Lecture 5
Covariance Estimation and Optimal Weighting Matrices[1]

In this lecture, we consider estimation of the asymptotic covariance matrix $B_0^{-1}\Omega_0 B_0^{-1}$ of the extremum estimator $\widehat{\theta}_n$.

1 Covariance Estimation

Lemma 4.1 and Assumptions EE2(i) and CF(iv)* combine to yield

$$\widehat{B}_n^{-1} = \left(\frac{\partial^2}{\partial\theta\partial\theta'}\hat{Q}_n(\widehat{\theta}_n)\right)^{-1} \to_p B_0^{-1}. \tag{1}$$

Hence, it remains to find a consistent estimator of Ω_0. The general principle employed is that of forming estimators by replacing expectations with sample averages and unknown parameters with consistent estimators of them. Then, Lemma 4.1 can be used to establish consistency of the resulting estimator $\widehat{\Omega}_n$.

We consider each of the examples in turn:

(1) ML Estimator: Let

$$\begin{aligned} \widehat{B}_n &= -\frac{1}{n}\textstyle\sum_{i=1}^n \dfrac{\partial^2}{\partial\theta\partial\theta'}\log f(W_i,\widehat{\theta}_n) \text{ and} \\ \widehat{\Omega}_n &= \frac{1}{n}\textstyle\sum_{i=1}^n \dfrac{\partial}{\partial\theta}\log f(W_i,\widehat{\theta}_n)\dfrac{\partial}{\partial\theta'}\log f(W_i,\widehat{\theta}_n). \end{aligned} \tag{2}$$

We obtain $\widehat{\Omega}_n \to_p \Omega_0$ by verifying conditions (i), (ii), and (iii) of Lemma 4.1. Condition (i) holds by consistency of $\widehat{\theta}_n$. Conditions (ii) and (iii) hold by the ULLN in Section 4 provided $\frac{\partial}{\partial\theta}\log f(w,\theta)$ (or, equivalently, $f(w,\theta)$ and $\frac{\partial}{\partial\theta}f(w,\theta)$) is continuous in θ on Θ_0 $\forall w \in \mathcal{W}$ (as is assumed in Lecture 4) and

$$E\sup_{\theta\in\Theta_0}\left|\left|\frac{\partial}{\partial\theta}\log f(W_i,\theta)\right|\right|^2 < \infty,$$

where Θ_0 is a compact neighborhood of θ_0.

If the model is correctly specified, then $B_0 = \Omega_0$ and the covariance matrix $B_0^{-1}\Omega_0 B_0^{-1}$ can be estimated by $\widehat{B}_n^{-1}\widehat{\Omega}_n\widehat{B}_n^{-1}$, $\widehat{B}_n^{-1}$, or $\widehat{\Omega}_n^{-1}$. Note that $\widehat{\Omega}_n$ only requires calculation of the first derivative of $f(w,\theta)$, whereas $\widehat{B}_n$ requires calculation of the second derivatives.

[1] The notes for this lecture is largely adapted from the notes of Donald Andrews on the same topic I am grateful for Professor Andrews' generosity and elegant exposition. All errors are mine.

(2) LS Estimator: Let

$$\begin{aligned}\widehat{B}_n &= \frac{1}{n}\sum_{i=1}^n \left(\frac{\partial}{\partial\theta}g(X_i,\widehat{\theta}_n)\frac{\partial}{\partial\theta'}g(X_i,\widehat{\theta}_n) - (Y_i - g(X_i,\widehat{\theta}_n))\frac{\partial^2}{\partial\theta\partial\theta'}g(X_i,\widehat{\theta}_n)\right) \text{ and} \\ \widehat{\Omega}_n &= \frac{1}{n}\sum_{i=1}^n (Y_i - g(X_i,\widehat{\theta}_n))^2 \frac{\partial}{\partial\theta}g(X_i,\widehat{\theta}_n)\frac{\partial}{\partial\theta'}g(X_i,\widehat{\theta}_n). \qquad (3)\end{aligned}$$

We obtain $\widehat{\Omega}_n \to_p \Omega_0$ by verifying the three conditions of Lemma 4.1. Condition (i) holds by consistency of $\widehat{\theta}_n$. Conditions (ii) and (iii) hold by ULLN provided $g(x,\theta)$ and $\frac{\partial}{\partial\theta}g(x,\theta)$ are continuous in θ on Θ_0 $\forall x \in \mathcal{X}$ (as is assumed in Lecture 4) and

$$E\sup_{\theta\in\Theta_0}\left|\left|(Y_i - g(X_i,\theta))\frac{\partial}{\partial\theta}g(X_i,\theta)\right|\right|^2 < \infty,$$

where Θ_0 is a compact neighborhood of θ_0.

If the regression model is correctly specified (i.e., $E(Y_i|X_i) = g(X_i,\theta_0)$ a.s.), then B_0 simplifies and $\widehat{B}_n$ can be simplified correspondingly. Let

$$\widetilde{B}_n = \frac{1}{n}\sum_{i=1}^n \frac{\partial}{\partial\theta}g(X_i,\widehat{\theta}_n)\frac{\partial}{\partial\theta'}g(X_i,\widehat{\theta}_n). \qquad (4)$$

In the correctly specified case, $\widetilde{B}_n \to_p B_0$ when CF(iv) holds. So, a consistent covariance matrix estimator for a correctly specified regression model is

$$\widetilde{B}_n^{-1}\widehat{\Omega}_n\widetilde{B}_n^{-1}. \qquad (5)$$

Note that this estimator allows for conditional heteroskedasticity of the errors — i.e., it is a "heteroskedasticity consistent" covariance matrix estimate.

If the model is correctly specified and the errors are conditionally homoskedastic, then $\Omega_0 = \sigma^2 B_0$ and $\widehat{\Omega}_n$ can be replaced by the estimator

$$\widehat{\sigma}^2\widetilde{B}_n^{-1}, \text{ where } \widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n (Y_i - g(X_i,\widehat{\theta}_n)^2. \qquad (6)$$

For a linear regression model, $\widehat{\sigma}^2\widetilde{B}_n^{-1}$ equals $\widehat{\sigma}^2\left(\frac{1}{n}\sum_{i=1}^n X_iX_i'\right)^{-1}$.

(3) GMM Estimators: Let

$$\begin{aligned}\widehat{B}_n &= \left[\frac{1}{n}\sum_{i=1}^n \frac{\partial}{\partial\theta'}g(W_i,\widehat{\theta}_n)\right]' A_n'A_n\frac{1}{n}\sum_{i=1}^n \frac{\partial}{\partial\theta'}g(W_i,\widehat{\theta}_n) \text{ and} \qquad (7) \\ \widehat{\Omega}_n &= \left[\frac{1}{n}\sum_{i=1}^n \frac{\partial}{\partial\theta'}g(W_i,\widehat{\theta}_n)\right]' A_n'A_n\frac{1}{n}\sum_{i=1}^n g(W_i,\widehat{\theta}_n)g(W_i,\widehat{\theta}_n)'A_n'A_n\frac{1}{n}\sum_{i=1}^n \frac{\partial}{\partial\theta'}g(W_i,\widehat{\theta}_n).\end{aligned}$$

Note that the definition of $\widehat{B}_n$ does not include the second summand of $\frac{\partial^2}{\partial\theta\partial\theta'}\hat{Q}_n(\widehat{\theta}_n)$ in Equation (33) in Lecture 4. The reason is that the second summand converges in probability to zero since $Eg(W_i,\theta_0)=0$ and, hence, can be omitted.

Each component of $\widehat{\Omega}_n$ has been shown in Lecture 4 to converge in probability to the corresponding component of Ω_0. The only exception is the component $\frac{1}{n}\sum_{i=1}^n g(W_i,\widehat{\theta}_n)g(W_i,\widehat{\theta}_n)'$. The latter converges in probability to $Eg(W_i,\theta_0)g(W_i,\theta_0)'$ by Lemma 12.1 and Theorem 11.3 provided $\widehat{\theta}_n \to_p \theta_0$, $g(w,\theta)$ is continuous in θ on Θ_0 $\forall w \in \mathcal{W}$ and

$$E \sup_{\theta\in\Theta_0} ||g(W_i,\theta)||^2 < \infty,$$

where Θ_0 is a compact neighborhood of θ_0.

(4) MD Estimators: Let

$$\begin{aligned}\widehat{B}_n &= \left(\frac{\partial}{\partial\theta'}g(\widehat{\theta}_n)\right)' A_n' A_n \frac{\partial}{\partial\theta'}g(\widehat{\theta}_n) \text{ and} \\ \widehat{\Omega}_n &= \left(\frac{\partial}{\partial\theta'}g(\widehat{\theta}_n)\right)' A_n' A_n \widehat{V}_n A_n' A_n \frac{\partial}{\partial\theta'}g(\widehat{\theta}_n),\end{aligned} \tag{8}$$

where $\widehat{V}_n$ is some consistent estimator of V_0, the asymptotic covariance matrix of $\sqrt{n}(\widehat{\pi}_n - \pi_0)$. Note that the definition of $\widehat{B}_n$ does not include the second summand of $\frac{\partial^2}{\partial\theta\partial\theta'}\hat{Q}_n(\widehat{\theta}_n)$ in Equation (37) in Lecture 4, because the latter converges in probability to zero given that $\pi_0 = g(\theta_0)$. Each component of $\widehat{\Omega}_n$ has been shown in Lecture 12 to converge in probability to the corresponding component of Ω_0, except $\widehat{V}_n$. We simply assume $\widehat{V}_n \to_p V_0$ here, because the specific form of V_0 has not been stated.

(5) TS Estimators: Let

$$\begin{aligned}\widehat{B}_n &= \left(\frac{\partial}{\partial\theta'}G_n(\widehat{\theta}_n,\widehat{\tau}_n)\right)' A_n' A_n \frac{\partial}{\partial\theta'}G_n(\widehat{\theta}_n,\widehat{\tau}_n) \text{ and} \\ \widehat{\Omega}_n &= \left(\frac{\partial}{\partial\theta'}G_n(\widehat{\theta}_n,\widehat{\tau}_n)\right)' A_n' A_n (\widehat{V}_{1n} + \widehat{\Lambda}_n\widehat{V}_{2n}' + \widehat{V}_{2n}\widehat{\Lambda}_n' + \widehat{\Lambda}_n\widehat{V}_{3n}\widehat{\Lambda}_n') \\ &\quad \times A_n' A_n \frac{\partial}{\partial\theta'}G_n(\widehat{\theta}_n,\widehat{\tau}_n), \text{ where} \\ \widehat{\Lambda}_n &= \frac{\partial}{\partial\tau'}G_n(\widehat{\theta}_n,\widehat{\tau}_n)\end{aligned} \tag{9}$$

and $\widehat{V}_{jn}$ is some consistent estimator of V_{j0} for $j=1,2,3$. If Λ_0 is zero, as occurs in some cases, such as feasible GLS estimation, then one can take $\widehat{\Lambda}_n = 0$ and the estimators $\widehat{V}_{2n}$ and $\widehat{V}_{3n}$ are not required.

2 Optimal Weight Matrices for GMM, MD, and TS Estimators

The GMM, MD, and TS estimators have asymptotic covariance matrices of the form

$$(\Gamma_0' C \Gamma_0)^{-1} \Gamma_0' C \Sigma_0 C \Gamma_0 (\Gamma_0' C \Gamma_0)^{-1}, \tag{10}$$

where $C = A'A$ and Σ_0 is a symmetric positive semi-definite (psd) matrix that depends on the estimator. We will show that the optimal choice of weight matrix A_n is a choice such that

$$A'A = \Sigma_0^{-1}, \text{ where } A_n \xrightarrow{p} A. \tag{11}$$

This choice minimizes the asymptotic covariance matrix of $\widehat{\theta}_n$.

When (11) holds, the asymptotic covariance matrix in (10) simplifies to $(\Gamma_0' \Sigma_0^{-1} \Gamma_0)^{-1}$. We will show that

$$(\Gamma_0' C \Gamma_0)^{-1} \Gamma_0' C \Sigma_0 C \Gamma_0 (\Gamma_0' C \Gamma_0)^{-1} - (\Gamma_0' \Sigma_0^{-1} \Gamma_0)^{-1} \geq 0, \tag{12}$$

where "≥ 0" denotes "is psd." Note that $F^{-1} - G^{-1} \geq 0$ if and only if $G - F \geq 0$. Thus, (12) holds if and only if

$$\Gamma_0' \Sigma_0^{-1} \Gamma_0 - \Gamma_0' C \Gamma_0 (\Gamma_0' C \Sigma_0 C \Gamma_0)^{-1} \Gamma_0' C \Gamma_0 \geq 0. \tag{13}$$

The left-hand side of (12) equals

$$\begin{aligned} & \Gamma_0' \Sigma_0^{-1/2} \left[I_k - \Sigma_0^{1/2} C \Gamma_0 (\Gamma_0' C \Sigma_0 C \Gamma_0)^{-1} \Gamma_0' C \Sigma_0^{1/2} \right] \Sigma_0^{-1/2} \Gamma_0 \\ = \; & HPH' \\ = \; & HP(HP)' \\ \geq \; & 0, \end{aligned} \tag{14}$$

where $H = \Gamma_0' \Sigma_0^{-1/2}$, $P = I_k - \Sigma_0^{1/2} C \Gamma_0 (\Gamma_0' C \Sigma_0 C \Gamma_0)^{-1} \Gamma_0' C \Sigma_0^{1/2}$, and the second equality uses the fact that P is a projection matrix (i.e., P is symmetric and idempotent, $P^2 = P$). A matrix of the form $HP(HP)'$ is necessarily psd, since $z'HP(HP)'z = ||PH'z||^2 \geq 0 \; \forall z \in R^d$.

In sum, the optimal weight matrix for the GMM, MD, and TS estimators depends on the asymptotic covariance matrix of $\sqrt{n} \frac{\partial}{\partial \theta} \hat{Q}_n(\theta_0)$, which is $\Omega_0 = \Gamma_0' C \Sigma_0 C \Gamma_0$. For the GMM and MD estimators, $\Sigma_0 = V_0$ and the optimal weight matrix A_n is such that

$$A_n' A_n \xrightarrow{p} A'A = V_0^{-1}. \tag{15}$$

For the TS estimator, the optimal weight matrix A_n is such that

$$A_n' A_n \xrightarrow{p} A'A = (V_{10} + \Lambda_0 V_{20}' + V_{20}\Lambda_0' + \Lambda_0 V_{30}\Lambda_0')^{-1}. \tag{16}$$

Two-Step GMM

It is usually desirable to use the optimal weight matrix rather than an arbitrary weight matrix when doing GMM and MD estimations. However, the optimal weight matrices depend on the covariance of the moment functions in the GMM case and the asymptotic variance of the initial estimator in the MD case. Either the covariance or the asymptotic variance is not known. Therefore, we need to obtain consistent estimators for them.

In the case of MD, the asymptotic variance of the initial estimator $\hat{\pi}_n$ can be estimated from the initial procedure used to obtain $\hat{\pi}_n$. Denote the estimator by V_n^*.

In the case of GMM, the covariance $\Sigma_0 = E\left(g\left(W_i, \theta_0\right) g\left(W_i, \theta_0\right)\right)$ can be consistently estimated by $\hat{\Sigma}_n^* = n^{-1}\sum_{i=1}^n g\left(W_i, \hat{\theta}_{n,1}\right) g\left(W_i, \hat{\theta}_{n,1}\right)'$ for some consistent estimator $\hat{\theta}_n$. The consistent estimator $\hat{\theta}_n$ may be obtained using GMM with the identity matrix as the weight matrix.

After estimating V_n^* and $\hat{\Sigma}_n^*$, we can use $A_n = sqrtm(V_n^*)$ and $A_n = sqrtm(\hat{\Sigma}_n^*)$ as the estimated optimal weight matrix to carry out GMM and MD estimation, respectively. The GMM/MD estimators obtained have the same asymptotic variance as GMM/MD estimators using the (infeasible) optimal weight matrices. The reason simply is that our estimators for the optimal weight matrices are consistent.

The GMM estimators obtained using the above procedure are called two-step GMM estimators because in this procedure, GMM estimation is carried out twice.

Multi-step GMM: Now that the two-step GMM estimators are "better" (in a second order asymptotic sense) than a one-step GMM estimator with identity weight matrix (or any other arbitrary weight matrix that is not the optimal weight matrix). Suppose that we estimate Σ_0 again using the two-step GMM estimators. The estimated Σ_0 can be reasonably believed to be better than the $\hat{\Sigma}_n^*$. You may wonder if we should run GMM again using the better covariance matrix estimator, and obtain a 3-step GMM estimator. You may also want to keep the iteration going. (**Iterative GMM**)

The multi-step GMM estimators do not improve upon the two-step version in first order (consistency) or second order (asymptotic variance) asymptotics. Some argues that finite sample property (like mean-bias or median-bias) might be better for iterated GMM.

Continuous Updating GMM (CUE): unlike the iterative or two-step GMM, CUE does not take the weight matrix as given. Instead, it treats the weight matrix as a function of θ, and minimizes:

$$\hat{Q}_n\left(\theta\right) = \bar{g}_n\left(\theta\right)' \hat{\Sigma}_n\left(\theta\right) \bar{g}_n\left(\theta\right), \tag{17}$$

where

$$\hat{\Sigma}_n(\theta) = n^{-1} \sum_{i=1}^{n} g(W_i, \theta) g(W_i, \theta)'. \tag{18}$$

The CUE is consistent and has the same asymptotic variance as the two-step or the iterative GMM estimators.

All three procedures are used in practice and none dominates the others. See Hansen, Heaton and Yaron (1994) for a Monte Carlo experiments that compare the three in finance applications. (Homework question: find a GMM problem in a cross-section setting and compare the three procedures by simulation)