Received: 19 September 2017 | Revised: 8 May 2018 | Accepted: 24 July 2018

DOI: 10.1111/sjos.12354

ORIGINAL ARTICLE

Scandinavian Journal of Statistics

On the asymptotic non-equivalence of efficient-GMM and MEL estimators in models with missing data

Xuerong Chen[1] | Yan Chen[2] | Alan T.K. Wan[3] | Yong Zhou[4]

[1]Center of Statistical Research, Southwestern University of Finance and Economics, Chengdu, China

[2]School of Statistics and Management, Shanghai University of Finance and Economics, Shanghai, China

[3]Department of Management Sciences, City University of Hong Kong, Kowloon, Hong Kong

[4]Faculty of Economics and Management, East China Normal University, Shanghai, China

Correspondence
Xuerong Chen, Center of Statistical Research, Southwestern University of Finance and Economics, 555 Liutai Avenue, Wenjiang District, 611130 Chengdu, Sichuan, China.
Email: chenxr522@foxmail.com

Funding information
National Natural Science Foundation of China (NSFC), Grant/Award Number: 11501461 and 71571113; Fundamental Research Funds for the Central Universities, Grant/Award Number: JBK140507, JBK1806002 and JBK1802070; State Key Program of the National Natural Science Foundation of China, Grant/Award Number: 71331006 and 91546202; Innovative Research Team in SUFE, Grant/Award Number: IRTSHUFE13122402

Abstract

The generalized method of moments (GMM) and empirical likelihood (EL) are popular methods for combining sample and auxiliary information. These methods are used in very diverse fields of research, where competing theories often suggest variables satisfying different moment conditions. Results in the literature have shown that the efficient-GMM (GMM_E) and maximum empirical likelihood (MEL) estimators have the same asymptotic distribution to order $n^{-1/2}$ and that both estimators are asymptotically semiparametric efficient. In this paper, we demonstrate that when data are missing at random from the sample, the utilization of some well-known missing-data handling approaches proposed in the literature can yield GMM_E and MEL estimators with nonidentical properties; in particular, it is shown that the GMM_E estimator is semiparametric efficient under all the missing-data handling approaches considered but that the MEL estimator is not always efficient. A thorough examination of the reason for the nonequivalence of the two estimators is presented. A particularly strong feature of our analysis is that we do not assume smoothness in the underlying moment conditions. Our results are thus relevant to situations involving non-smooth estimating functions, including quantile and rank regressions, robust estimation, the estimation of receiver operating characteristic (ROC) curves, and so on.

KEYWORDS

empirical likelihood, generalized method of moments, kernel, missing at random, non-smooth, semiparametric efficiency bound

1 INTRODUCTION

A general methodology that has found wide popularity recently, especially in econometrics and biostatistics, is to estimate parameters via estimating equations (EEs). Consider a set of l estimating functions (EFs) $\mathbf{g}(y, z, \boldsymbol{\theta}) = (g_1(y, z, \boldsymbol{\theta}), g_2(y, z, \boldsymbol{\theta}), \ldots, g_l(y, z, \boldsymbol{\theta}))^T$ that satisfy the unbiasedness condition

$$E\mathbf{g}(Y, \mathbf{Z}, \boldsymbol{\theta}_0) = 0, \tag{1}$$

where Y is an independent and identically distributed (i.i.d.) response variable with unknown distribution, $\mathbf{Z}$ is a covariate vector, and $\boldsymbol{\theta}$ is a q-dimensional ($q \le l$) parameter vector with true value $\boldsymbol{\theta}_0$. When the EEs are exactly identified, that is, $l = q$, $\boldsymbol{\theta}$ can be estimated by the standard method of moments (MoM). When there are more moment conditions than parameters, that is, $l > q$, the most widely applied methods of estimating $\boldsymbol{\theta}$ are the generalized method of moments (GMM) (Hansen, 1982) and empirical likelihood (EL) (Owen, 1988, 1990, 1991).

The flexibility offered by GMM and EL in combining sample and auxiliary information has made them popular in very diverse fields of research, especially in economics and finance. Whereas the maximum EL (MEL) estimator attempts to find the parameter values that maximize the EL function, the efficient-GMM (GMM_E) estimator is obtained by choosing the weight matrix that minimizes the asymptotic covariance of the estimator (Hansen, 1982). The GMM_E and MEL estimators are equivalent in many ways. For the exactly identified case where $l = q$, the GMM_E and MEL approaches yield the same solution as the MoM. As well, the results of Qin and Lawless (1994) and Imbens (1997, 2002) demonstrated that the empirical log-likelihood ratio (ELLR) is asymptotically Chi-square distributed, and if the moment conditions are correct, the GMM_E and MEL estimators have the same asymptotic distribution. Given the Chamberlain (1987) proof that the GMM_E estimator attains the asymptotic semiparametric efficiency bound, the latter result means that the MEL estimator is also asymptotically efficient. Until recently, the view has always been that GMM_E and MEL produce estimators with the equivalent asymptotic normality property. However, this long-held view has been altered by some recent results in the missing-data literature. There is evidence that when data are only partially observed for some variables, depending on the missing-data handling mechanism, the resultant GMM_E and MEL estimators' asymptotic properties are not always in accord.

Consider the situation where observations on the response or covariates may be missing at random (MAR), which means that the probability of missingness is only related to the fully observed variables and not the partially unobserved variables. Let the vector $(\mathbf{X}_i^T, \mathbf{X}_i^{cT})^T$ contain shuffled elements of $(Y_i, \mathbf{Z}_i^T)^T$ such that $\mathbf{X}_i$, a d-dimensional non-null vector, is observed for all i's, whereas $\mathbf{X}_i^c$ contains elements for which observations may or may not be available for some i's, and $\delta_i = 1$ if all values in $\mathbf{X}_i^c$ are observed, and $\delta_i = 0$ otherwise. Under the MAR assumption, the propensity score function is $P(\mathbf{X}_i)=Pr(\delta_i = 1 | Y_i, \mathbf{Z}_i)=Pr(\delta_i = 1 | \mathbf{X}_i)$. Zhou, Wan, and Wang (2008) studied the GMM and EL approaches for EFs with missing data based on the following EE projection (EEP) modified EF:

$$\tilde{\mathbf{g}}_{\mathcal{M}_1}(Y_i, \mathbf{Z}_i, \boldsymbol{\theta}) = \delta_i \mathbf{g}(Y_i, \mathbf{Z}_i, \boldsymbol{\theta}) + (1 - \delta_i)\mathbf{m}(\mathbf{X}_i, \boldsymbol{\theta}), \tag{2}$$

where $\mathbf{m}(\mathbf{X}, \boldsymbol{\theta}) = E[\mathbf{g}(Y, \mathbf{Z}, \boldsymbol{\theta}) \mid \mathbf{X}]$ is unknown and can be estimated by a kernel smoothing method. Zhou et al. (2008) showed that the GMM_E and MEL estimators do not produce the same asymptotic covariance of the estimator. More specifically, they showed that the GMM_E estimator is semiparametric efficient by achieving the semiparametric efficiency bound established by Chen, Hong, and Tarozzi (2008), whereas the MEL estimator does not possess the same property. As well, they demonstrated that the ELLR statistic is not asymptotically Chi-square distributed but converges instead in distribution to a weighted sum of Chi-square random variables.

Tang and Qin (2012) developed an EL approach based on the following augmented inverse probability weighted (AIPW) EF:

$$\tilde{\mathbf{g}}_{\mathcal{M}_2}(Y_i, \mathbf{Z}_i, \boldsymbol{\theta}) = \frac{\delta_i}{P(\mathbf{X}_i)}\mathbf{g}(Y_i, \mathbf{Z}_i, \boldsymbol{\theta}) + \left\{1 - \frac{\delta_i}{P(\mathbf{X}_i)}\right\}\mathbf{m}(\mathbf{X}_i, \boldsymbol{\theta}). \tag{3}$$

They found that the ELLR based on (3) has an asymptotic Chi-square distribution. More remarkably, their results indicated that the AIPW-based MEL estimator yields the same asymptotic covariance matrix as the GMM$_E$ estimator under (2), and this covariance matrix attains the semiparametric efficiency bound. Recently, Chen, Wan, and Zhou (2015) showed that when the data are MAR and the EEs are exactly identified, the MoM estimators (which coincide with GMM$_E$) obtained based on the EFs (2) and (3) and the following IPW EF

$$\tilde{\mathbf{g}}_{\mathcal{M}_3}(Y_i, \mathbf{Z}_i, \boldsymbol{\theta}) = \frac{\delta_i}{P(\mathbf{X}_i)}\mathbf{g}(Y_i, \mathbf{Z}_i, \boldsymbol{\theta}) \tag{4}$$

all have the same asymptotic properties and achieve the semiparametric efficiency bound.

Table 1 summarizes the established properties of the GMM$_E$ and MEL estimators as described above. The established results have raised some interesting, but also puzzling, questions about the properties of GMM and MEL estimators. First, it remains to be answered why the GMM$_E$ and MEL estimators under the EEP approach have different asymptotic variance. Second, the properties of the GMM estimator under the AIPW approach are yet to be explored; of particular interest is whether under the AIPW approach, the GMM$_E$ estimator can produce a semiparametric efficient estimator. Third, it is of interest to ascertain why the EEP and AIPW approaches produce ELLR statistics with nonidentical asymptotic distributions. Fourth, while the IPW, EEP, and AIPW methods produce estimators with the same asymptotic normality property when $l = q$, it is unclear if this property will continue to hold when $l \leq q$. The purpose of this paper is to take steps in addressing these questions.

Another objective of this paper is to examine GMM and EL inference for nonsmooth EEs under missing data. It should be noted that the overwhelming majority of the above studies assumes that the underlying EEs are smooth. This stringent requirement rules out the application

TABLE 1 Summary of existing results

Reference	Data type	EE	Main findings
Qin & Lawless (1994); Lawless (1997); Imbens (2002)	Fully observed	$\mathbf{g}$	The GMM$_E$ and MEL estimators under the same EEs have the same asymptotic distribution and are semiparametric efficient; the ELLR statistic is asymptotically Chi-square distributed.
Zhou et al. (2008)	MAR	$\tilde{\mathbf{g}}_{\mathcal{M}_1}$	The MEL estimator does not yield the same semiparametric efficiency as the GMM$_E$ estimator; the ELLR statistic converges in distribution to a weighted sum of Chi-square random variables.
Tang and Qin (2012)	MAR	$\tilde{\mathbf{g}}_{\mathcal{M}_2}$	MEL is semiparametric efficient; the ELLR statistic has an asymptotic Chi-square distribution; the MEL estimator under $\tilde{\mathbf{g}}_{\mathcal{M}_2}$ has the same asymptotic distribution as the GMM$_E$ estimator under $\tilde{\mathbf{g}}_{\mathcal{M}_1}$.
Chen et al. (2015)	MAR	$\tilde{\mathbf{g}}_{\mathcal{M}_1}$, $\tilde{\mathbf{g}}_{\mathcal{M}_2}$, $\tilde{\mathbf{g}}_{\mathcal{M}_3}$	When the EEs are exactly identified (i.e., $l = q$), estimators based on all three types of EEs result in the same asymptotic distribution.

Note. EE = estimating equation; ELLR = empirical log-likelihood ratio; GMM$_E$ = efficient generalized method of moments; MAR = missing at random; MEL = maximum empirical likelihood.

of the established methods to situations including quantile regression, rank regression, robust estimation, and the estimation of receiver operating characteristic (ROC) curves, distribution function, and differences of quantiles, where some or all of the underlying EEs are discontinuous. Chen et al. (2008) developed semiparametric efficient sieve-based GMM estimators for missing data when the underlying EEs are nonsmooth, but they did not examine EL-based inference. The extension of the IPW, EEP, and AIPW approaches of handling missing data to nonsmooth EEs is by no means straightforward because the Taylor series expansion for developing theoretical results under smooth EEs is inapplicable when the smoothness assumption is unfulfilled. Although Lopez, Van Keilegom, and Veraverbeke (2009) proved the asymptotic normality of the MEL estimator under nonsmooth EEs when no data are missing, they assumed that the MEL estimator is consistent without proving it. In this paper, we prove the consistency of the MEL estimator directly. We consider the latter a noteworthy aspect of our results.

The remainder of this paper is organized as follows. In Section 2, we describe the EEP-, AIPW-, and IPW-based missing-data handling approaches and outline the estimation methods by GMM and EL in conjunction with these three approaches. Section 3 contains an analysis of the asymptotic properties of the resultant GMM and MEL estimators when at least a subset of the underlying EFs is nonsmooth. Section 4 presents a thorough examination of the asymptotic nonequivalence of the GMM$_E$ and MEL estimators under the current setup and a comparison of the asymptotic efficiency of the proposed GMM$_E$ and MEL estimators with their parametric counterparts. Section 5 reports simulation findings on the finite-sample properties of the proposed estimators. Section 6 considers a real data application. Section 7 concludes. Proofs of results are contained in the Appendix.

2 | EE IMPUTATION AND ESTIMATION METHODS

2.1 | EE imputation

Throughout our analysis, we assume that the data are MAR, which is justified in many practical situations (Little & Rubin, 2002, chapter 1). There is a large amount of literature that adopts MAR as a baseline for analysis. The application of the EEs (2), (3), and (4) involves the imputation of $\mathbf{m}(\mathbf{x}, \boldsymbol{\theta}) = E\{\mathbf{g}(Y, \mathbf{Z}, \boldsymbol{\theta}) \mid \mathbf{X} = \mathbf{x}\}$ and $P(\mathbf{x})$. The idea of imputing the conditional expectation of EF was first explored by Zhou and Pepe (1995) and Paik (1997). Here, we impute $\mathbf{m}(\mathbf{x}, \boldsymbol{\theta})$ and $P(\mathbf{x})$ by kernel regression based on the observed data from the random sample $(Y_i, \mathbf{Z}_i, \delta_i)$, $i = 1, 2 \ldots, n$. Let the kernel regression estimators of $\mathbf{m}(\mathbf{x}, \boldsymbol{\theta})$ and $P(\mathbf{x})$ be

$$\hat{\mathbf{m}}(\mathbf{x}, \boldsymbol{\theta}) = \frac{\sum_{i=1}^{n} \boldsymbol{\mathcal{K}}_h(\mathbf{x} - \mathbf{X}_i)\mathbf{g}(Y_i, \mathbf{Z}_i, \boldsymbol{\theta})\delta_i}{\sum_{i=1}^{n} \boldsymbol{\mathcal{K}}_h(\mathbf{x} - \mathbf{X}_i)\delta_i} \quad \text{and} \quad \hat{P}(\mathbf{x}) = \frac{\sum_{i=1}^{n} \bar{K}_{\bar{h}}(\mathbf{x} - \mathbf{X}_i)\delta_i}{\sum_{i=1}^{n} \bar{K}_{\bar{h}}(\mathbf{x} - \mathbf{X}_i)},$$

respectively, where $\boldsymbol{\mathcal{K}}_h(u) = \text{diag}[K^{(1)}(./h_1)/h_1^d, \ldots, K^{(l)}(./h_l)/h_l^d]$, $K^{(i)}$ and $\bar{K}$ are d-variate kernel functions, and $h_i, i = 1, 2, \ldots, l$ and $\bar{h}$ are bandwidth parameters. Now, substituting $\hat{\mathbf{m}}(\mathbf{x}, \boldsymbol{\theta})$ and $\hat{P}(\mathbf{x})$ for $\mathbf{m}(\mathbf{x}, \boldsymbol{\theta})$ and $P(\mathbf{x})$ in (2) and (3), we obtain

$$\hat{\mathbf{g}}_{\mathcal{M}_1}(Y_i, \mathbf{Z}_i, \boldsymbol{\theta}) = \delta_i \mathbf{g}(Y_i, \mathbf{Z}_i, \boldsymbol{\theta}) + (1 - \delta_i)\hat{\mathbf{m}}(\mathbf{X}_i, \boldsymbol{\theta}), \tag{5}$$

$$\hat{\mathbf{g}}_{\mathcal{M}_2}(Y_i, \mathbf{Z}_i, \boldsymbol{\theta}) = \frac{\delta_i}{\hat{P}(\mathbf{X}_i)}\mathbf{g}(Y_i, \mathbf{Z}_i, \boldsymbol{\theta}) + \left\{1 - \frac{\delta_i}{\hat{P}(\mathbf{X}_i)}\right\}\hat{\mathbf{m}}(\mathbf{X}_i, \boldsymbol{\theta}), \tag{6}$$

and

$$\hat{\mathbf{g}}_{\mathcal{M}_3}(Y_i, \mathbf{Z}_i, \boldsymbol{\theta}) = \frac{\delta_i}{\hat{P}(\mathbf{X}_i)}\mathbf{g}(Y_i, \mathbf{Z}_i, \boldsymbol{\theta}) \tag{7}$$

as the imputed versions of (2), (3), and (4), that is, the EEP-, AIPW-, and IPW-based EFs, respectively.

It is readily seen that $\frac{1}{n}\sum_{i=1}^{n}\hat{\mathbf{g}}_{\mathcal{M}_k}(Y_i, \mathbf{Z}_i, \boldsymbol{\theta}) = 0, k = 1, 2, 3$, are asymptotically unbiased EEs of $\boldsymbol{\theta}$. These imputed EEs, which are not necessarily smooth in $\boldsymbol{\theta}$, will form the basis of our subsequent development of estimation and inference methods using the GMM and EL approaches.

2.2 | Estimation methods

This subsection outlines the estimation of unknowns by GMM and EL based on the imputed EEs of (5), (6), and (7) when l, the number of EEs, is greater than q, the number of unknown parameters.

Now, let

$$\hat{Q}_{kn}(\boldsymbol{\theta}) = \left\{\frac{1}{n}\sum_{i=1}^{n}\hat{\mathbf{g}}_{\mathcal{M}_k}(Y_i, \mathbf{Z}_i, \boldsymbol{\theta})\right\}^T \mathbf{W}_{kn}\left\{\frac{1}{n}\sum_{i=1}^{n}\hat{\mathbf{g}}_{\mathcal{M}_k}(Y_i, \mathbf{Z}_i, \boldsymbol{\theta})\right\},$$

where $\mathbf{W}_{kn}$ is some positive semidefinite symmetric weight matrix, and $\hat{\mathbf{g}}_{\mathcal{M}_k}(Y, \mathbf{Z}, \boldsymbol{\theta})$, $k = 1, 2, 3$, are the EFs given in (5), (6), and (7). The value of $\boldsymbol{\theta}$ that minimizes $\hat{Q}_{kn}(\boldsymbol{\theta})$, $\hat{\boldsymbol{\theta}}_k = \arg\min_{\boldsymbol{\theta}\in\Theta}\hat{Q}_{kn}(\boldsymbol{\theta})$, is the GMM estimator (Hansen, 1982) of $\boldsymbol{\theta}$, $k = 1, 2, 3$. EL estimation (Owen, 2001; Qin & Lawless, 1994), on the other hand, is based on the objective function

$$L_k(\boldsymbol{\theta}) = \max\left\{\prod_{i=1}^{n}p_{ki} \;\middle|\; \sum_{i=1}^{n}p_{ki}\hat{\mathbf{g}}_{\mathcal{M}_k}(Y_i, \mathbf{Z}_i, \boldsymbol{\theta}) = 0, p_{ki} \geq 0, \sum_{i=1}^{n}p_{ki} = 1\right\},$$

where p_{ki} is a nonnegative probability corresponding to observation i in the sample, and $\sum_{i=1}^{n}p_{ki} = 1$, $k = 1, 2, 3$. It is straightforward to show that

$$\hat{p}_{ki} = \frac{1}{n\left(1 + \boldsymbol{\lambda}_k^T\hat{\mathbf{g}}_{\mathcal{M}_k}(Y_i, \mathbf{Z}_i, \boldsymbol{\theta})\right)}$$

maximizes $L_k(\boldsymbol{\theta})$, where $\boldsymbol{\lambda}_k$ is a vector of Lagrange multipliers satisfying

$$\sum_{i=1}^{n}\frac{\hat{\mathbf{g}}_{\mathcal{M}_k}(Y_i, \mathbf{Z}_i, \boldsymbol{\theta})}{n\left(1 + \boldsymbol{\lambda}_k^T\hat{\mathbf{g}}_{\mathcal{M}_k}(Y_i, \mathbf{Z}_i, \boldsymbol{\theta})\right)} = 0. \tag{8}$$

This leads to the ELLR

$$\mathcal{R}_k(\boldsymbol{\theta}) = -2\log\prod_{i=1}^{n}(n\hat{p}_{ki}) = 2\sum_{i=1}^{n}\log\left(1 + \boldsymbol{\lambda}_k^T\hat{\mathbf{g}}_{\mathcal{M}_k}(Y_i, \mathbf{Z}_i, \boldsymbol{\theta})\right) \tag{9}$$

of $\boldsymbol{\theta}$. The MEL estimator of $\boldsymbol{\theta}$ is $\hat{\boldsymbol{\theta}}_{ke} = \arg\min_{\boldsymbol{\theta}}\mathcal{R}_k(\boldsymbol{\theta})$, $k = 1, 2, 3$.

3 | ASYMPTOTIC PROPERTIES OF ESTIMATORS

A major technical challenge of the current work is that the lack of smoothness in the EEs renders the Taylor series expansion, the main technical tool underlying Zhou et al. (2008) analysis, inapplicable in proving the consistency and asymptotic normality of estimators. In the present analysis, we make use of results of the empirical process of the Donsker class and stochastic equicontinuity stated under assumptions (C_5^k) and (C_6^k) in the Appendix. Assumption (C_5^k) permits the application of the uniform law of large numbers on the objective functions of the GMM and MEL estimators, which is crucial for establishing estimator consistency. When proving the asymptotic normality of estimators, the assumption of stochastic equicontinuity, as stated in (C_6^k), allows the application of the Taylor series expansion on $Eg(.)$ instead of $g(.)$.

Now, for notational convenience, let us write

$$\mathbf{m}(\mathbf{X}) = \mathbf{m}(\mathbf{X}, \boldsymbol{\theta}_0) = E\{\mathbf{g}(Y, \mathbf{Z}, \boldsymbol{\theta}_0) \mid \mathbf{X}\}, \quad \boldsymbol{\Sigma}_g(\mathbf{X}) = \boldsymbol{\Sigma}_g(X, \boldsymbol{\theta}_0) = \text{cov}\{\mathbf{g}(Y, \mathbf{Z}, \boldsymbol{\theta}_0) \mid \mathbf{X}\},$$
$$\mathbf{V}_1 = E[P(\mathbf{X})\boldsymbol{\Sigma}_g(\mathbf{X})] + E\{\mathbf{m}(\mathbf{X})\mathbf{m}^T(\mathbf{X})\}, \quad \boldsymbol{\Gamma} = E\{\nabla_{\boldsymbol{\theta}}\mathbf{m}(\mathbf{X})\} = \nabla_{\boldsymbol{\theta}}E\{\mathbf{g}(Y, \mathbf{Z}, \boldsymbol{\theta}_0)\},$$
$$\boldsymbol{\Sigma} = \mathbf{V}_2 = E\left\{\frac{\boldsymbol{\Sigma}_g(\mathbf{X})}{P(\mathbf{X})}\right\} + E\{\mathbf{m}(\mathbf{X})\mathbf{m}^T(\mathbf{X})\}, \mathbf{V}_3 = E\left\{\frac{\boldsymbol{\Sigma}_g(\mathbf{X})}{P(\mathbf{X})}\right\} + E\left\{\frac{\mathbf{m}(\mathbf{X})\mathbf{m}^T(\mathbf{X})}{P(\mathbf{X})}\right\},$$

and $\mathcal{W}_{kn} = \mathcal{W}_{kn}(\boldsymbol{\theta}_0) = n^{-1/2}\sum_{i=1}^n \hat{\mathbf{g}}_{\mathcal{M}_k}(Y_i, \mathbf{Z}_i, \boldsymbol{\theta}_0)$, where $\mathbf{V}_1, \mathbf{V}_2$, and $\mathbf{V}_3$ are the asymptotic second moments of EFs (5), (6), and (7), respectively, and $\boldsymbol{\Sigma}$, which is also equal to $\mathbf{V}_2$, is the asymptotic covariance of (5), (6), and (7). See Lemmas 2 and 3 in the Appendix.

Theorems 1 and 2 summarize the asymptotic properties of the GMM estimators $\hat{\boldsymbol{\theta}}_k, k = 1, 2, 3$, based on the imputed EEs (5), (6), and (7).

Theorem 1. *Assume that conditions (C_1^k), (C_2^k), (C_4^k), and (C_5^k) in the Appendix are satisfied, and there exists a unique $\boldsymbol{\theta}_0$ such that $E\mathbf{g}(Y,\mathbf{Z},\boldsymbol{\theta}_0)=0$. Then, $\hat{\boldsymbol{\theta}}_k \xrightarrow{P} \boldsymbol{\theta}_0$, $k = 1, 2, 3$.*

Theorem 2. *Assume that conditions $(C_1^k) - (C_2^k), (C_3)$ and $(C_4^k) - (C_6^k)$ in the Appendix are satisfied, and there exists a unique $\boldsymbol{\theta}_0$ such that $E\mathbf{g}(Y,\mathbf{Z},\boldsymbol{\theta}_0) = 0$, and $\boldsymbol{\Gamma}^T\mathbf{W}_{kn}\boldsymbol{\Gamma}$ is nonsingular. Then, we have*

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_k - \boldsymbol{\theta}_0) \xrightarrow{D} N\left(0, \left(\boldsymbol{\Gamma}^T\mathbf{W}_{kn}\boldsymbol{\Gamma}\right)^{-1}\boldsymbol{\Gamma}^T\mathbf{W}_{kn}\boldsymbol{\Sigma}\mathbf{W}_{kn}\boldsymbol{\Gamma}\left(\boldsymbol{\Gamma}^T\mathbf{W}_{kn}\boldsymbol{\Gamma}\right)^{-1}\right), k = 1, 2, 3.$$

Thus, the three GMM estimators $\hat{\boldsymbol{\theta}}_1$, $\hat{\boldsymbol{\theta}}_2$, and $\hat{\boldsymbol{\theta}}_3$, based on the EEP, AIPW, and IPW imputed EFs given in (5), (6), and (7), respectively, are $\sqrt{n}$-consistent. In fact, Theorem 2 shows that they are asymptotically equivalent. Note that these estimators are determined by the asymptotic distribution of $\mathcal{W}_{kn}$ and the derivative of $E\tilde{\mathbf{g}}_{\mathcal{M}_k}(Y, \mathbf{Z}, \boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$. From Lemma 2 in the Appendix, $\mathcal{W}_{kn}$'s have the same asymptotic distribution for $k = 1, 2, 3$. As well, under the MAR assumption, $E\tilde{\mathbf{g}}_{\mathcal{M}_k}(Y, \mathbf{Z}, \boldsymbol{\theta}) = E\mathbf{g}(Y, \mathbf{Z}, \boldsymbol{\theta}) = E\mathbf{m}(\mathbf{X}, \boldsymbol{\theta})$, for $k = 1, 2, 3$. Hence, $E\tilde{\mathbf{g}}_{\mathcal{M}_k}(Y, \mathbf{Z}, \boldsymbol{\theta})$'s have the same derivative with respect to $\boldsymbol{\theta}$ for $k = 1, 2, 3$.

From the standard GMM theory (Hansen, 1982), the most efficient GMM estimator results when $\mathbf{W}_{kn} = \boldsymbol{\Sigma}^{-1}$. We denote the GMM$_E$ estimator as $\hat{\boldsymbol{\theta}}_{kg}$, $k = 1, 2, 3$. Chen et al. (2008) showed that the semiparametric efficiency bound of estimators defined by the moment restriction given in model (1) when the data are MAR is

$$\boldsymbol{\Sigma}_0^{-1} := \left(\boldsymbol{\Gamma}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Gamma}\right)^{-1}.$$

In general, GMM efficiency does not necessarily imply semiparametric efficiency. A GMM-efficient estimator is semiparametric efficient only if its covariance matrix attains the above semiparametric efficiency bound. The following corollary shows that $\hat{\boldsymbol{\theta}}_{kg}$, $k = 1, 2, 3$, are semiparametric efficient in addition to being GMM efficient.

Corollary 1. *Under the assumptions of Theorem 2, if $\hat{\boldsymbol{\theta}}_k = \hat{\boldsymbol{\theta}}_{kg}$, the GMM$_E$ estimator that sets $\boldsymbol{W}_{kn}$ to $\boldsymbol{\Sigma}^{-1}$, then*

$$\sqrt{n}\left(\hat{\boldsymbol{\theta}}_{kg} - \boldsymbol{\theta}_0\right) \xrightarrow{D} N\left(0, \boldsymbol{\Sigma}_0^{-1}\right), \quad k = 1, 2, 3.$$

The following theorems provide results on the asymptotic properties of the MEL estimators $\hat{\boldsymbol{\theta}}_{ke}, k = 1, 2, 3$, based on the EEP, AIPW, and IPW imputed EFs, respectively.

Theorem 3. *Assume that conditions $(C_1^k) - (C_2^k), (C_3)$ and $(C_4^k) - (C_7^k)$ in the Appendix are satisfied, and there exists a unique $\boldsymbol{\theta}_0$ such that $E\mathbf{g}(Y,\mathbf{Z},\boldsymbol{\theta}_0) = 0$. Then, $\hat{\boldsymbol{\theta}}_{ke} \xrightarrow{p} \boldsymbol{\theta}_0$, $k = 1, 2, 3$.*

Theorem 4. *Assume that conditions* $(C_1^k) - (C_2^k), (C_3)$ *and* $(C_4^k) - (C_7^k)$ *in the Appendix are satisfied, and there exists a unique* $\boldsymbol{\theta}_0$ *such that* $E\mathbf{g}(Y, \mathbf{Z}, \boldsymbol{\theta}_0) = 0$*. Then,*

$$\sqrt{n}\left(\hat{\boldsymbol{\theta}}_{ke} - \boldsymbol{\theta}_0\right) \xrightarrow{D} N(0, \boldsymbol{\Sigma}_{ke}), \quad k = 1, 2, 3,$$

where $\boldsymbol{\Sigma}_{ke} = (\boldsymbol{\Gamma}^T\mathbf{V}_k^{-1}\boldsymbol{\Gamma})^{-1}\boldsymbol{\Gamma}^T\mathbf{V}_k^{-1}\boldsymbol{\Sigma}\mathbf{V}_k^{-1}\boldsymbol{\Gamma}(\boldsymbol{\Gamma}^T\mathbf{V}_k^{-1}\boldsymbol{\Gamma})^{-1}$*. Furthermore, the ELLRs, denoted by* $\mathcal{R}_k(\boldsymbol{\theta}_0)'s, k = 1, 2, 3$*, have the following asymptotic distributional properties:*

$$\mathcal{R}_1(\boldsymbol{\theta}_0) \xrightarrow{D} \sigma_1\omega_1^2 + \cdots + \sigma_l\omega_l^2, \quad \mathcal{R}_2(\boldsymbol{\theta}_0) \xrightarrow{D} \chi_l^2 \quad \text{and} \quad \mathcal{R}_3(\boldsymbol{\theta}_0) \xrightarrow{D} \varrho_1\omega_1^2 + \cdots + \varrho_l\omega_l^2,$$

where ω_i^2*'s,* $i = 1, \ldots, l$*, are Chi-square random variables, each with one degree of freedom, and distributed independently of one another, and the weights* σ_i*'s and* ϱ_i*'s,* $i = 1, \ldots, l$*, are eigenvalues of* $\mathbf{V}_1^{-1}\boldsymbol{\Sigma}$ *and* $\mathbf{V}_3^{-1}\boldsymbol{\Sigma}$*, respectively.*

Thus, the MEL estimators produced by all three missing-data handling approaches are $\sqrt{n}$-consistent. It is also seen that $\mathbf{V}_2 = \boldsymbol{\Sigma}$; hence, $\boldsymbol{\Sigma}_{2e} = \boldsymbol{\Sigma}_0^{-1}$. In other words, $\hat{\boldsymbol{\theta}}_{2e}$, the MEL estimator based on the AIPW imputed EF (6), attains the semiparametric efficiency bound of Chen et al. (2008). On the other hand, neither $\mathbf{V}_1$ nor $\mathbf{V}_3$ is equal to $\boldsymbol{\Sigma}$ unless $P(\mathbf{x}) = 1$ (i.e., no observation is missing); hence, $\hat{\boldsymbol{\theta}}_{1e}$ and $\hat{\boldsymbol{\theta}}_{3e}$, the MEL estimators based on the EEP and IPW imputed EFs (5) and (7), cannot attain the same semiparametric efficiency bound when data are MAR. Another disadvantage of the MEL estimators based on EEP and IPW methods is that they do not result in an ELLR with a central Chi-square distribution. One can reconcile this latter problem by applying an adjustment to $\mathcal{R}_k(\boldsymbol{\theta}_0), k = 1, 3$, as described in the following corollary.

Corollary 2. *Write* $\hat{\rho}_{1k}(\boldsymbol{\theta}_0) = [\mathcal{W}_{kn}^T\hat{V}_k^{-1}\mathcal{W}_{kn}]^{-1}\ [\mathcal{W}_{kn}^T\hat{\boldsymbol{\Sigma}}^{-1}\mathcal{W}_{kn}]$*. Under the conditions of Theorem 4,* $\mathcal{R}_k(\boldsymbol{\theta}_0)\hat{\rho}_{1k}(\boldsymbol{\theta}_0) \xrightarrow{D} \chi_l^2$*, k = 1,3.*

Hence, an α-level confidence region for $\boldsymbol{\theta}_0$ based on $\mathcal{R}_k(\boldsymbol{\theta}_0)$ is $I_{k\alpha} = \{\boldsymbol{\theta} : \hat{\rho}_k(\boldsymbol{\theta})\mathcal{R}_k(\boldsymbol{\theta}) \leq C_\alpha\}$ for $k = 1, 3$ or $I_{k\alpha} = \{\boldsymbol{\theta} : \mathcal{R}_k(\boldsymbol{\theta}) \leq C_\alpha\}$ for $k = 2$, where C_α is the upper α-quantile of the Chi-square density function with l degrees of freedom. The estimator $\hat{\boldsymbol{\theta}}_{ke}$ also facilitates the development of an ELLR test statistic for testing $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$. This statistic may be written as

$$R_k(\boldsymbol{\theta}_0) = 2\ell_k(\boldsymbol{\theta}_0) - 2\ell_k(\hat{\boldsymbol{\theta}}_{ke}),$$

where $\ell_k(\boldsymbol{\theta}) = \sum_{i=1}^n \log(1 + \boldsymbol{\lambda}_2^T\hat{\mathbf{g}}_{\mathcal{M}_k}(Y_i, \mathbf{Z}_i, \boldsymbol{\theta})), k = 1, 2, 3$.

The following theorem provides the asymptotic properties of $R_k(\boldsymbol{\theta}_0)$.

Theorem 5. *Assume that conditions* $(C_1^k) - (C_2^k), (C_3)$ *and* $(C_4^k) - (C_7^k)$ *in the Appendix are satisfied, and there exists a unique* $\boldsymbol{\theta}_0$ *such that* $E\mathbf{g}(Y, \mathbf{Z}, \boldsymbol{\theta}_0) = 0$*. Then, under* H_0*,*

$$R_1(\boldsymbol{\theta}_0) \xrightarrow{D} \tilde{\sigma}_1\omega_1^2 + \cdots + \tilde{\sigma}_q\omega_q^2, \quad R_2(\boldsymbol{\theta}_0) \xrightarrow{D} \chi_q^2 \quad \text{and} \quad R_3(\boldsymbol{\theta}_0) \xrightarrow{D} \tilde{\varrho}_1\omega_1^2 + \cdots + \tilde{\varrho}_q\omega_q^2, \tag{10}$$

where ω_i^2*'s,* $i = 1, \ldots, q$*, are Chi-square random variables, each with one degree of freedom, and distributed independently of one another, and the weights* $\tilde{\sigma}_i$*'s and* $\tilde{\varrho}_i$*'s,* $i = 1, \ldots, q$*, are eigenvalues of the matrix* $\mathbf{V}_1^{-1}\boldsymbol{\Gamma}(\boldsymbol{\Gamma}^T\mathbf{V}_1^{-1}\boldsymbol{\Gamma})^{-1}\boldsymbol{\Gamma}^T\mathbf{V}_1^{-1}\boldsymbol{\Sigma}$ *and* $\mathbf{V}_3^{-1}\boldsymbol{\Gamma}(\boldsymbol{\Gamma}^T\mathbf{V}_3^{-1}\boldsymbol{\Gamma})^{-1}\boldsymbol{\Gamma}^T\mathbf{V}_3^{-1}\boldsymbol{\Sigma}$*.*

Three corollaries of Theorem 5, labeled as Corollaries 3, 4, and 5, are presented in the Appendix. Corollary 3 considers the application of an adjustment factor, similar to that used on $\mathcal{R}_k(\boldsymbol{\theta}_0), k = 1, 3$, on $R_k(\boldsymbol{\theta}_0), k = 1, 3$, so that the resultant adjusted statistic converges to a Chi-square distribution. Corollaries 4 and 5 develop a profile ELLR (PELLR) testing approach for hypothesis testing when only a subset of $\boldsymbol{\theta}$ is of interest.

Remark 1. Our derivation of results does not assume smoothness in the EFs as in the works of Zhou et al. (2008) and Tang and Qin (2012). Instead, we require the conditional expectation of the EFs to be smooth in the parameters. Despite the fact that the results under the smooth and nonsmooth EF scenarios are derived under different technical conditions, we have found that the asymptotic covariances of GMM and MEL estimators under the two scenarios are similar, the only difference being that $\mathbf{\Gamma} = \nabla_{\theta} E\{\mathbf{g}(Y, \mathbf{Z}, \theta_0)\}$ when the EFs are nonsmooth and $\mathbf{\Gamma} = E\{\nabla_{\theta}\mathbf{g}(Y, \mathbf{Z}, \theta_0)\}$ when the EFs are smooth. The ELLR and PELLR statistics under smooth and nonsmooth EFs also have the same limiting distributions. When the underlying EFs are all smooth and the order of integration and differentiation are exchangeable, our results reduce to those of Zhou et al. (2008) and Tang and Qin (2012).

Remark 2. The relaxation of the smoothness assumption has significantly complicated the derivation of the asymptotic properties, especially with regard to the consistency of the MEL estimator. Note that in proving the asymptotic normality of the MEL estimator under nonsmooth EFs when no data are missing, Lopez et al. (2009) assumed the consistency of the MEL estimator without proving it.

In the following, we give three examples for which our proposed methods are applicable.

Example 1. (Quantile regression)
Consider the linear quantile regression model

$$Y = \mathbf{Z}^T \theta_\tau + \epsilon, \tag{11}$$

where $P(\epsilon < 0 \mid \mathbf{Z}) = \tau$ for the τth quantile ($0 < \tau < 1$). The corresponding EFs are

$$\mathbf{g}(Y, \mathbf{Z}, \theta) = \mathbf{Z}\left(I\left(Y - \mathbf{Z}^T\theta_\tau \le 0\right) - \tau\right). \tag{12}$$

Let $\mathbf{X}$ be a non-null subset of $(Y^T, \mathbf{Z})^T$, which is observed for all subjects, and f_ϵ be the density of ϵ. If ϵ satisfies $0 < f_{\epsilon|\mathbf{Z}}(0|\mathbf{Z}) < \infty$ (e.g., $\epsilon \sim N(\mu, \sigma^2)$ or $\epsilon \sim U(a, b)$), then by the Donsker property of indicator function classes, $(C_5^k) - (C_7^k)$ hold (see example 19.6 of Van der Vaart, 1998). By choosing a suitable kernel function, condition $(C_4^k), k = 1, 2, 3$ can be satisfied. For example, when $d = 1$, the Gaussian kernel satisfies these conditions. On the other hand, when $d > 1$, one can use a suitable higher-order kernel function (see Fan & Hu, 1992). The requirement associated with $f(\mathbf{X})$ in (C_1) can be fulfilled by joint distributions constructed from various continuous distributions including the normal, uniform, exponential, and other distributions. Conditions (C_2^k) and (C_3) can be satisfied by adding some moment condition to $\mathbf{Z}$ and Y.

Example 2. (Difference of conditional quantiles in a one-sample problem)
Consider again the quantile regression model in (11). Let

$$q_{\epsilon\,|\,\mathbf{z}} = F^{-1}_{\epsilon\,|\,\mathbf{z}}(\tau_1) - F^{-1}_{\epsilon\,|\,\mathbf{z}}(\tau), \tag{13}$$

where $1 > \tau_1 > \tau > 0$, $F^{-1}_{\epsilon\,|\,\mathbf{z}}(\tau) = \inf\{e, F_{\epsilon\,|\,\mathbf{z}}(e\,|\,\mathbf{z}) \ge \tau\}$. Then, we have the difference of conditional quantiles of response

$$q_{Y\,|\,\mathbf{z}} = F^{-1}_{Y\,|\,\mathbf{z}}(\tau_1) - F^{-1}_{Y\,|\,\mathbf{z}}(\tau) = q_{\epsilon\,|\,\mathbf{z}} + bZ^T\left(\theta_{\tau_1} - \theta_{\tau_2}\right). \tag{14}$$

Let $\tau = 0.25$ and $\tau_1 = 0.75$. Then, for obtaining the interquantile range, we have the EFs

$$\mathbf{g}(Y, \mathbf{Z}, \theta) = \begin{cases} \mathbf{Z}\left(I\left(Y - \mathbf{Z}^T\theta_\tau \le 0\right) - \tau\right) \\ \left(I\left(Y - q_{\epsilon\,|\,\mathbf{z}} - \mathbf{Z}^T\theta_{\tau_1} \le 0\right) - \tau_1\right) \\ \mathbf{Z}\left(I\left(Y - q_{\epsilon\,|\,\mathbf{z}} - \mathbf{Z}^T\theta_{\tau_1} \le 0\right) - \tau_1\right). \end{cases}$$

Similarly, let $\mathbf{X}$ be a random variable that is fully observed. We can verify conditions $(C_1^k)-(C_7^k)$ along the lines of Example 1, and we omit the details here for brevity. The difference of conditional quantiles in one sample can be generalized to the two-sample case.

Example 3. (ROC curves)
Consider independent responses Y_1 and Y_2 with distribution functions F_1 and F_2, respectively. Assume that observations are missing from Y_1 and Y_2, and the missing propensity score depends on the covariate $\mathbf{Z}$. For $0 < \tau < 1$, consider the ROC curve

$$\alpha =: \mathrm{ROC}(\tau) = 1 - F_1\left(F_2^{-1}(1-\tau)\right). \tag{15}$$

Now, for any $\mathbf{z}_0$, we have the EFs

$$\mathbf{g}(Y, \boldsymbol{\theta}, \alpha) = \begin{cases} I(Y_2 \le \theta) - (1-\tau) \\ I(Y_1 \le \theta) - (1-\alpha). \end{cases}$$

We can verify conditions $(C_1^k)-(C_7^k)$ along the lines of Example 1, and we omit the details here for brevity.

4 | EFFICIENCY COMPARISON

4.1 | Why are the GMM$_E$ and MEL estimators not always equivalent?

This subsection provides a thorough discussion on the asymptotic nonequivalence of the GMM$_E$ and MEL estimators based on EEP and IPW imputed EFs in (5) and (7). This matter deserves attention as the finding contradicts the well-known result that the GMM$_E$ estimator has the same asymptotic normal distribution as the MEL estimator under the moment restriction (1) when no data are missing. Now, from the proof of Theorem 3, the ELLR defined in (9) can be written as

$$\mathcal{R}_k(\boldsymbol{\theta}) = \left\{ \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \hat{\mathbf{g}}_{\mathcal{M}_k}(Y_i, \mathbf{Z}_i, \boldsymbol{\theta}) \right\}^T \mathbf{V}_k^{-1}(\boldsymbol{\theta}) \left\{ \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \hat{\mathbf{g}}_{\mathcal{M}_k}(Y_i, \mathbf{Z}_i, \boldsymbol{\theta}) \right\} + o_p(n^{-1}),$$

where $\mathbf{V}_k(\boldsymbol{\theta})$ is the asymptotic second moment of $\hat{\mathbf{g}}_{\mathcal{M}_k}(Y_i, \mathbf{Z}_i, \boldsymbol{\theta})$. This expression is equivalent to the GMM objective function with $\mathbf{V}_k^{-1}$ as the weight matrix. Thus, the MEL and GMM estimators based on the same weighted matrix $\mathbf{V}_k^{-1}$ have the same asymptotic distribution. Now, if the weight matrix used by the GMM$_E$ estimator, namely, the inverse of the asymptotic covariance of the EFs, is equal to $\mathbf{V}_k^{-1}$, then the GMM$_E$ estimator is asymptotically equivalent to the MEL estimator. That is, the two estimators have the same asymptotic efficiency if and only if the second moment and asymptotic covariance of the EF vector are the same. This same factor also determines if the ELLR and PELLR test statistics follow a central Chi-square asymptotic distribution. Note that $\mathcal{R}_k(\boldsymbol{\theta}_0)$ is central Chi-square distributed if and only if the second moment of the EF vector is identical to its asymptotic covariance.

Now, under the imputed EFs (5) and (7) based on the EEP and IPW approaches, respectively, $\mathbf{V}_k \neq \boldsymbol{\Sigma}$ for $k = 1, 3$. It thus follows that the MEL estimators $\hat{\boldsymbol{\theta}}_{1e}$ and $\hat{\boldsymbol{\theta}}_{3e}$ do not have the same asymptotic distribution as their corresponding GMM$_E$ counterparts $\hat{\boldsymbol{\theta}}_{1g}$ and $\hat{\boldsymbol{\theta}}_{3g}$. As $\mathbf{V}_k \neq \boldsymbol{\Sigma}$ for $k = 1, 3$, the corresponding $\mathcal{R}_k(\boldsymbol{\theta}_0)$, $R_k(\boldsymbol{\theta}_0)$ and $R_k(\boldsymbol{\theta}_{0A}, \hat{\boldsymbol{\theta}}_{keB}(\boldsymbol{\theta}_{0A}))$ do not follow an asymptotic central Chi-square distribution. Similar results have been observed in EL-based censored data studies (Qin & Jing, 2001; Qin & Tsao, 2003). On the other hand, under the imputed AIPW-based EF (6), $\mathbf{V}_2 = \boldsymbol{\Sigma}$, the GMM$_E$ and MEL estimators are asymptotically equally efficient, and the corresponding ELLR and PELLR test statistics are asymptotically Chi-square distributed. When no observation is missing, the second moment and covariance matrix of $\mathbf{g}(Y_i, \mathbf{Z}_i, \boldsymbol{\theta})$ are both equal to $E(\mathbf{g}\mathbf{g}^T)$. Hence, the MEL and GMM$_E$ estimators with $\mathbf{W}_{kn} = E^{-1}(\mathbf{g}\mathbf{g}^T)$ have the same asymptotic normal distribution and are asymptotically semiparametric efficient with a dispersion matrix that is equal to $[\boldsymbol{\Gamma}^T E^{-1}(\mathbf{g}\mathbf{g}^T)\boldsymbol{\Gamma}]^{-1}$.

4.2 | Efficiency comparison of parametric and nonparametric estimators

In this section, we compare the efficiency of the nonparametric GMM$_E$ and MEL estimators with that of their parametric counterparts, which are obtained based on the parametric versions of EFs (5), (6), and (7). From the work of Chen et al. (2008), the efficient score function for the moment restriction model (1) when the data are MAR is

$$\mathbf{g}^{\mathrm{ES}}(Y_i, \mathbf{Z}_i, \boldsymbol{\theta}) = \frac{\delta_i}{P(\mathbf{X}_i)}\mathbf{g}(Y_i, \mathbf{Z}_i, \boldsymbol{\theta}) + \left\{1 - \frac{\delta_i}{P(\mathbf{X}_i)}\right\}\mathbf{m}(\mathbf{X}_i, \boldsymbol{\theta}), \tag{16}$$

with $\boldsymbol{\Sigma}_0^{-1}$ being the corresponding semiparametric efficiency bound. Corollary 1 shows that the GMM$_E$ estimator is determined by the asymptotic covariance of its corresponding modified EFs. Hence, the GMM$_E$ estimator is semiparametric efficient if and only if its corresponding EF has the same asymptotic distribution as the efficient EF $n^{-1/2}\sum_{i=1}^n \mathbf{g}^{\mathrm{ES}}(Y_i, \mathbf{Z}_i, \boldsymbol{\theta}_0)$. On the other hand, Theorem 4 shows that the MEL estimator is semiparametric efficient if and only if its corresponding EF has the same asymptotic distribution as the efficient EF, and the resultant asymptotic covariance is equal to the second moment of the EF.

Let us assume that the propensity score $P(\mathbf{X}_i, \boldsymbol{\beta})$ and the conditional mean $\mathbf{m}(\mathbf{X}_i, \boldsymbol{\theta}, \boldsymbol{\eta})$ of these parametric methods are dependent on $\boldsymbol{\beta}$ and $\boldsymbol{\eta}$ through some parametric specifications. Now, denote the parametric counterparts of $\mathcal{W}_{kn}(\boldsymbol{\theta})$ as $\mathcal{W}^p_{kn}(\boldsymbol{\theta}) = n^{-1/2}\sum_{i=1}^n \hat{\mathbf{g}}^p_{\mathcal{M}_k}(Y_i, \mathbf{Z}_i, \boldsymbol{\theta}), k = 1, 2, 3,$ where $P(\mathbf{X}_i, \boldsymbol{\beta})$ and $\mathbf{m}(\mathbf{X}, \boldsymbol{\theta})$ are estimated parametrically by $P(\mathbf{X}_i, \hat{\boldsymbol{\beta}})$ and $\mathbf{m}(\mathbf{X}_i, \boldsymbol{\theta}, \hat{\boldsymbol{\eta}})$. Let $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\eta}}$ be $\sqrt{n}$-consistent estimators of $\boldsymbol{\beta}$ and $\boldsymbol{\eta}$.

After some calculations, we have

$$\begin{aligned}
\mathcal{W}^p_{1n}(\boldsymbol{\theta}) =& \frac{1}{\sqrt{n}}\sum_{i=1}^n \delta_i \mathbf{g}(Y_i, \mathbf{Z}_i, \boldsymbol{\theta}) + (1-\delta_i)\mathbf{m}(\mathbf{X}_i, \boldsymbol{\theta}, \boldsymbol{\eta}) \\
&+ E\left\{(1-\delta_i)\frac{\partial \mathbf{m}(\mathbf{X}_i, \boldsymbol{\theta}, \boldsymbol{\eta})}{\partial \boldsymbol{\eta}}\right\}\sqrt{n}(\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}) + o_p(1), \\
\mathcal{W}^p_{2n}(\boldsymbol{\theta}) =& \frac{1}{\sqrt{n}}\sum_{i=1}^n \frac{\delta_i}{P(\mathbf{X}_i, \boldsymbol{\beta})}\mathbf{g}(Y_i, \mathbf{Z}_i, \boldsymbol{\theta}) + \left\{1 - \frac{\delta_i}{P(\mathbf{X}_i, \boldsymbol{\beta})}\right\}\mathbf{m}(\mathbf{X}_i, \boldsymbol{\theta}, \boldsymbol{\eta}) \\
&+ E\left\{\left(1 - \frac{\delta_i}{P(\mathbf{X}_i, \boldsymbol{\beta})}\right)\frac{\partial \mathbf{m}(\mathbf{X}_i, \boldsymbol{\theta}, \boldsymbol{\eta})}{\partial \boldsymbol{\eta}}\right\}\sqrt{n}(\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}) \qquad (17) \\
&+ E\left\{\frac{\delta_i \partial P(\mathbf{X}_i, \boldsymbol{\beta})/\partial \boldsymbol{\beta}}{P^2(\mathbf{X}_i, \boldsymbol{\beta})}\left[\mathbf{m}(\mathbf{X}_i, \boldsymbol{\theta}, \boldsymbol{\eta}) - \mathbf{g}(Y_i, \mathbf{Z}_i, \boldsymbol{\theta})\right]\right\}\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) + o_p(1), \qquad (18)
\end{aligned}$$

and

$$\mathcal{W}^p_{3n}(\boldsymbol{\theta}) = \frac{1}{\sqrt{n}}\sum_{i=1}^n \frac{\delta_i}{P(\mathbf{X}_i, \boldsymbol{\beta})}\mathbf{g}(Y_i, \mathbf{Z}_i, \boldsymbol{\theta}) - E\left\{\frac{\delta_i \partial P(\mathbf{X}_i, \boldsymbol{\beta})/\partial \boldsymbol{\beta}}{P^2(\mathbf{X}_i, \boldsymbol{\beta})}\mathbf{g}(Y_i, \mathbf{Z}_i, \boldsymbol{\theta})\right\}\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) + o_p(1),$$

where $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ and $\sqrt{n}(\hat{\boldsymbol{\eta}} - \boldsymbol{\eta})$ can be replaced by their respective i.i.d. sum representation, when parametric functional forms are specified for $P(.)$ and $m(.)$. It is easily seen that when $k = 1, 3$, $\mathcal{W}^p_{kn}(\boldsymbol{\theta}_0)$ can never be asymptotically equivalent to the efficient EF irrespective of whether $P(\mathbf{X}_i, \boldsymbol{\beta})$ and $\mathbf{m}(\mathbf{X}_i, \boldsymbol{\theta}, \boldsymbol{\eta})$ are correctly specified. Thus, the GMM_E and MEL estimators obtained based on (5) and (7), that is, the EEP and IPW EFs, can never achieve the semiparametric efficiency bound. When the parametric models are correctly specified, the expectations in (17) and (18) are zero, resulting in the parametric AIPW EF $\mathcal{W}^p_{kn}(\boldsymbol{\theta}_0)$ having the same asymptotic distribution as the efficient EF, the second moment of $\hat{\mathbf{g}}^p_{\mathcal{M}_k}$ and asymptotic covariance of $\mathcal{W}^p_{kn}(\boldsymbol{\theta}_0)$ being identical. Hence, both the GMM_E and MEL estimators based on the parametric AIPW method are semiparametric efficient when the functional forms specified for $P(.)$ and $m(.)$ are correct. However, this efficiency property is immediately lost when either of the two parametric models is misspecified.

By the continuous mapping theorem and the dominated convergence theorem, it is readily seen that $\hat{\mathbf{g}}_{\mathcal{M}_k}$ and $\hat{\mathbf{g}}^p_{\mathcal{M}_k}$ have the same second moment. Hence, when the parametric models are correctly specified, the parametric GMM_E and MEL estimators based on the AIPW method have the same efficiency as their nonparametric counterparts. Except for this special case, assuming that the same $\boldsymbol{\Gamma}$ and weighted matrix are used in the parametric and nonparametric estimators, from Theorems 1 and 4, the nonparametric GMM_E and MEL estimators are more efficient than their corresponding parametric estimators. Based on the same EF, if the MEL estimator is semiparametric efficient, then so is the GMM_E estimator, but not vice versa.

Remark 3. The GMM_E estimators under the IPW, EEP, and AIPW missing-data handling methods all attain the semiparametric asymptotic efficiency bound and are equally efficient. For the MEL class of estimators, only the AIPW-based estimator can achieve the same asymptotic efficiency. In terms of computational difficulty, the AIPW-based estimators are the most difficult to compute because the AIPW method involves estimating both $P(X_i)$ and $m(X_i, \theta)$. One disadvantage of the IPW- and AIPW-based estimators is that they can lead to unstable estimates when $P(X_i)$ is close to zero. On the other hand, the EEP-based estimators do not suffer from the same drawback. All things considered, the EEP-based GMM_E estimator is the preferred estimator.

Remark 4. When the EFs involve plug-in elements (e.g., $\hat{P}(\mathbf{x})$ and $\hat{m}(\mathbf{x}, \boldsymbol{\theta})$ in $\hat{\mathbf{g}}_{\mathcal{M}_k}(Y_i, \mathbf{Z}_i, \boldsymbol{\theta})$), except when the plug-in elements converge at the same rate as or a slower rate than the EFs, the asymptotic covariance and the second moment of the EFs will be different due to the dependence of the asymptotic covariance on the plug-in elements. In general, it would not be possible to tell from the form of the EF if its asymptotic covariance and second moment are identical without performing a rigorous theoretical analysis.

5 | A SIMULATION STUDY

In this section, we report results of a simulation study on the finite-sample performance of the GMM_E and MEL estimators. We consider the following vector of EFs:

$$\mathbf{g}(Y, \mathbf{Z}, \boldsymbol{\theta}) = \begin{pmatrix} Z_1\left(\frac{1}{2} - I(Y \leq \theta_1 Z_1 + \theta_2 Z_2)\right) \\ Z_2\left(\frac{1}{2} - I(Y \leq \theta_1 Z_1 + \theta_2 Z_2)\right) \\ Y - \theta_1 Z_1 - \theta_2 Z_2 \end{pmatrix},$$

where $\boldsymbol{\theta} = (\theta_1, \theta_2)^T = (1, -2)^T$, $Z_1 \sim U(0,1)$, and $Z_2 \sim N(0,1)$. Given Z_1 and Z_2, Y has a normal distribution with mean $\theta_1 Z_1 + \theta_2 Z_2$ and variance 1. The first and second EFs in $\mathbf{g}(Y, \mathbf{Z}, \boldsymbol{\theta})$ correspond to the moment conditions for median regression and are discontinuous in θ_1 and θ_2; the third EF, on the other hand, is a condition for the mean of the errors and continuous. It is readily seen that $E\{\mathbf{g}(Y, \mathbf{Z}, \boldsymbol{\theta})\} = 0$, which means that all three EFs are unbiased.

The following scenarios of selection probability under the MAR assumption are considered:

- $S_1 : P(z_2) = 0.5 + 0.5 \sin(z_2 - 1)^2$ if $|z_2 - 1| \leq 1$; $P(z_2) = 1$ otherwise;
- $S_2 : P(z_2) = 0.5 + 0.2|z_2 - 1|$ if $|z_2 - 1| \leq 1$; $P(z_2) = 0.8$ otherwise;
- $S_3 : P(z_2, y) = 0.5 + 0.2|z_2 - y|$ if $|z_2 - y| \leq 1$; $P(z_2) = 0.7$ otherwise;
- $S_4 : P(z_2) = \exp(0.5 - z_2)/(1 + \exp(0.5 - z_2))$;
- $S_5 : P(z_2) = \Phi(-z_2)$, where Φ is the distribution function of the $N(0, 1)$ distribution;
- $S_6 : P(z_2) = 0.6$.

For all scenarios except S_3 and S_6, the data missing probability depends only on the fully observed covariate Z_2; for Scenario S_3, it depends on both Z_2 and the response Y, whereas for S_6, the data are missing completely at random. The average missing percentages for the six scenarios are 17.3%, 29.5%, 32.5%, 39.8%, 49.8%, and 40%, respectively.

We set the sample size n to 100 and the number of replications NS to 500. We adopt the Gaussian kernels $K(u) = \exp(-u^2/2)/(2\pi)^{1/2}$ and $K(u, t) = \exp(-u^2/2)\exp(-t^2/2)/(2\pi)$ for estimating $E(\mathbf{g}(Y, \mathbf{Z}, \boldsymbol{\theta}) \mid Z_2)$ and $E(\mathbf{g}(Y, \mathbf{Z}, \boldsymbol{\theta}) \mid Y, Z_2)$, respectively. Following Sepanski, Knickerbocker, and Carroll (1994), we let $h = \hat{\sigma}_{Z_2} n^{-1/3}$ and $h = \hat{\sigma}_{Z_2 Y} n^{-1/3}$ be the bandwidths for estimating $m(z_2, \boldsymbol{\theta})$ and $m(z_2, y, \boldsymbol{\theta})$, respectively, where $\hat{\sigma}_{Z_2}$ is the standard deviation of Z_2, and $\hat{\sigma}_{Z_2 Y}$ is the standard deviation of $[Z_2^T, Y^T]^T$. For the estimation of $P(z_2, y)$, we choose the Gaussian kernel $K(u, t)$ and set the bandwidth to $h = 1.5\hat{\sigma}_{Z_2 Y} n^{-1/5}$. The kernel and bandwidth for $P(x_2)$ are $K(u)$ and $h = 1.5\hat{\sigma}_{Z_2} n^{-1/5}$, respectively.

The estimators are evaluated with respect to bias (BIAS), standard deviation (SD), standard errors (SE), confidence interval coverage to the nominal target coverage of 0.95 (COV), and mean square errors (MSE). The results are reported in Table 2. Although neither the MEL nor the GMM_E estimator strictly dominates the other, that the GMM_E estimator is seen to be superior is about three quarters of the results presented. A notable exception occurs under Scenario S_6, where gains from using MEL in place of GMM_E are more frequently observed. The precise reason for the relative strong showing of the MEL estimator under S_6 is unclear, but we think it may be attributed to the data being missing completely at random for this scenario. If one excludes Scenario S_6, then in over 75% of the comparisons, the GMM_E estimator is seen to be the favored approach. The advantage of the GMM_E estimator is particularly evident with respect to MSE, by which the GMM_E estimator is superior to the MEL estimator in all but 4 of the 36 comparisons. In terms of BIAS, SD, and SE, GMM_E is the superior approach in between 70% and 80% of the comparisons. On the other hand, in terms of COV, the two estimators exhibit comparable performance, with the MEL estimator producing confidence interval coverage closer to 0.95 than does the GMM_E estimator in just under half of the instances.

Although our results in Section 3 show that asymptotically, the EEP, IPW, and AIPW approaches produce identical properties for the GMM_E estimator, Table 2 reveals that in finite samples, there are minor differences in the sampling behavior of the GMM_E estimator under the three approaches. These differences can be attributed to the sample size and the different bandwidths being used for the different missing-data handling approaches. For the same reason, the AIPW-based MEL estimator does not yield the same finite-sample characteristics as the GMM_E estimators, even though they are equally efficient asymptotically.

TABLE 2 Simulation results

	EEP				IPW				AIPW			
	MEL		GMM_E		MEL		GMM_E		MEL		GMM_E	
	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}_1$	$\hat{\theta}_2$
						S_1						
BIAS	−0.012	−0.002	−0.009	0.006	−0.010	0.004	−0.007	0.004	−0.003	−0.012	−0.003	−0.008
SD	0.241	0.140	0.241	0.139	0.231	0.141	0.236	0.131	0.231	0.133	0.233	0.135
SE	0.272	0.175	0.227	0.145	0.292	0.179	0.231	0.145	0.239	0.152	0.239	0.143
COV	0.952	0.940	0.922	0.915	0.944	0.957	0.934	0.935	0.956	0.982	0.956	0.974
MSE	0.058	0.020	0.057	0.019	0.053	0.020	0.055	0.017	0.053	0.018	0.054	0.018
						S_2						
BIAS	−0.002	0.004	0.000	0.004	−0.012	0.002	−0.009	0.005	−0.008	−0.003	−0.002	−0.006
SD	0.264	0.152	0.261	0.151	0.248	0.153	0.247	0.149	0.238	0.141	0.235	0.139
SE	0.317	0.194	0.253	0.158	0.295	0.183	0.268	0.176	0.249	0.161	0.251	0.162
COV	0.970	0.962	0.932	0.921	0.938	0.972	0.954	0.962	0.952	0.980	0.956	0.974
MSE	0.070	0.023	0.068	0.023	0.062	0.023	0.061	0.022	0.057	0.020	0.055	0.019
						S_3						
BIAS	0.006	−0.003	0.001	−0.000	0.003	0.007	−0.001	0.004	0.004	−0.003	0.011	0.001
SD	0.249	0.149	0.245	0.143	0.253	0.150	0.247	0.143	0.217	0.137	0.215	0.138
SE	0.268	0.198	0.254	0.168	0.283	0.184	0.284	0.179	0.233	0.159	0.227	0.160
COV	0.940	0.930	0.925	0.930	0.942	0.962	0.945	0.956	0.964	0.972	0.962	0.968
MSE	0.062	0.022	0.060	0.020	0.064	0.023	0.061	0.020	0.047	0.019	0.046	0.019
						S_4						
BIAS	−0.026	−0.017	−0.018	−0.011	−0.013	−0.009	0.004	−0.001	−0.022	−0.013	−0.013	−0.013
SD	0.288	0.195	0.268	0.185	0.259	0.165	0.245	0.159	0.273	0.186	0.266	0.172
SE	0.279	0.189	0.249	0.178	0.239	0.163	0.233	0.169	0.240	0.176	0.238	0.174
COV	0.936	0.938	0.930	0.956	0.926	0.962	0.922	0.962	0.918	0.950	0.922	0.960
MSE	0.083	0.038	0.072	0.034	0.067	0.027	0.060	0.025	0.075	0.035	0.071	0.030
						S_5						
BIAS	−0.014	−0.000	−0.002	0.004	−0.001	−0.006	0.016	0.010	−0.001	0.002	0.013	0.000
SD	0.367	0.258	0.314	0.226	0.281	0.169	0.263	0.161	0.348	0.274	0.310	0.219
SE	0.326	0.235	0.324	0.231	0.286	0.190	0.277	0.188	0.331	0.243	0.329	0.234
COV	0.938	0.946	0.962	0.952	0.954	0.980	0.958	0.978	0.944	0.938	0.964	0.966
MSE	0.134	0.067	0.098	0.051	0.079	0.029	0.069	0.026	0.121	0.075	0.096	0.048
						S_6						
BIAS	−0.012	−0.003	−0.013	−0.003	−0.011	−0.003	−0.013	0.000	−0.011	−0.002	−0.012	−0.007
SD	0.255	0.158	0.259	0.164	0.252	0.156	0.253	0.154	0.259	0.165	0.255	0.163
SE	0.261	0.175	0.264	0.175	0.226	0.158	0.228	0.161	0.263	0.174	0.264	0.174
COV	0.956	0.970	0.942	0.958	0.922	0.952	0.926	0.950	0.948	0.964	0.952	0.960
MSE	0.065	0.025	0.067	0.027	0.064	0.024	0.064	0.024	0.067	0.027	0.065	0.026

Note. AIPW = augmented inverse probability weighted; BIAS = bias; COV = confidence interval coverage to the nominal target coverage of 0.95; EEP = estimating equation projection; GMM_E = efficient generalized method of moments; IPW = inverse probability weighted; MEL = maximum empirical likelihood; MSE = mean square error; SD = standard deviation; SE = standard error.

6 | A REAL DATA EXAMPLE

In this section, we illustrate the proposed method using the data given in the work of Carpenter and Kenward (2005), extracted from the work of Blatchford, Goldstein, Martin, and Browne (2002). These data comprise 4,873 observations of a number of attributes of school children collected from 172 schools in the U.K. between 1996 and 1997. We randomly select 1,000 observations from this sample for the analysis. An objective of the study is to examine the extent to which the literacy ability of school children is affected by the different attributes. Our

TABLE 3 Results of real data analysis

			intercept	Z_1	Z_2	Z_3	Z_4
EEP	MEL	EST	0.253	0.664	−0.076	−0.110	−0.479
		SE	0.044	0.058	0.052	0.053	0.077
	GMM_E	EST	0.320	0.687	−0.144	−0.187	−0.511
		SE	0.041	0.034	0.053	0.043	0.074
IPW	MEL	EST	0.259	0.665	−0.127	−0.117	−0.479
		SE	0.045	0.042	0.047	0.056	0.0604
	GMM_E	EST	0.233	0.626	−0.112	−0.113	−0.468
		SE	0.042	0.038	0.054	0.046	0.059
AIPW	MEL	EST	0.246	0.696	−0.141	−0.102	−0.534
		SE	0.045	0.039	0.042	0.045	0.050
	GMM_E	EST	0.306	0.793	−0.145	−0.125	−0.646
		SE	0.043	0.039	0.051	0.049	0.045

Note. AIPW = augmented inverse probability weighted; EEP = estimating equation projection; EST = parameter estimator; GMM_E = efficient generalized method of moments; IPW = inverse probability weighted; MEL = maximum empirical likelihood; SE = standard error.

illustration is based on the following attributes: Z_1 (= prereception literacy score), Z_2 (= gender: 1 = male, 0 = female), Z_3 (= eligibility for free school meals: 1 = yes, 0 = no), and Z_4 (= term of school entry: 1 = spring or summer term, 0 = autumn term). The dependent variable of the study is Y (= postreception literacy score). Observations of all variables are complete except for Z_1, for which only 649 observations (or 64.9% of data) are available.

We examine the correlation between the missing indicator and the covariates for which the data are complete. It is found that all variables except Z_4 are correlated with the missing indicator. Hence, it seems reasonable to assume the dependence of the propensity function $P(.)$ on (Y, Z_2, Z_3). When analyzing the dependency of Y on the attributes, we utilize both linear mean and median regressions. From these imputed EEs, we obtain the least squares residuals; while using the medians of these residuals, we can obtain the moment conditions of the median regression. Hence, the imputed median regression EEs are based on (5) and (7).

We adopt the Gaussian kernel $K(u, v, t) = \exp(-u^2/2)\exp(-v^2/2)\exp(-t^2/2)/(2\pi)^{3/2}$ with bandwidth $h = 1.06\hat{\sigma}n^{-1/3}$ for estimating $P(Y, X_2, X_3)$ and $K(u, v, t, s) = \exp(-u^2/2)\exp(-v^2/2)\exp(-t^2/2)\exp(-s^2/2)/(2\pi)^2$ with $h = 1.06\hat{\sigma}n^{-1/3}$ for estimating $E[\mathbf{g}(Y, \mathbf{Z}, \boldsymbol{\theta})|Y, Z_2, Z_3, Z_4]$, where $\hat{\sigma}$ is the standard deviation of $[Y^T, Z_2^T, Z_3^T, Z_4^T]^T$.

The results are summarized in Table 3, where EST denotes the parameter estimator and SE denotes the standard error. From the Table, we observe that there is no significant discrepancy in results across the two imputed EE procedures and across the two methods of estimation. The estimation results indicate that the prereception literacy score has a positive association with the postreception score. There is a slight gender difference in favor of girls, a moderate disadvantage to those eligible for free school meals and a strong advantage to those entering school in the autumn term. Qualitatively, these conclusions are the same as those obtained by Carpenter and Kenward (2005). All the four covariates are significant here. On the whole, the GMM_E estimators are more efficient than the MEL estimators.

7 | CONCLUDING REMARKS

GMM and EL are routinely used for combining sample and auxiliary information. When sample data are completely observed and the moment conditions are correct, it is a well-established result that the GMM_E and MEL estimators are asymptotically equivalent as well as semiparametric efficient. This paper shows that this finding need not hold true when the data sample is incomplete. We have shown that of the IPW, AIPW, and EEP missing-data handling methods, only the AIPW method yields a MEL estimator that is semiparametric efficient, whereas the GMM_E estimators obtained under all three methods are semiparametric efficient. Our theoretical analysis demonstrates that the MEL estimator has the same asymptotic efficiency as the GMM_E estimator if and only if the second moment of the EF vector is identical to the asymptotic covariance of the EF. Of the situations we have considered, this arises either when no data are missing from the sample or the AIPW method is used as the missing-data handling procedure. This same condition also determines if the ELLR and PELLR test statistics follow a central Chi-square asymptotic distribution. We have also shown that the nonparametric GMM_E and MEL estimators are generally more efficient than their parametric counterparts. All our results are derived without assuming smoothness in the underlying EFs. We consider the relaxation of the smoothness assumption a significant theoretical advance.

The missing-data scenario considered in this paper is restricted to MAR. Studies by Hemvanich (2007) and Tang, Zhao, and Zhu (2014) have considered GMM and EL inference in EFs with nonignorably missing response data. Thorough comparisons of the two approaches as well as the development of an efficiency bound analogous to that derived by Chen et al. (2008) under a nonignorable missing-data situation are yet to be considered. These remain for future research.

ACKNOWLEDGEMENTS

The authors greatly acknowledge the Editors, Dr. Dalgaard and Dr. Hansen, the Associate Editor, and two referees for their many comments and suggestions that greatly improved the paper. Chen's work was supported by the National Natural Science Foundation of China (NSFC)(11501461) and the Fundamental Research Funds for the Central Universities (JBK140507, JBK1806002, and JBK1802070); Chen's and Zhou's work were supported by the NSFC (71571113), the State Key Program of the National Natural Science Foundation of China (71331006 and 91546202), and the Innovative Research Team in SUFE (IRTSHUFE13122402).

ORCID

Xuerong Chen http://orcid.org/0000-0001-7012-7312
Yong Zhou http://orcid.org/0000-0001-9751-3141

REFERENCES

Blatchford, P., Goldstein, H., Martin, C., & Browne, W. (2002). A study of class size effects in English school reception year classes. *British Educational Research Journal*, *28*, 169–185.

Carpenter, J., & Kenward, M. (2005). Example analyses using Winbugs 1.4. Available at http://www.missingdata.org.uk/

Chamberlain, G. (1987). Asymptotic efficiency in estimation with conditional moment restrictions. *Journal of Econometrics*, *34*, 305–334.

Chen, X. H., Hong, H., & Tarozzi, A. (2008). Semiparametric efficiency in GMM models with auxiliary data. *The Annals of Statistics*, *36*, 808–843.

Chen, X. R., Wan, A. T. K., & Zhou, Y. (2015). Efficient quantile regression analysis with missing observations. *Journal of the American Statistical Association*, *110*, 723–741.

Fan, J. Q., & Hu, T.-C. (1992). Bias correction and higher order kernel functions. *Statistics & Probability Letters*, *13*, 235–243.

Hansen, L. P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica*, *50*, 1029–1054.

Hemvanich, S. (2007). *GMM estimation for nonignorable missing data: Theory and practice* (PhD thesis). Coventry, UK: Department of Economics, University of Warwick.

Imbens, G. W. (1997). One-step estimators for over-identified generalized method of moments models. *The Review of Economic Studies*, *64*, 359–383.

Imbens, G. W. (2002). Generalized method of moments and empirical likelihood. *Journal of Business & Economic Statistics*, *20*, 493–506.

Kosorok, M. R. (2008). *Introduction to empirical processes and semiparametric inference*. New York, NY: Springer.

Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data* (2nd ed.). Hoboken, NJ: John Wiley & Sons.

Lopez, E. M. M., Van Keilegom, I. V., & Veraverbeke, N. (2009). Empirical likelihood for non-smooth criterion functions. *Scandinavian Journal of Statistics*, *36*, 413–432.

Newey, W. K., & McFadden, D. (1994). Large sample estimation and hypothesis testing. *Handbook of Econometrics*, *4*, 2111–2245.

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

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

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

Owen, A. B. (2001). *Empirical likelihood*. London, UK: Chapman and Hall/CRC.

Paik, M. C. (1997). The generalized estimating equation approach when data are not missing completely at random. *Journal of the American Statistical Association*, *92*, 1320–1329.

Qin, G., & Jing, B.-Y. (2001). Empirical likelihood for censored linear regression. *Scandinavian Journal of Statistics*, *28*, 661–673.

Qin, G., & Tsao, M. (2003). Empirical likelihood inference for median regression models for censored survival data. *Journal of Multivariate Analysis*, *85*, 416–430.

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

Sepanski, J. H., Knickerbocker, R., & Carroll, R. J. (1994). A semiparametric correction for attenuation. *Journal of the American Statistical Association*, *89*, 1366–1373.

Sherman, R. P. (1993). The limiting distribution of the maximum rank correlation estimator. *Econometrica*, *61*, 123–137.

Tang, C. Y., & Qin, Y. (2012). An efficient empirical likelihood approach for estimating equations with missing data. *Biometrica*, *99*, 1001–1007.

Tang, N., Zhao, P., & Zhu, H. (2014). Empirical likelihood for estimating equations with nonignorably missing data. *Statistica Sinica*, *24*, 723–747.

Van der Vaart, A. W. (1998). *Asymptotic statistics*. New York, NY: Cambridge University Press.

Van der Vaart, A. W., & Wellner, J. A. (1996). *Weak convergence and empirical processes: With applications to statistics*. New York, NY: Springer-Verlag.

Wang, D., & Chen, S. X. (2009). Empirical likelihood for estimating equations with missing values. *The Annals of Statistics*, *37*, 490–517.

Xue, L. G. (2009). Empirical likelihood confidence intervals for response mean with data missing at random. *Scandinavian Journal of Statistics*, *36*, 671–685.

Zhou, H., & Pepe, M. S. (1995). Auxiliary covariate data in failure time regression. *Biometrika*, *82*, 139–149.

Zhou, Y., Wan, A. T. K., & Wang, X. (2008). Estimating equation inference with missing data. *Journal of the American Statistical Association*, *103*, 1187–1199.

How to cite this article: Chen X, Chen Y, Wan ATK, Zhou Y. On the asymptotic non-equivalence of efficient-GMM and MEL estimators in models with missing data. *Scand J Statist*. 2019;46:361–388. https://doi.org/10.1111/sjos.12354

APPENDIX

For simplicity, we apply the same kernel function to $\mathbf{m}(\mathbf{x},\boldsymbol{\theta})$ and $P(\mathbf{x})$, that is, we set $K^{(i)}(.) = K(.)$ and $\bar{K}^{(i)} = \bar{K}(.), i = 1, 2, \dots, l$. We also assume that $K(.)$ and $\bar{K}(.)$ are symmetric probability density functions, and when $n \to \infty$, all bandwidths h_i's and $\bar{h}_i$'s approach 0 in the same order as h and $\bar{h}$, that is, $h_i = O(h)$ and $h_i = O(\bar{h}), i = 1, 2, \dots, l$. Write $a^{\otimes 2} = aa^T$, $\mu_k = \int u^k K(u)du$, and $\bar{\mu}_k = \int u^k \bar{K}(u)du$, and let $\mu_0 = 1, \bar{\mu}_0 = 1, \mu_2 = 1$, and $\bar{\mu}_2 = 1$. Now, Let $f(\mathbf{x})$ be the probability density function of $\mathbf{X}$, with $f(\mathbf{x})$ being bounded away from 0 and infinite in the support of $\mathbf{X}$. Denote $r(\mathbf{x}) = f(\mathbf{x})P(\mathbf{x})$. We further assume that the order of integration and differentiation can be exchanged, $\mathbf{V}_1, \boldsymbol{\Sigma}(= \mathbf{V}_2)$ and $\mathbf{V}_3$ are positive definite, and the parameter space Θ is compact. The following technical conditions are required for our proofs of results.

- (C_1) $f(\mathbf{x})$ and $P(\mathbf{x})$ have bounded partial derivatives with respect to x up to an order b with $b \geq 2, 2b > d$, $\inf_x r(\mathbf{x}) \geq c_0$, and $\inf_x P(\mathbf{x}) \geq \tilde{c}_0$, where c_0 and $\tilde{c}_0$ are arbitrarily small constants.
- (C_2^k) $\mathbf{m}(\mathbf{x},\boldsymbol{\theta})$ have bounded partial derivatives with respect to $\mathbf{x}$ up to an order b, and $\|\partial \mathbf{m}(\mathbf{x},\boldsymbol{\theta})/\partial\boldsymbol{\theta}\|$, $\|\partial^2 \mathbf{m}(\mathbf{x},\boldsymbol{\theta})/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T\|$, and $\|\tilde{\mathbf{g}}_{\mathcal{M}_k}(.)\|^3$ can be bounded by some integrable function $M(\mathbf{x})$ in a neighborhood of $\boldsymbol{\theta}_0$, $k = 1, 2, 3$.
- (C_3) $E\{[1 - P(\mathbf{X})][\nabla_x m(X,\boldsymbol{\theta}_0)/r(\mathbf{X})]^2\} < \infty$, where $\nabla_x \mathbf{m}(\mathbf{x},\boldsymbol{\theta}) = \frac{\partial \mathbf{m}(\mathbf{x},\boldsymbol{\theta})}{\partial x}$, $E[\boldsymbol{\Sigma}_g(\mathbf{X})/P(\mathbf{X})] < \infty$, $E[\mathbf{g}(Y,\mathbf{Z},\boldsymbol{\theta})]^{\otimes 2} < \infty$, and $E[\mathbf{m}(\mathbf{X},\boldsymbol{\theta})]^{\otimes 2} < \infty$.
- (C_4^1) $K(.)$ is a kernel function with compact support and order b and satisfies the Lipschitz condition; $h = h_n \to 0, nh^{2d} \to \infty, nh^{2b} \to 0$, and $nh^d/\log n \to \infty$ as $n \to \infty$.
- (C_4^2) $K(.)$ and $\bar{K}(.)$ are kernel functions with compact support and order b and satisfy the Lipschitz condition; $h \to 0, nh^{2d} \to \infty, nh^{2b} \to 0, nh^d/\log n \to \infty, \bar{h} = \bar{h}_n \to 0, n\bar{h}^{2d} \to \infty, n\bar{h}^{2b} \to 0$, and $n\bar{h}^d/\log n \to \infty$ as $n \to \infty$.
- (C_4^3) $\bar{K}(.)$ is a kernel function with compact support and order b and satisfies the Lipschitz condition; $\bar{h} = \bar{h}_n \to 0, n\bar{h}^{2d} \to \infty, n\bar{h}^{2b} \to 0$, and $n\bar{h}^d/\log n \to \infty$ as $n \to \infty$.
- (C_5^k) The class of functions $\{\tilde{\mathbf{g}}_{\mathcal{M}_k}(Y,\mathbf{Z},\boldsymbol{\theta}), \boldsymbol{\theta} \in \Theta\}$ is a Donsker class, $k = 1, 2, 3$.
- (C_6^k) $\sup_{\boldsymbol{\theta}} |n^{-1}\sum_{i=1}^n [\tilde{\mathbf{g}}_{\mathcal{M}_k}(Y_i,\mathbf{Z}_i,\boldsymbol{\theta}) - E\tilde{\mathbf{g}}_{\mathcal{M}_k}(Y,\mathbf{Z},\boldsymbol{\theta}) - \tilde{\mathbf{g}}_{\mathcal{M}_k}(Y_i,\mathbf{Z}_i,\boldsymbol{\theta}_0)]| = o_p(n^{-1/2})$ uniformly in $\boldsymbol{\theta}$ for $\boldsymbol{\theta} - \boldsymbol{\theta}_0 = O_p(n^{-1/2})$, $k = 1, 2, 3$.
- (C_7^k) $E[\tilde{\mathbf{g}}_{\mathcal{M}_k}(Y,\mathbf{Z},\boldsymbol{\theta})^{\otimes 2} \mid X]$ is continuous in a neighborhood of $\boldsymbol{\theta}_0$, $k = 1, 2, 3$.

Similar conditions have been used in other studies. For instance, conditions (C_1) are similar to, but slightly stronger than, conditions $(A1)$ and $(A2)$ of Zhou et al. (2008), condition $C1$ of Wang and Chen (2009), and condition (C_1) of Xue (2009). Conditions (C_2^k) and (C_7) are generalizations of condition $(C1)$ of Lopez et al. (2009) to missing-data situations. Unlike the work of Lopez et al. (2009), where smoothness conditions are required for $Eg(X, \mu_0, \nu)$, we require a smoothness condition for $\mathbf{m}(\mathbf{x},\boldsymbol{\theta}) = E\{\mathbf{g}(Y,\mathbf{Z},\boldsymbol{\theta}) \mid \mathbf{X} = \mathbf{x}\}$ as our EFs are imputed. Moreover, (C_2^k) is similar to condition (A_5) of Zhou et al. (2008) and conditions $C2$ and $C3$ of Wang and Chen (2009). Condition (C_3) guarantees a finite variance for the estimator, and (C_4) is common in the nonparametric literature.

Conditions (C_5^k) and (C_6^k) are common when the criterion functions are nonsmooth in the unknown parameters, and they ensure the consistency and asymptotic normality of estimators. By parts (i) and (v) of corollary 9.32 in the work of Kosorok (2008) (Donsker preservation results), (C_5^k) is satisfied if $\{\mathbf{g}(Y,\mathbf{Z},\boldsymbol{\theta}), \boldsymbol{\theta} \in \Theta\}$ and $\{\mathbf{m}(\mathbf{X},\boldsymbol{\theta}), \boldsymbol{\theta} \in \Theta\}$ are uniformly bounded Donsker classes because $|\delta| \leq 1$ and $|1/P(\mathbf{x})| \leq 1/\tilde{c}_0$ are uniformly bounded. Condition (C_6^k) is satisfied if $\{\tilde{\mathbf{g}}_{\mathcal{M}_k}(Y,\mathbf{Z},\boldsymbol{\theta}), \boldsymbol{\theta} \in \Theta\}$ is a Donsker class and $\tilde{\mathbf{g}}_{\mathcal{M}_k}(Y,\mathbf{Z},\boldsymbol{\theta})$ is L_2 continuous in $\boldsymbol{\theta}_0$ (see Lemma 3.3.5 of Van der Vaart & Wellner, 1996). A class of measurable functions $\mathcal{F}$ is a Donsker class if its bracketing integral $J_{[\,](\infty,\mathcal{F},L_2(P))}$ or its uniform entropy integral

$J(1, \mathcal{F}, L_2(P))$ is finite, where $J_{[]}(\delta, \mathcal{F}, L_r(P)) = \int_0^\delta [\log N_{[]}(\epsilon, \mathcal{F}, L_r(P))]^{\frac{1}{2}} d\epsilon$, and $J(\delta, \mathcal{F}, L_r(P)) = \int_0^\delta [\log \sup_P N(\epsilon \|F\|_{P,r}, \mathcal{F}, L_r(P))]^{\frac{1}{2}} d\epsilon$, with the bracketing number $N_{[]}(\epsilon, \mathcal{F}, L_r(P))$ being the minimum number of ϵ- brackets in $L_r(P)$ that guarantees every $f \in \mathcal{F}$ lies in at least one bracket and the covering number $N(\epsilon\|F\|_{P,r}, \mathcal{F}, L_r(P))$ being the minimum number of $L_r(P)\epsilon$-ball that covers $\mathcal{F}$. Note that many common function classes are of the Donsker class; for example, the class consists of all indicator functions with the form $f_t = I_{(\infty,t]}$, the Lipschitz continuous function class defined on compact support; the smooth function class $\{f : f(.,\theta) \in R, \theta \in \Theta, |\partial^\nu f(\theta)/\partial\theta^\nu| \leq M_0\}$ for $\Theta \in R^q$ being compact and $\nu > q/2$. Other examples of the Donsker class include the monotonic function, bounded variation function, and the Sobolev classes. Among these examples, the indicator Donsker class is most important for nonsmooth EE inference because most nonsmooth EEs are made up of indicator functions. More details can be found in the works of Van der Vaart (1998) and Kosorok (2008).

For certain special cases, some of the conditions given above are stronger than necessary. We provide the stronger conditions only for the purpose of unifying the theoretical analysis for all cases. For example, GMM generally requires weaker conditions than EL because the GMM objective function is a quadratic function with desirable properties that EL does not possess. If one is only interested in GMM inference, then (C_5^k) can be replaced by the following weaker condition $(\tilde{C}_5^k)$: $n^{-1}\sum_{i=1}^n [\mathbf{g}_{\mathcal{M}_k}(Y_i, \mathbf{Z}_i, \theta) - E\mathbf{g}_{\mathcal{M}_k}(Y, \mathbf{Z}, \theta)] = O_p(n^{-1/2})$ uniformly in θ in a neighborhood of θ_0 and $\sup_{\theta\in\Theta}|n^{-1}\sum_{i=1}^n \mathbf{g}_{\mathcal{M}_k}(Y_i, \mathbf{Z}_i, \theta) - E\mathbf{g}_{\mathcal{M}_k}(Y, \mathbf{Z}, \theta)| = o_p(1)$. Also, the conditions required for $\hat{\mathbf{g}}_{\mathcal{M}_2}$ are necessarily stronger than those for $\hat{\mathbf{g}}_{\mathcal{M}_1}$ due to the inclusion of $1/\hat{P}(\mathbf{X})$ in $\hat{\mathbf{g}}_{\mathcal{M}_2}$. If interest focuses only on $\hat{\mathbf{g}}_{\mathcal{M}_1}$, then $\tilde{\mathbf{g}}_{\mathcal{M}_1}(Y, \mathbf{Z}, \theta)$ in conditions $(C_6^1) - (C_7^1)$ can be replaced by $\mathbf{g}(Y, \mathbf{Z}, \theta)$, and the resultant conditions also guarantee our main results for $k = 1$. In addition, for GMM inference based on $\hat{\mathbf{g}}_{\mathcal{M}_1}$, $\tilde{\mathbf{g}}_{\mathcal{M}_1}(.)$ in $(\tilde{C}_5^1)$ can be replaced by $g(.)$.

It is worth mentioning that we do not assume the consistency of the MEL estimators; instead, we prove the consistency based on the conditions stated in the paper. We also do not require the EFs to be uniformly bounded. Again, unlike the second part of condition ($C1$) of Lopez et al. (2009), we do not make assumptions on $\zeta_k(\theta)$ defined in (A9) to be continuously differentiable in a neighborhood of θ_0, but we derive its differential property instead. This property is then used to guarantee the smoothness property of the expectation of the ELLR. The latter property is crucial for our present analysis concerning nonsmooth EEs. It is instructive to note that $\zeta_k(\theta)$, in fact, plays a similar role to that of $\lambda_k, k = 1, 2$; for example, while λ_k satisfies (8), $\zeta_k(\theta)$ satisfies its expectation (A9), and from (A11) and (A12), $\zeta_k(\theta)$ may be viewed as the limit of λ_k. Our conditions are introduced primarily to handle the nonsmooth characteristics of the EFs, and these conditions generally also hold for smooth EFs. Naturally, for the latter case, the conditions of Zhou et al. (2008) are also applicable.

Before proving the main results, we present the following three corollaries of Theorem 5.

Corollary 3. *Consider the adjustment factor*

$$\hat{\rho}_{2k}(\theta_0) = \left[\mathcal{W}_{kn}^T \hat{V}_k^{-1}\hat{\boldsymbol{\Gamma}}\left(\hat{\boldsymbol{\Gamma}}^T\hat{V}_k^{-1}\hat{\boldsymbol{\Gamma}}\right)^{-1}\hat{\boldsymbol{\Gamma}}^T\hat{V}_k^{-1}\mathcal{W}_{kn}\right)\right]^{-1}\left[\mathcal{W}_{kn}^T\hat{\boldsymbol{\Sigma}}^{-1}\hat{\boldsymbol{\Gamma}}\left(\hat{\boldsymbol{\Gamma}}^T\hat{\boldsymbol{\Sigma}}^{-1}\hat{\boldsymbol{\Gamma}}\right)^{-1}\hat{\boldsymbol{\Gamma}}^T\hat{\boldsymbol{\Sigma}}^{-1}\mathcal{W}_{kn}\right].$$

Under the conditions of Theorem 5, $R_k(\theta_0)\hat{\rho}_{2k}(\theta_0) \xrightarrow{D} \chi_q^2, k = 1, 2.$

By Corollary 3, an adjustment factor, similar to that used on $\mathcal{R}_k(\theta_0), k = 1, 3$, can be applied to $R_k(\theta_0), k = 1, 3$, so that the resultant adjusted statistic converges to a Chi-square distribution.

When only a subset of θ is of interest, a profile likelihood approach may be used for conducting hypothesis tests. Write $\theta = (\theta_A^T, \theta_B^T)^T$, with θ_A and θ_B being $q_1 \times 1$ and $(q - q_1) \times 1$

dimensional vectors, respectively. We assume that only $\boldsymbol{\theta}_A$ is of interest and $\boldsymbol{\theta}_B$ is a vector of nuisance parameters. Denote $\boldsymbol{\Gamma}_2 = \nabla_{\theta_B} E\{\mathbf{g}(Y, \mathbf{Z}, \boldsymbol{\theta}_0)\}$. The PELLR test statistic for testing $H_0^* : \boldsymbol{\theta}_A = \boldsymbol{\theta}_{0A}$ is defined as

$$R_k\left(\boldsymbol{\theta}_{0A}, \hat{\boldsymbol{\theta}}_{keB}(\boldsymbol{\theta}_{0A})\right) = 2\ell_k\left(\boldsymbol{\theta}_{0A}, \hat{\boldsymbol{\theta}}_{keB}(\boldsymbol{\theta}_{0A})\right) - 2\ell_k\left(\hat{\boldsymbol{\theta}}_{ke}\right),$$

where $\hat{\boldsymbol{\theta}}_{keB}$ is the maximizer of $L_k(\boldsymbol{\theta})$ subject to the constraint $\boldsymbol{\theta}_A = \boldsymbol{\theta}_{0A}$, $k = 1, 2, 3$. The asymptotic distribution of this PELLR test statistic is given in the following corollary.

Corollary 4. *Assume that the conditions of Theorem 5 are satisfied. Then, under H_0^*, we have*

$$R_1\left(\boldsymbol{\theta}_{0A}, \hat{\boldsymbol{\theta}}_{1eB}(\boldsymbol{\theta}_{0A})\right) \xrightarrow{D} \bar{\sigma}_1\omega_1^2 + \cdots + \bar{\sigma}_{q_1}\omega_{q_1}^2, \quad R_2\left(\boldsymbol{\theta}_{0A}, \hat{\boldsymbol{\theta}}_{2eB}(\boldsymbol{\theta}_{0A})\right) \xrightarrow{D} \chi_{q_1}^2, \tag{A1}$$

and

$$R_3\left(\boldsymbol{\theta}_{0A}, \hat{\boldsymbol{\theta}}_{3eB}(\boldsymbol{\theta}_{0A})\right) \xrightarrow{D} \bar{\varrho}_1\omega_1^2 + \cdots + \bar{\varrho}_{q_1}\omega_{q_1}^2, \tag{A2}$$

where ω_i^2's, $i = 1, \ldots, q_1$, are Chi-square random variables, each with one degree of freedom, and distributed independently of one another, and the weights $\bar{\sigma}_i$'s and $\bar{\varrho}_i$'s, $i = 1, \ldots, q_1$, are eigenvalues of $\boldsymbol{\Sigma}^{\frac{1}{2}}\mathbf{V}_k^{-1}[\boldsymbol{\Gamma}(\boldsymbol{\Gamma}^T\mathbf{V}_k^{-1}\boldsymbol{\Gamma})^{-1}\boldsymbol{\Gamma}^T - \boldsymbol{\Gamma}_2(\boldsymbol{\Gamma}_2^T\mathbf{V}_k^{-1}\boldsymbol{\Gamma}_2)^{-1}\boldsymbol{\Gamma}_2^T]\mathbf{V}_k^{-1}\boldsymbol{\Sigma}^{\frac{1}{2}}$ for $k = 1, 3$, respectively.

An adjusted version of $R_k(\boldsymbol{\theta}_{0A}, \hat{\boldsymbol{\theta}}_{1eB}(\boldsymbol{\theta}_{0A}))$, $k = 1, 3$ that converges to an asymptotic Chi-square distribution is provided in the following corollary.

Corollary 5. *Let $\hat{\rho}_{3k}(\boldsymbol{\theta}_{0A}, \hat{\boldsymbol{\theta}}_{keB}(\boldsymbol{\theta}_{0A})) = \{\mathcal{W}_{kn}^T\hat{V}_k^{-1}[\hat{\boldsymbol{\Gamma}}(\hat{\boldsymbol{\Gamma}}^T\hat{V}_k^{-1}\hat{\boldsymbol{\Gamma}})^{-1}\hat{\boldsymbol{\Gamma}}^T - \hat{\boldsymbol{\Gamma}}_2(\hat{\boldsymbol{\Gamma}}_2^T\mathbf{V}_k^{-1}\hat{\boldsymbol{\Gamma}}_2)^{-1}\ \hat{\boldsymbol{\Gamma}}_2^T]\hat{V}_k^{-1}\ \mathcal{W}_{kn})\}^{-1}\ \{\mathcal{W}_{kn}^T\hat{\boldsymbol{\Sigma}}^{-1}[\hat{\boldsymbol{\Gamma}} \times (\hat{\boldsymbol{\Gamma}}^T\hat{\boldsymbol{\Sigma}}^{-1}\hat{\boldsymbol{\Gamma}})^{-1}\hat{\boldsymbol{\Gamma}}^T - \hat{\boldsymbol{\Gamma}}_2(\hat{\boldsymbol{\Gamma}}_2^T\hat{\boldsymbol{\Sigma}}^{-1}\hat{\boldsymbol{\Gamma}}_2)^{-1}\hat{\boldsymbol{\Gamma}}_2^T]\hat{\boldsymbol{\Sigma}}^{-1}\mathcal{W}_{kn})\}$. Under the conditions of Theorem 5, we have*

$$R_k\left(\boldsymbol{\theta}_{0A}, \hat{\boldsymbol{\theta}}_{keB}(\boldsymbol{\theta}_{0A})\right)\hat{\rho}_{3k}\left(\boldsymbol{\theta}_{0A}, \hat{\boldsymbol{\theta}}_{keB}(\boldsymbol{\theta}_{0A})\right) \xrightarrow{D} \chi_{q_1}^2, k = 1, 3. \tag{A3}$$

Now, we prove the main results. In the interest of brevity, we will prove our main results when $\mathbf{X}$ is a scalar and $K(.)$ and $\bar{K}(.)$ are kernel functions with order 2. The extension from the univariate to the multivariate case is reasonably straightforward.

Lemma 1. *Suppose that $f(\boldsymbol{x})$ and $E\{\phi(X, U) \mid X = x\}$ are continuous and twice differentiable at x and $E|\phi(X, U)|^2 < \infty$. Then, as $n \to \infty$, we have*

$$\sup_{\mathbf{x}\in\mathcal{X}}\left|\frac{1}{nh}\sum_{i=1}^{n}K\left(\frac{\mathbf{X}_i - \mathbf{x}}{h}\right)\left(\frac{\mathbf{X}_i - \mathbf{x}}{h}\right)^k\phi(\mathbf{X}_i, U_i) - f(\mathbf{x})E(\phi(X, U) \mid X = \mathbf{x})\mu_k + h\nabla_{\mathbf{x}}\left[f(\mathbf{x})E(\phi(X, U)\,|X = \mathbf{x})\right]\mu_{k+1}\right| = O(\delta_n)\ a.s.,$$

where $\mathcal{X}$ is the support of $\mathbf{X}$, ∇_x denotes the first-order derivative with respect to $\mathbf{x}$, and $\delta_n = h^2 + (\frac{\log h^{-1}}{nh})^{1/2}$.

Proof. This lemma is similar to Lemma A.1 in the work of Zhou et al. (2008). □

Lemma 2. *For $k = 1, 2, 3$, assume that conditions $(C_1^k) - (C_2^k), (C_3)$, and (C_4^k) are satisfied, that is,*

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\hat{\mathbf{g}}_{\mathcal{M}_k}(Y_i, \mathbf{Z}_i, \boldsymbol{\theta}_0) \xrightarrow{D} N(0, \boldsymbol{\Sigma}).$$

Proof. The proof of this lemma is the same as that of Lemmas 8, 9, and 7 in the work of Chen et al. (2015) when $k = 1, 2, 3$, respectively. We omit it here for brevity. □

Lemma 3. *Assume that conditions* $(C_1^k) - (C_2^k), (C_3)$ *and* $(C_4^k) - (C_7^k)$ *are satisfied. Then, for* $k = 1, 2$, *we have*

$$\frac{1}{n}\sum_{i=1}^{n}\hat{\mathbf{g}}_{\mathcal{M}_k}(Y_i, \mathbf{Z}_i, \theta)\hat{\mathbf{g}}_{\mathcal{M}_k}^T(Y_i, \mathbf{Z}_i, \theta) \xrightarrow{P} \mathbf{V}_k(\theta).$$

Proof. The proof of this lemma is similar to that of Lemma A.6 of in the work of Zhou et al. (2008), and we omit the details here. □

Lemma 4. *Define*

$$Q_{k0}(\theta) =: -(E\mathbf{g}(Y, \mathbf{Z}, \theta))^T\mathbf{W}_{kn}E\mathbf{g}(Y, \mathbf{Z}, \theta) = -\left[E\tilde{\mathbf{g}}_{\mathcal{M}_k}(Y, \mathbf{Z}, \theta)\right]^T\mathbf{W}_{kn}E\tilde{\mathbf{g}}_{\mathcal{M}_k}(Y, \mathbf{Z}, \theta).$$

Assume that condition (C_2^k) *holds. Then, there exists a neighborhood* $\mathcal{N}$ *of* θ_0 *and a constant* $K_1 > 0$ *such that* $Q_{k0}(\theta) \leq -K_1\|\theta - \theta_0\|^2$ *for all* $\theta \in \mathcal{N}$ *and* $k = 1, 2, 3$, *where* $\|.\|$ *is the Euclidean norm.*

Proof. By a Taylor series expansion of $E\mathbf{m}(\mathbf{X}, \theta)$ about $\theta = \theta_0$, we have

$$\begin{aligned} Q_{k0}(\theta) &= -[E\mathbf{m}(\mathbf{X}, \theta)]^T\mathbf{W}_{kn}\,[E\mathbf{m}(\mathbf{X}, \theta)] \\ &= -\{\boldsymbol{\Gamma}(\theta - \theta_0) + o(\theta - \theta_0)\}^T\mathbf{W}_{kn}\,\{\boldsymbol{\Gamma}(\theta - \theta_0) + o(\theta - \theta_0)\}\,. \end{aligned}$$

Note that $-[\boldsymbol{\Gamma}(\theta - \theta_0)]^T\mathbf{W}_{kn}[\boldsymbol{\Gamma}(\theta - \theta_0)] \leq -K_1\|\theta - \theta_0\|^2$, where K_1 is the smallest eigenvalue of the matrix $\boldsymbol{\Gamma}^T\mathbf{W}_{kn}\boldsymbol{\Gamma}$. This completes the proof of Lemma 4. □

Lemma 5. *For* $k = 1, 2, 3$, *define* $\hat{Q}_{kn}(\theta) =: -\hat{\mathcal{Q}}_{kn}(\theta)$ *and*

$$Q_{kn}(\theta) =: -\left\{\frac{1}{n}\sum_{i=1}^{n}\tilde{\mathbf{g}}_{\mathcal{M}_k}(Y_i, \mathbf{Z}_i, \theta)\right\}^T\mathbf{W}_{kn}\left\{\frac{1}{n}\sum_{i=1}^{n}\tilde{\mathbf{g}}_{\mathcal{M}_k}(Y_i, \mathbf{Z}_i, \theta)\right\}.$$

Under conditions $(C_2^k), (C_4^k)$, *and* (C_5^k), *we have*

$$\begin{aligned} &\hat{Q}_{kn}(\theta) = Q_{k0}(\theta) + O_p\big(n^{-1/2}\|\theta - \theta_0\|\big) + o_p\big(\|\theta - \theta_0\|^2\big) + O_p(n^{-1}), k = 1, 3 \textit{ and} \\ &\hat{Q}_{2n}(\theta) = Q_{20}(\theta) + O_p\big(n^{-1/2}\|\theta - \theta_0\|\big) + o_p\big(\|\theta - \theta_0\|^2\big) + o_p(n^{-1}) \end{aligned}$$

uniformly in θ, *for* $\theta - \theta_0 = o_p(1)$.

Proof. By Lemmas 3-6 in the work of Chen et al. (2015), we know that

$$\frac{1}{n}\sum_{i=1}^{n}\left[\tilde{\mathbf{g}}_{\mathcal{M}_k}(Y_i, \mathbf{Z}_i, \theta) - \hat{\mathbf{g}}_{\mathcal{M}_k}(Y_i, \mathbf{Z}_i, \theta)\right] = O_p(n^{-1/2}), k = 1, 3, \tag{A4}$$

and

$$\frac{1}{n}\sum_{i=1}^{n}\left[\tilde{\mathbf{g}}_{\mathcal{M}_2}(Y_i, \mathbf{Z}_i, \theta) - \hat{\mathbf{g}}_{\mathcal{M}_2}(Y_i, \mathbf{Z}_i, \theta)\right] = o_p(n^{-1/2}). \tag{A5}$$

Thus, for $k = 1, 2, 3$, we have

$$\hat{Q}_{kn}(\boldsymbol{\theta}) - Q_{kn}(\boldsymbol{\theta})$$

$$= \left\{ \frac{1}{n}\sum_{i=1}^{n} \left[\tilde{\mathbf{g}}_{\mathcal{M}_k}(Y_i, \mathbf{Z}_i, \boldsymbol{\theta}) - \hat{\mathbf{g}}_{\mathcal{M}_k}(Y_i, \mathbf{Z}_i, \boldsymbol{\theta})\right] \right\}^T \mathbf{W}_{kn} \left\{ \frac{1}{n}\sum_{i=1}^{n} \left[\tilde{\mathbf{g}}_{\mathcal{M}_k}(Y_i, \mathbf{Z}_i, \boldsymbol{\theta}) - \hat{\mathbf{g}}_{\mathcal{M}_k}(Y_i, \mathbf{Z}_i, \boldsymbol{\theta})\right] \right\}$$

$$+ 2\left\{ \frac{1}{n}\sum_{i=1}^{n} \left[\tilde{\mathbf{g}}_{\mathcal{M}_k}(Y_i, \mathbf{Z}_i, \boldsymbol{\theta}) - \hat{\mathbf{g}}_{\mathcal{M}_k}(Y_i, \mathbf{Z}_i, \boldsymbol{\theta})\right] \right\}^T \mathbf{W}_{kn} \left\{ \frac{1}{n}\sum_{i=1}^{n} \hat{\mathbf{g}}_{\mathcal{M}_k}(Y_i, \mathbf{Z}_i, \boldsymbol{\theta}) \right\}.$$

Hence, $\hat{Q}_{kn}(\boldsymbol{\theta}) - Q_{kn}(\boldsymbol{\theta}) = O_p(n^{-1})$ for $k = 1, 3$ and $\hat{Q}_{2n}(\boldsymbol{\theta}) - Q_{2n}(\boldsymbol{\theta}) = o_p(n^{-1})$. By (C_5^k), we have

$$Q_{kn}(\boldsymbol{\theta}) - Q_{k0}(\boldsymbol{\theta})$$

$$= -\left\{ \frac{1}{n}\sum_{i=1}^{n} \tilde{\mathbf{g}}_{\mathcal{M}_k}(Y_i, \mathbf{Z}_i, \boldsymbol{\theta}) - E\tilde{\mathbf{g}}_{\mathcal{M}_k}(Y, \mathbf{Z}, \boldsymbol{\theta}) \right\}^T \mathbf{W}_{kn} \left\{ \frac{1}{n}\sum_{i=1}^{n} \tilde{\mathbf{g}}_{\mathcal{M}_k}(Y_i, \mathbf{Z}_i, \boldsymbol{\theta}) - E\tilde{\mathbf{g}}_{\mathcal{M}_k}(Y, \mathbf{Z}, \boldsymbol{\theta}) \right\}$$

$$- 2(E\tilde{\mathbf{g}}_{\mathcal{M}_k}(Y, \mathbf{Z}, \boldsymbol{\theta}))^T \mathbf{W}_{kn} \left\{ \frac{1}{n}\sum_{i=1}^{n} \tilde{\mathbf{g}}_{\mathcal{M}_k}(Y_i, \mathbf{Z}_i, \boldsymbol{\theta}) - E\tilde{\mathbf{g}}_{\mathcal{M}_k}(Y, \mathbf{Z}, \boldsymbol{\theta}) \right\}$$

$$= O_p(n^{-1}) - 2E\left[\nabla_{\boldsymbol{\theta}} m(X, \boldsymbol{\theta}_0)(\boldsymbol{\theta} - \boldsymbol{\theta}_0) + O\left(\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|^2\right)\right] O_p(n^{-1/2})$$

$$= O_p(n^{-1}) + O_p\big(n^{-1/2}\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|\big) + o_p\big(\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|^2\big), k = 1, 3.$$

Similarly, we can show that $Q_{2n}(\boldsymbol{\theta}) - Q_{20}(\boldsymbol{\theta}) = o_p(n^{-1}) + O_p(n^{-1/2}\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|) + o_p(\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|^2)$. These imply

$$\hat{Q}_{kn}(\boldsymbol{\theta}) = Q_{10}(\boldsymbol{\theta}) + O_p\big(n^{-1/2}\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|\big) + O_p(n^{-1}) + o_p\big(\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|^2\big), k = 1, 3 \text{ and}$$

$$\hat{Q}_{2n}(\boldsymbol{\theta}) = Q_{20}(\boldsymbol{\theta}) + O_p\big(n^{-1/2}\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|\big) + o_p(n^{-1}) + o_p\big(\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|^2\big)$$

uniformly in $\boldsymbol{\theta}$, for $\boldsymbol{\theta} - \boldsymbol{\theta}_0 = o(1)$. This completes the proof of Lemma 5. □

Proof of Theorem 1. Note that $\sup_{\boldsymbol{\theta}\in\Theta}|\frac{1}{n}\sum_{i=1}^{n}\{\hat{\mathbf{g}}_{\mathcal{M}_k}(Y_i, \mathbf{Z}_i, \boldsymbol{\theta}) - \tilde{\mathbf{g}}_{\mathcal{M}_k}(Y_i, \mathbf{Z}_i, \boldsymbol{\theta})\}| = o_p(1)$, for $k = 1, 2, 3$. Hence, we have

$$\sup_{\boldsymbol{\theta}\in\Theta}\left|\hat{Q}_{kn}(\boldsymbol{\theta}) - Q_{kn}(\boldsymbol{\theta})\right|$$

$$\le \sup_{\boldsymbol{\theta}\in\Theta}\left|\frac{1}{n}\sum_{i=1}^{n}\left\{\hat{\mathbf{g}}_{\mathcal{M}_k}(Y_i, \mathbf{Z}_i, \boldsymbol{\theta}) - \tilde{\mathbf{g}}_{\mathcal{M}_k}(Y_i, \mathbf{Z}_i, \boldsymbol{\theta})\right\}^T\right| \mathbf{W}_{kn} \sup_{\boldsymbol{\theta}\in\Theta}\left|\frac{1}{n}\sum_{i=1}^{n}\left\{\hat{\mathbf{g}}_{\mathcal{M}_k}(Y_i, \mathbf{Z}_i, \boldsymbol{\theta}) - \tilde{\mathbf{g}}_{\mathcal{M}_k}(Y_i, \mathbf{Z}_i, \boldsymbol{\theta})\right\}\right|$$

$$+ 2\sup_{\boldsymbol{\theta}\in\Theta}\left|\frac{1}{n}\sum_{i=1}^{n}\left\{\hat{\mathbf{g}}_{\mathcal{M}_k}(Y_i, \mathbf{Z}_i, \boldsymbol{\theta}) - \tilde{\mathbf{g}}_{\mathcal{M}_k}(Y_i, \mathbf{Z}_i, \boldsymbol{\theta})\right\}^T\right| \mathbf{W}_{kn} \sup_{\boldsymbol{\theta}\in\Theta}\left|\frac{1}{n}\sum_{i=1}^{n}\tilde{\mathbf{g}}_{\mathcal{M}_k}(Y_i, \mathbf{Z}_i, \boldsymbol{\theta})\right|$$

$$= o_p(1).$$

On the other hand, by (C_5^k), we have

$$
\begin{aligned}
&\sup_{\theta\in\Theta}|Q_{kn}(\theta)-Q_{k0}(\theta)| \\
&\quad\le \sup_{\theta\in\Theta}\left|\frac{1}{n}\sum_{i=1}^{n}\left\{\tilde{\mathbf{g}}_{\mathcal{M}_k}(Y_i,\mathbf{Z}_i,\theta)-E\tilde{\mathbf{g}}(Y,\mathbf{Z},\theta)\right\}^T\right|\mathbf{W}_{kn}\sup_{\theta\in\Theta}\left|\frac{1}{n}\sum_{i=1}^{n}\tilde{\mathbf{g}}_{\mathcal{M}_k}(Y_i,\mathbf{Z}_i,\theta)-E\tilde{\mathbf{g}}_{\mathcal{M}_k}(Y,\mathbf{Z},\theta)\right| \\
&\qquad+\sup_{\theta\in\Theta}2\left|(E\tilde{\mathbf{g}}_{\mathcal{M}_k}(Y,\mathbf{Z},\theta))^T\right|\mathbf{W}_{kn}\sup_{\theta\in\Theta}\left|\frac{1}{n}\sum_{i=1}^{n}\tilde{\mathbf{g}}_{\mathcal{M}_k}(Y_i,\mathbf{Z}_i,\theta)-E\tilde{\mathbf{g}}_{\mathcal{M}_k}(Y,\mathbf{Z},\theta)\right| \\
&\quad=o_p(1).
\end{aligned}
$$

This leads to $\sup_{\theta\in\Theta}|\hat{Q}_{kn}(\theta)-Q_{k0}(\theta)|=o_p(1)$. By (C_2^k), we see that $Q_{k0}(\theta)$ is continuous in θ. Using these results and Theorem 2.1 in the work of Newey and McFadden (1994), we can establish the consistency of the GMM estimator $\hat{\theta}_k$. □

Proof of Theorem 2. By Lemmas 4 and 5, the conditions of Theorem 1 in the work of Sherman (1993) can be shown to hold. Applying methods similar to the Proof of Theorem 1 of Sherman (1993), we can show that $|\hat{\theta}_k-\theta_0|=O_p(n^{-1/2}), k=1,2,3$. Hence, we focus on the $O_p(n^{-1/2})$ neighborhood of θ_0. Note that, for $k=1,2,3$, we have

$$
\begin{aligned}
\hat{Q}_{kn}(\theta)=&-\left\{\left[\frac{1}{n}\sum_{i=1}^{n}\hat{\mathbf{g}}_{\mathcal{M}_k}(Y_i,\mathbf{Z}_i,\theta)-E\tilde{\mathbf{g}}_{\mathcal{M}_k}(Y,\mathbf{Z},\theta)-\frac{1}{n}\sum_{i=1}^{n}\hat{\mathbf{g}}_{\mathcal{M}_k}(Y_i,\mathbf{Z}_i,\theta_0)\right]\right. \\
&\left.+\left[E\tilde{\mathbf{g}}_{\mathcal{M}_k}(Y,\mathbf{Z},\theta)+\frac{1}{n}\sum_{i=1}^{n}\hat{\mathbf{g}}_{\mathcal{M}_k}(Y_i,\mathbf{Z}_i,\theta_0)\right]\right\}^T\mathbf{W}_{kn} \\
&\times\left\{\left[\frac{1}{n}\sum_{i=1}^{n}\hat{\mathbf{g}}_{\mathcal{M}_k}(Y_i,\mathbf{Z}_i,\theta)-E\tilde{\mathbf{g}}_{\mathcal{M}_k}(Y,\mathbf{Z},\theta)-\frac{1}{n}\sum_{i=1}^{n}\hat{\mathbf{g}}_{\mathcal{M}_k}(Y_i,\mathbf{Z}_i,\theta_0)\right]\right. \\
&\left.+\left[E\tilde{\mathbf{g}}_{\mathcal{M}_k}(Y,\mathbf{Z},\theta)+\frac{1}{n}\sum_{i=1}^{n}\hat{\mathbf{g}}_{\mathcal{M}_k}(Y_i,\mathbf{Z}_i,\theta_0)\right]\right\} \\
=:&-\left\{D_{k1}^T\mathbf{W}_{kn}D_{k1}+2D_{k2}^T\mathbf{W}_{kn}D_{k1}+D_{k2}^T\mathbf{W}_{kn}D_{k2}\right\},
\end{aligned}
$$

where $D_{k1}=n^{-1}\sum_{i=1}^{n}\hat{\mathbf{g}}_{\mathcal{M}_k}(Y_i,\mathbf{Z}_i,\theta)-E\tilde{\mathbf{g}}_{\mathcal{M}_k}(Y,\mathbf{Z},\theta)-n^{-1}\sum_{i=1}^{n}\hat{\mathbf{g}}_{\mathcal{M}_k}(Y_i,\mathbf{Z}_i,\theta_0)$ and $D_{k2}=E\tilde{\mathbf{g}}_{\mathcal{M}_k}(Y,\mathbf{Z},\theta)+n^{-1}\sum_{i=1}^{n}\hat{\mathbf{g}}_{\mathcal{M}_k}(Y_i,\mathbf{Z}_i,\theta_0)$.

Applying Lemma 1 and (C_6^k) and noting that $nh^4\to0$, $\sqrt{n}h^2\to0$, $n\bar{h}^4\to0$, $\sqrt{n}\bar{h}^2\to0$, and $E\tilde{\mathbf{g}}_{\mathcal{M}_k}(Y,\mathbf{Z},\theta)=E\mathbf{g}(Y,\mathbf{Z},\theta)$, we can show that

$$
D_{k1}=\frac{1}{n}\sum_{i=1}^{n}\tilde{\mathbf{g}}_{\mathcal{M}_k}(Y_i,\mathbf{Z}_i,\theta)-E\tilde{\mathbf{g}}_{\mathcal{M}_k}(Y,\mathbf{Z},\theta)-\frac{1}{n}\sum_{i=1}^{n}\tilde{\mathbf{g}}_{\mathcal{M}_k}(Y_i,\mathbf{Z}_i,\theta_0)+O(h^2)=o_p(n^{-1/2}).
$$

Then, we have $D_{k1}^T\mathbf{W}_{kn}D_{k1}=o_p(n^{-1})$, and

$$
\begin{aligned}
D_{k2}^T\mathbf{W}_{kn}D_{k1}&=\left[E\mathbf{m}(\mathbf{X},\theta)+\frac{1}{n}\sum_{i=1}^{n}\hat{\mathbf{g}}_{\mathcal{M}_k}(Y_i,\mathbf{Z}_i,\theta_0)\right]^T\mathbf{W}_{kn}\left[o_p(n^{-1/2})\right] \\
&=\left[\boldsymbol{\Gamma}(\theta-\theta_0)+o(n^{-1/2})+O_p(n^{-1/2})\right]^T\mathbf{W}_{kn}\left[o_p(n^{-1/2})\right] \\
&=o_p(n^{-1}).
\end{aligned}
$$

Also, for $k = 1, 2, 3$, we have

$$D_{k2}^T \mathbf{W}_{kn} D_{2k} = \left[E\mathbf{m}(\mathbf{X}, \boldsymbol{\theta}) + \frac{1}{n}\sum_{i=1}^{n} \hat{\mathbf{g}}_{\mathcal{M}_k}(Y_i, \mathbf{Z}_i, \boldsymbol{\theta}_0) \right]^T \mathbf{W}_{kn} \left[E\mathbf{m}(\mathbf{X}, \boldsymbol{\theta}) + \frac{1}{n}\sum_{i=1}^{n} \hat{\mathbf{g}}_{\mathcal{M}_k}(Y_i, \mathbf{Z}_i, \boldsymbol{\theta}_0) \right]$$

$$= \left[\boldsymbol{\Gamma}(\boldsymbol{\theta} - \boldsymbol{\theta}_0) + \frac{1}{n}\sum_{i=1}^{n} \hat{\mathbf{g}}_{\mathcal{M}_k}(Y_i, \mathbf{Z}_i, \boldsymbol{\theta}_0) + o\left(n^{-\frac{1}{2}}\right) \right]^T$$

$$\mathbf{W}_{kn} \left[\boldsymbol{\Gamma}(\boldsymbol{\theta} - \boldsymbol{\theta}_0) + \frac{1}{n}\sum_{i=1}^{n} \hat{\mathbf{g}}_{\mathcal{M}_k}(Y_i, \mathbf{Z}_i, \boldsymbol{\theta}_0) + o\left(n^{-\frac{1}{2}}\right) \right].$$

Hence, for $k = 1, 2, 3$, we obtain

$$\hat{Q}_{kn}(\boldsymbol{\theta}) = -\left[\boldsymbol{\Gamma}(\boldsymbol{\theta} - \boldsymbol{\theta}_0) + \frac{1}{n}\sum_{i=1}^{n} \hat{\mathbf{g}}_{\mathcal{M}_k}(Y_i, \mathbf{Z}_i, \boldsymbol{\theta}_0) \right]^T$$

$$\mathbf{W}_{kn} \left[\boldsymbol{\Gamma}(\boldsymbol{\theta} - \boldsymbol{\theta}_0) + \frac{1}{n}\sum_{i=1}^{n} \hat{\mathbf{g}}_{\mathcal{M}_k}(Y_i, \mathbf{Z}_i, \boldsymbol{\theta}_0) \right] + o_p(n^{-1}).$$

As expected, $\hat{\boldsymbol{\theta}}_k = \arg\max_{\boldsymbol{\theta}\in\Theta} \hat{Q}_{kn}(\boldsymbol{\theta})$ satisfies

$$\mathbf{W}_{kn} \left[\boldsymbol{\Gamma}(\hat{\boldsymbol{\theta}}_k - \boldsymbol{\theta}_0) + \frac{1}{n}\sum_{i=1}^{n} \hat{\mathbf{g}}_{\mathcal{M}_k}(Y_i, \mathbf{Z}_i, \boldsymbol{\theta}_0) \right] + o_p(n^{-1/2}) = 0,$$

which implies

$$\left(\hat{\boldsymbol{\theta}}_k - \boldsymbol{\theta}_0\right) = -\left(\boldsymbol{\Gamma}^T \mathbf{W}_{kn} \boldsymbol{\Gamma}\right)^{-1} \boldsymbol{\Gamma}^T \mathbf{W}_{kn} \frac{1}{n}\sum_{i=1}^{n} \hat{\mathbf{g}}_{\mathcal{M}_k}(Y_i, \mathbf{Z}_i, \boldsymbol{\theta}_0) + o_p(n^{-1/2}). \tag{A6}$$

Using Lemma 2, we have

$$\sqrt{n}\left(\hat{\boldsymbol{\theta}}_k - \boldsymbol{\theta}_0\right) \xrightarrow{D} N\left(0, \left(\boldsymbol{\Gamma}^T \mathbf{W}_{kn} \boldsymbol{\Gamma}\right)^{-1} \boldsymbol{\Gamma}^T \mathbf{W}_{kn} \boldsymbol{\Sigma} \mathbf{W}_{kn} \boldsymbol{\Gamma} \left(\boldsymbol{\Gamma}^T \mathbf{W}_{kn} \boldsymbol{\Gamma}\right)^{-1}\right),$$

for $k = 1, 2, 3$. This completes the proof of Theorem 2. □

Proof of Corollary 1. The proof of Corollary 1 can be obtained directly from the proof of Theorem 2.

We now prove the results pertaining to the properties of the MEL estimator. Define

$$\hat{G}_{kn}(\boldsymbol{\theta}) =: -\frac{1}{n}\sum_{i=1}^{n} \log\left(1 + \lambda_k^T \hat{\mathbf{g}}_{\mathcal{M}_k}(Y_i, \mathbf{Z}_i, \boldsymbol{\theta})\right) \tag{A7}$$

and

$$G_{k0}(\boldsymbol{\theta}) =: -E\left\{\log\left(1 + \boldsymbol{\zeta}_k^T \tilde{\mathbf{g}}_{\mathcal{M}_k}(Y, \mathbf{Z}, \boldsymbol{\theta})\right)\right\}, \tag{A8}$$

where λ_k satisfies (8) and $\boldsymbol{\zeta}_k$ satisfies

$$E\left\{ \frac{\tilde{\mathbf{g}}_{\mathcal{M}_k}(Y, X, \boldsymbol{\theta})}{1 + \boldsymbol{\zeta}_k(\boldsymbol{\theta})\tilde{\mathbf{g}}_{\mathcal{M}_k}(Y, X, \boldsymbol{\theta})} \right\} = 0, \tag{A9}$$

for $k = 1, 2, 3$. □

Lemma 6. *Assume that (C_2^k), (C_3), (C_4^k), and (C_5^k) hold. Then, for $\boldsymbol{\theta} \in \Theta$, we have*

$$\lambda_k(\boldsymbol{\theta}) = O_p(n^{-1/2}) \text{ and } \boldsymbol{\zeta}_k(\boldsymbol{\theta}) = O(n^{-1/2}), \tag{A10}$$

$$\lambda_k(\boldsymbol{\theta}) = \mathbf{V}_k^{-1}(\boldsymbol{\theta})\frac{1}{n}\sum_{i=1}^{n}\hat{\mathbf{g}}_{\mathcal{M}_k}(Y_i,\mathbf{Z}_i,\boldsymbol{\theta}) + o_p(n^{-1/2}), \tag{A11}$$

and

$$\boldsymbol{\zeta}_k(\boldsymbol{\theta}) = \mathbf{V}_k^{-1}(\boldsymbol{\theta})E\tilde{\mathbf{g}}_{\mathcal{M}_k}(Y,\mathbf{Z},\boldsymbol{\theta}) + o(n^{-1/2}), \tag{A12}$$

for $k = 1, 2, 3$.

Proof. The proof may be constructed by following the proof of Theorem 3.2 in the work of Owen (2001) and combining Lemma 3 of Owen (1990) with conditions (C_2^k), (C_3), (C_4^k), and (C_5^k), and the uniform integrability of $\tilde{\mathbf{g}}_{\mathcal{M}_k}(Y,\mathbf{Z},\boldsymbol{\theta})$. The uniform integrability property can be easily shown by noting the existence of the second moment of $\tilde{\mathbf{g}}_{\mathcal{M}_k}(Y,\mathbf{Z},\boldsymbol{\theta})$ (see condition (C_3)). □

Lemma 7. *Under conditions* (C_2^k), (C_3), (C_5^k), (C_7^k), *and* (C_8^k), $\hat{\boldsymbol{\theta}}_{ke} = \arg\max_{\boldsymbol{\theta}}\hat{G}_{kn}(\boldsymbol{\theta})$ *and* $\boldsymbol{\theta}_0 = \arg\max_{\boldsymbol{\theta}} G_{k0}(\boldsymbol{\theta})$, *for* $k = 1, 2$, *and* 3.

Proof. By Lemma 6, we know that $\boldsymbol{\zeta}_k(\boldsymbol{\theta})$ is continuous in $\boldsymbol{\theta}$ and $\boldsymbol{\zeta}_k(\boldsymbol{\theta}_0) = 0$. The proof of the remaining part of this lemma is similar to that of Lemma 1 in the work of Lopez et al. (2009). □

Lemma 8. *Assume that* (C_2^k), (C_3), $(C_4^k)-(C_5^k)$, *and* (C_7^k) *hold. Then, there exists a neighborhood* $\mathcal{N}$ *of* $\boldsymbol{\theta}_0$ *and a constant* $K_{k2} > 0$ *such that for all* $\boldsymbol{\theta} \in \mathcal{N}$, $G_{k0}(\boldsymbol{\theta}) \leq -K_{k2}\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|^2$, $k = 1, 2, 3$.

Proof. From Lemma 6, we have $\boldsymbol{\zeta}_k(\boldsymbol{\theta}) = O(n^{-1/2})$ in a neighborhood $\mathcal{N}$ of $\boldsymbol{\theta}_0$. By Lemma 11.2 in the work of Owen (2001), we have $|\boldsymbol{\zeta}_k(\boldsymbol{\theta})^T\tilde{\mathbf{g}}_{\mathcal{M}_k}(Y,\mathbf{Z},\boldsymbol{\theta})| = o_p(1)$. Along the lines of the proof of Lemma 3 in the work of Lopez et al. (2009), we have $G_{k0}(\boldsymbol{\theta}) \leq -K_{k2}\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|^2$, where K_{k2} is the smallest eigenvalue of $\boldsymbol{\Gamma}^T\mathbf{V}_k(\boldsymbol{\theta})^{-1}\boldsymbol{\Gamma}$. □

Lemma 9. *Assume that* (C_2^k), (C_3), (C_4^k), *and* (C_5^k) *hold. Then, we have*

$$\hat{G}_{kn}(\boldsymbol{\theta}) = G_{10}(\boldsymbol{\theta}) + O_p\big(n^{-1/2}\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|\big) + o_p\big(\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|^2\big) + O_p(n^{-1}), k = 1, 3 \text{ and}$$
$$\hat{G}_{2n}(\boldsymbol{\theta}) = G_{20}(\boldsymbol{\theta}) + O_p\big(n^{-1/2}\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|\big) + o_p\big(\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|^2\big) + o_p(n^{-1})$$

uniformly in $\boldsymbol{\theta}$, *with* $\boldsymbol{\theta}-\boldsymbol{\theta}_0 = o_p(1)$.

Proof. Note that, for $k = 1, 2, 3$, we have

$$\begin{aligned}\hat{G}_{kn}(\boldsymbol{\theta}) - G_{k0}(\boldsymbol{\theta}) = &-\left\{\frac{1}{n}\sum_{i=1}^{n}\left[\log\big(1+\lambda_k^T\hat{\mathbf{g}}_{\mathcal{M}_k}(Y_i,\mathbf{Z}_i,\boldsymbol{\theta})\big) - \log\big(1+\boldsymbol{\zeta}_k^T\tilde{\mathbf{g}}_{\mathcal{M}_k}(Y_i,\mathbf{Z}_i,\boldsymbol{\theta})\big)\right]\right\}\\ &-\left\{\frac{1}{n}\sum_{i=1}^{n}\log\big(1+\boldsymbol{\zeta}_k^T\tilde{\mathbf{g}}_{\mathcal{M}_k}(Y_i,\mathbf{Z}_i,\boldsymbol{\theta})\big) - E\big(\log\big(1+\boldsymbol{\zeta}_k^T\tilde{\mathbf{g}}_{\mathcal{M}_k}(Y,\mathbf{Z},\boldsymbol{\theta})\big)\big)\right\}\\ =:&\ I_{k6} + I_{k7}.\end{aligned}$$

We can write

$$\begin{aligned}I_{k6} = &-\left\{\frac{1}{n}\sum_{i=1}^{n}\left[\log\big(1+\lambda_k^T\hat{\mathbf{g}}_{\mathcal{M}_k}(Y_i,\mathbf{Z}_i,\boldsymbol{\theta})\big) - \log\big(1+\boldsymbol{\zeta}_k^T\hat{\mathbf{g}}_{\mathcal{M}_k}(Y_i,\mathbf{Z}_i,\boldsymbol{\theta})\big)\right]\right\}\\ &-\left\{\frac{1}{n}\sum_{i=1}^{n}\left[\log\big(1+\boldsymbol{\zeta}_k^T\hat{\mathbf{g}}_{\mathcal{M}_k}(Y_i,\mathbf{Z}_i,\boldsymbol{\theta})\big) - \log\big(1+\boldsymbol{\zeta}_k^T\tilde{\mathbf{g}}_{\mathcal{M}_k}(Y_i,\mathbf{Z}_i,\boldsymbol{\theta})\big)\right]\right\}\\ =&\ I_{k61} + I_{k62}.\end{aligned}$$

Also, we obtain

$$I_{k61} = -\frac{1}{n}\sum_{i=1}^{n}\{(\lambda_k - \zeta_k)^T \frac{\hat{\mathbf{g}}_{\mathcal{M}_k}(Y_i, \mathbf{Z}_i, \theta)}{1 + \zeta_k^T \hat{\mathbf{g}}_{\mathcal{M}_k(Y_i, \mathbf{Z}_i, \theta)}} + \frac{1}{2}\frac{\left[(\lambda_k - \zeta_k)^T \hat{\mathbf{g}}_{\mathcal{M}_k}(Y_i, \mathbf{Z}_i, \theta)\right]^2}{\left(1 + \eta_{k1}^T \hat{\mathbf{g}}_{\mathcal{M}_k}(Y_i, \mathbf{Z}_i, \theta)\right)^2}\},$$

where η_{k1} lies between λ_k and ζ_k, for $k = 1, 2, 3$. By Lemma 11.2 of Owen (2001), we have $\max_{1\le i\le n}|\zeta_k^T \hat{\mathbf{g}}_{\mathcal{M}_k}(Y_i, \mathbf{Z}_i, \theta)| = o_p(1)$ and $\max_{1\le i\le n}|\eta_{k1}^T \hat{\mathbf{g}}_{\mathcal{M}_k}(Y_i, \mathbf{Z}_i, \theta)| = o_p(1)$. By Lemmas 3 and 6, $I_{k61} = O_p(n^{-1})$. Similarly, we have

$$I_{k62} = -\frac{1}{n}\sum_{i=1}^{n}\left\{\frac{\zeta_k^T\left(\hat{\mathbf{g}}_{\mathcal{M}_k}(Y_i, \mathbf{Z}_i, \theta) - \tilde{\mathbf{g}}_{\mathcal{M}_k}(Y_i, \mathbf{Z}_i, \theta)\right)}{1 + \zeta_k^T \tilde{\mathbf{g}}_{\mathcal{M}_k(Y_i, \mathbf{Z}_i, \theta)}} + \frac{1}{2}\frac{\left[\zeta_k^T\left(\hat{\mathbf{g}}_{\mathcal{M}_k}(Y_i, \mathbf{Z}_i, \theta) - \tilde{\mathbf{g}}_{\mathcal{M}_k}(Y_i, \mathbf{Z}_i, \theta)\right)\right]^2)}{\left(1 + \eta_{k2}^T \tilde{\mathbf{g}}_{\mathcal{M}_k}(Y_i, \mathbf{Z}_i, \theta)\right)^2}\right\},$$

where η_{k2} lies between 0 and ζ_k, $k = 1, 2, 3$. Note that $\frac{1}{n}\sum_{i=1}^{n}[\tilde{\mathbf{g}}_{\mathcal{M}_k}(Y_i, \mathbf{Z}_i, \theta) - \hat{\mathbf{g}}_{\mathcal{M}_k}(Y_i, \mathbf{Z}_i, \theta)] = O_p(n^{-1/2})$, $k = 1, 3$ and $\frac{1}{n}\sum_{i=1}^{n}[\tilde{\mathbf{g}}_{\mathcal{M}_2}(Y_i, \mathbf{Z}_i, \theta) - \hat{\mathbf{g}}_{\mathcal{M}_2}(Y_i, \mathbf{Z}_i, \theta)] = o_p(n^{-1/2})$. Thus, we obtain $I_{k62} = O_p(n^{-1})$, $k = 1, 3$ and $I_{262} = o_p(n^{-1})$, respectively. Now, write

$$\begin{aligned} I_{k7} = -&\left\{\frac{1}{n}\sum_{i=1}^{n}\left[\zeta_k^T \tilde{\mathbf{g}}_{\mathcal{M}_k}(Y_i, \mathbf{Z}_i, \theta) - E\left(\zeta_k^T \tilde{\mathbf{g}}_{\mathcal{M}_k}(Y, \mathbf{Z}, \theta)\right)\right]\right. \\ &-\frac{1}{2n}\sum_{i=1}^{n}\left[\left(\zeta_k^T \tilde{\mathbf{g}}_{\mathcal{M}_k}(Y_i, \mathbf{Z}_i, \theta)\right)^2 - E\left(\zeta_k^T \tilde{\mathbf{g}}_{\mathcal{M}_k}(Y, \mathbf{Z}, \theta)\right)^2\right] \\ &\left.+\frac{1}{3n}\sum_{i=1}^{n}\left[\frac{\left(\zeta_k^T \tilde{\mathbf{g}}_{\mathcal{M}_k}(Y_i, \mathbf{Z}_i, \theta)\right)^3}{\left(1 + \eta_{3k}^T \tilde{\mathbf{g}}_{\mathcal{M}_k}(Y_i, \mathbf{Z}_i, \theta)\right)^3} - E\left(\frac{\left(\zeta_k^T \tilde{\mathbf{g}}_{\mathcal{M}_k}(Y, \mathbf{Z}, \theta)\right)^3}{\left(1 + \eta_{3k}^T \tilde{\mathbf{g}}_{\mathcal{M}_k}(Y, \mathbf{Z}, \theta)\right)^3}\right)\right]\right\} \\ =:&\ I_{k71} + I_{k72} + I_{k73}, \end{aligned}$$

where η_{k3} lies between 0 and ζ_k, $k = 1, 2, 3$.

By Lemma 7 and (C_2^k), (C_5^k), and (C_8^k), we have

$$\begin{aligned} I_{k71} &= \left[\mathbf{V}_k^{-1}\Gamma(\theta - \theta_0) + O\left(\|\theta - \theta_0\|^2\right) + o_p(n^{-1/2})\right]^T \left[\frac{1}{n}\sum_{i=1}^{n}\tilde{\mathbf{g}}_{\mathcal{M}_k}(Y_i, \mathbf{Z}_i, \theta) - E\left(\tilde{\mathbf{g}}_{\mathcal{M}_k}(Y, \mathbf{Z}, \theta)\right)\right] \\ &= O_p\left(n^{-1/2}\|\theta - \theta_0\|\right) + o_p\left(\|\theta - \theta_0\|^2\right), \end{aligned}$$

and $I_{k72} = o_p(\|\theta - \theta_0\|^2)$. Note that $\frac{1}{n}\sum_{i=1}^{n}\tilde{\mathbf{g}}_{\mathcal{M}_k}(Y_i, \mathbf{Z}_i, \theta)^{\otimes 2}\tilde{\mathbf{g}}_{\mathcal{M}_k}(Y_i, \mathbf{Z}_i, \theta) = o_p(n^{1/2})$ by Lemma 3 in Owen (1990). Then, by (C_2^k), we have

$$\begin{aligned} |I_{k73}| \le \|\zeta_k\|^3 &\left\{\left\|\frac{1}{n}\sum_{i=1}^{n}\tilde{\mathbf{g}}_{\mathcal{M}_k}(Y_i, \mathbf{Z}_i, \theta)\tilde{\mathbf{g}}_{\mathcal{M}_k}^T(Y_i, \mathbf{Z}_i, \theta)\tilde{\mathbf{g}}_{\mathcal{M}_k}(Y_i, \mathbf{Z}_i, \theta)\right\| + E\left\|\tilde{\mathbf{g}}_{\mathcal{M}_k}(Y, \mathbf{Z}, \theta)\right\|^3\right\} \\ &\times \frac{1}{1 - o_p(1)} = o_p(n^{-1}). \end{aligned}$$

This completes the proof of Lemma 9. □

Proof of Theorem 3. Using Lemma 11.2 of Owen (2001) and Lemma 6, we can show that $\max_{1\le i\le n}|\zeta_k(\theta)^T \tilde{\mathbf{g}}_{\mathcal{M}_k}(Y_i, \mathbf{Z}_i, \theta_i)| = o(1)$ and $\max_{1\le i\le n}\ |\lambda_k(\theta)^T \tilde{\mathbf{g}}_{\mathcal{M}_k}(Y_i, \mathbf{Z}_i, \theta)| = o_p(1)$.

We then obtain

$$\hat{G}_{kn}(\theta) = -\frac{1}{n}\sum_{i=1}^{n}\left[\lambda_k(\theta)^T\hat{\mathbf{g}}_{\mathcal{M}_k}(Y_i,\mathbf{Z}_i,\theta) - \frac{1}{2}\left\{\lambda_k(\theta)^T\hat{\mathbf{g}}_{\mathcal{M}_k}(Y_i,\mathbf{Z}_i,\theta)\right\}^2\right] + R_{kn1}$$

and

$$G_{k0}(\theta) = -E\left[\zeta_k(\theta)^T\tilde{\mathbf{g}}_{\mathcal{M}_k}(Y,\mathbf{Z},\theta) - \frac{1}{2}\left\{\zeta_k(\theta)^T\tilde{\mathbf{g}}_{\mathcal{M}_k}(Y,\mathbf{Z},\theta)\right\}^2\right] + R_{kn2},$$

where

$$R_{kn1} = -\frac{1}{3n}\sum_{i=1}^{n}\frac{\left(\lambda_k(\theta)^T\hat{\mathbf{g}}_{\mathcal{M}_k}(Y_i,\mathbf{Z}_i,\theta)\right)^3}{\left(1+\lambda_k^{*T}\hat{\mathbf{g}}_{\mathcal{M}_k}(Y_i,\mathbf{Z}_i,\theta)\right)^3},\quad R_{kn2} = -\frac{1}{3}E\left\{\frac{\left(\zeta_k(\theta)^T\tilde{\mathbf{g}}_{\mathcal{M}_k}(Y,\mathbf{Z},\theta)\right)^3}{\left(1+\zeta_k^{*T}\tilde{\mathbf{g}}_{\mathcal{M}_k}(Y,\mathbf{Z},\theta)\right)^3}\right\},$$

λ_k^* lies between 0 and λ_k, and ζ_k^* lies between 0 and ζ_k. By Lemma 6, Lemma 3 in Owen (1990), and (C_2^k), it follows that $R_{kn1} = o_p(n^{-1})$ and $R_{kn2} = o(n^{-1})$. Thus, we have

$$\hat{G}_{kn}(\theta) = -\frac{1}{2}\left\{\frac{1}{n}\sum_{i=1}^{n}\hat{\mathbf{g}}_{\mathcal{M}_k}(Y_i,\mathbf{Z}_i,\theta)\right\}^T\mathbf{V}_k^{-1}(\theta)\left\{\frac{1}{n}\sum_{i=1}^{n}\hat{\mathbf{g}}_{\mathcal{M}_k}(Y_i,\mathbf{Z}_i,\theta)\right\} + o_p(n^{-1})$$

and $$G_{k0}(\theta) = -\frac{1}{2}\left[E\tilde{\mathbf{g}}_{\mathcal{M}_k}(Y,\mathbf{Z},\theta)\right]^T\mathbf{V}_k^{-1}(\theta)\left[E\tilde{\mathbf{g}}_{\mathcal{M}_k}(Y,\mathbf{Z},\theta)\right] + o(n^{-1}).$$

This implies that the EL method is almost identical to the GMM method that uses $\mathbf{V}_k^{-1}(\theta)$ as the weight matrix, because the terms of orders $o_p(n^{-1})$ and $o(n^{-1})$ have only negligible effects on the estimator of the parameter. The remainder of the proof is similar to that of Theorem 1, and we omit it here for brevity. □

Proof of Theorem 4. Note that by Lemmas 8 and 9 and Theorem 1 of Sherman (1993), we have $\hat{\theta}_{ke} - \theta_0 = O_p(n^{-1/2})$, $k = 1, 2, 3$. Thus, it makes sense to consider the $O_p(n^{-1/2})$ neighborhood of θ_0. Note that

$$\begin{aligned}\hat{G}_{kn}(\theta) &= -\frac{1}{n}\sum_{i=1}^{n}\lambda_k(\theta)^T\hat{\mathbf{g}}_{\mathcal{M}_k}(Y_i,\mathbf{Z}_i,\theta) + \frac{1}{2n}\sum_{i=1}^{n}\left(\lambda_k(\theta)^T\hat{\mathbf{g}}_{\mathcal{M}_k}(Y_i,\mathbf{Z}_i,\theta)\right)^2\\ &\quad -\frac{1}{3n}\sum_{i=1}^{n}\frac{\left(\lambda_k(\theta)^T\hat{\mathbf{g}}_{\mathcal{M}_k}(Y_i,\mathbf{Z}_i,\theta)\right)^3}{\left(1+\eta_{k4}^T\hat{\mathbf{g}}_{\mathcal{M}_k}(Y_i,\mathbf{Z}_i,\theta)\right)^3}\\ &=: I_{k8} + I_{k9} + I_{k10},\end{aligned}$$

where η_{k4} lies between 0 and λ_k.

By (A11), (A12), (C_6^k), and Lemma 6 and noting that D_{k1} defined in the proof of Theorem 2 is $o_p(n^{-1/2})$, we obtain

$$\begin{aligned}\lambda_k(\theta) &= \zeta_k(\theta) + \left(\zeta_k(\theta) - \lambda_k(\theta)\right)\\ &= \mathbf{V}_k^{-1}\Gamma(\theta-\theta_0) + \mathbf{V}_k^{-1}\left[\frac{1}{n}\sum_{i=1}^{n}\hat{\mathbf{g}}_{\mathcal{M}_k}(Y_i,\mathbf{Z}_i,\theta) - E\tilde{\mathbf{g}}_{\mathcal{M}_k}(Y,\mathbf{Z},\theta)\right.\\ &\quad\left. -\frac{1}{n}\sum_{i=1}^{n}\hat{\mathbf{g}}_{\mathcal{M}_k}(Y_i,\mathbf{Z}_i,\theta_0) + \frac{1}{n}\sum_{i=1}^{n}\hat{\mathbf{g}}_{\mathcal{M}_k}(Y_i,\mathbf{Z}_i,\theta_0)\right] + o_p(n^{-1/2})\\ &= \mathbf{V}_k^{-1}\left[\Gamma(\theta-\theta_0) + \frac{1}{n}\sum_{i=1}^{n}\hat{\mathbf{g}}_{\mathcal{M}_k}(Y_i,\mathbf{Z}_i,\theta_0)\right] + o_p(n^{-1/2}).\end{aligned} \tag{A13}$$

By (A13), $E\hat{\mathbf{g}}_{\mathcal{M}_k}(Y,\mathbf{Z},\boldsymbol{\theta}) = E\tilde{\mathbf{g}}_{\mathcal{M}_k}(Y,\mathbf{Z},\boldsymbol{\theta}) + o(n^{1/2})$. Write $I_{k8} =: I_{k81} + I_{k82}$. It follows from Lemma 6 that

$$
\begin{aligned}
I_{k81} &= -\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\lambda}_k(\boldsymbol{\theta})^T\left[\hat{\mathbf{g}}_{\mathcal{M}_k}(Y_i,\mathbf{Z}_i,\boldsymbol{\theta}) - \hat{\mathbf{g}}_{\mathcal{M}_k}(Y_i,\mathbf{Z}_i,\boldsymbol{\theta}_0)\right]\\
&= -\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\lambda}_k(\boldsymbol{\theta})^T\left\{\left[\hat{\mathbf{g}}_{\mathcal{M}_k}(Y_i,\mathbf{Z}_i,\boldsymbol{\theta}) - E\hat{\mathbf{g}}_{\mathcal{M}_k}(Y,\mathbf{Z},\boldsymbol{\theta}) - \hat{\mathbf{g}}_{\mathcal{M}_k}(Y_i,\mathbf{Z}_i,\boldsymbol{\theta}_0)\right] + E\hat{\mathbf{g}}_{\mathcal{M}_k}(Y,\mathbf{Z},\boldsymbol{\theta})\right\}\\
&= -\left\{\mathbf{V}_k^{-1}\left[\boldsymbol{\Gamma}(\boldsymbol{\theta}-\boldsymbol{\theta}_0) + \frac{1}{n}\sum_{i=1}^{n}\hat{\mathbf{g}}_{\mathcal{M}_k}(Y_i,\mathbf{Z}_i,\boldsymbol{\theta}_0)\right] + o_p(n^{-1/2})\right\}^T\left[\boldsymbol{\Gamma}(\boldsymbol{\theta}-\boldsymbol{\theta}_0) + o_p(n^{-1/2})\right]\\
&= -(\boldsymbol{\theta}-\boldsymbol{\theta}_0)\boldsymbol{\Gamma}^T\mathbf{V}_k^{-1}\boldsymbol{\Gamma}(\boldsymbol{\theta}-\boldsymbol{\theta}_0) + n^{-1/2}(\boldsymbol{\theta}-\boldsymbol{\theta}_0)\boldsymbol{\Gamma}^T\mathbf{V}_k^{-1}\mathcal{W}_{kn} + o_p(n^{-1})
\end{aligned}
$$

and

$$
I_{k82} = -\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\lambda}_k(\boldsymbol{\theta})^T\hat{\mathbf{g}}_{\mathcal{M}_k}(Y_i,\mathbf{Z}_i,\boldsymbol{\theta}_0) = -\left\{n^{-1/2}(\boldsymbol{\theta}-\boldsymbol{\theta}_0)\boldsymbol{\Gamma}^T\mathbf{V}_k^{-1}\mathcal{W}_{kn} + n^{-1}\mathcal{W}_{kn}^T\mathbf{V}_k^{-1}\mathcal{W}_{kn}\right\} + o_p(n^{-1}),
$$

with $\mathcal{W}_{kn} = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\hat{\mathbf{g}}_{\mathcal{M}_k}(Y_i,\mathbf{Z}_i,\boldsymbol{\theta}_0)$. Write $I_{k9} =: I_{k91} + I_{k92}$, where

$$
\begin{aligned}
I_{k91} &= \frac{1}{2n}\sum_{i=1}^{n}\left(\boldsymbol{\lambda}_k(\boldsymbol{\theta})^T\hat{\mathbf{g}}_{\mathcal{M}_k}(Y_i,\mathbf{Z}_i,\boldsymbol{\theta}_0)\right)^2\\
&= \frac{1}{2}\left[(\boldsymbol{\theta}-\boldsymbol{\theta}_0)\boldsymbol{\Gamma}^T\mathbf{V}_k^{-1}\boldsymbol{\Gamma}(\boldsymbol{\theta}-\boldsymbol{\theta}_0) + 2n^{-1/2}(\boldsymbol{\theta}-\boldsymbol{\theta}_0)\boldsymbol{\Gamma}^T\mathbf{V}_k^{-1}\mathcal{W}_{kn} + n^{-1}\mathcal{W}_{kn}^T\mathbf{V}_k^{-1}\mathcal{W}_{kn}\right] + o_p(n^{-1})
\end{aligned}
$$

and

$$
I_{k92} = \frac{1}{2n}\sum_{i=1}^{n}\left[\left(\boldsymbol{\lambda}_k(\boldsymbol{\theta})^T\hat{\mathbf{g}}_{\mathcal{M}_k}(Y_i,\mathbf{Z}_i,\boldsymbol{\theta})\right)^2 - \left(\boldsymbol{\lambda}_k(\boldsymbol{\theta})^T\tilde{\mathbf{g}}_{\mathcal{M}_k}(Y_i,\mathbf{Z}_i,\boldsymbol{\theta}_0)\right)^2\right] = o_p(n^{-1}),
$$

by the continuity of $\mathbf{V}_k$ (which can be easily derived by (C_2^k) and (C_7^k)) in a neighborhood of $\boldsymbol{\theta}_0$. For I_{k10}, by Lemma 3 in Owen (1990), we have

$$
|I_{k10}| \le \|\boldsymbol{\lambda}_k\|^3\left|\frac{1}{n}\sum_{i=1}^{n}\left\{\tilde{\mathbf{g}}_{\mathcal{M}_k}(Y_i,\mathbf{Z}_i,\boldsymbol{\theta})^{\otimes 2}\tilde{\mathbf{g}}_{\mathcal{M}_k}(Y_i,\mathbf{Z}_i,\boldsymbol{\theta})\right\}\right| = o_p(n^{-1}).
$$

Thus, we obtain

$$
\hat{G}_{kn}(\boldsymbol{\theta}) = -\frac{1}{2}\left[\mathbf{V}_k^{-1}\boldsymbol{\Gamma}(\boldsymbol{\theta}-\boldsymbol{\theta}_0) + n^{-1/2}\mathbf{V}_k^{-1}\mathcal{W}_{kn}\right]^T\mathbf{V}_k\left[\mathbf{V}_k^{-1}\boldsymbol{\Gamma}(\boldsymbol{\theta}-\boldsymbol{\theta}_0) + n^{-1/2}\mathbf{V}_k^{-1}\mathcal{W}_{kn}\right] + o_p(n^{-1}). \tag{A14}
$$

Hence, for $k = 1, 2, 3$, we have

$$
\left[\mathbf{V}_k^{-1}\boldsymbol{\Gamma}\left(\hat{\boldsymbol{\theta}}_{ke}-\boldsymbol{\theta}_0\right) + n^{-1/2}\mathbf{V}_k^{-1}\mathcal{W}_{kn} + o_p(n^{-1/2})\right] = 0
$$

or

$$
\sqrt{n}\left(\hat{\boldsymbol{\theta}}_{ke}-\boldsymbol{\theta}_0\right) = -\left(\boldsymbol{\Gamma}^T\mathbf{V}_k^{-1}\boldsymbol{\Gamma}\right)^{-1}\boldsymbol{\Gamma}^T\mathbf{V}_k^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\hat{\mathbf{g}}_{\mathcal{M}_k}(Y_i,\mathbf{Z}_i,\boldsymbol{\theta}_0) + o_p(n^{-1/2}). \tag{A15}
$$

We then obtain $\sqrt{n}(\hat{\boldsymbol{\theta}}_{ke}-\boldsymbol{\theta}_0) \xrightarrow{D} N(0, (\boldsymbol{\Gamma}^T\mathbf{V}_k^{-1}\boldsymbol{\Gamma})^{-1}\boldsymbol{\Gamma}^T\mathbf{V}_k^{-1}\boldsymbol{\Sigma}\mathbf{V}_k^{-1}\boldsymbol{\Gamma}(\boldsymbol{\Gamma}^T\mathbf{V}_k^{-1}\boldsymbol{\Gamma})^{-1})$. Noting that $\mathbf{V}_2 = \boldsymbol{\Sigma}$, we can easily show that $(\boldsymbol{\Gamma}^T\mathbf{V}_2^{-1}\boldsymbol{\Gamma})^{-1}\boldsymbol{\Gamma}^T\mathbf{V}_2^{-1}\boldsymbol{\Sigma}\mathbf{V}_2^{-1}\boldsymbol{\Gamma}(\boldsymbol{\Gamma}^T\mathbf{V}_2^{-1}\boldsymbol{\Gamma})^{-1} = \boldsymbol{\Sigma}_0^{-1}$. Moreover, by (A14), we have

$$
\mathcal{R}_k(\boldsymbol{\theta}_0) = -2n\hat{G}_{kn}(\boldsymbol{\theta}_0) = \left\{\boldsymbol{\Sigma}^{-\frac{1}{2}}\mathcal{W}_{kn}\right\}^T\left(\boldsymbol{\Sigma}^{\frac{1}{2}}\mathbf{V}_k^{-1}\boldsymbol{\Sigma}^{\frac{1}{2}}\right)\left\{\boldsymbol{\Sigma}^{-\frac{1}{2}}\mathcal{W}_{kn}\right\} + o_p(n^{-1}).
$$

This completes the proof of Theorem 4. □

Proof of Theorem 5. By (A14) and (A15), we have

$$2\ell_k(\hat{\boldsymbol{\theta}}_{ke}) = -2n\hat{G}_{kn}(\hat{\boldsymbol{\theta}}_{2e}) = \mathcal{W}_{kn}^T \left[\mathbf{V}_k^{-1} - \mathbf{V}_k^{-1}\boldsymbol{\Gamma}\left(\boldsymbol{\Gamma}^T\mathbf{V}_k^{-1}\boldsymbol{\Gamma}\right)^{-1}\boldsymbol{\Gamma}^T\mathbf{V}_k^{-1}\right]\mathcal{W}_{kn} \tag{A16}$$

and $2\ell_k(\boldsymbol{\theta}_0) = -2n\hat{G}_{2n}(\boldsymbol{\theta}_0) = \mathcal{W}_{kn}^T\mathbf{V}_k^{-1}\mathcal{W}_{kn} + o_p(n^{-1})$. Hence, we get

$$R_k(\boldsymbol{\theta}_0) = \left\{\boldsymbol{\Sigma}^{-\frac{1}{2}}\mathcal{W}_{kn}\right\}^T \boldsymbol{\Sigma}^{\frac{1}{2}}\mathbf{V}_k^{-1}\boldsymbol{\Gamma}\left(\boldsymbol{\Gamma}^T\mathbf{V}_k^{-1}\boldsymbol{\Gamma}\right)^{-1}\boldsymbol{\Gamma}^T\mathbf{V}_k^{-1}\boldsymbol{\Sigma}^{\frac{1}{2}}\left\{\boldsymbol{\Sigma}^{-\frac{1}{2}}\mathcal{W}_{kn}\right\}.$$

Recognizing the above result, Lemma 2, and the fact that $\mathbf{V}_2 = \boldsymbol{\Sigma}$, $(\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{\Gamma}(\boldsymbol{\Gamma}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Gamma})^{-1}\boldsymbol{\Gamma}^T\boldsymbol{\Sigma}^{-\frac{1}{2}})$ is idempotent with trace equal to q. On the basis of these observations, we can easily obtain Theorem 5. □

Proof of Corollary 2. Similarly, by (A14), (A15), and (A16), we have

$$2\ell_k(\boldsymbol{\theta}_{A0}, \hat{\boldsymbol{\theta}}_{keB}) = \mathcal{W}_{kn}^T\left[\mathbf{V}_k^{-1} - \mathbf{V}_k^{-1}\boldsymbol{\Gamma}_2\left(\boldsymbol{\Gamma}_2^T\mathbf{V}_k^{-1}\boldsymbol{\Gamma}_2\right)^{-1}\boldsymbol{\Gamma}_2^T\mathbf{V}_k^{-1}\right]\mathcal{W}_{kn} \text{ and}$$

$$\begin{aligned}
R_k(\boldsymbol{\theta}_{A0}, \hat{\boldsymbol{\theta}}_{keB}) &= \mathcal{W}_{kn}^T\mathbf{V}_k^{-1}\left[\boldsymbol{\Gamma}\left(\boldsymbol{\Gamma}^T\mathbf{V}_k^{-1}\boldsymbol{\Gamma}\right)^{-1}\boldsymbol{\Gamma}^T - \boldsymbol{\Gamma}_2\left(\boldsymbol{\Gamma}_2^T\mathbf{V}_k^{-1}\boldsymbol{\Gamma}_2\right)^{-1}\boldsymbol{\Gamma}_2^T\right]\mathbf{V}_k^{-1}\mathcal{W}_{kn} \\
&= \left\{\boldsymbol{\Sigma}^{-\frac{1}{2}}\mathcal{W}_{kn}^T\right\}\boldsymbol{\Sigma}^{\frac{1}{2}}\mathbf{V}_k^{-1}\left[\boldsymbol{\Gamma}\left(\boldsymbol{\Gamma}^T\mathbf{V}_k^{-1}\boldsymbol{\Gamma}\right)^{-1}\boldsymbol{\Gamma}^T - \boldsymbol{\Gamma}_2\left(\boldsymbol{\Gamma}_2^T\mathbf{V}_k^{-1}\boldsymbol{\Gamma}_2\right)^{-1}\boldsymbol{\Gamma}_2^T\right]\mathbf{V}_k^{-1}\boldsymbol{\Sigma}^{\frac{1}{2}}\left\{\boldsymbol{\Sigma}^{-\frac{1}{2}}\mathcal{W}_{kn}\right\}.
\end{aligned}$$

Noting that $\mathbf{V}_2 = \boldsymbol{\Sigma}$, the conclusion can be obtained immediately. □