where $\boldsymbol{f}_i(\cdot)$ is an n–vector of nonlinear functions, $\boldsymbol{u}_i$ is an n–vector of error terms, and $\boldsymbol{\theta}$ is a p–vector of parameters to be estimated. In general, subject to whatever restrictions need to be imposed for the system to be identified, all the endogenous and exogenous variables and all the parameters may appear in any equation.

The first step in any sort of IV procedure is to choose the instruments to be used. If the model is nonlinear only in the parameters, the matrix of optimal instruments is $\boldsymbol{X}$. As we have seen, however, there is no simple way to choose the instruments for models that are nonlinear in one or more of the endogenous variables. The theory of Section 17.4 can be applied, of course, but the result that it yields is not very practical. Under the usual assumptions about the error terms, namely, that they are homoskedastic and independent across observations but correlated across equations for each observation, one finds that a matrix of instruments $\boldsymbol{W}$ will be optimal if $\mathcal{S}(\boldsymbol{W})$ is equal to the subspace spanned by the union of the columns of the $E(\partial \boldsymbol{f}_i/\partial\boldsymbol{\theta})$. This result was originally derived by Amemiya (1977). It makes sense but is generally not very useful in practice. For now, we simply assume that *some* valid $n\times m$ matrix of instruments $\boldsymbol{W}$ is available, with $m \geq p$.

A nonlinear IV procedure for full-system estimation, similar in spirit to the single-equation NL2SLS procedure based on minimizing (18.78), was first proposed by Jorgenson and Laffont (1974) and called **nonlinear three-stage least squares**, or **NL3SLS**. The name is somewhat misleading, for the same reason that the name "NL2SLS" is misleading. By analogy with (18.60), the criterion function we would really like to minimize is

$$\sum_{i=1}^{g}\sum_{j=1}^{g}\sigma^{ij}\boldsymbol{f}_i^{\top}(\boldsymbol{Y},\boldsymbol{X},\boldsymbol{\theta})\boldsymbol{P}_W\boldsymbol{f}_j(\boldsymbol{Y},\boldsymbol{X},\boldsymbol{\theta}). \tag{18.80}$$

In practice, however, the elements σ^{ij} of the inverse of the contemporaneous covariance matrix $\boldsymbol{\Sigma}$ will not be known and will have to be estimated. This may be done in several ways. One possibility is to use NL2SLS for each equation separately. This will generally be easy, but it may not be possible if some parameters are identified only by cross-equation restrictions. Another approach which will work in that case is to minimize the criterion function

$$\sum_{i=1}^{g}\boldsymbol{f}_i^{\top}(\boldsymbol{Y},\boldsymbol{X},\boldsymbol{\theta})\boldsymbol{P}_W\boldsymbol{f}_i(\boldsymbol{Y},\boldsymbol{X},\boldsymbol{\theta}), \tag{18.81}$$

in which the unknown covariance matrix $\boldsymbol{\Sigma}$ is replaced by the identity matrix. The estimator obtained by minimizing (18.81) will evidently be a valid GMM estimator and thus will be consistent even though it is inefficient. Whichever inefficient estimator is used initially, it will yield g vectors of residuals $\acute{\boldsymbol{u}}_i$ from which the matrix $\boldsymbol{\Sigma}$ may be estimated consistently in exactly the same way as for linear models; see (18.62). Replacing the unknown σ^{ij}'s in (18.80) by the elements $\acute{\sigma}^{ij}$ of the inverse of the estimate of $\boldsymbol{\Sigma}$ then yields the criterion function

$$\sum_{i=1}^{g}\sum_{j=1}^{g}\acute{\sigma}^{ij}\boldsymbol{f}_i^{\top}(\boldsymbol{Y},\boldsymbol{X},\boldsymbol{\theta})\boldsymbol{P}_W\boldsymbol{f}_j(\boldsymbol{Y},\boldsymbol{X},\boldsymbol{\theta}), \tag{18.82}$$

which can actually be minimized in practice.

As usual, the minimized value of the criterion function (18.82) provides a test statistic for overidentifying restrictions; see Sections 7.8 and 17.6. If the model and instruments are correctly specified, this test statistic will be asymptotically distributed as $\chi^2(m-p)$; recall that m is the number of instruments and p is the number of free parameters. Moreover, if the model is estimated unrestrictedly and subject to r distinct restrictions, the difference between the two values of the criterion function will be asymptotically distributed as $\chi^2(r)$. If the latter test statistic is to be employed, it is important that the same estimate of $\boldsymbol{\Sigma}$ be used for both estimations, since otherwise the test statistic may not even be positive in finite samples.

When the sample size is large, it may be less computationally demanding to obtain one-step efficient estimates rather than actually to minimize (18.82). Suppose the initial consistent estimates, which may be either NL2SLS estimates or systems estimates based on (18.81), are denoted $\acute{\boldsymbol{\theta}}$. Then a first-order Taylor-series approximation to $\boldsymbol{f}_i(\boldsymbol{\theta}) \equiv \boldsymbol{f}_i(\boldsymbol{Y},\boldsymbol{X},\boldsymbol{\theta})$ around $\acute{\boldsymbol{\theta}}$ is

$$\boldsymbol{f}_i(\acute{\boldsymbol{\theta}}) + \boldsymbol{F}_i(\acute{\boldsymbol{\theta}})(\boldsymbol{\theta}-\acute{\boldsymbol{\theta}}),$$

where $\boldsymbol{F}_i$ is an $n\times p$ matrix of the derivatives of $\boldsymbol{f}_i(\boldsymbol{\theta})$ with respect to the p elements of $\boldsymbol{\theta}$. If certain parameters do not appear in the i^{th} equation, the corresponding columns of $\boldsymbol{F}_i$ will be identically zero. The one-step estimates, which will be asymptotically equivalent to NL3SLS estimates, are simply $\grave{\boldsymbol{\theta}} = \acute{\boldsymbol{\theta}} - \acute{\boldsymbol{t}}$, where $\acute{\boldsymbol{t}}$ denotes the vector of *linear* 3SLS estimates

$$\acute{\boldsymbol{t}} = \left[\sum_{i=1}^{g}\sum_{j=1}^{g}\acute{\sigma}^{ij}\acute{\boldsymbol{F}}_i^{\top}\boldsymbol{P}_W\acute{\boldsymbol{F}}_j\right]^{-1}\left[\sum_{i=1}^{g}\sum_{j=1}^{g}\acute{\sigma}^{ij}\acute{\boldsymbol{F}}_i^{\top}\boldsymbol{P}_W\acute{\boldsymbol{f}}_j\right]. \tag{18.83}$$

Compare expression (18.64), for the case with no cross-equation restrictions.

It is clear that NL3SLS can be generalized to handle heteroskedasticity of unknown form, serial correlation of unknown form, or both. For example, to handle heteroskedasticity one would simply replace the matrix $\boldsymbol{P}_W$ in (18.82) and (18.83) by the matrix

$$\boldsymbol{W}\big(\boldsymbol{W}^{\top}\acute{\boldsymbol{\Omega}}_{ij}\boldsymbol{W}\big)^{-1}\boldsymbol{W}^{\top},$$

where, by analogy with (18.76), $\acute{\boldsymbol{\Omega}}_{ij} = \operatorname{diag}(\acute{u}_{ti}\acute{u}_{tj})$ for $i,j=1,\ldots,g$. The initial estimates $\acute{\boldsymbol{\theta}}$ need not take account of heteroskedasticity. For a more detailed discussion of this sort of procedure, and of NL3SLS in general, see Gallant (1987, Chapter 6).