Schennach, Susanne M.; Starck, Vincent

Working Paper

Optimally-transported generalized method of moments

cemmap working paper, No. CWP17/22

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

Suggested Citation: Schennach, Susanne M.; Starck, Vincent (2022) : Optimally-transported generalized method of moments, cemmap working paper, No. CWP17/22, Centre for Microdata Methods and Practice (cemmap), London, https://doi.org/10.47004/wp.cem.2022.1722

This Version is available at:
https://hdl.handle.net/10419/272829

cemmap
centre for microdata methods and practice

Optimally-transported generalized method of moments

Susanne Schennach
Vincent Starck

The Institute for Fiscal Studies
Department of Economics, UCL

cemmap working paper CWP17/22

Economic and Social Research Council

Optimally-Transported Generalized Method of Moments

Susanne Schennach*†and Vincent Starck‡
Brown University

October 3, 2022

Abstract

We propose a novel optimal transport-based version of the Generalized Method of Moment (GMM). Instead of handling overidentified models by reweighting the data until all moment conditions are satisfied (as in Generalized Empirical Likelihood methods), this method proceeds by introducing measurement error of the least mean square magnitude necessary to simultaneously satisfy all moment conditions. This approach, based on the notion of optimal transport, aims to address the problem of assigning a logical interpretation to GMM results even when overidentification tests reject the null, a situation that cannot always be avoided in applications. Our approach thus introduces a practical alternative to standard GMM estimation to circumvent concerns regarding overidentification test rejections.

1 Introduction

The Generalized Method of Moment (GMM) (Hansen (1982)) has long been the workhorse of statistical modeling in economics and the social sciences. Its key distinguishing feature, relative to the basic method of moments, is the presence of overidentifying restrictions that enable the model's validity to be tested (Newey and McFadden (1994)). With this ability to test comes the obvious practical question of what one should do if an overidentified GMM model fails overidentification tests, a situation that is not uncommon (as noted in Hall and Inoue (2003), Hansen (2001), Masten and Poirier (2021), Conley, Hansen and Rossi (2012), Andrews and Kwon (2019)), even for perfectly reasonable, economically grounded, models.

A popular approach has been to find the "pseudo-true" value of the model parameter (Sawa (1978), White (1982)) that minimizes the *distance* or *discrepancy* between the data and the moment constraints implied by the model. This approach has gained further support since the introduction of Generalized Empirical Likelihoods (GEL) and Minimum Discrepancy estimators (Owen (1988), Qin and Lawless (1994), Newey and Smith (2004)), all of

*Support from NSF grants SES-1950969 and SES-2150003 is gratefully acknowledged. The authors would like to thank Heejun Lee, Florian Gunsilius, Alfred Galichon and seminar participants at NYU's Advanced mathematical modeling in economics seminar for useful comments.

†smschenn@brown.edu

‡vincent_starck@brown.edu

which provide more readily interpretable pseudo-true values (Imbens (1997), Kitamura and Stutzer (1997), Schennach (2007)).

This approach, however, faces a conceptual limitation: It implicitly attributes the mismatch in the moment conditions solely to an improper weighting of the data, i.e. a biased sampling of the population. While this is a possible explanation, it is not the only reason a valid model would fail overidentification tests, when taken to the data. Another natural possibility is the presence of measurement error (Aguiar and Kashaev (2021), Doraszelski and Jaumandreu (2013), Schennach (2020)). In this work, we develop an alternative to GMM that ensures, by construction, that overidentifying restrictions are satisfied by allowing for possible measurement error in the variables instead of sampling bias. In the same spirit as GEL, which does not require the form of the sampling bias to be explicitly specified, the measurement error process does not need to be explicitly specified in our approach, but is instead inferred from the requirement of satisfying the overidentifying constraints imposed by the GMM model.[1] Of course, the accuracy of the resulting estimated parameters will depend on the identifying power of the overidentifying restrictions (with a larger degree of overidentification being typically beneficial).

A fruitful way to accomplish this is to employ concepts from the general area of optimal transport problems (e.g., Galichon (2016), Villani (2009), Carlier, Chernozhukov and Galichon (2016), Chernozhukov et al. (2017), Gunsilius and Schennach (2021)). The idea is to find the parameter value that minimizes the measurement error (for instance, in an L_2 sense) needed to allow all moment conditions to be simultaneously satisfied. Formally, the true iid data z_i is assumed to satisfy $\mathbb{E}\left[g\left(z_i,\theta\right)\right]=0$, where $\mathbb{E}$ is the expectation operator, for some θ and some given d_g-dimensional vector $g\left(z_i,\theta\right)$ of moment functions. However, we instead observe a mismeasured counterpart x_i of the true vector z_i (both taking value in $\mathbb{R}^{d_x}$). We seek to exploit the model's over-identification to gain information regarding the measurement error in x_i. This setup suggests solving the following optimization problem:

$$\min_{\{z_i\}} \frac{1}{2}\hat{\mathbb{E}}\left[\left\|z-x\right\|^2\right] \tag{1}$$

subject to:

$$\hat{\mathbb{E}}\left[g\left(z,\theta\right)\right]=0, \tag{2}$$

where $\|\cdot\|$ is the Euclidean norm (potentially weighted) and where $\hat{\mathbb{E}}$ denotes sample averages (i.e. $\hat{\mathbb{E}}[a\left(x\right)]\equiv\frac{1}{n}\sum_{i=1}^{n}a\left(x_i\right)$, where n is sample size). It will be assumed throughout that this optimization problem is feasible, i.e., the constraint (2) holds for at least one choice of the z_j for a given θ.

This optimization problem is then nested into an optimization over θ, which delivers the estimated parameter value $\hat{\theta}$. We call this estimator an *Optimally-Transported* GMM (OTGMM) estimator. The Euclidean norm is chosen here for computational convenience, although one could imagine a whole class of related estimators obtained with different choices of metric.

[1] In the case where the statistical properties of the measurement error are well known a priori, it would be beneficial to impose such constraints using established methods (e.g. Schennach (2004), Schennach (2014)). However, we consider here the alternative setting where little is a priori known regarding the measurement error.

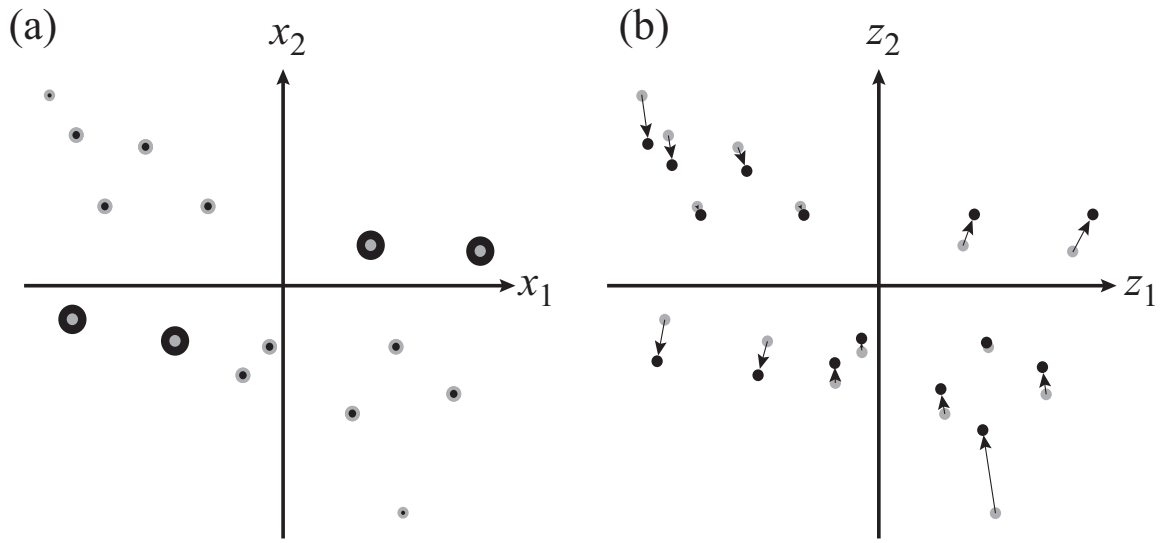

Figure 1: Comparison between the sample points adjustments for (a) Generalized Empirical Likelihood (GEL), where observation weights (shown by point size) are adjusted, and (b) Optimal Transport GMM, where point positions are adjusted. The original sample is shown in gray while the adjusted sample is shown in black. Both panels consider, for illustration, the simple case of an overidentified (parameter-free) model defined by a moment condition imposing no correlation between the two variables.

This type of optimization problem falls within the framework of optimal transport problems (Villani (2009), Galichon (2016), Carlier, Chernozhukov and Galichon (2016), Ekeland, Galichon and Henry (2011)), because it seeks to minimize the cost $\|z - x\|^2$ of "transporting" x to z. In other words, it minimizes the cost of "transporting" the observed distribution of the data onto another distribution that satisfies the moment conditions exactly. Researchers have the ability to indicate a priori information regarding the measurement error magnitude through the norm $\|\cdot\|$, e.g., using a weighted Euclidean norm to indicate the relative expected magnitude of the measurement error along the different dimensions of x. Our focus on Euclidean norms parallels the choice made in most common estimators (e.g. least squares regressions, classical minimum distance and even GMM). We discuss the choice of suitable weighted Euclidean norms $\|\cdot\|$ in Section 5.

Although one can recognize some similarity with GEL, in the sense that one minimizes some concept of distance under a moment constraint, one should realize that the notion of distance used is completely different. As shown in Figure 1, the distance here is measured along the "observation values" axis rather than the "observation weights" axis (as it would be in GEL). This feature arguably makes the method a hybrid between GEL and optimal transport methods, since GEL's goal of satisfying all the moment conditions is achieved through optimal transport instead of through optimal reweighting.

The remainder of the paper is organized as follows. We first formally define and solve the optimization problem defining our estimator, before considering the limit of small measurement error (in the spirit of Chesher (1991)) to gain some intuition. We then derive the resulting estimator's asymptotic properties for the general, large error, case. In particular,

we show asymptotic normality and root n consistency. We then contrast the method with other approaches that have been proposed to deal with model misspecification and propose some extensions. The empirical performance of the method is demonstrated through both Monte Carlo experiments. All proofs can be found in Appendix B.3.

2 The estimator

We now turn to a more formal definition of our estimator, as well as a description of some of the implementation aspects.

2.1 Definition

The Lagrangian associated with the constrained optimization problem defined in Equations (1) and (2) is

$$\frac{1}{2}\hat{\mathbb{E}}\left[\|z-x\|^2\right]-\lambda'\hat{\mathbb{E}}\left[g\left(z,\theta\right)\right]=0.$$

where $\hat{\mathbb{E}}[\ldots]$ denotes a sample average and where λ is a Lagrange multiplier. The first-order conditions of this dual problem with respect to θ, λ and z_j, respectively, are then

$$\begin{aligned}\hat{\mathbb{E}}\left[\partial_2 g'(z,\theta)\right]\lambda &= 0 &(3)\\ \hat{\mathbb{E}}\left[g(z,\theta)\right] &= 0 &(4)\\ (z_j-x_j)-\partial_1 g'(z_j,\theta)\lambda &= 0 \text{ for } j=1,\ldots,n &(5)\end{aligned}$$

where we let ∂_k denote a partial derivative with respect to the k^{th} function argument. We shall use $\partial_{k'}$ to denote a matrix of partial derivatives with respect to a transposed variable (e.g., $\partial_{2'}g(z,\theta)\equiv\partial g(z,\theta)/\partial\theta'$). This formulation of the problem assumes differentiability of $g\left(z,\theta\right)$ to a sufficiently high order, as shall be explicit formalized in our asymptotic analysis.

2.2 Implementation

The nonlinear system (3)-(5) of equations can be solved numerically. To this effect, we propose an iterative procedure to determine the z_j, λ for a given θ. This yields an objective function $\hat{Q}(\theta)$ that can be maximized to estimate θ. Let z_j^t and λ^t denote the approximations obtained after t steps. As shown in Appendix A.1, given tolerances ϵ,ϵ' and a given θ, the objective function $\hat{Q}(\theta)$ can be determined as follows:

Algorithm 2.1 *1. Start the iterations with $z_j^0=x_j$ and $t=0$.*

2. Let

$$\lambda^{t+1}=\left(\hat{\mathbb{E}}\left[H\left(z^t,\theta\right)H'\left(z^t,\theta\right)\right]\right)^{-1}\left(-\hat{\mathbb{E}}\left[g\left(z^t,\theta\right)\right]+\hat{\mathbb{E}}\left[H\left(z^t,\theta\right)\left(z^t-x\right)\right]\right)$$

$$z_j^{t+1}=x_j+H'\left(z_j^t,\theta\right)\lambda^{t+1},$$

where $H\left(z,\theta\right)=\partial_{1'}g\left(z,\theta\right)=\left(\partial_1 g'\left(z,\theta\right)\right)'$.

3. Increment t by 1 and repeat from step 2 until $\left\|z_j^{t+1} - z_j^t\right\| \leq \epsilon$ and $\|\lambda^{t+1} - \lambda^t\| \leq \epsilon'$.

4. The objective function is then:

$$\hat{Q}(\theta) = \frac{1}{2}(\lambda^t)'\hat{\mathbb{E}}\left[H\left(z^t, \theta\right) H'\left(z^t, \theta\right)\right]\lambda^t.$$

This algorithm is obtained by substituting $z_j = x_j + H'(z_j, \theta)\lambda$ obtained from Equation (5) into Equation (4) and expanding the resulting expression to linear order in λ. This linearized expression provides an improved approximation λ^t to the Lagrange multiplier which can, in turn, yield an improved approximation z_j^t. The process is then iterated to convergence. The expression for $\hat{Q}(\theta)$ is obtained by re-expressing $\hat{\mathbb{E}}\left[\|z - x\|^2\right]$ using Equation (5). Formal sufficient conditions for the convergence of this iterative procedure can be found in Appendix A.2.

To gain some intuition regarding the estimator, it is useful to consider the limit of small measurement error when solving Equations (1)-(2). In this limit, the estimator admits a closed form with an intuitive interpretation, as shown by the following result, established in Appendix B.1.

Proposition 2.2 *To the first order in $z_i - x_i$ ($i = 1, \ldots, n$) the estimator is equivalent to minimizing a GMM-like objective function with respect to θ with a non-standard weighting matrix:*

$$\hat{\theta} = \arg\min_{\theta} \hat{\mathbb{E}}\left[g'(x, \theta)\right]\left(\hat{\mathbb{E}}\left[H(x, \theta) H'(x, \theta)\right]\right)^{-1}\hat{\mathbb{E}}\left[g(x, \theta)\right]. \tag{6}$$

From this expression, it is clear that the estimator seeks to downweigh the moments that are the most sensitive to errors in x, as measured by $H(x, \theta) \equiv \partial_{1'} g(x, \theta)$. This accomplishes the desired goal of minimizing the effect of the measurement error, in a context where the properties of the process generating the measurement error are completely unknown.

Although this weighting matrix appears suboptimal (for a correctly specified GMM estimator), one should realize that the method is designed to address misspecification issues, in which case the notion of optimality is not clearly defined, since different estimators may have different pseudo-true values. It is also entirely expected that a model that assumes the absence of measurement error (GMM) would have a smaller variance than an estimator that allows for measurement error.

2.3 Constrained estimator

In some applications, it is useful to be able to constrain the measurement error, for instance to enforce the known fact that some variables are measured without error. The appropriate optimization problem then amounts to minimizing $\hat{\mathbb{E}}\left[\|z - x\|^2\right]$ subject to

$$\hat{\mathbb{E}}\left[g(z, \theta)\right] = 0 \tag{7}$$

$$C(z_i - x_i) = 0 \text{ for } i = 1, \ldots, n \tag{8}$$

for some known rectangular full row rank matrix C that selects the dimensions of the measurement error vector $x_i - z_i$ that should be constrained to be zero. Note that measurement

error constraint is imposed for each observation, not in an average sense. The Lagrangian for this problem is

$$\frac{1}{2}\hat{\mathbb{E}}\left\|z-x\right\|^{2}-\lambda'\hat{\mathbb{E}}\left[g\left(z,\theta\right)\right]-\sum_{i=1}^{n}\gamma_{i}'C\left(z_{i}-x_{i}\right) \tag{9}$$

where λ and γ_i are Lagrange multipliers. As shown in Appendix B.2, the first order conditions (3) and (4) are unchanged, while Equation (5) becomes:

$$\left(z_{j}-x_{j}\right)-PH'\left(x,\theta\right)\lambda=0 \tag{10}$$

where $P=\left(I-C'\left(CC'\right)^{-1}C\right)$ and $H'\left(x,\theta\right)\equiv\partial_{1}g'\left(z_{j},\theta\right)$. Thanks to linearity, the Lagrange multipliers γ_i can be explicitly solved for and the dimensionality of the problem is not increased. The only effect of the constraints is to "project out", through the matrix P, the dimensions where there is no measurement error.

The iterative Algorithm 2.1 can easily be adapted by replacing every instance of $H'\left(z^{t},\theta\right)$ by $PH'\left(z^{t},\theta\right)$. Similarly the linearized estimator of Equation (6) becomes:

$$\frac{1}{2}\hat{\mathbb{E}}\left[g'\left(x,\theta\right)\right]\left(\hat{\mathbb{E}}\left[H\left(x,\theta\right)PH'\left(x,\theta\right)\right]\right)^{-1}\hat{\mathbb{E}}\left[g\left(x,\theta\right)\right].$$

3 Asymptotics

In this section, we show that, despite the estimator's roots in the theory of optimal transport, its large sample behavior remains amenable to standard asymptotic tools since our focus is on an estimator of the parameter θ rather than on an estimator of a distribution. We first consider the case of small errors, a limiting case that may be especially important in the relatively common case of applications where overidentifying restrictions tests are near the rejection region boundary. This limit also parallel the approach taken in the GEL literature, where asymptotic properties are often derived in the case where the overidentifying restrictions hold (e.g., Newey and Smith (2004)).

3.1 Small errors limit

Our small error results enable us to illustrate that there is little risk in using our estimator instead of standard GMM when one is concerned about potential measurement error in the data. If the data were to, in fact, satisfy the moment conditions, using our approach does not sacrifice consistency, root n convergence or asymptotic normality. The only possible drawback would be a suboptimal weighting of overidentifying moment conditions leading to a small increase in variance if there happened to be no measurement error. Conversely, the optimal weighting of efficient GMM is only valid under the assumption that the only reason for not simultaneously satisfying all moment conditions is random sampling. If instead measurement error is the culprit, then the GMM weighting is no longer optimal and the priority becomes to minimize the effect of measurement error, which our approach seeks to accomplish. Hence, in that sense, the method provides a complementary alternative to standard GMM estimation.

Our consistency result requires a number of fairly standard primitive assumptions.

Assumption 3.1 *The random variables x_i form an iid sequence and take value in $\mathcal{X} \subset \mathbb{R}^{d_x}$.*

Assumption 3.2 $\mathbb{E}[g(x_i;\theta_0)] = 0$*, i.e. the observed data x_i satisfy the moment conditions and $\mathbb{E}[g(x_i;\theta)] \neq 0$ for any other $\theta \in \Theta$, a compact set.*

In other words, Assumption 3.2 indicates that we consider here the case where GMM would be consistent.

Assumption 3.3 $\mathbb{V}[g(x_i;\theta_0)] < \infty$*, where $\mathbb{V}$ denotes the variance operator.*

Assumption 3.4 *$g(x;\cdot)$ is almost surely continuous and $\|g(x;\theta)\| \leq h(x)$ for any $\theta \in \Theta$ and for some function h satisfying $\mathbb{E}\left[h\left(x_i\right)\right] < \infty$.*

Assumptions 3.1, 3.2, 3.3 and 3.4 directly parallel those needed to establish the asymptotic properties of a standard GMM estimator (e.g. Theorems 2.6 and 3.2 in Newey and McFadden (1994)). However, given that our estimator, in the small error limit (Equation (6)), contains a sample average involving derivative $H\left(x,\theta\right) \equiv \partial_{1'} g\left(x,\theta\right)$, we need to place some constraints on the behavior of that quantity as well.

Assumption 3.5 *g is differentiable in its first argument and the derivative satisfies $\mathbb{E}[\|\partial_{1'} g(x_i;\theta_0)\|^2] < \infty$. Moreover, $\|\partial_{1'} g(x_i;\theta)\| \neq 0$ almost surely for all $\theta \in \Theta$.*

Assumption 3.6 *$\partial_{1'} g(x;\theta_0)$ is Hölder continuous in x.*

Assumption 3.7 *Third-order partial derivatives in the first argument of g have finite variance.*

Assumption 3.8 $\mathbb{E}[\partial_{1'} g(x_i;\theta_0)\partial_1 g'(x_i;\theta_0)] < \infty$ *is of full rank.*

These assumptions ensure that the minimization problem defined by (1) and (2) is well-behaved, i.e., small changes in the values of x_i do not lead, with positive probability, to jumps in the solution z_i to the optimization problem. It is likely that these assumptions can be relaxed using empirical processes techniques. However, here we favor simply imposing more smoothness (compared to the standard GMM assumptions), because this leads to more transparent assumptions. They can all be stated in terms of the basic function $g(x;\theta)$ that defines the moment condition model, making them fairly primitive. We can then state our first consistency result.

Theorem 3.9 *Under assumptions 3.1-3.8, the OTGMM estimator is consistent for θ_0 and $\lambda = O_p(n^{-1/2})$.*

As a by-product, this theorem also secures a convergence rate on the Lagrange multiplier λ which proves useful for establishing our distributional results. The conditions needed to show asymptotic normality also closely mimic those of standard GMM estimators (e.g. Theorem 3.2 in Newey and McFadden (1994)):

Assumption 3.10 $\theta_0 \in \Theta^\circ$*, the interior of Θ.*

Assumption 3.11 $\mathbb{E}[\sup_{\theta\in\eta}||\partial_{2'}g(x_i;\theta)||] < \infty$ *where* $\eta \subset \Theta$ *is a neighborhood of* θ_0.

Assumption 3.12 $\left(\mathbb{E}[\partial_2 g(x_i;\theta_0)']\left(\mathbb{E}[\partial_1 g(x_i;\theta_0)\partial_1 g(x_i,\theta_0)']\right)^{-1}\mathbb{E}[\partial_2 g(x_i;\theta_0)]\right)$ *is invertible.*

We can then provide an explicit expression of the asymptotic variance of the estimator.

Theorem 3.13 *Under Assumptions 3.1-3.12, the OTGMM estimator is asymptotically normal with* $\sqrt{n}(\hat{\theta}_{OTGMM} - \theta_0) \to^d \mathcal{N}(0; V)$*, where*

$$\begin{aligned} V &= \left(\mathbb{E}[G_2']\left(\mathbb{E}[G_1G_1']\right)^{-1}\mathbb{E}[G_2]\right)^{-1} \times \\ &\quad \left(\mathbb{E}[G_2']\left(\mathbb{E}[G_1G_1']\right)^{-1}\mathbb{E}[gg']\left(\mathbb{E}[G_1G_1']\right)^{-1}\mathbb{E}[G_2]\right) \times \\ &\quad \left(\mathbb{E}[G_2']\left(\mathbb{E}[G_1G_1']\right)^{-1}\mathbb{E}[G_2]\right)^{-1}, \end{aligned}$$

where $G_j \equiv \partial_{j'} g(x_i;\theta_0)$ *and* $g \equiv g(x_i;\theta_0)$.

The variance has the expected "sandwich" form, since the reciprocal weights $\mathbb{E}[G_1G_1']$ differs from the moment variance $\mathbb{E}[gg']$. The asymptotic distribution under constraints on the measurement error follows from a straightforward adaptation of the previous theorem.

Corollary 3.14 *Theorem 3.13 holds under constraint (8), with all instances of* $\mathbb{E}[G_1G_1']$ *replaced by* $\mathbb{E}[G_1PG_1']$*, for* $P = \left(I - C'\left(CC'\right)^{-1}C\right)$.

3.2 Asymptotics under large errors

In some applications, it may be useful to relax the assumption of small measurement error for the method to uniformly handle all cases, whether overidentifying restrictions are violated or not. To handle this possibility more straightforwardly, it proves useful to observe the following equivalence, demonstrated in Appendix B.3.

Theorem 3.15 *If* $g(z,\theta)$ *is differentiable in its arguments, the OTGMM estimator is equivalent to a just-identified GMM estimator expressed in terms of the modified moment function*

$$\tilde{g}(x,\theta,\lambda) = \begin{bmatrix} \partial_2 g'\left(q(x,\theta,\lambda),\theta\right) & \lambda \\ g\left(q(x,\theta,\lambda),\theta\right) & \end{bmatrix} \tag{11}$$

that is a function of the observed data x *and the augmented parameter vector* $\tilde{\theta} \equiv (\theta',\lambda')'$ *and where*

$$q(x,\theta,\lambda) \equiv \arg\min_{z:z-\partial_1 g'(z,\theta)\lambda = x} \|z - x\|^2. \tag{12}$$

Note that $q(x,\theta,\lambda)$ is essentially the inverse of the mapping $z - \partial_1 g'(z,\theta)\lambda = x$ (from Equation (5)), augmented with a rule to select the appropriate branch in case the inverse is multivalued.

The equivalence result of Theorem 3.15 implies that much of asymptotic technical tools used in GMM-type estimators can be adapted to our setup, with the distinction that the function $q(x,\theta,\lambda)$ is defined only implicitly. Hence, many of our efforts below seek to recast necessary conditions on $q(x,\theta,\lambda)$ in terms of more primitive conditions on the moment function $g(z,\theta)$ whenever possible.

We first start with a standard identification condition:

Assumption 3.16 *For some compact sets Θ and Λ, there exists a unique $(\theta_0, \lambda_0) \in \Theta \times \Lambda$ solving $\mathbb{E}\left[\tilde{g}\left(x, \theta, \lambda\right)\right] = 0$ for $\tilde{g}\left(x, \theta, \lambda\right)$ defined in Theorem 3.15.*

Our identification condition could also be stated in terms of uniqueness of the solution of the primal optimization problem (Equations (1) and (2)), but the corresponding GMM formulation of Equation (11) we employ in Assumption 3.16 makes it easier to conceptualize the population limit (i.e. a continuum of observations).

Next, we consider standard continuity and dominance conditions that are used to establish uniform convergence of the GMM objective function. In a high-level form, this condition would read:

Assumption 3.17 *(i) $\tilde{g}\left(x, \theta, \lambda\right)$ is continuous in θ and λ for $(\theta, \lambda) \in \Theta \times \Lambda$ with probability one and (ii) $\mathbb{E}\left[\sup_{(\theta,\lambda)\in\Theta\times\Lambda} \left\|\tilde{g}\left(x, \theta, \lambda\right)\right\|\right] < \infty$.*

Alternatively, Assumption 3.17 can be replaced by more primitive conditions on $g\left(z, \theta\right)$ instead, as given below in Assumptions 3.18 and 3.19.

Assumption 3.18 *(i) $g\left(z, \theta\right)$ and $\partial_{1'} g\left(z, \theta\right)$ are differentiable in θ and (ii) $\partial_{2'} g\left(z, \theta\right)$ is continuous in both arguments.*

This assumption parallels continuity assumptions typically made for GMM, but higher order derivatives of $g\left(z, \theta\right)$ are needed because they enter the moment condition either directly or indirectly via the function $q\left(x, \theta, \lambda\right)$.

The next condition ensures that the function $q\left(x, \theta, \lambda\right)$ is well-behaved.

Assumption 3.19 *$\bar{\nu}\bar{\lambda} < 1$ where $\bar{\lambda} = \max_{\lambda\in\Lambda} \|\lambda\|$ and $\bar{\nu} = \sup_{\theta\in\Theta} \sup_{z\in\mathcal{X}} \max_{k\in\{1,\ldots,d_g\}}$ $\max \operatorname{eigval}\left(\partial_{11'} g_k\left(z, \theta\right)\right)$, in which $\partial_{11'} g_k\left(z, \theta\right)$ exists for $k = 1, \ldots, d_g$ and where $\operatorname{eigval}\left(M\right)$ for some matrix M denotes the set of its eigenvalues.*

The meaning of this assumption is perhaps best communicated through concepts drawn from convex analysis and optimal transport (Galichon (2016)). The idea is that the first order condition $z - \partial_1 g'\left(z, \theta\right)\lambda = x$, which implicitly defines $z = q\left(x, \theta, \lambda\right)$, can be written as

$$\frac{\partial}{\partial z}\left(\frac{z'z}{2} - g'\left(z, \theta\right)\lambda\right) = x.$$

Hence, in analogy with the calculation of Brenier maps, we seek a point z such that the slope of $\frac{z'z}{2} - g'\left(z, \theta\right)\lambda$ is x. A way to ensure that this point is unique and varies smoothly with x is to impose that $\frac{z'z}{2} - g'\left(z, \theta\right)\lambda$ is strictly convex. This is implied by ensuring that the Hessian of $\frac{z'z}{2} - g'\left(z, \theta\right)\lambda$ is positive definite, which is precisely what Assumption 3.19 requires. For notational simplicity, Assumption 3.19 is phrased as a global convexity condition, but our results would hold under a more local convexity condition.

In order to state our remaining regularity conditions, it is useful to introduce a notion of (nonuniform) Lipschitz continuity, combined with some standard dominance conditions.

Definition 3.20 *Let $\mathcal{L}$ be the set of functions $h(z,\theta)$ such that (i) $\mathbb{E}\left[\sup_{\theta\in\Theta}\|h(x,\theta)\|\right]<\infty$ and (ii) there exists a function $\bar{h}(x,\theta)$ satisfying*

$$\|h(z,\theta)-h(x,\theta)\| \leq \bar{h}(x,\theta)\|z-x\| \text{ for all } x,z\in\mathcal{X} \tag{13}$$

$$\mathbb{E}\left[\sup_{\theta\in\Theta}\bar{h}(x,\theta)\|\partial_{1'}g(x,\theta)\|\right] < \infty. \tag{14}$$

where $g(x,\theta)$ is the model's vector of moment conditions.

A Lipschitz continuity-type assumption is made here because it ensures that the behavior of the observed x and the underlying unobserved z will not differ to such an extent that moments of unobserved variables would be infinite, while the corresponding observed moments are finite. Clearly, without such an assumption, observable moments would be essentially uninformative. The idea underlying Definition 3.20 is that we want to define a property that is akin to Lipschitz continuity but that allows for some heterogeneity (through the function $\bar{h}(x,\theta)$ in Equation (13)). This heterogeneity proves particularly useful in the case where $\mathcal{X}$ is not compact (for compact $\mathcal{X}$, one can take $\bar{h}(x,\theta)$ to be constant in x with little loss of generality). For a given function $h(x,\theta)$ that is finite for finite x, membership in $\mathcal{L}$ is easy to check by inspecting the tail behavior (in x) of the given function $h(x,\theta)$. Polynomial tails will suggest a polynomial form for $\bar{h}(x,\theta)$, for instance. Equation (14) strengthen the dominance condition 3.20(i) to ensure that functions $h(x,\theta)$ in $\mathcal{L}$ also satisfy a dominance condition when interacted with other quantities entering the optimization problem, i.e. $\partial_{1'}g(x,\theta)$).

With this definition in hand, we can succinctly state a sufficient condition for $\tilde{g}(x,\theta,\lambda)$ to satisfy a dominance condition:

Assumption 3.21 *$g(\cdot,\cdot)$ and each element of $\partial_2 g'(\cdot,\cdot)$ belong to $\mathcal{L}$.*

We are now ready to state our general consistency result.

Theorem 3.22 *Under Assumptions 3.1, 3.16 and either Assumption 3.17 or Assumptions 3.18, 3.19, 3.21, the OTGMM estimator is consistent ($(\hat{\theta},\hat{\lambda})\xrightarrow{p}(\theta_0,\lambda_0)$).*

We now turn to asymptotic normality. We first need a conventional "interior solution" assumption.

Assumption 3.23 *(θ_0,λ_0) from Assumption 3.16 lies in the interior of $\Theta\times\Lambda$.*

Next, as in any GMM estimator, we need finite variance of the moment functions and their differentiability:

Assumption 3.24 *(i) $\mathbb{V}\left[\tilde{g}(x,\theta_0,\lambda_0)\right]\equiv\Omega$ exists and (ii) $\mathbb{E}\left[\partial\tilde{g}(x,\theta,\lambda)/\partial(\theta',\lambda')\right]\equiv\tilde{G}$ exists and is nonsingular.*

Assumption 3.24(ii) can be expressed in a more primitive fashion using the explicit form for $\tilde{G}$ provided in Theorem 3.27 below.

Next, we first state a high-level dominance condition that ensures uniform convergence of the Jacobian term $\partial\tilde{g}(x,\theta,\lambda)/\partial(\theta',\lambda')$.

Assumption 3.25 *(i)* $\tilde{g}(x,\theta,\lambda)$ *is continuously differentiable in* (θ,λ)*; (ii)* $\mathbb{E}[\sup_{(\theta,\lambda)\in\Theta\times\Lambda} \|\partial\tilde{g}(x,\theta,\lambda)/\partial(\theta',\lambda')\|] < \infty$.

This assumption is implied by the following, more primitive, condition:

Assumption 3.26 *(i)* $g(z,\theta)$ *and* $\partial_{2}g(z,\theta)$ *are continuously differentiable in* θ*, (ii) all elements of* $\partial_{2}g_{k}(z,\theta)$ *and* $\partial_{22'}g_{k}(z,\theta)$ *for* $k=1,\ldots,d_{g}$ *belong to* $\mathcal{L}$ *and (iii) Assumptions 3.18(i) and 3.19 hold.*

Theorem 3.27 *Let the assumptions of Theorem 3.22 hold as well as Assumptions 3.23, 3.24 and either Assumption 3.25 or 3.26. Then,*

$$\sqrt{n}\left(\begin{bmatrix}\hat{\theta}\\ \hat{\lambda}\end{bmatrix}-\begin{bmatrix}\theta_{0}\\ \lambda_{0}\end{bmatrix}\right)\xrightarrow{d}\mathcal{N}\left(0,W^{-1}\right)$$

where $W=\tilde{G}'\Omega^{-1}\tilde{G}$, $\Omega=\mathbb{E}[\tilde{g}\tilde{g}']$,

$$\tilde{g}\equiv\begin{bmatrix}\partial_{2}g'(z_{j},\theta_{0})\lambda_{0}\\ g(z_{j},\theta_{0})\end{bmatrix}\ \textit{and}\ \tilde{G}=\begin{bmatrix}\tilde{G}_{\theta\theta} & \tilde{G}_{\theta\lambda}\\ \tilde{G}_{\lambda\theta} & \tilde{G}_{\lambda\lambda}\end{bmatrix}$$

in which

$$\begin{aligned}
\tilde{G}_{\theta\theta} &\equiv \mathbb{E}[\partial_{22'}(\lambda_{0}'g(z_{j},\theta_{0}))+\partial_{21'}(\lambda_{0}'g(z_{j},\theta_{0}))\partial_{2'}q(x_{j},\theta_{0},\lambda_{0})]\\
\tilde{G}_{\lambda\theta} &\equiv \mathbb{E}[\partial_{2'}g(z_{j},\theta_{0})+\partial_{1'}g(z_{j},\theta_{0})\partial_{2'}q(x_{j},\theta_{0},\lambda_{0})]\\
\tilde{G}_{\theta\lambda} &\equiv \mathbb{E}[\partial_{2}(g'(z_{j},\theta_{0}))+\partial_{21'}(\lambda_{0}'g(z_{j},\theta_{0}))\partial_{3'}q(x_{j},\theta_{0},\lambda_{0})]\\
\tilde{G}_{\lambda\lambda} &\equiv \mathbb{E}[\partial_{1'}g(z_{j},\theta_{0})\partial_{3'}q(x_{j},\theta_{0},\lambda_{0})]
\end{aligned}$$

where $\partial_{k\ell'}$ *denote second derivatives with respect to the* k^{th} *and* ℓ^{th} *functional arguments, suitably transposed, and where* z_{j} *solves* $x_{j}=z_{j}-\partial_{1}g'(z_{j},\theta)\lambda$ *for given* x_{j},θ,λ *and where*

$$\partial_{2'}q(x,\theta,\lambda) = \left[\left(I-\partial_{11'}(\lambda'g(z,\theta))\right)^{-1}\partial_{12'}(\lambda'g(z,\theta))\right]_{z=q(x,\theta,\lambda)} \tag{15}$$

$$\partial_{3'}q(x,\theta,\lambda) = \left[\left(I-\partial_{11'}(\lambda'g(z,\theta))\right)^{-1}\partial_{1}g'(z,\theta)\right]_{z=q(x,\theta,\lambda)}. \tag{16}$$

In particular, for θ*, the partitioned inverse formula gives*

$$\sqrt{n}\left(\hat{\theta}-\theta_{0}\right)\xrightarrow{d}\mathcal{N}\left(0,\left(W_{\theta\theta}-W_{\theta\lambda}W_{\lambda\lambda}^{-1}W_{\lambda\theta}\right)^{-1}\right)$$

where W *is similarly partitioned as:*

$$W=\begin{bmatrix}W_{\theta\theta} & W_{\theta\lambda}\\ W_{\lambda\theta} & W_{\lambda\lambda}\end{bmatrix}$$

The asymptotic variance stated in Theorem 3.27 takes the familiar form expected from a just-identified GMM estimator ($\tilde{G}'\Omega^{-1}\tilde{G}$). The relatively lengthy expressions merely come from explicitly computing the first derivative matrix $\tilde{G}$ in terms of its constituents. This is accomplished by differentiating $\tilde{g}$ with respect to all parameters using the chain rule and calculating the derivative of $q(x,\theta,\lambda)$ using the implicit function theorem.

4 Discussion

Our approach should be contrasted with other proposals aimed at handling violations of overidentifying restrictions.

Masten and Poirier (2021) propose to replace moment equalities $\mathbb{E}[g(x,\theta)] = 0$ by approximate moment equalities $-\varepsilon \leq \mathbb{E}[g(x,\theta)] \leq \varepsilon$ (where the inequalities hold element-by-element). They consider any model where the tolerance vector ε is such that the model is not falsified and such that any tightening of the constraint would result in falsification. The identified set for θ is then the union of all the identified sets obtained for each ε considered. As this approach allows for "adjustments" of the values of the moments rather than of the data points, it is mostly useful if the source of misspecification is an invalid instrument, but less so if measurement error is the problem. Their method does not enforce that the tolerances on the different moments are all consistent with the same underlying measurement error structure. As such, their approach and ours differ in the type of misspecification it aims to address. Another difference is that our method transparently applies to any GMM model while theirs is difficult to apply to models beyond linear IV (The authors note: "In general, it is difficult to define a meaningful and tractable class of relaxations of one's baseline assumptions. In the linear model, however, there is a natural way to relax the exclusion restriction."). In a related vein, their method does not preserve optimal GMM's invariance to general linear transformation of the moment conditions,[2] while ours maintains it. Finally, the two approaches demand completely different estimators and asymptotic analysis techniques, since their method involves the concept of set-identification, while ours does not.

Conley, Hansen and Rossi (2012) also focus on linear IV and seek to relax strict moment inequalities by letting $\mathbb{E}[g(x,\theta)] = \varepsilon$, where ε is given a prior distribution and inference on θ is carried out through Bayesian methods. This method is useful to allow for violations of exclusions restrictions, but less so if the main problem is measurement error, for the same reason as the method of Masten and Poirier (2021). In recent work, Christensen and Connault (2022) instead maintain the moment equalities, but incorporate misspecification by allowing for deviations in the distribution used to evaluate the moments. The amount of deviation is controlled placing a bound on a discrepancy, in the spirit of GEL estimation. Once again, this approach is useful to handle generic misspecification, but is not specifically designed to handle a measurement error structure. It also requires, as an input, a priori information regarding the possible magnitude of the misspecification.

The problem of model misspecification has also been approached from the view point of sensitivity analysis (Andrews, Gentzkow and Shapiro (2017), Bonhomme and Weidner (2022)). However, this treatment is fundamentally limited to local misspecification (whose magnitude decreases with sample size). It mainly provides diagnostic tools and can only deliver a specific estimator (Bonhomme and Weidner (2022)), if one is willing to specify an a priori bound on the misspecification magnitude.

The literature on moment inequality models has also focused on the issue of misspecification (e.g. Andrews and Kwon (2019)), but the necessary methods differ significantly from ours due to the set-identified nature of the problem. In addition, the treatment of misspeci-

[2]The tolerance vector ε is applied element-by-element and the shape of the allowed values of the moment thus depends on the chosen coordinate system.

fication specifically due to measurement error has, to our knowledge, not been considered.

5 Extensions

An important feature of our proposed method is that it lets researchers allow for measurement error while remaining agnostic regarding its specific form. It could be classical or nonclassical, correlated over time or not, etc. Of course, this may not be the optimal approach for all applications. For instance, if researchers do have specific knowledge regarding the measurement error structure, then methods specifically designed for that purpose may be preferable (e.g., see Schennach (2016), Schennach (2020) and references therein). Alternatively, generic methods for handling latent variables may be helpful (Ekeland, Galichon and Henry (2010), Beresteanu, Molchanov and Molinari (2011), Schennach (2014)). Such approaches would, for instance, enable researchers to force the measurement error to have zero mean (perhaps conditionally on other variables). An even more interesting alternative would be a hybrid method in which (i) the possibility of general forms of measurement errors is accounted for with the current method by constructing the equivalent GMM formulation of the model via Theorem 3.15 and (ii) additional restrictions on the form of the measurement error are imposed via additional moment conditions involving some elements of z and x. This could prove a useful middle ground when a priori information regarding the measurement error is available for some, but not all, variables.

While we have mainly focused on the case where $\|\cdot\|$ is a standard Euclidean norm, a weighted Euclidean norm $\|u\| \equiv (u'Wu)^{1/2}$ could also be used to quantify the measurement error. The weighting matrix W reflects the expected relative variances of the errors along the different dimensions of x. (Note that the estimator is invariant to scaling all the weights by the same constant — only the relative magnitude of the weights matter.) The choice of weights is particularly important when the different dimensions have different units. However, simple rules can be used to naturally guide this choice. One possibility is to express all variables on a logarithmic scale, adjusting the moment conditions accordingly, and use an unweighted Euclidean norm. This approach effectively assumes multiplicative errors whose magnitudes (expressed as a percentage) are similar along all dimensions. If one prefers to maintain an additive error structure, one can simply scale all elements of x by their corresponding standard deviation (or some other measure of scale), again adjusting the moment conditions accordingly and using an unweighted norm. This approach then assumes that the error magnitudes for each element of x are a similar fraction of that variable's overall scale. The idea underlying these suggestions is basically to ensure that x is either dimensionless or contains elements expressed in comparable units. This being said, it is important to note that, in the small error limit (i.e., when the overidentifying restrictions hold), the estimator is consistent regardless of the choice of weight. This is entirely analogous to what happens with GEL methods, which are consistent for any choice of tilting function if the overidentifying restrictions hold, but have otherwise different pseudo-true values for different tilting functions when overidentifying restrictions do not hold.

It is fruitful to observe that the proposed method assigns the source of overidentification test failure entirely to measurement error, while standard GEL approaches would assign it entirely to sampling bias. However, one could also consider intermediate situations. When

our method is used with the constraint that measurement error is only present in some, but not all, variables, then the overidentifying restrictions are not automatically satisfied. This implies that there is still information to be extracted from the overidentifying restrictions and would suggest a hybrid method where both measurement error, handled via our approach, and re-weighting of the sample, handled via GEL, are simultaneously allowed.

6 Simulations

We conduct simulations to assess the performance of our estimator and compare it to efficient GMM. We consider both the OTGMM estimator (Equations (1) and (2)) and the GMM estimator obtained under the assumption of small errors (Equation (6)).

For a sample size of $n = 100$, we consider various moment conditions, underlying distributions and signal-to-noise ratios. There is an underlying random variable z_i which satisfies the moment conditions, but the researcher observes $x_i = z_i + \sigma e_i$ with $e_i \sim \mathcal{N}(0,1)$. We consider different values for the measurement error scale σ in order to assess the impact of magnitude of the measurement error on the performance of estimators that only use the observed x_i.

Specifically, we consider the following distributions for z_i: $z_i \sim \mathcal{N}(1.5, 2)$ $z_i \sim \text{Unif}[1,2]$ (uniform) $z_i \sim \mathcal{B}(5, 0.3)$ (binomial) and $\sigma = 0, 0.5, 1, 1.5, 2, 2.5$. The true parameter value is $\theta_0 = 1.5$, as obtained by the following moment conditions:

$$\mathbb{E}[z_i - \theta] = 0,\ \mathbb{E}[e^{z_i} - \frac{2}{3}\theta\mathbb{E}[e^{z_i}]] = 0 \tag{17}$$

$$\mathbb{E}[z_i - \theta] = 0,\ \mathbb{E}[\frac{e^{2z_i-3}}{1+e^{2z-3}} - \frac{2}{3}\theta\frac{e^{2z_i-3}}{1+e^{2z-3}}] = 0 \tag{18}$$

$$\mathbb{E}[e^{z_i} - \frac{2}{3}\theta\mathbb{E}[e^{z_i}]] = 0,\ \mathbb{E}[\frac{e^{2z_i-3}}{1+e^{2z-3}} - \frac{\theta e^{2z_i-3}}{(1.5)\,(1+e^{2z-3})}] = 0 \tag{19}$$

Finally, we consider a last process: $z_i \sim \text{Exp}(\frac{2}{3})$ with the moment conditions

$$\mathbb{E}[z_i - \theta] = 0,\ \mathbb{E}[{z_i}^2 - 2\theta^2] = 0. \tag{20}$$

In all cases, the model is correctly specified in the absence of measurement error ($\sigma = 0$) but starts to fail overidentifying restrictions when there is measurement error ($\sigma > 0$ so that $x \neq z$).

In Tables 1-4, we report the estimation error $\hat{\theta} - \theta_0$ and decompose it into its bias, standard deviation and the root mean square error (RMSE). These quantities are evaluated using averages over 5000 replications. We consider various estimators $\hat{\theta}$: the linear approximation to OTGMM in the small-error limit (leftmost columns), OTGMM in the general large-error case (middle columns) and efficient GMM ignoring the presence of measurement error (rightmost columns).

It is clear that OTGMM is, in general, preferable to its linear approximation. The key take-away from these simulations is that the OTGMM estimator exhibits the ability to substantially reduce bias while not substantially increasing the variance relative to efficient GMM. As a result, the overall RMSE criterion points in favor of OTGMM. This is exactly the

type of behavior one would expect for an effective measurement error-correcting method. The reduction in bias is especially important for inference and testing, as it significantly reduces size distortion. In contrast, a small increase in variance does not affect inference validity, as this variance can be straightforwardly accounted for in the asymptotics, unlike the bias, which is generally unknown.

Table 1: Simulation results: Equation(17)

										Bias								
	Linear approximation						OTGMM						Efficient GMM					
ME scale	0	0.5	1	1.5	2	2.5	0	0.5	1	1.5	2	2.5	0	0.5	1	1.5	2	2.5
Normal	0	-0.01	-0.05	-0.16	-0.5	-1.77	0	-0.01	-0.03	-0.05	-0.05	-0.05	-0.02	-0.03	-0.09	-0.18	-0.23	-0.26
Uniform	0	-0.23	-1.12	-3.58	-10.98	-37.34	0	-0.09	-0.12	-0.11	-0.09	-0.07	0	-0.13	-0.23	-0.27	-0.29	-0.29
Binomial	0	-0.02	-0.1	-0.33	-0.99	-3.31	0	-0.01	-0.04	-0.06	-0.06	-0.05	0	-0.04	-0.14	-0.21	-0.25	-0.27
										Standard deviation								
	Linear approximation						OTGMM						Efficient GMM					
ME scale	0	0.5	1	1.5	2	2.5	0	0.5	1	1.5	2	2.5	0	0.5	1	1.5	2	2.5
Normal	0.15	0.15	0.16	0.23	1.09	9.14	0.15	0.15	0.17	0.2	0.25	0.29	0.17	0.16	0.16	0.19	0.24	0.29
Uniform	0.01	0.05	0.28	1.48	9.07	69.12	0.01	0.04	0.1	0.15	0.2	0.25	0.01	0.04	0.09	0.15	0.21	0.26
Binomial	0.1	0.1	0.13	0.24	1.1	7.5	0.1	0.11	0.14	0.18	0.22	0.27	0.1	0.1	0.13	0.17	0.23	0.28
										RMSE								
	Linear approximation						OTGMM						Efficient GMM					
ME scale	0	0.5	1	1.5	2	2.5	0	0.5	1	1.5	2	2.5	0	0.5	1	1.5	2	2.5
Normal	0.15	0.15	0.17	0.28	1.2	9.31	0.15	0.15	0.17	0.21	0.25	0.29	0.17	0.16	0.19	0.26	0.33	0.39
Uniform	0.01	0.23	1.15	3.88	14.24	78.56	0.01	0.1	0.15	0.19	0.22	0.26	0.01	0.14	0.24	0.31	0.36	0.39
Binomial	0.1	0.11	0.16	0.41	1.48	8.2	0.1	0.11	0.14	0.19	0.23	0.28	0.1	0.11	0.19	0.27	0.34	0.39

Table 2: Simulation results: Equation(18)

										Bias								
	Linear approximation						OTGMM						Efficient GMM					
ME scale	0	0.5	1	1.5	2	2.5	0	0.5	1	1.5	2	2.5	0	0.5	1	1.5	2	2.5
Normal	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Uniform	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Binomial	0	0.01	0.01	0.02	0.02	0.02	0	0.01	0.02	0.03	0.03	0.04	-0.01	0.01	0.04	0.06	0.06	0.06
										Standard deviation								
	Linear approximation						OTGMM						Efficient GMM					
ME scale	0	0.5	1	1.5	2	2.5	0	0.5	1	1.5	2	2.5	0	0.5	1	1.5	2	2.5
Normal	0.11	0.12	0.12	0.13	0.15	0.16	0.11	0.12	0.12	0.12	0.13	0.13	0.1	0.1	0.1	0.1	0.09	0.09
Uniform	0	0.04	0.15	0.29	0.44	0.6	0	0.05	0.1	0.12	0.12	0.13	0	0.04	0.09	0.1	0.09	0.09
Binomial	0.1	0.11	0.12	0.15	0.17	0.2	0.1	0.11	0.12	0.12	0.13	0.13	0.1	0.1	0.1	0.1	0.1	0.1
										RMSE								
	Linear approximation						OTGMM						Efficient GMM					
ME scale	0	0.5	1	1.5	2	2.5	0	0.5	1	1.5	2	2.5	0	0.5	1	1.5	2	2.5
Normal	0.11	0.12	0.12	0.13	0.15	0.16	0.11	0.12	0.12	0.12	0.13	0.13	0.1	0.1	0.1	0.1	0.09	0.09
Uniform	0	0.04	0.15	0.29	0.44	0.6	0	0.05	0.1	0.12	0.12	0.13	0	0.04	0.09	0.1	0.09	0.09
Binomial	0.1	0.11	0.13	0.15	0.17	0.2	0.1	0.11	0.12	0.13	0.13	0.14	0.1	0.1	0.11	0.11	0.11	0.11

Table 3: Simulation results: Equation(19)

	Bias																	
	Linear approximation						OTGMM						Efficient GMM					
ME scale	0	0.5	1	1.5	2	2.5	0	0.5	1	1.5	2	2.5	0	0.5	1	1.5	2	2.5
Normal	-0.01	0	0.03	0.09	0.3	0.99	-0.01	0	0	0	0	0	-0.01	-0.01	-0.04	-0.07	-0.07	-0.07
Uniform	0	-0.13	-0.63	-2.03	-6.16	-20.38	0	-0.04	-0.04	-0.01	0	0	0	-0.12	-0.16	-0.15	-0.12	-0.09
Binomial	0	0.01	0.04	0.09	0.23	0.68	0	0.01	0.02	0.03	0.03	0.04	-0.01	-0.01	-0.05	-0.07	-0.06	-0.04
	Standard deviation																	
	Linear approximation						OTGMM						Efficient GMM					
ME scale	0	0.5	1	1.5	2	2.5	0	0.5	1	1.5	2	2.5	0	0.5	1	1.5	2	2.5
Normal	0.11	0.11	0.12	0.16	0.47	2.47	0.11	0.11	0.12	0.12	0.13	0.13	0.12	0.12	0.11	0.12	0.13	0.14
Uniform	0.04	0.06	0.27	1.1	5.05	31.07	0.04	0.06	0.09	0.11	0.12	0.13	0.04	0.06	0.09	0.11	0.13	0.13
Binomial	0.1	0.1	0.11	0.13	0.31	1.53	0.1	0.1	0.11	0.12	0.13	0.13	0.1	0.1	0.11	0.12	0.13	0.14
	RMSE																	
	Linear approximation						OTGMM						Efficient GMM					
ME scale	0	0.5	1	1.5	2	2.5	0	0.5	1	1.5	2	2.5	0	0.5	1	1.5	2	2.5
Normal	0.11	0.11	0.12	0.18	0.56	2.66	0.11	0.11	0.12	0.12	0.13	0.13	0.12	0.12	0.12	0.14	0.15	0.15
Uniform	0.04	0.14	0.69	2.31	7.96	37.15	0.04	0.08	0.1	0.11	0.12	0.13	0.04	0.13	0.19	0.19	0.17	0.16
Binomial	0.1	0.1	0.12	0.16	0.39	1.67	0.1	0.1	0.11	0.12	0.13	0.14	0.1	0.1	0.12	0.14	0.15	0.15

Table 4: Simulation results: Equation(20)

	Bias																	
	Linear approximation						OTGMM						Efficient GMM					
ME scale	0	0.5	1	1.5	2	2.5	0	0.5	1	1.5	2	2.5	0	0.5	1	1.5	2	2.5
Exponential	-0.02	0.02	0.15	0.37	0.66	0.98	-0.02	0.02	0.13	0.26	0.42	0.57	-0.05	-0.02	0.09	0.28	0.49	0.7
	Standard deviation																	
	Linear approximation						OTGMM						Efficient GMM					
ME scale	0	0.5	1	1.5	2	2.5	0	0.5	1	1.5	2	2.5	0	0.5	1	1.5	2	2.5
Exponential	0.17	0.16	0.16	0.17	0.17	0.18	0.17	0.16	0.16	0.18	0.2	0.23	0.16	0.16	0.17	0.2	0.24	0.28
	RMSE																	
	Linear approximation						OTGMM						Efficient GMM					
ME scale	0	0.5	1	1.5	2	2.5	0	0.5	1	1.5	2	2.5	0	0.5	1	1.5	2	2.5
Exponential	0.17	0.17	0.22	0.41	0.68	1	0.17	0.17	0.21	0.32	0.46	0.62	0.17	0.16	0.2	0.35	0.54	0.76

7 Conclusion

We have proposed a novel optimal transport-based version of the Generalized Method of Moment (GMM) that fulfills, by construction, the overidentifying restrictions by introducing the smallest amount of measurement error necessary to simultaneously satisfy all moment conditions. This approach conceptually merges the Generalized Empirical Likelihood (GEL) and optimal transport methodologies. It provides a theoretically motivated interpretation to GMM results when standard overidentification tests reject the null. GEL approaches handle model misspecification by re-weighting the data, which would be appropriate when misspecification arise from improper sampling of the population. In contrast, our optimal transport approach is appropriate when measurement error is the source of misspecification, which is arguably a common situation in applications. As a by-product, our approach provides insight into the measurement error structure of the variables.

References

Aguiar, V. H., and N. Kashaev. 2021. "Stochastic Dynamic Revealed Preferences with measurement error." *Review of Economic Studies*, 88: 2042–2093.

Andrews, D. W. K., and S. Kwon. 2019. "Inference in moment inequality models that is robust to spurious precision under model misspecification." cowles foundation discussion paper 2184R.

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.

Beresteanu, A., I. Molchanov, and F. Molinari. 2011. "Sharp identification regions in models with convex moment predictions." *Econometrica*, 79: 1785–1821.

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

Carlier, G., V. Chernozhukov, and A. Galichon. 2016. "Vector quantile regression: An optimal transport approach." *The Annals of Statistics*, 44: 1165–1192.

Chernozhukov, V., A. Galichon, M. Hallin, and M. Henry. 2017. "Monge–Kantorovich depth, quantiles, ranks and signs." *Annals of Statistics*, 45: 223–256.

Chesher, A. 1991. "The Effect of Measurement Error." *Biometrika*, 78: 451.

Christensen, T., and B. Connault. 2022. "Counterfactual Sensitivity and Robustness." *Econometrica*, forthcoming.

Conley, T. G., C. B. Hansen, and P. E. Rossi. 2012. "Plausibly Exogenous." *The Review of Economics and Statistics*, 94: 260–272.

Doraszelski, U., and J. Jaumandreu. 2013. "R&D and Productivity: Estimating Endogenous Productivity." *Review of Economic Studies*, 80: 1338–1383.

Ekeland, I., A. Galichon, and M. Henry. 2010. "Optimal transportation and the falsifiability of incompletely specified economic models." *Economic Theory*, 42: 355–374.

Ekeland, I., A. Galichon, and M. Henry. 2011. "Comonotonic Measures of Multivariate Risks." *Mathematical Finance*, 22: 109–132.

Galichon, A. 2016. *Optimal Transport Methods in Economics.* Princeton:Princeton University Press.

Gunsilius, F., and S. M. Schennach. 2021. "Independent Principal Component Analysis." *Journal of the American Statistical Association*, forthcoming.

Hall, A. R., and A. Inoue. 2003. "The large sample behavior of the generalized method of moments estimator in misspecified models." *Journal of Econometrics*, 114: 361–394.

Hansen, L. P. 1982. "Large sample properties of generalized method of moment estimators." *Econometrica*, 50: 1029–1054.

Hansen, L. P. 2001. "Acknowledging Misspecification in Macroeconomic Theory." *Review of Economic Dynamics*, 4: 519–535.

Imbens, G. W. 1997. "One-Step Estimators for Over-Identified Generalized Method of Moments Models." *Review of Economic Studies*, 64: 359–383.

Kitamura, Y., and M. Stutzer. 1997. "An Information-Theoretic Alternative to Generalized Method of Moment Estimation." *Econometrica*, 65: 861–874.

Masten, M. A., and A. Poirier. 2021. "Salvaging falsified instrumental variable models." *Econometrica*, 89: 1449–1469.

Newey, W., and D. McFadden. 1994. "Large Sample Estimation and Hypothesis Testing." In *Handbook of Econometrics*. Vol. IV, , ed. R. F. Engel and D. L. McFadden. Elsevier Science.

Newey, W., and R. J. Smith. 2004. "Higher-Order Properties of GMM and Generalized Empirical Likelihood Estimators." *Econometrica*, 72: 219–255.

Owen, A. B. 1988. "Empirical Likelihood Ratio Confidence Intervals for a Single Functional." *Biometrika*, 75: 237–249.

Qin, J., and J. Lawless. 1994. "Empirical Likelihood and General Estimating Equations." *Annals of Statistics*, 22: 300–325.

Sawa, T. 1978. "Information Criteria for Discriminating Among Alternative Regression Models." *Econometrica*, 46: 1273–1291.

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

Schennach, S. M. 2007. "Point Estimation with Exponentially Tilted Empirical Likelihood." *Annals of Statistics*, 35: 634–672.

Schennach, S. M. 2014. "Entropic Latent Variable Integration via Simulation." *Econometrica*, 82: 345–386.

Schennach, S. M. 2016. "Recent Advances in the Measurement Error Literature." *Annual Reviews of Economics*, 8: 341–377.

Schennach, S. M. 2020. "Mismeasured and unobserved variables." In *Handbook of Econometrics*. Vol. 7A, , ed. J. Heckman S. Durlauf, L. Hansen and R. Matzkin, Chapter 6, 487–565. Elsevier Science.

Villani, C. 2009. "Optimal transport: Old and New." In *Grundlehren der mathematischen Wissenschaften*. Heidelberg:Springer-Verlag.

White, H. 1982. "Maximum Likelihood Estimation of Misspecified Models." *Econometrica*, 50: 1–26.

A Algorithms

A.1 Iterative solution

The first order condition (5) can be re-written as

$$(z_j - x_j) = \partial_1 g'(z_j, \theta)\,\lambda. \tag{21}$$

We seek to construct a sequence z_j^t $(t = 0, 1, \ldots)$ that converges to z_j, starting with $z_j^t|_{t=0} = x_j$. From the moment conditions and (21), we have:

$$0 = \frac{1}{n}\sum_{i=1}^{n} g(z_i, \theta) = \frac{1}{n}\sum_{i=1}^{n} g\left(x_j + \partial_1 g'(z_j, \theta)\,\lambda, \theta\right).$$

Adding and subtracting z_j^t yields

$$\begin{aligned} 0 &= \frac{1}{n}\sum_{i=1}^{n} g\left(z_j^t + \left(x_j - z_j^t + \partial_1 g'(z_j, \theta)\,\lambda\right), \theta\right) \\ &\approx \frac{1}{n}\sum_{i=1}^{n} g\left(z_j^t, \theta\right) + \frac{1}{n}\sum_{i=1}^{n} \partial_{1'} g\left(z_j^t, \theta\right)\left(x_j - z_j^t + \partial_1 g'(z_j, \theta)\,\lambda\right) \end{aligned}$$

where the expansion is justified from the fact that $x_j - z_j^t + \partial_1 g'(z_j, \theta)\,\lambda \longrightarrow 0$ as $z_j^t \longrightarrow z_j$.

In the same limit, $\partial_1 g'\left(z_j^t, \theta\right) \longrightarrow \partial_1 g'(z_j, \theta)$, so

$$\begin{aligned} 0 &\approx \frac{1}{n}\sum_{i=1}^{n} g\left(z_j^t, \theta\right) + \frac{1}{n}\sum_{i=1}^{n} \partial_{1'} g\left(z_j^t, \theta\right)\left(x_j - z_j^t + \partial_1 g'\left(z_j^t, \theta\right)\lambda\right) \\ &= \frac{1}{n}\sum_{i=1}^{n} g\left(z_j^t, \theta\right) + \frac{1}{n}\sum_{i=1}^{n} \partial_{1'} g\left(z_j^t, \theta\right)\left(x_j - z_j^t\right) + \frac{1}{n}\sum_{i=1}^{n} \partial_{1'} g\left(z_j^t, \theta\right)\partial_1 g'\left(z_j^t, \theta\right)\lambda \\ &= \hat{\mathbb{E}}\left[g\left(z^t, \theta\right)\right] + \hat{\mathbb{E}}\left[H\left(z^t, \theta\right)\left(x - z^t\right)\right] + \hat{\mathbb{E}}\left[H\left(z^t, \theta\right)H'\left(z^t, \theta\right)\right]\lambda. \end{aligned}$$

Isolating λ gives the approximation to the Lagrange multiplier at step $t+1$:

$$\lambda^{t+1} = \left(\hat{\mathbb{E}}\left[H\left(z^t, \theta\right)H'\left(z^t, \theta\right)\right]\right)^{-1}\left(-\hat{\mathbb{E}}\left[g\left(z^t, \theta\right)\right] + \hat{\mathbb{E}}\left[H\left(z^t, \theta\right)\left(z^t - x\right)\right]\right). \tag{22}$$

From this, we can improve the approximation to z_j to go to the next step, using (21):

$$z_j^{t+1} = x_j + H'\left(z^t, \theta\right)\lambda^{t+1}. \tag{23}$$

It can be directly verified that the values of z_j and λ that satisfy the first order conditions are indeed a fixed point of this iterative rule. In the next subsection we shall provide conditions under which this fixed point is also attractive.

After iteration to convergence, the objective function can be written in term of the converged values of z and λ:

$$\begin{aligned} \hat{Q}(\theta) &= \frac{1}{2n}\sum_j \|z_j - x_j\|^2 = \frac{1}{2n}\sum_j \|H'(z_j, \theta)\,\lambda\|^2 = \frac{1}{2n}\sum_j \lambda' H(z_j, \theta)\,H'(z_j, \theta)\,\lambda \\ &= \frac{1}{2}\lambda'\hat{\mathbb{E}}\left[H(z, \theta)\,H'(z, \theta)\right]\lambda. \end{aligned}$$

A.2 Iterative procedure convergence

Substituting (22) into (23) yields an iterative rule expressed solely in terms of z_j^t:

$$z_j^{t+1} = x_j + H'\left(z_j^t, \theta\right)\left(\hat{\mathbb{E}}\left[H\left(z^t, \theta\right) H'\left(z^t, \theta\right)\right]\right)^{-1}\left(-\hat{\mathbb{E}}\left[g\left(z^t, \theta\right)\right] + \hat{\mathbb{E}}\left[H\left(z^t, \theta\right)\left(z^t - x\right)\right]\right). \tag{24}$$

This is an iterative rule of the form $\mathbf{z}^{t+1} = f\left(\mathbf{z}^t\right)$, for $\mathbf{z}^t = \left(z_1^{t\prime}, \ldots, z_n^{t\prime}\right)' \in \mathbb{R}^{nd_x}$ with fixed point denoted $\mathbf{z}^\infty$. We then have the following result.

Assumption A.1 *(i) $g\left(z, \theta\right)$ is twice continuously differentiable in z and (ii) $\hat{\mathbb{E}}\left[H\left(z, \theta\right) H'\left(z, \theta\right)\right]$ is nonsingular for $\mathbf{z}$ in a the closure of an open neighborhood η of the fixed point $\mathbf{z}^\infty$.*

Theorem A.2 *Under Assumption A.1, for a given sample $x_1, \ldots, x_n$, there exists a neighborhood η of $\mathbf{z}^\infty$, such that the iterative procedure defined by Equation (24) and starting at any $\mathbf{z}^0 \in \eta$ converges to the fixed point $\mathbf{z}^\infty$, provided $\|\lambda\|$ is sufficiently small (where λ solves the first order condition (21)).*

The condition that the initial point $\mathbf{z}^0$ should lie in a neighborhood of the solution is standard — most Newton-Raphson-type iterative refinements have a similar requirement. If necessary, this requirement can be met by simply attempting many different starting points in search for one that yields a convergent sequence. The condition that λ be small intuitively means that the measurement error should not be too large. This is a purely numerical condition which has nothing to do with sample size, statistical significance of specification tests. In particular, it does not mean that the measurement error magnitude must decreases with sample size or that the effect of the errors should be small relative to the estimator's standard deviation. Typically, the constraint on λ is relaxed as the starting point $\mathbf{z}^0$ is chosen closer to the solution $\mathbf{z}^\infty$.

Proof of Theorem A.2. For a rule of the form $\mathbf{z}^{t+1} = f\left(\mathbf{z}^t\right)$, Banach's Fixed Point Theorem applied to a neighborhood of $\mathbf{z}^\infty$ provides a simple sufficient condition for convergence: (i) f must be continuously differentiable in a neighborhood of $\mathbf{z}^\infty$ and (ii) all eigenvalues of the matrix $\left[\partial f\left(\mathbf{z}\right)/\partial \mathbf{z}'\right]_{\mathbf{z}=\mathbf{z}^\infty}$ must have a modulus strictly less than 1.

The smoothness condition (i) is trivially satisfied under Assumption A.1. Next, letting $z_{i,k}^t$ denote one element of the vector z_i^t, and $H_{\cdot k}\left(z_i^t, \theta\right)$ denote the k^{th} column of $H\left(z_i^t, \theta\right)$, we can express all blocks $\partial z_j^{t+1}/\partial z_{i,k}^t$ of the matrix of partial derivatives of $f\left(\mathbf{z}\right)$:

$$\begin{aligned}
\frac{\partial z_j^{t+1}}{\partial z_{i,k}^t} \;=\; & \left[\frac{\partial}{\partial z_{i,k}^t} H'\left(z_i^t, \theta\right)\left(\hat{\mathbb{E}}\left[H\left(z^t, \theta\right) H'\left(z^t, \theta\right)\right]\right)^{-1}\right]\left(-\hat{\mathbb{E}}\left[g\left(z^t, \theta\right)\right] + \hat{\mathbb{E}}\left[H\left(z^t, \theta\right)\left(z^t - x\right)\right]\right) \\
& + H'\left(z_i^t, \theta\right)\left(\hat{\mathbb{E}}\left[H\left(z^t, \theta\right) H'\left(z^t, \theta\right)\right]\right)^{-1} \times \\
& \left(-n^{-1} H_{\cdot k}\left(z_i^t, \theta\right) + n^{-1} H_{\cdot k}\left(z_i^t, \theta\right) + n^{-1}\left[\frac{\partial}{\partial z_{i,k}^t} H\left(z_i^t, \theta\right)\right]\left(z_i^t - x_i\right)\right),
\end{aligned}$$

where the two $n^{-1} H_{\cdot k}\left(z_i^t, \theta\right)$ terms cancel each other. At $\mathbf{z}^t = \mathbf{z}^\infty$, $\hat{\mathbb{E}}\left[g\left(z^\infty, \theta\right)\right] = 0$ and

$(z_i^\infty - x_i) = H'(z_i^\infty, \theta)\lambda$ and we have:

$$\begin{aligned}\frac{\partial z_j^{t+1}}{\partial z_{i,k}^t} &= \left[\frac{\partial}{\partial z_{i,k}^\infty} H'(z_i^\infty, \theta)\left(\hat{\mathbb{E}}\left[H(z^\infty, \theta) H'(z^\infty, \theta)\right]\right)^{-1}\right]\hat{\mathbb{E}}\left[H(z^\infty, \theta) H'(z^\infty, \theta)\right]\lambda \\ &\quad + H'(z_i^\infty, \theta)\left(\hat{\mathbb{E}}\left[H(z^\infty, \theta) H'(z^\infty, \theta)\right]\right)^{-1} n^{-1}\left[\frac{\partial}{\partial z_{i,k}^\infty} H(z_i^\infty, \theta)\right] H'(z_i^\infty, \theta)\lambda\end{aligned}$$

This expression (once all derivatives of products are expanded) has the general form of a product of λ with functions of $\mathbf{z}$ that contain terms of the form $\left(\hat{\mathbb{E}}\left[H(z, \theta) H'(z, \theta)\right]\right)^{-1}$, which is nonsingular for $\mathbf{z} \in \eta$ by Assumption A.1(ii), and derivatives of $g(z, \theta)$ up to order 2, which are bounded for z in the compact set $\{z_j : (z_1, \ldots, z_n) \in \eta \text{ and } j = 1, \ldots, n\}$ by Assumption A.1(i). Hence the elements $\partial z_j^{t+1} / \partial z_{i,k}^t$ are bounded by a constant times λ. It follows that the eigenvalues of the matrix of partial derivatives of $f(\mathbf{z})$ can be made strictly less than 1 for λ sufficiently small. ■

B Proofs

For a being a vector or matrix, let $\|a\| = \left(\sum_{i,j} a_{i,j}\right)^{1/2}$.

B.1 Linearized estimator

Proof of proposition 2.2. In the following derivation, the approximation denoted by "$\approx$" are exact to first order in $z_j - x_j$. In that limit, $\partial_1 g'(z_j, \theta) \approx \partial_1 g'(x_j, \theta)$. Therefore:

$$\begin{aligned} z_j - x_j &\approx \partial_1 g'(x_j, \theta)\lambda \\ z_j &\approx x_j + \partial_1 g'(x_j, \theta)\lambda \end{aligned} \tag{25}$$

Substituting into the constraint:

$$\sum_j g(x_j + \partial_1 g'(x_j, \theta)\lambda, \theta) \approx 0$$

Using a Taylor expansion:

$$\sum_j \left(g(x_j, \theta) + \partial_{1'} g(x_j, \theta)\partial_1 g'(x_j, \theta)\lambda\right) \approx 0$$

i.e.,

$$\frac{1}{n}\sum_j g(x_j, \theta) + \left(\frac{1}{n}\sum_j \partial_{1'} g(x_j, \theta)\left(\partial_{1'} g(x_j, \theta)\right)'\right)\lambda \approx 0$$

or

$$\hat{\mathbb{E}}\left[g(x, \theta)\right] + \left(\hat{\mathbb{E}}\left[H(x, \theta) H'(x, \theta)\right]\right)\lambda \approx 0$$

where

$$H(x, \theta) = \partial_{1'} g(x, \theta)$$

and where the operator $\hat{\mathbb{E}}$ denotes sample averages.

This implies that:

$$\lambda \approx -\left(\hat{\mathbb{E}}\left[H\left(x,\theta\right)H'\left(x,\theta\right)\right]\right)^{-1}\hat{\mathbb{E}}\left[g\left(x,\theta\right)\right] \tag{26}$$

Substituting (25) and (26) back into the objective function (1):

$$\begin{aligned}
& \frac{1}{2n}\sum_{j}\left\|z_j - x_j\right\|^2 \\
\approx\ & \frac{1}{2n}\sum_{j}\left\|x_j + H'\left(x_j,\theta\right)\lambda - x_j\right\|^2 \\
\approx\ & \frac{1}{2n}\sum_{j}\left\|-H'\left(x_j,\theta\right)\left(\hat{\mathbb{E}}\left[H\left(x,\theta\right)H'\left(x,\theta\right)\right]\right)^{-1}\hat{\mathbb{E}}\left[g\left(x,\theta\right)\right]\right\|^2 \\
=\ & \frac{1}{2n}\sum_{j}\hat{\mathbb{E}}\left[g'\left(x,\theta\right)\right]\left(\hat{\mathbb{E}}\left[H\left(x,\theta\right)H'\left(x,\theta\right)\right]\right)^{-1}H\left(x_j,\theta\right)H'\left(x_j,\theta\right)\left(\hat{\mathbb{E}}\left[H\left(x,\theta\right)H'\left(x,\theta\right)\right]\right)^{-1}\hat{\mathbb{E}}\left[g\left(x,\theta\right)\right] \\
=\ & \frac{1}{2}\hat{\mathbb{E}}\left[g'\left(x,\theta\right)\right]\left(\hat{\mathbb{E}}\left[H\left(x,\theta\right)H'\left(x,\theta\right)\right]\right)^{-1}\left(\hat{\mathbb{E}}\left[H\left(x,\theta\right)H'\left(x,\theta\right)\right]\right)\left(\hat{\mathbb{E}}\left[H\left(x,\theta\right)H'\left(x,\theta\right)\right]\right)^{-1}\hat{\mathbb{E}}\left[g\left(x,\theta\right)\right] \\
=\ & \frac{1}{2}\hat{\mathbb{E}}\left[g'\left(x,\theta\right)\right]\left(\hat{\mathbb{E}}\left[H\left(x,\theta\right)H'\left(x,\theta\right)\right]\right)^{-1}\hat{\mathbb{E}}\left[g\left(x,\theta\right)\right]
\end{aligned}$$

Therefore, for small errors the estimator is equivalent to minimizing:

$$\hat{\mathbb{E}}\left[g'\left(x,\theta\right)\right]\left(\hat{\mathbb{E}}\left[H\left(x,\theta\right)H'\left(x,\theta\right)\right]\right)^{-1}\hat{\mathbb{E}}\left[g\left(x,\theta\right)\right]$$

which is a GMM-like objective function but with a non-standard weighting matrix $\left(\hat{\mathbb{E}}\left[H\left(x,\theta\right)H'\left(x,\theta\right)\right]\right)^{-1}$.

■

B.2 Constrained optimization

The first order conditions of the Lagrangian (9) with respect to z_j is

$$\left(z_j - x_j\right) - \partial g_1'\left(z_j,\theta\right)\lambda - C'\gamma_j = 0. \tag{27}$$

Re-arranging and pre-multiplying both sides by the full column rank matrix C yields:

$$C\left(z_j - x_j\right) - CC'\gamma_j = C\partial g_1'\left(z_j,\theta\right)\lambda,$$

thus allowing us to solve for γ_j:

$$\gamma_j = -\left(CC'\right)^{-1}C\partial g_1'\left(z_j,\theta\right)\lambda.$$

Upon substitution of γ_i into (27) and simple re-arrangements, we obtain

$$\begin{aligned}
\left(z_j - x_j\right) &= \left(I - C'\left(CC'\right)^{-1}C\right)\partial g_1'\left(z_j,\theta\right)\lambda \\
&= PH\left(z,\theta\right)\lambda
\end{aligned}$$

where $P = \left(I - C'\left(CC'\right)^{-1}C\right)$ and $H\left(z,\theta\right) = \partial_{1'}g\left(z,\theta\right)$.

B.3 Asymptotics

Proof of Theorem 3.9. We maximize $\frac{1}{2}\sum_{i=1}^{n}||z_i - x_i||^2$ subject to $\sum_{i=1}^{n} g(z_i, \theta) = 0$. First-order conditions read

$$z_i - x_i = \partial_1 g'(z_i; \theta)\lambda \tag{28}$$

$$\sum_{i=1}^{n} g(z_i; \theta) = 0 \tag{29}$$

It is first shown that there exists a sequence z_i^* that matches the moment condition $\sum_{i=1}^{n} g(z_i^*; \theta_0)$ and converges uniformly to the x_i's, implying convergence of the z_i's by their definition in the optimization problem.

We now discuss how to eliminate observations that are too close to a zero gradient.

For some η and δ let A be the set of all x_i such that $\inf_{y \in B_\delta(x_i)} \|\partial_1 g(x_i; \theta_0)\| \geq \eta$. We must have $\mathbb{P}[A] > 0$ for some (η, δ) because otherwise $\{\inf_{y \in B_\delta(x_i)} \|\partial_1 g(x_i; \theta_0)\| \geq 1/n\}$ has probability 0 for all n, thus $\{\inf_{y \in B_\delta(x_i)} \|\partial_1 g(x_i; \theta_0)\| > 0\}$ has probability 0 for all δ by continuity from below, contradicting assumption 3.6 with continuity of $\partial_1 g$.

We now consider such a pair (η, δ), fix the resulting set A, and let A_s be the observations in sample that fall in it.

In order to get enough degrees of freedom to offset deviations of sample averages from 0, we make group of observations. Let $M \equiv dim(g(z_i; \theta_0))/dim(z_i)$, and assume for convenience it is an integer that divides $n - |A_s^c|$[3]. Without loss, let the x_i in A_s^c constitute the first $|A_s^c|$ observations and let $z_i^* = x_i$ for all $x_i \in A_s^c$. Then, for all $k \in \mathbb{N}$ (0 included) let $m_k \equiv \{|A_s^c| + Mk, \cdots, |A_s^c| + Mk + M - 1\}$ and solve wpa1 for z_i^* in $\sum_{i \in m_k}(g(z_i^*; \theta_0) - g(x_i; \theta)) = -M\frac{n}{|A_s|}\frac{1}{n}\sum_{i=1}^{n} g(x_i; \theta_0)$. By the LLN $\frac{1}{n}\sum_{i=1}^{n} g(x_i; \theta_0) \to^p 0$ and $|A_s|/n \to^p \mathbb{P}[A] > 0$ so that a sequence z_i^* with $z_i^* \to^p x_i$ will exist by continuity.

We also get $\sup_i \|z_i^* - x_i\| \to^p 0$ because $\sup_i \|z_i^* - x_i\| \leq \frac{\sup_i \|g(z_i^*; \theta) - g(x_i; \theta)\|}{\inf_{y \in A} \|\partial_1 g(y; \theta)\|} \leq \eta o_p(1)$.

By definition of z_i and the previous result, we have $\frac{1}{n}\sum_{i=1}^{n}||z_i - x_i||^2 \leq \frac{1}{n}\sum_{i=1}^{n}||z_i^* - x_i||^2 \leq \sup_i ||z_i^* - x_i||^2 \to^p 0$.

By properties of norms, Hölder continuity with exponent $\alpha \leq 1$, Cauchy-Schwartz, the LLN, and the previous convergence result

[3] If not, it suffices to set the remaining (components of) z_i^* to x_i and re-scale appropriately in what follows.

$$\begin{aligned}
\left\|\frac{1}{n}\sum_{i=1}^{n}\partial_1 g(z_i;\theta_0)(z_i-x_i)\right\| &\leq \frac{1}{n}\sum_{i=1}^{n}\|\partial_1 g(z_i;\theta_0)-\partial_1 g(x_i;\theta_0)\|\|z_i-x_i\| \\
&\quad+\frac{1}{n}\sum_{i=1}^{n}\|\partial_1 g(x_i;\theta_0)\|\|z_i-x_i\| \\
&\leq C\frac{1}{n}\sum_{i=1}^{n}\|z_i-x_i\|^{1+\alpha} \\
&\quad+\left(\frac{1}{n}\sum_{i=1}^{n}\|\partial_1 g(x_i;\theta_0)\|^2\right)^{1/2}\left(\frac{1}{n}\sum_{i=1}^{n}\|z_i-x_i\|^2\right)^{1/2} \\
&\to^p 0
\end{aligned}$$

Furthermore, proceeding component-wise with $(k\cdot)$ denoting the k^{th} row of a matrix and using assumption 3.7 together with previous results and proceeding as above for the term $\frac{1}{n}\sum_{i=1}^{n}\|[\partial_1 g(z_i;\theta_0)]_{k\cdot}\|\,\|z_i-x_i\|^{\alpha}$, we have

$$\begin{aligned}
&\left\|\frac{1}{n}\sum_{i=1}^{n}[\partial_1 g(z_i;\theta_0)]_{k\cdot}[\partial_1 g(z_i;\theta_0)]'_{j\cdot}-\frac{1}{n}\sum_{i=1}^{n}[\partial_1 g(x_i;\theta_0)]_{k\cdot}[\partial_1 g(x_i;\theta_0)]'_{j\cdot}\right\| \\
&\leq\left\|\frac{1}{n}\sum_{i=1}^{n}[\partial_1 g(z_i;\theta_0)]_{k\cdot}[\partial_1 g(z_i;\theta_0)]'_{j\cdot}-[\partial_1 g(z_i;\theta_0)]_{k\cdot}[\partial_1 g(x_i;\theta_0)]'_{j\cdot}\right\| \\
&+\left\|\frac{1}{n}\sum_{i=1}^{n}[\partial_1 g(z_i;\theta_0)]_{k\cdot}[\partial_1 g(x_i;\theta_0)]'_{j\cdot}-[\partial_1 g(x_i;\theta_0)]_{k\cdot}[\partial_1 g(x_i;\theta_0)]'_{j\cdot}\right\| \\
&=\left\|\frac{1}{n}\sum_{i=1}^{n}[\partial_1 g(z_i;\theta_0)]_{k\cdot}([\partial_1 g(z_i;\theta_0)]'_{j\cdot}-[\partial_1 g(x_i;\theta_0)]'_{j\cdot}\right\| \\
&+\left\|\frac{1}{n}\sum_{i=1}^{n}([\partial_1 g(z_i;\theta_0)]_{k\cdot}-[\partial_1 g(x_i;\theta_0)]_{k\cdot})[\partial_1 g(x_i;\theta_0)]'_{j\cdot}\right\| \\
&\leq C\frac{1}{n}\sum_{i=1}^{n}\|[\partial_1 g(z_i;\theta_0)]_{k\cdot}\|\,\|z_i-x_i\|^{\alpha} \\
&+C\left(\frac{1}{n}\sum_{i=1}^{n}\|z_i-x_i\|^{2\alpha}\right)^{1/2}\left(\frac{1}{n}\sum_{i=1}^{n}\|[\partial_1 g(x_i;\theta_0)]_{k\cdot}\|^2\right)^{1/2} \\
&\to^p 0+\left(\mathbb{E}[\|[\partial_1 g(x_i;\theta_0)]_{k\cdot}\|^2]\right)^{1/2}0=0
\end{aligned} \tag{30}$$

and thus

$$\begin{aligned}\|\lambda\| &= \left\| \left(\frac{1}{n} \sum_{i=1}^{n} \partial_1 g(z_i;\theta_0) \partial_1 g(z_i;\theta_0) \right)^{-1} \frac{1}{n} \sum_{i=1}^{n} \partial_1 g(z_i;\theta_0)(z_i - x_i) \right\| \\ &\leq \left\| \left(\frac{1}{n} \sum_{i=1}^{n} \partial_1 g(z_i;\theta_0) \partial_1 g(z_i;\theta_0)' \right)^{-1} \right\| \left\| \frac{1}{n} \sum_{i=1}^{n} \partial_1 g(z_i;\theta_0)(z_i - x_i) \right\| \\ &\to^p \left\| (\mathbb{E}[\partial_1 g(x_i;\theta_0) \partial_1 g(x_i;\theta_0)'])^{-1} \right\| \; 0 = 0\end{aligned}$$

Now we derive a precise rate of convergence and the resulting asymptotic distribution for λ.

Solving for z_i in equation (28) yields $z_i(\lambda)$, which can be plugged in the second equation to obtain

$$\sum_{i=1}^{n} g(z_i(\lambda);\theta) = 0 \tag{31}$$

By a Taylor expansion and assumption 3.8, we get

$$\frac{1}{n} \sum_{i=1}^{n} g(x_i,\theta) + \frac{1}{n} \sum_{i=1}^{n} \partial_1 g(x_i;\theta) \partial_1 g'(x_i;\theta) \lambda + O(||\lambda||^2) = 0 \tag{32}$$

Under assumptions 3.1, 3.3, and 3.5, $\frac{1}{\sqrt{n}} \sum_{i=1}^{n} g(x_i;\theta)$ converges in distribution to a normal random variables by the central limit theorem and thus the first term is $O_p(n^{-1/2})$.

Under assumptions 3.1 and 3.6 $\frac{1}{n} \sum_{i=1}^{n} \partial_1 g(x_i;\theta) \partial_1 g(x_i;\theta)' \to^p \mathbb{E}[\partial_1 g(x_i;\theta) \partial_1 g(x_i;\theta)']$ by the LLN and thus the second term is $O(\lambda)$. It follows that $\lambda = O_p(n^{1/2})$ with an asymptotically normal distribution.

Finally, we turn the situation where $\theta \neq \theta_0$. By the uniform Law of Large Numbers, using assumption 3.7, $\sup_{\theta \in \Theta} \| \frac{1}{n} \sum_{i=1}^{n} g(x_i;\theta) - \mathbb{E}[g(x_i;\theta)] \| \to^p 0$.

For any $\theta \in B_\varepsilon^c(\theta_0)$ we have by identification $\mathbb{E}[g(x_i;\theta)] \in B_\gamma^c(0)$ for some γ (otherwise, we can find a sequence whose mapping converges to 0 and by compactness there would be a convergent subsequence, implying existence of some $\theta^* \neq \theta_0$ that satisfies $\mathbb{E}[g(x_i;\theta^*)] = 0$).

With probability approaching one, we have by the mean value theorem and Cauchy-Schwartz $\frac{\gamma}{2} \leq \frac{1}{n} \sum_{i=1}^{n} \|g(z_i;\theta) - g(x_i;\theta)\| = \frac{1}{n} \sum_{i=1}^{n} \|g(\overline{z}_i;\theta)(z_i - x_i)\| \leq (\frac{1}{n} \sum_{i=1}^{n} \|g(\overline{z}_i;\theta)\|^2)^{1/2}$ $(\frac{1}{n} \sum_{i=1}^{n} \|z_i - x_i\|^2)^{1/2}$. As a result, $\frac{1}{n} \sum_{i=1}^{n} \|z_i - x_i\|^2 \to^p 0$ (or a subsequence) would imply $\frac{1}{n} \sum_{i=1}^{n} \|g(\overline{z}_i;\theta)\|^2 \to^p \mathbb{E}[\|g(x_i;\theta)\|]$ as before and thus $\gamma \leq o_p(1)$, which is impossible. Therefore, $\sum_{i=1}^{n} \|z_i - x_i\|^2 > O(n)$ with probability approaching one, and the probability that $\hat{\theta}$ lives outside any neighborhood of θ_0 decreases to 0.

Eventually, the first-order conditions read $z_i - x_i = \lambda' \partial_1 g(x_i;\theta_0) + o_p(n^{-1/2})$ and $\frac{1}{n} \sum_{i=1}^{n} g(z_i;\theta_0) = 0$ and the linearized version is asymptotically justified. ■

Proof of Theorem 3.13. We have by the first-order conditions, previous convergence results, assumption 3.11, and equation (32)

$$\begin{aligned}
F &\equiv \frac{1}{n}\sum_{i=1}^{n}\|z_i - x_i\|^2 \\
&= \lambda'\frac{1}{n}\sum_{i=1}^{n}\partial_1 g(z_i;\theta)\partial_1 g(z_i;\theta)'\lambda \\
&= \lambda'\frac{1}{n}\sum_{i=1}^{n}\partial_1 g(x_i;\theta)\partial_1 g(x_i;\theta)'\lambda + o_p(F) \\
&= \left(\left(\frac{1}{n}\sum_{i=1}^{n}\partial_1 g(x_i;\theta)\partial_1 g(x_i;\theta)'\right)^{-1}\frac{1}{n}\sum_{i=1}^{n}g(x_i;\theta)\right)'\frac{1}{n}\sum_{i=1}^{n}\partial_1 g(x_i;\theta)\partial_1 g(x_i;\theta)' \\
&\left(\left(\frac{1}{n}\sum_{i=1}^{n}\partial_1 g(x_i;\theta)\partial_1 g(x_i;\theta)'\right)^{-1}\frac{1}{n}\sum_{i=1}^{n}g(x_i;\theta)\right) + O_p(\|\lambda\|^3) + O(\|\lambda\|^4) + o_p(F)
\end{aligned}$$

Ignoring lower order terms, we can eventually reframe the problem as minimizing standard GMM:
$\sum_{i=1}^{n} g(x_i;\theta)'(\sum_{i=1}^{n}\partial_1 g(x_i;\theta_0)\partial_1 g'(x_i,\theta_0))^{-1}\sum_{i=1}^{n} g(x_i;\theta)$ to get the first-order conditions

$$\sum_{i=1}^{n}\partial_2 g(x_i;\theta)'\left(\sum_{i=1}^{n}\partial_1 g(x_i;\theta_0)\partial_1 g'(x_i,\theta_0)\right)^{-1}\sum_{i=1}^{n}g(x_i;\theta) = 0 \tag{33}$$

which are satisfied with probability approaching 1. By an expansion around θ_0 we have

$$\sum_{i=1}^{n}\partial_2 g(x_i;\theta)'\left(\sum_{i=1}^{n}\partial_1 g(x_i;\theta_0)\partial_1 g'(x_i,\theta_0)\right)^{-1}\sum_{i=1}^{n}[g(x_i;\theta_0) + \partial_2 g(x_i;\overline{\theta})(\theta-\theta_0)] = 0 \tag{34}$$

so that the estimator takes the form

$$\begin{aligned}
\hat{\theta}_{OTGMM} - \theta_0 = -&\left(\sum_{i=1}^{n}\partial_2 g(x_i;\hat{\theta})'\left(\sum_{i=1}^{n}\partial_1 g(x_i;\theta_0)\partial_1 g'(x_i,\theta_0)\right)^{-1}\sum_{i=1}^{n}\partial_2 g(x_i;\overline{\theta})\right)^{-1} \\
&\left(\sum_{i=1}^{n}\partial_2 g(x_i;\hat{\theta})'\left(\sum_{i=1}^{n}\partial_1 g(x_i;\theta_0)\partial_1 g'(x_i,\theta_0)\right)^{-1}\sum_{i=1}^{n}g(x_i;\theta_0)\right)
\end{aligned}$$

Noting that under assumptions 3.1, 3.3, and 3.5 $\frac{1}{\sqrt{n}}\sum_{i=1}^{n} g(x_i;\theta)$ converges in distribution to a normal random variables by the central limit theorem and that assumptions 3.10 and 3.12 together with consistency ensure convergence of sample averages to expectations, we obtain the asymptotic normality of $\sqrt{n}(\hat{\theta}_{OTGMM} - \theta_0)$ by Slutsky with asymptotic variance given in the theorem. ■

Proof of Theorem 3.15. The first-order conditions with respect to the z_j (Equation (5)) can be written as

$$x_j = z_j - \partial_1 g'(z_j, \theta)\lambda. \tag{35}$$

Under our assumptions, Equation (35) defines a direct relationship between z_j and x_j, and therefore an implicit reverse relationship between x_j and z_j. Since the latter may not be unique, we observe that our original optimization problem seeks to minimize the distance between x_j and z_j. Hence, in cases where (35) admits multiple solutions z_j for a given x_j, we identify the unique[4] solution that minimizes $\|z_j - x_j\|^2$. This is accomplished by defining the mapping (12).

With this definition, the first order conditions (3) and (4) of the Lagrangian optimization problem for θ and λ become, respectively,

$$\begin{aligned} \sum_i \partial_2 g'(q(x_j, \theta, \lambda), \theta)\lambda &= 0 \\ \sum_i g(q(x_j, \theta, \lambda), \theta) &= 0 \end{aligned}$$

This is a just-identified GMM estimator in terms of the modified moment function stated in the Theorem. ■

Lemma B.1 *Let $h(\cdot,\cdot,\cdot)$ be continuous in all of its arguments. Then, under Assumptions 3.18(i) and 3.19, $h(q(x,\theta,\lambda),\theta,\lambda)$ is continuous in (θ,λ).*

Proof of Lemma B.1. Since $h(z,\theta,\lambda)$ is assumed to be continuous in all of its arguments, there only remains to show that $q(x,\theta,\lambda)$ is continuous in (θ,λ). In fact, we can establish the stronger statement that $q(x,\theta,\lambda)$ is differentiable in (θ,λ). Differentiability in θ can be shown by the implicit function theorem

$$\begin{aligned} \partial_{2'} q(x,\theta,\lambda) &= \left[\left(\frac{\partial}{\partial z'}\left(z - \partial_1(\lambda' g(z,\theta))\right)\right)^{-1} \frac{\partial}{\partial \theta'}\partial_1\left(z - \lambda' g(z,\theta)\right)\right]_{z=q(x,\theta,\lambda)} \\ &= \left[\left(I - \partial_{11'}(\lambda' g(z,\theta))\right)^{-1} \partial_{12'}(\lambda' g(z,\theta))\right]_{z=q(x,\theta,\lambda)} \end{aligned}$$

since $q(x,\theta,\lambda)$ is the inverse of the mapping $z \mapsto z - \partial_1(\lambda' g(z,\theta))$. By the definition of $\bar{\lambda}$, $\bar{\nu}$,

$$\left\|\left(I - \partial_{11'}(\lambda' g(z,\theta))\right)^{-1} \partial_{12'}(\lambda' g(z,\theta))\right\| \leq \left(1 - \bar{\lambda}\bar{\nu}\right)^{-1} \left\|\partial_{12'}(\lambda' g(z,\theta))\right\|,$$

at $z = q(x,\theta,\lambda)$, where $\bar{\lambda}\bar{\nu} < 1$ by Assumption 3.19 and where $\partial_{12'}(\lambda' g(z,\theta))$ exists by Assumption 3.18. Thus $h(q(x,\theta,\lambda),\theta,\lambda)$ is continuous in θ.

[4]Having two solutions to the first order conditions that happen to have the same distance $\|z_j - x_j\|$ is an event of probability zero that can be safely neglected.

By a similar reasoning, we can show that $h\left(q\left(x,\theta,\lambda\right),\theta,\lambda\right)$ is continuous in λ if we can show that $\partial_{3'}q\left(x,\theta,\lambda\right)$ exists:

$$\begin{aligned}\left\|\partial_{3'}q\left(x,\theta,\lambda\right)\right\| &= \left\|\left[\left(\frac{\partial}{\partial z'}\left(z-\partial_1\left(\lambda' g\left(z,\theta\right)\right)\right)\right)^{-1}\frac{\partial}{\partial\lambda'}\left(\partial_1 g'\left(z,\theta\right)\lambda\right)\right]_{z=q(x,\theta,\lambda)}\right\| \\ &\leq \left(1-\bar{\lambda}\bar{\nu}\right)^{-1}\left\|\left[\partial_1 g'\left(z,\theta\right)\right]_{z=q(x,\theta,\lambda)}\right\|\end{aligned}$$

where $\bar{\lambda}\bar{\nu}<1$ by Assumption 3.19 and where $\partial_1 g'\left(z,\theta\right)$ exists by Assumption 3.18. ■

Lemma B.2 *Under Assumptions 3.16 and 3.19, if $h\in\mathcal{L}$, then*

$$\mathbb{E}\left[\sup_{(\theta,\lambda)\in\Theta\times\Lambda}\left\|h\left(q\left(x,\theta,\lambda\right),\theta\right)\right\|\right]<\infty$$

for $q\left(x,\theta,\lambda\right)$ defined in Theorem 3.15.

Proof of Lemma B.2. By the triangle inequality and Definition 3.20, there exists $\bar{h}\left(x,\theta\right)$ such that:

$$\begin{aligned}\left\|h\left(z,\theta\right)\right\| &\leq \left\|h\left(x,\theta\right)\right\|+\left\|h\left(z,\theta\right)-h\left(x,\theta\right)\right\| \\ &= \left\|h\left(x,\theta\right)\right\|+\bar{h}\left(x,\theta\right)\left\|z-x\right\|, \end{aligned}\tag{36}$$

for $z=q\left(x,\theta,\lambda\right)$. Next, using the first order conditions (Equation (5)), we have, by a mean value argument, the triangle inequality and the definitions of $\bar{\lambda}$ and $\bar{\nu}$ from Assumption 3.19,

$$\begin{aligned}\left\|z-x\right\| &= \left\|\partial_1\left(\lambda' g\left(z,\theta\right)\right)\right\| \\ &\leq \left\|\partial_1\left(\lambda' g\left(x,\theta\right)\right)\right\|+\left\|\partial_{11'}\left(\lambda' g\left(\tilde{x},\theta\right)\right)\left(z-x\right)\right\| \text{ for some mean value } \tilde{x} \\ &\leq \bar{\lambda}\left\|\partial_{1'}g\left(x,\theta\right)\right\|+\bar{\lambda}\bar{\nu}\left\|z-x\right\|. \end{aligned}\tag{37}$$

Re-arranging and using the fact that $\bar{\lambda}\bar{\nu}<1$ by Assumption 3.19 and $\bar{\lambda}<\infty$ by compactness of Λ,

$$\left\|z-x\right\|\leq\frac{\bar{\lambda}\left\|\partial_{1'}g\left(x,\theta\right)\right\|}{\left(1-\bar{\lambda}\bar{\nu}\right)}.\tag{38}$$

Combining (36) and (38) and noting that applying the $\mathbb{E}\left[\sup_{\theta\in\Theta}\ldots\right]$ operator does not alter the inequalities, we have

$$\mathbb{E}\left[\sup_{\theta\in\Theta}\left\|h\left(z,\theta\right)\right\|\right]\leq\mathbb{E}\left[\sup_{\theta\in\Theta}\left\|h\left(x,\theta\right)\right\|\right]+\frac{\bar{\lambda}}{\left(1-\bar{\lambda}\bar{\nu}\right)}\mathbb{E}\left[\sup_{\theta\in\Theta}\bar{h}\left(x,\theta\right)\left\|\partial_{1'}g\left(x,\theta\right)\right\|\right]$$

where the right-hand side quantities are finite by construction since $h\in\mathcal{L}$. ■

Proof of Theorem 3.22. Assumptions 3.1-3.17 directly imply consistency of our GMM estimator, by Theorem 2.6 in Newey and McFadden (1994). There remains to show that Assumption 3.17 is implied by Assumptions 3.18,3.19,3.21.

We first establish Assumption 3.17 (i): Continuity of $\tilde{g}(x,\theta,\lambda)$ in (θ,λ). To show that $g(q(x,\theta,\lambda),\theta)$ is continuous in (θ,λ), we can invoke Lemma B.1 for $h(z,\theta,\lambda) = g(z,\theta)$, under Assumptions 3.18(i) and 3.19. To show that $\partial_2 g'(q(x,\theta,\lambda),\theta)\lambda$ is continuous in (θ,λ), we can similarly invoke Lemma B.1 for $h(z,\theta,\lambda) = \partial_2 g'(z,\theta)\lambda$, where $\partial_2 g'(z,\theta)$ is continuous in both arguments by Assumption 3.18(ii).

We now establish Assumption 3.17 (ii): $\mathbb{E}\left[\sup_{(\theta,\lambda)\in\Theta\times\Lambda}\|\tilde{g}(x,\theta,\lambda)\|\right] < \infty$. Since $g(\cdot,\cdot) \in \mathcal{L}$ by Assumption 3.21, it follows that $\mathbb{E}\left[\sup_{(\theta,\lambda)\in\Theta\times\Lambda}\|g(q(x,\theta,\lambda),\theta)\|\right] < \infty$, by Lemma B.2. Next, we have, for $(\theta,\lambda)\in\Theta\times\Lambda$, $\|\partial_2 g'(q(x,\theta,\lambda),\theta)\ \lambda\| \leq \|\partial_2 g'(q(x,\theta,\lambda),\theta)\|\,\|\lambda\| \leq \|\partial_2 g'(q(x,\theta,\lambda),\theta)\|\,\bar{\lambda}$ by Assumption 3.19 and compactness of Λ. By Assumption 3.21 and Lemma B.2 we then also have that $\mathbb{E}\left[\sup_{(\theta,\lambda)\in\Theta\times\Lambda}\|\partial_2 g'(q(x,\theta,\lambda),\theta)\ \lambda\|\right] < \infty$. ■

Proof of Theorem 3.27. Theorem 3.22 implies consistency $(\theta,\lambda) \xrightarrow{p} (\theta_0,\lambda_0)$. This, in addition to Assumptions 3.23, 3.24 and 3.25 directly implies the stated asymptotic normality result, by Theorem 3.2 and Lemma 2.4 in Newey and McFadden (1994) and the Lindeberg-Levy Central Limit Theorem. There remains to show that Assumption 3.25 is implied by Assumption 3.26.

By Lemma B.1, Assumptions 3.26(i) and (iii) imply that both $g(q(x,\theta,\lambda),\lambda)$ and $\partial_2 g'(q(x,\theta,\lambda),\lambda)\lambda$ are continuously differentiable in (θ,λ), thus establishing Assumption 3.25(i).

By Lemma B.2, Assumptions 3.16, 3.26(ii) and (iii) imply Assumption 3.25(ii).

The asymptotic variance of the just-identified GMM estimator defined in Theorem 3.15 is then given by $\left(\tilde{G}'\Omega^{-1}\tilde{G}\right)^{-1}$ where

$$\begin{aligned}
\Omega &= \mathbb{E}\left[\tilde{g}(x_j,\theta,\lambda)\,\tilde{g}'(x_j,\theta,\lambda)\right] \\
\tilde{G} &= \mathbb{E}\left[\partial_{2'}\tilde{g}(x_j,\theta,\lambda)\right] = \mathbb{E}\left[\begin{array}{cc}\partial_{2'}\tilde{g}(x_j,\theta,\lambda) & \partial_{3'}\tilde{g}(x_j,\theta,\lambda)\end{array}\right] \equiv \left[\begin{array}{cc}\tilde{G}_{\theta\theta} & \tilde{G}_{\theta\lambda} \\ \tilde{G}_{\lambda\theta} & \tilde{G}_{\lambda\lambda}\end{array}\right]. \qquad (39)
\end{aligned}$$

where

$$\begin{aligned}
\tilde{G}_{\theta\theta} &= \mathbb{E}\left[\partial_{22'}\left(\lambda' g(q(x,\theta,\lambda),\theta)\right) + \partial_{21'}\left(\lambda' g(q(x,\theta,\lambda),\theta)\right)\partial_{2'} q(x,\theta,\lambda)\right] \\
\tilde{G}_{\lambda\theta} &= \mathbb{E}\left[\partial_{2'} g(q(x,\theta,\lambda),\theta) + \partial_{1'} g(q(x,\theta,\lambda),\theta)\,\partial_{2'} q(x,\theta,\lambda)\right] \\
\tilde{G}_{\theta\lambda} &= \mathbb{E}\left[\partial_2\left(g'(q(x,\theta,\lambda),\theta)\right) + \partial_{21'}\left(\lambda' g(q(x,\theta,\lambda),\theta)\right)\partial_{3'} q(x,\theta,\lambda)\right] \\
\tilde{G}_{\lambda\lambda} &= \mathbb{E}\left[\partial_{1'} g(q(x,\theta,\lambda),\theta)\,\partial_{3'} q(x,\theta,\lambda)\right]
\end{aligned}$$

where expressions of the form $\partial_{ij'}\left(\lambda' g(q(x,\theta,\lambda),\theta)\right)$ represent partial second derivatives of the scalar-valued function $\lambda' g(z,\theta)$ with respect to its i^{th} and j^{th} argument evaluated at $z = q(x,\theta,\lambda)$.

Finally, the explicit expressions for the derivatives of the function $z = q(x,\theta,\lambda)$ follow from the implicit function theorem after noting that $q(x,\theta,\lambda)$ is the inverse of the mapping $z \mapsto z - \partial_1\left(\lambda' g(z,\theta)\right)$. This can also be shown through an explicit calculation: To first order, (35) implies, for a small change $\Delta\theta$ in θ, a corresponding change Δz in z while keeping x and λ fixed, that:

$$0 = \Delta z - \partial_{11'}\left(\lambda' g(z,\theta)\right)\Delta z - \partial_{12'}\left(\lambda' g(z,\theta)\right)\Delta\theta.$$

Thus,

$$\Delta z = \left(I - \partial_{11'}\left(\lambda' g\left(z,\theta\right)\right)\right)^{-1} \partial_{12'}\left(\lambda' g\left(z,\theta\right)\right) \Delta\theta$$

and we have:

$$\partial_{2'} q\left(x,\theta,\lambda\right) = \left(I - \partial_{11'}\left(\lambda' g\left(z,\theta\right)\right)\right)^{-1} \partial_{12'}\left(\lambda' g\left(z,\theta\right)\right) \tag{40}$$

evaluated at $z = q\left(x,\theta,\lambda\right)$. A similar reasoning for λ and exploiting the fact that $\frac{\partial^2(\lambda' g(z,\theta))}{\partial z \partial \lambda'} = \frac{\partial g'(z,\theta)}{\partial z}$, yields:

$$\partial_{3'} q\left(x,\theta,\lambda\right) = \left(I - \partial_{11'}\left(\lambda' g\left(z,\theta\right)\right)\right)^{-1} \partial_1 g'\left(z,\theta\right). \tag{41}$$

Collecting (40), (41) (39) and its subblocks yields the expressions for $\tilde{G}$ in the statement of the theorem. ■