# Simple Estimation of Semiparametric Models with Measurement Errors

Kirill S. Evdokimov* Andrei Zeleneev†‡§

This version: November 10, 2025

**Abstract**

We develop a practical way of addressing the Errors-In-Variables (EIV) problem in the Generalized Method of Moments (GMM) framework. We focus on the settings in which the variability of the EIV is a fraction of that of the mismeasured variables, which is typical for empirical applications. For any initial set of moment conditions our approach provides a "corrected" set of moment conditions that are robust to the EIV. We show that the GMM estimator based on these moments is $\sqrt{n}$-consistent, with the standard tests and confidence intervals providing valid inference. This is true even when the EIV are so large that naive estimators (that ignore the EIV problem) are heavily biased with their confidence intervals having 0% coverage. Our approach involves no nonparametric estimation, which is especially important for applications with many covariates and settings with multivariate EIV. In particular, the approach makes it easy to use instrumental variables to address EIV in nonlinear models.

**Keywords:** errors-in-variables, nonstandard asymptotic approximation, nonparametric identification, instrumental variables

*Universitat Pompeu Fabra and Barcelona School of Economics: kirill.evdokimov@upf.edu.

†University College London: a.zeleneev@ucl.ac.uk.

‡First version: November 7, 2016. A part of the material of this paper was previously circulated as a part of Evdokimov and Zeleneev (2018).

§We thank the participants of the numerous seminars and conferences for helpful comments and suggestions. We are also grateful to the Gregory C. Chow Econometrics Research Program at Princeton University and the Department of Economics at the Massachusetts Institute of Technology for their hospitality and support. Evdokimov also gratefully acknowledges the support from the National Science Foundation via grant SES-1459993, and from the Spanish MCIN/AEI via grants RYC2020-030623-I, PID2019-107352GB-I00, PID2022-140825NB-I00, and Severo Ochoa Programme CEX2024-001476-S, funded by MICIU/AEI/10.13039/501100011033.

# 1 Introduction

Measurement errors are a common problem for empirical studies. Addressing the Errors-In-Variables (EIV) bias in nonlinear models requires elaborate strategies.[1] Despite the fundamental theoretical progress in identification and estimation of nonlinear models with EIV, the problem of EIV is still rarely addressed in empirical work outside of linear specifications.

The goal of this paper is to develop a simple and practical approach to estimation of nonlinear semiparametric models that can be expressed in the form of general moment conditions

$$\mathbb{E}[g(X_i^*, S_i, \theta)] = 0 \text{ iff } \theta = \theta_0, \tag{1}$$

where $g(\cdot)$ is a vector of moment functions and $\theta_0$ is the parameter vector of interest. The researcher has a random sample of $\{X_i, S_i\}_{i=1}^n$, where scalar or vector $X_i$ is a mismeasured version of unobserved $X_i^*$ with measurement error $\varepsilon_i$:

$$X_i = X_i^* + \varepsilon_i.$$

We will refer to $g(\cdot)$ as the *original* moment function, since it would have been valid had the researcher observed $X_i^*$. A naive GMM estimator (that ignores the EIV and uses $X_i$ in place of $X_i^*$) based on $g(\cdot)$ is biased because $\mathbb{E}[g(X_i, S_i, \theta_0)] \neq 0$, in contrast to equation (1).

**Example** (Nonlinear Regression, NLR)**.** Let $Y_i$ denote a scalar outcome, and let $X_i^*$ and $W_i$ be the covariates. Suppose

$$E[Y_i | X_i^*, W_i] = \rho(X_i^*, W_i, \theta_0) \tag{2}$$

for some function $\rho$ known up to the parameter $\theta$. For example, in the Logit model, $Y_i$ is binary, $\rho(x, w, \theta) \equiv 1/(1 + \exp(-(\theta_x' x + \theta_w' w)))$, and $\theta \equiv (\theta_x', \theta_w')'$.

Suppose the researcher has an instrumental variable $Z_i$. Then, they can use

$$g(y, x, w, z; \theta) \equiv (y - \rho(x, w, \theta))\, \varphi(x, w, z)$$

as the original moment function, where $\varphi(x, w, z)$ is a vector that, for example, can include $x$, $z$, $w$, their powers and/or interactions.[2] Here $S_i = (Y_i, W_i, Z_i)$. ■

[1] See Hausman, Ichimura, Newey, and Powell (1991); Hausman, Newey, and Powell (1995); Newey (2001); Schennach (2007); Li (2002); Schennach (2004); Chen, Hong, and Tamer (2005); Hu and Schennach (2008); Schennach (2014); Wilhelm (2019), among others.

[2] Note that the moment condition (1) is stated in terms of the true (correctly measured) $X_i^*$.

Even in this well-studied example of nonlinear regression, estimation in the presence of the EIV is a difficult problem. Importantly, nonlinear instrumental variable regression estimator cannot be used, since it is inconsistent in the presence of EIV (Amemiya, 1985). The existing approaches typically require nonparametric estimation that can be impractical in many empirical applications. In contrast, in this paper, we develop an alternative class of estimators, that are essentially GMM estimators that modify the original moment functions $g(\cdot)$ in a way that makes the moment conditions robust to the EIV. In particular, our approach makes it easy to use instrumental variables to address EIV in nonlinear models.

To provide a practical estimation approach for the general class of models (1), we focus on empirical settings in which the researcher believes the variability of the measurement error to be at most a fraction of the variability of the mismeasured variable, i.e., the noise-to-signal ratio $\tau \equiv \sigma_\varepsilon / \sigma_{X^*}$ to be moderate. The absolute magnitude of the measurement error $\sigma_\varepsilon$ does not need to be small. Existing validation studies provide insights into the magnitude of $\tau$ for some key economic variables and datasets. Bound and Krueger (1991) consider log-earnings in the Current Population Survey (CPS) data matched to the Social Security payroll records. Their estimates of the variance of the measurement errors correspond to $\tau$ of approximately 0.47 and 0.30 for the subsamples of men and women, respectively. Bound, Brown, Duncan, and Rodgers (1994) consider a validation study of Panel Study of Income Dynamics (PSID). Their estimates imply $\tau = 0.39 - 0.66$ for log-earnings and $\tau = 0.63 - 0.76$ for hours worked. Pischke (1995) estimates correspond to $\tau = 0.39 - 0.50$ for the log-earnings in PSID. Ashenfelter and Krueger (1994) assess the mismeasurement in the years of education; their estimates correspond to $\tau = 0.30 - 0.37$.

Focusing on these settings allows us to isolate the most important aspects of the problem and, as result, to develop a simple estimator, which does not require any nonparametric estimation or simulation. Such simple estimation becomes possible because in these settings we can obtain a simple approximation of the EIV bias of the moment conditions as a function of $\theta$.

We propose to bias correct the original moments $g(\cdot)$, which in turn removes the bias of the corresponding estimator of $\theta_0$. This bias correction depends on some moments of the distribution of the measurement errors that are unknown. Another

Determining what functions $g(\cdot)$ (or $h(\cdot)$ in the NLR model) satisfy this moment condition does not involve any consideration of the measurement errors and hence is straightforward.

difficulty is that the estimators of some components of the bias correction themselves may need to be bias corrected. To address these issues, we develop the *corrected moment conditions*, which depend on $\theta$ and additional parameters $\gamma$ that govern the bias correction. The true parameter value $\gamma_0$ is associated with (possibly conditional) low-order moments of $\varepsilon_i$. Despite some theoretical subtleties with the construction of the corrected moment conditions, their practical implementation is straightforward and they can be automatically computed for any original moment function $g(\cdot)$.

We introduce the Measurement Error Robust Moments (MERM) estimator, which is a GMM estimator that uses the corrected moment conditions to jointly estimate parameters $\theta_0$ and $\gamma_0$. The estimator can be computed using any standard software for GMM estimation. Joint estimation of parameters $\theta_0$ and $\gamma_0$ using the corrected moment conditions effectively robustifies moment conditions $g(\cdot)$ against the impact of the measurement errors.

To make these ideas precise and to study the properties of the proposed estimators, we develop an asymptotic theory using a nonstandard asymptotic approximation that models $\tau$ as slowly shrinking with the sample size. Standard asymptotics considers $\tau$ to be constant, which implies that as $n \to \infty$ the bias of a naive estimator dwarfs its sampling variability: the bias is constant while the standard errors shrink proportionally to $1/\sqrt{n}$. As a result, under the standard asymptotics, the problem of removing the EIV bias becomes central in the analysis, with relatively little attention paid to the sampling variability of estimators. However, this focus does not seem to be appropriate in many empirical applications, in which the researcher does not expect the potential EIV bias to be several orders of magnitude larger than the standard errors.[3] By considering $\tau$ as drifting towards zero with the sample size, our approach provides a better guidance on construction of EIV robust estimators with good finite sample properties when $\tau$ is small or moderate.[4]

Using this approximation, we show that the proposed estimation approach indeed

[3]Such empirical settings appear to be widespread. Although the concerns about measurement errors are often raised, the majority of applied work does not explicitly correct the EIV bias in nonlinear models, and instead implicitly or explicitly argues or conjectures that the EIV bias is likely not to be too large.

[4]Nonstandard asymptotic approximations with drifting parameters are often used to obtain better approximations of the finite sample behavior of estimators and tests. For example, in the instrumental variable regression settings, to consider the settings with relatively small first stage coefficients, Staiger and Stock (1997) model them as shrinking with $n$. It is important to keep in mind that such nonstandard asymptotic approximations are merely mathematical tools. One should not take them literally and think of parameters somehow changing if more data is collected.

addresses the EIV problem. The MERM estimator is shown to be $\sqrt{n}$-consistent and asymptotically normal and unbiased. The standard confidence intervals and tests for GMM estimators are also valid for the MERM estimator. Additionally, the standard GMM arsenal of assessment tools can be applied to the MERM estimator, allowing one to test model identification, conduct valid inference, and perform model specification diagnostics.

The usefulness of a large sample theory is measured by its ability to approximate the finite sample properties of the estimators and inference procedures. Thus, we study the MERM estimators in a variety of simulation experiments. The results confirm that the nonstandard asymptotic theory indeed provides a good approximation of the finite sample properties of the estimators even in the settings with relatively large EIV. In some of the simulation experiments, the EIV are so large that for the naive estimators' standard 95% confidence intervals have actual coverages of 0% in finite samples, due to the magnitude of the EIV bias. At the same time, even in these settings the MERM estimators perform well, removing the EIV bias and providing confidence intervals with the correct coverage. In particular, the simulation results show that despite the simplicity of implementation, the MERM estimators can compete with and outperform semi-nonparametric estimators.

The MERM estimator is structurally different from the existing approaches that require nonparametric estimation of some nuisance parameters, for example, of the density $f_{X^*|Z,W}$. Avoiding nonparametric estimation has at least two advantages. First, since the majority of empirical applications include additional covariates $W_i$, nonparametric estimation is often infeasible due to the curse of dimensionality. Because the MERM estimator does not involve any nonparametric estimation, it can be used in applications with a relatively large number of additional covariates $W_i$, and remains feasible even in the more complicated settings, including multi-equation and structural models, and applications with multiple mismeasured variables $X_i$. Second, estimation of infinite-dimensional nuisance parameters is typically more demanding towards the sources of identification available in the data, for example, requiring an instrumental variable with a large support (continuously distributed). In contrast, having a discrete instrument is sufficient for the MERM approach because the nuisance parameter $\gamma_0$ is finite-dimensional.

For example, in Section 3.3 we consider estimation of the model of multinomial choice among three modes of transportation. A leading alternative approach to the

errors-in-variables problem in this model is the semi-nonparametric sieve-MLE estimator advocated by Hu and Schennach (2008); Carroll, Chen, and Hu (2010), among others. This approach requires, among other things, estimating the conditional density $f_{X^*|Z,W}$ of $X_i^*$ given the instrument $Z_i$ and covariates $W_i$. In this empirical example, $W_i$ includes four continuously distributed covariates (two continuously distributed characteristics per choice) and a discrete one, while scalar $X_i^*$ and $Z_i$ are also continuous. Thus, $f_{X^*|Z,W}$ is a function of six continuous and one discrete variable. Hence, for typical sample sizes, estimating $f_{X^*|Z,W}$ in this example is infeasible due to the curse of dimensionality. In contrast, as the results of Section 3.3 demonstrate, the MERM approach is practical and effective in this application, in part because it avoids estimation of the high-dimensional nuisance functions like $f_{X^*|Z,W}$ altogether.

The simplicity and practicality of the MERM approach do come at a cost: there is a limit on the magnitude of the measurement errors it can handle. For example, one generally should not expect the MERM approach to work well when $\tau > 1$, i.e., when the noise dominates the signal; in this case the researcher should seek an alternative estimation method.

We view the MERM approach as providing a bridge between the settings in which the measurement errors are guaranteed to be absent or negligible, and the settings where the measurement errors are so large that one has to use the relatively more complicated estimators from the earlier literature (if they exist at all for the model of interest).

**Related Literature** Chen, Hong, and Nekipelov (2011), Schennach (2016), and Schennach (2020) provide excellent overviews of the measurement error literature.

The existing semiparametric approaches to estimation and inference in models with EIV involve nonparametric estimation of infinite-dimensional nuisance parameters (e.g., Chesher, 2000; Li, 2002; Schennach, 2004, 2007; Hu and Schennach, 2008; Schennach and Hu, 2013; Song, 2015), simulation (e.g., Schennach, 2014), or both (e.g., Newey, 2001; Wang and Hsiao, 2011). The exceptions include models with linear and polynomial regression functions (see Hausman et al., 1991, 1995), and Gaussian control variable models such as Probit and Tobit with endogeneity (see Smith and Blundell, 1986; Rivers and Vuong, 1988).

To the best of our knowledge, this paper is the first to provide an approach for $\sqrt{n}$-consistent and asymptotically normal and unbiased estimation of general GMM models with EIV that does not require any nonparametric estimation (or simulation).

We are able to provide such an estimator because we focus on the models with moderate measurement errors. Modeling the variance of the measurement error as shrinking to zero with the sample size is a popular approach in Statistics. The method has been proposed by Wolter and Fuller (1982), who used it to construct an approximate MLE estimator of a nonlinear regression model with Gaussian errors. Following their approach, the Statistics literature has mainly focused on the settings where the moments of the EIV needed to bias correct the estimators are either known or can be directly estimated from the available data such as repeated measurements (e.g., Carroll and Stefanski, 1990; Carroll, Ruppert, Stefanski, and Crainiceanu, 2006). In Economics, such data are relatively rare. The use of approximations with shrinking variance of measurement errors in Econometrics literature has been pioneered by Kadane (1971), Amemiya (1985), and Chesher (1991). Such approximations have been used to check the sensitivity of naive estimators to the EIV by considering how the estimates change as the unknown moments of the measurement errors vary within some set of plausible values, e.g., see Chesher and Schluter (2002), Chesher, Dumangane, and Smith (2002), Battistin and Chesher (2014), Chesher (2017), and Hong and Tamer (2003). Bound, Brown, and Mathiowetz (2001) review a broad list of validation studies matching standard economic dataset to administrative records. The estimates they report suggest that the measurement errors of moderate magnitude are typical for empirical applications. This suggests that the approach developed in this paper could prove valuable for a wide range of applied work.

This paper differs from the earlier literature in several ways. First, it presents a way to estimate the unknown nuisance parameters (moments of the measurement errors) jointly with the parameters of interest. As a result, the approach can, for example, use instrumental variables as a source of identification. Second, the method applies to a very general class of semiparametric models specified by moment conditions. Third, the MERM approach allows the measurement errors to have larger magnitudes than most of the papers in the earlier literature; this is achieved by the MERM approach recursively bias correcting the bias correction terms.

The most widespread approach to identification of the EIV models in economic applications is to use instrumental variables, e.g., see Hausman et al. (1991); Newey (2001); Schennach (2007); Wang and Hsiao (2011). In a recent paper, Hahn, Hausman, and Kim (2021) reconsider the regression model in Amemiya (1990) using a bias correction similar to ours. When proper excluded variables are not available, re-

searchers have considered using higher moments of $X_i$ as instruments, e.g., see Reiersøl (1950); Lewbel (1997); Erickson and Whited (2002); Schennach and Hu (2013); Ben-Moshe, D'Haultfœuille, and Lewbel (2017). When available, repeated measurements can also be used to identify the model, e.g., see Hausman et al. (1991); Li and Vuong (1998); Li (2002); Schennach (2004). The MERM estimator accommodates these identification approaches within a unified estimation framework.

The power of the general MERM approach can be illustrated in the NLR model. For example, when a candidate instrumental variable is available, the conditions it needs to satisfy are much weaker than what is required by many existing approaches. Availability of a discrete instrument is sufficient for identification; and the instrument is allowed to have heterogeneous impact on covariates $X_i^*$. One can also take a non-classical, nonlinear (e.g., discretized or censored), or biased measurement of $X_i^*$ as an instrument in the MERM approach. We discuss identification in Section 2.4. In addition, in a related paper Evdokimov and Zeleneev (2022) study nonparametric regression with EIV using the $\tau \to 0$ approximation, and demonstrate that the MERM approach can also be motivated from a nonparametric perspective.

Kitamura, Otsu, and Evdokimov (2013); Andrews, Gentzkow, and Shapiro (2017); Armstrong and Kolesár (2021); Bonhomme and Weidner (2022), among others, develop tools for estimation and inference in GMM, which are robust to general perturbation or misspecification of the true data generating process. They focus on the settings in which these perturbations are sufficiently small, so that naive estimators remain $\sqrt{n}$-consistent, and their biases are of the same order of magnitude as their standard errors. In contrast, we focus on more specific forms of data contamination due to the EIV. This allows the MERM approach to remain valid even in the settings with larger measurement errors, in which naive estimators may have slower than $\sqrt{n}$ rates of convergence.

The MERM approach also provides a useful foundation for dealing with EIV in more complicated settings. Evdokimov and Zeleneev (2018) utilize the MERM framework to address an issue of nonstandard inference, which turns out to arise generally when EIV models are identified using instrumental variables. Evdokimov and Zeleneev (2019) extend the analysis of this paper to long panel and network settings.

**Organization of the paper** Section 2 introduces the Moderate Measurement Error framework and the proposed MERM estimator. Section 3 presents several Monte

Carlo experiments that illustrate finite sample properties of the MERM estimators. Section 4 considers several extensions of the framework. A supplementary appendix contains all proofs and additional results for the numerical and empirical illustrations.

# 2 Moderate Measurement Errors Framework

To present the main ideas we first consider the case of univariate $X_i^*$. We will consider multivariate $X_i^*$ later. We assume that the measurement error is classical, i.e., that $\varepsilon_i$ is independent of $X_i^*$ and $S_i$; later we will discuss how this assumption can be relaxed. Following the rest of the literature, we assume that $\mathbb{E}[\varepsilon_i] = 0$.[5]

To develop a practical estimation approach for general moment condition models we focus on the settings in which $\tau \equiv \sigma_\varepsilon / \sigma_{X^*}$ is small or moderate. We consider an asymptotic approximation with $\tau_n \equiv \tau \to 0$ as $n \to \infty$.

Note that economically meaningful parameters are usually invariant to rescaling of $X_i^*$. Likewise, the extent of the EIV problem does not change with such rescaling. For simplicity of exposition, it is convenient to assume that $X_i^*$ is scaled so that $\sigma_{X^*}$ is of order one and, correspondingly, the moments $\mathbb{E}[|\varepsilon_i|^k] \propto \tau_n^k$ decrease with $k$ when $\tau_n < 1$. For example, this could be ensured by normalizing observed $X_i$ to have $\sigma_X = 1$. Let us stress that this normalization is used only to simplify the exposition; as we show in Appendix G, the proposed MERM estimator does not require any normalizations in practice.

## 2.1 Special Case: Quadratic Expansion

For clarity, we first consider a simple special case of the general approach. Let us denote $g_x^{(k)}(x, s, \theta) \equiv \partial^k g(x, s, \theta) / \partial x^k$. Since $\mathbb{E}[|\varepsilon_i|^k] \propto \tau_n^k \to 0$ as $n \to \infty$, under some regularity conditions, we can write the quadratic Taylor expansion of function $g(X_i, S_i, \theta) = g(X_i^* + \varepsilon_i, S_i, \theta)$ around $\varepsilon_i = 0$ as

$$\begin{aligned}\mathbb{E}[g(X_i, S_i, \theta)] &= \mathbb{E}\left[g(X_i^*, S_i, \theta) + g_x^{(1)}(X_i^*, S_i, \theta)\varepsilon_i + \frac{1}{2} g_x^{(2)}(X_i^*, S_i, \theta)\varepsilon_i^2\right] + O(\mathbb{E}[|\varepsilon_i|^3]) \\ &= \mathbb{E}[g(X_i^*, S_i, \theta)] + \frac{\mathbb{E}[\varepsilon_i^2]}{2}\mathbb{E}[g_x^{(2)}(X_i^*, S_i, \theta)] + O(\tau_n^3), \end{aligned} \tag{3}$$

[5] A location normalization such as $\mathbb{E}[\varepsilon_i] = 0$ is usually necessary because it is not possible to separately identify the means $\mathbb{E}[X_i^*]$ and $\mathbb{E}[\varepsilon_i]$.

where the second equality holds because $\varepsilon_i$ and $(X_i^*, S_i)$ are independent, and $\mathbb{E}[\varepsilon_i] = 0$.

Evaluating the expansion above at $\theta = \theta_0$ gives $\mathbb{E}[g(X_i, S_i, \theta_0)] = O(\sigma_\varepsilon^2) = O(\tau_n^2)$, because $\mathbb{E}[g(X_i^*, S_i, \theta_0)] = 0$. As a result, a naive estimator that ignores the EIV and uses $X_i$ in place of $X_i^*$ has EIV bias of order $\tau_n^2$.[6] The bias of the naive estimator should be compared with its standard error, which is of order $n^{-1/2}$. Thus, the bias of the naive estimator is not negligible, unless the measurement error is rather small (theoretically, unless $\tau_n^2 = o\left(n^{-1/2}\right)$). In particular, tests and confidence intervals based on the naive estimator are invalid and can provide highly misleading results. Moreover, if $\tau_n^2$ shrinks at a rate slower than $O\left(n^{-1/2}\right)$, the rate of convergence of the naive estimator is slower than $\sqrt{n}$.

Suppose $\tau_n = o\left(n^{-1/6}\right)$. Then, $O(\tau_n^3) = o\left(n^{-1/2}\right)$ and we can rearrange equation (3) as

$$\mathbb{E}[g(X_i^*, S_i, \theta)] = \mathbb{E}[g(X_i, S_i, \theta)] - \frac{\mathbb{E}[\varepsilon_i^2]}{2}\mathbb{E}\left[g_x^{(2)}(X_i^*, S_i, \theta)\right] + o(n^{-1/2}). \tag{4}$$

The left-hand side of this equation is exactly the moment condition (1) that we would like to use for estimation of $\theta_0$. The first term on the right-hand side involves only observed variables, and can be estimated by the sample average $\overline{g}(\theta) \equiv n^{-1}\sum_{i=1}^n g(X_i, S_i, \theta)$. The second term on the right-hand side can be thought of as a bias correction that removes the EIV-bias from the expected moment function $\mathbb{E}[g(X_i, S_i, \theta)]$.

The idea of the MERM estimator we propose is to make use of expansions such as (4) to bias correct the moment condition $\mathbb{E}[g(X_i, S_i, \theta)]$, which in turn removes the bias of the estimator of the parameters of interest $\theta_0$. To perform the bias correction we need to estimate two quantities: $\mathbb{E}[\varepsilon_i^2]$ and $\mathbb{E}[g_x^{(2)}(X_i^*, S_i, \theta)]$.

First, we show that in equation (4) we can substitute $\mathbb{E}[g_x^{(2)}(X_i^*, S_i, \theta)]$ with $\mathbb{E}[g_x^{(2)}(X_i, S_i, \theta)]$, which in turn can be estimated by $\overline{g}_x^{(2)}(\theta) \equiv n^{-1}\sum_{i=1}^n g_x^{(2)}(X_i, S_i, \theta)$. By the Taylor expansion around $\varepsilon_i = 0$ similar to equation (3), we can show that $\mathbb{E}[g_x^{(2)}(X_i^*, S_i, \theta)] = \mathbb{E}[g_x^{(2)}(X_i, S_i, \theta)] + O(\tau_n^2)$ and hence

$$\frac{1}{2}\mathbb{E}[\varepsilon_i^2]\left(\mathbb{E}\left[g_x^{(2)}(X_i^*, S_i, \theta)\right] - \mathbb{E}\left[g_x^{(2)}(X_i, S_i, \theta)\right]\right) = \mathbb{E}[\varepsilon_i^2]O\left(\tau_n^2\right) = O\left(\tau_n^4\right). \tag{5}$$

Here $O\left(\tau_n^4\right) = o\left(n^{-1/2}\right)$ because we assume that $\tau_n = o\left(n^{-1/6}\right)$. The idea behind

[6] For example, consider a linear regression with a scalar mismeasured regressor. The bias of the naive OLS estimator of the slope parameter $\theta_{01}$ is $-\theta_{01}\frac{\tau_n^2}{1+\tau_n^2} = -\theta_{01}\tau_n^2 + O\left(\tau_n^4\right)$.

this substitution is that the bias of order $O(\tau_n^2)$ in $\mathbb{E}[g_x^{(2)}(X_i, S_i, \theta)]$ can be ignored because it is multiplied by $E[\varepsilon_i^2] = O(\tau_n^2)$.[7] With the substitution, we can rearrange equation (4) and write it as

$$\mathbb{E}[g(X_i^*, S_i, \theta)] = \mathbb{E}\left[g(X_i, S_i, \theta) - \frac{\mathbb{E}[\varepsilon_i^2]}{2} g_x^{(2)}(X_i, S_i, \theta)\right] + o(n^{-1/2}). \tag{6}$$

Second, we propose estimating the unknown $\mathbb{E}[\varepsilon_i^2]$ together with the parameter of interest $\theta$. Specifically, let $\gamma_{02} \equiv \mathbb{E}[\varepsilon_i^2]/2$ denote the true value of parameter $\gamma_2$, and consider the following *corrected moment function*:

$$\psi(X_i, S_i, \theta, \gamma) \equiv g(X_i, S_i, \theta) - \gamma_2 g_x^{(2)}(X_i, S_i, \theta). \tag{7}$$

Function $\psi$ is a moment function parameterized by $\theta$ and $\gamma$, and

$$\mathbb{E}[\psi(X_i, S_i, \theta_0, \gamma_{02})] = \mathbb{E}[g(X_i^*, S_i, \theta_0)] + o\left(n^{-1/2}\right) = o\left(n^{-1/2}\right), \tag{8}$$

where the first equality follows from equation (6) and the definition of $\gamma_{02}$, and the second equality follows from equation (1). Hence, the corrected moment conditions $\psi$ can be used to jointly estimate the true parameters $\theta_0$ and $\gamma_{02}$ by a GMM estimator.[8]

**Remark 1.** If $\mathbb{E}[\varepsilon_i^3] = 0$ (e.g., if the distribution of $\varepsilon_i$ is symmetric), the remainder in equation (3) is of a smaller order $O(\tau_n^4)$. Hence, the corrected moments (8) remain valid for larger values of $\tau_n$, requiring only the weaker condition $\tau_n = o(n^{-1/8})$. The bias of the naive estimators in this case can be as large as $o(n^{-1/4})$.

## 2.2 General Case: Expansion of order $K$

The quadratic expansion of equation (3) can be extended to general order $K \geq 2$. Considering larger $K$ theoretically allows $\tau_n$ converging to zero at a slower rate. In finite samples this corresponds to the asymptotics providing good approximations for larger values of $\tau_n$, i.e., large measurement errors. Expanding $g(X_i^* + \varepsilon_i, S_i, \theta)$ around $\varepsilon_i = 0$ we have,

$$\mathbb{E}[g(X_i, S_i, \theta)] = \mathbb{E}\left[g(X_i^*, S_i, \theta) + \sum_{k=1}^{K} \frac{\varepsilon_i^k}{k!} g_x^{(k)}(X_i^*, S_i, \theta)\right] + O\left(\mathbb{E}\left[|\varepsilon_i|^{K+1}\right]\right). \tag{9}$$

[7] Such substitutions of $X^*$ with $X$ have been used in other contexts, e.g., Chesher and Schluter (2002).

[8] In the moment condition settings, having $o\left(n^{-1/2}\right)$ is equivalent to having 0 on the right-hand side of equation (8).

The above special case of quadratic expansion corresponds to $K = 2$.

The approximation we consider is formalized by the following assumption.

**Assumption MME.** (Moderate Measurement Errors) *(i)* $\tau_n = o(n^{-1/(2K+2)})$ *for some integer* $K \geq 2$*; and (ii)* $\mathbb{E}[|\varepsilon_i|^L] \leq C\sigma_\varepsilon^L$ *for some* $L \geq K+1$ *and* $C > 0$.

Assumption MME(i) limits the magnitude of the measurement errors and implies that $\tau_n^{K+1} = o\left(n^{-1/2}\right)$. Assumption MME(ii) implies that $\mathbb{E}[|\varepsilon_i|^k] = O\left(\sigma_\varepsilon^k\right)$, and requires the tails of $\varepsilon_i/\sigma_\varepsilon$ to be sufficiently thin. Together, parts (i) and (ii) imply that $\mathbb{E}[|\varepsilon_i|^{K+1}] = O\left(\tau_n^{K+1}\right) = o\left(n^{-1/2}\right)$, and hence ensure that the remainder in equation (9) is negligible. Using $\mathbb{E}\left[\varepsilon_i | X_i^*, S_i\right] = 0$ to further simplify this expansion and rearranging the terms we obtain

$$\mathbb{E}[g(X_i^*, S_i, \theta)] = \mathbb{E}[g(X_i, S_i, \theta)] - \sum_{k=2}^{K} \frac{\mathbb{E}[\varepsilon_i^k]}{k!} \mathbb{E}\left[g_x^{(k)}(X_i^*, S_i, \theta)\right] + o(n^{-1/2}). \tag{10}$$

This equation is the general expansion analog of equation (4). The summation on the right hand side is the bias correction term, which we use to construct the MERM estimator.[9]

It turns out that for $K \geq 4$, estimation of $\mathbb{E}[g_x^{(k)}(X_i^*, S_i, \theta)]$ is more intricate than in the case of $K = 2$, and the substitution we made in equation (6) no longer works. Larger values of $K$ allow for larger values of $\tau_n$ and hence larger EIV biases of naive estimators $n^{-1}\sum_{i=1}^n g_x^{(k)}(X_i, S_i, \theta)$. The expansion of order $K$ includes terms up to the order $\tau_n^K$, with the asymptotically negligible remainder of order $O\left(\tau_n^{K+1}\right)$. For $K \geq 4$, terms of order $\tau_n^4$ are not negligible. This implies that we cannot ignore the EIV bias that would arise from substituting $\mathbb{E}[g_x^{(2)}(X_i^*, S_i, \theta)]$ with $\mathbb{E}[g_x^{(2)}(X_i, S_i, \theta)]$ in equation (10), because this bias is of order $O\left(\tau_n^4\right)$ according to equation (5). To address this problem, we instead replace $\mathbb{E}[g_x^{(2)}(X_i^*, S_i, \theta)]$ with the bias corrected expression $\mathbb{E}[g_x^{(2)}(X_i, S_i, \theta)] - (\mathbb{E}[\varepsilon_i^2]/2)\,\mathbb{E}[g_x^{(4)}(X_i, S_i, \theta)]$. Thus, for $K \geq 4$, one needs to bias correct the estimator of the bias correction term. Moreover, for larger $K$ one needs to bias correct the bias correction of the bias correction term and so on.

Fortunately, we show that these bias corrections can be constructed as linear combinations of the expectations of the higher order derivatives of $g_x^{(k)}(X_i, S_i, \theta)$. Let

[9] It is useful to get a sense of the magnitudes of the coefficients $\mathbb{E}\left[\varepsilon_i^k\right]/k!$ in equation (10). Suppose $\varepsilon_i \sim N\left(0, \sigma_\varepsilon^2\right)$, $\sigma_\varepsilon = 0.5$, and $\sigma_{X^*} = 1$, so $\tau = \sigma_\varepsilon = 0.5$. Then the coefficients in front of $g_x^{(2)}$, $g_x^{(4)}$, and $g_x^{(6)}$ are $\mathbb{E}\left[\varepsilon_i^2\right]/2! = 0.125$, $\mathbb{E}\left[\varepsilon_i^4\right]/4! \approx 0.008$, and $\mathbb{E}\left[\varepsilon_i^6\right]/6! \approx 0.0003$.

us define the following *corrected moment function*:

$$\psi(X_i, S_i, \theta, \gamma) \equiv g(X_i, S_i, \theta) - \sum_{k=2}^{K} \gamma_k g_x^{(k)}(X_i, S_i, \theta), \tag{11}$$

where $\gamma = (\gamma_2, \ldots, \gamma_K)'$ is a $K-1$ dimensional vector of parameters. Let $\gamma_0 \equiv (\gamma_{02}, \ldots, \gamma_{0K})'$ denote the vector of true parameters $\gamma_{0k}$, defined as

$$\gamma_{02} \equiv \frac{\mathbb{E}\left[\varepsilon_i^2\right]}{2}, \ \gamma_{03} \equiv \frac{\mathbb{E}\left[\varepsilon_i^3\right]}{6}, \ \text{ and } \ \gamma_{0k} \equiv \frac{\mathbb{E}\left[\varepsilon_i^k\right]}{k!} - \sum_{\ell=2}^{k-2} \frac{\mathbb{E}\left[\varepsilon_i^{k-\ell}\right]}{(k-\ell)!}\gamma_{0\ell} \ \text{ for } \ k \geq 4. \tag{12}$$

We formalize this discussion below.

**Assumption CME.** (Classical Measurement Error) *$\varepsilon_i$ is independent from $(X_i^*, S_i)$ and $\mathbb{E}[\varepsilon_i] = 0$.*

The following lemma establishes validity of the corrected moment conditions under Assumptions MME, CME, and some mild regularity conditions provided in Appendix A.

**Lemma 1.** *Under Assumptions MME, CME and A.1 in Appendix A,*

$$\mathbb{E}[\psi(X_i, S_i, \theta_0, \gamma_0)] = \mathbb{E}[g(X_i^*, S_i, \theta_0)] + o\left(n^{-1/2}\right) = o\left(n^{-1/2}\right).$$

Lemma 1 implies that the corrected moment conditions $\psi$ are valid and can potentially be used to jointly estimate parameters $\theta_0$ and $\gamma_0$. The total number of parameters to be estimated is now $\dim(\theta) + K - 1$. Thus, joint estimation of $\theta_0$ and $\gamma_0$ requires that $\dim(\psi) = \dim(g) \geq \dim(\theta) + K - 1$, i.e., that the original moment conditions $g$ include sufficiently many overidentifying restrictions. For example, the overidentifying restrictions can be constructed by using an instrumental variable; we discuss this in more detail below.

**Remark 2.** Construction of the corrected moment conditions $\psi$ requires the original moment function $g(x, s, \theta)$ to have a sufficient number of derivatives with respect to $x$. Thus, the proposed correction method does not apply to settings with non-differentiable moment functions, for example, those arising in the instrumental variable quantile regression (IVQR).

## 2.3 Measurement Error Robust Moments (MERM) estimator

The MERM estimator jointly estimates the parameters $\theta_0$ and $\gamma_0$ using moment conditions $\psi$. It is convenient to define the joint vector of parameters

$$\beta \equiv (\theta', \gamma')', \ \beta_0 \equiv (\theta_0', \gamma_0')', \ \hat{\beta} \equiv (\hat{\theta}', \hat{\gamma}')',$$

and the parameter space $\mathcal{B} \equiv \Theta \times \Gamma$, where $\Theta$ and $\Gamma$ are the parameter spaces for $\theta$ and $\gamma$. Then, MERM estimator is the GMM estimator (Hansen, 1982):

$$\hat{\beta} \equiv \underset{\beta \in \mathcal{B}}{\operatorname{argmin}} \, \hat{Q}(\beta), \qquad \hat{Q}(\beta) \equiv \overline{\psi}(\beta)' \hat{\Xi} \overline{\psi}(\beta), \tag{13}$$

where $\overline{\psi}(\beta) \equiv n^{-1} \sum_{i=1}^{n} \psi_i(\beta)$, $\psi_i(\beta) \equiv \psi(X_i, S_i, \beta)$, $\hat{\Xi}$ is a weighting matrix, and $\hat{Q}(\beta)$ is the standard GMM objective function.

While Lemma 1 establishes validity of the corrected moment restrictions $\psi$, the MERM estimator also relies on $\beta_0$ being identified from $\psi$. This requirement is formalized by the following assumption.

**Assumption ID.** (Identification)

*(i) the Jacobian $\Psi^*$ has full column rank, where*

$$\begin{aligned} \Psi^* &\equiv \mathbb{E}\left[\nabla_\theta \psi(X_i^*, S_i, \theta_0, 0), \nabla_\gamma \psi(X_i^*, S_i, \theta_0, 0)\right] \\ &= \mathbb{E}\left[\nabla_\theta g(X_i^*, S_i, \theta_0), -g_x^{(2)}(X_i^*, S_i, \theta_0), \ldots, -g_x^{(K)}(X_i^*, S_i, \theta_0)\right]; \end{aligned}$$

*(ii)* $\mathbb{E}\left[\psi(X_i^*, S_i, \theta, \gamma)\right] = 0$ *iff* $\theta = \theta_0$ *and* $\gamma = 0$.

Assumptions ID(i) and (ii) are the standard GMM local and global identification conditions applied to the moment function $\psi(X_i^*, S_i, \theta, \gamma)$. These are high-level conditions, which we will return to later in the paper. The moment conditions formulation is sufficiently general to encompass a wide variety of sources of identification. In Section 2.4, we discuss identification in detail and illustrate the construction of the moment function using an instrumental variable or a second measurement.

Under some additional regularity conditions, estimator $\hat{\beta}$ behaves as a standard GMM-type estimator: it is $\sqrt{n}$-consistent and asymptotically normal and unbiased. This result is formalized by the following theorem.

**Theorem 2** (Asymptotic Normality)**.** *Suppose that* $\{(X_i^*, S_i', \varepsilon_i)\}_{i=1}^n$ *are i.i.d.. Then, under Assumptions MME, CME, ID, and A.1-A.3 in Appendix A,*

$$n^{1/2}\Sigma^{-1/2}(\hat{\beta} - \beta_0) \xrightarrow{d} N(0, I_{\dim(\beta)}), \text{ where} \tag{14}$$
$$\Sigma \equiv (\Psi'\Xi\Psi)^{-1}\Psi'\Xi\Omega_{\psi\psi}\Xi\Psi(\Psi'\Xi\Psi)^{-1}.$$

Theorem 2 shows that the MERM approach addresses the EIV bias problem, and in particular provides a $\sqrt{n}$-consistent asymptotically normal and unbiased estimator $\hat{\theta}$, which can be used to conduct inference about the true parameters $\theta_0$. The asymptotic variance $\Sigma$ takes the standard sandwich form, with $\Psi \equiv \mathbb{E}[\nabla_\beta \psi_i(\beta_0)]$, $\Omega_{\psi\psi} \equiv \mathbb{E}[\psi_i(\beta_0)\psi_i'(\beta_0)]$, and $\hat{\Xi} \to_p \Xi$.

**Remark 3.** Notice that the bias of naive estimators (such as a GMM estimator based on the original moment conditions) is $O(\tau_n^2)$, so their rate of convergence is $O_p(\tau_n^2 + n^{-1/2})$. The bias dominates sampling variability and naive estimators are not $\sqrt{n}$-consistent unless $\tau_n = O(n^{-1/4})$, i.e., unless the magnitude of the measurement error is rather small. At the same time, the MERM estimator remains $\sqrt{n}$-consistent for much larger values of $\tau_n$, up to $\tau_n = O(n^{-1/(2K+2)})$, whereas the rate of convergence of naive estimators is only $O_p(n^{-1/(K+1)})$ in this case.

Once the corrected moment condition $\psi$ is constructed, estimation of and inference about parameters $\beta_0$ can be performed using any standard software package for GMM estimation. In other words, the proposed estimator can be simply treated as a standard GMM estimator based on the corrected moment conditions $\psi$, and the conventional standard errors, tests, and confidence intervals are valid.

In addition to the estimation of the parameters $\theta_0$, researchers are often interested in average effects of the form $\lambda_0 \equiv \mathbb{E}[\lambda(X_i^*, S_i, \theta_0)]$. For instance, in the NLR model, one may be interested in the average partial effect of $x$ (i.e., $\lambda_0 \equiv \mathbb{E}[\nabla_x \rho(X_i^*, S_i, \theta_0)]$) or another covariate. The naive average partial effect estimator $\hat{\lambda}_{\text{Naive}} \equiv \frac{1}{n}\sum_{i=1}^n \lambda(X_i, S_i, \hat{\theta})$ suffers from the EIV bias, unless function $\lambda$ is linear in $X_i^*$. Instead, one should use estimates $\hat{\gamma}$ to construct the bias-corrected estimator $\hat{\lambda}_{\text{MERM}} \equiv \frac{1}{n}\sum_{i=1}^n \left\{\lambda(X_i, S_i, \hat{\theta}) - \sum_{k=2}^K \hat{\gamma}_k \lambda_x^{(k)}(X_i, S_i, \hat{\theta})\right\}$.

**Remark 4.** The standard $J$-test of overidentifying restrictions remains valid in the MERM settings, and can be used to check the model specification. The $J$-test jointly tests the following hypotheses: *(i)* $K$ is sufficiently large to correct the EIV bias;

*(ii)* assumptions on the EIV are valid; and *(iii)* the original moment conditions $g$ are correctly specified so equation (1) holds, i.e., that the original economic model is correctly specified aside from the presence of the EIV in $X_i$. Thus, if the $J$-test rejects the validity of the corrected moment conditions $\psi$, the researcher might want to *(i)* consider taking a larger $K$; *(ii)* employ a different correction method; or *(iii)* consider an alternative specification of the original moments $g$.

**Remark 5.** Considering larger $K$ allows for $\tau_n$ converging to zero at a slower rate, which in finite samples corresponds to the asymptotics providing better approximations for larger magnitudes of measurement errors. On the other hand, taking a larger $K$ increases the dimension of the nuisance parameter $\gamma_0$ and thus typically increases the variance of $\hat{\theta}$. We consider this issue in more detail and provide a data-driven method for choosing $K$ in Section 4.1.

**Remark 6.** The MERM framework can be extended to the case of non-classical measurement errors; see Evdokimov and Zeleneev (2022) for details and a fully nonparametric analysis. In this paper, we focus on the classical measurement errors, developing a practical bias correction approach, which can be easily implemented in a wide range of economic applications. Even when Assumption CME is violated, the deviations from it are often limited in magnitude, so the corrections based on the CME assumption remove most of the EIV bias. Thus, in practice, using the estimator designed for classical measurement errors is typically preferable to ignoring mismeasurement altogether.

**Remark 7.** It is important to note that $\gamma_{0k} \neq \mathbb{E}\left[\varepsilon_i^k\right]/k!$ for $k \geq 4$, contrary to what equation (10) might suggest. For example, $\gamma_{04} = \left(\mathbb{E}\left[\varepsilon_i^4\right] - 6\sigma_\varepsilon^4\right)/24$ is negative for many distributions, including normal. The reason that generally $\gamma_{0k} \neq \mathbb{E}\left[\varepsilon_i^k\right]/k!$ is that the estimators of the correction terms themselves need a correction, which is accounted for by the form of $\gamma_{0k}$. Since there is a one-to-one relationship between $\gamma_0$ and the moments $\mathbb{E}\left[\varepsilon_i^\ell\right]$, parameter space $\Gamma$ for $\gamma_0$ can incorporate restrictions that the moments must satisfy (e.g., $\sigma_\varepsilon^2 \geq 0$ and $\mathbb{E}\left[\varepsilon_i^4\right] \geq \sigma_\varepsilon^4$). Such restrictions can increase the efficiency of the estimator and the power of tests.

**Remark 8.** No parametric assumptions are imposed on the distribution of $\varepsilon_i$, i.e. the distribution of $\varepsilon_i$ is treated nonparametrically. The regularity conditions restrict only the magnitude of the moments of $\varepsilon_i$. The approach imposes no restrictions

on the smoothness of the distributions of $X_i^*$ and $\varepsilon_i$, which are not even required to be continuous. Examples in which this can be useful include individual wages (whose distributions may have point masses at round numbers), and allowing the measurement error $\varepsilon_i$ to have a point mass at zero (a fraction of the population may have a zero measurement or recall error).

**Remark 9.** The formulas of the derivatives $g_x^{(k)}(\cdot)$ are typically easy to compute analytically or using symbolic algebra software. Alternatively, these derivatives can be computed using numerical differentiation. Thus, the corrected set of moments can be automatically produced for a generic moment function $g(\cdot)$ provided by the user.

## 2.4 Model Identification: Jacobian $\Psi$

Theorem 2 requires $\beta_0$ to be identified and the Jacobian matrix $\Psi$ to be full rank. Notably, the MERM framework encompasses many possible sources of identification at once, including instrumental variables, additional measurements, or nonlinearities of the functional form. The identifying information is incorporated in the moment functions. Essentially, our approach first characterizes in what directions the measurement errors can bias the moment conditions $\mathbb{E}\left[g\left(X_i, S_i, \theta\right)\right]$, and then uses the moments orthogonal to those directions for identification of $\theta_0$. To be more specific, we will now consider identification when in addition to the error-laden $X_i$ we have either (i) a general instrument $Z_i$ or (ii) a second measurement $Q_i$.

**Identification Using A General Instrument $Z_i$** Many applications can be formulated as the following conditional moment restriction:

$$\mathbb{E}\left[u\left(X_i^*, S_i, \theta\right)|X_i^*\right] = 0 \text{ iff } \theta = \theta_0, \tag{15}$$

for some moment function $u$. For example, consider the nonlinear regression model $\mathbb{E}\left[Y_i|X_i^* = x\right] = \rho\left(x, \theta_0\right)$, then $u\left(x, y, \theta\right) = \rho\left(x, \theta\right) - y$.[10]

In applications, identification of the models with EIV would typically rely on an instrumental variable $Z_i$. Suppose the instrument satisfies the exclusion restriction $\mathbb{E}\left[u\left(X_i^*, S_i, \theta\right)|X_i^*, Z_i\right] = \mathbb{E}\left[u\left(X_i^*, S_i, \theta\right)|X_i^*\right]$, i.e., conditional on the true $X_i^*$

[10]For simplicity of the exposition, in the expectation in equation (15) we only condition on $X_i^*$. The discussion applies in a straightforward way to the settings with additional correctly measured variables $W_i$ in the conditioning set, i.e., the model $\mathbb{E}\left[u\left(X_i^*, S_i, \theta_0\right)|X_i^*, W_i\right] = 0$. For example, in the nonlinear regression example with additional covariates $W_i$ we have $\mathbb{E}\left[Y_i|X_i^* = x, W_i = w\right] = \rho\left(x, w, \theta_0\right)$, so $u\left(x, y, w, \theta\right) = \rho\left(x, w, \theta\right) - y$.

the instrument has no further effect on the moment conditions $u$. Consider the moment functions $h(x,s,\theta) \equiv u(x,s,\theta) \otimes \varphi_X(x)$, where $\varphi_X(x)$ is a vector of functions of $x$, e.g., $\varphi_X(x) \equiv \left(1, x, \ldots, x^J\right)'$. By the Law of Iterated Expectations, $\mathbb{E}\left[h\left(X_i^*, S_i, \theta_0\right) | Z_i = z\right] = 0$ for all $z$. However, the same expectation with $X_i^*$ replaced by the observed $X_i$, $\mathbb{E}\left[h\left(X_i, S_i, \theta_0\right) | Z_i = z\right]$, will depend on $z$. One way to see this is to consider a Taylor expansion similar to equation (3):

$$\mathbb{E}\left[h\left(X_i, S_i, \theta_0\right) | Z_i = z\right] = \underbrace{\mathbb{E}\left[h\left(X_i^*, S_i, \theta_0\right) | Z_i = z\right]}_{=0 \text{ by the LIE}} + \gamma_0 \mathbb{E}\left[h_x^{(2)}\left(X_i^*, S_i, \theta_0\right) | Z_i = z\right] + O\left(\tau_n^3\right),$$

which shows that $\mathbb{E}\left[h\left(X_i, S_i, \theta_0\right) | Z_i = z\right]$ is zero for all $z$ (up to a negligible remainder) unless $\gamma_0 \neq 0$, where $\gamma_0 = \sigma^2/2$. Thus, $\mathbb{E}\left[h\left(X_i, S_i, \theta_0\right) | Z_i = z\right]$ varies with $z$ only because of the presence of the measurement error. Intuitively, the magnitude of this variation then identifies the nuisance parameters $\gamma_0$. Thus, one can rely on the original moment functions of the typical form $g(x,s,\theta) = h(x,s,\theta) \otimes \varphi_Z(z)$, where $\varphi_Z(z)$ is a vector of functions of $z$.

The above discussion provides the intuition for identification of the nonlinear moment condition models with EIV. It is important to note that identification of general nonlinear moment condition models is a complicated problem. Even in the settings without measurement errors, it is generally not possible to give low-level conditions guaranteeing that a specific set of nonlinear moment conditions identifies the parameter vector. The presence of EIV makes the question of identification even harder.

We attempt to address this concern and make the above intuitions more precise in two ways. First, in the following subsection we consider a specific (but frequently employed) kind of an instrument: a second measurement (possibly non-classical). The specific form of the excluded variable allows us to provide more transparent identification conditions. Second, in Evdokimov and Zeleneev (2022) we study nonparametric regression model with EIV using the $\tau_n \to 0$ approximation. We show that the model is identified using an instrument (even a discrete one), and motivate the MERM approach from a nonparametric perspective. Finally, since MERM estimator is a standard GMM estimator, one can test the strength of identification of the model parameters, or conduct identification-robust inference using the standard methods (e.g., Stock and Wright, 2000; Kleibergen, 2005; Guggenberger and Smith, 2005; Guggenberger, Ramalho, and Smith, 2012; Andrews and Mikusheva, 2016; An-

drews, 2016; Andrews and Guggenberger, 2019).

**Identification Using A Second Measurement** Suppose we observe a second measurement

$$Q_i = \alpha_1 X_i^* + \varepsilon_{Q,i},$$

where $\alpha_1$ may not be known. Assume that $\alpha_1 \neq 0$ and $\mathbb{E}\left[\varepsilon_{Q,i} | X_i^*, S_i, \varepsilon_i\right] = 0$. The variance of $\varepsilon_{Q,i}$ does not need to be small. Note that the measurement error in $Q_i$ can be non-classical: $Q_i - X_i^*$ and $X_i^*$ are correlated unless $\alpha_1 = 1$.

Consider the conditional moment restrictions (15). If $X_i^*$ were observed, we could have constructed the unconditional moments

$$\mathbb{E}\left[h\left(X_i^*, S_i, \theta_0\right)\right] = 0, \qquad h\left(x, s, \theta\right) \equiv u\left(x, s, \theta\right) \times \left(1, x, \ldots, x^J\right)',$$

for some $J \geq \dim\left(\theta\right) - 1$. Suppose that the model is identified if $X_i^*$ observed, which means that the Jacobian of these moment conditions has full rank:

$$\mathrm{Rk}\left(H^*\right) = \dim\left(\theta\right), \text{ where } H^* \equiv E\left[\nabla_\theta h\left(X_i^*, S_i, \theta_0\right)\right].$$

To deal with the error-laden $X_i$, consider the MERM estimator with $K = 2$ based on the following moment function

$$g\left(x, s, q, \theta\right) \equiv \begin{pmatrix} h\left(x, s, \theta\right) \\ u\left(x, s, \theta\right) q \times \left(1, x, \ldots, x^{J-1}\right)' \end{pmatrix}. \tag{16}$$

Here the total number of moments is $m = 2J + 1$. The additional $J$ moments added in equation (16) use $Q_i$, which will allow identifying $\gamma_0 = \mathbb{E}[\varepsilon_i^2]/2$.

It turns out that in these settings there is a simple sufficient condition for Assumption ID(i) to hold. Appendix E demonstrates that $\Psi^*$ will have full rank if

$$\mathbb{E}\left[u_x^{(1)}\left(X_i^*, S_i, \theta_0\right) \times \left(1, X_i^*, \ldots, \left(X_i^*\right)^{J-1}\right)'\right] \neq 0. \tag{17}$$

Condition (17) has a very simple interpretation: it essentially it means that $\mathbb{E}\left[u_x^{(1)}\left(X_i^*, S_i, \theta_0\right) \middle| X_i^*\right]$ should not be identically zero. For example, in the nonlinear regression model $u\left(x, y, \theta\right) = \rho\left(x, \theta\right) - y$, and condition (17) is satisfied as long as $\mathbb{E}\left[\rho_x^{(1)}\left(X_i^*, \theta_0\right)\left(X_i^*\right)^j\right] \neq 0$ for some $j \in \{0, \ldots, J-1\}$.

# 3 Numerical Evidence

## 3.1 Comparison with a Semi-Nonparametric Estimation Approach

We compare MERM estimator with the state-of-the-art semiparametric estimator of Schennach (2007, henceforth S07) for nonlinear regression models. The Monte Carlo designs are taken from S07, and include a polynomial, rational fraction, and Probit nonlinear regression models. Identification of the model is ensured by the availability of an instrument.

$$Y_i = \rho(X_i^*, \theta_0) + U_i, \quad X_i^* = \pi_1 Z_i + V_i, \quad X_i = X_i^* + \varepsilon_i, \tag{18}$$

$(Z_i, V_i, \varepsilon_i)' \sim N\left((0,0,0)', \text{Diag}(1, 1/4, 1/4)\right)$, $\pi_1 = 1$, and $n = 1000$. The conditional expectation function $\rho$, the true value of the parameter of interest $\theta_0$, and the conditional distribution of the regression error $U_i$ are design-specific and reported in Tables 1-3 below. In all designs, $\tau = \sigma_\varepsilon / \sigma_X^* \approx 0.45$, so the measurement error is "fairly large" (Schennach, 2007).

We report simulation results for the MERM estimator considering correction schemes with $K = 2$ and $K = 4$. The original moment function is

$$g(x, y, z, \theta) = (y - \rho(x, \theta))\varphi(x, z),$$

where we use $\varphi(x, z) = (1, x, z, x^2, z^2, x^3, z^3)'$ for $K = 2$ and $\varphi(x, z) = (1, x, z, x^2, xz, z^2, x^3, x^2 z, xz^2, z^3)'$ for $K = 4$.

The finite sample properties of the MERM estimators (evaluated based on 5,000 replications) are reported in Tables 1-3 below. For comparison, we also provide the same statistics for naive estimators (OLS/NLLS) and for the benchmark estimator of S07 (as reported in the original paper). For the polynomial model (Table 1), both $K = 2$ and $K = 4$ MERM estimators effectively remove the EIV bias. Componentwise, the MERM estimators perform similarly (for $\theta_2$ and $\theta_4$) or better (for $\theta_1$ and $\theta_3$) compared to the benchmark estimator of S07. For the rational fraction model (Table 2), both the MERM estimators are vastly superior to the benchmark estimator both in terms of the bias and the standard deviation. For the probit model (Table 3), the MERM estimator with $K = 2$ removes a large fraction of the EIV bias compared to the NLLS estimator. However, the EIV bias remains non-negligible when this

simplest correction scheme is used. Employing a higher order correction scheme with $K = 4$ completely eliminates the remaining EIV bias, while at the same time having smaller standard deviations (than the benchmark estimator of S07) . Overall, in the considered designs, the MERM estimator with $K = 4$ consistently outperforms the benchmark estimator. It also proves to be more effective in removing the EIV bias compared to the $K = 2$ estimator, especially in the highly nonlinear settings of the considered probit design.

Table 1: Simulation results for the polynomial model of S07

| | Bias | | | | Std. Dev. | | | | RMSE | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $\theta_1$ | $\theta_2$ | $\theta_3$ | $\theta_4$ | $\theta_1$ | $\theta_2$ | $\theta_3$ | $\theta_4$ | $\theta_1$ | $\theta_2$ | $\theta_3$ | $\theta_4$ | All |
| OLS | -0.00 | -0.43 | 0.00 | 0.21 | 0.07 | 0.13 | 0.06 | 0.04 | 0.07 | 0.45 | 0.06 | 0.22 | 0.51 |
| S07 | -0.05 | -0.07 | -0.02 | 0.05 | 0.17 | 0.19 | 0.24 | 0.05 | 0.17 | 0.20 | 0.24 | 0.07 | 0.36 |
| $K = 2$ | -0.00 | 0.10 | 0.00 | 0.00 | 0.10 | 0.23 | 0.10 | 0.08 | 0.10 | 0.25 | 0.10 | 0.08 | 0.29 |
| $K = 4$ | -0.00 | 0.00 | 0.00 | 0.02 | 0.09 | 0.21 | 0.10 | 0.08 | 0.09 | 0.21 | 0.10 | 0.08 | 0.27 |

The DGP is as in (18) with $\rho(x, \theta) = \theta_1 + \theta_2 x + \theta_3 x^2 + \theta_4 x^3$, $\theta_0 = (1, 1, 0, -0.5)'$, and $U_i \sim N(0, 1/4)$.

Table 2: Simulation results for the rational fraction model of S07

| | Bias | | | Std. Dev. | | | RMSE | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | $\theta_1$ | $\theta_2$ | $\theta_3$ | $\theta_1$ | $\theta_2$ | $\theta_3$ | $\theta_1$ | $\theta_2$ | $\theta_3$ | All |
| OLS | 0.339 | -0.167 | -0.644 | 0.040 | 0.020 | 0.076 | 0.341 | 0.168 | 0.648 | 0.752 |
| S07 | 0.107 | 0.117 | -0.150 | 0.146 | 0.139 | 0.328 | 0.181 | 0.182 | 0.361 | 0.443 |
| $K = 2$ | -0.004 | -0.018 | 0.014 | 0.062 | 0.026 | 0.139 | 0.062 | 0.032 | 0.139 | 0.156 |
| $K = 4$ | 0.014 | -0.002 | -0.024 | 0.062 | 0.031 | 0.154 | 0.063 | 0.031 | 0.156 | 0.171 |

The DGP is as in (18) with $\rho(x, \theta) = \theta_1 + \theta_2 x + \frac{\theta_3}{(1+x^2)^2}$, $\theta_0 = (1, 1, 2)'$, and $U_i \sim N(0, 1/4)$.

## 3.2 Estimation and Inference in a Multinomial Choice Model

Consider the standard multinomial logit model, in which an agent chooses between 3 available options. For an agent $i$ with characteristics $(X_i^*, W_i)$, the utility of option $j$ is given by

$$U_{ij} = \theta_{0j1} X_i^* + \theta_{0j2} W_{ij} + \theta_{0j3} + \epsilon_{ij} \quad \text{for } j \in \{1, 2\},$$

Table 3: Simulation results for the Probit model of S07

| | Bias | | Std. Dev. | | RMSE | | |
|---|---|---|---|---|---|---|---|
| | $\theta_1$ | $\theta_2$ | $\theta_1$ | $\theta_2$ | $\theta_1$ | $\theta_2$ | All |
| NLLS | 0.38 | -0.97 | 0.06 | 0.08 | 0.39 | 0.98 | 1.05 |
| S07 | 0.05 | -0.06 | 0.39 | 0.53 | 0.39 | 0.53 | 0.69 |
| $K = 2$ | 0.11 | -0.31 | 0.18 | 0.34 | 0.21 | 0.46 | 0.51 |
| $K = 4$ | -0.01 | -0.01 | 0.23 | 0.42 | 0.23 | 0.42 | 0.48 |

The DGP is as in (18) with $\rho(x,\theta) = \frac{1}{2}(1 + \text{erf}(\theta_1 + \theta_2 x))$, $\theta_0 = (-1, 2)'$, and $U_i = 1 - \rho(X_i^*, \theta_0)$ with probability $\rho(X_i^*, \theta_0)$ and $-\rho(X_i^*, \theta_0)$ otherwise.

and $U_{i0} = \epsilon_{i0}$ for the outside option $j = 0$, where $\epsilon_{ij}$ are i.i.d. (across $i$ and $j$) draws from a standard type-1 extreme value distribution. The researcher observes $\{(X_i, W_i, Y_{i1}, Y_{i2}, Y_{i0})\}_{i=1}^n$, where $Y_{ij}$ is a binary variable indicating whether agent $i$ chooses option $j$, i.e. $Y_{ij} = 1$ if and only if $j = \text{argmax}_{j' \in \{0,1,2\}} U_{ij'}$. In addition,

$$X_i^* = V_{i1} Z_i + V_{i0}, \quad X_i = X_i^* + \varepsilon_i, \quad W_{ij} = \rho X_i^* / \sigma_X^* + \sqrt{1 - \rho^2}\nu_{ij},$$

and $(V_{i1}, V_{i0}, Z_i, \varepsilon_i, \nu_{i1}, \nu_{i2})' \sim N\left((1,0,0,0,0,0)', \text{Diag}(\sigma_{V1}^2, \sigma_{V0}^2, \sigma_Z^2, \sigma_\varepsilon^2, \sigma_\nu^2, \sigma_\nu^2)\right)$. In all of the designs, we fix $(\theta_{011}, \theta_{012}, \theta_{013}, \theta_{021}, \theta_{022}, \theta_{023}, \rho, \sigma_{V1}^2, \sigma_{V0}^2, \sigma_Z^2, \sigma_\nu^2) = (1, 0, 0, 0, 0, 0, 0.7, 1/2, 1/2, 1, 1)$ and $n = 2000$. We consider $\tau = \sigma_\varepsilon / \sigma_{X^*} \in \{1/4, 1/2, 3/4\}$. Setting $\sigma_{V1} = 0$ would correspond to the additive control variable model. We omit such simulation results for brevity.

Similarly to Section 3.1, we report results for the MERM estimators with $K = 2$ and $K = 4$ based on the following original moment function

$$g(x, w, y, z, \theta) = \left((y_1 - p_1(x, w, \theta))\, \varphi_1(x, z, w)', (y_2 - p_2(x, w, \theta))\, \varphi_2(x, z, w)'\right)',$$
$$p_j(x, w, \theta) = \frac{\exp(\theta_{j1} x + \theta_{j2} w_j + \theta_{j3})}{1 + \exp(\theta_{11} x + \theta_{12} w_1 + \theta_{13}) + \exp(\theta_{21} x + \theta_{22} w_2 + \theta_{23})},$$

where $\varphi_j(x, z, w) = (1, x, z, x^2, z^2, x^3, z^3, w_j)'$ for $K = 2$ and $\varphi_j(x, z, w) = (1, x, z, x^2, xz, z^2, x^3, x^2 z, x z^2, z^3, w_j)'$ for $K = 4$.

We report the results on estimation and inference on the partial derivatives of the conditional choice probabilities $p_j(x, w_1, w_2)$ with respect to $x$, $w_1$, and $w_2$, evaluated at the population means.

Table 4 reports the finite sample biases, standard deviations, and RMSE of the MERM estimators, as well as the sizes of the corresponding t-tests with nominal size of 5%. To illustrate the importance of dealing with EIV, we also report the same

statistics for the standard (naive) MLE estimator that ignores the presence of the measurement errors.

In all designs, the MLE estimator is biased, and the corresponding t-tests over-reject. Note that failing to account for the EIV in the mismeasured variable $X_i^*$ generally biases estimators of all of the parameter, including those corresponding to the correctly measured variables $W_{i1}$ and $W_{i2}$. In particular, the t-tests may falsely reject true null hypotheses $\partial p_j/\partial w_\ell = 0$ up to nearly 100% of the time.

The MERM estimator with $K = 2$ removes a large fraction of the EIV bias in all of the designs. While this proves to be enough to achieve accurate size control when the magnitude of the measurement error is moderate ($\tau = 1/4$), the remaining EIV bias may still result in size distortions of the t-tests with larger measurement errors, especially $\tau = 3/4$. Using the higher order correction scheme with $K = 4$ effectively removes the EIV bias in all of the simulation designs for all of the parameters. Remarkably, the corresponding finite sample null rejection probabilities remain close to the nominal 5% rate even when the standard deviation of the measurement error is as large as 75% of the standard deviation of the mismeasured $X^*$.

To further check the limits of applicability of our method, we also consider larger values of $\tau \in \{1, 3/2, 2\}$. The numerical results analogous to the ones reported in Table 4 are provided in Table 7 in Appendix F.1. Specifically, we find that inference results based on the correction scheme with $K = 4$ remain accurate even for $\tau = 1$. Unsurprisingly, inference becomes less reliable for bigger $\tau = 3/2$ and $\tau = 2$. At the same time, while the $K = 4$ correction scheme fails to entirely eliminate the EIV bias in these designs, it still removes a big fraction of the bias and greatly improves on MLE in terms of the RMSE. Thus, while inference based on our estimator might be less reliable in extreme settings when the measurement error overwhelms the signal, the MERM estimator using $K = 4$ appears to be sufficiently accurate over a wide range of $\tau$.

## 3.3 Empirical Illustration: Choice of Transportation Mode

In this section, we illustrate the finite sample properties of the MERM estimator in the context of a classical multinomial choice application: choice of transportation mode (e.g., McFadden, 1974).

To calibrate the numerical experiment, we use the ModeCanada dataset, a survey

Table 4: Simulation results for the multinomial logit model

| | MLE | | | | $K=2$ | | | | $K=4$ | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | bias, $10^{-2}$ | std, $10^{-2}$ | rmse, $10^{-2}$ | size | bias, $10^{-2}$ | std, $10^{-2}$ | rmse, $10^{-2}$ | size | bias, $10^{-2}$ | std, $10^{-2}$ | rmse, $10^{-2}$ | size |
| | | | | | | $\tau = 1/4$ | | | | | | |
| $\partial p_1/\partial x$ | -3.24 | 1.36 | 3.51 | 66.98 | 0.74 | 2.63 | 2.74 | 4.30 | 1.13 | 2.73 | 2.95 | 7.86 |
| $\partial p_1/\partial w_1$ | 2.32 | 1.64 | 2.84 | 30.74 | -0.11 | 2.30 | 2.30 | 4.82 | -0.31 | 2.29 | 2.31 | 6.54 |
| $\partial p_1/\partial w_2$ | 0.48 | 0.75 | 0.90 | 9.40 | -0.04 | 0.87 | 0.87 | 4.82 | -0.08 | 0.87 | 0.87 | 5.36 |
| $\partial p_2/\partial x$ | 1.96 | 1.17 | 2.28 | 39.44 | -0.40 | 1.88 | 1.92 | 4.72 | -0.63 | 1.93 | 2.03 | 6.74 |
| $\partial p_2/\partial w_1$ | -1.16 | 0.82 | 1.42 | 30.66 | 0.06 | 1.15 | 1.15 | 4.84 | 0.15 | 1.15 | 1.16 | 6.48 |
| $\partial p_2/\partial w_2$ | -0.96 | 1.50 | 1.78 | 9.48 | 0.09 | 1.74 | 1.74 | 4.98 | 0.17 | 1.74 | 1.75 | 5.44 |
| $\partial p_0/\partial x$ | 1.28 | 1.01 | 1.63 | 25.28 | -0.34 | 1.43 | 1.47 | 5.08 | -0.50 | 1.48 | 1.56 | 7.36 |
| $\partial p_0/\partial w_1$ | -1.16 | 0.82 | 1.42 | 30.60 | 0.06 | 1.15 | 1.15 | 4.82 | 0.15 | 1.15 | 1.16 | 6.46 |
| $\partial p_0/\partial w_2$ | 0.48 | 0.74 | 0.88 | 9.46 | -0.05 | 0.87 | 0.87 | 4.90 | -0.09 | 0.87 | 0.88 | 5.32 |
| | | | | | | $\tau = 1/2$ | | | | | | |
| $\partial p_1/\partial x$ | -8.97 | 1.09 | 9.04 | 100.00 | -1.69 | 2.60 | 3.10 | 9.84 | 0.97 | 2.89 | 3.05 | 6.04 |
| $\partial p_1/\partial w_1$ | 6.44 | 1.53 | 6.62 | 98.96 | 1.39 | 2.44 | 2.81 | 12.14 | -0.21 | 2.41 | 2.42 | 5.54 |
| $\partial p_1/\partial w_2$ | 1.28 | 0.72 | 1.47 | 42.54 | 0.29 | 0.90 | 0.94 | 7.00 | -0.06 | 0.92 | 0.92 | 5.00 |
| $\partial p_2/\partial x$ | 5.22 | 0.97 | 5.31 | 99.98 | 1.05 | 1.89 | 2.16 | 10.16 | -0.53 | 2.08 | 2.15 | 6.14 |
| $\partial p_2/\partial w_1$ | -3.21 | 0.77 | 3.30 | 98.96 | -0.69 | 1.22 | 1.40 | 12.10 | 0.10 | 1.21 | 1.21 | 5.52 |
| $\partial p_2/\partial w_2$ | -2.52 | 1.41 | 2.88 | 42.82 | -0.56 | 1.79 | 1.87 | 7.20 | 0.13 | 1.84 | 1.84 | 4.98 |
| $\partial p_0/\partial x$ | 3.75 | 0.86 | 3.85 | 98.78 | 0.64 | 1.45 | 1.59 | 8.18 | -0.44 | 1.59 | 1.65 | 6.50 |
| $\partial p_0/\partial w_1$ | -3.23 | 0.78 | 3.32 | 98.96 | -0.70 | 1.22 | 1.41 | 12.06 | 0.10 | 1.21 | 1.21 | 5.48 |
| $\partial p_0/\partial w_2$ | 1.23 | 0.69 | 1.41 | 42.90 | 0.28 | 0.89 | 0.93 | 7.20 | -0.07 | 0.92 | 0.92 | 4.94 |
| | | | | | | $\tau = 3/4$ | | | | | | |
| $\partial p_1/\partial x$ | -13.35 | 0.86 | 13.38 | 100.00 | -6.83 | 2.64 | 7.32 | 80.32 | 0.71 | 3.22 | 3.29 | 4.74 |
| $\partial p_1/\partial w_1$ | 9.69 | 1.45 | 9.80 | 100.00 | 4.95 | 2.65 | 5.61 | 65.52 | 0.01 | 2.62 | 2.62 | 5.34 |
| $\partial p_1/\partial w_2$ | 1.81 | 0.69 | 1.94 | 75.30 | 1.01 | 0.89 | 1.35 | 26.08 | -0.01 | 0.98 | 0.98 | 5.24 |
| $\partial p_2/\partial x$ | 7.48 | 0.79 | 7.52 | 100.00 | 4.06 | 1.83 | 4.45 | 68.82 | -0.37 | 2.32 | 2.35 | 5.74 |
| $\partial p_2/\partial w_1$ | -4.83 | 0.73 | 4.88 | 100.00 | -2.47 | 1.32 | 2.81 | 65.46 | -0.01 | 1.31 | 1.31 | 5.28 |
| $\partial p_2/\partial w_2$ | -3.51 | 1.33 | 3.76 | 75.60 | -1.99 | 1.76 | 2.66 | 26.38 | 0.03 | 1.97 | 1.97 | 5.32 |
| $\partial p_0/\partial x$ | 5.87 | 0.73 | 5.92 | 100.00 | 2.77 | 1.47 | 3.14 | 56.28 | -0.34 | 1.77 | 1.80 | 5.82 |
| $\partial p_0/\partial w_1$ | -4.87 | 0.75 | 4.93 | 100.00 | -2.48 | 1.33 | 2.81 | 65.40 | -0.01 | 1.31 | 1.31 | 5.32 |
| $\partial p_0/\partial w_2$ | 1.70 | 0.64 | 1.82 | 75.76 | 0.98 | 0.87 | 1.31 | 26.50 | -0.02 | 0.99 | 0.99 | 5.30 |

This table reports the simulated finite sample bias, standard deviation, RMSE, and size of the MLE and the MERM estimators and the corresponding t-tests for the partial derivatives $\partial p_j(x, w, \theta_0)/\partial x$, $\partial p_j(x, w, \theta_0)/\partial w_1$, $\partial p_j(x, w, \theta_0)/\partial w_2$ for $j \in \{1, 2, 0\}$ evaluated at the population mean. The true values of the marginal effects are $(\partial p_1/\partial x, \partial p_2/\partial x, \partial p_0/\partial x) = (0.222, -0.111, -0.111)$ and zeros for the rest. The results are based on 5,000 replications.

of business travelers for the Montreal-Toronto corridor. We focus on the subset of travelers choosing between train, air, and car ($n = 2769$), and estimate the conditional logit model with traveler $i$'s utilities given in the table below.

| Mode | Utility |
|---|---|
| Air | $U_{i1} = \theta_{01}\, Income_i^* + \theta_{02}\, Urban_i + \theta_{03} + \theta_{07}\, Price_{i1} + \theta_{08}\, InTime_{i1} + \epsilon_{i1}$ |
| Car | $U_{i2} = \theta_{04}\, Income_i^* + \theta_{05}\, Urban_i + \theta_{06} + \theta_{07}\, Price_{i2} + \theta_{08}\, InTime_{i2} + \epsilon_{i2}$ |
| Train | $U_{i0} = \theta_{07}\, Price_{i0} + \theta_{08}\, InTime_{i0} + \epsilon_{i0}$ |

To generate the simulated samples, we randomly draw covariates from their joint empirical distribution. To generate the simulated outcomes, we draw $\epsilon_{ij}$ from the standard type-I extreme value distribution. The true value of $\theta_0$ is set to be the MLE estimate based on the original dataset. More details about this numerical experiment are given in Appendix I.

To evaluate the performance of the MERM estimator in these settings, we generate mismeasured $Income_i = Income_i^* + \varepsilon_i$. We focus on the individual income because it is often mismeasured.We report the results for $\tau = \sigma_\varepsilon / \sigma_{Income^*} \in \{1/4, 1/2, 3/4\}$.

Table 5 reports the simulation results for the (naive) MLE estimator and for the MERM estimators with $K = 2$ and $K = 4$. We focus on estimation of and inference on the income elasticities (evaluated at the population mean of the covariates). The MLE estimator is considerably biased for $\tau \in \{1/2, 3/4\}$, which results in substantial size distortions of the MLE based t-tests. The MERM estimator with $K = 4$ effectively eliminates the EIV bias and the corresponding t-tests provide accurate size control in all of the considered designs. The estimator with $K = 2$ is more precise, while successfully removing the EIV bias for $\tau \leq 1/2$.

Overall, the MERM estimators perform well in the considered empirical context, providing a basis for estimation and inference even for quite large values of $\tau$.

Table 5: Simulation results for the empirically calibrated conditional logit model

| | MLE | | | | $K=2$ | | | | $K=4$ | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | bias | std | rmse | size | bias | std | rmse | size | bias | std | rmse | size |
| | | | | | $\tau = 1/4$ | | | | | | | |
| $\partial \ln p_1 / \partial \ln I$ | -0.07 | 0.12 | 0.14 | 9.00 | 0.01 | 0.14 | 0.14 | 5.68 | 0.02 | 0.19 | 0.19 | 7.02 |
| $\partial \ln p_2 / \partial \ln I$ | 0.03 | 0.07 | 0.08 | 5.84 | -0.00 | 0.08 | 0.08 | 5.68 | -0.01 | 0.10 | 0.10 | 6.40 |
| $\partial \ln p_0 / \partial \ln I$ | 0.05 | 0.13 | 0.13 | 6.10 | 0.00 | 0.14 | 0.14 | 5.42 | -0.00 | 0.17 | 0.17 | 7.66 |
| | | | | | $\tau = 1/2$ | | | | | | | |
| $\partial \ln p_1 / \partial \ln I$ | -0.24 | 0.11 | 0.27 | 61.84 | -0.05 | 0.14 | 0.15 | 6.96 | 0.02 | 0.21 | 0.21 | 6.16 |
| $\partial \ln p_2 / \partial \ln I$ | 0.09 | 0.07 | 0.11 | 24.76 | 0.02 | 0.09 | 0.09 | 5.96 | -0.01 | 0.10 | 0.10 | 6.16 |
| $\partial \ln p_0 / \partial \ln I$ | 0.16 | 0.12 | 0.20 | 25.36 | 0.04 | 0.15 | 0.15 | 6.06 | -0.00 | 0.18 | 0.18 | 6.86 |
| | | | | | $\tau = 3/4$ | | | | | | | |
| $\partial \ln p_1 / \partial \ln I$ | -0.43 | 0.09 | 0.44 | 99.50 | -0.19 | 0.14 | 0.24 | 27.46 | 0.02 | 0.22 | 0.22 | 5.84 |
| $\partial \ln p_2 / \partial \ln I$ | 0.16 | 0.06 | 0.17 | 71.88 | 0.07 | 0.08 | 0.11 | 13.78 | -0.01 | 0.11 | 0.11 | 6.32 |
| $\partial \ln p_0 / \partial \ln I$ | 0.29 | 0.11 | 0.31 | 73.20 | 0.12 | 0.15 | 0.19 | 13.40 | 0.00 | 0.19 | 0.19 | 6.32 |

This table reports the simulated finite sample bias, standard deviation, RMSE, and size of the MLE and the MERM estimators and the corresponding t-tests for the income elasticities $\partial \ln p_j(I, w, \theta_0)/\partial \ln I$, $j \in \{1, 2, 0\}$, evaluated at the population mean. The true values of the income elasticities are $(\partial \ln p_1/\partial \ln I, \partial \ln p_2/\partial \ln I, \partial \ln p_0/\partial \ln I) = (1.11, -0.39, -0.82)$. The results are based on 5,000 replications.

# 4 Extensions

## 4.1 Data-driven choice of $K$

Making an appropriate choice of the expansion order $K$ is important for the estimation procedure. One has to be cautious not to take $K$ too small, as this may result in an estimator that only partially removes the EIV bias. On the other hand, picking a larger $K$ than needed might inflate standard errors and result in less powerful inference.

In this section, we address this issue by providing a data-dependent procedure for selecting $K$. We demonstrate that our procedure has desirable theoretical properties. We also find that the procedure has good finite sample properties in a set of Monte Carlo simulation experiments across different values of $\tau$ and sample sizes.

Consider two alternative values of the expansion order: $L$ and $K$, where $2 \leq L < K$. In practice, even-order biases tend to dominate, so to reduce the set of choices, it is useful to focus on even values of the expansion orders $L$ and $K$. For example, one

would typically be interested in choosing between $L = 2$ and $K = 4$.

Let $\hat{\beta}_L$ and $\hat{\beta}_K$ denote the corresponding MERM estimators. The estimator $\hat{\beta}_L$ should be preferred as having smaller asymptotic variance provided that its remaining EIV bias is negligible relative to its standard error. Otherwise, the more conservative $\hat{\beta}_K$ should be used instead.

Note that Lemma 1 suggests that the remaining asymptotic bias of $\hat{\beta}_L$ (due to the additional terms accounted for when the expansion of higher order $K$ is used) is given by (up to an $o(n^{-1/2})$ remainder)

$$\text{AsyB}(\hat{\beta}_L) \equiv B \sum_{k=L+1}^{K} \gamma_{0k} \mathbb{E}[g_x^{(k)}(X_i, S_i, \theta_0)], \quad B \equiv -(\Psi' \Xi \Psi)^{-1} \Psi' \Xi,$$

where matrix $B$ is based on the moments used for estimation of $\hat{\beta}_L$.

Importantly, $\gamma_{0k} = O\left(\sigma_\varepsilon^k\right)$, and hence for $k > 2$ we can estimate a bound on $\sqrt{n}\gamma_{0k}$ sufficiently quickly to provide a valid procedure for choosing $K$. To this end, we first estimate the model using the larger $K$. Let $\hat{\beta}_K = (\hat{\theta}', \hat{\gamma}')'$ and $\hat{\sigma}_\varepsilon^2 \equiv 2\hat{\gamma}_2$, where we dropped the additional subscripts $K$ for notation simplicity. Then, we can estimate $\sigma_\varepsilon^K$ by $\hat{\sigma}_\varepsilon^K \equiv (\hat{\sigma}_\varepsilon^2)^{K/2}$. Using $\hat{\sigma}_\varepsilon^2 = \sigma_\varepsilon^2 + O_p(n^{-1/2})$, in the appendix we show that

$$\sqrt{n}\left(\hat{\sigma}_\varepsilon^K - \sigma_\varepsilon^K\right) = O_p\left(\sigma_\varepsilon^{K-2} + n^{-(K-2)/4}\right) = o_p(1). \tag{19}$$

Next, consider a sequence $\varkappa_n \to 0$, which we will specify precisely later, and let

$$\delta_n = 1\left\{\max_{1 \le \ell \le \dim(\beta)} \left|\hat{\Sigma}_{\ell\ell}^{-1/2} \sqrt{n} \hat{\sigma}_\varepsilon^K \hat{B}_{\ell\cdot} \overline{g}_x^{(K)}(\hat{\theta})\right| \le c_K \varkappa_n\right\}, \tag{20}$$

where $\hat{\Sigma}$ and $\hat{B}$ are consistent estimators of the asymptotic variance of $\hat{\beta}_L$ and of $B$, with $\hat{\Sigma}_{\ell\ell}$ denoting its $\ell$-th diagonal element, $\hat{B}_{\ell\cdot}$ denoting its $\ell$-th row, $\overline{g}_x^{(K)}(\hat{\theta}) \equiv n^{-1}\sum_{i=1}^n g^{(K)}(X_i, S_i, \hat{\theta})$, and $c_K > 0$ is a constant that we will calibrate below after we state the main theoretical result of this section.

If $\delta_n = 1$, the researchers should select $\hat{\beta}_L$. Otherwise, $\hat{\beta}_K$ should be used. The following lemma demonstrates that the proposed selection procedure has desirable theoretical properties.

**Lemma 3.** *Suppose the hypotheses of Theorem 2 hold for some even $K \ge 4$, and consider $L$ satisfying $2 \le L < K$. Suppose $\varkappa_n n^{(K-L-1)/(2L+2)} \to 0$. Then $\sqrt{n}\text{AsyB}(\hat{\beta}_L)\delta_n = o_p(1)$.*

*Moreover, consider $\varkappa_n = n^{-(K-L-1)/(2L+2)} (\ln n)^{-a}$ for any $a > 0$. Then the crite-*

*rion is consistent, in the sense that if* $\tau_n = o(n^{-\frac{1}{2L+2}-\epsilon})$ *for any* $\epsilon > 0$*, the criterion will choose* $\hat{\beta}_L$ *with probability approaching one.*

The first part of Lemma 3 shows that, provided that $\varkappa_n$ goes to zero sufficiently fast, $\hat{\beta}_L$ is selected (i.e., $\delta_n = 1$) only when its asymptotic bias is negligible. Next, note that $\hat{\beta}_L$ is asymptotically unbiased as long as $\tau_n = (n^{-\frac{1}{2L+2}})$. The second part of the lemma shows that, for the suggested choices of $\varkappa_n$, the criterion is non-vacuous, i.e., that it does select the MERM estimator with a smaller expansion order $L$ when this is appropriate. Typically, one would pick $L = K - 2$, so the lemma suggests taking $\varkappa_n = n^{-1/(2K-2)} (\ln n)^{-a}$ in this case.

In practice, it is important to pick an appropriate constant $c_K$ used in the construction of $\delta_n$ in equation (20). When constructing $\delta_n$, we used $|\gamma_{0K}| \propto \sigma_\varepsilon^K$, which motivates choosing $c_K = \sigma_\varepsilon^K / |\gamma_{0K}|$. It is convenient to use a rule-of-thumb approach, using a reference distribution to determine $c_K$. It turns out that the normal distribution is not only convenient, but also sufficiently conservative (note that the bigger $\gamma_{0K}$ is, the smaller $c_K$ is, resulting in a more conservative selection procedure). For example, suppose $K = 4$, and consider using Student's $t(\nu)$ distribution as the reference distribution. Then, for $\nu \geq 5$, the biggest $|\gamma_{04}| / \sigma_\varepsilon^4$ and the smallest $c_4$ correspond to $\nu = \infty$ matching the normal distribution.[11]

Since for the normal distribution we have $\gamma_{0K} = \sigma_\varepsilon^K / K!!$ for even $K$, we recommend using $c_K = K!!$ in equation (20). Finally, while we recommend using normal distribution as the reference distribution, we also stress that the results of Lemma 3 hold even if the measurement error is not normal or if it is skewed.

We summarize the proposed procedure in the following suggested algorithm.

**Algorithm 1.**

1. *Pick a large enough even* $K \geq 4$ *and compute* $\hat{\beta}_K = (\hat{\theta}', \hat{\gamma}')'$ *and* $\hat{\sigma}_\varepsilon^2 = 2\hat{\gamma}_2$.
2. *Compute* $\delta_n$ *in* (20) *with* $L = K - 2$, $c_K = K!!$, *and* $\varkappa_n = (n \ln n)^{-1/(2K-2)}$.
3. *Pick the length of expansion* $L = K - 2$ *if* $\delta_n = 1$, *otherwise keep the initial* $K$.

*This procedure can be iterated if desired.*

[11] Note that Student's $t(\nu)$ distribution does not have a finite 5-th moment $\nu \leq 5$.

To illustrate the performance of the algorithm provided above, we revisit the numerical experiment considered in Section 3.2. We focus on choosing between $K = 4$ and $L = 2$, and, as in Section 3.2, we report results for $n = 2000$ in Table 6 below. To ensure that the proposed algorithm performs well in a variety of sample sizes, we also consider $n = 1000$ and $n = 4000$ and report the corresponding results in Tables 8 and 9 in Appendix F.2.

We report the finite sample bias and RMSE for the naive MLE estimator, as well as for the MERM estimators using $K = 2$ and $K = 4$, and for the adaptive MERM estimator using data-driven $K$ following Algorithm 1. Notice that for the considered sample sizes, $K = 2$ is preferred when $\tau = 1/4$, and $K = 4$ is preferred when $\tau = 3/4$. We find that in both of these regimes and for all sample sizes, the adaptive estimator using data-driven $K$ is essentially equivalent to the preferred estimators, i.e., our procedure selects the appropriate $K$. Interestingly, in the intermediate regime with $\tau = 1/2$, the adaptive estimator also has the smallest bias among the considered estimators, coming at the cost of a slightly bigger RMSE compared to the MERM estimator using $K = 4$. Thus, the considered numerical experiment suggests that our algorithm has good finite sample properties supporting the findings of Lemma 3.

## 4.2 Multiple Mismeasured Variables

It is easy to use the MERM framework to deal with multiple mismeasured variables. This is useful in many applications, including not only settings with multiple mismeasured covariates, but also settings with serially correlated measurement errors, settings where repeated measurements are available, and panel data models. Using the MERM approach is particularly advantageous in such applications, since it avoids nonparametric estimation of multivariate unobserved distributions.

Suppose $X_i^*$, $\varepsilon_i$, and $X_i$ are $d \times 1$ vectors. Let $\tau_n \equiv \max_{j \leq d} \sigma_{\varepsilon_j} / \sigma_{X_j^*}$, where $\sigma_{\varepsilon_j}$ and $\sigma_{X_j^*}$ denote the standard deviations of the $j$-th components of $\varepsilon_i$ and $X_i^*$, so $\mathbb{E}\left[|\varepsilon_{ij}|^k\right] = O(\tau_n^k)$ for $k \in \{1, \ldots, K\}$.

For a $d \times 1$ vector of non-negative integers $\kappa = (\kappa_1, \ldots, \kappa_d) \in \mathbb{Z}_+^d$, let

$$\partial_\kappa \equiv \frac{\partial^{|\kappa|}}{\partial {x_1}^{\kappa_1} \ldots \partial {x_d}^{\kappa_d}}, \quad \text{where } |\kappa| \equiv \sum_{j=1}^{d} \kappa_j.$$

Also, for a positive integer $k$, let $\mathcal{K}_k = \{\kappa \in \mathbb{Z}_+^d : |\kappa| = k\}$. Then, we consider the

Table 6: Choice of $K$ simulation results for the multinomial logit model, $n = 2000$

| | MLE | | $K = 2$ | | $K = 4$ | | data-driven $K$ | |
|---|---|---|---|---|---|---|---|---|
| | bias, $10^{-2}$ | rmse, $10^{-2}$ | bias, $10^{-2}$ | rmse, $10^{-2}$ | bias, $10^{-2}$ | rmse, $10^{-2}$ | bias, $10^{-2}$ | rmse, $10^{-2}$ |
| | | | | $\tau = 1/4$ | | | | |
| $\partial p_1/\partial x$ | -3.24 | 3.51 | 0.74 | 2.74 | 1.13 | 2.95 | 0.75 | 2.75 |
| $\partial p_1/\partial w_1$ | 2.32 | 2.84 | -0.11 | 2.30 | -0.31 | 2.31 | -0.11 | 2.30 |
| $\partial p_1/\partial w_2$ | 0.48 | 0.90 | -0.04 | 0.87 | -0.08 | 0.87 | -0.04 | 0.87 |
| $\partial p_2/\partial x$ | 1.96 | 2.28 | -0.40 | 1.92 | -0.63 | 2.03 | -0.41 | 1.93 |
| $\partial p_2/\partial w_1$ | -1.16 | 1.42 | 0.06 | 1.15 | 0.15 | 1.16 | 0.06 | 1.15 |
| $\partial p_2/\partial w_2$ | -0.96 | 1.78 | 0.09 | 1.74 | 0.17 | 1.75 | 0.09 | 1.74 |
| $\partial p_0/\partial x$ | 1.28 | 1.63 | -0.34 | 1.47 | -0.50 | 1.56 | -0.34 | 1.48 |
| $\partial p_0/\partial w_1$ | -1.16 | 1.42 | 0.06 | 1.15 | 0.15 | 1.16 | 0.06 | 1.15 |
| $\partial p_0/\partial w_2$ | 0.48 | 0.88 | -0.05 | 0.87 | -0.09 | 0.88 | -0.05 | 0.87 |
| | | | | $\tau = 1/2$ | | | | |
| $\partial p_1/\partial x$ | -8.97 | 9.04 | -1.69 | 3.10 | 0.97 | 3.05 | 0.88 | 3.15 |
| $\partial p_1/\partial w_1$ | 6.44 | 6.62 | 1.39 | 2.81 | -0.21 | 2.42 | -0.15 | 2.48 |
| $\partial p_1/\partial w_2$ | 1.28 | 1.47 | 0.29 | 0.94 | -0.06 | 0.92 | -0.05 | 0.93 |
| $\partial p_2/\partial x$ | 5.22 | 5.31 | 1.05 | 2.16 | -0.53 | 2.15 | -0.48 | 2.20 |
| $\partial p_2/\partial w_1$ | -3.21 | 3.30 | -0.69 | 1.40 | 0.10 | 1.21 | 0.08 | 1.24 |
| $\partial p_2/\partial w_2$ | -2.52 | 2.88 | -0.56 | 1.87 | 0.13 | 1.84 | 0.11 | 1.85 |
| $\partial p_0/\partial x$ | 3.75 | 3.85 | 0.64 | 1.59 | -0.44 | 1.65 | -0.40 | 1.68 |
| $\partial p_0/\partial w_1$ | -3.23 | 3.32 | -0.70 | 1.41 | 0.10 | 1.21 | 0.08 | 1.24 |
| $\partial p_0/\partial w_2$ | 1.23 | 1.41 | 0.28 | 0.93 | -0.07 | 0.92 | -0.06 | 0.93 |
| | | | | $\tau = 3/4$ | | | | |
| $\partial p_1/\partial x$ | -13.35 | 13.38 | -6.83 | 7.32 | 0.71 | 3.29 | 0.71 | 3.29 |
| $\partial p_1/\partial w_1$ | 9.69 | 9.80 | 4.95 | 5.61 | 0.01 | 2.62 | 0.01 | 2.62 |
| $\partial p_1/\partial w_2$ | 1.81 | 1.94 | 1.01 | 1.35 | -0.01 | 0.98 | -0.01 | 0.98 |
| $\partial p_2/\partial x$ | 7.48 | 7.52 | 4.06 | 4.45 | -0.37 | 2.35 | -0.37 | 2.35 |
| $\partial p_2/\partial w_1$ | -4.83 | 4.88 | -2.47 | 2.81 | -0.01 | 1.31 | -0.01 | 1.31 |
| $\partial p_2/\partial w_2$ | -3.51 | 3.76 | -1.99 | 2.66 | 0.03 | 1.97 | 0.03 | 1.97 |
| $\partial p_0/\partial x$ | 5.87 | 5.92 | 2.77 | 3.14 | -0.34 | 1.80 | -0.34 | 1.80 |
| $\partial p_0/\partial w_1$ | -4.87 | 4.93 | -2.48 | 2.81 | -0.01 | 1.31 | -0.01 | 1.31 |
| $\partial p_0/\partial w_2$ | 1.70 | 1.82 | 0.98 | 1.31 | -0.02 | 0.99 | -0.02 | 0.99 |

This table reports the simulated finite sample bias and RMSE of the MLE and the MERM estimators for the partial derivatives $\partial p_j(x, w, \theta_0)/\partial x$, $\partial p_j(x, w, \theta_0)/\partial w_1$, $\partial p_j(x, w, \theta_0)/\partial w_2$ for $j \in \{1, 2, 0\}$ evaluated at the population mean. The true values of the marginal effects are $(\partial p_1/\partial x, \partial p_2/\partial x, \partial p_0/\partial x) = (0.222, -0.111, -0.111)$ and zeros for the rest. The results are based on 5,000 replications.

following corrected moment function

$$\psi(x, s, \theta, \gamma) = g(x, s, \theta) - \sum_{k=2}^{K} \sum_{\kappa \in \mathcal{K}_k} \gamma_\kappa \partial_\kappa g(x, s, \theta),$$

where, with some abuse of notation, $\gamma$ is a collection of all $\gamma_\kappa$ with $\kappa \in \mathcal{K}_k$ and $k \in \{2, \dots, K\}$.

Under mild smoothness conditions

$$\mathbb{E}\left[\psi(X_i, S_i, \theta_0, \gamma_0)\right] = \mathbb{E}\left[g(X_i^*, S_i, \theta_0)\right] + O(\tau_n^{K+1}) = o(n^{-1/2}),$$

where the second equality holds provided that $O(\tau_n^{K+1}) = o(n^{-1/2})$. Similarly to the

scalar case, components of $\gamma_0$ are determined by the moments of $\varepsilon_i$. Specifically, let $\mu_\kappa \equiv \mathbb{E}\left[\varepsilon_{i1}^{\kappa_1} \ldots \varepsilon_{id}^{\kappa_d}\right]$, then

$$\gamma_{0\kappa} = \frac{\mu_\kappa}{\kappa!}, \quad \text{for } \kappa \in \{\mathcal{K}_2, \mathcal{K}_3\}, \tag{21}$$

where $\kappa! \equiv \kappa_1! \ldots \kappa_d!$. For $|\kappa| \geq 4$, the coefficients can be computed by the following formulas. For example, for $\kappa \in \mathcal{K}_4$, let $\mathcal{K}_{2,\kappa} = \{\tilde{\kappa} \in \mathcal{K}_2 : \kappa - \tilde{\kappa} \in \mathcal{K}_2\}$. Then,

$$\gamma_{0\kappa} = \frac{\mu_\kappa}{\kappa!} - \sum_{\tilde{\kappa} \in \mathcal{K}_{2,\kappa}} \frac{\mu_{\kappa-\tilde{\kappa}}}{(\kappa - \tilde{\kappa})!} \gamma_{0\tilde{\kappa}}, \quad \text{for } \kappa \in \mathcal{K}_4.$$

More generally, for $\kappa \in \mathcal{K}_k$ with $k \geq 4$, let $\mathcal{K}_{\ell,\kappa} = \{\tilde{\kappa} \in \mathcal{K}_\ell, \kappa - \tilde{\kappa} \in \mathcal{K}_{|\kappa|-\ell}\}$ for $\ell \leq |\kappa| - 2$. Then,

$$\gamma_{0\kappa} = \frac{\mu_\kappa}{\kappa!} - \sum_{\ell=2}^{k-2} \sum_{\tilde{\kappa} \in \mathcal{K}_{\ell,\kappa}} \frac{\mu_{\kappa-\tilde{\kappa}}}{(\kappa - \tilde{\kappa})!} \gamma_{0\tilde{\kappa}}.$$

**Example** (Bivariate $X$, $K = 4$)**.**
Suppose $X$ is bivariate (i.e., $d = 2$) and $K = 4$. For $\kappa \in \mathcal{K}_2 = \{(2,0), (1,1), (0,2)\}$ and $\kappa \in \mathcal{K}_3 = \{(3,0), (2,1), (1,2), (0,3)\}$, $\gamma_{0\kappa}$ is given by (21). For $\kappa \in \mathcal{K}_4$, $\gamma_{0\kappa}$ is given by

| $\kappa$ | $\gamma_{0\kappa}$ |
|---|---|
| (4,0) | $\left(\mathbb{E}[\varepsilon_{i1}^4] - 6\mathbb{E}[\varepsilon_{i1}^2]^2\right)/24$ |
| (3,1) | $\left(\mathbb{E}[\varepsilon_{i1}^3\varepsilon_{i2}] - 6\mathbb{E}[\varepsilon_{i1}^2]\mathbb{E}[\varepsilon_{i1}\varepsilon_{i2}]\right)/6$ |
| (2,2) | $\left(\mathbb{E}[\varepsilon_{i1}^2\varepsilon_{i2}^2] - 2\mathbb{E}[\varepsilon_{i1}^2]\mathbb{E}[\varepsilon_{i2}^2] - 4\mathbb{E}[\varepsilon_{i1}\varepsilon_{i2}]^2\right)/4$ |
| (1,3) | $\left(\mathbb{E}[\varepsilon_{i1}\varepsilon_{i2}^3] - 6\mathbb{E}[\varepsilon_{i2}^2]\mathbb{E}[\varepsilon_{i1}\varepsilon_{i2}]\right)/6$ |
| (0,4) | $\left(\mathbb{E}[\varepsilon_{i2}^4] - 6\mathbb{E}[\varepsilon_{i2}^2]^2\right)/24$ |

If in addition measurement errors $\varepsilon_{i1}$ and $\varepsilon_{i2}$ are independent, $\gamma_{0\kappa} = 0$ for $\kappa \in \{(1,1), (2,1), (1,2), (3,1), (1,3)\}$. In this case, the total number of the nuisance parameters to be estimated is 6. ■

# References


Amemiya, Y. (1985): "Instrumental variable estimator for the nonlinear errors-in-variables model," *Journal of Econometrics*, 28, 273–289.

——— (1990): "Two-stage instrumental variables estimators for the nonlinear errors-in-variables model," *Journal of Econometrics*, 44, 311–332.

Andrews, D. W. K. and P. Guggenberger (2019): "Identification-and singularity-robust inference for moment condition models," *Quantitative Economics*, 10, 1703–1746.

Andrews, I. (2016): "Conditional Linear Combination Tests for Weakly Identified Models," *Econometrica*, 84, 2155–2182.

Andrews, I., M. Gentzkow, and J. M. Shapiro (2017): "Measuring the Sensitivity of Parameter Estimates to Estimation Moments," *Quarterly Journal of Economics*, 132, 1553–1592.

Andrews, I. and A. Mikusheva (2016): "Conditional Inference With a Functional Nuisance Parameter," *Econometrica*, 84, 1571–1612.

Armstrong, T. B. and M. Kolesár (2021): "Sensitivity analysis using approximate moment condition models," *Quantitative Economics*, 12, 77–108.

Ashenfelter, O. and A. Krueger (1994): "Estimates of the Economic Return to Schooling from a New Sample of Twins," *The American Economic Review*, 84, 1157–1173.

Battistin, E. and A. Chesher (2014): "Treatment effect estimation with covariate measurement error," *Journal of Econometrics*, 178, 707–715.

Ben-Moshe, D., X. D'Haultfœuille, and A. Lewbel (2017): "Identification of additive and polynomial models of mismeasured regressors without instruments," *Journal of Econometrics*, 200, 207–222.

Bonhomme, S. and M. Weidner (2022): "Minimizing sensitivity to model misspecification," *Quantitative Economics*, 13, 907–954.

Bound, J., C. Brown, G. J. Duncan, and W. L. Rodgers (1994): "Evidence on the Validity of Cross-Sectional and Longitudinal Labor Market Data," *Journal of Labor Economics*, 12, 345–368.

Bound, J., C. Brown, and N. Mathiowetz (2001): "Measurement Error in Survey Data," in *Handbook of Econometrics*, Elsevier, 3705–3843.

Bound, J. and A. B. Krueger (1991): "The Extent of Measurement Error in Longitudinal Earnings Data: Do Two Wrongs Make a Right?" *Journal of Labor Economics*, 9, 1–24.

Carroll, R. J., X. Chen, and Y. Hu (2010): "Identification and estimation of nonlinear models using two samples with nonclassical measurement errors," *Journal of Nonparametric Statistics*, 22, 379–399.

Carroll, R. J., D. Ruppert, L. A. Stefanski, and C. M. Crainiceanu (2006): *Measurement Error in Nonlinear Models*, Chapman and Hall/CRC.

Carroll, R. J. and L. A. Stefanski (1990): "Approximate quasi-likelihood estimation in models with surrogate predictors," *Journal of the American Statistical Association*, 85, 652–663.

Chen, X., H. Hong, and D. Nekipelov (2011): "Nonlinear Models of Measurement Errors," *Journal of Economic Literature*, 49, 901–937.

Chen, X., H. Hong, and E. Tamer (2005): "Measurement Error Models with Auxiliary Data," *The Review of Economic Studies*, 72, 343–366.

Chesher, A. (1991): "The effect of measurement error," *Biometrika*, 78, 451–462.

——— (2000): "Measurement Error Bias Reduction," Working paper.

——— (2017): "Understanding the effect of measurement error on quantile regressions," *Journal of Econometrics*, 200, 223–237.

Chesher, A., M. Dumangane, and R. J. Smith (2002): "Duration response measurement error," *Journal of Econometrics*, 111, 169–194.

Chesher, A. and C. Schluter (2002): "Welfare measurement and measurement error," *The Review of Economic Studies*, 69, 357–378.

Erickson, T. and T. M. Whited (2002): "Two-Step GMM Estimation Of The Errors-In-Variables Model Using High-Order Moments," *Econometric Theory*, 18.

Evdokimov, K. S. and A. Zeleneev (2018): "Issues of Nonstandard Inference in Measurement Error Models," Working paper.

——— (2019): "Errors-In-Variables in Large Nonlinear Panel and Network Models," Working paper.

——— (2022): "Nonparametric Identification and Estimation with Non-Classical Errors-in-Variables," Working paper.

Guggenberger, P., J. J. Ramalho, and R. J. Smith (2012): "GEL statistics under weak identification," *Journal of Econometrics*, 170, 331–349.

Guggenberger, P. and R. J. Smith (2005): "Generalized empirical likelihood estimators and tests under partial, weak, and strong identification," *Econometric Theory*, 21, 667–709.

Hahn, J., J. Hausman, and J. Kim (2021): "A small sigma approach to certain problems in errors-in-variables models," *Economics Letters*, 208, 110094.

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

Hansen, L. P. (1982): "Large Sample Properties of Generalized Method of Moments Estimators," *Econometrica*, 50, 1029–1054.

Hausman, J. A., H. Ichimura, W. K. Newey, and J. L. Powell (1991): "Identification and estimation of polynomial errors-in-variables models," *Journal of Econometrics*, 50, 273–295.

Hausman, J. A., W. K. Newey, and J. L. Powell (1995): "Nonlinear errors in variables Estimation of some Engel curves," *Journal of Econometrics*, 65, 205–233.

Hong, H. and E. Tamer (2003): "A simple estimator for nonlinear error in variable models," *Journal of Econometrics*, 117, 1–19.

Hu, Y. and S. M. Schennach (2008): "Instrumental Variable Treatment of Nonclassical Measurement Error Models," *Econometrica*, 76, 195–216.

Kadane, J. B. (1971): "Comparison of k-Class Estimators When the Disturbances Are Small," *Econometrica*, 39, 723.

Kitamura, Y., T. Otsu, and K. Evdokimov (2013): "Robustness, infinitesimal neighborhoods, and moment restrictions," *Econometrica*, 81, 1185–1201.

Kleibergen, F. (2005): "Testing Parameters in GMM Without Assuming that They are Identified," *Econometrica*, 73, 1103–1123.

Koppelman, F. S. and C.-H. Wen (2000): "The paired combinatorial logit model: properties, estimation and application," *Transportation Research Part B: Methodological*, 34, 75–89.

Lewbel, A. (1997): "Constructing Instruments for Regressions With Measurement Error When no Additional Data are Available, with An Application to Patents and R&D," *Econometrica*, 65, 1201–1213.

Li, T. (2002): "Robust and consistent estimation of nonlinear errors-in-variables models," *Journal of Econometrics*, 110, 1–26.

Li, T. and Q. Vuong (1998): "Nonparametric Estimation of the Measurement Error Model Using Multiple Indicators," *Journal of Multivariate Analysis*, 65, 139–165.

McFadden, D. (1974): "The measurement of urban travel demand," *Journal of public economics*, 3, 303–328.

Newey, W. K. (2001): "Flexible Simulated Moment Estimation of Nonlinear Errors-in-Variables Models," *The Review of Economics and Statistics*, 83, 616–627.

Newey, W. K. and D. McFadden (1994): "Large sample estimation and hypothesis testing," *Handbook of Econometrics*, IV.

Pischke, J.-S. (1995): "Measurement Error and Earnings Dynamics: Some Estimates From the PSID Validation Study," *Journal of Business & Economic Statistics*, 13, 305–314.

Reiersøl, O. (1950): "Identifiability of a Linear Relation between Variables Which Are Subject to Error," *Econometrica*, 18, 375–389.

Rivers, D. and Q. H. Vuong (1988): "Limited Information Estimators And Exogeneity Tests For Simultaneous Probit Models," *Journal of Econometrics*, 39, 347–366.

Schennach, S. M. (2004): "Estimation of Nonlinear Models with Measurement Error," *Econometrica*, 72, 33–75.

——— (2007): "Instrumental Variable Estimation of Nonlinear Errors-in-Variables Models," *Econometrica*, 75, 201–239.

——— (2014): "Entropic Latent Variable Integration via Simulation," *Econometrica*, 82, 345–385.

——— (2016): "Recent Advances in the Measurement Error Literature," *Annual Review of Economics*, 8, 341–377.

——— (2020): "Mismeasured and unobserved variables," in *Handbook of Econometrics*, Elsevier, 487–565.

Schennach, S. M. and Y. Hu (2013): "Nonparametric Identification and Semiparametric Estimation of Classical Measurement Error Models Without Side Information," *Journal of the American Statistical Association*, 108, 177–186.

Smith, R. J. and R. W. Blundell (1986): "An Exogeneity Test for a Simultaneous Equation Tobit Model with an Application to Labor Supply," *Econometrica*, 54, 679.

Song, S. (2015): "Semiparametric estimation of models with conditional moment restrictions in the presence of nonclassical measurement errors," *Journal of Econometrics*, 185, 95–109.

Staiger, D. and J. H. Stock (1997): "Instrumental Variables Regression with Weak Instruments," *Econometrica*, 65, 557–586.

Stock, J. H. and J. H. Wright (2000): "GMM with Weak Identification," *Econometrica*, 68, 1055–1096.

Wang, L. and C. Hsiao (2011): "Method of moments estimation and identifiability of semiparametric nonlinear errors-in-variables models," *Journal of Econometrics*, 165, 30–44.

Wen, C.-H. and F. S. Koppelman (2001): "The generalized nested logit model," *Transportation Research Part B: Methodological*, 35, 627–641.

Wilhelm, D. (2019): "Testing for the presence of measurement error," CeMMAP working papers CWP48/19, Centre for Microdata Methods and Practice, IFS.

Wolter, K. M. and W. A. Fuller (1982): "Estimation of Nonlinear Errors-in-Variables Models," *The Annals of Statistics*, 10, 539–548.

# A Regularity Conditions

**Notation.** Let $\mathcal{X} \subseteq \mathbb{R}$ be some closed convex set containing the union of the supports of $X_i^*$ and $X_i$, and $\mathcal{S} = \operatorname{supp}(S_i)$.

**Assumption A.1.** (Moment function) *Suppose that the moment restrictions* (1) *are satisfied and the following conditions hold:*

*(i) For all $s \in \mathcal{S}$ and $\theta \in \Theta$, $g_x^{(K)}(x, s, \theta)$ exists and is continuous on $\mathcal{X}$. Moreover, there exist functions $b_1, b_2 : \mathcal{X} \times \mathcal{S} \times \Theta \to \mathbb{R}_+$ and integer $M \geqslant K+1$ such that for all $x, x' \in \mathcal{X}$, $s \in \mathcal{S}$, and $\theta \in \Theta$,*

$$\left\| g_x^{(K)}(x', s, \theta) - g_x^{(K)}(x, s, \theta) \right\| \leqslant b_1(x, s, \theta)|x' - x| + b_2(x, s, \theta)|x' - x|^{M-K}; \quad \text{(A.1)}$$

*(ii) Assumption MME holds with $L \geqslant M$;*

*(iii) $\mathbb{E}\left[g_x^{(k)}(X_i^*, S_i, \theta_0)\right]$, $k \in \{1, \ldots, K\}$, and $\mathbb{E}\left[b_j(X_i^*, S_i, \theta_0)\right]$, $j \in \{1, 2\}$, exist and are bounded.*

Assumption A.1 allows us to bound the remainder of the Taylor expansion of $g(X_i, S_i, \theta)$ around $X_i^*$ by a polynomial in $|X_i - X_i^*| = |\varepsilon_i|$. Combined with Assumption MME (which bounds the moments of $\varepsilon_i$), it ensures that this remainder is $o(n^{-1/2})$, which is crucial for establishing validity of the corrected moment function $\psi$ (Lemma 1).

Notice that if $\mathcal{X}$ is compact, condition (A.1) is satisfied if $g_x^{(K+1)}(x, s, \theta)$ is bounded on $\mathcal{X}$ (for all $s \in \mathcal{S}$ and $\theta \in \Theta$). If $\mathcal{X}$ is unbounded, condition (A.1) is satisfied if

for some $J$, such that $K < J \leqslant M$, $\sup_{x\in\mathcal{X}} \left\|g_x^{(J)}(x,s,\theta)\right\| \leqslant B(s,\theta)$ for some function $B(s,\theta)$. Also notice that condition (A.1) is stronger than the standard Lipschitz continuity because in applications $\left\|g_x^{(K)}(x,s,\theta)\right\|$ may behave like a polynomial in $x$ for large $x$.

**Assumption A.2.** (Parameter space)

*(i)* $\Theta \subset \mathbb{R}^{\dim(\theta)}$ *and* $\Gamma \subset \mathbb{R}^{K-1}$ *are compact,* $\theta_0 \in int(\Theta)$ *and* $\gamma_{0n} \in \Gamma$*;*

*(ii)* $0_{K-1} \in int(\Gamma)$*.*

**Assumption A.3.** (Regularity and smoothness conditions)

*(i) For all* $s \in \mathcal{S}$*,* $G_x^{(K)}(x,s,\theta)$ *exists and is continuous on* $\mathcal{X}\times\Theta$*; moreover, there exist functions* $b_{G1}, b_{G2} : \mathcal{X}\times\mathcal{S}\times\Theta \to \mathbb{R}_+$ *and* $\delta > 0$ *and for all* $x, x' \in \mathcal{X}$*,* $s \in \mathcal{S}$*, and* $\theta \in B_\delta(\theta_0)$

$$\left\|G_x^{(K)}(x',s,\theta) - G_x^{(K)}(x,s,\theta)\right\| \leqslant b_{G1}(x,s,\theta)|x'-x| + b_{G2}(x,s,\theta)|x'-x|^{M-K}$$

*(ii)* $\mathbb{E}\left[\left\|g_x^{(k)}(X_i^*,S_i,\theta_0)\right\|^2\right]$*,* $\mathbb{E}\left[\sup_{\theta\in\Theta}\left\|g_x^{(k)}(X_i^*,S_i,\theta)\right\|\right]$*, for* $k \in \{0,\dots,K\}$*, and* $\mathbb{E}\left[b_j(X_i^*,S_i,\theta_0)^2\right]$*,* $\mathbb{E}\left[\sup_{\theta\in\Theta} b_j(X_i^*,S_i,\theta)\right]$*, for* $j \in \{1,2\}$*, are bounded;*

*(iii) for some* $\delta > 0$*,* $\mathbb{E}\left[\sup_{\theta\in B_\delta(\theta_0)}\left\|G_x^{(k)}(X_i^*,S_i,\theta)\right\|\right]$*, for* $k \in \{0,\dots,K\}$*, and* $\mathbb{E}\left[\sup_{\theta\in B_\delta(\theta_0)} b_{Gj}(X_i^*,S_i,\theta)\right]$*, for* $j \in \{1,2\}$*, are bounded;*

*(iv)* $\hat{\Xi} \xrightarrow{p} \Xi$*, where* $\Xi$ *is a symmetric positive definite matrix;*

*(v) Assumption MME holds with* $L \geqslant 2M$*.*

Assumptions A.2 and A.3 include basic regularity conditions, which help to ensure $\sqrt{n}$-consistency and asymptotic normality of the suggested estimator $\hat{\theta}$. Specifically, Assumption A.3(i) is a counterpart of Assumption A.1(i) applied to the Jacobian function. It ensures that the effect of the measurement error on the Jacobian is localized and allows us to establish $G \to G^*$, so $\Psi \to \Psi^*$. As a result, local identification of $\theta_0$ and the asymptotic properties of $\hat{\theta}$ are controlled by $G^*$ (and $\Psi^*$), the Jacobian associated with the correctly measured variables (see Assumption ID(i)).

# B Proof of Lemma 1

To stress that in our asymptotic approximation the variance and the higher moments of $\varepsilon_i$ depend on $n$, we will use $\sigma_n^2 \equiv \mathbb{E}[\varepsilon_i^2]$, $\gamma_{0n} \equiv \gamma_0$.

Making use of Assumption A.1(i) , we expand $g(X_i, S_i, \theta_0)$ around $X_i^*$ as

$$\begin{aligned} g(X_i, S_i, \theta_0) =& g(X_i^*, S_i, \theta_0) + g_x^{(1)}(X_i^*, S_i, \theta_0)\varepsilon_i + \sum_{k=2}^{K} \frac{1}{k!} g_x^{(k)}(X_i^*, S_i, \theta_0)\varepsilon_i^k \\ &+ \frac{1}{K!}\left(g_x^{(K)}(\tilde{X}_i, S_i, \theta_0) - g_x^{(K)}(X_i^*, S_i, \theta_0)\right)\varepsilon_i^K, \end{aligned} \tag{B.1}$$

where $\tilde{X}_i$ lies between $X_i^*$ and $X_i$ (and hereafter $\tilde{X}_i$ is allowed to be component specific). Similarly, for $k' \in \{2, \ldots, K\}$, we have

$$\begin{aligned} g_x^{(k)}(X_i, S_i, \theta_0) =& g_x^{(k)}(X_i^*, S_i, \theta_0) + \sum_{\ell=k+1}^{K} \frac{1}{(\ell-k)!} g_x^{(\ell)}(X_i^*, S_i, \theta_0)\varepsilon_i^\ell \\ &+ \frac{1}{(K-k)!}\left(g_x^{(K)}(\tilde{X}_{ki}, S_i, \theta_0) - g_x^{(K)}(X_i^*, S_i, \theta_0)\right)\varepsilon_i^{K-k}, \end{aligned} \tag{B.2}$$

where $\tilde{X}_{ki}$ lies between $X_i^*$ and $X_i$. Hence, combining these expressions and rearranging the terms, we obtain

$$\begin{aligned} \psi(X_i, S_i, \theta_0, \gamma) =& g(X_i, S_i, \theta_0) - \sum_{k=2}^{K} \gamma_k g_x^{(k)}(X_i, S_i, \theta_0) \\ =& g(X_i^*, S_i, \theta_0) + g_x^{(1)}(X_i^*, S_i, \theta_0)\varepsilon_i \\ &+ \sum_{k=2}^{K} g_x^{(k)}(X_i^*, S_i, \theta_0)\left(\frac{1}{k!}\varepsilon_i^k - \sum_{\ell=2}^{k} \frac{1}{(k-\ell)!}\varepsilon_i^{k-\ell}\gamma_\ell\right) \\ &+ \frac{1}{K!}\left(g_x^{(K)}(\tilde{X}_i, S_i, \theta_0) - g_x^{(K)}(X_i^*, S_i, \theta_0)\right)\varepsilon_i^K \\ &- \sum_{k=2}^{K} \frac{\gamma_k}{(K-k)!}\left(g_x^{(K)}(\tilde{X}_{ki}, S_i, \theta_0) - g_x^{(K)}(X_i^*, S_i, \theta_0)\right)\varepsilon_i^{K-k}. \end{aligned} \tag{B.3}$$

We want to show that for a properly chosen $\gamma = \gamma_{0n}$, $\mathbb{E}[\psi(X_i, S_i, \theta_0, \gamma_{0n})] = o(n^{-1/2})$. Note that the first two terms in (B.3) are mean zero, i.e. we have

$$\mathbb{E}[g(X_i^*, S_i, \theta_0)] = 0, \quad \mathbb{E}\left[g_x^{(1)}(X_i^*, S_i, \theta_0)\varepsilon_i\right] = 0, \tag{B.4}$$

where the latter is guaranteed by Assumptions CME.

Second, we argue that for a properly chosen $\gamma = \gamma_{0n}$, we have

$$\mathbb{E}\left[\frac{1}{k!}\varepsilon_i^k - \sum_{\ell=2}^{k}\frac{1}{(k-\ell)!}\varepsilon_i^{k-\ell}\gamma_{0\ell n}\right] = 0, \tag{B.5}$$

for all $k \in \{2,\ldots,K\}$. Let us reparameterize $\gamma_{0n} = (\gamma_{02n},\ldots,\gamma_{0Kn})'$ using $\gamma_{0kn} = \sigma_n^k a_{kn}$. Then, (B.5) can be rewritten as

$$\mathbb{E}\left[\frac{1}{k!}(\varepsilon_i/\sigma_n)^k - \sum_{\ell=2}^{k}\frac{1}{(k-\ell)!}(\varepsilon_i/\sigma_n)^{k-\ell}a_{\ell n}\right] = 0,$$

which can also be represented as

$$B_n a_n = c_n \tag{B.6}$$

where $a_n = (a_{2n},\ldots,a_{Kn})'$, and

$$B_n = \begin{bmatrix} 1 & 0 & \ldots & 0 & 0 \\ \mathbb{E}[\varepsilon_i/\sigma_n] & 1 & \ldots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \frac{\mathbb{E}[(\varepsilon_i/\sigma_n)^{K-3}]}{(K-3)!} & \frac{\mathbb{E}[(\varepsilon_i/\sigma_n)^{K-4}]}{(K-4)!} & \ldots & 1 & 0 \\ \frac{\mathbb{E}[(\varepsilon_i/\sigma_n)^{K-2}]}{(K-2)!} & \frac{\mathbb{E}[(\varepsilon_i/\sigma_n)^{K-3}]}{(K-3)!} & \ldots & \mathbb{E}[(\varepsilon_i/\sigma_n)] & 1 \end{bmatrix}, \quad c_n = \begin{bmatrix} \mathbb{E}[(\varepsilon_i/\sigma_n)^2]/2! \\ \mathbb{E}[(\varepsilon_i/\sigma_n)^3]/3! \\ \vdots \\ \mathbb{E}[(\varepsilon_i/\sigma_n)^{K-1}]/(K-1)! \\ \mathbb{E}[(\varepsilon_i/\sigma_n)^K]/K! \end{bmatrix}.$$

Since $B_n$ is invertible, (B.6) has a unique solution $a_n = B_n^{-1}c_n$. Moreover, $a_n$ is bounded since both $B_n^{-1}$ and $c_n$ are bounded (Assumption MME). Hence, we conclude that (B.5) has a unique solution $\gamma_{0n} = \left(\sigma_n^2 a_{2n},\ldots,\sigma_n^K a_{Kn}\right)'$. Since (B.5) is satisfied, using Assumption CME, we also conclude that

$$\mathbb{E}\left[\sum_{k=2}^{K} g_x^{(k)}(X_i^*, S_i, \theta_0)\left(\frac{1}{k!}\varepsilon_i^k - \sum_{\ell=2}^{k}\frac{1}{(k-\ell)!}\varepsilon_i^{k-\ell}\gamma_{0\ell n}\right)\right] = 0. \tag{B.7}$$

To complete the proof of $\mathbb{E}\left[\psi(X_i, S_i, \theta_0, \gamma_{0n})\right] = o_n(n^{-1/2})$, it is sufficient to show that

$$\mathbb{E}\left[\left(g_x^{(K)}(\tilde{X}_i, S_i, \theta_0) - g_x^{(K)}(X_i^*, S_i, \theta_0)\right)\varepsilon_i^K\right] = o(n^{-1/2}), \tag{B.8}$$

$$\gamma_{0kn}\mathbb{E}\left[\left(g_x^{(K)}(\tilde{X}_{ki}, S_i, \theta_0) - g_x^{(K)}(X_i^*, S_i, \theta_0)\right)\varepsilon_i^{K-k}\right] = o(n^{-1/2}) \tag{B.9}$$

for $k \in \{2,\ldots,K\}$. We start with (B.8). Using Assumption A.1(i), we obtain

$$\left\|\left(g_x^{(K)}(\tilde{X}_i, S_i, \theta_0) - g_x^{(K)}(X_i^*, S_i, \theta_0)\right)\varepsilon_i^K\right\| \leqslant b_1(X_i^*, S_i, \theta_0)\left|\varepsilon_i\right|^{K+1} + b_2(X_i^*, S_i, \theta_0)\left|\varepsilon_i\right|^M. \tag{B.10}$$

Hence, using Assumption CME, and the fact $\left|\tilde{X}_i - X_i^*\right| \leqslant \varepsilon_i$, we get

$$\begin{aligned}
&\mathbb{E}\left[\left(g_x^{(K)}(\tilde{X}_i, S_i, \theta_0) - g_x^{(K)}(X_i^*, S_i, \theta_0)\right)\varepsilon_i^K\right] \\
&\leqslant \sigma_n^{K+1}\mathbb{E}\left[b_1(X_i^*, S_i, \theta_0)\right]\mathbb{E}\left[\left|\varepsilon_i/\sigma_n\right|^{K+1}\right] + \sigma_n^M\mathbb{E}\left[b_2(X_i^*, S_i, \theta_0)\right]\mathbb{E}\left[\left|\varepsilon_i/\sigma_n\right|^M\right].
\end{aligned}$$

Since (i) the expectations above are bounded (Assumptions MME, A.1(ii), and A.1(iii)) and (ii) $\sigma_n^{K+1} = o(n^{-1/2})$ and $\sigma_n^M = o(n^{-1/2})$ (Assumption MME), this implies that (B.8) holds. To inspect (B.9), recall that $\gamma_{0kn} = \sigma_n^k a_{kn}$. As a result, using Assumptions A.1(i) and CME, and $\left|\tilde{X}_{ki} - X_i^*\right| \leqslant \varepsilon_i$ again, we also have

$$\begin{aligned}
&\gamma_{0kn}\mathbb{E}\left[\left(g_x^{(K)}(\tilde{X}_{ki}, S_i, \theta_0) - g_x^{(K)}(X_i^*, S_i, \theta_0)\right)\varepsilon_i^{K-k}\right] \\
&\leqslant a_{kn}\left(\sigma_n^{K+1}\mathbb{E}\left[b_1(X_i^*, S_i, \theta_0)\right]\mathbb{E}\left[\left|\varepsilon_i/\sigma_n\right|^{K+1-k}\right] + \sigma_n^M\mathbb{E}\left[b_2(X_i^*, S_i, \theta_0)\right]\mathbb{E}\left[\left|\varepsilon_i/\sigma_n\right|^{M-k}\right]\right).
\end{aligned}$$

Since $a_{kn}$ is bounded, we conclude that (B.9) holds analogously to (B.8).

Combining (B.3) with (B.4), and (B.7)-(B.9), we conclude that $\mathbb{E}\left[\psi(X_i, S_i, \theta_0, \gamma_{0n})\right] = o(n^{-1/2})$.

Finally, we want to verify the recursive expressions for the components of $\gamma_{0n}$ using (B.5). First, $\gamma_{02n} = \mathbb{E}\left[\varepsilon_i^2\right]/2$ and $\gamma_{03n} = \mathbb{E}\left[\varepsilon_i^3\right]/6$ (since $\mathbb{E}\left[\varepsilon_i\right] = 0$). For $k \geqslant 4$, suppose that $\gamma_{0\ell n}$ are known for $\ell \in \{2, \ldots, k-1\}$. Then $\gamma_{0kn}$ can be directly computed from (B.5):

$$\sum_{\ell=2}^{k} \frac{\mathbb{E}\left[\varepsilon_i^{k-\ell}\right]}{(k-\ell)!}\gamma_{0\ell n} = \frac{\mathbb{E}\left[\varepsilon_i^k\right]}{k!}.$$

Plugging $\mathbb{E}\left[\varepsilon_i\right] = 0$ and rearranging the terms give the expession in (12). Q.E.D.

# C Proof of Theorem 2

**Notation.** To stress that in our asymptotic approximation the variance and the higher moments of $\varepsilon_i$ depend on $n$, we will use $\sigma_n^2 \equiv \mathbb{E}\left[\varepsilon_i^2\right]$, $\gamma_{0n} \equiv \gamma_0$, and $\beta_{0n} \equiv \beta_0 \equiv (\theta_0', \gamma_{0n}')'$.

All vectors are columns. For some generic parameter vector $\alpha$ and a vector (or matrix) valued function $a(x, s, \alpha)$ and , let $a_i(\beta) \equiv a(X_i, S_i, \alpha)$, $\overline{a}(\alpha) \equiv n^{-1}\sum_{i=1}^n a_i(\alpha)$, $a(\alpha) \equiv \mathbb{E}[a_i(\alpha)]$. Similarly, we let $a_i^*(\alpha) \equiv a(X_i^*, S_i, \alpha)$, $\overline{a}^*(\alpha) \equiv n^{-1}\sum_{i=1}^n a_i^*(\alpha)$, $a^*(\alpha) \equiv \mathbb{E}[a_i^*(\alpha)]$.

For the true value of the parameter $\alpha_0$, we often write $a_i \equiv a(\alpha_0)$, $\overline{a} \equiv \overline{a}(\alpha_0)$,

$a \equiv a(\alpha_0)$, $a_i^* \equiv a(\alpha_0)$, $\overline{a}^* \equiv \overline{a}^*(\alpha_0)$, $a^* \equiv a^*(\alpha_0)$.

## C.1 Auxiliary lemmas

**Lemma C.1.** *Suppose that* $\{(X_i^*, S_i', \varepsilon_i)\}_{i=1}^n$ *are i.i.d.. Then, under Assumptions MME, CME, A.1, A.2*(i)*, and A.3*(i)-(iii)*, we have*

*(i)*

$$\sup_{\theta \in \Theta} \left\| \overline{g}_x^{(k)}(\theta) - g_x^{(k)*}(\theta) \right\| = o_p(1)$$

*and* $g_x^{(k)*}(\theta)$ *is continuous on* $\Theta$ *for* $k \in \{0, \ldots, K\}$*;*

*(ii) for some* $\delta > 0$*,*

$$\sup_{\theta \in B_\delta(\theta_0)} \left\| \overline{G}_x^{(k)}(\theta) - G_x^{(k)*}(\theta) \right\| = o_p(1),$$

*and* $G_x^{(k)*}(\theta)$ *is continuous on* $B_\delta(\theta_0)$ *for* $k \in \{0, \ldots, K\}$*.*

*Proof of Lemma C.1.* First, we show

$$\sup_{\theta \in \Theta} \|\overline{g}(\theta) - g^*(\theta)\| = o_p(1).$$

By the triangle inequality,

$$\sup_{\theta \in \Theta} \|\overline{g}(\theta) - g^*(\theta)\| \leqslant \sup_{\theta \in \Theta} \|\overline{g}(\theta) - \overline{g}^*(\theta)\| + \sup_{\theta \in \Theta} \|\overline{g}^*(\theta) - g^*(\theta)\|.$$

Then, it is sufficient to show that both terms on the right hand side of the inequality above are $o_p(1)$. Expanding $g(X_i, S_i, \theta_0)$ around $X_i^*$ as in (B.1) and invoking

Assumption A.1(i),

$$\begin{aligned}\sup_{\theta\in\Theta}\|\overline{g}(\theta)-\overline{g}^*(\theta)\| &= \sup_{\theta\in\Theta}\left\|\sum_{k=1}^{K-1}\frac{1}{k!}\frac{1}{n}\sum_{i=1}^{n}g_x^{(k)}(X_i^*,S_i,\theta)\varepsilon_i^k+\frac{1}{K!}\frac{1}{n}\sum_{i=1}^{n}g_x^{(K)}(\tilde{X}_i^*,S_i,\theta)\varepsilon_i^K\right\| \\ &\leqslant \sum_{k=1}^{K}\frac{1}{k!}\underbrace{\frac{1}{n}\sum_{i=1}^{n}\sup_{\theta\in\Theta}\left\|g_x^{(k)}(X_i^*,S_i,\theta)\right\|\,|\varepsilon_i|^k}_{o_p(1)} \\ &\quad+\frac{1}{K!}\underbrace{\frac{1}{n}\sum_{i=1}^{n}\sup_{\theta\in\Theta}b_1(X_i^*,S_i,\theta)|\varepsilon_i|^{K+1}}_{o_p(1)} \\ &\quad+\frac{1}{K!}\underbrace{\frac{1}{n}\sum_{i=1}^{n}\sup_{\theta\in\Theta}b_2(X_i^*,S_i,\theta)|\varepsilon_i|^{M}}_{o_p(1)},\end{aligned}$$

where $\tilde{X}_i$ lies in between of $X_i^*$ and $X_i$. Now observe that all the terms following the inequality sign are $o_p(1)$. Indeed, this is guaranteed by Markov's inequality paired with Assumptions MME, CME, and A.3(ii). Hence, $\sup_{\theta\in\Theta}\|\overline{g}(\theta)-\overline{g}^*(\theta)\| = o_p(1)$, and we are left to show $\sup_{\theta\in\Theta}\|\overline{g}^*(\theta)-g^*(\theta)\| = o_p(1)$. This, in turn, follows from the standard ULLN (e.g., Lemma 2.4 in Newey and McFadden, 1994), which also ensures continuity of $g^*(\theta)$ on $\Theta$. Hence, we conclude that the assertion of the lemma holds for $g$.

Applying nearly identical arguments, one can also establish the desired results for $g_x^{(k)}$ for $k\in\{1,\dots,K\}$ and for $G_x^{(k)}$ for $k\in\{0,\dots,K\}$ (for the latter, Assumptions A.3(i) and (iii) take the places of Assumptions A.1(i) and A.3(ii), respectively). Q.E.D.

**Lemma C.2.** *Suppose that the hypotheses of Lemma C.1 are satisfied. Then, $g_x^{(k)}\to g_x^{(k)*}$ and $G_x^{(k)}\to G_x^{(k)*}$ for $k\in\{0,\dots,K\}$. Suppose also $\hat{\theta}\xrightarrow{p}\theta_0$. Then, $\overline{g}_x^{(k)}(\hat{\theta})\xrightarrow{p} g_x^{(k)*}$ and $\overline{G}_x^{(k)}(\hat{\theta})\xrightarrow{p} G_x^{(k)*}$ for $k\in\{0,\dots,K\}$.*

*Proof of Lemma C.2.* First, we prove the assertions of the lemma for $g_x^{(k)}$. Note that, by the standard expansion of $g_x^{(k)}(X_i,S_i,\theta_0)$ around $X_i^*$ (see Eq. (B.2) above), we

have

$$\begin{aligned}\left\|g_x^{(k)} - g_x^{(k)*}\right\| \leqslant & \mathbb{E}\left[\left\|g(X_i, S_i, \theta_0) - g_x^{(k)}(X_i^*, S_i, \theta)\right\|\right] \\ \leqslant & \sum_{\ell=k+1}^{K} \frac{1}{(\ell-k)!}\mathbb{E}\left[\left\|g_x^{(\ell)}(X_i^*, S_i, \theta_0)\right\| |\varepsilon_i|^{\ell}\right] \\ & + \frac{1}{(K-k)!}\mathbb{E}\left[\left\|\left(g_x^{(K)}(\tilde{X}_i, S_i, \theta_0) - g_x^{(K)}(X_i^*, S_i, \theta_0)\right)\right\| |\varepsilon_i|^{K-k}\right].\end{aligned}$$

By Assumptions MME, CME, and A.3(ii), $\mathbb{E}\left[\left\|g_x^{(\ell)}(X_i^*, S_i, \theta_0)\right\| |\varepsilon_i|^{\ell}\right] \to 0$ for all $\ell \in \{1, \ldots, K\}$. Next, using Assumptions A.1(i) and CME,

$$\begin{aligned}& \mathbb{E}\left[\left\|\left(g_x^{(K)}(\tilde{X}_i, S_i, \theta_0) - g_x^{(K)}(X_i^*, S_i, \theta_0)\right)\right\| |\varepsilon_i|^{K-k}\right] \\ & \leqslant \mathbb{E}\left[b_1(X_i^*, S_i, \theta_0)\right]\mathbb{E}\left[|\varepsilon_i|^{K+1-k}\right] + \mathbb{E}\left[b_2(X_i^*, S_i, \theta_0)\right]\mathbb{E}\left[|\varepsilon_i|^{M-k}\right] \to 0,\end{aligned}$$

where the convergence follows from Assumptions MME, A.1(ii) and A.3(ii). Hence, we conclude $g_x^{(k)} \to g_x^{(k)*}$.

Next, we show $\overline{g}_x^{(k)}(\hat{\theta}) \xrightarrow{p} g_x^{(k)*}$. By the triangle inequality,

$$\left\|\overline{g}_x^{(k)}(\hat{\theta}) - g_x^{(k)*}\right\| \leqslant \sup_{\theta \in B_\delta(\theta_0)} \left\|\overline{g}_x^{(k)}(\theta) - g_x^{(k)*}(\theta)\right\| + \left\|g_x^{(k)*}(\hat{\theta}) - g_x^{(k)*}(\theta_0)\right\|,$$

where the inequality holds with probability approaching one since $\hat{\theta} \in B_\delta(\theta_0)$ with probability approaching one. Note that, By Lemma C.1, $\sup_{\theta \in B_\delta(\theta_0)} \left\|\overline{g}_x^{(k)}(\theta) - g_x^{(k)*}(\theta)\right\| = o_p(1)$ and $\left\|g_x^{(k)*}(\hat{\theta}) - g_x^{(k)*}(\theta_0)\right\| = o_p(1)$, where the second result follows from consistency of $\hat{\theta}$ and continuity of $g_x^{(k)*}(\theta)$. Hence, $\overline{g}_x^{(k)}(\hat{\theta}) \xrightarrow{p} g_x^{(k)*}$, which completes the proof of the results for $g_x^{(k)}$ for all $k \in \{0, \ldots, K\}$.

A nearly identical argument, can be invoked to establish the same results for $G_x^{(k)}$ for $k \in \{0, \ldots, K\}$, with Assumptions A.3(i) and (iii) taking the places of Assumptions A.1(i) and A.3(ii), respectively. Q.E.D.

**Lemma C.3.** *Suppose that the hypotheses of Lemma C.1 are satisfied. Then, under additional Assumptions A.3(iv) and ID(ii), we have $\hat{\theta} \xrightarrow{p} \theta_0$, $\hat{\gamma} \xrightarrow{p} 0$ and $\hat{\gamma} \xrightarrow{p} \gamma_{0n}$.*

*Proof of Lemma C.3.* First, we argue that $\sup_{\beta \in \mathcal{B}} \left\|\overline{\psi}(\beta) - \psi^*(\beta)\right\| = o_p(1)$. Notice that, by the triangle inequality,

$$\sup_{\beta \in \mathcal{B}} \left\|\overline{\psi}(\beta) - \psi^*(\beta)\right\| \leqslant \sup_{\theta \in \Theta} \left\|\overline{g}(\theta) - g^*(\theta)\right\| + \sum_{k=2}^{K} |\gamma_k| \sup_{\theta \in \Theta} \left\|\overline{g}_x^{(k)}(\theta) - g_x^{(k)*}(\theta)\right\| = o_p(1), \tag{C.1}$$

where the equality follows from Lemma C.1(i) and boundedness of $\gamma$ (Assumption A.2(i)). Moreover, Lemma C.1(i) also ensures that $\psi^*(\beta)$ is continuous on compact $\mathcal{B}$ and, consequently, is bounded.

Let $\hat{Q}(\beta) = \overline{\psi}(\beta)'\hat{\Xi}\overline{\psi}(\beta)$ and $Q^*(\beta) = \psi^*(\beta)'\Xi\psi^*(\beta)$. Notice that (C.1), boundedness of $\psi^*(\beta)$, and Assumption A.3(iv) together guarantee that $\sup_{\beta\in\mathcal{B}}\left|\hat{Q}(\beta) - Q^*(\beta)\right| = o_p(1)$. Next, recall that $\gamma_{0n} \to 0_{K-1}$ (Lemma 1). Since $\Gamma$ is compact and $\gamma_{0n} \in \Gamma$ (Assumption A.2(i)), $0_{K-1} \in \Gamma$. Consequently, Assumptions ID(ii) and A.3(iv) together guarantee that $Q^*(\beta)$ is uniquely minimized at $\theta = \theta_0$ and $\gamma = 0_{K-1}$. Consequently, applying the standard consistency argument (e.g., Theorem 2.1 of Newey and McFadden, 1994), we conclude that $\hat{\theta} \to \theta_0$ and $\hat{\gamma} \to 0_{K-1}$. Finally, since $\gamma_{0n} \to 0$ (Lemma 1), we also have $\hat{\gamma} \xrightarrow{p} \gamma_{0n}$. Q.E.D.

**Lemma C.4.** *Suppose that $\{(X_i^*, S_i', \varepsilon_i)\}_{i=1}^n$ are i.i.d.. Then, under Assumptions MME, CME, A.1, and A.3(ii) and (v), we have*

$$n^{1/2}\overline{\psi}(\beta_{0n}) \xrightarrow{d} N(0, \Omega_{gg}^*),$$

*where $\Omega_{gg}^* \equiv \mathbb{E}\left[g(X_i, S_i, \theta_0)g(X_i, S_i, \theta_0)'\right]$.*

*Proof of Lemma C.4.* Using expansion (B.3), we obtain

$$\begin{aligned} n^{1/2}\overline{\psi}(\beta_{0n}) =& n^{-1/2}\sum_{i=1}^n g(X_i^*, S_i, \theta_0) + n^{-1/2}\sum_{i=1}^n g_x^{(1)}(X_i^*, S_i, \theta_0)\varepsilon_i \\ &+ \sum_{k=2}^K n^{-1/2}\sum_{i=1}^n g_x^{(k)}(X_i^*, S_i, \theta_0)\left(\frac{1}{k!}\varepsilon_i^k - \sum_{\ell=2}^k \frac{1}{(k-\ell)!}\varepsilon_i^{k-\ell}\gamma_{0kn}\right) \\ &+ \frac{1}{K!}n^{-1/2}\sum_{i=1}^n \left(g_x^{(K)}(\tilde{X}_i, S_i, \theta_0) - g_x^{(K)}(X_i^*, S_i, \theta_0)\right)\varepsilon_i^K \\ &- \sum_{k=2}^K \frac{\gamma_{0kn}}{(K-k)!}n^{-1/2}\sum_{i=1}^n \left(g_x^{(K)}(\tilde{X}_{ki}, S_i, \theta_0) - g_x^{(K)}(X_i^*, S_i, \theta_0)\right)\varepsilon_i^{K-k}. \end{aligned} \tag{C.2}$$

First, note that, by the standard CLT, $n^{-1/2}\sum_{i=1}^n g(X_i^*, S_i, \theta_0) \xrightarrow{d} N(0, \Omega_{gg}^*)$. The rest of the proof is to show that the remaining terms are $o_p(1)$. By Assumptions MME, CME, A.3(ii), Chebyshev's inequality guarantees

$$n^{-1/2}\sum_{i=1}^n g_x^{(1)}(X_i^*, S_i, \theta_{0n})\varepsilon_i = o_p(1)$$

Next, (B.7) ensures that we can similarly apply Chebyshev's inequality (combined with Assumptions MME, CME, A.3(ii) and (v)) to ensure that for $k \in \{2, \ldots, K\}$

$$n^{-1/2} \sum_{i=1}^{n} g_x^{(k)}(X_i^*, S_i, \theta_0) \left( \frac{1}{k!} \varepsilon_i^k - \sum_{\ell=2}^{k} \frac{1}{(k-\ell)!} \varepsilon_i^{k-\ell} \gamma_{0kn} \right) = o_p(1).$$

Next, using (B.10),

$$\begin{aligned}
&\left\| n^{-1/2} \sum_{i=1}^{n} \left( g_x^{(K)}(\tilde{X}_i, S_i, \theta_0) - g_x^{(K)}(X_i^*, S_i, \theta_0) \right) \varepsilon_i^K \right\| \\
\leqslant & n^{-1/2} \sum_{i=1}^{n} b_1(X_i^*, S_i, \theta_0) \left| \varepsilon_i \right|^{K+1} + n^{-1/2} \sum_{i=1}^{n} b_2(X_i^*, S_i, \theta_0) \left| \varepsilon_i \right|^{M} \\
\leqslant & \underbrace{n^{1/2} \sigma_n^{K+1}}_{\to 0} \underbrace{\left( n^{-1} \sum_{i=1}^{n} b_1(X_i^*, S_i, \theta_0) \left| \varepsilon_i / \sigma_n \right|^{K+1} \right)}_{O_p(1)} \\
& + \underbrace{n^{1/2} \sigma_n^{M}}_{\to 0} \underbrace{\left( n^{-1} \sum_{i=1}^{n} b_2(X_i^*, S_i, \theta_0) \left| \varepsilon_i / \sigma_n \right|^{M} \right)}_{O_p(1)} = o_p(1),
\end{aligned}$$

where both $n^{1/2}\sigma_n^{K+1}$ and $n^{1/2}\sigma_n^{M}$ converge to zero by Assumption MME, and the terms in the brackets are $O_p(1)$ by Markov's inequality (ensured by Assumptions MME, CME, A.1(ii) and A.3(ii)). Recall that in the proof of Lemma 1, we have demonstrated that $\gamma_{0kn} = \sigma_n^k a_{kn}$, where $a_{kn}$ are bounded, for $k \in \{2, \ldots, K\}$. Hence, similarly, we have

$$\begin{aligned}
&\left\| \gamma_{0kn} n^{-1/2} \sum_{i=1}^{n} \left( g_x^{(K)}(\tilde{X}_i, S_i, \theta_0) - g_x^{(K)}(X_i^*, S_i, \theta_0) \right) \varepsilon_i^{K-k} \right\| \\
\leqslant & a_{kn} \sigma_n^k \left[ n^{-1/2} \sum_{i=1}^{n} b_1(X_i^*, S_i, \theta_0) \left| \varepsilon_i \right|^{K-k+1} + n^{-1/2} \sum_{i=1}^{n} b_2(X_i^*, S_i, \theta_0) \left| \varepsilon_i \right|^{M-k} \right] \\
\leqslant & a_{kn} \underbrace{n^{1/2} \sigma_n^{K+1}}_{\to 0} \underbrace{\left( n^{-1} \sum_{i=1}^{n} b_1(X_i^*, S_i, \theta_0) \left| \varepsilon_i / \sigma_n \right|^{K-k+1} \right)}_{O_p(1)} \\
& + a_{kn} \underbrace{n^{1/2} \sigma_n^{M}}_{\to 0} \underbrace{\left( n^{-1} \sum_{i=1}^{n} b_2(X_i^*, S_i, \theta_0) \left| \varepsilon_i / \sigma_n \right|^{M-k} \right)}_{O_p(1)} = o_p(1).
\end{aligned}$$

Hence, we have demonstrated that all the remaining terms in (C.2) are $o_p(1)$, i.e. we have

$$\begin{aligned} n^{1/2}\overline{\psi}(\beta_{0n}) =& n^{-1/2}\sum_{i=1}^{n} g(X_i^*, S_i, \theta_0) + o_p(1) \\ &\xrightarrow{d} N(0, \Omega_{gg}^*), \end{aligned}$$

which completes the proof. Q.E.D.

## C.2 Proof of Theorem 2

Equipped with Lemmas C.1-C.4, we are ready to prove Theorem 2.

Since (i) $\hat{\theta}$ and $\hat{\gamma}$ are consistent for $\theta_0$ and $\gamma_{0n}$, respectively (Lemma C.3) and (ii) both $\theta_0$ and $\gamma_{0n} \to 0$ (Assumption MME) are bounded away from the boundaries of $\Theta$ and $\Gamma$ respectively (Assumption A.2), the standard GMM FOC is satisfied with probability approaching one, i.e., we have (with probability approaching one)

$$\overline{\Psi}(\hat{\beta})'\hat{\Xi}\overline{\psi}(\hat{\beta}) = 0.$$

Expanding $\overline{\psi}(\hat{\beta})$ around $\overline{\psi}(\beta_{0n})$ gives

$$\overline{\Psi}(\hat{\beta})'\hat{\Xi}\left(\overline{\psi}(\beta_{0n}) + \overline{\Psi}(\tilde{\beta})(\hat{\beta} - \beta_0)\right) = 0, \tag{C.3}$$

where $\tilde{\beta}$ lies between $\beta_{0n}$ and $\hat{\beta}$ (and, consequently, $\tilde{\theta} \xrightarrow{p} \theta_0$ and $\tilde{\gamma} \xrightarrow{p} 0$). Next, we argue that $\overline{\Psi}(\hat{\beta}) = \Psi^* + o_p(1)$. Observe

$$\overline{\Psi}(\hat{\beta}) = \left[\overline{G}(\hat{\theta}) - \sum_{k=2}^{K} \hat{\gamma}_k \overline{G}_x^{(k)}(\hat{\theta}), -\overline{g}_x^{(2)}(\hat{\theta}), \ldots, -\overline{g}_x^{(K)}(\hat{\theta})\right].$$

Since $\hat{\theta} \xrightarrow{p} \theta_0$ (Lemma C.3), we can invoke the result of Lemma C.2 to argue that $\overline{g}_x^{(k)}(\hat{\theta}) \xrightarrow{p} g_x^{(k)*}$ and $\overline{G}_x^{(k)}(\hat{\theta}) \xrightarrow{p} G_x^{(k)*}$ for all $k \in \{0, \ldots, K\}$. This, combined with $\hat{\gamma} \to 0$ (Lemma C.3), ensures that $\overline{\Psi}(\hat{\beta}) = \Psi^* + o_p(1)$ and, analogously, $\overline{\Psi}(\tilde{\beta}) = \Psi^* + o_p(1)$. Coupling these result with Assumption A.3(iv), we conclude that $\overline{\Psi}(\hat{\beta})'\hat{\Xi}\overline{\Psi}(\tilde{\beta}) \xrightarrow{p} \Psi^{*\prime}\Xi\Psi^*$, which is invertible by Assumption ID(i). Hence, (C.3) can be rearranged as (with probability approaching one)

$$\begin{aligned} n^{1/2}(\hat{\beta} - \beta_{0n}) &= -\left(\overline{\Psi}(\hat{\beta})'\hat{\Xi}\overline{\Psi}(\tilde{\beta})\right)^{-1}\overline{\Psi}(\hat{\beta})'\hat{\Xi}n^{1/2}\overline{\psi}(\beta_{0n}) \\ &= -\left(\Psi^{*\prime}\Xi\Psi^*\right)^{-1}\Psi^{*\prime}\Xi n^{1/2}\overline{\psi}(\beta_{0n}) + o_p(1), \end{aligned}$$

where, by Lemma C.4, $n^{1/2}\overline{\psi}(\beta_{0n}) \xrightarrow{d} N(0, \Omega^*_{gg})$. Hence, we conclude

$$n^{1/2}(\hat{\beta} - \beta_{0n}) \xrightarrow{d} N(0, \Sigma^*),$$

where

$$\Sigma^* = \left(\Psi^{*\prime}\Xi\Psi^*\right)^{-1} \Psi^{*\prime}\Xi\Omega^*_{gg}\Psi^*\Xi \left(\Psi^{*\prime}\Xi\Psi^*\right)^{-1}.$$

To complete the proof, we need to show that $\Sigma \to \Sigma^*$. First, note that, by Lemma C.2 and $\gamma_{0n} \to 0$ (Assumption MME)

$$\Psi = \left[G - \sum_{k=2}^{K} \gamma_{0kn} G_x^{(k)}, -g_x^{(2)}, \ldots, -g_x^{(K)}\right] \to \left[G^*, -g_x^{(2)*}, \ldots, -g_x^{(K)*}\right] = \Psi^*.$$

Next, we want to argue that $\Omega_{\psi\psi} \to \Omega^*_{gg}$. Observe that

$$\Omega_{\psi\psi} = \mathbb{E}\left[\left(g_i - \sum_{k=2}^{K} \gamma_{0kn} g_{xi}^{(k)}\right)\left(g_i - \sum_{k=2}^{K} \gamma_{0kn} g_{xi}^{(k)}\right)'\right] = \mathbb{E}\left[g_i g_i'\right] + o(1),$$

where the equality follows since (i) $\gamma_{0kn} \to 0$ for all $k \in \{2, \ldots, K\}$ (Assumption MME) and (ii) $\mathbb{E}\left[g_{xi}^{(k)}\left(g_{xi}^{(k')}\right)'\right]$ is bounded for all $k, k' \in \{0, \ldots, K\}$. In particular, (ii) can be inspected by expanding $g_x^{(k)}(X_i, S_i, \theta_0)$ and $g_x^{(k')}(X_i, S_i, \theta_0)$ around $X_i^*$ as in (B.2) and bounding the expectations as in the proof of Lemma C.2 (using Assumptions MME, CME, A.1(i), A.3(ii), and A.3(v)). Similarly, by expanding $g(X_i, S_i, \theta_0)$ around $X_i^*$ and bounding the residual terms as in the proof of Lemma C.2 (again, using Assumptions MME, CME, A.1(i), A.3(ii), and A.3(v)), we verify that $\mathbb{E}\left[g_i g_i'\right] \to \mathbb{E}\left[g_i^* g_i^{*\prime}\right] = \Omega^*_{gg}$. Hence, $\Omega_{\psi\psi} \to \Omega^*_{gg}$ and, consequently, we verified that $\Sigma \to \Sigma^*$. Finally, we conclude

$$n^{1/2}\Sigma^{-1/2}(\hat{\beta} - \beta_{0n}) \to N(0, I_{\dim(\theta)+K-1}),$$

which completes the proof. Q.E.D.

# D Proof of Lemma 3

As in the previous proofs, we use notation $\sigma_n^2 \equiv \mathbb{E}[\varepsilon_i^2] \equiv \sigma_\varepsilon^2$, $\hat{\sigma}_n^2 \equiv \hat{\sigma}_\varepsilon^2$, and $\hat{\sigma}_n^K \equiv \hat{\sigma}_\varepsilon^K$.

First, we show that (19) holds. By Theorem 2, we have $\hat{\sigma}_n^2 = \sigma_n^2 + O_p(n^{-1/2})$.

Since $K$ is even, (19) can be obtained by expanding

$$\begin{aligned}\hat{\sigma}_n^K &= (\sigma_n^2 + O_p(n^{-1/2}))^{K/2} = \sum_{\ell=0}^{K/2} \binom{n}{\ell} (\sigma_n^2)^{K/2-\ell} (O_p(n^{-1/2}))^\ell \\ &= \sigma_n^K + \sum_{\ell=1}^{K/2} \binom{n}{\ell} (\sigma_n^2)^{K/2-\ell} O_p(n^{-\ell/2}) \\ &= \sigma_n^K + O_p(\max\{\sigma_n^{K-2} n^{-1/2}, n^{-K/4}\}).\end{aligned}$$

Next, we prove the first statement of the lemma. First, using Lemma C.2, we have $\overline{g}_x^{(K)}(\hat{\theta}) = g_x^{(K)*} + o_p(1)$. Here, $\|g_x^{(K)*}\|$ is bounded away from zero and above, with the former implied by Assumption ID(i). Likewise, since the Jacobian corresponding to $\hat{\beta}_L$ (involved in the construction of $B$) is a submatrix of the Jacobian corresponding to $\hat{\beta}_K$, we conclude that $\|Bg_x^{(K)*}\|$ is also bounded from zero and above. Hence, using consistency of $\hat{B}$ and $\hat{\Sigma}$, we conclude that, with probability approaching one, for some $\underline{C}, \overline{C} > 0$, we have

$$\underline{C}\sqrt{n}\hat{\sigma}_n^K \leq \max_{1\leq\ell\leq\dim(\beta)} |\hat{\Sigma}_{\ell\ell}^{-1/2}\sqrt{n}\hat{\sigma}^K \hat{B}_{\ell\cdot}\overline{g}_x^{(K)}(\hat{\theta})| \leq \overline{C}\sqrt{n}\hat{\sigma}_n^K. \tag{D.1}$$

To prove the first part of the lemma it is sufficient to show that $\delta_n \sigma_n^{L+1} = o_p(n^{-1/2})$. Let $n_j$ index a subsequence of $\delta_{n_j}$ such that $\delta_{n_j} = 1$. If this subsequence is finite, the statement is trivial. Otherwise, we need to show that $\sigma_{n_j}^{L+1} = o(n_j^{-1/2})$ along this subsequence. We prove this by contradiction. Suppose $\sigma_{n_j}^{L+1} = o(n_j^{-1/2})$ does not hold meaning that there exists another subsequence within $n_j$, denoted by $n_m$ such that for all sufficiently large $m$ we have $\sqrt{n_m}\sigma_{n_m}^{L+1} \geq C > 0$ and $\delta_{n_m} = 1$. Since we have $\sigma_{n_m} \geq C n_m^{-\frac{1}{2(L+1)}}$ and $\sqrt{n_m}\sigma_{n_m}^K \geq C n_m^{-(K-L-1)/2(L+1)}$, using (19), we conclude that, for some $C > 0$, with probability approaching one, we have

$$\sqrt{n_m}\hat{\sigma}_{n_m}^K = \sqrt{n_m}\sigma_{n_m}^K + O_p\left(\sigma_{n_m}^{K-2} + n_{n_m}^{-(K-2)/4}\right) \geq C n_m^{-(K-L-1)/2(L+1)} \tag{D.2}$$

because the remainder $O_p(\sigma_{n_m}^{K-2} + n_{n_m}^{-(K-2)/4})$ is dominated by $\sqrt{n_m}\sigma_{n_m}^K$ in this case. At the same time, $\delta_{n_m} = 1$ together with (D.1) implies that $\sqrt{n_m}\hat{\sigma}_{n_m}^K \leq C\varkappa_{n_m}$ for some $C > 0$. This contradicts (D.2) since $\varkappa_n n^{(K-L-1)/(2L+2)} \to 0$. The contradiction completes the proof of the first part.

We now turn to the second part of the lemma. First, note that, using Jensen's

inequality, we obtain

$$\sqrt{n}\hat{\sigma}_n^K = \sqrt{n}(\sigma_n^2 + O_p(n^{-1/2}))^{K/2} \leq C\sqrt{n}\sigma_n^K + O_p(n^{-(K-2)/4}).$$

In particular, for $\varkappa_n = n^{-(K-L-1)/(2L+2)} (\ln n)^{-a}$ and $\sigma_n = o(n^{-\frac{1}{2L+2}-\epsilon})$, the above implies that $\sqrt{n}\hat{\sigma}_n^K/\varkappa_n = o_p(1)$. This, together with (D.1), implies that $\delta_n = 1$ with probability approaching one, which completes the proof. Q.E.D.

# E Details on the sufficient condition (17)

In this section, we demonstrate that the Jacobian $\Psi^*$ associated with the moment conditions (16) in Section 2.4 is guaranteed to have a full rank (for $K = 2$) when the sufficient condition (17) holds.

It is convenient to define $\varrho_i(x,\theta) \equiv u(x, S_i, \theta) \times \left(1, x, \ldots, x^{J-1}\right)'$ and $u_i(x,\theta) \equiv (x, S_i, \theta)$. Then

$$\begin{aligned} g(x, S_i, q, \theta) &\equiv \begin{pmatrix} u_i(x,\theta) \\ \varrho_i(x,\theta)\, x \\ \varrho_i(x,\theta)\, q \end{pmatrix}, \text{ so} \\ g_x^{(2)}(x, S_i, q, \theta) &= \begin{pmatrix} u_{x,i}^{(2)}(x,\theta) \\ \varrho_{x,i}^{(2)}(x,\theta)\, x + 2\varrho_{x,i}^{(1)}(x,\theta) \\ \varrho_{x,i}^{(2)}(x,\theta)\, q \end{pmatrix}. \end{aligned}$$

First, notice that

$$\Psi^* = \mathbb{E}\begin{pmatrix} u_{\theta,i}(X_i^*,\theta) & -u_{x,i}^{(2)}(X_i^*,\theta) \\ \varrho_{\theta,i}(X_i^*,\theta_0)\, X_i^* & -\varrho_{x,i}^{(2)}(X_i^*,\theta_0)\, X_i^* - 2\varrho_{x,i}^{(1)}(X_i^*,\theta_0) \\ \varrho_{\theta,i}(X_i^*,\theta_0)\,(\alpha_1 X_i^*) & -\varrho_{x,i}^{(2)}(X_i^*,\theta_0)\,(\alpha_1 X_i^*) \end{pmatrix},$$

where we used $\mathbb{E}[\varepsilon_{Q,i}|X_i^*, S_i, \varepsilon_i] = 0$. Dividing moments $J+2, \ldots, 2J+1$ by $\alpha_1$ and

subtracting them from the moments $2, \ldots, J+1$, we obtain

$$\begin{aligned}
\operatorname{Rk}(\Psi^*) &= \operatorname{Rk}\mathbb{E}\begin{pmatrix} u_{\theta,i}(X_i^*,\theta) & -u_{x,i}^{(2)}(X_i^*,\theta) \\ 0 & -2\varrho_{x,i}^{(1)}(X_i^*,\theta_0) \\ \varrho_{\theta,i}(X_i^*,\theta_0)X_i^* & -\varrho_{x,i}^{(2)}(X_i^*,\theta_0)X_i^* \end{pmatrix} \\
&= \operatorname{Rk}\mathbb{E}\begin{pmatrix} u_{\theta,i}(X_i^*,\theta) & -u_{x,i}^{(2)}(X_i^*,\theta) \\ \varrho_{\theta,i}(X_i^*,\theta_0)X_i^* & -\varrho_{x,i}^{(2)}(X_i^*,\theta_0)X_i^* \\ 0 & -2\varrho_{x,i}^{(1)}(X_i^*,\theta_0) \end{pmatrix} \\
&= \operatorname{Rk}\begin{pmatrix} H^* & \mathbb{E}\begin{pmatrix} u_{x,i}^{(2)}(X_i^*,\theta) \\ \varrho_{x,i}^{(2)}(X_i^*,\theta_0)X_i^* \end{pmatrix} \\ 0 & 2\mathbb{E}\left[\varrho_{x,i}^{(1)}(X_i^*,\theta_0)\right] \end{pmatrix}.
\end{aligned}$$

Here $\operatorname{Rk}(H^*) = \dim(\theta)$, because this is the rank identification condition for $\theta_0$ in the model without EIV. Thus, for $\Psi^*$ to have full rank $\dim(\theta)+1$, it is sufficient to have $\mathbb{E}\left[\varrho_{x,i}^{(1)}(X_i^*,\theta_0)\right] \neq 0$. Note that since $\mathbb{E}\left[u(X_i^*,S_i,\theta_0)\,|X_i^*\right] = 0$, we have

$$\mathbb{E}\left[\varrho_{x,i}^{(1)}(X_i^*,\theta_0)\right] = \mathbb{E}\left[u_x^{(1)}(X_i^*,S_i,\theta_0)\left(1,X_i^*,\ldots,(X_i^*)^{J-1}\right)'\right] \neq 0,$$

where the last equality follows from (17).

# F Additional Numerical Results

In this section, we provide additional numerical results for the experiments considered in Sections 3.2 and 4.1.

## F.1 Additional numerical results for Section 3.2

In this section, we provide additional simulation results for the same experiment as considered in Section 3.2 but for larger values of $\tau \in \{1, 1.5, 2\}$. The results are provided in Table 7 below, reporting the same statistics as in Table 4 in the main text.

Table 7: Simulation results for the multinomial logit model

| | MLE | | | | $K=2$ | | | | $K=4$ | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | bias, $10^{-2}$ | std, $10^{-2}$ | rmse, $10^{-2}$ | size | bias, $10^{-2}$ | std, $10^{-2}$ | rmse, $10^{-2}$ | size | bias, $10^{-2}$ | std, $10^{-2}$ | rmse, $10^{-2}$ | size |
| | | | | | | $\tau = 1$ | | | | | | |
| $\partial p_1/\partial x$ | -16.14 | 0.70 | 16.15 | 100.00 | -11.45 | 2.21 | 11.66 | 99.16 | 0.21 | 3.53 | 3.54 | 5.44 |
| $\partial p_1/\partial w_1$ | 11.82 | 1.40 | 11.91 | 100.00 | 8.49 | 2.46 | 8.84 | 96.14 | 0.48 | 2.88 | 2.92 | 6.30 |
| $\partial p_1/\partial w_2$ | 2.08 | 0.67 | 2.19 | 89.28 | 1.63 | 0.84 | 1.84 | 56.78 | 0.10 | 1.04 | 1.05 | 5.94 |
| $\partial p_2/\partial x$ | 8.78 | 0.65 | 8.80 | 100.00 | 6.59 | 1.50 | 6.76 | 97.64 | -0.04 | 2.53 | 2.53 | 6.42 |
| $\partial p_2/\partial w_1$ | -5.87 | 0.71 | 5.91 | 100.00 | -4.24 | 1.23 | 4.42 | 96.08 | -0.24 | 1.44 | 1.46 | 6.26 |
| $\partial p_2/\partial w_2$ | -4.03 | 1.27 | 4.22 | 89.50 | -3.20 | 1.62 | 3.59 | 57.34 | -0.18 | 2.09 | 2.10 | 6.24 |
| $\partial p_0/\partial x$ | 7.36 | 0.61 | 7.39 | 100.00 | 4.86 | 1.33 | 5.04 | 94.72 | -0.17 | 1.95 | 1.96 | 5.30 |
| $\partial p_0/\partial w_1$ | -5.95 | 0.73 | 6.00 | 100.00 | -4.24 | 1.24 | 4.42 | 96.14 | -0.24 | 1.44 | 1.46 | 6.28 |
| $\partial p_0/\partial w_2$ | 1.94 | 0.61 | 2.04 | 89.52 | 1.57 | 0.79 | 1.76 | 57.96 | 0.08 | 1.05 | 1.05 | 6.24 |
| | | | | | | $\tau = 3/2$ | | | | | | |
| $\partial p_1/\partial x$ | -19.01 | 0.50 | 19.01 | 100.00 | -16.59 | 1.32 | 16.64 | 100.00 | -2.29 | 3.67 | 4.33 | 16.56 |
| $\partial p_1/\partial w_1$ | 14.07 | 1.34 | 14.13 | 100.00 | 12.83 | 1.90 | 12.97 | 99.98 | 2.70 | 3.40 | 4.34 | 24.14 |
| $\partial p_1/\partial w_2$ | 2.31 | 0.64 | 2.40 | 95.70 | 2.20 | 0.75 | 2.32 | 87.72 | 0.56 | 1.18 | 1.31 | 12.18 |
| $\partial p_2/\partial x$ | 9.97 | 0.47 | 9.98 | 100.00 | 9.05 | 0.95 | 9.10 | 100.00 | 1.58 | 2.70 | 3.13 | 16.10 |
| $\partial p_2/\partial w_1$ | -6.96 | 0.69 | 6.99 | 100.00 | -6.40 | 0.96 | 6.48 | 99.98 | -1.35 | 1.71 | 2.18 | 24.08 |
| $\partial p_2/\partial w_2$ | -4.44 | 1.21 | 4.60 | 95.88 | -4.29 | 1.43 | 4.52 | 87.88 | -1.09 | 2.35 | 2.59 | 12.68 |
| $\partial p_0/\partial x$ | 9.04 | 0.46 | 9.05 | 100.00 | 7.54 | 0.94 | 7.60 | 100.00 | 0.71 | 2.05 | 2.18 | 9.46 |
| $\partial p_0/\partial w_1$ | -7.11 | 0.71 | 7.15 | 100.00 | -6.42 | 0.98 | 6.50 | 99.98 | -1.35 | 1.70 | 2.17 | 23.90 |
| $\partial p_0/\partial w_2$ | 2.13 | 0.58 | 2.21 | 95.90 | 2.09 | 0.69 | 2.20 | 88.08 | 0.53 | 1.17 | 1.29 | 13.00 |
| | | | | | | $\tau = 2$ | | | | | | |
| $\partial p_1/\partial x$ | -20.28 | 0.38 | 20.29 | 100.00 | -18.86 | 0.88 | 18.88 | 100.00 | -5.98 | 3.91 | 7.15 | 55.08 |
| $\partial p_1/\partial w_1$ | 15.09 | 1.32 | 15.15 | 100.00 | 14.91 | 1.66 | 15.00 | 100.00 | 5.87 | 3.87 | 7.03 | 55.08 |
| $\partial p_1/\partial w_2$ | 2.39 | 0.63 | 2.47 | 97.42 | 2.37 | 0.71 | 2.47 | 94.38 | 1.12 | 1.30 | 1.72 | 26.44 |
| $\partial p_2/\partial x$ | 10.45 | 0.36 | 10.46 | 100.00 | 9.97 | 0.69 | 9.99 | 100.00 | 3.80 | 2.82 | 4.73 | 47.94 |
| $\partial p_2/\partial w_1$ | -7.45 | 0.68 | 7.48 | 100.00 | -7.43 | 0.85 | 7.48 | 100.00 | -2.95 | 1.95 | 3.54 | 54.96 |
| $\partial p_2/\partial w_2$ | -4.58 | 1.18 | 4.73 | 97.42 | -4.61 | 1.36 | 4.81 | 94.54 | -2.20 | 2.54 | 3.36 | 27.46 |
| $\partial p_0/\partial x$ | 9.83 | 0.36 | 9.84 | 100.00 | 8.88 | 0.70 | 8.91 | 100.00 | 2.18 | 2.21 | 3.11 | 28.50 |
| $\partial p_0/\partial w_1$ | -7.64 | 0.71 | 7.68 | 100.00 | -7.48 | 0.87 | 7.53 | 100.00 | -2.92 | 1.93 | 3.50 | 54.80 |
| $\partial p_0/\partial w_2$ | 2.19 | 0.57 | 2.26 | 97.42 | 2.24 | 0.66 | 2.34 | 94.62 | 1.08 | 1.26 | 1.65 | 27.78 |

This table reports the simulated finite sample bias, standard deviation, RMSE, and size of the MLE and the MERM estimators and the corresponding t-tests for the partial derivatives $\partial p_j(x, w, \theta_0)/\partial x$, $\partial p_j(x, w, \theta_0)/\partial w_1$, $\partial p_j(x, w, \theta_0)/\partial w_2$ for $j \in \{1, 2, 0\}$ evaluated at the population mean. The true values of the marginal effects are $(\partial p_1/\partial x, \partial p_2/\partial x, \partial p_0/\partial x) = (0.222, -0.111, -0.111)$ and zeros for the rest. The results are based on 5,000 replications.

## F.2 Additional numerical results for Section 4.1

In this section, we provide additional numerical results for the experiment considered in Section 4.1. In particular, Tables 8 and 9 below report the same statistics as Table 6 in the main text, but for smaller and larger sample sizes $n = 1000$ and $n = 4000$.

Table 8: Choice of $K$ simulation results for the multinomial logit model, $n = 1000$

| | MLE | | $K = 2$ | | $K = 4$ | | data-driven $K$ | |
|---|---|---|---|---|---|---|---|---|
| | bias, $10^{-2}$ | rmse, $10^{-2}$ | bias, $10^{-2}$ | rmse, $10^{-2}$ | bias, $10^{-2}$ | rmse, $10^{-2}$ | bias, $10^{-2}$ | rmse, $10^{-2}$ |
| | | | | $\tau = 1/4$ | | | | |
| $\partial p_1/\partial x$ | -3.22 | 3.74 | 1.34 | 3.96 | 2.10 | 4.56 | 1.37 | 4.05 |
| $\partial p_1/\partial w_1$ | 2.29 | 3.27 | -0.13 | 3.32 | -0.57 | 3.39 | -0.15 | 3.34 |
| $\partial p_1/\partial w_2$ | 0.52 | 1.19 | -0.01 | 1.22 | -0.10 | 1.26 | -0.02 | 1.23 |
| $\partial p_2/\partial x$ | 1.93 | 2.55 | -0.73 | 2.78 | -1.18 | 3.11 | -0.75 | 2.82 |
| $\partial p_2/\partial w_1$ | -1.15 | 1.64 | 0.07 | 1.66 | 0.29 | 1.70 | 0.07 | 1.67 |
| $\partial p_2/\partial w_2$ | -1.02 | 2.36 | 0.03 | 2.46 | 0.21 | 2.53 | 0.04 | 2.46 |
| $\partial p_0/\partial x$ | 1.28 | 1.93 | -0.61 | 2.16 | -0.92 | 2.41 | -0.62 | 2.20 |
| $\partial p_0/\partial w_1$ | -1.14 | 1.63 | 0.07 | 1.66 | 0.29 | 1.70 | 0.07 | 1.67 |
| $\partial p_0/\partial w_2$ | 0.51 | 1.17 | -0.02 | 1.24 | -0.11 | 1.27 | -0.02 | 1.24 |
| | | | | $\tau = 1/2$ | | | | |
| $\partial p_1/\partial x$ | -8.97 | 9.10 | -1.82 | 4.10 | 1.76 | 4.57 | 1.35 | 4.82 |
| $\partial p_1/\partial w_1$ | 6.42 | 6.79 | 1.88 | 4.04 | -0.35 | 3.52 | -0.10 | 3.73 |
| $\partial p_1/\partial w_2$ | 1.32 | 1.67 | 0.42 | 1.33 | -0.05 | 1.32 | 0.00 | 1.34 |
| $\partial p_2/\partial x$ | 5.21 | 5.39 | 1.16 | 2.91 | -0.97 | 3.20 | -0.73 | 3.30 |
| $\partial p_2/\partial w_1$ | -3.21 | 3.39 | -0.94 | 2.02 | 0.18 | 1.76 | 0.05 | 1.87 |
| $\partial p_2/\partial w_2$ | -2.59 | 3.26 | -0.83 | 2.65 | 0.11 | 2.65 | 0.01 | 2.68 |
| $\partial p_0/\partial x$ | 3.76 | 3.95 | 0.66 | 2.18 | -0.79 | 2.50 | -0.62 | 2.56 |
| $\partial p_0/\partial w_1$ | -3.21 | 3.40 | -0.94 | 2.02 | 0.18 | 1.76 | 0.05 | 1.86 |
| $\partial p_0/\partial w_2$ | 1.27 | 1.60 | 0.41 | 1.33 | -0.06 | 1.33 | -0.01 | 1.35 |
| | | | | $\tau = 3/4$ | | | | |
| $\partial p_1/\partial x$ | -13.36 | 13.41 | -7.17 | 7.97 | 1.07 | 4.74 | 1.07 | 4.75 |
| $\partial p_1/\partial w_1$ | 9.68 | 9.90 | 5.59 | 6.67 | 0.17 | 3.79 | 0.17 | 3.79 |
| $\partial p_1/\partial w_2$ | 1.85 | 2.09 | 1.15 | 1.69 | 0.07 | 1.41 | 0.07 | 1.41 |
| $\partial p_2/\partial x$ | 7.48 | 7.57 | 4.26 | 4.91 | -0.54 | 3.36 | -0.54 | 3.36 |
| $\partial p_2/\partial w_1$ | -4.82 | 4.94 | -2.80 | 3.34 | -0.09 | 1.90 | -0.09 | 1.90 |
| $\partial p_2/\partial w_2$ | -3.59 | 4.05 | -2.27 | 3.33 | -0.13 | 2.82 | -0.13 | 2.82 |
| $\partial p_0/\partial x$ | 5.87 | 5.96 | 2.91 | 3.53 | -0.53 | 2.67 | -0.53 | 2.68 |
| $\partial p_0/\partial w_1$ | -4.86 | 4.97 | -2.79 | 3.34 | -0.08 | 1.90 | -0.08 | 1.90 |
| $\partial p_0/\partial w_2$ | 1.74 | 1.96 | 1.12 | 1.64 | 0.06 | 1.41 | 0.06 | 1.42 |

This table reports the simulated finite sample bias and RMSE of the MLE and the MERM estimators for the partial derivatives $\partial p_j(x, w, \theta_0)/\partial x$, $\partial p_j(x, w, \theta_0)/\partial w_1$, $\partial p_j(x, w, \theta_0)/\partial w_2$ for $j \in \{1, 2, 0\}$ evaluated at the population mean. The true values of the marginal effects are $(\partial p_1/\partial x, \partial p_2/\partial x, \partial p_0/\partial x) = (0.222, -0.111, -0.111)$ and zeros for the rest. The results are based on 5,000 replications.

Table 9: Choice of $K$ simulation results for the multinomial logit model, $n = 4000$

| | MLE | | $K = 2$ | | $K = 4$ | | data-driven $K$ | |
|---|---|---|---|---|---|---|---|---|
| | bias, $10^{-2}$ | rmse, $10^{-2}$ | bias, $10^{-2}$ | rmse, $10^{-2}$ | bias, $10^{-2}$ | rmse, $10^{-2}$ | bias, $10^{-2}$ | rmse, $10^{-2}$ |
| | | | | $\tau = 1/4$ | | | | |
| $\partial p_1/\partial x$ | -3.28 | 3.41 | 0.39 | 1.90 | 0.56 | 1.93 | 0.39 | 1.90 |
| $\partial p_1/\partial w_1$ | 2.32 | 2.57 | -0.08 | 1.59 | -0.15 | 1.57 | -0.08 | 1.59 |
| $\partial p_1/\partial w_2$ | 0.48 | 0.72 | -0.04 | 0.62 | -0.05 | 0.61 | -0.04 | 0.62 |
| $\partial p_2/\partial x$ | 1.97 | 2.13 | -0.23 | 1.35 | -0.33 | 1.37 | -0.23 | 1.35 |
| $\partial p_2/\partial w_1$ | -1.16 | 1.28 | 0.04 | 0.80 | 0.08 | 0.78 | 0.04 | 0.79 |
| $\partial p_2/\partial w_2$ | -0.96 | 1.43 | 0.07 | 1.23 | 0.11 | 1.23 | 0.07 | 1.23 |
| $\partial p_0/\partial x$ | 1.31 | 1.49 | -0.16 | 1.01 | -0.23 | 1.03 | -0.16 | 1.01 |
| $\partial p_0/\partial w_1$ | -1.16 | 1.29 | 0.04 | 0.80 | 0.08 | 0.78 | 0.04 | 0.80 |
| $\partial p_0/\partial w_2$ | 0.47 | 0.71 | -0.04 | 0.62 | -0.05 | 0.62 | -0.04 | 0.62 |
| | | | | $\tau = 1/2$ | | | | |
| $\partial p_1/\partial x$ | -9.00 | 9.03 | -1.52 | 2.41 | 0.49 | 2.03 | 0.48 | 2.04 |
| $\partial p_1/\partial w_1$ | 6.43 | 6.52 | 1.06 | 1.99 | -0.11 | 1.65 | -0.10 | 1.65 |
| $\partial p_1/\partial w_2$ | 1.28 | 1.38 | 0.21 | 0.68 | -0.04 | 0.65 | -0.04 | 0.65 |
| $\partial p_2/\partial x$ | 5.22 | 5.27 | 0.90 | 1.64 | -0.30 | 1.46 | -0.29 | 1.47 |
| $\partial p_2/\partial w_1$ | -3.21 | 3.25 | -0.53 | 1.00 | 0.05 | 0.82 | 0.05 | 0.83 |
| $\partial p_2/\partial w_2$ | -2.51 | 2.70 | -0.42 | 1.35 | 0.09 | 1.30 | 0.09 | 1.30 |
| $\partial p_0/\partial x$ | 3.78 | 3.83 | 0.62 | 1.19 | -0.20 | 1.10 | -0.19 | 1.10 |
| $\partial p_0/\partial w_1$ | -3.22 | 3.26 | -0.53 | 1.00 | 0.05 | 0.82 | 0.05 | 0.83 |
| $\partial p_0/\partial w_2$ | 1.23 | 1.32 | 0.21 | 0.67 | -0.05 | 0.65 | -0.05 | 0.65 |
| | | | | $\tau = 3/4$ | | | | |
| $\partial p_1/\partial x$ | -13.37 | 13.38 | -6.49 | 6.79 | 0.41 | 2.25 | 0.41 | 2.25 |
| $\partial p_1/\partial w_1$ | 9.68 | 9.73 | 4.44 | 4.84 | -0.02 | 1.80 | -0.02 | 1.80 |
| $\partial p_1/\partial w_2$ | 1.80 | 1.87 | 0.91 | 1.12 | -0.03 | 0.69 | -0.03 | 0.69 |
| $\partial p_2/\partial x$ | 7.48 | 7.50 | 3.83 | 4.08 | -0.24 | 1.63 | -0.24 | 1.63 |
| $\partial p_2/\partial w_1$ | -4.82 | 4.85 | -2.22 | 2.42 | 0.01 | 0.90 | 0.01 | 0.90 |
| $\partial p_2/\partial w_2$ | -3.50 | 3.63 | -1.80 | 2.22 | 0.06 | 1.39 | 0.06 | 1.39 |
| $\partial p_0/\partial x$ | 5.89 | 5.91 | 2.66 | 2.87 | -0.16 | 1.22 | -0.16 | 1.22 |
| $\partial p_0/\partial w_1$ | -4.86 | 4.89 | -2.22 | 2.42 | 0.01 | 0.90 | 0.01 | 0.90 |
| $\partial p_0/\partial w_2$ | 1.70 | 1.76 | 0.89 | 1.09 | -0.03 | 0.69 | -0.03 | 0.69 |

This table reports the simulated finite sample bias and RMSE of the MLE and the MERM estimators for the partial derivatives $\partial p_j(x, w, \theta_0)/\partial x$, $\partial p_j(x, w, \theta_0)/\partial w_1$, $\partial p_j(x, w, \theta_0)/\partial w_2$ for $j \in \{1, 2, 0\}$ evaluated at the population mean. The true values of the marginal effects are $(\partial p_1/\partial x, \partial p_2/\partial x, \partial p_0/\partial x) = (0.222, -0.111, -0.111)$ and zeros for the rest. The results are based on 5,000 replications.

# G MERM derivation when $\sigma_\varepsilon$ is not small

Note that $\tau$ can be small without $\sigma_\varepsilon$ being small in absolute magnitude. For example, suppose $\sigma_\varepsilon = 10$ and $\sigma_{X^*} = 100$. Then $\tau = 0.1$, so the measurement error is quite small relative to $\sigma_{X^*}$, and relying on the approximation $\tau \to 0$ is reasonable. At the same time, approximation $\sigma_\varepsilon \to 0$ may not be suitable for this example.

In this Appendix we show that the corrected moment conditions and the MERM estimator are valid without assuming that $\sigma_\varepsilon$ is small in absolute magnitude. In Section 2 we used Taylor expansions in $\varepsilon_i$ around $\varepsilon_i = 0$, with the remainder of order $\mathbb{E}\left[|\varepsilon_i|^{K+1}\right]$. When $\sigma_\varepsilon > 1$, term $O\left(\mathbb{E}\left[|\varepsilon_i|^{K+1}\right]\right)$ in equation (9) cannot be viewed

as a negligible remainder, because $\mathbb{E}\left[|\varepsilon_i|^{K+1}\right] > 1$ and, moreover, terms $\mathbb{E}\left[|\varepsilon_i|^k\right]$ increase rather than decrease with $k$.

In Section 2, to simplify the exposition, we have assumed that $X^*$ is scaled so that $\sigma_{X^*}$ is of order one. This in particular ensures that $\mathbb{E}\left[|\varepsilon_i|^k\right]$ decrease with $k$. We will now show that this assumption about the scale of $X^*$ is not necessary, and that the procedure remains valid without any such scaling.

We will show that by rescaling the Taylor expansions in Section 2 can be written in terms of powers of $\tau^k$, which necessarily decrease with $k$ when $\tau < 1$.

Remember the model of Section 2:

$$\mathbb{E}[g(X_i^*, S_i, \theta_0)] = 0, \quad X_i = X_i^* + \varepsilon_i, \quad \mathbb{E}[\varepsilon_i] = 0. \tag{G.1}$$

Let $\xi_i$ denote a random variable with $\mathbb{E}[\xi_i] = 0$ and $\mathbb{E}\left[\xi_i^2\right] = 1$, $\mathbb{E}\left[|\xi_i|^{L+1}\right]$ is bounded, and $\varepsilon_i \equiv \sigma_\varepsilon \xi_i$. Also, let us denote

$$\tau \equiv \sigma_\varepsilon / \sigma_{X^*}, \quad \widetilde{X}_i \equiv X_i / \sigma_{X^*}, \quad \widetilde{X}_i^* \equiv X_i^* / \sigma_{X^*}, \quad \widetilde{g}(\widetilde{x}, s, \theta) \equiv g(\sigma_{X^*}\widetilde{x}, s, \theta).$$

Then, we can rewrite equation (G.1) as

$$\mathbb{E}[\widetilde{g}(\widetilde{X}_i^*, S_i, \theta_0)] = 0, \quad \widetilde{X}_i = \widetilde{X}_i^* + \tau\xi_i, \quad \mathbb{E}[\xi_i] = 0.$$

Expand $\widetilde{g}(\widetilde{X}_i, S_i, \theta) = \widetilde{g}(\widetilde{X}_i^* + \tau\xi_i, S_i, \theta)$ around $\tau = 0$ to obtain

$$\mathbb{E}[\widetilde{g}(\widetilde{X}_i, S_i, \theta)] = \mathbb{E}[\widetilde{g}(\widetilde{X}_i^*, S_i, \theta)] + \sum_{k=2}^{K} \frac{\tau^k \mathbb{E}\left[\xi_i^k\right]}{k!} \mathbb{E}\left[\widetilde{g}_x^{(k)}(\widetilde{X}_i^*, S_i, \theta)\right] + O(\tau^{K+1}),$$

which is similar to equation (9), except $\mathbb{E}\left[\varepsilon_i^k\right]$ is replaced by $\tau^k \mathbb{E}\left[\xi_i^k\right]$, and $\widetilde{X}_i$, $\widetilde{X}_i^*$, $\widetilde{g}$ are replaced by $X_i$, $X_i^*$, $g$. Then, the corrected moment condition has the form

$$\widetilde{\psi}(\widetilde{X}_i, S_i, \theta, \widetilde{\gamma}) = \widetilde{g}(\widetilde{X}_i, S_i, \theta) - \sum_{k=2}^{K} \widetilde{\gamma}_k \widetilde{g}_x^{(k)}(\widetilde{X}_i, S_i, \theta), \tag{G.2}$$

where true parameter values $\widetilde{\gamma}_0$ are $\widetilde{\gamma}_{02} = \tau^2 \mathbb{E}\left[\xi_i^2\right]/2 = \tau^2/2$, $\widetilde{\gamma}_{03} = \tau^3 \mathbb{E}\left[\xi_i^3\right]/6$, and $\widetilde{\gamma}_{0k} = \frac{\tau^k \mathbb{E}\left[\xi_i^k\right]}{k!} - \sum_{\ell=2}^{k-2} \frac{\tau^{k-\ell} \mathbb{E}\left[\xi_i^{k-\ell}\right]}{(k-\ell)!} \widetilde{\gamma}_{0\ell}$ for $k \geq 4$.

We will now show that

$$\gamma_{0k} = \sigma_{X^*}^k \widetilde{\gamma}_{0k} \text{ for all } k \geq 2.$$

First, $\gamma_{02} = \mathbb{E}[\varepsilon_i^2]/2 = \mathbb{E}\left[(\sigma_\varepsilon \xi_i)^2\right]/2 = \sigma_{X^*}^2 \widetilde{\gamma}_{02}$, $\gamma_{03} = \mathbb{E}[\varepsilon_i^3]/6 = \sigma_{X^*}^3 \widetilde{\gamma}_{03}$ by defini-

tion. Then, for $k \geq 4$, by induction we have

$$\begin{aligned}
\gamma_{0k} &= \frac{\mathbb{E}\left[\varepsilon_i^k\right]}{k!} - \sum_{\ell=2}^{k-2} \frac{\mathbb{E}\left[\varepsilon_i^{k-\ell}\right]}{(k-\ell)!}\gamma_{0\ell} \\
&= \sigma_{X^*}^k \left( \frac{\left(\sigma_\varepsilon/\sigma_{X^*}\right)^k \mathbb{E}\left[\xi_i^k\right]}{k!} - \sum_{\ell=2}^{k-2} \frac{\left(\sigma_\varepsilon/\sigma_{X^*}\right)^{k-\ell} \mathbb{E}\left[\xi_i^{k-\ell}\right]}{(k-\ell)!} \frac{\gamma_{0\ell}}{\sigma_{X^*}^\ell} \right) \\
&= \sigma_{X^*}^k \left( \frac{\tau^k \mathbb{E}\left[\xi_i^k\right]}{k!} - \sum_{\ell=2}^{k-2} \frac{\tau^{k-\ell} \mathbb{E}\left[\xi_i^{k-\ell}\right]}{(k-\ell)!} \widetilde{\gamma}_{0\ell} \right) = \sigma_{X^*}^k \widetilde{\gamma}_{0k}.
\end{aligned}$$

Finally, let us now show that moment condition $\widetilde{\psi}$ in equation (G.2) is numerically identical to $\psi$ in equation (11) with $\gamma_k = \sigma_{X^*}^k \widetilde{\gamma}_k$. Note that for $\widetilde{x} = x/\sigma_{X^*}$ we have $\widetilde{g}_{\widetilde{x}}^{(k)}(\widetilde{x}, s, \theta) \equiv \nabla_{\widetilde{x}}^k g(\sigma_{X^*}\widetilde{x}, s, \theta) = \sigma_{X^*}^k g_a^{(k)}(a, s, \theta)|_{a=\sigma_{X^*}\widetilde{x}} = \sigma_{X^*}^k g_x^{(k)}(x, s, \theta)$, and hence

$$\begin{aligned}
\widetilde{\psi}(\widetilde{X}_i, S_i, \theta, \widetilde{\gamma}) &= g(\sigma_{X^*}\widetilde{X}_i, S_i, \theta) - \sum_{k=2}^{K} \left(\widetilde{\gamma}_k \sigma_{X^*}^k\right) g_x^{(k)}(\sigma_{X^*}\widetilde{X}_i, S_i, \theta) \\
&= g(X_i, S_i, \theta) - \sum_{k=2}^{K} \left(\widetilde{\gamma}_k \sigma_{X^*}^k\right) g_x^{(k)}(X_i, S_i, \theta) \\
&= \psi(X_i, S_i, \theta, \gamma).
\end{aligned}$$

# H Some Implementation Details

**Numerical Optimization**

Since $\overline{\psi}(\theta, \gamma)$ is a linear function of $\gamma$ it can be profiled out of the quadratic form $\hat{Q}(\theta, \gamma)$. Thus, the criterion function only needs to be minimized numerically over $\theta$.

**Choice of the weighting matrix $\hat{\Xi}$**

As for the standard GMM estimator, the optimal weighting matrix can be estimated by

$$\hat{\Xi}_{\text{eff}} \equiv \hat{\Omega}_{\psi\psi}^{-1}(\tilde{\theta}, \tilde{\gamma}),$$

where $\tilde{\theta}$ and $\tilde{\gamma}$ are some preliminary estimators of $\theta_0$ and $\gamma_0$, and $\hat{\Omega}_{\psi\psi}(\theta, \gamma) \equiv n^{-1}\sum_{i=1}^n \psi_i(\theta, \gamma)\psi_i(\theta, \gamma)'$. One example of such a preliminary estimator would be the 1-step (GMM-)MERM estimator using $\hat{\Xi}_{\text{GMM1}} \equiv \hat{\Omega}_{\psi\psi}^{-1}(\hat{\theta}_{\text{Naive}}, 0)$ as the first-step GMM weighting matrix, where $\hat{\theta}_{\text{Naive}}$ is a naive estimator of $\theta_0$ that ignores EIV. Note that $\hat{\Omega}_{\psi\psi}(\hat{\theta}_{\text{Naive}}, 0) = \hat{\Omega}_{gg}\left(\hat{\theta}_{\text{Naive}}\right)$, where $\hat{\Omega}_{gg}(\theta) \equiv n^{-1}\sum_{i=1}^n g_i(\theta)g_i(\theta)'$.

One may also consider the regularized version of the efficient weighting matrix estimator $\hat{\Xi}_{\text{eff,R}} \equiv \hat{\Omega}_{\psi\psi}^{-1}(\tilde{\theta}, 0)$. Since $\gamma_0 \to 0$, using the regularized version $\hat{\Xi}_{\text{eff,R}}$ does not lead to a loss of efficiency. Moreover, our simulation studies suggest that using the regularized weighting matrix $\hat{\Xi}_{\text{eff,R}}$ results in better finite sample performance of the MERM estimator and, hence, is recommended in practice.

Although not indicated by the notation in equation (13), the weighting matrix $\hat{\Xi} \equiv \hat{\Xi}(\theta, \gamma)$ is allowed to be a function of $\theta$ and $\gamma$. For example, Continuously Updating GMM Estimator (CUE) corresponds to taking $\hat{\Xi}_{\text{CUE}}(\theta, \gamma) \equiv \hat{\Omega}_{\psi\psi}^{-1}(\theta, \gamma)$. Similarly to $\hat{\Xi}_{\text{eff,R}}$, one may also consider $\hat{\Xi}_{\text{CUE,R}}(\theta, \gamma) \equiv \hat{\Omega}_{\psi\psi}^{-1}(\theta, 0)$ without introducing any loss of efficiency. In contrast to the criterion function of the CUE estimator, criterion function of $\hat{Q}_{\text{CUE,R}}(\theta, \gamma)$ is quadratic in $\gamma$. This implies that $\gamma$ can be profiled out analytically. This simplifies the numerical optimization problem reducing it to minimizing $\hat{Q}_{\text{CUE,R}}(\theta, \hat{\gamma}(\theta))$ over $\theta \in \Theta$. Then, the dimension of the optimization parameter $\theta$ for the corrected moment condition problem remains the same as for the original (naive) estimation problem without the EIV correction.

**Estimation of the asymptotic variance $\Sigma$**

Theorem 2 shows that the MERM estimator $\hat{\beta} = (\hat{\theta}', \hat{\gamma}')'$ behaves like a standard GMM estimator based on the corrected moment function $\psi(\theta, \gamma)$. The researcher can rely on the standard GMM inference procedures. The asymptotic variance of $\hat{\beta}$ can be consistently estimated by

$$\hat{\Sigma} \equiv (\hat{\Psi}'\hat{\Xi}\hat{\Psi})^{-1}\hat{\Psi}'\hat{\Xi}\hat{\Omega}_{\psi\psi}\hat{\Xi}\hat{\Psi}(\hat{\Psi}'\hat{\Xi}\hat{\Psi})^{-1},$$

where, $\hat{\Xi}$ is the chosen weighting matrix, and $\hat{\Psi} \equiv \overline{\Psi}(\hat{\theta}, \hat{\gamma}) = n^{-1}\sum_{i=1}^{n}\Psi_i(\hat{\theta}, \hat{\gamma})$ and $\hat{\Omega}_{\psi\psi} = \hat{\Omega}_{\psi\psi}(\hat{\theta}, \hat{\gamma})$ are estimators of $\Psi$ and $\Omega_{\psi\psi}$.

# I Implementation Details of the Empirical Illustration

In this section, we provide additional details on the implementation of the numerical experiment in Section 3.3.

**Data**

The original dataset is the ModeCanada dataset supplied with the R package `mlogit`.

This dataset has been extensively used in transportation research. For a detailed description of the dataset see, for example, Koppelman and Wen (2000), Wen and Koppelman (2001), and Hansen (2022). As in Koppelman and Wen (2000), we use only the subset of travelers who chose train, air, or car (and had all of those alternatives available for them), which leaves $n = 2769$ observations.

**Monte-Carlo design**

We choose $\theta_0$ to be the MLE estimates using the considered dataset, which are reported in the table bellow.

| | $\theta_1$ | $\theta_2$ | $\theta_3$ | $\theta_4$ | $\theta_5$ | $\theta_6$ | $\theta_7$ | $\theta_8$ |
|---|---|---|---|---|---|---|---|---|
| Estimates | 0.0355 | 0.2976 | -2.0891 | 0.0079 | -0.9900 | 1.8794 | -0.0223 | -0.0149 |
| Std. Err. | 0.0036 | 0.0844 | 0.4674 | 0.0036 | 0.0876 | 0.2037 | 0.0038 | 0.0008 |

To generate the simulated samples, we randomly draw the covariates (with replacement) from their joint empirical distribution. To ensure identification of the model, we also generate an instrumental variable $Z_i$ as

$$Z_i = \kappa \, Income_i^* / \sigma_{Income^*} + \sqrt{1-\kappa^2}\zeta_i,$$

where $\sigma_{Income^*} \approx 17.5$ is the standard deviation of $Income^*$, $\kappa = 0.5$, and $\zeta_i$ are i.i.d. draws from $N(0,1)$ (which are also independent from all the over variables). Note that the instrument $Z_i$ is "caused by $X_i^*$". For example, $Z_i$ can be some (noisy) measure of individual consumption.

**Moments**

To simplify the notation, let $X_i^* \equiv Income_i^*$, $X_i \equiv Income_i$, $R_i \equiv Urban_i$, $R_{ij} \equiv (Price_{ij}, InTime_{ij})'$ for $j \in \{0,1,2\}$, and $W_i \equiv (R_i, R_{i1}', R_{i2}', R_{i0}')'$. Also let $Y_{ij} \equiv \mathbb{1}\{j = \text{argmax}_{j' \in \{0,1,2\}} U_{ij'}\}$ for $j \in \{0,1,2\}$, $Y_i \equiv (Y_{i1}, Y_{i2}, Y_{i0})'$, and $p_j(x,w,\theta) \equiv \mathbb{P}(Y_{ij} = 1 | X_i^* = x, W_i = w; \theta)$ with $w \equiv (r, r_1', r_2', r_0')$, so

$$p_1(x,w,\theta) = \frac{e^{\theta_1 x + \theta_2 r + \theta_3 + (\theta_7,\theta_8) r_1}}{e^{\theta_1 x + \theta_2 r + \theta_3 + (\theta_7,\theta_8) r_1} + e^{\theta_4 x + \theta_5 r + \theta_6 + (\theta_7,\theta_8) r_2} + e^{(\theta_7,\theta_8) r_0}},$$

$$p_2(x,w,\theta) = \frac{e^{\theta_4 x + \theta_5 r + \theta_6 + (\theta_7,\theta_8) r_2}}{e^{\theta_1 x + \theta_2 r + \theta_3 + (\theta_7,\theta_8) r_1} + e^{\theta_4 x + \theta_5 r + \theta_6 + (\theta_7,\theta_8) r_2} + e^{(\theta_7,\theta_8) r_0}},$$

and $p_0(x, w, \theta) = 1 - p_1(x, w, \theta) - p_2(x, w, \theta)$. Then, the original moment function takes the form of

$$g(x, w, y, z, \theta) = \left((y_1 - p_1(x, w, \theta))\, \varphi_1(x, z, w)', (y_2 - p_2(x, w, \theta))\, \varphi_2(x, z, w)'\right)'.$$

and $\varphi_j(x, z, w) = (1, x, z, x^2, z^2, x^3, z^3, r, (r_j - r_0)')'$ for $K = 2$ and $\varphi_j(x, z, w) = (1, x, z, x^2, xz, z^2, x^3, x^2z, xz^2, z^3, r, (r_j - r_0)')'$ for $K = 4$.

**Income Elasticities**

In Section 3.3, we focus on estimation of and inference on the income elasticities

$$\frac{\partial \ln p_j}{\partial \ln x}(x, w, \theta) = \frac{x}{p_j(x, w, \theta)} \frac{\partial p_j(x, w, \theta)}{\partial x}.$$

We report the results are for the income elasticities evaluated at the sample mean of $X^*$ and $W$ in the original sample.

**Estimation of and Inference on the $\theta_0$**

In Table 10 below, we also report the estimation and inference results for the vector of parameters $\theta_0$ underlying the reported results about elasticities.

Table 10: Simulation results for the empirically calibrated conditional logit model

| | MLE | | | | $K=2$ | | | | $K=4$ | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | bias | std | rmse | size | bias | std | rmse | size | bias | std | rmse | size |
| | | | | | | $\tau = 1/4$ | | | | | | |
| $\theta_1$ | -0.0021 | 0.0035 | 0.0041 | 8.70 | 0.0001 | 0.0042 | 0.0042 | 5.48 | 0.0005 | 0.0057 | 0.0058 | 7.38 |
| $\theta_2$ | 0.0047 | 0.0932 | 0.0933 | 5.10 | 0.0028 | 0.0957 | 0.0957 | 5.36 | 0.0022 | 0.0960 | 0.0960 | 5.40 |
| $\theta_3$ | 0.1152 | 0.4452 | 0.4599 | 6.00 | -0.0048 | 0.4821 | 0.4821 | 5.94 | -0.0251 | 0.5336 | 0.5342 | 6.68 |
| $\theta_4$ | -0.0004 | 0.0031 | 0.0031 | 4.52 | -0.0001 | 0.0034 | 0.0034 | 5.32 | -0.0001 | 0.0036 | 0.0036 | 6.86 |
| $\theta_5$ | -0.0023 | 0.0894 | 0.0895 | 5.18 | -0.0088 | 0.0918 | 0.0922 | 5.54 | -0.0113 | 0.0922 | 0.0929 | 5.96 |
| $\theta_6$ | 0.0232 | 0.1821 | 0.1836 | 4.64 | 0.0250 | 0.1982 | 0.1998 | 5.72 | 0.0329 | 0.2089 | 0.2115 | 6.74 |
| $\theta_7$ | -0.0001 | 0.0035 | 0.0035 | 5.58 | -0.0002 | 0.0036 | 0.0036 | 6.24 | -0.0003 | 0.0036 | 0.0037 | 6.04 |
| $\theta_8$ | -0.0001 | 0.0007 | 0.0007 | 4.82 | -0.0001 | 0.0007 | 0.0007 | 5.48 | -0.0002 | 0.0007 | 0.0007 | 5.58 |
| | | | | | | $\tau = 1/2$ | | | | | | |
| $\theta_1$ | -0.0073 | 0.0032 | 0.0080 | 60.08 | -0.0016 | 0.0043 | 0.0046 | 6.86 | 0.0005 | 0.0061 | 0.0061 | 6.60 |
| $\theta_2$ | 0.0109 | 0.0930 | 0.0936 | 5.18 | 0.0050 | 0.0959 | 0.0960 | 5.54 | 0.0026 | 0.0964 | 0.0965 | 5.36 |
| $\theta_3$ | 0.4080 | 0.4452 | 0.6039 | 17.12 | 0.0936 | 0.4874 | 0.4963 | 6.58 | -0.0263 | 0.5475 | 0.5481 | 6.38 |
| $\theta_4$ | -0.0012 | 0.0029 | 0.0031 | 6.46 | -0.0003 | 0.0035 | 0.0035 | 5.22 | -0.0002 | 0.0038 | 0.0038 | 6.52 |
| $\theta_5$ | -0.0006 | 0.0894 | 0.0894 | 5.16 | -0.0083 | 0.0919 | 0.0923 | 5.52 | -0.0110 | 0.0924 | 0.0930 | 5.92 |
| $\theta_6$ | 0.0655 | 0.1752 | 0.1870 | 6.22 | 0.0348 | 0.2035 | 0.2064 | 5.86 | 0.0326 | 0.2158 | 0.2183 | 6.42 |
| $\theta_7$ | -0.0003 | 0.0035 | 0.0036 | 5.64 | -0.0003 | 0.0036 | 0.0037 | 6.34 | -0.0003 | 0.0037 | 0.0037 | 6.06 |
| $\theta_8$ | -0.0001 | 0.0007 | 0.0007 | 5.06 | -0.0001 | 0.0007 | 0.0007 | 5.54 | -0.0002 | 0.0007 | 0.0007 | 5.48 |
| | | | | | | $\tau = 3/4$ | | | | | | |
| $\theta_1$ | -0.0132 | 0.0029 | 0.0135 | 99.34 | -0.0056 | 0.0043 | 0.0071 | 25.12 | 0.0003 | 0.0065 | 0.0065 | 6.06 |
| $\theta_2$ | 0.0180 | 0.0923 | 0.0940 | 5.36 | 0.0102 | 0.0961 | 0.0966 | 5.76 | 0.0033 | 0.0973 | 0.0973 | 5.44 |
| $\theta_3$ | 0.7336 | 0.4496 | 0.8604 | 41.66 | 0.3203 | 0.4859 | 0.5820 | 12.00 | -0.0130 | 0.5666 | 0.5667 | 6.20 |
| $\theta_4$ | -0.0024 | 0.0026 | 0.0035 | 14.00 | -0.0009 | 0.0035 | 0.0036 | 5.94 | -0.0002 | 0.0041 | 0.0041 | 5.76 |
| $\theta_5$ | 0.0021 | 0.0890 | 0.0891 | 5.08 | -0.0071 | 0.0921 | 0.0924 | 5.68 | -0.0109 | 0.0926 | 0.0932 | 5.82 |
| $\theta_6$ | 0.1204 | 0.1654 | 0.2046 | 9.76 | 0.0648 | 0.2048 | 0.2148 | 6.56 | 0.0334 | 0.2294 | 0.2318 | 5.98 |
| $\theta_7$ | -0.0004 | 0.0036 | 0.0036 | 6.00 | -0.0004 | 0.0036 | 0.0037 | 6.34 | -0.0003 | 0.0037 | 0.0037 | 6.06 |
| $\theta_8$ | -0.0001 | 0.0007 | 0.0007 | 5.54 | -0.0002 | 0.0007 | 0.0008 | 6.00 | -0.0002 | 0.0007 | 0.0008 | 5.42 |

This table reports the simulated finite sample bias, standard deviation, RMSE, and size of the MLE and the MERM estimators and the corresponding t-tests for the components of $\theta_0$. The true value of the parameters of interest are $\theta_0 = (0.0355, 0.2976, -2.0891, 0.0079, -0.9900, 1.8794, -0.0223, -0.0149)'$. The results are based on 5,000 replications.