Generalized method of moments

April 5, 2026 1:28 AM

Generalized method of moments (GMM)

Overview

- Estimation techniques: maximum likelihood, method of moments
- GMM is a generalization of method of moments
- Pros
  - Weak assumptions on distribution, full shape of data's distribution may not be known
- Cons
  - The solution set may not be unique
- Content page
  - Review method of moments
  - GMM in 2 parts
    - Moment conditions
    - Loss function
  - Bonus content

Original method of moments

Definition of moments (old):

- $$\mu'_n = \langle X^n \rangle \stackrel{\text{def}}{=} \begin{cases} \sum_i x_i^n f(x_i), & \text{discrete distribution} \\ \int x^n f(x)\, dx, & \text{continuous distribution} \end{cases}$$
- Examples: Mean ($E[X]$), Variance ($E[X^2]$), Skewness ($E[X^3]$), Kurtosis ...

Method of moments (old):

1. Find the **relation/equation** between the moments (old) and the distribution parameters, and setup the following system of equations
   a. $$\begin{aligned} \mu_1 &\equiv \mathrm{E}[W] = g_1(\theta_1, \theta_2, \ldots, \theta_k), \\ \mu_2 &\equiv \mathrm{E}[W^2] = g_2(\theta_1, \theta_2, \ldots, \theta_k), \\ &\vdots \\ \mu_k &\equiv \mathrm{E}[W^k] = g_k(\theta_1, \theta_2, \ldots, \theta_k). \end{aligned}$$
2. Substitute the moments with sample estimations
   a. $$\begin{aligned} \hat{\mu}_1 &= g_1(\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_k), \\ \hat{\mu}_2 &= g_2(\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_k), \\ &\vdots \\ \hat{\mu}_k &= g_k(\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_k). \end{aligned}$$
3. Solve the system of equations for parameters (θ)

Notice!!

- We **don't need** $\mu = g(\theta)$, $\mu - g(\theta) = 0$ works too
- We **don't need** equalities, inequalities would also make sense
- We **don't need** to use the old moments, these equations can be any property that the

distribution satisfies

Generalized method of moments

Moments (new): any property that the data distribution satisfies

Moment conditions

- Moment equalities
  - $m(\theta_0) \equiv \mathrm{E}[\, g(Y_t, \theta_0)\,] = 0,$
- Moment inequalities
  - $E[g(X_i, \theta)] \geq 0$
- The set of moment conditions form a system of equations (under a statistical framework)
- You can use those equations to describe any phenomena you want. They are just any properties that the distribution satisfies. e.x. game rules, physics laws etc.

How to solve

- Since these general systems are often harder to solve, we may not have analytical solutions. Instead, we solve algorithmically
- Notice
  - In equality case, deviations from our condition takes the form of $\mathrm{m} = \mathrm{E}\big[\mathrm{g}\big(\mathrm{Y_t}, \theta_0\big)\big] > 0$ or $\mathrm{m} = \mathrm{E}\big[\mathrm{g}\big(\mathrm{Y_t}, \theta_0\big)\big] < 0$
  - Thus, we can penalize this from happening by minimizing m^2
- Thus, we take m^2 as our loss function
- Since some moments are more important than others, we want to weight them, thus we want to introduce some weight into m^2
- Thus, we get our final loss function

$$\|\hat{m}(\theta)\|_W^2 = \hat{m}(\theta)^\mathsf{T}\, W \hat{m}(\theta),$$

  - Note: setting W = I makes the weights equal for all moments, so this is the same as m^2

Method of moments (general)

1. Identify your moment conditions
2. Substitute expectation with empirical analog, sample average

$$\hat{m}(\theta) \equiv \frac{1}{T}\sum_{t=1}^{T} g(Y_t, \theta)$$

3. Solve algorithmically based on loss function

$$\hat{\theta} = \arg\min_{\theta \in \Theta} \left(\frac{1}{T}\sum_{t=1}^{T} g(Y_t, \theta)\right)^\mathsf{T} \hat{W} \left(\frac{1}{T}\sum_{t=1}^{T} g(Y_t, \theta)\right)$$

System of equations properties

- L: number of moment conditions
- K: number of parameters

Scenario	Status	Outcome
$L < K$	Under-identified	No unique solution (Incomplete model).
$L = K$	Just-identified	**OMM.** Unique solution, but no way to test if your moments are "true."
$L > K$ **(Small)**	Over-identified	**Ideal GMM.** Higher efficiency and you can run the J-test.
$L \gg K$ **(Huge)**	"Many Moments"	High bias, unstable weighting matrix, and unreliable results.

1. The "Fitting the Noise" Problem

In the GMM framework, every moment condition is a sample average: .
Even if the true population expectation is zero, the sample average will almost never be exactly zero due to random sampling error. When you add a large number of these conditions:

- The optimizer tries to satisfy all of them simultaneously.
- If you have 50 moments and only 100 observations, the model starts treating the **random fluctuations** in those 50 moments as if they are meaningful signals.
- Instead of finding the true , the model "twists" the estimate to cancel out the random errors in the moments. This is essentially **overfitting** the moment conditions.

Bonus content

Comparison with other estimation techniques: Many other popular estimation techniques can be cast in terms of GMM optimization:

- Ordinary least squares (OLS) is equivalent to GMM with moment conditions:
  $\mathrm{E}[\, x_t(y_t - x_t^{\mathrm{T}}\beta)\,] = 0$
- Weighted least squares (WLS)
  $\mathrm{E}[\, x_t(y_t - x_t^{\mathrm{T}}\beta)/\sigma^2(x_t)\,] = 0$
- Instrumental variables regression (IV)
  $\mathrm{E}[\, z_t(y_t - x_t^{\mathrm{T}}\beta)\,] = 0$
- Non-linear least squares (NLLS):
  $\mathrm{E}\left[\, \nabla_\beta\, g(x_t, \beta) \cdot \left(y_t - g(x_t, \beta)\right)\right] = 0$
- Maximum likelihood estimation (MLE):
  $\mathrm{E}[\, \nabla_\theta \ln f(x_t, \theta)\,] = 0$

Two-step GMM

- *Step 1*: Take $W = I$ (the identity matrix) or some other positive-definite matrix, and compute preliminary GMM estimate $\hat{\theta}_{(1)}$. This estimator is consistent for θ_0, although not efficient.
- *Step 2*: $\hat{W}_T(\hat{\theta}_{(1)})$converges in probability to Ω^{-1} and therefore if we compute $\hat{\theta}$with this weighting matrix, the estimator will be asymptotically efficient

Example

The Economic Intuition

Imagine a consumer deciding how much to eat today versus how much to save for tomorrow. If they are optimizing their utility, the "marginal cost" of saving one dollar today must equal the "marginal benefit" of consuming that dollar (plus interest) tomorrow.
Mathematically, for a rational agent, the expectation of their discounted marginal utility should be zero based on all information available today.

1. The Moment Condition

We start with the Euler Equation:

$$E_t\left[\beta\left(\frac{C_{t+1}}{C_t}\right)^{-\gamma}R_{t+1}-1\right]=0$$

Where:

- C_t: Consumption at time t
- R_{t+1}: The gross return on an asset
- β: The discount factor (patience)
- γ: The coefficient of relative risk aversion

In GMM, we define the "error term" as the expression inside the brackets. If the model is correct, should be `uncorrelated` with any information known at time (like past returns or past consumption).

2. Instrumental Variables

To actually estimate and , we use `instruments` (Z_t). These are variables known at time t (e.g.,). Because these are known, the following must hold

$$E[(\beta(\frac{C_{t+1}}{C_t})^{-\gamma}R_{t+1}-1)\otimes Z_t]=0$$

This gives us our `sample moments`. If we have more instruments than parameters to estimate (which we usually do), the system is "overidentified."

3. The GMM Objective Function

Since we have more equations than unknowns, we can't make every sample moment exactly zero. Instead, GMM minimizes a weighted quadratic form:

$$J(\theta)=\bar{g}(\theta)^T W\bar{g}(\theta)$$

Why use GMM here instead of OLS?

- **Endogeneity:** Consumption and returns are determined simultaneously. OLS would be biased.
- **No Distributional Assumptions:** Unlike Maximum Likelihood Estimation (MLE), we don't have to assume the data follows a Normal distribution. We only assume the agents are rational (the moment condition holds).
- **Testability:** Because the model is overidentified, we can use the `Sargan-Hansen J-test` to see if our instruments are actually valid and if the model "fits" the data.