Adversarial Method of Moments

Ignacio Cigliutti
NYU

Elena Manresa
NYU

March 18, 2022

Abstract

We introduce Adversarial Method of Moments (AMM), for models defined with moment conditions. The estimator is asymptotically equivalent to optimally–weighted 2–step GMM, but outperforms the GMM estimator in finite samples. We show this both in theory and in simulations. In our theoretical results, we exploit the relationship between AMM and GEL estimators to show, using stochastic expansions, that AMM has smaller bias than optimally–weighted GMM. In our simulation experiments, we consider different models, including estimation of variance as in Altonji and Segal (1996) and a dynamic panel data model. We compare the estimator's performance to other commonly–used procedures in the literature, and find that AMM outperforms in cases where other estimators fail. In the appendix, we extend AMM to simulation–based settings, with an application to the estimation of DSGE models by matching IRF.

1 Introduction

It is well known, at least since Altonji and Segal (1996), that GMM suffers from important finite sample bias. Since then, many alternatives have been proposed, including the Empirical likelihood (EL), Continuous–Updating (CUE) and Exponential Tilting (ET) estimators. In a seminal paper, Newey and Smith (2004) shown that all these estimators share a common structure, being members of a class of generalized empirical likelihood (GEL) estimators. They also showed that these estimators overcome many flaws of GMM. However, mostly due to implementation hurdles, GEL estimators have rarely been used in empirical research.

The goal of this paper is to introduce a new estimator, Adversarial Method of Moments (AMM), as an easy–to–implement alternative to GEL estimators. As we will show, AMM share a similar structure to GEL, and thus inherits some of its desirable finite sample properties. Moreover, AMM is much easier to implement, since computationally it amounts to run a Logit regression.

The AMM estimator is inspired on Generative Adversarial Networks (GAN) Goodfellow et al. (2014). GAN approaches estimation via a min–max optimization criterion, in which two models compete over the loss: A generative model G, from which we may sample synthetic data, and a discriminative model D that estimates the probability that an observation came from the original data rather than the synthetic data. The maximizer of the parameters corresponds to the value for which D find it more difficult to distinguish both kinds of data.

AMM primarily considers models that are defined by moment conditions, $\mathbb{E}[g(x_i, \theta)] = 0$, and restricts the class of discriminators to logistic regression. The estimator, $\hat{\theta}$, is defined as the value for which the probability (according to the predictions of the logistic regression) that an observation $g\left(\hat{\theta}, x_i\right)$ is drawn from a 0 mean variable is the same for all i and equal to $1/2$.

Most related to this paper is Kaji, Manresa, and Pouliot (2020) (KMP), who introduced adversarial estimation for structural models. There, KMP focus on the case where D is a non–parametric estimator of an oracle discriminator, and showed that in that case, the estimator of the structural model attains efficiency. Moreover, KMP showed that if D is set to be a neural network with zero hidden layers and the activation function is chosen to be logistic, the estimator asymptotically equivalent to SMM (Gourieroux, Monfort, and Renault (1993)).[1] Instead, in this paper we focus on models described by a finite set of moment conditions, so our procedure does not involve solving (and simulating) a structural model. Moreover, we are interested in understanding the finite–sample properties of adversarial estimators, whereas KMP was all about asymptotics.

The paper is structured as follows: In Section 2, we set the notation and describe the different estimators we will consider. Section 3 describes the AMM estimator in detail. We distinguish two cases: Models based on moment conditions, and structural models from which it is possible to draw simulations. Section 4 provides Monte–Carlo evidence of the performance of the AMM estimator: For the first case, we consider the frameworks of Altonji and Segal (1996) and Arellano and Bond (1991); for the second, we consider matching the impulse–response functions of a Structural VAR and a New Keynesian model. Section 5 develops the theory for the asymptotic and finite–sample properties of the AMM estimator. Section 6 concludes.

2 The Set Up

We consider models with a finite number of moment restrictions. To describe it, let $z_i \;\; (i = 1, \ldots, n)$ be i.i.d. observations on a data vector z. Also, let θ be a $(p \times 1)$ vector of parameters of interest and $g\,(z;\theta)$ be a $(m \times 1)$ vector of functions of the data, where $m \geq p$. The identification requirement is that true parameter value θ_0 will satisfy $\mathbb{E}\,[g\,(z;\theta_0)] = 0$, where $\mathbb{E}\,[\cdot]$ denotes expectation taken with respect to the

[1] Relatedly, Liang, 2021 study the optimal rate of convergence of a general class of adversarial methods and particularizes its findings to GMM/SMM estimation. Our work differs from his in that we characterize the finite sample bias of AMM and show it is smaller than that of otpimally–weighted GMM/SMM.

distribution of z_i. Within this set of models, we consider GMM and GEL estimators, which we describe below.

2.1 GMM estimation

A widely used estimator is the two–step GMM estimator of Hansen (1982). To describe it, let $g_i(\theta) \equiv g(x_i;\theta), \widehat{g}(\theta) \equiv n^{-1}\sum g_i(\theta)$ and $\widehat{\Omega}(\theta) \equiv n^{-1}\sum g_i(\theta) g_i(\theta)'$. Also, let $\tilde{\theta}$ be the one–step GMM estimator, given by $\tilde{\theta} = \arg\min_{\theta\in\Theta} \widehat{g}(\theta)' W^{-1} \widehat{g}(\theta)$ where Θ denotes the parameter space, and $W = I_m$. The two–step GMM estimator is given by

$$\widehat{\theta}_{GMM} = \arg\min_{\theta\in\Theta} \widehat{g}(\theta)' \widehat{\Omega}\left(\tilde{\theta}\right)^{-1} \widehat{g}(\theta)$$

We also consider the continuous–updating estimator (CUE) of Hansen, Heaton, and Yaron (1996), which is analogous to GMM except that the objective function is simultaneously minimized over θ in $\widehat{\Omega}(\theta)^{-1}$. It is given by

$$\widehat{\theta}_{CUE} = \arg\min_{\theta\in\Theta} \widehat{g}(\theta)' \widehat{\Omega}(\theta)^{-1} \widehat{g}(\theta)$$

2.2 GEL estimation

A second set of estimators is the generalized empirical likelihood (GEL) estimators (from which CUE estimator is a particular case). These estimators solve a min–max problem, and are described by a function $\rho(v)$ of a scalar v, which is concave on its domain $\mathcal{V}$, an open interval containing zero. If we define the admissible set in the inner maximization as $\hat{\mathcal{B}}_n(\theta) = \{\lambda : \lambda' g_i(\theta) \in \mathcal{V}, i = 1, \dots, n\}$, then the GEL estimator is the solution to a saddle point problem

$$\hat{\theta}_{\text{GEL}} = \arg\min_{\theta\in\Theta} \sup_{\lambda\in\hat{\mathcal{B}}_n(\theta)} \sum_{i=1}^{n} \rho(\lambda' g_i(\theta)) \tag{1}$$

As mentioned before, GEL nests different estimators, by coosing ρ appropiately: $\rho(v) = \log(1-v)$ and $\mathcal{V} = (-\infty, 1)$ delivers the empirical likelihood (EL) estimator; $\rho(v) = -e^v$ describes the exponential tilting (ET) estimator, and the case of quadratic $\rho(v)$ corresponds to CUE.

3 The AMM estimator

Our estimator uses two sources of inputs: Observations from the true data (which are related to the moment conditions), denoted by $\mathbf{g}(\theta) \equiv \{g_i(\theta)\}_{i=1}^n$, and some random draws, $\widetilde{\mathbf{g}} \equiv \{\widetilde{g}_i\}_{i=1}^m$. Throughout this paper we'll let $\widetilde{g}_i = \nu\varepsilon_i$, with $\epsilon \sim \mathcal{N}(0,1)$. [2] Moreover, we denote the stacked inputs as $\mathbf{X}(\theta) = \left(\mathbf{g}(\theta)', \widetilde{\mathbf{g}}'\right)'$. In addition to $\mathbf{X}(\theta)$,

[2] However, any generator with $\mathbb{E}[\widetilde{g}_i] = 0$ would work.

there's also a binary output variable $\mathbf{d} = \{d_i\}_{i=1}^{n+m}$, which reflects whether an observation comes from the true data ($d_i = 1$) or not ($d_i = 0$).

The AMM estimator involves two models, a discriminator D and a generator G. These models, in turn, compete over the loss through a min–max problem: On the one hand, D uses $(\mathbf{X}(\theta), \mathbf{d})$ to solve a classification problem by estimating the probability that a given observation comes either from the true data or not, in the same way as Logit regression. On the other hand, G seeks to confuse D, so that it's incapable of distinguishing true and random observations. To do so, G is allowed to choose θ.

At a broad level, our estimator can be described as the solution to the following optimization problem

$$\min_{\theta\in\Theta}\left\{\max_{D\in\mathcal{D}}\frac{1}{n}\sum_{i=1}^{n}\log D\left(g_i\left(\theta\right)\right)+\frac{1}{m}\sum_{i=1}^{m}\log\left(1-D\left(\widetilde{g}_i\right)\right)\right\}$$

where $\mathcal{D}$ is some family of discriminators. In principle, $\mathcal{D}$ could be very rich, case in which the discriminator may be able to leverage on many features of the likelihood. (KMP), for example, focus their attention on models with vast heterogeneity, and consider a sieve of neural networks for efficient estimation. As hinted before, however, in this paper we'll focus on the family of logistic discriminators

$$\mathcal{D}\equiv\left\{D : D\left(x\right)=\Lambda\left(\lambda' x\right),\ \ \lambda\in\mathbb{R}^d\right\}$$

where $\Lambda(t)\equiv\left(1+e^{-t}\right)^{-1}$. This restriction allows us to better study its finite–sample properties, and thus we can describe the AMM estimator as

$$\widehat{\theta}_{AMM}=\arg\min_{\theta\in\Theta}\left\{\max_{\lambda\in\mathbb{R}^d}\frac{1}{n}\sum_{i=1}^{n}\log\Lambda\left(\lambda' g_i\left(\theta\right)\right)+\frac{1}{m}\sum_{i=1}^{m}\log\left(1-\Lambda\left(\lambda'\widetilde{g}_i\right)\right)\right\}$$

3.1 The Objective function

The motivation for this loss function can be stated as follows: Suppose we want to describe the probability of a binary outcome. In particular, we are interested in whether the observation i comes from the true data or not. We proceed by parametrizing the probability with the Logistic function: $\Pr\left(i\ true\right)=\Pr\left(d_i=1\right)=\Lambda\left(\lambda' g\left(x_i,\theta\right)\right)$, and use the features $g\left(\theta\right)$ as regressors to solve this classification problem (for a fixed θ). In such case, the likelihood is

$$\mathcal{L}=\prod_{i=1}^{n+m}\Lambda\left(\lambda' g\left(x,\theta\right)\right)^{d_i}\left[1-\Lambda\left(\lambda' g\left(\widetilde{x}_i\left(\theta\right),\theta\right)\right)\right]^{1-d_i}$$

Our loss function is just the log–likelihood in that case.

3.2 The First Order Conditions

We can also build intuition for the estimator by considering the FOC of the inner maximization (D's problem) for a given θ:

$$\frac{1}{n}\sum_{i=1}^{n}\left[1-\Lambda\left(\lambda' g_i\left(\theta\right)\right)\right] g_i\left(\theta\right)=\frac{1}{m}\sum_{i=1}^{m}\Lambda\left(\lambda'\widetilde{g}_i\right)\widetilde{g}_i$$

The idea is that D searches for the value of λ that matches the weighted averages of $g_i\left(\theta\right)$ and $\widetilde{g}_i$. Under correct specification of the moment conditions, we have $\mathbb{E}\left[g\left(x_i,\theta_0\right)\right]=0$, so $\lambda\left(\theta_0\right)=0$ would be a solution. In fact, by concavity of the objective function with respect to λ, it is the only solution. In that case, since $\Lambda\left(0\right)=1-\Lambda\left(0\right)=1/2$, the FOC boils down to

$$\frac{1}{n}\sum_{i=1}^{n} g\left(z_i,\theta_0\right)=\frac{1}{m}\sum_{i=1}^{m}\widetilde{g}_i\simeq 0$$

so that the AMM estimator enforces the moment conditions. Finally, and as will be laid out below, AMM can be seen as a GEL estimator with a particular choice of the function ρ: $\rho(v)=\log\Lambda(v)$, $\log\left[1-\Lambda(v)\right]$. However, the AMM estimator circumvents computational issues that arise in GEL estimators due to the nature of the constrained optimization problem. In AMM we do not need to restrict λ in any way and solving the inner maximization problem amounts to estimate a standard (and convex!) logistic regression model.

3.3 A simple example – OLS

As an example, consider the case of a linear regression model

$$y_i=\beta' x_i+\eta_i$$

We are interested in β, and our moment condition is $\mathbb{E}(x_i(y_i-\beta x_i))=0$. In this case, our dataset would be

$$\left(\mathbf{X}\left(\theta\right),\mathbf{d}\right)=\begin{bmatrix} x_1\left(y_1-\beta' x_1\right) & 1\\ x_2\left(y_2-\beta' x_2\right) & 1\\ \vdots & \vdots\\ x_n\left(y_n-\beta' x_n\right) & 1\\ \nu\varepsilon_1 & 0\\ \nu\varepsilon_2 & 0\\ \vdots & \vdots\\ \nu\varepsilon_m & 0\end{bmatrix}$$

and proceed to estimation. We consider a simple case where $x_i\sim N\left(\mu_x,\sigma_x\right)$ and $\eta_i\sim N\left(0,1\right)$. Moreover, we fix $\epsilon_i\sim N\left(0,1\right)$ and work with different values of ν to study how different values of the dispersion coeffcient may impact estimation. In particular, we run $S = 500$ simulations of size $N = 200$.

	σ	Bias	STD	RMSE
OLS	–	0.001	0.076	0.076
AMM	0	0.001	0.076	0.076
AMM	0.05	0.001	0.076	0.076
AMM	0.1	0.001	0.077	0.077
AMM	0.5	0	0.085	0.085
AMM	1	-0.001	0.107	0.107

Table 1: 500 simulations, sample size = 200

For this simple example, results remain unchanged. However, as we will see later on, this is not a general result.

4 Simulation Exercises

In this section, we present several applications that will highlight the finite–sample performance of the AMM estimator in multiple frameworks: First, we consider the model in Altonji and Segal (1996). This is a natural starting point, since it's the first paper that showed that stacking many moment conditions may yield poor finite–sample properties in optimally–weighted procedures (due to the fact that the same data is used both for estimating the moment conditions and the weighting matrix). As a second example, we consider the estimation of the autorregresive coefficient in a dynamic panel data model with the moment conditions described in Arellano and Bond (1991). We conclude this section with an application of the simulation–based AMM: We estimate a DSGE model by matching impulse–response functions in the model and the data.

4.1 Altonji–Segal (1996)

Altonji and Segal (1996) consider a panel of individuals ($i = 1, \dots, N$) observed across T periods. In their model, an observation from agent i is $x_i = (x_{i1}, x_{i2}, \dots, x_{iT})'$. Moreover, observations are independent across (i, t) and all have equal mean and variance, $\mathbb{E}[x_{it}] = \mu$, $\mathbb{E}\left[(x_{it} - \mu)^2\right] = \sigma^2$.

Since our focus is to estimate variance, we look at the case with $\mu = 0$ and $\sigma^2 = 1$. The set of moment conditions are the T cross–sectional variances (linear in σ^2):

$$g\left(x_i,\theta\right)=\left[\begin{array}{c} x_{i1}^2-\sigma^2 \\ x_{i2}^2-\sigma^2 \\ \vdots \\ x_{iT}^2-\sigma^2 \end{array}\right]$$

In this example, it can be shown that GMM is equivalent to OLS estimator of $s_t^2=\theta+\eta_t$, where $s_t^2=n^{-1}\sum_{i=1}^n x_{it}^2$

In what follows, we consider different parametric distributions of the data in order to assess the robustness of the AMM estimator with normal errors. We also consider different values for (N,T) and different dispersion coefficients for the AMM estimator $\nu=(0,0.05,0.1,0.5,1)$

Results

Here we present our simulation results. In the following table, we present results for bias and RMSE for GMM and AMM. In particular, we consider $N=500$ and $T=5,10,15,20,30$ (which determines the number of moment conditions). For the DGPs, we consider the Student and Log–Normal distributions[3]. In the following table, each column represents a different value of T. We present results for various cases of GMM: 1–step, 2–step, iterated (IT), continuous–updating (CUE), and diagonal (sets off-diagonal elements of the weighting matrix in the 2–step case to zero).

	ν	T=1	T=5	T=10	T=15	T=20	T=30
AMM	0	**0.006**	-0.011	-0.014	-0.015	-0.016	-0.016
	0.05	0.006	**-0.009**	**0.008**	**0.004**	**-0.002**	**-0.003**
	0.1	0.006	-0.009	-0.012	-0.011	-0.013	-0.01
	0.5	0.006	-0.01	-0.012	-0.014	-0.014	-0.014
	1	0.007	-0.007	-0.009	-0.01	-0.011	-0.011
GMM	**1–step**	**-0.003**	**0.001**	**0**	**0**	**-0.001**	**0**
	2–step	-0.003	-0.038	-0.042	-0.044	-0.046	-0.045
	IT	-0.003	-0.038	-0.042	-0.044	-0.046	-0.045
	CUE	-0.003	-0.039	-0.042	-0.044	-0.046	-0.045
	Diagonal W	-0.003	-0.039	-0.042	-0.044	-0.046	-0.045

Table 2: Bias based on 500 simulations, sample size = 500. Student–t with 3 degrees of freedom. ν denotes the noise coefficient of the AMM estimator

[3] more results can be found in the appendix

	ν	T=1	T=5	T=10	T=15	T=20	T=30
AMM	0	**-0.006**	-0.058	-0.236	-0.324	-0.28	-0.159
	0.05	-0.006	-0.049	-0.097	**-0.088**	**-0.063**	**-0.039**
	0.1	-0.006	-0.017	-0.091	-0.11	-0.073	-0.048
	0.5	-0.006	**0.003**	**-0.088**	-0.11	-0.117	-0.112
	1	-0.006	-0.052	-0.096	-0.108	-0.111	-0.116
GMM	**1–step**	**-0.003**	**-0.002**	**-0.013**	**-0.006**	**-0.007**	**-0.006**
	2–step	-0.003	-0.209	-0.236	-0.238	-0.243	-0.247
	IT	-0.003	-0.209	-0.236	-0.238	-0.243	-0.247
	CUE	-0.003	-0.211	-0.24	-0.242	-0.247	-0.253
	Diagonal W	-0.003	-0.211	-0.24	-0.242	-0.247	-0.253

Table 3: Bias based on 500 simulations, sample size = 500. Log–Normal distribution. ν denotes the noise coefficient of the AMM estimator

From this first of results, we see in the case of the Student–t, all AMM estimators regardless of ν display at most 0.01 of bias. GMM estimators, on the other hand, display consistently a bias that is five times larger (except for 1–step GMM, which we already discussed is the MLE estimator in this setting). Moving on to the more challenging case of the Log–Normal distribution, we see that AMM's performance vary with ν. In particular, we see that small but strictly positive values of ν (0.05 and 0.1) yield parameter estimates that have almost no bias. In contrast, most GMM estimators display severe bias in this case, which again is almost five times larger than AMM.

	ν	T=1	T=5	T=10	T=15	T=20	T=30
AMM	0	0.059	0.026	0.022	0.021	0.021	0.019
	0.05	0.059	0.048	0.133	0.113	0.084	0.07
	0.1	0.059	0.047	0.047	0.051	0.04	0.048
	0.5	0.059	0.025	0.021	0.02	0.019	0.019
	1	0.059	0.024	0.019	0.017	0.016	0.015
GMM	1–step	0.121	0.054	0.038	0.032	0.028	0.024
	2–step	0.121	0.06	0.053	0.052	0.052	0.049
	IT	0.121	0.06	0.053	0.052	0.052	0.049
	CUE	0.121	0.06	0.053	0.052	0.052	0.05
	Diagonal W	0.121	0.06	0.053	0.052	0.052	0.05

Table 4: RMSE based on 500 simulations, sample size = 500. Student–t with 3 degrees of freedom. ν denotes the noise coefficient of the AMM estimator

	ν	T=1	T=5	T=10	T=15	T=20	T=30
AMM	0	0.169	0.203	0.348	0.415	0.36	0.197
	0.05	0.169	0.222	0.181	0.22	0.267	0.282
	0.1	0.169	0.261	0.186	0.171	0.247	0.263
	0.5	0.169	0.296	0.185	0.154	0.151	0.159
	1	0.169	0.214	0.159	0.128	0.13	0.127
GMM	1–step	0.369	0.198	0.133	0.104	0.099	0.093
	2–step	0.369	0.238	0.25	0.247	0.25	0.252
	IT	0.369	0.238	0.25	0.247	0.25	0.252
	CUE	0.369	0.241	0.254	0.251	0.255	0.258
	Diagonal W	0.369	0.241	0.254	0.251	0.255	0.258

Table 5: RMSE based on 500 simulations, sample size = 500. Log–Normal distribution. ν denotes the noise coefficient of the AMM estimator

Moving on to the second set of results, we see that, for the Student–t case, the performance of both sets of estimators is quite similar, except for AMM with $\nu = 1$. This last case provides an important insight: The choice of the dispersion coefficient entails a bias–variance tradeoff. In particular, larger values of ν reduce the variance of the estimator at the expense of some bias inflation.

4.2 Dynamic Panel Data Model

Next, we consider the following dynamic panel data model,

$$Y_{it} = \rho Y_{it-1} + \alpha_i + \epsilon_{it}$$

where $\alpha_i, \epsilon_{it} \overset{iid}{\sim} N(0,1)$. Our parameter of interest is $\theta = \rho$. For estimation, we consider the moment conditions from Arellano and Bond (1991), which require differencing out the fixed effects:

$$\Delta Y_{it} = \rho \Delta Y_{it-1} + \Delta \epsilon_{it}$$

and use the set of lagged levels and differences as instruments. This procedure then yields the following moment conditions:

$$g_L\left(Y_i, t, \theta\right) = Y_i^{t-2} \cdot \left(\Delta Y_{i,t} - \theta \Delta Y_{i,t-1}\right)$$
$$g_D\left(Y_i, t, \theta\right) = \Delta Y_i^{t-2} \cdot \left(\Delta Y_{i,t} - \theta \Delta Y_{i,t-1}\right)$$

where $Y_i^s \equiv (Y_{i,1}, Y_{i,2}, \ldots, Y_{i,s})$ and the dot product $(\cdot)$ represents an element–by–element multiplication of the vector Y_i^{t-2} with the scalar $(\Delta Y_{i,t} - \theta \Delta Y_{i,t-1})$. Moreover, these moment conditions are defined for $t \geq 3$. This, in turn, implies that moment conditions increase at a quadratic rate in T, which makes this also a good framework to test our estimator. For this exercise, we consider a sample size of $N = 500$ and different values of the autoregressive coefficient $\rho = (0.7, 0.9, 0.95)$ and of $T = (10, 15, 20)$

Results

Below we present the estimation results for the bias and RMSE for $N = 500$ observations over $S = 500$ simulations. The main insight to take from this exercise is that, first, a small amount of noise is preferred for an optimal performance of AMM. Second, and most importantly, AMM improves its performance as we increase the number of moment conditions, whereas GMM's performance is clearly hurt as we increase T.

		ν	T=10	T=15	T=20
$\rho = 0.7$	GMM	1–step	-0.249	-0.306	-0.377
	GMM	2–step	-0.054	-0.101	-0.186
	AMM	0	-0.029	**-0.012**	-0.207
	AMM	0.05	**-0.028**	**-0.012**	-0.135
	AMM	0.1	-0.029	**-0.012**	-0.025
	AMM	0.5	-0.039	-0.026	**-0.007**
	AMM	1	-0.058	-0.046	-0.021
$\rho = 0.9$	GMM	1–step	-0.556	-0.636	-0.649
	GMM	2–step	-0.303	-0.371	-0.426
	AMM	0	-0.177	-0.079	-0.063
	AMM	0.05	**-0.175**	**-0.077**	-0.038
	AMM	0.1	-0.177	-0.079	**-0.027**
	AMM	0.5	-0.211	-0.12	-0.057
	AMM	1	-0.26	-0.168	-0.09
$\rho = 0.95$	GMM	1–step	-0.82	-0.809	-0.86
	GMM	2–step	-0.679	-0.659	-0.739
	AMM	0	-0.495	-0.27	-0.165
	AMM	0.05	**-0.491**	**-0.266**	**-0.153**
	AMM	0.1	-0.497	-0.271	-0.159
	AMM	0.5	-0.544	-0.355	-0.235
	AMM	1	-0.593	-0.443	-0.34

Table 6: Bias based on 500 simulations, sample size = 500. ν denotes the noise coefficient of the AMM estimator

		ν	T=10	T=15	T=20
$\rho = 0.7$	GMM	1–step	0.293	0.332	0.396
	GMM	2–step	0.076	0.116	0.2
	AMM	0	**0.048**	**0.03**	0.285
	AMM	0.05	**0.048**	**0.03**	0.23
	AMM	0.1	**0.048**	0.031	0.101
	AMM	0.5	0.061	0.043	**0.028**
	AMM	1	0.083	0.064	0.044
$\rho = 0.9$	GMM	1–step	0.656	0.691	0.682
	GMM	2–step	0.387	0.431	0.459
	AMM	0	0.217	0.095	0.177
	AMM	0.05	**0.215**	**0.093**	0.113
	AMM	0.1	0.216	0.096	**0.047**
	AMM	0.5	0.25	0.137	0.076
	AMM	1	0.309	0.191	0.113
$\rho = 0.95$	GMM	1–step	0.914	0.863	0.894
	GMM	2–step	0.776	0.724	0.779
	AMM	0	0.611	**0.363**	0.304
	AMM	0.05	**0.608**	**0.363**	**0.29**
	AMM	0.1	0.612	0.366	0.298
	AMM	0.5	0.64	0.43	0.321
	AMM	1	0.677	0.51	0.436

Table 7: RMSE based on 500 simulations, sample size = 500. ν denotes the noise coefficient of the AMM estimator

5 Theory

Asymptotic properties of the AMM estimator were first derived by Kaji, Manresa, and Pouliot (2020), providing conditions for consistency and asymptotic normality. Moreover, they showed that AMM is asymptotically equivalent to optimally–weighted SMM. In this section, we will derive this result under weaker conditions, leveraging in the link between AMM and GEL estimators.

The intuition for the asymptotic equivalence with SMM can be seen from the FOC of the discriminator, which emphasizes that D searches for the value of λ that matches the weighted averages of $g\left(x_i, \theta\right)$ and $g\left(\widetilde{x}_i\left(\theta\right), \theta\right)$. Under correct specification of the moment conditions, we have $\mathbb{E}\left[g\left(x_i, \theta_0\right)\right] = \mathbb{E}\left[g\left(\widetilde{x}_i\left(\theta_0\right), \theta_0\right)\right]$, so $\lambda\left(\theta_0\right) = 0$ would be a solution. In fact, by concavity of the objective function with respect to λ, it is the only solution. In that case, since $\Lambda\left(0\right) = 1 - \Lambda\left(0\right) = 1/2$, the FOC boils down to

$$\frac{1}{n}\sum_{i=1}^{n} g\left(x_i, \theta_0\right) = \frac{1}{m}\sum_{i=1}^{m} g\left(\widetilde{x}_i\left(\theta_0\right), \theta_0\right)$$

so that $\widehat{\theta}_{AMM}$ will indeed match the two sets of moments.

5.1 The link between AMM and GEL estimators

Recall that the GEL estimator is the solution to the saddle point problem

$$\hat{\theta}_{\text{GEL}} = \arg\min_{\theta\in\Theta}\left\{\sup_{\lambda\in\hat{\mathcal{B}}_n(\theta)}\sum_{i=1}^{n}\rho\left(\lambda' g_i\left(\theta\right)\right)\right\}$$

The AMM estimator, on the other hand, solves

$$\widehat{\theta}_{AMM} = \arg\min_{\theta\in\Theta}\left\{\max_{\lambda\in\mathbb{R}^d}\frac{1}{n}\sum_{i=1}^{n}\log\Lambda\left(\lambda' g\left(x_i, \theta\right)\right) + \frac{1}{m}\sum_{i=1}^{m}\log\left(1 - \Lambda\left(\lambda' g\left(\widetilde{x}_i\left(\theta\right), \theta\right)\right)\right)\right\}$$

In what follows, we will state conditions for consistency and asymptotic normality, which are necessary for stochastic expansions. Since the two versions of the estimators require slightly different sets of assumptions, we will first characterize the ones that are shared by both, and then develop each in more detail.

5.2 Assumptions and Results

In what follows, we will provide conditions for consistency, asymptotic normality and existence of the stochastic expansion for AMM estimators. For consistency, we require only standard assumptions

Assumption 1 *(a) $\theta_0 \in \Theta$ is the unique solution to $\mathbb{E}\left[g\left(x,\theta\right)\right] = \mathbb{E}\left[g\left(x^{\theta},\theta\right)\right]$ (b) Θ is compact; (c) $g\left(x,\theta\right)$ is continuous at each $\theta \in \Theta$ with probability one; (d) $\mathbb{E}\left[\sup_{\theta\in\Theta}\|g\left(x,\theta\right)\|^{\alpha}\right] < \infty$ for some $\alpha > 2$; (e) Ω is nonsingular.*

Theorem 5.1 *If assumption A1 is satified, then $\widehat{\theta} \xrightarrow{p} \theta_0$*

Additional assumptions are needed for asymptotic normality:

Assumption 2 *(a) $\theta_0 \in int\left(\Theta\right)$; (b) $g\left(x,\theta\right)$ is continuously differentiable in a neighborhood $\mathcal{N}$ of θ_0 and $\mathbb{E}\left[\sup_{\theta\in\mathcal{N}}\|\partial g_i\left(\theta\right)/\partial\theta'\|\right] < \infty$; (c) $rank\left(G\right) = p$.*

Theorem 5.2 *If assumptions A1 and A2 are satisfied, moment–based AMM*

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{\mathcal{D}} N\left(0,\Sigma\right)$$

Moreover, simulation–based AMM is asymptotically equivalent to SMM

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{\mathcal{D}} N\left(0,(1+\tau)\Sigma\right)$$

where $\tau = \frac{n}{m}$

The reason we don't need to impose additional assumption is that the particular structure of GAN with a logistic discriminator guarantees a good behavior of our problem.

Consistency and asymptotic normality are necessary conditions for stochastic expansions, the device to analyze finite–sample properties of our estimator. Under further additional assumptions, the AMM estimator admits the following expansion:

$$\sqrt{n}\left(\hat{\theta} - \theta_0\right) = \tilde{\psi} + Q_1\left(\tilde{\psi},\tilde{a},F_0\right)/\sqrt{n} + Q_2\left(\tilde{\psi},\tilde{a},\tilde{b},F_0\right)/n + R_n$$

where Q_1 is quadratic in its first two arguments, Q_2 is cubic in its first three arguments, and $R_n = O_p\left(n^{-3/2}\right)$. The particular expression for this expansion is given in the appendix, but the asymptotic (higher order) bias formula is given by

$$\text{Bias}(\hat{\theta}) = \mathbb{E}\left[Q_1\left(\psi_i,a_i,F_0\right)\right]/n$$

The specific assumptions required for this result can be found in the appendix. The way we are able to prove the above results is by casting the AMM estimator as a simulation–based GEL estimator. With this, we leverage on the vast set of results in Newey and Smith (2004) to arrive to our theoretical results.

6 Conclusion

In this paper, we investigated properties of the AMM estimator for models characterized by moments conditions. We describe the intuition behind the estimator, and provided asymptotic as well as finite sample results. We used stochastic expansions to characterize the finite–sample properties of AMM. In particular, we showed that AMM is asymptotically equivalent to optimally–weighted GMM, but it displays better finite performance. In the second part of the paper, we put our estimator to work under different frameworks to illustrate its performance. In particular, we showed that, although AMM and optimally–weighted SMM are asymptotically equivalent, their performance differs in small samples.

A Asymptotic Normality

In what follows, we'll derive expressions for the asymptotic distribution of the AMM estimator. In general, we denote

$$\begin{aligned} &\Sigma = \left(G'\Omega^{-1}G\right)^{-1},\ H = \Sigma G'\Omega^{-1} \\ &P = \Omega^{-1} - \Omega^{-1}G\Sigma G'\Omega^{-1} \end{aligned}$$

Moreover, we'll use the following results:

$$H\Omega H' = \Sigma,\ P\Omega H' = 0,\ P\Omega P = P$$

A mean–value expansion of the FOC gives

$$0 = -\left(\begin{array}{c} 0 \\ \widehat{g}\left(\theta_0\right) - \nu\widehat{\epsilon} \end{array}\right) + \overline{M}\left(\widehat{\varphi} - \varphi_0\right)$$

where

$$\begin{aligned} \overline{M} = &\left(\begin{array}{cc} 0 & \sum_{i=1}^{n}\rho_1^D\left(\widehat{\lambda}'\widehat{g}_i\right)G_i\left(\widehat{\theta}\right)'/n \\ \sum_{i=1}^{n}\rho_1^D\left(\overline{\lambda}'\widehat{g}_i\right)G_i\left(\overline{\theta}\right)/n & \sum_{i=1}^{n}\rho_2^D\left(\overline{\lambda}'\widehat{g}_i\right)g_i\left(\overline{\theta}\right)\widehat{g}_i'/n \end{array}\right) \\ &+\left(\begin{array}{cc} 0 & 0 \\ 0 & \sum_{j=1}^{m}\rho_2^S\left(\overline{\lambda}'\nu\epsilon_j\right)\nu^2\epsilon_j\epsilon_j'/m \end{array}\right) \end{aligned}$$

where the last term is the additional piece from the random draws, with the property that $\sum_{j=1}^{m}\rho_2^S\left(\bar{\lambda}'\nu\epsilon_j\right)\nu^2\epsilon_j\epsilon_j'/m \to \nu^2 I$ as $m \to \infty$. Moreover, we have that $\overline{M} \to M$ with

$$M = -\left(\begin{array}{cc} 0 & G' \\ G & \Omega_M \end{array}\right), \quad M_M^{-1} = -\left(\begin{array}{cc} -\Sigma_M & H_M \\ H_M' & P_M \end{array}\right)$$

where

$$\begin{aligned} &\Omega_M = \Omega + \nu^2 I,\ \Sigma_M = \left(G'\Omega_M^{-1}G\right)^{-1},\ H_M = \Sigma_M G'\Omega_M^{-1} \\ &P_M = \Omega_M^{-1} - \Omega_M^{-1}G\Sigma_M G'\Omega_M^{-1} \end{aligned}$$

Under standard assumptions, we have that

$$\sqrt{n}\left(\begin{array}{c} \widehat{g} \\ \nu\widehat{\epsilon} \end{array}\right) \longrightarrow N\left(\left(\begin{array}{c} 0 \\ 0 \end{array}\right), \left(\begin{array}{cc} \Omega & 0 \\ 0 & \tau\nu^2 I \end{array}\right)\right)$$

Under this framework, increasing the number of synthetic observations do not improve accuracy, since we already have $(\mu_\epsilon, \sigma_\epsilon^2) = (0, 1)$. We thus restrict to the case where $m = n$, so that $\tau = 1$. By means of Slutsky's Theorem, we get

$$\sqrt{n}\left(\widehat{g}-\nu\widehat{\epsilon}\right)\longrightarrow N\left(0,\Omega_M\right)$$

which implies that

$$\sqrt{n}\left(\widehat{\varphi}-\varphi_0\right)=M^{-1}\left(\begin{array}{c}0\\ \sqrt{n}\left(\widehat{g}-\nu\widehat{\epsilon}\right)\end{array}\right)+o_p\left(1\right)$$

and thus

$$\sqrt{n}\left(\widehat{\varphi}-\varphi_0\right)\longrightarrow N\left(0,\operatorname{diag}\left(\Sigma_M,P_M\right)\right)$$

Moreover, the case with $\nu=0$ is asymptotically equivalent to optimally–weighted GMM, since $\Omega_M\left(\nu=0\right)=\Omega$ so that $\Sigma_M\left(\nu=0\right)=\left(G'\Omega^{-1}G\right)^{-1}$

B Stochastic expansion for AMM

B.1 Taylor Expansion

The stochastic expansion for AMM is similar to Newey and Smith (2004), Lemma A4. We start with the following notation:

$$\begin{array}{l}M_j=\mathbb{E}\left[\partial^2 m\left(x,\gamma_0\right)/\partial\gamma_j\partial\gamma\right],\quad M_{jk}=\mathbb{E}\left[\partial^3 m\left(x,\gamma_0\right)/\partial\gamma_k\partial\gamma_j\partial\gamma\right]\\ A\left(x\right)=\partial m\left(x,\gamma_0\right)/\partial\gamma-M,\quad B_j\left(x\right)=\partial^2 m\left(x,\gamma_0\right)/\partial\gamma_j\partial\gamma-M_j\\ \psi\left(x\right)=-M^{-1}m\left(x,\gamma_0\right),\quad a\left(x\right)=vec\left(A\left(x\right)\right),\quad b\left(x\right)=vec\left[B_1\left(x\right),\ldots,B_q\left(x\right)\right]\end{array}$$

Also, for any operator $T\left(x;\gamma\right)$, let

$$\begin{array}{l}\widehat{T}\left(\gamma\right)\equiv n^{-1}\sum T\left(x_i;\gamma\right),\ \ \widetilde{T}\left(\gamma\right)\equiv n^{-1/2}\widehat{T}\left(\gamma\right)\\ T\left(\gamma\right)\equiv\mathbb{E}\left[T\left(x;\gamma\right)\right],\ \ T\equiv T\left(\gamma_0\right)\end{array}$$

A Taylor expansion with Lagrange remainder gives

$$\begin{aligned}0=&\widehat{m}+\widehat{M}\left(\widehat{\gamma}-\gamma_0\right)+\sum_{j=1}^{q}\left(\widehat{\gamma}_j-\gamma_{j0}\right)\left[\partial\widehat{M}\left(\gamma_0\right)/\partial\gamma_j\right]\left(\widehat{\gamma}-\gamma_0\right)/2\\ &+\sum_{j,k=1}^{q}\left(\widehat{\gamma}_j-\gamma_{j0}\right)\left(\widehat{\gamma}_k-\gamma_{k0}\right)\left[\partial^2\widehat{M}\left(\overline{\gamma}\right)/\partial\gamma_k\partial\gamma_j\right]\left(\widehat{\gamma}-\gamma_0\right)/6\end{aligned}$$

Adding and substracting in the second and third terms gives

$$\begin{aligned}\widehat{\gamma}-\gamma_0=&\,n^{-1/2}\widetilde{\psi}-M^{-1}\left(n^{-1/2}\widetilde{A}\right)\left(\widehat{\gamma}-\gamma_0\right)\\ &-M^{-1}\sum_{j=1}^{q}\left(\widehat{\gamma}_j-\gamma_{j0}\right)\left[M_j+n^{-1/2}\widetilde{B}_j\right]\left(\widehat{\gamma}-\gamma_0\right)/2\\ &-M^{-1}\sum_{j,k=1}^{q}\left(\widehat{\gamma}_j-\gamma_{j0}\right)\left(\widehat{\gamma}_k-\gamma_{k0}\right)M_{jk}\left(\widehat{\gamma}-\gamma_0\right)/6+O_p\left(n^{-2}\right)\end{aligned}$$

which follows from the fact that

$$\begin{aligned} -M^{-1}\widehat{m} &= n^{-1/2}\left[n^{-1/2}\sum\left(-M^{-1}m\left(x_i;\gamma_0\right)\right)\right] = n^{-1/2}\widetilde{\psi} \\ \left(\widehat{M}-M\right) &= n^{-1/2}\left[n^{-1/2}\sum\left(\partial m\left(x_i;\gamma\right)/\partial\gamma - M\right)\right] = n^{-1/2}\widetilde{A} \\ \partial\widehat{M}\left(\gamma_0\right)/\partial\gamma_j \pm M_j &= M_j + n^{-1/2}\left[n^{-1/2}\sum\left(\partial^2 M\left(\gamma_0\right)/\partial\gamma_j\partial\gamma - M_j\right)\right] = M_j + n^{-1/2}\widetilde{B}_j \\ \left\|\partial^2\widehat{M}\left(\overline{\gamma}\right)/\partial\gamma_k\partial\gamma_j - M_{jk}\right\| &= O_p\left(n^{-1/2}\right) \end{aligned}$$

As all the terms except $n^{-1/2}\widetilde{\psi}$ are $O_p\left(n^{-1}\right)$, it follows that

$$\widehat{\gamma}-\gamma_0 = n^{-1/2}\widetilde{\psi} + O_p\left(n^{-1}\right)$$

Next, since the last three terms (including the remainder) are $O_p\left(n^{-3/2}\right)$, and replacing $\widehat{\gamma}-\gamma_0$ by $n^{-1/2}\widetilde{\psi}$ in the second and third terms also generates an error that is $O_p\left(n^{-3/2}\right)$, we obtain

$$\begin{aligned} \widehat{\gamma}-\gamma_0 &= n^{-1/2}\widetilde{\psi} - n^{-1}M^{-1}\left[\widetilde{A}\widetilde{\psi} + \sum_{j=1}^{q}\widetilde{\psi}_j M_j\widetilde{\psi}/2\right] + O_p\left(n^{-3/2}\right) \\ &= n^{-1/2}\widetilde{\psi} + n^{-1}Q_1\left(\widetilde{\psi},\widetilde{a}\right) + O_p\left(n^{-3/2}\right) \end{aligned}$$

Finally, replacing $\widehat{\gamma}-\gamma_0$ in the second and third terms by the above expression, and in the fourth and fifth terms by $n^{-1/2}\widetilde{\psi}$ gives

$$n^{-1/2}\left(\widehat{\gamma}-\gamma_0\right) = \widetilde{\psi} + n^{-1/2}Q_1\left(\widetilde{\psi},\widetilde{a},F_0\right) + n^{-1}Q_2\left(\widetilde{\psi},\widetilde{a},\widetilde{b},F_0\right) + R_n$$

where Q_1 is quadratic in its first two arguments, Q_2 is cubic in its first three arguments, and $R_n = O_p\left(n^{-3/2}\right)$. In particular

$$\begin{aligned} Q_1\left(\widetilde{\psi},\widetilde{a}\right) &= -M^{-1}\left[\widetilde{A}\widetilde{\psi} + \textstyle\sum_{j=1}^{q}\widetilde{\psi}_j M_j\widetilde{\psi}/2\right] \\ Q_2\left(\widetilde{\psi},\widetilde{a},\widetilde{b}\right) &= -M^{-1}\left[\widetilde{A}Q_1\left(\widetilde{\psi},\widetilde{a},\widetilde{b}\right) + \textstyle\sum_{j,k=1}^{q}\widetilde{\psi}_j\widetilde{\psi}_k M_{jk}\widetilde{\psi}/6\right] \\ -M^{-1}\textstyle\sum_{j=1}^{q} & \left\{\widetilde{\psi}_j M_j Q_1\left(\widetilde{\psi},\widetilde{a}\right) + Q_{1j}\left(\widetilde{\psi},\widetilde{a}\right)M_j\widetilde{\psi} + \widetilde{\psi}_j\widetilde{B}_j\widetilde{\psi}\right\}/2 \end{aligned}$$

B.2 Characterization for AMM

Denote $\varphi = \left(\lambda,\theta\right), G_i\left(\theta\right) = \frac{\partial g_i(\theta)}{\partial\theta}$ and $m^{\ell}\left(x_i,\varphi\right), \ell = T, S$ denote each of the FONC terms of the AMM estimator,

$$m^{\ell}\left(x_i^{\ell}\left(\theta\right),\varphi\right) = \rho_1^{\ell}\left(\lambda' g_i\left(\theta\right)\right)\begin{pmatrix} G_i^{\ell}\left(\theta\right)'\lambda \\ g_i^{\ell}\left(\theta\right) \end{pmatrix}$$

with $\rho^T\left(x\right) \equiv \log\Lambda\left(x\right)$, $\rho^S\left(x\right) \equiv \log\left(1-\Lambda\left(x\right)\right)$ and $\Lambda\left(x\right) = \left(1+e^{-x}\right)^{-1}$. We can write the FONC for AMM similarly as the ones for GEL estimators in Newey and

Smith (2004), the only difference being that the appropriate moment condition is now the sum of the true and synthetic moments

$$m\left(x_i, x_i\left(\theta\right), \varphi\right) = m^T\left(x_i, \varphi\right) + m^S\left(\widetilde{x}_i\left(\theta\right), \varphi\right)$$

where $\rho^T\left(0\right) = 1 - \rho^S\left(0\right) = 1/2$ Moreover, from Newey and Smith (2004), Theorem 4.2, we have that

$$\begin{aligned}
\mathbb{E}\left[A\left(z_i\right)\psi_i\right] &= \begin{pmatrix} \mathbb{E}\left[G_i' P g_i\right] \\ \mathbb{E}\left[G_i H g_i + g_i g_i' P g_i\right] \end{pmatrix} \\
\sum_{j=1}^{q} M_j \mathbb{E}\left[\psi_i \psi_i'\right] e_j/2 &= \sum_{j=1}^{p} M_j[\Sigma, 0]' e_j/2 + \sum_{j=1}^{m} M_{j+p}[0, P]' e_j/2 \\
&= -\sum_{j=1}^{p} \begin{pmatrix} 0 \\ \mathbb{E}\left[G_i^j\right] \Sigma e_j/2 \end{pmatrix} - \sum_{j=1}^{m} \begin{pmatrix} \mathbb{E}\left[G_i' e_j g_i' + g_{ij} G_i'\right] P e_j/2 \\ 0 \end{pmatrix} \\
&= \begin{pmatrix} -\mathbb{E}\left[G_i' P g_i\right] \\ -a \end{pmatrix}
\end{aligned}$$

thus, Bias($\hat{\theta}$) are the first p elements of

$$\mathbb{E}\left[Q_1\left(\psi_i, a_i, F_0\right)\right]/n = -M^{-1}\left(-a + \mathbb{E}\left[G_i H g_i\right] + \mathbb{E}\left[g_i g_i' P g_i\right]\right)/n$$

B.3 Further details

Let $\rho_i^\ell\left(x\right) \equiv \partial^i \rho\left(x\right)/\partial x^i, \ell = \left(T, S\right)$ denote the i^{th} derivative of ρ^j, and $\rho_i^\ell = \rho_i^\ell\left(0\right)$. Denote also $v_i^\ell\left(\varphi\right) \equiv \lambda' g_i^\ell\left(\theta\right)$, $h_i^\ell\left(\varphi\right) \equiv \partial v_i^\ell\left(\varphi\right)/\partial\varphi$, so each FONC term can be expressed as $m_i^\ell\left(\varphi\right) = \rho_1^\ell\left(v_i^\ell\left(\varphi\right)\right) h_i^\ell\left(\varphi\right)$. From this characterization we have that

$$\begin{aligned}
\partial m_i\left(\varphi\right)/\partial\varphi &= \rho_2\left(v_i\left(\varphi\right)\right) h_i\left(\varphi\right) h_i\left(\varphi\right)' + \rho_1\left(v_i\left(\varphi\right)\right) \partial h_i\left(\varphi\right)/\partial\varphi \\
\partial^2 m_i\left(\varphi\right)/\partial\varphi_j\partial\varphi &= \rho_3\left(v_i\left(\varphi\right)\right) h_i\left(\varphi\right){}_j h_i\left(\varphi\right) h_i\left(\varphi\right)' + \rho_2\left(v_i\left(\varphi\right)\right) \partial\left[h_i\left(\varphi\right) h_i\left(\varphi\right)'\right]/\partial\varphi_j \\
&\quad + \rho_2\left(v_i\left(\varphi\right)\right) h_i\left(\varphi\right)_j \partial h_i\left(\varphi\right)/\partial\varphi + \rho_1\left(v_i\left(\varphi\right)\right) \partial^2 h_i\left(\varphi\right)/\partial\varphi_j\partial\varphi \\
\partial^3 m_i\left(\varphi\right)/\partial\varphi_k\partial\varphi_j\partial\varphi &= \rho_4\left(v_i\left(\varphi\right)\right) h_i\left(\varphi\right)_k h_i\left(\varphi\right)_j h_i\left(\varphi\right) h_i\left(\varphi\right)' + \rho_3\left(v_i\left(\varphi\right)\right) \partial\left[h_i\left(\varphi\right)_j h_i\left(\varphi\right) h_i\left(\varphi\right)'\right]/\partial\varphi_k \\
&\quad + \rho_3\left(v_i\left(\varphi\right)\right) h_i\left(\varphi\right)_k \partial\left[h_i\left(\varphi\right) h_i\left(\varphi\right)'\right]/\partial\varphi_j + \rho_2\left(v_i\left(\varphi\right)\right) \partial^2\left[h_i\left(\varphi\right) h_i\left(\varphi\right)'\right]/\partial\varphi_k\partial\varphi_j \\
&\quad + \rho_3\left(v_i\left(\varphi\right)\right) h_i\left(\varphi\right)_k h_i\left(\varphi\right)_j \partial h_i\left(\varphi\right)/\partial\varphi + \rho_2\left(v_i\left(\varphi\right)\right) \partial\left[h_i\left(\varphi\right)_j \partial h_i\left(\varphi\right)/\partial\varphi\right]/\partial\varphi_k \\
&\quad + \rho_2\left(v_i\left(\varphi\right)\right) h_i\left(\varphi\right)_k \partial^2 h_i\left(\varphi\right)/\partial\varphi_j\partial\varphi + \rho_1\left(v_i\left(\varphi\right)\right) \partial^3 h_i\left(\varphi\right)/\partial\varphi_k\partial\varphi_j\partial\varphi
\end{aligned}$$

which are the key terms that enter the stochastic expansion.

Before we continue, let us impose the following normalization: $\rho^\ell\left(x\right) = \rho^\ell\left(2x\right)$. In turn, this implies that

$$\begin{aligned}
\rho_1 &\equiv \rho_1^S = -\rho_1^D = -1 \\
\rho_2 &\equiv \rho_2^S = \rho_2^D = -1 \\
\rho_3 &\equiv \rho_3^S = \rho_3^D = 0 \\
\rho_4 &\equiv \rho_4^S = \rho_4^D = 2
\end{aligned}$$

so that the first two moments are normalized to unity, as in Newey and Smith (2004), proof of Theorem 3.4, with the only difference on the sign of ρ_1^D. This will be taken care of below.

We need to compute two terms: the one involving the true data, and the one involving the random draws. We start by computing the terms of $m_i^D(\varphi_0)$. Given our normalization, the derivation of this term is exactly the same as in GEL.

The second term, which involves the random draws, is similar to m_i^D terms in the simulation–based approach; since neither involve θ, only few terms survive. We have that, for $u = k - p, t = j - p$

$$\begin{aligned}
h_i^S \left(h_i^S\right)' &= \begin{bmatrix} 0 & 0 \\ 0 & \nu^2 \epsilon_i \epsilon_i' \end{bmatrix} \\
\left(h_i^S\right)_j \cdot h_i^S \left(h_i^S\right)' &= \begin{bmatrix} 0 & 0 \\ 0 & \nu^3 \epsilon_{ij} \epsilon_i \epsilon_i' \end{bmatrix} \quad j > p \\
\left(h_i\right)_k \left(h_i^S\right)_j \cdot h_i^S \left(h_i^S\right)' &= \nu^4 \epsilon_{iu} \epsilon_{it} \begin{bmatrix} 0 & 0 \\ 0 & \epsilon_i \epsilon_i' \end{bmatrix} \quad j, k > p
\end{aligned}$$

thus

$$\begin{aligned}
\partial m_i^S(\varphi) / \partial \varphi &= -\begin{bmatrix} 0 & 0 \\ 0 & \nu^2 \epsilon_i \epsilon_i' \end{bmatrix} \\
\partial^2 m_i^S(\varphi) / \partial \varphi_j \partial \varphi &= 0 \\
\partial^3 m_i^S(\varphi) / \partial \varphi_k \partial \varphi_j \partial \varphi &= 2\nu^2 \epsilon_{iu} \epsilon_{it} \begin{bmatrix} 0 & 0 \\ 0 & \nu^2 \epsilon_i \epsilon_i' \end{bmatrix} \quad j, k > p
\end{aligned}$$

And recall that $\mathbb{E}\left[\epsilon_i \epsilon_i'\right]$ is just the identity matrix.

C AMM implementation

We now describe the procedure for AMM estimation. Before we begin, we must first choose a distribution $\mathcal{F}$ for the synthetic observations $\{\epsilon_i\}_{i=1}^n$. For simplicity, we choose the standard normal, so $\mathcal{F} = \mathcal{N}(0, 1)$, but any distribution such that $\mathbb{E}[\epsilon] = 0, \mathbb{V}[\epsilon] = 1$ would work. Moreover, we must choose an initial guess $\theta_{(0)}$.

1. Fix the random number generator by drawing a sample of i.i.d. shocks $\epsilon_i \sim \mathcal{F}$ with the same size as the true data. Scale the shocks with ν to get our synthetic data. These observations are held fix throughout iterations.

2. At each step s, construct the dataset

$$\left[\mathbf{X}_{(s)}|\mathbf{d}\right] = \left[\begin{array}{cc|c} 1 & g\left(x_1, \theta_{(s)}\right)' & 1 \\ 1 & g\left(x_2, \theta_{(s)}\right)' & 1 \\ \vdots & \vdots & \vdots \\ 1 & g\left(x_n, \theta_{(s)}\right)' & 1 \\ 1 & \nu\varepsilon_1' & 0 \\ 1 & \nu\varepsilon_2' & 0 \\ \vdots & \vdots & \vdots \\ 1 & \nu\varepsilon_n' & 0 \end{array}\right]$$

   and run the discriminator to get $\lambda_{(s)}$.

3. Update each parameter θ_j $j \leq p$ using gradient descent: Construct new moment observations $\left\{g\left(x_i, \theta^{\pm}_{(s)}\right)\right\}_{i=1}^{n}$, where $\theta^{\pm}_{j,(s)} \equiv \theta_{j,(s)} \pm k_{(s)}$. Here, $k_{(s)}$ is a tuning parameter that describes the step size. We compute the numerical gradient of the loss function as

$$\widehat{\nabla\mathcal{L}}_{(s)} = \frac{\mathcal{L}_n\left(\theta^{+}_{(s)}\right) - \mathcal{L}_n\left(\theta^{-}_{(s)}\right)}{2k_{(s)}}$$

   keeping fixed the discriminator's parameters at $\lambda_{(s)}$.

4. If all numerical gradients are (close to) zero, stop. If not, update $\theta_{(s+1)} = \theta_{(s)} + \eta_{(s+1)}\widehat{\nabla\mathcal{L}}_{(s)}$ and go back to step 2.

D Extension to simulation–based estimation

Often times with structural models, moment–based estimation is not possible. In such cases, we proceed with simulated minimum–distance (SMD) estimation: We look for parameter values such that moments in the data and the model are as close as possible. To be precise, we start with some data $\{x_i\}_{i=1}^{n}$, together with certain moments $\mathbb{E}\left[g\left(x\right)\right]$ that we would like to be replicated by our model. Moreover, we can draw simulated observations, $\{\widetilde{x}_i\left(\theta\right)\}_{i=1}^{m}$, and we can construct model counterparts of $g\left(x_i\right)$, which will be denoted as $g\left(\widetilde{x}_i\left(\theta\right)\right)$. In this case, the AMM estimator can be defined as

$$\widehat{\theta}_{AMM} = \underset{\theta\in\Theta}{\arg\min}\left\{\max_{\lambda\in\mathbb{R}^d}\frac{1}{n}\sum_{i=1}^{n}\log\Lambda\left(\lambda' g\left(x_i\right)\right) + \frac{1}{m}\sum_{i=1}^{m}\log\left(1-\Lambda\left(\lambda' g\left(\widetilde{x}_i\left(\theta\right)\right)\right)\right)\right\}$$

D.1 Asymptotics

A mean–value expansion of the FOC in this case gives

$$0 = -\begin{pmatrix} 0 \\ \widehat{g} - \widehat{g}\left(\widetilde{x}\right) \end{pmatrix} + \overline{M}_S\left(\widehat{\varphi} - \varphi_0\right)$$

where $\widehat{g}\left(\widetilde{x}\right) = m^{-1}\sum_{i=1}^{m} g\left(\widetilde{x}_i\left(\theta_0\right)\right)$.

Since θ only enters when computing the simulated sample, we have that

$$\begin{aligned} \overline{M}_S = & \begin{pmatrix} 0 & 0 \\ 0 & \sum_{i=1}^{n}\rho_2^D\left(\overline{\lambda}' g_i\right) g_i g_i'/n \end{pmatrix} \\ & + \begin{pmatrix} 0 & \sum_{i=1}^{m}\rho_1^S\left(\overline{\lambda}' g_i^{\hat{\theta}}\right)\left(G_i^{\overline{\theta}}\right)'/m \\ \sum_{i=1}^{m}\rho_1^S\left(\overline{\lambda}' g_i^{\hat{\theta}}\right) G_i^{\overline{\theta}}/m & \sum_{i=1}^{m}\rho_2^S\left(\overline{\lambda}' g_i^{\hat{\theta}}\right) g_i^{\overline{\theta}}\left(g_i^{\hat{\theta}}\right)'/m \end{pmatrix} \end{aligned}$$

where $g_i^{\theta} = g\left(\widetilde{x}_i\left(\theta\right)\right), G_i^{\theta} = \frac{\partial g}{\partial x}\left(\widetilde{x}_i\left(\theta\right)\right)\cdot\frac{\partial\widetilde{x}_i}{\partial\theta}\left(\theta\right)$

In this case, we have that $\overline{M}_S \to M_S$ with

$$M_S = -\begin{pmatrix} 0 & G' \\ G & 2\Omega \end{pmatrix}, \quad M_S^{-1} = -\begin{pmatrix} -2\Sigma & H \\ H' & P/2 \end{pmatrix}$$

As it's standard in this case, we assume conditions such that

$$\sqrt{n}\begin{pmatrix} \widehat{g} - \gamma\left(\theta_0\right) \\ \widehat{g}\left(\widetilde{x}\right) - \gamma\left(\theta_0\right) \end{pmatrix} \longrightarrow N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \Omega & 0 \\ 0 & \tau\Omega \end{pmatrix}\right)$$

so that

$$\sqrt{n}\left(\widehat{g} - \widehat{g}\left(\theta_0\right)\right) \longrightarrow N\left(0, \left(1+\tau\right)\Omega\right)$$

Proceeding in the same way as before, we get that

$$\sqrt{n}\left(\widehat{\varphi} - \varphi_0\right) = M_S^{-1}\begin{pmatrix} 0 \\ \sqrt{n}\left(\widehat{g} - \widehat{g}\left(\widetilde{x}\right)\right) \end{pmatrix} + o_p\left(1\right)$$

and so

$$\sqrt{n}\left(\widehat{\varphi} - \varphi_0\right) \longrightarrow N\left(0, \left(1+\tau\right)\operatorname{diag}\left(\Sigma, P/4\right)\right)$$

since

$$\begin{aligned} M_S^{-1}\mathbb{AV}\left(\sqrt{n}\left(\widehat{g} - \widehat{g}\left(\theta_0\right)\right)\right) M_S^{-1} &= \begin{pmatrix} \left(1+\tau\right) H\Omega H' & \left(\frac{1+\tau}{2}\right) H\Omega P \\ \left(\frac{1+\tau}{2}\right) P\Omega H' & \left(\frac{1+\tau}{4}\right) P\Omega P \end{pmatrix} \\ &= \left(1+\tau\right)\operatorname{diag}\left(\Sigma, P/4\right) \end{aligned}$$

We thus conclude that AMM is asymptotically equivalent to optimally–weighted SMM.

D.2 Stochastic Expansion

We start by computing the terms of $m_i^S(\varphi_0)$. Note that since $h_i(\varphi_0) = (0', g_i')'$ and $\rho_1 = \rho_2 = -1$, we get

$$\partial m_i^S(\varphi_0)/\partial\varphi = -\begin{pmatrix} 0 & G_i' \\ G_i & g_i g_i' \end{pmatrix}$$

Now let $G_i^j = \partial^2 g_i(\theta_0)/\partial\theta_j\partial\theta$, $g_i^j = \partial g_i(\theta_0)/\partial\theta_j, t = j - p$ for $j > p$, let e_t denote the t th unit vector, and a t subscript denote the t th element of a vector. Then evaluate at $\varphi = \varphi_0$ to obtain

$$\begin{aligned}\partial^2 m_i^S(\varphi_0)/\partial\varphi_j\partial\varphi &= -\begin{pmatrix} 0 & G_i^{j\prime} \\ G_i^j & g_i^j g_i' + g_i g_i^{j'} \end{pmatrix} \quad (j \le p) \\ &= -\begin{pmatrix} \partial^2\left[e_t' g_i(\theta_0)\right]/\partial\theta\partial\theta' & G_i' e_t g_i' + g_{it} G_i' \\ g_i e_t' G_i + g_{it} G_i & 0 \end{pmatrix} \quad (j > p)\end{aligned}$$

Next, let $G_i^{jk} = \partial^3 g_i(\theta_0)/\partial\theta_k\partial\theta_j\partial\theta$ and $g_i^{jk} = \partial^2 g_i(\theta_0)/\partial\theta_k\partial\theta_j$. Then for the second derivatives corresponding to θ, with $j \le p$ and $k \le p$,

$$\partial^3 m_i^S(\varphi_0)/\partial\varphi_k\partial\varphi_j\partial\varphi = -\begin{pmatrix} 0 & G_i^{jk\prime} \\ G_i^{jk} & g_i^{jk} g_i' + g_i^j g_i^{k\prime} + g_i^k g_i^{j\prime} + g_i g_i^{jk\prime} \end{pmatrix}$$

For the cross partial between λ_t and θ_j, with $j \le p, k > p$, and $t = k - p$,

$$\begin{aligned}&\partial^3 m_i^S(\varphi_0)/\partial\varphi_k\partial\varphi_j\partial\varphi \\ &\quad = -\begin{pmatrix} \partial^3 g_{it}(\theta_0)/\partial\theta_j\partial\theta\partial\theta' & G_i' e_t g_i^{j'} + G_i^{j'} e_t g_i' + G_{itj} G_i' + g_{it} G_i^{j\prime} \\ g_i^j e_t G_i + g_i e_t G_i^j + G_{itj} G_i + g_{it} G_i^j & -\rho_3\left[G_{itj} g_i g_i' + g_{it}\left(g_i^j g_i + g_i g_i^{j\prime}\right)\right] \end{pmatrix}\end{aligned}$$

For the second partial derivatives between λ_t and λ_u, with $j > p, k > p, t = j - p$, and $u = k - p$

$$\begin{aligned}\partial^3 m_i^S(\varphi_0)/\partial\varphi_k\partial\varphi_j\partial\varphi = &\begin{pmatrix} -G_i' e_t e_u' G_i - G_i' e_u e_t' G_i & \rho_3\left(g_{it} G_i' e_u + g_{iu} G_i' e_t\right) g_i' \\ \rho_3 g_i\left(g_{it} e_u' G_i + g_{iu} e_t' G_i\right) & \rho_4 g_{it} g_{iu} g_i g_i' \end{pmatrix} \\ &- \begin{pmatrix} g_{it}\partial^2 g_{iu}(\theta_0)/\partial\theta\partial\theta' + g_{iu}\partial^2 g_{it}(\theta_0)/\partial\theta\partial\theta' & -\rho_3 g_{it} g_{iu} G_i' \\ -\rho_3 g_{it} g_{iu} G_i & 0 \end{pmatrix}\end{aligned}$$

Then we compute the terms of the derivative for $\ell = D$. Given that h_i^D does not depend on parameters, only few terms survive. We have that, for $u = k - p, t = j - p$

$$h_i^D h_i^{D\prime} = \begin{bmatrix} 0 & 0 \\ 0 & g_i g_i' \end{bmatrix}$$

$$\left(h_i^D\right)_j \cdot h_i^D\left(h_i^D\right)' = \begin{bmatrix} 0 & 0 \\ 0 & g_{ij} g_i g_i' \end{bmatrix} \quad j > p$$

$$\left(h_i^D\right)_k\left(h_i^D\right)_j \cdot h_i^D\left(h_i^D\right)' = g_{iu} g_{it}\begin{bmatrix} 0 & 0 \\ 0 & g_i g_i' \end{bmatrix} \quad j, k > p$$

thus

$$\begin{aligned}
\partial m_i^D\left(\varphi\right)/\partial\varphi &= -\begin{bmatrix} 0 & 0 \\ 0 & g_i g_i' \end{bmatrix} \\
\partial^2 m_i^D\left(\varphi\right)/\partial\varphi_j\partial\varphi &= 0 \\
\partial^3 m_i^D\left(\varphi\right)/\partial\varphi_k\partial\varphi_j\partial\varphi &= 2g_{iu}g_{it}\begin{bmatrix} 0 & 0 \\ 0 & g_i g_i' \end{bmatrix} \qquad j,k>p
\end{aligned}$$

Putting all together, we get

$$\begin{aligned}
\partial m_i\left(\varphi_0\right)/\partial\varphi &= -\begin{pmatrix} 0 & G_i' \\ G_i & g_i g_i' + g_i^D\left(g_i^D\right)' \end{pmatrix} \\
\partial^2 m_i\left(\varphi_0\right)/\partial\varphi_j\partial\varphi &= -\begin{pmatrix} 0 & G_i^{j\prime} \\ G_i^j & g_i^j g_i' + g_i g_i^{j\prime} \end{pmatrix} \qquad (j \le p) \\
&= -\begin{pmatrix} \partial^2\left[e_t' g_i\left(\theta_0\right)\right]/\partial\theta\partial\theta' & G_i' e_t g_i' + g_{it}G_i' \\ g_i e_t' G_i + g_{it}G_i & 0 \end{pmatrix} \qquad (j>p)
\end{aligned}$$

and

$$\begin{aligned}
\partial^3 m_i\left(\varphi_0\right)/\partial\varphi_k\partial\varphi_j\partial\varphi &= -\begin{pmatrix} 0 & G_i^{jk\prime} \\ G_i^{jk} & g_i^{jk}g_i' + g_i^j g_i^{k\prime} + g_i^k g_i^{j\prime} + g_i g_i^{jk\prime} \end{pmatrix} \\
\partial^3 m_i\left(\varphi_0\right)/\partial\varphi_k\partial\varphi_j\partial\varphi &= -\begin{pmatrix} \partial^3 g_{it}\left(\theta_0\right)/\partial\theta_j\partial\theta\partial\theta' & G_i' e_t g_i^{j'} + G_i^{j'} e_t g_i' + G_{itj}G_i' + g_{it}G_i^{j\prime} \\ g_i^j e_t G_i + g_i e_t G_i^j + G_{itj}G_i + g_{it}G_i^j & 0 \end{pmatrix} \\
\partial^3 m_i\left(\varphi_0\right)/\partial\varphi_k\partial\varphi_j\partial\varphi &= -\begin{pmatrix} G_i' e_t e_u' G_i + G_i' e_u e_t' G_i & 0 \\ 0 & -2g_{it}g_{iu}g_i g_i' \end{pmatrix} \\
&\quad -\begin{pmatrix} g_{it}\partial^2 g_{iu}\left(\theta_0\right)/\partial\theta\partial\theta' + g_{iu}\partial^2 g_{it}\left(\theta_0\right)/\partial\theta\partial\theta' & 0 \\ 0 & -2g_{it}^D g_{iu}^D g_i^D\left(g_i^D\right)' \end{pmatrix}
\end{aligned}$$

for each of the possible cases.

D.3 An application: Matching IRF of a DSGE model

We conclude the Monte Carlo simulations with a structural model. We borrow the example from Guerron-Quintana, Inoue, and Kilian (2017), and focus on a small–scale New Keynesian model, which serves as an illustrative example in the macro literature. This model consists of a Phillips curve, a Taylor rule, an investment– savings relationship, and the exogenous driving processes z_t and ξ_t

$$\begin{aligned}
\pi_t &= \kappa x_t + \beta\mathbb{E}\left(\pi_{t+1} \mid \mathcal{I}_{t-1}\right) \\
R_t &= \rho_r R_{t-1} + \left(1-\rho_r\right)\phi_\pi \pi_t + \left(1-\rho_r\right)\phi_x x_t + \xi_t \\
x_t &= \mathbb{E}\left(x_{t+1} \mid \mathcal{I}_{t-1}\right) - \sigma\left(\mathbb{E}\left(R_t \mid \mathcal{I}_{t-1}\right) - \mathbb{E}\left(\pi_{t+1} \mid \mathcal{I}_{t-1}\right) - z_t\right) \\
z_t &= \rho_z z_{t-1} + \sigma^z\varepsilon_t^z \\
\xi_t &= \sigma^r\varepsilon_t^r
\end{aligned}$$

where x_t, π_t and R_t denote the output gap, inflation rate, and interest rate, respectively. The structural shocks ε_t^z and ε_t^r are assumed to be distributed $\mathcal{N}(0,1)$. The model parameters are the discount factor β, the intertemporal elasticity of substitution $1/\sigma$, the probability α of not adjusting prices for a given firm, the elasticity of substitution across varieties of goods, θ, the parameter ω controlling disutility of labor supply; ϕ_π and ϕ_x capture the central bank's reaction to changes in inflation and the output gap, respectively, and $\kappa = \frac{(1-\alpha)(1-\alpha\beta)}{\alpha}\frac{\omega+\sigma}{\sigma(\omega+\theta)}$.

In this model, inflation and real output do not react contemporaneously to the monetary policy shock, ξ_t, but they do respond contemporaneously to a shock to the investment-savings relationship, z_t. These restrictions are required for us to be able to identify the structural shocks of interest in the VAR model based on short-run identifying restrictions. Given this informational constraint, household and firms form expectations based on the information set $\mathcal{I}_{t-1}$.

We focus on the estimation of one parameter only in the simulation study: The probability of not adjusting prices, α, by matching the impulse responses of inflation and of the interest rate with the remaining parameters set to their population values in estimation. The population parameters in the data generating process are $\sigma = 1, \alpha = 0.75, \beta = 0.99, \varphi_\pi = 1.5, \varphi_x = 0.125, \omega = 1, \rho_r = 0.75, \rho_z = 0.90, \theta = 6, \sigma_z = 0.30, \sigma_r = 0.20$.

To proceed to estimation, we write the DSGE model in it's state– space representation form,

$$\begin{aligned} x_t &= Ax_{t-1} + B\epsilon_t \\ y_t &= Cx_t \end{aligned}$$

where x_t is a vector of state variables, ε_t is a vector that consists of the technology shock and the monetary policy shock, and y_t is a vector that consists of inflation and the interest rate. Moreover, A, B and C are matrices of suitable dimensions. In turn, given the parameter values, this system has an invertible moving average representation, so we can write it as a VAR(∞), which may be approximated by a finite–order structural VAR model. Finally, given our assumptions, the structural shocks can be recovered by applying a lower triangular Cholesky decomposition to the residual covariance matrix with the diagonals of the decomposition normalized to be positive.

Let $\Gamma_j = E\left(y_t y_{t-j}'\right)$ denote the population autocovariances implied by the state space representation given a structural parameter value. Then the population parameter values of the VAR(p) model fitted to data generated by the model may be expressed as:

$$\underset{[2p\times 2]}{\Phi} = \begin{bmatrix} \Gamma_0 & \Gamma_1 & \cdots & \Gamma_{p-1} \\ \Gamma_1' & \Gamma_0 & \cdots & \Gamma_{p-2} \\ \vdots & \vdots & \ddots & \vdots \\ \Gamma_{p-1}' & \Gamma_{p-2}' & \cdots & \Gamma_0 \end{bmatrix}^{-1} \begin{bmatrix} \Gamma_1' \\ \Gamma_2' \\ \vdots \\ \Gamma_p' \end{bmatrix}$$

$$\underset{[2\times 2]}{\Sigma} = \Gamma_0 - \begin{bmatrix} \Gamma_1 & \Gamma_2 & \cdots & \Gamma_p \end{bmatrix} \times \Phi$$

The population structural impulse responses can be written as functions of Φ and Σ, so we can use $\{\Gamma_0, \Gamma_1, \dots, \Gamma_p\}$ as the moment conditions in AMM estimation. It's important to note that since we are matching the moments directly, the horizon is irrelevant for our estimation.

D.3.1 Results

Here, we compare the results from our estimation and the bootstrap approach in Guerron-Quintana, Inoue, and Kilian (2017) (they report two versions of their estimator: one with a diagonal weighting matrix, and another with optimal weighting).

		AMM		Diagonal W		Optimal W	
T	L	Bias	RMSE	Bias	RMSE	Bias	RMSE
100	2	-0.007	0.038	0.009	0.022	0.003	0.012
100	4	-0.011	0.077	0.011	0.022	0.004	0.013
100	6	-0.021	0.101	0.012	0.023	0.006	0.014
232	2	-0.001	0.021	0.004	0.015	0.001	0.007
232	4	-0.001	0.021	0.004	0.014	0.002	0.007
232	6	-0.002	0.024	0.005	0.014	0.002	0.008

Table 8: Based on 500 simulations. T denotes the number of time periods and L is the SVAR lag length

From the previous table, we see that, even though the AMM estimator is not tailored to this framework (as the bootstrap estimator of Guerron-Quintana, Inoue, and Kilian (2017)), we can see that the performance of both is quite similar. Moreover, it's important to highlight that our procedure is much less computationally intensive, since we don't require computation of many bootstrap samples, as the other estimator does.

References

Altonji, Joseph and Lewis M Segal (1996). "Small-Sample Bias in GMM Estimation of Covariance Structures". In: *Journal of Business & Economic Statistics* 14.3, pp. 353–66. URL: https://EconPapers.repec.org/RePEc:bes:jnlbes:v:14:y:1996:i:3:p:353-66.

Arellano, Manuel and Stephen Bond (Apr. 1991). “Some Tests of Specification for Panel Data: Monte Carlo Evidence and an Application to Employment Equations”. In: *The Review of Economic Studies* 58.2, pp. 277–297. ISSN: 0034-6527. DOI: 10.2307/2297968. eprint: https://academic.oup.com/restud/article-pdf/58/2/277/4454458/58-2-277.pdf. URL: https://doi.org/10.2307/2297968.

Goodfellow, Ian J. et al. (2014). “Generative Adversarial Networks”. In: arXiv: 1406.2661 [stat.ML].

Gourieroux, C., A. Monfort, and E. Renault (1993). “Indirect inference”. In: *Journal of Applied Econometrics* 8.S1, S85–S118. DOI: 10.1002/jae.3950080507. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/jae.3950080507.

Guerron-Quintana, Pablo, Atsushi Inoue, and Lutz Kilian (2017). “Impulse response matching estimators for DSGE models”. In: *Journal of Econometrics* 196.1, pp. 144–155. DOI: 10.1016/j.jeconom.2016.09. URL: https://ideas.repec.org/a/eee/econom/v196y2017i1p144-155.html.

Hansen, Lars (1982). “Large Sample Properties of Generalized Method of Moments Estimators”. In: *Econometrica* 50.4, pp. 1029–1054. ISSN: 00129682, 14680262. URL: http://www.jstor.org/stable/1912775.

Hansen, Lars, John Heaton, and Amir Yaron (1996). “Finite-Sample Properties of Some Alternative GMM Estimators”. In: *Journal of Business & Economic Statistics* 14.3, pp. 262–80. URL: https://EconPapers.repec.org/RePEc:bes:jnlbes:v:14:y:1996:i:3:p:262-80.

Kaji, Tetsuya, Elena Manresa, and Guillaume Pouliot (July 2020). *An Adversarial Approach to Structural Estimation.* Papers 2007.06169. arXiv.org. URL: https://ideas.repec.org/p/arx/papers/2007.06169.html.

Liang, Tengyuan (2021). “How Well Generative Adversarial Networks Learn Distributions”. In: *Journal of Machine Learning Research* 22.228, pp. 1–41. URL: http://jmlr.org/papers/v22/20-911.html.

Newey, Whitney and Richard Smith (2004). “Higher Order Properties of GMM and Generalized Empirical Likelihood Estimators”. In: *Econometrica* 72.1, pp. 219–255. DOI: 10.1111/j.1468-0262.2004.00482.x. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1468-0262.2004.00482.x.