

# Chapter 1

## Generalized Method of Moments

**Generalized method of moments** (GMM) is an estimation principle that extends **method of moments**. It seeks the parameter value that minimizes a quadratic form of the sample moments. It is particularly useful in estimating structural models in which moment conditions can be (hopefully) derived from economic theory. GMM emerges as one of the most popular estimators in modern econometrics.

### 1.1 Estimating Equations

We are interested in a moment function  $g_i(\theta) = g(z_i, \theta) \in \mathbb{R}^m$ , where  $\theta$  is a  $p$ -dimensional parameter. Motivated by some economic theory, we look for the true value  $\theta_0 \in \Theta$  as the solution to the following  $m$  moment equations

$$\mathbb{E}[g_i(\theta)] = \mathbf{0}_m.$$

These are the **estimating equations**.

The true value  $\theta_0$  is identified if it uniquely satisfies  $\mathbb{E}[g_i(\theta_0)] = 0$ . To measure the distance of  $\mathbb{E}[g_i(\theta)]$  from  $\mathbf{0}_m$ , define

$$S(\theta; W) = \mathbb{E}[g_i(\theta)]' W \mathbb{E}[g_i(\theta)],$$

where  $W$  an  $m \times m$  symmetric positive-definite matrix  $W$ , which is called the **weighting matrix**. The choice of  $W$  is arbitrary, and one of the valid candidates is just the identity matrix  $I_m$ . Formally, **identification** means that for any  $\varepsilon > 0$ , there exists a  $\delta > 0$  such that

$$\inf_{\theta \in \Theta \setminus \mathcal{N}_\delta(\theta_0)} S(\theta; W) > S(\theta_0; W) + \varepsilon.$$

Next, we provide more conditions under which we can verify identification. Suppose  $g(\cdot, \theta)$  is twice differentiable with respect to  $\theta$  over the support of  $z_i$ , and the differentiation and expectation is exchangeable. Denote the  $m \times p$  matrix  $D_i(\theta) = \frac{\partial}{\partial \theta'} g_i(\theta)$ , and  $D_0 = \mathbb{E}[D_i(\theta_0)]$ . Define

$$\begin{aligned} \psi(\theta; W) &= \frac{\partial}{\partial \theta} S(\theta; W) = 2\mathbb{E}[D_i'(\theta)] W \mathbb{E}[g_i(\theta)] \\ H(\theta; W) &= \frac{\partial^2}{\partial \theta \partial \theta'} S(\theta; W) = 2 \sum_{k=1}^m (W \mathbb{E}[g_i(\theta)])_k \mathbb{E} \left[ \frac{\partial^2 g_k(z_i, \theta)}{\partial \theta \partial \theta'} \right] + 2\mathbb{E}[D_i'(\theta)] W \mathbb{E}[D_i(\theta)]. \end{aligned}$$

Take a Taylor expansion of  $S(\theta; W)$  around  $\theta_0$ :

$$S(\theta; W) = S(\theta_0; W) + \psi(\theta_0; W)(\theta - \theta_0) + \frac{1}{2}(\theta - \theta_0)' H(\theta_0; W)(\theta - \theta_0) + h.o.t.,$$

where the higher-order term (h.o.t.) is  $o(\|\theta - \theta_0\|^2)$ . Evaluated at  $\theta = \theta_0$ , since  $\mathbb{E}[g_i(\theta_0)] = 0$  we have

$$\begin{aligned} S(\theta_0; W) &= 0 \\ \psi(\theta_0; W) &= \mathbf{0}_m \\ H(\theta_0; W) &= 2D_0'WD_0. \end{aligned}$$

If we take a heuristic approach by ignoring the h.o.t., for any  $\theta$  such that  $\|\theta - \theta_0\| > \delta$ , we obtain

$$S(\theta; W) = \frac{1}{2}(\theta - \theta_0)' H(\theta_0; W)(\theta - \theta_0) = (\theta - \theta_0)' D_0' W D_0 (\theta - \theta_0).$$

It entails a full rank  $D_0$  to guarantee that there exists a constant  $c$  such that

$$S(\theta; W) > \|\theta - \theta_0\|^2 c \cdot \lambda_{\min}(W) \geq \delta^2 c \cdot \lambda_{\min}(W).$$

Otherwise, there exists a  $\theta^\dagger$  such that  $\theta^\dagger \neq \theta_0$  but  $D_0(\theta - \theta_0) = \mathbf{0}_m$ ; we cannot identify  $\theta_0$  because  $S(\theta^\dagger; W) = 0$  as well.

**Exercise 1.1.** Argue that a necessary condition for identification is  $p \leq m$ . In other words, the number of moment conditions must be no smaller than the number of parameters.

**Example 1.1.** Consider the linear IV model  $y_i = x_i'\theta + \varepsilon_i$ , where  $x_i$  is a  $p$ -dimensional endogenous variable and  $w_i$  is an  $m$ -dimensional instrumental variable. The moment function  $g_i(\theta) = w_i(y_i - x_i'\theta)$ . The orthogonality between  $z_i$  and  $\varepsilon_i$  ensures

$$\mathbb{E}[g_i(\theta_0)] = \mathbb{E}[w_i(y_i - x_i'\theta_0)] = \mathbb{E}[w_i\varepsilon_i] = \mathbf{0}_m.$$

The necessary condition for identification is a full rank  $D_0 = -\mathbb{E}[w_i x_i']$ , which is the relevant condition between  $w_i$  and  $x_i$ .

The above heuristic argument for identification only ensures **local identification**, because we neglect the h.o.t. in a small neighborhood around  $\theta_0$ . Sufficient conditions for **global identification** is challenging and under active research.

## 1.2 GMM Estimator

Given an iid sample  $z_i$ ,  $i = 1, \dots, n$ , we expect the sample moments

$$\frac{1}{n} \sum_{i=1}^n g_i(\theta_0) \approx \mathbf{0}_m.$$

Therefore, it is reasonable to estimate  $\theta$  by finding some value in  $\Theta$  to set  $\frac{1}{n} \sum_{i=1}^n g_i(\theta)$  close to  $\mathbf{0}_m$ .

**Example 1.2.** Consider again the linear IV model. If  $p = m$  (just identified), then we can solve

$$\frac{1}{n} \sum_{i=1}^n g_i(\theta) = \frac{1}{n} \sum_{i=1}^n w_i(y_i - x_i'\theta) = \frac{1}{n} \sum_{i=1}^n w_i y_i - \left( \frac{1}{n} \sum_{i=1}^n w_i x_i' \right) \theta = \mathbf{0}_m$$

with  $\hat{\theta} = \left( \frac{1}{n} \sum_{i=1}^n w_i x_i' \right)^{-1} \frac{1}{n} \sum_{i=1}^n w_i y_i = (W'X)^{-1} W'Y$ . If  $m > p$  (over-identified), the  $m$ -equation system

$$\left( \frac{1}{n} \sum_{i=1}^n w_i x_i' \right) \theta = \frac{1}{n} \sum_{i=1}^n w_i y_i$$

has no solution because there are few than  $m$  free parameters in  $\theta$ .

To work with just identified cases and the over-identified cases in a unified framework, we use the sample moment  $\bar{g}_n(\theta) = \frac{1}{n} \sum_{i=1}^n g_i(\theta)$  to mimic the population moment  $\mathbb{E}[g_i(\theta)]$ , and define the sample criterion function

$$S_n(\theta) = \bar{g}_n'(\theta) W \bar{g}_n(\theta)$$

as the counterpart of the population distance  $S(\theta; W)$ . The **GMM estimator**

$$\hat{\theta} = \arg \min_{\theta \in \Theta} S_n(\theta)$$

is the minimizer of the sample criterion function  $S_n(\theta)$ . Notice that GMM is not an M-function as it is not a sample average of the loss function of each individual observation. Instead, it is a quadratic form of a sample average of the moment function  $g_i(\theta)$ .

The proof of consistency of the GMM under identification and ULLN is identical to that of the M-estimator. We leave it as an exercise.

**Exercise 1.2.** Show that  $\hat{\theta} \xrightarrow{P} \theta_0$  if (i)  $\theta_0$  is identified, (ii)  $\sup_{\theta \in \Theta_0} |S_n(\theta) - S(\theta)| \xrightarrow{P} 0$ .

Next, given consistency we further check the asymptotic distribution of the GMM estimator under conditions of differentiability. The first-order condition of the optimality of GMM yields

$$\mathbf{0}_m = \frac{\partial}{\partial \theta} S_n(\hat{\theta}; W) = 2\bar{D}_n'(\theta) W \bar{g}_n(\theta),$$

where  $\bar{D}_n(\theta) = \frac{1}{n} \sum D_i(\theta)$ . Denote

$$\bar{\psi}_n(\theta; W) = 2\bar{D}_n'(\theta) W \bar{g}_n(\theta).$$

Take a Taylor expansion of  $\bar{\psi}_n(\hat{\theta}; W)$  around  $\theta_0$ :

$$\mathbf{0}_m = \bar{\psi}_n(\hat{\theta}; W) = \bar{\psi}_n(\theta_0; W) + \bar{H}_n(\theta_0; W)(\hat{\theta} - \theta_0) + h.o.t.,$$

where

$$\bar{H}_n(\theta; W) = 2 \sum_{k=1}^m (W \bar{g}_n(\theta))_k \frac{\partial^2 \bar{g}_{n,k}(z_i, \theta)}{\partial \theta \partial \theta'} + 2\bar{D}_n'(\theta) W \bar{D}_n(\theta).$$

We again ignore the higher-order term and rearrange:

$$\sqrt{n}(\hat{\theta} - \theta_0) = -\bar{H}_n^{-1}(\theta_0; W) \times \sqrt{n}\bar{\psi}_n(\theta_0; W). \quad (1.1)$$

We analyze the numerator  $\sqrt{n}\bar{\psi}_n(\theta_0; W) = 2\bar{D}_n'(\theta_0) W \frac{1}{\sqrt{n}}g_i(\theta_0)$ . Since  $\mathbb{E}[g_i(\theta_0)] = 0$ , if the variance is finite we can apply an CLT:

$$\frac{1}{\sqrt{n}}g_i(\theta_0) \xrightarrow{d} N(0, \Omega_0)$$

where  $\Omega_0 = \text{var}[g_i(\theta_0)] = \mathbb{E}[g_i(\theta_0)g_i'(\theta_0)]$ . If in addition  $\bar{D}_n(\theta_0) \xrightarrow{P} D_0$  by some LLN, we have

$$\sqrt{n}\bar{\psi}_n(\theta_0; W) \xrightarrow{d} N(0, 4D_0'W\Omega_0WD_0).$$

The denominator

$$\bar{H}_n(\theta_0; W) = 2 \sum_{k=1}^m (W \bar{g}_n(\theta_0))_k \frac{\partial^2 \bar{g}_{n,k}(z_i, \theta_0)}{\partial \theta \partial \theta'} + 2 \bar{D}'_n(\theta_0) W \bar{D}_n(\theta_0) \xrightarrow{p} 2 D_0' W D_0,$$

where the first term vanishes as  $\bar{g}_n(\theta_0) \xrightarrow{p} \mathbf{0}_m$ . Put the numerator and the denominator together, we conclude

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} N\left(0, (D_0' W D_0)^{-1} D_0' W \Omega_0 W D_0 (D_0' W D_0)^{-1}\right).$$

It is obvious that the asymptotic variance of the GMM estimator depends on the choice of the weight matrix  $W$ . If we choose  $W = \Omega_0^{-1}$ , the asymptotic variance is simplified as

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} N\left(0, (D_0' \Omega_0^{-1} D_0)^{-1}\right).$$

It turns out that this choice of  $W$  achieves semiparametric efficiency.

### 1.3 Over-identification Test

The definition of  $\hat{\theta}$  entails that  $S_n(\hat{\theta}; W)$  is the minimum of  $S_n(\theta; W)$ . It turns out that the scaled (by  $n$ ) criterion function  $nS_n(\theta; \Omega_0^{-1})$  can serve as a test statistic for model specification.

The scaling factor  $n$  is essential because  $S_n(\hat{\theta}; W) \leq S_n(\theta_0; W) \xrightarrow{p} 0$  due to  $\bar{g}_n(\theta_0) \xrightarrow{p} \mathbf{0}_m$ , whereas

$$nS_n(\theta_0; W) = (\sqrt{n}\bar{g}'_n(\theta_0)) W (\sqrt{n}\bar{g}_n(\theta_0))$$

is a quadratic form of an asymptotically normal random vector  $\sqrt{n}\bar{g}'_n(\theta_0)$ . In particular, the choice  $W = \Omega_0^{-1}$  gives

$$nS_n(\theta_0; \Omega_0^{-1}) \xrightarrow{d} \chi^2(p), \quad (1.2)$$

but this is different from the asymptotic distribution of

$$nS_n(\hat{\theta}; \Omega_0^{-1}) = n\bar{g}'_n(\hat{\theta}) \Omega_0^{-1} \bar{g}_n(\hat{\theta}) = \left\| \Omega_0^{-1/2} \sqrt{n}\bar{g}_n(\hat{\theta}) \right\|^2. \quad (1.3)$$

Decompose

$$\sqrt{n}\bar{g}_n(\hat{\theta}) = \sqrt{n}\bar{g}_n(\theta_0) + \sqrt{n}(\bar{g}_n(\hat{\theta}) - \bar{g}_n(\theta_0)) \quad (1.4)$$

The second term, by a Taylor expansion around  $\theta_0$  and putting the h.o.t. into  $o_p(1)$ , becomes

$$\begin{aligned} \sqrt{n}(\bar{g}_n(\hat{\theta}) - \bar{g}_n(\theta_0)) &= \sqrt{n}\bar{D}'_n(\theta_0) \times (\hat{\theta} - \theta_0) + o_p(1) \\ &= -\bar{D}'_n(\theta_0) \times \bar{H}_n^{-1}(\theta_0; \Omega_0^{-1}) \times \sqrt{n}\bar{\psi}_n(\theta_0; \Omega_0^{-1}) + o_p(1) \\ &= -\bar{D}'_n(\theta_0) \bar{H}_n^{-1}(\theta_0; \Omega_0^{-1}) 2\bar{D}'_n(\theta_0) \Omega_0^{-1} \sqrt{n}\bar{g}_n(\theta_0) + o_p(1) \\ &= -D_0 (D_0' \Omega_0^{-1} D_0)^{-1} D_0' \Omega_0^{-1} \sqrt{n}\bar{g}_n(\theta_0) + o_p(1). \end{aligned} \quad (1.5)$$

where the second line follows by (1.1), the third line by the definition of  $\bar{\psi}_n$ , and the last line by citing the probabilistic limit of the matrices  $\bar{H}_n$  and  $\bar{D}_n$ . (1.4) and (1.5) imply

$$\begin{aligned} \Omega_0^{-1/2} \sqrt{n}\bar{g}_n(\hat{\theta}) &= \Omega_0^{-1/2} \left( I_m - D_0 (D_0' \Omega_0^{-1} D_0)^{-1} D_0' \Omega_0^{-1} \right) \sqrt{n}\bar{g}_n(\theta_0) + o_p(1) \\ &\xrightarrow{d} \Omega_0^{-1/2} \left( I_m - D_0 (D_0' \Omega_0^{-1} D_0)^{-1} D_0' \Omega_0^{-1} \right) \times N(0, \Omega_0) \\ &\sim \Omega_0^{-1/2} \left( I_m - D_0 (D_0' \Omega_0^{-1} D_0)^{-1} D_0' \Omega_0^{-1} \right) \Omega_0^{1/2} \times N(0, I_m) \\ &\sim \left( I_m - \Omega_0^{-1/2} D_0 (D_0' \Omega_0^{-1} D_0)^{-1} D_0' \Omega_0^{-1/2} \right) \times N(0, I_m). \end{aligned}$$

Since  $\left(I_m - \Omega^{-1/2} D_0 (D_0' \Omega_0^{-1} D_0)^{-1} D_0' \Omega^{-1/2}\right)$  is idempotent with rank  $(p - m)$ , we have

$$nS_n \left( \hat{\theta}; \Omega_0^{-1} \right) = \left\| \Omega_0^{-1/2} \sqrt{n} \bar{g}_n \left( \hat{\theta} \right) \right\| \xrightarrow{d} \chi^2 (m - p).$$

Compared with (1.2), the estimated  $p$ -dimensional parameter  $\hat{\theta}$  deducts  $p$  degrees of freedom.

To make the test statistic feasible, we can use any consistent estimator  $\hat{\Omega}$  for  $\Omega_0$ , for example,

$$\hat{\Omega} = \frac{1}{n} \sum_{i=1}^n g_i \left( \hat{\theta} \right) g_i' \left( \hat{\theta} \right) - \bar{g}_n \left( \hat{\theta} \right) \bar{g}_n' \left( \hat{\theta} \right).$$

The feasible statistic  $nS_n \left( \hat{\theta}; \hat{\Omega}^{-1} \right)$  is called the  $J$ -statistic.

Zhentao Shi. January 31, 2024