require of course that $l \geq k$, where the parameter vector $\hat{\boldsymbol{\theta}}$ has k elements. The empirical moment conditions that we use for estimation can be expressed as

$$\boldsymbol{W}^\top \boldsymbol{f}(\boldsymbol{\theta}) = \mathbf{0}, \tag{17.47}$$

where $\boldsymbol{f}$ is an n–vector with typical component f_t. If $l = k$, the estimator $\hat{\boldsymbol{\theta}}$ is obtained by solving the k equations (17.47). If $l > k$, it is obtained by minimizing the quadratic form constructed from the components of the left-hand side of (17.47) and an estimate of their covariance matrix. Let $\boldsymbol{\Omega}$ denote the covariance matrix of the f_t's. Thus, if the DGP is denoted by μ and the true parameter vector by $\boldsymbol{\theta}_0$,

$$\Omega_{ts} = E_\mu\big(f_t(\boldsymbol{\theta}_0) f_s(\boldsymbol{\theta}_0) \,|\, \Omega_t\big) \quad \text{for all } t \leq s.$$

Then the conditional covariance matrix of the empirical moments in (17.47) is $\boldsymbol{\Phi} \equiv \boldsymbol{W}^\top \boldsymbol{\Omega} \boldsymbol{W}$.

In the usual case, with $l > k$, the criterion function used for obtaining parameter estimates is

$$\boldsymbol{f}(\boldsymbol{\theta})^\top \boldsymbol{W} \big(\boldsymbol{W}^\top \boldsymbol{\Omega} \boldsymbol{W}\big)^{-1} \boldsymbol{W}^\top \boldsymbol{f}(\boldsymbol{\theta}).$$

The asymptotic covariance matrix of this estimator is given by the probability limit of $(\boldsymbol{D}^\top \boldsymbol{\Phi}^{-1} \boldsymbol{D})^{-1}$, where

$$\boldsymbol{D}_{ij} = \operatorname*{plim}_{n \to \infty} \left(\frac{1}{n} \sum_{t=1}^{n} W_{ti} \frac{\partial f_t}{\partial \theta_j} \right). \tag{17.48}$$

Let $\boldsymbol{J}(\boldsymbol{y}, \boldsymbol{\theta})$ denote the $n \times k$ matrix with typical element $\partial f_t(y_t, \boldsymbol{\theta}) / \partial \theta_j$.[1] Then the right-hand side of (17.48) is the limit of $n^{-1} \boldsymbol{W}^\top \boldsymbol{J}$. Thus the asymptotic covariance matrix of $n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ reduces to the limit of

$$\left(\Big(\frac{1}{n} \boldsymbol{J}^\top \boldsymbol{W}\Big) \Big(\frac{1}{n} \boldsymbol{W}^\top \boldsymbol{\Omega} \boldsymbol{W}\Big)^{-1} \Big(\frac{1}{n} \boldsymbol{W}^\top \boldsymbol{J}\Big) \right)^{-1}. \tag{17.49}$$

The first result about how to choose the instruments $\boldsymbol{W}$ optimally is simple and intuitive. It is that if we increase the number of instruments, the limiting covariance matrix (17.49) cannot increase. Imagine that instead of the empirical moment conditions (17.47) we use a set of linear combinations of them. That is, we replace (17.47) by

$$\boldsymbol{B}^\top \boldsymbol{W}^\top \boldsymbol{f}(\boldsymbol{\theta}) = \mathbf{0},$$

[1] The notation $\boldsymbol{J}$ was chosen because the matrix is the *Jacobian* of $\boldsymbol{f}$ with respect to $\boldsymbol{\theta}$ and because $\boldsymbol{F}$ was previously used to denote something else.

for some $l \times p$ matrix $\boldsymbol{B}$, where $p \leq l$. It is easy to see that this corresponds to replacing $\boldsymbol{D}$ by $\boldsymbol{B}^\top\boldsymbol{D}$ and $\boldsymbol{\Phi}$ by $\boldsymbol{B}^\top\boldsymbol{\Phi}\boldsymbol{B}$. Consider the difference

$$\boldsymbol{D}^\top\boldsymbol{\Phi}^{-1}\boldsymbol{D} - \boldsymbol{D}^\top\boldsymbol{B}\big(\boldsymbol{B}^\top\boldsymbol{\Phi}\boldsymbol{B}\big)^{-1}\boldsymbol{B}^\top\boldsymbol{D}$$

between the inverses of the $k\times k$ asymptotic covariance matrices corresponding to the instruments $\boldsymbol{W}$ and $\boldsymbol{WB}$, respectively. If, as before, we denote by $\boldsymbol{\Psi}$ a symmetric $l \times l$ matrix such that $\boldsymbol{\Psi}^2 = \boldsymbol{\Phi}^{-1}$, this difference is

$$\boldsymbol{D}^\top\boldsymbol{\Psi}\Big(\mathbf{I} - \boldsymbol{\Psi}^{-1}\boldsymbol{B}\big(\boldsymbol{B}^\top\boldsymbol{\Psi}^{-2}\boldsymbol{B}\big)^{-1}\boldsymbol{B}^\top\boldsymbol{\Psi}^{-1}\Big)\boldsymbol{\Psi}\boldsymbol{D}. \tag{17.50}$$

This matrix is clearly positive semidefinite, because the matrix in large parentheses is the orthogonal projection off the columns of $\boldsymbol{\Psi}^{-1}\boldsymbol{B}$. For any two symmetric, positive definite matrices $\boldsymbol{P}$ and $\boldsymbol{Q}$ of the same dimension, $\boldsymbol{P} - \boldsymbol{Q}$ is positive semidefinite if and only if $\boldsymbol{Q}^{-1} - \boldsymbol{P}^{-1}$ is positive semidefinite (see Appendix A). Thus the fact that (17.50) is positive semidefinite establishes our first result.

This result might seem to suggest that one should always use as many instruments as possible in order to get as efficient estimates as possible. Such a conclusion is generally wrong, however. Recall the discussion in Section 7.5, illustrated by Figure 7.1. There we saw that, in the ordinary IV context, there is a trade-off between asymptotic efficiency and bias in finite samples. The same trade-off arises in the GMM case as well. Using a large number of overidentifying restrictions may lead to a smaller asymptotic covariance matrix, but the estimates may be seriously biased. Another argument against the use of too many instruments is simply that there are inevitably diminishing returns, on account of the existence of the GMM bound.

The second result shows how to choose the instruments $\boldsymbol{W}$ optimally. It says that if we set $\boldsymbol{W} = \boldsymbol{\Omega}^{-1}\boldsymbol{J}$ in (17.47), then the asymptotic covariance matrix that results is smaller than the one given by any other choice. From (17.49) it then follows that the GMM bound for the asymptotic covariance matrix is $\operatorname{plim}\,(n^{-1}\boldsymbol{J}^\top\boldsymbol{\Omega}^{-1}\boldsymbol{J})^{-1}$. Unfortunately, as we will see, this result is not always useful in practice.

The proof is very simple. As with the first result, it is easiest to work with the inverses of the relevant covariance matrices. Let the symmetric $n \times n$ matrix $\boldsymbol{\Upsilon}$ be defined so that $\boldsymbol{\Upsilon}^2 \equiv \boldsymbol{\Omega}$. Then, suppressing limits and factors of n for the moment, we see that

$$\begin{aligned} &\boldsymbol{J}^\top\boldsymbol{\Omega}^{-1}\boldsymbol{J} - \boldsymbol{J}^\top\boldsymbol{W}\,\big(\boldsymbol{W}^\top\boldsymbol{\Omega}\boldsymbol{W}\big)^{-1}\,\boldsymbol{W}^\top\boldsymbol{J} \\ &= \boldsymbol{J}^\top\boldsymbol{\Upsilon}^{-1}\Big(\mathbf{I} - \boldsymbol{\Upsilon}\boldsymbol{W}\,\big(\boldsymbol{W}^\top\boldsymbol{\Upsilon}^2\boldsymbol{W}\big)^{-1}\boldsymbol{W}^\top\boldsymbol{\Upsilon}\Big)\boldsymbol{\Upsilon}^{-1}\boldsymbol{J}. \end{aligned} \tag{17.51}$$

Since the matrix in large parentheses is the orthogonal projection off the columns of $\boldsymbol{\Upsilon}\boldsymbol{W}$, this expression is positive semidefinite, and the second result is established.