Optimally-Transported Generalized Method of Moments

Susanne Schennach*† and Vincent Starck‡§
Brown University, LMU München

October 1, 2025

Abstract

We propose a novel optimal transport-based version of the Generalized Method of Moment (GMM). Instead of handling overidentification by reweighting the data to satisfy the moment conditions (as in Generalized Empirical Likelihood methods), this method proceeds by allowing for errors in the variables of the least mean-square magnitude necessary to simultaneously satisfy all moment conditions. This approach, based on the notions of optimal transport and Wasserstein metric, 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. We illustrate the method by revisiting Duranton, Morrow and Turner's (2014) study of the relationship between a city's exports and the extent of its transportation infrastructure. Our results corroborate theirs under weaker assumptions and provide insight into the error structure of the variables.

Keywords: Wasserstein metric, GMM, overidentification, misspecification.

*Support from NSF grants SES-1950969 and SES-2150003 is gratefully acknowledged. The authors would like to thank Alfred Galichon, Florian Gunsilius, Heejun Lee and seminar participants at NYU's Advanced mathematical modeling in economics seminar, at the Econometrics and Optimal Transport Workshop at the University of Washington, at the Women in Econometrics Conference at the University of Toronto, and at the Workshop on Optimal Transport in Econometrics at Collegio Carlo Alberto for useful comments.

†smschenn@brown.edu

‡Financial support from the European Research Council (Starting Grant No. 852332) is gratefully acknowledged.

§V.Starck@lmu.de

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

GEL implicitly attributes the mismatch in the moment conditions solely to 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 errors in the variables (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 errors in the variables instead of sampling bias. We employ the generic term *error* to include, not only measurement error, but anything that could cause the recorded data to differ from the value they should have if the model were fully correct, i.e., this could include some model errors. More generally, we allow distortions in the data generating process whose magnitude is quantified by a Wasserstein-type metric (Villani (2009)), in the spirit of distributionally robust methods (Christensen and Connault (2022), Blanchet, Murthy, and Si (2022)). In analogy with GEL, which does not require the form of the sampling bias to be explicitly specified, the 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. Of course, the accuracy of the resulting estimated parameters will typically improve with the degree of overidentification.[1]

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), Ekeland, Galichon, and Henry (2011), Chernozhukov, Galichon, Hallin, and Henry (2017), Gunsilius and Schennach (2021)). The idea is to find the parameter value that minimizes cost of "transporting" the observed distribution of the data μ_x onto another distribution μ_z that satisfies the moment conditions exactly. 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 a parameter value θ in some set Θ and some given d_g-dimensional vector $g\left(z_i,\theta\right)$ of moment functions. However, we instead observe an error-contaminated counterpart x_i of the true vector z_i (both taking value in $\mathcal{X}\subseteq\mathbb{R}^{d_x}$). We seek to exploit the model's over-identification to gain information regarding the error in x_i. The Euclidean norm $\|(z-x)\|$ is chosen here for computational convenience, although one could imagine a whole class of related estimators obtained with different choices of metric. Our focus on Euclidean norms parallels the choice made in common estimators (e.g. least squares regressions, classical minimum distance and even GMM). Considering a *weighted* Euclidean norm can be useful to indicate the relative expected error magnitudes along different dimensions of x.

Given a probability measure μ_x for the random variable x, this setup suggests solving the following population optimization problem, for a given θ:

$$\min_{\mu_{zx}}\mathbb{E}_{\mu_{zx}}\left[\|z-x\|^2\right] \tag{1}$$

subject to μ_{zx}, supported on $\mathcal{X}\times\mathcal{X}$, having marginal μ_x and $\mathbb{E}_{\mu_{zx}}\left[g\left(z,\theta\right)\right]=0$, where $\mathbb{E}_\mu$ denotes an expectation under the measure μ. (This problem is guaranteed to have a solution if there exists at least one measure μ_z^* such that $\mathbb{E}_{\mu_z^*}\left[g\left(z,\theta\right)\right]=0$.) This setup covers the most general case, including both discrete and continuous variables, and can be handled using linear programming techniques (e.g., Santambrogio (2015), Section 6.4.1), after observing that the moment constraint is easy to incorporate since it is linear in the probability measure. However, we shall focus on the purely continuous case in the remainder of this paper, because it enables us to express the

[1] In the case where the statistical properties of the errors are in fact known a priori, other methods may be more appropriate (e.g. Schennach (2004), Schennach (2014)), Schennach (2016), Schennach (2020) and references therein.

main ideas more transparently. The fully continuous case indeed admits a convenient treatment, under the following regularity condition:

Assumption 1.1 *The marginals μ_z (arising from the solution μ_{zx} at each $\theta \in \Theta$) and μ_x have finite variance and μ_x is absolutely continuous with respect to the Lebesgue measure.*

Under this condition, by Theorem 1.22 in Santambrogio (2015), there exists a unique μ_{zx} implied by a deterministic transport map $z = q(x)$ that solves the constrained optimization problem (1) and yielding a transport cost $\mathbb{E}_{\mu_x}\left[\|q(x) - x\|^2\right]$. Since determining the function q amounts to finding which value z each point x should be mapped to, the sample version of this problem can be stated as

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

subject to:

$$\hat{\mathbb{E}}\left[g\left(z, \theta\right)\right] = 0, \tag{3}$$

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 and $a(\cdot)$ a given function). 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.

Our approach is conceptually similar to GEL, in that it minimizes some concept of distributional distance under moment constraints. Yet, the notion of distance used differs significantly. 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, since GEL's goal of satisfying all the moment conditions is achieved through optimal transport instead of optimal reweighting. (The distinction from GEL applies to the discrete case as well, since the OTGMM objective function depends on both the amount and location of probability mass transfers, while the GEL objective function is only sensitive on the amount, but not on the specific locations, of the probability mass transfers.) Most of our regularity conditions will parallel those of optimal GMM and GEL, but some will not (they are tied to the optimal transport nature of the problem and involve assumptions regarding higher order derivatives). In analogy with the behavior of GEL estimators, the OTGMM estimator will be shown

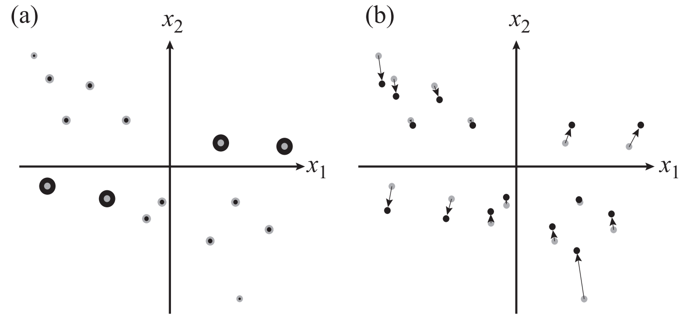

Figure 1: Comparison between the sample points adjustments for (a) Generalized Empirical Likelihood (GEL), where observation weights (shown by point size) are adjusted, and (b) Optimal Transport GMM, where point positions are adjusted. Simple case of overidentified (parameter-free) model imposing no correlation shown, with original sample in gray and adjusted sample in black.

to be root-n consistent and asymptotically normal, despite involving an optimization problem having an infinite-dimensional nuisance parameter (the z_i). In general, however, OTGMM's asymptotic variance does not coincide with that of GEL or efficiently weighted GMM under correct specification. Hence, OTGMM is most useful when errors in the variables or Wasserstein-type deviations in the data distribution constitute a primary concern.

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 errors (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. We show asymptotic normality and root n consistency in both cases. We then discuss related approaches and some extensions. The method's practical usefulness is illustrated by revisiting an influential study of the relationship between a city's exports and the extent of its transportation infrastructure (Duranton, Morrow, and Turner (2014)). Our results corroborate that study under weaker assumptions and provide insight into the error structure of the variables.

2 The estimator

2.1 Definition

The Lagrangian associated with the constrained optimization problem defined in Equations (2) and (3) is

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

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

$$\hat{\mathbb{E}}\left[\partial_\theta g'(z,\theta)\right]\lambda = 0 \tag{4}$$

$$\hat{\mathbb{E}}\left[g(z,\theta)\right] = 0 \tag{5}$$

$$(z_j-x_j)-\partial_z g'(z_j,\theta)\lambda = 0 \text{ for } j=1,\ldots,n \tag{6}$$

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

2.2 Implementation

The nonlinear system (4)-(6) 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 minimized to estimate θ. Let z_j^t and λ^t denote the approximations obtained after t steps. As shown in Supplement S.1.1, 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 $\lambda^{t+1}=\left(\hat{\mathbb{E}}\left[H\left(z^t,\theta\right)H'\left(z^t,\theta\right)\right]\right)^{-1}\left(-\hat{\mathbb{E}}\left[g\left(z^t,\theta\right)\right]+\hat{\mathbb{E}}\left[H\left(z^t,\theta\right)\left(z^t-x\right)\right]\right)$ and $z_j^{t+1}=x_j+H'\left(z_j^t,\theta\right)\lambda^{t+1}$, where $H\left(z,\theta\right)=\partial_{z'}g\left(z,\theta\right)=\left(\partial_z g'\left(z,\theta\right)\right)'$.

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

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 $z_j=x_j+H'\left(z_j,\theta\right)\lambda$ obtained from Equation (6) into Equation (5) 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 (6). Formal sufficient conditions for the convergence of this iterative procedure can be found in Supplement S.1.2. In cases where this simple approach fails to converge, one can employ robust schemes based on a combination of discretization and linear programming (see Section 6.4.1 in Santambrogio (2015)).

To gain some intuition regarding the estimator, it is useful to consider the limit of small errors when solving Equations (2)-(3), in the spirit of Chesher (1991)). This limit corresponds to assuming that higher-order powers of $\|z_i - x_i\|$ are negligible relative to $\|z_i - x_i\|$ itself. In this limit, the estimator admits a closed form with an intuitive interpretation, as shown by the following result, shown in Appendix A.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 specific choice of weighting matrix:*

$$\hat{\theta} = \arg\min_{\theta} \hat{\mathbb{E}}\left[g'\left(x,\theta\right)\right]\left(\hat{\mathbb{E}}\left[H\left(x,\theta\right)H'\left(x,\theta\right)\right]\right)^{-1}\hat{\mathbb{E}}\left[g\left(x,\theta\right)\right]. \tag{7}$$

From this expression, it is clear that the estimator downweights the moments that are the most sensitive to errors in x, as measured by $H\left(x,\theta\right) \equiv \partial_{z'} g\left(x,\theta\right)$. This accomplishes the desired goal of minimizing the effect of the errors when the properties of the error process are unknown.

Although this weighting matrix appears suboptimal (relative to a correctly specified optimally weighted GMM estimator), one should realize that the notion of optimality depends on what class of data generating processes the "true model" encompasses. Optimal GMM and GEL can be seen as Maximum Likelihood estimators (Chamberlain (1987), Newey and Smith (2004)) under moment conditions expressed in terms of the observed x. In contrast, OTGMM can be interpreted as a Maximum Likelihood estimator for homoskedastic and normally distributed errors $(x - z)$ under moment constraints on the unobserved z. There is therefore a clear efficiency-robustness trade-off: OTGMM is less efficient if the observed x satisfy the moment constraints, but allows for additional error terms that maintain the model's correct specification even if the observed x do not satisfy the overidentified moment constraints, a more general setting where optimally weighted GMM or GEL offers no efficiency guarantees.

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 parallels 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 efficient GMM when one is concerned about overidentification test failure. 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, potentially leading to an increase in variance if the model happened to be correctly specified. Conversely, if the model is misspecified, e.g., because the data is error-contaminated, the optimal weighting of efficient GMM is no longer the optimal weighting (since random deviations due to sampling variability are not the main reason for the failure to simultaneously satisfy all moment conditions). For instance, if the errors are such that there is an unknown bias in the moment conditions that decays to zero asymptotically but at a rate possibly slower than $n^{-1/2}$, then the model is still correctly specified asymptotically but the bias dominates the random sampling error. Then, the optimal weighting should seek to minimize the effect of error-induced bias, which our approach seeks to accomplish by weighting based on the effect of errors in the variables on the moment conditions. Hence, in that sense, the method provides a complementary alternative to standard GMM estimation offering a different trade-off between efficiency and robustness to misspecification.

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

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

Assumption 3.2 $\mathbb{E}[g(x_i;\theta_0)] = 0$*, and* $\mathbb{E}[g(x_i;\theta)] \neq 0$ *for 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, in analogy with the setup traditionally considered in the GEL literature (e.g. Newey and Smith (2004)).

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* $\mathbb{E}[h(x_i)] < \infty$*.*

While 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)), our estimator requires a few more low-level regularity conditions. Given that our estimator, in the small error limit (Equation (7)), involves a sample average involving derivative $H(x,\theta) \equiv \partial_{z'} g(x,\theta)$, we need to place some constraints on the behavior of that quantity as well. Below, we let $\|a\| = \left(\sum_{i,j} a_{i,j}^2\right)^{1/2}$ for a matrix a.

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

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

Assumption 3.7 $\mathbb{E}[\partial_{z'} g(x_i;\theta_0)\partial_z g'(x_i;\theta_0)]$ *exists and is of full rank.*

These assumptions ensure that the minimization problem defined by (2) and (3) is well-behaved, i.e., small changes in the values of x_i do not lead to jumps in the solution z_i to the optimization problem (aside from zero-probability events). 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 3.8 *Under assumptions 3.1-3.7, 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 3.9 *$\theta_0 \in \Theta^\circ$, the interior of Θ.*

Assumption 3.10 *$\mathbb{E}[\sup_{\theta\in\eta} \|\partial_{\theta'} g(x_i;\theta)\|] < \infty$ where $\eta \subset \Theta$ is a neighborhood of θ_0.*

Assumption 3.11 *$\left(\mathbb{E}[\partial_{\theta'} g(x_i;\theta_0)'] \left(\mathbb{E}[\partial_{z'} g(x_i;\theta_0)\partial_{z'} g(x_i,\theta_0)']\right)^{-1} \mathbb{E}[\partial_{\theta'} g(x_i;\theta_0)]\right)$ is invertible.*

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

Theorem 3.12 *Under Assumptions 3.1-3.11, the OTGMM estimator is asymptotically normal with $\sqrt{n}(\hat{\theta}_{OTGMM} - \theta_0) \to^d \mathcal{N}(0;V)$, where*

$$\begin{aligned} V \;=\; & \left(\mathbb{E}[G_i'] \left(\mathbb{E}[H_i H_i']\right)^{-1} \mathbb{E}[G_i]\right)^{-1} \times \left(\mathbb{E}[G_i'] \left(\mathbb{E}[H_i H_i']\right)^{-1} \mathbb{E}[g_i g_i'] \left(\mathbb{E}[H_i H_i']\right)^{-1} \mathbb{E}[G_i]\right) \times \\ & \left(\mathbb{E}[G_i'] \left(\mathbb{E}[H_i H_i']\right)^{-1} \mathbb{E}[G_i]\right)^{-1}, \end{aligned}$$

where $H_i \equiv \partial_{z'} g(x_i;\theta_0)$, $G_i \equiv \partial_{\theta'} g(x_i;\theta_0)$ and $g_i \equiv g(x_i;\theta_0)$.

This simple normal limiting distribution with root n convergent behavior is somewhat unexpected from an estimator that involves a high-dimensional optimization over $O(n)$ latent variables. As in GEL, this is made possible thanks to the existence of an equivalent low-dimensional dual optimization problem, which is, in turn, equivalent to a simple GMM estimator, albeit with a nonstandard weighting matrix. Thus, the variance has the expected "sandwich" form, since the reciprocal weights $\mathbb{E}[H_i H_i']$ differs from the moment variance $\mathbb{E}[g_i g_i']$. For comparison, an optimally weighted GMM estimator would have an asymptotic variance of $(\mathbb{E}[G_i]'(\mathbb{E}[g_i g_i'])^{-1}\mathbb{E}[G_i])^{-1}$.

It is difficult to formulate a simple expression for the difference $V_{OTGMM} - V_{GMM}$ between the asymptotic variance of OTGMM and that of optimal GMM, but a simple example suffices to illustrate that this difference could be any non-negative-definite matrix.

Let x_i take value in $\mathbb{R}^{d_x}$ ($d_x \geq 2$) with $\mathbb{E}[x_{i\ell}] = 0$ and $\mathbb{E}[x_{i\ell} x_{i\ell'}] \equiv \omega_\ell$ when $\ell = \ell'$ and zero otherwise, for $\ell, \ell' = 1, \ldots, d_x$. For $\theta \in \mathbb{R}$, let the moment functions be given by $g_\ell(x_i, \theta) = x_{i\ell} - \theta$. We then have

$$V_{OTGMM} = \frac{1}{d_x^2} \sum_{\ell=1}^{d_x} \omega_\ell \text{ and } V_{GMM} = \left(\sum_{\ell=1}^{d_x} \omega_\ell^{-1} \right)^{-1}.$$

Then, by the harmonic-arithmetic mean inequality, $V_{OTGMM} \geq V_{GMM}$ with equality when $\omega_\ell = c$ for all ℓ. If $\omega_1 \longrightarrow \infty$ leaving all the other ω_ℓ finite, $V_{OTGMM} \longrightarrow \infty$ while V_{GMM} remains finite. (General non-negative-definite differences can be obtained by stacking such moment conditions for different parameters θ_k ($k = 1, \ldots, d_\theta$) and possibly linearly transforming the resulting parameter vector θ.) Generally, we expect the difference to be large when some moments have a disproportionally large variance while not being disproportionally sensitive to changes in the underlying variables.

3.2 Asymptotics under large errors

In some applications, there may be considerable misspecification or its magnitude may be a priori unknown. It thus proves useful to relax the assumption of small errors in deriving the estimator's asymptotic properties. To handle this more general setup, we employ the following equivalence, demonstrated in Appendix A.2.

Theorem 3.13 *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*

$$\tilde{g}(x, \theta, \lambda) = \begin{bmatrix} \partial_\theta g'(q(x, \theta, \lambda), \theta) \;\; \lambda \\ g(q(x, \theta, \lambda), \theta) \end{bmatrix} \tag{8}$$

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

$$q(x, \theta, \lambda) \equiv \arg \min_{z: z - \partial_z g'(z, \theta)\lambda = x} \|z - x\|^2. \tag{9}$$

Note that $q(x, \theta, \lambda)$ is essentially the inverse of the mapping $z - \partial_z g'(z, \theta)\lambda = x$ (from Equation (6)), augmented with a rule to select the appropriate branch in case the inverse is multivalued.

The equivalence result of Theorem 3.13 implies that many of the 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 or in terms of regularity conditions drawn from optimal transport theory.

We start with a standard GMM-like identification condition:

Assumption 3.14 *For some compact sets Θ and Λ, there exists a unique $(\theta_0, \lambda_0) \in \Theta \times \Lambda$ solving $\mathbb{E}[\tilde{g}(x, \theta, \lambda)] = 0$ for $\tilde{g}(x, \theta, \lambda)$ defined in Theorem 3.13.*

This condition is implied by a natural uniqueness and regularity condition on the solution to the primal optimal transportation problem (Equations (1)):

Assumption 3.15 *Let $\mu_{zx;\theta}$ denote the solution to Problem (1) for a given $\theta \in \Theta$. (i) $\mathbb{E}_{\mu_{zx;\theta}}[\|z - x\|^2]$ is uniquely minimized at $\theta = \theta_0$ (ii) The corresponding marginals $\mu_{z;\theta}$ and μ_x are absolutely continuous with respect to the Lebesgue measure with a density that is finite, nonvanishing and Hölder continuous on their convex support.*

Indeed, by Theorem [C3], part b) and d), in Caffarelli (1996), Assumption 3.15(ii) implies that, at each θ, there exists a unique invertible transport map $z = q(x)$ from μ_x to $\mu_{z;\theta}$ and both q and its inverse are equal to the gradient of a twice differentiable strictly convex function. The fact that q^{-1} is the gradient of a twice differentiable strictly convex function ensures that the first-order condition in Assumption 3.14 has a unique solution (making a rule to handle multivalued inverses unnecessary).

Next, we consider standard continuity and dominance conditions that are used to establish uniform convergence of the GMM objective function. These assumptions constitute a superset of those needed for standard GMM because the modified moment conditions include the additional parameter λ and higher-order derivatives of the original moment conditions. In a high-level form, these conditions read:

Assumption 3.16 *(i) $\tilde{g}(x, \theta, \lambda)$ is continuous in θ and λ for $(\theta, \lambda) \in \Theta \times \Lambda$ with probability one and (ii) $\mathbb{E}\left[\sup_{(\theta,\lambda)\in\Theta\times\Lambda} \|\tilde{g}(x, \theta, \lambda)\|\right] < \infty$.*

Alternatively, Assumption 3.16 can be replaced by more primitive conditions on $g(z, \theta)$ instead, as given below in Assumptions 3.17, 3.18 and 3.20.

Assumption 3.17 *(i) $g(z, \theta)$ and $\partial_{z'} g(z, \theta)$ are differentiable in θ and (ii) $\partial_{\theta'} g(z, \theta)$ is continuous in both arguments.*

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

Assumption 3.18 *$\bar{\nu}\bar{\lambda} < 1$ where $\bar{\lambda} = \max_{\lambda\in\Lambda}\|\lambda\|$ and $\bar{\nu} = \sup_{\theta\in\Theta}\sup_{z\in\mathcal{X}}\max_{k\in\{1,\ldots,d_g\}}$ $\max\operatorname{eigval}(\partial_{zz'}g_k(z,\theta))$, in which $\partial_{zz'}g_k(z,\theta)$ exists for $k = 1,\ldots,d_g$ and where $\operatorname{eigval}(M)$ for some matrix M denotes the set of its eigenvalues.*

Once again, this condition can be alternatively phrased in terms of optimal transport concepts. The first order condition which implicitly defines $z = q(x,\theta,\lambda)$ can be written in terms of the derivative of a *potential function* $\psi(z,\theta,\lambda) = z'z/2 - g'(z,\theta)\lambda$: $\nabla_z\psi(z,\theta,\lambda) = x$. With the help of Theorem [C3], part b) and d), in Caffarelli (1996), Assumption 3.15(ii) implies that, at each θ, the above potential $\psi(z,\theta,\lambda)$ has a positive-definite Hessian (with respect to z), which implies Assumption 3.18.

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

Definition 3.19 *Let $\mathcal{L}$ be the set of functions $h(z,\theta)$ such that (i) $\mathbb{E}[\sup_{\theta\in\Theta}\|h(x,\theta)\|] < \infty$ and (ii) there exists a function $\bar{h}(x,\theta)$ satisfying*

$$\mathbb{E}\left[\sup_{\theta\in\Theta}\bar{h}(x,\theta)\|\partial_{z'}g(x,\theta)\|\right] < \infty. \tag{10}$$

$$\|h(z,\theta) - h(x,\theta)\| \leq \bar{h}(x,\theta)\|z - x\| \tag{11}$$

for all $x, z \in \mathcal{X}$ and $\theta \in \Theta$, and where $g(x,\theta)$ is as in the moment conditions.

This Lipschitz continuity-type assumption has no parallel in conventional GMM. It 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 3.19 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 (11)). 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 (10) strengthens the dominance condition 3.19(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. $\partial_{z'} g(x,\theta)$).

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

Assumption 3.20 *$g(\cdot,\cdot)$ and each element of $\partial_\theta g'(\cdot,\cdot)$ belong to $\mathcal{L}$.*

We are now ready to state our general consistency result.

Theorem 3.21 *Under Assumptions 3.1, 3.14 and either Assumption 3.16 or Assumptions 3.17, 3.18, 3.20, 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 3.22 *(θ_0,λ_0) from Assumption 3.14 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 3.23 *(i) $\mathbb{V}[\tilde{g}(x,\theta_0,\lambda_0)] \equiv \Omega$ exists and (ii) $\mathbb{E}[\partial\tilde{g}(x,\theta,\lambda)/\partial(\theta',\lambda')] \equiv \tilde{G}$ exists and is nonsingular.*

Assumption 3.23(ii) can be expressed in a more primitive fashion using the explicit form for $\tilde{G}$ provided in Theorem 3.26 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 3.24 *(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 3.25 *(i) $g(z,\theta)$ and $\partial_\theta g(z,\theta)$ are continuously differentiable in θ, (ii) all elements of $\partial_\theta g_k(z,\theta)$ and $\partial_{\theta\theta'} g_k(z,\theta)$ for $k = 1,\ldots,d_g$ belong to $\mathcal{L}$ and (iii) Assumptions 3.17(i) and 3.18 hold.*

We can now state our general asymptotic normality and root-n consistency result, shown in Appendix A.2.

Theorem 3.26 *Let the assumptions of Theorem 3.21 hold as well as Assumptions 3.22, 3.23 and either Assumption 3.24 or 3.25. 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 $W = \tilde{G}'\Omega^{-1}\tilde{G}$, $\Omega = \mathbb{E}\left[\tilde{g}\tilde{g}'\right]$,

$$\tilde{g} \equiv \tilde{g}(z,(\theta_0,\lambda_0)) = \begin{bmatrix}\partial_\theta g'(z,\theta_0)\,\lambda_0\\ g(z,\theta_0)\end{bmatrix} \text{ and } \tilde{G} \equiv \mathbb{E}[\partial_\theta \tilde{g}'] = \begin{bmatrix}\tilde{G}_{\theta\theta} & \tilde{G}_{\theta\lambda}\\ \tilde{G}_{\lambda\theta} & \tilde{G}_{\lambda\lambda}\end{bmatrix}$$

in which

$$\begin{aligned}
\tilde{G}_{\theta\theta} &\equiv \mathbb{E}\left[\partial_{\theta\theta'}\left(\lambda_0' g(z,\theta_0)\right) + \partial_{\theta z'}\left(\lambda_0' g(z,\theta_0)\right)\partial_{\theta'} q(x,\theta_0,\lambda_0)\right]\\
\tilde{G}_{\lambda\theta} &\equiv \mathbb{E}\left[\partial_{\theta'} g(z,\theta_0) + \partial_{z'} g(z,\theta_0)\,\partial_{\theta'} q(x,\theta_0,\lambda_0)\right]\\
\tilde{G}_{\theta\lambda} &\equiv \mathbb{E}\left[\partial_\theta\left(g'(z,\theta_0)\right) + \partial_{\theta z'}\left(\lambda_0' g(z,\theta_0)\right)\partial_{\lambda'} q(x,\theta_0,\lambda_0)\right]\\
\tilde{G}_{\lambda\lambda} &\equiv \mathbb{E}\left[\partial_{z'} g(z,\theta_0)\,\partial_{\lambda'} q(x,\theta_0,\lambda_0)\right]
\end{aligned}$$

where z solves $x = z - \partial_z g'(z,\theta)\,\lambda$ for given x,θ,λ and where

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

$$\partial_{\lambda'} q(x,\theta,\lambda) = \left[\left(I - \partial_{zz'}\left(\lambda' g(z,\theta)\right)\right)^{-1}\partial_z g'(z,\theta)\right]_{z=q(x,\theta,\lambda)}. \tag{13}$$

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 3.26 takes the familiar form expected from a just-identified GMM estimator: $(\tilde{G}'\Omega^{-1}\tilde{G})^{-1}$. 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.

We thus have now completely characterized the first-order asymptotic properties of our estimator in the most general settings of large (i.e., non-local) misspecification. This result thus allows researcher to directly replace their GMM estimator which may happen to fail overidentification tests by another, logically consistent and easy-to-interpret, estimator where the overidentification failure is naturally accounted for by errors in the variables. In addition, researchers can further document the presence of errors via Theorem 3.26, as it enables, as a by-product, a formal test of the absence of error. Under this null hypothesis, which can be stated as $\lambda = 0$, we have

$$n\hat{\lambda}' \left(W_{\lambda\lambda} - W_{\lambda\theta} W_{\theta\theta}^{-1} W_{\theta\lambda} \right)^{-1} \hat{\lambda} \xrightarrow{d} \chi^2_{d_g}, \tag{14}$$

where the above expression can be straightforwardly derived from the partitioned inverse formula applied to the λ sub-block of the asymptotic variance.

4 Discussion and extensions

On a conceptual level, our use of a so-called Wasserstein metric to measure distance between distributions does provide some desirable theoretical properties. For instance, the Wasserstein metric metrizes convergence in distribution (see Theorem 6.9 in Villani (2009)) under some simple bounded moment assumptions. In contrast, the discrepancies which generate GEL estimators do not admit such an interpretation. In fact, most discrepancies are not metrics, as they lack symmetry. The Kullback-Leibler discrepancy, which is perhaps the best known among them, does not allow comparison between distributions that are not absolutely continuous with respect to one another, whereas the Wassertein metric does. (Of course, leveraging this advantage requires considering the most general transport problem of Equation (1).) Finally, it is arguably logical to penalize probability transfer over larger distances more than the same probability transfer over smaller distances, as the Wasserstein metric does, while none of GEL-related discrepancies do.

An interesting extension of our approach would be a hybrid method in which (i) the possibility of general forms of errors is accounted for with the current method by constructing the equivalent GMM formulation of the model via Theorem 3.13 and (ii) additional restrictions on the form of the errors 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 errors is available for some, but not all, variables.

As shown in Supplement S.2, it is straightforward to extend our approach to allow error in some, but not all variables. When our method is used while allowing for errors in only a few variables, it may not be possible to simultaneously satisfy all moment conditions for any θ. In such cases, it could make sense to consider a hybrid method where both errors in the variables, handled via our approach, and re-weighting of the sample, handled via GEL, are simultaneously allowed.

Finally, we should mention other approaches aimed at handling violations of overidentifying restrictions, which include the use of set identification combined with relaxed moment constraints (Masten and Poirier (2021)), placing a Bayesian prior on the magnitude of the deviations from correct specification of the moments (Conley, Hansen, and Rossi (2012)), distributionally robust approaches that allow for deviations from the data generating process up to a given bound (Christensen and Connault (2022), Blanchet, Murthy, and Si (2022)), sensitivity analysis (Andrews, Gentzkow, and Shapiro (2017), Bonhomme and Weidner (2022)) and misspecified moment inequality models (e.g. Andrews and Kwon (2019)).

5 Application

We revisit the study of Duranton, Morrow, and Turner (2014), who documented evidence that the quantity of goods a city exports is strongly related to the extent of interstate highways present in that city. Due to simultaneity concerns, the authors adopt an instrumental variable approach to recover the causal effect of building highways. Although the authors mitigate instrument validity concerns with controls, instrument exogeneity or exclusion could remain a concern (as noted by Masten and

Poirier (2021)) and this problem would manifest itself by specification test failure.

Table 1: Main results

	GMM	OTGMM
log highway km	0.39	0.40
se	(0.12)	(0.11)
log employment	0.47	1.24
se	(0.32)	(0.31)
market access (export)	-0.63	-0.66
se	(0.11)	(0.10)
log 1920 population	-0.29	-0.57
se	(0.23)	(0.23)
log 1950 population	0.65	1.14
se	(0.37)	(0.37)
log 2000 population	-0.20	-1.25
se	(0.44)	(0.35)
log % manuf empl	0.64	0.58
se	(0.12)	(0.12)
Overidentification P-value	0.30	

Main results from Table 5 in Duranton, Morrow, and Turner (2014). Original GMM estimates and OTGMM estimates. Heteroskedasticity-robust standard errors (GMM) and small-error asymptotic standard error (OTGMM) in parentheses.

We apply our method to further assess the robustness of the results to potential model misspecification. We seek to recover point estimates that remain interpretable under potential misspecification and account for misspecification by viewing the model's variables as potentially measured with error. Not only do we consider errors in the regressors, but we also think of potentially invalid instruments as simply mismeasured versions of an underlying valid instrument that is unfortunately not available. Alternatively, one can think of the underlying valid instrument as the counterfactual value of the instrument in a world where the mechanism causing this instrument to be invalid would be absent. This broader interpretation of what could constitute an "error" under our framework considerably expands the scope of models that are conceptually consistent with our approach.

Duranton, Morrow, and Turner (2014) consider three instruments: (log) kilometers of railroads in 1898, quantity of historical exploration routes, and planned (log) highway kilometers according to a 1947 construction map. The validity of these instruments could be criticized, for instance, in a situation where some cities are in proximity to key natural resources, which could cause higher exports and, at the same time, more transportation routes (or plans to build them). If this mechanism is active both in the present and in the past, the causal effect of highways on exports would be overestimated. In our framework, the true but unavailable instrument could represent a measure of past transportation routes in a counterfactual world where natural resources would be evenly distributed among cities. The actual available instruments represent an approximation to this ideal instrument, a situation which we represent as a potentially non-classical errors-in-variables model.

The model's moment conditions are written in terms of $g(z_i, \theta) = w_i(y_i - r_i'\theta)$, where the vector of observables $z_i = (y_i, r_i', w_i')'$ contains the dependent variable y_i, the vector of regressors r_i and the vector of instruments w_i. In Duranton, Morrow, and Turner (2014), the dependent variable y_i measures "propensity to export" and is constructed from an auxiliary panel data model, which regresses volume of exports between given cities on distance and trading partner characteristics, modeled as fixed effects. It is the value of these fixed effects that is used as y_i. Following Masten and Poirier (2021), we take this construction as given and focus on export volume measured by weight. Explicitly accounting for errors in y_i is superfluous since, in a regression setting, errors in the dependent variable are already allowed for.

In the analysis below, r_i always includes the regressor of main interest: the (logarithm of) the number of kilometers of highway. It also contains a number of controls, which may differ in the different models considered. These controls include: log employment, log market access, log population in 1920, 1950, and 2000, and log manufacturing share in 2003, average distance to the nearest body of water, land gradient, dummy variables for census regions, log share of the fraction of adult population with a college degree or more, log average income per capita, log share of employment in wholesale trade, and log average daily traffic on the interstate highways in 2005. We allow for errors in all of these variables, except for the constant term and dummies.

We consider two of the specifications employed by Duranton, Morrow, and Turner (2014): The specification with many covariates that most obviously passes the overidentifying test and the specification with few covariates that most clearly fails this test. (The other specifications fall in between these extreme cases and are thus not reported here, for conciseness.) The results from GMM, replicated from the original study, and from OTGMM (allowing for large errors) are reported in Tables 1 and 2.

The OTGMM estimates are similar to those of Duranton, Morrow, and Turner

Table 2: Specification that fails the test for over-identifying restriction

	GMM	OTGMM
log highway km	0.57	0.65
se	(0.16)	(0.19)
log employment	0.52	0.49
se	(0.11)	(0.09)
market access (export)	-0.45	-0.46
se	(0.13)	(0.14)
Overidentification P-value	0.043	

Specification from column 2, Table 5, in Duranton, Morrow, and Turner (2014). Replicated estimates from the original paper and OTGMM estimates. Heteroskedasticity-robust standard errors (GMM) and small-error asymptotic standard error (OTGMM) in parentheses.

(2014). Although some coefficients of the controls are larger in magnitude, the main coefficient — the elasticity of export weight relative to kilometers of highway — is almost unchanged and remains statistically significant. As in the original study, the 95% confidence intervals of the main elasticity of interest obtained with different sets of controls overlap significantly.

It is also instructive to look at the correction of the underlying variables implied by OTGMM. Table 3, columns 1 and 2, documents this by looking at the standard deviation of different elements of the correction $z_i - x_i$. The corresponding quantities for the models of Tables 1 and 2 are reported in columns 1 and 2, respectively. To get an idea of the scale, the last column reports the standard deviations of the observed variables. As expected, the errors in column 1 are exceedingly small, reflecting the fact that the model passes the overidentification test. In contrast, column 2 is particularly interesting for our purposes because it corresponds to a specification that fails the test of over-identifying restrictions. While this may, at first, cast doubt regarding the GMM estimates, OTGMM shows that errors of the order of only 10% of the observed regressors' magnitude are sufficient to eliminate the mispecification. Such small error magnitudes are highly plausible empirically, thus supporting the plausibility of the OTGMM estimate. As the GMM and OTGMM estimates of the main elasticity of interest are very close in Table 2, this also corroborates the authors' GMM estimates. Overall, our approach strongly supports the conclusions of the original study and thus provides an effective robustness test.

The fact that the model with more controls leads to smaller error magnitudes is very consistent with our interpretation: to the extent that including more controls reduces the magnitude of the potentially endogenous residuals, one would expect that our estimator has to perform smaller alterations to the variables to arrive at valid instruments and/or non-endogeneous regressors. More quantitatively, suppose that the instrument w and the error term ϵ can be decomposed into separate components, say $w = w_1 + w_2$ and $\epsilon = \epsilon_1 + \epsilon_2$, where w_1 and ϵ_1 are correlated, but the latter is well explained by additional controls. In a model with fewer controls, our method has to identify w_1 and/or ϵ_1 as error components to obtain a correctly specified model. However, if including controls already accounts for the ϵ_1 term, then our estimator no longer needs to correct the corresponding components and thus achieves orthogonality with smaller errors.

In contrast, using GEL to address misspecification would effectively assume that the misspecification originates from some form of selection bias. The fact that adding more control eliminates the misspecification indicates that the controls incorporate the variables that explain selection bias. Yet, these controls are not included in the model with few controls, precisely the model where the largest amount of sample reweighting would take place when using GEL. This paradox makes the use of GEL as a remedy difficult to rationalize in a setting where misspecification arises primarily from the unavailability of adequate control variables.

Table 3: Standard deviations of errors and corresponding regressors

Variables	$\sqrt{\hat{\mathbb{V}}[z-x]}$ (main)	$\sqrt{\hat{\mathbb{V}}[z-x]}$ (failed over-ID)	$\sqrt{\hat{\mathbb{V}}[x]}$
log highway km	0.0034	0.0634	0.5884
log railroads km 1898	0.0048	0.0267	0.6031
exploration routes	0.0005	0.0418	0.8339
plan 1947	0.0075	0.0598	0.7049
log employment	0.0144	0.0473	0.8573
market access (export)	0.0056	0.0444	0.4864
log 1920 population	0.0052		1.0417
log 1950 population	0.0097		0.9253
log 2000 population	0.0163		0.8083
log % manuf empl	0.0050		0.3707

In Supplement S.4.1, we perform a number of robustness tests along the lines suggested by Masten and Poirier (2021). We further replicate specifications with many more controls and find that the GMM and OTGMM are still in close agreement. In Supplement S.4.2, we also consider alternative specifications that allow for the possibility that some instrumental variables should be included as regressors. The performance of OTGMM in other simple models is also investigated via simulations in Supplement S.5.

6 Conclusion

We have proposed a novel optimal transport-based version of the Generalized Method of Moment (GMM) that fulfills, by construction, the overidentifying moment restrictions by introducing the smallest possible amount of error in the variables or, equivalently, by allowing for the smallest possible Wasserstein metric distortions in the data distribution. 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 the sources of misspecification are errors or, more generally, Wasserstein-type distortions in the data distribution, which is arguably a common situation in applications.

A Proofs

A.1 Linearized estimator

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

$$\begin{aligned} z_j - x_j &\approx \partial_z g'(x_j, \theta)\,\lambda \\ z_j &\approx x_j + \partial_z g'(x_j, \theta)\,\lambda \end{aligned} \tag{15}$$

Substituting into the constraint yields: $\sum_j g(x_j + \partial_z g'(x_j, \theta)\,\lambda, \theta) \approx 0.$, while using a Taylor expansion gives: $\sum_j (g(x_j, \theta) + \partial_{z'} g(x_j, \theta)\,\partial_z g'(x_j, \theta)\,\lambda) \approx 0$, or:

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

where $H(x,\theta) = \partial_{z'} g(x,\theta)$, thus implying:

$$\lambda \approx -\left(\hat{\mathbb{E}}\left[H(x,\theta)\, H'(x,\theta)\right]\right)^{-1} \hat{\mathbb{E}}\left[g(x,\theta)\right] \tag{16}$$

Substituting (15) and (16) back into the objective function (2) yields (with $M \equiv \hat{\mathbb{E}}\left[H(x,\theta)\, H'(x,\theta)\right]$) :

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

A.2 Asymptotics

Proof of Theorem 3.8. We minimize $\frac{1}{2}\sum_{i=1}^n \|z_i - x_i\|^2$ subject to $\sum_{i=1}^n g(z_i,\theta) = 0$. First-order conditions for z_i read

$$z_i - x_i = \partial_z g'(z_i;\theta)\lambda \tag{17}$$

It is first shown that there exists a sequence z_i^* that matches the moment condition $\sum_{i=1}^n g(z_i^*;\theta_0) = 0$ 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 $\inf_{y\in B_\delta(x_i)} \|\partial_z g(y;\theta_0)\| \geq \eta$. We must have $\mathbb{P}[A] > 0$ for some (η,δ) because otherwise $\{\inf_{y\in B_\delta(x_i)} \|\partial_z g(y;\theta_0)\| \geq 1/n\}$ has probability 0 for all n, thus $\{\inf_{y\in B_\delta(x_i)} \|\partial_z g(y;\theta_0)\| > 0\}$ has probability 0 for all δ by continuity from below, contradicting assumption 3.5 with continuity of $\partial_z g$.

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 $n - |A_s^c|$[2]. 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 $m_k \equiv \{|A_s^c| +$

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

$Mk, \cdots, |A_s^c| + Mk + M - 1\}$ 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 $\sup_i \|z_i^* - x_i\| \leq \frac{\sup_i \|g(z_i^*;\theta) - g(x_i;\theta)\|}{\inf_{y \in A}\|\partial_z g(y;\theta)\|} \leq \eta o_p(1)$. 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, assumptions 3.5-3.6 with Hölder continuity exponent $\alpha \leq 1$, Cauchy-Schwartz, the LLN, and the previous convergence result

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

Furthermore, proceeding component-wise with $(k\cdot)$ denoting the k^{th} row of a matrix and using assumptions 3.5-3.6 together with previous results and proceeding as above for the term $\frac{1}{n}\sum_{i=1}^n \|[\partial_z g(z_i;\theta_0)]_{k\cdot}\| \, \|z_i - x_i\|^\alpha$, we have

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

and thus, using assumption 3.7, there exists $C_n = O_p(1)$ such that

$$\|\lambda\| \leq C_n \left\|\frac{1}{n}\sum_{i=1}^n \partial_z g(z_i;\theta_0)(z_i - x_i)\right\| \xrightarrow{p} 0$$

Now we derive a precise rate of convergence and the resulting asymptotic distribution for λ. Solving for z_i in equation (17) yields $z_i(\lambda)$, which can be plugged in the second equation to obtain $\sum_{i=1}^n g(z_i(\lambda);\theta) = 0$. By a Taylor expansion and assumption 3.7, we get

$$\frac{1}{n}\sum_{i=1}^{n} g(x_i,\theta_0) + \frac{1}{n}\sum_{i=1}^{n} \partial_z g(x_i;\theta_0)\partial_z g'(x_i;\theta_0)\lambda + O(\|\lambda\|^2) = 0 \tag{18}$$

By assumptions 3.1, 3.2, 3.3 and the central limit theorem, the first term is $O_p(n^{-1/2})$.

Under assumptions 3.1 and 3.7, we have $\frac{1}{n}\sum_{i=1}^{n}\partial_z g(x_i;\theta_0)\partial_z g(x_i;\theta_0)' \to^p \mathbb{E}[\partial_z g(x_i;\theta_0)\partial_z g(x_i;\theta_0)']$ 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 to the situation where $\theta \neq \theta_0$. By the uniform Law of Large Numbers, using assumption 3.4, $\sup_{\theta\in\Theta}\|\frac{1}{n}\sum_{i=1}^{n} g(x_i;\theta) - \mathbb{E}[g(x_i;\theta)]\| \to^p 0$.

For any $\theta \in B^c_\varepsilon(\theta_0)$ we have by identification $\mathbb{E}[g(x_i;\theta)] \in B^c_\gamma(0)$ 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 $\frac{1}{n}\sum_{i=1}^{n}\|g(\overline{z}_i;\theta)\|^2 \to^p \mathbb{E}[\|g(x_i;\theta)\|^2]$ 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 $z_i - x_i = \lambda'\partial_z g(x_i;\theta_0) + o_p(n^{-1/2})$ and $\frac{1}{n}\sum_{i=1}^{n} g(z_i;\theta_0) = 0$ and the linearized version is asymptotically justified. ■

Proof of Theorem 3.12. Using first-order conditions, previous results, and equation (18), we have

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

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

$$\sum_{i=1}^n \partial_\theta g(x_i;\theta)'\left(\sum_{i=1}^n \partial_z g(x_i;\theta_0)\partial_z g'(x_i,\theta_0)\right)^{-1}\sum_{i=1}^n g(x_i;\theta)=0$$

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

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

so that the estimator takes the form

$$\hat{\theta}_{OTGMM}-\theta_0=-\left(\sum_{i=1}^n \partial_\theta g(x_i;\hat{\theta})'\left(\sum_{i=1}^n \partial_z g(x_i;\theta_0)\partial_z g'(x_i,\theta_0)\right)^{-1}\sum_{i=1}^n \partial_\theta g(x_i;\overline{\theta})\right)^{-1}$$
$$\left(\sum_{i=1}^n \partial_\theta g(x_i;\hat{\theta})'\left(\sum_{i=1}^n \partial_z g(x_i;\theta_0)\partial_z g'(x_i,\theta_0)\right)^{-1}\sum_{i=1}^n g(x_i;\theta_0)\right)$$

Noting that under assumptions 3.1, 3.2, and 3.3 $\frac{1}{\sqrt{n}}\sum_{i=1}^n g(x_i;\theta_0)$ converges in distribution to a normal random variables by the central limit theorem and that assumptions 3.9-3.11 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 3.13. The first-order conditions with respect to the z_j (Equation (6)) can be written as

$$x_j = z_j - \partial_z g'\left(z_j,\theta\right)\lambda. \tag{19}$$

Under our assumptions, Equation (19) 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 (19) admits multiple solutions z_j for a given x_j, we identify the unique (with probability one) solution that minimizes $\|z_j - x_j\|^2$. This is accomplished by defining the mapping (9). With this definition, the first-order conditions (4) and (5) of the Lagrangian optimization problem for θ

and λ yield the just-identified GMM estimator stated in the Theorem. ■

The following Lemmas are shown in Supplement S.3.

Lemma A.1 *Let $h(\cdot,\cdot,\cdot)$ be continuous in all of its arguments. Then, under Assumptions 3.17(i) and 3.18, $h(q(x,\theta,\lambda),\theta,\lambda)$ is continuous in (θ,λ).*

Lemma A.2 *Under Assumptions 3.14 and 3.18, if $h \in \mathcal{L}$, then, for $q(x,\theta,\lambda)$ defined in Theorem 3.13, $\mathbb{E}\left[\sup_{(\theta,\lambda)\in\Theta\times\Lambda}\|h(q(x,\theta,\lambda),\theta)\|\right] < \infty$.*

Proof of Theorem 3.21. Assumptions 3.1-3.16 directly imply consistency of our GMM estimator, by Theorem 2.6 in Newey and McFadden (1994). There remains to show that Assumption 3.16 is implied by Assumptions 3.17,3.18,3.20.

We first establish Assumption 3.16 (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 A.1 for $h(z,\theta,\lambda) = g(z,\theta)$, under Assumptions 3.17(i) and 3.18. To show that $\partial_\theta g'(q(x,\theta,\lambda),\theta)\lambda$ is continuous in (θ,λ), we can similarly invoke Lemma A.1 for $h(z,\theta,\lambda) = \partial_\theta g'(z,\theta)\lambda$, where $\partial_\theta g'(z,\theta)$ is continuous in both arguments by Assumption 3.17(ii).

We now establish Assumption 3.16 (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 3.20, it follows that $\mathbb{E}\left[\sup_{(\theta,\lambda)\in\Theta\times\Lambda}\|g(q(x,\theta,\lambda),\theta)\|\right] < \infty$, by Lemma A.2. Next, we have, for $(\theta,\lambda) \in \Theta\times\Lambda$, $\|\partial_\theta g'(q(x,\theta,\lambda),\theta)\ \lambda\| \le \|\partial_\theta g'(q(x,\theta,\lambda),\theta)\|\,\|\lambda\| \le \|\partial_\theta g'(q(x,\theta,\lambda),\theta)\|\,\bar{\lambda}$ by Assumption 3.18 and compactness of Λ. By Assumption 3.20 and Lemma A.2 we then also have that $\mathbb{E}\left[\sup_{(\theta,\lambda)\in\Theta\times\Lambda}\|\partial_\theta g'(q(x,\theta,\lambda),\theta)\ \lambda\|\right] < \infty$. ■

Proof of Theorem 3.26. Theorem 3.21 implies consistency $(\theta,\lambda) \xrightarrow{p} (\theta_0,\lambda_0)$. This, in addition to Assumptions 3.22, 3.23 and 3.24 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 3.24 is implied by Assumption 3.25.

By Lemma A.1, Assumptions 3.25(i) and (iii) imply that both $g(q(x,\theta,\lambda),\lambda)$ and $\partial_\theta g'(q(x,\theta,\lambda),\lambda)\lambda$ are continuously differentiable in (θ,λ), thus establishing Assumption 3.24(i). By Lemma A.2, Assumptions 3.14, 3.25(ii) and (iii) imply Assumption 3.24(ii).

The asymptotic variance of the just-identified GMM estimator defined in Theorem 3.13 is then given by $\left(\tilde{G}'\Omega^{-1}\tilde{G}\right)^{-1}$ for Ω and $\tilde{G}$ as defined in the Theorem statement.

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 $z \mapsto z - \partial_z (\lambda' g(z, \theta))$. This can also be shown through an explicit calculation: To first order, (19) implies, for a small change $\Delta\theta$ in θ, a corresponding change Δz in z while keeping x and λ fixed, that:

$$0 = \Delta z - \partial_{zz'} (\lambda' g(z, \theta)) \Delta z - \partial_{z\theta'} (\lambda' g(z, \theta)) \Delta\theta.$$

Thus, $\Delta z = (I - \partial_{zz'} (\lambda' g(z, \theta)))^{-1} \partial_{z\theta'} (\lambda' g(z, \theta)) \Delta\theta$ and we have:

$$\partial_{\theta'} q(x, \theta, \lambda) = (I - \partial_{zz'} (\lambda' g(z, \theta)))^{-1} \partial_{z\theta'} (\lambda' g(z, \theta))$$

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

$$\partial_{\lambda'} q(x, \theta, \lambda) = (I - \partial_{zz'} (\lambda' g(z, \theta)))^{-1} \partial_z g'(z, \theta). \qquad \blacksquare$$

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," Discussion Paper 2184R, cowles foundation discussion paper.

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.

Blanchet, J., K. Murthy, and N. Si (2022): "Confidence regions inWasserstein distributionally robust estimation," *Biometrika*, 109, 295–315.

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

Caffarelli, L. A. (1996): "Boundary Regularity of Maps with Convex Potentials–II," *Annals of Mathematics, Second Series*, 144, 453–496.

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

Chamberlain, G. (1987): "Asymptotic efficiency in estimation with conditional moment restrictions," *Journal of Econometrics*, 34, 305–334.

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.

Duranton, G., P. M. Morrow, and M. A. Turner (2014): "Roads and Trade: Evidence from the US," *Review of Economic Studies*, 81(2), 681–724.

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 University Press, Princeton.

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.

——— (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(3), 1449–1469.

Newey, W., and D. McFadden (1994): "Large Sample Estimation and Hypothesis Testing," in *Handbook of Econometrics*, ed. by R. F. Engel, and D. L. McFadden, vol. IV. 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.

SANTAMBROGIO, F. (2015): *Optimal Transport for Applied Mathematicians.* Springer, New York.

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.

——— (2007): "Point Estimation with Exponentially Tilted Empirical Likelihood," *Annals of Statistics*, 35, 634–672.

——— (2014): "Entropic Latent Variable Integration via Simulation," *Econometrica*, 82, 345–386.

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

——— (2020): "Mismeasured and unobserved variables," in *Handbook of Econometrics*, ed. by J. H. S. Durlauf, L. Hansen, and R. Matzkin, vol. 7A, chap. 6, pp. 487–565. Elsevier Science.

VILLANI, C. (2009): "Optimal transport: Old and New," in *Grundlehren der mathematischen Wissenschaften.* Springer-Verlag, Heidelberg.

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

Online Supplement for:

Optimally-Transported Generalized Method of Moments

by Susanne Schennach and Vincent Starck

S.1 Algorithms

S.1.1 Iterative solution

The first order condition (6) can be re-written as

$$(z_j - x_j) = \partial_z g'(z_j, \theta)\lambda. \tag{20}$$

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 (20), we have:

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

Adding and subtracting z_j^t yields

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

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

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

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

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

$$\lambda^{t+1} = \left(\hat{\mathbb{E}}\left[H\left(z^t,\theta\right)H'\left(z^t,\theta\right)\right]\right)^{-1}\left(-\hat{\mathbb{E}}\left[g\left(z^t,\theta\right)\right]+\hat{\mathbb{E}}\left[H\left(z^t,\theta\right)\left(z^t-x\right)\right]\right). \quad (21)$$

From this, we can improve the approximation to z_j to go to the next step, using (20):

$$z_j^{t+1} = x_j + H'\left(z^t,\theta\right)\lambda^{t+1}. \quad (22)$$

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:

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

S.1.2 Iterative procedure convergence

Substituting (21) into (22) yields an iterative rule expressed solely in terms of z_j^t:

$$z_j^{t+1} = x_j + H'\left(z_j^t,\theta\right)\left(\hat{\mathbb{E}}\left[H\left(z^t,\theta\right)H'\left(z^t,\theta\right)\right]\right)^{-1}\left(-\hat{\mathbb{E}}\left[g\left(z^t,\theta\right)\right]+\hat{\mathbb{E}}\left[H\left(z^t,\theta\right)\left(z^t-x\right)\right]\right). \quad (23)$$

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 S.1.1 *(i) $g(z,\theta)$ 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 the closure of an open neighborhood η of the fixed point $\mathbf{z}^\infty$.*

Theorem S.1.2 *Under Assumption S.1.1, 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 (23) and starting at any $\mathbf{z}^0 \in \eta$ converges to the unique fixed point $\mathbf{z}^\infty \in \eta$, provided $\|\lambda\|$ is sufficiently small (where λ solves the first order condition (20)).*

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 errors 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 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 S.1.2. For a rule of the form $\mathbf{z}^{t+1} = f(\mathbf{z}^t)$, 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 $[\partial f(\mathbf{z})/\partial \mathbf{z}']_{\mathbf{z}=\mathbf{z}^\infty}$ must have a modulus strictly less than 1.

The smoothness condition (i) is trivially satisfied under Assumption S.1.1. Next, letting $z_{i,k}^t$ denote one element of the vector z_i^t, and $H_{\cdot k}(z_i^t,\theta)$ denote the k^{th} column of $H(z_i^t,\theta)$, we can express all blocks $\partial z_j^{t+1}/\partial z_{i,k}^t$ of the matrix of partial derivatives of $f(\mathbf{z})$:

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

where the two $n^{-1}H_{\cdot k}(z_i^t,\theta)$ terms cancel each other. At $\mathbf{z}^t = \mathbf{z}^\infty$, $\hat{\mathbb{E}}[g(z^\infty,\theta)] = 0$ and $(z_i^\infty - x_i) = H'(z_i^\infty,\theta)\lambda$ and we have:

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

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 $\left(\hat{\mathbb{E}}\left[H\left(z,\theta\right)H'\left(z,\theta\right)\right]\right)^{-1}$, which is nonsingular for $\mathbf{z}\in\eta$ by Assumption S.1.1(ii), and derivatives of $g\left(z,\theta\right)$ 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 S.1.1(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 $f\left(\mathbf{z}\right)$ can be made strictly less than 1 for λ sufficiently small. ■

S.2 Constrained estimator

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

$$\begin{aligned}\hat{\mathbb{E}}\left[g\left(z,\theta\right)\right] &= 0 &(24)\\ C\left(z_i-x_i\right) &= 0 \text{ for } i=1,\ldots,n &(25)\end{aligned}$$

for some known rectangular full row rank matrix C that selects the dimensions of the error vector x_i-z_i that should be constrained to be zero. Note that error constraint is imposed for each observation, not in an average sense. Naturally, we assume that the number of constraints imposed is not so large that it is no longer possible to satisfy all moment conditions simultaneously.

Proposition S.2.1 *The first order conditions (4) and (5) are unchanged, while Equation (6) becomes:*

$$\left(z_j-x_j\right)-PH'\left(x,\theta\right)\lambda=0 \tag{26}$$

where $P=\left(I-C'\left(CC'\right)^{-1}C\right)$ *and* $H'\left(x,\theta\right)\equiv\partial_z g'\left(z_j,\theta\right)$.

Thanks to linearity, the Lagrange multipliers of the error constraints 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 are no errors.

Proof. The Lagrangian for this problem is

$$\frac{1}{2}\hat{\mathbb{E}}\|z-x\|^2-\lambda'\hat{\mathbb{E}}\left[g\left(z,\theta\right)\right]-\sum_{i=1}^{n}\gamma_i'C\left(z_i-x_i\right) \tag{27}$$

where λ and γ_i are Lagrange multipliers. The first order conditions of the Lagrangian (27) with respect to z_j is

$$(z_j - x_j) - \partial_z g'(z_j, \theta)\,\lambda - C'\gamma_j = 0. \tag{28}$$

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

$$C\,(z_j - x_j) - CC'\gamma_j = C\partial_z g'(z_j, \theta)\,\lambda,$$

thus allowing us to solve for γ_j:

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

Upon substitution of γ_i into (28) and simple re-arrangements, we obtain

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

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

■

The iterative Algorithm 2.1 can easily be adapted by replacing every instance of $H'(z^t, \theta)$ by $PH'(z^t, \theta)$. Similarly the linearized estimator of Equation (7) becomes:

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

Note that this expression assumes that the matrix being inverted remains full rank, a condition that can be interpreted as requiring the constraints to not be so strong as to make impossible to simultaneously satisfy all the moment conditions.

Asymptotic results can also be straightforwardly adapted.

Corollary S.2.2 *Theorem 3.12 holds under constraint (25), with all instances of* $\mathbb{E}[H_i H_i']$ *replaced by* $\mathbb{E}[H_i P H_i']$, *for* $P = \left(I - C'\left(CC'\right)^{-1} C\right)$.

S.3 Asymptotics: Proofs of Lemmas

Proof of Lemma A.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

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

since $q(x,\theta,\lambda)$ is the inverse of $z\mapsto z-\partial_z\left(\lambda' g(z,\theta)\right)$. By the definition of $\bar\lambda$, $\bar\nu$,

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

at $z=q(x,\theta,\lambda)$, where $\bar\lambda\bar\nu<1$ by Assumption 3.18 and where $\partial_{z\theta'}\left(\lambda' g(z,\theta)\right)$ exists by Assumption 3.17. Thus $h(q(x,\theta,\lambda),\theta,\lambda)$ is continuous in θ.

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

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

where $\bar\lambda\bar\nu<1$ by Assumption 3.18 and $\partial_z g'(z,\theta)$ exists by Assumption 3.17. ■

Proof of Lemma A.2. By the triangle inequality and Definition 3.19, there exists $\bar h(x,\theta)$ such that:

$$\|h(z,\theta)\| \le \|h(x,\theta)\| + \|h(z,\theta)-h(x,\theta)\| \le \|h(x,\theta)\| + \bar h(x,\theta)\|z-x\|, \quad (29)$$

for $z=q(x,\theta,\lambda)$. Next, using the first order conditions (Equation (6)), we have, by a mean value argument, the triangle inequality and the definitions of $\bar\lambda$ and $\bar\nu$ from Assumption 3.18, for some mean value $\tilde x$,

$$\begin{aligned}\|z-x\| &= \left\|\partial_z\left(\lambda' g(z,\theta)\right)\right\| \le \left\|\partial_z\left(\lambda' g(x,\theta)\right)\right\| + \left\|\partial_{zz'}\left(\lambda' g(\tilde x,\theta)\right)(z-x)\right\| \\ &\le \bar\lambda\left\|\partial_{z'} g(x,\theta)\right\| + \bar\lambda\bar\nu\|z-x\|. \end{aligned} \quad (30)$$

Re-arranging and using the fact that $\bar\lambda\bar\nu<1$ by Assumption 3.18 and $\bar\lambda<\infty$ by compactness of Λ,

$$\|z - x\| \leq \frac{\bar{\lambda} \|\partial_{z'} g(x, \theta)\|}{\left(1 - \bar{\lambda}\bar{\nu}\right)}. \tag{31}$$

Combining (29) and (31) and noting that applying the $\mathbb{E}\left[\sup_{\theta\in\Theta} \ldots\right]$ operator does not alter the inequalities, we have

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

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

S.4 Application: Additional results

S.4.1 Regression with all controls

In this section, we estimate a regression with controls, which now also includes average distance to the nearest body of water; land gradient; dummy variables for census regions; log share of the fraction of adult population with a college degree or more; log average income per capita; log share of employment in wholesale trade; and log average daily traffic on the interstate highways in 2005. The results, reported in Table S.1, are similar to those of the main regression shown in Table 1. In Table S.2, we also show how the main coefficient of interest changes when including various subsets of controls.

S.4.2 Relaxing exclusion restrictions

As Masten and Poirier (2021), we consider the possibility that some instrumental variables have non-zero regression coefficient, i.e., have a direct impact on the outcome. Since our estimator relies on over-identifying restrictions to recover meaningful variable corrections, we only relax one of the exclusion-restrictions at a time. This leads to the following results.

As in Masten and Poirier (2021), we found that removing railroads has the largest impact on coefficients and is likely an improper instrument.

Table S.1: Results with all controls

	GMM	OTGMM
Dependent variable	Exporter fixed effect weight	
log highway km	0.45	0.47
se	(0.16)	(0.17)
log employment	0.67	1.06
se	(0.42)	(1.12)
market access (export)	−0.49	−0.50
se	(0.10)	(0.10)
log 1920 population	−0.29	−0.40
se	(0.24)	(0.30)
log 1950 population	0.71	0.88
se	(0.39)	(0.50)
log 2000 population	−0.65	−1.17
se	(0.47)	(1.30)
log % manuf empl	0.57	0.54
se	(0.14)	(0.14)
Water	0.07	0.06
se	(0.05)	(0.05)
Land gradient	−0.21	−0.20
se	(0.08)	(0.09)
College	−0.56	−0.77
se	(0.47)	(0.54)
price	0.34	0.59
se	(0.57)	(0.54)
wholesale	0.74	0.73
se	(0.26)	(0.63)
traffic	0.37	0.38
se	(0.30)	(0.29)

Results with all controls. Replicated estimates from the original paper and OT-GMM estimates. Heteroskedasticity-robust standard errors (GMM) and small-error asymptotic standard error (OTGMM) in parentheses. The regression also include an intercept and census region dummies.

Table S.2: Coefficient on log highway kilometers for various control sets

Controls	GMM	OTGMM
none	1.13 (0.14)	1.17 (0.15)
2	0.57 (0.16)	0.65 (0.19)
1-5	0.47 (0.13)	0.46 (0.12)
1-6	0.39 (0.12)	0.40 (0.11)
1-7	0.34 (0.15)	0.37 (0.13)
1-8	0.28 (0.14)	0.31 (0.13)
1-9	0.27 (0.14)	0.30 (0.13)
1-10	0.27 (0.13)	0.29 (0.13)
1-11	0.38 (0.13)	0.45 (0.14)
1-12	0.44 (0.16)	0.48 (0.17)
All	0.45 (0.16)	0.47 (0.17)

Coefficient on log highway kilometers for various sets of controls, numbered as follows: employment (1), market access (2), population 1920 (3), population 1950 (4), population 2000 (5), log % manuf employment (6), water (7), land gradient (8), college (9), price (10), wholesale (11), traffic (12), region census dummies (13-20). All regressions include an intercept.

Table S.3: Including log 1947 highway kilometers as a regressor

	GMM	OTGMM
Dependent variable	Exporter fixed effect weight	
log highway km	0.52	0.46
se	(0.51)	(0.52)
log 1947 highway km	−0.09	−0.04
se	(0.36)	(0.36)
log employment	0.44	1.24
se	(0.33)	(0.37)
market access (export)	−0.63	−0.65
se	(0.11)	(0.10)
log 1920 population	−0.30	−0.58
se	(0.24)	(0.23)
log 1950 population	0.64	1.15
se	(0.38)	(0.38)
log 2000 population	−0.19	−1.26
se	(0.46)	(0.38)
log % manuf empl	0.64	0.58
se	(0.12)	(0.12)

Including log 1947 highway kilometers as a regressor. Replicated estimates from the original paper and OTGMM estimates. Heteroskedasticity-robust standard errors (GMM) and small-error asymptotic standard error (OTGMM) in parentheses

Table S.4: Including log exploration as a regressor

	GMM	OTGMM
Dependent variable	Exporter fixed effect weight	
log highway km	0.42	0.44
se	(0.16)	(0.16)
log exploration	−0.02	−0.03
se	(0.06)	(0.06)
log employment	0.47	1.18
se	(0.34)	(0.36)
market access (export)	−0.63	−0.65
se	(0.11)	(0.11)
log 1920 population	−0.29	−0.52
se	(0.23)	(0.26)
log 1950 population	0.64	1.05
se	(0.38)	(0.43)
log 2000 population	−0.20	−1.16
se	(0.45)	(0.43)
log % manuf empl	0.63	0.57
se	(0.13)	(0.12)

Including log exploration as a regressor. Replicated estimates from the original paper and OTGMM estimates. Heteroskedasticity-robust standard errors (GMM) and small-error asymptotic standard error (OTGMM) in parentheses

Table S.5: Including log 1998 railroad as a regressor

	GMM	OTGMM
Dependent variable	Exporter fixed effect weight	
log highway km	0.23	0.24
se	(0.14)	(0.14)
log 1998 railroad	0.16	0.15
se	(0.10)	(0.10)
log employment	0.44	0.52
se	(0.31)	(0.44)
market access (export)	−0.63	−0.63
se	(0.11)	(0.11)
log 1920 population	−0.41	−0.41
se	(0.24)	(0.30)
log 1950 population	0.76	0.77
se	(0.37)	(0.48)
log 2000 population	−0.16	−0.26
se	(0.41)	(0.48)
log % manuf empl	0.62	0.61
se	(0.12)	(0.56)

Including log 1998 railroad as a regressor. Replicated estimates from the original paper and OTGMM estimates. Heteroskedasticity-robust standard errors (GMM) and small-error asymptotic standard error (OTGMM) in parentheses

S.5 Simulations

We conduct simulations to assess the performance of our estimator and compare it to efficient GMM. We consider both the OTGMM estimator (Equations (2) and (3)) and the GMM estimator obtained under the assumption of small errors (Equation (7)).

Given the nonlinear nature of our estimator, we deliberately select a small sample size ($n = 100$) to explore the regime where asymptotic results do not trivially hold. 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 error scale σ in order to assess the impact of magnitude of the errors on the performance of estimators that only use the observed x_i.

Specifically, we consider the following distributions for z_i: $z_i \sim \mathcal{N}(1.5, 2)$ $z_i \sim$ 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:

$$\mathbb{E}[z_i - \theta] = 0, \ \mathbb{E}\left[e^{z_i} - \frac{2}{3}\theta\mathbb{E}[e^{z_i}]\right] = 0 \tag{32}$$

$$\mathbb{E}[z_i - \theta] = 0, \ \mathbb{E}\left[\frac{e^{2z_i-3}}{1+e^{2z-3}} - \frac{2}{3}\theta\frac{e^{2z_i-3}}{1+e^{2z-3}}\right] = 0 \tag{33}$$

$$\mathbb{E}[e^{z_i} - \frac{2}{3}\theta\mathbb{E}[e^{z_i}]] = 0, \ \mathbb{E}\left[\frac{e^{2z_i-3}}{1+e^{2z-3}} - \frac{\theta e^{2z_i-3}}{(1.5)\left(1+e^{2z-3}\right)}\right] = 0 \tag{34}$$

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

$$\mathbb{E}[z_i - \theta] = 0, \ \mathbb{E}[{z_i}^2 - 2\theta^2] = 0. \tag{35}$$

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

In Tables S.6-S.9, 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, indicated by "Linearized OTGMM"), OTGMM in the general large-error case (middle columns) and efficient GMM ignoring the presence of errors (rightmost columns).

Table S.6: Simulation results: Equation (32)

	Bias																	
	Linearized OTGMM						OTGMM						Efficient GMM					
error 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																	
	Linearized OTGMM						OTGMM						Efficient GMM					
error 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																	
	Linearized OTGMM						OTGMM						Efficient GMM					
error 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 S.7: Simulation results: Equation (33)

	Bias																	
	Linearized OTGMM						OTGMM						Efficient GMM					
error 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																	
	Linearized OTGMM						OTGMM						Efficient GMM					
error 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																	
	Linearized OTGMM						OTGMM						Efficient GMM					
error 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

While these tables consider a large number of specifications and data generating processes, we here highlight a few values that illustrate the typical trends present throughout the simulations. First, it is clear that OTGMM is, in general, preferable to linearized OTGMM in terms of bias, and hence we focus our discussion on the former. As an example, let us compare the OTGMM and GMM results in Table S.6, for the column that corresponds to an error scale of $\sigma = 1.5$ for the normal model. The OTGMM estimator reduces the bias magnitude to only 0.05, as compared to 0.18 for GMM. At the same time, the variance only increases from 0.19 (for GMM) to 0.20 (for OTGMM), which is a negligible increase, thus supporting the idea that GMM's use of the optimal weighting matrix is not particularly beneficial in this context. As a result, the overall root mean square error (RMSE) is 0.21 for OTGMM down from

Table S.8: Simulation results: Equation (34)

	Bias																	
	Linearized OTGMM						OTGMM						Efficient GMM					
error 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																	
	Linearized OTGMM						OTGMM						Efficient GMM					
error 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																	
	Linearized OTGMM						OTGMM						Efficient GMM					
error 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 S.9: Simulation results: Equation (35)

	Bias																	
	Linearized OTGMM						OTGMM						Efficient GMM					
error 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																	
	Linearized OTGMM						OTGMM						Efficient GMM					
error 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																	
	Linearized OTGMM						OTGMM						Efficient GMM					
error 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

0.26 for GMM.

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.