).} \quad \text{(A.3)}
\]
By replacing \( \theta \) with \( \theta_0 \) in (A.3), we know that due to (A.1), (A.2) and Lemma A.2,
\[
\| \eta_n(\theta_0) \| = O_p(n^{-1/3}).
\]
By the similar arguments as in Lemma A.2, one can show that
\[
\sup_{n \geq 1} \| S_n(\theta) - S_n(\theta_0) \| r_n \rightarrow 0
\]
in probability and consequently there exists a \( a > 0 \) s.t. \( \alpha^T S_n(\theta) \alpha > a \) uniformly in \( \| \theta - \theta_0 \| \leq n^{-1/3} \). Moreover, by expanding \( \tilde{\mathbf{M}}_n(\theta) \) in the \( n^{-1/3} \)-neighborhood of \( \theta_0 \), we have
\[
\tilde{\mathbf{m}}_n(\theta) = \tilde{\mathbf{m}}_n(\theta_0) + O_p(\| \theta - \theta_0 \|) = O_p(n^{-1/3})
\]
uniformly in \( \| \theta - \theta_0 \| \leq n^{-1/3} \). Again by (A.3), we can conclude that
\[
\sup_{n \geq 1} \| \eta_n(\theta) \| = O_p(n^{-1/3}).
\]
Finally, from (7) we know that \( \eta_n(\theta) \) satisfies the constraint
\[
0 = \frac{1}{n} \sum_{i=1}^{n} \frac{\mathbf{m}_i(\theta)}{1 + \eta_n(\theta)^T \mathbf{m}_i(\theta)} \quad \text{(A.4)}
\]
Consequently,
\[
\eta_n(\theta) = S_n(\theta)^{-1} \tilde{\mathbf{m}}_n(\theta) + S_n(\theta)^{-1} \frac{1}{n} \sum_{i=1}^{n} \frac{\alpha^T \mathbf{m}_i(\theta) \mathbf{m}_i(\theta)^T}{1 + \eta_n(\theta)^T \mathbf{m}_i(\theta)} \| \eta_n(\theta) \|^2.
\]
By triangular inequality and some simple algebra, the final term in (A.4) can be bounded by
\[
\sup_{n \geq 1} \| \eta_n(\theta) - S_n(\theta)^{-1} \tilde{\mathbf{m}}_n(\theta) \| = o_p(n^{-1/3}).
\]

**Lemma A.4.** Under (C.1)–(C.6),
\[
A_{n,m}^T \Sigma_n^{-1} A_{n,m} \rightarrow A^T \Sigma^*^{-1} A.
\]

**Proof.** Since \( B_n \mathcal{W}_n \mathcal{b}_n(\tilde{X}) \) does not involve \( \theta \), we have
\[
E \left( \frac{\partial \tilde{\mathbf{m}}_n(\theta_0)}{\partial \theta^T} \right) = A^T \Sigma_n^{-1} A.
\]
Under (C.4), we have
\[ A^T \Sigma_n^{-1} A \rightarrow A^T \Sigma^{-1} A . \] (A.5)

For any n, let \( \mathbf{m}_n^\text{opt}(\theta) = A_{\text{opt}} \mathbf{m}_n(\theta) \), where \( A_{\text{opt}} = A_{n,m}^T \Sigma_n^{-1} \) is a \( p \times r_n \) matrix, and \( \Sigma_{n,\text{opt}} = E[\mathbf{m}_n^\text{opt}(\theta_0) \mathbf{m}_n^\text{opt}(\theta_0)^T] \). It is easy to see that
\[ A_{n,m}^T \Sigma_n^{-1} A_{n,m} = E \left( \frac{\partial \mathbf{m}_n^\text{opt}(\theta)}{\partial \theta^T} \right)^T \Sigma_{n,\text{opt}}^{-1} E \left( \frac{\partial \mathbf{m}_n^\text{opt}(\theta)}{\partial \theta^T} \right) . \]

According to Qin and Lawless [10], \( \mathbf{m}_n^\text{opt}(\theta_0) \) is optimal among all linear combinations of \( A \mathbf{m}_n(\theta_0) \). Note that
\[ E \left( \frac{\partial \mathbf{m}_n^\text{opt}(\theta_0)}{\partial \theta^T} \right)^T \Sigma_{n,\text{opt}}^{-1} E \left( \frac{\partial \mathbf{m}_n^\text{opt}(\theta_0)}{\partial \theta^T} \right) = A^T \Sigma^{-1} A \]
is positive definite due to optimality. Moreover,
\[ E \left( \frac{\partial \mathbf{m}_n^\text{opt}(\theta_0)}{\partial \theta^T} \right)^T \Sigma_{n,\text{opt}}^{-1} E \left( \frac{\partial \mathbf{m}_n^\text{opt}(\theta_0)}{\partial \theta^T} \right) - A^T \Sigma^{-1} A = E \left( \frac{\partial \mathbf{m}_n^\text{opt}(\theta_0)}{\partial \theta^T} \right)^T \Sigma_{n,\text{opt}}^{-1} E \left( \frac{\partial \mathbf{m}_n^\text{opt}(\theta_0)}{\partial \theta^T} \right) \]
\[ - A^T \Sigma^{-1} A + A^T \Sigma^{-1} A - A^T \Sigma^{-1} A . \]

As we have discussed in Section 2, Cattaneo [2] showed that the semiparametric efficiency bound is given by \((A^T \Sigma^{-1} A)^{-1}\), which implies that
\[ E \left( \frac{\partial \mathbf{m}_n^\text{opt}(\theta_0)}{\partial \theta^T} \right)^T \Sigma_{n,\text{opt}}^{-1} E \left( \frac{\partial \mathbf{m}_n^\text{opt}(\theta_0)}{\partial \theta^T} \right) = A^T \Sigma^{-1} A \]
is non-positive definite. From (A.5), we can conclude that
\[ E \left( \frac{\partial \mathbf{m}_n^\text{opt}(\theta_0)}{\partial \theta^T} \right)^T \Sigma_{n,\text{opt}}^{-1} E \left( \frac{\partial \mathbf{m}_n^\text{opt}(\theta_0)}{\partial \theta^T} \right) = A^T \Sigma^{-1} A = A_{n,m}^T \Sigma_n^{-1} A_{n,m} \]
and the conclusion follows immediately. \(\square\)

**Lemma A.5.** Under (C.1), (C.2), (C.3) and (C.5), \( \|\hat{\theta}_E - \theta_0\| < n^{-1/3} \) with probability tending to 1.

**Proof.** Consider \( \theta \) such that \( \hat{\theta}_E - \theta_0 = u_n^{-1/3} \), where \( \|u\| = 1 \). By the Taylor series expansion and Lemma A.3, we have
\[ 2l_n(\theta) = 2n \eta_n(\theta)^T \mathbf{m}_n(\theta) - n \eta_n(\theta)^T S_n(\theta) \eta_n(\theta) + O_p \left( \|\eta_n(\theta)\|^2 \sum_{i=1}^n \frac{\|\mathbf{m}_{i,n}(\theta)\|^2}{1 + \eta_n(\theta)^T \mathbf{m}_{i,n}(\theta)} \right) \]
\[ = 2n \eta_n(\theta)^T \mathbf{m}_n(\theta) - n \eta_n(\theta)^T S_n(\theta) \eta_n(\theta) + o_p(n^{1/3}) . \]

By plugging in the expression of \( \eta_n(\theta) \) given in Lemma A.3, it becomes
\[ \tilde{n} \mathbf{m}_n(\theta)^T S_n(\theta)^{-1} \mathbf{m}_n(\theta) + o_p(n^{1/3}) . \]

Expanding at \( \theta_0 \), it equals to
\[ n \left( \mathbf{m}_n(\theta_0) + \frac{\partial \mathbf{m}_n(\theta_0)}{\partial \theta} u_n^{-1/3} \right)^T S_n(\theta_0)^{-1} \left( \mathbf{m}_n(\theta_0) + \frac{\partial \mathbf{m}_n(\theta_0)}{\partial \theta} u_n^{-1/3} \right) + o_p(n^{1/3}) \]
\[ = u_n^T A_{n,m}^T \Sigma_n^{-1} A_{n,m} u_n^{1/3} + o_p(n^{1/3}) = u_n^T A_n^T \Sigma^{-1} A u_n^{1/3} + o_p(n^{1/3}) \]
\[ \geq \min \left( A_n^T \Sigma^{-1} A \right) n^{1/3} + o_p(n^{1/3}) = O_p(n^{1/3}) . \]
On the other hand,
\[ 2l_n(\theta_0) = 2n\eta_n(\theta_0)'\widehat{m}_n(\theta_0) - n\eta_n(\theta_0)'S_n(\theta_0)\eta_n(\theta_0) + o_p(n^{1/3}) = O_p(n). \]

Consequently, \( l_n(\theta_0) \) is strictly less than \( l_4(\theta) \) when \( \|\theta - \theta_0\| = n^{-1/3} \) with probability tending to 1. By definition of the MELE, we obtain that \( \|\hat{\theta}_{EL} - \theta_0\| < n^{-1/3} \) with probability tending to 1. \( \square \)

**Lemma A.6.** Under (C.1)–(C.6),
\[ A^T_{n,m} \Sigma^{-1}_{n,m} \sqrt{n}\widehat{m}_n(\theta_0) \rightarrow N \left( 0, A^T \Sigma^{-1} A \right) \]
in distribution as \( n \rightarrow \infty \).

**Proof.** It is sufficient to show that for any \( p \) constant vector \( t \),
\[ t^T A^T_{n,m} \Sigma^{-1}_{n,m} \sqrt{n}\widehat{m}_n(\theta_0) \rightarrow N \left( 0, t^T A^T \Sigma^{-1} A t \right) \]  
(A.6)
in distribution as \( n \rightarrow \infty \). First, the variance of the left hand side of (A.6)
\[ \sum_{i=1}^{n} E \left[ |n^{-1/2}t^T A^T_{n,m} \Sigma^{-1}_{n,m} \widehat{m}_n(\theta_0)|^2 \right] = t^T A^T_{n,m} \Sigma^{-1}_{n,m} A_{n,m} t, \]
which converges to \( t^T A^T \Sigma^{-1} A t \) by Lemma A.4.

Second, the Lindeberg condition needs to be verified. Note that
\[ P \left( |t^T A^T_{n,m} \Sigma^{-1}_{n,m} \widehat{m}_n(\theta_0)| > n^{1/2} \epsilon \right) \leq \frac{E \left[ |t^T A^T_{n,m} \Sigma^{-1}_{n,m} \widehat{m}_n(\theta_0)|^2 \right]}{n\epsilon^2}, \]
which goes to 0 since the numerator is asymptotically bounded by Lemma A.4. Therefore,
\[ \sum_{i=1}^{n} E \left[ |n^{-1/2}t^T A^T_{n,m} \Sigma^{-1}_{n,m} \widehat{m}_n(\theta_0)|^2 \right] \rightarrow 0. \]
Hence the conclusion follows by the Lindeberg–Feller Central Limit Theorem. \( \square \)

Now we are ready to prove the three theorems.

**Proof of Theorem 1.** Let \( \hat{\eta} = \eta_n(\hat{\theta}_{EL}) \). It can be found that \( (\hat{\theta}_{EL}, \hat{\eta}) \) satisfies \( U_{1n}(\hat{\theta}_{EL}, \hat{\eta}) = 0 \), \( U_{2n}(\hat{\theta}_{EL}, \hat{\eta}) = 0 \), where

\[ U_{1n}(\theta, \eta) = \frac{1}{n} \sum_{i=1}^{n} \frac{m_{i,n}(\theta)}{1 + \eta^T m_{i,n}(\theta)}, \quad U_{2n}(\theta, \eta) = \frac{1}{n} \sum_{i=1}^{n} \eta \left( \partial m_{i,n}(\theta) / \partial \theta \right)^T.\]

By taking Taylor series expansion, we have
\[ 0 = U_{1n}(\hat{\theta}_{EL}, \hat{\eta}) \]
\[ = U_{1n}(\theta_0, 0) + \frac{\partial U_{1n}(\theta_0, 0)}{\partial \theta} (\hat{\theta}_{EL} - \theta_0) + \frac{\partial U_{1n}(\theta_0, 0)}{\partial \eta} \hat{\eta} + o_p(\|\hat{\theta}_{EL} - \theta_0\| + \|\hat{\eta}\|) \]
\[ 0 = U_{1n}(\hat{\theta}_{EL}, \hat{\eta}) \]
\[ = U_{2n}(\theta_0, 0) + \frac{\partial U_{2n}(\theta_0, 0)}{\partial \theta} (\hat{\theta}_{EL} - \theta_0) + \frac{\partial U_{2n}(\theta_0, 0)}{\partial \eta} \hat{\eta} + o_p(\|\hat{\theta}_{EL} - \theta_0\| + \|\hat{\eta}\|). \]

Solving the above two equations gives out
\[ \sqrt{n} (\hat{\theta}_{EL} - \theta_0) = -\sqrt{n} \left[ \left( \frac{\partial \widehat{m}_n(\theta_0)}{\partial \theta} \right)^T S_n(\theta_0)^{-1} \left( \frac{\partial \widehat{m}_n(\theta_0)}{\partial \theta} \right) \right]^{-1} \left( \frac{\partial \widehat{m}_n(\theta_0)}{\partial \eta} \right)^T S_n(\theta_0)^{-1} \widehat{m}_n(\theta_0) + o_p(1). \]  
(A.7)
By triangular inequality, we have
\[ \left| A^T \Sigma^{-1} A - \left( \frac{\partial \overline{m}(\theta_0)}{\partial \theta^T} \right)^T S_n(\theta_0)^{-1} \left( \frac{\partial \overline{m}(\theta_0)}{\partial \theta^T} \right) \right| \]
\[ \leq \left| A^T \Sigma^{-1} A - A_{n,m}^T \Sigma^{-1} A_{n,m} \right| + \left| A_{n,m}^T \Sigma^{-1} A_{n,m} - \left( \frac{\partial \overline{m}(\theta_0)}{\partial \theta^T} \right)^T S_n(\theta_0)^{-1} \left( \frac{\partial \overline{m}(\theta_0)}{\partial \theta^T} \right) \right|. \]
The first term on the right hand side converges to 0 as \( n \to \infty \) by Lemma A.4 and the second term is of order \( o_p(1) \). Therefore
\[ \left| A^T \Sigma^{-1} A - \left( \frac{\partial \overline{m}(\theta_0)}{\partial \theta^T} \right)^T S_n(\theta_0)^{-1} \left( \frac{\partial \overline{m}(\theta_0)}{\partial \theta^T} \right) \right| \to 0. \]
Finally, by Lemma A.6, we have
\[ \left( \frac{\partial \overline{m}(\theta_0)}{\partial \theta^T} \right)^T S_n(\theta_0)^{-1} \sqrt{n} \overline{m}(\theta_0) = A_{n,m}^T \Sigma^{-1} \sqrt{n} \overline{m}(\theta_0) + o_p(1) \]
\[ \to N \left( 0, \left( A^T \Sigma^{-1} A \right) \right). \]
By applying the Slutsky Theorem, the conclusion of Theorem 1 follows immediately from (A.7).

**Proof of Theorem 2.** Note that there are only finitely many terms in \( \overline{m} \) and \( S_n \) that contain \( \theta \), by the delta-method and Lemma A.5, we obtain that
\[ \left| \left[ \frac{\partial \overline{m}(\theta_0)}{\partial \theta^T} \right] S_n(\theta_0)^{-1} \left( \frac{\partial \overline{m}(\theta_0)}{\partial \theta^T} \right) \right| \to o_p(1). \]
Combining this with the fact that
\[ \left| A^T \Sigma^{-1} A - \left( \frac{\partial \overline{m}(\theta_0)}{\partial \theta^T} \right)^T S_n(\theta_0)^{-1} \left( \frac{\partial \overline{m}(\theta_0)}{\partial \theta^T} \right) \right| \to o_p(1) \]
we get the conclusion of the theorem.

**Proof of Theorem 3.** By Taylor series expansion at \( \theta_0 \), we obtain that
\[ 2 l_n(\hat{\theta}_L) = 2 l_n(\theta_0) + \sqrt{n} \left( \hat{\theta}_L - \theta_0 \right)^T \left( \frac{1}{n} \frac{\partial^2 l_n(\theta_0)}{\partial \theta \partial \theta^T} \right) \sqrt{n} \left( \hat{\theta}_L - \theta_0 \right) + o_p(1). \]
Consequently, we have
\[ T_{1n}(\theta_0) = 2 l_n(\theta_0) - 2 l_n(\hat{\theta}_L) = \sqrt{n} \left( \hat{\theta}_L - \theta_0 \right)^T \left( A^T \Sigma^{-1} A \right) \sqrt{n} \left( \hat{\theta}_L - \theta_0 \right) + o_p(1). \]
By Theorem 1, the conclusion follows immediately.

**References**