A GMM Approach in Coupling Internal Data and External Summary Information with Heterogeneous Data Populations

Jun Shao[1] [2], Jinyi Wang[2], and Lei Wang[*3]

Abstract

Because of advances in data collection and storage, statistical analysis in modern scientific research and practice now has opportunities to utilize external information such as summary statistics from similar studies. A likelihood approach based on a parametric model assumption has been developed in the literature to utilize external summary information when the populations for external data and the main internal data are assumed to be the same. In this article we instead consider the generalized estimation equation (GEE) approach for statistical inference, which is semiparametric or nonparametric, and show how to utilize external summary information even when internal and external data populations are not the same. Our approach is coupling the internal data and external summary information to form additional estimation equations, and then applying the generalized method of moments (GMM). We show that the proposed GMM estimator is asymptotically normal and, under some conditions, is more efficient than the GEE estimator without using external summary information. Estimators of asymptotic covariance matrix of the GMM estimators are also proposed. Simulation results are obtained to confirm our theory and to quantify the improvements from utilizing external data. An example is also included for illustration.

Keywords: Adjustment for heterogeneity, constraints, data integration, generalized method of moments, summary statistics.

[1]KLATASDS-MOE, School of Statistics, East China Normal University

[2]Department of Statistics, University of Wisconsin-Madison

[3]School of Statistics and Data Science, KLMDASR, LEBPS and LPMC, Nankai University

Corresponding to Dr. Lei Wang. Email: `lwangstat@nankai.edu.cn`

1 Introduction

In modern statistical analyses we have not only primary individual-level data (referred to as the internal data in what follows) carefully collected from a population of interest but also summary or aggregated information from some independent external datasets, for example, population-based census, administrative datasets, and data from past investigations or other similar studies. Due to various practical reasons, individual-level data from external sources are not available. For simplicity of notation, we consider a single external dataset, since extensions to multiple external datasets are straightforward. In both internal and external datasets, Y denotes a univariate response of interest and $\boldsymbol{X}$ is an associated covariate vector. The internal dataset contains an additional covariate vector $\boldsymbol{Z}$ (not in external dataset) because of new technology and/or new scientific relevance. The growing need for research in internal data analysis utilizing external information fits into the general framework of data integration (Merkouris, 2004; Chatterjee et al., 2016; Lohr and Raghunathan, 2017; Zhang et al., 2017; Yang et al., 2020; Yang and Kim, 2020; Zhang et al., 2020; Kim et al., 2021; Li et al., 2021; Rao, 2021; Tian and Feng, 2022) and is different from the meta-analysis, e.g., Lin and Zeng (2010), He et al. (2016), Kundu et al. (2019), Li et al. (2022), in which the analysis focuses on the same parameter in multiple datasets with summary statistics or individual data, not on a parameter in an internal individual-level dataset with an additional covariate $\boldsymbol{Z}$. Our goal is to couple the internal data and external summary information to improve estimation efficiency over the analysis using internal data only.

A distinction of our work from the existing papers in the literature is that we consider situations where only external summary statistics (not individual-level data) are available and $\boldsymbol{Z}$ is measured in internal dataset only but not external dataset, except for Chatterjee et al. (2016) and Zhang et al. (2020) whose difference with our work is described next.

To analyze internal data with additional covariate $\boldsymbol{Z}$ and external summary statistics,

Chatterjee et al. (2016) proposed a constrained maximum likelihood estimation under the following two key assumptions:

$$\text{The internal and external data have the same population:} \quad f(y, \boldsymbol{x}, \boldsymbol{z} \,|\, D = 1) = f(y, \boldsymbol{x}, \boldsymbol{z} \,|\, D = 0). \tag{A1}$$

$$\text{The internal population has a correctly specified parametric model} \quad f(y \,|\, \boldsymbol{x}, \boldsymbol{z}, D = 1) = f_{\boldsymbol{\theta}}(y \,|\, \boldsymbol{x}, \boldsymbol{z}), \tag{B1}$$

where D is a binary indicator with $D = 1$ for internal datum and $D = 0$ for external datum, $f(\,\cdot \,|\, \cdot\,)$ is a generic notation for conditional probability density, $\boldsymbol{\theta}$ is a vector of unknown parameters, and $f_{\boldsymbol{\theta}}$ is known when $\boldsymbol{\theta}$ is known. Their approach is to maximize the parametric likelihood based on $f_{\boldsymbol{\theta}}(y \,|\, \boldsymbol{x}, \boldsymbol{z})$ and internal data, subject to the constraint

$$0 = \iiint u(y, \boldsymbol{x}, \boldsymbol{\varphi}) f_{\boldsymbol{\theta}}(y \,|\, \boldsymbol{x}, \boldsymbol{z}) f(\boldsymbol{x}, \boldsymbol{z} \,|\, D = 0)\, dy d\boldsymbol{x} d\boldsymbol{z}, \tag{C1}$$

where $u(\,\cdot\,)$ is a known function (based on a working model for external data) and $\boldsymbol{\varphi}$ is an unknown parameter vector. Under assumption (A1), conditioning $D = 0$ in (C1) can be ignored. To use (C1) as a constraint, we replace all integrals by empirical integrals based on internal data and $\boldsymbol{\varphi}$ by an estimate $\widehat{\boldsymbol{\varphi}}$ available as a summary statistic based on external data independent of internal data. Zhang et al. (2020) developed an improved approach, under basically the same setting and assumptions (A1) and (B1).

The first purpose of our paper is to relax the strong parametric model assumption (B1). We consider the generalized estimation equation (GEE) for estimating a parameter $\boldsymbol{\beta}$ of interest in the internal data population $f(y, \boldsymbol{x}, \boldsymbol{z} \,|\, D = 1)$. In the last three decades, the GEE approach has shown its great success in analysis without a fully parametric likelihood assumption. Our main effort is to derive a constraint relating the external summary information to the estimation of $\boldsymbol{\beta}$ in GEE (B2) specified in Section 2, which serves as a replacement of (C1) as (C1) depends on (B1). Details are presented in Section 2.

Since heterogeneity often exists among datasets, especially when internal data are col-

lected under a carefully designed study whereas external data are from past or different studies, it is crucial to relax assumption (A1) for a wider scope of application, which is the second purpose of our paper. It is challenging to do data coupling with different internal and external populations, similar to the problem with missing data in which the population of completed data may be different from the population of incomplete data (i.e., missingness is not at random), especially when we do not have external individual-level data. Chatterjee et al. (2016) and Yang et al. (2020) discussed the population heterogeneity, but assumed that individual-level external data are available. In Section 3 we start with a discussion on what is the assumption really needed for the success of method assuming (A1), i.e., the robustness against violation of (A1). We then relax assumption (A1) in two different ways, by linking the internal and external data with techniques in treating missing data (although data coupling has a different study goal from analysis with incomplete data). This link enables us to derive some constraints based on external summary statistics that can be utilized in the GEE for $\boldsymbol{\beta}$ with internal data.

In Section 4, some simulation results are presented to illustrate finite sample performance of the proposed method. We also illustrate our method using a real data example.

2 GMM under Homogeneous Data Populations

We start with a description of data, following the notation in Section 1. Consider two scenarios different in whether the internal and external sample sizes are random or nonrandom. In the first scenario, we have a random sample of size n (a known nonrandom integer), $(Y_i, \boldsymbol{X}_i, \boldsymbol{Z}_i, D_i)$, $i = 1, ..., n$, where $(Y_i, \boldsymbol{X}_i, \boldsymbol{Z}_i, D_i) \sim (Y, \boldsymbol{X}, \boldsymbol{Z}, D)$ with probability density $f(y, \boldsymbol{x}, \boldsymbol{z}, d)$, $D_i = 1$ indicates the observed internal data, and $D_i = 0$ indicates the unavailable external data, $i = 1, ..., n$. Although $\boldsymbol{Z}_i$ is not measured in the external dataset, we still include it as a potentially observable quantity. In this scenario, the observed internal sample size is $n_1 = \sum_{i=1}^{n} D_i$, which is random with expectation πn, $\pi = P(D = 1)$,

and the external sample size is $n_0 = n - n_1$. In the second scenario, the internal dataset is a random sample $(Y_i, \boldsymbol{X}_i, \boldsymbol{Z}_i, D_i = 1)$, $i = 1, ..., \pi n$, with a known nonrandom size πn and $(Y_i, \boldsymbol{X}_i, \boldsymbol{Z}_i) \sim f(y, \boldsymbol{x}, \boldsymbol{z} \,|\, D = 1)$, the external dataset is another independent random sample $(Y_i, \boldsymbol{X}_i, \boldsymbol{Z}_i, D_i = 0)$, $i = 1, ..., (1-\pi)n$, $(Y_i, \boldsymbol{X}_i, \boldsymbol{Z}_i) \sim f(y, \boldsymbol{x}, \boldsymbol{z} \,|\, D = 0)$, and the total sample size is n. In any scenario, we consider large sample analysis with $n \to \infty$, in which π and n_1 depend on n but the subscript n is omitted for simplicity. We focus on the situation where the dimension of $\boldsymbol{W} = (\boldsymbol{X}^\top, \boldsymbol{Z}^\top)^\top$ is fixed (does not depend on n) and less than n, where $\boldsymbol{A}^\top$ denotes the transpose $\boldsymbol{A}$. A discussion about high-dimensional $\boldsymbol{W}$ is given in the end of this section.

In this section, we assume (A1), i.e., the internal population $f(y, \boldsymbol{x}, \boldsymbol{z} \,|\, D = 1)$ and the external population $f(y, \boldsymbol{x}, \boldsymbol{z} \,|\, D = 0)$ are the same, but we replace the parametric model assumption (B1) with the following GEE to estimate an unknown parameter vector $\boldsymbol{\beta}$ of interest:

$$\mathbf{0} = E\left[\{Y - \phi\left(\boldsymbol{W}^\top \boldsymbol{\beta}\right)\}\boldsymbol{W} \,|\, D = 1\right], \tag{B2}$$

where ϕ is a known function and $\mathbf{0}$ denotes a vector of all zeros with an appropriate order. Conditioning $D = 1$ can be ignored in (B2) since (A1) is assumed but it is kept for extensions to the case without (A1) considered in Section 3.

This GEE approach is actually nonparametric as the expectation E in (B2) is jointly on $(Y, \boldsymbol{W})$ so that $\boldsymbol{\beta}$ is almost always defined. A semi-parametric GEE assumes a conditional mean model $E(Y \,|\, \boldsymbol{W}) = \phi(\boldsymbol{W}^\top \boldsymbol{\beta})$, which is stronger than (B2) but still much weaker than the parametric likelihood assumption (B1) described in Section 1, since it only specifies the conditional mean model.

Consider the use of external summary statistic, a function of $(Y_i, \boldsymbol{X}_i)$ with $D_i = 0$ for all i. The question is how to derive a constraint that relates the external information to the estimation of $\boldsymbol{\beta}$ via GEE (B2). From the description in Section 1, constraint (C1) relates external information to $\boldsymbol{\theta}$ through the correctly specified parametric likelihood $f_{\boldsymbol{\theta}}(y \,|\, \boldsymbol{x}, \boldsymbol{z})$

under (B1), which we cannot use since $f_{\boldsymbol{\theta}}(y\,|\,\boldsymbol{x},\boldsymbol{z})$ is not available without (B1).

Suppose that the external summary statistic is an estimate $\widehat{\boldsymbol{\gamma}}$ of an unknown parameter vector $\boldsymbol{\gamma}$, using GEE

$$\mathbf{0} = E[\{Y - \psi(\boldsymbol{X}^\top\boldsymbol{\gamma})\}\boldsymbol{X}\,|\,D=0] \tag{1}$$

derived under a working model (not necessarily correct) based on external data without $\boldsymbol{Z}$, where ψ is a known function. We only have the value of $\widehat{\boldsymbol{\gamma}}$ and knowledge about (1), not the individual-level external data. Ignoring the condition $D=0$ in (1) as (A1) is assumed in this section, we obtain $E(Y\boldsymbol{X}) = E\{\psi(\boldsymbol{X}^\top\boldsymbol{\gamma})\boldsymbol{X}\}$, which together with $E(Y\boldsymbol{X}) = E\{\phi(\boldsymbol{W}^\top\boldsymbol{\beta})\boldsymbol{X}\}$ from (B2) show that (B2)-(1) is equivalent to (B2) and

$$\mathbf{0} = E[\{\psi(\boldsymbol{X}^\top\boldsymbol{\gamma}) - \phi(\boldsymbol{W}^\top\boldsymbol{\beta})\}\boldsymbol{X}\,|\,D=1], \tag{C2}$$

where the condition $D=1$ can be ignored in this section but it is kept for extensions in Section 3. Note that (C2) relates external information to the estimation of $\boldsymbol{\beta}$ in GEE (B2).

Thus, we replace assumptions (B1)-(C1) by (B2)-(C2) and apply GEE utilizing external summary information $\widehat{\boldsymbol{\gamma}}$ and (1). Using internal data $(Y_i, \boldsymbol{W}_i, D_i=1)$, $i=1,...,n$, and (C2) as a constraint in GEE (B2), we propose the following GEE estimator $\widehat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$, which is a solution to $\bar{\boldsymbol{g}}(\widehat{\boldsymbol{\gamma}},\boldsymbol{\beta}) = \mathbf{0}$, where

$$\begin{aligned}
&\bar{\boldsymbol{g}}(\boldsymbol{\gamma},\boldsymbol{\beta}) = \frac{1}{n_1}\sum_{i=1}^{n} D_i\,\boldsymbol{g}(Y_i,\boldsymbol{W}_i,\boldsymbol{\gamma},\boldsymbol{\beta}),\\
&\boldsymbol{g}(y,\boldsymbol{w},\boldsymbol{\gamma},\boldsymbol{\beta}) = \begin{pmatrix} \boldsymbol{g}_1(y,\boldsymbol{w},\boldsymbol{\beta}) \\ \boldsymbol{g}_2(\boldsymbol{w},\boldsymbol{\gamma},\boldsymbol{\beta}) \end{pmatrix} = \begin{pmatrix} \{y-\phi(\boldsymbol{w}^\top\boldsymbol{\beta})\}\boldsymbol{w} \\ \{\psi(\boldsymbol{x}^\top\boldsymbol{\gamma}) - \phi(\boldsymbol{w}^\top\boldsymbol{\beta})\}\boldsymbol{x} \end{pmatrix},
\end{aligned} \tag{2}$$

$\boldsymbol{w} = \binom{\boldsymbol{x}}{\boldsymbol{z}}$, and n_1 is the internal sample size. The number of equations in $\bar{\boldsymbol{g}}(\widehat{\boldsymbol{\gamma}},\boldsymbol{\beta}) = \mathbf{0}$ is the dimension of $\boldsymbol{W}$ plus the dimension of $\boldsymbol{X}$, more than the dimension of $\boldsymbol{\beta}$ that is the same as the dimension of $\boldsymbol{W}$. Thus, no single $\widehat{\boldsymbol{\beta}}$ satisfies $\bar{\boldsymbol{g}}(\widehat{\boldsymbol{\gamma}},\widehat{\boldsymbol{\beta}}) = \mathbf{0}$. We therefore apply the two-step generalized method of moments (GMM) (Hansen, 1982) to obtain an estimator $\widehat{\boldsymbol{\beta}}$

as

$$\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \left\{ \bar{\boldsymbol{g}}(\widehat{\boldsymbol{\gamma}}, \boldsymbol{\beta})^\top \widehat{\boldsymbol{\Sigma}}^{-1} \bar{\boldsymbol{g}}(\widehat{\boldsymbol{\gamma}}, \boldsymbol{\beta}) \right\}, \tag{3}$$

where $\bar{\boldsymbol{g}}(\boldsymbol{\gamma}, \boldsymbol{\beta})$ is defined in (2),

$$\widehat{\boldsymbol{\Sigma}} = \frac{1}{n_1} \sum_{i=1}^{n} D_i \, \boldsymbol{g}(Y_i, \boldsymbol{W}_i, \widehat{\boldsymbol{\gamma}}, \widehat{\boldsymbol{\beta}}^{(1)}) \boldsymbol{g}(Y_i, \boldsymbol{W}_i, \widehat{\boldsymbol{\gamma}}, \widehat{\boldsymbol{\beta}}^{(1)})^\top,$$

and $\widehat{\boldsymbol{\beta}}^{(1)} = \arg\min_{\boldsymbol{\beta}} \{\bar{\boldsymbol{g}}(\widehat{\boldsymbol{\gamma}}, \boldsymbol{\beta})^\top \bar{\boldsymbol{g}}(\widehat{\boldsymbol{\gamma}}, \boldsymbol{\beta})\}$.

The following result establishes the asymptotic normality of GMM estimator $\widehat{\boldsymbol{\beta}}$ in (3) and provides an explicit form of its asymptotic covariance matrix. The proof is given in the Appendix.

Theorem 1. *Assume* (A1) *and* (B2). *Suppose that the true value* $(\boldsymbol{\gamma}_*, \boldsymbol{\beta}_*)$ *of the parameter* $(\boldsymbol{\gamma}, \boldsymbol{\beta})$ *defined in* (B2) *and* (C2) *is an interior point of the parameter space, the function* $\boldsymbol{g}(\boldsymbol{x}, \boldsymbol{w}, \boldsymbol{\gamma}, \boldsymbol{\beta})$ *defined in (2) is continuously differentiable in* $(\boldsymbol{\gamma}, \boldsymbol{\beta})$ *in a neighborhood* $\mathcal{N}$ *of* $(\boldsymbol{\gamma}_*, \boldsymbol{\beta}_*)$, $\boldsymbol{\Sigma} = E\{\boldsymbol{g}(Y, \boldsymbol{W}, \boldsymbol{\gamma}_*, \boldsymbol{\beta}_*)\boldsymbol{g}(Y, \boldsymbol{W}, \boldsymbol{\gamma}_*, \boldsymbol{\beta}_*)^\top\}$ *exists and is positive definite,* $\boldsymbol{M} = E\{\nabla_{\boldsymbol{\beta}} \boldsymbol{g}(Y, \boldsymbol{W}, \boldsymbol{\gamma}_*, \boldsymbol{\beta}_*)\}$ *exists and is of full rank, and* $E\{\sup_{(\boldsymbol{\gamma}, \boldsymbol{\beta}) \in \mathcal{N}} \|\nabla_{\boldsymbol{\gamma}, \boldsymbol{\beta}} \boldsymbol{g}(Y, \boldsymbol{W}, \boldsymbol{\gamma}, \boldsymbol{\beta})\|\} < \infty$, *where* $\nabla_{\boldsymbol{\xi}}$ *denotes the vector of partial derivatives with respect to* $\boldsymbol{\xi}$ *and* $\|\boldsymbol{C}\|^2 = \mathrm{trace}(\boldsymbol{C}\boldsymbol{C}^\top)$. *Assume that the internal sample size* $n_1 \to \infty$ *and the internal and external size ratio* $n_1/n_0 \to r \in [0, \infty)$ *almost surely as* $n \to \infty$. *Assume further that the external summary statistic* $\widehat{\boldsymbol{\gamma}}$ *is a GEE estimator using (1),* $\boldsymbol{\Lambda} = E[\{Y - \psi(\boldsymbol{X}^\top \boldsymbol{\gamma}_*)\}^2 \boldsymbol{X}\boldsymbol{X}^\top]$ *exists and is positive definite, the function* $\psi(\,\cdot\,)$ *is continuously differentiable with derivative* ψ', $E\{\psi'(\boldsymbol{X}^\top \boldsymbol{\gamma}_*)\boldsymbol{X}\boldsymbol{X}^\top\}$ *exists and is of full rank, and* $E\{\sup_{\boldsymbol{\gamma} \in \mathcal{N}_{\boldsymbol{\gamma}_*}} \|\psi'(\boldsymbol{X}^\top \boldsymbol{\gamma})\boldsymbol{X}\boldsymbol{X}^\top\|\} < \infty$, *where* $\mathcal{N}_{\boldsymbol{\gamma}_*}$ *is a neighborhood of* $\boldsymbol{\gamma}_*$. *Then, for the estimator* $\widehat{\boldsymbol{\beta}}$ *defined by (3),*

$$\sqrt{n_1}\,(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_*) \xrightarrow{d} N\big(\boldsymbol{0}, \boldsymbol{V}\big), \tag{4}$$

where $\xrightarrow{d}$ *denotes convergence in distribution as* $n \to \infty$,

$$\boldsymbol{V} = (\boldsymbol{M}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{M})^{-1} + r\,(\boldsymbol{M}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{M})^{-1} \boldsymbol{M}^\top \boldsymbol{\Sigma}^{-1} \left(\begin{smallmatrix} \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{\Lambda} \end{smallmatrix}\right) \boldsymbol{\Sigma}^{-1} \boldsymbol{M} (\boldsymbol{M}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{M})^{-1}, \tag{5}$$

and $\boldsymbol{0}$ *denotes a vector or matrix of all zeros with an appropriate order.*

When the size of external data is much larger than the size of internal data, i.e., the ratio $n_1/n_0 \to r = 0$, the asymptotic covariance matrix of $\sqrt{n_1}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_*)$ is $\boldsymbol{V} = (\boldsymbol{M}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{M})^{-1}$ and the asymptotic distribution of $\widehat{\boldsymbol{\beta}}$ is not affected by the estimation of $\boldsymbol{\gamma}$. If $r > 0$, then the second term on the right side of (5) is the price for estimating $\boldsymbol{\gamma}$ by $\widehat{\boldsymbol{\gamma}}$.

The GEE estimator without using external information, denoted as $\widehat{\boldsymbol{\beta}}_I$, has the following asymptotic distribution,

$$\sqrt{n_1}\,(\widehat{\boldsymbol{\beta}}_I - \boldsymbol{\beta}_*) \xrightarrow{d} N\big(\mathbf{0}, (\boldsymbol{M}_1^\top \boldsymbol{\Sigma}_1^{-1} \boldsymbol{M}_1)^{-1}\big), \tag{6}$$

where $\boldsymbol{\Sigma}_1 = E\{\boldsymbol{g}_1(Y, \boldsymbol{W}, \boldsymbol{\beta}_*)\boldsymbol{g}_1(Y, \boldsymbol{W}, \boldsymbol{\beta}_*)^\top\}$, $\boldsymbol{M}_1 = E\{\nabla_{\boldsymbol{\beta}}\, \boldsymbol{g}_1(Y, \boldsymbol{W}, \boldsymbol{\beta}_*)\}$, and $\boldsymbol{g}_1$ is defined in (2). It can be seen from (4) and (6) that $\widehat{\boldsymbol{\beta}}$ does not improve $\widehat{\boldsymbol{\beta}}_I$ in terms of convergence rate, i.e., they both have convergence rate $n_1^{-1/2}$. This is due to the fact that (i) we do not have external individual-level data but have only summary statistics not necessary in the form for estimating $\boldsymbol{\beta}$ and/or (ii) we do not have external information from $\boldsymbol{Z}$ which may be useful in estimating $\boldsymbol{\beta}$.

In the following we show that

$$\boldsymbol{M}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{M} - \boldsymbol{M}_1^\top \boldsymbol{\Sigma}_1^{-1} \boldsymbol{M}_1 \ \text{ is semi-positive definite,} \tag{7}$$

i.e., the GMM estimator $\widehat{\boldsymbol{\beta}}$ in (3) is asymptotically more efficient than $\widehat{\boldsymbol{\beta}}_I$ when $r = 0$, although the convergence rates of $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\beta}}_I$ are the same. From the definition of $\boldsymbol{\Sigma}$ and $\boldsymbol{g}_1$ and $\boldsymbol{g}_2$ in (2),

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_1 & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{12}^\top & \boldsymbol{\Sigma}_2 \end{pmatrix},$$

where $\boldsymbol{\Sigma}_{12} = E\{\boldsymbol{g}_1(Y, \boldsymbol{W}, \boldsymbol{\beta}_*)\boldsymbol{g}_2(\boldsymbol{W}, \boldsymbol{\gamma}_*, \boldsymbol{\beta}_*)^\top\}$ and $\boldsymbol{\Sigma}_2 = E\{\boldsymbol{g}_2(\boldsymbol{W}, \boldsymbol{\gamma}_*, \boldsymbol{\beta}_*)\boldsymbol{g}_2(\boldsymbol{W}, \boldsymbol{\gamma}_*, \boldsymbol{\beta}_*)^\top\}$. From the partition of $\boldsymbol{\Sigma}$, we obtain that

$$\boldsymbol{\Sigma}^{-1} = \begin{pmatrix} \boldsymbol{\Sigma}_1^{-1} + \boldsymbol{\Sigma}_1^{-1}\boldsymbol{\Sigma}_{12}(\boldsymbol{\Sigma}_2 - \boldsymbol{\Sigma}_{12}^\top \boldsymbol{\Sigma}_1^{-1}\boldsymbol{\Sigma}_{12})^{-1}\boldsymbol{\Sigma}_{12}^\top \boldsymbol{\Sigma}_1^{-1} & -\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\Sigma}_{12}(\boldsymbol{\Sigma}_2 - \boldsymbol{\Sigma}_{12}^\top \boldsymbol{\Sigma}_1^{-1}\boldsymbol{\Sigma}_{12})^{-1} \\ -(\boldsymbol{\Sigma}_2 - \boldsymbol{\Sigma}_{12}^\top \boldsymbol{\Sigma}_1^{-1}\boldsymbol{\Sigma}_{12})^{-1}\boldsymbol{\Sigma}_{12}^\top \boldsymbol{\Sigma}_1^{-1} & (\boldsymbol{\Sigma}_2 - \boldsymbol{\Sigma}_{12}^\top \boldsymbol{\Sigma}_1^{-1}\boldsymbol{\Sigma}_{12})^{-1} \end{pmatrix}.$$

Note that $\boldsymbol{M}^\top = (\boldsymbol{M}_1^\top, \boldsymbol{M}_2^\top)$, where $\boldsymbol{M}_2 = E\{\nabla_{\boldsymbol{\beta}} \boldsymbol{g}_2(\boldsymbol{W}, \boldsymbol{\gamma}_*, \boldsymbol{\beta}_*)\}$. Then

$$\boldsymbol{M}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{M} = \boldsymbol{M}_1^\top \boldsymbol{\Sigma}_1^{-1} \boldsymbol{M}_1 + \boldsymbol{A}^\top \boldsymbol{A} + \boldsymbol{B}^\top \boldsymbol{B} - \boldsymbol{A}^\top \boldsymbol{B} - \boldsymbol{B}^\top \boldsymbol{A},$$

where $\boldsymbol{A} = (\boldsymbol{\Sigma}_2 - \boldsymbol{\Sigma}_{12}^\top \boldsymbol{\Sigma}_1^{-1} \boldsymbol{\Sigma}_{12})^{-1/2} \boldsymbol{\Sigma}_{12}^\top \boldsymbol{\Sigma}_1^{-1} \boldsymbol{M}_1$ and $\boldsymbol{B} = (\boldsymbol{\Sigma}_2 - \boldsymbol{\Sigma}_{12}^\top \boldsymbol{\Sigma}_1^{-1} \boldsymbol{\Sigma}_{12})^{-1/2} \boldsymbol{M}_2$, and result (7) follows because $\boldsymbol{A}^\top \boldsymbol{A} + \boldsymbol{B}^\top \boldsymbol{B} - \boldsymbol{A}^\top \boldsymbol{B} - \boldsymbol{B}^\top \boldsymbol{A}$ is semi-positive definite. In fact, for any vector $\boldsymbol{c}$ with the same dimension as $\boldsymbol{\beta}$,

$$\boldsymbol{c}^\top (\boldsymbol{A}^\top \boldsymbol{A} + \boldsymbol{B}^\top \boldsymbol{B} - \boldsymbol{A}^\top \boldsymbol{B} - \boldsymbol{B}^\top \boldsymbol{A}) \boldsymbol{c} = (\boldsymbol{A}\boldsymbol{c} - \boldsymbol{B}\boldsymbol{c})^2 \geq 0,$$

with the equality holding if and only if $\boldsymbol{A}\boldsymbol{c} = \boldsymbol{B}\boldsymbol{c}$, i.e., $(\boldsymbol{\Sigma}_{12}^\top \boldsymbol{\Sigma}_1^{-1} \boldsymbol{M}_1 - \boldsymbol{M}_2)\boldsymbol{c} = 0$.

For assessing accuracy or statistical inference, we need a consistent (as $n \to \infty$) estimator of the covariance matrix $\boldsymbol{V}$ in (5). Using substitution based on (5) and assuming the conditions in Theorem 1, we obtain the following consistent estimator,

$$\widehat{\boldsymbol{V}} = (\widehat{\boldsymbol{M}}^\top \widehat{\boldsymbol{\Sigma}}^{-1} \widehat{\boldsymbol{M}})^{-1} + \tfrac{n_1}{n_0} (\widehat{\boldsymbol{M}}^\top \widehat{\boldsymbol{\Sigma}}^{-1} \widehat{\boldsymbol{M}})^{-1} \widehat{\boldsymbol{M}}^\top \widehat{\boldsymbol{\Sigma}}^{-1} \left(\begin{smallmatrix} \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \widehat{\boldsymbol{\Lambda}} \end{smallmatrix}\right) \widehat{\boldsymbol{\Sigma}}^{-1} \widehat{\boldsymbol{M}} (\widehat{\boldsymbol{M}}^\top \widehat{\boldsymbol{\Sigma}}^{-1} \widehat{\boldsymbol{M}})^{-1},$$

where $\widehat{\boldsymbol{\Sigma}}$ is given by (3) with $\widehat{\boldsymbol{\beta}}^{(1)}$ replaced by $\widehat{\boldsymbol{\beta}}$,

$$\widehat{\boldsymbol{M}} = \frac{1}{n_1} \sum_{i=1}^{n} D_i \nabla_{\boldsymbol{\beta}} \boldsymbol{g}(Y_i, \boldsymbol{W}_i, \widehat{\boldsymbol{\gamma}}, \widehat{\boldsymbol{\beta}}) \qquad \text{and} \qquad \widehat{\boldsymbol{\Lambda}} = \frac{1}{n_1} \sum_{i=1}^{n} D_i \{Y_i - \psi(\boldsymbol{X}_i^\top \widehat{\boldsymbol{\gamma}})\}^2 \boldsymbol{X}_i \boldsymbol{X}_i^\top.$$

The matrix $\boldsymbol{\Lambda}$ can also be estimated by an accuracy measure for $\widehat{\boldsymbol{\gamma}}$ if it is provided as external summary information.

We end this section with a discussion about the extension to situation where the dimension of $\boldsymbol{W}$ depends on n and is high, which often occurs in modern statistical studies. When the dimension of $\boldsymbol{W}$ diverges as $n \to \infty$, we can apply penalized GMM (Caner, 2009; Liao, 2013) instead of the non-penalized GMM defined by (3), i.e., we obtain estimators

$$\widehat{\boldsymbol{\beta}}_{\lambda_n} = \arg\min_{\boldsymbol{\beta}} \left\{ \bar{\boldsymbol{g}}(\widehat{\boldsymbol{\gamma}}, \boldsymbol{\beta})^\top \widehat{\boldsymbol{\Sigma}}^{-1} \bar{\boldsymbol{g}}(\widehat{\boldsymbol{\gamma}}, \boldsymbol{\beta}) + \lambda_n(\boldsymbol{\beta}) \right\},$$
$$\widehat{\boldsymbol{\beta}}^{(1)}_{\lambda_n} = \arg\min_{\boldsymbol{\beta}} \left\{ \bar{\boldsymbol{g}}(\widehat{\boldsymbol{\gamma}}, \boldsymbol{\beta})^\top \bar{\boldsymbol{g}}(\widehat{\boldsymbol{\gamma}}, \boldsymbol{\beta}) + \lambda_n(\boldsymbol{\beta}) \right\},$$

where $\lambda_n(\boldsymbol{\beta})$ is a suitably chosen penalty such as LASSO, SCAD, or MCP (Caner, 2009; Liao, 2013). Asymptotic results for the penalized GMM estimator $\widehat{\boldsymbol{\beta}}_{\lambda_n}$ can be established under some conditions along the lines of Caner (2009), Liao (2013), He et al. (2016), Li et al. (2021), and Tian and Feng (2022), which will be our future research.

3 GMM under Heterogeneous Data Populations

We focus on $\boldsymbol{W}$ with a dimension not varying with n. The extension to high dimensional $\boldsymbol{W}$ is as discussed in the end of Section 2.

3.1 Robustness of GMM (3)

When (A1) may not hold, i.e., the internal and external distributions are possibly different, we still consider GEE (B2) for internal data and GEE (1) for external data.

Our first question is whether the GMM estimator $\widehat{\boldsymbol{\beta}}$ given by (3) is robust against assumption (A1). Consider the following assumption weaker than (A1),

$$E(Y \,|\, \boldsymbol{X}, D=1) = E(Y \,|\, \boldsymbol{X}, D=0) \quad \text{and} \quad f(\boldsymbol{x} \,|\, D=1) = f(\boldsymbol{x} \,|\, D=0). \tag{A2}$$

Note that the first condition in (A2) is on conditional means and is implied by the condition $f(y \,|\, \boldsymbol{x}, D=1) = f(y \,|\, \boldsymbol{x}, D=0)$. Also, (A2) is implied by $f(y, \boldsymbol{x} \,|\, D=1) = f(y, \boldsymbol{x} \,|\, D=0)$, which is still weaker than (A1). By the second condition in (A2),

$$E\{\psi(\boldsymbol{X}^\top \boldsymbol{\gamma})\boldsymbol{X} \,|\, D=0\} = E\{\psi(\boldsymbol{X}^\top \boldsymbol{\gamma})\boldsymbol{X} \,|\, D=1\}.$$

Also,

$$\begin{aligned} E(Y\boldsymbol{X} \,|\, D=0) &= E\{E(Y \,|\, \boldsymbol{X}, D=0)\boldsymbol{X} \,|\, D=0\} \\ &= E\{E(Y \,|\, \boldsymbol{X}, D=1)\boldsymbol{X} \,|\, D=0\} \\ &= E\{E(Y \,|\, \boldsymbol{X}, D=1)\boldsymbol{X} \,|\, D=1\} \\ &= E(Y\boldsymbol{X} \,|\, D=1), \end{aligned}$$

where the second equality is from the first condition in (A2) and the third equality follows from the second condition in (A2) as $E(Y \,|\, \boldsymbol{X}, D = 1)\boldsymbol{X}$ is a function of $\boldsymbol{X}$. This together with (B2) imply that (C2) in Section 2 holds even when internal and external populations are different. Consequently, under (A2), we can use constraint (C2) in GEE (B2) and apply the GMM in (3) as it involves internal data only, and the result in Section 2 still holds.

A similar argument shows that (C2) is satisfied and the result in Section 2 holds under the following alternative assumption,

$$E(Y \,|\, \boldsymbol{W}, D = 1) = E(Y \,|\, \boldsymbol{W}, D = 0) \quad \text{and} \quad f(\boldsymbol{w} \,|\, D = 1) = f(\boldsymbol{w} \,|\, D = 0). \quad \text{(A2}')$$

Note that there is no definite relationship between the first conditions in (A2) and (A2′), although both (A2) and (A2′) are weaker than (A1).

This shows that the GMM estimator $\widehat{\boldsymbol{\beta}}$ in (3) is robust against (A1) to some extent.

3.2 Results under weaker assumptions on populations

We next consider extensions when neither (A2) nor (A2′) holds. Can a valid estimator using external summary information be derived under only the first condition in (A2)? The following analysis indicates that we need some additional conditions.

As in the previous argument, we still just need to consider (1) to derive an appropriate constraint for GEE (B2). From (1),

$$\begin{aligned}
\mathbf{0} &= E\left[E\{Y\boldsymbol{X} - \psi(\boldsymbol{X}^\top\boldsymbol{\gamma})\boldsymbol{X} \,|\, \boldsymbol{X}, D = 0\} \,|\, D = 0\right] \\
&= E\left[\{E(Y \,|\, \boldsymbol{X}, D = 0) - \psi(\boldsymbol{X}^\top\boldsymbol{\gamma})\}\boldsymbol{X} \,|\, D = 0\right] \\
&= E\left[\{E(Y \,|\, \boldsymbol{X}, D = 1) - \psi(\boldsymbol{X}^\top\boldsymbol{\gamma})\}\boldsymbol{X} \,|\, D = 0\right] \\
&= \int \{E(Y \,|\, \boldsymbol{X} = \boldsymbol{x}, D = 1) - \psi(\boldsymbol{x}^\top\boldsymbol{\gamma})\}\, \boldsymbol{x} f(\boldsymbol{x} \,|\, D = 0) d\boldsymbol{x} \\
&= \int \kappa(\boldsymbol{x})\{E(Y \,|\, \boldsymbol{X} = \boldsymbol{x}, D = 1) - \psi(\boldsymbol{x}^\top\boldsymbol{\gamma})\}\, \boldsymbol{x} f(\boldsymbol{x} \,|\, D = 1) d\boldsymbol{x} \\
&= E\{\kappa(\boldsymbol{X})YX \,|\, D = 1\} - E\{\kappa(\boldsymbol{X})\psi(\boldsymbol{X}^\top\boldsymbol{\gamma})\boldsymbol{X} \,|\, D = 1\},
\end{aligned}$$

where the third equality follows from the first condition in (A2), the fifth equality is from

$$\kappa(\boldsymbol{x}) = \frac{f(\boldsymbol{x} \,|\, D=0)}{f(\boldsymbol{x} \,|\, D=1)}, \tag{8}$$

and $\kappa(\,\cdot\,)$ is typically an unknown function. In the scenario where D_i's are random as described in the beginning of Section 2, we have

$$\kappa(\boldsymbol{x}) = \frac{P(D=0 \,|\, \boldsymbol{x})}{P(D=1 \,|\, \boldsymbol{x})} \frac{P(D=1)}{P(D=0)}. \tag{9}$$

However, $E\{\kappa(\boldsymbol{X})Y\boldsymbol{X} \,|\, D=1\}$ is not directly related with $\boldsymbol{\beta}$ in (B2) because of the extra $\kappa(\boldsymbol{X})$. If we strengthen (B2) to

$$E[\kappa(\boldsymbol{X})\{Y - \phi(\boldsymbol{W}^\top\boldsymbol{\beta})\}\boldsymbol{X} \,|\, D=1] = \boldsymbol{0}, \tag{B2+}$$

which holds if the semiparametric mean model $E(Y \mid \boldsymbol{W}, D=1) = \phi(\boldsymbol{W}^\top\boldsymbol{\beta})$ is correct, then $E\{\kappa(\boldsymbol{X})Y\boldsymbol{X} \,|\, D=1\} = E\{\kappa(\boldsymbol{X})\phi(\boldsymbol{W}^\top\boldsymbol{\beta})\boldsymbol{X} \,|\, D=1\}$ and we obtain that

$$\boldsymbol{0} = E[\kappa(\boldsymbol{X})\{\psi(\boldsymbol{X}^\top\boldsymbol{\gamma}) - \phi(\boldsymbol{W}^\top\boldsymbol{\beta})\}\boldsymbol{X} \,|\, D=1], \tag{C3}$$

which can be used as a constraint replacing (C2). We can view $\kappa(\boldsymbol{x})$ as an adjustment for the difference in internal and external populations, in order to use the external information. The idea here is similar to the use of propensity score in dealing with missing data, although $\kappa(\boldsymbol{x})$ is an odds ratio of propensities in view of (9).

It remains to estimate $\kappa(\boldsymbol{x})$ in (8), as it is unknown. If we have external individual-level data, then the estimation of $\kappa(\boldsymbol{x})$ in (8) is simple. As we only have external summary statistics, we need another condition to estimate $\kappa(\boldsymbol{x})$. Assume that

> $\kappa(\boldsymbol{X}) = q(\boldsymbol{X}^\top\boldsymbol{\eta})$, $q(\cdot)$ is known, $\boldsymbol{\eta}$ is an unknown parameter vector, and there is a vector function $\boldsymbol{S}$ of $\boldsymbol{X}$ with dimension $\geq$ the dimension of $\boldsymbol{\eta}$ such that the sample mean of $\boldsymbol{S}$ is provided as an external summary statistic or the population mean vector $E(\boldsymbol{S})$ is known. (10)

As an example, components of $\boldsymbol{S}$ can be the vector of indicators of gender and age group with the number of groups $\geq$ the dimension of $\boldsymbol{\eta}$. If $\boldsymbol{X} = X$ is continuous and univariate,

then $\boldsymbol{S}$ can be the vector whose jth component is X^j, $j = 0, 1, ..., l-1$ with $l \geq$ the dimension of $\boldsymbol{\eta}$.

We propose to estimate $\boldsymbol{\eta}$ by using GEE or GMM with estimation equation

$$\frac{1}{n_1}\sum_{i=1}^{n} D_i\, q(\boldsymbol{X}_i^\top \boldsymbol{\eta})\, \boldsymbol{S}_i = \bar{\boldsymbol{S}}_0, \tag{11}$$

where $\bar{\boldsymbol{S}}_0$ is the available sample mean of $\boldsymbol{S}_i$'s in external dataset and n_1 is the internal sample size. This is supported by the fact that, under the law of large numbers, the left side of (11) converges in probability to $E\{\kappa(\boldsymbol{X})\boldsymbol{S} \,|\, D=1\} = E(\boldsymbol{S} \,|\, D=0)$ and the right side of (11) converges in probability to $E(\boldsymbol{S} \,|\, D=0)$. The situation where the population mean of $\boldsymbol{S}$ is known can be treated similarly.

If the second condition in (A2) holds, then $\kappa(\boldsymbol{x}) \equiv 1$ and (10) automatically holds with $\boldsymbol{S}$ to be a constant. Thus, we consider the following assumption weaker than (A2),

$$E(Y \,|\, \boldsymbol{X}, D=1) = E(Y \,|\, \boldsymbol{X}, D=0) \quad \text{and} \quad (10) \text{ holds.} \tag{A3}$$

Under conditions (A3) and (B2+), we use (C3) as a constraint to obtain the GMM estimator

$$\begin{pmatrix} \widehat{\boldsymbol{\beta}} \\ \widehat{\boldsymbol{\eta}} \end{pmatrix} = \arg\min_{\boldsymbol{\beta},\boldsymbol{\eta}} \left\{ \bar{\boldsymbol{g}}(\widehat{\boldsymbol{\gamma}}, \bar{\boldsymbol{S}}_0, \boldsymbol{\beta}, \boldsymbol{\eta})^\top\, \widehat{\boldsymbol{\Sigma}}_+^{-1}\, \bar{\boldsymbol{g}}(\widehat{\boldsymbol{\gamma}}, \bar{\boldsymbol{S}}_0, \boldsymbol{\beta}, \boldsymbol{\eta}) \right\}, \tag{12}$$

where

$$\begin{aligned} \bar{\boldsymbol{g}}(\boldsymbol{\gamma}, \boldsymbol{\varsigma}, \boldsymbol{\beta}, \boldsymbol{\eta}) &= \frac{1}{n_1}\sum_{i=1}^{n} D_i\, \boldsymbol{g}(Y_i, \boldsymbol{W}_i, \boldsymbol{\gamma}, \boldsymbol{\varsigma}, \boldsymbol{\beta}, \boldsymbol{\eta}), \\ \boldsymbol{g}(y, \boldsymbol{w}, \boldsymbol{\gamma}, \boldsymbol{\varsigma}, \boldsymbol{\beta}, \boldsymbol{\eta}) &= \begin{pmatrix} \{y - \phi(\boldsymbol{w}^\top \boldsymbol{\beta})\}\, \boldsymbol{w} \\ q(\boldsymbol{x}^\top \boldsymbol{\eta})\, \boldsymbol{s} - \boldsymbol{\varsigma} \\ q(\boldsymbol{x}^\top \boldsymbol{\eta})\, \{\psi(\boldsymbol{x}^\top \boldsymbol{\gamma}) - \phi(\boldsymbol{w}^\top \boldsymbol{\beta})\}\, \boldsymbol{x} \end{pmatrix}, \end{aligned} \tag{13}$$

$$\widehat{\boldsymbol{\Sigma}}_+ = \frac{1}{n_1}\sum_{i=1}^{n} D_i\, \boldsymbol{g}(Y_i, \boldsymbol{W}_i, \widehat{\boldsymbol{\gamma}}, \bar{\boldsymbol{S}}_0, \widehat{\boldsymbol{\beta}}^{(1)}, \widehat{\boldsymbol{\eta}}^{(1)}) \boldsymbol{g}(Y_i, \boldsymbol{W}_i, \widehat{\boldsymbol{\gamma}}, \bar{\boldsymbol{S}}_0, \widehat{\boldsymbol{\beta}}^{(1)}, \widehat{\boldsymbol{\eta}}^{(1)})^\top,$$

$\boldsymbol{\varsigma}$ denotes $E(\boldsymbol{S} \,|\, D=0)$, and $(\widehat{\boldsymbol{\beta}}^{(1)}, \widehat{\boldsymbol{\eta}}^{(1)})^\top = \arg\min_{\boldsymbol{\beta},\boldsymbol{\eta}} \{\bar{\boldsymbol{g}}(\widehat{\boldsymbol{\gamma}}, \boldsymbol{\beta}, \boldsymbol{\eta})^\top \bar{\boldsymbol{g}}(\widehat{\boldsymbol{\gamma}}, \boldsymbol{\beta}, \boldsymbol{\eta})\}$.

The following theorem establishes the asymptotic distribution of $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\eta}}$ in (12). The proof is in the Appendix.

Theorem 2. *Assume* (A3) *and* (B2+). *Suppose that the true value* $(\boldsymbol{\gamma}_*, \boldsymbol{\varsigma}_*, \boldsymbol{\beta}_*, \boldsymbol{\eta}_*)$ *of the parameter* $(\boldsymbol{\gamma}, \boldsymbol{\varsigma}, \boldsymbol{\beta}, \boldsymbol{\eta})$ *is an interior point of the parameter space,* $\boldsymbol{g}(\boldsymbol{x}, \boldsymbol{w}, \boldsymbol{\gamma}, \boldsymbol{\varsigma}, \boldsymbol{\beta}, \boldsymbol{\eta})$ *defined in (13) is continuously differentiable in* $(\boldsymbol{\gamma}, \boldsymbol{\varsigma}, \boldsymbol{\beta}, \boldsymbol{\eta})$ *in a neighborhood* $\mathcal{N}$ *of the true value* $(\boldsymbol{\gamma}_*, \boldsymbol{\varsigma}_*, \boldsymbol{\beta}_*, \boldsymbol{\eta}_*)$, $\boldsymbol{\Sigma}_+ = E\{\boldsymbol{g}(Y, \boldsymbol{W}, \boldsymbol{\gamma}_*, \boldsymbol{\varsigma}_*, \boldsymbol{\beta}_*, \boldsymbol{\eta}_*)\boldsymbol{g}(Y, \boldsymbol{W}, \boldsymbol{\gamma}_*, \boldsymbol{\varsigma}_*, \boldsymbol{\beta}_*, \boldsymbol{\eta}_*)^\top \,|\, D = 1\}$ *exists and is positive definite,* $\boldsymbol{M}_+ = E\{\nabla_{\boldsymbol{\beta},\boldsymbol{\eta}}\, \boldsymbol{g}(Y, \boldsymbol{W}, \boldsymbol{\gamma}_*, \boldsymbol{\varsigma}_*, \boldsymbol{\beta}_*, \boldsymbol{\eta}_*) \,|\, D = 1\}$ *exists and is of full rank, and* $E\{\sup_{(\boldsymbol{\gamma},\boldsymbol{\varsigma},\boldsymbol{\beta},\boldsymbol{\eta})\in\mathcal{N}} \|\nabla_{\boldsymbol{\gamma},\boldsymbol{\varsigma},\boldsymbol{\beta},\boldsymbol{\eta}}\, \boldsymbol{g}(Y, \boldsymbol{W}, \boldsymbol{\gamma}, \boldsymbol{\varsigma}, \boldsymbol{\beta}, \boldsymbol{\eta})\| \mid D = 1\} < \infty$. *Assume that* $n_1 \to \infty$ *and the ratio* $n_1/n_0 \to r \in [0, \infty)$ *almost surely as* $n \to \infty$. *Assume further that the external summary statistic* $\widehat{\boldsymbol{\gamma}}$ *is a GEE estimator using (1),* $\boldsymbol{\Lambda}_0 = E[\{Y - \psi(\boldsymbol{X}^\top\boldsymbol{\gamma}_*)\}^2\boldsymbol{X}\boldsymbol{X}^\top \mid D = 0]$ *exists and is positive definite, the function* $\psi(\,\cdot\,)$ *is continuously differentiable with derivative* ψ', $\boldsymbol{H}_0 = E\{\psi'(\boldsymbol{X}^\top\boldsymbol{\gamma}_*)\boldsymbol{X}\boldsymbol{X}^\top \mid D = 0\}$ *exists and is of full rank, and* $E\{\sup_{\boldsymbol{\gamma}\in\mathcal{N}_{\boldsymbol{\gamma}_*}} \|\psi'(\boldsymbol{X}^\top\boldsymbol{\gamma})\boldsymbol{X}\boldsymbol{X}^\top\| \,|\, D = 0\} < \infty$, *where* $\mathcal{N}_{\boldsymbol{\gamma}_*}$ *is a neighborhood of* $\boldsymbol{\gamma}_*$. *Then,*

$$\sqrt{n_1}\begin{pmatrix} \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_* \\ \widehat{\boldsymbol{\eta}} - \boldsymbol{\eta}_* \end{pmatrix} \xrightarrow{d} N\big(\boldsymbol{0}, \boldsymbol{V}_+\big),$$

where

$$\boldsymbol{V}_+ = (\boldsymbol{M}_+^\top\boldsymbol{\Sigma}_+^{-1}\boldsymbol{M}_+)^{-1} + r\,(\boldsymbol{M}_+^\top\boldsymbol{\Sigma}_+^{-1}\boldsymbol{M}_+)^{-1}\boldsymbol{M}_+^\top\boldsymbol{\Sigma}_+^{-1}\begin{pmatrix} \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{C}_0 & \boldsymbol{C}_{01}^\top \\ \boldsymbol{0} & \boldsymbol{C}_{01} & \boldsymbol{\Lambda}_0 \end{pmatrix}\boldsymbol{\Sigma}_+^{-1}\boldsymbol{M}_+(\boldsymbol{M}_+^\top\boldsymbol{\Sigma}_+^{-1}\boldsymbol{M}_+)^{-1},$$

$\boldsymbol{C}_0 = E\{(\boldsymbol{S} - \boldsymbol{\varsigma}_*)(\boldsymbol{S} - \boldsymbol{\varsigma}_*)^\top \,|\, D = 0\}$, *and* $\boldsymbol{C}_{01} = E[\{\psi(\boldsymbol{X}^\top\boldsymbol{\gamma}_*) - Y\}\boldsymbol{X}(\boldsymbol{S} - \boldsymbol{\varsigma}_*)^\top \,|\, D = 0]$.

As an alternative to (A3), we consider the following different relaxation of (A2′). If $E(Y \,|\, \boldsymbol{W}, D = 1) = E(Y \,|\, \boldsymbol{W}, D = 0)$, then

$$\begin{aligned} E(\boldsymbol{X}Y \,|\, D = 0) &= E\{\boldsymbol{X}E(Y \,|\, \boldsymbol{W}, D = 0) \,|\, D = 0\} \\ &= E\{\boldsymbol{X}E(Y \,|\, \boldsymbol{W}, D = 1) \,|\, D = 0\} \\ &= E\{\kappa(\boldsymbol{W})E(\boldsymbol{X}Y \,|\, \boldsymbol{W}, D = 1) \,|\, D = 1\} \\ &= E\{\kappa(\boldsymbol{W})\boldsymbol{X}Y \,|\, D = 1\}, \end{aligned}$$

where the third equality follows from

$$\kappa(\boldsymbol{w}) = \frac{f(\boldsymbol{w} \,|\, D = 0)}{f(\boldsymbol{w} \,|\, D = 1)}. \tag{14}$$

Under (B2+) with $\kappa(\boldsymbol{X})$ replaced by $\kappa(\boldsymbol{W})$, we obtain constraint

$$\mathbf{0} = E[\kappa(\boldsymbol{W})\boldsymbol{X}\{\psi(\boldsymbol{X}^\top\boldsymbol{\gamma}) - \phi(\boldsymbol{W}^\top\boldsymbol{\beta})\}\,|\,D=1],$$

which can be used to replace (C3) and obtain the GMM estimator (12), if $\kappa(\boldsymbol{w})$ in (14) can be estimated. Note that $\kappa(\boldsymbol{w})$ in (14) plays the same role as $\kappa(\boldsymbol{x})$ in (8). To estimate $\kappa(\boldsymbol{w})$, we apply (11) with $\boldsymbol{X}$ replaced by $\boldsymbol{W}$.

Thus, we consider the following assumption weaker than (A2′):

$$E(Y\,|\,\boldsymbol{W}, D=1) = E(Y\,|\,\boldsymbol{W}, D=0) \text{ and (10) holds with } \kappa(\boldsymbol{W}) = q(\boldsymbol{W}^\top\boldsymbol{\eta}). \quad \text{(A3′)}$$

Note that (A3′) does not have a definite relationship with (A3), because $E(Y\,|\,\boldsymbol{W}, D=1) = E(Y\,|\,\boldsymbol{W}, D=0)$ has no definite relationship with $E(Y\,|\,\boldsymbol{X}, D=1) = E(Y\,|\,\boldsymbol{X}, D=0)$.

Under (A3′) and (B2+), we can apply the GMM estimator in (12). A result similar to Theorem 2 can be established for the asymptotic normality of this GMM estimator.

To end this subsection we provide the following diagram of the relationships among different assumptions on populations, where $\Rightarrow$ indicates "stronger than":

$$\text{(A1)} \Rightarrow \begin{cases} \text{(A2)} \Rightarrow \text{(A3)}, \\ \\ \text{(A2′)} \Rightarrow \text{(A3′)}. \end{cases}$$

3.3 Estimation of covariance matrix $\boldsymbol{V}_+$

The asymptotic covariance matrix of $\widehat{\boldsymbol{\beta}}$ is the first diagonal sub-matrix of $\boldsymbol{V}_+$ in Theorem 2 with dimension the same as that of $\boldsymbol{\beta}$. A consistent estimator of $\boldsymbol{V}_+$ can be obtained by substituting $\boldsymbol{\Sigma}_+$, $\boldsymbol{M}_+$, $\boldsymbol{\Lambda}_0$, $\boldsymbol{C}_0$, and $\boldsymbol{C}_{01}$ in $\boldsymbol{V}_+$ by consistent estimators. Matrices $\boldsymbol{\Sigma}_+$ and $\boldsymbol{M}_+$ can be consistently estimated by $\widehat{\boldsymbol{\Sigma}}_+$ and $\widehat{\boldsymbol{M}}_+ = n_1^{-1}\sum_{i=1}^n D_i\nabla_{\boldsymbol{\beta},\boldsymbol{\eta}}\,\boldsymbol{g}(Y_i, \boldsymbol{W}_i, \widehat{\boldsymbol{\gamma}}, \bar{\boldsymbol{S}}_0, \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\eta}})$,

respectively. Matrices $\boldsymbol{\Lambda}_0$, $\boldsymbol{C}_0$, and $\boldsymbol{C}_{01}$ can be estimated by consistent estimators

$$\widehat{\boldsymbol{\Lambda}}_0 = \frac{1}{n_1}\sum_{i=1}^{n} D_i\, q(\boldsymbol{W}_i^\top \widehat{\boldsymbol{\eta}})\{Y_i - \psi(\boldsymbol{X}_i^\top \widehat{\boldsymbol{\gamma}})\}^2 \boldsymbol{X}_i \boldsymbol{X}_i^\top,$$

$$\widehat{\boldsymbol{C}}_0 = \frac{1}{n_1}\sum_{i=1}^{n} D_i\, q(\boldsymbol{W}_i^\top \widehat{\boldsymbol{\eta}})(\boldsymbol{S}_i - \bar{\boldsymbol{S}}_0)(\boldsymbol{S}_i - \bar{\boldsymbol{S}}_0)^\top,$$

$$\widehat{\boldsymbol{C}}_{01} = \frac{1}{n_1}\sum_{i=1}^{n} D_i\, q(\boldsymbol{W}_i^\top \widehat{\boldsymbol{\eta}})\{\psi(\boldsymbol{X}_i^\top \widehat{\boldsymbol{\gamma}}) - Y_i\}\boldsymbol{X}_i(\boldsymbol{S}_i - \bar{\boldsymbol{S}}_0)^\top,$$

respectively, where $\boldsymbol{W}_i$ can be replaced by $\boldsymbol{X}_i$ under (A3). The consistency of $\widehat{\boldsymbol{C}}_0$ and $\widehat{\boldsymbol{C}}_{01}$ follows from the law of large numbers, (A3) or (A3′), and the fact that $E\{\kappa(\boldsymbol{W})(\boldsymbol{S}-\boldsymbol{\varsigma}_*)(\boldsymbol{S}-\boldsymbol{\varsigma}_*)^\top \,|\, D=1\} = E\{(\boldsymbol{S}-\boldsymbol{\varsigma}_*)(\boldsymbol{S}-\boldsymbol{\varsigma}_*)^\top \,|\, D=0\} = \boldsymbol{C}_0$ and $E[\kappa(\boldsymbol{W})\{\psi(\boldsymbol{X}^\top\boldsymbol{\gamma}_*)-Y\}\boldsymbol{X}\boldsymbol{S}^\top \,|\, D=1] = E[\{\psi(\boldsymbol{X}^\top\boldsymbol{\gamma}_*)-Y\}\boldsymbol{X}\boldsymbol{S}^\top \,|\, D=0] = \boldsymbol{C}_{01}$. For $\boldsymbol{\Lambda}_0$, the consistency of $\widehat{\boldsymbol{\Lambda}}_0$ is shown in the Appendix under an additional minor condition that $E(Y^2 \,|\, \boldsymbol{X}, D=1) = E(Y^2 \,|\, \boldsymbol{X}, D=0)$.

Although the substitution estimator of $\boldsymbol{V}_+$ is consistent, it often underestimates variances (MacKinnon and White, 1985). Thus, we consider a bootstrap procedure (Efron and Tibshirani, 1993) as an alternative to the substitution method. Internal bootstrap data $(Y_i^*, \boldsymbol{W}_i^*)$, $i = 1, ..., n_1$, are generated as a simple random sample with replacement from $(Y_i, \boldsymbol{W}_i)$, $i = 1, ..., n_1$. Since we do not have external individual level data, we generate bootstrap analog of $\bar{\boldsymbol{S}}_0$ and $\widehat{\boldsymbol{\gamma}}$ using

$$\begin{pmatrix} \bar{\boldsymbol{S}}_0^* \\ \widehat{\boldsymbol{\gamma}}^* \end{pmatrix} \sim N\left(\begin{pmatrix} \bar{\boldsymbol{S}}_0 \\ \widehat{\boldsymbol{\gamma}} \end{pmatrix}, \frac{1}{n_0}\begin{pmatrix} \widehat{\boldsymbol{C}}_0 & -\widehat{\boldsymbol{C}}_{01}^\top \widehat{\boldsymbol{H}}_1^{-1} \\ -\widehat{\boldsymbol{H}}_1^{-1}\widehat{\boldsymbol{C}}_{01} & \widehat{\boldsymbol{H}}_1^{-1}\widehat{\boldsymbol{\Lambda}}_0\widehat{\boldsymbol{H}}_1^{-1} \end{pmatrix}\right),$$

where $\widehat{\boldsymbol{C}}_0$, $\widehat{\boldsymbol{C}}_{01}$ and $\widehat{\boldsymbol{\Lambda}}_0$ are the same as previously defined and

$$\widehat{\boldsymbol{H}}_1 = \frac{1}{n_1}\sum_{i=1}^{n} D_i q(\boldsymbol{X}_i^\top \widehat{\boldsymbol{\eta}})\psi'(\boldsymbol{X}_i^\top \widehat{\boldsymbol{\gamma}})\boldsymbol{X}_i\boldsymbol{X}_i^\top.$$

The bootstrap GMM estimator $(\widehat{\boldsymbol{\beta}}^*, \widehat{\boldsymbol{\eta}}^*)$ is calculated according to (12) based on bootstrap internal data and $\bar{\boldsymbol{S}}_0^*$ and $\widehat{\boldsymbol{\gamma}}^*$. This bootstrap procedure is repeated independently B (e.g., 100) times and the bootstrap variance estimator for $\widehat{\boldsymbol{\beta}}$ is the sample variance of B $\widehat{\boldsymbol{\beta}}^*$'s.

Note that this bootstrap method requires extra repeated computations. Thus, when ϕ in (B2) is nonlinear, we suggest a linearized bootstrap calculating

$$\begin{pmatrix} \widehat{\boldsymbol{\beta}}^* - \widehat{\boldsymbol{\beta}} \\ \widehat{\boldsymbol{\eta}}^* - \widehat{\boldsymbol{\eta}} \end{pmatrix} = -(\widehat{\boldsymbol{M}}_+^{*\top} \widehat{\boldsymbol{\Sigma}}_+^{*-1} \widehat{\boldsymbol{M}}_+^*)^{-1} \widehat{\boldsymbol{M}}_+^{*\top} \widehat{\boldsymbol{\Sigma}}_+^{*-1} \left\{ \bar{\boldsymbol{g}}^*(\widehat{\boldsymbol{\gamma}}, \bar{\boldsymbol{S}}_0, \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\eta}}) + \begin{pmatrix} \mathbf{0} \\ -(\bar{\boldsymbol{S}}_0^* - \bar{\boldsymbol{S}}_0) \\ \widehat{\boldsymbol{H}}_1^*(\widehat{\boldsymbol{\gamma}}^* - \widehat{\boldsymbol{\gamma}}) \end{pmatrix} \right\},$$

where $\bar{\boldsymbol{g}}^*(\widehat{\boldsymbol{\gamma}}, \bar{\boldsymbol{S}}_0, \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\eta}})$, $\widehat{\boldsymbol{H}}_1^*$, $\widehat{\boldsymbol{\Sigma}}_+^*$, and $\widehat{\boldsymbol{M}}_+^*$ are $\bar{\boldsymbol{g}}(\widehat{\boldsymbol{\gamma}}, \bar{\boldsymbol{S}}_0, \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\eta}})$, $\widehat{\boldsymbol{H}}_1$, $\widehat{\boldsymbol{\Sigma}}_+$, and $\widehat{\boldsymbol{M}}_+$, respectively, with $(Y_i, \boldsymbol{W}_i)$'s replaced by $(Y_i^*, \boldsymbol{W}_i^*)$'s.

4 Empirical Results

4.1 Simulations

We first present some simulation results to examine the finite-sample performance of the proposed GMM estimators (3) and (12) and to compare them with the GEE estimator without using external summary information. Also, our simulation studies the performance of variance estimators described in Section 3.3 and the related asymptotic confidence intervals for components of $\boldsymbol{\beta}$ in (B2).

We consider a 4-dimensional normally distributed covariate vector

$$\begin{pmatrix} U_1 \\ U_2 \\ U_3 \\ U_4 \end{pmatrix} \sim N\left(\begin{pmatrix} 1 \\ 0 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & -0.4 & 0.4 & 0.3 \\ -0.4 & 1 & 0.3 & 0.4 \\ 0.4 & 0.3 & 1 & 0.4 \\ 0.3 & 0.4 & 0.4 & 1 \end{pmatrix} \right).$$

Following the notation in Sections 2-3, in internal analysis we use $\boldsymbol{W} = (\boldsymbol{X}^\top, \boldsymbol{Z}^\top)^\top$ with $\boldsymbol{X} = (1, U_1, U_2, U_3)^\top$ and $\boldsymbol{Z} = (U_4, U_1U_2, U_1U_4)^\top$. In the external dataset, we only have $\boldsymbol{X}$, not $\boldsymbol{Z}$, where $\boldsymbol{Z}$ contains not only U_4, but also the cross-products U_1U_2 and U_1U_4.

Two GEE models (B2) are considered in internal analysis.

(i) Linear model $E(Y \,|\, \boldsymbol{W}) = \boldsymbol{W}^\top \boldsymbol{\beta}$, where $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6)^\top = (-1.6, 0.7, 0.4, 0.5, 0.4, 0.4, 0.4)^\top$ is a 7-dimensional parameter vector and $Y \mid \boldsymbol{W}$ is normally distributed but we do not use the normality information in analysis.

(ii) Logistic model with binary Y and $P(Y = 1 \,|\, \boldsymbol{W}) = 1/\{1 + \exp(\boldsymbol{W}^\top \boldsymbol{\beta})\}$, where $\boldsymbol{\beta}$ is the 7-dimensional parameter vector in (i).

For the external data, we consider GEE (1) with $\psi(t) = t$ for linear model and $\psi(t) = 1/\{1 + \exp(t)\}$ for the logistic setting. Although the model for internal data is correct, the model for external data is wrong because, under the logistic setting, $P(Y = 1 \,|\, \boldsymbol{X}) = E\{P(Y = 1 \,|\, \boldsymbol{W}) \,|\, \boldsymbol{X}\}$ is no longer logistic and, under the linear model setting, the product terms U_1U_2 and U_1U_4 in $\boldsymbol{W}$ are nonlinear.

To generate internal and external data, we consider random D with

$$P(D = 1 \,|\, \boldsymbol{X}, \boldsymbol{Z}, Y) = \frac{1}{1 + \exp(\eta_0 + \eta_1 U_1 + \eta_2 U_4)}.$$

Consequently,

$$q(\boldsymbol{W}^\top \boldsymbol{\eta}) = \exp\left(\log r + \eta_0 + \eta_1 U_1 + \eta_2 U_4\right),$$

where $r = P(D = 1)/P(D = 0)$. Three sets of $\boldsymbol{\eta}$'s are considered; (1) $\eta_0 = 2.3$, $\eta_1 = \eta_2 = 0$; (2) $\eta_0 = 3.6$, $\eta_1 = -1$, $\eta_2 = 0$; (3) $\eta_0 = 3.8$, $\eta_1 = -1$, $\eta_2 = 1$. These values are chosen so that $r \approx 0.1$. When $\eta_2 = 0$, $q(\boldsymbol{W}^\top \boldsymbol{\eta}) = q(\boldsymbol{X}^\top \boldsymbol{\eta})$. Throughout, we consider $\boldsymbol{S} = \boldsymbol{X}$.

We consider the total sample size $n = 5,500$. Since $r \approx 0.1$, the internal sample size n_1 is around 500 and the external sample size n_0 is around $5,000$.

Based on 1,000 simulation runs, Tables 1-2 provide the bias, standard deviation (SD), average standard error (SE), and coverage probability (CP) of 95% asymptotic confidence interval for the estimation of β_j's under the linear and logistic model settings (i)-(ii), respectively. The SE of GEE estimator using internal data only is based on substitution. The SE of GMM (12) is computed using both the substitution and bootstrap described in Section 3.3. The SE of GMM (3) is computed using the substitution; the bootstrap is used

only in the case where $\eta_1 = \eta_2 = 0$, because the GMM (3) is incorrect when either η_1 or η_2 is nonzero.

The following is a summary of the results in Tables 1-2.

1. The empirical results confirm our asymptotic theory, i.e., using external summary information substantially improves efficiency and the confidence intervals based on asymptotic theory work well. The SD of GMM (12), or GMM (3) when it is correct, is much smaller than the SD of GEE without using external information, for the estimation of $\boldsymbol{\beta}$-coefficients in front of $\boldsymbol{X}$. The improvement in many cases is over 50%. Although in some cases using internal data only leads to a CP closer to 95%, this advantage is built on a much longer confidence interval. For estimation of $\boldsymbol{\beta}$-coefficients in front of $\boldsymbol{Z}$, the GMM does not improve or has a slight improvement, because $\boldsymbol{Z}$-information is not in the external data.

2. If the heterogeneity between internal and external populations exists and is not well addressed, the use of external information may create a non-negligible bias of GMM (3) leading to a very low CP. It also affects the convergence of GMM (3), for example, the SD of GMM (3) is extremely large in some cases under the logistic setting. On the other hand, the correctness of GMM (12) is not affected by whether heterogeneity in population exists or not, and two GMM estimators have comparable performance when populations are homogeneous.

3. For GMM (12), the SE based on substitution may underestimate, especially when η_1 or η_2 is not 0. When this occurs, the SE based on bootstrap is better, although the bootstrap SE sometimes leads to a conservative CP.

4.2 An example

We illustrate our approach with data from the National Health and Nutrition Examination Survey (CDCP, 2018), a program collecting the health and nutritional status of nationally

representative adults and children across the United States. The internal dataset consists of 500 sampled units from 2017-2018 survey cycle, in which the response Y is the systolic blood pressure and associated covariates of interest are gender, age, total cholesterol (mmol/L) and triglycerides (mmol/L). The external dataset contains results from 6,755 sampled units in the 2015-2016 survey with systolic blood pressure, gender, age, and total cholesterol, but not triglycerides. With the notation in Section 2, $\boldsymbol{X} = (1, \text{gender, age, total cholesterol})^\top$ and $Z =$ triglycerides. We only use the summary information from the external dataset, i.e., the mean values of components of $\boldsymbol{X}$ and the GEE estimates of the parameters in a working linear model between Y and $\boldsymbol{X}$.

We compute three estimates of $\boldsymbol{\beta}$ in (B2) with $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3, \beta_4)^\top$ and $\boldsymbol{W} = (\boldsymbol{X}^\top, Z)^\top$, which are the GEE estimate with internal data only, the GMM (3) using external summary information but assuming identical internal and external populations, and the GMM (12) with possibly heterogeneous internal and external populations. The estimates together with standard errors (SE) calculated by substitution are shown in Table 3. Both GMM estimates (3) and (12) have considerably smaller standard errors than the GEE estimate using internal data only, except for the estimation of β_4 for triglycerides as the external dataset has no information about triglycerides. The two GMM estimates are not significantly different in the estimation of $\boldsymbol{\beta}$-coefficients in front of gender, age, and triglycerides, but they are different in the estimation of intercept β_0 and β_3 for cholesterol, which indicates that the assumption of identical internal and external populations is questionable.

Supplementary Material

The supplementary material contains computer codes.

Acknowledgements

The authors would like to thank Editor-in-Chief, Associate Editor-in-Chief, and two anonymous referees for helpful comments and suggestions. Jun Shao's research was partially supported by the National Natural Science Foundation of China (11831008) and the U.S. National Science Foundation (DMS-1914411). Lei Wang's research was supported by the Fundamental Research Funds for the Central Universities and the National Natural Science Foundation of China (12271272, 11871287, 11771144, 11801359).

Appendix

Proof of Theorem 1. Since $\widehat{\boldsymbol{\beta}}$ is a solution to (3),

$$0 = \nabla_{\boldsymbol{\beta}}\, \bar{\boldsymbol{g}}(\widehat{\boldsymbol{\gamma}}, \widehat{\boldsymbol{\beta}})^\top \widehat{\boldsymbol{\Sigma}}^{-1} \bar{\boldsymbol{g}}(\widehat{\boldsymbol{\gamma}}, \widehat{\boldsymbol{\beta}}).$$

From Taylor's expansion of $\bar{\boldsymbol{g}}(\boldsymbol{\gamma}, \boldsymbol{\beta})$ at $\boldsymbol{\beta} = \boldsymbol{\beta}_*$ and $\boldsymbol{\gamma} = \boldsymbol{\gamma}_*$,

$$\bar{\boldsymbol{g}}(\widehat{\boldsymbol{\gamma}}, \widehat{\boldsymbol{\beta}}) = \bar{\boldsymbol{g}}(\boldsymbol{\gamma}_*, \boldsymbol{\beta}_*) + \nabla_{\boldsymbol{\beta}}\, \bar{\boldsymbol{g}}(\tilde{\boldsymbol{\gamma}}, \tilde{\boldsymbol{\beta}})(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_*) + \nabla_{\boldsymbol{\gamma}}\, \bar{\boldsymbol{g}}(\tilde{\boldsymbol{\gamma}}, \tilde{\boldsymbol{\beta}})(\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_*),$$

where $\tilde{\boldsymbol{\beta}}$ is between $\boldsymbol{\beta}_*$ and $\widehat{\boldsymbol{\beta}}$, and $\tilde{\boldsymbol{\gamma}}$ is between $\boldsymbol{\gamma}_*$ and $\widehat{\boldsymbol{\gamma}}$. Combining the two equations we obtain that

$$\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_* = -\Big\{\nabla_{\boldsymbol{\beta}}\, \bar{\boldsymbol{g}}(\widehat{\boldsymbol{\gamma}}, \widehat{\boldsymbol{\beta}})^\top \widehat{\boldsymbol{\Sigma}}^{-1} \nabla_{\boldsymbol{\beta}}\, \bar{\boldsymbol{g}}(\tilde{\boldsymbol{\gamma}}, \tilde{\boldsymbol{\beta}})\Big\}^{-1} \nabla_{\boldsymbol{\beta}}\, \bar{\boldsymbol{g}}(\widehat{\boldsymbol{\gamma}}, \widehat{\boldsymbol{\beta}})^\top \widehat{\boldsymbol{\Sigma}}^{-1} \Big\{\bar{\boldsymbol{g}}(\boldsymbol{\gamma}_*, \boldsymbol{\beta}_*) + \nabla_{\boldsymbol{\gamma}}\, \bar{\boldsymbol{g}}(\tilde{\boldsymbol{\gamma}}, \tilde{\boldsymbol{\beta}})(\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_*)\Big\}.$$

Under the assumed conditions in Theorem 1, $\sqrt{n_0}(\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_*)\,|\,n_0 \xrightarrow{d} N(\mathbf{0}, \boldsymbol{H}^{-1}\boldsymbol{\Lambda}\boldsymbol{H}^{-1})$, where $n_0 = n - n_1$ and $\boldsymbol{H} = E\{\psi'(\boldsymbol{X}^\top\boldsymbol{\gamma}_*)\boldsymbol{X}\boldsymbol{X}^\top\}$. Hence, $\sqrt{n_1}\,\boldsymbol{H}(\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_*) \xrightarrow{d} N(\mathbf{0}, r\boldsymbol{\Lambda})$ since $n_1/n_0 \to r$. Further, under the assumed conditions, $\widehat{\boldsymbol{\Sigma}}^{-1} \xrightarrow{p} \boldsymbol{\Sigma}^{-1}$, $\nabla_{\boldsymbol{\beta}}\, \bar{\boldsymbol{g}}(\widehat{\boldsymbol{\gamma}}, \widehat{\boldsymbol{\beta}}) \xrightarrow{p} \boldsymbol{M}$, $\nabla_{\boldsymbol{\beta}}\, \bar{\boldsymbol{g}}(\tilde{\boldsymbol{\gamma}}, \tilde{\boldsymbol{\beta}}) \xrightarrow{p} \boldsymbol{M}$, and $\nabla_{\boldsymbol{\gamma}}\, \bar{\boldsymbol{g}}(\tilde{\boldsymbol{\gamma}}, \tilde{\boldsymbol{\beta}}) \xrightarrow{p} E\{\nabla_{\boldsymbol{\gamma}}\, \bar{\boldsymbol{g}}(\boldsymbol{\gamma}_*, \boldsymbol{\beta}_*)\} = \binom{\mathbf{0}}{\boldsymbol{H}}$, where $\xrightarrow{p}$ denotes convergence in probability. Hence,

$$\sqrt{n_1}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_*) = -\sqrt{n_1}(\boldsymbol{M}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{M})^{-1}\boldsymbol{M}^\top\boldsymbol{\Sigma}^{-1}\{\bar{\boldsymbol{g}}(\boldsymbol{\gamma}_*, \boldsymbol{\beta}_*) + \binom{\mathbf{0}}{\boldsymbol{H}(\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_*)}\} + o_p(1),$$

where $o_p(1)$ denotes a term converging to 0 in probability. From the central limit theorem, $\sqrt{n_1}\,\bar{\boldsymbol{g}}(\boldsymbol{\gamma}_*, \boldsymbol{\beta}_*) \xrightarrow{d} N(\mathbf{0}, \boldsymbol{\Sigma})$. Then the result follows from the independence between $\widehat{\boldsymbol{\gamma}}$ and internal data and

$$\sqrt{n_1}\begin{pmatrix} \mathbf{0} \\ \boldsymbol{H}(\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_*) \end{pmatrix} \xrightarrow{d} N\left(\mathbf{0},\ r\begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Lambda} \end{pmatrix}\right).$$

In the previous argument, we use the following result repeatedly to handle the case where n_1 is random. If a quantity $\boldsymbol{Q}_n$ satisfies $\sqrt{n_1}\boldsymbol{Q}_n \,|\, n_1 \xrightarrow{d} N(\mathbf{0}, \boldsymbol{\Omega})$ for some fixed matrix $\boldsymbol{\Omega}$ not depending on n, then unconditionally we also have $\sqrt{n_1}\boldsymbol{Q}_n \xrightarrow{d} N(\mathbf{0}, \boldsymbol{\Omega})$.

Proof of Theorem 2. Following the proof of Theorem 1, we obtain that

$$\sqrt{n_1}\begin{pmatrix} \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_* \\ \widehat{\boldsymbol{\eta}} - \boldsymbol{\eta}_* \end{pmatrix} = -\sqrt{n_1}(\boldsymbol{M}_+^\top \boldsymbol{\Sigma}_+^{-1} \boldsymbol{M}_+)^{-1} \boldsymbol{M}_+^\top \boldsymbol{\Sigma}_+^{-1} \left\{ \bar{\boldsymbol{g}}(\boldsymbol{\gamma}_*, \boldsymbol{\varsigma}_*, \boldsymbol{\beta}_*, \boldsymbol{\eta}_*) + \begin{pmatrix} \mathbf{0} \\ -(\bar{\boldsymbol{S}}_0 - \boldsymbol{\varsigma}_*) \\ \boldsymbol{H}_1(\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_*) \end{pmatrix} \right\} + o_p(1),$$

where $\boldsymbol{H}_1 = E[q(\boldsymbol{X}^\top \boldsymbol{\eta}_*)\psi'(\boldsymbol{X}^\top \boldsymbol{\gamma}_*)\boldsymbol{X}\boldsymbol{X}^\top \mid D = 1] = \boldsymbol{H}_0$ under (A3). Then the result follows from the independence between $(\widehat{\boldsymbol{\gamma}}, \bar{\boldsymbol{S}}_0)$ and internal data and

$$\sqrt{n_1}\begin{pmatrix} \mathbf{0} \\ -(\bar{\boldsymbol{S}}_0 - \boldsymbol{\varsigma}_*) \\ \boldsymbol{H}_1(\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_*) \end{pmatrix} \xrightarrow{d} N\left(\mathbf{0},\ r\begin{pmatrix} \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{C}_0 & \boldsymbol{C}_{01}^\top \\ \mathbf{0} & \boldsymbol{C}_{01} & \boldsymbol{\Lambda}_0 \end{pmatrix}\right).$$

Proof of the consistency of $\widehat{\boldsymbol{\Lambda}}_0$ in Section 3.3. Under the assumed conditions and the law of large numbers, $\widehat{\boldsymbol{\Lambda}}_0$ converges in probability to

$$\begin{aligned}
& E[\kappa(\boldsymbol{W})\{Y - \psi(\boldsymbol{X}^\top \boldsymbol{\gamma}_*)\}^2 \boldsymbol{X}\boldsymbol{X}^\top \,|\, D = 1] \\
&= E\left\{\kappa(\boldsymbol{W}) E[\{Y - \psi(\boldsymbol{X}^\top \boldsymbol{\gamma}_*)\}^2 \boldsymbol{X}\boldsymbol{X}^\top \,|\, \boldsymbol{X}, D = 1] \,|\, D = 1\right\} \\
&= E\left\{E[\{Y - \psi(\boldsymbol{X}^\top \boldsymbol{\gamma}_*)\}^2 \boldsymbol{X}\boldsymbol{X}^\top \,|\, \boldsymbol{X}, D = 1] \,|\, D = 0\right\} \\
&= E\left\{E[\{Y - \psi(\boldsymbol{X}^\top \boldsymbol{\gamma}_*)\}^2 \boldsymbol{X}\boldsymbol{X}^\top \,|\, \boldsymbol{X}, D = 0] \,|\, D = 0\right\} \\
&= E[\{Y - \psi(\boldsymbol{X}^\top \boldsymbol{\gamma}_*)\}^2 \boldsymbol{X}\boldsymbol{X}^\top \,|\, D = 0] \\
&= \boldsymbol{\Lambda}_0.
\end{aligned}$$

References

Caner, M. (2009). Lasso-type gmm estimator. *Econometric Theory*, 25(1):270–290.

CDCP (2018). National health and nutrition examination survey data. *Hyattsville, MD: US Department of Health and Human Services, Centers for Disease Control and Prevention (CDCP), 2020*.

Chatterjee, N., Chen, Y. H., Maas, P., and Carroll, R. J. (2016). Constrained maximum likelihood estimation for model calibration using summary-level information from external big data sources. *Journal of the American Statistical Association*, 111(513):107–117.

Efron, B. and Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*. Chapman and Halll/CRC, Boca Raton, FL, U.S.

Hansen, L. (1982). Large sample properties of generalized method of moments estimators. *Econometrica*, 50(4):1029–1054.

He, Q., Zhang, H. H., Avery, C. L., and Lin, D. (2016). Sparse meta-analysis with high-dimensional data. *Biostatistics*, 17(2):205–220.

Kim, H. J., Wang, Z., and Kim, J. K. (2021). Survey data integration for regression analysis using model calibration. *arXiv 2107.06448*.

Kundu, P., Tang, R., and Charterjee, N. (2019). Generalized meta-analysis for multiple regression models across studies with disparate covariate information. *Biometrika*, 106(2):567–585.

Li, S., Cai, T. T., and Li, H. (2021). Transfer learning for high-dimensional linear regression: prediction, estimation and minimax optimality. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), to appear.*

Li, S., Cai, T. T., and Li, H. (2022). Estimation and inference with proxy data and its genetic applications. *arXiv preprint arXiv:2201.03727*.

Liao, Z. (2013). Adaptive gmm shrinkage estimation with consistent moment selection. *Econometric Theory*, 29(5):857–904.

Lin, D. and Zeng, D. (2010). On the relative efficiency of using summary statistics versus individual-level data in meta-analysis. *Biometrika*, 97(2):321–332.

Lohr, S. L. and Raghunathan, T. E. (2017). Combining survey data with other data sources. *Statistical Science*, 32(2):293–312.

MacKinnon, J. G. and White, H. (1985). Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties. *Journal of Econometrics*, 29(3):305–325.

Merkouris, T. (2004). Combining independent regression estimators from multiple surveys. *Journal of the American Statistical Association*, 99(468):1131–1139.

Rao, J. (2021). On making valid inferences by integrating data from surveys and other sources. *Sankhya B*, 83(1):242–272.

Tian, Y. and Feng, Y. (2022). Transfer learning under high-dimensional generalized linear models. *Journal of the American Statistical Association, to appear.*

Yang, S. and Kim, J. K. (2020). Statistical data integration in survey sampling: a review. *Japanese Journal of Statistics and Data Science*, 3(2):625–650.

Yang, S., Zeng, D., and Wang, X. (2020). Elastic integrative analysis of randomized trial and real-world data for treatment heterogeneity estimation. *arXiv preprint arXiv:2005.10579*.

Zhang, H., Deng, L., Schiffman, M., Qin, J., and Yu, K. (2020). Generalized integration model for improved statistical inference by leveraging external summary data. *Biometrika*, 107(3):689–703.

Zhang, Y., Ouyang, Z., and Zhao, H. (2017). A statistical framework for data integration through graphical models with application to cancer genomics. *The Annals of Applied Statistics*, 11(1):161–184.

Table 1. Bias, standard deviation (SD), average standard error (SE), and coverage probability (CP) of 95% asymptotic confidence interval based on 1,000 simulations under linear model setting

							substitution		bootstrap	
η_0	η_1	η_2	estimand	method	bias	SD	SE	CP	SE	CP
2.3	0	0	β_0	GEE (internal)	-0.001	0.167	0.169	0.950		
				GMM (12)	-0.007	0.111	0.110	0.954	0.110	0.965
				GMM (3)	-0.011	0.110	0.112	0.954	0.114	0.957
			β_1	GEE (internal)	0.005	0.147	0.144	0.951		
				GMM (12)	0.004	0.096	0.097	0.958	0.100	0.966
				GMM (3)	0.006	0.096	0.099	0.965	0.100	0.970
			β_2	GEE (internal)	0.008	0.166	0.170	0.955		
				GMM (12)	0.008	0.138	0.134	0.941	0.142	0.954
				GMM (3)	0.010	0.136	0.136	0.945	0.140	0.951
			β_3	GEE (internal)	-0.003	0.123	0.122	0.947		
				GMM (12)	0.004	0.065	0.062	0.932	0.064	0.940
				GMM (3)	0.004	0.064	0.062	0.942	0.062	0.938
			β_4	GEE (internal)	-0.010	0.150	0.154	0.951		
				GMM (12)	-0.011	0.153	0.153	0.955	0.161	0.963
				GMM (3)	-0.011	0.150	0.153	0.952	0.157	0.956
			β_5	GEE (internal)	0.002	0.090	0.090	0.943		
				GMM (12)	-0.001	0.093	0.089	0.937	0.096	0.955
				GMM (3)	-0.001	0.091	0.089	0.940	0.090	0.951
			β_6	GEE (internal)	0.001	0.091	0.093	0.944		
				GMM (12)	-0.000	0.094	0.092	0.944	0.098	0.955
				GMM (3)	-0.001	0.092	0.093	0.944	0.096	0.946

Table 1 (continued)

							substitution		bootstrap	
η_0	η_1	η_2	estimand	method	bias	SD	SE	CP	SE	CP
3.6	-1.0	0	β_0	GEE (internal)	0.013	0.217	0.216	0.947		
				GMM (12)	-0.004	0.129	0.119	0.926	0.129	0.947
				GMM (3)	0.117	0.138	0.139	0.880		
			β_1	GEE (internal)	-0.009	0.151	0.148	0.944		
				GMM (12)	0.003	0.106	0.097	0.930	0.104	0.946
				GMM (3)	-0.112	0.114	0.111	0.817		
			β_2	GEE (internal)	-0.007	0.218	0.220	0.951		
				GMM (12)	-0.024	0.154	0.142	0.927	0.157	0.948
				GMM (3)	-0.362	0.204	0.207	0.597		
			β_3	GEE (internal)	0.001	0.119	0.118	0.946		
				GMM (12)	-0.001	0.072	0.065	0.931	0.072	0.947
				GMM (3)	0.003	0.065	0.067	0.954		
			β_4	GEE (internal)	-0.003	0.205	0.211	0.944		
				GMM (12)	-0.001	0.191	0.187	0.940	0.205	0.962
				GMM (3)	-0.037	0.209	0.213	0.945		
			β_5	GEE (internal)	0.000	0.097	0.094	0.937		
				GMM (12)	0.006	0.078	0.070	0.914	0.078	0.937
				GMM (3)	0.027	0.111	0.095	0.891		
			β_6	GEE (internal)	0.001	0.098	0.096	0.938		
				GMM (12)	-0.002	0.087	0.081	0.926	0.090	0.955
				GMM (3)	0.016	0.102	0.097	0.931		

Table 1 (continued)

							substitution		bootstrap	
η_0	η_1	η_2	estimand	method	bias	SD	SE	CP	SE	CP
3.8	-1.0	1.0	β_0	GEE (internal)	0.006	0.258	0.254	0.944		
				GMM (12)	-0.028	0.169	0.144	0.900	0.170	0.931
				GMM (3)	0.167	0.211	0.215	0.895		
			β_1	GEE (internal)	0.002	0.164	0.162	0.946		
				GMM (12)	0.018	0.116	0.100	0.902	0.114	0.934
				GMM (3)	0.187	0.137	0.133	0.713		
			β_2	GEE (internal)	-0.001	0.203	0.207	0.953		
				GMM (12)	-0.003	0.168	0.149	0.909	0.167	0.953
				GMM (3)	-0.369	0.235	0.191	0.538		
			β_3	GEE (internal)	-0.006	0.123	0.121	0.950		
				GMM (12)	0.001	0.081	0.070	0.921	0.079	0.945
				GMM (3)	0.007	0.081	0.072	0.921		
			β_4	GEE (internal)	0.007	0.189	0.191	0.953		
				GMM (12)	-0.005	0.184	0.173	0.929	0.191	0.953
				GMM (3)	0.046	0.198	0.195	0.941		
			β_5	GEE (internal)	0.004	0.098	0.096	0.043		
				GMM (12)	0.005	0.088	0.080	0.908	0.088	0.938
				GMM (3)	0.212	0.157	0.100	0.475		
			β_6	GEE (internal)	-0.004	0.100	0.094	0.932		
				GMM (12)	-0.002	0.099	0.090	0.925	0.098	0.948
				GMM (3)	-0.049	0.115	0.096	0.872		

Table 2. Bias, standard deviation (SD), average standard error (SE), and coverage probability (CP) of 95% asymptotic confidence interval based on 1,000 simulations under logistic model setting

							substitution		bootstrap	
η_0	η_1	η_2	estimand	method	bias	SD	SE	CP	SE	CP
2.3	0	0	β_0	GEE (internal)	-0.043	0.244	0.245	0.960		
				GMM (12)	-0.014	0.121	0.121	0.953	0.125	0.960
				GMM (3)	-0.012	0.122	0.122	0.956	0.133	0.965
			β_1	GEE (internal)	0.032	0.207	0.199	0.944		
				GMM (12)	0.009	0.116	0.112	0.945	0.114	0.950
				GMM (3)	0.006	0.116	0.113	0.944	0.121	0.956
			β_2	GEE (internal)	-0.012	0.239	0.235	0.941		
				GMM (12)	-0.027	0.172	0.166	0.946	0.171	0.949
				GMM (3)	-0.029	0.172	0.168	0.944	0.184	0.956
			β_3	GEE (internal)	0.012	0.174	0.165	0.935		
				GMM (12)	0.010	0.080	0.075	0.931	0.076	0.932
				GMM (3)	0.009	0.077	0.075	0.942	0.080	0.954
			β_4	GEE (internal)	0.023	0.233	0.222	0.937		
				GMM (12)	0.023	0.235	0.217	0.928	0.224	0.935
				GMM (3)	0.027	0.236	0.218	0.926	0.237	0.926
			β_5	GEE (internal)	0.026	0.146	0.141	0.939		
				GMM (12)	0.029	0.151	0.140	0.933	0.143	0.938
				GMM (3)	0.029	0.147	0.140	0.936	0.157	0.937
			β_6	GEE (internal)	0.003	0.175	0.166	0.919		
				GMM (12)	0.004	0.177	0.163	0.916	0.168	0.923
				GMM (3)	0.002	0.177	0.163	0.917	0.192	0.958

Table 2 (continued)

							substitution		bootstrap	
η_0	η_1	η_2	estimand	method	bias	SD	SE	CP	SE	CP
3.6	-1.0	0	β_0	GEE (internal)	-0.043	0.315	0.309	0.947		
				GMM (12)	-0.014	0.143	0.136	0.938	0.147	0.950
				GMM (3)	-0.165	3.523	5.265	0.968		
			β_1	GEE (internal)	0.017	0.212	0.210	0.950		
				GMM (12)	0.007	0.130	0.120	0.936	0.127	0.948
				GMM (3)	0.074	2.472	3.694	0.942		
			β_2	GEE (internal)	-0.027	0.307	0.306	0.951		
				GMM (12)	-0.050	0.193	0.174	0.926	0.188	0.939
				GMM (3)	-0.548	3.224	4.839	0.632		
			β_3	GEE (internal)	0.016	0.170	0.165	0.936		
				GMM (12)	0.005	0.082	0.080	0.949	0.085	0.960
				GMM (3)	0.020	0.109	0.204	0.947		
			β_4	GEE (internal)	0.056	0.355	0.335	0.937		
				GMM (12)	0.040	0.330	0.302	0.925	0.317	0.935
				GMM (3)	0.034	0.523	0.793	0.927		
			β_5	GEE (internal)	0.025	0.151	0.145	0.938		
				GMM (12)	0.031	0.119	0.109	0.927	0.115	0.946
				GMM (3)	0.122	2.296	3.415	0.930		
			β_6	GEE (internal)	-0.013	0.187	0.181	0.940		
				GMM (12)	-0.001	0.174	0.164	0.938	0.170	0.947
				GMM (3)	0.018	0.511	0.829	0.933		

Table 2 (continued)

							substitution		bootstrap	
η_0	η_1	η_2	estimand	method	bias	SD	SE	CP	SE	CP
3.8	-1.0	1.0	β_0	GEE (internal)	-0.050	0.364	0.353	0.941		
				GMM (12)	-0.030	0.195	0.166	0.893	0.189	0.929
				GMM (3)	-6.187	25.43	2329	0.917		
			β_1	GEE (internal)	0.030	0.240	0.231	0.943		
				GMM (12)	0.021	0.140	0.124	0.917	0.138	0.937
				GMM (3)	6.296	22.02	2328	0.748		
			β_2	GEE (internal)	0.011	0.333	0.318	0.941		
				GMM (12)	-0.035	0.209	0.194	0.047	0.212	0.956
				GMM (3)	-8.351	37.03	3779	0.797		
			β_3	GEE (internal)	0.012	0.184	0.176	0.940		
				GMM (12)	0.002	0.097	0.087	0.932	0.096	0.954
				GMM (3)	1.430	5.916	412.7	0.936		
			β_4	GEE (internal)	0.043	0.322	0.309	0.939		
				GMM (12)	0.036	0.308	0.272	0.922	0.292	0.947
				GMM (3)	5.981	26.77	2424	0.902		
			β_5	GEE (internal)	0.013	0.172	0.163	0.946		
				GMM (12)	0.022	0.136	0.129	0.952	0.139	0.959
				GMM (3)	6.683	27.06	2854	0.881		
			β_6	GEE (internal)	-0.008	0.183	0.174	0.934		
				GMM (12)	-0.001	0.180	0.162	0.921	0.172	0.933
				GMM (3)	0.934	13.62	877.8	0.900		

Table 3. Estimate and standard error (SE) in the example with data from the National Health and Nutrition Examination Survey (Section 4.2)

estimand	method	estimate	SE
β_0: intercept	GEE (internal)	95.5358	3.3002
	GMM (12)	95.4741	0.9782
	GMM (3)	98.5944	0.9857
β_1: gender	GEE (internal)	-2.6235	1.4682
	GMM (12)	-2.7205	0.4953
	GMM (3)	-2.8063	0.4982
β_2: age	GEE (internal)	0.4239	0.0372
	GMM (12)	0.4499	0.0106
	GMM (3)	0.4532	0.0105
β_3: cholesterol	GEE (internal)	1.9175	0.7426
	GMM (12)	1.5935	0.3172
	GMM (3)	0.7960	0.3164
β_4: triglycerides	GEE (internal)	1.0296	0.4954
	GMM (12)	0.7727	0.4927
	GMM (3)	1.0780	0.4951