Preliminary and Incomplete

A Generalized Focused Information Criterion for GMM Model and Moment Selection

Francis J. DiTraglia

University of Pennsylvania

May 29, 2013

Abstract

In this paper I propose a novel criterion for simultaneous GMM model and moment selection: the generalized focused information criterion (GFIC). Rather than attempting to identify the correct specification, the GFIC chooses from a set of potentially mis-specified overidentifying and parameter restrictions to minimize the mean-squared error of a target parameter specified by the user. Both the focused moment selection criterion (FMSC) of DiTraglia (2012) and the focused information criterion (FIC) of Claeskens and Hjort (2003) and the can be viewed as special cases of the GFIC. In addition to the general theory, I specialize the GFIC to a simple dynamic panel model with correlated individual effects. In a simulation study based on this example, the GFIC performs well relative to alternatives from the literature.

1 Introduction

An econometric model is a tool for answering a particular research question: different questions may suggest different models for the same data. And the fact that a model is wrong, as the old saying goes, does not prevent it from being

useful. This paper proposes a novel selection criterion for GMM estimation that takes both of these points to heart: the generalized focused information criterion (GFIC). Rather than attempting to identify the correct specification, the GFIC chooses from a set of potentially mis-specified moment conditions and parameter restrictions to yield the smallest mean squared error (MSE) estimator of a user-specified scalar target parameter. I derive the GFIC under local mis-specification, using asymptotic mean squared error (AMSE) to approximate finite-sample MSE. In this framework mis-specification, while present for any fixed sample size, disappears in the limit so that asymptotic variance and squared bias remain comparable. GMM estimators remain consistent under local mis-specification but their limit distributions show an asymptotic bias. Adding an additional moment condition or imposing a parameter restriction generally reduces asymptotic variance but, if incorrectly specified, introduces a source of bias. The GFIC trades off these two effects in the first-order asymptotic expansion of an estimator to approximate its finite sample behavior.

The GFIC requires two assumptions. First, it must be possible to write all candidate models as parameter restrictions applied to a "wide" or "unrestricted" model. In a regression setting, for example, the unrestricted model would include all regressors while the other candidate models might restrict certain regression coefficients to be zero. As long as all candidate models nest inside the unrestricted model, the GFIC allows us to make non-nested model comparisons. Second, we must have a minimal set of moment conditions that are known to be valid and identify the parameters of the unrestricted model. When these conditions hold, the GFIC provides an asymptotically unbiased estimator of AMSE for each candidate specification. When the second condition does not hold, it remains possible to use the GFIC to carry out a sensitivity analysis (c.f. DiTraglia (2012), Section 3.2), although I will not consider this situation further here.

As its name suggests, the GFIC extends the focused information criterion (FIC) of Claeskens and Hjort (2003), a model selection procedure for maximum likelihood estimators that uses local mis-specification to approximate the MSE of a target

parameter. The idea of targeted, risk-based model selection has proved popular in recent years, leading to a number of interesting extensions. Hjort and Claeskens (2006), for example, propose an FIC for the Cox proportional hazards model while Claeskens and Carroll (2007) extend the FIC more generally to problems in which the likelihood involves an infinite-dimensional parameter but selection is carried out over the parametric part. More recently, Zhang and Liang (2011) extend the FIC to generalized additive partially linear models and Behl et al. (2012) develop an FIC for quantile regression.

While MSE is a natural risk-function for asymptotically normal estimators, different applications of model selection may call for different risk functions. Claeskens et al. (2006), for example, suggest combining local mis-specification with L_p-risk or mis-classification error rates to derive an FIC better-suited to prediction in logistic regression models. In a similar vein, the weighted FIC (wFIC) of Claeskens and Hjort (2008) provides a potentially important tool for policy analysis, allowing researchers to choose the model that minimizes weighted average risk for generalized linear models. While the FIC can be used, for example, to choose the best model for estimating the mean response at a given set of covariate values, the wFIC allows us to minimize the expected mean response over a *distribution* of covariate values corresponding to some target population. In time series problems, predictive MSE is typically more interesting than estimator MSE. Accordingly, Claeskens et al. (2007) develop an FIC to minimize forecast MSE in autoregressive models where the true order of the process is infinite. Independently of the FIC literature, Schorfheide (2005) likewise uses local mis-specification to suggest a procedure for using finite order vector autoregressions to forecast an infinite-order vector moving average process with minimum quadratic loss. This idea shares similarities with Skouras (2001).

Like the FIC and related proposals, the GFIC uses local mis-specification to derive a risk-based selection criterion. Unlike them, however, the GFIC provides both moment and model selection for general GMM estimators. The focused moment selection criterion (FMSC) of DiTraglia (2012) represents a special case of

the GFIC in which model specification is fixed and selection carried out over moment conditions only. Thus, the GFIC extends both the FIC and the FMSC. Comparatively few papers propose criteria for simultaneous GMM model and moment selection under mis-specification.[1] Andrews and Lu (2001) propose a family of selection criteria by adding appropriate penalty and "bonus" terms to the J-test statistic, yielding analogues of AIC, BIC, and the Hannan-Quinn information criterion. Hong et al. (2003) extend this idea to generalized empirical likelihood (GEL). The principal goal of both papers is consistent selection: they state conditions under which the correct model and all correct moment conditions are chosen in the limit. As a refinement to this approach, Lai et al. (2008) suggest a two-step procedure: first consistently eliminate incorrect models using an empirical log-likelihood ratio criterion, and then select from the remaining models using a bootstrap covariance matrix estimator. The point of the second step is to address a shortcoming in the standard limit theory. While first-order asymptotic efficiency requires that we use all available correctly specified moment conditions, this can lead to a deterioration in finite sample performance if some conditions are only weakly informative. Hall and Peixe (2003) make a similar point about the dangers of including "redundant" moment conditions while Caner (2009) proposes a lasso-type GMM estimator to consistently remove redundant parameters.

In contrast to these suggestions, the GFIC does not aim to identify the correct model and moment conditions: its goal is a low MSE estimate of a quantity of interest, even if this entails using a specification that is not exactly correct. Although their combined moments (CM) estimator is not strictly a selection procedure, Judge and Mittelhammer (2007) take a similar perspective, emphasizing that incoporating the information from an incorrect specification could lead to favorable bias-variance tradeoff under the right circumstances. Their proposal uses a Cressie-Read divergence measure to combine the information from competing moment specifications, for example OLS versus two-stage least squares (2SLS), yielding a data-driven compromise estimator. Unlike the GFIC, however, the CM

[1] See Smith (1992) for an approach to GMM model selection based on non-nested hypothesis testing.

estimator is not targeted to a particular research goal.

The remainder of this paper is organized as follows. Section 2 derives the asymptotic distribution of GMM estimators under locally mis-specified moment conditions and parameter restrictions. Section 3 uses this information to calculate the AMSE of a user-specified target parameter and provides asymptotically unbiased estimators of the required bias parameters, yielding the GFIC. Section 5 specializes the GFIC to a dynamic panel example, and explores its performance in a simulation study. Section 7 concludes. Proofs appear in the Appendix.

2 Notation and Asymptotic Framework

Let $f(\cdot,\cdot)$ be a $(p+q)$-vector of moment functions of a random vector Z and an $(r+s)$-dimensional parameter vector β. To represent moment selection, we partition the moment functions according to $f(\cdot,\cdot) = (g(\cdot,\cdot)', h(\cdot,\cdot)')'$ where $g(\cdot,\cdot)$ and $h(\cdot,\cdot)$ are p- and q-vectors. The moment condition associated with $g(\cdot,\cdot)$ is assumed to be correct, while that associated with $h(\cdot,\cdot)$ is locally mis-specified. The moment selection problem is to choose which, if any, of the elements of h to use in estimation. To represent model selection, we partition the full parameter vector according to $\beta = (\gamma', \theta')'$, where γ is an r-vector and θ an s-vector of parameters. The model selection problem is to decide which if any of the elements of γ to estimate, and which to set equal to the corresponding elements of γ_0, an r-vector of known constants. The parameters contained in θ are those that we always estimate, the "protected" parameters. Any specification that does not estimate the full parameter vector β is locally mis-specified. The precise form of the local mis-specification, over parameter restrictions and moment conditions, is as follows.

Assumption 2.1 (Local Mis-specification)**.** *Let* $\{Z_{ni}\colon 1 \leq i \leq n, n = 1, 2, \ldots\}$ *be a triangular array of random vectors defined on a probability space* $(\Upsilon, \mathcal{F}, \mathbb{P})$ *satisfying*

(a) $\mathbb{E}[g(Z_{ni}, \gamma_n, \theta_0)] = 0$

(b) $\mathbb{E}[h\left(Z_{ni}, \gamma_n, \theta_0\right)] = \tau_n$

(c) $\{f(Z_{ni}, \gamma_n, \theta_0) \colon 1 \leq i \leq n, n = 1, 2, \ldots\}$ *is uniformly integrable, and*

(d) $Z_{ni} \to_d Z_i$*, where the* Z_i *are identically distributed.*

where $\gamma_n = \gamma_0 + n^{-1/2}\delta$ *with* δ *an unknown* r*-vector of constants and* $\tau_n = n^{-1/2}\tau$ *with* τ *an unknown* q*-vector of constants.*

Under Assumption 2.1, the true parameter vector $\beta_n = (\gamma_n', \theta_0')'$, changes with sample size but converges to $\beta_0 = (\gamma_0', \theta_0')'$ as $n \to \infty$. Unless some elements of δ are zero, any estimator that restricts γ is mis-specified for fixed n. In the limit, however, the restriction $\gamma = \gamma_0$ holds. Similarly, for any fixed sample size n, the expectation of h evaluated at the true parameter value β_n depends on the unknown constant vector τ, but this source of mis-specification disappears in the limit. Thus, under Assumption 2.1, only estimators that use moment conditions from g to estimate the full parameter vector β are correctly specified. In the limit, however, *every* estimator is correctly specified, regardless of which elements of γ it restricts and which elements of h it includes. The purpose of local mis-specification is to ensure that squared asymptotic bias is of the same order as asymptotic variance: Assumption 2.1 is a device rather than literal description of real-world data. To simplify the proofs we make the following further assumption concerning the triangular array Assumption 2.1, although it is not strictly necessary.

Assumption 2.2. $\{Z_{ni} \colon 1 \leq i \leq n, n = 1, 2, \ldots\}$ *is iid over* i *for fixed* n*.*

Note that, by Assumptions 2.1–2.2, the limiting random variable Z_i satisfies the population moment condition $\mathbb{E}[f\left(Z_i, \gamma_0, \theta_0\right))] = 0$. Since the Z_i are assumed to have a common marginal law, we will use the shorthand Z for Z_i throughout.

Before defining the estimators under consideration, we require some further notation. Let b be a *model selection vector*, an r-vector of ones and zeros indicating which elements of γ we have chosen to estimate. When $b = 1_r$, where 1_m represents an m-vector of ones, we estimate both θ and the full vector γ. When $b = 0_r$, where 0_m denotes an m-vector of zeros, we estimate only θ, setting $\gamma = \gamma_0$.

More generally, we estimate $|b|$ components of γ and set the others equal to the corresponding elements of γ_0. Let $\gamma^{(b)}$ be the $|b|$-dimensional subvector of γ corresponding to those elements selected for estimation. Similarly, let $\gamma_0^{(-b)}$ denote the $(r-|b|)$-dimensional subvector containing the values to which we set those components of γ that are *not* estimated. Analogously, let $c = \left(c_g', c_h'\right)'$ be a *moment selection vector*, a $(p+q)$-vector of ones and zeros indicating which of the moment conditions we have chosen to use in estimation. We denote by $|c|$ the total number of moment conditions used in estimation. Let $\mathcal{BC}$ denote the collection of all model and moment selection pairs (b, c) under consideration.

To express moment and model selection in matrix form, we define the selection matrices Ξ_b and Ξ_c. Multiplying β by the $(|b|+s)\times(r+s)$ *model selection matrix* Ξ_b extracts the elements corresponding to θ and the subset of γ indicated by the model selection vector b. Thus $\Xi_b\beta = \left(\gamma^{(b)'}, \theta'\right)'$. Similarly, multiplying a vector by the $|c|\times(p+q)$ moment selection matrix Ξ_c extracts the components corresponding to the moment conditions indicated by the moment selection vector c.

To express the estimators themselves, define the sample analogue of the expectations in Assumption 2.1 as follows,

$$f_n(\beta) = \frac{1}{n}\sum_{i=1}^{n} f(Z_{ni}, \gamma, \theta) = \begin{bmatrix} g_n(\beta) \\ h_n(\beta) \end{bmatrix} = \begin{bmatrix} n^{-1}\sum_{i=1}^{n} g(Z_{ni}, \gamma, \theta) \\ n^{-1}\sum_{i=1}^{n} h(Z_{ni}, \gamma, \theta) \end{bmatrix} \tag{2.1}$$

and let $\widetilde{W}$ be a $(q+p)\times(q+p)$ positive semi-definite weighting matrix

$$\widetilde{W} = \begin{bmatrix} \widetilde{W}_{gg} & \widetilde{W}_{gh} \\ \widetilde{W}_{hg} & \widetilde{W}_{hh} \end{bmatrix} \tag{2.2}$$

partitioned conformably to the partition of $f(Z,\beta)$ by $g(Z,\beta)$ and $h(Z,\beta)$. Each model and moment selection pair $(b,c) \in \mathcal{BC}$ defines a $(|b|+s)$-dimensional estimator $\widehat{\beta}(b,c) = (\widehat{\gamma}^{(b)}(b,c)', \widehat{\theta}(b,c)')'$ of $\beta^{(b)} = \left(\gamma^{(b)'}, \theta'\right)'$ according to

$$\widehat{\beta}(b,c) = \underset{\beta^{(b)}\in\mathbf{B}^{(b)}}{\arg\min} \left[\Xi_c f_n\left(\beta^{(b)}, \gamma_0^{(-b)}\right)\right]' \left[\Xi_c \widetilde{W} \Xi_c'\right] \left[\Xi_c f_n\left(\beta^{(b)}, \gamma_0^{(-b)}\right)\right]. \tag{2.3}$$

A particularly important special case is the estimator using only the moment conditions in g to estimate the full parameter vector $\beta = (\theta', \gamma')'$, the *valid* estimator:

$$\widehat{\beta}_v = \begin{bmatrix} \widehat{\gamma}_v \\ \widehat{\theta}_v \end{bmatrix} = \underset{\beta \in \mathbf{B}}{\arg\min}\ g_n(\beta)' \widetilde{W}_{gg}\ g_n(\beta). \tag{2.4}$$

Because it is correctly specified both for finite n and in the limit, the valid estimator contains the information we use to identify τ and δ, and thus carry out moment and model selection. For estimation based on g alone to be possible, we require $p \geq r + s$. This is assumed throughout.

Because Assumption 2.1 ensures that they are correctly specified in the limit, *all* candidate specifications $(b, c) \in \mathcal{BC}$ provide consistent estimators of θ_0 under standard, high level regularity conditions.[2] Essential differences arise, however, when we consider their respective asymptotic distributions. Let

$$F = \begin{bmatrix} \nabla_{\gamma'} g\left(Z, \gamma_0, \theta_0\right) & \nabla_{\theta'} g\left(Z, \gamma_0, \theta_0\right) \\ \nabla_{\gamma'} h\left(Z, \gamma_0, \theta_0\right) & \nabla_{\theta'} h\left(Z, \gamma_0, \theta_0\right) \end{bmatrix} \tag{2.5}$$

partitioned according to

$$F = \begin{bmatrix} F_\gamma & F_\theta \end{bmatrix} = \begin{bmatrix} G_\gamma & G_\theta \\ H_\gamma & H_\theta \end{bmatrix} = \begin{bmatrix} G \\ H \end{bmatrix} \tag{2.6}$$

and define

$$\Omega = Var \begin{bmatrix} g(Z, \gamma_0, \theta_0) \\ h(Z, \gamma_0, \theta_0) \end{bmatrix} = \begin{bmatrix} \Omega_{gg} & \Omega_{gh} \\ \Omega_{hg} & \Omega_{hh} \end{bmatrix}. \tag{2.7}$$

Notice that each of these expressions involves the limiting random variable Z rather than Z_{ni}. Thus, the corresponding expectations are taken with respect to a distribution for which all moment conditions have expectation zero evaluated at (γ_0, θ_0). Finally, let $F(b, c) = \Xi_c F \Xi_b'$ and similarly define $\Omega_c = \Xi_c \Omega \Xi_c'$ and $W_c = \Xi_c W \Xi_c'$ where W is the positive definite probability limit of $\widetilde{W}$. Under

[2]The required high-level sufficient conditions are essentially identical to Assumption 2.2 of DiTraglia (2012).

Assumption 2.1, both δ and τ induce a bias term in the limiting distribution of $\sqrt{n}\left(\widehat{\beta}(b,c) - \beta_0^{(b)}\right)$. The key results is as follows.

Theorem 2.1 (Asymptotic Distribution)**.** *Under Assumptions 2.1–2.2 and standard regularity conditions,*

$$\sqrt{n}\left(\widehat{\beta}(b,c) - \beta_0^{(b)}\right) \to_d -K(b,c)\Xi_c\left(\mathscr{N} + \begin{bmatrix} 0 \\ \tau \end{bmatrix} - F_\gamma\delta\right) \tag{2.8}$$

where $\beta_0^{(b)'} = (\theta_0, \gamma_0^{(b)})$,

$$K(b,c) = \left[F(b,c)'W_cF(b,c)\right]^{-1}F(b,c)'W_c \tag{2.9}$$

and

$$\mathscr{N} = \begin{bmatrix} \mathscr{N}_g \\ \mathscr{N}_h \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \Omega_{gg} & \Omega_{gh} \\ \Omega_{hg} & \Omega_{hh} \end{bmatrix}\right). \tag{2.10}$$

Because it employs the correct specification, the valid estimator of θ shows no asymptotic bias. Moreover, the valid estimator of γ has an asymptotic distribution that is centered around δ, suggesting an estimator of this bias parameter.

Corollary 2.1 (Asymptotic Distribution of Valid Estimator)**.** *Under Assumptions 2.1–2.2 and standard regularity conditions,*

$$\sqrt{n}\left(\widehat{\beta}_v - \beta_0\right) = \sqrt{n}\begin{pmatrix} \widehat{\theta}_v - \theta_0 \\ \widehat{\gamma}_v - \gamma_0 \end{pmatrix} \to_d \begin{bmatrix} 0 \\ \delta \end{bmatrix} - K_v\mathscr{N}_g$$

where $K_v = \left[G'W_{gg}G\right]^{-1}G'W_{gg}$ *and* $W_{gg} = plim_{n\to\infty}\widetilde{W}_{gg}$.

We use these results in the following section to construct the GFIC.

3 The GFIC

The GFIC chooses among potentially incorrect moment conditions and parameter restrictions to minimize estimator AMSE for a scalar target parameter. Denote

this target parameter by $\mu = \varphi(\gamma, \theta)$, where φ is a real-valued, almost surely continuous function of the underlying model parameters θ and γ. Let $\mu_n = \varphi(\gamma_n, \theta_0)$ and define μ_0 and $\widehat{\mu}(b, c)$ analogously. By Theorem 2.1 and the delta method, we have the following result.

Corollary 3.1. *Under the hypotheses of Theorem 2.1,*

$$\sqrt{n}\left(\widehat{\mu}(b,c) - \mu_0\right) \to_d -\nabla_\beta \varphi_0' \Xi_b' K(b,c) \Xi_c \left(\mathscr{N} + \begin{bmatrix} 0 \\ \tau \end{bmatrix} - F_\gamma \delta \right)$$

where $\varphi_0 = \varphi(\gamma_0, \theta_0)$.

The true value of μ, however, is μ_n rather than μ_0 under Assumption 2.1. Accordingly, to calculate AMSE we recenter the limit distribution as follows.

Corollary 3.2. *Under the hypotheses of Theorem 2.1,*

$$\sqrt{n}\left(\widehat{\mu}(b,c) - \mu_n\right) \to_d -\nabla_\beta \varphi_0' \Xi_b' K(b,c) \Xi_c \left(\mathscr{N} + \begin{bmatrix} 0 \\ \tau \end{bmatrix} - F_\gamma \delta \right) - \nabla_\gamma \varphi_0' \delta$$

where $\varphi_0 = \varphi(\gamma_0, \theta_0)$.

We see that the limiting distribution of $\widehat{\mu}(b, c)$ is not, in general, centered around zero: both τ and δ induce an asymptotic bias. Note that, while τ enters the limit distribution only once, δ has two distinct effects. First, like τ, it shifts the limit distribution of $\sqrt{n} f_n(\gamma_0, \theta_0)$ away from zero, thereby influencing the asymptotic behavior of $\sqrt{n}\left(\widehat{\mu}(b, c) - \mu_0\right)$. Second, unless the derivative of φ with respect to γ is zero at (γ_0, θ_0), δ induces a second source of bias when $\widehat{\mu}(b, c)$ is recentered around μ_n. Crucially, this second source of bias exactly cancels the asymptotic bias present in the limit distribution of $\widehat{\gamma}_v$. Thus, the valid estimator of μ is asymptotically unbiased and its AMSE equals its asymptotic variance.

Corollary 3.3. *Under the hypotheses of Theorem 2.1,*

$$\sqrt{n}\left(\widehat{\mu}_v - \mu_n\right) \to_d -\nabla_\beta \varphi(\theta_0, \gamma_0)' K_v \mathscr{N}_g$$

where $\widehat{\mu}_v = \varphi(\widehat{\theta}_v, \widehat{\gamma}_v)$*. Thus, the valid estimator* $\widehat{\mu}_v$ *shows no asymptotic bias and has asymptotic variance* $\nabla_\beta \varphi(\theta_0, \gamma_0)' K_v \Omega_{gg} K_v' \nabla_\beta \varphi(\theta_0, \gamma_0)$*.*

Using Corollary 3.2, the AMSE of $\widehat{\mu}(b,c)$ is as follows,

$$\text{AMSE}\left(\widehat{\mu}(b,c)\right) = \text{AVAR}\left(\widehat{\mu}\,(b,c)\right) + \text{BIAS}\left(\widehat{\mu}\,(b,c)\right)^2 \tag{3.1}$$

where

$$\begin{aligned} \text{AVAR}\left(\widehat{\mu}\,(b,c)\right) &= \nabla_\beta \varphi_0' \Xi_b' K(b,c) \Omega_c K(b,c)' \Xi_b \nabla_\beta \varphi_0 & (3.2) \\ \text{BIAS}\left(\widehat{\mu}\,(b,c)\right) &= -\nabla_\beta \phi_0' M(b,c) \begin{bmatrix} \delta \\ \tau \end{bmatrix} & (3.3) \end{aligned}$$

and

$$M(b,c) = \Xi_b' K(b,c) \Xi_c \begin{bmatrix} -G_\gamma & 0 \\ -H_\gamma & I \end{bmatrix} + \begin{bmatrix} I_r & 0_{r\times q} \\ 0_{p\times r} & 0_{s\times q} \end{bmatrix} \tag{3.4}$$

The idea behind the GFIC is to construct an estimate $\widehat{\text{AMSE}}\left(\widehat{\mu}(b,c)\right)$ and choose the specification $(b^*, c^*) \in \mathcal{BC}$ that makes this quantity as small as possible. As a side-effect of the consistency of the estimators $\widehat{\beta}(b,c)$, the usual sample analogues provide consistent estimators of $K(b,c)$ and $F_\gamma' = (G_\gamma', H_\gamma')$ under Assumption 2.1, and $\varphi(\widehat{\theta}_v, \gamma_0)$ is consistent for φ_0. Consistent estimators of Ω are also readily available under local mis-specification although the best choice may depend on the situation.[3] Since γ_0 is known, as are Ξ_b and Ξ_c, only δ and τ remain to be estimated. Unfortunately, neither of these quantities is consistently estimable under local mis-specification. Intuitively, the data become less and less informative about τ and δ as the sample size increases since each term is divided by $\sqrt{n}$. Multiplying through by $\sqrt{n}$ counteracts this effect, but also stabilizes the variance of our estimators. Hence, the best we can do is to construct *asymptotically unbiased* estimators of τ and δ. Corollary 2.1 provides the required estimator for δ, namely $\widehat{\delta} = \sqrt{n}\left(\widehat{\gamma}_v - \gamma_0\right)$.

[3]For more on this point, see DiTraglia (2012) Section 3.3.

Corollary 3.4 (Asymptotically Unbiased Estimator of δ)**.** *Under the hypotheses of Theorem 2.1,*

$$\widehat{\delta} = \sqrt{n}\left(\widehat{\gamma}_v - \gamma_0\right) \to_d \delta - K_v^{\gamma}\mathscr{N}_g$$

where $K_v = \left[G'W_{gg}G\right]^{-1}G'W_{gg} = \left(K_v^{\gamma\prime}, K_v^{\theta\prime}\right)'$*. Hence,* $\widehat{\delta}$ *is an asymptotically unbiased estimator of* δ*.*

To estimate τ, we simply plug $\widehat{\beta}_v$ into the locally mis-specified moment conditions contained in h.

Lemma 3.1 (Asymptotically Unbiased Estimator of τ)**.** *Under the hypotheses of Theorem 2.1,*

$$\widehat{\tau} = \sqrt{n}h_n\left(\widehat{\beta}_v\right) \to_d \tau - HK_v\mathscr{N}_g + \mathscr{N}_h$$

where, as above, $K_v = \left[G'W_{gg}G\right]^{-1}G'W_{gg}$*. Hence,* $\widehat{\tau}$ *is an asymptotically unbiased estimator of* τ*.*

Combining Corollary 3.4 and Lemma 3.1, we can express the joint distribution of $\widehat{\delta}$ and $\widehat{\tau}$ as follows.

Theorem 3.1. *Under the hypotheses of Theorem 2.1,*

$$\begin{bmatrix} \widehat{\delta} \\ \widehat{\tau} \end{bmatrix} = \sqrt{n}\begin{bmatrix} (\widehat{\gamma}_v - \gamma_0) \\ h_n(\widehat{\beta}_v) \end{bmatrix} \to_d \begin{bmatrix} \delta \\ \tau \end{bmatrix} + \Psi\mathscr{N}.$$

where

$$\Psi = \begin{bmatrix} -K_v^{\gamma} & \mathbf{0} \\ -HK_v & I \end{bmatrix}$$

and K_v *is partitioned according to* $K_v' = (K_v^{\gamma\prime}, K_v^{\theta\prime})$

From Equation 3.3,

$$\text{BIAS}\left(\widehat{\mu}\left(b,c\right)\right)^2 = \nabla_{\beta}\varphi_0' M(b,c)\begin{bmatrix} \tau\tau' & \tau\delta' \\ \delta\tau' & \delta\delta' \end{bmatrix} M(b,c)'\nabla_{\beta}\varphi_0$$

Thus, the bias parameters τ and δ enter the AMSE expression in Equation 3.1 as outer products: $\tau\tau'$, $\delta\delta'$ and $\tau\delta'$. Although $\widehat{\tau}$ and $\widehat{\delta}$ are asymptotically unbiased

estimators of τ and δ, it does *not* follow that $\widehat{\tau}\widehat{\tau}$, $\widehat{\delta}\widehat{\delta}$ and $\widehat{\tau}\widehat{\delta}'$ are asymptotically unbiased estimators of $\tau\tau'$, $\delta\delta'$, and $\tau\delta'$. The following result shows how to adjust these quantities to provide the required asymptotically unbiased estimates.

Corollary 3.5. *Suppose that $\widehat{\Psi}$ and $\widehat{\Omega}$ are consistent estimators of Ψ and Ω. Then,*

$$\widehat{B} = \begin{bmatrix} \widehat{\tau}\widehat{\tau}' & \widehat{\tau}\widehat{\delta}' \\ \widehat{\delta}\widehat{\tau}' & \widehat{\delta}\widehat{\delta}' \end{bmatrix} - \widehat{\Psi}\widehat{\Omega}\widehat{\Psi}' \tag{3.5}$$

is an asymptotically unbiased estimator of the squared bias matrix

$$\begin{bmatrix} \tau\tau' & \tau\delta' \\ \delta\tau' & \delta\delta' \end{bmatrix}.$$

Combining Corollary 3.5 with consistent estimates of the remaining quantities yields the GFIC, an asymptotically unbiased estimator of the AMSE of our estimator of a target parameter μ under each specification $(b, c) \in \mathcal{BC}$

$$\text{GFIC}(b,c) = \nabla_\beta \widehat{\varphi}_0' \left[\Xi_b' \widehat{K}(b,c) \widehat{\Omega}_c \widehat{K}(b,c)' \Xi_b + \widehat{M}(b,c)\ \widehat{B}\ \widehat{M}(b,c)' \right] \nabla_\beta \widehat{\varphi}_0. \tag{3.6}$$

We choose the specification (b^*, c^*) that minimizes the value of the GFIC over the candidate set $\mathcal{BC}$.

4 Averaging and Post-Selection Inference

Assumption 4.1 (Data-Dependent Weights)**.** *Let $\widehat{\omega}(b,c)$ be a function of the data $Z_{n1}, \ldots, Z_{nn}$ and (b,c) satisfying*

(a) $\sum_{(b,c)\in\mathcal{BC}} \widehat{\omega}(b,c) = 1$

(b) $\widehat{\omega}(b,c) \to_d \psi(\mathcal{N}, \delta, \tau | b, c)$ *jointly for all $(b,c) \in \mathcal{BC}$ where ψ is a function of the normal random vector $\mathcal{N}$, the bias parameters δ and τ, and consistently estimable quantities only.*

Corollary 4.1 (Limit Distribution of Averaging Estimators)**.** *Let $\widehat{\omega}(b,c)$ be a set of weights satisfying Assumption 4.1 and define*

$$\widehat{\mu} = \sum_{(b,c)\in\mathcal{BC}} \widehat{\omega}(b,c)\widehat{\mu}(b,c).$$

Then, under the hypotheses of Theorem 2.1,

$$\sqrt{n}\,(\widehat{\mu} - \mu_n) \to_d \Lambda(\tau,\delta)$$

where

$$\Lambda(\tau,\delta) = -\nabla_\beta\varphi_0' \sum_{(b,c)\in\mathcal{BC}} \psi(\mathscr{N},\delta,\tau|b,c)\left\{\Xi_b' K(b,c)\Xi_c\mathscr{N} + M(b,c)\begin{bmatrix}\delta\\ \tau\end{bmatrix}\right\} \quad (4.1)$$

Algorithm 4.1 (Simulation-based Confidence Interval for $\widehat{\mu}$)**.**

1. Construct $R(\alpha_1)$, a $(1-\alpha_1)\times 100\%$ joint confidence region for (δ,τ)
2. For each $(\delta,\tau)\in R(\alpha_1)$:
   (i) For each $j = 1,2,\ldots,B$, generate $\mathscr{N}_j \sim N(0,\widehat{\Omega})$
   (ii) For each for $j = 1,2,\ldots,B$ set

$$\Lambda_j(\tau,\delta) = -\nabla_\beta\widehat{\varphi}_0' \sum_{(b,c)\in\mathcal{BC}} \widehat{\psi}(\mathscr{N}_j,\delta,\tau|b,c)\left\{\Xi_b'\widehat{K}(b,c)\Xi_c\mathscr{N}_j + \widehat{M}(b,c)\begin{bmatrix}\delta\\ \tau\end{bmatrix}\right\}$$

   (iii) Using $\{\Lambda_j(\delta,\tau)\}_{j=1}^B$, calculate $\widehat{a}(\delta,\tau)$, $\widehat{b}(\delta,\tau)$ such that

$$\mathbb{P}\left\{\widehat{a}(\delta,\tau) \le \Lambda(\delta,\tau) \le \widehat{b}(\delta,\tau)\right\} = 1-\alpha_2$$

3. Define

$$\begin{aligned}\widehat{a}_{min}(\widehat{\delta},\widehat{\tau}) &= \min_{(\delta,\tau)\in R(\alpha_1)} \widehat{a}(\delta,\tau)\\ \widehat{b}_{max}(\widehat{\delta},\widehat{\tau}) &= \max_{(\delta,\tau)\in R(\alpha_1)} \widehat{b}(\delta,\tau)\end{aligned}$$

4. The confidence interval for is μ given by

$$\text{CI}_{sim} = \left[\widehat{\mu} - \frac{\widehat{b}_{max}(\widehat{\delta}, \widehat{\tau})}{\sqrt{n}}, \quad \widehat{\mu} - \frac{\widehat{a}_{min}(\widehat{\delta}, \widehat{\tau})}{\sqrt{n}}\right]$$

5 Dynamic Panel Example

I now specialize the GFIC to a dynamic panel model with unobserved individual effects similar to the example from Andrews and Lu (2001) Section 6. For simplicity, and to avoid weak and many instruments problems, I consider 2SLS estimators similar to those suggested by Anderson and Hsiao (1982).[4] To keep the presentation transparent, I likewise restrict attention to a model with only one regressor besides the lagged dependent variable and no constant terms. Introducing additional regressors merely complicates the notation.

5.1 Models and Moment Conditions

Our aim is to estimate θ, the effect of a regressor x_{it} on an outcome y_{it}, with minimum MSE. The true data generating process is

$$y_{it} = \gamma y_{it-1} + \theta x_{it} + u_{it} \tag{5.1}$$

where $i = 1, \ldots, n$ indexes individuals and $t = 1, \ldots, T$ indexes time periods. We assume stationarity of x_{it} and u_{it} and $|\gamma| < 1$ so that y_{it} is stationary. The error term u_{it} follows a one-way error components model

$$u_{it} = \eta_i + v_{it} \tag{5.2}$$

with idiosyncratic component v_{it} and individual effect η_i. The individual effect η_i is correlated with x_{it} according to $\mathbb{E}[x_i t \eta_i] = \sigma_{x\eta}$. Under the true DGP, x_{it} is predetermined but may not be strictly exogenous. That is, $\mathbb{E}[x_{it} v_{is}] = 0$ for all

[4] Although a system-GMM approach is asymptotically more efficient, it can lead to serious finite sample problems. See, for example, Roodman (2009).

$s \geq t$ but $\mathbb{E}[x_{it}v_{is}]$ may be nonzero for $s < t$. To remove the correlated individual effects, we take first differences, yielding

$$\Delta y_{it} = \gamma \Delta y_{it-1} + \theta \Delta x_{it} + \Delta v_{it}. \tag{5.3}$$

Under the true data generating process, x_{it-1} and y_{it-2} are both valid instruments for period t. Although x_{it-1} is a strong instrument, using both x_{it-1} and x_{it} to instrument for Δx_{it} would be far more efficient. Unless $\mathbb{E}[x_{it}v_{it-1}] = 0$, however, x_{it} is correlated with Δv_{it}, and including it will bias our estimates. Yet if σ_{xv} is *nearly* zero, this bias may be small relative to the reduction in variance that including x_{it} provides. Our moment selection decision is whether or not to use x_{it} as an instrument for period t.

Because we observe only $t = 1, \ldots, T$, estimation in differences with a lagged dependent variable uses information from $T - 2$ time periods: $t = 3, \ldots, T$. In contrast, estimation without a lagged dependent variable uses information from $T - 1$ time periods: $t = 2, \ldots, T$. When T is small, as in many micro-data applications, including an unnecessary lagged dependent variable could result in a huge loss in information, substantially increasing the variance of our estimate of θ. On the other hand, unless γ is zero, failing to include a lagged dependent variable will bias our estimates. If γ is *nearly* zero, however, this bias may be small compared to the reduction in variance achieved by using an additional time period and estimating one fewer parameter. Our model selection decision is whether or not to set $\gamma = 0$.

Taking these considerations together, we consider four specifications: LW, LS, W, and S. Both LW and LS include a lagged dependent variable – hence the designation "L" – while W and S do not. LW and W assume only that x_{it} is predetermined – hence the designation "W" for "weak exogeneity assumption" – while LS and S impose the stronger assumption of *strict* exogeneity. Thus, LW and LS estimate the correct model while LW and W use the correct instrument sets. The correct specification is LW.

Estimation based on LW uses the $2(T-2)$ moment conditions

$$\mathbb{E}\left[\begin{pmatrix} y_{i,t-2} \\ x_{i,t-1} \end{pmatrix} (\Delta y_{it} - \gamma \Delta y_{i,t-1} - \theta \Delta x_{it})\right] = 0, \text{ for } t = 3, \ldots, T \tag{5.4}$$

to which LS adds

$$\mathbb{E}\left[x_{it} (\Delta y_{it} - \gamma \Delta y_{i,t-1} - \theta \Delta x_{it})\right] = 0, \text{ for } t = 3, \ldots, T \tag{5.5}$$

for a total of $3(T-2)$ moment conditions. The additional $T-2$ conditions in Equation 5.5, however, may be incorrect: $\mathbb{E}[x_{it}\Delta v_{it}] = -\mathbb{E}[x_{it}v_{it-1}]$ since x_{it} is only predetermined. Since it is the only violation of strict exogeneity that is relevant for the specifications under consideration, we let $\mathbb{E}[x_{it}v_{it-1}] = \sigma_{xv}$. When $\sigma_{xv} \neq 0$, the moment conditions in Equation 5.5 are mis-specified.

Estimation based on specification W uses the $T-1$ moment conditions

$$\mathbb{E}\left[x_{i,t-1} (\Delta y_{it} - \theta \Delta x_{it})\right] = 0, \text{ for } t = 2, \ldots, T \tag{5.6}$$

to which specification S adds a further $T-1$ moment conditions, namely

$$\mathbb{E}\left[x_{it} (\Delta y_{it} - \theta \Delta x_{it})\right] = 0, \text{ for } t = 2, \ldots, T \tag{5.7}$$

for a total of $2(T-1)$ conditions. Because specifications W and S use the wrong model, however, these moment conditions are mis-specified:

$$\mathbb{E}\left[\begin{pmatrix} x_{i,t-1} \\ x_{it} \end{pmatrix} (\Delta y_{it} - \theta \Delta x_{it})\right] = \begin{bmatrix} \gamma \mathbb{E}[x_{it-1}\Delta y_{it-1}] \\ \gamma \mathbb{E}[x_{it}\Delta y_{it-1}] - \sigma_{xv} \end{bmatrix} \tag{5.8}$$

which are non-zero unless $\sigma_{xv} = \gamma = 0$.

5.2 Estimators and Local Mis-specification

Our aim is to use the GFIC to choose between competing estimators of θ on the basis of AMSE. To do so we must first specify the appropriate form of local mis-

specification by analogy with Assumption 2.1. In this example, the parameters γ and σ_{xv} control the degree of mis-specification present in LS, W and S. When $\gamma = 0$, both models, with and without a lag, are correctly specified; when $\sigma_{xv} = 0$ all instruments under consideration are valid. Accordingly, we let $\gamma = \delta/\sqrt{n}$ and $-\sigma_{xv} = \tau/\sqrt{n}$ so that, in the limit, all four specifications are correct. In this framework the true parameter vector is $\beta_n = (\delta/\sqrt{n}, \theta_0)'$ which converges to $\beta_0 = (0, \theta_0)'$.

Assumption 5.1 (Local Mis-specification for Dynamic Panel Example)**.** *Assume that $\gamma = \delta/\sqrt{n}$ and $-\sigma_{xv} = \tau/\sqrt{n}$ where δ and τ are unknown constants.*

To define the estimators corresponding to specifications LW, LS, W and S we first require some additional notation. The symbol "+" used as a superscript indicates the inclusion of the extra time period $t = 2$ that becomes available when we exclude the lagged dependent variable. Using this convention, let $\Delta y_i = (\Delta y_{i3}, \ldots, \Delta y_{iT})'$ and $\Delta y_i^+ = (\Delta y_{i2}, \ldots, \Delta y_{iT})'$. Define $\Delta x_i, \Delta x_i^+$ and $\Delta v_i, \Delta v_i^+$ analogously. Similarly, let $\Delta y_{i,-1} = (\Delta y_{i2}, \ldots, \Delta y_{i,T-1})'$ and $\Delta y_{i,-1}^+ = (\Delta y_{i1}, \ldots, \Delta y_{i,T-1})'$. Note that the first element of $\Delta y_{i,-1}^+$ is not observed as $t = 1$ is the first available time period. Stacking over individuals in the usual way, define $\Delta y = (\Delta y_1', \ldots, \Delta y_n')'$ and so on.

The specifications LW and LS share the same model, and hence a design matrix. We denote this as:

$$X_L = \begin{bmatrix} \Delta y_{-1} & \Delta x \end{bmatrix} \tag{5.9}$$

where the subscript L indicates that both of these specifications include a lagged dependent variable. Similarly, let

$$X_L^+ = \begin{bmatrix} \Delta y_{-1}^+ & \Delta x^+ \end{bmatrix}. \tag{5.10}$$

Although X_L^+ is not observed, we use it in the derivations that follow as it allows

us to represent the true data generating process in matrix form. Specifically,

$$\Delta y = X_L \beta_n + \Delta v \tag{5.11}$$

$$\Delta y^+ = X_L^+ \beta_n + \Delta v^+ \tag{5.12}$$

We now turn our attention to the instrument matrices. For ease of notation, define the $(T-k+1) \times 1$ column vector

$$\{z_t\}_{t=k}^T = (z_k, z_{k+1}, \ldots, z_{T-1}, z_T)' \tag{5.13}$$

and the $(T-k+1) \times (T-k+1)$ diagonal matrix

$$D\{z_t\}_{t=k}^T = \begin{bmatrix} z_k & & 0 \\ & \ddots & \\ 0 & & z_T \end{bmatrix}. \tag{5.14}$$

To construct the instrument matrices, first define the $(T-2) \times (T-2)$ submatrices

$$Z(y_{i,-2}) = D\{y_{i,t-2}\}_{t=3}^T \tag{5.15}$$

$$Z(x_{i,-1}) = D\{x_{i,t-1}\}_{t=3}^T \tag{5.16}$$

$$Z(x_i) = D\{x_{it}\}_{t=3}^T \tag{5.17}$$

and the $(T-1) \times (T-1)$ submatrices

$$Z(x_{i,-1}^+) = D\{x_{i,t-1}\}_{t=2}^T \tag{5.18}$$

$$Z(x_i^+) = D\{x_{it}\}_{t=2}^T. \tag{5.19}$$

As above, the symbol "+" used as a superscript indicates the addition of an

additional time period. Combining these, define

$$
\begin{aligned}
Z_{LS,i} &= (Z(y_{i,-2}), Z(x_{i,-1}), Z(x_i))' && (5.20)\\
Z_{LW,i} &= (Z(y_{i,-2}), Z(x_{i,-1}))' && (5.21)\\
Z_{S,i} &= \left(Z(x^+_{i,-1}), Z(x^+_i)\right)' && (5.22)\\
Z_{W,i} &= Z(x^+_{i,-1}). && (5.23)
\end{aligned}
$$

Stacking over individuals, let $Z'_{LS} = (Z_{LS,1}, \ldots, Z_{LS,N})$ and so on. Finally, define the shorthand

$$
\widehat{K} = \left[\left(\frac{X'Z}{n}\right)\left(\frac{Z'Z}{n}\right)^{-1}\left(\frac{Z'X}{n}\right)\right]^{-1}\left(\frac{X'Z}{n}\right)\left(\frac{Z'Z}{n}\right)^{-1}. \tag{5.24}
$$

Using this notation, our four estimators are:

$$
\begin{aligned}
\widehat{\beta}_{LS} &= \widehat{K}_{LS}\left(\frac{Z'_{LS}\Delta y}{n}\right) && (5.25)\\
\widehat{\beta}_{LW} &= \widehat{K}_{LW}\left(\frac{Z'_{LW}\Delta y}{n}\right) && (5.26)\\
\widehat{\theta}_{S} &= \widehat{K}_{S}\left(\frac{Z'_{S}\Delta y^+}{n}\right) && (5.27)\\
\widehat{\theta}_{W} &= \widehat{K}_{W}\left(\frac{Z'_{W}\Delta y^+}{n}\right). && (5.28)
\end{aligned}
$$

which can be expanded as

$$
\begin{aligned}
\sqrt{n}(\widehat{\beta}_{LS} - \beta_0) = \sqrt{n}\begin{bmatrix}\widehat{\gamma}_{LS}\\ \widehat{\theta}_{LS} - \theta_0\end{bmatrix} &= \begin{bmatrix}\delta\\ 0\end{bmatrix} + \widehat{K}_{LS}\left(\frac{Z'_{LS}\Delta v}{n^{1/2}}\right) && (5.29)\\
\sqrt{n}(\widehat{\beta}_{LW} - \beta_0) = \sqrt{n}\begin{bmatrix}\widehat{\gamma}_{LW}\\ \widehat{\theta}_{LW} - \theta_0\end{bmatrix} &= \begin{bmatrix}\delta\\ 0\end{bmatrix} + \widehat{K}_{LW}\left(\frac{Z'_{LW}\Delta v}{n^{1/2}}\right) && (5.30)
\end{aligned}
$$

and

$$\sqrt{n}\left(\widehat{\theta}_S - \theta_0\right) = \widehat{K}_S\left[\delta\left(\frac{Z_S'\Delta y_{-1}^+}{n}\right) + \left(\frac{Z_S'\Delta v^+}{n^{1/2}}\right)\right] \tag{5.31}$$

$$\sqrt{n}\left(\widehat{\theta}_W - \theta_0\right) = \widehat{K}_W\left[\delta\left(\frac{Z_W'\Delta y_{-1}^+}{n}\right) + \left(\frac{Z_W'\Delta v^+}{n^{1/2}}\right)\right] \tag{5.32}$$

by substituting Equation 5.11. Combining these expressions with the Lindeberg-Feller central limit theorem and standard regularity conditions gives

$$\sqrt{n}\begin{bmatrix}\widehat{\gamma}_{LS}\\ \widehat{\theta}_{LS}-\theta_0\end{bmatrix} \to_d \begin{bmatrix}\delta\\ 0\end{bmatrix} + K_{LS}\left\{\begin{bmatrix}0_2\\ \tau\end{bmatrix}\otimes\iota_{T-2} + N\left(0,\mathcal{V}_{LS}\right)\right\} \tag{5.33}$$

$$\sqrt{n}\begin{bmatrix}\widehat{\gamma}_{LW}\\ \widehat{\theta}_{LW}-\theta_0\end{bmatrix} \to_d \begin{bmatrix}\delta\\ 0\end{bmatrix} + K_{LW}\ N\left(0,\mathcal{V}_{LW}\right) \tag{5.34}$$

and

$$\sqrt{n}\left(\widehat{\theta}_S - \theta_0\right) \to_d K_S\left[\left(\delta\begin{bmatrix}\psi_0\\ \psi_1\end{bmatrix} + \begin{bmatrix}0\\ \tau\end{bmatrix}\right)\otimes\iota_{T-1} + N\left(0,\mathcal{V}_S\right)\right] \tag{5.35}$$

$$\sqrt{n}\left(\widehat{\theta}_W - \theta_0\right) \to_d K_W\left[\delta\psi_0\otimes\iota_{T-1} + N\left(0,\mathcal{V}_W\right)\right]. \tag{5.36}$$

where K denotes the probability limit of $\widehat{K}$ (see Equation 5.24) and

$$\psi_0 = \mathbb{E}[x_{it}\Delta y_{it}] \tag{5.37}$$

$$\psi_1 = \mathbb{E}[x_{it}\Delta y_{it-1}] \tag{5.38}$$

with the expectations taken with respect to the limiting DGP, in which all four specifications are correct. These expressions immediately yield the AMSE of each estimator of θ. To implement the GFIC, we simply estimate the unknowns, as described below.

5.3 GFIC for the Dynamic Panel Example

To operationalize the GFIC, we need estimates of the unknowns in Equations 5.33–5.36. To estimate K_{LS}, K_{LW}, K_W and K_S we use $\widehat{K}_{LS}$, $\widehat{K}_{LW}$, $\widehat{K}_W$ and $\widehat{K}_S$, which remain consistent under local mis-specification. There are many consistent estimators of the variance matrices $\mathcal{V}_{LS}$, $\mathcal{V}_{LW}$, $\mathcal{V}_S$ and $\mathcal{V}_W$ under local mis-specification. For robustness, we use the centered, panel robust estimator that allows for heteroscedasticity. We do not center the estimator for LW because this specification is assumed correct, and this yields a more efficient estimator. Using the assumption of stationarity,

$$\begin{aligned}\widehat{\psi}_0 &= \frac{1}{n(T-1)}\sum_{t=2}^{T}\sum_{i=1}^{N} x_{it}\Delta y_{it} \\ \widehat{\psi}_1 &= \frac{1}{n(T-2)}\sum_{t=3}^{T}\sum_{i=1}^{N} x_{it}\Delta y_{it-1}\end{aligned}$$

provide consistent estimators of ψ_0 and ψ_1. The only remaining quantities needed to calculate the GFIC involve the bias parameters τ and δ. As described above, no consistent estimators of these quantities exist under local mis-specification. It remains possible, however, to construct asymptotically unbiased estimators. We can read off an asymptotically unbiased estimator of δ directly from Equation 5.34, namely $\widehat{\delta} = \sqrt{n}\ \widehat{\gamma}_{LW}$. To construct an asymptotically unbiased estimator of τ, we define $Z'(x) = (Z(x_1), \ldots, Z(x_n))$, see Equation 5.17, and expand the quantity $n^{-1/2}(\Delta y - X_L\widehat{\beta}_{LW})$ as follows:

$$n^{-1/2}(\Delta y - X_L\widehat{\beta}_{LW}) = \left[\begin{array}{cc} -n^{-1}Z'(x)X_L\widehat{K}_{LW} & I \end{array}\right] n^{-1/2}Z'_{LS}\Delta v$$

Now, by the Lindeberg-Feller Central Limit Theorem (c.f. Equation 5.33) we have

$$n^{-1/2}Z'_{LS}\Delta v \rightarrow_d \left[\begin{array}{c} 0_2 \\ \tau \end{array}\right] \otimes \iota_{T-2} + N\left(0, \mathcal{V}_{LS}\right)$$

and by a Law of Large Numbers,

$$n^{-1}Z'(x)X_L \to_p \mathbb{E}\left[\begin{array}{cc} x_{it}\Delta y_{it-1} & x_{it}\Delta x_{it} \end{array}\right] \otimes \iota_{T-2}$$

where the expectations are taken with respect to the limiting DGP. Thus,

$$n^{-1/2}(\Delta y - X_L\widehat{\beta}_{LW}) \to_d \tau \otimes \iota_{T-2} + \left[\begin{array}{cc} \Psi & I \end{array}\right] N(0, \mathcal{V}_{LS}) \tag{5.39}$$

where

$$\Psi = -\mathbb{E}\left[\begin{array}{cc} x_{it}\Delta y_{it-1} & x_{it}\Delta x_{it} \end{array}\right] \otimes \iota_{T-2}K_{LW} \tag{5.40}$$

Using stationarity to gain efficiency we take the time average

$$\widetilde{\tau} = \left(\frac{\iota'_{T-2}}{T-2}\right) n^{-1/2}Z'(x)(\Delta y - X_L\widehat{\beta}_{LW}) \tag{5.41}$$

as our estimator of τ. It follows from above that

$$\widetilde{\tau} \to_d \tau + \left(\frac{\iota'_{T-2}}{T-2}\right)\left[\begin{array}{cc} \Psi & I \end{array}\right] N(0, \mathcal{V}_{LS}) \tag{5.42}$$

As describe above for the general GMM case, asymptotically unbiased estimators of τ and δ require a bias correction to provide asymptotically unbiased estimators of the quantities τ^2, δ^2 and $\tau\delta$ needed to estimate AMSE. To carry out this correction, we use the joint distribution of the bias parameter estimators:

$$\left[\begin{array}{c} \widehat{\delta} \\ \widetilde{\tau} \end{array}\right] \to_d \left[\begin{array}{c} \widehat{\delta} \\ \widetilde{\tau} \end{array}\right] + \left[\begin{array}{cc} K^{\gamma}_{LW} & 0 \\ \left(\frac{\iota'_{T-2}}{T-2}\right)\Psi & \left(\frac{\iota'_{T-2}}{T-2}\right) I \end{array}\right] N(0, \mathcal{V}_{LS}) \tag{5.43}$$

where K^{γ}_{LW} denotes the first row of K_{LW} (i.e. the row corresponding to γ). Asymptotically unbiased estimators of δ^2, τ^2 and $\tau\delta$ are given by

$$\delta^2\colon \qquad \widehat{\delta}^2 - \widehat{\sigma}^2_{\delta} \tag{5.44}$$

$$\tau^2\colon \qquad \widetilde{\tau}^2 - \widehat{\sigma}^2_{\tau} \tag{5.45}$$

$$\tau\delta\colon \qquad \widetilde{\tau}\widehat{\delta} - \widehat{\sigma}_{\tau\delta} \tag{5.46}$$

where $\widehat{\sigma}_\delta^2$, $\widehat{\sigma}_\tau^2$ and $\widehat{\sigma}_{\tau\delta}$ are consistent estimators of the elements of

$$\begin{bmatrix} K_{LW}^{\gamma} & 0 \\ \left(\frac{\iota'_{T-2}}{T-2}\right)\Psi & \left(\frac{\iota'_{T-2}}{T-2}\right) I \end{bmatrix} \mathcal{V}_{LS} \begin{bmatrix} K_{LW}^{\gamma} & 0 \\ \left(\frac{\iota'_{T-2}}{T-2}\right)\Psi & \left(\frac{\iota'_{T-2}}{T-2}\right) I \end{bmatrix}'$$

We have already described how to consistently estimate K_{LW} and $\mathcal{V}_{LS}$ above, so the only quantities for which we still require consistent estimators are

$$\omega_1 = \mathbb{E}[x_{it}\Delta y_{it-1}] \tag{5.47}$$

$$\omega_2 = \mathbb{E}[x_{it}\Delta x_{it}] \tag{5.48}$$

which appear in the expression for Ψ. Under stationarity, the following estimators are consistent:

$$\widehat{\omega}_1 = \frac{1}{n(T-2)} \sum_{t=3}^{T} \sum_{i=1}^{n} x_{it}\Delta y_{it-1} \tag{5.49}$$

$$\widehat{\omega}_2 = \frac{1}{n(T-1)} \sum_{t=2}^{T} \sum_{i=1}^{n} x_{it}\Delta x_{it} \tag{5.50}$$

Substituting these estimators into the AMSE expressions implied by Equations 5.33–5.36 yields the GFIC.

6 Simulation Study

We now evaluate the performance of the GFIC in a simulation experiment based on the dynamic panel example from the preceding section. The details of this simulation are similar those of Andrews and Lu (2001).[5] The simulated covariates

[5]Unlike Andrews and Lu (2001) we do not generate "extra" presample observations for use with estimators that include a lagged dependent variable. This is for two reasons. First, in real-world applications such additional observations would not be available. Second, we are explicitly interested in how the loss of time periods for estimation affects finite sample MSE.

and error terms are jointly normal with mean zero and unit variance. Specifically,

$$\begin{bmatrix} x_i \\ \eta_i \\ v_i \end{bmatrix} \sim \text{iid } N\left(\begin{bmatrix} 0_T \\ 0 \\ 0_T \end{bmatrix}, \begin{bmatrix} I_T & \sigma_{x\eta}\iota_T & \sigma_{xv}\Gamma_T \\ \sigma_{x\eta}\iota_T' & 1 & 0_T' \\ \sigma_{xv}\Gamma_T' & 0_T & I_T \end{bmatrix} \right) \tag{6.1}$$

where 0_m denotes an m-vector of zeros, I_m the $(m \times m)$ identity matrix, ι_m an m-vector of ones, and Γ_m an $m \times m$ matrix with ones on the subdiagonal and zeros elsewhere, namely

$$\Gamma_m = \begin{bmatrix} 0_{m-1}' & 0 \\ I_{m-1} & 0_{m-1} \end{bmatrix}. \tag{6.2}$$

Under this covariance matrix structure, η_i and v_i are uncorrelated with each other, but both are correlated with x_i: $\mathbb{E}[x_{it}\eta_i] = \sigma_{x\eta}$ and x_{it} is predetermined but not strictly exogenous with respect to v_{it}. Specifically, $\mathbb{E}[x_{it}v_{it-1}] = \sigma_{xv}$, while $\mathbb{E}[x_{it}v_{is}] = 0$ for $s \neq t-1$.

We initialize the presample observations y_{i0} to zero, the mean of their stationary distribution, and generate the remaining time periods according to

$$y_{it} = \gamma y_{it-1} + \theta x_{it} + \eta_i + v_{it}$$

In the simulation we take $\theta = 0.5$, $\sigma_{x\eta} = 0.2$ and vary γ, σ_{xv}, T and N over a grid. Each grid point is based on 2000 simulation replications.

The first question is how the finite sample MSE of the 2SLS estimators of θ based on specifications LW, LS, W, and S (see Section 5) changes with γ and σ_{xv}. Figures 1 and 2 present RMSE comparisons for these four estimators over a simulation grid with $\gamma, \sigma_{xv} \in \{0, 0.005, 0.01, \ldots, 0.195, 0.2\}$, $N \in \{250, 500\}$, $T \in \{4, 5\}$.[6] For each point in the parameter space, the color in Figure 1 indicates the estimator of θ with the *lowest* finite sample RMSE. The saturation of the color indicates the relative difference in RMSE of the best estimator at that point measured against the second-best estimator: darker indicates a larger advantage;

[6]Taking T no smaller than 4 ensures that MSE exists for all four estimators: the finite sample moments of the 2SLS estimator only exist up to the order of over-identification.

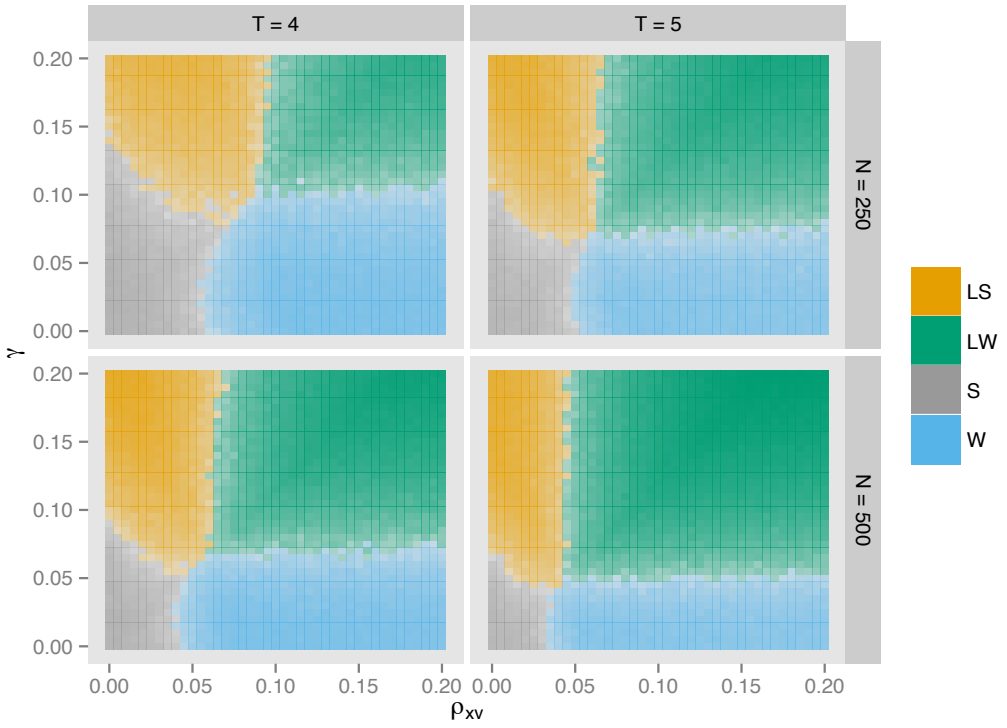

Figure 1: Minimum RMSE Specification at each combination of parameter values. Shading gives RMSE relative to second best specification.

lighter indicates a smaller advantage. While Figure 1 indicates *which* estimator is best, Figure 2 indicates how much of an advantage in RMSE can be gained over the correct specification, LW. These plots indicate that, provided γ and σ_{xv} are not too large, there are potentially large gains to be had by intentionally using an incorrectly specified estimator. The question remains, can the GFIC identify such situations?

To provide a basis for comparison, we consider a number of other selection procedures. The first is a "naïve" Downard J-test. To implement this procedure, we select the *most restrictive* specification that is not rejected by the over-identifying restrictions test at a fixed significance level, either 5% or 10%. Specifically, we proceed as follows:

1. Use S unless the J-test rejects it.

2. If S is rejected, use W unless the J-test rejects it.

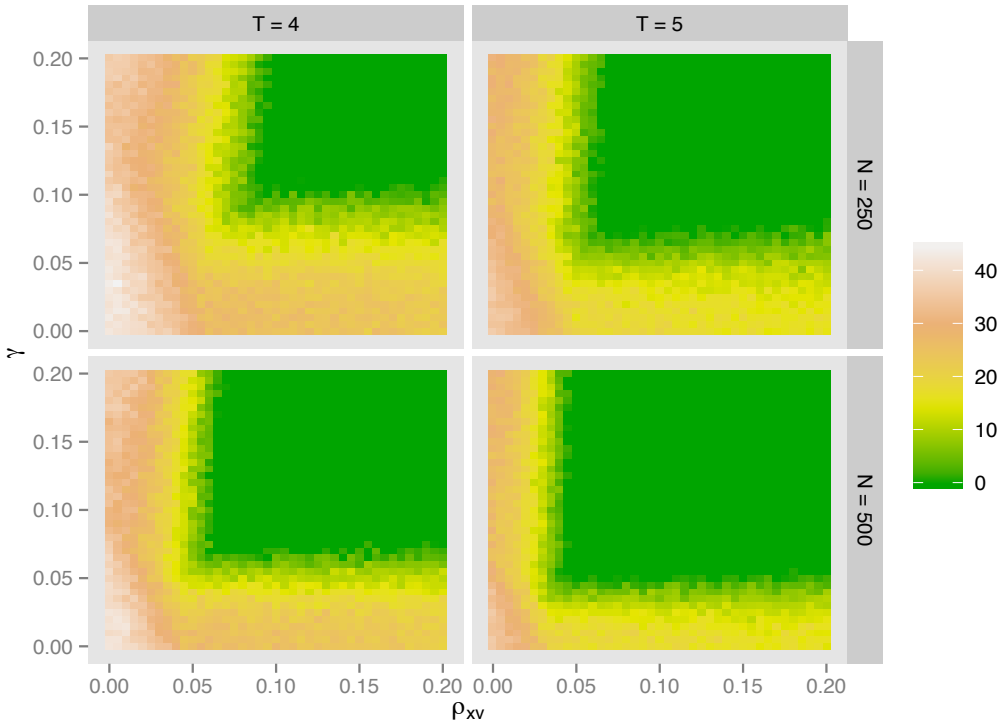

Figure 2: % RMSE Advantage of Best Specification (vs. LW)

3. If W is rejected, use LS unless the J-test rejects it.

4. Only use LW if all others specifications are rejected.

This procedure is "naïve" because the significance thresholds are chosed arbitrarily rather than with a view towards some kind of selection optimality. We also consider the GMM model and moment selection criteria of Andrews and Lu (2001):

$$\begin{aligned} \text{GMM-BIC} \quad & J - (|c| - |b|) \log n \\ \text{GMM-AIC} \quad & J - 2(|c| - |b|) \\ \text{GMM-HQ} \quad & J - 2.01(|c| - |b|) \log \log n \end{aligned}$$

where $|b|$ is the number of parameters estimated, and $|c|$ the number of moment conditions used. Under certain assumptions, it can be shown that both the GMM-BIC and GMM-HQ are consistent: they select the maximal correctly specified estimator with probability approaching one in the limit. To implement these criteria, we calculate the J-test based on the optimal, two-step GMM estimator with a

panel robust, heteroscedasticity-consistent, centered covariance matrix estimator for each specification.

To compare selection procedures we use the same simulation grid as above, namely γ and σ_{xv}, namely $\gamma, \sigma_{xv} \in \{0, 0.005, 0.01, \ldots, 0.195, 0.20\}$. Again, each point on the simulation grid is calculated from 2000 simulation replications. Tables 1 and 2 compare the performance of GFIC selection against each of the fixed specifications LW, LS, W, and S as well as the Downward J-test and the GMM moment and model selection criteria of Andrews and Lu (2001). Table 2 gives average and maximum, i.e. worst-case, RMSE over the parameter space for γ, σ_{xv} while Table 1 gives *relative* RMSE comparisons. Specifically, the values in the panel "Average" of Table 1 tell how much larger, in percentage points, the average RMSE of a given estimator or selection procedure is than that of the pointwise oracle. The pointwise oracle is the infeasible procedure that uses the true minimum RMSE estimator at each point on the parameter grid. In contrast, the values in the panel "Worst-Case" of Table 1 tell how much larger, in percentage points, the maximum RMSE of a given estimator or selection procedure is than that of the fixed specification LW. Over this parameter grid, LW is the minimax estimator.

Compared both to the fixed specifications and the other selection procedures, the GFIC performs well. In particular, it has a substantially lower average and worst-case RMSE than any of the other selection procedures. Compared to simply using the correct specification, LW, the GFIC also performs relatively well. When T and N are small, the GFIC outperforms LW in average RMSE. As they grow it performs slightly worse, but only by a small amount.

7 Conclusion

This paper has introduced the GFIC, a proposal to choose moment conditions and parameter restrictions based on the quality of the estimates they provide. While I focus here on a dynamic panel application, the GFIC can be applied to any GMM problem in which a minimal set of correctly specified moment conditions identifies an unrestricted model. A natural extension of this work would be to consider risk

Average	$N = 250$		$N = 500$	
	$T = 4$	$T = 5$	$T = 4$	$T = 5$
LW	19	10	13	7
LS	30	44	54	79
W	24	34	46	64
S	31	50	64	94
GFIC	17	13	15	10
Downward J-test (10%)	32	45	55	74
Downward J-test (5%)	31	47	57	79
GMM-BIC	32	48	62	87
GMM-HQ	32	46	57	77
GMM-AIC	31	39	47	57

Worst-Case	$N = 250$		$N = 500$	
	$T = 4$	$T = 5$	$T = 4$	$T = 5$
LW	0	0	0	0
LS	42	81	94	154
W	49	88	105	158
S	48	92	107	171
GFIC	3	8	6	11
Downward J-test (10%)	43	78	91	140
Downward J-test (5%)	45	83	98	153
GMM-BIC	48	89	106	168
GMM-HQ	46	85	102	154
GMM-AIC	39	68	81	118

Table 1: Average and Worst-case RMSE Relative to Oracle (% points)

Average	$N = 250$		$N = 500$	
	$T = 4$	$T = 5$	$T = 4$	$T = 5$
LW	0.073	0.057	0.051	0.040
LS	0.079	0.074	0.070	0.066
W	0.075	0.069	0.066	0.061
S	0.080	0.077	0.074	0.072
GFIC	0.071	0.058	0.052	0.041
Downward J-test (10%)	0.080	0.074	0.070	0.065
Downward J-test (5%)	0.080	0.075	0.071	0.067
GMM-BIC	0.080	0.076	0.073	0.069
GMM-HQ	0.080	0.075	0.071	0.066
GMM-AIC	0.080	0.071	0.066	0.058

Worst-Case	$N = 250$		$N = 500$	
	$T = 4$	$T = 5$	$T = 4$	$T = 5$
LW	0.084	0.064	0.059	0.045
LS	0.120	0.116	0.115	0.113
W	0.125	0.120	0.122	0.115
S	0.125	0.123	0.122	0.121
GFIC	0.087	0.069	0.063	0.049
Downward J-test (10%)	0.120	0.114	0.113	0.107
Downward J-test (5%)	0.122	0.117	0.117	0.113
GMM-BIC	0.125	0.121	0.122	0.119
GMM-HQ	0.123	0.118	0.120	0.113
GMM-AIC	0.117	0.107	0.107	0.097

Table 2: Average and Worst-case RMSE.

functions other than MSE, by analogy to Claeskens et al. (2006) and Claeskens and Hjort (2008). Another possibility would be to derive a version of the GFIC for GEL estimators. Although first-order equivalent to GMM, GEL estimators often exhibit superior finite-sample properties and may thus improve the quality of the selection criterion (Newey and Smith, 2004).

References

Anderson, T., Hsiao, C., 1982. Formulation and estimation of dynamic models using panel data. Journal of Econometrics 18, 47–82.

Andrews, D. W. K., Lu, B., 2001. Consistent model and moment selection procedures for GMM estimation with application to dynamic panel data models. Journal of Econometrics 101, 123–164.

Behl, P., Claeskens, G., Dette, H., March 2012. Focused model selection in quantile regression, Working Paper.

Caner, M., 2009. Lasso-type GMM estimator. Econometric Theory 25, 270–290.

Claeskens, G., Carroll, R. J., 2007. An asymptotic theory for model selection inference in general semiparametric problems. Biometrika 94, 249–265.

Claeskens, G., Croux, C., Jo, 2006. Variable selection for logistic regression using a prediction-focused information criterion. Biometrics 62, 972–979.

Claeskens, G., Croux, C., Kerckhoven, J. V., 2007. Prediction-focused model selection for autoregressive models. Australian and New Zealand Journal of Statistics 49, 359–379.

Claeskens, G., Hjort, N. L., 2003. The focused information criterion. Journal of the American Statistical Association 98 (464), 900–945.

Claeskens, G., Hjort, N. L., 2008. Minimizing average risk in regression models. Econometric Theory 24, 493–527.

DiTraglia, F. J., 2012. Using invalid instruments on purpose: Focused moment selection and averaging for GMM, Working Paper.

Hall, A. R., Peixe, F. P., 2003. A consistent method for the selection of relevant instruments in linear models. Econometric Reviews 22, 269–288.

Hjort, N. L., Claeskens, G., 2006. Focused information criteria and model averaging for the cox hazard regression model. Journal of the American Statistical Association 101 (476), 1449–1464.

Hong, H., Preston, B., Shum, M., 2003. Generalized empirical likelihood-based model selection for moment condition models. Econometric Theory 19, 923–943.

Judge, G. G., Mittelhammer, R. C., 2007. Estimation and inference in the case of competing sets of estimating equations. Journal of Econometrics 138, 513–531.

Lai, T. L., Small, D. S., Liu, J., 2008. Statistical inference in dynamic panel data models. Journal of Statistical Panning and Inference 138, 2763–2776.

Newey, W. K., Smith, R. J., 2004. Higher order properties of gmm and generalized empirical likelihood. Econometrica 72 (1), 219–255.

Roodman, D., 2009. A note on the theme of too many instruments. Oxford Bulletin of Economics and Statistics 71 (1), 135–158.

Schorfheide, F., 2005. VAR forecasting under misspecification. Journal of Econometrics 128, 99–136.

Skouras, S., November 2001. Decisionmetrics: A decision-based approach to econometric modelling, Working Paper.

Smith, R. J., July 1992. Non-nested tests for competing models estimated by generalized method of moments. Econometrica 60 (4), 973–980.

Zhang, X., Liang, H., 2011. Focused information criterion and model averaging for generalized additive partial linear models. Annals of Statistics 39 (1), 174–200.

A Proofs

Results from Section 2

Proof of Theorem 2.1. By a mean-value expansion around (γ_0, θ_0),

$$\sqrt{n}\left(\widehat{\beta}(b,c) - \beta_0^{(b)}\right) = -K(b,c)\Xi_c\sqrt{n}f_n(\gamma_0,\theta_0) + o_p(1)$$

and by the Lindeberg-Feller central limit theorem,

$$\sqrt{n}f_n(\gamma_0,\theta_0) - \sqrt{n}\mathbb{E}\left[f(Z_{ni},\gamma_0,\theta_0)\right] \to_d \begin{bmatrix} \mathscr{N}_g \\ \mathscr{N}_h \end{bmatrix}.$$

Now, by a mean-value expansion around γ_n,

$$\begin{aligned}
\sqrt{n}\mathbb{E}\left[f(Z_{ni},\gamma_0,\theta_0)\right] &= \sqrt{n}\mathbb{E}\left[f(Z_{ni},\gamma_n,\theta_0)\right] + \sqrt{n}\nabla_{\gamma'}\mathbb{E}\left[f(Z_{ni},\bar{\gamma},\theta_0)\right](\gamma_0-\gamma_n) \\
&= \begin{bmatrix} 0 \\ \tau \end{bmatrix} - \nabla_{\gamma'}\mathbb{E}\left[f(Z_{ni},\bar{\gamma},\theta_0)\right]\delta \\
&\to \begin{bmatrix} 0 \\ \tau \end{bmatrix} - F_\gamma\delta.
\end{aligned}$$

Hence,

$$\sqrt{n}f_n(\gamma_0,\theta_0) \to_d \begin{bmatrix} \mathscr{N}_g \\ \mathscr{N}_h \end{bmatrix} + \begin{bmatrix} 0 \\ \tau \end{bmatrix} - F_\gamma\delta \tag{A.1}$$

and the result follows by the continuous mapping theorem. □

Proof of Corollary 2.1. For the valid estimator,

$$\Xi_c\left(\begin{bmatrix} \mathscr{N}_g \\ \mathscr{N}_h \end{bmatrix} + \begin{bmatrix} 0 \\ \tau \end{bmatrix} - F_\gamma\delta\right) = \mathscr{N}_g - G_\gamma\delta$$

since Ξ_c picks out only components corresponding to g. Thus,

$$\sqrt{n}\left(\widehat{\beta}_v - \beta_0\right) \to_d -K_v\left(\mathscr{N}_g - G_\gamma\delta\right).$$

Finally

$$\begin{aligned} K_v G_\gamma \delta &= \left(\begin{bmatrix} G'_\gamma \\ G'_\theta \end{bmatrix} W_{gg} \begin{bmatrix} G_\gamma & G_\theta \end{bmatrix} \right) \begin{bmatrix} G'_\gamma \\ G'_\theta \end{bmatrix} W_{gg} G_\gamma \delta \\ &= \begin{bmatrix} G'_\gamma W_{gg} G_\gamma & G'_\gamma W_{gg} G_\theta \\ G'_\theta W_{gg} G_\gamma & G'_\theta W_{gg} G_\theta \end{bmatrix}^{-1} \begin{bmatrix} G'_\gamma W_{gg} G_\gamma \\ G'_\theta W_{gg} G_\gamma \end{bmatrix} \delta = \begin{bmatrix} 0 \\ \delta \end{bmatrix} \end{aligned}$$

by the definition of the matrix inverse. □

Results from Section 3

Proof of Corollary 3.2. By a mean-value expansion around γ_0,

$$\begin{aligned} \mu_n &= \varphi(\gamma_0, \theta_0) + \nabla_\gamma \varphi(\bar{\gamma}, \theta_0)'(\gamma_n - \gamma_0) \\ &= \mu_0 + \nabla_\gamma \varphi(\bar{\gamma}, \theta_0)' \delta / \sqrt{n} \end{aligned}$$

for some $\bar{\gamma}$ between γ_0 and $\gamma_n = \gamma_0 + \delta/\sqrt{n}$. Hence,

$$\sqrt{n}(\mu_n - \mu_0) = \nabla_\gamma \varphi(\bar{\gamma}, \theta_0)' \delta \to \nabla_\gamma \varphi(\gamma_0, \theta_0)' \delta$$

The result follows since

$$\sqrt{n}\left(\widehat{\mu}(b,c) - \mu_n\right) = \sqrt{n}\left(\widehat{\mu}(b,c) - \mu_0\right) - \sqrt{n}\left(\mu_n - \mu_0\right).$$

□

Proof of Corollary 3.3. The result follows from Corollaries 2.1 and 3.2 since,

$$\begin{aligned} \sqrt{n}\left(\widehat{\mu}_v - \mu_n\right) &\to_d \nabla_\beta \varphi'_0 \left\{ \begin{bmatrix} 0 \\ \delta \end{bmatrix} - K_v \mathscr{N}_g \right\} - \nabla_\gamma \varphi'_0 \delta \\ &= -\nabla_\beta \varphi'_0 K_v \mathscr{N}_g + \begin{bmatrix} \nabla_\theta \varphi'_0 & \nabla_\gamma \varphi'_0 \end{bmatrix} \begin{bmatrix} 0 \\ \delta \end{bmatrix} - \nabla_\gamma \varphi'_0 \delta \\ &= -\nabla_\beta (\gamma_0, \theta_0)' K_v \mathscr{N}_g. \end{aligned}$$

□

Proof of Lemma 3.1. By a mean-value expansion around $\beta_0 = (\gamma_0', \theta_0')'$,

$$\sqrt{n}h_n\left(\widehat{\beta}_v\right) = \sqrt{n}h_n(\beta_0) + H\sqrt{n}\left(\widehat{\beta}_v - \beta_0\right) + o_p(1).$$

Now, since

$$\sqrt{n}f_n(\gamma_0, \theta_0) \to_d \begin{bmatrix} \mathscr{N}_g \\ \mathscr{N}_h \end{bmatrix} + \begin{bmatrix} 0 \\ \tau \end{bmatrix} - \begin{bmatrix} G_\gamma \\ H_\gamma \end{bmatrix} \delta$$

we have

$$\sqrt{n}h_n(\gamma_0, \theta_0) \to_d \mathscr{N}_h + \tau - H_\gamma\delta.$$

Substituting,

$$\begin{aligned} \sqrt{n}h_n(\widehat{\beta}_v) &\to_d \mathscr{N}_h + \tau - H_\gamma\delta + H\left(-K_v\mathscr{N}_g + \begin{bmatrix} 0 \\ \delta \end{bmatrix}\right) \\ &= \mathscr{N}_h + \tau - H_\gamma\delta - HK_v\mathscr{N}_g + \begin{bmatrix} H_\theta & H_\gamma \end{bmatrix} \begin{bmatrix} 0 \\ \delta \end{bmatrix} \\ &= \mathscr{N}_h + \tau - H_\gamma\delta - HK_v\mathscr{N}_g + H_\gamma\delta \\ &= \tau - HK_v\mathscr{N}_g + \mathscr{N}_h \end{aligned}$$

as claimed. □

Proof of Corollary 3.5. Define

$$\begin{bmatrix} U \\ V \end{bmatrix} = \begin{bmatrix} \delta \\ \tau \end{bmatrix} + \Psi \begin{bmatrix} \mathscr{N}_g \\ \mathscr{N}_h \end{bmatrix}.$$

By the Continuous Mapping Theorem and Theorem 3.1,

$$\begin{bmatrix} \widehat{\delta} \\ \widehat{\tau} \end{bmatrix} \begin{bmatrix} \widehat{\delta}' & \widehat{\tau}' \end{bmatrix} \to_d \begin{bmatrix} U \\ V \end{bmatrix} \begin{bmatrix} U' & V' \end{bmatrix}$$

The result follows since

$$\Psi\Omega\Psi' = Var \begin{bmatrix} U \\ V \end{bmatrix} = \mathbb{E} \begin{bmatrix} UU' & UV' \\ VU' & VV' \end{bmatrix} - \begin{bmatrix} \delta\delta' & \delta\tau' \\ \tau\delta' & \tau\tau' \end{bmatrix}.$$

□