arXiv:2108.04852v4 [stat.ME] 22 Feb 2022

MULTIWAY EMPIRICAL LIKELIHOOD

HAROLD D. CHIANG, YUKITOSHI MATSUSHITA, AND TAISUKE OTSU

Abstract. This paper develops a general methodology to conduct statistical inference for observations indexed by multiple sets of entities. We propose a novel multiway empirical likelihood statistic that converges to a chi-square distribution under the non-degenerate case, where corresponding Hoeffding type decomposition is dominated by linear terms. Our methodology is related to the notion of jackknife empirical likelihood but the leave-out pseudo values are constructed by leaving out columns or rows. We further develop a modified version of our multiway empirical likelihood statistic, which converges to a chi-square distribution regardless of the degeneracy, and discover its desirable higher-order property compared to the t-ratio by the conventional Eicker-White type variance estimator. The proposed methodology is illustrated by several important statistical problems, such as bipartite network, two-stage sampling, generalized estimating equations, and three-way observations.

1. Introduction

Many important statistical problems feature multiway data, in which observations are indexed by multiple sets of entities, often arranged as rows and columns. Examples include longitudinal data (Liang and Zeger, 1986), classical random effect models (Searle et al., 2009, Ch 5), row-column exchangeable models (McCullagh, 2000), nonnested multilevel data (Miglioretti and Heagerty, 2007), bipartite networks (Choi and Wolfe, 2014), multi-stage sampling (Fuller, 2011, Ch 3), and multiway clustering (Cameron et al., 2011), to list a few. Observations in such datasets, when correspond to a same set of entity, can exhibit strong dependence that does not diminish as certain distance measure increases, which invalidates conventional asymptotic theory. Although Eicker-White type multiway cluster robust standard errors have been developed for statistical inference on multiway data, and are frequently used in empirical research[1], (i) their derivations are largely case-by-case, (ii) the resulting inference may not be reliable in finite samples, especially when one of the index dimensions contains only a moderate number of units, (iii) in situations with weak or no cluster dependence, the resulting inference often demonstrates significantly less precision than in the cases with strong dependence, and (iv) they often underestimate the variance in finite samples and lead to distortions in the size or coverage properties.

This paper develops a general framework to conduct inference on statistical models for various multiway data. In particular, inspired by the idea of jackknife empirical likelihood of Jing et al. (2009), we propose a novel multiway empirical likelihood (MEL). Unlike the conventional leave-one-out operation for jackknifing, one leaves all the observations in a column or a row out at a time to construct the leave-out pseudo values. The resulting MEL function is computationally

H. Chiang is supported by Office of the Vice Chancellor for Research and Graduate Education at UW-Madison with funding from Wisconsin Alumni Research Foundation.

[1]For example, according to Google Scholar, as of 13th of October, 2021, Cameron et al. (2011) receives over 3, 100 citations, while Thompson (2011) has over 1, 400 citations.

attractive and is shown to be asymptotically pivotal when the linear terms of the Hoeffding type decomposition of the statistical object of interest dominate the quadratic terms (called the non-degenerate case). In multiway data and models, however, degeneracy often occurs so that the quadratic terms in the Hoeffding type decomposition emerge in the first-order, and the MEL statistic loses its asymptotic pivotalness. This phenomenon can be understood as an analogy of emergence of the Efron-Stein bias for the jackknife variance estimator in the multiway context even though in the original setup of Efron and Stein (1981), the bias is of second-order. To recover asymptotic pivotalness, we modify the baseline MEL statistic by incorporating leave one column and row out adjustments, which may be considered as an extension of the leave-two-out bias correction idea in Hinkley (1978) and Efron and Stein (1981) to our two-way setup. Under mild regularity conditions, this modified MEL statistic converges to a chi-square distribution regardless of the degeneracy.

To further motivate our modified MEL approach, we investigate higher-order properties of the modified MEL statistic in a simplified setup. In particular, we find that the second-order term of the asymptotic expansion for the modified MEL statistic is closer to zero than that of the Wald statistic with the Eicker-White type robust standard error, which is always negative. This illustrates an advantage of our modified MEL inference and also explains the oversize phenomenon of the Wald test based on Eicker-White cluster robust standard errors that has been well-documented in the literature (Cameron et al., 2011; Thompson, 2011; MacKinnon et al., 2021, for example). To the best of our knowledge, this is the first result on higher-order properties of the leave-out-based empirical likelihood methods, and this result endorses desirable finite sample properties of the modified MEL method, as showcased in our simulation study.

We illustrate wide-applicability of the modified MEL by various statistical applications, including sparse bipartite network formation models (Bickel et al., 2011; Graham, 2020), generalized estimating equations (Liang and Zeger, 1986; Xie and Yang, 2003; Balan and Schiopu-Kratina, 2005), and statistics under two-stage sampling (Bhattacharya, 2005; Berger and De La Riva Torres, 2016; Chauvet and Vallée, 2020). Throughout these different contexts, the proposed methodology can be applied without modification. We also generalize the modified MEL method to a three-way index setting. Finally, we conduct simulation studies over various settings for classical random effect models and bipartite stochastic block models. The results suggest that the finite sample performance of the modified MEL significantly dominates the Eicker-White multiway cluster robust standard error. The difference is especially profound when cluster dependence is weak or absent. Furthermore, in contrast to the Eicker-White procedure, the modified MEL delivers reliable coverage probabilities even when one of the index dimensions contains only a moderate number of observations.

This paper also contributes to the literature of empirical likelihood (Owen (1988); see Owen (2001) for an overview). After the seminal work by Jing et al. (2009), jackknife empirical likelihood and its variants have been extended to various statistical problems, e.g., Gong et al. (2010), Zhang and Zhao (2013), Zhong and Chen (2014), among others. In particular, Matsushita and Otsu (2021) proposed modified jackknife empirical likelihood to cope with the Efron-Stein bias and established its asymptotic pivotalness under both conventional and non-standard asymptotics.

Under the conventional asymptotics, empirical likelihood inference has been studied and extended to various contexts; see e.g., Bertail (2006), Zhu and Xue (2006), Hjort et al. (2009), Bravo et al. (2020), and a review by Chen and Van Keilegom (2009), among many others.

This paper is organized as follows. Section 2 presents our basic theoretical results on the MEL and its modification for inference on the means of two-way data. Sections 2.1 and 2.2 study the first and higher order asymptotic properties, respectively. In Section 3, we extend our MEL approach to a bipartite network model (Section 3.1), two-stage or cluster sampling model (Section 3.2), generalized estimating equations (Section 3.3), and three-way data (Section 3.4). Section 4 illustrates the proposed method by two simulation examples. All proofs are contained in the Appendix.

2. Benchmark case: Two-way empirical likelihood for mean

2.1. **First-order asymptotic theory.** As a benchmark, we first consider a two-way sample of d-dimensional random vectors $\{X_{ij} : i = 1, \ldots, N, j = 1, \ldots, M\}$ generated following

$$X_{ij} = \tau_n(U, U_{i0}, U_{0j}, U_{ij}), \tag{2.1}$$

for $i = 1, \ldots, N$ and $j = 1, \ldots, M$, where $\{U, U_{i0}, U_{0j}, U_{ij} : i = 1, \ldots, N, j = 1, \ldots, M\}$ are i.i.d. unobservable latent shocks that can be normalized to $U[0,1]$,[2] and τ_n is an unknown real-valued Borel-measurable map that may vary with n. Hereafter all population objects, such as the distribution and moments of X_{ij}, are conditional on U and depend on n through $\tau_n(\cdot)$, but we suppress the conditioning on U and dependence on n for notational brevity. In this benchmark setup, we consider statistical inference on the mean vector $\theta = E[X_{11}]$ by using the MEL method.

A common sufficient condition for the representation in (2.1) is that $\{X_{ij} : i = 1, \ldots, N, j = 1, \ldots, M\}$ is embedded into an infinite two-way separately exchangeable array $(X_{ij})_{(i,j)\in\mathbb{N}^2}$. An infinite array $(X_{ij})_{(i,j)\in\mathbb{N}^2}$ is called separately exchangeable if for any two permutations of positive integers $\pi_1, \pi_2 : \mathbb{N} \to \mathbb{N}$ and any finite subset $A \subset \mathbb{N}^2$, it holds $(X_{ij})_{(i,j)\in A} \overset{d}{=} (X_{\pi_1(i)\pi_2(j)})_{(i,j)\in A}$. Under such condition, (2.1) follows from the celebrated Aldous-Hoover-Kallenberg representation for separately exchangeable arrays (e.g., Corollary 7.23 of Kallenberg, 2006). This representation is widely used in modern statistics, e.g., Diaconis and Janson (2008), Bickel et al. (2011), Choi and Wolfe (2014), Bhattacharyya and Bickel (2015), Gao et al. (2015), Caron and Fox (2017), Choi (2017), Zhang et al. (2017), Lauritzen et al. (2018), Veitch and Roy (2019), Davezies et al. (2021), MacKinnon et al. (2021), Menzel (2021), and many more. See also the review by Orbanz and Roy (2014).

The representation in (2.1) is useful for our theoretical development since it allows us to establish a Hoeffding type decomposition for the estimation error of the point estimator $\hat{\theta} = \frac{1}{NM}\sum_{i=1}^N \sum_{j=1}^M X_{ij}$, that is

$$\hat{\theta} - \theta = \frac{1}{N}\sum_{i=1}^N L_{i0} + \frac{1}{M}\sum_{j=1}^M L_{0j} + \frac{1}{NM}\sum_{i=1}^N\sum_{j=1}^M (W_{ij} + R_{ij}), \tag{2.2}$$

[2] Independence over $(U_{0j})_{j\in\mathbb{N}}$ is assumed for simplicity, and can be replaced by martingale difference sequence type conditions.

where

$$\begin{aligned} L_{i0} &= E[X_{i1}|U_{i0}] - E[X_{11}], \qquad L_{0j} = E[X_{1j}|U_{0j}] - E[X_{11}], \\ W_{ij} &= E[X_{ij}|U_{i0}, U_{0j}] - E[X_{i1}|U_{i0}] - E[X_{1j}|U_{0j}] + E[X_{11}], \\ R_{ij} &= X_{ij} - E[X_{ij}|U_{i0}, U_{0j}]. \end{aligned}$$

Observe that $\{L_{i0} : i = 1, \dots, N\}$ and $\{L_{0j} : j = 1, \dots, M\}$ are i.i.d., and $\{R_{ij} : i = 1, \dots, N, j = 1, \dots, M\}$ is also i.i.d. conditional on $\{U_{i0}, U_{0j} : i = 1, \dots, N, j = 1, \dots, M\}$.

Throughout this paper, we regard $S(\theta) = \hat{\theta} - \theta$ as an estimating equation for θ and construct the MEL function to conduct inference on θ. More precisely, we introduce the leave-out pseudo value:

$$V_l(\theta) = nS(\theta) - (n-1)S_l(\theta),$$

for $l = 1, \dots, n$, where $S_l(\theta) = \hat{\theta}^{(l)} - \theta$ and $\hat{\theta}^{(l)}$ is the leave one column or row out counterpart of $\hat{\theta}$ defined as

$$\hat{\theta}^{(l)} = \begin{cases} \frac{1}{(N-1)M} \sum_{i \neq l} \sum_{j=1}^{M} X_{ij} & \text{if } l \leq N, \\ \frac{1}{N(M-1)} \sum_{i=1}^{N} \sum_{j \neq l} X_{ij} & \text{otherwise.} \end{cases}$$

Unlike the conventional leave-one-out operation for jackknifing, we leave all the observations that have a specific i or j out at a time. Thus the number of leave-out pseudo values is $n = N + M$ instead of NM, which indicates computational attractiveness of our MEL method.

By considering the leave-out pseudo value $V_l(\theta)$ as an estimating function for θ, the MEL function for θ is constructed as

$$\ell(\theta) = -2 \sup_{w_1, \dots, w_n} \sum_{l=1}^{n} \log(nw_l) \qquad \text{s.t.} \quad w_l \geq 0, \quad \sum_{l=1}^{n} w_l = 1, \quad \sum_{l=1}^{n} w_l V_l(\theta) = 0.$$

Although this optimization involves n variables $\{w_1 \dots, w_n\}$, in practice $\ell(\theta)$ can be computed by the dual form

$$\ell(\theta) = 2 \sup_{\lambda \in \mathbb{R}^d} \sum_{l=1}^{n} \log(1 + \lambda' V_l(\theta)), \tag{2.3}$$

where λ' denotes the transpose of λ.

To study asymptotic properties of the MEL statistic $\ell(\theta)$, we impose the following assumptions. Let $\underline{n} = N \wedge M$, $\sigma_L^2 = n\{Var(L_{10})/N + Var(L_{01})/M\}$ and $\sigma_R^2 = Var(R_{11})$ be the variance matrices of the components in the Hoeffding type decomposition in (2.2), and $\lambda_{\min}(A)$ be the minimum eigenvalue of a matrix A. We say the sequence of data-generating processes is *non-degenerate* if $\lambda_{\min}(\underline{n}\sigma_L^2) \to \infty$ and *nearly degenerate* if $\lambda_{\min}(\underline{n}\sigma_L^2) = O(1)$.

Assumption 1. *(i) $\{X_{ij} : i = 1, \dots, N, j = 1, \dots, M\}$ is generated following (2.1). (ii) For some $q > 4$, $E[\|X_{11}\|^q]$ is bounded from above uniformly in n. Also $\lim_{n\to\infty} \underline{n}/(N \vee M) \in (0, 1)$. (iii) Under the nearly degenerate case, $\lambda_{\min}(\sigma_R^2) \geq c > 0$ for a constant c independent of n and $\| \sum_{i=1}^{N} \sum_{j=1}^{M} W_{ij}\| / \| \sum_{i=1}^{N} \sum_{j=1}^{M} R_{ij}\| = o_p(1)$.*

Remark 1. Assumption 1(i) assumes that the data have a two-way dependence structure. Assumption 1(ii) requires the observables to have more than four moments, as well as limiting the growth rates of N and M to be similar. This can be loosen by imposing alternative assumptions to control the growth rates of $Var(L_{10})$ and $Var(L_{01})$. Assumption 1(iii) imposes some high level conditions on the asymptotic behaviour of the term W_{ij} in the Hoeffding type decomposition for the nearly degenerate case. It assumes that the term R_{ij} in (2.2) remains random asymptotically, but the term W_{ij} is asymptotically negligible compared to the R_{ij} term. This condition holds automatically in several representative applications, such as the classical random effect models (Searle et al., 2009, Ch 5), asymptotics of sparse networks (Bickel et al., 2011; Graham, 2020), and kernel-type estimation of directed dyadic regression models (Graham et al., 2019).

Under these assumptions, the limiting distribution of the MEL statistic $\ell(\theta)$ is obtained as follows.

Theorem 1. *Under Assumption 1, it holds*

$$\ell(\theta) \xrightarrow{d} \begin{cases} \chi_d^2 & \text{for non-degenerate case,} \\ \xi'\Omega^{-1}\xi & \text{for nearly degenerate case,} \end{cases}$$

where $\xi \sim N(0, \lim_{n\to\infty}(\underline{n}\sigma_L^2 + \sigma_R^2))$ *and* $\Omega = \lim_{n\to\infty}(\underline{n}\sigma_L^2 + 2\sigma_R^2)$.

Remark 2. This theorem says that the asymptotic distribution of the MEL statistic $\ell(\theta)$ depends on the behaviour of the variance component σ_L^2 for the linear term in (2.2). If it is asymptotically non-negligible in the sense that σ_L^2 does not converge to zero at least as fast as $\underline{n}^{-1}$, then the MEL statistic is asymptotically pivotal. However, when σ_L^2 converges to zero at a rate of $\underline{n}^{-1}$ or faster, then the MEL statistic is no longer pivotal and its asymptotic distribution depends on $\underline{n}$, σ_L^2, and σ_R^2. If τ_n, $(U_{i0})_i$, and $(U_{0j})_j$ in (2.1) are known to the statistician and $(U_{ij})_{i,j}$ do not enter τ_n, then the discrepancy between "$2\sigma_R^2$" in Ω and "σ_R^2" in the variance of ξ can be understood as a two-sample U-statistics generalization of the Efron-Stein bias for the second-order bias. Nonetheless, under our asymptotic framework, this bias emerges in the first-order.

In order to conduct statistical inference based on the MEL statistic $\ell(\theta)$, we need to employ different critical values for the different cases. In particular, for the nearly degenerate case, we need to estimate $Var(\xi)$ and Ω. Thus, it is desirable to modify the MEL statistic to have the same limiting distribution for both cases.

Motivated by the bias correction method in Efron and Stein (1981), we develop a modified version of the MEL statistic as follows. For each $l = 1, \ldots, N$ and $l_1 = 1, \ldots, M$, let $\hat{\theta}^{(l,l_1)} = \frac{1}{(N-1)(M-1)} \sum_{i\neq l} \sum_{j \neq l_1} X_{ij}$ be the leave one-column and one-row out counterparts of $\hat{\theta}$, $S_{l,l_1}(\theta) = \hat{\theta}^{(l,l_1)} - \theta$, and

$$Q_{ll_1} = \mathcal{C}(N, M) \cdot [nS(\theta) - (n-1)\{S_l(\theta) + S_{N+l_1}(\theta)\} + (n-2)S_{l,l_1}(\theta)], \tag{2.4}$$

where $\mathcal{C}(N, M) = \frac{(N-1)(M-1)n}{(NM)(n-2)}$. Note that the term Q_{ll_1} is different from the leave-two-out counterpart employed in Efron and Stein (1981) to correct the higher-order bias of the jackknife variance estimator since we delete a whole column and row of the data matrix (X_{ij}). Therefore, the total number of leave-out estimators required here is of order $O(n^2)$, similar to the usual leave-one-out procedures, in contrast to $\binom{NM}{2} = O(n^4)$ of the conventional leave-two-out methods. The factor $\mathcal{C}(N, M)$ is a finite sample adjustment to make the coefficient of the leading term

$(W_{ll_1}+R_{ll_1})$ of Q_{ll_1} to be "$\frac{n}{NM}$" as in (A.8) in the Appendix so that the leading term of the adjustment term $\frac{MN}{n^2}\sum_{l=1}^{N}\sum_{l_1=1}^{M}Q_{ll_1}Q_{ll_1}'$ will be $\frac{1}{NM}\sum_{l=1}^{N}\sum_{l_1=1}^{M}(W_{ll_1}+R_{ll_1})^2$, which is unbiased for $Var(W_{ll_1}+R_{ll_1})$.

Based on Q_{ll_1}, the modified estimating function is defined as

$$V_l^m(\theta)=V_l(\hat{\theta})-\hat{\Gamma}\tilde{\Gamma}^{-1}\{V_l(\hat{\theta})-V_l(\theta)\},$$

for $l=1,\dots,n$, where $\hat{\Gamma}$ and $\tilde{\Gamma}$ are so that

$$\hat{\Gamma}\hat{\Gamma}'=\frac{1}{n}\sum_{l=1}^{n}V_l(\hat{\theta})V_l(\hat{\theta})',\qquad \tilde{\Gamma}\tilde{\Gamma}'=\frac{1}{n}\sum_{l=1}^{n}V_l(\theta)V_l(\theta)'-\frac{1}{n}\sum_{l=1}^{N}\sum_{l_1=1}^{M}Q_{ll_1}Q_{ll_1}'. \tag{2.5}$$

By using $V_l^m(\theta)$ as a moment function, the modified MEL statistic is defined as

$$\ell^m(\theta)=2\sup_{\lambda\in\mathbb{R}^d}\sum_{l=1}^{n}\log\{1+\lambda' V_l^m(\theta)\}, \tag{2.6}$$

and the asymptotic property of this statistic is obtained as follows.

Theorem 2. *Under Assumption 1, it holds (for both non-degenerate and nearly degenerate cases)*

$$\ell^m(\theta)\xrightarrow{d}\chi_d^2.$$

This theorem shows that the modified MEL statistic $\ell^m(\theta)$ has the asymptotically pivotal distribution of χ_d^2 for both asymptotic regimes. We emphasize that the modified MEL inference only requires the estimators, $\hat{\theta}$, $\hat{\theta}^{(l)}$, and $\hat{\theta}^{(l,l_1)}$, and circumvents estimation of $Var(\xi)$ and Ω in Theorem 1. Based on this theorem, the asymptotic $100(1-\alpha)\%$ modified MEL confidence set can be constructed as $\{\theta:\ell^m(\theta)\le\chi_{d,\alpha}^2\}$, where $\chi_{d,\alpha}^2$ is the $(1-\alpha)$-th quantile of the χ_d^2 distribution.

Remark 3. A by-product of the proposed modified MEL procedure is the modified multiway variance estimator $\tilde{\Gamma}\tilde{\Gamma}'$ in (2.5) evaluated at $\theta=\hat{\theta}$, an alternative to the Eicker-White multiway cluster robust variance estimators. It can be considered as an analogy of the bias-corrected jackknife variance estimator (Efron and Stein, 1981) for our multiway context. Based this variance estimator, we can also construct a confidence interval $\left[\hat{\theta}_j\pm n^{-1/2}z_{\alpha/2}\sqrt{[\tilde{\Gamma}\tilde{\Gamma}']_{(j,j)}}\right]$ for the j-th element θ_j of θ. This confidence interval seems to be new in the literature, and in contrast to Efron and Stein (1981), the correction term in the new standard error $\sqrt{[\tilde{\Gamma}\tilde{\Gamma}']_{(j,j)}}$ is not asymptotically negligible in the first-order.

2.2. **Higher-order properties.** In this subsection, we provide some theoretical justification for desirable accuracy of the modified MEL statistic by the asymptotic χ^2 calibration based on the higher-order property. In particular, we compare the second-order terms of the distributions of the modified MEL statistic $\ell^m(\theta)$ and the Wald statistic or t-ratio based on the Eicker-White type cluster robust variance estimator, i.e., $T(\theta)=(\hat{\theta}-\theta)'\hat{\Sigma}^{-1}(\hat{\theta}-\theta)$, where

$$\hat{\Sigma}\;=\;\frac{1}{N^2M^2}\sum_{i=1}^{N}\sum_{j=1}^{M}\sum_{j_1=1}^{M}(X_{ij}-\hat{\theta})(X_{ij_1}-\hat{\theta})'+\frac{1}{N^2M^2}\sum_{i=1}^{N}\sum_{i_1=1}^{N}\sum_{j=1}^{M}(X_{ij}-\hat{\theta})(X_{i_1j}-\hat{\theta})'$$

$$-\frac{1}{N^2M^2}\sum_{i=1}^{N}\sum_{j=1}^{M}(X_{ij}-\hat{\theta})(X_{ij}-\hat{\theta})'.$$

The variance estimator $\hat{\Sigma}$ is a two-way version of the cluster robust variance estimator of Liang and Zeger (1986) from Miglioretti and Heagerty (2007), Cameron et al. (2011), and Thompson (2011). Its asymptotic properties are subsequently investigated in Davezies et al. (2021) and MacKinnon et al. (2021). In terms of the first-order asymptotic property, both $\ell^m(\theta)$ and $T(\theta)$ converge in distribution to the χ^2_d distribution.

Here we compare higher-order accuracy of these statistics in a simplified setup, where X_{ij} is scalar and a stronger version of the degeneracy condition in Assumption 1 (iii) is imposed. Let Φ and ϕ be the standard normal cumulative distribution and density functions, respectively. Higher order properties of $\ell^m(\theta)$ and $T(\theta)$ are presented as follows.

Theorem 3. *Suppose that Assumption 1 holds true with scalar X_{ij}, $E[X_{ij}^{10}]<\infty$, $\frac{1}{N}\sum_{i=1}^{N}L_{i0}+\frac{1}{M}\sum_{j=1}^{M}L_{0j}=o_p(n^{-2})$, $\frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M}W_{ij}=o_p(n^{-2})$, and $\limsup_{|t|\to\infty}|E[e^{\mathrm{i}tX_{ij}}]|<1$. Furthermore suppose $\sigma_R^2=1$ to simplify the presentation. Then for each $t>0$,*

$$\Pr\{\sqrt{T(\theta)}\le t\}=\Phi(t)-\left\{\frac{3}{2}\left(\frac{1}{N}+\frac{1}{M}\right)t+\frac{1}{2}\left(\frac{1}{N}+\frac{1}{M}\right)t^3\right\}\phi(t)+o(n^{-1}),$$

and

$$\begin{aligned}\Pr\{\sqrt{\ell^m(\theta)}\le t\} &= \Phi(t)-\left\{\left(\frac{3}{2}\left(\frac{1}{N}+\frac{1}{M}\right)-\frac{3}{N}-\frac{3}{M}+\frac{5}{n}\right)t+\frac{1}{2}\left(\frac{1}{N}+\frac{1}{M}-\frac{2(\sqrt{2}-1)}{n}\right)t^3\right\}\phi(t)\\ &\quad +o(n^{-1}).\end{aligned}$$

Several remarks follow. First, the asymptotic expansion for the (signed root of) Wald statistic $T(\theta)$ based on the Eicker-White type cluster robust variance estimator shows that its second-order term is of order $O(n^{-1})$ and takes a negative value. This result suggests undercoverage of the Wald-type confidence interval based on $T(\theta)$ in finite samples as illustrated in our simulation studies in Section 4.

Second, the asymptotic expansion for the modified MEL statistic $\ell^m(\theta)$ shows that the second-order term is also of order $O(n^{-1})$ but is closer to zero. This result indicates desirable accuracy of the modified MEL statistic by the χ^2 calibration for the degenerate case.

Third, even without the factor $\mathcal{C}(N,M)$ in the adjustment term in (2.4), the corresponding modified MEL (say, $\tilde{\ell}^m(\theta)$) yields a smaller second-order term than that of the Wald statistic as

$$\begin{aligned}\Pr\{\sqrt{\tilde{\ell}^m(\theta)}\le t\} &= \Phi(t)-\left\{\left(\frac{3}{2}\left(\frac{1}{N}+\frac{1}{M}\right)-\frac{1}{N}-\frac{1}{M}+\frac{1}{n}\right)t+\frac{1}{2}\left(\frac{1}{N}+\frac{1}{M}-\frac{2(\sqrt{2}-1)}{n}\right)t^3\right\}\phi(t)\\ &\quad +o(n^{-1}).\end{aligned}$$

However, the refinement of $\tilde{\ell}^m(\theta)$ in the coefficient of $t\phi(t)$ is smaller than that of $\ell^m(\theta)$.

Fourth, this theorem covers the case where M and N can be of same order. When these orders are different, e.g., $M = o(N)$, our expansions are simplified to

$$\begin{aligned}
\Pr\{\sqrt{T(\theta)} \leq t\} &= \Phi(t) - \left(\frac{3}{2}\frac{1}{M}t + \frac{1}{2}\frac{1}{M}t^3\right)\phi(t) + o(M^{-1}), \\
\Pr\{\sqrt{\ell^m(\theta)} \leq t\} &= \Phi(t) - \left(\frac{1}{2}\frac{1}{M}t + \frac{1}{2}\frac{1}{M}t^3\right)\phi(t) + o(M^{-1}),
\end{aligned}$$

which also imply that the second-order term of $\ell^m(\theta)$ is closer to zero than that of $T(\theta)$.

Finally, if we replace the degeneracy condition $\frac{1}{NM}\sum_{i=1}^N \sum_{j=1}^M W_{ij} = o_p(n^{-2})$ with Assumption 1 (iii), we can still establish analogous expansions as in (A.9) and (A.10) with an additional term, $\frac{1}{\sqrt{NM}}\sum_{i=1}^N \sum_{j=1}^M W_{ij}$, and the cumulants can be computed in an analogous way. However, due to the non-normal limiting distribution of $\frac{1}{\sqrt{NM}}\sum_{i=1}^N \sum_{j=1}^M W_{ij}$, its Edgeworth expansion will be more involved.

3. Generalizations

3.1. **Bipartite network.** In this section, we extend our (modified) MEL inference method for slope parameters in the logistic regression model for sparse bipartite network models investigated by Graham (2020). While the asymptotic properties of the maximum composite likelihood estimator under sparse network asymptotics has been studied in the literature, no inference method has been proposed for this estimator.

Let μ, $\{(W_{i0}, A_{i0}) : i = 1, \dots, N\}$, $\{(W_{0j}, A_{0j}) : j = 1, \dots, M\}$, and $\{V_{ij} : i = 1, \dots, N, j = 1, \dots, M\}$ be i.i.d. sequences, where W_{i0} and W_{0j} are observed attributes with supports $\mathcal{W}_1$ and $\mathcal{W}_2$, respectively, and $(\mu, A_{i0}, A_{0j}, V_{ij})$ are unobserved shocks. Suppose the random bipartite graph $\{Y_{ij} \in \{0, 1\} : i = 1, \dots, N, j = 1, \dots, M\}$ is generated according to

$$Y_{ij} = h_{N,M}(\mu, W_{i0}, W_{0j}, A_{i0}, A_{j0}, V_{ij}), \tag{3.1}$$

where $h_{N,M} : [0,1] \times \mathcal{W}_1 \times \mathcal{W}_2 \times [0,1]^3 \to \{0,1\}$ is a graphon unknown to the statistician. Suppose the statistician observes $\{Y_{ij}, Z_{ij} : i = 1, \dots, N, j = 1, \dots, M\}$, where $Z_{ij} = z(W_{i0}, W_{0j})$ is a vector of known transformations of W_{i0} and W_{0j}. Consider the logistic network formation model

$$\Pr(Y_{ij} = 1 \mid W_{i0}, W_{0j}) = \Lambda(\alpha_n + Z_{ij}'\beta), \tag{3.2}$$

where $\Lambda(u) = \exp(u)/(1 + \exp(u))$, and α_n is an intercept, which may vary with n. The asymptotics is understood as $n \to \infty$. Suppose $M/n \to \phi \in (0, 1)$. Let $\rho_n = E[\Lambda(\alpha_n + Z_{11}'\beta)]$ be the marginal link formation probability, and $\lambda_n^1 = M\rho_n$ and $\lambda_n^2 = N\rho_n$ be the average degrees of the first and second cluster dimensions, respectively. Suppose $\alpha_n = \log(\eta/n)$ for some constant η. Note that under such setting, $\alpha_n \to -\infty$ and $\lambda_n^1 \to \lambda^1 \in (0, \infty)$ as $n \to \infty$, that is, both average degrees stay finite in the limit.

We estimate the model by the composite maximum likelihood

$$(\hat{\alpha}, \hat{\beta}) = \arg\max_{\alpha,\beta} \sum_{i=1}^N \sum_{j=1}^M \mathcal{L}_{ij}(\alpha, \beta), \tag{3.3}$$

where $\mathcal{L}_{ij}(\alpha, \beta) = Y_{ij} \log \Lambda(\alpha_n + Z_{ij}'\beta) + (1 - Y_{ij}) \log(1 - \Lambda(\alpha_n + Z_{ij}'\beta))$. Suppose the object of interest is a d-dimensional subvector θ of $(\alpha, \beta')'$. The MEL function $\ell(\theta)$ for θ is obtained as in (2.3) by setting $S(\theta) = \hat{\theta} - \theta$ and $S^{(l)}(\theta) = \hat{\theta}^{(l)} - \theta$, where $\hat{\theta}$ is the corresponding subvector of the estimator in (3.3) and $\hat{\theta}^{(l)}$ is the corresponding subvector of

$$(\hat{\alpha}^{(l)}, \hat{\beta}^{(l)}) = \begin{cases} \arg\max_{\alpha,\beta} \sum_{i \neq l} \sum_{j=1}^{M} \mathcal{L}_{ij}(\alpha, \beta) & \text{if } i \leq N, \\ \arg\max_{\alpha,\beta} \sum_{i=1}^{N} \sum_{j \neq l} \mathcal{L}_{ij}(\alpha, \beta) & \text{otherwise.} \end{cases}$$

Similarly, the modified MEL function $\ell^m(\theta)$ can be obtained as in (2.6) by setting $S_{l,l_1}(\theta) = \hat{\theta}^{(l,l_1)} - \theta$, where $\hat{\theta}^{(l,l_1)}$ is the corresponding subvector of $\arg\max_{\alpha,\beta} \sum_{i \neq l} \sum_{j \neq l_1} \mathcal{L}_{ij}(\alpha, \beta)$.

The asymptotic property of $\ell^m(\theta)$ is obtained under the following conditions. Note that we do not need to impose a high-level condition that corresponds to Assumption 1(iii), as it holds automatically under the current setting (Graham, 2020, pp. 15).

Assumption 2. *(i) (3.1) and (3.2) hold true. (ii) $(\eta, \beta')'$ lies in the interior of a compact parameter space. (iii) $z(\cdot, \cdot)$ is compactly supported. (iv) $M/n \to \phi \in (0, 1)$ as $n \to \infty$. (v) $H = -\eta E[\exp(Z_{11}'\beta)(1, Z_{11}')'(1, Z_{11}')]$ is of full rank.*

Theorem 4. *Under Assumption 2, it holds*

$$\ell^m(\theta) \xrightarrow{d} \chi_d^2.$$

Similar comments to Theorem 2 apply. As shown in Graham (2020), the asymptotic variance of the maximum composite likelihood estimator $\hat{\theta}$ involves several terms due to non-negligible contributions from the higher-order terms in the Hoeffding type decomposition (2.2). In contrast, our modified MEL approach only requires $\hat{\theta}$, $\hat{\theta}^{(l)}$'s, and $\hat{\theta}^{(l,l_1)}$'s, and circumvents estimation of such variance components.

3.2. **Two-stage or cluster sampling.** Suppose the sample $\{Y_{ij} : i = 1, \ldots, N, j = 1, \ldots, M\}$ is obtained from a two-stage or cluster sampling design. We could consider (i) one-stage sampling, where j's are treated as clusters and the j-th column is drawn with probability, say π_j; or (ii) two-stage sampling, where we first draw clusters (or primary sampling units, PSUs) and then for each cluster, we do another sampling to obtain secondary sampling units (SSUs) within the cluster. So Y_{ij} is drawn with probability, say π_{ij}. See, e.g., Fuller (2011) for an overview of sampling statistics.

In this subsection, let us illustrate our MEL approach by the Horvitz-Thompson type estimator for the population mean:

$$\hat{\theta}_\pi = \frac{\sum_{i=1}^{N} \sum_{j=1}^{M} \pi_{ij}^{-1} Y_{ij}}{\sum_{i=1}^{N} \sum_{j=1}^{M} \pi_{ij}^{-1}}, \tag{3.4}$$

which is routinely applied in the analysis of survey data. There are many studies on this estimator in design-based context; see, e.g., Chauvet (2015), Berger and De La Riva Torres (2016), Chauvet and Vallée (2020), Zhao et al. (2020), and references therein. To adapt the asymptotic theory in the last section, here we take the model-based approach based on a hypothetical infinite population (Fisher, 1922). This complements the recent results of, e.g., Chauvet and Vallée (2020), where a design-based approach is employed and the numbers of sampled primary and secondary sampling units are proportional to the size of the finite population. We leave the design-based asymptotic analysis of the MEL approach for future research.

To be specific, we consider the two-stage sampling design for household surveys, which is studied by Bhattacharya (2005) for the case of fixed M (i.e., the size of SSUs). We complement the analysis in Bhattacharya (2005) by investigating the case where both N (i.e., the size of PSUs) and M are large. Let i and j be indexes for a PSU and a SSU in the PSU, respectively. Let J be the total number of PSUs in the population, H_i be the number of SSUs in PSU i, and v_{ij} be the size of the j-th SSU in the i-th PSU. All the following expectation in this subsection are taken with respect to the sampling distribution. Let $F(y, v)$ be the joint cumulative distribution function of the per capita characteristic Y and SSU size v. Then the population mean θ of Y can be defined as the solution of

$$J \int v(y-\theta)dF(y,v) = JE\left[\sum_{j=1}^{H_i} v_{ij}(Y_{ij}-\theta)\right] = 0.$$

Suppose we observe a sample of N PSUs and M SSUs from each of the sampled PSU. Then the estimator $\hat{\theta}_\pi$ in (3.4) is derived as the method of moments estimator with the weight $\pi_{ij} = (JH_i v_{ij})^{-1}$. Also the leave one column or row out version of $\hat{\theta}_\pi$ is defined as

$$\hat{\theta}_\pi^{(l)} = \begin{cases} \frac{\sum_{i\neq l}\sum_{j=1}^{M}\pi_{ij}^{-1}Y_{ij}}{\sum_{i\neq l}\sum_{j=1}^{M}\pi_{ij}^{-1}} & \text{if } l \leq N, \\ \frac{\sum_{i=1}^{N}\sum_{j\neq l}\pi_{ij}^{-1}Y_{ij}}{\sum_{i=1}^{N}\sum_{j\neq l}\pi_{ij}^{-1}} & \text{otherwise.} \end{cases}$$

Similarly we can define $\hat{\theta}_\pi^{(l,l_1)} = (\sum_{i\neq l}\sum_{j\neq l_1}\pi_{ij}^{-1}Y_{ij})/(\sum_{i\neq l}\sum_{j\neq l_1}\pi_{ij}^{-1})$ for $l = 1, \ldots, N$ and $l_1 = 1, \ldots, M$. Then the modified MEL function $\ell^m(\theta)$ can be obtained as in (2.6) by setting $S(\theta) = \hat{\theta}_\pi - \theta$, $S_l(\theta) = \hat{\theta}_\pi^{(l)} - \theta$, and $S_{l,l_1}(\theta) = \hat{\theta}_\pi^{(l,l_1)} - \theta$.

Assumption 3. *(i) $X_{ij} = (Y_{ij}, v_{ij})$, $\{X_{ij} : i = 1, \ldots, N, j = 1, \ldots, M\}$ can be embedded into a structure of the form (2.1). (ii) $\sup_n E[|\pi_{11}^{-1}Y_{11}|^4]/Var(\pi_{11}^{-1}Y_{11})^2 < \infty$, $\sup_n E[\pi_{11}^{-4}]/Var(\pi_{11}^{-1})^2 < \infty$ and $Var(\pi_{11}^{-1}Y_{11}) \vee Var(\pi_{11}^{-1}) \geq c > 0$. (iii) $\lim_{n\to\infty} \underline{n}/(N \vee M) \in (0, 1)$. (iv) Under the nearly degenerate case, $\|\sum_{i=1}^{N}\sum_{j=1}^{M} W_{ij}\|/\|\sum_{i=1}^{N}\sum_{j=1}^{M} R_{ij}\| = o_p(1)$.*

The following result is a straightforward implication of the proof of Theorem 2 combined with the delta method.

Proposition 1. *Under Assumption 3, it holds*

$$\ell^m(\theta) \xrightarrow{d} \chi_1^2.$$

It is remarkable that our modified MEL inference based on the Horvitz-Thompson type estimator $\hat{\theta}_\pi$ is free from variance estimation, which is typically involved for two-stage sampling designs. Under the model-based asymptotics, Bhattacharya (2005) derived the asymptotic variance of his estimator, which requires estimation of several components.

3.3. **Generalized estimating equations under cluster dependence.** In Section 3.1, we consider inference on a logistic regression model for bipartite network data, where the modified MEL function is constructed based on the composite maximum likelihood estimator. More generally, our MEL method can be applied to conduct inference on parameters defined via generalized estimating equations (GEEs) for longitudinal data (Liang and Zeger, 1986). The existing literature on the GEE mostly focuses on the case where the cluster size is fixed and there is no dependence across clusters. A notable exception is Xie and Yang (2003) who investigated the asymptotic properties of the GEE estimators under the asymptotic regime of $N \to \infty$ and M being either fixed or diverges to infinity at some appropriate rates while maintaining independence across clusters. Thus it is an interesting open question whether we can conduct valid inference for parameters under both growing cluster sizes and dependence across clusters.

To fix the idea, consider a generalized linear model based on the density $f(Y_{ij}|Z_{ij},\theta,\phi) = \exp[\{Y_{ij}u(Z_{ij}'\theta) - a(u(Z_{ij}'\theta)) + b(Y_{ij})\}/\phi]$ for $i = 1,\dots,N$ and $j = 1,\dots,M$, where u, a, and b are known functions and ϕ is a known constant. To conduct inference on θ when $M \to \infty$ and (Y_{ij}, Z_{ij}) is embedded into a separately exchangeable array, we employ the estimating equations using the independent working correlation matrix

$$\sum_{i=1}^{N}\sum_{j=1}^{M} u^{(1)}(Z_{ij}'\theta)Z_{ij}\{Y_{ij} - a^{(1)}(u(Z_{ij}'\theta))\} = 0,$$

where $u^{(1)}$ and $a^{(1)}$ are the derivatives of u and a, respectively. Letting $X_{ij}(\theta) = u^{(1)}(Z_{ij}'\theta)Z_{ij}\{Y_{ij} - a^{(1)}(u(Z_{ij}'\theta))\}$, the modified MEL function $\ell^m(\theta)$ is defined as in (2.6) by setting $S(\theta) = \frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M} X_{ij}(\theta)$,

$$\begin{aligned} S_l(\theta) &= \begin{cases} \frac{1}{(N-1)M}\sum_{i\neq l}\sum_{j=1}^{M} X_{ij}(\theta) & \text{if } l \leq N, \\ \frac{1}{N(M-1)}\sum_{i=1}^{N}\sum_{j\neq l} X_{ij}(\theta) & \text{otherwise,} \end{cases} \\ S_{l,l_1}(\theta) &= \frac{1}{(N-1)(M-1)}\sum_{i\neq l}\sum_{j\neq l_1} X_{ij}(\theta). \end{aligned}$$

Then as far as the assumptions for Theorem 2 are satisfied for $X_{ij}(\theta)$, we obtain $\ell^m(\theta) \xrightarrow{d} \chi^2_{\dim(\theta)}$ at the true value of θ. Also the modified MEL statistic for the composite null hypothesis $H_0: r(\theta) = 0$ can be obtained by $\min_{\theta: r(\theta)=0} \ell^m(\theta)$, which converges to $\chi^2_{\dim(r(\theta))}$ by adapting the argument in Qin and Lawless (1994).

3.4. **Multiway MEL.** Let us now extend the (modified) MEL approach to the three way case $\{X_{ijt} : i = 1,\dots,N, j = 1,\dots,M, t = 1,\dots,T\}$. Similar modification works for K-way for any fixed $K \in \mathbb{N}$. Consider the sample mean $\hat{\theta} = (NMT)^{-1}\sum_{i=1}^{N}\sum_{j=1}^{M}\sum_{t=1}^{T} X_{ijt}$ for the population mean $\theta = E[X_{111}]$. Note that following the Aldous-Hoover representation as in (2.1) and the Hoeffding type decomposition (Chiang et al., 2021, Lemma 1) for general K-way mean, we have

$$\begin{aligned} \hat{\theta} - \theta &= \frac{1}{N}\sum_{i=1}^{N} L_{i00} + \frac{1}{M}\sum_{j=1}^{M} L_{0j0} + \frac{1}{T}\sum_{t=1}^{T} L_{00t} \\ &\quad + \frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M} A_{ij0} + \frac{1}{MT}\sum_{j=1}^{M}\sum_{t=1}^{T} A_{0jt} + \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} A_{i0t} + \frac{1}{NMT}\sum_{i=1}^{N}\sum_{j=1}^{M}\sum_{t=1}^{T} A_{ijt}, \end{aligned}$$

where

$$L_{i00} = E[X_{i11} \mid U_{i00}] - \theta, \quad L_{0j0} = E[X_{1j1} \mid U_{0j0}] - \theta, \quad L_{00t} = E[X_{11t} \mid U_{00t}] - \theta,$$

$$\begin{aligned}
A_{ij0} &= E[X_{ij1} \mid U_{i00}, U_{0j0}, U_{ij0}] - L_{i00} - L_{0j0} - \theta, \\
A_{i0t} &= E[X_{i1t} \mid U_{i00}, U_{00t}, U_{i0t}] - L_{i00} - L_{00t} - \theta, \\
A_{0jt} &= E[X_{1jt} \mid U_{0j0}, U_{00t}, U_{0jt}] - L_{0j0} - L_{00t} - \theta, \\
A_{ijt} &= X_{ijt} - A_{ij0} - A_{i0t} - A_{0jt} - L_{i00} - L_{0j0} - L_{00t} - \theta.
\end{aligned}$$

Suppose it holds that

$$\sum_{i=1}^{N}\sum_{j=1}^{M} A_{ij0} = \sum_{i=1}^{N}\sum_{j=1}^{M} R_{ij0}\{1 + o_p(1)\},$$

where $R_{ij0} = g_{12}(U_{ij0})$, $g_{12} : [0,1] \to \mathbb{R}^{\dim(\theta)}$ is an unknown Borel-measurable mapping. Similarly, suppose that $\sum_{i=1}^{N}\sum_{t=1}^{T} A_{i0t} = \sum_{i=1}^{N}\sum_{t=1}^{T}\{R_{i0t} + o_p(1)\}$ with $R_{i0t} = g_{13}(U_{i0t})$, $\sum_{j=1}^{M}\sum_{t=1}^{T} A_{0jt} = \sum_{j=1}^{M}\sum_{t=1}^{T}\{R_{0jt} + o_p(1)\}$ with $R_{0jt} = g_{23}(U_{0jt})$, and $\sum_{i=1}^{N}\sum_{j=1}^{M}\sum_{t=1}^{T} A_{ijt} = \sum_{i=1}^{N}\sum_{j=1}^{M}\sum_{t=1}^{T}\{R_{ijt} + o_p(1)\}$ with $R_{ijt} = g_{123}(U_{ijt})$ for some unknown Borel-measurable mappings g_{13}, g_{23}, and g_{123}. This is analogous to Assumption 1(iii). Let $n = N + M + T$. Then a central limit theorem implies that as $\min\{N, M, T\} \to \infty$, we have $\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} N(0, \sigma_*^2)$, where

$$\sigma_*^2 = n\left\{\frac{Var(L_{100})}{N} + \frac{Var(L_{010})}{M} + \frac{Var(L_{001})}{T} + \frac{Var(R_{110})}{NM} + \frac{Var(R_{101})}{NT} + \frac{Var(R_{011})}{MT} + \frac{Var(R_{111})}{NMT}\right\}.$$

Now, define $S(\theta) = \hat{\theta} - \theta$, $S_l(\theta) = \hat{\theta}^{(l)} - \theta$, and the leave-one-index-out estimators

$$\hat{\theta}^{(l)} = \begin{cases} \frac{1}{(N-1)MT}\sum_{i\neq l}\sum_{j=1}^{M}\sum_{t=1}^{T} X_{ijt} & \text{if } l \leq N, \\ \frac{1}{N(M-1)T}\sum_{i=1}^{N}\sum_{j\neq l}\sum_{t=1}^{T} X_{ijt} & \text{if } N < l \leq N+M, \\ \frac{1}{NM(T-1)}\sum_{i=1}^{N}\sum_{j=1}^{M}\sum_{t\neq l} X_{ijt} & \text{otherwise.} \end{cases}$$

Further, define $V_l(\theta) = nS(\theta) - (n-1)S^{(l)}(\theta)$, $V_l^m(\theta) = V_l(\hat{\theta}) - \hat{\Gamma}\tilde{\Gamma}^{-1}\{V_l(\hat{\theta}) - V_l(\theta)\}$, for $l = 1, \ldots, N$, where $\hat{\Gamma}$ and $\tilde{\Gamma}$ are so that

$$\begin{aligned}
\hat{\Gamma}\hat{\Gamma}' &= \frac{1}{n}\sum_{l=1}^{n} V_l(\hat{\theta})V_l(\hat{\theta})', \\
\tilde{\Gamma}\tilde{\Gamma}' &= \frac{1}{n}\sum_{l=1}^{n} V_l(\hat{\theta})V_l(\hat{\theta})' - \frac{1}{n}\left\{\sum_{l=1}^{N}\sum_{l_1=1}^{M} Q_{ll_10}Q'_{ll_10} + \sum_{l_1=1}^{M}\sum_{l_2=1}^{T} Q_{0l_1l_2}Q'_{0l_1l_2} + \sum_{l=1}^{N}\sum_{l_2=1}^{T} Q_{l0l_2}Q'_{l0l_2}\right\} \\
&\quad + \frac{1}{n}\sum_{l=1}^{N}\sum_{l_1=1}^{M}\sum_{l_2=1}^{T} Q_{ll_1l_2}Q'_{ll_1l_2},
\end{aligned}$$

where

$$\begin{aligned}
Q_{ll_10} &= [n\hat{\theta} - (n-1)(\hat{\theta}^{(l)} + \hat{\theta}^{(N+l_1)}) + (n-2)\hat{\theta}^{(l,l_1,0)}], \\
Q_{l0l_2} &= [n\hat{\theta} - (n-1)(\hat{\theta}^{(l)} + \hat{\theta}^{(N+M+l_2)}) + (n-2)\hat{\theta}^{(l,0,l_2)}], \\
Q_{0l_1l_2} &= [n\hat{\theta} - (n-1)(\hat{\theta}^{(N+l_1)} + \hat{\theta}^{(N+M+l_2)}) + (n-2)\hat{\theta}^{(0,l_1,l_2)}], \\
Q_{ll_1l_2} &= [n\hat{\theta} - (n-1)(\hat{\theta}^{(l)} + \hat{\theta}^{(N+l_1)} + \hat{\theta}^{(N+M+l_2)}) \\
&\quad + (n-2)(\hat{\theta}^{(l,l_1,0)} + \hat{\theta}^{(0,l_1,l_2)} + \hat{\theta}^{(l,0,l_2)}) - (n-3)\hat{\theta}^{(l,l_1,l_2)}],
\end{aligned}$$

and the leave-two-index-out and leave-three-index-out estimators are defined by

$$\hat{\theta}^{(l,l_1,0)} = \frac{1}{(N-1)(M-1)T}\sum_{i\neq l}\sum_{j\neq l_1}\sum_{t=1}^{T} X_{ijt} \quad \text{for } l\leq N, l_1\leq M,$$

$$\hat{\theta}^{(0,l_1,l_2)} = \frac{1}{N(M-1)(T-1)}\sum_{i=1}^{N}\sum_{j\neq l_1}\sum_{t\neq l_2} X_{ijt} \quad \text{for } l_1\leq M, l_2\leq T,$$

$$\hat{\theta}^{(l,0,l_2)} = \frac{1}{(N-1)M(T-1)}\sum_{i\neq l}\sum_{j=1}^{M}\sum_{t\neq l_2} X_{ijt} \quad \text{for } l\leq N, l_2\leq T,$$

$$\hat{\theta}^{(l,l_1,l_2)} = \frac{1}{(N-1)(M-1)(T-1)}\sum_{i\neq l}\sum_{j\neq l_1}\sum_{t\neq l_2} X_{ijt} \quad \text{for } l\leq N, l_1\leq M, l_2\leq T.$$

Under some regularity conditions, it can be shown similarly as in the proof of Theorem 1 that

$$\frac{1}{n}\sum_{l=1}^{n} V_l(\theta)V_l(\theta)' = n\Bigg\{\left(\frac{Var(L_{100})}{N}+\frac{Var(L_{010})}{M}+\frac{Var(L_{001})}{T}\right) +2\left(\frac{Var(R_{110})}{NM}+\frac{Var(R_{101})}{NT}+\frac{Var(R_{011})}{MT}\right)+3\left(\frac{Var(R_{111})}{NMT}\right)\Bigg\}+o_p(1).$$

We can then obtain $\ell^m(\theta)$ by following the same definition as in (2.6) with corresponding components replaced by those defined in this section. It can be shown that under regularity conditions, the modified MEL statistic has the pivotal asymptotic distribution

$$\ell^m(\theta) \xrightarrow{d} \chi^2_{\dim(\theta)}.$$

In fact, this proposed procedure is much less computationally intensive in comparison with the corresponding jackknife procedures for U-statistics since the number of all leave-out estimators in the proposed procedure is $O(n^3)$, the same as the order of all different leave-one-out estimators for i.i.d. data. On the other hand, the conventional leave-three-out estimators with sample size NMT consist of $\binom{NMT}{3} = O(n^9)$ possibilities.

4. Simulation

This section conducts a simulation study to evaluate the finite sample properties of the proposed MEL inference methods. In particular, we consider a random effect model (Section 4.1) and bipartite stochastic block model (Section 4.2). We shall focus on simple means based on the reasoning as in Owen (2007) that one can expect a method that gives the correct variance for a mean to be reliable for more complicated statistics such as smooth functions of means and estimating equation parameters.

4.1. **Random effect model.** We first consider the random effect model studied in Owen (2007) and Searle et al. (2009, Ch 5):

$$X_{ij} = \theta + a_i + b_j + \varepsilon_{ij},$$

where $\theta = 1$ and $(a_i, b_j, \varepsilon_{ij})$ are mutually independent random variables with $a_i, b_j \sim N(0,\sigma^2)$ and $\varepsilon_{ij} \sim N(0,1)$. The estimator considered here is the sample mean. We vary $\sigma^2 \in \{1, 0.1, 0\}$

to examine the performance under non-degenerate, nearly degenerate, and degenerate cases, respectively. We set $N = 50$ and $M \in \{5, 10, 15, 20, 30, 50\}$.

We compare five methods of constructing confidence intervals: (i) multiway empirical likelihood (MEL), (ii) modified MEL (mMEL), (iii) Wald confidence interval with modified multiway variance estimator from Remark 3 (mMW), (iv) Wald with Eicker-White type multiway cluster robust variance estimator (EWW), and (v) Wald with i.i.d. variance estimator (IID). The methods (i)-(iii) are our developments, (iv) is a conventional method, and (v) is asymptotically invalid (except for the degenerate case) but included for comparison. The nominal coverage is set as 0.95.

Table 1 reports empirical coverages of the methods (i)-(v) based on 5,000 Monte Carlo replications. Our findings are summarized as follows. First, IID does not work at all except for the degenerate case (i.e., $\sigma^2 = 0$). Since IID is asymptotically invalid, its size distortion remains even for $N, M = 50$. Second, EWW exhibits severe under-coverages when M is small, as predicted by the higher order asymptotic theory. Third, mMEL, which is asymptotically valid for all cases, outperforms in almost all cases. Even if M is small, mMEL performs well. Fourth, MEL works well for non-degenerate case (i.e., $\sigma^2 = 1$) but over-covers for nearly degenerate and degenerate cases. This result is expected from Theorem 1. As shown in Theorem 2, mMEL recovers asymptotic pivotalness for all cases and our simulation result clearly illustrate this point. Fifth, for Wald-type confidence intervals, the proposed mMW works better than the conventional EWW but slightly under-covers.

Overall we recommend mMEL, which exhibits accurate coverages and is robust for all cases.

M	σ^2	MEL	mMEL	mMW	EWW	IID
5	1	0.946	0.950	0.922	0.857	0.332
	0.1	0.955	0.944	0.928	0.857	0.587
	0	0.988	0.941	0.934	0.818	0.947
10	1	0.948	0.953	0.936	0.907	0.326
	0.1	0.965	0.958	0.944	0.911	0.571
	0	0.992	0.948	0.943	0.887	0.953
15	1	0.955	0.962	0.951	0.930	0.329
	0.1	0.963	0.957	0.948	0.924	0.558
	0	0.991	0.941	0.939	0.903	0.946
20	1	0.947	0.951	0.943	0.927	0.305
	0.1	0.964	0.958	0.951	0.932	0.544
	0	0.994	0.953	0.949	0.921	0.954
30	1	0.952	0.956	0.952	0.938	0.286
	0.1	0.958	0.951	0.948	0.933	0.535
	0	0.993	0.946	0.944	0.920	0.950
50	1	0.955	0.957	0.954	0.946	0.266
	0.1	0.956	0.951	0.949	0.941	0.490
	0	0.993	0.949	0.946	0.935	0.953

TABLE 1. Coverage rates for random effect model with $N = 50$

4.2. **Bipartite stochastic block model.** We next consider a stochastic block model, which is an adapted version of Bhattacharyya and Bickel (2015, Sec. 5.2) for bipartite graphs with two different community dimensions. First, each i is randomly assigned to a membership $a \in \{1, 2\}$ of the first community dimension with probabilities $\pi_1 = (0.7, 0.3)'$ and each j is randomly assigned to a membership $b \in \{1, 2\}$ of the second community dimension with probabilities $\pi_2 = (0.2, 0.8)'$. Then consider the following edge formation probabilities

$$F_{ab} = \Pr(X_{ij} = 1 | i \in A_a, j \in B_b) = s_\theta S_{ab}, \text{ for } a \in \{1, 2\} \text{ and } b \in \{1, 2\},$$

where the blocks A_a and B_b satisfy $A_1 \cup A_2 = \{1, ..., N\}$, $A_1 \cap A_2 = \emptyset$, $B_1 \cup B_2 = \{1, ..., M\}$, $B_1 \cap B_2 = \emptyset$, S_{ab}'s are elements of $S = \begin{bmatrix} 0.6 & 0.4 \\ 0.3 & 0.7 \end{bmatrix}$, and s_θ is chosen to satisfy $\theta = \pi_1' F \pi_2 \in \{0.5, 0.1, 0.05\}$.

Similar to the last subsection, we consider the five confidence intervals (i)-(v) for θ with the nominal coverage 0.95 for the cases of $N = 50$ and $M \in \{5, 10, 15, 20, 30, 50\}$. Table 2 reports empirical coverages of the methods (i)-(v) based on 5,000 Monte Carlo replications. The results are qualitatively similar to the ones in the last subsection. IID has size distortions, EWW exhibits under-coverages when M is small, and mMEL outperforms the rest for almost all cases. In this simulation study on a bipartite network, MEL shows over-coverages for all cases including the relatively dense case (i.e., $\theta = 0.5$). Therefore, we recommend to use mMEL for this simulation study, too.

M	θ	MEL	mMEL	mMW	EWW	IID
5	0.5	0.989	0.938	0.926	0.826	0.928
	0.1	0.986	0.948	0.931	0.822	0.928
	0.05	0.989	0.951	0.926	0.834	0.930
10	0.5	0.988	0.951	0.939	0.900	0.906
	0.1	0.991	0.943	0.938	0.884	0.940
	0.05	0.990	0.950	0.931	0.891	0.934
15	0.5	0.990	0.955	0.945	0.923	0.902
	0.1	0.994	0.952	0.943	0.917	0.946
	0.05	0.991	0.948	0.937	0.908	0.949
20	0.5	0.986	0.957	0.946	0.928	0.876
	0.1	0.993	0.947	0.937	0.920	0.946
	0.05	0.993	0.950	0.934	0.919	0.942
30	0.5	0.986	0.950	0.940	0.930	0.848
	0.1	0.993	0.948	0.940	0.924	0.939
	0.05	0.991	0.940	0.930	0.918	0.932
50	0.5	0.983	0.957	0.949	0.943	0.798
	0.1	0.990	0.950	0.944	0.935	0.929
	0.05	0.993	0.950	0.941	0.931	0.941

TABLE 2. Coverage rates for stochastic block model with $N = 50$

Appendix A. Mathematical appendix

A.1. **Proof of Theorem 1.** In this subsection, we use the following lemma.

Lemma 1. *Suppose that* $E[\|X_{11}\|^q] \le C < \infty$ *and* $\rho = n/\underline{n} = o(n^{1/2-2/q})$ *for some* $q > 4$. *Then it holds* $\max_{1\le l\le n} \|V_l(\theta)\| = o_p(n^{1/2})$.

Proof of Lemma 1. For notational simplicity, assume that X_{11} is a scalar. For each $l = 1, \dots, N$, it holds

$$S(\theta) = \frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M} X_{ij} - \theta = \frac{1}{NM}\sum_{j=1}^{M}(X_{lj} - \theta) + \frac{N-1}{N}S_l(\theta),$$

and thus $V_l(\theta) = nS(\theta) - (n-1)S_l(\theta)$ satisfies

$$\begin{aligned}\max_{1\le l\le N} |V_l(\theta)| &= \max_{1\le l\le N}\left|\frac{n}{NM}\sum_{j=1}^{M}(X_{lj}-\theta) + \left(n\frac{N-1}{N} - (n-1)\right)S_l(\theta)\right| \\ &= O_p\left(\frac{n}{N}\max_{1\le i\le N}\max_{1\le j\le N}|X_{ij}|\right) + O_p\left(\max_{1\le l\le N}|S_l(\theta)|\right) = O_p\left(\rho \max_{1\le i\le N}\max_{1\le j\le N}|X_{ij}|\right).\end{aligned}$$

Similarly, it holds that $\max_{N+1\le l\le n} |V_l(\theta)| = O_p\left(\rho \max_{1\le i\le N}\max_{1\le j\le N}|X_{ij}|\right)$. Now, by Jensen's inequality,

$$E\left[\max_{1\le i\le N}\max_{1\le j\le N}|X_{ij}|\right] \le \left(E\left[\max_{1\le i\le N}\max_{1\le j\le N}|X_{ij}|^q\right]\right)^{1/q} \le C^{1/q}(NM)^{1/q} \le C^{1/q}n^{2/q}.$$

Combining these results, we obtain

$$\max_{1\le l\le n}|V_l(\theta)| = O_p(\rho n^{2/q}) = o_p(n^{1/2}).$$

Proof of Theorem 1. For notational simplicity, assume that X_{11} is a scalar, as the general result follows analogously with an application of a multivariate central limit theorem. In this proof, we will frequently utilize the U-statistic form of the sample variance: for $\bar{x} = n^{-1}\sum_{l=1}^{n} x_l$

$$\frac{1}{n}\sum_{l=1}^{n}(x_l - \bar{x})^2 = \frac{1}{n^2}\sum_{1\le l<l_1\le n}(x_l - x_{l_1})^2. \tag{A.1}$$

We will also make use of the algebraic identity $S(\theta) = n^{-1}\sum_{l=1}^{n} S_l(\theta) = n^{-1}\sum_{l=1}^{n} V_l(\theta)$. The rest of this proof is divided into three steps.

Step 1. In this step, we shall establish the main asymptotic result for the MEL statistic. Let $\hat{\lambda} = \arg\max_\lambda \sum_{l=1}^{n}\log(1+\lambda V_l(\theta))$. The first-order condition for $\hat{\lambda}$ and the fact that $(1+x)^{-1} = 1 - x + x^2(1+x)^{-1}$ give

$$0 = \frac{1}{n}\sum_{l=1}^{n}\frac{V_l(\theta)}{1+\hat{\lambda}V_l(\theta)} = \frac{1}{n}\sum_{l=1}^{n}V_l(\theta) - \frac{1}{n}\sum_{l=1}^{n}V_l(\theta)^2\hat{\lambda} + \frac{1}{n}\sum_{l=1}^{n}\frac{V_l(\theta)^3\hat{\lambda}^2}{1+\hat{\lambda}V_l(\theta)}.$$

Since Lemma 1 guarantees $\max_{1\le l\le n}|V_l(\theta)| = o_p(n^{1/2})$, a standard argument as in (Owen, 1990, eq. (2.14)) yields $\hat{\lambda} = O_p(n^{-1/2})$. Thus, one has

$$\hat{\lambda} = \frac{\sum_{l=1}^{n}V_l(\theta)}{\sum_{l=1}^{n}V_l(\theta)^2} + o_p(n^{-1/2}).$$

A Taylor expansion yields

$$\begin{aligned}\ell(\theta) &= 2\sum_{l=1}^{n}\log(1+\hat{\lambda}V_l(\theta)) = 2\sum_{l=1}^{n}\left(\hat{\lambda}V_l(\theta)-\frac{1}{2}\{\hat{\lambda}V_l(\theta)\}^2\right)+o_p(1)\\ &= \frac{\left\{n^{-1/2}\sum_{l=1}^{n}V_l(\theta)\right\}^2}{n^{-1}\sum_{l=1}^{n}V_l(\theta)^2}+o_p(1). \qquad (A.2)\end{aligned}$$

Therefore, the conclusion follows from the convergence in distribution of the numerator and the consistency of the denominator for the non-degenerate case (Step 2) and nearly degenerate case (Step 3).

Step 2. We shall now investigate the limiting behaviours of the numerator and denominator in (A.2) under non-degeneracy. Under this scenario, it holds that $\underline{n}\sigma_L^2 \to \infty$, which implies $\sigma_R^2 = o(\underline{n}\sigma_L^2)$. Also note that $\lim_{n\to\infty}\sigma_L^2/n < \infty$ under Assumption 1(ii). Define $\sigma^2 = \sigma_L^2/n + \sigma_R^2/(NM)$. In this case, it is sufficient for the conclusion of the theorem to show that

$$\frac{1}{n\sigma}\sum_{l=1}^{n}V_l(\theta) \overset{d}{\to} N(0,1), \qquad (A.3)$$

$$\frac{1}{n^2\sigma^2}\sum_{l=1}^{n}V_l(\theta)^2 \overset{p}{\to} 1. \qquad (A.4)$$

We first show (A.3). Eq. (2.2) and Assumption 1(iii) imply

$$\frac{1}{n}\sum_{l=1}^{n}V_l(\theta) = \{1+o_p(1)\}\left\{\frac{1}{N}\sum_{i=1}^{N}L_{i0}+\frac{1}{M}\sum_{j=1}^{M}L_{0j}\right\}+\frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M}R_{ij}.$$

For each n, define the sequence $\{Y_{n,t} : t = 1,\dots,n+NM\}$ by

$$Y_{n,t} = \begin{cases} L_{t0}/N & \text{for } t \le N,\\ L_{0(t-N)}/M & \text{for } N+1 \le t \le n,\\ R_{ij}/NM & \text{for } t > n, \text{ where } j = \lceil (t-n)/N \rceil, i = (t-n)-(j-1)N. \end{cases} \qquad (A.5)$$

Notice that $n^{-1}\sum_{l=1}^{n}V_l(\theta) = \sum_{t=1}^{n+NM}Y_{n,t}+o_p(1)$. It is straightforward to verify that it has zero-mean and $E[Y_{n,t}|\{Y_{n,t_1} : t_1 = 1,\dots,t-1\}] = 0$, i.e., $\{Y_{n,t} : t = 1,\dots,n+NM\}$ is a martingale difference sequence. Moreover, observe that

$$\sum_{t=1}^{n+NM}E[(Y_{n,t}/\sigma)^{2+\delta}] \to 0 \text{ for some } \delta > 0, \text{ and } \sum_{t=1}^{n+NM}Y_{n,t}^2-\sigma^2 \overset{p}{\to} 0,$$

where the first result (Lyapunov's condition) is implied by Assumption 1(ii) and the condition that $\underline{n}\sigma_L^2 \to \infty$, and the second result follows from the weak law of large numbers and conditional independence of $\{R_{ij} : i = 1,\dots,N, j = 1,\dots,M\}$. Therefore, the central limit theorem for martingale difference triangular arrays (White, 2001, Corollary 5.26) implies

$$\sum_{t=1}^{n+NM}\frac{Y_{n,t}}{\sigma} = \frac{1}{n\sigma}\sum_{l=1}^{n}V_l(\theta)+o_p(1) \overset{d}{\to} N(0,1),$$

i.e., the result in (A.3) follows.

We now show (A.4). Note that

$$\frac{1}{n^2\sigma^2}\sum_{l=1}^{n} V_l(\theta)^2 = \frac{1}{n^2\sigma^2}\sum_{l=1}^{n}[S(\theta)-(n-1)\{S_l(\theta)-S(\theta)\}]^2$$

$$= \frac{1}{n\sigma^2}S(\theta)^2 + \frac{(n-1)^2}{n^2\sigma^2}\sum_{l=1}^{n}\{S_l(\theta)-S(\theta)\}^2 = \frac{(n-1)^2}{n^2\sigma^2}\sum_{l=1}^{n}\{S_l(\theta)-S(\theta)\}^2 + o_p(1)$$

$$= \frac{(n-1)^2}{n^2\sigma^2}\left[\sum_{i=1}^{N}\left\{S_i(\theta)-\frac{1}{N}\sum_{i=1}^{N}S_i(\theta)\right\}^2 + \sum_{j=1}^{M}\left\{S_{N+j}(\theta)-\frac{1}{M}\sum_{j=1}^{M}S_{N+j}(\theta)\right\}^2\right] + o_p(1)$$

$$= \frac{(n-1)^2}{n^2\sigma^2}\left[\frac{1}{N}\sum_{i=1}^{N}\sum_{i'=i+1}^{N}\{S_i(\theta)-S_{i'}(\theta)\}^2 + \frac{1}{M}\sum_{j=1}^{M}\sum_{j'=j+1}^{M}\{S_{N+j}(\theta)-S_{N+j'}(\theta)\}^2\right] + o_p(1)$$

$$= \frac{(n-1)^2}{n^2N\sigma^2}\sum_{i=1}^{N}\sum_{i'=i+1}^{N}\left[\frac{1}{N-1}(L_{i'0}-L_{i0}) + \frac{1}{(N-1)M}\sum_{l=1}^{M}\{(W_{i'l}-W_{il})+(R_{i'l}-R_{il})\}\right]^2$$

$$+\frac{(n-1)^2}{n^2M\sigma^2}\sum_{j=1}^{M}\sum_{j'=j+1}^{M}\left[\frac{1}{M-1}(L_{0j'}-L_{0j}) + \frac{1}{N(M-1)}\sum_{k=1}^{N}\{(W_{kj'}-W_{kj})+(R_{kj'}-R_{kj})\}\right]^2 + o_p(1)$$

$$= \frac{\sigma_L^2}{n\sigma^2} + o_p(1) = 1 + o_p(1),$$

where the second equality follows from the identity $n^{-1}\sum_{l=1}^{n} S_l(\theta) = S(\theta)$, the third equality follows from $S(\theta) = O_p(\underline{n}^{-1/2})$, the fourth equality follows from $\frac{1}{N}\sum_{i=1}^{N} S_i(\theta) = \frac{1}{M}\sum_{j=1}^{M} S_{N+j}(\theta) = S(\theta)$, the fifth equality follows from (A.1), the seventh equality follows from the fact that each of $(L_{i0})_i$, $(L_{0j})_j$, $(W_{ij})_{ij}$, $(R_{ij})_{ij}$ is i.i.d. and centred, as well as an application of the law of large numbers, and the last equality follows from the fact $n\sigma^2 = \sigma_L^2 + o(1)$ when $\underline{n}\sigma_L^2 \to \infty$.

Step 3. We shall now investigate the limiting behaviours of (A.2) under the nearly degenerate scenario. In this case, it holds that $\sigma_L^2 = O(1/\underline{n})$, which implies $\sigma_L^2/n = O(\sigma_R^2/(NM))$. Notice that under Assumption 1(iii) and (2.2), we have

$$\frac{1}{n}\sum_{l=1}^{n} V_l(\theta) = \frac{1}{N}\sum_{i=1}^{N} L_{i0} + \frac{1}{M}\sum_{j=1}^{M} L_{0j} + \frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M} R_{ij}(1+o_p(1)).$$

For each n, we define the sequence $\{Y_{n,t} : 1 \le t \le n + NM\}$ as in (A.5) again, which is a martingale difference sequence. Using the same argument as in Step 2, the central limit theorem for martingale difference triangular arrays yields the same result in (A.3).

Thus, it remains to show

$$\frac{1}{n^2\sigma^2}\sum_{l=1}^{n} V_l(\theta)^2 \xrightarrow{p} \lim_{n\to\infty}\frac{\sigma_L^2/n + 2\sigma_R^2/(NM)}{\sigma_L^2/n + \sigma_R^2/(NM)}. \tag{A.6}$$

To this end, analogous arguments as in Step 2 (using the fact that $S(\theta) = O_p((NM)^{-1/2})$) yield

$$\frac{1}{n^2\sigma^2}\sum_{i=1}^{n} V_i(\theta)^2 = \frac{1}{n^2\sigma^2}\sum_{i=1}^{n}[S(\theta)-(n-1)\{S^{(i)}(\theta)-S(\theta)\}]^2$$

$$= \frac{1}{n\sigma^2}S(\theta)^2 + \frac{(n-1)^2}{n^2\sigma^2}\sum_{i=1}^n\{S^{(i)}(\theta)-S(\theta)\}^2 = \frac{(n-1)^2}{n^2\sigma^2}\sum_{i=1}^n\{S^{(i)}(\theta)-S(\theta)\}^2 + o_p(1)$$

$$= \frac{(n-1)^2}{n^2\sigma^2}\left[\frac{1}{N}\sum_{i=1}^N\sum_{i'=i+1}^N\{S^{(i)}(\theta)-S^{(i')}(\theta)\}^2 + \frac{1}{M}\sum_{j=1}^M\sum_{j'=j+1}^M\{S^{(N+j)}(\theta)-S^{(N+j')}(\theta)\}^2\right] + o_p(1)$$

$$= \frac{(n-1)^2}{n^2N\sigma^2}\sum_{i=1}^N\sum_{i'=i+1}^N\left[\frac{1}{N-1}(L_{i'0}-L_{i0}) + \frac{1}{(N-1)M}\sum_{l=1}^M\{(W_{i'l}-W_{il})+(R_{i'l}-R_{il})\}\right]^2$$

$$+\frac{(n-1)^2}{n^2M\sigma^2}\sum_{j=1}^M\sum_{j'=j+1}^M\left[\frac{1}{M-1}(L_{0j'}-L_{0j}) + \frac{1}{N(M-1)}\sum_{k=1}^N\{(W_{kj'}-W_{kj})+(R_{kj'}-R_{kj})\}\right]^2 + o_p(1)$$

$$= \frac{\sigma_L^2/n + 2\sigma_R^2/(NM)}{\sigma_L^2/n+\sigma_R^2/(NM)} + o_p(1),$$

where the second equality follows from the identity $n^{-1}\sum_{l=1}^n S_l(\theta) = S(\theta)$, the third equality follows from $S(\theta) = O_p((NM)^{-1/2})$, the fourth equality follows from (A.1), and the last equality follows from the fact that each of $(L_{i0})_i$, $(L_{0j})_j$, $(W_{ij})_{ij}$, $(R_{ij})_{ij}$ is i.i.d. and centred, as well as an application of the law of large numbers. Therefore, the conclusion follows.

A.2. **Proof of Theorem 2.** For brevity, we focus on the case where X_{11} is scalar. Also we present the proof only for the nearly degenerate case (i.e., $\sigma_L^2 = O(1/\underline{n})$ and $\sum_{i=1}^N\sum_{j=1}^M W_{ij} = o_p(\sum_{i=1}^N\sum_{j=1}^M R_{ij})$), as the non-degenerate case can be shown similarly. As in Step 1 in the proof of Theorem 1, one can obtain the asymptotic expansion

$$\ell^m(\theta) = \frac{\left\{n^{-1/2}\sum_{l=1}^n V_l^m(\theta)\right\}^2}{n^{-1}\sum_{l=1}^n V_l^m(\theta)^2} + o_p(1).$$

Define $\sigma^2 = \sigma_L^2/n + \sigma_R^2/(NM)$. A similar argument to Step 3 in the proof of Theorem 1 together with the consistency of $\hat\theta$ imply $(n^2\sigma^2)^{-1}\sum_{l=1}^n V_l^m(\hat\theta)^2 \xrightarrow{p} \lim_{n\to\infty}(\underline{n}\sigma_L^2 + 2\sigma_R^2)/(\underline{n}\sigma_L^2+\sigma_R^2)$. It now suffices to show

$$\frac{1}{n^2\sigma^2}\sum_{l=1}^N\sum_{l_1=1}^M Q_{ll_1}^2 = \frac{\sigma_R^2}{NM\sigma^2} + o_p(1), \tag{A.7}$$

as the desired result is then implied by Step 3 in the proof of Theorem 1 as well as

$$\frac{1}{n\sigma}\sum_{l=1}^n V_l^m(\theta) = \left\{\frac{\sum_{l=1}^n V_l(\hat\theta)^2}{\sum_{l=1}^n V_l(\hat\theta)^2 - \sum_{l=1}^N\sum_{l_1=1}^M Q_{ll_1}^2}\right\}^{1/2}\frac{1}{n\sigma}\sum_{l=1}^n V_l(\theta)$$

$$= \left\{\frac{\underline{n}\sigma_L^2 + 2\sigma_R^2}{\underline{n}\sigma_L^2+\sigma_R^2}\right\}^{1/2}\frac{1}{n\sigma}\sum_{l=1}^n V_l(\theta) + o_p(1).$$

Notice that for each $l = 1,\dots,N$ and $l_1 = 1,\dots,M$, eq. (2.2) yields

$$nS(\theta) = \frac{n}{N}\sum_{i=1}^N L_{i0} + \frac{n}{M}\sum_{j=1}^M L_{0j} + \frac{n}{NM}\sum_{i=1}^N\sum_{j=1}^M\{W_{ij}+R_{ij}\},$$

$$(n-1)S_l(\theta) = \frac{n-1}{N-1}\sum_{i\neq l} L_{i0} + \frac{n-1}{M}\sum_{j=1}^M L_{0j} + \frac{n-1}{(N-1)M}\sum_{i\neq l}\sum_{j=1}^M\{W_{ij}+R_{ij}\},$$

$$(n-1)S_{N+l_1}(\theta) = \frac{n-1}{N}\sum_{i=1}^{N} L_{i0} + \frac{n-1}{M-1}\sum_{j\neq l_1} L_{0j} + \frac{n-1}{N(M-1)}\sum_{i=1}^{N}\sum_{j\neq l_1}^{M}\{W_{ij}+R_{ij}\},$$

$$(n-2)S_{l,l_1}(\theta) = \frac{n-2}{N-1}\sum_{i\neq l} L_{i0} + \frac{n-2}{M-1}\sum_{j\neq l_1} L_{0j} + \frac{n-2}{(N-1)(M-1)}\sum_{i\neq l}\sum_{j\neq l_1}\{W_{ij}+R_{ij}\}.$$

By plugging in these expressions, a direct calculation yields

$$\begin{aligned} Q_{ll_1} &= \mathcal{C}(N,M)\left\{-\frac{1}{N-1}\left(L_{l0} - \frac{1}{N}\sum_{i=1}^{N} L_{i0}\right) - \frac{1}{M-1}\left(L_{0l_1} - \frac{1}{M}\sum_{j=1}^{M} L_{0j}\right)\right\} \\ &\quad -\frac{1}{M^2}\sum_{j=1}^{M}(W_{lj}+R_{lj}) - \frac{1}{N^2}\sum_{i=1}^{N}(W_{il_1}+R_{il_1}) + \frac{n}{NM}(W_{ll_1}+R_{ll_1}) + o_p(1). \quad \text{(A.8)} \end{aligned}$$

By inserting this expression to the left hand side of (A.7), applications of the weak law of large numbers under Assumption 1 (ii)-(iii) imply (A.7).

A.3. **Proof of Theorem 3.** By proceeding as in Chapter 2 of Hall (1992), we have

$$\Pr\{\sqrt{T(\theta)} \le t\} = \Phi(t) - \left\{(\kappa_{2,n}-1)t + \frac{\kappa_{4,n}}{12}(t^3-3t)\right\}\phi(t) + o((NM)^{-1/2}),$$

where $\kappa_{q,n}$ is the q-th cumulant of $T(\theta)$. By an expansion, $T(\theta)$ can be written as

$$T(\theta) = A\{1 + B_1 + B_2 + B_3 + B_4 + B_5\} + o_p((NM)^{-1/2}) \quad \text{(A.9)}$$

where

$$\begin{aligned} A &= \frac{1}{\sqrt{NM}}\sum_{a=1}^{N}\sum_{b=1}^{M} R_{ab}, \quad B_1 = -\frac{1}{2}\frac{1}{NM}\sum_{i=1}^{N}\sum_{k=1}^{M}\sum_{l\neq k}^{M} R_{ik}R_{il}, \quad B_2 = -\frac{1}{2}\frac{1}{NM}\sum_{i=1}^{N}\sum_{j\neq i}^{N}\sum_{k=1}^{M} R_{ik}R_{jk}, \\ B_3 &= -\frac{1}{2}\left(\frac{1}{NM}\sum_{i=1}^{N}\sum_{k=1}^{M} R_{ik}^2 - \sigma_R^2\right), \quad B_4 = \frac{1}{2}\frac{N+M-1}{NM}\left(1 + \frac{1}{NM}\sum_{i=1}^{N}\sum_{j\neq i}^{N}\sum_{k=1}^{M}\sum_{l\neq k}^{M} R_{ik}R_{jl}\right), \\ B_5 &= \frac{3}{8}\left(\frac{1}{NM}\sum_{i=1}^{N}\sum_{k=1}^{M}\sum_{l\neq k}^{M} R_{ik}R_{il} + \frac{1}{NM}\sum_{i=1}^{N}\sum_{j\neq i}^{N}\sum_{k=1}^{M} R_{ik}R_{jk}\right)^2. \end{aligned}$$

Then tedious but direct calculations yield

$$\begin{aligned} \kappa_{1,n} &= -\frac{1}{2}(NM)^{-1/2}E[R_{ik}^3] + o((NM)^{-1/2}), \quad \kappa_{2,n} = 1 + \frac{3}{N} + \frac{3}{M} + o((NM)^{-1/2}), \\ \kappa_{3,n} &= -2(NM)^{-1/2}E[R_{ij}^3] + o((NM)^{-1/2}), \quad \kappa_{4,n} = \frac{6}{N} + \frac{6}{M} + o((NM)^{-1/2}), \end{aligned}$$

and the conclusion for $T(\theta)$ follows.

We next consider the modified MEL statistic $\ell^m(\theta)$. By proceeding as in DiCiccio et al. (1994) with $\sigma_R^2 = 1$, we obtain the signed root expansion

$$\sqrt{\ell^m(\theta)} = A\{1 + B_1 + B_2 + B_3 + B_4 + B_5 + C\} + o_p((NM)^{-1/2}), \quad \text{(A.10)}$$

where

$$C = -\frac{1}{2}\left(\frac{3}{N} + \frac{3}{M} - \frac{5}{n}\right) + \frac{E[R_{ij}^3]}{3\sqrt{2}}\frac{1}{\sqrt{NM}}A - \left(\frac{\sqrt{2}-1}{2}\right)\frac{1}{n}A^2.$$

Then the cumulants of $\sqrt{\ell^m(\theta)}$ are obtained as

$$\begin{aligned}\kappa_{1,n}^m &= -\frac{1}{2}\frac{1}{\sqrt{NM}}E[R_{ij}^3]+o((NM)^{-1/2}), \quad \kappa_{2,n}^m = 1+\frac{8-3\sqrt{2}}{n}+o((NM)^{-1/2}),\\ \kappa_{3,n}^m &= -\frac{2}{\sqrt{NM}}E[R_{ij}^3]+o((NM)^{-1/2}), \quad \kappa_{4,n}^m = \frac{6}{N}+\frac{6}{M}-\frac{12(\sqrt{2}-1)}{n}+o((NM)^{-1/2}).\end{aligned}$$

By proceeding as in Chapter 2 of Hall (1992), we have

$$\Pr\{\sqrt{\ell^m(\theta)}\le t\} = \Phi(t)-\left\{(\kappa_{2,n}^m-1)t+\frac{\kappa_{4,n}^m}{12}(t^3-3t)\right\}\phi(t)+o((NM)^{-1/2}),$$

and the conclusion for $\sqrt{\ell^m(\theta)}$ follows.

A.4. **Proof of Theorem 4.** Here we focus on the case of $\theta=(\alpha,\beta')'$. The case where θ is a subvector of $(\alpha,\beta')'$ can be shown analogously by applying the argument in Qin and Lawless (1994). Let $\hat{\theta}=(\hat{\alpha},\hat{\beta}')'$, $D_{ij}=(1,Z'_{ij})'$, $s_{ij}(\theta)=(Y_{ij}-\Lambda(D'_{ij}\theta))D_{ij}$, and

$$H_n(\theta)=\frac{1}{NM}\sum_{i=1}^N\sum_{j=1}^M \nabla_\theta s_{ij}(\theta).$$

Step 1. In this step, we shall derive the asymptotic distribution of $S(\theta)=\hat{\theta}-\theta$. Following Section 3 in Graham (2020), using the first-order condition for $\hat{\theta}$ and a mean value expansion yield that for some $\overline{\theta}$ that has each of its components lie between the corresponding components of $\hat{\theta}$ and θ, it holds

$$S(\theta)=\hat{\theta}-\theta=\{nH_n(\bar{\theta})\}^{-1}\frac{n}{NM}\sum_{i=1}^N\sum_{j=1}^M s_{ij}(\theta).$$

For the inverse factor in this expression, Appendix A of Graham (2020) under Assumption 2 gives $\{nH_n(\bar{\theta})\}^{-1}\overset{p}{\to}H^{-1}$. For the linear component, we need further notation. Define $U_{i0}=(W_{i0},A_{i0})$, $U_{0j}=(W_{0j},A_{0j})$, and

$$\begin{aligned}L_{i0} &= E[s_{i1}(\theta_0)|U_{i0}], \qquad L_{0j}=E[s_{1j}(\theta_0)|U_{0j}],\\ W_{ij} &= E[s_{ij}(\theta_0)|U_{i0},U_{0j}]-E[s_{i1}(\theta_0)|U_{i0}]-E[s_{1j}(\theta_0)|U_{0j}],\\ R_{ij} &= s_{ij}(\theta_0)-E[s_{ij}(\theta_0)|U_{i0},U_{0j}],\\ \sigma_L^2 &= n\{Var(L_{10})/N+Var(L_{01})/M\}, \qquad \sigma_R^2=Var(R_{11})\\ \Sigma &= \lim_{n\to\infty} n^2H^{-1}\{\sigma_L^2+\sigma_R^2/\underline{n}\}H^{-1}, \qquad \Sigma_m=\lim_{n\to\infty} n^2H^{-1}\{\sigma_L^2+2\sigma_R^2/\underline{n}\}H^{-1}.\end{aligned}$$

Under Assumption 2, by Theorem 1 in Graham (2020) and the discussion there-before, we obtain

$$\begin{aligned}\sqrt{n}S(\theta) &= \frac{n^{3/2}}{NM}H^{-1}\sum_{i=1}^N\sum_{j=1}^M s_{ij}(\theta)+o_p(1)\\ &= n^{3/2}H^{-1}\left\{\frac{1}{N}\sum_{i=1}^N L_{i0}+\frac{1}{M}\sum_{j=1}^M L_{0j}+\frac{1}{NM}\sum_{i=1}^N\sum_{j=1}^M R_{ij}\right\}+o_p(1)\\ &\overset{d}{\to} N(0,\Sigma), \qquad\qquad (A.11)\end{aligned}$$

as well as that $n^{3/2}(NM)^{-1}H^{-1}\sum_{i=1}^{N}\sum_{j=1}^{M}W_{ij} = o_p(1)$.

Step 2. In this step, we shall prove

$$\frac{1}{\sqrt{n}}\sum_{l=1}^{n}V_l(\theta) = \sqrt{n}S(\theta) + o_p(1). \tag{A.12}$$

Since

$$\frac{1}{\sqrt{n}}\sum_{l=1}^{n}V_l(\theta) = \sqrt{n}S(\theta) - \frac{n-1}{\sqrt{n}}\sum_{l=1}^{n}\{S_l(\theta) - S(\theta)\},$$

it sufficient for (A.12) to show $\sum_{l=1}^{n}S_l(\theta) = nS(\theta) + o_p(n^{-1/2})$. By a fourth-order Taylor expansion, it holds

$$\begin{aligned}
S(\theta) &= \hat{\theta} - \theta \\
&= \{nH_n(\theta)\}^{-1}\frac{n}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M}\Bigg\{s_{ij}(\theta) + \sum_{k=1}^{d}\frac{\nabla_{\theta\theta_k}s_{ij}(\theta)}{2}(\hat{\theta}-\theta)(\hat{\theta}_k-\theta_k) \\
&\quad + \sum_{k=1}^{d}\sum_{\ell=1}^{d}\frac{\nabla_{\theta\theta_k\theta_\ell}s_{ij}(\theta)}{3!}(\hat{\theta}-\theta)(\hat{\theta}_k-\theta_k)(\hat{\theta}_\ell-\theta_\ell) \\
&\quad + \sum_{k=1}^{d}\sum_{\ell=1}^{d}\sum_{m=1}^{d}\frac{\nabla_{\theta\theta_k\theta_\ell\theta_m}s_{ij}(\tilde{\theta})}{4!}(\hat{\theta}-\theta)(\hat{\theta}_k-\theta_k)(\hat{\theta}_\ell-\theta_\ell)(\hat{\theta}_m-\theta_m)\Bigg\} \\
&= \{nH_n(\theta)\}^{-1}\frac{n}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M}\Bigg\{s_{ij}(\theta) + \sum_{k=1}^{d}\frac{\nabla_{\theta\theta_k}s_{ij}(\theta)}{2}(\hat{\theta}-\theta)(\hat{\theta}_k-\theta_k) \\
&\quad + \sum_{k=1}^{d}\sum_{\ell=1}^{d}\frac{\nabla_{\theta\theta_k\theta_\ell}s_{ij}(\theta)}{3!}(\hat{\theta}-\theta)(\hat{\theta}_k-\theta_k)(\hat{\theta}_\ell-\theta_\ell)\Bigg\} + O_p(n^{-2}) \\
&= \frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M}X_{ij} + O_p(n^{-2}),
\end{aligned} \tag{A.13}$$

for $\tilde{\theta} = (\tilde{\theta}_k)_{k=1}^{d}$ with each $\tilde{\theta}_k$ lies between $\hat{\theta}_k$ and θ_k, where

$$X_{ij} = n\{nH_n(\theta)\}^{-1}\left(s_{ij}(\theta) + \sum_{k=1}^{d}\frac{\nabla_{\theta\theta_k}s_{ij}(\theta)}{2}(\hat{\theta}-\theta)(\hat{\theta}_k-\theta_k) + \sum_{k=1}^{d}\sum_{\ell=1}^{d}\frac{\nabla_{\theta\theta_k\theta_\ell}s_{ij}(\theta)}{3!}(\hat{\theta}-\theta)(\hat{\theta}_k-\theta_k)(\hat{\theta}_\ell-\theta_\ell)\right).$$

We shall now establish similar asymptotic representations for the leave one column or row out counterparts. Denote

$$H_{n-1}^{l}(\theta) = \begin{cases} \frac{1}{(N-1)M}\sum_{i\neq l}\sum_{j=1}^{M}\nabla_\theta s_{ij}(\theta), & \text{if } l \leq N, \\ \frac{1}{N(M-1)}\sum_{i=1}^{N}\sum_{j\neq l}\nabla_\theta s_{ij}(\theta) & \text{otherwise.} \end{cases}$$

For each $l \leq N$, we have

$$\begin{aligned}
S_l(\theta) &= \hat{\theta}^{(l)} - \theta \\
&= \{nH_{n-1}^{l}(\theta)\}^{-1}\frac{n}{(N-1)M}\sum_{i\neq l}\sum_{j=1}^{M}\Bigg\{s_{ij}(\theta) + \sum_{k=1}^{d}\frac{\nabla_{\theta\theta_k}s_{ij}(\theta)}{2}(\hat{\theta}^{(l)}-\theta)(\hat{\theta}_k^{(l)}-\theta_k)
\end{aligned}$$

$$+\sum_{k=1}^{d}\sum_{\ell=1}^{d}\frac{\nabla_{\theta\theta_k\theta_\ell}s_{ij}(\theta)}{3!}(\hat{\theta}^{(l)}-\theta)(\hat{\theta}_k^{(l)}-\theta_k)(\hat{\theta}_\ell^{(l)}-\theta_\ell)$$

$$+\sum_{k=1}^{d}\sum_{\ell=1}^{d}\sum_{m=1}^{d}\frac{\nabla_{\theta\theta_k\theta_\ell\theta_m}s_{ij}(\tilde{\theta}^{(l)})}{4!}(\hat{\theta}^{(l)}-\theta)(\hat{\theta}_k^{(l)}-\theta_k)(\hat{\theta}_\ell^{(l)}-\theta_\ell)(\hat{\theta}_m^{(l)}-\theta_m)\Bigg\}$$

$$= \{nH_n(\theta)\}^{-1}\frac{n}{(N-1)M}\sum_{i\neq l}\sum_{j=1}^{M}\Bigg\{s_{ij}(\theta)+\sum_{k=1}^{d}\frac{\nabla_{\theta\theta_k}s_{ij}(\theta)}{2}(\hat{\theta}^{(l)}-\theta)(\hat{\theta}_k^{(l)}-\theta_k)$$

$$+\sum_{k=1}^{d}\sum_{\ell=1}^{d}\frac{\nabla_{\theta\theta_k\theta_\ell}s_{ij}(\theta)}{3!}(\hat{\theta}^{(l)}-\theta)(\hat{\theta}_k^{(l)}-\theta_k)(\hat{\theta}_\ell^{(l)}-\theta_\ell)\Bigg\}+O_p(n^{-2})$$

$$-\left(\{nH_n(\theta)\}^{-1}-\{nH_{n-1}^l(\theta)\}^{-1}\right)\frac{n}{(N-1)M}\sum_{i\neq l}\sum_{j=1}^{M}s_{ij}(\theta)(1+o_p(1)), \quad \text{(A.14)}$$

where $\tilde{\theta}^{(l)}$ are vectors of mean values. Note that the last term of (A.14) has an order of $O_p(n^{-2})$ following the fact that $(B^{-1}-A^{-1})=B^{-1}(A-B)A^{-1}$ for A and B invertible, the fact that $nH_n(\theta)$ and $nH_{n-1}(\theta)$ have their eigenvalues bounded and bounded away from zero with probability approaching one as H is of full rank, the asymptotic order $(n/NM)\sum_{i\neq l}\sum_{j=1}^{M}s_{ij}(\theta)=O_p(n^{-1/2})$ implied by the asymptotic normality in Step 1, and the claim that $n(H_n(\theta)-H_{n-1}^l(\theta))=O_p(n^{-3/2})$, which we shall show next. To show this claim, note that for each $l\leq N$,

$$n(H_n(\theta)-H_{n-1}^l(\theta)) = -\frac{n}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M}\nabla_\theta s_{ij}(\theta)+\frac{n}{(N-1)M}\sum_{i\neq l}\sum_{j=1}^{M}\nabla_\theta s_{ij}(\theta)$$

$$= -\frac{n}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M}\nabla_\theta s_{ij}(\theta)+\frac{n}{(N-1)M}\sum_{i=1}^{N}\sum_{j=1}^{M}\nabla_\theta s_{ij}(\theta)-\frac{n}{(N-1)M}\sum_{j=1}^{M}\nabla_\theta s_{lj}(\theta)$$

$$= -\frac{1}{N-1}\left(\frac{1}{M}\sum_{j=1}^{M}\{n\nabla_\theta s_{lj}(\theta)-H\}-\frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M}\{n\nabla_\theta s_{ij}(\theta)-H\}\right)$$

$$= O_p(n^{-1})O_p(n^{-1/2})=O_p(n^{-3/2}), \quad \text{(A.15)}$$

where the last equality follows from applications of the maximal inequalities for separately exchangeable arrays (Chiang et al., 2021, Corollary 3) and for i.i.d. random variables (van der Vaart and Wellner, 1996, Theorem 2.14.1), as well as the fact that $nE[\nabla_\theta s_{11}(\theta)]=H$. Hence (A.14) becomes

$$\hat{\theta}^{(l)}-\theta = \{nH_n(\theta)\}^{-1}\frac{n}{(N-1)M}\sum_{i\neq l}\sum_{j=1}^{M}\Bigg\{s_{ij}(\theta)+\sum_{k=1}^{d}\frac{\nabla_{\theta\theta_k}s_{ij}(\theta)}{2}(\hat{\theta}^{(l)}-\theta)(\hat{\theta}_k^{(l)}-\theta_k)$$

$$+\sum_{k=1}^{d}\sum_{\ell=1}^{d}\frac{\nabla_{\theta\theta_k\theta_\ell}s_{ij}(\theta)}{3!}(\hat{\theta}^{(l)}-\theta)(\hat{\theta}_k^{(l)}-\theta_k)(\hat{\theta}_\ell^{(l)}-\theta_\ell)\Bigg\}+O_p(n^{-2})$$

$$= \{nH_n(\theta)\}^{-1}\frac{n}{(N-1)M}\sum_{i\neq l}\sum_{j=1}^{M}\Bigg\{s_{ij}(\theta)+\sum_{k=1}^{d}\frac{\nabla_{\theta\theta_k}s_{ij}(\theta)}{2}(\hat{\theta}-\theta)(\hat{\theta}_k-\theta_k)$$

$$+\sum_{k=1}^{d}\sum_{\ell=1}^{d}\frac{\nabla_{\theta\theta_k\theta_\ell}s_{ij}(\theta)}{3!}(\hat{\theta}-\theta)(\hat{\theta}_k-\theta_k)(\hat{\theta}_\ell-\theta_\ell)\Bigg\}$$

$$
\begin{aligned}
&+O_p(n^{-2}) + O_p(1) \cdot \left| \|\hat{\theta}^{(l)} - \theta\|^2 - \|\hat{\theta} - \theta\|^2 \right| \\
= \quad & \frac{1}{(N-1)M} \sum_{i \neq l} \sum_{j=1}^{M} X_{ij} + O_p(n^{-2}) + O_p(1) \cdot \left| \|\hat{\theta}^{(l)} - \theta\|^2 - \|\hat{\theta} - \theta\|^2 \right|. \qquad \text{(A.16)}
\end{aligned}
$$

To obtain a concrete bound for the asymptotic order in (A.16), note that by the fact that $a^2 - b^2 = (a+b)(a-b)$, we have

$$
\begin{aligned}
\left| \|\hat{\theta}^{(l)} - \theta\|^2 - \|\hat{\theta} - \theta\|^2 \right| \leq & \left| \|\hat{\theta}^{(l)} - \theta\| - \|\hat{\theta} - \theta\| \right| \cdot \left| \|\hat{\theta}^{(l)} - \theta\| + \|\hat{\theta} - \theta\| \right| \\
\leq & \|\hat{\theta}^{(l)} - \hat{\theta}\| \cdot \left| \|\hat{\theta}^{(l)} - \theta\| + \|\hat{\theta} - \theta\| \right| = \|\hat{\theta}^{(l)} - \hat{\theta}\| \cdot O_p(n^{-1/2}), \qquad \text{(A.17)}
\end{aligned}
$$

where the second inequality is implied by the reverse triangle inequality for norms $|\|a\| - \|b\|| \leq \|a - b\|$. We now proceed by the following recursive argument. Observe that using the coarse representations implied by the earlier Taylor expansions, it holds that

$$
\begin{aligned}
\hat{\theta} - \theta \quad &= \quad \{nH_n(\theta)\}^{-1} \frac{n}{NM} \sum_{i=1}^{N} \sum_{j=1}^{M} s_{ij}(\theta) + O_p(n^{-1}), \\
\hat{\theta}^{(l)} - \theta \quad &= \quad \begin{cases} \{nH_n(\theta)\}^{-1} \frac{n}{(N-1)M} \sum_{i \neq l} \sum_{j=1}^{M} s_{ij}(\theta) + O_p(n^{-1}) & \text{if } l \leq N, \\ \{nH_n(\theta)\}^{-1} \frac{n}{N(M-1)} \sum_{i=1}^{N} \sum_{j \neq l} s_{ij}(\theta) + O_p(n^{-1}) & \text{otherwise.} \end{cases}
\end{aligned}
$$

Following the same argument as (A.15), we have $\|\hat{\theta}^{(l)} - \hat{\theta}\| = O_p(n^{-3/2}) + O_p(n^{-1}) = O_p(n^{-1})$. This implies that (A.17) is of order $O_p(n^{-3/2})$ and thus (A.16) gives

$$
\hat{\theta}^{(l)} - \theta = \frac{1}{(N-1)M} \sum_{i \neq l} \sum_{j=1}^{M} X_{ij} + O_p(n^{-3/2}).
$$

Now, using this representation and (A.13), again following the same argument as (A.15), we have $\|\hat{\theta}^{(l)} - \hat{\theta}\| = O_p(n^{-3/2})$ and thus (A.17) becomes $O_p(n^{-2})$. Plugging this into (A.16) once more, we have the representation

$$
S_l(\theta) = \hat{\theta}^{(l)} - \theta = \frac{1}{(N-1)M} \sum_{i \neq l} \sum_{j=1}^{M} X_{ij} + O_p(n^{-2}).
$$

Likewise, for $l > N$, a symmetric argument shows that

$$
S_l(\theta) = \hat{\theta}^{(l)} - \theta = \frac{1}{N(M-1)} \sum_{i=1}^{N} \sum_{j \neq l} X_{ij} + O_p(n^{-2}).
$$

Utilizing these asymptotic representations yield that

$$
\begin{aligned}
\sum_{l=1}^{n} S_l(\theta) = & \frac{1}{(N-1)M} \sum_{l=1}^{N} \left(\sum_{i=1}^{N} \sum_{j=1}^{M} X_{ij} - \sum_{j=1}^{M} X_{lj} \right) + \frac{1}{N(M-1)} \sum_{l_1=1}^{M} \left(\sum_{i=1}^{N} \sum_{j=1}^{M} X_{ij} - \sum_{i=1}^{N} X_{il_1} \right) + O_p(n \cdot n^{-2}) \\
= & \left\{ \frac{NM(N-1)}{(N-1)M} + \frac{NM(M-1)}{(M-1)N} \right\} S(\theta) + O_p(n^{-1}) = nS(\theta) + o_p(n^{-1/2}),
\end{aligned}
$$

as required.

Step 3. In this step, we shall verify

$$\frac{1}{n}\sum_{l=1}^{n} V_l(\theta)V_l(\theta)' \xrightarrow{p} \Sigma_m, \qquad \frac{1}{n}\sum_{l=1}^{N}\sum_{l_1=1}^{M} Q_{ll_1}Q_{ll_1}' \xrightarrow{p} \lim_{n\to\infty} n^2H^{-1}(\sigma_R^2/\underline{n})H^{-1}.$$

Using the Hoeffding type decomposition in (A.11), we have

$$S_l(\theta)-S_{l_1}(\theta)=\begin{cases}\frac{nH^{-1}}{N-1}(L_{l_10}-L_{l0})+\frac{nH^{-1}}{(N-1)M}\sum_{j=1}^{M}(R_{l_1j}-R_{lj})+o_p(n^{-1/2}) & \text{if } l<l_1\le N,\\ \frac{nH^{-1}}{M}L_{0(l_1-N)}-\frac{nH^{-1}}{N}L_{l0}+\frac{nH^{-1}}{N(M-1)}\sum_{i=1}^{N}R_{i(l_1-N)}-\frac{nH^{-1}}{(N-1)M}\sum_{j=1}^{M}R_{lj}+o_p(n^{-1/2}) & \text{if } l\le N<l_1,\\ \frac{nH^{-1}}{M-1}(L_{0(l_1-N)}-L_{0(l-N)})+\frac{nH^{-1}}{N(M-1)}\sum_{i=1}^{N}(R_{i(l_1-N)}-R_{i(l-N)})+o_p(n^{-1/2}) & \text{if } N<l<l_1.\end{cases}$$

Using this and (A.12), it holds

$$\begin{aligned}
&\frac{1}{n}\sum_{l=1}^{n}V_l(\theta)V_l(\theta)'=\frac{1}{n}\sum_{l=1}^{n}\{S(\theta)-(n-1)(S_l(\theta)-S(\theta))\}\{S(\theta)-(n-1)(S_l(\theta)-S(\theta))\}'\\
=\;&\frac{(n-1)^2}{n}\sum_{l=1}^{n}\{S_l(\theta)-S(\theta)\}\{S_l(\theta)-S(\theta)\}'+o_p(1)\\
=\;&\frac{(n-1)^2}{n^2}\sum_{l<l_1}\{S_l(\theta)-S_{l_1}(\theta)\}\{S_l(\theta)-S_{l_1}(\theta)\}'+o_p(1)\\
=\;&\frac{2n^2}{N^2}\sum_{i<j\le N}\frac{H^{-1}(L_{i0}-L_{j0})(L_{i0}-L_{j0})'H^{-1}}{2}+\frac{2n^2}{M^2}\sum_{i<j\le M}\frac{H^{-1}(L_{0i}-L_{0j})(L_{0i}-L_{0j})'H^{-1}}{2}\\
&+n^2\sum_{i=1}^{N}\sum_{j=1}^{M}H^{-1}\left(\frac{L_{0j}L_{0j}'}{M^2}+\frac{L_{i0}L_{i0}'}{N^2}\right)H^{-1}+\frac{n^2}{N^2M^2}\sum_{i<j\le N}\sum_{l=1}^{M}H^{-1}(R_{jl}-R_{il})(R_{jl}-R_{il})'H^{-1}\\
&+\frac{n^2}{N^2M^2}\sum_{i<j\le M}\sum_{k=1}^{N}H^{-1}(R_{kj}-R_{ki})(R_{kj}-R_{ki})'H^{-1}\\
&+\frac{n^2}{N^2M^2}\sum_{i=1}^{N}\sum_{j=1}^{M}H^{-1}\left(\sum_{k=1}^{N}R_{kj}-\sum_{l=1}^{M}R_{il}\right)\left(\sum_{k=1}^{N}R_{kj}-\sum_{l=1}^{M}R_{il}\right)'H^{-1}+o_p(1)\\
=\;&n^2H^{-1}\left\{\left(1+\frac{N}{M}\right)Var(L_{10})+\left(1+\frac{M}{N}\right)Var(L_{01})+\left(\frac{1}{M}+\frac{1}{N}+\frac{n}{NM}\right)Var(R_{11})\right\}H^{-1}+o_p(1)\\
\xrightarrow{p}\;&\Sigma_m,
\end{aligned}$$

where the last equality follows from the weak law of large numbers. On the other hand, using an analogous argument as in the proof of Theorem 2, we have

$$\begin{aligned}
&\frac{1}{n}\sum_{l=1}^{N}\sum_{l_1=1}^{M}Q_{ll_1}Q_{ll_1}'\\
=\;&\frac{1}{n}\sum_{l=1}^{N}\sum_{l_1=1}^{M}\left\{\frac{n^2}{NM}H^{-1}R_{ll_1}+O_p(\underline{n}^{-3/2})\right\}\left\{\frac{n^2}{NM}H^{-1}R_{ll_1}+O_p(\underline{n}^{-3/2})\right\}'+o_p(1)\\
\xrightarrow{p}\;&\lim_{n\to\infty}n^2H^{-1}(\sigma_R^2/\underline{n})H^{-1}.
\end{aligned}$$

Step 4. Combining the results in Steps 1-3 with the same argument in the proof of Theorem 2 and the fact that $\sum_{l=1}^{n} S_l(\hat{\theta}) = \sum_{l=1}^{n}\{S_l(\theta) - S(\theta)\}$, we obtain

$$\frac{1}{\sqrt{n}}\sum_{l=1}^{n} V_l^m(\theta) \xrightarrow{d} N(0, \Sigma_m), \qquad \frac{1}{n}\sum_{l=1}^{n} V_l^m(\theta)V_l^m(\theta)' \xrightarrow{p} \Sigma_m.$$

The rest then follows straightforwardly from the linearization argument as in Step 1 in the proof of Theorem 1.

References

Balan, R. M. and I. Schiopu-Kratina (2005): "Asymptotic results with generalized estimating equations for longitudinal data," *Annals of Statistics*, 33, 522–541.

Berger, Y. G. and O. De La Riva Torres (2016): "Empirical likelihood confidence intervals for complex sampling designs," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78, 319–341.

Bertail, P. (2006): "Empirical likelihood in some semiparametric models," *Bernoulli*, 12, 299–331.

Bhattacharya, D. (2005): "Asymptotic inference from multi-stage samples," *Journal of Econometrics*, 126, 145–171.

Bhattacharyya, S. and P. J. Bickel (2015): "Subsampling bootstrap of count features of networks," *Annals of Statistics*, 43, 2384–2411.

Bickel, P. J., A. Chen, and E. Levina (2011): "The method of moments and degree distributions for network models," *Annals of Statistics*, 39, 2280–2301.

Bravo, F., J. C. Escanciano, and I. Van Keilegom (2020): "Two-step semiparametric empirical likelihood inference," *Annals of Statistics*, 48, 1–26.

Cameron, C. A., J. B. Gelbach, and D. L. Miller (2011): "Robust inference with multiway clustering," *Journal of Business & Economic Statistics*, 29, 238–249.

Caron, F. and E. B. Fox (2017): "Sparse graphs using exchangeable random measures," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79, 1295.

Chauvet, G. (2015): "Coupling methods for multistage sampling," *Annals of Statistics*, 43, 2484–2506.

Chauvet, G. and A.-A. Vallée (2020): "Inference for two-stage sampling designs," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82, 797–815.

Chen, S. X. and I. Van Keilegom (2009): "A review on empirical likelihood methods for regression," *Test*, 18, 415–447.

Chiang, H. D., K. Kato, and Y. Sasaki (2021): "Inference for high-dimensional exchangeable arrays," *arXiv preprint arXiv:2009.05150*.

Choi, D. (2017): "Co-clustering of nonsmooth graphons," *Annals of Statistics*, 45, 1488–1515.

Choi, D. and P. J. Wolfe (2014): "Co-clustering separately exchangeable network data," *Annals of Statistics*, 42, 29–63.

Davezies, L., X. D'Haultfoeuille, and Y. Guyonvarch (2021): "Empirical process results for exchangeable arrays," *Annals of Statistics*, 49, 845–862.

Diaconis, P. and S. Janson (2008): "Graph limits and exchangeable random graphs," *Rendiconti di Matematica e delle sue Applicazioni*, 28, 33–61.

DiCiccio, T. J., P. Hall, and J. Romano (1994): "Empirical likelihood is Bartlett-correctable," *Annals of Statistics*, 19, 1053–1061.

Efron, B. and C. Stein (1981): "The jackknife estimate of variance," *Annals of Statistics*, 586–596.

Fisher, R. A. (1922): "On the mathematical foundations of theoretical statistics," *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 222, 309–368.

Fuller, W. A. (2011): *Sampling Statistics*, vol. 560, John Wiley & Sons.

Gao, C., Y. Lu, and H. H. Zhou (2015): "Rate-optimal graphon estimation," *Annals of Statistics*, 43, 2624–2652.

Gong, Y., L. Peng, and Y. Qi (2010): "Smoothed jackknife empirical likelihood method for ROC curve," *Journal of Multivariate Analysis*, 101, 1520–1531.

Graham, B. S. (2020): "Sparse network asymptotics for logistic regression," *working paper*.

Graham, B. S., F. Niu, and J. L. Powell (2019): "Kernel density estimation for undirected dyadic data," *arXiv preprint arXiv:1907.13630*.

Hall, P. (1992): *The bootstrap and Edgeworth expansion*, Springer.

Hinkley, D. V. (1978): "Improving the jackknife with special reference to correlation estimation," *Biometrika*, 65, 13–21.

Hjort, N. L., I. W. McKeague, and I. Van Keilegom (2009): "Extending the scope of empirical likelihood," *Annals of Statistics*, 37, 1079–1111.

Jing, B.-Y., J. Yuan, and W. Zhou (2009): "Jackknife empirical likelihood," *Journal of the American Statistical Association*, 104, 1224–1232.

Kallenberg, O. (2006): *Probabilistic Symmetries and Invariance Principles*, Springer Science & Business Media.

Lauritzen, S., A. Rinaldo, and K. Sadeghi (2018): "Random networks, graphical models and exchangeability," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80, 481–508.

Liang, K.-Y. and S. L. Zeger (1986): "Longitudinal data analysis using generalized linear models," *Biometrika*, 73, 13–22.

MacKinnon, J. G., M. Ø. Nielsen, and M. D. Webb (2021): "Wild bootstrap and asymptotic inference with multiway clustering," *Journal of Business & Economic Statistics*, 39, 505–519.

Matsushita, Y. and T. Otsu (2021): "Jackknife empirical likelihood: small bandwidth, sparse network and high-dimension asymptotic," *Biometrika*, forthcoming.

McCullagh, P. (2000): "Resampling and exchangeable arrays," *Bernoulli*, 6, 285–301.

Menzel, K. (2021): "Bootstrap with clustering in two or more dimensions," *Econometrica*, To appear.

Miglioretti, D. L. and P. J. Heagerty (2007): "Marginal modeling of nonnested multilevel data using standard software," *American Journal of Epidemiology*, 165, 453–463.

Orbanz, P. and D. M. Roy (2014): "Bayesian models of graphs, arrays and other exchangeable random structures," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37, 437–461.

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

——— (1990): "Empirical likelihood ratio confidence regions," *Annals of Statistics*, 90–120.

——— (2001): *Empirical Likelihood*, CRC press.

——— (2007): "The pigeonhole bootstrap," *Annals of Applied Statistics*, 1, 386–411.

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

Searle, S. R., G. Casella, and C. E. McCulloch (2009): *Variance Components*, John Wiley & Sons.

Thompson, S. B. (2011): "Simple formulas for standard errors that cluster by both firm and time," *Journal of Financial Economics*, 99, 1–10.

van der Vaart, A. W. and J. A. Wellner (1996): *Weak Convergence and Empirical Processes*, Springer.

Veitch, V. and D. M. Roy (2019): "Sampling and estimation for (sparse) exchangeable graphs," *Annals of Statistics*, 47, 3274–3299.

White, H. (2001): *Asymptotic Theory for Econometricians*, Academic press.

Xie, M. and Y. Yang (2003): "Asymptotics for generalized estimating equations with large cluster sizes," *Annals of Statistics*, 31, 310–347.

Zhang, Y., E. Levina, and J. Zhu (2017): "Estimating network edge probabilities by neighbourhood smoothing," *Biometrika*, 104, 771–783.

Zhang, Z. and Y. Zhao (2013): "Empirical likelihood for linear transformation models with interval-censored failure time data," *Journal of Multivariate Analysis*, 116, 398–409.

Zhao, P., M. Ghosh, J. Rao, and C. Wu (2020): "Bayesian empirical likelihood inference with complex survey data," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82, 155–174.

Zhong, P.-S. and S. Chen (2014): "Jackknife empirical likelihood inference with regression imputation and survey data," *Journal of Multivariate Analysis*, 129, 193–205.

Zhu, L. and L. Xue (2006): "Empirical likelihood confidence regions in a partially linear single-index model," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68, 549–570.

Department of Economics, University of Wisconsin-Madison, 1180 Observatory Drive Madison, WI 53706-1393, USA.

Email address: `hdchiang@wisc.edu`

Graduate School of Economics, Hitotsubashi University, 2-1 Naka, Kunitachi, Tokyo 186-8601, Japan.

Email address: `matsushita.y@r.hit-u.ac.jp`

Department of Economics, London School of Economics, Houghton Street, London, WC2A 2AE, UK.

Email address: `t.otsu@lse.ac.uk`