Statistical Theory and Related Fields

ISSN: 2475-4269 (Print) 2475-4277 (Online) Journal homepage: www.tandfonline.com/journals/tstf20

Two-stage least squares model averaging for instrumental variable models with exogenous variables

Wenjun Shen & Xiaochao Xia

To cite this article: Wenjun Shen & Xiaochao Xia (07 Mar 2026): Two-stage least squares model averaging for instrumental variable models with exogenous variables, Statistical Theory and Related Fields, DOI: 10.1080/24754269.2026.2635747

To link to this article: https://doi.org/10.1080/24754269.2026.2635747

Published online: 07 Mar 2026.

STATISTICAL THEORY AND RELATED FIELDS
https://doi.org/10.1080/24754269.2026.2635747

Two-stage least squares model averaging for instrumental variable models with exogenous variables

Wenjun Shen[a] and Xiaochao Xia [a,b]

[a]College of Mathematics and Statistics, Chongqing University, Chongqing, People's Republic of China; [b]Key Laboratory of Nonlinear Analysis and Its Applications (Ministry of Education), Chongqing University, Chongqing, People's Republic of China

ABSTRACT

Instrumental variable (IV) methods are widely used to address unmeasured confoundings in structural equation models. In this paper, we focus on the settings where a possibly large number of instruments and a weak correlation between the instruments and the endogenous variable exist. Specifically, we propose a novel two-stage least squares (2SLS) model averaging approach to estimate the coefficient of an endogenous variable. Differing from existing literature, our model averaging estimation allows multiple exogenous variables to be included in both stages simultaneously. Theoretically, we study the consistency and asymptotic distributions of the estimated weights and the proposed model averaging estimator. Importantly, we discover that the proposed model averaging estimator produces an asymptotic bias when the endogenous variable and exogenous variables are correlated. Then, we construct a debiased estimator and establish its consistency and asymptotic normality to make statistical inference. Furthermore, we present an equivalent interpretation of the debiased estimator from another construction. Finally, numerical simulations and a real data analysis are conducted to illustrate our proposal.

ARTICLE HISTORY
Received 21 February 2025
Revised 22 June 2025
Accepted 12 February 2026

KEYWORDS
Exogenous variables; model averaging; instrumental variables

1. Introduction

Structural equation model (SEM) generally adopts instrumental variable (IV) strategy to address the issue of unmeasured confoundings in observational studies (Duncan, 1975; Kline, 1998). A qualified IV has three key features: (i) relevance: the IV must be correlated to the endogenous variable; (ii) exogeneity: the IV is independent of the unmeasured confoundings involved in the error term conditional on the covariates; and (iii) exclusion restriction (ER): the IV has an effect on the response variable only through its effect on the endogenous variable (Angrist et al., 1996). In reality, it is possible that the relevance feature is violated, which is known as the weak instrumental variable problem. The problem often occurs in situations where a large number of instruments are selected within a model, while most are weakly correlated with the endogenous variable. In the sense of being weakly correlated with

CONTACT Xiaochao Xia xxc@cqu.edu.cn College of Mathematics and Statistics, Chongqing University, Chongqing 401331, People's Republic of China; Key Laboratory of Nonlinear Analysis and Its Applications (Ministry of Education), Chongqing University, Chongqing 401331, People's Republic of China

the endogenous variables, the use of an IV estimator can lead to large inconsistencies, even further from the true value than OLS estimators in finite samples (Bound et al., 1995; Nelson & Startz, 1990).

Under the setting where $p \leq n$, the existence of weak instruments causes misleading results in causal inference (Stock et al., 2002). Asymptotic theories and statistical inference are discussed under the assumption of weak IVs (Bekker, 1994; C. Hansen et al., 2008). For high-dimensional data analysis where $p > n$, the regularization-based methods are quite appealing to researchers. For instance, when instruments are high-dimensional, Okui (2011) develops the shrinkage two-stage least squares (2SLS) and shrinkage limited information maximum likelihood (LIML) methods to select a subset of instruments. Belloni et al. (2012) suggest applying lasso to the first stage of the 2SLS method to select optimal instruments. Various penalties are adopted to achieve variable selection (Fan & Zhong, 2018; Kang et al., 2016). It is known that regularized methods select a single optimal model as the final model and the resulting model may have a great change when data are slightly permutated.

Model averaging serves as an important alternative to model selection and has many merits. For example, it can reduce the risk of selecting a single misspecified model by combining multiple working models and it is less sensitive to disturbed datasets. Model averaging has been successfully extended to various models, such as parametric models (Ando & Li, 2017; Feng et al., 2022; B. E. Hansen, 2007; Liu, 2015; Zhang & Liu, 2023; Zhang et al., 2014), nonparametric models (J. Chen et al., 2023; C. Li et al., 2018; Zhu et al., 2023), and semiparametric models (J. Chen et al., 2018; Fang et al., 2022; J. Li et al., 2022, 2018; Zhang & Wang, 2019; Zhu et al., 2019). In recent years, model averaging methods are also applied to causal studies (Canay, 2010; Kuersteiner & Okui, 2010). For instance, Martins and Gabriel (2014) propose a smoothing procedure to apply empirical weights to IV selection. B. E. Hansen (2017) proposes a Stein-like 2SLS estimator, whose asymptotic distribution and asymptotic risk are also discussed. More recently, Seng and Li (2022) introduce the model averaging into SEM to solve endogenous problems, where the optimal weights are determined in the sense of minimizing an empirical least squares function, and the consistency of the weights and the asymptotic normality of the model averaging estimator are established. J. Chen et al. (2023) further extend the method of Seng and Li (2022) to nonparametric instrumental variable models. This paper aims to extend the method of Seng and Li (2022) in the presence of the exogenous variables.

By omitting exogenous variables that are wrongly treated as unmeasured confoundings in the error term, one may get biased estimation for the coefficients of endogenous variables, which might lead to a misleading conclusion. Kok et al. (2021) indicate that in empirical studies of behavioural psychology and sentiment analysis, the exogenous variables have an effect on the endogenous variables, and the bias arises from including irrelevant exogenous variables in latent constructs. They propose sparse extended redundancy analysis (SERA) with exclusive LASSO regularization to address this bias in practical settings. A limitation of their proposed method is that the SERA method they propose is not robust when the model is misspecified. More details about the role that the exogenous variables play in both models of SEM can be found in B. E. Hansen (2022). It is worth noting that Kuersteiner and Okui (2010) have investigated the model averaging for IV models by taking into account exogenous variables; however, they require the summation of weights to one, i.e., $\sum_{i=1}^{M} w_i = 1$, where w_i is the weight assigned to the ith submodel. In this paper, we impose no constraint on the weights and our weights are chosen by minimizing the empirical squared loss function, which

is also different from Kuersteiner and Okui (2010). Moreover, when exogenous variables are included in both stages of the 2SLS method, it is not clear whether the theoretical results of Seng and Li (2022) can still apply.

To track the problem, we propose a two-stage least squares model averaging method for IV models to take into account exogenous variables in both stages. The main contribution of this paper is as follows. Firstly, we theoretically establish the asymptotic properties of the estimated weights and the proposed model averaging estimator, including the consistency and asymptotic distributions. Compared to the results of Seng and Li (2022), we find that the proposed model averaging estimator has a non-ignorable asymptotic bias when the exogenous variables and the endogenous variable are correlated and the sum of weights is not equal to one. Secondly, in order to correct the bias term, we propose a debiased version of the model averaging estimator. The consistency and asymptotic normality of the debiased model averaging estimator are accordingly derived under regularity conditions. Thirdly, we provide an intuitive interpretation of the debiased estimator from another perspective of the construction of the estimator. Last, we construct a consistent estimator of the variance of the error term in the SEM based on the proposed model averaging method. Numerical results further empirically show that the theories of the proposed method are valid.

The rest of the paper is organized as follows. We introduce the model averaging methodology in Section 2 and provide some theoretical results in Section 3. In Section 4, we present a debiased estimator and its another interpretation, and give the estimation of the variance of the error term in the SEM. Section 5 gives extensive simulation results. A real-world data set is analysed in Section 6. Concluding remarks are given in Section 7. All the proofs of results are relegated in the Appendices.

2. Methodology

2.1. Model setup

In this subsection, we consider the standard SEM as follows:

$$Y = \beta_e X_e + \boldsymbol{\beta}_o^\top \boldsymbol{X}_o + \varepsilon, \tag{1}$$

where Y is the response variable, X_e is the endogenous variable whose regression coefficient β_e is of major interest, $\boldsymbol{X}_o$ includes p observed covariates with regression coefficients $\boldsymbol{\beta}_o$, and ε is a disturbance with mean zero and variance σ^2, which may involve unmeasured confounders correlated with X_e. A consistent estimator of β_e can be obtained by traditional 2SLS method under the qualified IV conditions. In this paper, we allows that some of the IVs may be weakly correlated with the endogenous variable X_e. Additionally, we assume that there are some exogenous variables $\boldsymbol{X}_o$ correlated with Y.

Let $\boldsymbol{X}_I = (X_{I1}, \ldots, X_{Iq})^\top$ denote the vector of q observed instrumental variables, independent of ε. Assume that X_e and $\boldsymbol{X}_I$ satisfy the reduced form equation:

$$X_e = \boldsymbol{\alpha}_I^\top \boldsymbol{X}_I + \boldsymbol{\alpha}_o^\top \boldsymbol{X}_o + e, \tag{2}$$

where $\boldsymbol{\alpha}_I$ is the vector of the coefficients for $\boldsymbol{X}_I$, $\boldsymbol{\alpha}_o$ measures the effect of $\boldsymbol{X}_o$ on X_e, and e is an error term with mean zero and finite variance.

Without loss of generality, suppose that all the observable variables $(Y, X_e, \boldsymbol{X}_o, \boldsymbol{X}_I)$ involved are centred and obey Equations (1) and (2). We assume that the data set $\{(Y_i, X_{e,i}, \boldsymbol{X}_{o,i}, \boldsymbol{X}_{I,i})\}_{i=1}^n$ consists of n independent copies of $(Y, X_e, \boldsymbol{X}_o, \boldsymbol{X}_I)$. To ease

our presentation, denote $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^\top \in \mathbb{R}^n$, $\boldsymbol{X}_e = (X_{e,1}, \ldots, X_{e,n})^\top \in \mathbb{R}^n$, $\mathcal{X}_o = (\boldsymbol{X}_{o,1}, \ldots, \boldsymbol{X}_{o,n})^\top \in \mathbb{R}^{n\times p}$, $\mathcal{X}_I = (\boldsymbol{X}_{I,1}, \ldots, \boldsymbol{X}_{I,n})^\top \in \mathbb{R}^{n\times q}$, which is a matrix for instruments, and $\boldsymbol{\beta} = (\beta_e, \boldsymbol{\beta}_o^\top)^\top \in \mathbb{R}^{1+p}$, which is a vector of regression coefficients for $(X_e, \boldsymbol{X}_o)$ in Equation (1). Our main interest lies in estimating the coefficient β_e of the endogenous variable X_e.

2.2. 2SLS estimator

In this section, we briefly review the commonly used two-stage least squares (2SLS) method (e.g., B. E. Hansen, 2022) to estimate β_e in Equation (1). Let $\mathcal{Z} = (\mathcal{X}_{\boldsymbol{I}}, \mathcal{X}_{\boldsymbol{o}})$ be an $n \times (q + p)$ matrix, and denote its projection matrix by $\boldsymbol{P}_Z = \mathcal{Z}(\mathcal{Z}^\top\mathcal{Z})^{-1}\mathcal{Z}^\top$. In the first stage of 2SLS, we perform a least squares regression of $\boldsymbol{X}_e$ on $\mathcal{Z}$ in Equation (2). Thus, a least squares estimator of $\boldsymbol{X}_e$ can be obtained as $\widehat{\boldsymbol{X}}_e = \boldsymbol{P}_Z\boldsymbol{X}_e$. In the second stage of 2SLS, we use the resulting estimator $\widehat{\boldsymbol{X}}_e$ to substitute $\boldsymbol{X}_e$ in Equation (1), and again perform a least squares regression of $\boldsymbol{Y}$ on $\widehat{\boldsymbol{X}}_e$ and $\mathcal{X}_o$ to get an estimator of $\boldsymbol{\beta}$. This is known as the 2SLS method. The resulting 2SLS estimator of $\boldsymbol{\beta}$ can be formulated as $\widehat{\boldsymbol{\beta}}_{2\text{SLS}} = [(\widehat{\boldsymbol{X}}_e, \mathcal{X}_o)^\top(\widehat{\boldsymbol{X}}_e, \mathcal{X}_o)]^{-1}(\widehat{\boldsymbol{X}}_e, \mathcal{X}_o)^\top\boldsymbol{Y} = [\mathcal{X}^\top\boldsymbol{P}_Z\mathcal{X}]^{-1}\mathcal{X}^\top\boldsymbol{P}_Z\boldsymbol{Y}$, where $\mathcal{X} = (\boldsymbol{X}_e, \mathcal{X}_{\boldsymbol{o}})$ is an $n \times (1 + p)$ matrix.

Let $\boldsymbol{P}_1 = \mathcal{X}_o(\mathcal{X}_o^\top\mathcal{X}_o)^{-1}\mathcal{X}_o^\top$ be a projection matrix, and define $\mathcal{Z}_I = (\boldsymbol{I}_n - \boldsymbol{P}_1)\mathcal{X}_I$. Clearly, $\boldsymbol{P}_Z\boldsymbol{P}_1 = \boldsymbol{P}_1$ and $\mathcal{Z}_I$ is $\mathcal{X}_I$ projected orthogonal to $\mathcal{X}_o$. Then, by an application of the Frisch–CWaugh–CLovell (FWL) theorem (B. E. Hansen, 2022, Theorem 3.5), one can obtain the 2SLS estimator of β_e as

$$\widehat{\beta}_{e,2\text{SLS}} = \left[\widehat{\boldsymbol{X}}_e^\top(\boldsymbol{I}_n - \boldsymbol{P}_1)\widehat{\boldsymbol{X}}_e\right]^{-1}\widehat{\boldsymbol{X}}_e^\top(\boldsymbol{I}_n - \boldsymbol{P}_1)\boldsymbol{Y}. \tag{3}$$

Since $\mathcal{X}_o$ and $\mathcal{Z}_I$ are orthogonal, we could get $\boldsymbol{P}_Z = \boldsymbol{P}_1 + \boldsymbol{P}_2$, where $\boldsymbol{P}_2 = \mathcal{Z}_I(\mathcal{Z}_I^\top\mathcal{Z}_I)^{-1}\mathcal{Z}_I^\top$. Therefore, we also have

$$\widehat{\beta}_{e,2\text{SLS}} = (\boldsymbol{X}_e^\top\boldsymbol{P}_2\boldsymbol{X}_e)^{-1}\boldsymbol{X}_e^\top\boldsymbol{P}_2\boldsymbol{Y} = [\boldsymbol{X}_e^\top\mathcal{Z}_I(\mathcal{Z}_I^\top\mathcal{Z}_I)^{-1}\mathcal{Z}_I^\top\boldsymbol{X}_e]^{-1}\boldsymbol{X}_e^\top\mathcal{Z}_I(\mathcal{Z}_I^\top\mathcal{Z}_I)^{-1}\mathcal{Z}_I^\top\boldsymbol{Y}.$$

When the linear reduced form Model (2) is correctly specified, under some regularity conditions, it can be shown that $\widehat{\beta}_{e,2\text{SLS}}$ enjoys the consistency and asymptotic normality (see Lemma A.1 in Appendix 2).

2.3. 2SLS-based model averaging estimator

It is worth noting that in the above 2SLS method, the reduced form Equation (2) should be correctly specified. Otherwise, the 2SLS estimator of β_e may not be consistent. However, this equation is unlikely to hold in reality due to the possible complex nonlinear relationship between X_e and $\boldsymbol{X}_I$. Furthermore, the observed IVs could be incomplete, which also causes Equation (2) inaccurate. Indeed, it is often difficult to specify a correct reduced form equation. When some observed IVs (i.e., $\boldsymbol{X}_I$) are available, we may consider a model averaging method to quantify the uncertainty from $\boldsymbol{X}_I$ through combining a set of approximate working models. In this subsection, we develop a novel 2SLS-based model averaging approach to obtain an estimated value of X_e and an estimator of β_e accordingly. Our idea is as follows. As in B. E. Hansen (2007) and Seng and Li (2022), we firstly consider a least squares model averaging method in the first stage of 2SLS, and then estimate β_e using the weighted averaging predicted value of X_e in the second stage.

Specifically, in the first stage, we construct M candidate working submodels, where each candidate assumes a linear reduced form consisting of a distinct subset of the q IVs. To be more specific, for $m = 1, \ldots, M$, we let $\boldsymbol{X}_I^{(m)} = (X_{I1}^{(m)}, \ldots, X_{It_m}^{(m)})^\top$ and $X_{Ii}^{(m)} \in \{X_{I1}, \ldots, X_{Iq}\}$, where $X_{Ii}^{(m)} \neq X_{Ij}^{(m)}$ with $i \neq j$. Note that $\boldsymbol{X}_I^{(m)}$ involves t_m IVs with $t_m \leq q$. In the mth submodel with $m = 1, \ldots, M$, we use X_e as the response variable, $\boldsymbol{X}_I^{(m)}$ and $\boldsymbol{X}_o$ as the explanatory variables to fit a linear regression model. In other words, the mth submodel $\mathcal{M}_m$ has the following form

$$X_e = (\boldsymbol{\alpha}_I^{(m)})^\top \boldsymbol{X}_I^{(m)} + (\boldsymbol{\alpha}_o^{(m)})^\top \boldsymbol{X}_o + e^{(m)}, \tag{4}$$

where $\boldsymbol{\alpha}_I^{(m)} = (\alpha_{I1}^{(m)}, \ldots, \alpha_{It_m}^{(m)})^\top$ is a vector of coefficients of $\boldsymbol{X}_I^{(m)}$, $\boldsymbol{\alpha}_o^{(m)} = (\alpha_{o1}^{(m)}, \ldots, \alpha_{op}^{(m)})^\top$ is a vector of coefficients of $\boldsymbol{X}_o$ and $e^{(m)}$ corresponds to the approximation error term.

Write $\boldsymbol{Z}^{(m)} = ((\boldsymbol{X}_I^{(m)})^\top, \boldsymbol{X}_o^\top)^\top$, a $(t_m + p)$-vector. Thus, the reduced form (4) can be written as $X_e = (\boldsymbol{\eta}^{(m)})^\top \boldsymbol{Z}^{(m)} + e^{(m)}$, where the parameter vector $\boldsymbol{\eta}^{(m)} = ((\boldsymbol{\alpha}_I^{(m)})^\top, (\boldsymbol{\alpha}_o^{(m)})^\top)^\top$ is defined as the minimizer of $h(\widetilde{\boldsymbol{\eta}}) = E(X_e - \widetilde{\boldsymbol{\eta}}^\top \boldsymbol{Z}^{(m)})^2$, where $\widetilde{\boldsymbol{\eta}}$ is a $(t_m + p)$-vector. For $m, m' = 1, \ldots, M$, we denote

$$\boldsymbol{\Sigma}_{mm'} = E(\boldsymbol{Z}^{(m)}(\boldsymbol{Z}^{(m')})^\top), \boldsymbol{\nu}_m = E\left(\boldsymbol{Z}^{(m)} X_e\right), \boldsymbol{d}_m = E\left(\boldsymbol{Z}^{(m)} \boldsymbol{X}_o^\top\right),$$

where $\boldsymbol{\Sigma}_{mm'} \in \mathbb{R}^{(t_m+p)\times(t_{m'}+p)}$, $\boldsymbol{\nu}_m \in \mathbb{R}^{(t_m+p)}$ and $\boldsymbol{d}_m \in \mathbb{R}^{(t_m+p)\times p}$. Then, $\boldsymbol{\eta}^{(m)} = \boldsymbol{\Sigma}_{mm}^{-1} \boldsymbol{\nu}_m$. As a result, the least squares estimator of $\boldsymbol{\eta}^{(m)}$ is given by $\widehat{\boldsymbol{\eta}}^{(m)} = ((\mathcal{Z}^{(m)})^\top \mathcal{Z}^{(m)})^{-1} (\mathcal{Z}^{(m)})^\top) \boldsymbol{X}_e$, where $\mathcal{Z}^{(m)} = (\boldsymbol{Z}_1^{(m)}, \ldots, \boldsymbol{Z}_n^{(m)})^\top \in \mathbb{R}^{n\times(t_m+p)}$. Accordingly, the predicted value of X_e in the mth submodel is obtained as $\widehat{X}_{em} = (\widehat{\boldsymbol{\eta}}^{(m)})^\top \boldsymbol{Z}^{(m)}$ at new covariates $\boldsymbol{Z}^{(m)}$. In matrix notation, at all observations of $\boldsymbol{Z}^{(m)}$, we can use the mth fitted submodel to obtain the fitted values of X_e as

$$\widehat{\boldsymbol{X}}_{em} \triangleq (\widehat{X}_{em,1}, \ldots, \widehat{X}_{em,n})^\top = \mathcal{Z}^{(m)}((\mathcal{Z}^{(m)})^\top \mathcal{Z}^{(m)})^{-1} (\mathcal{Z}^{(m)})^\top \boldsymbol{X}_e, \quad m = 1, \ldots, M.$$

To combine the results from all candidate models, we may consider the weighted average of $\{\widehat{X}_{em}, m = 1, \ldots, M\}$ as the final estimator of X_e, that is, $\widehat{X}_e(\boldsymbol{w}) = \sum_{m=1}^M w_m \widehat{X}_{em}$, where w_m is the weight assigned to the mth submodel. In order to determine the optimal weights for a model averaging estimator of X_e, we consider $f(\boldsymbol{w}) = E(X_e - \sum_{m=1}^M w_m (\boldsymbol{\eta}^{(m)})^\top \boldsymbol{Z}^{(m)})^2$ as the risk function, where $\boldsymbol{w} = (w_1, \ldots, w_M)^\top$ is the weight vector. We define the optimal weight vector as $\boldsymbol{w}_0 = (w_{01}, \ldots, w_{0M})^\top = \arg\min_{\boldsymbol{w}} f(\boldsymbol{w})$. It can be shown that

$$\boldsymbol{w}_0 = \boldsymbol{\Psi}^{-1} \boldsymbol{u},$$

where $\boldsymbol{u} = (\boldsymbol{\nu}_m^\top \boldsymbol{\Sigma}_{mm}^{-1} \boldsymbol{\nu}_m) \in \mathbb{R}^{M\times 1}$ and $\boldsymbol{\Psi} = (\boldsymbol{\nu}_m^\top \boldsymbol{\Sigma}_{mm}^{-1} \boldsymbol{\Sigma}_{mm'} \boldsymbol{\Sigma}_{m'm'}^{-1} \boldsymbol{\nu}_{m'}) \in \mathbb{R}^{M\times M}$. For the M prepared submodels, a natural idea is to optimize $\boldsymbol{w}_0$ by minimizing the empirical squared loss function $Q(\boldsymbol{w}) = \|\boldsymbol{X}_e - \sum_{m=1}^M w_m \widehat{\boldsymbol{X}}_{em}\|^2$. It seems that solving this optimization may cause overfitting due to the repeated use of samples. Ideally, if another independent and identically distributed data set is available, we could estimate $\widehat{\boldsymbol{X}}_{em}$ based on this independent data set to avoid overfitting. However, we usually do not have an additional data set, independent of the original data, in real-world problems. To handle this issue, we do not minimize $Q(\boldsymbol{w})$ to obtain $\hat{\boldsymbol{w}}$. Instead, from the relationship $\boldsymbol{w}_0 = \boldsymbol{\Psi}^{-1} \boldsymbol{u}$, we can obtain a plug-in estimator of

$\boldsymbol{w}_0$ using the plug-in estimators of $\boldsymbol{u}$ and $\boldsymbol{\Psi}$ as

$$\hat{\boldsymbol{w}} = \widehat{\boldsymbol{\Psi}}^{-1}\hat{\boldsymbol{u}} = (\widetilde{\mathcal{X}}_e^\top \boldsymbol{H}\boldsymbol{H}^\top \widetilde{\mathcal{X}}_e)^{-1}\widetilde{\mathcal{X}}_e^\top \boldsymbol{H}\boldsymbol{X}_e, \tag{5}$$

where $\widetilde{\mathcal{X}}_e = \boldsymbol{X}_e \otimes \boldsymbol{I}_M$ is an $Mn \times M$ matrix, $\boldsymbol{I}_M$ is an $M \times M$ identity matrix, $\boldsymbol{H} = (\boldsymbol{H}_1, \ldots, \boldsymbol{H}_M)^\top$ is an $Mn \times n$ matrix, and $\boldsymbol{H}_m = \mathcal{Z}^{(m)}((\mathcal{Z}^{(m)})^\top \mathcal{Z}^{(m)})^{-1}(\mathcal{Z}^{(m)})^\top$ is an $n \times n$ idempotent matrix. We will establish the consistency and asymptotic normality of $\hat{\boldsymbol{w}}$ in Theorem 3.1. With the optimal weights $\boldsymbol{w}_0$ and estimated optimal weights $\hat{\boldsymbol{w}}$ in Equation (5), we can obtain the model averaging estimator for $\boldsymbol{X}_e$, respectively, as $\widehat{\boldsymbol{X}}_e(\boldsymbol{w}_0) = \boldsymbol{H}^\top \widetilde{\mathcal{X}}_e \boldsymbol{\Psi}^{-1}\boldsymbol{u}$ and $\widehat{\boldsymbol{X}}_e(\hat{\boldsymbol{w}}) = \boldsymbol{H}^\top \widetilde{\mathcal{X}}_e(\widetilde{\mathcal{X}}_e^\top \boldsymbol{H}\boldsymbol{H}^\top \widetilde{\mathcal{X}}_e)^{-1}\widetilde{\mathcal{X}}_e^\top \boldsymbol{H}\boldsymbol{X}_e$.

Next, in the second stage of 2SLS, we first replace $\boldsymbol{X}_e$ in the SEM (1) with $\widehat{\boldsymbol{X}}_e(\hat{\boldsymbol{w}})$, and then perform a least squares regression of $\boldsymbol{Y}$ on the model averaging estimator $\widehat{\boldsymbol{X}}_e(\hat{\boldsymbol{w}})$ and $\boldsymbol{X}_o$. Hence, we can get a model averaging estimator of β_e, denoted as $\widehat{\beta}_{e,\mathrm{MA}}(\hat{\boldsymbol{w}})$. By a simple calculation, we have

$$\begin{aligned}
\widehat{\beta}_{e,\mathrm{MA}}(\hat{\boldsymbol{w}}) &= \left[\widehat{\boldsymbol{X}}_e(\hat{\boldsymbol{w}})^\top (\boldsymbol{I}_n - \boldsymbol{P}_1)\widehat{\boldsymbol{X}}_e(\hat{\boldsymbol{w}})\right]^{-1}\left[\widehat{\boldsymbol{X}}_e(\hat{\boldsymbol{w}})^\top (\boldsymbol{I}_n - \boldsymbol{P}_1)\boldsymbol{Y}\right] \\
&= \left[\boldsymbol{X}_e^\top \boldsymbol{H}^\top \widetilde{\mathcal{X}}_e \left(\widetilde{\mathcal{X}}_e^\top \boldsymbol{H}\boldsymbol{H}^\top \widetilde{\mathcal{X}}_e\right)^{-1} \widetilde{\mathcal{X}}_e^\top \boldsymbol{H}(\boldsymbol{I}_n - \boldsymbol{P}_1)\boldsymbol{H}^\top \widetilde{\mathcal{X}}_e \left(\widetilde{\mathcal{X}}_e^\top \boldsymbol{H}\boldsymbol{H}^\top \widetilde{\mathcal{X}}_e\right)^{-1} \widetilde{\mathcal{X}}_e^\top \boldsymbol{H}\boldsymbol{X}_e\right]^{-1} \\
&\quad \times \boldsymbol{X}_e^\top \boldsymbol{H}^\top \widetilde{\mathcal{X}}_e \left(\widetilde{\mathcal{X}}_e^\top \boldsymbol{H}\boldsymbol{H}^\top \widetilde{\mathcal{X}}_e\right)^{-1} \widetilde{\mathcal{X}}_e^\top \boldsymbol{H}(\boldsymbol{I}_n - \boldsymbol{P}_1)\boldsymbol{Y},
\end{aligned}$$

where $\boldsymbol{P}_1 = \mathcal{X}_o(\mathcal{X}_o^\top \mathcal{X}_o)^{-1}\mathcal{X}_o^\top$ is a projection matrix. We will investigate the asymptotic properties for $\widehat{\beta}_{e,\mathrm{MA}}(\hat{\boldsymbol{w}})$ in the next section.

3. Theoretical properties

3.1. Basic assumptions

In order to derive the theoretical properties of the estimated weight vector $\hat{\boldsymbol{w}}$ and the model averaging estimator $\widehat{\beta}_{e,\mathrm{MA}}(\hat{\boldsymbol{w}})$, we need the conditions listed below. We use $\xrightarrow{p}$ to denote convergence in probability and $\xrightarrow{d}$ to denote convergence in law.

(i) $\boldsymbol{X}_I$ and $\boldsymbol{X}_o$ are random vectors with zero mean.
(ii) $\boldsymbol{\alpha}_I \neq 0$, which requires $\boldsymbol{X}_I$ to be correlated with X_e conditional on e.
(iii) The error term ε has mean zero and variance σ^2 and is uncorrelated with $\boldsymbol{X}_I$ and $\boldsymbol{X}_o$. The error term e has mean zero.
(iv) For each $i = 1, \ldots, n$, $E\|X_{e,i}\boldsymbol{X}_{o,i}\| < \infty$, $E\|\boldsymbol{X}_{o,i}\boldsymbol{X}_{o,i}^\top\| < \infty$, $E\|X_{e,i}\boldsymbol{X}_{I,i}\| < \infty$, $E\|\boldsymbol{X}_{I,i}\boldsymbol{X}_{I,i}^\top\| < \infty$ and $E\|\boldsymbol{X}_{I,i}\boldsymbol{X}_{o,i}^\top\| < \infty$. Denote $\boldsymbol{M}_{oo} = E(\boldsymbol{X}_o\boldsymbol{X}_o^\top)$ and $\boldsymbol{M}_{II} = E(\boldsymbol{X}_I\boldsymbol{X}_I^\top)$. Assume that the rank of $\boldsymbol{M}_{oo}$ is p and the rank of $\boldsymbol{M}_{II}$ is q.
(iv') For each $i = 1, \ldots, n$, $E\|X_{e,i}\boldsymbol{X}_{o,i}\| < \infty$, $E\|\boldsymbol{X}_{o,i}\boldsymbol{X}_{o,i}^\top\| < \infty$, $E\|X_{e,i}\boldsymbol{X}_{I,i}^{(m)}\| < \infty$, $E\|\boldsymbol{X}_{I,i}^{(m)}(\boldsymbol{X}_{I,i}^{(m)})^\top\| < \infty$ and $E\|\boldsymbol{X}_{I,i}^{(m)}\boldsymbol{X}_{o,i}^\top\| < \infty$.
(v) For each $i = 1, \ldots, n$, $E\|\boldsymbol{\varphi}_i\|^2 < \infty$, $E|\tilde{\varphi}_i|^2 < \infty$ and $E\|\boldsymbol{\phi}_i\|^2 < \infty$ where $\boldsymbol{\varphi}_i$, $\tilde{\varphi}_i$ and $\boldsymbol{\phi}_i$ are denoted in Appendix 1.

Assumptions (i)–(iii) are standard assumptions, which are mild. Assumption (i) holds by centralizing variables. Assumptions (ii) and (iii) are requirements on IVs. Assumptions (iv)

and (v) are indeed some moments assumptions on IVs, endogenous variable, and exogenous variables. Assumption (iv) is used for the derivations of the consistency of the 2SLS estimator in Section 2.2. Similar to Assumption (iv), Assumption (iv') is imposed for the proposed model averaging estimator. Assumption (v) is assumed to derive the limiting distribution of the estimated optimal weight vector and the proposed model averaging estimator in Section 2.3. Under Assumption (iv'), we have $n^{-1}\boldsymbol{X}_e^\top \mathcal{Z}^{(m)} \xrightarrow{p} \boldsymbol{v}_m^\top$, $n^{-1}(\mathcal{Z}^{(m)})^\top \mathcal{Z}^{(m')} \xrightarrow{p} \boldsymbol{\Sigma}_{mm'}$ and $n^{-1}(\mathcal{Z}^{(m)})^\top \mathcal{X}_o \xrightarrow{p} \boldsymbol{d}_m$ by the law of large numbers. Similar assumptions are also made in Corbae et al. (2006), Hong (2020) and Seng and Li (2022).

We verify the standard validity conditions for instrumental variables through Assumptions (ii), (iii), and (iv). In particular, Assumption (ii) ensures the instrument relevance condition ($\boldsymbol{\alpha}_I \neq 0$), while Assumption (iii) guarantees instrument exogeneity. Moment conditions in (iv) and (iv') further ensure finite variance and identification. The assumed correlations between the endogenous and exogenous variables are addressed in Assumptions (iv) and (iv'), which allows for general correlations between endogenous and exogenous variables through finite moment conditions, note that it could be zero as discussed in Remark 4.1.

3.2. Asymptotic properties

With the previous assumption conditions, we can provide some theoretical results to justify the 2SLS estimator in Section 2.2 and the model averaging estimator in Section 2.3, respectively. The consistency and asymptotic normality of the 2SLS estimator and the derivations are in the Appendices.

The second result is about the large sample property of the estimator of the weight vector involved in the model averaging estimator.

Theorem 3.1: *Under Assumptions* (i)–(iii), (iv') *and* (v), *as* $n \to \infty$, *we have* (i) *(Consistency)* $\hat{\boldsymbol{w}} \xrightarrow{p} \boldsymbol{w}_0$, *and* (ii) *(Asymptotic Normality)* $\sqrt{n}(\hat{\boldsymbol{w}} - \boldsymbol{w}_0) \xrightarrow{d} N(0, \boldsymbol{\Xi})$, *where* $\boldsymbol{\Xi} = \mathrm{cov}(\boldsymbol{\varphi}_i, \boldsymbol{\varphi}_i)$ *and* $\boldsymbol{\varphi}_i = \boldsymbol{\Psi}^{-1}(\boldsymbol{\xi}_i - \boldsymbol{\gamma}_i \boldsymbol{w}_0)$ *are defined in Appendix* 1.

Remark 3.1: From Theorem 3.1, we know that $\hat{\boldsymbol{w}}$ is consistent with the true weight vector $\boldsymbol{w}_0$, and has a normal limiting distribution with mean zero and variance $\boldsymbol{\Xi}$. With this result, we can make statistical inferences on $\boldsymbol{w}_0$ such as constructing a 95%-level confidence interval for $\boldsymbol{w}_0$. Note that this result is similar to those in the literature of the weights with no constraint imposed (J. Chen et al., 2018; Seng & Li, 2022), but is distinct from the results with the weights constrained in a simplex space (Fang et al., 2022; Zhang & Zhang, 2023).

The third result is associated with the asymptotic theory of the resulting model averaging estimator, $\widehat{\beta}_{e,\mathrm{MA}}(\hat{\boldsymbol{w}})$, stated in Section 2.3.

Theorem 3.2: *Under Assumptions* (i)–(iii), (iv') *and* (v), *as* $n \to \infty$, *we have* (i)

$$\widehat{\beta}_{e,\mathrm{MA}}(\hat{\boldsymbol{w}}) = \beta_e + \mathbf{bias} + o_p(1), \tag{6}$$

where $\mathbf{bias} = \frac{a\boldsymbol{w}_0^\top \mathbf{1}_M(\boldsymbol{w}_0^\top \mathbf{1}_M - 1)}{\boldsymbol{w}_0^\top \boldsymbol{\Psi} \boldsymbol{w}_0 - a(\boldsymbol{w}_0^\top \mathbf{1}_M)^2}\beta_e$, $a = \boldsymbol{M}_{oe}^\top \boldsymbol{M}_{oo}^{-1} \boldsymbol{M}_{oe} = E(X_e \boldsymbol{X}_o^\top)[E(\boldsymbol{X}_o \boldsymbol{X}_o^\top)]^{-1} E(\boldsymbol{X}_o X_e)$, *and* (ii)

$$\sqrt{n}(\widehat{\beta}_{e,\mathrm{MA}}(\hat{\boldsymbol{w}}) - \beta_e - \mathbf{bias}) \xrightarrow{d} N\left(0, \mathrm{cov}(\tilde{\varphi}_i, \tilde{\varphi}_i) s_3^{-2}\right), \tag{7}$$

where $\tilde{\varphi}_i$ *and* s_3 *are given in Appendix* 1.

Remark 3.2: In Theorem 3.2, Part (i) reveals that the proposed model averaging estimator of β_e produces a bias, which can not be negligible in the asymptotic sense if **bias** $\neq 0$. However, this bias term can vanish in two special cases as follows. (a) In the case where the exogenous variables $\boldsymbol{X}_o$ and the endogenous variable X_e are uncorrelated, that is, $E(\boldsymbol{X}_{o,i}X_{e,i}) = 0$ (i.e., $\boldsymbol{M}_{oe} = 0$), we have $a = 0$ and, accordingly, **bias** $= 0$ and $\zeta_i = 0$, which implies $\tilde{\varphi}_i = \boldsymbol{w}_0^\top \boldsymbol{\phi}_i$. Hence, in this case, it can be derived that $\text{cov}(\tilde{\varphi}_i, \tilde{\varphi}_i)s_3^{-2} = \boldsymbol{w}_0^\top \boldsymbol{\Psi} \boldsymbol{w}_0 \sigma^2 (\boldsymbol{w}_0^\top \boldsymbol{\Psi} \boldsymbol{w}_0)^{-2} = (\boldsymbol{u}^\top \boldsymbol{\Psi}^{-1} \boldsymbol{u})^{-1} \sigma^2$. Thus, from Part (ii), we can obtain $\sqrt{n}(\widehat{\beta}_{e,\text{MA}}(\hat{\boldsymbol{w}}) - \beta_e) \xrightarrow{d} N(0, (\boldsymbol{u}^\top \boldsymbol{\Psi}^{-1} \boldsymbol{u})^{-1} \sigma^2)$ when $E(\boldsymbol{X}_{o,i}X_{e,i}) = 0$. This result coincides with the result obtained in Seng and Li (2022). (b) It is interesting to see that in the case where $\boldsymbol{w}_0^\top \mathbf{1}_M = 1$, we can also get **bias** $= 0$, indicating the asymptotic bias of our proposed estimator is eliminated when the optimal weights sum to one. In another word, under the constraint of the weights summation to one, the proposed model averaging estimator $\widehat{\beta}_{e,\text{MA}}(\hat{\boldsymbol{w}})$ converges in probability to the true value β_e. Meanwhile, in this case, its asymptotic variance can be simplified as $\text{cov}(\tilde{\varphi}_i, \tilde{\varphi}_i)s_3^{-2}$ with $s_3 = \boldsymbol{w}_0^\top \boldsymbol{\Psi} \boldsymbol{w}_0 - a$ and $\tilde{\varphi}_i = a\beta_e \mathbf{1}_M^\top \boldsymbol{\Psi}^{-1} \boldsymbol{\xi}_i - a\beta_e \mathbf{1}_M^\top \boldsymbol{\Psi}^{-1} \boldsymbol{\gamma}_i \boldsymbol{w}_0 + \boldsymbol{w}_0^\top \boldsymbol{\phi}_i$. Moreover, we notice that if neither Case (a) nor Case (b) occurs, the bias term generally can not be cancelled. Hence, if using Part (ii) of Theorem 3.2 to further make statistical inference on β_e, we have to estimate the bias term. However, because the bias contains the unknown coefficient β_e, it is often difficult to estimate the bias directly. In the next section, we introduce an approach to adjust the bias term.

4. Asymptotic consistent estimation

4.1. Debiased estimator

As indicated in Theorem 3.2 in the previous section, the proposed estimator $\widehat{\beta}_{e,\text{MA}}(\hat{\boldsymbol{w}})$ generally yields a non-ignorable bias, **bias** $= c_{\text{bias}}\beta_e$, where $c_{\text{bias}} = \frac{a\boldsymbol{w}_0^\top \mathbf{1}_M (\boldsymbol{w}_0^\top \mathbf{1}_M - 1)}{\boldsymbol{w}_0^\top \boldsymbol{\Psi} \boldsymbol{w}_0 - a(\boldsymbol{w}_0^\top \mathbf{1}_M)^2}$ is the coefficient of β_e in the *bias* term. Since the **bias** term involves the unknown parameter β_e, we cannot directly construct an estimator of **bias** although the sample estimator of c_{bias} can be easily obtained as $\hat{c}_{\text{bias}} = \frac{\hat{a}\hat{\boldsymbol{w}}^\top \mathbf{1}_M (\hat{\boldsymbol{w}}^\top \mathbf{1}_M - 1)}{\hat{\boldsymbol{w}}^\top \hat{\boldsymbol{\Psi}} \hat{\boldsymbol{w}} - \hat{a}(\hat{\boldsymbol{w}}^\top \mathbf{1}_M)^2}$, where $\hat{a}$ is a plug-in estimator using sample moments to replace the expectations in the expression. If one substitutes the biased estimator $\widehat{\beta}_{e,\text{MA}}(\hat{\boldsymbol{w}})$ into **bias**, the resulting estimator of **bias** still inherits a non-ignorable bias.

Fortunately, according to Theorem 3.2, we know that $\widehat{\beta}_{e,\text{MA}}(\hat{\boldsymbol{w}}) = (1 + c_{\text{bias}})\beta_e + o_p(1)$. This relationship indicates that we can estimate β_e immediately based on the estimator $\widehat{\beta}_{e,\text{MA}}(\hat{\boldsymbol{w}})$ and the estimator of c_{bias}. Thus, we propose the following debiased estimator of β_e,

$$\widehat{\beta}_e^{(\text{debias})}(\hat{\boldsymbol{w}}) = (1 + \hat{c}_{\text{bias}})^{-1} \hat{\beta}_{e,\text{MA}}(\hat{\boldsymbol{w}}), \tag{8}$$

where $\hat{\boldsymbol{w}}$ is given in Equation (5). The consistency and asymptotic normality of this debiased estimator $\widehat{\beta}_e^{(\text{debias})}(\hat{\boldsymbol{w}})$ are established below.

Theorem 4.1: *Under Assumptions* (i)–(iii), (iv') *and* (v), *as* $n \to \infty$, *we have* (i) *(Consistency)* $\widehat{\beta}_e^{(\text{debias})}(\hat{\boldsymbol{w}}) = \beta_e + o_p(1)$, *and* (ii) *(Asymptotic Normality)* $\sqrt{n}(\hat{\beta}_e^{(\text{debias})}(\hat{\boldsymbol{w}}) - \beta_e) \xrightarrow{d} N(0, \nu)$, *where* $\nu = \boldsymbol{w}_0^\top (\boldsymbol{\Psi} - \boldsymbol{d}\mathbf{1}_M^\top - \mathbf{1}_M \boldsymbol{d}^\top + a\mathbf{1}_M\mathbf{1}_M^\top)\boldsymbol{w}_0 \sigma^2 / (\boldsymbol{w}_0^\top \boldsymbol{\Psi} \boldsymbol{w}_0 - a\boldsymbol{w}_0^\top \mathbf{1}_M)^2$, *in which the definition of* $\boldsymbol{d}$ *is given in Appendix* 1.

Remark 4.1: In the special case where $E(\boldsymbol{X}_{o,i}X_{e,i}) = 0$ (i.e., $\boldsymbol{M}_{oe} = 0$), we have $a = 0$ and $\boldsymbol{d} = 0$. This results in the asymptotic variance $\nu = (\boldsymbol{w}_0^\top \boldsymbol{\Psi} \boldsymbol{w}_0)^{-1}\sigma^2 = (\boldsymbol{u}^\top \boldsymbol{\Psi}^{-1}\boldsymbol{u})^{-1}\sigma^2$. Thus, when $E(\boldsymbol{X}_{o,i}X_{e,i}) = 0$, the above result is the same as that in Seng and Li (2022).

Remark 4.2: This theorem reveals that the debiased estimator is consistent. Thus, we use $\widehat{\beta}_e^{(\text{debias})}(\hat{\boldsymbol{w}})$ as the final estimator of β_e, which is illustrated in both our simulation and real data analysis. According to the asymptotic normality in this theorem, one can construct the confidence interval of β_e. To this end, we only need to consistently estimate the asymptotic variance ν. This can be achieved by substituting $\boldsymbol{u}$, $\boldsymbol{\Psi}$, a and $\boldsymbol{d}$ with their moment estimators, and σ^2 with $\hat{\sigma}^2_{\text{MA}}$ illustrated later in Remark 4.4 in Section 4.2. Concretely, the estimator of ν is formulated as $\hat{\nu} = \hat{\boldsymbol{w}}^\top(\widehat{\boldsymbol{\Psi}} - \hat{\boldsymbol{d}}\mathbf{1}_M^\top - \mathbf{1}_M\hat{\boldsymbol{d}}^\top + \hat{a}\mathbf{1}_M\mathbf{1}_M^\top)\hat{\boldsymbol{w}}\hat{\sigma}^2_{MA}/[\hat{\boldsymbol{w}}^\top(\widehat{\boldsymbol{\Psi}}\hat{\boldsymbol{w}} - \hat{a}\mathbf{1}_M)]^2$. The 95% confidence interval of β_e can be constructed as $[\widehat{\beta}_e^{(\text{debias})}(\hat{\boldsymbol{w}}) - 1.96\sqrt{\hat{\nu}/n}, \widehat{\beta}_e^{(\text{debias})}(\hat{\boldsymbol{w}}) + 1.96\sqrt{\hat{\nu}/n}]$.

4.2. Another interpretation

In this subsection, we present an alternative construction procedure that provides a more intuitive interpretation of the debiased model averaging estimator $\widehat{\beta}_e^{(\text{debias})}(\hat{\boldsymbol{w}})$ proposed in Section 4.1. To proceed, in terms of Equation (3) in Section 2.2 and using the facts that $\boldsymbol{P}_1\boldsymbol{P}_Z = \boldsymbol{P}_1$ and $\boldsymbol{P}_Z\boldsymbol{P}_1 = \boldsymbol{P}_1$, we obtain

$$\widehat{\beta}_{e,\text{2SLS}} = [\widehat{\boldsymbol{X}}_e^\top(\boldsymbol{I}_n - \boldsymbol{P}_1)\boldsymbol{X}_e]^{-1}\widehat{\boldsymbol{X}}_e^\top(\boldsymbol{I}_n - \boldsymbol{P}_1)\boldsymbol{Y}. \tag{9}$$

On the basis of Equation (9), a natural idea is to utilize some model averaging estimator of $\boldsymbol{X}_e$ as a surrogate of $\widehat{\boldsymbol{X}}_e$ in order to form a model averaging estimator of β_e. For this purpose, following the previous arguments in Section 2.3, one can use the model averaging estimator $\widehat{\boldsymbol{X}}_e(\hat{\boldsymbol{w}})$ of $\boldsymbol{X}_e$ to replace the term $\widehat{\boldsymbol{X}}_e$ in Equation (9). As a result, another model averaging estimator of β_e can be formulated as

$$\widetilde{\beta}_{e,\text{MA}}(\hat{\boldsymbol{w}}) = [\widehat{\boldsymbol{X}}_e(\hat{\boldsymbol{w}})^\top(\boldsymbol{I}_n - \boldsymbol{P}_1)\boldsymbol{X}_e]^{-1}\widehat{\boldsymbol{X}}_e(\hat{\boldsymbol{w}})^\top(\boldsymbol{I}_n - \boldsymbol{P}_1)\boldsymbol{Y}, \tag{10}$$

where the estimated weight vector $\hat{\boldsymbol{w}}$ is given in (5).

The following theorem gives the equivalence relationship between $\widetilde{\beta}_{e,\text{MA}}(\hat{\boldsymbol{w}})$ in (10) and $\widehat{\beta}_e^{(\text{debias})}(\hat{\boldsymbol{w}})$ in (8).

Theorem 4.2: *It follows that* $\widetilde{\beta}_{e,\text{MA}}(\hat{\boldsymbol{w}}) = \widehat{\beta}_e^{(\text{debias})}(\hat{\boldsymbol{w}})$.

Remark 4.3: This theorem indicates that the two model averaging estimators, $\widetilde{\beta}_{e,\text{MA}}(\hat{\boldsymbol{w}})$ and $\widehat{\beta}_e^{(\text{debias})}(\hat{\boldsymbol{w}})$, are exactly in the same form, but they are constructed in two different ways. This result together with Theorem 4.1 implies that $\widetilde{\beta}_{e,\text{MA}}(\hat{\boldsymbol{w}})$ is also a consistent estimator of β_e and has the same asymptotic distribution as that of $\widehat{\beta}_e^{(\text{debias})}(\hat{\boldsymbol{w}})$. That is, under the conditions of Theorem 4.1, $\widetilde{\beta}_{e,\text{MA}}(\hat{\boldsymbol{w}}) = \beta_e + o_p(1)$ and $\sqrt{n}(\widetilde{\beta}_{e,\text{MA}}(\hat{\boldsymbol{w}}) - \beta_e) \xrightarrow{d} N(0, \nu)$.

Remark 4.4: As stated previously, the estimated variance $\hat{\nu}$ involves an estimator of σ^2, $\hat{\sigma}^2_{\text{MA}}$, which is detailed as follows. Similar to the construction of $\widetilde{\beta}_{e,\text{MA}}(\hat{\boldsymbol{w}})$, we can also obtain a consistent model averaging estimator of $\boldsymbol{\beta} = (\beta_e, \boldsymbol{\beta}_o^\top)^\top$. To be specific, denote $\widehat{\mathcal{X}}(\hat{\boldsymbol{w}}) =$

$(\widehat{\boldsymbol{X}}_e(\hat{\boldsymbol{w}}), \mathcal{X}_o)$, an $n \times (1+p)$ matrix. Then, the model averaging estimator of $\boldsymbol{\beta}$ can be constructed as $\widetilde{\boldsymbol{\beta}}_{\mathrm{MA}}(\hat{\boldsymbol{w}}) \triangleq (\widehat{\mathcal{X}}(\hat{\boldsymbol{w}})^\top \mathcal{X})^{-1} \widehat{\mathcal{X}}(\hat{\boldsymbol{w}})^\top \boldsymbol{Y}$. Hence, with this, we can obtain an estimator of σ^2 as $\hat{\sigma}^2_{\mathrm{MA}} = \frac{1}{n-p-1} \|\boldsymbol{Y} - \mathcal{X}\widetilde{\boldsymbol{\beta}}_{\mathrm{MA}}(\hat{\boldsymbol{w}})\|^2$, which is adopted in the subsequent simulations and real data analysis. The following result gives the theoretical justification for $\boldsymbol{\beta}$ and $\hat{\sigma}^2_{\mathrm{MA}}$.

Theorem 4.3: *Suppose that* $E\|\widetilde{\boldsymbol{\phi}}_i\|^2 < \infty$, *where* $\widetilde{\boldsymbol{\phi}}_i$ *is given in Appendix* 1. *Under Assumptions* (i)–(iii) *and* (iv'), *we have*

(i) *(Consistency)* $\widetilde{\boldsymbol{\beta}}_{\mathrm{MA}}(\hat{\boldsymbol{w}}) = \boldsymbol{\beta} + o_p(1)$;

(ii) *(Asymptotic Normality)* $\sqrt{n}[\widetilde{\boldsymbol{\beta}}_{\mathrm{MA}}(\hat{\boldsymbol{w}}) - \boldsymbol{\beta}] \xrightarrow{d} N(0, \boldsymbol{\Lambda}_1^{-1}\boldsymbol{\Lambda}_2(\boldsymbol{\Lambda}_1^\top)^{-1}\sigma^2)$, *where* $\boldsymbol{\Lambda}_1$ *and* $\boldsymbol{\Lambda}_2$ *are given in Appendix* 1;

(iii) *(Consistency of* $\hat{\sigma}^2_{\mathrm{MA}}$*) Furthermore, we have* $\hat{\sigma}^2_{\mathrm{MA}} = \sigma^2 + o_p(1)$.

5. Simulation

5.1. Simulation examples

In this section, we conduct some simulation studies to illustrate the validity and efficiency of our proposed model averaging methods. To be specific, we consider six simulation examples with different settings of parameters (p, q, σ^2, τ^2). In Examples 1–5 below, the response variable Y is generated from Model (1), where $\varepsilon \sim N(0, \sigma^2)$. The endogenous variable X_e is generated based on Model (2), where $e \sim N(0, \tau^2)$. In Example 6, we generate Y from Model (1) except that the component X_{o5} is replaced with $e^{X_{o5}}$. This example allows nonlinear exogenous variables to be contained in the model, and thus the linear IV models used as candidates are all misspecified. In all examples, the components of $\boldsymbol{\alpha}_I$ in Model (2), $\{\alpha_{Ij}\}_{j=1}^q$, are independent and generated from the same uniform distribution, $U(0, b)$, namely, $\alpha_{Ij} \sim_{\mathrm{iid}} U(0, b)$ for $j = 1, \ldots, q$, where the value of b may be distinct in different examples. In the generation of α_{Ij}, we allow the correlation between $\boldsymbol{X}_I$ and X_e to be weak when b is small. For the coefficients of $\boldsymbol{X}_o$ in Model (2), we simulate $\alpha_{oj} \sim_{\mathrm{iid}} U(-c, c)$ for $j = 1, \ldots, p$ with distinct values of c in various examples. The coefficients of $\boldsymbol{X}_o = (X_{o1}, \ldots, X_{op})^\top$ in Model (1), $\{\beta_{oj}\}_{j=1}^p$, are sampled with replacement from the set $\{-5, -4, -3, -2, -1, 1, 2, 3, 4, 5\}$. We generate the instrumental variable $\boldsymbol{X}_I \sim N(0, \boldsymbol{\Omega})$ and the exogenous variable $\boldsymbol{X}_o \sim N(0, \boldsymbol{\Pi})$.

- **Example 1** Set $q = 10$, $p = 5$, $\sigma^2 = 3.25$, $\tau^2 = 5.69$, $\mathrm{cov}(X_e, \varepsilon) = 3$, $\beta_e = -1$, $b = 3.5$, $c = 5$, $\rho_o = 0$ and $\rho_{\mathrm{cs}} = 0, 0.2, 0.5$.
- **Example 2** Same as Example 1 except that $\rho = 0.2$ and $\rho_o = 0.2, 0.5$.
- **Example 3** Set $q = 450$ with 45 coefficients of the IVs being non-zero. $p = 20$, $\sigma^2 = 5$, $\tau^2 = 2$, $\mathrm{cov}(X_e, \varepsilon) = -2$, $\beta_e = 1$, $b = 2.3$, $c = 4$, $\rho_o = 0$ and $\rho_{\mathrm{cs}} = 0$.
- **Example 4** Set $q = 450$ with 45 coefficients of the IVs being non-zero. $p = 20$, $\sigma^2 = 5$, $\tau^2 = 5.14$, $\mathrm{cov}(X_e, \varepsilon) = -3.4$, $\beta_e = 1$, $b = 2.5$, $c = 5$, $\rho_o = 0$ and $\rho_{\mathrm{cs}} = 0.3, 0.5$.
- **Example 5** Same as Example 4 except that $\rho_{\mathrm{cs}} = 0.2$ and $\rho_o = 0.2, 0.5$.
- **Example 6** Same as Example 1 except that two of the instrument variables and one of the exogenous variables are nonlinear. Specifically, we replace X_{I4} with $\sin(X_{I4})$, X_{I5} with X_{I5}^2 and X_{o5} with $e^{X_{o5}}$ in Model (2). Also we replace X_{o5} with $e^{X_{o5}}$ in Model (1).

Let $\boldsymbol{\Omega} = (\rho_{ij})_{i,j=1}^q$. In Examples 1, 3, 4, 5 and 6, we set $\rho_{ij} = 1$ if $i = j$ and $\rho_{ij} = \rho_{\mathrm{cs}}$ if $i \neq j$ for some $-1 < \rho_{cs} < 1$. In Examples 2, we set $\rho_{ij} = 1$ if $i = j$ and $\rho_{ij} = \rho_{ji}$, sampled

uniformly within the interval $(0, \rho)$, where the value of ρ can change in different examples. We use `make.positive.definite()` in R package *corpcor* to ensure the generated matrix $\boldsymbol{\Omega}$ to be a covariance matrix. In Examples 1–6, the diagonal entries of $\boldsymbol{\Pi}$ are ones, and all off-diagonal entries are fixed as the same value ρ_o for $-1 \leq \rho_o \leq 1$.

Throughout, we set $t_m = t$ for convenience, in which way, each submodel under the same parameter setting includes the same number of IVs. We obtain the optimal (t, M) under BIC_M criterion using grid search in a reasonable range. In Examples 1, 2 and 6, we search t and M in range $[2, 10]$. In Examples 3–5, we search t in $[10, 100]$ and M in $[20, 200]$. We find in our coding process that the simulation results are not sensitive to the choice of t and M in these ranges. Thus, we set $(t, M) = (9, 2)$ for Examples 1, 2 and 6, and $(t, M) = (10, 20)$ for Examples 3–5.

5.2. Methods and evaluation criteria

To evaluate the finite-sample performance of our approach, we compare the following six methods.

- (2SLS): The standard 2SLS method using all q IVs in the first stage of 2SLS.
- ($\mathrm{MA}^{(t,M)}$): The model averaging estimator involving M submodels in the first stage of 2SLS, in which the mth submodel includes $\boldsymbol{X}_o$ and different $\boldsymbol{X}_I^{(m)} = (X_{I1}^{(m)}, \ldots, X_{It}^{(m)})^\top \in \mathbb{R}^t$, with each $X_{Ii}^{(m)}$ drawn from $\{X_{I1}, \ldots, X_{Iq}\}$ with equal probability and no replacement. We restrict that $\{X_{I1}^{(m_1)}, \ldots, X_{It}^{(m_1)}\} \neq \{X_{I1}^{(m_2)}, \ldots, X_{It}^{(m_2)}\}$ for $m_1 \neq m_2$ to ensure that the subset of IVs involved in each submodel is not exactly the same.
- ($\mathrm{MA}^{+(t,M)}$): Resembling $\mathrm{MA}^{(t,M)}$ except that each $X_{Ii}^{(m)}$ is drawn from $\{X_{I1}, \ldots, X_{Iq}\}$ with *unequal* probability. The sampling probability of X_{Ii} is computed proportional to the absolute value of the Pearson correlation between X_{Ii} and X_e.
- ($\mathfrak{s}^{(t,M)}$): The model averaging method in Seng and Li (2022) regressing X_e on only the t IVs $\boldsymbol{X}_I^{(m)} = (X_{I1}^{(m)}, \ldots, X_{It}^{(m)})^\top \in \mathbb{R}^t$ in Stage one of every submodel.
- ($\mathfrak{s}^{+(t,M)}$): Resembling $\mathfrak{s}^{(t,M)}$ except that each $X_{Ii}^{(m)}$ is drawn from $\{X_{I1}, \ldots, X_{Iq}\}$ with *unequal* probability. The sampling probability of X_{Ii} is computed similarly to that in $\mathrm{MA}^{+(t,M)}$.
- (pLasso): Applying Lasso to all IVs and the exogenous variables in Stage one as a variable selection method, regressing Y on $\widehat{X}_{e,\mathrm{Lasso}}$ and the exogenous variables selected in the previous stage.
- ($\mathrm{pEL}_{0.5}$): Resembling Lasso except that applying elastic net with mixing parameter $a = 0.5$ to Stage one.
- (Naive): Standard OLS method regressing the response variable Y on the endogenous X_e directly.

Denote the 2SLS estimator of β_e by $\widehat{\beta}_{e,\mathrm{2SLS}}$ and the Lasso estimator of β_e by $\widehat{\beta}_{e,\mathrm{Lasso}}$. Similarly, denote the least squares estimator of $\boldsymbol{X}_e$ in the first stage of 2SLS method by $\widehat{\boldsymbol{X}}_{e,\mathrm{2SLS}}$, the Lasso estimator of $\boldsymbol{X}_e$ in the first stage of 2SLS method by $\widehat{\boldsymbol{X}}_{e,\mathrm{Lasso}}$, and the least squares estimators of $\boldsymbol{\beta}_o$ in the second stage are $\widehat{\boldsymbol{\beta}}_{o,\mathrm{2SLS}}$ and $\hat{\boldsymbol{\beta}}_{o,Lasso}$ respectively. Note that $\mathcal{X}_{o,\mathrm{Lasso}}$ are the exogenous variables selected in the first stage, and p_{Lasso} is the number of the exogenous variables selected. In our simulations, σ^2 is estimated by $\hat{\sigma}^2_{\mathrm{2SLS}} =$

$\frac{1}{n-p-1}\|\boldsymbol{Y}-\widehat{\boldsymbol{X}}_{e,\text{2SLS}}\widehat{\beta}_{e,\text{2SLS}}-\mathcal{X}_o\widehat{\boldsymbol{\beta}}_{o,\text{2SLS}}\|^2$ for 2SLS method, and by $\hat{\sigma}^2_{\text{Lasso}}=\frac{1}{n-p_{\text{Lasso}}-1}\|\boldsymbol{Y}-\widehat{\boldsymbol{X}}_{e,\text{Lasso}}\widehat{\beta}_{e,Lasso}-\mathcal{X}_{o,\text{Lasso}}\widehat{\boldsymbol{\beta}}_{o,\text{Lasso}}\|^2$ for Lasso method.

We utilize the evaluation criteria including bias, standard deviation (SD), standard error (SE), coverage probability (CP) of the β_e estimators and the average execution time of 500 independent simulation processes for the first stage of 2SLS in seconds ($\mathcal{T}$). All results are based on 500 independent simulations.

5.3. Simulation results and analysis

Tables 1–6 present the simulation results of Examples 1–6. Specifically, note that the results of

Table 1. Results for Example 1 (with $\rho_{\text{cs}}=0, 0.2, 0.5$).

ρ_{cs}	n		2SLS	MA$^{(t,M)}$	MA$^{+(t,M)}$	$\mathfrak{s}^{(t,M)}$	$\mathfrak{s}^{+(t,M)}$	pLasso	pEL$_{0.5}$	Naive
0	200	Bias	**0.0033**	**0.0033**	**0.0033**	0.0483	0.0474	0.0034	0.0035	0.0675
		SD	0.0207	0.021	**0.0208**	0.0709	0.071	0.0214	0.0214	0.0266
		SE	0.0197	0.0236	**0.0204**	0.0269	0.0243	0.0204	0.0204	0.0186
		CP	0.938	0.966	**0.944**	0.436	0.412	0.936	0.936	0.074
		$\mathcal{T}$	**0.0034**	0.011	0.0195	0.0104	0.0162	0.1339	0.1222	–
	400	Bias	0.0028	0.0028	0.0028	0.0234	0.0231	0.0027	**0.0026**	0.0692
		SD	0.0153	0.0153	**0.0153**	0.0526	0.0514	0.0154	0.0155	0.0249
		SE	0.0139	0.0165	**0.0142**	0.0176	0.0157	0.0141	0.0141	0.0131
		CP	0.926	**0.956**	0.934	0.472	0.444	0.928	0.928	0.002
		$\mathcal{T}$	**0.0088**	0.0186	0.0365	0.0148	0.0344	0.2011	0.2011	–
	800	Bias	0.001	0.001	**0.0009**	0.0102	0.0101	0.001	**0.0009**	0.0674
		SD	0.0107	**0.0109**	0.0106	0.0391	0.0387	0.0117	0.0118	0.0206
		SE	0.0097	**0.0115**	0.0099	0.012	0.0106	0.01	0.01	0.0092
		CP	0.934	**0.956**	0.936	0.484	0.428	0.932	0.932	0
		$\mathcal{T}$	**0.009**	0.0297	0.0349	0.0178	0.0316	0.2988	0.2933	–
0.2	200	Bias	0.0029	**0.0028**	**0.0028**	0.0202	0.0207	0.0029	0.0031	0.0339
		SD	0.0143	**0.0144**	0.0143	0.0519	0.0514	0.0144	0.0146	0.018
		SE	0.013	**0.0148**	0.0138	0.0173	0.0166	0.0135	0.0135	0.013
		CP	0.91	**0.942**	0.934	0.486	0.478	0.912	0.91	0.306
		$\mathcal{T}$	**0.0073**	0.0236	0.0228	0.022	0.0221	0.285	0.2827	–
	400	Bias	**0.0006**	**0.0006**	**0.0006**	0.003	0.0033	**0.0006**	**0.0006**	0.0334
		SD	0.0072	**0.0072**	0.0072	0.0267	0.0265	0.0072	0.0072	0.0143
		SE	0.0065	**0.0074**	0.0069	0.0078	0.0074	0.0068	0.0068	0.0065
		CP	0.938	0.968	**0.948**	0.474	0.47	0.94	0.94	0.004
		$\mathcal{T}$	**0.0148**	0.0465	0.0428	0.0328	0.0361	0.48	0.4533	–
	800	Bias	**0.0002**	**0.0002**	**0.0002**	0.0036	0.0039	**0.0002**	0.0003	0.0328
		SD	0.0069	0.0069	**0.007**	0.0263	0.0265	0.0072	0.0073	0.0143
		SE	0.0065	0.0074	**0.0069**	0.0078	0.0074	0.0067	0.0067	0.0065
		CP	0.924	**0.954**	0.942	0.484	0.446	0.922	0.922	0.008
		$\mathcal{T}$	**0.0143**	0.0481	0.046	0.0303	0.0352	0.5014	0.5127	–
0.5	200	Bias	**0.0004**	**0.0004**	**0.0004**	0.0085	0.0089	0.0005	**0.0004**	0.0179
		SD	0.0103	**0.0102**	**0.0103**	0.0368	0.0367	0.0106	0.0107	0.0122
		SE	0.0096	**0.0104**	**0.0101**	0.0123	0.0121	0.0102	0.0101	0.0098
		CP	0.928	**0.948**	0.942	0.498	0.496	0.936	0.936	0.562
		$\mathcal{T}$	**0.0077**	0.0235	0.026	0.0197	0.0218	0.2664	0.2827	–
	400	Bias	0.0004	0.0004	0.0004	0.0049	0.0052	**0.0003**	**0.0003**	0.0185
		SD	0.0071	0.0071	**0.0071**	0.0264	0.0264	0.0075	0.0073	0.0105
		SE	0.0068	0.0074	**0.0071**	0.0082	0.008	0.0071	0.0071	0.0069
		CP	**0.944**	0.962	0.958	0.46	0.44	0.94	0.942	0.286
		$\mathcal{T}$	**0.0113**	0.0346	0.0377	0.0284	0.0275	0.3566	0.3493	–
	800	Bias	**0**	**0**	**0**	0.0014	0.0015	**0**	**0**	0.0183
		SD	0.005	0.005	**0.005**	0.0178	0.0178	0.0054	0.0054	0.0087
		SE	0.0048	0.0052	**0.005**	0.0055	0.0053	0.005	0.005	0.0049
		CP	0.938	0.958	**0.95**	0.464	0.446	0.942	0.942	0.092
		$\mathcal{T}$	**0.0204**	0.055	0.046	0.0302	0.0359	0.5333	0.5508	–

Note: True value $\beta_e=-1$.

Table 2. Results for Example 2 (with $\rho_{cs} = 0.2$, $\rho_o = 0.2, 0.5$).

ρ_o	n		2SLS	MA$^{(t,M)}$	MA$^{+(t,M)}$	$\mathfrak{s}^{(t,M)}$	$\mathfrak{s}^{+(t,M)}$	pLasso	pEL$_{0.5}$	Naive
0.2	200	Bias	0.0028	**0.0026**	0.0028	0.0266	0.0271	0.0028	0.0028	0.0423
		SD	**0.0153**	0.0154	**0.0153**	0.0557	0.055	0.0166	0.0166	0.0224
		SE	**0.0149**	0.0171	**0.0157**	0.0197	0.0186	0.0154	0.0153	0.0147
		CP	0.928	0.964	**0.942**	0.468	0.454	0.928	0.928	0.238
		$\mathcal{T}$	**0.0101**	0.0332	0.0326	0.0415	0.0331	0.3851	0.3837	–
	400	Bias	**0.0003**	**0.0003**	0.0004	0.0154	0.0155	0.0004	0.0004	0.0448
		SD	0.0119	0.0121	**0.0119**	0.0422	0.042	0.012	0.012	0.0189
		SE	0.011	0.0116	**0.0116**	0.0137	0.0127	0.0112	0.0112	0.0108
		CP	0.932	0.944	**0.946**	0.48	0.45	0.934	0.932	0.044
		$\mathcal{T}$	**0.0108**	0.048	0.0375	0.0249	0.0326	0.4113	0.3667	–
	800	Bias	**0.0004**	**0.0004**	**0.0004**	0.0093	0.0092	0.0005	0.0005	0.0431
		SD	0.008	**0.0081**	**0.008**	0.0287	0.0287	0.0082	0.0082	0.0161
		SE	0.0075	**0.008**	**0.0079**	0.009	0.0084	0.0077	0.0076	0.0074
		CP	0.942	**0.952**	0.954	0.456	0.424	0.942	0.942	0
		$\mathcal{T}$	**0.0087**	0.0277	0.0237	0.023	0.0226	0.3137	0.3238	–
0.5	200	Bias	0.0025	0.0024	0.0025	0.0256	0.0261	0.0022	**0.0021**	0.0429
		SD	0.0169	**0.017**	**0.0169**	0.0573	0.057	0.0186	0.0184	0.0212
		SE	0.0152	**0.0159**	**0.0158**	0.0203	0.0191	0.0159	0.0157	0.0148
		CP	0.93	0.936	**0.94**	0.508	0.502	0.924	0.926	0.22
		$\mathcal{T}$	**0.0056**	0.0184	0.0227	0.0149	0.0207	0.2157	0.2192	–
	400	Bias	0.0009	0.0008	0.0009	0.0143	0.0146	**0.0007**	0.0009	0.0417
		SD	0.0109	**0.0111**	0.0109	0.0417	0.0412	0.0119	0.0113	0.0176
		SE	0.0106	**0.0112**	0.0111	0.0135	0.0123	0.0111	0.011	0.0104
		CP	**0.95**	0.952	0.956	0.502	0.47	0.946	0.948	0.04
		$\mathcal{T}$	**0.0065**	0.0263	0.0251	0.0132	0.0142	0.2034	0.2123	–
	800	Bias	**0.0002**	0.0003	**0.0002**	0.0062	0.0066	0.0004	0.0003	0.0431
		SD	0.0082	**0.0083**	**0.0082**	0.0291	0.0288	0.0085	0.0085	0.0173
		SE	0.0076	**0.0081**	**0.008**	0.0093	0.0084	0.0078	0.0078	0.0075
		CP	**0.95**	0.956	0.954	0.526	0.45	0.948	0.948	0.002
		$\mathcal{T}$	**0.0113**	0.034	0.0288	0.0339	0.0251	0.3166	0.3284	–

Note: True value $\beta_e = -1$.

our proposed methods reported in all tables are calculated based on the debiased estimator $\widehat{\beta}_e^{(\text{debias})}(\hat{\boldsymbol{w}})$ in Section 4.1. Besides, under high-dimensional settings, as in Tables 3–5, 2SLS methods can not be applied when $n = 200$ and $n = 400$ because the number of IVs and exogenous variables is larger than the sample size in Stage one.

From the tables, we can see that, in terms of bias, the proposed methods MA$^{(t,M)}$ and MA$^{+(t,M)}$, 2SLS and the variable selection methods generally outperform $\mathfrak{s}^{(t,M)}$ and $\mathfrak{s}^{+(t,M)}$ under the settings of either low-dimensional or sparse high-dimensional IVs. This is reasonable since these models include exogenous variables in two stages. Similarly, under low-dimensional conditions, the 2SLS estimator and all variable selection methods also obviously outperform $\mathfrak{s}^{(t,M)}$ and $\mathfrak{s}^{+(t,M)}$, especially when the sample size n is smaller. As n increases, the bias of the naive method hardly changes while the biases of all other methods generally decrease. When q is greater with sparse settings as in Examples 3–5, our methods outperform all other methods. The bias of $\mathfrak{s}^{(t,M)}$ and $\mathfrak{s}^{+(t,M)}$ is even larger than that of the naive OLS method, whose bias is generally the largest under low-dimensional conditions. Another observation from Table 1 is that the biases of all model averaging methods become smaller as the correlation of IVs gets larger, which is common in realistic problems where the IVs are usually correlated.

From the results of Example 6, which are obtained under misspecified models, we can see that the bias of our model averaging methods is slightly larger than that in Example 1 under the same settings but much smaller than $\mathfrak{s}^{(t,M)}$, $\mathfrak{s}^{+(t,M)}$ and the OLS method. It seems that the variable selection methods are comparable to our methods in this case but take much

Table 3. Results for Example 3.

n		2SLS	MA$^{(t,M)}$	MA$^{+(t,M)}$	$\mathfrak{s}^{(t,M)}$	$\mathfrak{s}^{+(t,M)}$	pLasso	pEL$_{0.5}$	Naive
200	Bias	–	0.0148	**0.0118**	0.3478	0.3063	0.0216	0.0221	0.0242
	SD	–	0.0315	**0.0268**	0.0764	0.0671	0.0351	0.0361	0.0187
	SE	–	0.03	**0.0261**	0.0511	0.0419	0.0363	0.0374	0.0183
	CP	–	0.902	**0.932**	0	0	0.878	0.886	0.736
	S($\boldsymbol{X}_I$)	–	165	148.878	165	148.878	145.18	152.37	–
	$\mathcal{T}$	–	0.2419	0.1606	**0.1196**	0.1377	0.8679	0.6941	–
400	Bias	–	0.0066	**0.0059**	0.3466	0.287	0.017	0.0174	0.0241
	SD	–	0.0244	**0.0185**	0.0736	0.0644	0.0176	0.0172	0.0129
	SE	–	0.023	**0.0178**	0.045	0.033	0.0161	0.0159	0.0126
	CP	–	**0.936**	0.92	0	0	0.76	0.746	0.538
	S($\boldsymbol{X}_I$)	–	165	149.37	165	149.37	165.344	177.766	–
	$\mathcal{T}$	–	0.265	0.2003	**0.1349**	0.1596	1.0013	0.8527	–
800	Bias	0.0139	0.0048	**0.002**	0.2882	0.2059	0.0087	0.0093	0.0241
	SD	0.0095	0.0154	**0.0117**	0.0733	0.0529	0.0114	0.0111	0.01
	SE	0.0081	0.0162	**0.0116**	0.0376	0.0236	0.0098	0.0096	0.0088
	CP	0.594	0.944	**0.952**	0	0	0.776	0.764	0.252
	S($\boldsymbol{X}_I$)	450	165	145.098	165	145.098	155.27	169.548	–
	$\mathcal{T}$	0.3745	0.4645	0.344	**0.2124**	0.234	1.9088	1.6307	–

Note: True value $\beta_e = 1$.

Table 4. Results for Example 4 (with $\rho_{cs} = 0.3, 0.5$).

ρ_{cs}	n		2SLS	MA$^{(t,M)}$	MA$^{+(t,M)}$	$\mathfrak{s}^{(t,M)}$	$\mathfrak{s}^{+(t,M)}$	pLasso	pEL$_{0.5}$	Naive
0.3	200	Bias	–	0.0003	**0.0002**	0.0205	0.0219	0.0016	0.0018	0.0031
		SD	–	0.0055	**0.0054**	0.027	0.0267	0.0121	0.0119	0.0053
		SE	–	0.0054	**0.0054**	0.0188	0.0185	0.0118	0.0118	0.0053
		CP	–	**0.95**	0.96	0.702	0.682	0.952	0.952	0.922
		S($\boldsymbol{X}_I$)	–	165	162.304	165	162.304	106.092	109.066	–
		$\mathcal{T}$	–	0.3045	0.3166	**0.2227**	0.2666	1.3217	1.3475	–
	400	Bias	–	**0.0001**	0.0002	0.0116	0.0132	0.001	0.0009	0.0032
		SD	–	**0.0038**	0.0038	0.0205	0.0203	0.0065	0.0065	0.0037
		SE	–	**0.0038**	0.0037	0.0134	0.0133	0.0059	0.0059	0.0036
		CP	–	0.932	**0.942**	0.726	0.702	0.926	0.924	0.842
		S($\boldsymbol{X}_I$)	–	165	162.142	165	162.142	110.786	114.16	–
		$\mathcal{T}$	–	0.2897	0.3083	**0.188**	0.2062	1.0884	1.0013	–
	800	Bias	0.002	**0.0001**	**0.0001**	0.0063	0.0073	0.0007	0.0008	0.0034
		SD	0.0025	**0.0026**	**0.0026**	0.0142	0.0142	0.0037	0.0036	0.0026
		SE	0.0024	**0.0026**	**0.0026**	0.0095	0.0094	0.003	0.003	0.0025
		CP	0.844	0.946	**0.948**	0.782	0.77	0.902	0.898	0.726
		S($\boldsymbol{X}_I$)	450	165	163.996	165	163.996	139.042	145.198	–
		$\mathcal{T}$	0.5978	0.8191	0.5395	**0.3335**	0.3993	3.3789	3.1002	–
0.5	200	Bias	–	**0.0001**	**0.0001**	0.0132	0.0137	0.001	0.0011	0.0019
		SD	–	**0.0041**	**0.0041**	0.021	0.0207	0.0115	0.0113	0.004
		SE	–	**0.0042**	**0.0042**	0.0132	0.0132	0.0106	0.0105	0.0042
		CP	–	**0.964**	**0.964**	0.708	0.7	0.948	0.952	0.954
		S($\boldsymbol{X}_I$)	–	165	163.146	165	163.146	98.262	101.384	–
		$\mathcal{T}$	–	0.3311	0.3096	**0.2147**	0.271	1.3929	1.4136	–
	400	Bias	–	0.0002	**0.0001**	0.0078	0.0083	0.0006	0.0007	0.0022
		SD	–	**0.0029**	**0.0029**	0.0144	0.0145	0.0064	0.0065	0.0029
		SE	–	**0.0029**	**0.0029**	0.0091	0.0091	0.0058	0.0058	0.0028
		CP	–	0.942	**0.944**	0.722	0.712	0.936	0.932	0.878
		S($\boldsymbol{X}_I$)	–	165	161.984	165	161.984	108.812	113.218	–
		$\mathcal{T}$	–	0.3125	0.2878	**0.1956**	0.2229	1.1287	1.0981	–
	800	Bias	0.0012	**0.0001**	**0.0001**	0.0035	0.0038	0.0006	0.0005	0.0021
		SD	0.002	**0.002**	**0.002**	0.0113	0.0113	0.0027	0.0026	0.002
		SE	0.0018	**0.002**	**0.002**	0.0064	0.0064	0.0022	0.0022	0.002
		CP	0.89	0.946	**0.95**	0.716	0.722	0.902	0.908	0.84
		S($\boldsymbol{X}_I$)	450	165	161.77	165	161.77	134.948	140.108	–
		$\mathcal{T}$	0.6357	0.8214	0.5464	**0.359**	0.3811	3.4726	3.1713	–

Note: True value $\beta_e = 1$.

Table 5. Results for Example 5 (with $\rho_o = 0.2, 0.5$).

ρ_o	n		2SLS	MA$^{(t,M)}$	MA$^{+(t,M)}$	$\mathfrak{s}^{(t,M)}$	$\mathfrak{s}^{+(t,M)}$	pLasso	pEL$_{0.5}$	Naive
0.2	200	Bias	–	**0.0012**	0.0013	0.0342	0.0375	0.0052	0.0053	0.0052
		SD	–	0.0064	**0.0065**	0.0315	0.0316	0.0148	0.0151	0.0062
		SE	–	0.0066	**0.0066**	0.0237	0.0232	0.0136	0.0137	0.0063
		CP	–	**0.94**	**0.94**	0.638	0.6	0.93	0.928	0.88
		S($\boldsymbol{X}_I$)	–	165	160.832	165	160.832	116.474	120.49	–
		$\mathcal{T}$	–	0.343	0.3049	**0.2282**	0.271	1.4037	1.3641	–
	400	Bias	–	**0.0006**	**0.0006**	0.0192	0.022	0.0019	0.002	0.0049
		SD	–	**0.0046**	0.0047	0.0225	0.0224	0.0075	0.0076	0.0045
		SE	–	**0.0046**	0.0046	0.0171	0.0168	0.0067	0.0067	0.0043
		CP	–	0.944	**0.946**	0.748	0.696	0.922	0.918	0.802
		S($\boldsymbol{X}_I$)	–	165	161.708	165	161.708	117.23	120.366	–
		$\mathcal{T}$	–	0.2657	0.2732	**0.1875**	0.2096	1.0936	1.0177	–
	800	Bias	0.0027	**0.0002**	0.0003	0.0105	0.012	0.0011	0.0012	0.0048
		SD	0.0031	0.0033	**0.0032**	0.0175	0.0176	0.0044	0.0044	0.0031
		SE	0.0028	0.0032	**0.0032**	0.0122	0.0119	0.0036	0.0036	0.003
		CP	0.806	**0.954**	**0.954**	0.786	0.748	0.894	0.892	0.656
		S($\boldsymbol{X}_I$)	450	165	161.81	165	161.81	145.63	152.232	–
		$\mathcal{T}$	0.4607	0.6399	0.458	**0.2828**	0.3235	2.4548	2.1332	–
0.5	200	Bias	–	**0.0016**	**0.0016**	0.0334	0.0371	0.0048	0.0048	0.0054
		SD	–	**0.0066**	**0.0066**	0.0363	0.036	0.0146	0.0149	0.0063
		SE	–	**0.0066**	**0.0066**	0.0237	0.0231	0.0139	0.014	0.0063
		CP	–	**0.948**	0.938	0.674	0.66	0.93	0.93	0.88
		S($\boldsymbol{X}_I$)	–	165	160.716	165	160.716	120.69	125.188	–
		$\mathcal{T}$	–	0.244	0.2273	**0.1606**	0.1814	0.976	0.9084	–
	400	Bias	–	**0.0006**	**0.0006**	0.0188	0.0216	0.0018	0.0019	0.005
		SD	–	**0.0046**	**0.0046**	0.0289	0.03	0.0074	0.0074	0.0044
		SE	–	**0.0046**	**0.0046**	0.0171	0.0167	0.0069	0.0069	0.0043
		CP	–	**0.95**	**0.95**	0.74	0.696	0.916	0.912	0.8
		S($\boldsymbol{X}_I$)	–	165	161.642	165	161.642	118.89	122.322	–
		$\mathcal{T}$	–	0.2576	0.2534	**0.1658**	0.1846	1.0992	1.0224	–
	800	Bias	0.0028	**0.0002**	0.0003	0.0096	0.0112	0.0016	0.0017	0.0048
		SD	0.0032	**0.0033**	**0.0033**	0.0181	0.0184	0.0041	0.0042	0.0033
		SE	0.0028	**0.0032**	**0.0032**	0.0122	0.012	0.0036	0.0036	0.003
		CP	0.796	**0.946**	0.942	0.78	0.758	0.894	0.884	0.664
		S($\boldsymbol{X}_I$)	450	165	161.868	165	161.868	157.216	165.498	–
		$\mathcal{T}$	0.4442	0.6363	0.4453	**0.2609**	0.3097	2.4621	2.3338	–

Note: True value $\beta_e = 1$.

longer in computing time. Meanwhile, the bias generally decreases as the sample size n or the correlation of $\boldsymbol{X}_I$ becomes larger. These results indicate the effectiveness and robustness of our proposed methods.

In terms of the CPs of the 95% confidence interval, the OLS model, $\mathfrak{s}^{(t,M)}$ and $\mathfrak{s}^{+(t,M)}$ are significantly lower than other methods. The proposed methods perform best since the associated CPs are closest to the nominal level (95%) among all the comparable methods. Under low-dimensional settings, the 2SLS method and two variable selection methods perform as well as our methods while they unlikely perform better when the number q of IVs gets greater. In Examples 3–5, the CP of either $\mathfrak{s}^{(t,M)}$ or $\mathfrak{s}^{+(t,M)}$ is even lower than the naive OLS method. In Example 6, despite the fact that the CP has decreased for all methods, our proposed model averaging methods still outperform other methods. The results indicate that our proposed methods generally perform well in terms of CP, verifying the validity of Theorem 3.2 in finite samples. Furthermore, Figure 1 showcases that although the points at both ends in each subfigure tend to depart from the reference line for all the methods, however, our proposed methods perform most satisfactorily since their points are closest to the reference line.

Table 6. Results for Example 6 (with $\rho_{cs} = 0, 0.2$).

ρ_o	n		2SLS	MA$^{(t,M)}$	MA$^{+(t,M)}$	$\mathfrak{s}^{(t,M)}$	$\mathfrak{s}^{+(t,M)}$	pLasso	pEL$_{0.5}$	Naive
0	200	Bias	0.0039	**0.0038**	0.004	0.0596	0.0618	0.0046	0.0047	0.0304
		SD	0.1048	0.1035	**0.1039**	0.1379	0.139	0.1056	0.1052	0.4425
		SE	0.1082	0.1087	**0.1085**	0.1065	0.1061	0.1095	0.1091	0.0453
		CP	0.89	**0.894**	0.892	0.772	0.764	0.892	0.89	0.118
		$\mathcal{T}$	**0.0053**	0.0198	0.0339	0.0154	0.0256	0.2209	0.2068	–
	400	Bias	0.0023	0.0024	0.0023	0.0361	0.0378	**0.0021**	**0.0021**	0.0044
		SD	0.0743	0.0732	0.0738	0.0959	0.0955	0.0743	**0.0749**	0.422
		SE	0.0807	0.0559	0.0558	0.0794	0.0793	0.0811	**0.081**	0.031
		CP	0.92	0.892	0.884	0.782	0.772	**0.922**	0.918	0.06
		$\mathcal{T}$	**0.0065**	0.019	0.042	0.0171	0.0307	0.2161	0.2101	–
	800	Bias	0.0024	**0.0022**	0.0023	0.018	0.0197	0.0025	0.0024	0.0418
		SD	0.0492	**0.0489**	0.0492	0.0694	0.0693	0.0493	0.0494	0.4486
		SE	0.0563	**0.0424**	0.0424	0.0563	0.0562	0.0566	0.0565	0.0223
		CP	0.912	**0.934**	0.93	0.784	0.778	0.912	0.91	0.04
		$\mathcal{T}$	**0.012**	0.0381	0.037	0.0242	0.0318	0.3707	0.4163	–
0.2	200	Bias	0.0022	**0.0021**	**0.0021**	0.0083	0.0089	0.0023	0.0022	0.0696
		SD	0.0324	0.0322	0.0323	0.0439	0.0439	0.0328	**0.0329**	0.3181
		SE	0.0368	0.0368	0.0369	0.0369	0.037	0.0372	**0.0371**	0.0192
		CP	0.892	0.892	**0.894**	0.812	0.812	0.89	0.892	0.062
		$\mathcal{T}$	**0.0113**	0.0329	0.0239	0.0232	0.0223	0.3203	0.2984	–
	400	Bias	0.0003	0.0004	**0.0001**	0.0155	0.0166	0.0006	0.0004	0.0352
		SD	0.046	**0.0456**	0.0457	0.064	0.064	0.0471	0.0473	0.3164
		SE	0.0541	**0.0408**	0.0407	0.0538	0.0538	0.0547	0.0547	0.0274
		CP	0.91	**0.926**	0.924	0.816	0.812	0.914	0.912	0.088
		$\mathcal{T}$	**0.0116**	0.0198	0.0257	0.0162	0.0222	0.2351	0.238	–
	800	Bias	0.0004	**0.0003**	**0.0003**	0.0086	0.0093	**0.0003**	**0.0003**	0.0026
		SD	0.033	0.0329	0.033	**0.0461**	0.0463	0.0339	0.0336	0.3165
		SE	0.0402	0.0402	0.0403	**0.0401**	0.0401	0.0405	0.0405	0.0191
		CP	0.918	**0.92**	**0.92**	0.808	0.806	**0.92**	**0.92**	0.07
		$\mathcal{T}$	**0.0093**	0.0319	0.0273	0.0207	0.0208	0.3008	0.302	–

Note: True value $\beta_e = -1$.

In addition, in terms of the execution time in Stage one, we can see that the execution time of all model average methods is slightly longer than the 2SLS method in low-dimensional settings, while the gap gets much wider in high-dimensional settings, see Tables 3–5. Besides, it is noted that the execution time of the variable selection methods is the longest among all methods, generally $3 \sim 10$ times longer than our methods. This is reasonable because variable selection methods usually involve some additional tuning parameters to be determined using some data-driven methods such as BIC and cross-validation, which burdens the computation.

In Examples 3–5, the number of IVs used in different methods is shown and compared. The model averaging procedure tends to choose a comparably smaller set of IVs than the 2SLS method, which shortens the computing time when the sample size gets larger. Moreover, the model averaging methods generally choose more variables than the variable selection methods when the exogenous variables or IVs are correlated. Accordingly, instead of abandoning the IVs with comparably lower marginal correlation with X_e, the model averaging methods tend to attribute non-zero sampling probability to them when constructing candidate submodels.

To further complement the simulation study, we provide additional results in Appendix 4. Specifically, Appendix A.1 investigates small-sample scenarios by reducing the sample size to $n = 50$ or $n = 100$ under dense or sparse designs. These results demonstrate that our proposed methods remain competitive even in small samples. Appendix A.2 conducts a sensitivity analysis with respect to the correlation between the endogenous and exogenous

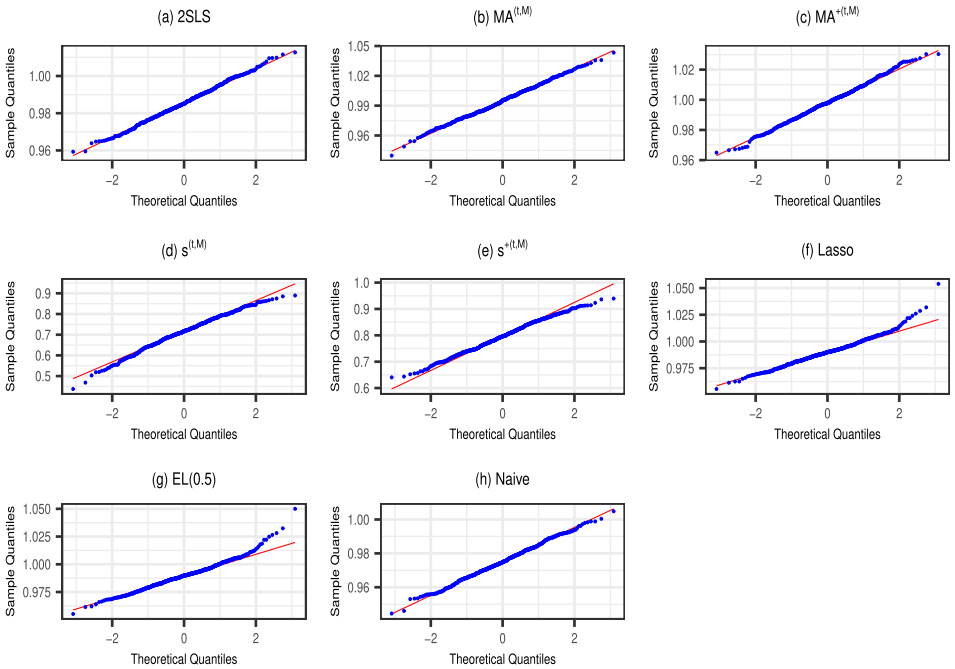

Figure 1. The quantile-quantile (QQ) plots for the different methods using the standard normal as the comparison distribution for Example 3 with sample size 800.

variables, denoted by ρ_{oe}. We vary this correlation level systematically and report the corresponding estimation biases. These additional results further support the robustness of our proposed estimators.

All codes are uploaded in https://github.com/wenjuns/2sls_code.git.

6. Real data analysis

We apply our proposed model averaging methods to the data set analysed by Belloni et al. (2012) and Belloni et al. (2014), which is published in the supplementary material on https://www.tandfonline.com/doi/suppl/10.1080/07350015.2020.1870479?scroll=top. This data set was used to study the effect of the government's exercise of eminent domain on home price, in which case the response variable Y denotes the log of Case-Shiller home price index depicting the numerical trend of home price. The endogenous variable X_e we are interested in is the number of pro-plaintiff appellate decisions, which is the number of the court ruling that a taking was illegal and judging the property in favour of a private owner over the government. X_e reflects the attitude of the court's towards the government's seizure of a possible private owner's property. If the value of X_e rises, it may indicate that the regime becomes more protective and supportive of individual property rights. Accordingly, β_e stands for the effect of an additional pro-plaintiff appellate decision supporting the private owner on the home price index. Referring to historical literature, we consider the exogenous variables, $\boldsymbol{X}_o$, including dummy variables such as whether there was any case in that circuit-year, numerical variables such as the number of appellate decisions, relevant factors including certain features

Table 7. Stage one performance of the different methods for the Case–Shiller data.

Method	$S(\boldsymbol{X}_I)$	F	pF	R^2	$\mathcal{T}_r$
$MA^{(10,12)}$	83	52.3512	7.6347×10^{-51}	0.786	0.0453
$MA^{+(10,12)}$	76	39.8719	2.809×10^{-43}	0.7367	0.0272
$\mathfrak{s}^{(10,12)}$	83	17.3996	5.5795×10^{-24}	0.5498	0.0117
$\mathfrak{s}^{+(10,12)}$	76	15.6873	3.4527×10^{-21}	0.5041	0.0101
pLasso	4	34.5186	2.389×10^{-21}	0.4355	0.2123
$pEL_{0.5}$	9	15.2486	3.6955×10^{-18}	0.4409	0.1175

Note: $S(\boldsymbol{X}_I)$ is the number of unique $\boldsymbol{X}_{Ii}$ used, F is the F value, pF is the p-value of the F statistic obtained, R^2 is the R-squared value and $\mathcal{T}_r$ is the execution time for Stage one in seconds.

of federal circuit court judges in a given circuit-year, circuit-relevant features, time-relevant features, and circuit-relevant time trends.

Although Guo et al. (2018) reject the endogeneity in the study of the causal effect between pro-plaintiff appellate decisions and home price under DWH test, other researches including D. L. Chen and Yeh (2012), Belloni et al. (2012), and Belloni et al. (2014) approve of the existence of possible endogeneity in this case. Thus, we regard the pro-plaintiff appellate decisions as the endogenous variable X_e in this study. The selection of IVs is based on the fact that the judges who deal with these cases are randomly assigned. Thus, the identity and characteristics of the judges are potential instrumental variables, in that they have influence on the response variable only through the effect of the endogenous variable X_e. Referring to past research, we also take the individual demographics of the federal circuit court judges in a given circuit-year as potential instruments, such as basic identical information including gender, race, religion, political affiliation, and other dummy variables including whether the judge obtained the bachelor's degree in-state, type of university from which the judge obtained the bachelor's degree, and possible interactions are constructed among these features (Belloni et al., 2012, 2014). There are a total of 147 variables selected as instruments. We first compute the Pearson correlations of these potential instruments with the endogenous variable X_e before constructing the instruments set. The result shows that there exist some very weak instruments, and we keep them to investigate whether our methods perform well under the influence of weak instruments. All 71 exogenous variables are also included as exogenous variable $\boldsymbol{X}_o$ in Models (1) and (2). The sample size of this data is $n = 183$. All variables involved in the model including Y, X_e, $\boldsymbol{X}_I$, and $\boldsymbol{X}_o$ are standardized to have mean zero and standard deviation one.

We employ all the methods in the simulation study to analyse this data and make a comparison. We fix $t_m = t$ for convenience such that each submodel has the same number of IVs. We take $(t, M) = (10, 12)$ in this case study referring to the BIC_M criterion. We consider M submodels in the first stage of 2SLS, in which the mth submodel includes $\boldsymbol{X}_o$ and different $\boldsymbol{X}_I^{(m)} = (X_{I1}^{(m)}, \ldots, X_{It}^{(m)})^\top \in \mathbb{R}^t$, with each $X_{Ii}^{(m)}$ being sampled from $\{X_{I1}, \ldots, X_{Iq}\}$ with *equal* or *unequal* probability in two model averaging methods, respectively (Table 7).

Since the true value of β_e is unknown in reality, we take the results of Belloni et al. (2014), which proposed a double selection procedure, as the benchmark. We use the method of Belloni et al. (2014) to select only one IV, and the resulting estimate of β_e is 0.0648. Table 8 reports the estimates of β_e of various methods. In terms of bias, $MA^{+(10,12)}$ produces the closest result to the benchmark among all comparing methods. Another observation is that $\mathfrak{s}^{+(10,12)}$ and model selection methods tend to obtain relatively larger estimates of β_e than our

Table 8. Results of estimation for the Case–Shiller data.

Method	$\widehat{\beta_e}$	SE	p-value
$MA^{(10,12)}$	0.0359	0.0343	0.2959
$MA^{+(10,12)}$	0.0656	0.0397	0.0982
$\mathfrak{s}^{(10,12)}$	0.0538	0.0195	0.0058
$\mathfrak{s}^{+(10,12)}$	0.0842	0.0198	2.073×10^{-5}
pLasso	0.0922	0.036	0.0104
$pEL_{0.5}$	0.099	0.0343	0.0039
naive	0.0237	0.0224	0.2918

Note: $\hat{\beta}_e$ is the estimated coefficient for X_e and SE is the estimated standard error of the estimator.

methods. However, $\mathfrak{s}^{+(10,12)}$ may overestimate β_e without considering the exogenous variables in Stage one. The model selection methods may remove too many variables in the first stage referring to the value of $S(\boldsymbol{X}_I)$. Furthermore, we also find that all estimates except the naive and $MA^{(10,12)}$ are significant at the 0.1 level. Overall, $MA^{+(10,12)}$ gives a relatively less bias and has an advantage in execution time over the model selection methods.

7. Concluding remarks

In this paper, we investigate a two-stage least squares model averaging method for the estimation of the coefficient of the endogenous variable in the structural equation models. Differing from the existing work, we allow exogenous variables to be included in both stages. Theoretically, we show that the proposed model averaging estimator can produce a bias when the exogenous variables considered are correlated with the endogenous variable. To make statistical inference, we then propose a debiased estimator, which is consistent and asymptotic normal. Meanwhile, another interpretation of the debiased estimator is provided in Section 4.2 from a distinct construction perspective of the estimator. Extensive numerical results are carried out to illustrate our proposal.

In the future work, several issues on the extension of this paper may deserve further study. First, when the number of IVs is ultrahigh-dimensional, one may develop some IV screening method to reduce the dimension and then perform model averaging. For instance, in a Mendelian randomization study, millions of genetic markers or SNPs can be qualified as IVs. Second, as preparing submodels is an important issue in model averaging, one may construct submodels based on some sampling criteria. Third, when the endogenous variable is binary, one may consider logistic regression to handle this case. Moreover, extending this approach to nonlinear or semiparametric IV models would be interesting as well.

In terms of optimization of weights, while many traditional model averaging methods impose constraints such as non-negative weights and a simplex constraint $\sum_{i=1}^{M} w_i = 1$, there also exists a substantial body of literature that does not require such constraints (J. Chen et al., 2018; D. Li et al., 2015). In our framework, we do not explicitly enforce these constraints. However, as noted in Remark 3.2, if the weight vector satisfies $\sum_{i=1}^{M} w_i = 1$, the bias term can be eliminated under certain conditions. This observation suggests that imposing such a constraint may be beneficial in reducing bias, which is left to be explored in our future research.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

The authors were supported by National Natural Science Foundation of China [No. 11801202] and Fundamental Research Funds for the Central Universities [No. 2025CDJZKPT-09].

ORCID

Xiaochao Xia http://orcid.org/0000-0002-9414-355X

References

Ando, T., & Li, K.-C. (2017). A weight-relaxed model averaging approach for high-dimensional generalized linear models. *The Annals of Statistics*, *45*(6), 2654–2679. https://doi.org/10.1214/17-AOS1538

Angrist, J. D., Imbens, G. W., & Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, *91*(434), 468–472.

Bekker, P. A. (1994). Alternative approximations to the distributions of instrumental variable estimators. *Econometrica*, *62*(3), 657–681. https://doi.org/10.2307/2951662

Belloni, A., Chen, D., Chernozhukov, V., & Hansen, C. (2012). Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica*, *80*(6), 2369–2429. https://doi.org/10.3982/ECTA9626

Belloni, A., Chernozhukov, V., & Hansen, C. (2014). High-dimensional methods and inference on structural and treatment effects. *Journal of Economic Perspectives*, *28*(2), 29–50. https://doi.org/10.1257/jep.28.2.29

Bound, J., Jaeger, D. A., & Baker, R. M. (1995). Problems with instrumental variables estimation when the correlation between the instruments and the endogeneous explanatory variable is weak. *Journal of the American Statistical Association*, *90*(430), 443–450.

Canay, I. A. (2010). Simultaneous selection and weighting of moments in GMM using a trapezoidal kernel. *Journal of Econometrics*, *156*(2), 284–303. https://doi.org/10.1016/j.jeconom.2009.10.036

Chen, D. L., & Yeh, S. (2012). Growth under the shadow of expropriation? The economic impacts of eminent domain. https://doi.org/10.2139/ssrn.2977074

Chen, J., Jiang, B., & Li, J. (2023). Nonparametric instrument model averaging. *Journal of Nonparametric Statistics*, *35*(4), 905–926. https://doi.org/10.1080/10485252.2023.2215339

Chen, J., Li, D., Linton, O., & Lu, Z. (2018). Semiparametric ultra-high dimensional model averaging of nonlinear dynamic time series. *Journal of the American Statistical Association*, *113*(522), 919–932. https://doi.org/10.1080/01621459.2017.1302339

Corbae, D., Durlauf, S., & Hansen, B. (2006). *Econometric Theory and Practice*. Cambridge University Press.

Duncan, O. D. (1975). *Introduction to Structural Equation Models*. Academic Press.

Fan, Q., & Zhong, W. (2018). Variable selection for structural equation with endogeneity. *Journal of Systems Science and Complexity*, *31*(3), 787–803. https://doi.org/10.1007/s11424-017-6195-4

Fang, F., Li, J., & Xia, X. (2022). Semiparametric model averaging prediction for dichotomous response. *Journal of Econometrics*, *229*(2), 219–245. https://doi.org/10.1016/j.jeconom.2020.09.008

Feng, Y., Liu, Q., Yao, Q., & Zhao, G. (2022). Model averaging for nonlinear regression models. *Journal of Business & Economic Statistics*, *40*(2), 785–798. https://doi.org/10.1080/07350015.2020.1870477

Guo, Z., Kang, H., Cai, T. T., & Small, D. S. (2018). Testing endogeneity with high dimensional covariates. *Journal of Econometrics*, *207*(1), 175–187. https://doi.org/10.1016/j.jeconom.2018.07.002

Hansen, B. E. (2007). Least squares model averaging. *Econometrica*, *75*(4), 1175–1189. https://doi.org/10.1111/ecta.2007.75.issue-4

Hansen, B. E. (2017). Stein-like 2SLS estimator. *Econometric Reviews*, *36*(6–9), 840–852. https://doi.org/10.1080/07474938.2017.1307579

Hansen, B. E. (2022). *Econometrics*. Princeton University Press.

Hansen, C., Hausman, J., & Newey, W. (2008). Estimation with many instrumental variables. *Journal of Business & Economic Statistics*, *26*(4), 398–422. https://doi.org/10.1198/073500108000000024
Hong, Y. (2020). *Foundations of Modern Econometrics*. World Scientific.
Kang, H., Zhang, A., Cai, T. T., & Small, D. S. (2016). Instrumental variables estimation with some invalid instruments and its application to mendelian randomization. *Journal of the American Statistical Association*, *111*(513), 132–144. https://doi.org/10.1080/01621459.2014.994705
Kline, R. B. (1998). *Principles and Practice of Structural Equation Modeling*. The Guilford Press.
Kok, B. C., Choi, J. S., Oh, H., & Choi, J. Y. (2021). Sparse extended redundancy analysis: Variable selection via the exclusive Lasso. *Multivariate Behavioral Research*, *56*(3), 426–446. https://doi.org/10.1080/00273171.2019.1694477
Kuersteiner, G., & Okui, R. (2010). Constructing optimal instruments by first-stage prediction averaging. *Econometrica*, *78*(2), 697–718. https://doi.org/10.3982/ECTA7444
Li, C., Li, Q., Racine, J. S., & Zhang, D. (2018). Optimal model averaging of varying coefficient models. *Statistica Sinica*, *28*(4, SI), 1017–0405.
Li, D., Linton, O., & Lu, Z. (2015). A flexible semiparametric forecasting model for time series. *Journal of Econometrics*, *187*(1), 495–509.
Li, J., Lv, J., Wan, A. T. K., & Liao, J. (2022). Adaboost semiparametric model averaging prediction for multiple categories. *Journal of the American Statistical Association*, *117*(537), 495–509. https://doi.org/10.1080/01621459.2020.1790375
Li, J., Xia, X., Wong, W. K., & Nott, D. (2018). Varying-coefficient semiparametric model averaging prediction. *Biometrics*, *74*(4), 1417–1426. https://doi.org/10.1111/biom.12904
Liu, C.-A. (2015). Distribution theory of the least squares averaging estimator. *Journal of Econometrics*, *186*(1), 142–159. https://doi.org/10.1016/j.jeconom.2014.07.002
Martins, L. F., & Gabriel, V. J. (2014). Linear instrumental variables model averaging estimation. *Computational Statistics & Data Analysis*, *71*, 709–724. https://doi.org/10.1016/j.csda.2013.05.008
Nelson, C. R., & Startz, R. (1990). The distribution of the instrumental variables estimator and its t-ratio when the instrument is a poor one. *The Journal of Business*, *63*(S1), S125–S140. https://doi.org/10.1086/jb.1990.63.issue-S1
Okui, R. (2011). Instrumental variable estimation in the presence of many moment conditions. *Journal of Econometrics*, *165*(1), 70–86. https://doi.org/10.1016/j.jeconom.2011.05.007
Seng, L., & Li, J. (2022). Structural equation model averaging: Methodology and application. *Journal of Business & Economic Statistics*, *40*(2), 815–828. https://doi.org/10.1080/07350015.2020.1870479
Stock, J. H., Wright, J. H., & Yogo, M. (2002). A survey of weak instruments and weak identification in generalized method of moments. *Journal of Business & Economic Statistics*, *20*(4), 518–529. https://doi.org/10.1198/073500102288618658
Zhang, X., & Liu, C. (2023). Model averaging prediction by *K*-fold cross-validation. *Journal of Econometrics*, *235*(1), 280–301. https://doi.org/10.1016/j.jeconom.2022.04.007
Zhang, X., & Wang, W. (2019). Optimal model averaging estimation for partially linear models. *Statistica Sinica*, *29*(2), 693–718.
Zhang, X., & Zhang, X. (2023). Optimal model averaging based on forward-validation. *Journal of Econometrics*, *237*(2),105295. https://doi.org/10.1016/j.jeconom.2022.03.010
Zhang, X., Zou, G., & Liang, H. (2014). Model averaging and weight choice in linear mixed-effects models. *Biometrika*, *101*(1), 205–218. https://doi.org/10.1093/biomet/ast052
Zhu, R., Wan, A. T. K., Zhang, X., & Zou, G. (2019). A Mallows-type model averaging estimator for the varying-coefficient partially linear model. *Journal of the American Statistical Association*, *114*(526), 882–892. https://doi.org/10.1080/01621459.2018.1456936
Zhu, R., Zhang, X., Wan, A. T. K., & Zou, G. (2023). Kernel averaging estimators. *Journal of Business & Economic Statistics*, *41*(1), 157–169. https://doi.org/10.1080/07350015.2021.2006668

Appendices

Appendix 1. List of notations

Before proving the theoretical results stated in Sections 3 and 4, we list some notations to simplify our proof. For $m, m' = 1, \ldots, M$, $i = 1, \ldots, n$,

$$\tilde{\varepsilon}_i \triangleq [\boldsymbol{X}_{I,i} - \mathcal{X}_I^\top \mathcal{X}_o (\mathcal{X}_o^\top \mathcal{X}_o)^{-1} \boldsymbol{X}_{o,i}]\varepsilon_i,$$

$$\hat{\boldsymbol{v}}_m \triangleq \frac{1}{n} \mathcal{Z}^{(m)\top} \boldsymbol{X}_e, \quad \hat{\boldsymbol{\Sigma}}_{mm'} \triangleq \frac{1}{n} \mathcal{Z}^{(m)\top} \mathcal{Z}^{(m')}, \quad \hat{\boldsymbol{\Sigma}}_{mm}^{-1} \triangleq \left(\frac{1}{n} \mathcal{Z}^{(m)\top} \mathcal{Z}^{(m)} \right)^{-1},$$

$$\boldsymbol{w}_0 \triangleq \boldsymbol{\Psi}^{-1} \boldsymbol{u}, \quad \boldsymbol{u} \triangleq \left(\boldsymbol{v}_m^\top \boldsymbol{\Sigma}_{mm}^{-1} \boldsymbol{v}_m \right) \in \mathbb{R}^M, \quad \boldsymbol{\Psi} \triangleq \left(\boldsymbol{v}_m^\top \boldsymbol{\Sigma}_{mm}^{-1} \boldsymbol{\Sigma}_{mm'} \boldsymbol{\Sigma}_{m'm'}^{-1} \boldsymbol{v}_{m'} \right) \in \mathbb{R}^{M \times M},$$

$$\tilde{\boldsymbol{w}} \triangleq \hat{\boldsymbol{\Psi}}^{-1} \hat{\boldsymbol{u}}, \quad \hat{\boldsymbol{u}} \triangleq \frac{1}{n} \tilde{\mathcal{X}}_e^\top \boldsymbol{H} \boldsymbol{X}_e, \quad \hat{\boldsymbol{\Psi}} \triangleq \frac{1}{n} \tilde{\mathcal{X}}_e^\top \boldsymbol{H} \boldsymbol{H}^\top \tilde{\mathcal{X}}_e,$$

$$\boldsymbol{Q} \triangleq a \mathbf{1}_M \mathbf{1}_M^\top \quad \text{with } a \triangleq \boldsymbol{M}_{oe}^\top \boldsymbol{M}_{oo}^{-1} \boldsymbol{M}_{oe},$$

$$\hat{\boldsymbol{Q}} \triangleq \hat{a} \mathbf{1}_M \mathbf{1}_M^\top \quad \text{with } \hat{a} \triangleq \frac{1}{n} \boldsymbol{X}_e^\top \mathcal{X}_o (\mathcal{X}_o^\top \mathcal{X}_o)^{-1} \mathcal{X}_o^\top \boldsymbol{X}_e,$$

$$\boldsymbol{\xi}_i \triangleq (\xi_{1,i}, \ldots, \xi_{M,i})^\top, \quad \boldsymbol{\gamma}_i \triangleq (\gamma_{mm',i})_{m,m'=1,\ldots,M} \in \mathbb{R}^{M \times M},$$

$$\zeta_i \triangleq 2 \boldsymbol{M}_{oe}^\top \boldsymbol{M}_{oo}^{-1} (\boldsymbol{X}_{o,i} X_{e,i} - \boldsymbol{M}_{oe}) - \boldsymbol{M}_{oe}^\top \boldsymbol{M}_{oo}^{-1} (\boldsymbol{X}_{o,i} \boldsymbol{X}_{o,i}^\top - \boldsymbol{M}_{oo}) \boldsymbol{M}_{oo}^{-1} \boldsymbol{M}_{oe},$$

$$\boldsymbol{\varphi}_i \triangleq \boldsymbol{\Psi}^{-1} (\boldsymbol{\xi}_i - \boldsymbol{\gamma}_i \boldsymbol{w}_0), \quad \boldsymbol{\Xi} \triangleq \operatorname{cov}(\boldsymbol{\varphi}_i, \boldsymbol{\varphi}_i) = \boldsymbol{\Psi}^{-1} \operatorname{cov}(\boldsymbol{\xi}_i - \boldsymbol{\gamma}_i \boldsymbol{w}_0, \boldsymbol{\xi}_i - \boldsymbol{\gamma}_i \boldsymbol{w}_0) \boldsymbol{\Psi}^{-1},$$

$$s_1 \triangleq \boldsymbol{w}_0^\top \boldsymbol{\Psi} \boldsymbol{w}_0 \beta_e - \boldsymbol{w}_0^\top \mathbf{1}_M a \beta_e, \quad s_2 \triangleq \left[\boldsymbol{w}_0^\top \boldsymbol{\Psi} \boldsymbol{w}_0 - a (\boldsymbol{w}_0^\top \mathbf{1}_M)^2 \right] \beta_e,$$

$$s_3 \triangleq \boldsymbol{w}_0^\top \boldsymbol{\Psi} \boldsymbol{w}_0 - a (\boldsymbol{w}_0^\top \mathbf{1}_M)^2, \quad s_4 \triangleq \boldsymbol{w}_0^\top \boldsymbol{\Psi} \boldsymbol{w}_0 - \boldsymbol{w}_0^\top \mathbf{1}_M a,$$

$$\kappa_i \triangleq 2 \boldsymbol{w}_0^\top (\boldsymbol{I}_{M \times M} - a \mathbf{1}_M \mathbf{1}_M^\top \boldsymbol{\Psi}^{-1}) \boldsymbol{\xi}_i + \boldsymbol{w}_0^\top (2a \mathbf{1}_M \mathbf{1}_M^\top \boldsymbol{\Psi}^{-1} - \boldsymbol{I}_{M \times M}) \boldsymbol{\gamma}_i \boldsymbol{w}_0 - \boldsymbol{w}_0^\top \mathbf{1}_M \mathbf{1}_M^\top \boldsymbol{w}_0 \zeta_i,$$

$$\boldsymbol{\phi}_i \triangleq \left(\phi_{1,i}, \ldots, \phi_{M,i} \right)^\top, \quad \text{with } \phi_{m,i} = (\boldsymbol{v}_m^\top \boldsymbol{\Sigma}_{mm}^{-1} \boldsymbol{Z}_{m,i} - \boldsymbol{M}_{oe}^\top \boldsymbol{M}_{oo}^{-1} \boldsymbol{X}_{o,i}) \varepsilon_i, m = 1, \ldots, M,$$

$$\begin{aligned} \tilde{\varphi}_i \triangleq\ & [(\beta_e + \text{bias}) 2a \boldsymbol{w}_0^\top \mathbf{1}_M \mathbf{1}_M^\top \boldsymbol{\Psi}^{-1} - a \beta_e \mathbf{1}_M^\top \boldsymbol{\Psi}^{-1} - 2\text{bias} \boldsymbol{w}_0^\top] \boldsymbol{\xi}_i \\ & + [\beta_e a \mathbf{1}_M^\top \boldsymbol{\Psi}^{-1} - 2a \boldsymbol{w}_0^\top \mathbf{1}_M \mathbf{1}_M^\top \boldsymbol{\Psi}^{-1} (\beta_e + \text{bias}) + \text{bias} \boldsymbol{w}_0^\top] \boldsymbol{\gamma}_i \boldsymbol{w}_0 \\ & + [(\beta_e + \text{bias}) (\boldsymbol{w}_0^\top \mathbf{1}_M)^2 - \beta_e \boldsymbol{w}_0^\top \mathbf{1}_M] \zeta_i + \boldsymbol{w}_0^\top \boldsymbol{\phi}_i, \end{aligned}$$

$$\boldsymbol{d} \triangleq \left(\boldsymbol{v}_1^\top \boldsymbol{\Sigma}_{11}^{-1} \boldsymbol{d}_1 \boldsymbol{M}_{oo}^{-1} \boldsymbol{M}_{oe}, \ldots, \boldsymbol{v}_M^\top \boldsymbol{\Sigma}_{MM}^{-1} \boldsymbol{d}_M \boldsymbol{M}_{oo}^{-1} \boldsymbol{M}_{oe} \right)^\top, \quad \text{with } \boldsymbol{d}_m \triangleq E(\boldsymbol{Z}_{m,i} \boldsymbol{X}_o^\top) \in \mathbb{R}^{(t_m + p) \times p},$$

$$c_{\text{bias}} \triangleq s_3^{-1} \boldsymbol{w}_0^\top \boldsymbol{Q} \boldsymbol{w}_0 - a \boldsymbol{w}_0^\top \mathbf{1}_M), \quad \hat{c}_{\text{bias}} \triangleq \hat{s}_3^{-1} (\hat{\boldsymbol{w}}^\top \hat{\boldsymbol{Q}} \hat{\boldsymbol{w}} - \hat{a} \hat{\boldsymbol{w}}^\top \mathbf{1}_M),$$

$$\text{bias} \triangleq c_{\text{bias}} \beta_e,$$

$$\boldsymbol{\Lambda}_1 \triangleq \begin{pmatrix} \boldsymbol{u}^\top \boldsymbol{\Psi}^{-1} \boldsymbol{u} & \boldsymbol{u}^\top \boldsymbol{\Psi}^{-1} \tilde{\boldsymbol{u}} \\ \boldsymbol{M}_{oe} & \boldsymbol{M}_{oo} \end{pmatrix} \quad \text{with } \tilde{\boldsymbol{u}} = \left(\boldsymbol{v}_m^\top \boldsymbol{\Sigma}_{mm} E(\boldsymbol{Z}^{(m)} X_e) \right)_{M \times 1},$$

$$\tilde{\boldsymbol{\phi}}_i \triangleq \begin{pmatrix} g_i \\ \boldsymbol{X}_{o,i} \end{pmatrix} \varepsilon_i \text{ with } g_i = \boldsymbol{u}^\top \boldsymbol{\Psi}^{-1} \begin{pmatrix} \boldsymbol{v}_1^\top \boldsymbol{\Sigma}_{11}^{-1} \boldsymbol{Z}_{1,i} \\ \vdots \\ \boldsymbol{v}_M^\top \boldsymbol{\Sigma}_{MM}^{-1} \boldsymbol{Z}_{M,i} \end{pmatrix},$$

$$\boldsymbol{\Lambda}_2 \triangleq \begin{pmatrix} \Lambda_{2,1} & \boldsymbol{\Lambda}_{2,2} \\ \boldsymbol{\Lambda}_{2,2}^\top & \boldsymbol{\Lambda}_{2,3} \end{pmatrix} \quad \text{with } \Lambda_{2,1} = \boldsymbol{u}^\top \boldsymbol{\Psi}^{-1} \boldsymbol{u}, \boldsymbol{\Lambda}_{2,2} = \boldsymbol{u}^\top \boldsymbol{\Psi}^{-1} (\boldsymbol{v}_m^\top \boldsymbol{\Sigma}_{mm}^{-1} \boldsymbol{d}_m)_{M \times p} \text{ and } \boldsymbol{\Lambda}_{2,3} = \boldsymbol{M}_{oo},$$

where the mth element of $\boldsymbol{\xi}_i$ is $\xi_{m,i} = 2 \boldsymbol{v}_m^\top \boldsymbol{\Sigma}_{mm}^{-1} (\boldsymbol{Z}_{m,i} X_{e,i} - \boldsymbol{v}_m) - \boldsymbol{v}_m^\top \boldsymbol{\Sigma}_{mm}^{-1} (\boldsymbol{Z}_{m,i} \boldsymbol{Z}_{m,i}^\top - \boldsymbol{\Sigma}_{mm}) \boldsymbol{\Sigma}_{mm}^{-1} \boldsymbol{v}_m$, and the (m, m')th element of $\boldsymbol{\gamma}_i$ is

$$\gamma_{mm',i} = [(\boldsymbol{Z}_{m,i} X_{e,i} - \boldsymbol{v}_m)^\top \boldsymbol{\Sigma}_{mm}^{-1} \boldsymbol{\Sigma}_{mm'} \boldsymbol{\Sigma}_{m'm'}^{-1} \boldsymbol{v}_{m'}$$

$$
\begin{aligned}
&+ \boldsymbol{\nu}_m^\top \boldsymbol{\Sigma}_{mm}^{-1} \boldsymbol{\Sigma}_{mm'} \boldsymbol{\Sigma}_{m'm'}^{-1} (\boldsymbol{Z}_{m',i} X_{e,i} - \boldsymbol{\nu}_{m'})] \\
&- [\boldsymbol{\nu}_m^\top \boldsymbol{\Sigma}_{mm}^{-1} (\boldsymbol{Z}_{m,i} \boldsymbol{Z}_{m,i}^\top - \boldsymbol{\Sigma}_{mm}) \boldsymbol{\Sigma}_{mm}^{-1} \boldsymbol{\Sigma}_{mm'} \boldsymbol{\Sigma}_{m'm'}^{-1} \boldsymbol{\nu}_{m'} \\
&+ \boldsymbol{\nu}_m^\top \boldsymbol{\Sigma}_{mm}^{-1} \boldsymbol{\Sigma}_{mm'} \boldsymbol{\Sigma}_{m'm'}^{-1} (\boldsymbol{Z}_{m',i} \boldsymbol{Z}_{m',i}^\top - \boldsymbol{\Sigma}_{m'm'}) \boldsymbol{\Sigma}_{m'm'}^{-1} \boldsymbol{\nu}_{m'}] \\
&+ \boldsymbol{\nu}_m^\top \boldsymbol{\Sigma}_{mm}^{-1} (\boldsymbol{Z}_{m,i} \boldsymbol{Z}_{m',i}^\top - \boldsymbol{\Sigma}_{mm'}) \boldsymbol{\Sigma}_{m'm'}^{-1} \boldsymbol{\nu}_{m'}.
\end{aligned}
$$

Appendix 2. Proofs of some useful lemmas

Lemma A.1: *Suppose that* $E|\tilde{\varepsilon}_i|^2 < \infty$, *where* $\tilde{\varepsilon}_i$ *is denoted in Appendix* 1. *Under Assumptions* (i)–(iv), *as* $n \to \infty$, *we have (i) (Consistency)* $\widehat{\beta}_{e,2\text{SLS}} \xrightarrow{p} \beta_e$, *and* (ii) *(Asymptotic Normality)* $\sqrt{n}(\widehat{\beta}_{e,2\text{SLS}} - \beta_e) \xrightarrow{d} N(0, [\widetilde{\boldsymbol{M}}_{Ie}^\top \widetilde{\boldsymbol{M}}_{II}^{-1} \widetilde{\boldsymbol{M}}_{Ie}]^{-1} \sigma^2)$, *where* $\widetilde{\boldsymbol{M}}_{Ie} = \boldsymbol{M}_{Ie} - \boldsymbol{M}_{Io} \boldsymbol{M}_{oo}^{-1} \boldsymbol{M}_{oe}$, $\widetilde{\boldsymbol{M}}_{II} = \boldsymbol{M}_{II} - \boldsymbol{M}_{Io} \boldsymbol{M}_{oo}^{-1} \boldsymbol{M}_{Io}^\top$, $\boldsymbol{M}_{oe}^\top = E(X_e \boldsymbol{X}_o^\top)$, $\boldsymbol{M}_{Ie}^\top = E(X_e \boldsymbol{X}_I^\top)$ *and* $\boldsymbol{M}_{Io} = E(\boldsymbol{X}_I \boldsymbol{X}_o^\top)$.

Remark A.1: This lemma provides a conventional result for the 2SLS estimator when all observed IVs are valid and available and the IV model is correctly specified in both stages. The consistency and asymptotic normality derived in Lemma A.1 could help us to make statistical inference on the 2SLS estimator, $\widehat{\beta}_{e,2\text{SLS}}$. For example, one may construct the confidence interval with the asymptotic variance of $\widehat{\beta}_{e,2\text{SLS}}$ being replaced with the sample moment estimator of $[\widetilde{\boldsymbol{M}}_{Ie}^\top \widetilde{\boldsymbol{M}}_{II}^{-1} \widetilde{\boldsymbol{M}}_{Ie}]$, which can be easily computed as in our simulation and empirical studies.

Proof: (i) First we derive the consistency of $\widehat{\beta}_{e,2\text{SLS}}$. Note that

$$
\hat{\beta}_{e,2\text{SLS}} = \beta_e + \left[\left(\frac{1}{n} \boldsymbol{X}_e^\top \mathcal{Z}_I \right) \left(\frac{1}{n} \mathcal{Z}_I^\top \mathcal{Z}_I \right)^{-1} \left(\frac{1}{n} \mathcal{Z}_I^\top \boldsymbol{X}_e \right) \right]^{-1} \left(\frac{1}{n} \boldsymbol{X}_e^\top \mathcal{Z}_I \right) \left(\frac{1}{n} \mathcal{Z}_I^\top \mathcal{Z}_I \right)^{-1} \left(\frac{1}{n} \mathcal{Z}_I^\top \boldsymbol{\varepsilon} \right). \tag{A1}
$$

Then we have

$$
\begin{aligned}
\frac{1}{n} \boldsymbol{X}_e^\top \mathcal{Z}_I &= \frac{1}{n} \boldsymbol{X}_e^\top (\boldsymbol{I}_n - \boldsymbol{P}_1) \mathcal{X}_I \\
&= \frac{1}{n} \boldsymbol{X}_e^\top (\boldsymbol{I}_n - \mathcal{X}_o (\mathcal{X}_o^\top \mathcal{X}_o)^{-1} \mathcal{X}_o^\top) \mathcal{X}_I \\
&\xrightarrow{p} \boldsymbol{M}_{Ie}^\top - \boldsymbol{M}_{oe}^\top \boldsymbol{M}_{oo}^{-1} \boldsymbol{M}_{Io}^\top = \widetilde{\boldsymbol{M}}_{Ie}^\top,
\end{aligned}
$$

where the last line is obtained by Assumption (iv) and the law of large numbers. Similarly we have

$$
\begin{aligned}
\frac{1}{n} \mathcal{Z}_I^\top \mathcal{Z}_I &= \frac{1}{n} \mathcal{X}_I^\top (\boldsymbol{I}_n - \boldsymbol{P}_1) \mathcal{X}_I \\
&\xrightarrow{p} \boldsymbol{M}_{II} - \boldsymbol{M}_{Io} \boldsymbol{M}_{oo}^{-1} \boldsymbol{M}_{Io}^\top = \widetilde{\boldsymbol{M}}_{II}.
\end{aligned}
$$

Additionally we have

$$
\begin{aligned}
\frac{1}{n} \mathcal{Z}_I^\top \boldsymbol{\varepsilon} &= \frac{1}{n} \mathcal{X}_I^\top (\boldsymbol{I}_n - \boldsymbol{P}_1) \boldsymbol{\varepsilon} \\
&= \frac{1}{n} \mathcal{X}_I^\top (\boldsymbol{I}_n - \mathcal{X}_o (\mathcal{X}_o^\top \mathcal{X}_o)^{-1} \mathcal{X}_o^\top) \boldsymbol{\varepsilon} \\
&= \frac{1}{n} \mathcal{X}_I^\top \boldsymbol{\varepsilon} - \frac{1}{n} \mathcal{X}_I^\top \mathcal{X}_o \left(\frac{1}{n} \mathcal{X}_o^\top \mathcal{X}_o \right)^{-1} \left(\frac{1}{n} \mathcal{X}_o^\top \boldsymbol{\varepsilon} \right) \\
&\xrightarrow{p} 0
\end{aligned}
$$

under Assumptions (i), (iii) and (iv), where the last line is obtained by continuous mapping theorem. Again applying continuous mapping theorem and the above results to Equation (A1), thus

$$\widehat{\beta}_{e,\text{2SLS}} - \beta_e \xrightarrow{p} 0.$$

(ii) Under Assumptions (i)-(iv), by continuous mapping theorem, we have

$$\begin{aligned}\sqrt{n}(\widehat{\beta}_{e,\text{2SLS}} - \beta_e) &= \left[\left(\frac{1}{n}\boldsymbol{X}_e^\top \mathcal{Z}_I\right)\left(\frac{1}{n}\mathcal{Z}_I^\top \mathcal{Z}_I\right)^{-1}\left(\frac{1}{n}\mathcal{Z}_I^\top \boldsymbol{X}_e\right)\right]^{-1}\left(\frac{1}{n}\boldsymbol{X}_e^\top \mathcal{Z}_I\right)\left(\frac{1}{n}\mathcal{Z}_I^\top \mathcal{Z}_I\right)^{-1}\left(\frac{1}{\sqrt{n}}\mathcal{Z}_I^\top \boldsymbol{\varepsilon}\right)\\ &= \widetilde{\boldsymbol{M}}_{Ie}^\top \widetilde{\boldsymbol{M}}_{II}^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^n [\boldsymbol{X}_{I,i} - \mathcal{X}_I^\top \mathcal{X}_o(\mathcal{X}_o^\top \mathcal{X}_o)^{-1}\boldsymbol{X}_{o,i}]\varepsilon_i \times \{o_p(1)+1\}\\ &= \widetilde{\boldsymbol{M}}_{Ie}^\top \widetilde{\boldsymbol{M}}_{II}^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^n \tilde{\varepsilon}_i,\end{aligned}$$

where $\tilde{\varepsilon}_i = [\boldsymbol{X}_{I,i} - \mathcal{X}_I^\top \mathcal{X}_o(\mathcal{X}_o^\top \mathcal{X}_o)^{-1}\boldsymbol{X}_{o,i}]\varepsilon_i$. It is easy to obtain that $\text{cov}(\tilde{\varepsilon}_i, \tilde{\varepsilon}_i) = \widetilde{\boldsymbol{M}}_{II}\sigma^2$. Under the second-order moment condition that $E\|\tilde{\varepsilon}_i\|^2 < \infty$, by multivariate Linderberg-lévy central limit theorem (B. E. Hansen, 2022, Theorem 6.3, p. 160), we have

$$\sqrt{n}(\widehat{\beta}_{e,\text{2SLS}} - \beta_e) \xrightarrow{d} N(0, (\widetilde{\boldsymbol{M}}_{Ie}^\top \widetilde{\boldsymbol{M}}_{II}^{-1}\widetilde{\boldsymbol{M}}_{Ie})^{-1}\sigma^2).$$

■

Lemma A.2: *Under Assumptions* (i)–(v), *for all* $m, m' = 1, \ldots, M$, *we have* (i) $\hat{\boldsymbol{\nu}}_m = \boldsymbol{\nu}_m + o_p(1)$; (ii) $\widehat{\boldsymbol{\Sigma}}_{mm'} = \boldsymbol{\Sigma}_{mm'} + o_p(1)$; (iii) $\widehat{\boldsymbol{\Sigma}}_{mm}^{-1} = \boldsymbol{\Sigma}_{mm}^{-1} + o_p(1)$; (iv) $\hat{\boldsymbol{u}} = \boldsymbol{u} + o_p(1)$; *(v)* $\widehat{\boldsymbol{\Psi}} = \boldsymbol{\Psi} + o_p(1)$; (v') $\widehat{\boldsymbol{\Psi}}^{-1} = \boldsymbol{\Psi}^{-1} + o_p(1)$; (vi) $\widehat{\boldsymbol{Q}} = \boldsymbol{Q} + o_p(1)$; (vi') $\hat{a} = a + o_p(1)$; (vii) $\hat{\boldsymbol{w}} = \widehat{\boldsymbol{\Psi}}^{-1}\hat{\boldsymbol{u}} = \boldsymbol{\Psi}^{-1}\boldsymbol{u} = \boldsymbol{w}_0 + o_p(1)$.

Proof: The proof follows from the law of large numbers and continuous mapping theorem. ■

Lemma A.3: *Under Assumptions* (i)–(v), *for all* $m, m' = 1, \ldots, M$, *we have*

(i) $\sqrt{n}(\hat{\boldsymbol{u}} - \boldsymbol{u}) = \frac{1}{\sqrt{n}}\sum_{i=1}^n \boldsymbol{\xi}_i\{1 + o_p(1)\}$;
(ii) $\sqrt{n}(\widehat{\boldsymbol{\Psi}} - \boldsymbol{\Psi}) = \frac{1}{\sqrt{n}}\sum_{i=1}^n \boldsymbol{\gamma}_i\{1 + o_p(1)\}$;
(iii) $\sqrt{n}(\hat{a} - a) = \frac{1}{\sqrt{n}}\sum_{i=1}^n \zeta_i\{1 + o_p(1)\}$;
(iii') $\sqrt{n}(\widehat{\boldsymbol{Q}} - \boldsymbol{Q}) = \mathbf{1}_M\mathbf{1}_M^\top\frac{1}{\sqrt{n}}\sum_{i=1}^n \zeta_i\{1 + o_p(1)\}$.

Proof: (i) Recall that $\sqrt{n}(\hat{\boldsymbol{\nu}}_m - \boldsymbol{\nu}_m) = \frac{1}{\sqrt{n}}\sum_{i=1}^n(\boldsymbol{Z}_{m,i}X_{e,i} - \boldsymbol{\nu}_m)$ and $\sqrt{n}(\widehat{\boldsymbol{\Sigma}}_{mm'} - \boldsymbol{\Sigma}_{mm'}) = \frac{1}{\sqrt{n}}\sum_{i=1}^n(\boldsymbol{Z}_{m,i}\boldsymbol{Z}_{m',i}^\top - \boldsymbol{\Sigma}_{mm'})$. Also we have $(\hat{\boldsymbol{u}} - \boldsymbol{u})$ is an $M \times 1$ vector with the mth element being $\hat{\boldsymbol{\nu}}_m^\top\widehat{\boldsymbol{\Sigma}}_{mm}^{-1}\hat{\boldsymbol{\nu}}_m - \boldsymbol{\nu}_m^\top\boldsymbol{\Sigma}_{mm}^{-1}\boldsymbol{\nu}_m$. Hence,

$$\begin{aligned}&\sqrt{n}(\hat{\boldsymbol{\nu}}_m^\top\widehat{\boldsymbol{\Sigma}}_{mm}^{-1}\hat{\boldsymbol{\nu}}_m - \boldsymbol{\nu}_m^\top\boldsymbol{\Sigma}_{mm}^{-1}\boldsymbol{\nu}_m)\\ &= \sqrt{n}(\hat{\boldsymbol{\nu}}_m - \boldsymbol{\nu}_m)^\top\widehat{\boldsymbol{\Sigma}}_{mm}^{-1}\hat{\boldsymbol{\nu}}_m + \sqrt{n}\boldsymbol{\nu}_m^\top(\widehat{\boldsymbol{\Sigma}}_{mm}^{-1}\hat{\boldsymbol{\nu}}_m - \boldsymbol{\Sigma}_{mm}^{-1}\boldsymbol{\nu}_m)\\ &= \frac{1}{\sqrt{n}}\sum_{i=1}^n 2\boldsymbol{\nu}_m^\top\boldsymbol{\Sigma}_{mm}^{-1}(\boldsymbol{Z}_{m,i}X_{e,i} - \boldsymbol{\nu}_m)\{1 + o_p(1)\}\\ &\quad - \frac{1}{\sqrt{n}}\sum_{i=1}^n \boldsymbol{\nu}_m^\top\boldsymbol{\Sigma}_{mm}^{-1}(\boldsymbol{Z}_{m,i}\boldsymbol{Z}_{m,i}^\top - \boldsymbol{\Sigma}_{mm})\boldsymbol{\Sigma}_{mm}^{-1}\boldsymbol{\nu}_m\{1 + o_p(1)\}\\ &= \frac{1}{\sqrt{n}}\sum_{i=1}^n \xi_{m,i}\{1 + o_p(1)\}.\end{aligned}$$

Then we have $\sqrt{n}(\hat{\boldsymbol{u}}-\boldsymbol{u})=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\boldsymbol{\xi}_i\{1+o_p(1)\}$.

(ii) $(\widehat{\boldsymbol{\Psi}}-\boldsymbol{\Psi})$ is an $M\times M$ matrix with the (m,m')th element being $\hat{\boldsymbol{v}}_m^\top\widehat{\boldsymbol{\Sigma}}_{mm}^{-1}\widehat{\boldsymbol{\Sigma}}_{mm'}\widehat{\boldsymbol{\Sigma}}_{m'm'}^{-1}\hat{\boldsymbol{v}}_{m'}-\boldsymbol{v}_m^\top\boldsymbol{\Sigma}_{mm}^{-1}\boldsymbol{\Sigma}_{mm'}\boldsymbol{\Sigma}_{m'm'}^{-1}\boldsymbol{v}_{m'}$. We also have

$$\begin{aligned}
&\sqrt{n}(\hat{\boldsymbol{v}}_m^\top\widehat{\boldsymbol{\Sigma}}_{mm}^{-1}\widehat{\boldsymbol{\Sigma}}_{mm'}\widehat{\boldsymbol{\Sigma}}_{m'm'}^{-1}\hat{\boldsymbol{v}}_{m'}-\boldsymbol{v}_m^\top\boldsymbol{\Sigma}_{mm}^{-1}\boldsymbol{\Sigma}_{mm'}\boldsymbol{\Sigma}_{m'm'}^{-1}\boldsymbol{v}_{m'})\\
&=\sqrt{n}(\hat{\boldsymbol{v}}_m-\boldsymbol{v}_m)^\top\widehat{\boldsymbol{\Sigma}}_{mm}^{-1}\widehat{\boldsymbol{\Sigma}}_{mm'}\widehat{\boldsymbol{\Sigma}}_{m'm'}^{-1}\hat{\boldsymbol{v}}_{m'}\\
&\quad+\sqrt{n}\boldsymbol{v}_m^\top(\widehat{\boldsymbol{\Sigma}}_{mm}^{-1}-\boldsymbol{\Sigma}_{mm}^{-1})\widehat{\boldsymbol{\Sigma}}_{mm'}\widehat{\boldsymbol{\Sigma}}_{m'm'}^{-1}\hat{\boldsymbol{v}}_{m'}\\
&\quad+\sqrt{n}\boldsymbol{v}_m^\top\boldsymbol{\Sigma}_{mm}^{-1}(\widehat{\boldsymbol{\Sigma}}_{mm'}-\boldsymbol{\Sigma}_{mm'})\widehat{\boldsymbol{\Sigma}}_{m'm'}^{-1}\hat{\boldsymbol{v}}_{m'}\\
&\quad+\sqrt{n}\boldsymbol{v}_m^\top\boldsymbol{\Sigma}_{mm}^{-1}\boldsymbol{\Sigma}_{mm'}(\widehat{\boldsymbol{\Sigma}}_{m'm'}^{-1}-\boldsymbol{\Sigma}_{m'm'}^{-1})\hat{\boldsymbol{v}}_{m'}\\
&\quad+\sqrt{n}\boldsymbol{v}_m^\top\boldsymbol{\Sigma}_{mm}^{-1}\boldsymbol{\Sigma}_{mm'}\boldsymbol{\Sigma}_{m'm'}^{-1}(\hat{\boldsymbol{v}}_{m'}-\boldsymbol{v}_{m'})\\
&=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}[(\boldsymbol{Z}_{m,i}X_{e,i}-\boldsymbol{v}_m)^\top\boldsymbol{\Sigma}_{mm}^{-1}\boldsymbol{\Sigma}_{mm'}\boldsymbol{\Sigma}_{m'm'}^{-1}\boldsymbol{v}_{m'}\\
&\quad-\boldsymbol{v}_m^\top\boldsymbol{\Sigma}_{mm}^{-1}(\boldsymbol{Z}_{m,i}\boldsymbol{Z}_{m,i}^\top-\boldsymbol{\Sigma}_{mm})\boldsymbol{\Sigma}_{mm}^{-1}\boldsymbol{\Sigma}_{mm'}\boldsymbol{\Sigma}_{m'm'}^{-1}\boldsymbol{v}_{m'}\\
&\quad+\boldsymbol{v}_m^\top\boldsymbol{\Sigma}_{mm}^{-1}(\boldsymbol{Z}_{m,i}\boldsymbol{Z}_{m',i}^\top-\boldsymbol{\Sigma}_{mm})\boldsymbol{\Sigma}_{mm'}^{-1}\boldsymbol{v}_{m'}\\
&\quad-\boldsymbol{v}_m^\top\boldsymbol{\Sigma}_{mm}^{-1}\boldsymbol{\Sigma}_{mm'}\boldsymbol{\Sigma}_{m'm'}^{-1}(\boldsymbol{Z}_{m',i}\boldsymbol{Z}_{m',i}^\top-\boldsymbol{\Sigma}_{m'm'})\boldsymbol{\Sigma}_{m'm'}^{-1}\boldsymbol{v}_{m'}\\
&\quad+\boldsymbol{v}_m^\top\boldsymbol{\Sigma}_{mm}^{-1}\boldsymbol{\Sigma}_{mm'}\boldsymbol{\Sigma}_{m'm'}^{-1}(\boldsymbol{Z}_{m',i}X_{e,i}-\boldsymbol{v}_{m'})]\times\{1+o_p(1)\}\\
&=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\gamma_{mm',i}\{1+o_p(1)\}.
\end{aligned}$$

Then we have $\sqrt{n}(\widehat{\boldsymbol{\Psi}}-\boldsymbol{\Psi})=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\boldsymbol{\gamma}_i\{1+o_p(1)\}$.

(iii) First we have

$$\begin{aligned}
\sqrt{n}(\hat{a}-a)&=\sqrt{n}(\widehat{\boldsymbol{M}}_{oe}^\top\widehat{\boldsymbol{M}}_{oo}^{-1}\widehat{\boldsymbol{M}}_{oe}-\boldsymbol{M}_{oe}^\top\boldsymbol{M}_{oo}^{-1}\boldsymbol{M}_{oe})\\
&=\sqrt{n}(\widehat{\boldsymbol{M}}_{oe}-\boldsymbol{M}_{oe})^\top\widehat{\boldsymbol{M}}_{oo}^{-1}\widehat{\boldsymbol{M}}_{oe}+\sqrt{n}\boldsymbol{M}_{oe}^\top(\widehat{\boldsymbol{M}}_{oo}^{-1}-\boldsymbol{M}_{oo}^{-1})\widehat{\boldsymbol{M}}_{oe}\\
&\quad+\boldsymbol{M}_{oe}^\top\boldsymbol{M}_{oo}^{-1}\sqrt{n}(\widehat{\boldsymbol{M}}_{oe}-\boldsymbol{M}_{oe})\\
&=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}[2\boldsymbol{M}_{oe}^\top\boldsymbol{M}_{oo}^{-1}(\boldsymbol{X}_{o,i}X_{e,i}-\boldsymbol{M}_{oe})\\
&\quad-\boldsymbol{M}_{oe}^\top\boldsymbol{M}_{oo}^{-1}(\boldsymbol{X}_{o,i}\boldsymbol{X}_{o,i}^\top-\boldsymbol{M}_{oo})\boldsymbol{M}_{oo}^{-1}\boldsymbol{M}_{oe}]\times\{1+o_p(1)\}\\
&=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\zeta_i\{1+o_p(1)\},
\end{aligned}$$

and hence, $\sqrt{n}(\widehat{\boldsymbol{Q}}-\boldsymbol{Q})=(\sqrt{n}(\hat{a}-a))_{M\times M}=\mathbf{1}_M\mathbf{1}_M^\top(\sqrt{n}(\hat{a}-a)=\mathbf{1}_M\mathbf{1}_M^\top\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\zeta_i\{1+o_p(1)\}$.

(iii') Note that $\sqrt{n}(\widehat{\boldsymbol{Q}}-\boldsymbol{Q})=(\sqrt{n}(\hat{a}-a))_{M\times M}=\mathbf{1}_M\mathbf{1}_M^\top\sqrt{n}(\hat{a}-a)$. Thus, (iii') follows from (iii). ■

Lemma A.4: *Under Assumptions* (i)–(v)*, we have*

(i) $\sqrt{n}(\hat{\boldsymbol{w}}^\top\widehat{\boldsymbol{\Psi}}\hat{\boldsymbol{w}}-\boldsymbol{w}_0^\top\boldsymbol{\Psi}\boldsymbol{w}_0)=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\boldsymbol{w}_0^\top(2\boldsymbol{\xi}_i-\boldsymbol{\gamma}_i\boldsymbol{w}_0)\{1+o_p(1)\}$.

(ii) $\sqrt{n}(\widehat{\boldsymbol{w}}^\top\widehat{\boldsymbol{Q}}\hat{\boldsymbol{w}}-\boldsymbol{w}_0^\top\boldsymbol{Q}\boldsymbol{w}_0)=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}[2a\boldsymbol{w}_0^\top\mathbf{1}_M\mathbf{1}_M^\top\boldsymbol{\Psi}^{-1}(\boldsymbol{\xi}_i-\boldsymbol{\gamma}_i\boldsymbol{w}_0)+\boldsymbol{w}_0^\top\mathbf{1}_M\mathbf{1}_M^\top\boldsymbol{w}_0\zeta_i]\times\{1+o_p(1)\}$.

(iii) $\sqrt{n}(\hat{\boldsymbol{w}}^\top\mathbf{1}_M\hat{a}-\boldsymbol{w}_0^\top\mathbf{1}_Ma)=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}[a\mathbf{1}_M^\top\boldsymbol{\Psi}^{-1}(\boldsymbol{\xi}_i-\boldsymbol{\gamma}_i\boldsymbol{w}_0)+\boldsymbol{w}_0^\top\mathbf{1}_M\zeta_i]\times\{1+o_p(1)\}$.

Proof: (i) It follows that

$$\begin{aligned}
&\sqrt{n}(\hat{\boldsymbol{w}}^\top \widehat{\boldsymbol{\Psi}} \hat{\boldsymbol{w}} - \boldsymbol{w}_0^\top \boldsymbol{\Psi} \boldsymbol{w}_0) \\
&\quad = \sqrt{n}(\hat{\boldsymbol{w}} - \boldsymbol{w}_0)^\top \widehat{\boldsymbol{\Psi}} \hat{\boldsymbol{w}} + \sqrt{n} \boldsymbol{w}_0^\top (\widehat{\boldsymbol{\Psi}} \hat{\boldsymbol{w}} - \boldsymbol{\Psi} \boldsymbol{w}_0) \\
&\quad = 2\boldsymbol{w}_0^\top \boldsymbol{\Psi} \sqrt{n}(\hat{\boldsymbol{w}} - \boldsymbol{w}_0)\{1 + o_p(1)\} + \sqrt{n} \boldsymbol{w}_0^\top (\hat{\boldsymbol{\Psi}} - \boldsymbol{\Psi}) \boldsymbol{w}_0 \{1 + o_p(1)\} \\
&\quad = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (2\boldsymbol{w}_0^\top \boldsymbol{\Psi} \boldsymbol{\varphi}_i + \boldsymbol{w}_0^\top \boldsymbol{\gamma}_i \boldsymbol{w}_0)\{1 + o_p(1)\} \\
&\quad = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \boldsymbol{w}_0^\top (2\boldsymbol{\xi}_i - \boldsymbol{\gamma}_i \boldsymbol{w}_0)\{1 + o_p(1)\},
\end{aligned}$$

where the third equality is obtained by Theorem 3.1 (ii) and Lemma A.3 (ii).

(ii) It follows that

$$\begin{aligned}
&\sqrt{n}(\hat{\boldsymbol{w}}^\top \widehat{\boldsymbol{Q}} \hat{\boldsymbol{w}} - \boldsymbol{w}_0^\top \boldsymbol{Q} \boldsymbol{w}_0) \\
&\quad = \sqrt{n}(\hat{\boldsymbol{w}} - \boldsymbol{w}_0)^\top \widehat{\boldsymbol{Q}} \hat{\boldsymbol{w}} + \sqrt{n} \boldsymbol{w}_0^\top (\widehat{\boldsymbol{Q}} \hat{\boldsymbol{w}} - \boldsymbol{Q} \boldsymbol{w}_0) \\
&\quad = 2\boldsymbol{w}_0^\top \boldsymbol{Q} \sqrt{n}(\hat{\boldsymbol{w}} - \boldsymbol{w}_0)\{1 + o_p(1)\} + \sqrt{n} \boldsymbol{w}_0^\top (\widehat{\boldsymbol{Q}} - \boldsymbol{Q}) \boldsymbol{w}_0 \{1 + o_p(1)\} \\
&\quad = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (2\boldsymbol{w}_0^\top \boldsymbol{Q} \boldsymbol{\varphi}_i + \boldsymbol{w}_0^\top \mathbf{1}_M \mathbf{1}_M^\top \zeta_i \boldsymbol{w}_0)\{1 + o_p(1)\} \\
&\quad = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} [2a\boldsymbol{w}_0^\top \mathbf{1}_M \mathbf{1}_M^\top \boldsymbol{\Psi}^{-1} (\boldsymbol{\xi}_i - \boldsymbol{\gamma}_i \boldsymbol{w}_0) + \boldsymbol{w}_0^\top \mathbf{1}_M \mathbf{1}_M^\top \boldsymbol{w}_0 \zeta_i] \times \{1 + o_p(1)\},
\end{aligned}$$

where the third equality is obtained by Theorem 3.1 (ii) and Lemma A.3 (iii).

(iii) It follows that

$$\begin{aligned}
&\sqrt{n}(\hat{\boldsymbol{w}}^\top \mathbf{1}_M \hat{a} - \boldsymbol{w}_0^\top \mathbf{1}_M a) \\
&\quad = \sqrt{n}(\hat{\boldsymbol{w}} - \boldsymbol{w}_0)^\top \mathbf{1}_M \hat{a} + \boldsymbol{w}_0^\top \mathbf{1}_M (\hat{a} - a) \\
&\quad = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} [a\mathbf{1}_M^\top \boldsymbol{\Psi}^{-1} (\boldsymbol{\xi}_i - \boldsymbol{\gamma}_i \boldsymbol{w}_0) + \boldsymbol{w}_0^\top \mathbf{1}_M \zeta_i] \times \{1 + o_p(1)\},
\end{aligned}$$

where the second equality is obtained by Theorem 3.1 (ii) and Lemma A.3 (iii).

■

Appendix 3. Proofs of Theorems 3.1–4.3

Proof: (i) $\hat{\boldsymbol{w}} \xrightarrow{p} \boldsymbol{w}_0$ can be easily derived by Lemma A.2 (iv) and (v).

(ii) To show the asymptotic normality of $\hat{\boldsymbol{w}}$, we note that

$$\begin{aligned}
\sqrt{n}(\hat{\boldsymbol{w}} - \boldsymbol{w}_0) \quad &= \sqrt{n}(\widehat{\boldsymbol{\Psi}}^{-1} \hat{\boldsymbol{u}} - \boldsymbol{\Psi}^{-1} \boldsymbol{u}) \\
&= -\boldsymbol{\Psi}^{-1} \sqrt{n}(\widehat{\boldsymbol{\Psi}}^{-1} - \boldsymbol{\Psi}^{-1}) \boldsymbol{\Psi}^{-1} \{1 + o_p(1)\} \boldsymbol{u} \{1 + o_p(1)\} + \boldsymbol{\Psi}^{-1} \sqrt{n}(\hat{\boldsymbol{u}} - \boldsymbol{u}) \\
&= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \boldsymbol{\Psi}^{-1} (\boldsymbol{\xi}_i - \boldsymbol{\gamma}_i \boldsymbol{w}_0) + o_p(1) \\
&= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \boldsymbol{\varphi}_i + o_p(1) \\
&\xrightarrow{d} N(0, \boldsymbol{\Xi}),
\end{aligned}$$

where the second equality is obtained by Lemma A.2, the third equality is obtained by Lemma A.3 and the last line is by multivariate Linderberg-lévy central limit theorem (B. E. Hansen, 2022, Theorem 6.3, p. 160) under Assumption (v). ■

Proof: (i) We firstly prove $\widehat{\beta}_{e,\mathrm{MA}}(\hat{\boldsymbol{w}}) = \beta_e + \mathbf{bias} + o_p(1)$. Note that $\widehat{\beta}_{e,\mathrm{MA}}(\hat{\boldsymbol{w}}) = \beta_e + [\widehat{\boldsymbol{X}}_e(\hat{\boldsymbol{w}})^\top(\boldsymbol{I}_n - \boldsymbol{P}_1)\widehat{\boldsymbol{X}}_e(\hat{\boldsymbol{w}})]^{-1}\widehat{\boldsymbol{X}}_e(\hat{\boldsymbol{w}})^\top(\boldsymbol{I}_n - \boldsymbol{P}_1)(\boldsymbol{Y} - \widehat{\boldsymbol{X}}_e(\hat{\boldsymbol{w}})\beta_e)$, so

$$\widehat{\beta}_{e,\mathrm{MA}}(\hat{\boldsymbol{w}}) - \beta_e = \Delta_1^{-1}\Delta_2, \tag{A2}$$

where $\Delta_1 = \widehat{\boldsymbol{X}}_e(\hat{\boldsymbol{w}})^\top(\boldsymbol{I}_n - \boldsymbol{P}_1)\widehat{\boldsymbol{X}}_e(\hat{\boldsymbol{w}})$, $\Delta_2 = \widehat{\boldsymbol{X}}_e(\hat{\boldsymbol{w}})^\top(\boldsymbol{I}_n - \boldsymbol{P}_1)(\boldsymbol{Y} - \widehat{\boldsymbol{X}}_e(\hat{\boldsymbol{w}})\beta_e)$. On the one hand,

$$\begin{aligned}
\frac{1}{n}\Delta_1 &= \frac{1}{n}\widehat{\boldsymbol{X}}_e(\hat{\boldsymbol{w}})^\top(\boldsymbol{I}_n - \boldsymbol{P}_1)\widehat{\boldsymbol{X}}_e(\hat{\boldsymbol{w}}) \\
&= \hat{\boldsymbol{u}}^\top\widehat{\boldsymbol{\Psi}}^{-1}\hat{\boldsymbol{u}} - \hat{\boldsymbol{u}}^\top\widehat{\boldsymbol{\Psi}}^{-1}\widehat{\boldsymbol{Q}}\widehat{\boldsymbol{\Psi}}^{-1}\hat{\boldsymbol{u}} \\
&= \boldsymbol{w}_0^\top\boldsymbol{\Psi}\boldsymbol{w}_0 - a(\boldsymbol{w}_0^\top\mathbf{1}_M)^2 + o_p(1) \\
&= s_3 + o_p(1),
\end{aligned}$$

where the second equation is obtained by Lemma A.2.

On the other hand

$$\begin{aligned}
\frac{1}{n}\Delta_2 &= \frac{1}{n}\hat{\boldsymbol{X}}_e(\hat{\boldsymbol{w}})^\top(\boldsymbol{I}_n - \boldsymbol{P}_1)[\boldsymbol{Y} - \hat{\boldsymbol{X}}_e(\hat{\boldsymbol{w}})\beta_e] \\
&= \frac{1}{n}\hat{\boldsymbol{X}}_e(\hat{\boldsymbol{w}})^\top(\boldsymbol{I}_n - \boldsymbol{P}_1)\boldsymbol{X}_e\beta_e + \frac{1}{n}\hat{\boldsymbol{X}}_e(\hat{\boldsymbol{w}})(\boldsymbol{I}_n - \boldsymbol{P}_1)\boldsymbol{\varepsilon} - \frac{1}{n}\Delta_1\beta_e \\
&\triangleq I_1 + I_2 - I_3
\end{aligned} \tag{A3}$$

where the first term on the right-hand side of Equation (A3) is

$$\begin{aligned}
I_1 &\triangleq \frac{1}{n}\widehat{\boldsymbol{X}}_e(\hat{\boldsymbol{w}})^\top(\boldsymbol{I}_n - \boldsymbol{P}_1)\boldsymbol{X}_e\beta_e \\
&= \hat{\boldsymbol{u}}^\top\widehat{\boldsymbol{\Psi}}^{-1}\hat{\boldsymbol{u}}\beta_e - \hat{\boldsymbol{u}}^\top\widehat{\boldsymbol{\Psi}}^{-1}\mathbf{1}_M\hat{a}\beta_e \\
&= \boldsymbol{w}_0^\top\boldsymbol{\Psi}\boldsymbol{w}_0\beta_e - \boldsymbol{w}_0^\top\mathbf{1}_M a\beta_e + o_p(1) = s_1 + o_p(1),
\end{aligned}$$

the third term on the right-hand side of Equation (A3) is

$$I_3 \triangleq \frac{1}{n}\Delta_1\beta_e = \left[\boldsymbol{w}_0^\top\boldsymbol{\Psi}\boldsymbol{w}_0 - a(\boldsymbol{w}_0^\top\mathbf{1}_M)^2\right]\beta_e + o_p(1) = s_2 + o_p(1),$$

and the second term on the right-hand side of Equation (A3) is

$$I_2 \triangleq \frac{1}{n}\widehat{\boldsymbol{X}}_e(\hat{\boldsymbol{w}})(\boldsymbol{I}_n - \boldsymbol{P}_1)\boldsymbol{\varepsilon} = \hat{\boldsymbol{u}}^\top\widehat{\boldsymbol{\Psi}}^{-1}\frac{1}{n}\widetilde{\mathcal{X}}_e^\top\boldsymbol{H}(\boldsymbol{I}_n - \boldsymbol{P}_1)\boldsymbol{\varepsilon} \xrightarrow{p} 0,$$

which means $I_2 = o_p(1)$.

Then we have

$$\begin{aligned}
\frac{1}{n}\Delta_2 &= I_1 + I_2 - I_3 \\
&= \boldsymbol{w}_0^\top\boldsymbol{\Psi}\boldsymbol{w}_0\beta_e - \boldsymbol{w}_0^\top\mathbf{1}_M a\beta_e - \left[\boldsymbol{w}_0^\top\boldsymbol{\Psi}\boldsymbol{w}_0 - a(\boldsymbol{w}_0^\top\mathbf{1}_M)^2\right]\beta_e + o_p(1) \\
&= a\boldsymbol{w}_0^\top\mathbf{1}_M(\boldsymbol{w}_0^\top\mathbf{1}_M - 1)\beta_e + o_p(1).
\end{aligned}$$

Replacing Δ_1 and Δ_2 in Equation (A2), we have

$$\begin{aligned}
\widehat{\beta}_{e,\mathrm{MA}}(\hat{\boldsymbol{w}}) &= \beta_e + \left(\frac{1}{n}\Delta_1\right)^{-1}\left(\frac{1}{n}\Delta_2\right) \\
&= \beta_e + [\boldsymbol{w}_0^\top\boldsymbol{\Psi}\boldsymbol{w}_0 - a(\boldsymbol{w}_0^\top\mathbf{1}_M)^2]^{-1}a\boldsymbol{w}_0^\top\mathbf{1}_M(\boldsymbol{w}_0^\top\mathbf{1}_M - 1)\beta_e + o_p(1)
\end{aligned}$$

$$\triangleq \beta_e + \mathbf{bias} + o_p(1),$$

in which $\mathbf{bias} = s_3^{-1}(s_1 - s_2)$.

(ii) Secondly, we verify the asymptotic property $\sqrt{n}[\widehat{\beta}_{e,\mathrm{MA}}(\hat{\boldsymbol{w}}) - \beta_e - \mathbf{bias}] \xrightarrow{d} N(0, v)$.

From the arguments previously stated, we observe that

$$\begin{aligned}\sqrt{n}\left[\widehat{\beta}_{e,\mathrm{MA}}(\hat{\boldsymbol{w}}) - \beta_e - \mathbf{bias}\right] &= \sqrt{n}\left[\left(\frac{1}{n}\Delta_1\right)^{-1}\left(\frac{1}{n}\Delta_2\right) - \mathbf{bias}\right] \\ &= \sqrt{n}\left[(\frac{1}{n}\Delta_1)^{-1} - s_3^{-1}\right](I_1 + I_2 - I_3) \\ &\quad + \sqrt{n}s_3^{-1}(I_1 + I_2 - I_3 - s_1 + s_2) \\ &\triangleq II_1 + II_2.\end{aligned}$$

The first term on the right-hand side is

$$\begin{aligned}II_1 &= \sqrt{n}\left[\left(\frac{1}{n}\Delta_1\right)^{-1} - s_3^{-1}\right](I_1 + I_2 - I_3) \\ &= -s_3^{-1}\sqrt{n}\left(\frac{1}{n}\Delta_1 - s_3\right)s_3^{-1}(s_1 + o_p(1) - s_2)\{1 + o_p(1)\} \\ &= -\frac{\mathbf{bias}}{s_3}\sqrt{n}\left(\frac{1}{n}\Delta_1 - s_3\right)\{1 + o_p(1)\},\end{aligned}$$

where the second equality uses the result $\frac{1}{n}\Delta_1 = s_3 + o_p(1)$.

We also have

$$\begin{aligned}\sqrt{n}\left(\frac{1}{n}\Delta_1 - s_3\right) &= \sqrt{n}\left(\hat{\boldsymbol{u}}^\top\widehat{\boldsymbol{\Psi}}^{-1}\hat{\boldsymbol{u}} - \hat{\boldsymbol{u}}^\top\widehat{\boldsymbol{\Psi}}^{-1}\widehat{\boldsymbol{Q}}\widehat{\boldsymbol{\Psi}}^{-1}\hat{\boldsymbol{u}} - s_3\right) \\ &= \sqrt{n}(\hat{\boldsymbol{w}}^\top\widehat{\boldsymbol{\Psi}}\hat{\boldsymbol{w}} - \boldsymbol{w}_0^\top\boldsymbol{\Psi}\boldsymbol{w}_0) - \sqrt{n}(\hat{\boldsymbol{w}}^\top\widehat{\boldsymbol{Q}}\hat{\boldsymbol{w}} - \boldsymbol{w}_0^\top\boldsymbol{Q}\boldsymbol{w}_0) \\ &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}[2\boldsymbol{w}_0^\top(\boldsymbol{I}_{M\times M} - a\mathbf{1}_M\mathbf{1}_M^\top\boldsymbol{\Psi}^{-1})\boldsymbol{\xi}_i \\ &\quad + \boldsymbol{w}_0^\top(2a\mathbf{1}_M\mathbf{1}_M^\top\boldsymbol{\Psi}^{-1} - \boldsymbol{I}_{M\times M})\boldsymbol{\gamma}_i\boldsymbol{w}_0 \\ &\quad - \boldsymbol{w}_0^\top\mathbf{1}_M\mathbf{1}_M^\top\boldsymbol{w}_0\zeta_i] \times \{1 + o_p(1)\} \\ &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\kappa_i\{1 + o_p(1)\},\end{aligned} \tag{A4}$$

where the third equality is obtained by Lemma A.4 (i) and (ii).

Hence,

$$\begin{aligned}II_1 &= -\frac{\mathbf{bias}}{s_3}\sqrt{n}(\frac{1}{n}\Delta_1 - s_3)\{1 + o_p(1)\} \\ &= -s_3^{-1}\mathbf{bias}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\kappa_i\{1 + o_p(1)\}.\end{aligned}$$

Moreover,

$$\begin{aligned}II_2 &= \sqrt{n}s_3^{-1}(I_1 + I_2 - I_3 - s_1 + s_2) \\ &= s_3^{-1}\sqrt{n}(I_1 - s_1) + s_3^{-1}\sqrt{n}I_2 - s_3^{-1}\sqrt{n}(I_3 - s_2) \\ &\triangleq II_2^{(1)} + II_2^{(2)} - II_2^3,\end{aligned}$$

where the first term on the right-hand side

$$\begin{aligned}
II_2^{(1)} &= s_3^{-1}\sqrt{n}(I_1 - s_1) \\
&= s_3^{-1}\sqrt{n}(\hat{\boldsymbol{w}}^\top \widehat{\boldsymbol{\Psi}}\hat{\boldsymbol{w}} - \boldsymbol{w}_0^\top \boldsymbol{\Psi}\boldsymbol{w}_0)\beta_e - s_3^{-1}\sqrt{n}(\hat{\boldsymbol{w}}^\top \mathbf{1}_M\hat{a} - \boldsymbol{w}_0^\top \mathbf{1}_M a) \\
&= \frac{s_3^{-1}\beta_e}{\sqrt{n}}\sum_{i=1}^{n}[\boldsymbol{w}_0^\top(2\boldsymbol{\xi}_i - \boldsymbol{\gamma}_i\boldsymbol{w}_0) - a\mathbf{1}_M^\top\boldsymbol{\Psi}^{-1}(\boldsymbol{\xi}_i - \boldsymbol{\gamma}_i\boldsymbol{w}_0) - \boldsymbol{w}_0^\top\mathbf{1}_M\zeta_i]\{1 + o_p(1)\}.
\end{aligned}$$

Next, we consider $II_2^{(3)}$,

$$\begin{aligned}
II_2^{(3)} &= \sqrt{n}s_3^{-1}(I_3 - s_2) \\
&= \sqrt{n}s_3^{-1}\beta_e\left[(\hat{\boldsymbol{w}}^\top\widehat{\boldsymbol{\Psi}}\hat{\boldsymbol{w}} - \boldsymbol{w}_0^\top\boldsymbol{\Psi}\boldsymbol{w}_0) - (\hat{\boldsymbol{w}}^\top\widehat{\boldsymbol{Q}}\hat{\boldsymbol{w}} - \boldsymbol{w}_0^\top\boldsymbol{Q}\boldsymbol{w}_0)\right] \\
&= s_3^{-1}\beta_e\frac{1}{\sqrt{n}}\sum_{i=1}^{n}[2\boldsymbol{w}_0^\top(\boldsymbol{I}_{M\times M} - a\mathbf{1}_M\mathbf{1}_M^\top\boldsymbol{\Psi}^{-1})\boldsymbol{\xi}_i \\
&\quad + \boldsymbol{w}_0^\top(2a\mathbf{1}_M\mathbf{1}_M^\top\boldsymbol{\Psi}^{-1} - \boldsymbol{I}_{M\times M})\boldsymbol{\gamma}_i\boldsymbol{w}_0 - \boldsymbol{w}_0^\top\mathbf{1}_M\mathbf{1}_M^\top\boldsymbol{w}_0\zeta_i] \times \{1 + o_p(1)\}.
\end{aligned}$$

Last, we consider

$$\begin{aligned}
II_2^{(2)} &= \sqrt{n}s_3^{-1}I_2 \\
&= \sqrt{n}s_3^{-1}\hat{\boldsymbol{w}}^\top\frac{1}{n}\widetilde{\mathcal{X}}_e^\top\boldsymbol{H}(\boldsymbol{I}_n - \boldsymbol{P}_1)\boldsymbol{\varepsilon} \\
&= s_3^{-1}\boldsymbol{w}_0^\top\begin{pmatrix}
\boldsymbol{v}_1^\top\boldsymbol{\Sigma}_{11}^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\boldsymbol{Z}_{1,i}\varepsilon_i - \boldsymbol{M}_{oe}^\top\boldsymbol{M}_{oo}^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\boldsymbol{X}_{o,i}\varepsilon_i \\
\vdots \\
\boldsymbol{v}_M^\top\boldsymbol{\Sigma}_{MM}^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\boldsymbol{Z}_{M,i}\varepsilon_i - \boldsymbol{M}_{oe}^\top\boldsymbol{M}_{oo}^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\boldsymbol{X}_{o,i}\varepsilon_i
\end{pmatrix} \times \{1 + o_p(1)\} \\
&= s_3^{-1}\boldsymbol{w}_0^\top\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\boldsymbol{\phi}_i \times \{1 + o_p(1)\}.
\end{aligned}$$

Therefore,

$$\begin{aligned}
II_2 &= II_2^{(1)} + II_2^{(2)} - II_2^3 \\
&= \frac{s_3^{-1}\beta_e}{\sqrt{n}}\sum_{i=1}^{n}[(2\boldsymbol{w}_0^\top a\mathbf{1}_M\mathbf{1}_M^\top\boldsymbol{\Psi}^{-1} - a\mathbf{1}_M^\top\boldsymbol{\Psi}^{-1})\boldsymbol{\xi}_i \\
&\quad + (a\mathbf{1}_M^\top\boldsymbol{\Psi}^{-1} - 2\boldsymbol{w}_0^\top a\mathbf{1}_M\mathbf{1}_M^\top\boldsymbol{\Psi}^{-1})\boldsymbol{\gamma}_i\boldsymbol{w}_0 \\
&\quad + (\boldsymbol{w}_0^\top\mathbf{1}_M)(\boldsymbol{w}_0^\top\mathbf{1}_M - 1)\zeta_i \\
&\quad + \beta_e^{-1}\boldsymbol{w}_0^\top\boldsymbol{\phi}_i] \times \{1 + o_p(1)\}.
\end{aligned}$$

In summary,

$$\begin{aligned}
&\sqrt{n}\left[\widehat{\beta}_{e,\mathrm{MA}}(\hat{\boldsymbol{w}}) - \beta_e - \mathbf{bias}\right] \\
&= II_1 + II_2 \\
&= \frac{s_3^{-1}}{\sqrt{n}}\sum_{i=1}^{n}\{[(\beta_e + \mathbf{bias})2a\boldsymbol{w}_0^\top\mathbf{1}_M\mathbf{1}_M^\top\boldsymbol{\Psi}^{-1} - a\beta_e\mathbf{1}_M^\top\boldsymbol{\Psi}^{-1} - 2\mathbf{bias}\boldsymbol{w}_0^\top]\boldsymbol{\xi}_i \\
&\quad + [\beta_e a\mathbf{1}_M^\top\boldsymbol{\Psi}^{-1} - 2a\boldsymbol{w}_0^\top\mathbf{1}_M\mathbf{1}_M^\top\boldsymbol{\Psi}^{-1}(\beta_e + \mathbf{bias}) + \mathbf{bias}\boldsymbol{w}_0^\top]\boldsymbol{\gamma}_i\boldsymbol{w}_0 \\
&\quad + [(\beta_e + \mathbf{bias})(\boldsymbol{w}_0^\top\mathbf{1}_M)^2 - \beta_e\boldsymbol{w}_0^\top\mathbf{1}_M]\zeta_i + \boldsymbol{w}_0^\top\boldsymbol{\phi}_i\} \times \{1 + o_p(1)\}
\end{aligned}$$

$$= \frac{s_3^{-1}}{\sqrt{n}} \sum_{i=1}^{n} \widetilde{\varphi}_i \{1 + o_p(1)\}$$

$$\xrightarrow{d} N\left(0, \operatorname{cov}(\widetilde{\varphi}_i, \widetilde{\varphi}_i) s_3^{-2}\right),$$

where the last line follows from multivariate Linderberg-lévy central limit theorem (B. E. Hansen, 2022, Theorem 6.3, p. 160) under Assumption (v). ■

Proof: We have

$$\widehat{\beta}_e = (1 + \hat{c}_{\text{bias}})^{-1} \widehat{\beta}_{e,\text{MA}}(\hat{\boldsymbol{w}}) = \frac{1 + c_{\text{bias}}}{1 + \hat{c}_{\text{bias}}} \beta_e + o_p(1) = \beta_e + o_p(1),$$

where $\hat{c}_{\text{bias}} = \hat{s}_3^{-1}(\hat{\boldsymbol{w}}^\top \widehat{\boldsymbol{Q}} \hat{\boldsymbol{w}} - \hat{a} \hat{\boldsymbol{w}}^\top \mathbf{1}_M)$. Next we discuss the asymptotic distribution of $\widehat{\beta}_e$,

$$\begin{aligned}
\sqrt{n}(\widehat{\beta}_e^{(\text{debias})}(\hat{\boldsymbol{w}}) - \beta_e) &= \frac{1}{1 + \hat{c}_{\text{bias}}} \left[\sqrt{n}(\widehat{\beta}_{e,\text{MA}}(\hat{\boldsymbol{w}}) - \beta_e - \mathbf{bias}) + \sqrt{n}(c_{\text{bias}} - \hat{c}_{\text{bias}})\beta_e\right] \\
&= \frac{1}{1 + c_{\text{bias}}} \frac{s_3^{-1}}{\sqrt{n}} \sum_{i=1}^{n} \widetilde{\varphi}_i \{1 + o_p(1)\} - \frac{\beta_e}{1 + c_{\text{bias}}} \sqrt{n}(\hat{c}_{\text{bias}} - c_{\text{bias}}).
\end{aligned}$$

It can be shown that

$$\begin{aligned}
\sqrt{n}(\hat{c}_{\text{bias}} - c_{\text{bias}}) &= \sqrt{n}[\hat{s}_3^{-1}(\hat{\boldsymbol{w}}^\top \widehat{\boldsymbol{Q}} \hat{\boldsymbol{w}} - \hat{a} \hat{\boldsymbol{w}}^\top \mathbf{1}_M) - s_3^{-1}(\boldsymbol{w}_0^\top \boldsymbol{Q} \boldsymbol{w}_0 - a \boldsymbol{w}_0^\top \mathbf{1}_M)] \\
&= -\sqrt{n} s_3^{-1}(\hat{s}_3 - s_3) \hat{s}_3^{-1} \hat{\boldsymbol{w}}^\top \widehat{\boldsymbol{Q}} \hat{\boldsymbol{w}} + \sqrt{n} s_3^{-1}(\hat{\boldsymbol{w}}^\top \widehat{\boldsymbol{Q}} \hat{\boldsymbol{w}} - \boldsymbol{w}_0^\top \boldsymbol{Q} \boldsymbol{w}_0) \\
&\quad - [-\sqrt{n} s_3^{-1}(\hat{s}_3 - s_3) \hat{s}_3^{-1} \hat{a} \hat{\boldsymbol{w}}^\top \mathbf{1}_M + \sqrt{n} s_3^{-1}(\hat{a} \hat{\boldsymbol{w}}^\top \mathbf{1}_M - a \boldsymbol{w}_0^\top \mathbf{1}_M)] \\
&= \frac{s_3^{-1}}{\sqrt{n}} \sum_{i=1}^{n} [(\boldsymbol{w}_0^\top \mathbf{1}_M)(\boldsymbol{w}_0^\top \mathbf{1}_M - 1)\zeta_i \\
&\quad + a(2\boldsymbol{w}_0^\top \mathbf{1}_M - 1) \mathbf{1}_M \boldsymbol{\Psi}^{-1}(\boldsymbol{\xi}_i - \boldsymbol{\gamma}_i \boldsymbol{w}_0) \\
&\quad - s_3^{-1} a \boldsymbol{w}_0^\top \mathbf{1}_M (\boldsymbol{w}_0^\top \mathbf{1}_M - 1)\kappa_i] \times \{1 + o_p(1)\},
\end{aligned}$$

where the third equation is obtained by Equation (A4) and Lemma A.4 (ii) and (iii). Simplifying the equation, we have

$$\begin{aligned}
\sqrt{n}(\widehat{\beta}_e^{(\text{debias})}(\hat{\boldsymbol{w}}) - \beta_e) &= \frac{1}{1 + c_{\text{bias}}} \frac{s_3^{-1}}{\sqrt{n}} \sum_{i=1}^{n} \widetilde{\varphi}_i \{1 + o_p(1)\} - \frac{\beta_e}{1 + c_{\text{bias}}} \sqrt{n}(\hat{c}_{\text{bias}} - c_{\text{bias}}) \\
&= \frac{s_3^{-1}}{1 + c_{\text{bias}}} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \boldsymbol{w}_0^\top \boldsymbol{\phi}_i + o_p(1) \\
&= \frac{1}{\boldsymbol{u}^\top \boldsymbol{\Psi}^{-1}(\boldsymbol{u} - a\mathbf{1}_M)} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \boldsymbol{w}_0^\top \boldsymbol{\phi}_i + o_p(1).
\end{aligned}$$

Now we consider $\operatorname{cov}(\boldsymbol{\phi}_i, \boldsymbol{\phi}_i)$. Note that $\boldsymbol{\phi}_i = (\phi_{1,i}, \ldots, \phi_{M,i})^\top$, where $\phi_{m,i} = (\boldsymbol{v}_m^\top \boldsymbol{\Sigma}_{mm}^{-1} \boldsymbol{Z}_{m,i} - \boldsymbol{M}_{oe}^\top \boldsymbol{M}_{oo}^{-1} \boldsymbol{X}_{o,i})\varepsilon_i$, $m = 1, \ldots, M$. Then we have

$$\operatorname{cov}(\boldsymbol{\phi}_i, \boldsymbol{\phi}_i) = \left(\boldsymbol{\Psi} - \boldsymbol{d}\mathbf{1}_M^\top - \mathbf{1}_M \boldsymbol{d} + a\mathbf{1}_M \mathbf{1}_M^\top\right) \sigma^2.$$

By Slutsky's theorem and multivariate Linderberg-lévy central limit theorem (B. E. Hansen, 2022, Theorem 6.3, p. 160), under Assumption (v), we have

$$\sqrt{n}(\widehat{\beta}_e^{(\text{debias})}(\hat{\boldsymbol{w}}) - \beta_e) = \frac{1}{\boldsymbol{u}^\top \boldsymbol{\Psi}^{-1}(\boldsymbol{u} - a\mathbf{1}_M)} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \boldsymbol{w}_0^\top \boldsymbol{\phi}_i + o_p(1)$$

$$\xrightarrow{d} N(0, \nu),$$

where $\nu = \frac{\boldsymbol{u}^\top \boldsymbol{\Psi}^{-1}(\boldsymbol{\Psi} - \boldsymbol{d}\mathbf{1}_M^\top - \mathbf{1}_M \boldsymbol{d}^\top + a\mathbf{1}_M\mathbf{1}_M^\top)\boldsymbol{\Psi}^{-1}\boldsymbol{u}}{[\boldsymbol{u}^\top \boldsymbol{\Psi}^{-1}(\boldsymbol{u} - a\mathbf{1}_M)]^2}\sigma^2$. ■

Proof: Note that $\widetilde{\beta}_{e,\mathrm{MA}}(\hat{\boldsymbol{w}}) = [\widehat{\boldsymbol{X}}_e(\hat{\boldsymbol{w}})^\top(\boldsymbol{I}_n - \boldsymbol{P}_1)\boldsymbol{X}_e]^{-1}\widehat{\boldsymbol{X}}_e(\hat{\boldsymbol{w}})^\top(\boldsymbol{I}_n - \boldsymbol{P}_1)\boldsymbol{Y}$ and $\widehat{\beta}_{e,\mathrm{MA}}(\hat{\boldsymbol{w}}) = [\widehat{\boldsymbol{X}}_e(\hat{\boldsymbol{w}})^\top(\boldsymbol{I}_n - \boldsymbol{P}_1)\widehat{\boldsymbol{X}}_e(\hat{\boldsymbol{w}})]^{-1}\widehat{\boldsymbol{X}}_e(\hat{\boldsymbol{w}})^\top(\boldsymbol{I}_n - \boldsymbol{P}_1)\boldsymbol{Y}$. We also have $\widehat{\boldsymbol{X}}_e(\hat{\boldsymbol{w}})^\top(\boldsymbol{I}_n - \boldsymbol{P}_1)\boldsymbol{X}_e = \hat{\boldsymbol{u}}^\top\widehat{\boldsymbol{\Psi}}^{-1}(\hat{\boldsymbol{u}} - \hat{a}\mathbf{1}_M)$ and $\widehat{\boldsymbol{X}}_e(\hat{\boldsymbol{w}})^\top(\boldsymbol{I}_n - \boldsymbol{P}_1)\widehat{\boldsymbol{X}}_e(\hat{\boldsymbol{w}}) = \hat{\boldsymbol{u}}^\top\widehat{\boldsymbol{\Psi}}^{-1}\hat{\boldsymbol{u}} - \hat{a}(\hat{\boldsymbol{u}}^\top\widehat{\boldsymbol{\Psi}}^{-1}\mathbf{1}_M)^2$. Recalling that $\hat{c}_{\mathrm{bias}} = \frac{\hat{a}(\hat{\boldsymbol{u}}^\top\widehat{\boldsymbol{\Psi}}^{-1}\mathbf{1}_M)^2 - \hat{a}\hat{\boldsymbol{u}}^\top\widehat{\boldsymbol{\Psi}}^{-1}\mathbf{1}_M}{\hat{\boldsymbol{u}}^\top\widehat{\boldsymbol{\Psi}}^{-1}\hat{\boldsymbol{u}} - \hat{a}(\hat{\boldsymbol{u}}^\top\widehat{\boldsymbol{\Psi}}^{-1}\mathbf{1}_M)^2}$, we easily obtain $[\widehat{\boldsymbol{X}}_e(\hat{\boldsymbol{w}})^\top(\boldsymbol{I}_n - \boldsymbol{P}_1)\boldsymbol{X}_e]^{-1} = \frac{1}{1+\hat{c}_{\mathrm{bias}}}[\widehat{\boldsymbol{X}}_e(\hat{\boldsymbol{w}})^\top(\boldsymbol{I}_n - \boldsymbol{P}_1)\widehat{\boldsymbol{X}}_e(\hat{\boldsymbol{w}})]^{-1}$, which implies $\widetilde{\beta}_{e,\mathrm{MA}}(\hat{\boldsymbol{w}}) = \frac{1}{1+\hat{c}_{\mathrm{bias}}}\widehat{\beta}_{e,\mathrm{MA}}(\hat{\boldsymbol{w}}) = \widehat{\beta}_e^{(\mathrm{debias})}(\hat{\boldsymbol{w}})$. ■

Proof: (i) (Consistency of $\widetilde{\boldsymbol{\beta}}_{\mathrm{MA}}(\hat{\boldsymbol{w}})$) Since $\widetilde{\boldsymbol{\beta}}_{\mathrm{MA}}(\hat{\boldsymbol{w}}) = [\widehat{\mathcal{X}}(\hat{\boldsymbol{w}})^\top\mathcal{X}]^{-1}\widehat{\mathcal{X}}(\hat{\boldsymbol{w}})^\top\boldsymbol{Y}$, we have

$$\widetilde{\boldsymbol{\beta}}_{\mathrm{MA}}(\hat{\boldsymbol{w}}) - \boldsymbol{\beta} = [\widehat{\mathcal{X}}(\hat{\boldsymbol{w}})^\top\mathcal{X}]^{-1}\widehat{\mathcal{X}}(\hat{\boldsymbol{w}})^\top\boldsymbol{\varepsilon}. \tag{A5}$$

In the above equation, by the law of large numbers we have

$$\frac{1}{n}\widehat{\mathcal{X}}(\hat{\boldsymbol{w}})^\top\mathcal{X} = \begin{pmatrix} \frac{1}{n}\widehat{\boldsymbol{X}}_e(\hat{\boldsymbol{w}})^\top\boldsymbol{X}_e & \frac{1}{n}\widehat{\boldsymbol{X}}_e(\hat{\boldsymbol{w}})^\top\mathcal{X}_o \\ \frac{1}{n}\mathcal{X}_o^\top\boldsymbol{X}_e & \frac{1}{n}\mathcal{X}_o^\top\mathcal{X}_o \end{pmatrix} \xrightarrow{p} \boldsymbol{\Lambda}_1. \tag{A6}$$

Additionally we have

$$\begin{aligned}\frac{1}{n}\widehat{\boldsymbol{X}}_e(\hat{\boldsymbol{w}})^\top\boldsymbol{\varepsilon} &= \frac{1}{n}\boldsymbol{X}_e^\top\boldsymbol{H}^\top\widetilde{\mathcal{X}}_e(\widetilde{\mathcal{X}}_e^\top\boldsymbol{H}\boldsymbol{H}^\top\widetilde{\mathcal{X}}_e)^{-1}\widetilde{\mathcal{X}}_e^\top\boldsymbol{H}\boldsymbol{\varepsilon} \\ &= \boldsymbol{u}^\top\boldsymbol{\Psi}^{-1}\begin{pmatrix} \boldsymbol{v}_1^\top\boldsymbol{\Sigma}_{11}^{-1}\sum_{i=1}^n \frac{1}{n}\mathbf{Z}_{1,i}\varepsilon_i \\ \vdots \\ \boldsymbol{v}_M^\top\boldsymbol{\Sigma}_{MM}^{-1}\sum_{i=1}^n \frac{1}{n}\mathbf{Z}_{M,i}\varepsilon_i \end{pmatrix} \times \{1 + o_p(1)\} = o_p(1),\end{aligned}$$

and

$$\frac{1}{n}\mathcal{X}_o^\top\boldsymbol{\varepsilon} = \frac{1}{n}\sum_{i=1}^n \boldsymbol{X}_{o,i}\varepsilon_i = o_p(1).$$

It follows that

$$\frac{1}{n}\widehat{\mathcal{X}}(\hat{\boldsymbol{w}})^\top\boldsymbol{\varepsilon} = \begin{pmatrix} \frac{1}{n}\widehat{\boldsymbol{X}}_e(\hat{\boldsymbol{w}})^\top\boldsymbol{\varepsilon} \\ \frac{1}{n}\mathcal{X}_o^\top\boldsymbol{\varepsilon} \end{pmatrix} = \begin{pmatrix} \frac{1}{n}\boldsymbol{X}_e^\top\boldsymbol{H}^\top\widetilde{\mathcal{X}}_e(\frac{1}{n}\widetilde{\mathcal{X}}_e^\top\boldsymbol{H}\boldsymbol{H}^\top\widetilde{\mathcal{X}}_e)^{-1}\frac{1}{n}\widetilde{\mathcal{X}}_e^\top\boldsymbol{H}\boldsymbol{\varepsilon} \\ \frac{1}{n}\mathcal{X}_o^\top\boldsymbol{\varepsilon} \end{pmatrix} \xrightarrow{p} 0, \tag{A7}$$

under Assumptions (i)–(iii) and (v), where the last line is obtained by continuous mapping theorem. Again we apply continuous mapping theorem and the above results to Equation (A5), thus

$$\widetilde{\boldsymbol{\beta}}_{\mathrm{MA}}(\hat{\boldsymbol{w}}) - \boldsymbol{\beta} \xrightarrow{p} 0.$$

(ii) (Asymptotic Normality of $\widetilde{\boldsymbol{\beta}}_{\mathrm{MA}}(\hat{\boldsymbol{w}})$) By Equation (A5), we have

$$\begin{aligned}\sqrt{n}[\widetilde{\boldsymbol{\beta}}_{\mathrm{MA}}(\hat{\boldsymbol{w}}) - \boldsymbol{\beta}] &= \sqrt{n}[\widehat{\mathcal{X}}(\hat{\boldsymbol{w}})^\top\mathcal{X}]^{-1}\widehat{\mathcal{X}}(\hat{\boldsymbol{w}})^\top\boldsymbol{\varepsilon} \\ &= \left[\frac{1}{n}\widehat{\mathcal{X}}(\hat{\boldsymbol{w}})^\top\mathcal{X}\right]^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^n \widehat{\boldsymbol{X}}_i(\hat{\boldsymbol{w}})^\top\varepsilon_i \\ &= (\boldsymbol{\Lambda}_1)^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^n \begin{pmatrix} g_i \\ \boldsymbol{X}_{o,i} \end{pmatrix}\varepsilon_i \times \{1 + o_p(1)\} \\ &= (\boldsymbol{\Lambda}_1)^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^n \widetilde{\boldsymbol{\phi}}_i \times \{1 + o_p(1)\},\end{aligned}$$

where $\widetilde{\boldsymbol{\phi}}_i = (g_i, \boldsymbol{X}_{o,i}^\top)^\top \varepsilon_i$ and

$$g_i = \boldsymbol{u}^\top \boldsymbol{\Psi}^{-1} \begin{pmatrix} \boldsymbol{v}_1^\top \boldsymbol{\Sigma}_{11}^{-1} \boldsymbol{Z}_{1,i} \\ \vdots \\ \boldsymbol{v}_M^\top \boldsymbol{\Sigma}_{MM}^{-1} \boldsymbol{Z}_{M,i} \end{pmatrix}.$$

Now we consider $\operatorname{cov}(\widetilde{\boldsymbol{\phi}}_i, \widetilde{\boldsymbol{\phi}}_i)$. Since

$$E\left[\begin{pmatrix} g_i \\ \boldsymbol{X}_{o,i} \end{pmatrix} \begin{pmatrix} g_i & \boldsymbol{X}_{o,i}^\top \end{pmatrix}\right] = \boldsymbol{\Lambda}_2,$$

we have $\operatorname{cov}(\widetilde{\boldsymbol{\phi}}_i, \widetilde{\boldsymbol{\phi}}_i) = \boldsymbol{\Lambda}_2 \sigma^2$. By Slutsky's theorem and multivariate Linderberg-lévy central limit theorem (B. E. Hansen, 2022, Theorem 6.3, p. 160), under Assumptions (i)–(v), we have

$$\sqrt{n}[\widetilde{\boldsymbol{\beta}}_{\mathrm{MA}}(\hat{\boldsymbol{w}}) - \boldsymbol{\beta}] \xrightarrow{d} N(0, \boldsymbol{\Lambda}_1^{-1} \boldsymbol{\Lambda}_2 (\boldsymbol{\Lambda}_1^\top)^{-1} \sigma^2).$$

(iii) (Consistency of $\hat{\sigma}^2_{\mathrm{MA}}$) Denote $\boldsymbol{W} = \mathcal{X}[\widehat{\mathcal{X}}(\hat{\boldsymbol{w}})^\top \mathcal{X}]^{-1} \widehat{\mathcal{X}}(\hat{\boldsymbol{w}})^\top$. By Equation (A5), we have

$$\begin{aligned} \hat{\sigma}^2_{\mathrm{MA}} &= \frac{1}{n-p-1} \sum_{i=1}^{n} \left[y_i - \boldsymbol{X}_i^\top \widetilde{\boldsymbol{\beta}}_{MA}(\hat{\boldsymbol{w}}) \right]^2 \\ &= \frac{1}{n-p-1} \left[\boldsymbol{Y} - \mathcal{X} \widetilde{\boldsymbol{\beta}}_{\mathrm{MA}}(\hat{\boldsymbol{w}}) \right]^\top \left[\boldsymbol{Y} - \mathcal{X} \widetilde{\boldsymbol{\beta}}_{MA}(\hat{\boldsymbol{w}}) \right] \\ &= \frac{1}{n-p-1} \left[\boldsymbol{\varepsilon}^\top \boldsymbol{\varepsilon} - \boldsymbol{\varepsilon}^\top \boldsymbol{W} \boldsymbol{\varepsilon} - (\boldsymbol{\varepsilon}^\top \boldsymbol{W} \boldsymbol{\varepsilon})^\top + \boldsymbol{\varepsilon}^\top \boldsymbol{W}^\top \boldsymbol{W} \boldsymbol{\varepsilon} \right] \\ &= \frac{1}{n-p-1} \left[\boldsymbol{\varepsilon}^\top \boldsymbol{\varepsilon} - 2\boldsymbol{\varepsilon}^\top \boldsymbol{W} \boldsymbol{\varepsilon} + \boldsymbol{\varepsilon}^\top \boldsymbol{W}^\top \boldsymbol{W} \boldsymbol{\varepsilon} \right]. \end{aligned} \tag{A8}$$

Now we separately deal with the three terms on the right-hand side of Equation (A8). To this end, we first note that $\frac{1}{n} \widehat{\mathcal{X}}(\hat{\boldsymbol{w}})^\top \boldsymbol{\varepsilon} = o_p(1)$ follows from Equation (A7). We further have $\frac{1}{n} \mathcal{X}^\top \boldsymbol{\varepsilon} = \begin{pmatrix} \frac{1}{n} \boldsymbol{X}_e^\top \boldsymbol{\varepsilon} \\ \frac{1}{n} \mathcal{X}_o^\top \boldsymbol{\varepsilon} \end{pmatrix} = \begin{pmatrix} E X_{e,i} \varepsilon_i \\ 0 \end{pmatrix} + o_p(1)$ by the law of large numbers and using $E \boldsymbol{X}_{o,i} \varepsilon_i = 0$. Besides, it follows from Equation (A6) that $\frac{1}{n} \widehat{\mathcal{X}}(\hat{\boldsymbol{w}})^\top \mathcal{X} = \boldsymbol{\Lambda}_1 + o_p(1)$. With the above arguments, we can obtain

$$\begin{aligned} \frac{1}{n} \boldsymbol{\varepsilon}^\top \boldsymbol{W} \boldsymbol{\varepsilon} &= \frac{1}{n} \boldsymbol{\varepsilon}^\top \mathcal{X} \left[\frac{1}{n} \widehat{\mathcal{X}}(\hat{\boldsymbol{w}})^\top \mathcal{X} \right]^{-1} \frac{1}{n} \widehat{\mathcal{X}}(\hat{\boldsymbol{w}})^\top \boldsymbol{\varepsilon} \\ &= \left[(E X_{e,i} \varepsilon_i, 0^\top)^\top + o_p(1) \right] \times (\boldsymbol{\Lambda}_1 + o_p(1))^{-1} \times o_p(1) = o_p(1). \end{aligned}$$

Similarly, we have

$$\frac{1}{n} \boldsymbol{\varepsilon}^\top \boldsymbol{W}^\top \boldsymbol{W} \boldsymbol{\varepsilon} = \frac{1}{n} \boldsymbol{\varepsilon}^\top \widehat{\mathcal{X}}(\hat{\boldsymbol{w}}) [\frac{1}{n} \mathcal{X}^\top \widehat{\mathcal{X}}(\hat{\boldsymbol{w}})]^{-1} \frac{1}{n} \mathcal{X}^\top \mathcal{X} [\frac{1}{n} \widehat{\mathcal{X}}(\hat{\boldsymbol{w}})^\top \mathcal{X}]^{-1} \frac{1}{n} \widehat{\mathcal{X}}(\hat{\boldsymbol{w}})^\top \boldsymbol{\varepsilon} = o_p(1).$$

Therefore, by noting that $\frac{n}{n-p-1} \to 1$ as $n \to \infty$, we can write Equation (A8) as

$$\hat{\sigma}^2_{\mathrm{MA}} = \frac{n}{n-p-1} \left[\frac{1}{n} \boldsymbol{\varepsilon}^\top \boldsymbol{\varepsilon} + o_p(1) \right] = E \varepsilon_i^2 + o_p(1) = \sigma^2 + o_p(1).$$

Thus, the proof is finished. ■

Appendix 4. Small-sample scenario and sensitivity analysis

A.1 Small-sample scenario

Table A1. Small sample senario ($n = 50$) for Example 1 (with $\rho_{cs} = 0, 0.2, 0.5$).

ρ_{cs}		2SLS	MA$^{(t,M)}$	MA$^{+(t,M)}$	$\mathfrak{s}^{(t,M)}$	$\mathfrak{s}^{+(t,M)}$	pLasso	pEL$_{0.5}$	Naive
0	Bias	0.015	0.015	**0.0148**	0.1443	0.1432	0.0165	0.015	0.068
	SD	0.0431	**0.0435**	0.0436	0.1181	0.1149	0.0478	0.0457	0.0442
	SE	0.0414	**0.0429**	0.0426	0.0624	0.0594	0.0453	0.044	0.039
	CP	0.914	0.918	0.914	0.406	0.396	0.918	**0.92**	0.592
	$\mathcal{T}$	**0.001**	0.002	0.0034	0.0016	0.003	0.0261	0.0247	–
0.2	Bias	0.0094	**0.0089**	0.0093	0.0634	0.063	0.0097	0.0094	0.0356
	SD	0.029	0.0291	**0.0289**	0.0798	0.0806	0.0373	0.0361	0.0326
	SE	0.0268	0.0284	**0.0283**	0.0436	0.0425	0.0313	0.0302	0.0272
	CP	0.91	**0.912**	0.91	0.58	0.578	**0.912**	0.908	0.756
	$\mathcal{T}$	**0.001**	0.0021	0.0031	0.0018	0.0027	0.0279	0.025	–
0.5	Bias	0.0041	**0.0039**	0.004	0.0309	0.0315	0.0049	0.0044	0.0184
	SD	0.0241	**0.0239**	0.0242	0.0649	0.0648	0.0323	0.0308	0.024
	SE	0.0204	**0.021**	0.021	0.0323	0.0324	0.0238	0.0241	0.0205
	CP	0.91	0.918	**0.92**	0.614	0.622	0.896	0.908	0.832
	$\mathcal{T}$	**0.001**	0.0018	0.0021	0.0017	0.0018	0.027	0.025	–

Note: True value $\beta_e = -1$.

Table A2. Small Sample Senario ($n = 100$) for Example 4 (with $\rho_{cs} = 0.3, 0.5$).

ρ_{cs}		2SLS	MA$^{(t,M)}$	MA$^{+(t,M)}$	$\mathfrak{s}^{(t,M)}$	$\mathfrak{s}^{+(t,M)}$	pLasso	pEL$_{0.5}$	Naive
0.3	Bias	–	**0.0013**	0.0014	0.0354	0.0372	0.0035	0.0033	0.0033
	SD	–	**0.008**	0.0079	0.0344	0.033	0.0301	0.0302	0.0077
	SE	–	**0.0081**	0.0081	0.0257	0.0254	0.0299	0.0296	0.0079
	CP	–	0.946	**0.95**	0.672	0.644	0.93	0.904	0.94
	S($\boldsymbol{X}_I$)	–	165	160.912	165	160.912	76.288	81.094	–
	$\mathcal{T}$	–	0.0359	0.0355	**0.031**	0.0314	0.1556	0.1561	–
0.5	Bias	–	**0.0007**	0.0008	0.0216	0.022	0.0015	0.0018	0.0021
	SD	–	**0.0065**	**0.0065**	0.026	0.0263	0.0232	0.0231	0.0065
	SE	–	**0.0063**	**0.0063**	0.0186	0.0184	0.0227	0.0224	0.0063
	CP	–	**0.944**	0.942	0.712	0.712	0.944	0.942	0.928
	S($\boldsymbol{X}_I$)	–	165	161.844	165	161.844	69.96	74.194	–
	$\mathcal{T}$	–	0.0359	0.0353	**0.0305**	0.031	0.1666	0.1714	–

Note: True value $\beta_e = 1$.

A.2 Sensitivity analysis

We simulate a setting with 100 instrumental variables, among which 10 have coefficients equal to 1 and the rest are set to 0, along with 5 exogenous variables. Since the correlation between the endogenous and exogenous variables forms a vector, for ease of sensitivity analysis, we define a common correlation level denoted by ρ_{oe}, and set each entry of the correlation vector between the endogenous variable and the exogenous variables equal to ρ_{oe}. Based on this assumption, we compute the corresponding coefficient values $\beta_{o,j}$, for $j = 1, \ldots, 5$, such that the marginal correlation between X_e and each $X_o^{(j)}$ equals ρ_{oe}. Sample size $n = 400$.

Due to the structure of this parameterization, ρ_{oe} cannot exceed $1/\sqrt{p} \approx 0.447$ when there are $p = 5$ exogenous variables. All other parameter settings are kept the same as in Example 4. The simulation results of this sensitivity analysis are summarized in Figure A1.

It is not surprising to observe that the bias is not highly sensitive to the correlation between the endogenous and exogenous variables. As explained in Remark 3.2, the bias term can vanish under two

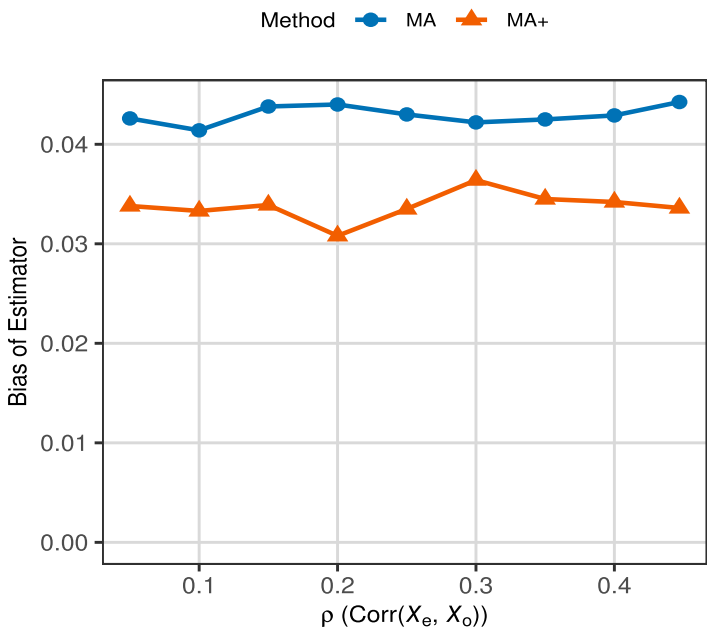

Figure A1. Sensitivity of estimation bias to the correlation between the endogenous and exogenous variables.

special cases: (a) when the exogenous variables $\boldsymbol{X}_o$ and the endogenous variable X_e are uncorrelated, that is, $E(\boldsymbol{X}_{o,i}X_{e,i}) = 0$; (b) when the model averaging weight vector satisfies $\boldsymbol{w}_0^\top \mathbf{1}_M = 1$. In our simulations, we find that the estimated weight vector $\hat{\boldsymbol{w}}$ satisfies $\hat{\boldsymbol{w}}^\top \mathbf{1}_M \approx 1$ and the value of the weight vector varies in each simulation, which makes the bias term intrinsically small and less affected by the value of $E(\boldsymbol{X}_{o,i}X_{e,i})$. This partially explains the weak sensitivity pattern observed in Figure A1. More in-depth sensitivity investigations may be conducted in future research.