Generalized Method of Moments (GMM)

Magíster en Economía

Teoría Econométrica (Econometric Theory)

Prof. Luis Chancí

www.luischanci.com

Outline

1. From classical method of moments to GMM
2. The GMM criterion and identification
3. OLS, IV, and MLE as special cases
4. Weighting matrices and efficient GMM
5. Asymptotic theory
6. Hansen's J-test

> **Main idea: many estimators are built from population moments of the form**
>
> $$\mathbb{E}[h(\theta_0, w_i)] = 0.$$
>
> **GMM turns those moment restrictions into an estimator.**

1. From Classical Method of Moments to GMM

Why GMM?

In previous topics, we saw three estimation strategies:

- **OLS**: uses orthogonality between regressors and errors
- **IV**: uses instrument-based orthogonality conditions
- **MLE**: uses score conditions from a full likelihood

> **GMM unifies all of them by starting from a vector of moment conditions:**
>
> $$\mathbb{E}[h(\theta_0, w_i)] = 0.$$

Compared with MLE, GMM requires **less structure**: we do not need the full distribution of the data.
Compared with OLS and IV, it is **more general**: it can combine several moments at once.

Classical Method of Moments

Suppose the model implies k population moments:

$$\mathbb{E}[m_j(y_i;\theta)] = 0, \qquad j = 1, \ldots, k.$$

If the model is **exactly identified**, a natural estimator solves the sample analogues:

$$\frac{1}{n}\sum_{i=1}^{n} m_j(y_i;\hat{\theta}) = 0, \qquad j = 1, \ldots, k.$$

> **This is the classical method of moments: match sample moments to population moments.**

A Simple Example: Student-t Degrees of Freedom

Suppose $y_i \overset{\text{i.i.d.}}{\sim} t(\nu)$, and the parameter of interest is ν.

If $\nu > 2$, we know that the second moment is $\mathbb{E}[y_i^2] = \frac{\nu}{\nu-2}$.

Thus, let

$$\hat{\mu}_2 = \frac{1}{n}\sum_{i=1}^{n} y_i^2.$$

Method of moments sets

$$\hat{\mu}_2 = \frac{\nu}{\nu - 2},$$

Therefore, the MM estimator would be:

$$\hat{\nu}_{MM} = \frac{2\hat{\mu}_2}{\hat{\mu}_2 - 1}.$$

Why Method of Moments Is Not Enough

Suppose we also know that, for $\nu > 4$,

$$\mathbb{E}[y_i^4] = \frac{3\nu^2}{(\nu-2)(\nu-4)}.$$

We can stack the two conditions into one moment vector:

$$h(\nu, y_i) = \begin{bmatrix} y_i^2 - \dfrac{\nu}{\nu-2} \\ \\ y_i^4 - \dfrac{3\nu^2}{(\nu-2)(\nu-4)} \end{bmatrix}, \qquad \nu > 4.$$

> **Here $r = 2$ moments but only $k = 1$ parameter (ν): the model is overidentified. We cannot force both sample moments to zero at once, so GMM makes them collectively as close to zero as possible.**

2. The GMM Setup

Population and Sample Moments

Let w_i denote the observed data and let $\theta \in \Theta \subset \mathbb{R}^k$ be the unknown parameter vector.

Suppose the true parameter satisfies

$$\mathbb{E}[h(\theta_0, w_i)] = 0, \qquad h(\theta, w_i) \in \mathbb{R}^r.$$

The sample analogue is

$$g_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} h(\theta, w_i).$$

> **Keep the two objects separate: $h(\theta, w_i)$ is one observation's contribution (random at the unit level); $g_n(\theta)$ is the sample average that enters the criterion. The CLT applies to $\sqrt{n}\, g_n(\theta_0)$, not to a single $h(\theta_0, w_i)$.**

The GMM Criterion Function

Definition — GMM estimator

The GMM estimator is

$$\hat{\theta}_{GMM} = \arg\min_{\theta \in \Theta} Q_n(\theta),$$

where

$$Q_n(\theta) = g_n(\theta)' W_n g_n(\theta),$$

and W_n is a positive definite weighting matrix.

Interpretation: choose θ so that the sample moments are as close to zero as possible, using W_n to decide how much each moment matters.

Exact Identification vs. Overidentification

Let k = number of parameters and r = number of moment conditions. Then:

- if $r = k$, the model is **exactly identified** — often solve $g_n(\hat{\theta}) = 0$
- if $r > k$, the model is **overidentified** — minimize $Q_n(\theta) = g_n(\theta)' W_n g_n(\theta)$

> **In the exactly identified case, if $g_n(\hat{\theta}) = 0$ can be solved and W_n is positive definite, the estimator does not depend on W_n (the criterion hits zero at the same point). This is not automatic under weak identification or local minima.**

Why Overidentification Is Useful

Overidentification means we have **more valid information than parameters**.

That brings two advantages:

- the estimator may become more efficient
- the extra moments create a natural specification test

> **This is one of the major attractions of GMM: it turns extra moments into both precision and testable restrictions.**

A Concrete Overidentified Example

Suppose y_i is i.i.d. with **mean** = variance = λ (the Poisson case). Then λ satisfies two moments:

$$\mathbb{E}[y_i - \lambda] = 0, \qquad \mathbb{E}\left[(y_i - \lambda)^2 - \lambda\right] = 0.$$

Stacking them,

$$h(\lambda, y_i) = \begin{bmatrix} y_i - \lambda \\ (y_i - \lambda)^2 - \lambda \end{bmatrix}, \qquad g_n(\lambda) = \begin{bmatrix} \bar{y} - \lambda \\ \frac{1}{n}\sum_i (y_i - \lambda)^2 - \lambda \end{bmatrix}.$$

> **$r = 2, k = 1$: GMM minimizes $Q_n(\lambda) = g_n(\lambda)' W_n g_n(\lambda)$. The mean alone gives $\hat{\lambda} = \bar{y}$; the variance adds information, and the J-test will check the equidispersion restriction.**

3. OLS, IV, and MLE as GMM

OLS as GMM

Start from the linear model

$$y_i = x_i'\beta + u_i,$$

with exogeneity

$$\mathbb{E}[x_i u_i] = 0.$$

Since $u_i = y_i - x_i'\beta$, the moment condition is

$$\mathbb{E}[x_i(y_i - x_i'\beta)] = 0.$$

So define

$$h(\beta, w_i) = x_i(y_i - x_i'\beta), \qquad g_n(\beta) = \frac{1}{n}X'(y - X\beta).$$

If the model is exactly identified, GMM solves $X'(y - X\hat{\beta}) = 0$ — exactly the **OLS normal equations**.

IV / 2SLS as GMM

With instruments z_i, IV is built on $\mathbb{E}[z_i u_i] = 0$, so

$$h(\beta, w_i) = z_i(y_i - x_i'\beta), \qquad g_n(\beta) = \frac{1}{n} Z'(y - X\beta).$$

Choosing

$$W_n = \left(\frac{Z'Z}{n} \right)^{-1}$$

yields the usual IV / 2SLS estimator.

> **Note: $W_n = (Z'Z/n)^{-1}$ is optimal only under conditional homoskedasticity. Under heteroskedasticity, efficient GMM uses $\hat{S}^{-1}$ with $S = \mathrm{E}[z_i z_i' u_i^2]$ and is asymptotically more efficient than standard 2SLS. This is the main practical reason to prefer GMM over 2SLS.**

MLE as Moment-Based Estimation

Under correct specification, the score satisfies

$$\mathbb{E}[\boldsymbol{s}_i(\theta_0)] = 0,$$

which is itself a moment condition. Moreover, the **information equality** holds:

$$\mathbb{E}[\boldsymbol{s}_i(\theta_0)\boldsymbol{s}_i(\theta_0)'] = -\mathbb{E}\left[\frac{\partial \boldsymbol{s}_i(\theta_0)}{\partial \theta'}\right] = \mathcal{I}(\theta_0).$$

> **So MLE is essentially GMM on the score moments with the efficient weighting already built in — which is why correctly specified MLE is asymptotically efficient. OLS, IV, and many MLEs are all special cases of one moment-based view.**

4. Weighting Matrices and Efficient GMM

Why the Weighting Matrix Matters

When $r > k$, not all moments can be satisfied exactly, and moments differ in precision and are correlated with one another.

> **Efficient GMM is a matrix problem, not a one-by-one ranking: it accounts for the covariance across moments, downweighting noisy or redundant linear combinations — not just individual moments.**

The Covariance Matrix of the Moments

Define

$$S = \lim_{n \to \infty} n \, \mathbb{V}\big(g_n(\theta_0)\big).$$

Under i.i.d. sampling,

$$S = \mathbb{E}[h(\theta_0, w_i) h(\theta_0, w_i)'].$$

- With **dependent data**, S is the **long-run covariance** of the moments, and $\hat{S}$ must account for serial correlation (e.g. HAC / Newey–West estimators).

> **S tells us how noisy the moments are, jointly.**

The Optimal Weighting Matrix

Efficient GMM weighting

Among all positive definite weighting matrices, the asymptotically optimal choice is

$$W = S^{-1}.$$

If we use $W = S^{-1}$, the GMM estimator has the smallest asymptotic variance within the class of estimators based on the chosen moments. This is the matrix analogue of generalized least squares.

Two-Step GMM

In practice, S is unknown, so efficient GMM is implemented in two steps:

1. Choose an initial weighting matrix $W_n^{(0)}$ (often I)
2. Compute a preliminary consistent estimator $\hat{\theta}^{(0)}$
3. Estimate $\hat{S}$ **(assuming i.i.d. data):**

$$\hat{S} = \frac{1}{n} \sum_{i=1}^{n} h(\hat{\theta}^{(0)}, w_i) h(\hat{\theta}^{(0)}, w_i)'$$

(time-series / dependent data: use a HAC estimator such as Newey–West here)

4. Set $W_n^{(1)} = \hat{S}^{-1}$
5. Re-estimate using the efficient criterion

> **The first-step weighting matrix need not be optimal — it only needs to give a consistent $\hat{\theta}^{(0)}$. Efficiency is recovered in step 2 after estimating S. (This is why the identity matrix is fine to start.)**

Practical GMM Workflow

A map of the whole procedure before we turn to theory:

1. Specify economically meaningful moments: $\mathbb{E}[h(\theta_0, w_i)] = 0$
2. Build sample moments: $g_n(\theta) = \frac{1}{n}\sum_i h(\theta, w_i)$
3. Choose an initial weighting matrix $W_n^{(0)}$
4. Estimate the preliminary $\hat{\theta}^{(0)}$
5. Estimate $\hat{S}$ and set $W_n^{(1)} = \hat{S}^{-1}$
6. Re-estimate, then compute standard errors and the J-test

> **Every applied GMM exercise is some version of these six steps.**

5. Asymptotic Theory of GMM

Regularity Conditions

Let

$$D = \mathbb{E}\left[\frac{\partial h(\theta_0, w_i)}{\partial \theta'}\right]$$

be the $r \times k$ expected Jacobian (sensitivity of the moments to θ).

Regularity conditions for GMM

- identification: $\mathbb{E}[h(\theta, w_i)] = 0$ iff $\theta = \theta_0$
- smoothness of $h(\theta, w_i)$ in θ
- uniform LLN for $g_n(\theta)$ and its Jacobian
- CLT for the sample moments: $\sqrt{n}\, g_n(\theta_0) \xrightarrow{d} \mathcal{N}(0, S)$
- $W_n \xrightarrow{p} W$ with W positive definite
- D has full column rank k

Consistency of GMM

The population criterion is

$$Q(\theta) = g(\theta)' W g(\theta), \qquad g(\theta) = \mathbb{E}[h(\theta, w_i)].$$

Because $g(\theta_0) = 0$, we have $Q(\theta_0) = 0$. If identification holds, θ_0 is the **unique** minimizer of the population criterion.

Under uniform convergence $Q_n(\theta) \xrightarrow{p} Q(\theta)$, so

$$\hat{\theta}_{GMM} \xrightarrow{p} \theta_0.$$

Asymptotic Normality: Main Expansion

(Assumes a differentiable criterion and an interior solution; nonsmooth moments or boundary parameters need modified arguments.)

The first-order condition is

$$G_n(\hat{\theta})' W_n g_n(\hat{\theta}) = 0, \qquad G_n(\theta) = \frac{\partial g_n(\theta)}{\partial \theta'}.$$

Expand $g_n(\hat{\theta})$ around θ_0:

$$g_n(\hat{\theta}) = g_n(\theta_0) + G_n(\bar{\theta})(\hat{\theta} - \theta_0).$$

Solving for the scaled error,

$$\sqrt{n}(\hat{\theta} - \theta_0) = -\Big[G_n(\hat{\theta})' W_n G_n(\bar{\theta})\Big]^{-1} G_n(\hat{\theta})' W_n \sqrt{n}\, g_n(\theta_0).$$

Asymptotic Normality: Final Result

Component limits: $G_n(\hat{\theta}) \to_p D,\ G_n(\bar{\theta}) \to_p D,\ W_n \to_p W,\ \sqrt{n}\, g_n(\theta_0) \to_p \mathcal{N}(0, S)$.

By Slutsky,

$$\sqrt{n}(\hat{\theta}_{GMM} - \theta_0) \xrightarrow{d} \mathcal{N}(0, V(W)), \quad V(W) = (D'WD)^{-1} D'WSWD(D'WD)^{-1}.$$

Logic: linearize the sample moments around θ_0, use the CLT for $\sqrt{n}\, g_n(\theta_0)$, and transfer that normality to $\hat{\theta}$ through the Jacobian D. The sandwich stacks three effects: sensitivity D, weighting W, and moment noise S.

Efficient GMM Variance

If the optimal weighting matrix is used, $W = S^{-1}$, the sandwich collapses to

$$V_{opt} = (D' S^{-1} D)^{-1}.$$

This is the efficient GMM variance bound for the chosen set of moments.

> **Efficiency here is relative to the chosen moments, not global as in correctly specified MLE. More valid moments can expand the information set.**

Feasible Variance Estimation

In practice, estimate

$$\hat{D} = \frac{1}{n}\sum_{i=1}^{n} \frac{\partial h(\hat{\theta}, w_i)}{\partial \theta'}, \qquad \hat{S} = \frac{1}{n}\sum_{i=1}^{n} h(\hat{\theta}, w_i) h(\hat{\theta}, w_i)' \;\; \text{(i.i.d.)},$$

and for efficient two-step GMM,

$$\widehat{\mathrm{Var}}(\hat{\theta}_{GMM}) = \frac{1}{n}(\hat{D}'\hat{S}^{-1}\hat{D})^{-1}.$$

> **Reading the sandwich: $\hat{D}$ = sensitivity of moments to parameters, $\hat{S}$ = covariance of the moments, $\hat{W}$ = how moments are weighted. With dependent data, $\hat{S}$ must be a HAC estimator.**

6. Hansen's J-Test

Testing Overidentifying Restrictions

In overidentified models, valid moments imply that at the GMM estimate, $g_n(\hat{\theta})$ should be close to zero. This leads to Hansen's J-test.

Definition — Hansen's J-statistic

Let $\hat{\theta}_{GMM}$ be the **efficient** two-step (or iterated) GMM estimator. Then

$$J = n\, Q_n(\hat{\theta}_{GMM}) = n\, g_n(\hat{\theta}_{GMM})' \hat{S}^{-1} g_n(\hat{\theta}_{GMM}).$$

The J-statistic is literally the minimized GMM objective Q_n, scaled by n — the same criterion we optimized, now read as a test.

Limiting Distribution of the J-Test

Under the null that all overidentifying restrictions are valid,

$$J \xrightarrow{d} \chi^2_{r-k},$$

with $r - k$ = number of overidentifying restrictions.

> **The J-test checks the joint validity of the moments, not the coefficients. A rejection does not say which moment is invalid; a non-rejection is not proof of validity (the test can have low power in small samples or under weak identification).**

Why Efficient GMM Matters for the J-Test

The standard χ^2 limit holds only when the weighting matrix converges to the optimal choice $W = S^{-1}$.

If a non-optimal weighting matrix is used, the minimized objective does **not** generally have a standard χ^2 distribution.

> **So the usual J-test must be based on efficient GMM, not on an arbitrary first-step criterion. In exactly identified models ($r = k$) the test is degenerate: the moments are forced to zero.**

Final Summary

1. GMM starts from moment conditions: $\mathbb{E}[h(\theta_0, w_i)] = 0$
2. It nests **OLS**, **IV**, and many **MLEs** as special cases
3. In overidentified models, the weighting matrix matters; the optimal choice is $W = S^{-1}$
4. Under regularity, $\sqrt{n}(\hat{\theta}_{GMM} - \theta_0) \xrightarrow{d} \mathcal{N}(0, V(W))$
5. Overidentification gives a natural specification test through Hansen's J

> **The flexibility of GMM is conditional on the validity of the moments: invalid moments do not merely reduce efficiency — they change the probability limit of the estimator.**

Looking Back and Looking Ahead

- **OLS** used orthogonality between regressors and errors
- **IV** used instruments to generate valid moments
- **MLE** used score conditions from a likelihood
- **GMM** puts all of them into one common framework

> **Once you understand GMM, you can recognize many estimators as variations on one idea: choose parameters so that economically meaningful sample moments are as close to zero as possible.**

Cierre

¿Preguntas?

O vía E-mail:

luis.chanci@usach.cl
luischanci@santotomas.cl