

EC241a
Fall 2004

Econometric Theory
UC Berkeley Department of Economics

GENERALIZED METHOD OF MOMENTS ESTIMATION II:
CONSISTENCY AND ASYMPTOTIC NORMALITY

The case we consider here has a M -dimensional vector of moment functions $\psi(Z, \theta)$, with the dimension of θ equal to K . We first consider the estimator $\hat{\theta}$ obtained by minimizing $Q_C(\theta) = (\sum \psi(z_i; \theta))' \cdot C \cdot (\sum \psi(z_i; \theta))$ for an arbitrary strictly positive definite symmetric matrix C .

We make the following assumptions:

Assumption 1 (COMPACTNESS)

The true value of the parameter θ_0 is an element of the interior of Θ , a compact subset of R^K .

Assumption 2 (IDENTIFICATION)

$\mathbb{E}[\psi(Z, \theta)] = 0$ implies $\theta = \theta_0$.

Assumption 3 (INDEPENDENCE)

Z_1, Z_2, \dots is a sequence of independent identically distributed random variables.

(Note that this is not necessary. Hansen (1984) deals with the more general case where the Z_i are allowed to follow a covariance stationary process.)

Assumption 4 (FULL RANK) *The expectations*

$$\Gamma = \mathbb{E} \left[\frac{\partial \psi}{\partial \theta'}(Z, \theta_0) \right],$$

and

$$\Delta = \mathbb{E} \left[\psi(Z, \theta_0) \cdot \psi(Z, \theta_0)' \right],$$

are finite and have full rank.

Assumption 5 (CONTINUITY)

$\psi(z, \theta)$ is continuous in θ and z .

Assumption 6 (DOMINANCE)

There is a function $k(z)$ such that $\|\psi(z, \theta)\| \leq k(z)$ for all z and θ and $\mathbb{E}[k(Z)]$ is finite.

Result 1 Suppose Assumptions 1–6 hold. Then

$$\hat{\theta} \xrightarrow{p} \theta_0,$$

where

$$\hat{\theta} = \operatorname{argmax}_{\theta} Q_C(\theta).$$

The argument goes as follows.¹ The dominance condition, together with continuity and compactness guarantees that $\sum \psi(z_i, \theta)/N \rightarrow \mathbb{E}[\psi(Z; \theta)]$ (and therefore $Q_C(\theta)/N^2 \rightarrow Q_C^*(\theta) = \mathbb{E}[\psi(Z; \theta)]' C \mathbb{E}[\psi(Z; \theta)]$ uniformly in θ). The limit $Q_C^*(\theta)$ has a unique minimum at θ_0 by the identification assumption. Since the limit is also continuous, minimizing the sequence $Q_C(\theta)$ leads to a consistent estimator for θ_0 , by the same argument used for consistency of maximum likelihood estimators.

We wish to weaken this a little to allow for a weight matrix that depends on the data.

Assumption 7 (SEQUENCE OF WEIGHT MATRICES)

Let C_1, C_2, \dots be a sequence of symmetric, positive definite matrices, and C_0 be a symmetric positive definite matrix such that $C_N \xrightarrow{p} C_0$.

Result 2 Suppose Assumptions 1–7 hold. Then

$$\hat{\theta} \xrightarrow{p} \theta_0,$$

¹For a formal proof see Hansen (1984), or Newey and McFadden (1994).

where

$$\hat{\theta} = \operatorname{argmax}_{\theta} Q_{C_N}(\theta).$$

Given consistency we are interested in the large sample distribution of $\hat{\theta}$. Because θ_0 is in the interior of the parameter space the first order conditions for a minimum of $Q_{C_n}(\theta)$ must be satisfied, at least in large samples. This implies that in large samples,

$$g(\hat{\theta}) = 2 \cdot \left[\sum_{i=1}^N \frac{\partial \psi'}{\partial \theta}(z_i, \hat{\theta})/N \right] \cdot C_N \cdot \left[\sum_{i=1}^N \psi(z_i, \hat{\theta})/\sqrt{N} \right] = 0.$$

As $N \rightarrow \infty$, and therefore $\hat{\theta} \rightarrow \theta_0$, the first factor converges in probability to Γ' , and the second factor converges to C_0 . Now expand the third factor in $g(\hat{\theta})$ around θ_0 to get

$$0 = \Gamma' \cdot C_0 \cdot \left[\sum_{i=1}^N \psi(z_i, \theta_0)/\sqrt{N} \right] + \Gamma' \cdot C_0 \cdot \left[\sum_{i=1}^N \frac{\partial \psi}{\partial \theta'}(z_i, \tilde{\theta})/N \right] \cdot \sqrt{N} \cdot (\hat{\theta} - \theta_0).$$

The factor multiplying $\sqrt{N} \cdot (\hat{\theta} - \theta_0)$ converges to $\Gamma' C_0 \Gamma$, so we get

$$\sqrt{N} \cdot (\hat{\theta} - \theta_0) \approx -(\Gamma' C_0 \Gamma)^{-1} \cdot \left(\Gamma' \cdot C_0 \cdot \left[\frac{1}{\sqrt{N}} \sum_{i=1}^N \psi(z_i, \theta_0) \right] \right).$$

The sum $\sum_{i=1}^N \psi(z_i, \theta_0)/\sqrt{N}$ satisfies a central limit theorem:

$$\frac{1}{\sqrt{N}} \sum_{i=1}^N \psi(z_i, \theta_0) \xrightarrow{d} \mathcal{N}(0, \Delta),$$

so we end up with:

Result 3 *Suppose Assumptions 1–7 hold. Then:*

$$\sqrt{N} \cdot (\hat{\theta} - \theta_0) \xrightarrow{d} \mathcal{N}(0, (\Gamma' C_0 \Gamma)^{-1} (\Gamma' C_0 \Delta C_0 \Gamma) (\Gamma' C_0 \Gamma)^{-1}).$$

For a formal proof see Hansen (1984) (who discusses the case with dependent, that is, autocorrelated, variables), or Newey and McFadden (1994).

There are two special cases. If M , the dimension of ψ is equal to K , the dimension of θ , both Δ and Γ are invertible $K \times K$ matrices by the full rank assumption. The choice of C_0 , or

that of the sequence C_N , does not matter now because $\hat{\theta}$ sets all the average moments exactly equal to zero, at least in large samples. Hence the variance/covariance matrix simplifies to $(\Gamma\Delta^{-1}\Gamma)^{-1} = \Gamma^{-1}\Delta(\Gamma')^{-1}$.

In the over-identified case, if we choose a sequence C_n such that $C_0 = \Delta^{-1}$, the variance/covariance matrix simplifies to $(\Gamma\Delta^{-1}\Gamma)^{-1}$. This is the optimal choice for C_0 (Hansen, 1984). Moreover, as we will see in the Chamberlain (1986) paper the resulting estimator is efficient in the sense of achieving something akin to the Cramér–Rao bound for maximum likelihood estimators, with the provision that here we are in a semi-parametric context where the distribution is not fully parametrized. To see the first part of this claim, that Δ^{-1} is the optimal choice, consider a simpler case with a scalar parameter (this is not necessary). Consider the choice $C_0 = \Delta^{-1} + a \cdot A$ for any symmetric matrix A . The variance covariance matrix for this choice of weight matrix is

$$\begin{aligned} V &= (\Gamma' C_0 \Gamma)^{-1} (\Gamma' C_0 \Delta C_0 \Gamma) (\Gamma' C_0 \Gamma)^{-1} \\ &= (\Gamma' (\Delta^{-1} + a \cdot A \Gamma)^{-1} (\Gamma' (\Delta^{-1} + a \cdot A \Delta (\Delta^{-1} + a \cdot A \Gamma) (\Gamma' (\Delta^{-1} + a \cdot A \Gamma)^{-1})) \\ &= (\Gamma' \Delta^{-1} \Gamma + a \Gamma' A \Gamma)^{-1} (\Gamma' \Delta^{-1} \Gamma + 2a \Gamma' A \Gamma + a^2 \Gamma' A \Delta A \Gamma) (\Gamma' \Delta^{-1} \Gamma + a \Gamma' A \Gamma)^{-1} \\ &= \frac{\Gamma' \Delta^{-1} \Gamma + 2a \Gamma' A \Gamma + a^2 \Gamma' A \Delta A \Gamma}{(\Gamma' \Delta^{-1} \Gamma + a \Gamma' A \Gamma)^2}. \end{aligned}$$

The derivative of this with respect to a is zero at $a = 0$, and the second derivative is positive. Hence $a = 0$ is the minimizing value, and since this is true for any matrix A , $C_0 = \Delta^{-1}$ is the optimal choice for the weight matrix.

The way Hansen (1984) suggests implementing this estimator is in a two-step procedure. First we minimize $Q_C(\theta)$ for some arbitrary positive definite matrix C . One option is the identity matrix. Alternatively we can use a diagonal matrix with zeros on $M - K$ of the diagonal elements. (In that case the matrix is obviously not positive definite, but its rank is high enough that it should not matter). Let $\hat{\theta}_{\text{init}}$ be the estimate from this procedure. Then we estimate the inverse of the optimal weight matrix as $\hat{\Delta} = \sum \psi(z_i, \hat{\theta}_{\text{init}}) \cdot \psi(z_i, \hat{\theta}_{\text{init}}) / N$. followed by minimizing $Q_{\hat{\Delta}^{-1}}(\theta)$. The asymptotic distribution is not affected by the first round choice of C . The numerical value of the final estimates, as well as the small sample

properties are obviously affected by this choice. A disadvantage of this procedure is the degree of arbitrariness introduced by the choice of the initial weight matrix. We return to this issue later. Note, however that for exactly identified models with $M = K$ this issue is moot.

Another result for overidentified models is the basis of specification tests:

Result 4 *If $M > K$, as $N \rightarrow \infty$, and with $C_N \rightarrow C_0 = \Delta$, then*

(i)

$$\frac{1}{N} \cdot Q_{C_N}(\theta_0) = \left[\frac{1}{\sqrt{N}} \sum_{i=1}^N \psi(z_i, \theta_0) \right]' \cdot C_N \cdot \left[\frac{1}{\sqrt{N}} \sum_{i=1}^N \psi(z_i, \theta_0) \right] \xrightarrow{d} \mathcal{X}^2(M),$$

and, (ii),

$$\frac{1}{N} \cdot Q_{C_N}(\hat{\theta}) = \left[\frac{1}{\sqrt{N}} \sum_{i=1}^N \psi(z_i, \hat{\theta}) \right]' \cdot C_N \cdot \left[\frac{1}{\sqrt{N}} \sum_{i=1}^N \psi(z_i, \hat{\theta}) \right] \xrightarrow{d} \mathcal{X}^2(M - K).$$

Obviously $\sum \psi(z_i, \theta_0)/\sqrt{N}$ satisfies a central limit theorem and we have a Chi-squared distribution with M degrees of freedom for the normalized objective function. Every parameter we have to estimate reduces the degrees of freedom by one. Once there are as many unknown parameters as moments there is nothing left to test and the statistic is equal to zero. To see this note that

$$\frac{1}{\sqrt{N}} \sum_{i=1}^N \psi(z_i, \hat{\theta}) \approx \frac{1}{\sqrt{N}} \sum_{i=1}^N \psi(z_i, \theta_0) + \Gamma \sqrt{N}(\hat{\theta} - \theta_0).$$

Also,

$$\sqrt{N}(\hat{\theta} - \theta_0) \approx -(\Gamma' C_0 \Gamma)^{-1} \cdot \left(\Gamma' \cdot C_0 \cdot \left[\frac{1}{\sqrt{N}} \sum_{i=1}^N \psi(z_i, \theta_0) \right] \right).$$

Hence,

$$\frac{1}{\sqrt{N}} \sum_{i=1}^N \psi(z_i, \hat{\theta}) \approx \left(\mathcal{I} - \Gamma(\Gamma' C_0 \Gamma)^{-1} \cdot \left(\Gamma' \cdot C_0 \right) \right) \frac{1}{\sqrt{N}} \sum_{i=1}^N \psi(z_i, \theta_0).$$

If $C_0 = \Delta^{-1}$, this is approximately

$$\frac{1}{\sqrt{N}} \sum_{i=1}^N \psi(z_i, \hat{\theta}) \approx (\mathcal{I} - \Gamma(\Gamma' \Delta^{-1} \Gamma)^{-1} \cdot \Gamma' \cdot \Delta^{-1}) \frac{1}{\sqrt{N}} \sum_{i=1}^N \psi(z_i, \theta_0).$$

Thus the variance of $\sum_i \psi(z_i, \hat{\theta})/\sqrt{N}$ is approximately equal to

$$\begin{aligned} & (\mathcal{I} - \Gamma(\Gamma'\Delta^{-1}\Gamma)^{-1} \cdot \Gamma' \cdot \Delta^{-1}) \Delta (\mathcal{I} - \Delta^{-1}\Gamma(\Gamma'\Delta^{-1}\Gamma)^{-1} \cdot \Gamma') \\ &= \Delta - \Gamma (\Gamma'\Delta^{-1}\Gamma)^{-1} \Gamma' \end{aligned}$$

Defining $\Delta^{1/2}$ so that $\Delta = \Delta^{1/2}(\Delta^{1/2})'$, and $B' = \Gamma'\Delta^{-1/2}$ (an $M \times K$ matrix), this is equal to

$$\Delta^{1/2} (\mathcal{I} - B(B'B)^{-1}B') (\Delta^{1/2})'.$$

Then the matrix in the middle, $\mathcal{I} - B(B'B)^{-1}B'$, is a projection matrix with rank equal to $M - K$, the difference between the number of moments and the number of estimated parameters. Hence the quadratic form is going to have a chi-squared distribution with degrees of freedom equal to the number of overidentifying restrictions.

Next, consider the special case where the moments are the score from the likelihood:

$$\psi(Z, \theta) = \frac{\partial \ln f}{\partial \theta}(Z; \theta).$$

Then we do have a moment with expectation zero, so we can go ahead with the GMM procedures. Consider the large sample covariance matrix:

$$\begin{aligned} [\Gamma'\Delta^{-1}\Gamma]^{-1} &= \left[\mathbb{E} \frac{\partial \psi}{\partial \theta'}(Z, \theta_0) \right]' \left[\mathbb{E} \psi(Z, \theta_0) \psi(Z, \theta_0)' \right]^{-1} \left[\mathbb{E} \frac{\partial \psi}{\partial \theta'}(Z, \theta_0) \right]^{-1} \\ &= \left(\left[\mathbb{E} \frac{\partial^2 \ln f}{\partial \theta \partial \theta'}(Z, \theta_0) \right]' \left[\mathbb{E} \frac{\partial \ln f}{\partial \theta}(Z, \theta_0) \frac{\partial \ln f}{\partial \theta'}(Z, \theta_0) \right]^{-1} \left[\mathbb{E} \frac{\partial^2 \ln f}{\partial \theta \partial \theta'}(Z, \theta_0) \right] \right)^{-1}. \end{aligned}$$

Under the information matrix equality this variance simplifies to

$$\mathcal{I}(\theta_0)^{-1} = - \left(\mathbb{E} \left[\frac{\partial^2 \ln f}{\partial \theta \partial \theta'}(Z, \theta_0) \right] \right)^{-1} = \left(\mathbb{E} \left[\frac{\partial \ln f}{\partial \theta}(Z, \theta_0) \frac{\partial \ln f}{\partial \theta'}(Z, \theta_0) \right] \right)^{-1}.$$

Under general miss-specification the maximum likelihood estimator will, under the GMM regularity conditions, be consistent for the value of θ that sets $\mathbb{E}[\frac{\partial \ln f}{\partial \theta}(Z, \theta)]$ equal to zero, with the three matrix formulation of the variance. This is the focus of the paper by White (1984). A question that is difficult to answer in general is what the interest is in the parameter that is being estimated under this general form of miss-specification.

Another case of interest is the standard linear model:

$$\hat{\beta}_{ols} = \operatorname{argmin}_{\beta} \sum_{i=1}^N (y_i - x_i' \beta)^2.$$

This corresponds to the gmm estimator with moments

$$\psi(y, x, \beta) = x \cdot (y - x' \beta).$$

The parameter being estimated is the best linear predictor, equal to

$$\beta_{blp} = (\mathbb{E}[XX'])^{-1} \mathbb{E}[XY].$$

The asymptotic variance of the normalized estimator is

$$(\mathbb{E}[XX'])^{-1} \mathbb{E}[X \varepsilon^2 X'] (\mathbb{E}[XX'])^{-1},$$

where $\varepsilon = Y - X' \beta_{blp}$. If the disturbances are homoskedastic, the variance reduces to the standard $\mathbb{E}[\varepsilon^2] (\mathbb{E}[XX'])^{-1}$. The GMM variance, for this case also known as the Eicker-White heteroskedasticity-consistent variance does not rely on distributional or functional form assumptions: there is in general a β_{blp} that sets the average moments equal to zero.

ADDITIONAL REFERENCES

WHITE, H., (1982), "Maximum Likelihood Estimation of Misspecified Models" *Econometrica*, Vol. 48, 817–838.

WHITE, H., (1980), "A Heteroskedasticity Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity" *Econometrica*, Vol 50, 1–25.