Lecture 5: "Focused" Model Selection

Francis J. DiTraglia

March 23, 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:

$$\begin{aligned}\text{Power}(T_N) &= P\left(\sqrt{N}\bar{X}_{2N} > z_{1-\alpha}\right) = P\left(Z + \mu\sqrt{N} > z_{1-\alpha}\right) \\ &= P\left(Z > z_{1-\alpha} - \mu\sqrt{N}\right) = 1 - \Phi\left(z_{1-\alpha} - \mu\sqrt{2N}\right)\end{aligned}$$

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

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

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

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

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

Put together a simple example (iid) showing what's going on here. Perhaps a little L_2 WLLN. Also, give the Lindeberg-Feller CLT. Point them to the Andrews papers that give technical conditions.

2 Focused Evaluation

Best model for a *particular purpose* rather than "one-size-fits all." Once we acknowledge that are models are probably wrong, we should ask "how wrong" and it might depend on what we want to use the model for. More generally, it's not clear that we would want to use the true model *even if we knew its form*: bias-variance tradeoff that comes from estimation uncertainty of more complicated models.

This example comes from Hansen (2005). Consider 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 $k = 0$.

2.1 The FIC for Hansen's (2005) Example

3 The Focused Information Criterion (FIC)

Portable, like AIC and BIC, based on risk minimization, like FPE and Mallow's C_p, but *focused* in the sense of Hansen (2005). Want to be able to select different models for different purposes. 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 major area of my research in the past few years!

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. Can look at various loss functions, but the original papers use MSE so that's what we'll look at here. I'll talk about extensions below. Roughly speaking, the idea 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:

$$J_S = Var_0 \left[\begin{array}{c} U(y) \\ V_S(y) \end{array} \right] = \left[\begin{array}{cc} J_{00} & J_{01,S} \\ J_{10,S} & J_{11,S} \end{array} \right] = \left[\begin{array}{cc} J_{00} & J_{01}\pi_S' \\ \pi_S J_{10} & \pi_S J_{11} \pi_S' \end{array} \right]$$

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):

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

$$\left[\begin{array}{c} \frac{1}{\sqrt{n}} \sum_{i=1}^n U(Y_i) \\ \frac{1}{\sqrt{n}} \sum_{i=1}^n V(Y_i) \end{array} \right] \xrightarrow{d} \left(\begin{array}{c} J_{01}\delta \\ J_{11}\delta \end{array} \right) + \left(\begin{array}{c} M \\ N \end{array} \right)$$

where

$$\left(\begin{array}{c} M \\ N \end{array} \right) \sim N_{p+q}(0, J_{Full})$$

Proof. To prove this result, we apply the Lindeberg-Feller CLT to the trian-

gular 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,

$$-E_0 \left[\begin{array}{c} \nabla_{\theta\gamma'} \log f(Y_i, \theta_0, \gamma_0) \\ \nabla_{\gamma\gamma'} \log f(Y_i, \theta_0, \gamma_0) \end{array} \right] = \left[\begin{array}{c} J_{01} \\ J_{11} \end{array} \right]$$

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

Lemma 3.2

$$\left[\begin{array}{c} \sqrt{n}(\hat{\theta} - \theta_0) \\ \sqrt{n}(\hat{\gamma} - \gamma_0) \end{array} \right] \xrightarrow{d} \left[\begin{array}{c} C_S \\ D_S \end{array} \right] = J_S^{-1} \left(\begin{array}{c} J_{01}\delta + M \\ \pi_S J_{11}\delta + N_S \end{array} \right) \sim \mathcal{N}_{p+|S|} \left(J_S^{-1} \left[\begin{array}{c} J_{01} \\ \pi_S J_{11} \end{array} \right] \delta, J_S^{-1} \right)$$

First Claim: Define $W \equiv J^{10}M + J^{11}N$. Then, $W = K(N - J_{10}J_{00}^{-1}M)$ and M and W are indep. with $W \sim \mathcal{N}_q(0, K)$ and $M \sim \mathcal{N}_p(0, J_{00})$.

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,

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

as required. Now we need to show the independence of W and M. Write

$$\left[\begin{array}{c} M \\ W \end{array}\right] = \left[\begin{array}{c} M \\ J^{10}M + J^{11}N \end{array}\right] = \left[\begin{array}{cc} I_p & 0_{p\times q} \\ J^{10} & J^{11} \end{array}\right] \left[\begin{array}{c} M \\ N \end{array}\right] \equiv A \left[\begin{array}{c} M \\ N \end{array}\right]$$

Since $\left[\begin{array}{c} M \\ N \end{array}\right] \sim \mathcal{N}_{p+q}(0, J_{Full})$, we have $A \left[\begin{array}{c} M \\ N \end{array}\right] \sim \mathcal{N}_{p+q}(0, AJ_{Full}A')$. Multiplying through, we find that

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

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 $J_{00}J^{01} + J_{01}J^{11} = 0$. Therefore

$$\begin{bmatrix} M \\ W \end{bmatrix} \sim \mathcal{N}_{p+q}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} J_{00} & 0 \\ 0 & J^{11} \end{bmatrix} \right)$$

Thus, $W \sim \mathcal{N}_q(0, J^{11})$ independent of $M \sim \mathcal{N}_p(0, J_{00})$. □

Second Claim: 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:

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

where, as before, $K \equiv J^{11}$.

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:

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

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

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

□

More Notation:

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

Notice that:

1. ω depends on the choice of focus parameter μ
2. H_S is symmetric and idempotent, thus it is a projection matrix.
3. From (2) it follows that H_S is orthogonal to $I - H_S$
4. Define $H_\emptyset$ as a $q \times q$ null matrix.

Lemma 3.3 If μ has continuous partial derivatives in a neighborhood of (θ_0, γ_0), then:

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

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

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

Proof. Applying the delta-method to Lemma 3.2 along with a mean value expansion for

$$\mu(\hat{\theta}_S, \hat{\gamma}_S) - \mu(\theta_0, \gamma_0 + \delta/\sqrt{n})$$

gives[1]

$$\sqrt{n}\left(\hat{\mu}_S - \mu_{true}\right) \xrightarrow{d} \nabla_\theta \mu(\theta_0,\gamma_0)' C_S + \left[\pi_S \nabla_\gamma \mu(\theta_0,\gamma_0)\right]' D_S - \nabla_\gamma \mu(\theta_0,\gamma_0)'\delta \equiv \Lambda_S$$

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 express the limiting RV Λ_S in the alternative formulation given in Lemma 3.3. This is because M and $D = \delta + W$ are *independent* normal random vectors! 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:

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

[1] for details see Section 2.5 of my First-Year Paper

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,

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

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

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

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,

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

and

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

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

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

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

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

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

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

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

□

Compromise Estimators: From Lemma 3.2, we know that

$$\hat{\delta}_S \equiv \sqrt{n}\left(\hat{\gamma}_S - \gamma_{0,S}\right) \xrightarrow{d} D_S = K_S \pi_S K^{-1} (\delta + W)$$

In the case of the full model, that is $S = \left\{1, 2, \xrightarrow{d} ots, q\right\}$ and $\pi_S = I_q$ so that $K_S = K$, this gives

$$D_n \equiv \hat{\delta}_{full} \equiv \sqrt{n}\left(\hat{\gamma}_{full} - \gamma_0\right) \xrightarrow{d} D = (\delta + W)$$

Thus, *any* submodel estimator $\hat{\delta}_S$ of δ converges in distribution to a linear combination of D, while the full model estimator of D_n of δ simply converges in distribution to D. In other words, the behavior of $\hat{\delta}_S$ is "essentially determined" by that of D_n. More precisely, the difference between $\hat{\delta}_S$ and $K_S \pi_S K^{-1} D_n$ is

at most $o_p(1)$. Now consider a **Compromise Estimator** of the form:

$$\hat{\mu} = \sum_S c\left(S|D_n\right)\hat{\mu}_S$$

that is, we weight and sum submodel estimators where the weights are a function of $D_n = \hat{\delta}_{full} = \sqrt{n}(\hat{\gamma}_{full} - \gamma_0)$. *To ensure consistency, the weights must sum to one.*

Notation: Define G, a $q \times q$ matrix of functions, by

$$G(\xrightarrow{d} ot) = K^{-1/2}\left\{\sum_S c(S|\xrightarrow{d} ot)H_S\right\}K^{1/2}$$

and $\hat{\delta}(D)$, an estimator of δ based on D, by

$$\hat{\delta}(D) = G(D)'D$$

Since H_S is symmetric and the weights $c(\xrightarrow{d} ot|\xrightarrow{d} ot)$ are scalars,

$$\hat{\delta}(D) = \left[K^{-1/2}\left\{\sum_S c(S|D)H_S\right\}K^{1/2}\right]' D = K^{1/2}\left\{\sum_S c(S|D)H_S\right\}K^{-1/2}D$$

Theorem 4.1 As long as the weight functions $c(\xrightarrow{d} ot|\xrightarrow{d} ot)$ sum to one and have at most a countable number of discontinuities, then

$$\sqrt{n}\left(\hat{\mu} - \mu_{true}\right) \xrightarrow{d} \sum_S c(S|D)\Lambda_S \equiv \Lambda$$

and

$$\Lambda = \nabla_\theta \mu(\theta_0, \gamma_0)' J_{00}^{-1} M + \omega'\left[\delta - \hat{\delta}(D)\right]$$

This is, in general, a **non-normal distribution** with

$$\begin{aligned} \text{mean} &= \omega' \left\{ \delta - E\left[\hat{\delta}(D)\right] \right\} \\ \text{variance} &= \tau_0^2 + \omega' Var\left[\hat{\delta}(D)\right] \omega \end{aligned}$$

where

$$\begin{aligned} \tau_0^2 &\equiv \nabla + \theta\mu(\theta_0, \gamma_0)' J_{00}^{-1} \\ \omega &\equiv J_{10} J_{00}^{-1} \nabla_\theta \mu(\theta_0, \gamma_0) - \nabla_\gamma \mu(\theta_0, \gamma_0) \end{aligned}$$

and the MSE of Λ is

$$E[\Lambda^2] = \tau_0^2 + R(\delta)$$

where

$$R(\delta) = \omega' \left[\left\{ \hat{\delta}(D) - \delta \right\} \left\{ \hat{\delta}(D) - \delta \right\}' \right] \omega$$

Proof. First, using the fact that $\sum_S c(S|D_n) = 1$, we have

$$\begin{aligned} \sqrt{n}\left(\hat{\mu} - \mu_{true}\right) &= \sqrt{n}\left[\sum_S c(S|D_n)\hat{\mu}_S - \mu_{true} \right] \\ &= \sqrt{n}\left[\sum_S c(S|D_n)\hat{\mu}_S - \left\{ \sum_S c(S|D_n) \right\} \mu_{true} \right] \\ &= \sum_S \left[c(S|D_n)\sqrt{n}\left(\hat{\mu}_S - \mu_{true}\right) \right] \end{aligned}$$

So we see that $\sqrt{n}\left(\hat{\mu} - \mu_{true}\right)$ is an almost-surely continuous function of the submodel estimators $\sqrt{n}\left(\hat{\mu}_S - \mu_{true}\right)$ and $D_n = \sqrt{n}\left(\hat{\gamma}_{full} - \gamma_0\right)$. Thus, to find the limiting distribution of the compromise estimator, we can apply the continuous mapping theorem, provided we have *joint* convergence in distribution of the submodel estimators and D_n.

Fortunately, we have already established precisely this joint convergence! In Lemma 3.3, we showed that the limit distribution of each submodel esti-

mator $\sqrt{n}\,(\hat{\mu}_S - \mu_{true})$ is a linear combination of

$$\begin{pmatrix} M \\ N \end{pmatrix} \sim \mathcal{N}_{p+q}(0, J_{full})$$

Further the limit distribution of $D_n = \sqrt{n}\,(\hat{\gamma}_{full} - \gamma_0)$ is another linear combination of M and N, namely

$$D = (\delta + W) = \delta + K\left(N - J_{10}J_{00}^{-1}M\right)$$

Therefore, the limiting distribution of all the submodel estimators *jointly* with D_n can be written as the appropriate linear combination of $(M', N')'$, so the joint distribution is multivariate normal. Now we can apply the continuous mapping theorem as desired, to find:

$$\sqrt{n}\,(\hat{\mu} - \mu_{true}) = \sum_S \left[c(S|D_n)\sqrt{n}\,(\hat{\mu}_S - \mu_{true})\right] \xrightarrow{d} \sum_S c(S|D_n)\Lambda_S$$

where Λ_S is the limit distribution of $\sqrt{n}\,(\hat{\mu}_S - \mu_{true})$ defined above. Let

$$\Lambda \equiv \sum_S c(S|D_n)\Lambda_S$$

We want to express Λ in a more convenient form using the fact that

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

as shown in Lemma 3.3. Substituting,

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

where we have used the following facts:

1. Only H_S depends on S.
2. The weights sum to one.
3. As scalars, the weights commute and are (trivially) symmetric.
4. H_S, $K^{1/2}$, and $K^{-1/2}$ are symmetric.

along with the definitions of $G(\xrightarrow{d} ot)$ and $\hat{\delta}(D)$. Now we have shown that

$$\Lambda = \nabla_\theta \mu(\theta_0,\gamma_0)' J_{00}^{-1} M + \omega'\left[\delta - \hat{\delta}(D)\right]$$

Notice that, since $\hat{\delta}(D)$ depends only on $D = \delta + W$, and M is independent of W, it follows that the two terms in this expression are likewise independent.

The first follows a normal distribution but the second is, in general, non-normal.

Now, since M and $D = \delta + W$ are independent, it follows that the distribution of $M|D$ is the same as that of M. Thus,

$$\Lambda|(D = d) \sim \nabla_\theta \mu(\theta_0, \gamma_0)' J_{00}^{-1} M + \omega' \left[\delta - \hat{\delta}(d)\right]$$

which is a normal distribution, since $\hat{\delta}(d)$ is a constant taking into account the conditioning. The mean and variance are as follows:

$$\begin{aligned} E\left[\Lambda|D = d\right] &= E\left[\nabla_\theta \mu(\theta_0, \gamma_0)' J_{00}^{-1} M\right] + \omega' \left[\delta - \hat{\delta}(d)\right] \\ &= \nabla_\theta \mu(\theta_0, \gamma_0)' J_{00}^{-1} E\left[M\right] + \omega' \left[\delta - \hat{\delta}(d)\right] \\ &= \omega' \left[\delta - \hat{\delta}(d)\right] \end{aligned}$$

$$\begin{aligned} Var\left[\Lambda|D = d\right] &= Var\left[\nabla_\theta \mu(\theta_0, \gamma_0)' J_{00}^{-1} M\right] \\ &= \nabla_\theta \mu(\theta_0, \gamma_0)' J_{00}^{-1} Var[M] J_{00}^{-1} \nabla_\theta \mu(\theta_0, \gamma_0) \\ &= \nabla_\theta \mu(\theta_0, \gamma_0)' J_{00}^{-1} J_{00} J_{00}^{-1} \nabla_\theta \mu(\theta_0, \gamma_0) \\ &= \nabla_\theta \mu(\theta_0, \gamma_0)' J_{00}^{-1} \nabla_\theta \mu(\theta_0, \gamma_0) \\ &\equiv \tau_0^2 \end{aligned}$$

since $\omega'(\delta - \hat{\delta}(d))$ is a constant. Note that τ_0^2 is the *minimal variance* of the estimators under consideration. Although the *unconditional distribution* of Λ is non-normal, we can still calculate its mean and variance using our decomposition into two independent terms and the linearity of expectation:

$$\begin{aligned} E\left[\Lambda\right] &= E\left[\nabla_\theta \mu(\theta_0, \gamma_0)' J_{00}^{-1} M\right] + E\left[\omega' \left\{\delta - \hat{\delta}(d)\right\}\right] \\ &= \omega'\delta - \omega' E\left[\hat{\delta}(d)\right] \\ &= \omega' \left\{\delta - E\left[\hat{\delta}(d)\right]\right\} \end{aligned}$$

$$
\begin{aligned}
Var\left[\Lambda\right] &= Var\left[\nabla_\theta \mu(\theta_0,\gamma_0)' J_{00}^{-1} M\right] + Var\left[\omega'\left\{\delta - \hat{\delta}(d)\right\}\right] \\
&= \tau_0^2 + \omega' Var\left[\hat{\delta}(D)\right]\omega
\end{aligned}
$$

Now, Λ is the limit distribution of $\sqrt{n}(\hat{\mu} - \mu_{true})$ where $\hat{\mu}$ is the compromise estimator, thus if asymptotically unbiased, it should be centered around zero. Accordingly we find the MSE of Λ as follows:

$$
\begin{aligned}
MSE(\Lambda) &= E\left[(\Lambda - 0)^2\right] = E\left[\Lambda^2\right] \\
&= E\left[\left(\nabla_\theta \mu(\theta_0,\gamma_0)' J_{00}^{-1} M + \omega'\left\{\delta - \hat{\delta}(D)\right\}\right)^2\right] \\
&= E\left[\left\{\nabla_\theta \mu(\theta_0,\gamma_0)' J_{00}^{-1} M\right\}^2\right] + E\left[\left(\omega'\left\{\delta - \hat{\delta}(D)\right\}\right)^2\right] \\
&\qquad +2E\left[\nabla_\theta \mu(\theta_0,\gamma_0)' J_{00}^{-1} M \omega'\left\{\delta - \hat{\delta}(D)\right\}\right] \\
&= E\left[\nabla_\theta \mu(\theta_0,\gamma_0)' J_{00}^{-1} M M' J_{00}^{-1} \nabla_\theta \mu(\theta_0,\gamma_0)\right] \\
&\qquad +E\left[\omega'\{\delta - \hat{\delta}(D)\}\{\delta - \hat{\delta}(D)\}'\omega\right] \\
&\qquad +2E\left[\nabla_\theta \mu(\theta_0,\gamma_0)' J_{00}^{-1} M \omega'\left\{\delta - \hat{\delta}(D)\right\}\right] \\
&= \nabla_\theta \mu(\theta_0,\gamma_0)' J_{00}^{-1} E\left[MM'\right] J_{00}^{-1} \nabla_\theta \mu(\theta_0,\gamma_0) \\
&\qquad +\omega' E\left[\{\delta - \hat{\delta}(D)\}\{\delta - \hat{\delta}(D)\}'\right]\omega \\
&\qquad +2\nabla_\theta \mu(\theta_0,\gamma_0)' J_{00}^{-1} E\left[M\right] E\left[\omega'\left\{\delta - \hat{\delta}(D)\right\}\right] \\
&= \nabla_\theta \mu(\theta_0,\gamma_0)' J_{00}^{-1} J_{00} J_{00}^{-1} \nabla_\theta \mu(\theta_0,\gamma_0) \\
&\qquad +\omega' E\left[\{\delta - \hat{\delta}(D)\}\{\delta - \hat{\delta}(D)\}'\right]\omega \\
&= \tau_0^2 + \omega' E\left[\{\delta - \hat{\delta}(D)\}\{\delta - \hat{\delta}(D)\}'\right]\omega \\
&= \tau_0^2 + R(\delta)
\end{aligned}
$$

Where we have used the fact that $E[M] = 0$ and hence $Var[M] = E[MM']$ and the independence of M and D and hence of any measurable functions thereof. □

IMPORTANT: For convenience, define $\Lambda_0 = \nabla_\theta \mu(\theta_0, \gamma_0)' J_{00}^{-1} M$. The *key point* here is that the distribution of

$$\Lambda = \sum_S c(S|D)\Lambda_S = \Lambda_0 + \omega'\{\delta - \hat{\delta}(D)\}$$

is often **dramatically non-normal**. To find the density of Λ, first condition on D using the result from above:

$$\begin{aligned} \Lambda|(D = x) &= \Lambda_0 + \omega'\{\delta - \hat{\delta}(x)\} \\ &\sim \mathcal{N}(0, \tau_0^2) + \omega'\{\delta - \hat{\delta}(x)\} \end{aligned}$$

Now, let $h(z)$ denote the density of Λ. We can calculate h by integrating D out of the joint density of (Λ, D). Let $f(x)$ denote the density of D. We have

$$\begin{cases} h(z|D = x) \sim \mathcal{N}\left(\omega'\{\delta - \hat{\delta}(x)\}, \tau_0^2\right) \\ D = \delta + W \sim \mathcal{N}_q(\delta, K) \end{cases}$$

Now factor the joint density according to $h(z|D = x)f(x)$ and integrate out D as follows:

$$h(z) = \int h(z|D = x)f(x)\, dx$$

We can then substitute the two normal distributions and then either numerically integrate or simulate. *Notice*, however, that **the result depends on the unknown constant** δ.

Using the Full Model Variance One approach to constructing a confidence interval that takes account of model selection uncertainty is to essentially

use the variance of the full model. Define

$$\begin{aligned}
\tau_{full}^2 &= \text{AVAR}(\hat{\mu}_{full}) = Var[\Lambda_{full}] \\
&= \nabla_\theta \mu(\theta_0, \gamma_0)' J_{00}^{-1} \nabla_\theta \mu(\theta_0, \gamma_0) + \omega' K^{1/2} H_{full} K^{1/2} \omega \\
&= \tau_0^2 + \omega' K^{1/2} H_{full} K^{1/2} \omega \\
&= \tau_0^2 + \omega' K^{1/2} \left\{ K^{-1/2} (\pi_{full}' K_{full} \pi_{full}) K^{-1/2} \right\} K^{1/2} \omega \\
&= \tau_0^2 + \omega' K^{1/2} K^{-1/2} K K^{-1/2} K^{1/2} \omega \\
&= \tau_0^2 + \omega' K \omega
\end{aligned}$$

And accordingly $\tau_{full} = (\tau_0^2 + \omega' K \omega)^{1/2}$. Now let $\hat{\omega}$ be a consistent estimator of ω and $\widehat{\kappa}$ be a consistent estimator of τ_{full}. Define

$$T_n = \left[\sqrt{n}(\hat{\mu} - \mu_{true}) - \widehat{\omega}' \left\{ D_n - \sum_{S \in \mathcal{A}} c(S|D_n) G_S D_n \right\} \right] \Big/ \widehat{\kappa}$$

From above, we know that the following converges jointly in distribution:

$$\begin{bmatrix} \sqrt{n}(\hat{\mu} - \mu_{true}) \\ D_n \end{bmatrix} \xrightarrow{d} \begin{bmatrix} \Lambda_0 + \omega' \left\{ \delta - \sum_{S \in \mathcal{A}} c(S|D) G_S D \right\} \\ D \end{bmatrix}$$

Thus

$$\begin{aligned}
T_n &\xrightarrow{d} \left[\Lambda_0 + \omega' \left\{ \delta - \sum_{S \in \mathcal{A}} c(S|D) G_S D \right\} - \omega' \left\{ D - \sum_{S \in \mathcal{A}} c(S|D) G_S D \right\} \right] \Big/ \tau_{full} \\
&= (\tau_0^2 + \omega' K \omega)^{-1/2} \left[\Lambda_0 + \omega' (\delta - D) \right]
\end{aligned}$$

We know from above that

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

where $K = J^{11}$, so

$$\begin{aligned} \Lambda_0 &\equiv \nabla_\theta \mu(\theta_0, \gamma_0)' J_{00}^{-1} M \\ &\sim \mathcal{N}\left(0, \tau_0^2\right) \end{aligned}$$

independently of

$$\begin{aligned} \omega'(\delta - D) &\equiv -\omega' W \\ &\sim \mathcal{N}(0, \omega' K \omega) \end{aligned}$$

Therefore

$$\begin{aligned} T_n &\xrightarrow{d} (\tau_0^2 + \omega' K \omega)^{-1/2} \left[\Lambda_0 + \omega'(\delta - D)\right] \\ &= (\tau_0^2 + \omega' K \omega)^{-1/2} \times \mathcal{N}\left(0, \tau_0^2 + \omega' K \omega\right) \\ &\sim \mathcal{N}(0, 1) \end{aligned}$$

which is a **standard normal**! We can use this result to create a approximate confidence interval for $\hat{\mu}$ as follows. For large n,

$$T_n = \widehat{\kappa}^{-1} \left[\sqrt{n}(\hat{\mu} - \mu_{true}) - \omega' \left\{ D_n - \sum_{S \in \mathcal{A}} c(S|D_n) G_S D_n \right\} \right] \approx \mathcal{N}(0, 1)$$

To create a $(1-\alpha) \times 100\%$ interval, define $z_{\alpha/2}$ as the appropriate quantile of a standard normal random variable so that

$$P\left[-z_{\alpha/2} \leq T_n \leq z_{\alpha/2}\right] \approx 1 - \alpha$$

Then,

$$
\begin{aligned}
1-\alpha &\approx P\left[-\frac{z_{\alpha/2}\widehat{\kappa}}{\sqrt{n}} \leq (\hat{\mu}-\mu_{true}) - \frac{\widehat{\omega}'}{\sqrt{n}}\left\{D_n - \sum_{S\in\mathcal{A}} c(S|D_n)G_S D_n\right\} \leq \frac{z_{\alpha/2}\widehat{\kappa}}{\sqrt{n}}\right] \\
&= P\left[\frac{z_{\alpha/2}\widehat{\kappa}}{\sqrt{n}} \geq (\mu_{true}-\hat{\mu}) + \frac{\widehat{\omega}'}{\sqrt{n}}\left\{D_n - \sum_{S\in\mathcal{A}} c(S|D_n)G_S D_n\right\} \geq -\frac{z_{\alpha/2}\widehat{\kappa}}{\sqrt{n}}\right] \\
&= P\left[-\frac{z_{\alpha/2}\widehat{\kappa}}{\sqrt{n}} \leq (\mu_{true}-\hat{\mu}) + \frac{\widehat{\omega}'}{\sqrt{n}}\left\{D_n - \sum_{S\in\mathcal{A}} c(S|D_n)G_S D_n\right\} \leq \frac{z_{\alpha/2}\widehat{\kappa}}{\sqrt{n}}\right] \\
&= P\left[\hat{\mu}-\frac{z_{\alpha/2}\widehat{\kappa}}{\sqrt{n}} \leq \mu_{true} + \frac{\widehat{\omega}'}{\sqrt{n}}\left\{D_n - \sum_{S\in\mathcal{A}} c(S|D_n)G_S D_n\right\} \leq \hat{\mu}+\frac{z_{\alpha/2}\widehat{\kappa}}{\sqrt{n}}\right] \\
&= P\left[\hat{\mu}-\frac{z_{\alpha/2}\widehat{\kappa}}{\sqrt{n}} \leq \mu_{true} + \Delta_n \leq \hat{\mu}+\frac{z_{\alpha/2}\widehat{\kappa}}{\sqrt{n}}\right] \\
&= P\left[\left(\hat{\mu}-\frac{z_{\alpha/2}\widehat{\kappa}}{\sqrt{n}}\right) - \Delta_n \leq \mu_{true} \leq \left(\hat{\mu}+\frac{z_{\alpha/2}\widehat{\kappa}}{\sqrt{n}}\right) - \Delta_n\right]
\end{aligned}
$$

where

$$\Delta_n \equiv \frac{\widehat{\omega}'}{\sqrt{n}}\left[D_n - \sum_{S\in\mathcal{A}} c(S|D_n)G_S D_n\right]$$

According to Claeskens and Hjort: "this method is first-order equivalent to using the full model for confidence interval construction, with a modification for location."

Correcting Confidence Intervals: Simulation Another possibility is to simulate from the limiting distribution of Λ *for a range fixed value of* δ, using consistent estimates of all other unknown quantities. This procedure can then be repeated for a variety of choices of δ. To make this clearer, first rewrite Λ

using an expression from the proof Theorem 4.1

$$\begin{aligned}
\Lambda &= \nabla_\theta \mu(\theta_0, \gamma_0)' J_{00}^{-1} M + \omega' \delta - \omega' K^{1/2} \left[\sum_S c(S|D) H_S \right] K^{-1/2} D \\
&= \Lambda_0 + \omega' \left[\delta - K^{1/2} \sum_S c(S|D) H_S K^{-1/2} D \right] \\
&= \Lambda_0 + \omega' \left[\delta - \sum_S c(S|D) K^{1/2} H_S K^{-1/2} D \right] \\
&= \Lambda_0 + \omega' \left[\delta - \sum_S c(S|D) G_S D \right]
\end{aligned}$$

where we have defined $G_S = K^{1/2} H_S K^{-1/2}$. We know from above that $\Lambda_0 \sim \mathcal{N}(0, \tau_0^2)$ independent of $D \sim \mathcal{N}_q(\delta, K)$. The simulation procedure is as follows:

1. Calculate consistent estimates of G_S, τ_0^2, ω and K using the estimation results and fix a value of δ.

2. Using the result of step one, generate:

   (a) $\Lambda_{0,j} \sim \mathcal{N}(0, \hat{\tau}_0^2)$ independently of

   (b) $D_j \sim \mathcal{N}_q(\delta, \hat{K})$

3. Calculate the weights c using D_j and set

$$\Lambda_j = \Lambda_{0,j} + \hat{\omega}' \left[\delta - \sum_S c(S|D_j) \hat{G}_S D_j \right]$$

4. Repeat steps 1 and 2 for $j = 1, 2, \ldots, B$

5. Using the samples $\{\Lambda_1, \Lambda_2, \ldots, \Lambda_B\}$ generated in steps 3 and 4, calculate quantiles $a(\delta)$ and $b(\delta)$ that satisfy:

$$P\left[a(\delta) \leq \Lambda(\delta) \leq b(\delta)\right] = 0.95$$

6. Repeat steps 2–5 for varying choices of δ.

Now, suppose we know the values $[a(\delta), b(\delta)]$. Since Λ is the limit distribution of $\sqrt{n}(\hat{\mu} - \mu_{true})$, it follows that

$$P\left[a(\delta) \leq \sqrt{n}(\hat{\mu} - \mu_{true}) \leq b(\delta)\right] \to P\left[a(\delta) \leq \Lambda(\delta) \leq b(\delta)\right] = 0.95$$

Thus, $[\hat{\mu} - b(\delta)/\sqrt{n}, \hat{\mu} - a(\delta)/\sqrt{n}]$ covers μ_{true} with probability 0.95 asymptotically.

But We Don't Know δ! A naive approach would be to substitute our estimate $D_n = \hat{\delta}_{full} = \sqrt{n}(\hat{\gamma}_{full} - \gamma_0)$, carrying about the above simulations at this value and creating an interval based on $\hat{a} = a(D_n)$ and $\hat{b} = b(D_n)$. This is simple, but may not always work well. Let $p_n(\delta)$ be the coverage probability for this procedure. Its limit is

$$p(\delta) = P\left[a(D) \leq \Lambda(\delta) \leq b(D)\right]$$

which can be simulated by the method described above. It turns out that this method sometimes gives coverage that is *far too low*.

A Better Procedure: Instead of simply substituting D_n for δ in the simulation described above, we could first construct a confidence region for δ and use this region to create an interval for $\hat{\mu}$. Since

$$D_n \xrightarrow{d} D = \delta + W \sim \mathcal{N}_q\left(\delta, K\right)$$

where $K = J^{11}$, we have

$$\left(D_n - \delta\right)' \widehat{K}^{-1} \left(D_n - \delta\right) \xrightarrow{d} \chi^2_q$$

Now, define

$$\rho_n(D_n, \delta) = \left[\left(D_n - \delta\right)' \widehat{K}^{-1} \left(D_n - \delta\right)\right]^{1/2}$$

and the event

$$A_n(c) = \{\rho_n(D_n, \delta) \leq c\}$$

Now, since $\rho_n(D_n, \delta)^2 \approx \chi^2_q$ we have

$$P\{A_n(c)\} = P\{\rho_n(D_n, \delta) \leq c\} = P\left\{\rho_n(D_n, \delta)^2 \leq c^2\right\} \approx P\left\{\chi^2_q \leq c^2\right\}$$

where we have used the fact that x^2 is strictly increasing for $x \geq 0$ and that $\rho_n \geq 0$. Now define $z = (\chi^2_{q,0.95})^{1/2}$ and $A_n = A_n(z)$, so that $P\{A_n\} \approx 0.95$. In the simulations described above in which we assumed that δ was known, we defined $a(\delta)$ and $b(\delta)$ so that

$$P[a(\delta) \leq \Lambda(\delta) \leq b(\delta)] = 0.95$$

Now, let

$$\begin{aligned} \widehat{a}_0(D_n) &= \min\{a(\delta)\colon \rho_n(D_n, \delta) \leq z\} \\ \widehat{b}_0(D_n) &= \max\{b(\delta)\colon \rho_n(D_n, \delta) \leq z\} \end{aligned}$$

The claim is that the limit coverage level of

$$\mathrm{CI}^*_n = \left[\widehat{\mu} - \frac{\widehat{b}_0(D_n)}{\sqrt{n}}, \quad \widehat{\mu} - \frac{\widehat{a}_0(D_n)}{\sqrt{n}}\right]$$

is always *above* 0.90, resulting in a conservative procedure. To see why this is the case, we return to the limit experiment, in which we have joint convergence of all the necessary random variables, as described above. This implies that the coverage probability $r_n(\delta)$ to which $\{\mu_{true} \in \mathrm{CI}^*_n\}$ converges is given by

$$r(\delta) = P\{a_0(D) \leq \Lambda(\delta) \leq b_0(D)\}$$

where $\rho(D,\delta)^2 = (D-\delta)' K^{-1} (D-\delta)$ and

$$\begin{aligned} a_0(D) &= \min\{a(\delta)\colon \rho(D,\delta) \leq z\} \\ b_0(D) &= \min\{b(\delta)\colon \rho(D,\delta) \leq z\} \end{aligned}$$

How does this work? The interval CI_n^* is based on

$$\begin{aligned} 0.9 &\leq P\left[\widehat{\mu} - \frac{\widehat{b}_0(D_n)}{\sqrt{n}} \leq \mu_{true} \leq \widehat{\mu} - \frac{\widehat{a}_0(D_n)}{\sqrt{n}}\right] \\ &= P\left[-\frac{\widehat{b}_0(D_n)}{\sqrt{n}} \leq \mu_{true} - \widehat{\mu} \leq -\frac{\widehat{a}_0(D_n)}{\sqrt{n}}\right] \\ &= P\left[-\widehat{b}_0(D_n) \leq \sqrt{n}\,(\mu_{true} - \widehat{\mu}) \leq -\widehat{a}_0(D_n)\right] \\ &= P\left[\widehat{b}_0(D_n) \geq \sqrt{n}\,(\widehat{\mu} - \mu_{true}) \geq \widehat{a}_0(D_n)\right] \\ &= P\left[\widehat{a}_0(D_n) \leq \sqrt{n}\,(\widehat{\mu} - \mu_{true}) \leq \widehat{b}_0(D_n)\right] \end{aligned}$$

Now, we know that

$$\begin{bmatrix} \sqrt{n}(\hat{\mu} - \mu_{true}) \\ D_n \end{bmatrix} \xrightarrow{d} \begin{bmatrix} \Lambda_0 + \omega' \left\{\delta - \sum_{S\in\mathcal{A}} c(S|D) G_S D\right\} \\ D \end{bmatrix}$$

so, by the Continuous Mapping Theorem,

$$\begin{bmatrix} \sqrt{n}(\hat{\mu} - \mu_{true}) \\ \rho_n(D_n,\delta) \\ \widehat{a}_0(D_n) \\ \widehat{b}_0(D_n) \end{bmatrix} \xrightarrow{d} \begin{bmatrix} \Lambda(\delta) \\ \rho(D,\delta) \\ a_0(D) \\ b_0(D) \end{bmatrix}$$

and thus,

$$P\left[\widehat{a}_0(D_n) \leq \sqrt{n}\,(\widehat{\mu} - \mu_{true}) \leq \widehat{b}_0(D_n)\right] \to P\left\{a_0(D) \leq \Lambda(\delta) \leq b_0(D)\right\} = r(\delta)$$

Now, let $A = \{\rho(D, \delta) \leq z\}$ where, as before, $z = (\chi^2_{q,0.95})^{1/2}$ implying that $P\{A\} = 0.95$. Then,

$$\begin{aligned} 0.95 &= P\{a(\delta) \leq \Lambda(\delta) \leq b(\delta)\} \\ &= P[\{a(\delta) \leq \Lambda(\delta) \leq b(\delta)\} \cap A] + P[\{a(\delta) \leq \Lambda(\delta) \leq b(\delta)\} \cap A^c] \end{aligned}$$

Now, since

$$\begin{aligned} a_0(D) &= \min\{a(\delta) \colon \rho(D, \delta) \leq z\} \\ b_0(D) &= \min\{b(\delta) \colon \rho(D, \delta) \leq z\} \end{aligned}$$

we have

$$\{a(\delta) \leq \Lambda(\delta) \leq b(\delta)\} \cap A \Rightarrow \{a_0(D) \leq \Lambda(\delta) \leq b_0(D)\}$$

and hence

$$P[\{a(\delta) \leq \Lambda(\delta) \leq b(\delta)\} \cap A] \leq P\{a_0(D) \leq \Lambda(\delta) \leq b_0(D)\}$$

Further, since

$$\{a(\delta) \leq \Lambda(\delta) \leq b(\delta)\} \cap A^c \Rightarrow A^c$$

we have

$$P[\{a(\delta) \leq \Lambda(\delta) \leq b(\delta)\} \cap A^c] \leq P(A^c)$$

Combining:

$$\begin{aligned} 0.95 &= P[\{a(\delta) \leq \Lambda(\delta) \leq b(\delta)\} \cap A] + P[\{a(\delta) \leq \Lambda(\delta) \leq b(\delta)\} \cap A^c] \\ &\leq P\{a_0(D) \leq \Lambda(\delta) \leq b_0(D)\} + P(A^c) \\ &= P\{a_0(D) \leq \Lambda(\delta) \leq b_0(D)\} + 0.05 \end{aligned}$$

since A is defined with reference to a 95% confidence interval. Subtracting,

$$P\left\{a_0(D) \leq \Lambda(\delta) \leq b_0(D)\right\} \geq 0.90$$

as claimed.

Here's the intuition for what just happened. Λ is a random variable whose distribution depends on the unknown constant δ. The constants $a(\delta)$ and $b(\delta)$ are quantiles of the distribution of Λ such that

$$P\left[a(\delta) \leq \Lambda(\delta) \leq b(\delta)\right] = 0.95$$

Since Λ depends on δ, so do its quantiles: different values of δ would result in different intervals.

This is the procedure. First we use $\rho(D, \delta) \leq z$ to get a confidence interval for δ. Then we plug each point in this interval for δ into $\Lambda(\delta)$ and calculate the corresponding bounds $a(\delta)$ and $b(\delta)$. For each value of δ such that $\rho(D, \delta) \leq z$ we get a *different* confidence interval for Λ. The lower bound of all these intervals is $a_0(D)$ while the upper bound is $b_0(D)$. The point here is to assess the coverage of the resulting interval.

The confusion here comes from bad notation: sometimes δ is being treated as fixed, other times as variable. Need to come up with some clearer notation...

4 Schorfheide (2005)