ECONSTOR
Make Your Publications Visible.

A Service of

ZBW Leibniz-Informationszentrum Wirtschaft
Leibniz Information Centre for Economics

Zeleneev, Andrei; Evdokimov, Kirill S.

Working Paper

Simple estimation of semiparametric models with measurement errors

cemmap working paper, No. CWP10/23

Provided in Cooperation with:
Institute for Fiscal Studies (IFS), London

Suggested Citation: Zeleneev, Andrei; Evdokimov, Kirill S. (2023) : Simple estimation of semiparametric models with measurement errors, cemmap working paper, No. CWP10/23, Centre for Microdata Methods and Practice (cemmap), London, https://doi.org/10.47004/wp.cem.2023.1023

This Version is available at:
https://hdl.handle.net/10419/284134

Standard-Nutzungsbedingungen:

Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden.

Sie dürfen die Dokumente nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, öffentlich zugänglich machen, vertreiben oder anderweitig nutzen.

Sofern die Verfasser die Dokumente unter Open-Content-Lizenzen (insbesondere CC-Lizenzen) zur Verfügung gestellt haben sollten, gelten abweichend von diesen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.

Terms of use:

Documents in EconStor may be saved and copied for your personal and scholarly purposes.

You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.

If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.

WWW.ECONSTOR.EU

Mitglied der Leibniz-Gemeinschaft

cemmap
centre for microdata methods and practice

Simple estimation of semiparametric models with measurement errors

Andrei Zeleneev
Kirill S. Evdokimov

The Institute for Fiscal Studies
Department of Economics, UCL

cemmap working paper CWP10/23

Economic and Social Research Council

Simple Estimation of Semiparametric Models with Measurement Errors

Kirill S. Evdokimov* Andrei Zeleneev†‡§

This version: May 10, 2023

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) may be heavily biased with the confidence intervals having 0% coverage. Our approach involves no nonparametric estimation, which is particularly important for applications with multiple covariates, and settings with multivariate, serially correlated, or non-classical EIV.

Keywords: errors-in-variables, nonstandard asymptotic approximation, non-classical measurement errors, nonparametric identification

*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, and Severo Ochoa Programme CEX2019-000915-S.

1 Introduction

Measurement errors are a common problem for empirical studies. While the standard instrumental variables approach can be used to remove the Errors-In-Variables (EIV) bias in linear models, as pointed out by Amemiya (1985), nonlinear models require more 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 general nonlinear moment condition models

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

where $g(\cdot)$ is a vector of moment functions and θ_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 ε_i:

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

The measurement error can be classical or non-classical. 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 $\mathbb{E}[g(X_i, S_i, \theta_0)] \neq 0$.[2]

The general moment condition model (1) encompasses a wide variety of semi-parametric models. Most of the existing literature on EIV focuses on the nonlinear regression (NLR) model:

Example (NLR)**.** Let Y_i denote the 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 ρ known up to the parameter θ. 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 $\theta \equiv (\theta_x', \theta_w')'$.

[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] Mismeasured variables X_i do not need to be covariates. In general nonlinear models, measurement errors in any of the variables (including outcomes) may bias naive estimators.

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, includes powers of x, their interactions with z, and w.[3] ■

To provide a practical estimation approach for such a general class of models, we focus on the 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, e.g., $\tau \lesssim 0.5$ (we later discuss how the appropriateness of this assumption can be checked). The absolute magnitude of the measurement error σ_ε does not need to be small. 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 θ.

We propose to bias correct the original moments $g(\cdot)$, which in turn removes the bias of the corresponding estimator of θ_0. This bias correction depends on some moments of the distribution of the measurement errors that are unknown. Another 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 θ and additional parameters γ that govern the bias correction. The true parameter value γ_0 is associated with (possibly conditional) low-order moments of ε_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 θ_0 and γ_0. The estimator can be computed using any standard software for GMM estimation. Joint estimation of parameters θ_0 and γ_0 using the corrected moment conditions effectively robustifies moment conditions $g(\cdot)$ against the impact of the measurement errors.

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

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 τ as slowly shrinking with the sample size. Standard asymptotics considers τ 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.[4] By considering τ 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 τ is small or moderate.[5]

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

[4]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. See also the review of Bound, Brown, and Mathiowetz (2001).

[5]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.

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 $f_{X_i^*|W}$. Avoiding nonparametric estimation has at least two advantages. First, since the majority of empirical applications include at least a handful of 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 γ_0 is finite-dimensional.

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).

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. Fourth, our approach allows relaxing the assumption of classical measurement errors.

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, researchers 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^*.[6] One can also take a nonclassical, nonlinear (e.g., discretized or censored), or biased measurement of X_i^* as an instrument in the MERM approach. In Section 4, we study identification of the NLR model in the MERM framework, and show that this model is nonparametrically identified.

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 $\sqrt{n}$

[6]The importance of heterogeneity of the effects of the instruments in empirical applications has been widely recognized, e.g., see Imbens and Angrist (1994); Heckman and Vytlacil (1998); Imbens and Newey (2009). Note that such heterogeneity is ruled out by the EIV-robust methods that rely on the additive control variable assumption for identification, i.e., assume that $X_i^* = m(Z_i) + V_i$ with the control variable V_i independent from Z_i. In contrast, in Section 4 we illustrate identification in a random coefficient first stage model.

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 establishes nonparametric identification of the nonlinear regression in our setting and motivates the MERM approach from a nonparametric perspective. Section 5 considers several extensions of the framework.

2 Moderate Measurement Errors Framework

To present the main ideas we first consider the case of univariate X_i^* and classical measurement error ε_i. Later we consider multivariate X_i^* and non-classical measurement errors ε_i.

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.

The magnitude of the EIV bias of such parameters is also invariant to the scaling; changing the units of measurement of X_i does not meaningfully change the EIV problem. For simplicity of exposition, it is convenient to assume that X_i^* is scaled so that σ_{X^*} is of order one and, correspondingly, 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 D, the proposed MERM estimator does not require any normalizations in practice. Following the rest of the literature, we assume that $\mathbb{E}[\varepsilon_i] = 0$.[7]

[7] A location normalization such as $\mathbb{E}[\varepsilon_i] = 0$ is usually necessary because it is not possible to

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

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

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

This expansion implies that $\mathbb{E}[g(X_i, S_i, \theta_0)] = O(\sigma_\varepsilon^2) = O(\tau_n^2)$. As a result, a naive estimator that ignores the EIV and uses X_i in place of X_i^* has EIV bias of order τ_n^2.[8] Bias of the naive estimator should be compared with its standard error, which is of order $n^{-1/2}$. 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 τ_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 θ_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

separately identify the means $\mathbb{E}[X_i^*]$ and $\mathbb{E}[\varepsilon_i]$.

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

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 θ_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 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 $E\left[\varepsilon_i^2\right] = O\left(\tau_n^2\right)$.[9] 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 θ. Specifically, let $\gamma_{02} \equiv \mathbb{E}[\varepsilon_i^2]/2$ denote the true value of parameter γ_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 ψ is a moment function parameterized by θ and γ, 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 γ_{02}, and the second equality follows from equation (1). Hence, the corrected moment conditions ψ can be used to jointly estimate the true parameters θ_0 and γ_{02} by a GMM estimator.[10]

Remark 1. If $\mathbb{E}[\varepsilon_i^3] = 0$ (e.g., if the distribution of ε_i is symmetric), the remainder

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

[10] 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).

in equation (3) is of a smaller order $O(\tau_n^4)$. Hence, the corrected moments (8) remain valid for larger values of τ_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 $o(n^{-1/4})$.

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 τ_n converging to zero at a slower rate. In finite samples this corresponds to the asymptotics providing good approximations for larger values of τ_n, i.e., large measurement errors. Expanding $g(X_i^* + \varepsilon_i, S_i, \theta)$ around $\varepsilon_i = 0$ we have,

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

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.

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 τ_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 τ_n^K, with the asymptotically negligible remainder of order $O\left(\tau_n^{K+1}\right)$. For $K \geq 4$, terms of order τ_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 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 γ_{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. *ε_i is independent from (X_i^*, S_i) and $\mathbb{E}[\varepsilon_i] = 0$.*

Assumption CME is the classical measurement error assumption. We relax this assumption later in Section 5.2. 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 ψ can be used to jointly estimate parameters θ_0 and γ_0. The total number of parameters to be estimated is now $\dim(\beta) = \dim(\theta) + K - 1$. Thus, joint estimation of θ_0 and γ_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.

Measurement Error Robust Moments (MERM) estimator The MERM estimator jointly estimates the parameters θ_0 and γ_0 using moment conditions ψ. 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 Θ and Γ are the parameter spaces for θ and γ. 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.

Under some 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, and A.1-A.4 in Appendix A,*

$$n^{1/2} \Sigma^{-1/2} (\hat{\beta} - \beta_0) \xrightarrow{d} N(0, I_{\dim(\beta)}), \ \textit{where} \tag{14}$$

$$\Sigma \equiv (\Psi' \Xi \Psi)^{-1} \Psi' \Xi \Omega_{\psi\psi} \Xi \Psi (\Psi' \Xi \Psi)^{-1}.$$

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 θ_0. The asymptotic variance Σ 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 $\hat{\Xi} \to_p \Xi$. The key condition for the validity of the estimator is that the Jacobian matrix Ψ has full rank; we discuss this in detail below.

Remark 2. 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 τ_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 ψ is constructed, estimation of and inference about parameters β_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 ψ, and the conventional standard errors, tests, and confidence intervals are valid.

Remark 3. Researchers may be interested in average effects of the form $\lambda_0 \equiv \mathbb{E}[\lambda(X_i^*, S_i, \theta_0)]$. In the NLR, one may be interested in the average partial effect 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 λ is linear in X_i^*. Instead, one can 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. 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 $\varepsilon_i \sim N(0, \sigma_\varepsilon^2)$, $\sigma_\varepsilon = 0.5$, 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 $\mathbb{E}[\varepsilon_i^2]/2! = 0.125$, $\mathbb{E}[\varepsilon_i^4]/4! \approx 0.008$, and $\mathbb{E}[\varepsilon_i^6]/6! \approx 0.0003$.

Remark 5. 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, $\gamma_{04} = (\mathbb{E}[\varepsilon_i^4] - 6\sigma_\varepsilon^4)/24$ is negative for many distributions, including normal. For instance, in the example of Remark 4, $\gamma_{04} \approx -0.0026$. 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 γ_{0k}. Since there is a one-to-one relationship between γ_0 and the moments $\mathbb{E}\left[\varepsilon_i^\ell\right]$, parameter space Γ for γ_0 can incorporate restrictions that the moments must satisfy (e.g., $\sigma_\varepsilon^2 \geq 0$ and $\mathbb{E}[\varepsilon_i^4] \geq \sigma_\varepsilon^4$). Such restrictions can increase the efficiency of the estimator and the power of tests.

Remark 6. No parametric assumptions are imposed on the distribution of ε_i, i.e. the distribution of ε_i is treated nonparametrically. The regularity conditions restrict only the magnitude of the moments of ε_i. The approach imposes no restrictions

on the smoothness of the distributions of X_i^* and ε_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 ε_i to have a point mass at zero (a fraction of the population may have a zero measurement or recall error).

Remark 7. Considering larger K allows τ_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 γ_0 and thus typically increases the variance of $\hat{\theta}$.

Remark 8. 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. In either way, the corrected set of moments can be automatically produced for a generic moment function $g(\cdot)$ provided by the user.

Model Identification: Jacobian Ψ Theorem 2 requires β_0 to be identified and the Jacobian matrix Ψ to be full rank. Notably, the MERM framework encompasses many possible sources of identification at once, including instrumental variables, repeated measurements, and 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 θ_0.

In the example below, we describe how the parameters θ_0 and γ_0 can be identified when in addition to the error-laden X_i we have a second measurement Q_i. Identification using an instrumental variable is more elaborate and will be discussed in Section 4. Specifically, in Section 4, we will establish nonparametric identification of the nonlinear regression model using a (possibly discrete) instrument. Moreover, we will show that the corrected moments and the corresponding semiparametric MERM estimator naturally arise from the nonparametric characterization of the problem.

Example 2.1 (Identification Using a Second Measurement)**.** Suppose we have a conditional moment restriction

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

For simplicity, in this example, we only condition on X_i^*. It is straightforward to add additional (exogenous) variables to the conditioning set.

If X_i^* were observed we could have constructed 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 $J \geq \dim\left(\theta\right) - 1$. Assume that the model is identified if X_i^* observed, which means that the Jacobian of these moment conditions $H^* \equiv E\left[\nabla_\theta h\left(X_i^*, S_i, \theta_0\right)\right]$ has full rank, i.e., $\text{Rk}\left(H^*\right) = \dim\left(\theta\right)$.

Suppose we observe a second measurement $Q_i = \alpha_1 X_i^* + \varepsilon_{Q,i}$, where α_1 may not be known. Suppose 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 MERM estimator with $K = 2$ based on the following moment function

$$g\left(x, s, q, \theta\right) \equiv \begin{pmatrix} h\left(x, s, \theta\right) \\ u\left(x, s, \theta\right) q \times \left(1, x, \ldots, x^{J-1}\right)' \end{pmatrix}. \tag{15}$$

Here the total number of moments is $m = 2J + 1$. The additional J moments added in equation (15) use Q_i to identify $\gamma_0 = \mathbb{E}[\varepsilon_i^2]/2$.

Appendix E demonstrates that a sufficient condition for Ψ to have full rank is

$$\mathbb{E}\left[u_x^{(1)}\left(X_i^*, S_i, \theta_0\right) \times \left(1, X_i^*, \ldots, \left(X_i^*\right)^{J-1}\right)'\right] \neq 0. \tag{16}$$

Condition (16) 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, consider the nonlinear regression model $\mathbb{E}\left[Y_i|X_i^* = x\right] = \rho\left(x, \theta_0\right)$. Then $u\left(x, y, \theta\right) = \rho\left(x, \theta\right) - y$, and condition (16) 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 $j \in \{0, \ldots, J-1\}$. ■

Assessing model validity The standard J-test 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. Issue *(i)* can be address by taking a larger K. In Section 5 we extend the framework to obtain corrected moments that are valid under weaker assumptions on the EIV, which can help addressing issue *(ii)*.

It is possible to obtain an explicit bound on the higher-order EIV bias of the MERM estimator. Suppose the researcher is interested in estimating a linear combination of parameters $v'\beta_0$ for some non-zero $v \in \mathbb{R}^{\dim(\beta)}$, e.g., the first component of θ_0. Since $\hat{\beta}$ is a standard GMM estimator,

$$\text{Bias}(v'\hat{\beta}) \approx -v'(\Psi'\Xi\Psi)^{-1}\Psi'\Xi\,\mathbb{E}\left[\psi(X_i, S_i, \theta_0, \gamma_{0n})\right]. \tag{17}$$

Here $\mathbb{E}\left[\psi(X_i, S_i, \theta_0, \gamma_{0n})\right] \approx \gamma_{0\overline{K}}\mathbb{E}[g_x^{(\overline{K})}(X_i, S_i, \theta_0)]$ with $\overline{K} = K + 2$ if K is even and ε_i is symmetric, and $\overline{K} = K + 1$ otherwise. Thus,

$$\text{Bias}(v'\hat{\beta}) \approx \gamma_{0\overline{K}} b_v,$$

where b_v can be easily estimated. The unknown $\gamma_{0\overline{K}}$ can be bounded using equation (12) and some (a priori) bounds on the $\overline{K}$-th moment of ε_i. Then the bias of $v'\hat{\beta}$ is approximately bounded between $\underline{\gamma}_{0\overline{K}}\hat{b}_v$ and $\overline{\gamma}_{0\overline{K}}\hat{b}_v$, where $\underline{\gamma}_{0\overline{K}}$ and $\overline{\gamma}_{0\overline{K}}$ are the lower and upper bounds on $\gamma_{0\overline{K}}$.

The bounds on $\text{Bias}(v'\hat{\beta})$ can also be used to assess the appropriateness of Assumption MME and the reliability of the inference based on the normal approximation in equation (14). If the (worst case) bias of $v'\hat{\beta}$ is substantial, relative to its standard error, the researcher may want to account for this and consider choosing a higher K. Appendix C provides a further discussion of this issue and additional details on the calculation of $\text{Bias}(v'\hat{\beta})$.

Finally, one can test the strength of identification of the model parameters or even conduct identification-robust inference (e.g., Stock and Wright, 2000; Kleibergen, 2005; Guggenberger and Smith, 2005; Guggenberger, Ramalho, and Smith, 2012; Andrews and Mikusheva, 2016; Andrews, 2016; Andrews and Guggenberger, 2019).

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.

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

$(Z_i, V_i, \varepsilon_i)' \sim N\left((0,0,0)', \text{Diag}(1, 1/4, 1/4)\right)$ and $n = 1000$. The conditional expectation function ρ, the true value of the parameter of interest θ_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 θ_2 and θ_4) or better (for θ_1 and θ_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				
	θ_1	θ_2	θ_3	θ_4	θ_1	θ_2	θ_3	θ_4	θ_1	θ_2	θ_3	θ_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			
	θ_1	θ_2	θ_3	θ_1	θ_2	θ_3	θ_1	θ_2	θ_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		
	θ_1	θ_2	θ_1	θ_2	θ_1	θ_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 ϵ_{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^*.

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 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 ϵ_{ij} from the

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.

standard type-I extreme value distribution. The true value of θ_0 is set to be the MLE estimate based on the original dataset. More details about this numerical experiment are given in Appendix J.

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 τ.

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 Identification of the Nonlinear Regression Model

It is important to understand under what conditions the parameters of interest are identified in our settings. This section considers identification and estimation of the nonlinear regression function $\rho(x) \equiv \mathbb{E}[Y_i|X_i^* = x]$. To address the problem of EIV, the researcher has an instrument Z_i, which can be discrete or continuous. In this section, we do not impose any functional form assumptions on the true regression function $\rho(\cdot)$, i.e., the analysis is nonparametric.

Consider the model

$$Y_i = \rho(X_i^*) + U_i, \qquad X_i = X_i^* + \varepsilon_i.$$

The joint distribution of observables (Y_i, X_i, Z_i) satisfies the following assumptions.

Assumption 4.1. $\mathbb{E}[Y_i|X_i^*, Z_i] = \mathbb{E}[Y_i|X_i^*]$.

Assumption 4.2. *Functions $\rho(\cdot)$ and $f_{X^*|Z}(\cdot|z)$ have at least $p \geq 3$ bounded derivatives, and $\mathbb{E}[|\varepsilon_i|^p] \leq C\tau^p$ for some constant C.*

Assumption 4.1 is a standard exclusion restriction on the instrument Z_i. Assumption 4.2 collects some weak regularity conditions. We will study the properties of this model using the approximation $\tau \to 0$. The identification analysis views the unknown function ρ and the joint distribution of (Y_i, X_i^*, Z_i) as fixed. Thus, the joint distribution of the observables (Y_i, X_i, Z_i) is implicitly indexed by τ, but varies with τ only due to the changes in the distribution of ε_i.

The naive population regression $q(x) \equiv \mathbb{E}[Y_i|X_i = x]$ suffers from the EIV bias

$$\mathbb{E}[Y_i|X_i = x] = \rho(x) + O\left(\tau^2\right).$$

Similar to the semiparametric case of Section 2, identifying σ^2 or τ^2 allows dealing with this bias.[11]

[11] Note that σ^2 can be approximated by 0. However, to be meaningful, identification of σ^2 needs to characterize it up to an error of an order smaller than σ^2 itself.

Let us define

$$q(x,z) \equiv \mathbb{E}[Y_i | X_i = x, Z_i = z], \quad s_{X|Z}(x|z) \equiv \frac{f'_{X|Z}(x|z)}{f_{X|Z}(x|z)}, \quad s_{X^*|Z}(x|z) \equiv \frac{f'_{X^*|Z}(x|z)}{f_{X^*|Z}(x|z)}. \tag{19}$$

Let $\mathcal{S}_{X^*}(z) \equiv \{x : f_{X^*|Z}(x|z) > 0\}$, so that $s_{X^*|Z}(x|z)$ is well-defined for all $x \in \mathcal{S}_{X^*}(z)$.

The following functions can be identified directly from the joint distribution of the observed (Y_i, X_i, Z_i):

$$\widetilde{\sigma}^2(x) \equiv \frac{q(x,z_1) - q(x,z_2)}{q'(x)\left[s_{X|Z}(x|z_1) - s_{X|Z}(x|z_2)\right]}, \tag{20}$$

$$\widetilde{\rho}(x,z,v) \equiv q(x,z) - \widetilde{v}\left[q'(x)\, s_{X|Z}(x|z) + \tfrac{1}{2}q''(x)\right]. \tag{21}$$

Theorem 3. *Suppose Assumptions CME, 4.1, 4.2 hold and either (i) $p = 3$ or (ii) $\mathbb{E}[\varepsilon_i^3] = 0$ and $p = 4$. Suppose there exist z_1, z_2, and a point $x \in \mathcal{S}_{X^*}(z_1) \cap \mathcal{S}_{X^*}(z_2)$, such that $\rho'(x)\left[s_{X^*|Z}(x|z_1) - s_{X^*|Z}(x|z_2)\right] \neq 0$. Then, as $\tau \to 0$,*

$$\widetilde{\sigma}^2(x) \equiv \sigma^2 + O(\tau^p). \tag{22}$$

Moreover, for any $\widetilde{v} = \sigma^2 + O(\tau^p)$ (including $\widetilde{v} = \widetilde{\sigma}^2(x)$), any z, and any $x \in S_{X^*}(z)$,

$$\widetilde{\rho}(x,z,\widetilde{v}) \equiv \rho(x) + O(\tau^p) \quad \text{for all } x \in \mathcal{S}_{X^*}(z). \tag{23}$$

Equation (22) shows that we can identify σ^2 up to an error of a smaller order $O(\sigma^p) = O(\tau^p)$. As a consequence, $\rho(\cdot)$ is identified up to the same order $O(\tau^p)$. Below we will consider the implications of this theorem for estimation.

Theorem 3 requires the rank condition $\rho'(x)\left[s_{X^*|Z}(x|z_1) - s_{X^*|Z}(x|z_2)\right] \neq 0$ to hold for some x. The key here is the instrument relevance condition that requires $s_{X^*|Z}(x|z)$ to vary with z. The proof of the theorem shows that

$$q(x,z) = \rho(x) + \sigma^2\rho'(x)\, s_{X^*|Z}(x|z) + \tfrac{1}{2}\sigma^2\rho''(x) + O(\tau^p). \tag{24}$$

This equation is used to identify σ^2 by varying z, since only the second term on the right-hand side depends on z. Note that for $q(x,z)$ to vary with z we need the additional rank condition $\rho'(x) \neq 0$. Requiring that there exists a point x with

$\rho'(x) \neq 0$ is a weak condition, since $\rho'(x) = 0$ for all x means that $\rho(x)$ is constant, in which case EIV do not bias the naive regression estimator and $q(x) = \rho(x)$.[12,13] The rank condition $\rho'(x)\left[s_{X^*|Z}(x|z_1) - s_{X^*|Z}(x|z_2)\right] \neq 0$ in Theorem 3 can be replaced by the condition that $q'(x)\left[s_{X|Z}(x|z_1) - s_{X|Z}(x|z_2)\right]$ is bounded away from zero, which is stated in terms of the observables. Likewise, set $\mathcal{S}_{X^*}(z)$ above can be replaced with $\mathcal{S}_X(z) \equiv \{x : f_{X|Z}(x|z) > C_f\}$ for any positive constant C_f.

The relevance condition imposed on the instrumental variable is weak. The equality $s_{X^*|Z}(x|z_1) = s_{X^*|Z}(x|z_2)$ can only hold for all x, z_1, and z_2 if X^* and Z are independent. Finally, consider the following example:

Example. Suppose X_i^* follows a Gaussian random coefficient model:

$$X_i^* = \Pi_{1i} + \Pi_{2i} Z_i, \qquad \text{where} \quad \begin{pmatrix} \Pi_{1i} \\ \Pi_{2i} \end{pmatrix} \sim N\left(\begin{pmatrix} \pi_1 \\ \pi_2 \end{pmatrix}, \begin{pmatrix} \omega_{\pi 1}^2 & 0 \\ 0 & \omega_{\pi 2}^2 \end{pmatrix} \right),$$

and $Z_i \in \{0, 1\}$ is a binary instrument. In many applications the instruments are likely to have heterogeneous effects on the covariate, e.g., Angrist, Graddy, and Imbens (2000) and Heckman and Vytlacil (1998), which corresponds to $\omega_{\pi 2}^2 > 0$. Then, $X_i^*|Z_i = z \sim N(\mu(z), \omega^2(z))$ with $\mu(z) = \pi_1 + \pi_2 z$ and $\omega^2(z) = \omega_{\pi 1}^2 + \omega_{\pi 2}^2 z^2$. Thus, the instrument is relevant unless $\pi_2 = \omega_{\pi 2} = 0$. Notably, the instrument is relevant even if $\pi_2 = 0$ (so $\text{corr}(X_i^*, Z_i) = 0$), as long as Z_i has a heterogeneous effect on X_i^*. ■

Importantly, in contrast to many approaches to EIV in nonlinear models, an instrument itself can be mismeasured, and a variable that is caused by X_i^* can serve as an instrument. In particular, Z_i can be a nonclassical second measurement of X_i^*.

Remark 9. Note that σ^2 is overidentified, which can be used to test the model assumptions.

To clarify the implications of Theorem 3 it is useful to consider nonparametric estimation of the regression function $\rho(x)$. The approach of Theorem 3 is constructive, and using equations (19)-(21) we can construct a nonparametric measurement error

[12] See Evdokimov and Zeleneev (2018) for more details on the role of $\rho'(x)$ in the literature on measurement errors.

[13] Equation (24) extends the calculation of the conditional expectation in equation (4.6) in Chesher (1991) by introducing the instrumental variables and obtaining more precise bounds on the approximation error.

robust analog estimator $\widehat{\rho}^{\text{MER}}(x)$. We compare such $\widehat{\rho}^{\text{MER}}(x)$ with the naive nonparametric regression estimator $\widehat{\rho}^{\text{Naive}}(x)$ of $\mathbb{E}[Y|X_i = x]$, which ignores the presence of EIV in the data. Both estimators can be implemented using standard nonparametric estimation methods (e.g., kernel or sieve estimators). In the following discussion assume that the tuning parameters are chosen optimally for each of the estimators. For brevity we focus only on the case (ii) in Theorem 3.

Proposition 4. *Suppose the conditions of Theorem 3 hold,* $\mathbb{E}[\varepsilon_i^3] = 0$, *functions* $\rho(\cdot)$ *and* $f_{X^*|Z}(\cdot|z)$ *have* $m \geq 4$ *bounded derivatives, and* Z_i *is discrete. Suppose* $\tau_n = O\left(n^{-\frac{1}{4}\frac{m}{2m+1}}\right)$, *then*

$$\begin{aligned}\widehat{\rho}^{MER}(x) - \rho(x) &= O_p\left(n^{-\frac{m}{2m+1}}\right), \\ \widehat{\rho}^{Naive}(x) - \rho(x) &= O_p\left(n^{-\frac{1}{2}\frac{m}{2m+1}}\right).\end{aligned}$$

The proposition provides a nonparametric analog of the semiparametric results in Section 2. In particular, $\widehat{\rho}^{\text{MER}}$ generally has a faster rate of convergence than $\widehat{\rho}^{\text{Naive}}$. For example, if $m = 4$ the rates of convergence of $\widehat{\rho}^{\text{MER}}(x)$ and $\widehat{\rho}^{\text{Naive}}(x)$ are $O_p\left(n^{-4/9}\right)$ and $O_p\left(n^{-2/9}\right)$, respectively. Note that the rate of convergence of $\widehat{\rho}^{\text{MER}}(x)$ in Proposition 4 is optimal and cannot be improved: even if had data on (Y_i, X_i^*) without EIV, under the smoothness assumptions of the proposition, function $\rho(x)$ cannot be estimated nonparametrically at a rate faster than $O_p\left(n^{-\frac{m}{2m+1}}\right)$, see Stone (1980). Note also that for the models with large m, the rates of convergence in the Proposition approach those in Remark 1.

Remark 10. Our identification analysis is fully nonparametric (we make no parametric form assumptions on the regression function or the distributions, including the distributions of the measurement errors). The identification result of Theorem 3 is also global: it shows that joint distribution of the data allows distinguishing true function ρ from all other possible (suitably smooth) functions up to an error of order $O(\tau^p)$. However, our analysis is confined to the approximations $\tau \to 0$, and in this is different from the nonparametric identification approaches with "large" measurement errors.

Connection to the MERM Estimator for the Nonlinear Regression Model It turns out that the above nonparametric identification analysis is directly connected

to the MERM estimator for the semiparametric nonlinear regression model. Suppose the family of regression functions is parameterized as $\rho(x,\theta)$, where θ is a finite dimensional parameter of interest, and $\rho(x) \equiv \rho(x,\theta_0)$. The standard original moment function for the nonlinear regression is

$$g(x,y,z,\theta) \equiv (\rho(x,\theta) - y)\,\varphi(x,z) \tag{25}$$

for a vector of smooth functions $\varphi(x,z)$. The original moment condition $\mathbb{E}[g(X_i^*, Y_i, Z_i, \theta_0)] = 0$ would have been satisfied at the true parameter value θ_0 had we observed X_i^*.

Suppose Assumption MME holds with $K = 2$ and we use the corrected moment conditions from Section 2 to deal with the EIV. For the original moment conditions (25), evaluated at the true parameter values θ_0 and $\gamma_{02} = \sigma^2/2$, the corrected moment conditions satisfy

$$\begin{aligned} & \mathbb{E}[\psi(X_i, Y_i, Z_i, \theta_0, \gamma_{02})] \\ = \;& \mathbb{E}\left[(\rho(X_i) - Y_i)\,\varphi(X_i, Z_i) - \frac{\sigma^2}{2}\left(2\rho_x(X_i)\,\varphi_x(X_i, Z_i) + \rho_x^{(2)}(X_i)\,\varphi(X_i, Z_i)\right)\right] + O(\tau^4) \\ = \;& o\left(n^{-1/2}\right), \end{aligned} \tag{26}$$

where the first equality follows by equation (7) and the fact that $\mathbb{E}\left[(\rho(X_i) - Y_i)\,\frac{\sigma^2}{2}\varphi_x^{(2)}(X_i, Z_i)\right] = O(\tau^4)$, and the second equality follows by Lemma 1.

Alternatively, we can consider developing a semiparametric estimator based directly on Theorem 3 with $\rho(x) \equiv \rho(x,\theta_0)$. We can restate equation (24) as

$$q(x,z) = \rho(x) + \sigma^2 \rho_x(x)\, s_{X|Z}(x|z) + \tfrac{1}{2}\sigma^2 \rho_x^{(2)}(x) + o\left(n^{-1/2}\right),$$

since $\sigma^2\left(s_{X|Z}(x|z) - s_{X^*|Z}(x|z)\right) = O(\tau^4) = o\left(n^{-1/2}\right)$. By definition, $q(x,z) = \mathbb{E}[Y_i | X_i = x, Z_i = z]$, so the above equation can be viewed as the conditional moment restriction

$$\mathbb{E}\left[\nu\left(X_i, Y_i, Z_i, \theta_0, \sigma^2\right) | X_i = x, Z_i = z\right] = o\left(n^{-1/2}\right),$$

where

$$\nu\left(x, y, z, \theta_0, \sigma^2\right) \equiv \rho(x) + \sigma^2 \rho_x(x)\, s_{X|Z}(x|z) + \frac{\sigma^2}{2}\rho_x^{(2)}(x) - y.$$

This conditional moment restriction cannot be directly used for estimation of θ_0, because it depends on function $s_{X|Z}(x|z)$. One could consider estimating function $s_{X|Z}$ nonparametrically. However, it turns out that this is unnecessary.

Let us transform the conditional moment ν into an unconditional moment using the vector of smooth functions $\varphi(x,z)$. Using generalized Stein's lemma (or integration by parts) we obtain

$$\begin{aligned}
& \mathbb{E}\left[\nu\left(X_i, Z_i, Y_i, \theta_0, \sigma^2\right)\varphi\left(X_i, Z_i\right)\right] \\
= \; & \mathbb{E}\left[\left(\rho\left(X_i\right) - Y_i\right)\varphi\left(X_i, Z_i\right) + \sigma^2\left\{\rho_x\left(X_i\right)\varphi\left(X_i, Z_i\right)s_{X|Z}\left(X_i|Z_i\right) + \frac{1}{2}\rho_x^{(2)}\left(X_i\right)\varphi\left(X_i, Z_i\right)\right\}\right] \\
= \; & \mathbb{E}\left[\left(\rho\left(X_i\right) - Y_i\right)\varphi\left(X_i, Z_i\right) + \sigma^2\left\{-\nabla_x\left\{\rho_x\left(X_i\right)\varphi\left(X_i, Z_i\right)\right\} + \frac{1}{2}\rho_x^{(2)}\left(X_i\right)\varphi\left(X_i, Z_i\right)\right\}\right] \\
= \; & \mathbb{E}\left[\left(\rho\left(X_i\right) - Y_i\right)\varphi\left(X_i, Z_i\right) + \sigma^2\left\{-\rho_x\left(X_i\right)\varphi_x\left(X_i, Z_i\right) - \frac{1}{2}\rho_x^{(2)}\left(X_i\right)\varphi\left(X_i, Z_i\right)\right\}\right] \\
= \; & \mathbb{E}\left[\left(\rho\left(X_i\right) - Y_i\right)\varphi\left(X_i, Z_i\right) - \frac{\sigma^2}{2}\left\{2\rho_x\left(X_i\right)\varphi_x\left(X_i, Z_i\right) + \rho_x^{(2)}\left(X_i\right)\varphi\left(X_i, Z_i\right)\right\}\right].
\end{aligned}$$

Notice that the unknown function $s_{X|Z}(x|z)$ has disappeared, and the last line of the equation matches the corrected moment condition (26).

Thus, the MERM estimator for nonlinear regression can be viewed as a semiparametric implementation of the nonparametric characterization in equation (24) that avoids any nonparametric estimation of the nuisance parameters.

5 Extensions

5.1 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^*, ε_i, and X_i are $d \times 1$ vectors. Let $\tau_n \equiv \max_{j \leq d} \sigma_{\varepsilon_j} / \sigma_{X_j^*}$, where σ_{ε_j} and $\sigma_{X_j^*}$ denote the standard deviations of the j-th components of ε_i and X_i^*, so $\mathbb{E}\left[\left|\varepsilon_{ij}\right|^k\right] = O(\tau_n^k)$ 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 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, γ is a collection of all γ_κ with $\kappa \in \mathcal{K}_k$ and $k \in \{2, \ldots, K\}$.

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 γ_0 are determined by the moments of ε_i. Specifically, let $\mu_\kappa \equiv \mathbb{E}\left[\varepsilon_{i1}^{\kappa_1} \ldots \varepsilon_{id}^{\kappa_d}\right]$, then

$$\gamma_{0\kappa} = \frac{\mu_\kappa}{\kappa!}, \quad \text{for } \kappa \in \{\mathcal{K}_2, \mathcal{K}_3\}, \tag{27}$$

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 (27). For $\kappa \in \mathcal{K}_4$, $\gamma_{0\kappa}$ is

given by

κ	$\gamma_{0\kappa}$
(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 ε_{i1} and ε_{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. ■

5.2 Weakly Classical and Non-Classical Measurement Errors

In some applications, assuming that the measurement error is classical, i.e., ε_i is independent from (X_i^*, S_i), may be too restrictive. In this section, we demonstrate how the MERM framework can be used to address this issue.

First, in Section 5.2.1, we consider estimation of models with weakly classical measurement errors satisfying $\mathbb{E}\left[\varepsilon_i | X_i^*, S_i\right] = 0$. This restriction on ε_i is much weaker than the classical measurement error assumption. For example, the variance of the measurement error $\mathbb{E}\left[\varepsilon_i^2 | X_i^*, S_i\right]$ (and other moments) could depend on X_i^* and S_i. Building on this, in Section 5.2.2, we consider estimation with general non-classical measurement errors that may be correlated with the mismeasured variable.

5.2.1 Weakly Classical Measurement Errors

In this section, we relax Assumption CME and assume that the measurement error is weakly classical. Specifically, assume that $\mathbb{E}\left[\varepsilon_i | X_i^*, S_i\right] = 0$ or, equivalently, $\mathbb{E}\left[X_i | X_i^*, S_i\right] = X_i^*$.

Suppose that $\mathbb{E}\left[|\varepsilon_i|^k | X_i^*, S_i\right] = O(\tau_n^k)$ for $k \in \{2, \ldots, K\}$. Conditional on X_i^* and S_i, the moment function can be expanded as

$$\mathbb{E}[g(X_i, S_i, \theta) | X_i^*, S_i] = g(X_i^*, S_i, \theta) + \sum_{k=2}^{K} \frac{\mathbb{E}\left[\varepsilon_i^k | X_i^*, S_i\right]}{k!} g_x^{(k)}(X_i^*, S_i, \theta) + O(\tau_n^{K+1}). \quad (28)$$

Suppose $\mathbb{E}\left[\varepsilon_i^k | X_i^*, S_i\right] = v_k(X_i^*, S_i, \omega_0)$, $k \in \{2, \ldots, K\}$ where $\omega_0 \in \mathbb{R}^{\dim(\omega)}$ are unknown parameters. Typically, functions v_k are assumed to depend only on some of the components of (X_i^*, S_i). For example, in the nonlinear regression (2), it may be natural for v_k to depend on X_i^* and/or W_i, but not on Y_i. Equation (28) then implies that

$$\mathbb{E}\left[g(X_i^*, S_i, \theta)\right] = \mathbb{E}\left[g(X_i, S_i, \theta) - \sum_{k=2}^{K} \frac{v_k(X_i^*, S_i, \omega_0)}{k!} g_x^{(k)}(X_i^*, S_i, \theta)\right] + O(\tau_n^{K+1}). \quad (29)$$

Equation (29) is an analog of equation (10), which motivates correcting the moment conditions by estimating and subtracting the bias terms $\frac{v_k(X_i^*, S_i, \omega_0)}{k!} g_x^{(k)}(X_i^*, S_i, \theta)$. The corrected moment function takes the form

$$\psi(X_i, S_i, \theta, \omega) = g(X_i, S_i, \theta) - \sum_{k=2}^{K} f_k(X_i, S_i, \theta, \omega), \quad (30)$$

where

$$f_2(x, s, \theta, \omega) = \frac{v_2(x, s, \omega)}{2} g_x^{(2)}(x, s, \theta), \quad f_3(x, s, \theta, \omega) = \frac{v_3(x, s, \omega)}{6} g_x^{(3)}(x, s, \theta),$$
$$f_k(x, s, \theta, \omega) = \frac{v_k(x, s, \omega)}{k!} g_x^{(k)}(x, s, \theta) - \sum_{\ell=2}^{k-2} \frac{v_{k-\ell}(x, s, \omega)}{(k-\ell)!} \frac{\partial^{k-\ell}}{\partial x^{k-\ell}} f_\ell(x, s, \theta, \omega), \text{ for } k \geq 4.$$

Notice that for $K \leq 3$ the corrected moment functions are similar to the classical EIV case except for $v_2(x, s, \omega)$ and $v_3(x, s, \omega)$ taking the places of $\mathbb{E}\left[\varepsilon_i^2\right]$ and $\mathbb{E}\left[\varepsilon_i^3\right]$ respectively. For $K \geq 4$, we need to bias correct the bias correction terms such as $f_2(x, s, \theta, \omega)$ accounting for $v_2(X_i, S_i, \theta, \omega)$ being evaluated at X_i instead of X_i^*. Also, when the conditional moments of ε_i do not depend on X_i^*, the corrected moment function matches the classical EIV case with $v_k(s, \omega)$ replacing $\mathbb{E}\left[\varepsilon_i^k\right]$.

Under the regularity conditions

$$\mathbb{E}\left[\psi(X_i, S_i, \theta_0, \omega_0)\right] = \mathbb{E}\left[g(X_i^*, S_i, \theta_0)\right] + O(\tau_n^{K+1}) = o(n^{-1/2}),$$

provided that $O(\tau_n^{K+1}) = o(n^{-1/2})$.

Thus, the model can be estimated using GMM estimator with the corrected moment function (30), where parameters θ_0 and ω_0 are estimated together, i.e., the estimator in equation (13) with $\beta \equiv (\theta', \omega')'$. The following example provides a con-

venient parameterization of functions v.

Example 5.1 (Exponential Specification, $K = 4$)**.**
Suppose $\varepsilon_i = \exp(\omega_{01} X_i^*)\zeta_i$, where ζ_i is mean zero and independent of (X_i^*, S_i). In this case, $\mathbb{E}\left[\varepsilon_i^k | X_i^*, S\right] = \mathbb{E}\left[\zeta_i^k\right] \exp(k\omega_{01} X_i^*)$. Then, we can take $v_k(x, s, \omega) = \omega_k \exp(k\omega_1 x)$, $k \in \{2, \ldots, 4\}$, with $\omega_0 = (\omega_{01}, \mathbb{E}\left[\zeta_i^2\right], \mathbb{E}\left[\zeta_i^3\right], \mathbb{E}\left[\zeta_i^4\right])'$. Functions f_k in equation (30) are

$$f_2(x, s, \theta, \omega) = \frac{\omega_2}{2} e^{2\omega_1 x} g_x^{(2)}(x, s, \theta), \quad f_3(x, s, \theta, \omega) = \frac{\omega_3}{6} e^{3\omega_1 x} g_x^{(3)}(x, s, \theta),$$
$$f_4(x, s, \theta, \omega) = e^{4\omega_1 x} \left(\tfrac{\omega_4 - 6\omega_2^2}{24} g_x^{(4)}(x, s, \theta) - \omega_1 \omega_2^2 g_x^{(3)}(x, s, \theta) - \omega_1^2 \omega_2^2 g_x^{(2)}(x, s, \theta) \right). \quad \blacksquare$$

Identifiability of parameters ω_0 depends on the specific model and moment conditions g. For the nonlinear regression model with weakly classical EIV Evdokimov and Zeleneev (2022) establish nonparametric identification using instrumental variables.

A numerical experiment illustrating the finite sample properties of the MERM estimator with weakly-classical measurement errors is provided in Section F.

5.2.2 Non-Classical Measurement Errors

Since X_i^* is not observed, in the absence of some a priori information about its distribution, some assumptions such as $\mathbb{E}\left[X_i | X_i^*, S_i\right] = X_i^*$ appear to be necessary, and are routinely imposed in the EIV literature. For example, see Hu and Schennach (2008), Hu and Sasaki (2015), Freyberger (2021), and Schennach (2021) for recent discussions of such assumptions. In the absence of such assumptions one cannot determine even the scale of the true unobserved covariate.

To be specific, suppose the true covariate is denoted by $\mathcal{X}_i^*$ and is mismeasured according to the following model

$$X_i = \alpha_0 + \alpha_1 \mathcal{X}_i^* + \mathcal{E}_i, \qquad \mathbb{E}\left[\mathcal{E}_i | \mathcal{X}_i^*, S_i\right] = 0, \tag{31}$$

so $\mathbb{E}\left[X_i | \mathcal{X}_i^*\right] = \alpha_0 + \alpha_1 \mathcal{X}_i^*$ for some unknown α_0 and $\alpha_1 > 0$. Note that

$$\mathbb{E}\left[X_i - \mathcal{X}_i^* | \mathcal{X}_i^*\right] \neq 0,$$

and the measurement error is non-classical. For concreteness, consider the nonlinear

regression model:

$$\mathbb{E}\left[Y_i|\mathcal{X}_i^*, W_i\right] = \rho\left(\theta_{01} + \theta_{0\mathcal{X}}\mathcal{X}_i^* + \theta_{0W}' W_i\right) \tag{32}$$

for some known function ρ.

Since $\mathcal{X}_i^*$ is unobservable, we can write an observationally equivalent model in terms of another unobservable X_i^* *defined* as

$$X_i^* \equiv \alpha_0 + \alpha_1 \mathcal{X}_i^*, \tag{33}$$

$$\mathbb{E}\left[Y_i|X_i^*, W_i\right] = \rho\left(\tilde{\theta}_{01} + \tilde{\theta}_{0X} X_i^* + \theta_{0W}' W_i\right), \tag{34}$$

where $\tilde{\theta}_{0X} \equiv \theta_{0\mathcal{X}}/\alpha_1$, $\quad \tilde{\theta}_{01} \equiv \theta_{01} - \tilde{\theta}_{0X}\alpha_0$. Thus, parameters $(\theta_{01}, \theta_{0\mathcal{X}})$ and (α_0, α_1) are not separately identified without additional information about the location and scale of $\mathcal{X}_i^*$. Note that neither instrumental variables nor multiple measurements of the form (31) provide such information.

Equation (31) and the definition (33) of X_i^* imply that condition $\mathbb{E}\left[X_i|\, X_i^*, S_i\right] = X_i^*$ holds for the observationally equivalent model (34). This observationally equivalent model can be identified and estimated using the approach of Section 5.2.1. This then allows identification of many parameters of interest in the original model (32) because they do not depend on the (unidentified) scale of $\mathcal{X}_i^*$.

First, parameters θ_{0W} are the same in both models. Moreover, we have $\left(\mathcal{X}_i^* - \mu_{\mathcal{X}^*}\right)/\sigma_{\mathcal{X}^*} = \left(X_i^* - \mu_{X^*}\right)/\sigma_{X^*}$, and hence the following average structural function

$$m(t, w) \equiv \mathbb{E}\left[Y_i|\mathcal{X}_i^* = \mu_{\mathcal{X}^*} + t\sigma_{\mathcal{X}^*}, W_i = w\right]$$

is identified, since $m(t, w) = \mathbb{E}\left[Y_i|X_i^* = \mu_{X^*} + t\sigma_{X^*}, W_i = w\right]$ is identified. As a result, we can identify marginal effects such as $\frac{\partial}{\partial w}\mathbb{E}\left[Y_i|\mathcal{X}_i^* = \mu_{\mathcal{X}^*}, W_i = w\right] = \frac{\partial}{\partial w} m(0, w)$, or the effect of increasing $\mathcal{X}_i^*$ from $\mu_{\mathcal{X}^*}$ to $\mu_{\mathcal{X}^*} + \lambda\sigma_{\mathcal{X}^*}$, i.e., $m(\lambda, w) - m(0, w)$. The corresponding marginal effects averaged with respect to the distribution of $(\mathcal{X}_i^*, W_i)$ are also identified and can be estimated following Remark 3. Thus, in this model, the MERM estimator from Section 5.2.1 allows estimation of many economically meaningful marginal effects, even though the true measurement model is given by equation (31), and the measurement error $X_i - \mathcal{X}_i^*$ is non-classical.[14]

At the same time, identification of the true values of θ_{01} and $\theta_{0\mathcal{X}}$ or some marginal

[14] In the constructed observationally equivalent model (34) we have $\varepsilon_i = \mathcal{E}_i$, and the MERM approach is justified when $\sigma_{\mathcal{E}}^2/\sigma_X^2$ is moderate. The true measurement error $X_i - \mathcal{X}_i^*$ does not need to be small.

effects such as $\frac{\partial}{\partial \varkappa}\mathbb{E}\left[Y_i|\mathcal{X}_i^* = \varkappa, W_i = w\right]$ requires identifying α_0 and α_1.

Using the summary statistics for $\mathcal{X}^*$ In some applications the researchers may have information about the *marginal* distribution of $\mathcal{X}^*$ such as its mean $\mu_{\mathcal{X}^*}$ and variance $\sigma^2_{\mathcal{X}^*}$, e.g., from an administrative dataset.[15,16] Then, the MERM approach can be used to estimate α_0 and α_1 together with parameters θ_0. For any original moment condition $\mathbb{E}\left[g\left(\mathcal{X}_i^*, S_i, \theta_0\right)\right] = 0$ consider the augmented moment condition $\mathbb{E}\left[g_A\left(\mathcal{X}_i^*, S_i, \theta_0\right)\right] = 0$, where

$$g_A\left(\mathcal{X}_i^*, S_i, \theta_0\right) \equiv \left(g\left(\mathcal{X}_i^*, S_i, \theta_0\right)', \mathcal{X}_i^* - \mu_{\mathcal{X}^*}, \left(\mathcal{X}_i^* - \mu_{\mathcal{X}^*}\right)^2 - \sigma^2_{\mathcal{X}^*}\right)'.$$

Define $X_i^* \equiv \alpha_0 + \alpha_1 \mathcal{X}_i^*$, $\tilde{\theta}_0 \equiv (\theta_0', \alpha_0, \alpha_1)'$, and $\tilde{g}_A(X_i^*, S_i, \tilde{\theta}_0) \equiv g_A\left(\left(X_i^* - \alpha_0\right)/\alpha_1, S_i, \theta_0\right)$. Then one can take $\mathbb{E}[\tilde{g}_A(X_i^*, S_i, \tilde{\theta}_0)] = 0$ as the new "original" moment condition (1). Since the condition $\mathbb{E}\left[X_i|X_i^*, S_i\right] = X_i^*$ holds, $\tilde{\theta}_0$ can be estimated using the MERM approach and the corrected moments (30) applied to $\tilde{g}_A$ in place of g.

Evdokimov and Zeleneev (2022) show how an identification argument based on the assumption of weakly classical measurement errors can be used for nonparametric identification in the settings with non-classical measurement errors. The linearity of equation (31) simplifies the estimation procedure, but the underlying approach is valid nonparametrically. Using a parsimonious model, this section provides a practical way of dealing with non-classical measurement errors in empirical applications, avoiding any nonparametric estimation.

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

[15] It is important that $\mathcal{X}_j^*$ and (X_i, S_i) are drawn from the same population.

[16] Knowing $\sigma^2_{\mathcal{X}^*}$ does not identify even the unconditional variance σ^2_ε, since $\sigma^2_X = \alpha_1^2\sigma^2_{\mathcal{X}^*} + \sigma^2_\varepsilon$.

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.

Angrist, J. D., K. Graddy, and G. W. Imbens (2000): "The Interpretation of Instrumental Variables Estimators in Simultaneous Equations Models with an Application to the Demand for Fish," *Review of Economic Studies*, 67, 499–527.

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

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

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.

Freyberger, J. (2021): "Normalizations and misspecification in skill formation models," 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.

Heckman, J. and E. Vytlacil (1998): "Instrumental Variables Methods for the Correlated Random Coefficient Model: Estimating the Average Rate of Return to Schooling When the Return is Correlated with Schooling," *The Journal of Human Resources*, 33, 974.

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

Hu, Y. and Y. Sasaki (2015): "Closed-form estimation of nonparametric models with non-classical measurement errors," *Journal of Econometrics*, 185, 392–408.

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

Imbens, G. and W. Newey (2009): "Identification and Estimation of Triangular Simultaneous Equations Models without Additivity," *Econometrica*, 77, 1481–1512.

Imbens, G. W. and J. D. Angrist (1994): "Identification and Estimation of Local Average Treatment Effects," *Econometrica*, 62, 467–475.

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.

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.

——— (2021): "Measurement systems," *Journal of Economic Literature*, forthcoming.

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.

Stone, C. J. (1980): "Optimal Rates of Convergence for Nonparametric Estimators," *Annals of Statistics*, 8, 1348–1360.

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 ε_i), it ensures that this remainder is $o(n^{-1/2})$, which is crucial for establishing validity of the corrected moment function ψ (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* $k \in \{0, \ldots, K\}$*, 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* $k \in \{0, \ldots, K\}$*, 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* Ξ *is a symmetric positive definite matrix;*

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

Assumption A.4. (Global and local identification)

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

(ii) $\Psi^{*\prime}\Xi\Psi^*$ *is invertible, where*

$$\begin{aligned}\Psi^* &\equiv \mathbb{E}\left[\Psi(X_i^*, S_i, \theta_0, 0)\right] \\ &= \mathbb{E}\left[G(X_i^*, S_i, \theta_0), -g_x^{(2)}(X_i^*, S_i, \theta_0), \ldots, -g_x^{(K)}(X_i^*, S_i, \theta_0)\right].\end{aligned}$$

Assumption A.2-A.4 is a collection of 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, the asymptotic properties of $\hat{\theta}$ are controlled by G^* (and Ψ^*), the Jacobian associated with the

correctly measured variables. Assumptions A.4 (i) and (ii) are the standard GMM global and local identification conditions applied to the "limiting" moment function $\psi(X_i^*, S_i, \theta, \gamma)$.

B Proof of Lemma 1

To stress that in our asymptotic approximation the variance and the higher moments of ε_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 $\gamma = \gamma_{0n}$, $\mathbb{E}\left[\psi(X_i, S_i, \theta_0, \gamma_{0n})\right] = o(n^{-1/2})$. Note that the first two terms in (B.3) are mean zero, i.e. we have

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

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 γ_{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 γ_{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.

Online Apendix

C Evaluating the remaining bias of the MERM estimator

In this section, we provide additional details on the derivation of the (higher-order) bias of the MERM estimator of $v'\hat{\beta}$ and discuss potential implications for inference.

The derivation of the expression for $\text{Bias}(v'\hat{\beta})$ provided in equation (17) is based on the standard expansion of the GMM first order conditions around β_0. Following the lines of the proof of Lemma 1, we can consider a higher order expansion of $\psi(X_i, S_i, \theta_0, \gamma_{0n})$ to the order $o(\tau^{\overline{K}})$, which implies

$$\mathbb{E}\left[\psi(X_i, S_i, \theta_0, \gamma_{0n})\right] \approx \gamma_{0\overline{K}} \mathbb{E}\left[g_x^{(\overline{K})}(X_i, S_i, \theta_0)\right],$$

where $\overline{K} = K+2$ if K is even and ε_i is symmetric, and $\overline{K} = K+1$ otherwise. Hence,

$$\text{Bias}(v'\hat{\beta}) \approx -v'(\Psi'\Xi\Psi)^{-1}\Psi'\Xi\, \mathbb{E}\left[g_x^{(\overline{K})}(X_i, S_i, \theta_0)\right] \gamma_{0\overline{K}}. \tag{C.1}$$

Notice that, except for $\gamma_{0\overline{K}}$, all the remaining elements on the right hand side of (C.1) are either known or can be consistently estimated. The value of $\gamma_{0\overline{K}}$ is given by the recursive relationship (12) and is determined by the unknown $\mathbb{E}[\varepsilon_i^{\overline{K}}]$ (and by the lower order moments of ε_i, which can be recovered from $\hat{\gamma}$ with a sufficient precision). Hence, the range of the possible values of $\gamma_{0\overline{K}}$ and, hence, the range of $\text{Bias}(v'\hat{\beta})$ can be determined based on an a priori range of possible values of $\mathbb{E}[\varepsilon_i^{\overline{K}}]$ (or, equivalently, $\mathbb{E}[(\varepsilon_i/\sigma_\varepsilon)^{\overline{K}}]$) specified by the researcher.

Example (Evaluating the bias with $K = 2$)**.** Suppose the researcher uses the MERM estimator with $K = 2$, and believes that $\mathbb{E}[\varepsilon_i^3] = 0$. Then, the bound on the bias depends on the $\mathbb{E}[\varepsilon_i^4]$. The variance σ_ε^2 is estimated by the MERM estimator, hence it is convenient to bound the kurtosis of ε_i as $\mathbb{E}[\varepsilon_i^4]/\sigma_\varepsilon^4 \in [1, \overline{\kappa}]$ for some $\overline{\kappa}$. For example, one could take $\overline{\kappa} = 10$ as a conservative bound for the kurtosis.[17] Since $\hat{\gamma}$ estimates $\gamma_0 = \gamma_{02} = \sigma_\varepsilon^2/2$, the range of possible values of γ_{04} can be approximately bounded

[17] For Gaussian ε_i the kurtosis κ is 3. Student's $t(\nu)$ distribution has $\kappa < 10$ for all $\nu \geq 5$ (the kurtosis and the 4th moment do not exist for $\nu \leq 4$).

using

$$\gamma_{04} = \frac{\mathbb{E}\left[\varepsilon_i^4\right] - 6\sigma_\varepsilon^4}{24} \in \left[-\frac{5}{24}\sigma_\varepsilon^4,\ \frac{\overline{\kappa}-6}{24}\sigma_\varepsilon^4\right] = \left[-\frac{5}{6}\gamma_{02}^2,\ \frac{\overline{\kappa}-6}{6}\gamma_{02}^2\right].$$

Hence, the range of possible values of $\text{Bias}(v'\hat{\beta})$ can be approximated as

$$\text{Bias}(v'\hat{\beta}) \in \begin{cases} \left[-5\overline{b}\hat{\gamma}^2/6,\ \overline{b}\hat{\gamma}^2(\overline{\kappa}-6)/6\right], & \text{if } \overline{b} \geq 0, \\ \left[\overline{b}\hat{\gamma}^2(\overline{\kappa}-6)/6,\ -5\overline{b}\hat{\gamma}^2/6\right], & \text{otherwise,} \end{cases}$$

where

$$\overline{b} \equiv -v'(\hat{\Psi}'\hat{\Xi}\hat{\Psi})^{-1}\hat{\Psi}'\hat{\Xi}\ \left(\frac{1}{n}\sum_{i=1}^{n} g_x^{(4)}(X_i, S_i, \hat{\theta})\right).$$

Notice that if $\overline{\kappa} \leq 6$, we have $\gamma_{04} \leq 0$, and the sign of $\text{Bias}(v'\hat{\beta})$ becomes known.

If (the absolute value of the worst case) $\text{Bias}(v'\hat{\beta})$ is sufficiently small relative to its standard error $\sqrt{v'\hat{\Sigma}v/n}$, the estimator $v'\hat{\beta}$ is approximately unbiased and the inference on $v'\beta_0$ can be based on equation (14). Otherwise, the researcher could consider using a higher K and/or possibly employing inference tools that take into account the remaining bias (e.g., Armstrong and Kolesár, 2021).

D MERM derivation when σ_ε is not small

Note that τ can be small without σ_ε 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 σ_{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 σ_ε is small in absolute magnitude. In Section 2 we used Taylor expansions in ε_i around $\varepsilon_i = 0$, with the remainder of order $\mathbb{E}\left[\left|\varepsilon_i\right|^{K+1}\right]$. When $\sigma_\varepsilon > 1$, term $O\left(\mathbb{E}\left[\left|\varepsilon_i\right|^{K+1}\right]\right)$ in equation (9) cannot be viewed as a negligible remainder, because $\mathbb{E}\left[\left|\varepsilon_i\right|^{K+1}\right] > 1$ and, moreover, terms $\mathbb{E}\left[\left|\varepsilon_i\right|^{k}\right]$ increase rather than decrease with k.

In Section 2, to simplify the exposition, we have assumed that X^* is scaled so that σ_{X^*} is of order one. This in particular ensures that $\mathbb{E}\left[\left|\varepsilon_i\right|^{k}\right]$ 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 τ^k, which necessarily decrease with k when $\tau < 1$.

Remember the model of Section 2:

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

Let ξ_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 (D.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

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

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 (D.2) is numerically identical to ψ 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

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

E Details for Example 2.1

In this section, we demonstrate that the Jacobian Ψ associated with the moment conditions (15) is guaranteed to have a full rank (for $K = 2$) when the sufficient condition (16) 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

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

From Lemma G.2 (see Section G.1 below), we have $\Psi \to \Psi^* \equiv \mathbb{E}[\nabla_\beta \psi(X_i^*, S_i, Q_i, \theta_0, 0)]$. Hence, it is sufficient to demonstrate that Ψ^* has full rank.

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 α_1 and subtracting them from the moments $2, \ldots, J+1$, we obtain

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

Here $\mathrm{Rk}(H^*) = \dim(\theta)$, because this is the rank identification condition for θ_0 in the model without EIV. Thus, for Ψ^* 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 $\mathbb{E}[u(X_i^*, S_i, \theta_0) | X_i^*] = 0$, 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 (16).

F Numerical Illustration with Weakly Classical Measurement Errors

Consider the following simple Logit model

$$Y_i = \mathbb{1}\{\theta_{01} X_i^* + \theta_{02} W_i + \theta_{03} - \eta_i \geq 0\},$$
$$X_i^* = Z_i + V_i, \quad X_i = X_i^* + \varepsilon_i, \quad \varepsilon_i = \exp(\omega_{01} X_i^*)\zeta_i, \quad W_i = \rho X_i^*/\sigma_{X^*} + \sqrt{1-\rho^2}\nu_i,$$

where $\eta_i \sim$ *Logistic* and $(Z_i, V_i, \zeta_i, \nu_i)' \sim N\left((0,0,0,0)', \text{Diag}(\sigma_Z^2, \sigma_V^2, \sigma_\zeta^2, \sigma_\nu^2)\right)$ are independent from each other.

We fix $(\theta_{01}, \theta_{02}, \theta_{03}, \omega_{01}, \rho, \sigma_Z^2, \sigma_V^2, \sigma_\nu^2) = (1, 0, 2, 0.3, 0.7, 1, 1, 1)$ and $n = 2000$. By adjusting σ_ζ^2 accordingly, we consider $\tau = \sigma_\varepsilon/\sigma_{X^*} \in \{1/4, 1/2, 3/4\}$, where, as before, σ_ε denotes the (unconditional) standard deviation of ε_i.

We report results for the MERM estimator based on the corrected moment function (30) with $K = 2$ and $K = 4$. The original moment function is

$$g(x, w, y, z, \theta) = (y - \Lambda(\theta_1 x + \theta_2 w + \theta_3))\, \varphi(x, z, w),$$

where we use $\varphi(x, z, w) = (1, x, z, x^2, z^2, x^3, z^3, w)'$ for $K = 2$ and $\varphi(x, z, w) = (1, x, z, x^2, xz, z^2, x^3, x^2 z, xz^2, z^3, w)'$ for $K = 4$. The corrected moment function is as in Example 5.1, where we set $\omega_3 = 0$ (using $\mathbb{E}\left[\zeta_i^3\right] = 0$).

Table 6 reports the simulation results. Both of the correction schemes effectively remove the EIV bias for $\tau \in \{1/4, 1/2\}$. However, employing the higher order correction scheme with $K = 4$ is needed to achieve accurate size control for larger values of τ ($\tau = 3/4$).

Table 6: Simulation results for the logit model with weakly classical measurement errors

	MLE				$K=2$				$K=4$			
	bias	std	rmse	size	bias	std	rmse	size	bias	std	rmse	size
						$\tau=1/4$						
θ_1	-0.062	0.073	0.096	14.20	0.041	0.113	0.120	3.50	0.065	0.120	0.137	5.82
θ_2	0.063	0.090	0.110	10.86	-0.014	0.119	0.120	2.70	-0.027	0.125	0.128	3.96
∂_x	-0.006	0.007	0.010	16.10	0.001	0.010	0.010	2.92	0.002	0.010	0.011	4.08
∂_w	0.007	0.009	0.012	10.74	-0.001	0.012	0.012	3.06	-0.003	0.012	0.013	4.36
						$\tau=1/2$						
θ_1	-0.224	0.068	0.234	89.84	0.019	0.134	0.136	3.50	0.043	0.141	0.147	5.78
θ_2	0.219	0.083	0.235	75.32	0.019	0.138	0.139	3.82	-0.003	0.141	0.141	4.10
∂_x	-0.022	0.006	0.023	91.66	-0.003	0.012	0.012	4.14	-0.001	0.012	0.012	4.52
∂_w	0.024	0.009	0.025	74.86	0.002	0.014	0.014	3.82	-0.000	0.014	0.014	4.52
						$\tau=3/4$						
θ_1	-0.419	0.058	0.423	100.00	-0.116	0.129	0.173	12.18	0.031	0.154	0.157	5.96
θ_2	0.396	0.076	0.403	99.98	0.172	0.131	0.217	31.82	0.033	0.142	0.146	5.86
∂_x	-0.040	0.006	0.041	100.00	-0.017	0.011	0.020	40.72	-0.005	0.012	0.013	7.64
∂_w	0.044	0.009	0.045	99.96	0.017	0.013	0.022	28.20	0.003	0.014	0.014	5.74

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 θ_{01} and θ_{02}, and the marginal effects associated with X^* and W evaluated at the population means. The true values of the considered parameters are $(\theta_{01}, \theta_{02}, \partial_x, \partial_w) = (1, 0, 0.105, 0)$. The results are based on 5,000 replications.

G Proof of Theorem 2

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

All vectors are columns. For some generic parameter vector α 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 α_0, we often write $a_i \equiv a(\alpha_0)$, $\overline{a} \equiv \overline{a}(\alpha_0)$, $a \equiv a(\alpha_0)$, $a_i^* \equiv a(\alpha_0)$, $\overline{a}^* \equiv \overline{a}^*(\alpha_0)$, $a^* \equiv a^*(\alpha_0)$.

G.1 Auxiliary lemmas

Lemma G.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* Θ *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 G.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),

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

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 Θ. 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 $k\in\{1,\ldots,K\}$ and for $G_x^{(k)}$ for $k\in\{0,\ldots,K\}$ (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 G.2. *Suppose that the hypotheses of Lemma G.1 are satisfied. Then,* $g_x^{(k)}\to g_x^{(k)*}$ *and* $G_x^{(k)}\to G_x^{(k)*}$ *for* $k\in\{0,\ldots,K\}$. *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* $k\in\{0,\ldots,K\}$.

Proof of Lemma G.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

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

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 G.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 G.3. *Suppose that the hypotheses of Lemma G.1 are satisfied. Then, under additional Assumptions A.3 (iv) and A.4 (i), we have $\hat{\theta} \xrightarrow{p} \theta_0$, $\hat{\gamma} \xrightarrow{p} 0$ and $\hat{\gamma} \xrightarrow{p} \gamma_{0n}$.*

Proof of Lemma G.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,

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

where the equality follows from Lemma G.1 (i) and boundedness of γ (Assumption A.2 (i)). Moreover, Lemma G.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 (G.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 Γ is compact and $\gamma_{0n} \in \Gamma$ (Assumption A.2 (i)), $0_{K-1} \in \Gamma$. Consequently, Assumptions A.4 (i) and A.3 (iv) together guarantee that $\hat{Q}^*(\beta)$ 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 G.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 G.4. Using expansion (B.3), we obtain

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

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 (G.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.

G.2 Proof of Theorem 2

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

Proof of Theorem 2. Since (i) $\hat{\theta}$ and $\hat{\gamma}$ are consistent for θ_0 and γ_{0n}, respectively (Lemma G.3) and (ii) both θ_0 and $\gamma_{0n} \to 0$ (Assumption MME) are bounded away from the boundaries of Θ and Γ 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

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

where $\tilde{\beta}$ lies between β_{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 G.3), we can invoke the result of Lemma G.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 G.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 A.4 (ii). Hence, (G.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 G.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 G.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 G.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 G.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.

H Proofs of the Results in Section 4

H.1 Proof of Theorem 3

1. First, in parts 1-4 below, we prove the theorem in the case $p = 3$. Then, in part 5 of this proof, we consider the case $p = 4$. The proof in the latter case is identical, except some of the remainder terms are of smaller orders. In parts 1-4 of the proof, it will be convenient to state the resulting bounds that depend on p in the general form using p, but to avoid confusion, until reaching part 5 of the proof the reader should only consider the case $p = 3$.

Let us obtain some preliminary results about $f_{X|Z}(x|z)$ and $f_{\varepsilon|XZ}(\varepsilon|x,z)$. Using Assumption CME, we obtain

$$\begin{aligned} f_{\varepsilon X|Z}(\varepsilon, x|z) &= f_{\varepsilon}(\varepsilon) f_{X^*|Z}(x-\varepsilon|z), \\ f_{\varepsilon|XZ}(\varepsilon|x,z) &= \frac{f_{\varepsilon X|Z}(\varepsilon,x|z)}{f_{X|Z}(x|z)} = \frac{f_{\varepsilon}(\varepsilon) f_{X^*|Z}(x-\varepsilon|z)}{f_{X|Z}(x|z)}, \\ f_{X|Z}(x|z) &= \int f_{\varepsilon}(\varepsilon) f_{X^*|Z}(x-\varepsilon|z)\, d\varepsilon. \end{aligned}$$

Since $f_{X^*|Z}(x|z)$ has 3 bounded derivatives in x, $f_{X|Z}(x|z)$ also has 3 bounded derivatives in x. Moreover,

$$f_{X|Z}(x|z) = \int \left[f_{X^*|Z}(x|z) - \varepsilon f'_{X^*|Z}(x|z) + \varepsilon^2 \tfrac{1}{2} f''_{X^*|Z}(x|z) \right] f_{\varepsilon}(\varepsilon)\, d\varepsilon + R_{f|Z}(x|z), \tag{H.1}$$

where $R_{f|Z} \equiv -(1/6)\,\mathbb{E}\left[\varepsilon_i^3 f'''_{X^*|Z}(x - \widetilde{\varepsilon}(\varepsilon_i)) \,|Z_i = z\right]$, $\widetilde{\varepsilon}(\varepsilon_i)$ is a point between 0 and ε_i, and

$$\left|R_{f|Z}(x|z)\right| \leq \mathbb{E}\left[|\varepsilon_i|^3 \left| f'''_{X^*|Z}(x - \widetilde{\varepsilon}(\varepsilon_i)) \right| \middle| Z_i = z \right] = O\left(\mathbb{E}\left[|\varepsilon_i|^3\right]\right) = O\left(\sigma^3\right).$$

Since $\mathbb{E}[\varepsilon_i] = 0$ and $\mathbb{E}[\varepsilon_i^2] = \sigma^2$, we have

$$\begin{aligned} f_{X|Z}(x|z) &= f_{X^*|Z}(x|z) - \mathbb{E}[\varepsilon_i] f'_{X^*|Z}(x|z) + \mathbb{E}\left[\varepsilon_i^2\right] \tfrac{1}{2} f''_{X^*|Z}(x|z) + O(\sigma^p) \\ &= f_{X^*|Z}(x|z) + \left(\sigma^2/2\right) f''_{X^*|Z}(x|z) + O(\sigma^p). \end{aligned} \tag{H.2}$$

Since $f_{X^*|Z}(x|z) > C$ for some $C > 0$, and $f''_{X^*|Z}(x|z)$ is bounded, $f_{X|Z}(x|z) > C/2$ for small enough σ^2. Thus,

$$f_{\varepsilon|XZ}(\varepsilon|x,z) = \frac{f_\varepsilon(\varepsilon) f_{X^*|Z}(x-\varepsilon|z)}{f_{X^*|Z}(x|z) + (\sigma^2/2) f''_{X^*|Z}(x|z) + O(\sigma^p)}.$$

Similarly,

$$f'_{X|Z}(x|z) = \int f_\varepsilon(\varepsilon) f'_{X^*|Z}(x-\varepsilon|z)\, d\varepsilon = f'_{X^*|Z}(x|z) + O\left(\sigma^2\right), \tag{H.3}$$

$$f''_{X|Z}(x|z) = \int f_\varepsilon(\varepsilon) f''_{X^*|Z}(x-\varepsilon|z)\, d\varepsilon = f''_{X^*|Z}(x|z) + O\left(\sigma^{p-2}\right). \tag{H.4}$$

2. Consider any function $a(x)$ that has 3 bounded derivatives. Since $X_i^* = X_i - \varepsilon_i$,

$$\begin{aligned}
& \mathbb{E}\left[a\left(X_i^*\right)|X_i = x, Z_i = z\right] \\
= \; & \mathbb{E}\left[a\left(x-\varepsilon_i\right)|X_i = x, Z_i = z\right] \\
= \; & a(x) - a'(x)\,\mathbb{E}\left[\varepsilon_i|X_i = x, Z_i = z\right] \\
& + \tfrac{1}{2} a''(x)\,\mathbb{E}\left[\varepsilon_i^2|X_i = x, Z_i = z\right] + R_{a|XZ},
\end{aligned} \tag{H.5}$$

where $R_{a|XZ} \equiv -(1/6)\,\mathbb{E}\left[\varepsilon_i^3 a'''\left(x - \widetilde{\varepsilon}\left(\varepsilon_i\right)\right)|X_i = x, Z_i = z\right]$.

We now consider $\mathbb{E}\left[\varepsilon_i^\ell|X_i = x, Z_i = z\right]$ for $\ell \in \{1,2\}$:

$$\begin{aligned}
\mathbb{E}\left[\varepsilon_i|X_i = x, Z_i = z\right] &= \int \varepsilon f_{\varepsilon|XZ}(\varepsilon|x,z)\, d\varepsilon = \int \varepsilon \frac{f_\varepsilon(\varepsilon) f_{X^*|Z}(x-\varepsilon|z)}{f_{X|Z}(x|z)} d\varepsilon \\
&= \frac{\int \varepsilon \left\{ f_{X^*|Z}(x|z) - \varepsilon f'_{X^*|Z}(x|z) + \varepsilon^2 \frac{1}{2} f''_{X^*|Z}\left(x - \widetilde{\varepsilon}(\varepsilon)|z\right)\right\} f_\varepsilon(\varepsilon)\, d\varepsilon}{f_{X^*|Z}(x|z) + O(\sigma^2)} \\
&= 0 - \sigma^2 \frac{f'_{X^*|Z}(x|z)}{f_{X^*|Z}(x|z) + O(\sigma^2)} + O(\sigma^p) = -\sigma^2 s_{X^*|Z}(x|z) + O(\sigma^p),
\end{aligned}$$

$$\begin{aligned}
\mathbb{E}\left[\varepsilon_i^2|X_i = x, Z_i = z\right] &= \frac{\int \varepsilon^2 \left\{ f_{X^*|Z}(x|z) - \varepsilon f'_{X^*|Z}\left(x - \widetilde{\varepsilon}(\varepsilon)|z\right)\right\} f_\varepsilon(\varepsilon)\, d\varepsilon}{f_{X^*|Z}(x|z) + O(\sigma^2)} \\
&= \sigma^2 + O(\sigma^p).
\end{aligned}$$

Next,

$$\mathbb{E}\left[\left|\varepsilon_i^p\right| | X_i = x, Z_i = z\right] = \frac{\int \left|\varepsilon^p\right| f_{X^*|Z}\left(x - \widetilde{\varepsilon}\left(\varepsilon\right) | z\right) f_\varepsilon\left(\varepsilon\right) d\varepsilon}{f_{X^*|Z}\left(x|z\right) + O\left(\sigma^2\right)} = O\left(\sigma^p\right)$$

and hence $\left|R_{a|XZ}\right| \leq \mathbb{E}\left[\left|\varepsilon_i^3\right| \left|a'''\left(x - \widetilde{\varepsilon}\left(\varepsilon_i\right)\right)\right| | X_i = x, Z_i = z\right] = O\left(\sigma^3\right)$.

Combining these with equation (H.5), we obtain

$$\mathbb{E}\left[a\left(X_i^*\right) | X_i = x, Z_i = z\right] = a\left(x\right) + \sigma^2 a'\left(x\right) s_{X^*|Z}\left(x|z\right) + \left(\sigma^2/2\right) a''\left(x\right) + O\left(\sigma^p\right).$$

3. Next, consider $\nabla_x^\ell \mathbb{E}\left[a\left(X_i^*\right) | X_i = x\right]$ for $\ell \in \{1, 2\}$. We have

$$\mathbb{E}\left[a\left(X_i^*\right) | X_i = x\right] = \frac{1}{f_X\left(x\right)} \int a\left(x - \varepsilon\right) f_{X^*}\left(x - \varepsilon\right) f_\varepsilon\left(\varepsilon\right) d\varepsilon.$$

Let $\varphi\left(x\right)$ and $\eta\left(x\right)$ be any functions, possibly changing with σ, with 3 bounded derivatives. Suppose $\eta\left(x\right) \geq C$ for some $C > 0$ for all small enough σ, and let $\zeta\left(x\right) \equiv \frac{1}{\eta(x)} \int \varphi\left(x - \varepsilon\right) f_\varepsilon\left(\varepsilon\right) d\varepsilon$. Then,

$$\begin{aligned}
\zeta'\left(x\right) &= \frac{1}{\eta\left(x\right)} \int \varphi'\left(x - \varepsilon\right) f_\varepsilon\left(\varepsilon\right) d\varepsilon - \frac{\eta'\left(x\right)}{\left(\eta\left(x\right)\right)^2} \int \varphi\left(x - \varepsilon\right) f_\varepsilon\left(\varepsilon\right) d\varepsilon \\
&= \frac{\varphi'\left(x\right) + O\left(\sigma^2\right)}{\eta\left(x\right)} - \frac{\eta'\left(x\right)\left(\varphi\left(x\right) + O\left(\sigma^2\right)\right)}{\left(\eta\left(x\right)\right)^2} = \frac{\varphi'\left(x\right)}{\eta\left(x\right)} - \frac{\eta'\left(x\right)\varphi\left(x\right)}{\left(\eta\left(x\right)\right)^2} + O\left(\sigma^2\right).
\end{aligned}$$

Taking $\varphi\left(t\right) \equiv a\left(t\right) f_{X^*}\left(t\right)$ and $\eta\left(x\right) \equiv f_X\left(x\right)$, and using equations (H.2) and (H.3), we obtain

$$\nabla_x \mathbb{E}\left[a\left(X_i^*\right) | X_i = x\right] = a'\left(x\right) + O\left(\sigma^2\right).$$

Similarly,

$$\zeta''\left(x\right) = \frac{\varphi''\left(x\right)}{\eta\left(x\right)} - 2\frac{\eta'\left(x\right)\varphi'\left(x\right)}{\left(\eta\left(x\right)\right)^2} - \left(\frac{\eta''\left(x\right)}{\left(\eta\left(x\right)\right)^2} - \frac{2\left(\eta'\left(x\right)\right)^2}{\left(\eta\left(x\right)\right)^3}\right)\varphi\left(x\right) + O\left(\sigma^{p-2}\right),$$

and using identical substitutions, and equations (H.2), (H.3), and (H.4), we obtain

$$\nabla_x^2 \mathbb{E}\left[a\left(X_i^*\right) | X_i = x\right] = a''\left(x\right) + O\left(\sigma^{p-2}\right).$$

4. Consider

$$q(x,z) = \mathbb{E}\left[\rho\left(X_i^*\right) + U_i | X_i = x, Z_i = z\right] = \mathbb{E}\left[\rho\left(X_i^*\right) | X_i = x, Z_i = z\right].$$

Part 2 of the proof shows that

$$q(x,z) = \rho(x) + \sigma^2 \rho'(x)\, s_{X^*|Z}(x|z) + \tfrac{1}{2}\sigma^2 \rho''(x) + O\left(\sigma^p\right). \tag{H.6}$$

Therefore,

$$q(x,z_1) - q(x,z_2) = \sigma^2 \rho'(x)\left[s_{X^*|Z}(x|z_1) - s_{X^*|Z}(x|z_2)\right] + O\left(\sigma^p\right). \tag{H.7}$$

Part 1 implies that $s_{X^*|Z}(x|z) = s_{X|Z}(x|z) + O(\sigma^2)$. Let $q(x) \equiv \mathbb{E}[Y_i|X_i = x]$. Applying part 3 with $a(x) = q(x)$ we obtain $q'(x) = \rho'(x) + O(\sigma^2)$. Substituting these into equation (H.7), we obtain

$$q(x,z_1) - q(x,z_2) = \sigma^2 q'(x)\left[s_{X|Z}(x|z_1) - s_{X|Z}(x|z_2)\right] + O\left(\sigma^p\right), \tag{H.8}$$

and hence, for any x with $q'(x)\left[s_{X|Z}(x|z_1) - s_{X|Z}(x|z_2)\right] \neq 0$,

$$\widetilde{\sigma}^2(x) = \sigma^2 + O\left(\sigma^p\right), \quad \text{where } \widetilde{\sigma}^2(x) \equiv \frac{q(x,z_1) - q(x,z_2)}{q'(x)\left[s_{X|Z}(x|z_1) - s_{X|Z}(x|z_2)\right]},$$

which identifies σ^2 up to $O(\sigma^p)$.

Next, from part 3 we also have $q''(x) = \rho''(x) + O(\sigma^{p-2})$. Thus, equation (H.6) implies that

$$q(x,z) = \rho(x) + \sigma^2 q'(x)\, s_{X|Z}(x,z) + \tfrac{1}{2}\sigma^2 q''(x) + O\left(\sigma^p\right),$$

and hence we obtain

$$\widetilde{\rho}(x,z_1) = \rho(x) + O\left(\sigma^p\right), \quad \text{where } \widetilde{\rho}(x,z) \equiv q(x,z) - \widetilde{\sigma}^2(x)\left[q'(x)\, s_{X|Z}(x|z) + \tfrac{1}{2}q''(x)\right],$$

which identifies $\rho(x)$ up to $O(\sigma^p)$. This completes the proof for the case $p = 3$.

Note that $\widetilde{\rho}(x,z_1) - \widetilde{\rho}(x,z_2) = q(x,z_1) - q(x,z_2) - \widetilde{\sigma}^2(x)\left[q'(x)\left(s_{X|Z}(x|z_1) - s_{X|Z}(x|z_2)\right)\right] = 0$, i.e., $\widetilde{\rho}(x,z_1) = \widetilde{\rho}(x,z_2)$.

5. When $p = 4$ and $\mathbb{E}[\varepsilon_i^3] = 0$, the above Taylor expansions can be extended to the next order, providing the corresponding improvements in the remainder terms.

For example, in part 1 of the proof, expansion (H.1) becomes

$$f_{X|Z}(x|z) = \int \{ f_{X^*|Z}(x|z) - \varepsilon f'_{X^*|Z}(x|z) + \varepsilon^2 \tfrac{1}{2} f''_{X^*|Z}(x|z) - \varepsilon^3 \tfrac{1}{6} f'''_{X^*|Z}(x|z) \} f_\varepsilon(\varepsilon)\, d\varepsilon + R_{f|Z}(x|z),$$

where $R_{f|Z}(x|z) \equiv (1/24)\,\mathbb{E}\left[\varepsilon_i^4 f''''_{X^*|Z}(x - \widetilde{\varepsilon}(\varepsilon_i))\,|Z_i = z\right]$, so $\left|R_{f|Z}(x|z)\right| = O(\sigma^4)$. Combining the expansion above with $\mathbb{E}[\varepsilon_i^3] = 0$, we verify that (H.2) also holds with $p = 4$.

In addition to the calculations in part 2 of the proof, we also use

$$\begin{aligned} \mathbb{E}\left[\varepsilon_i^3 | X_i = x, Z_i = z\right] &= \frac{\int \varepsilon^3 \left\{ f_{X^*|Z}(x|z) - \varepsilon f'_{X^*|Z}(x - \widetilde{\varepsilon}(\varepsilon)\,|z) \right\} f_\varepsilon(\varepsilon)\, d\varepsilon}{f_{X^*|Z}(x|z) + O(\sigma^2)} \\ &= \mathbb{E}\left[\varepsilon_i^3\right] + O\left(\sigma^4\right) = O\left(\sigma^4\right). \end{aligned}$$

In parts 2 and 3 of the proof, we also require functions $a(x)$, $\varphi(x)$, and $\eta(x)$ to have 4 bounded derivatives.

Then, the previous steps of the proof and the conclusions of the theorem hold with $p = 4$. Q.E.D.

H.2 Proof of Proposition 4

The proof of Theorem 3 shows that $q(x,z)$ and $f_{X|Z}(x|z)$ have m bounded derivatives. Construct $\widehat{\rho}^{\text{MER}}(x)$ using equations (20)-(21) nonparametrically estimating $q(x,z)$, $q(x)$, $f_{X|Z}(x|z)$, and their derivatives, e.g., using standard kernel or sieve estimators. If the tuning parameters are chosen optimally, under the usual regularity conditions, the rates of convergence of these estimators are $\widehat{q}(x,z) - q(x,z) = O_p\left(n^{-\frac{m}{2m+1}}\right)$, $\widehat{q}(x) - q(x) = O_p\left(n^{-\frac{m}{2m+1}}\right)$, $\widehat{q}'(x) - q'(x) = O_p\left(n^{-\frac{m-1}{2m+1}}\right)$, $\widehat{q}''(x) - q''(x) = O_p\left(n^{-\frac{m-2}{2m+1}}\right)$, and $\widehat{s}_{X|Z}(x|z) - s_{X|Z}(x|z) = O_p\left(n^{-\frac{m-1}{2m+1}}\right)$ for $x \in S_{X^*}(z)$, where $\widehat{s}_{X|Z}(x|z) \equiv \widehat{f}'_{X|Z}(x|z) \big/ \widehat{f}_{X|Z}(x|z)$. Note also that by equation (H.6), $q(x,z_1) - q(x,z_2) = O(\tau_n^2)$.

Then, since for the analog estimator $\widehat{\sigma}^2(x)$ of $\widetilde{\sigma}^2(x)$ we have

$$\begin{aligned}\widehat{\sigma}^2(x) &= \widetilde{\sigma}^2(x) + O_p\left(n^{-\frac{m}{2m+1}} + \tau_n^2 n^{-\frac{m-1}{2m+1}}\right) = \widetilde{\sigma}^2(x) + O_p\left(n^{-\frac{m}{2m+1}}\right) \\ &= \sigma^2 + O_p\left(\tau_n^4 + n^{-\frac{m}{2m+1}}\right) = \sigma^2 + O_p\left(n^{-\frac{m}{2m+1}}\right), \end{aligned} \tag{H.9}$$

where the first equality follows from equation (20), using $\widehat{a}/\widehat{b} - a/b = (\widehat{a}-a)/\widehat{b} + a(1/\widehat{b} - 1/b)$, equation (H.8), and the rates of convergence listed above, the second equality holds because $\frac{m}{2m+1} \leq \frac{1}{2}\frac{m}{2m+1} + \frac{m-1}{2m+1}$ for $m \geq 2$, and the third equality holds by equation (22) in Theorem 3.

Next, consider the analog estimator $\widehat{\rho}^{\text{MER}}(x)$ of $\rho(x)$ based on equation (21),

$$\begin{aligned}\widehat{\rho}^{\text{MER}}(x) &= \widehat{q}(x,z) - \widehat{\sigma}^2(x)\left[\widehat{q}'(x)\widehat{s}(x,z) + \tfrac{1}{2}\widehat{q}''(x)\right] \\ &= \widetilde{\rho}(x) + O_p\left(n^{-\frac{m}{2m+1}} + \tau_n^2\left(n^{-\frac{m-1}{2m+1}} + n^{-\frac{m-2}{2m+1}}\right)\right) \\ &= \rho(x) + O_p\left(n^{-\frac{m}{2m+1}} + \tau_n^2 n^{-\frac{m-2}{2m+1}} + \tau_n^4\right) \\ &= \rho(x) + O_p\left(n^{-\frac{m}{2m+1}}\right), \end{aligned}$$

where the first equality is the definition of the analog estimator, the second equality follows from the rates of convergence listed above and equation (H.9), the third equality holds by equation (23) in Theorem 3, and the fourth equality holds because $\frac{m}{2m+1} \leq \frac{1}{2}\frac{m}{(2m+1)} + \frac{m-2}{2m+1}$ for $m \geq 4$.

The naive estimator is

$$\widehat{\rho}^{\text{Naive}}(x) = \widehat{q}(x) = \rho(x) + O_p\left(\tau_n^2 + n^{-\frac{m}{2m+1}}\right) = \rho(x) + O_p\left(n^{-\frac{1}{2}\frac{m}{2m+1}}\right),$$

where the second equality follows from equation (H.6). Q.E.D

I Some Implementation Details

Numerical Optimization

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

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 θ_0 and γ_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 θ_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 θ and γ. 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 γ. This implies that γ 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 θ 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 Σ

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 Ψ and $\Omega_{\psi\psi}$.

J 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 θ_0 to be the MLE estimates using the considered dataset, which are reported in the table bellow.

	θ_1	θ_2	θ_3	θ_4	θ_5	θ_6	θ_7	θ_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 ζ_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 $\varphi_j(x,z,w) = (1, x, z, x^2, xz, z^2, x^3, x^2 z, xz^2, z^3, r, (r_j - r_0)')'$ 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 θ_0

In Table 7 below, we also report the estimation and inference results for the vector of parameters θ_0 underlying the reported results about elasticities.

Table 7: 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
						$\tau = 1/4$						
θ_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
θ_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
θ_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
θ_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
θ_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
θ_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
θ_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
θ_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
						$\tau = 1/2$						
θ_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
θ_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
θ_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
θ_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
θ_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
θ_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
θ_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
θ_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
						$\tau = 3/4$						
θ_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
θ_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
θ_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
θ_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
θ_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
θ_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
θ_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
θ_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 θ_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.