$_f ocused_s election.tex Francis Di Tragli a f ditraglia 15 hours ago that's all for tonight!$

Lecture 5: "Focused" Model Selection

Francis J. DiTraglia

March 27, 2014

1 Local Mis-specification

1.1 Introduction

In this lecture we'll be using a kind of asymptotic thought experiment that may be unfamiliar to you, so I'd like to spend a bit of time motivating it before proceeding. Roughly speaking, the idea is to consider a parameter whose value *changes with sample size*. This basic idea is widely used in econometrics and statistics and is known by several different names. Among them are "local alternatives," "Pitman Drift," and "local mis-specification." Although it may seem strange at first, "drifting parameters" are actually the natural asymptotic setting for certain problems, as I hope to convince you with the following two simple examples.

1.2 What's Wrong with Asymptotic Power?

Consider the following simple testing problem. Suppose we observe N observations from the following DGP

$$X_1, X_2, \ldots, X_N \overset{iid}{\sim} \mathcal{N}(\mu, 1)$$

and want to test $H_0\colon \mu = 0$ against the one-sided alternative $H_1\colon \mu > 0$. In this admittedly very simple example, the obvious test statistic is

$$T_N = \sqrt{N}\bar{X}_N \sim N\left(\mu\sqrt{N}, 1\right)$$

where $\bar{X}_N$ is the sample mean. We reject when $\sqrt{N}\bar{X}_N > z_{1-\alpha}$ where $z_{1-\alpha}$ is the $1-\alpha$ quantile of a standard normal distribution. We can calculate the power of this test as follows:

$$\begin{aligned}\text{Power}(T_N) &= P\left(\sqrt{N}\bar{X}_N > z_{1-\alpha}\right) = P\left(Z + \mu\sqrt{N} > z_{1-\alpha}\right) \\ &= P\left(Z > z_{1-\alpha} - \mu\sqrt{N}\right) = 1 - \Phi\left(z_{1-\alpha} - \mu\sqrt{N}\right)\end{aligned}$$

where Z is a standard normal random variable and Φ is the corresponding CDF. Now suppose we decided to do something completely crazy: throw away half our sample. Let $\bar{X}_{N/2}$ denote the sample mean based on observations $1, 2, \ldots, \lfloor N/2 \rfloor$ *only*. We can still construct a perfectly valid test with size α as follows. Define

$$T_{N/2} = \sqrt{\lfloor N/2 \rfloor}\bar{X}_{N/2} \sim N\left(\mu\sqrt{\lfloor N/2 \rfloor}, 1\right)$$

and reject if $\sqrt{N}\bar{X}_N > z_{1-\alpha}$. But there's an obvious problem here: there *must* be a cost for throwing away perfectly good data. Indeed, if we calculate the power for this crazy test, we'll find that it's *strictly lower* than that of the sensible test based on the full sample. In particular,

$$\text{Power}(T_{N/2}) = 1 - \Phi\left(z_{1-\alpha} - \mu\sqrt{\lfloor N/2 \rfloor}\right)$$

using the same argument as above with $\lfloor N/2 \rfloor$ in place of N.

Now, for an example this simple we'd never resort to asymptotics, but suppose we did. How do these two tests compare as the sample size goes to infinity? The asymptotic size in this example is the same as the finite-sample

size since we know the exact sampling distribution of the test statistics under the null and neither depends on sample size. But what about the power? We have,

$$\begin{aligned}
\lim_{N\to\infty} \text{Power}(T_N) &= \lim_{N\to\infty} \left[1 - \Phi\left(z_{1-\alpha} - \mu\sqrt{N}\right)\right] = 1 \\
\lim_{N\to\infty} \text{Power}(T_{N/2}) &= \lim_{N\to\infty} \left[1 - \Phi\left(z_{1-\alpha} - \mu\sqrt{\lfloor N/2 \rfloor}\right)\right] = 1
\end{aligned}$$

In other words, both of these tests are *consistent*: as the sample size goes to infinity, the power goes to one. Think about this for a moment: we know that for *any* fixed sample size a test based on the full sample is *strictly more powerful* but in the limit this difference disappears. This strongly suggests that something is wrong with our asymptotic thought experiment in this setting.

You might object that I've cooked up a particularly perverse example, but it turns out that this phenomenon is quite general. It's easy to find consistent tests, in fact it's difficult to find tests that *aren't* consistent. But we know from simulation studies that not all consistent tests are created equal: some have *much* better finite sample power than others. One way around this problem would be to only compare the finite-sample properties of different tests and never use asymptotics. But we almost *never* know the exact sampling distribution of our test statistics.

This is where *local alternatives* come in. Rather than evaluating our tests against a *fixed* alternative μ, suppose we were to evaulate it against a *sequence* of *local* alternatives that *drift towards the null* at rate $N^{-1/2}$. In other words, our alternative becomes $H_1 \colon \mu = \delta/\sqrt{N}$ where, for this one-sided test, $\delta > 0$. If we substitute $\delta/\sqrt{N}$ for μ and take the limit as $N \to \infty$, we find

$$\begin{aligned}
\lim_{N\to\infty} \text{Power}(T_N) &= \lim_{N\to\infty} \left[1 - \Phi\left(z_{1-\alpha} - \frac{\delta}{\sqrt{N}}\sqrt{N}\right)\right] \\
&= 1 - \Phi\left(z_{1-\alpha} - \delta\right)
\end{aligned}$$

and similarly

$$\begin{aligned}\lim_{N\to\infty} \text{Power}(T_{N/2}) &= \lim_{N\to\infty}\left[1-\Phi\left(z_{1-\alpha}-\frac{\delta}{\sqrt{N}}\sqrt{\lfloor N/2\rfloor}\right)\right] \\ &= 1-\Phi\left(z_{1-\alpha}-\frac{\delta}{\sqrt{2}}\right)\end{aligned}$$

Wow! Our problem has disappeared! The asymptotic power of the two tests now differs in essentially the same way as the finite sample power. Also note that the power no longer converges to one. Intuitively, this is because the drifting sequence of alternatives $\delta/\sqrt{n}$ makes it "harder and harder" to reject the null as the sample size grows by shrinking *just fast enough* but not so fast that the power goes to zero. This type of calculation is called a *local power analysis*. A test that has asymptotic power greater than zero in such a setting is said to have "power against local alternatives."

1.3 Weak Identification

Drifting parameter sequences of the kind described above are also used in the weak instruments and weak identification literature.

Possibly add a simple example later.

1.4 A Bias-Variance Tradeoff in the Limit

When we derived Mallow's C_p, the idea was to compare models on the basis of predictive mean-squared error. Bigger models generally have a lower bias but a higher variance because there are more parameters to estimate. In the example we considered in class, everything was linear and we made enough assumptions about the finite sample distribution that we could deduce the *exact* MSE conditional on X. In many settings, however, finite sample results unavailable and we are forced to rely on asymptotic approximations. We know there is a tradeoff between bias and variance in the finite sample and we'd like to capture this idea in our limit results. The question is how?

Suppose that $\widehat{\mu}$ is a *potentially biased* estimator of μ. Then we have

$$MSE(\widehat{\mu}) = E[(\widehat{\mu} - \mu)^2] = (E[\widehat{\mu} - \mu])^2 + Var(\widehat{\mu})$$

Now, if we don't know the finite sample distribution of $\widehat{\mu}$, we can't calculate the proceeding expression. So what can we do instead? If $\widehat{\mu}$ is asymptotically normal, then we might try to use the features of its limit distribution to calculate the *asymptotic* mean-squared error and use this to as a "stand-in" for the exact, finite-sample quantity. Let μ_0 be the probability limit of $\widehat{\mu}$ and μ be the "true" parameter value. Suppose that

$$\sqrt{T}\left(\widehat{\mu} - \mu_0\right) \xrightarrow{d} \mathcal{N}(0, \sigma^2)$$

In maximum likelihood estimation, μ_0 would be the pseudo-true value that minimizes the KL divergence and σ^2 would be a diagonal element of $J^{-1}KJ^{-1}$. Now, an obvious idea is estimate $Var(\widehat{\mu})$ using the *asymptotic variance*, namely $\text{AVAR}(\widehat{\mu}) = \sigma^2$. But what about the bias term $E[\widehat{\mu} - \mu]$? The limit distribution of $\widehat{\mu}$ is centered around μ_0, the pseudo-true value, but we need to evaluate the bias relative to μ. Let's try recentering by adding and subtracting $\sqrt{T}\mu$ as follows:

$$\begin{aligned}
\sqrt{T}\left(\widehat{\mu} - \mu_0\right) &= \sqrt{T}\widehat{\mu} - \sqrt{T}\mu_0 \\
&= \sqrt{T}\widehat{\mu} - \sqrt{T}\mu_0 - \sqrt{T}\mu + \sqrt{T}\mu \\
&= \sqrt{T}\left(\widehat{\mu} - \mu\right) + \sqrt{T}\left(\mu - \mu_0\right)
\end{aligned}$$

Rearranging, we can write

$$\sqrt{T}\left(\widehat{\mu} - \mu_0\right) = \sqrt{T}\left(\widehat{\mu} - \mu_0\right) - \sqrt{T}\left(\mu - \mu_0\right)$$

Now we have an expression for $\widehat{\mu}$ centered around μ, so the obvious thing to do is look at the mean of the limiting distribution of $\sqrt{T}\left(\widehat{\mu} - \mu\right)$ and call this the "asymptotic bias." Unfortunately, we have a problem. By assumption, the

first term $\sqrt{T}(\widehat{\mu} - \mu_0)$ is $O_p(1)$ but the second term *diverges*! We recentered $\widehat{\mu}$ around μ *precisely because* we thought that μ_0 was potentially different from μ. But if this is the case, then $\sqrt{T}(\mu - \mu_0) = O(T^{1/2})$. So what's going on here? The problem is that the asymptotic variance is of a *different order* than the asymptotic bias. We need to scale $\widehat{\mu}$ up by $\sqrt{T}$ to get a result that has non-zero asymptotic variance, but this same scaling causes the bias to explode. In other words, there is no way to get a meaningful bias-variance tradeoff in the limit under conventional asymptotics.

So how can we fix this problem? Above we had $\sqrt{T}(\mu - \mu_0) = O(T^{1/2})$ but what we want is $\sqrt{T}(\mu - \mu_0) = O(1)$, so somehow or other we need to ensure that $(\mu - \mu_0) = O(T^{-1/2})$. This is where local mis-specification makes its grand appearance. Suppose that we have a DGP under which the true parameter value is $\mu_T = \mu_0 + \delta/\sqrt{T}$ where δ is a constant. That is, suppose we assume that the true parameter value *changes with sample size* and drifts towards μ_0 at rate $T^{-1/2}$. This may sound like a crazy idea, but there's no arguing with the fact that it solves our problem. We have,

$$\begin{aligned}
\sqrt{T}(\widehat{\mu} - \mu_T) &= \sqrt{T}(\widehat{\mu} - \mu_0) - \sqrt{T}(\mu_T - \mu_0) \\
&= \sqrt{T}(\widehat{\mu} - \mu_0) - \sqrt{T}\left(\mu_0 + \delta/\sqrt{T} - \mu_0\right) \\
&= \sqrt{T}(\widehat{\mu} - \mu_0) - \delta \\
&\xrightarrow{d} \mathcal{N}(0, \sigma^2) - \delta
\end{aligned}$$

hence, the asymptotic mean-squared error of $\widehat{\mu}$ is $\text{AMSE}(\widehat{\mu}) = \delta^2 + \sigma^2$. But what does it mean to have a parameter that changes with sample size? It's important to be clear that this does *not* mean that we think real-world datasets follow a DGP that changes with sample size. This is a *thought experiment*: we also don't believe that it's possible to have an infinite sample size! When we use asymptotics, the point is to derive tractable expressions that approximate the effects that actually occur in finite samples. We know that there is a bias-variance tradeoff in finite samples but we showed above that the conventional

asymptotics can't capture this. In other words, local mis-specification is a *device* to get a limiting theory that provides a better approximation to what's really going on in finite samples. For more on the sense in which local mis-specification provides a much more realistic portrait of the effects of model selection, see Leeb and Pötscher (2005).

1.5 Triangular Array Asymptotics

When parameter values change with sample size, we no longer have iid random variables. Instead we have what is called a "triangular array DGP" and we need to index random variables *by sample size* in addition to the usual index:

$$\begin{aligned} &Y_{11} \\ &Y_{21}, Y_{22} \\ &\vdots \\ &Y_{n1}, Y_{n2}, \ldots, Y_{nn} \end{aligned}$$

When we want to avoid the double subscript on the random variables, it's common to add one to the expectation and variance operators so indicate the distribution with respect to which the given moment is being evaulated.

To give you a sense of how triangular array DGPs work, I'll show you some very simple results. For much more general, and also much more technical, results for triangular array DGPs, see Andrews (1988) and Andrews (1992).

A Very Simple LLN for Triangular Arrays Suppose $Y_1, \ldots, Y_n \sim$ iid with mean $\mu + \delta/\sqrt{n}$ and variance σ_n^2. Can we still establish a LLN for the sample mean $\bar{Y}_n = n^{-1} \sum_{i=1}^n Y_i$? If so how? By Chebyshev's Inequality, we know that one simple way to establish a WLLN is via an L_2 argument. In this case, it is sufficient to show that $E_n[\bar{Y}_n] \to \mu$ and $Var_n[\bar{Y}_n] \to 0$. Although the triangular array of RVs in this example is not identically distributed in the

strict sense, it *is* identically distributed *for fixed* n. Thus, we have,

$$E_n[\bar{Y}_n] = \frac{1}{n}\sum_{i=1}^{n} E_n[Y_i] = \mu + \delta/\sqrt{n} \to \mu$$

Using independence, we have

$$Var_n(\bar{Y}_n) = \frac{1}{n^2}\sum_{i=1}^{n} Var_n(Y_i) = \frac{\sigma_n^2}{n}$$

Thus, as long as σ_n^2 is *uniformly bounded* by some constant M, we have $Var_n(\bar{Y}_n) \to 0$ and it follows that $\bar{Y}_n \xrightarrow{p} \mu$. Although this example as so simple as to be nearly trivial it illustrates the basic flavor of triangular array asymptotics: they're very similar to the usual asymptotics you see in first year, but typically require some kind of uniform bound on the array.

Lindeberg-Feller CLT The previous example showed a simple LLN for triangular arrays. What about a CLT? The simplest case assumes independent data and is called the Lindeberg-Feller CLT. For each n, let $Y_{n,1}, Y_{n,2}, \ldots, Y_{n,k_n}$ be independent random vectors with finite variances such that

$$\sum_{i=1}^{k_n} E\left[\|Y_{n,i}\|^2 \mathbf{1}\left\{\|Y_{n,i}\| > \epsilon\right\}\right] \to 0$$

for every $\epsilon > 0$ and

$$\sum_{i=1}^{k_n} Var(Y_{n,i}) \to \Sigma$$

Then $\sum_{i=1}^{k_n}\left(Y_{n,i} - E\left[Y_{n,i}\right]\right) \xrightarrow{d} \mathrm{N}(0, \Sigma)$.

2 Focused Evaluation

The idea behind focused model selection is to choose the model that is best for a *particular purpose* rather than seeking "one-size-fits all" best model.

In general, "best" means minimum risk relative to some loss function: it is *not* a matter of searching for the "true." There are two main ideas here. First, even if we knew what the true DGP model was, up to some unknown parameters that we need to estimate, it's not clear that we should use it. In most interesting settings there is a bias-variance trade-off. If the true model is somewhat complicated, we may be better off fitting a simpler model. Although this introduces a bias, it could lead to a large reduction in variance, depending on sample size. Second, different modeling goals may call for different models *of the same data*. Estimating a structural parameter and creating a forecast are two very different goals. It is far from obvious that we should use the same model for both.

The following example comes from Hansen (2005). Consider an AR(k) model

$$y_t = \mu + \beta_1 y_{t-1} + \cdots + \beta_k y_{t-k} + \epsilon_t$$

where $\{\epsilon_t\}$ is a martingale difference sequence, that is $E[\epsilon_t | I_{t-1}] = 0$. We're interested in learning about a scalar "focus parameter" $\theta = g(\beta)$. This could be for example, one of the individual coefficients β_j, the long-run variance, or an impulse response at some specified horizon. The point is that it's a scalar and a *function* of the underlying model parameters $\beta_1, \ldots, \beta_k$. So what constitutes a "good" model for learning about θ? The natural way to proceed is to specify a loss function and try to find the estimator $\widehat{\theta}$ that minimizes the expectation of the loss. For this example we'll use mean-squared error and search for a model that minimizes $E[(\widehat{\theta} - \theta)^2]$

Hansen (2005) uses a simple simulation experiment to show that different focus parameters can lead to *very different* selected models. The setup is as follows. We consider the family of AR(k) models for $k = 0, 1, \ldots, k_{\max}$ but the true DGP is in fact an ARMA(1,1) model, namely

$$\begin{aligned} y_t &= \alpha y_{t-1} + \epsilon_t - \gamma \epsilon_{t-1} \\ \epsilon_t &\sim \text{iid } N(0,1) \end{aligned}$$

Thus *none* of the models under consideration is correctly specified since the true DGP can be expressed as an AR(∞) model. Now suppose we're interested in the impulse responses. A little algebra reveals that the true impulse responses for the DGP are

$$\theta_m = (\alpha - \gamma)\alpha^{m-1}$$

where m denotes the horizon. The estimated impulse responses for the class of models we are considering can be calculated recursively from the estimated AR parameters. By simulating the DGP with $T = 200$ for a range of parameter values (α, γ) Hansen (2005) shows that the optimal AR order for approximating the impulse response of the true DGP in a minumum mean-squared error sense is *highly* sensitive to m, the horizon of interest. To take a particularly stark example, when $\alpha = 0.5$ and $\beta = 0.9$ the optimal AR order for $m = 2$ is $k = 10$ but the optimal AR order for $m = 6$ is $k = 0$.

3 The Focused Information Criterion (FIC)

The motivation behind the FIC is to create a model selection criterion that is portable like AIC and BIC, based on risk minimization like FPE and C_p, but *focused* in the sense of Hansen (2005). The result turns out to be even *more* portable than AIC and BIC: although originally derived in a likelihood framework, the idea behind the FIC can be easily extended to any situation in which it is possible to derive a limiting distribution. Indeed extending the idea behing the FIC idea to novel settings has been a topic of my recent research!

Although it has been extended in a number of ways, here I'll follow the notation and framework of the original two papers: Claeskens & Hjort (2003) and Hjort & Claeskens (2003). These papers appear in the same issue of JASA and the derivations and explanations are split between them. One can look at various loss functions, but the original papers use MSE so that's what we'll discuss here.

Roughly speaking, the idea behind the FIC is to estimate a user-specified target parameter μ with minimum mean-square error. Since finite-sample MSE can only be calculated in very simple examples, the FIC uses an asymptotic MSE to approximate finite-sample behavior. As discussed above, this requires an asymptotic framework based on drifting sequences of parameters.

Local Mis-specification Framework: Suppose $Y_1, \ldots, Y_n$ are independent with density

$$f_{true}(y) = f(y, \theta_0, \gamma_0 + \delta/\sqrt{n})$$

This could be a regression model, in which case the likelihood is conditional on x but we'll suppress this in the notation. The p-vector θ contains the "protected parameters." These are the parameters that we have decided in advance we definitely want to estimate. In contrast, the q-vector γ contains the parameters over which we will carry out model selection: we consider the restriction $\gamma = \gamma_0$ where γ_0 is a *known* parameter. When we restrict a component of γ we *do not estimate it*: we simply substitute the restriction into the likelihood. In a linear regression problem, for example, we might have something like

$$y_i = x_i'\theta + z_i'\gamma + \epsilon_i$$

and consider setting some or all of the elements of γ equal to zero rather than estimating them. The true value of γ is *changing with sample size* according to $\gamma_n = \gamma_0 + \delta/\sqrt{n}$ where δ is a fixed but unknown constant q-vector. Thus, any specification that does not estimate γ is *locally mis-specified* but the mis-specification disappears in the limit as $n \to \infty$.

N.B. There's something slightly awkward in the notation here: θ_0 is the true value of θ but γ_0 is *not* the true value of γ. It is only *in the limit* that $\gamma = \gamma_0$. Unlike θ_0, which is unknown, γ_0 is *known* since it's the restriction we're considering. This is something the econometrician chooses based on the specifics of the problem at hand.

The Focus Parameter: The FIC is not a specific model selection criterion. Instead it is a *procedure* that allows the *user* to create her own model selection criterion for a particular problem. Let $\mu = \mu(\theta, \gamma)$ be the user-specified parameter of interest. Under local mis-specification, the true value of μ is changing with sample size according to

$$\mu_{\text{true}} = \mu\left(\theta_0, \gamma_0 + \delta/\sqrt{n}\right)$$

The goal is to estimate μ with minimum mean-squared error. But since we are considering general ML models, it's not possible to work out the exact finite-sample distributions of the various estimators. Instead, we calculate the *asymptotic mean-squared error* (AMSE) of our estimators of μ and attempt to select a model to minimize this quantity. The key innovation here is that we are *not* interested in γ for its own sake: all that matters is how our modeling decisions about γ affect our estimates of μ.

Candidate Models: Considered in full generality, we could restrict any number of components of γ. Since this parameter is q-dimensional, we could consider a total of 2^q candidate models if desired. Alternatively, we could decide to consider only particular groups of restrictions. The simplest case considers only two models: the *wide* model estimates *all* elements of γ and the *narrow* model estimates *none* of the elements of gamma. However we choose to restrict the set of candidates, each model is indexed by S which is a subset of $\{1, \ldots, q\}$ that indicates which elements of γ we estimate. Its complement, S^c, indicates which elements of γ we set equal to the corresponding elements of γ_0. Each candidate model S implies a maximum likelihood estimator for the underlying model parameters θ and γ_S, where γ_S denotes the elements of γ that are esetimated under model S. The corresponding ML estimator $\widehat{\mu}_S = \mu\left(\widehat{\mu}_S, \widehat{\gamma}_S\right)$ for the target parameter μ

$$\widehat{\mu}_S = \mu\left(\widehat{\theta}_S, \widehat{\gamma}_S, \gamma_{0,S^c}\right)$$

where γ_{0,S^c} denotes a vector containing the elements of γ_0 whose indices are in S^c. These are the elements of γ that are *not estimated*.

The "Full" Model The *full*, aka *wide*, model is the specification in which we estimate all elements of γ. Under the local mis-specificiation assumption, this model is *correctly specified*. Any model selection criterion relies on some form of over-identification to evaluate the quality of a candidate model relative to alternatives. In the FIC framework this is achieved by comparing the results of each candidate S to those of the full model. We denote the **score function** of the full model by

$$\begin{bmatrix} U(y) \\ V(y) \end{bmatrix} = \begin{bmatrix} \nabla_\theta \log f(y, \theta_0, \gamma_0) \\ \nabla_\gamma \log f(y, \theta_0, \gamma_0) \end{bmatrix} \begin{matrix} (p \times 1) \\ (q \times 1) \end{matrix}$$

Note that the score is evaluated at the *null point* (θ_0, γ_0). This is *not* the true parameter vector for *any finite sample size*, but it is the true parameter vector in the limit. Similarly, the **information matrix** of the full model by

$$J_{Full} = Var_0 \begin{bmatrix} U(y) \\ V(y) \end{bmatrix} = \begin{bmatrix} J_{00} & J_{01} \\ J_{10} & J_{11} \end{bmatrix} \begin{matrix} (p \times p) & (p \times q) \\ (q \times p) & (q \times q) \end{matrix}$$

where the the zero subscript indicates that the expectation is being taken with respect to the distribution in which $\gamma = \gamma_0$. This is the *limiting* DGP which is *different* from the DGP for any finite sample size under local mis-specification. We partition the inverset of the information matrix for the full model as follows

$$J_{Full}^{-1} = \begin{bmatrix} J^{00} & J^{01} \\ J^{10} & J^{11} \end{bmatrix}$$

where

$$K \equiv J^{11} = (J_{11} - J_{10} J_{00}^{-1} J_{01})^{-1}$$

by the partitioned matrix inverse formula. The quantity J^{11} appears so frequently in the derivation of the FIC that it is called K to keep the superscripts from getting out of control.

Selection Matrices In various matrix manipulations in the paper, it turns out to be helpful to define a matrix that *selects* the elements of γ that are estimated under model S. Let π_S be the $|S| \times q$ matrix that "selects" only those elements of a q-vector that correspond to the indices in the set S. For example, suppose $q = 3$ and $S = \{1, 3\}$. Then,

$$\pi_S = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

In this case $\gamma = (\gamma_1, \gamma_2, \gamma_3)'$ and $\pi_S\gamma = (\gamma_1, \gamma_3)'$. For the *wide* or *full* model, i.e. the model that estimates all components of γ, we have $S = \{1, \ldots, q\}$ and hence π_S is simply the identity matrix of order q. An extremely useful fact about π_S is that we can use it to transform the information matrix for the *full* aka *wide* model – the model that estimates all components of γ – into the information matrix for a candidate model S as follows:

$$J_S = Var_0 \begin{bmatrix} U(y) \\ V_S(y) \end{bmatrix} = \begin{bmatrix} J_{00} & J_{01,S} \\ J_{10,S} & J_{11,S} \end{bmatrix} = \begin{bmatrix} J_{00} & J_{01}\pi_S' \\ \pi_S J_{10} & \pi_S J_{11}\pi_S' \end{bmatrix}$$

By the partitioned matrix inverse formula:

$$\begin{aligned} K_S \equiv J^{11,S} &= \left(\pi_S K^{-1}\pi_S'\right)^{-1} = \left[\pi_S\left(J_{11} - J_{10}J_{00}^{-1}J_{01}\right)\pi_S'\right]^{-1} \\ J^{01,S} &= -J_{00}^{-1}J_{01}\pi_S' K_S \\ J^{00,S} &= J_{00}^{-1} + J_{00}^{-1}J_{01}\left(\pi_S' K_S \pi_S\right) J_{10}J_{00}^{-1} \end{aligned}$$

Again, the quantity $J^{11,S}$ appears so many times in the derivation of the FIC that it is called K_S for short.

CLT for the Score of the Full Model The first step in deriving the FIC is to calculate the limiting distribution of the score for the full model evaluated at (θ_0, γ_0). This appears as Lemma 3.1 in Hjort & Claeskens (2003). Before stating it, we'll define the following notation:

$$\begin{bmatrix} \bar{U}_n \\ \bar{V}_n \end{bmatrix} = \frac{1}{n} \sum_{i=1}^{n} \begin{bmatrix} U(Y_i) \\ V(Y_i) \end{bmatrix}$$

Lemma 1 (CLT for Score of Full Model)**.** *Under local mis-specification,*

$$\begin{bmatrix} \sqrt{n}\bar{U}_n \\ \sqrt{n}\bar{V}_n \end{bmatrix} \xrightarrow{d} \begin{pmatrix} J_{01}\delta \\ J_{11}\delta \end{pmatrix} + \begin{pmatrix} M \\ N \end{pmatrix}$$

where

$$\begin{pmatrix} M \\ N \end{pmatrix} \sim N_{p+q}(0, J_{Full})$$

Proof. To prove this result, we apply the Lindeberg-Feller CLT to the triangular array of random variables

$$\begin{bmatrix} U(Y_i)/\sqrt{n} \\ V(Y_i)/\sqrt{n} \end{bmatrix}$$

Since the Y_i are iid *for fixed* n, we have

$$Var \sum_{i=1}^{n} \begin{bmatrix} U(Y_i)/\sqrt{n} \\ V(Y_i)/\sqrt{n} \end{bmatrix} = Var_n \begin{bmatrix} U(Y_i) \\ V(Y_i) \end{bmatrix} \to Var_0 \begin{bmatrix} U(Y_i) \\ V(Y_i) \end{bmatrix} = J_{full}$$

under appropriate regularity conditions. Thus, assuming the Lindeberg condition is satisfied, we have

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left(\begin{bmatrix} U(Y_i) \\ V(Y_i) \end{bmatrix} - E_n \begin{bmatrix} U(Y_i) \\ V(Y_i) \end{bmatrix} \right) \xrightarrow{d} \begin{bmatrix} M \\ N \end{bmatrix}$$

where $(M', N')' \sim N_{p+q}(0, J_{full})$. Again, since the Y_i are iid for fixed n,

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left(\begin{bmatrix} U(Y_i) \\ V(Y_i) \end{bmatrix} - E_n \begin{bmatrix} U(Y_i) \\ V(Y_i) \end{bmatrix} \right) = \left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \begin{bmatrix} U(Y_i) \\ V(Y_i) \end{bmatrix} \right) - \sqrt{n} E_n \begin{bmatrix} U(Y_i) \\ V(Y_i) \end{bmatrix}$$

And by a mean-value expansion around $\gamma_n = \gamma_0 + \delta/\sqrt{n}$,

$$E_n \begin{bmatrix} U(Y_i) \\ V(Y_i) \end{bmatrix} = E_n \begin{bmatrix} \nabla_\theta \log f(Y_i, \theta_0, \gamma_n) \\ \nabla_\gamma \log f(Y_i, \theta_0, \gamma_n) \end{bmatrix} + E_n \begin{bmatrix} \nabla_{\theta\gamma'} \log f(Y_i, \theta_0, \gamma^*) \\ \nabla_{\gamma\gamma'} \log f(Y_i, \theta_0, \gamma^*) \end{bmatrix} (\gamma_0 - \gamma_n)$$

where γ^* is between γ_0 and γ_n. The first term is simply the population moment condition for ML estimation and hence equals zero: thanks to the mean-value expansion, the expectation is now evaluated at (θ_0, γ_n) which is the true parameter value for the DGP based on a sample size of n. Thus, since $\gamma_0 - \gamma_n = -\delta/\sqrt{n}$, we have

$$\sqrt{n} E_n \begin{bmatrix} U(Y_i) \\ V(Y_i) \end{bmatrix} = -E_n \begin{bmatrix} \nabla_{\theta\gamma'} \log f(Y_i, \theta_0, \gamma^*) \\ \nabla_{\gamma\gamma'} \log f(Y_i, \theta_0, \gamma^*) \end{bmatrix} \delta \to -E_0 \begin{bmatrix} \nabla_{\theta\gamma'} \log f(Y_i, \theta_0, \gamma_0) \\ \nabla_{\gamma\gamma'} \log f(Y_i, \theta_0, \gamma_0) \end{bmatrix} \delta$$

under appropriate regularity conditions. Recall that, in the limit, (θ_0, γ_0) are the *true* parameter values. Hence,

$$-E_0 \begin{bmatrix} \nabla_{\theta\gamma'} \log f(Y_i, \theta_0, \gamma_0) \\ \nabla_{\gamma\gamma'} \log f(Y_i, \theta_0, \gamma_0) \end{bmatrix} = \begin{bmatrix} J_{01} \\ J_{11} \end{bmatrix}$$

by the information matrix equality, yielding the desired result. □

Asymptotic Normality of the Estimators The next step in the derivation of the FIC is to work out the limiting distribution of the ML estimators $(\widehat{\theta}_S, \widehat{\gamma}_S)$ under model S. This is Lemma 3.2 in Hjort & Claeskens (2003).

Lemma 2. *Under local mis-specification,*

$$\begin{bmatrix} \sqrt{n}(\hat{\theta}_S - \theta_0) \\ \sqrt{n}(\hat{\gamma}_S - \gamma_0) \end{bmatrix} \xrightarrow{d} \begin{bmatrix} C_S \\ D_S \end{bmatrix}$$

where

$$\begin{bmatrix} C_S \\ D_S \end{bmatrix} = J_S^{-1} \begin{pmatrix} J_{01}\delta + M \\ \pi_S J_{11}\delta + N_S \end{pmatrix} \sim N_{p+|S|} \left(J_S^{-1} \begin{bmatrix} J_{01} \\ \pi_S J_{11} \end{bmatrix} \delta, J_S^{-1} \right)$$

and $N_S = \pi_S N$.

Proof. The usual Taylor Expansion argument for ML continues to apply under local mis-specification. Furthermore, the information matrix equality holds in the limit since all the models under consideration are asymptotically correctly specified. Thus, we have

$$\begin{bmatrix} \widehat{\theta}_S \\ \widehat{\gamma}_S \end{bmatrix} = \begin{bmatrix} \theta_0 \\ \gamma_{0,S} \end{bmatrix} + J_S^{-1} \begin{bmatrix} \bar{U}_n \\ \pi_S \bar{V}_n \end{bmatrix} + o_p(n^{-1/2})$$

Restricting Lemma 1 to model S, we have

$$\begin{bmatrix} \sqrt{n}\bar{U}_n \\ \sqrt{n}\pi_S \bar{V}_n \end{bmatrix} \xrightarrow{d} \begin{pmatrix} J_{01}\delta \\ \pi_S J_{11}\delta \end{pmatrix} + \begin{pmatrix} M \\ \pi_S N \end{pmatrix}$$

so the result follows by the Continuous Mapping Theorem. □

Important Point Notice that the only place the mis-specification showed up in the preceding proof was in the CLT for the score. This means that *all* of the models under consideration yield *consistent estimators*.

Some Additional Notation To make the final results a bit more compact, Hjort & Claeskens (2003) introduce some additional notation:

$$W = J^{10}M + J^{11}N$$

The random variable W is simply a linear combination of the random variables M and N that emerged from applying a CLT to the score of the full model. The reason it's worth naming this quantity is because of the following result[1]

Lemma 3. *Define $W \equiv J^{10}M + J^{11}N$. Then, $W = K(N - J_{10}J_{00}^{-1}M)$ and M and W are independent with $W \sim N_q(0, K)$ and $M \sim N_p(0, J_{00})$.*

Proof. By the formula for the inverse of a partitioned matrix,

$$\begin{aligned} J^{11} &= \left(J_{11} - J_{10}J_{00}^{-1}J_{01}\right)^{-1} \\ J^{01} &= -J_{00}^{-1}J_{01}J^{11} \\ J^{10} &= -J^{11}J_{10}J_{00}^{-1} \\ J^{00} &= J_{00}^{-1} + J_{00}^{-1}J_{01}J^{11}J_{10}J_{00}^{-1} \end{aligned}$$

Thus,

$$\begin{aligned} W \equiv J^{10}M + J^{11}N &= \left(-J^{11}J_{10}J_{00}^{-1}\right)M + J^{11}N \\ &= J^{11}\left(N - J_{10}J_{00}^{-1}M\right) \\ &= K\left(N - J_{10}J_{00}^{-1}M\right) \end{aligned}$$

Now we need to show the independence of W and M. Because they're jointly normal, it is sufficient to show that they are uncorrelated. Write

$$\begin{bmatrix} M \\ W \end{bmatrix} = \begin{bmatrix} M \\ J^{10}M + J^{11}N \end{bmatrix} = \begin{bmatrix} I_p & 0_{p\times q} \\ J^{10} & J^{11} \end{bmatrix} \begin{bmatrix} M \\ N \end{bmatrix} \equiv A \begin{bmatrix} M \\ N \end{bmatrix}$$

Since $\begin{bmatrix} M \\ N \end{bmatrix} \sim \mathcal{N}_{p+q}(0, J_{Full})$, we have $A \begin{bmatrix} M \\ N \end{bmatrix} \sim \mathcal{N}_{p+q}(0, AJ_{Full}A')$. Multi-

[1] This doesn't actually appear as a Lemma in the paper: it's one of those "it's not difficult to show" assertions and appears immediately after Lemma 3.2.

plying through, we find that

$$AJ_{Full}A' = \begin{bmatrix} J_{00} & J_{00}J^{01} + J_{01}J^{11} \\ J^{10}J_{00} + J^{11}J_{10} & J^{10}\left(J_{00}J^{01} + J_{01}J^{11}\right) + J^{11}\left(J_{10}J^{01} + J_{11}J^{11}\right) \end{bmatrix}$$

Now,

$$\begin{aligned} J_{00}J^{01} + J_{01}J^{11} &= J_{00}\left(-J_{00}^{-1}J_{01}J^{11}\right) + J_{01}J^{11} \\ &= -J_{01}J^{11} + J_{01}J^{11} = 0 \end{aligned}$$

and similarly

$$\begin{aligned} J^{10}J_{00} + J^{11}J_{10} &= \left(-J^{11}J_{10}J_{00}^{-1}\right)J_{00} + J^{11}J_{10} \\ &= -J^{11}J_{10} + J^{11}J_{10} = 0 \end{aligned}$$

Finally,

$$\begin{aligned} J^{10}\left(J_{00}J^{01} + J_{01}J^{11}\right) + J^{11}\left(J_{10}J^{01} + J_{11}J^{11}\right) &= J^{11}\left(J_{10}J^{01} + J_{11}J^{11}\right) \\ &= J^{11}\left(J_{10}\left[-J_{00}^{-1}J_{01}J^{11}\right] + J_{11}J^{11}\right) \\ &= J^{11}\left(J_{11} - J_{10}J_{00}^{-1}J_{01}\right)J^{11} \\ &= J^{11}\left(J_{11}\right)^{-1}J^{11} = J^{11} \end{aligned}$$

where the first equality uses the fact that $J_{00}J^{01} + J_{01}J^{11} = 0$. □

Estimating δ As we saw in Lemma 2, the limiting distribution of the ML estimators depends on the local mis-specification parameter, δ. Since this is unknown we will, ultimately, need to estimate it. To this end, define

$$\begin{aligned} \widehat{\delta}_S &= \sqrt{n}\left(\widehat{\gamma}_S - \gamma_{0,S}\right) \\ D_S &= K_S\pi_S K^{-1}(\delta + W) \end{aligned}$$

where W is the random variable described in Lemma 3. The key result concerning these quantities is as follows[2]

Lemma 4. *Lemma 3.2 and some algebra imply that*

$$\hat{\delta}_S \equiv \sqrt{n}(\hat{\gamma}_S - \gamma_{0,S}) \xrightarrow{d} D_S$$

where $D_S = K_S \pi_s K^{-1}(\delta + W) = K_S \pi_s K^{-1} D$, *defining* $D = \delta + W$. *In particular:*

$$D_n \equiv \hat{\delta}_{Full} = \sqrt{n}(\hat{\gamma}_{Full} - \gamma_0) \xrightarrow{d} D = (\delta + W) \sim \mathcal{N}_q(\delta, K)$$

Proof. Lemma 3.2 establishes that

$$\begin{bmatrix} \sqrt{n}(\hat{\theta} - \theta_0) \\ \sqrt{n}(\hat{\gamma} - \gamma_0) \end{bmatrix} \xrightarrow{d} \begin{bmatrix} C_S \\ D_S \end{bmatrix} = J_S^{-1} \begin{pmatrix} J_{01}\delta + M \\ \pi_S J_{11}\delta + N_S \end{pmatrix}$$

so we know immediately that $\hat{\delta}_S \equiv \sqrt{n}(\hat{\gamma}_S - \gamma_{0,S}) \xrightarrow{d} D_S$. We need to show that $D_S = K_S \pi_S K^{-1} D$ where $D = \delta + W$. We have:

$$\begin{aligned} \begin{bmatrix} C_S \\ D_S \end{bmatrix} &= J_S^{-1} \begin{bmatrix} J_{01}\delta + M \\ \pi_S J_{11}\delta + N_S \end{bmatrix} = \begin{bmatrix} J^{00,S} & J^{01,S} \\ J^{10,S} & J^{11,S} \end{bmatrix} \begin{bmatrix} J_{01}\delta + M \\ \pi_S J_{11}\delta + N_S \end{bmatrix} \\ &= \begin{bmatrix} J_{00}^{-1} + J_{00}^{-1} J_{01} \left(\pi_S' K_S \pi_S\right) J_{10} J_{00}^{-1} & -J_{00}^{-1} J_{01} \pi_S' K_S \\ -K_S \pi_S J_{10} J_{00}^{-1} & K_S \end{bmatrix} \begin{bmatrix} J_{01}\delta + M \\ \pi_S J_{11}\delta + N_S \end{bmatrix} \end{aligned}$$

[2]This does not appear as a lemma in the paper: "it follows from Lemma 3.2 and a little algebra."

where $K_S = (\pi_S K^{-1} \pi_S')^{-1}$ and $K \equiv J^{11}$. Thus, we have

$$\begin{aligned}
D_S &= -K_S \pi_S J_{10} J_{00}^{-1} (J_{01}\delta + M) + K_S (\pi_S J_{11}\delta + N_S) \\
&= K_S \left[(\pi_S J_{11}\delta + N_S) - \pi_S J_{10} J_{00}^{-1} (J_{01}\delta + M) \right] \\
&= K_S \left[\pi_S J_{11}\delta + \pi_S N - \pi_S J_{10} J_{00}^{-1} (J_{01}\delta + M) \right] \\
&= K_S \pi_S \left[\left(J_{11} - J_{10} J_{00}^{-1} J_{01} \right) \delta + N - J_{10} J_{00}^{-1} M \right] \\
&= K_S \pi_S \left[K^{-1}\delta + K^{-1} K \left(N - J_{10} J_{00}^{-1} M \right) \right] \\
&= K_S \pi_S K^{-1} \left[\delta + K \left(N - J_{10} J_{00}^{-1} M \right) \right] \\
&= K_S \pi_S K^{-1} (\delta + W)
\end{aligned}$$

□

Estimating the Focus Parameter We're finally ready to work out the limiting distribution of $\widehat{\mu}_S$. First two final items of notation. Define

$$\begin{aligned}
H_S &= K^{-1/2} \pi_S' K_S \pi_S K^{-1/2} \\
\omega &= J_{10} J_{00}^{-1} \nabla_\theta \mu(\theta_0, \gamma_0) - \nabla_\gamma \mu(\theta_0, \gamma_0)
\end{aligned}$$

Notice that:

1. ω depends on the choice of focus parameter μ but *not* on the model S.
2. H_S is symmetric and idempotent, thus it is a projection matrix.
3. H_S is orthogonal to $I - H_S$
4. Define $H_\emptyset$ as a $q \times q$ null matrix.

When $S = \emptyset$, i.e. when we consider a submodel that estimates *none* of the componentes of γ, we define $H_\emptyset$ as a $q \times q$ matrix of zeros. The key result, which appears as Lemma 3.3 in the Paper, is as follows

Lemma 5. *If μ has continuous partial derivatives in a neighborhood of (θ_0, γ_0),*

$$\sqrt{n}\,(\hat{\mu}_S - \mu_{true}) \xrightarrow{d} \Lambda_S$$

where $\mu_{true} = \mu(\theta_0, \gamma_0 + \delta/\sqrt{n})$ and

$$\Lambda_S = \nabla_\theta \mu(\theta_0, \gamma_0)' J_{00}^{-1} M + \omega' \left(\delta - K^{1/2} H_S K^{-1/2} D\right)$$

Thus, the the scalar random variable Λ_S follows a normal distribution with

$$\begin{aligned} Mean &= \omega'(I - K^{1/2} H_S K^{-1/2})\delta \\ Variance &= \nabla_\theta \mu(\theta_0, \gamma_0)' J_{00}^{-1} \nabla_\theta \mu(\theta_0, \gamma_0) + \omega' K^{1/2} H_S K^{1/2} \omega \end{aligned}$$

Proof. The first thing to notice is that the limiting result distribution given in the Lemma is centered around $\mu_{\text{true}} = \mu(\theta_0, \gamma_n)$ where $\gamma_n = \gamma_0 + \delta/\sqrt{n}$. It is *not* centered around $\mu_0 = \mu(\theta_0, \gamma_0)$. This means that we cannot immediately apply the Delta Method to Lemma 2 since the limit distributions given there are centered around (θ_0, γ_0). By a mean-value expansion around γ_0,

$$\mu_{\text{true}} = \mu(\theta_0, \gamma_0 + \delta/\sqrt{n}) = \mu(\theta_0, \gamma_0) + \nabla_\gamma \mu(\theta_0, \bar{\gamma})' \frac{\delta}{\sqrt{n}}$$

where $\bar{\gamma}$ is between γ_0 and $\gamma_0 + \delta/\sqrt{n}$. Thus, we have

$$\begin{aligned} \sqrt{n}\,(\hat{\mu}_S - \mu_{true}) &= \sqrt{n}\,(\widehat{\mu}_S - \mu_0) - \sqrt{n}\,(\mu_{\text{true}} - \mu_0) \\ &= \sqrt{n}\,(\widehat{\mu}_S - \mu_0) - \nabla_\gamma \mu(\theta_0, \bar{\gamma})' \delta \end{aligned}$$

Applying the Delta Method to the first term via Lemma 2 and using the fact that $\bar{\gamma} \to \gamma_0$ for the second term, we have $\sqrt{n}\,(\hat{\mu}_S - \mu_{true}) \xrightarrow{d} \Lambda_S$ where

$$\Lambda_S = \nabla_\theta \mu(\theta_0, \gamma_0)' C_S + \left[\pi_S \nabla_\gamma \mu(\theta_0, \gamma_0)\right]' D_S - \nabla_\gamma \mu(\theta_0, \gamma_0)' \delta$$

From here, it is immediate that Λ_S is MV normal, as it is a linear combination

of a normal random vector. Although we *could* find its mean and variance directly using this result, it will be helpful to simplify the expression for Λ_S. The point is that M and $D = \delta + W$ are *independent* normal random vectors, so if we can isolate them, we have a much easier expression to deal with. We established above that:

$$\begin{aligned}
\begin{bmatrix} C_S \\ D_S \end{bmatrix} &= J_S^{-1} \begin{bmatrix} J_{01}\delta + M \\ \pi_S J_{11}\delta + N_S \end{bmatrix} = \begin{bmatrix} J^{00,S} & J^{01,S} \\ J^{10,S} & J^{11,S} \end{bmatrix} \begin{bmatrix} J_{01}\delta + M \\ \pi_S J_{11}\delta + N_S \end{bmatrix} \\
&= \begin{bmatrix} J_{00}^{-1} + J_{00}^{-1} J_{01} \left(\pi_S' K_S \pi_S\right) J_{10} J_{00}^{-1} & -J_{00}^{-1} J_{01} \pi_S' K_S \\ -K_S \pi_S J_{10} J_{00}^{-1} & K_S \end{bmatrix} \begin{bmatrix} J_{01}\delta + M \\ \pi_S J_{11}\delta + N_S \end{bmatrix}
\end{aligned}$$

and, multiplying this out, found $D_S = K_S \pi_S K^{-1}(\delta + W)$. Now we will do the same for C_S. To begin:

$$\begin{aligned}
C_S &= J^{00,S}\left(J_{01}\delta + M\right) + J^{01,S}\left(\pi_S J_{11}\delta + N_S\right) \\
&= \left(J^{00,S} J_{01} + J^{01,S}\pi_S J_{11}\right)\delta + \left(J^{00,S} M + J^{01,S} N_S\right) \\
&\equiv A\delta + B
\end{aligned}$$

Now,

$$\begin{aligned}
A &\equiv J^{00,S} J_{01} + J^{01,S}\pi_S J_{11} \\
&= \left(J_{00}^{-1} + J_{00}^{-1} J_{01}\left[\pi_S' K_S \pi_S\right] J_{10} J_{00}^{-1}\right) J_{01} + \left(-J_{00}^{-1} J_{01}\pi_S' K_S\right)\pi_S J_{11} \\
&= J_{00}^{-1} J_{01}\left(I + \left[\pi_S' K_S \pi_S\right] J_{10} J_{00}^{-1} J_{01} - \left[\pi_S' K_S \pi_S\right] J_{11}\right) \\
&= J_{00}^{-1} J_{01}\left[I - \left(\pi_S' K_S \pi_S\right)\left(J_{11} - J_{10} J_{00}^{-1} J_{01}\right)\right] \\
&= J_{00}^{-1} J_{01}\left[I - \left(\pi_S' K_S \pi_S\right) K^{-1}\right] \\
&= J_{00}^{-1} J_{01}\left[I - K^{1/2} K^{-1/2}\left(\pi_S' K_S \pi_S\right) K^{-1/2} K^{-1/2}\right] \\
&= J_{00}^{-1} J_{01}\left[I - K^{1/2}\left(K^{-1/2}\pi_S' K_S \pi_S K^{-1/2}\right) K^{-1/2}\right] \\
&= J_{00}^{-1} J_{01}\left[I - K^{1/2} H_S K^{-1/2}\right]
\end{aligned}$$

$$
\begin{aligned}
B &\equiv J^{00,S}M + J^{01,S}N_S \\
&= \left(J_{00}^{-1} + J_{00}^{-1}J_{01}\pi_S' K_S \pi_S J_{10} J_{00}^{-1}\right) M + \left(-J_{00}^{-1}J_{01}\pi_S' K_S\right)\pi_S N \\
&= J_{00}^{-1}M + J_{00}^{-1}J_{01}\pi_S' K_S \pi_S \left(J_{10}J_{00}^{-1}M - N\right) \\
&= J_{00}^{-1}M - J_{00}^{-1}J_{01}\pi_S' K_S \pi_S \left(N - J_{10}J_{00}^{-1}M\right) \\
&= J_{00}^{-1}M - J_{00}^{-1}J_{01}\left(K^{1/2}K^{-1/2}\right)\pi_S' K_S \pi_S \left(K^{-1}K\right)\left(N - J_{10}J_{00}^{-1}M\right) \\
&= J_{00}^{-1}M - J_{00}^{-1}J_{01}\left(K^{1/2}K^{-1/2}\right)\pi_S' K_S \pi_S \left(K^{-1}\right)\left[K\left(N - J_{10}J_{00}^{-1}M\right)\right] \\
&= J_{00}^{-1}M - J_{00}^{-1}J_{01}\left(K^{1/2}K^{-1/2}\right)\pi_S' K_S \pi_S \left(K^{-1/2}K^{-1/2}\right)\left[K\left(N - J_{10}J_{00}^{-1}M\right)\right] \\
&= J_{00}^{-1}M - J_{00}^{-1}J_{01}K^{1/2}\left(K^{-1/2}\pi_S' K_S \pi_S K^{-1/2}\right)K^{-1/2}\left[K\left(N - J_{10}J_{00}^{-1}M\right)\right] \\
&= J_{00}^{-1}M - J_{00}^{-1}J_{01}K^{1/2}H_S K^{-1/2}W
\end{aligned}
$$

where we have substituted the definition of H_S and used the fact that, as we showed above, $K(N - J_{10}J_{00}^{-1}M) = W$. Combining these,

$$
\begin{aligned}
C_S &= J_{00}^{-1}J_{01}\left(I - K^{1/2}H_S K^{-1/2}\right)\delta + J_{00}^{-1}M - J_{00}^{-1}J_{01}K^{1/2}H_S K^{-1/2}W \\
&= J_{00}^{-1}J_{01}\delta - \left(J_{00}^{-1}J_{01}K^{1/2}H_S K^{-1/2}\right)\delta + J_{00}^{-1}M - \left(J_{00}^{-1}J_{01}K^{1/2}H_S K^{-1/2}\right)W \\
&= \left(J_{00}^{-1}J_{01}\right)\delta - \left(J_{00}^{-1}J_{01}\right)K^{1/2}H_S K^{-1/2}(\delta + W) + J_{00}^{-1}M \\
&= J_{00}^{-1}M + J_{00}^{-1}J_{01}\left[\delta - K^{1/2}H_S K^{-1/2}(\delta + W)\right] \\
&= J_{00}^{-1}M + J_{00}^{-1}J_{01}\left(\delta - K^{1/2}H_S K^{-1/2}D\right)
\end{aligned}
$$

Thus, expressing everything in terms of the independent normal random vectors M and $D = \delta + W$, we have

$$
\begin{bmatrix} C_S \\ D_S \end{bmatrix} = \begin{bmatrix} J_{00}^{-1}M + J_{00}^{-1}J_{01}\left(\delta - K^{1/2}H_S K^{-1/2}D\right) \\ K_S \pi_S K^{-1} D \end{bmatrix}
$$

Now, recall that

$$
\Lambda_S = \nabla_\theta \mu(\theta_0, \gamma_0)' C_S + \left[\pi_S \nabla_\gamma \mu(\theta_0, \gamma_0)\right]' D_S - \nabla_\gamma \mu(\theta_0, \gamma_0)'\delta
$$

Multiplying through,

$$\nabla_\theta \mu(\theta_0, \gamma_0)' C_S = \nabla_\theta \mu(\theta_0, \gamma_0)' \left[J_{00}^{-1} M + J_{00}^{-1} J_{01} \left(\delta - K^{1/2} H_S K^{-1/2} D \right) \right]$$

and

$$\begin{aligned}
\left[\pi_S \nabla_\gamma \mu(\theta_0, \gamma_0) \right]' D_S &= \nabla_\gamma \mu(\theta_0, \gamma_0)' \pi_S' D_S \\
&= \nabla_\gamma \mu(\theta_0, \gamma_0)' \pi_S' K_S \pi_S K^{-1} D \\
&= \nabla_\gamma \mu(\theta_0, \gamma_0)' \left(K^{1/2} K^{-1/2} \right) \pi_S' K_S \pi_S \left(K^{-1/2} K^{-1/2} \right) D \\
&= \nabla_\gamma \mu(\theta_0, \gamma_0)' K^{1/2} \left(K^{-1/2} \pi_S' K_S \pi_S K^{-1/2} \right) K^{-1/2} D \\
&= \nabla_\gamma \mu(\theta_0, \gamma_0)' K^{1/2} H_S K^{-1/2} D
\end{aligned}$$

Therefore,

$$\begin{aligned}
\Lambda_S &= \nabla_\theta \mu(\theta_0, \gamma_0)' C_S + \left[\pi_S \nabla_\gamma \mu(\theta_0, \gamma_0) \right]' D_S - \nabla_\gamma \mu(\theta_0, \gamma_0)' \delta \\
&= \nabla_\theta \mu(\theta_0, \gamma_0)' \left[J_{00}^{-1} M + J_{00}^{-1} J_{01} \left(\delta - K^{1/2} H_S K^{-1/2} D \right) \right] \\
&\qquad + \left[\nabla_\gamma \mu(\theta_0, \gamma_0)' K^{1/2} H_S K^{-1/2} D \right] - \nabla_\gamma \mu(\theta_0, \gamma_0)' \delta \\
&= \nabla_\theta \mu(\theta_0, \gamma_0)' J_{00}^{-1} M + \nabla_\theta \mu(\theta_0, \gamma_0)' J_{00}^{-1} J_{01} \left(\delta - K^{1/2} H_S K^{-1/2} D \right) \\
&\qquad - \nabla_\gamma \mu(\theta_0, \gamma_0)' \left(\delta - K^{1/2} H_S K^{-1/2} D \right) \\
&= \nabla_\theta \mu(\theta_0, \gamma_0)' J_{00}^{-1} M + \left[\nabla_\theta \mu(\theta_0, \gamma_0)' J_{00}^{-1} J_{01} - \nabla_\gamma \mu(\theta_0, \gamma_0)' \right] \left(\delta - K^{1/2} H_S K^{-1/2} D \right) \\
&= \nabla_\theta \mu(\theta_0, \gamma_0)' J_{00}^{-1} M + \left[J_{10} J_{00}^{-1} \nabla_\theta \mu(\theta_0, \gamma_0) - \nabla_\gamma \mu(\theta_0, \gamma_0) \right]' \left(\delta - K^{1/2} H_S K^{-1/2} D \right) \\
&= \nabla_\theta \mu(\theta_0, \gamma_0)' J_{00}^{-1} M + \omega' \left(\delta - K^{1/2} H_S K^{-1/2} D \right)
\end{aligned}$$

Now we can easily calculate the mean and variance of the scalar random variable Λ_S as we have expressed it as a linear combination of two independent normal random vectors: M and $D = \delta + W$. Recall that

$$\begin{bmatrix} M \\ W \end{bmatrix} \sim \mathrm{N}_{p+q} \left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} J_{00} & 0 \\ 0 & K \end{bmatrix} \right)$$

where $K = J^{11}$. Exploiting the symmetry of variance matrices in several places

as well as the symmetry and idempotency of H_S, we have

$$\begin{aligned}
E[\Lambda_S] &= E\left[\nabla_\theta \mu(\theta_0,\gamma_0)' J_{00}^{-1} M\right] + E\left[\omega'\left(\delta - K^{1/2} H_S K^{-1/2} D\right)\right] \\
&= \nabla_\theta \mu(\theta_0,\gamma_0)' J_{00}^{-1} E\left[M\right] + \omega'\delta - \omega' K^{1/2} H_S K^{-1/2} E[\delta + W] \\
&= \omega'\delta - \omega' K^{1/2} H_S K^{-1/2}\left(\delta + E[W]\right) \\
&= \omega'\delta - \omega' K^{1/2} H_S K^{-1/2}\delta \\
&= \omega'\left(I - K^{1/2} H_S K^{-1/2}\right)\delta
\end{aligned}$$

$$\begin{aligned}
Var\left[\nabla_\theta \mu(\theta_0,\gamma_0)' J_{00}^{-1} M\right] &= \left[\nabla_\theta \mu(\theta_0,\gamma_0)' J_{00}^{-1}\right] Var[M] \left[\nabla_\theta \mu(\theta_0,\gamma_0)' J_{00}^{-1}\right]' \\
&= \nabla_\theta \mu(\theta_0,\gamma_0)' J_{00}^{-1} J_{00} J_{00}^{-1} \nabla_\theta \mu(\theta_0,\gamma_0) \\
&= \nabla_\theta \mu(\theta_0,\gamma_0)' J_{00}^{-1} \nabla_\theta \mu(\theta_0,\gamma_0)
\end{aligned}$$

$$\begin{aligned}
Var\left[\omega'\left(\delta - K^{1/2} H_S K^{-1/2} D\right)\right] &= \left(\omega' K^{1/2} H_S K^{-1/2}\right) Var[D] \left(\omega' K^{1/2} H_S K^{-1/2}\right)' \\
&= \omega' K^{1/2} H_S K^{-1/2} K K^{-1/2} H_S K^{1/2} \omega \\
&= \omega' K^{1/2} H_S \left(K^{-1/2} K^{1/2}\right)\left(K^{1/2} K^{-1/2}\right) H_S K^{1/2} \omega \\
&= \omega' K^{1/2} H_S H_S K^{1/2} \omega \\
&= \omega' K^{1/2} H_S K^{1/2} \omega
\end{aligned}$$

$$\begin{aligned}
Var[\Lambda_S] &= Var\left[\nabla_\theta \mu(\theta_0,\gamma_0)' J_{00}^{-1} M\right] + Var\left[\omega'\left(\delta - K^{1/2} H_S K^{-1/2} D\right)\right] \\
&= \nabla_\theta \mu(\theta_0,\gamma_0)' J_{00}^{-1} \nabla_\theta \mu(\theta_0,\gamma_0) + \omega' K^{1/2} H_S K^{1/2} \omega
\end{aligned}$$

□

Estimating AMSE So far, all we have done, admittedly at great length, is derive the limit distribution of $\widehat{\mu}_S$. Now we're *finally* ready to state our model selection criterion: the FIC. From Lemma 5, the asymptotic mean-squared

error of $\sqrt{n}\,(\widehat{\mu}_S - \mu_{\text{true}})$ is

$$\begin{aligned} r(S) &= \text{Bias}^2 + \text{Variance} \\ &= \left[\omega'(I - K^{1/2}H_S K^{-1/2})\delta\right]\left[\omega'(I - K^{1/2}H_S K^{-1/2})\delta\right]' \\ &\qquad + \left[\nabla_\theta \mu(\theta_0,\gamma_0)' J_{00}^{-1} \nabla_\theta \mu(\theta_0,\gamma_0) + \omega' K^{1/2} H_S K^{1/2}\omega\right] \\ &= \omega'(I - K^{-1/2}H_S K^{1/2})\delta\delta'(I - K^{1/2}H_S K^{-1/2})\omega \\ &\qquad + \omega' K^{1/2} H_S K^{1/2}\omega + \tau_0^2 \end{aligned}$$

Where

$$\tau_0^2 = \nabla_\theta \mu(\theta_0,\gamma_0)' J_{00}^{-1} \nabla_\theta \mu(\theta_0,\gamma_0)$$

which is non-negative and does *not* vary across models. Ideally, we would simply choose S to minimize AMSE(S) but the formula depends on various unknowns. The solution is, of course, to estimate them. Under local mis-specification, consistent estimators of all quantities *except* δ are readily available: they're just the usual ML estimators.[3]

So what can we do about δ? Notice from above that we actually need to estimate $\delta\delta'$, *not* δ. If we had a consistent estimator $\widetilde{\delta}$ of δ, then $\widetilde{\delta}\widetilde{\delta}'$ would be a consistent estimator of $\delta\delta'$. Unfortunately *no consistent estimator of* δ *exists under local mis-specification.* Intuitively, the problem is that the data become "less and less informative" about δ as the sample size grows. Instead, the FIC substitutes an **asymptotically unbiased estimator** of this quantity, constructed as follows. First, we know from Lemma 3 that

$$D_n = \widehat{\delta}_{Full} \xrightarrow{d} D = \delta + W \sim \text{N}_q(\delta, K)$$

Thus, $\widehat{\delta}_{Full}$ is an *asymptotically unbiased estimator of* δ. By the Continuous Mapping Theorem,

$$D_n D_n' \xrightarrow{d} DD'$$

[3]There's a slight issue about whether it makes more sense to use the estimates from the wide model or from a given submodel but this doesn't show up anywhere in the asymptotics. For more discussion on this point, see Claeskens & Hjort (2003).

But, by the shortcut formula

$$E[DD'] = Var(D) + E[D]E[D'] = K + \delta\delta'$$

which means that $D_n D_n'$ is an asymptotically *biased* estimator of $\delta\delta'$. Fortunately, to remove the bias we simply need to subtract K. Thus, our asymptotically unbiased estimator of $\delta\delta'$ is

$$D_n D_n' - \widehat{K}$$

Substituting this quantity along with consistent estimators of everything else provides an **asymptotically unbiased estimator of AMSE**.

The FIC We could really just stop here, but in the paper Claeskens and Hjort express the FIC in a slightly different (and simpler) way by removing constants that do not vary across models. First they construct the *limit experiment* version of the AMSE by substituting $DD' - K$ for $\delta\delta'$. This yields

$$\begin{aligned}
\widehat{r}(S) &= \omega'(I - K^{1/2}H_S K^{-1/2})\,(DD' - K)\,(I - K^{-1/2}H_S K^{1/2})\omega \\
&\qquad +\omega' K^{1/2}H_S K^{1/2}\omega + \tau_0^2 \\
&= \omega'(I - K^{1/2}H_S K^{-1/2})DD'(I - K^{-1/2}H_S K^{1/2})\omega \\
&\quad -\omega' K\omega + \omega' K^{1/2}H_S K^{-1/2}K\omega + \omega' K K^{-1/2}H_S K^{1/2}\omega \\
&\quad -\omega' K^{1/2}H_S K^{-1/2}K K^{-1/2}H_S K^{1/2}\omega \\
&\quad +\omega' K^{1/2}H_S K^{1/2}\omega + \tau_0^2 \\
&= \omega'(I - K^{1/2}H_S K^{-1/2})DD'(I - K^{-1/2}H_S K^{1/2})\omega \\
&\quad -\omega' K\omega + \omega' K^{1/2}H_S K^{1/2}\omega + \omega' K^{1/2}H_S K^{1/2}\omega \\
&\quad -\omega' K^{1/2}H_S K^{1/2}\omega \\
&\quad +\omega' K^{1/2}H_S K^{1/2}\omega + \tau_0^2 \\
&= \omega'(I - K^{1/2}H_S K^{-1/2})DD'(I - K^{-1/2}H_S K^{1/2})\omega \\
&\quad +2\omega' K^{1/2}H_S K^{1/2}\omega + (\tau_0^2 - \omega' K\omega)
\end{aligned}$$

Next they write the limiting (i.e. infeasible) version of the FIC by subtracting $\tau_0^2 - \omega' K \omega$ since this is constant across models. This gives

$$\begin{aligned} FIC &= \omega'(I - K^{1/2} H_S K^{-1/2}) DD' (I - K^{-1/2} H_S K^{1/2}) \omega \\ &\quad + 2\omega' K^{1/2} H_S K^{1/2} \omega \\ &= \omega'(I - K^{1/2} H_S K^{-1/2}) DD' (I - K^{-1/2} H_S K^{1/2}) \omega + 2\omega_S' K_S \omega_S \end{aligned}$$

Where $\omega_S = \pi_S \omega$. Finally, the FIC substitutes estimators as follows

$$\widehat{FIC} = \widehat{\omega}'(I - \widehat{K}^{1/2} \widehat{H}_S \widehat{K}^{-1/2}) \widehat{\delta}_{Full} \widehat{\delta}_{Full}' (I - \widehat{K}^{-1/2} \widehat{H}_S \widehat{K}^{1/2}) \widehat{\omega} + 2\widehat{\omega}_S' \widehat{K}_S \widehat{\omega}_S$$

This formula may look somewhat complicated, but calculating it only requires quantities that we get automatically from fitting the full model. Thus, the FIC does *not* require us to fit each of the candidate models.

4 Extensions of FIC Idea

The FIC idea turns out to be extremely general, and has been extended in a number of directions by the original authors, among others. Claeskens, Croux and Van Kerckhoven (2006) adapt the FIC idea to a number of loss functions besides MSE in the case of logistic regression, while Claeskens, Croux and Van Kerckhoven (2007) consider the problem of model selection for autoregressive models. Claeskens & Hjort (2008) consider both more general loss functions and focus parameters that depend on the data through some kind of average. In a more theoretical contribution, Claeskens & Carroll work out the asymptotics necessary to extend the FIC to semiparametric problems More recently, Brownlees and Gallo (2011) use the FIC to choose the amount of shrinkage used in estimation of the deterministic component of a conditional duration model, while Zhang, Wan and Zhou (2012) derive an FIC-type criterion for Tobit model selection. The idea behind the FIC can even be extended to GMM models. This is the topic of our next lecture.

5 Schorfheide (2005)

Although developed independently, Schorfheide (2005) shares many similarities with Claeskens & Hjort (2003). Working in a local asymptotic framework, this paper proposes for choosing VAR lag length and deciding between maximum likelihood and loss function-based estimation in multistep forecasting problems.