A Generalized Focused Information Criterion for GMM with Applications to Panel Data Models[*]

Minsu Chang   Francis J. DiTraglia[†]

University of Pennsylvania

This Version: January 29, 2017   First Version: February 15, 2013

Abstract

In this paper we propose a 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 misspecified moment conditions and parameter restrictions to minimize the mean-squared error of a user-specified target parameter. In addition to presenting the general theory, we specialize the GFIC to the problem of choosing between random and fixed effects estimators, and propose a novel averaging estimator that combines the two. In addition, we consider an application to a dynamic panel model, in which the user wishes to select over both the exogeneity assumptions used to generate instruments, and the lag specification. The GFIC performs well in simulations for both examples.

Keywords: Model Selection, Moment selection, Model averaging, Panel Data, GMM Estimation, Focused Information Criterion, Post-selection estimators

JEL Codes: C23, C52

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. We 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 takes its motivation from a situation that is common in empirical practice. A researcher who hopes to estimate a parameter of interest μ must decide which assumptions to use. On the one hand is a set of relatively uncontroversial "baseline" assumptions. We suppose that the baseline assumptions are correct and identify μ. But the very fact that that the baseline assumptions do not raise eyebrows suggests that they may not be especially informative about μ. On the other hand are one or more stronger controversial "suspect" assumptions. These stronger assumptions are expected to be much more informative about μ. If we were certain that they were correct, we would definitely choose to impose them in estimation. Indeed, by continuity, even if they were *nearly* correct, imposing the suspect assumptions could yield a favorable bias-variance tradeoff. This is the essential idea behind the GFIC.

The focused moment selection criterion (FMSC) of DiTraglia (2015) can be viewed as a special case of the GFIC. While the FMSC considers the problem of selecting moment conditions while holding the model specification *fixed*, the GFIC allows us to select over both aspects of our specification simultaneously. This extension is particularly valuable in panel data applications, where we may, for example, wish to carry out selection over the lag specification as well as the exogeneity assumptions used to estimate a dynamic panel model. We specialize the GFIC to this example below, in addition to another that involves selecting between random and fixed effects estimators. For this latter example, we further derive a novel averaging estimator that optimally combines the information contained in the random effects and fixed effects estimators. The GFIC and averaging estimators perform well in simulation studies for both examples. In addition to extending the FMSC to a broader class

of problems and deriving specific results for well-known panel data problems, we also extend the results of DiTraglia (2015) on post-selection and moment-average estimators to the more general setting of the GFIC.

As its name suggests, the GFIC is related to 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 (2015) 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 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 4 extends the results on averaging estimators and post-selection inference from DiTraglia (2015) to the more general setting of this paper. Section 5 specializes the GFIC to the problem of choosing between random effects and fixed effects estimators, and proposes an estimator that optimally averages the two while Section ??? considers a dynamic panel example. Section ??? presents the results of simulation studies for each of the two examples and Section ??? concludes. Proofs appear in the Appendix.

---

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

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* $\{Z_{ni} \colon 1 \le i \le n, n = 1, 2, \ldots\}$ *be a triangular array of random vectors defined on a probability space* $(\Upsilon, \mathcal{F}, P)$ *satisfying*

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

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

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

(d) $Z_{ni} \xrightarrow{d} Z_i$, *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. $\{Z_{ni} \colon 1 \le i \le n, n = 1, 2, \ldots\}$ *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 $E[f(Z_i, \gamma_0, \theta_0))] = 0$. 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 $\Xi_b\beta = \left(\gamma^{(b)\prime}, \theta'\right)'$. 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,

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

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

$$\widetilde{W} = \begin{bmatrix} \widetilde{W}_{gg} & \widetilde{W}_{gh} \\ \widetilde{W}_{hg} & \widetilde{W}_{hh} \end{bmatrix} \tag{2}$$

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 $\beta^{(b)} = \left(\gamma^{(b)\prime}, \theta'\right)'$ according to

$$\widehat{\beta}(b, c) = \underset{\beta^{(b)} \in \mathbf{B}^{(b)}}{\arg \min} \ \left[\Xi_c f_n \left(\beta^{(b)}, \gamma_0^{(-b)}\right)\right]' \left[\Xi_c \widetilde{W} \Xi_c'\right] \left[\Xi_c f_n \left(\beta^{(b)}, \gamma_0^{(-b)}\right)\right]. \tag{3}$$

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:

$$\widehat{\beta}_v = \begin{bmatrix} \widehat{\gamma}_v \\ \widehat{\theta}_v \end{bmatrix} = \underset{\beta \in \mathbf{B}}{\arg \min} \ g_n(\beta)' \widetilde{W}_{gg} \ g_n(\beta). \tag{4}$$

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

$$F = \begin{bmatrix} \nabla_{\gamma'} g \left(Z, \gamma_0, \theta_0\right) & \nabla_{\theta'} g \left(Z, \gamma_0, \theta_0\right) \\ \nabla_{\gamma'} h \left(Z, \gamma_0, \theta_0\right) & \nabla_{\theta'} h \left(Z, \gamma_0, \theta_0\right) \end{bmatrix} \tag{5}$$

partitioned according to

$$F = \begin{bmatrix} F_\gamma & F_\theta \end{bmatrix} = \begin{bmatrix} G_\gamma & G_\theta \\ H_\gamma & H_\theta \end{bmatrix} = \begin{bmatrix} G \\ H \end{bmatrix} \tag{6}$$

and define

$$\Omega = Var \begin{bmatrix} g(Z, \gamma_0, \theta_0) \\ h(Z, \gamma_0, \theta_0) \end{bmatrix} = \begin{bmatrix} \Omega_{gg} & \Omega_{gh} \\ \Omega_{hg} & \Omega_{hh} \end{bmatrix}. \tag{7}$$

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 Assumption 2.1, both δ and τ induce a bias term in the

---

[2]The required high-level sufficient conditions are essentially identical to Assumption 2.2 of DiTraglia (2015).

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,*

$$\sqrt{n}\left(\widehat{\beta}(b,c) - \beta_0^{(b)}\right) \xrightarrow{d} -K(b,c)\Xi_c\left(\mathcal{N} + \begin{bmatrix} 0 \\ \tau \end{bmatrix} - F_\gamma \delta\right) \tag{8}$$

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

$$K(b,c) = \left[F(b,c)'W_c F(b,c)\right]^{-1} F(b,c)'W_c \tag{9}$$

and

$$\mathcal{N} = \begin{bmatrix} \mathcal{N}_g \\ \mathcal{N}_h \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \Omega_{gg} & \Omega_{gh} \\ \Omega_{hg} & \Omega_{hh} \end{bmatrix}\right). \tag{10}$$

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,*

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

where $K_v = [G'W_{gg}G]^{-1} G'W_{gg}$ and $W_{gg} = plim\,\widetilde{W}_{gg}$.

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,*

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

8

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,*

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

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,*

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

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,

$$\text{AMSE}\left(\widehat{\mu}(b,c)\right) = \text{AVAR}\left(\widehat{\mu}\left(b,c\right)\right) + \text{BIAS}\left(\widehat{\mu}\left(b,c\right)\right)^2 \tag{11}$$

where

$$\text{AVAR}\left(\widehat{\mu}\left(b,c\right)\right) = \nabla_\beta \varphi_0' \Xi_b' K(b,c) \Omega_c K(b,c)' \Xi_b \nabla_\beta \varphi_0 \tag{12}$$

$$\text{BIAS}\left(\widehat{\mu}\left(b,c\right)\right) = -\nabla_\beta \phi_0' M(b,c) \begin{bmatrix} \delta \\ \tau \end{bmatrix} \tag{13}$$

and

$$M(b,c) = \Xi_b' K(b,c) \Xi_c \begin{bmatrix} -G_\gamma & 0 \\ -H_\gamma & I \end{bmatrix} + \begin{bmatrix} I_r & 0_{r \times q} \\ 0_{p \times r} & 0_{s \times q} \end{bmatrix} \tag{14}$$

9

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 $\widehat{\delta} = \sqrt{n}\left(\widehat{\gamma}_v - \gamma_0\right)$.

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

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

where $K_v = [G'W_{gg}G]^{-1} G'W_{gg} = \left(K_v^{\gamma'}, K_v^{\theta'}\right)'$. 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,*

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

where, as above, $K_v = [G'W_{gg}G]^{-1} G'W_{gg}$. 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,*

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

where

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

---

[3] For more on this point, see DiTraglia (2015) Section 3.3.

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

From Equation 13,

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

Thus, the bias parameters τ and δ enter the AMSE expression in Equation 11 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,*

$$\widehat{B} = \begin{bmatrix} \widehat{\tau}\widehat{\tau}' & \widehat{\tau}\widehat{\delta}' \\ \widehat{\delta}\widehat{\tau}' & \widehat{\delta}\widehat{\delta}' \end{bmatrix} - \widehat{\Psi}\widehat{\Omega}\widehat{\Psi}' \tag{15}$$

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 $(b,c) \in \mathcal{BC}$

$$\text{GFIC}(b,c) = \nabla_\beta \widehat{\varphi}_0' \left[\Xi_b' \widehat{K}(b,c) \widehat{\Omega}_c \widehat{K}(b,c)' \Xi_b + \widehat{M}(b,c) \, \widehat{B} \, \widehat{M}(b,c)' \right] \nabla_\beta \widehat{\varphi}_0. \tag{16}$$

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

While we are primarily concerned in this paper with the mean-squared error performance of our proposed selection techniques, it is important to have tools for carrying out valid inference post-selection. To this end, we now show how to extend the results from Section 4 of DiTraglia (2015) to the more general setting considered in this paper, one that allows for

simultaneous model and moment selection.[4] Consider an estimator of the form

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

where $\widehat{\mu}$ denotes the target parameter under the moment conditions and parameter restrictions indexed by (b,c), $\mathcal{BC}$ denotes the full set of candidate specifications, and $\widehat{\omega}(b,c)$ denotes a collection of data-dependent weights satisfying the following assumption.

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) $\widehat{\omega}(b,c) \xrightarrow{d} \psi(\mathcal{N}, \delta, \tau | b, c)$ *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.*

Assumption 4.1 is quite weak, covering a broad range of examples, including genuine averaging estimators, post-GFIC estimators, and pre-test estimators based on the J-statistic. Under this assumption, we can characterize the limit distribution of $\widehat{\mu}$ as follows.

Corollary 4.1 (Limit Distribution of Averaging Estimators)**.** *Let $\widehat{\omega}(b,c)$ be a set of weights satisfying Assumption 4.1. Then, under the hypotheses of Theorem 2.1,*

$$\sqrt{n}\left(\widehat{\mu} - \mu_n\right) \xrightarrow{d} \Lambda(\tau, \delta)$$

where

$$\Lambda(\tau, \delta) = -\nabla_\beta \varphi_0' \sum_{(b,c) \in \mathcal{BC}} \psi(\mathcal{N}, \delta, \tau | b, c) \left\{ \Xi_b' K(b,c) \Xi_c \mathcal{N} + M(b,c) \begin{bmatrix} \delta \\ \tau \end{bmatrix} \right\} \quad (17)$$

Note that the limit distribution from the preceding corollary is highly non-normal: it is a *randomly* weighted average of a normal random vector, $\mathcal{N}$. To tabulate this distribution for the purposes of inference, we will in general need to resort to simulation. If τ and δ were known, the story would end here. We could simply substitute consistent estimators of K and M, and then repeatedly draw $\mathcal{N} \sim N(0, \widehat{\Omega})$, where $\widehat{\Omega}$ is a consistent estimator of Ω, and thus tabulate the distribution of Λ to arbitrary precision. Unfortunately, no

---

[4]Because the conceptual issues are largely the same as in the case where one considers only moment selection, we direct the reader to DiTraglia (2015) for more discussion.

consistent estimators of τ or δ exist: all we have at our disposal are asymptotically unbiased estimators. Simply plugging in these estimators $\widehat{\tau}$ and $\widehat{\delta}$ and proceeding with the simulation is not guaranteed to lead to valid confidence intervals.[5] In contrast, the following two-step procedure, is guaranteed to yield confidence intervals with asymptotic coverage probability *no less than* $1 - (\alpha_1 + \alpha_2)$.

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 $j = 1, 2, \ldots, J$ 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 $\{\Lambda_j(\delta, \tau)\}_{j=1}^J$, calculate $\widehat{a}(\delta, \tau), \widehat{b}(\delta, \tau)$ such that

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

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

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

Theorem 4.1 (Simulation-based Confidence Interval for $\widehat{\mu}$)**.** *Let* $\nabla_\beta \widehat{\varphi}_0$, $\widehat{\psi}(\cdot | b, c)$, $\widehat{K}(b, c)$ *and* $\widehat{M}(b, c)$ *be consistent estimators of* $\nabla_\beta \varphi$, $\psi(\cdot | b, c)$, $K(b, c)$ *and* $M(b, c)$ *and let* $R(\alpha_1)$ *be a* $(1 - \alpha_1) \times 100\%$ *joint confidence region for* (δ, τ) *constructed from Theorem 3.1. Then the interval* CI_{sim} *defined in Algorithm 4.1 has asymptotic coverage probability no less than* $1 - (\alpha_1 + \alpha_2)$ *as* $J, n \to \infty$.

[5]Although it does not work in general, in particular examples this plug-in procedure may perform well. For more discussion of this point, see Section 4.4 of DiTraglia (2015).

5 Random Effects versus Fixed Effects Example

In this section we consider a simple example in which the the GFIC is used to choose between and average over alternative assumptions about individual heterogeneity: Random Effects versus Fixed Effects. For simplicity we consider the homoskedastic case and assume that any strictly exogenous regressors, including a constant term, have been "projected out" so we may treat all random variables as mean zero. To avoid triple subscripts in the notation, we further suppress the dependence of random variables on the cross-section dimension n except within statements of theorems. Suppose that

$$
\begin{align}
y_{it} &= \beta x_{it} + v_{it} \tag{18}\\
v_{it} &= \alpha_i + \varepsilon_{it} \tag{19}
\end{align}
$$

for $i = 1, \ldots, n$, $t = 1, \ldots, T$ where ε_{it} is iid across i, t with $Var(\varepsilon_{it}) = \sigma_\varepsilon^2$ and α_i is iid across i with $Var(\alpha_i) = \sigma_\alpha^2$. Stacking observations for a given individual over time in the usual way, let $\mathbf{y}_i = (y_{i1}, \ldots, y_{iT})'$ and define $\mathbf{x}_i, \mathbf{v}_i$ and $\boldsymbol{\varepsilon}_i$ analogously. Our goal in this example is to estimate β, the effect of x on y. Although x_{it} is uncorrelated with the time-varying portion of the error term, $Cov(x_{it}, \varepsilon_{it}) = 0$, we are unsure whether or not it is correlated with the individual effect α_i. If we knew for certain that $Cov(x_{it}, \alpha_i) = 0$, we would prefer to report the "random effects" generalized least squares (GLS) estimator given by

$$
\widehat{\beta}_{GLS} = \left(\sum_{i=1}^n \mathbf{x}_i' \widehat{\Omega}^{-1} \mathbf{x}_i \right)^{-1} \left(\sum_{i=1}^n \mathbf{x}_i' \widehat{\Omega}^{-1} \mathbf{y}_i \right) \tag{20}
$$

where $\widehat{\Omega}^{-1}$ is a preliminary consistent estimator of

$$
\Omega^{-1} = [Var(\mathbf{v}_i)]^{-1} = \frac{1}{\sigma_\epsilon^2} \left[I_T - \frac{\sigma_\alpha^2}{(T\sigma_\alpha^2 + \sigma_\epsilon^2)} \boldsymbol{\iota}_T \boldsymbol{\iota}_T' \right] \tag{21}
$$

and I_T denotes the $T \times 1$ identity matrix and $\boldsymbol{\iota}_T$ a T-vector of ones. This estimator makes efficient use of the variation between and within individuals, resulting in an estimator with a lower variance. When $Cov(x_{it}, \alpha_i) \neq 0$, however, the random effects estimator is biased. Although its variance is higher than that of the GLS estimator, the "fixed effects" estimator given by

$$
\widehat{\beta}_{FE} = \left(\sum_{i=1}^n \mathbf{x}_i' Q \mathbf{x}_i \right)^{-1} \left(\sum_{i=1}^n \mathbf{x}_i' Q \mathbf{y}_i \right), \tag{22}
$$

where $Q = I_T - \boldsymbol{\iota}\boldsymbol{\iota}'/T$, remains unbiased even when x_{it} is correlated with α_i.

The conventional wisdom holds that one should use the fixed effects estimator whenever

$Cov(x_{it}, \alpha_i) \neq 0$. If the correlation between the regressor of interest and the individual effect is *sufficiently small*, however, the lower variance of the random effects estimator could more than compensate for its bias in a mean-squared error sense. This is precisely the possibility that we consider here using the GFIC. In this example, the local mis-specification assumption takes the form

$$\sum_{t=1}^{T} E\left[x_{it}\alpha_i\right] = \frac{\tau}{\sqrt{n}} \tag{23}$$

where τ is fixed, unknown constant. In the limit the random effects assumption that $Cov(x_{it}, \alpha_i) = 0$ holds, since $\tau/\sqrt{n} \to 0$. Unless $\tau = 0$, however, this assumption *fails* to hold for any finite sample size. An asymptotically unbiased estimator of τ for this example is given by

$$\widehat{\tau} = (T\widehat{\sigma}_\alpha^2 + \widehat{\sigma}_\epsilon^2)\left[\frac{1}{\sqrt{n}}\sum_{i=1}^{n} \mathbf{x}_i'\widehat{\Omega}^{-1}(\mathbf{y}_i - \mathbf{x}_i\widehat{\beta}_{FE})\right] \tag{24}$$

leading to the following result, from which we will construct the GFIC for this example.

Theorem 5.1 (Fixed versus Random Effects Limit Distributions). *Let $(\mathbf{x}_{ni}, \alpha_{ni}, \boldsymbol{\varepsilon}_{ni})$ be an iid triangular array of random variables such that $Var(\boldsymbol{\varepsilon}_i|\mathbf{x}_{ni}, \alpha_{ni}) \to \sigma_\varepsilon^2 I_T$, $E[\mathbf{x}_{it}'Q\boldsymbol{\varepsilon}_{it}] = 0$, and $E\left[\alpha_i \boldsymbol{\iota}_T'\mathbf{x}_{it}\right] = \tau/\sqrt{n}$ for all n. Then, under standard regularity conditions,*

$$\begin{bmatrix} \sqrt{n}(\widehat{\beta}_{RE} - \beta) \\ \sqrt{n}(\widehat{\beta}_{FE} - \beta) \\ \widehat{\tau} \end{bmatrix} \xrightarrow{d} N\left(\begin{bmatrix} c\tau \\ 0 \\ \tau \end{bmatrix}, \begin{bmatrix} \eta^2 & \eta^2 & 0 \\ \eta^2 & c^2\sigma^2 + \eta^2 & -c\sigma^2 \\ 0 & -c\sigma^2 & \sigma^2 \end{bmatrix}\right)$$

where $\eta^2 = E[\mathbf{x}_i'\Omega^{-1}\mathbf{x}_i]$, $c = E[\mathbf{x}_i'Q\mathbf{x}_i]/(T\sigma_\alpha^2 + \sigma_\varepsilon^2)$, and

$$\sigma^2 = \frac{(T\sigma_\alpha^2 + \sigma_\varepsilon^2)^2}{E\left[\mathbf{x}_i'\Omega^{-1}\mathbf{x}_i\right]}\left(\frac{\sigma_\varepsilon^2}{E\left[\mathbf{x}_i\Omega^{-1}\mathbf{x}_i\right]E\left[\mathbf{x}_iQ\mathbf{x}_i\right]} - 1\right).$$

We see from Theorem 5.1 that $AMSE(\widehat{\beta}_{RE}) = c^2\tau^2 + \eta^2$, $AMSE(\widehat{\beta}_{FE}) = c^2\sigma^2 + \eta^2$, and $\widehat{\tau}^2 - \sigma^2$ provides an asymptotically unbiased estimator of τ. Thus, substituting $\widehat{\tau}^2 - \sigma^2$ for τ and rearranging the preceding AMSE expressions, the GFIC tells us that we should select the random effects estimator whenever $|\widehat{\tau}| \leq \sqrt{2}\sigma$. To implement this rule in practice, we construct a consistent estimator of σ^2, for which we require estimators of $\sigma_\alpha^2, \sigma_\varepsilon^2$ and $\sigma_v^2 = Var(\alpha_i + \varepsilon_{it})$. We estimate these from the residuals

$$\widehat{\epsilon}_{it} = (y_{it} - \bar{y}_i) - (x_{it} - \bar{x}_i)\widehat{\beta}_{FE}; \quad \widehat{v}_{it} = y_{it} - x_{it}\widehat{\beta}_{OLS}$$

where $\widehat{\beta}_{OLS}$ denotes the *pooled* OLS estimator of β, leading to the variance estimators

$$\widehat{\sigma}_\alpha^2 = \widehat{\sigma}_v^2 - \widehat{\sigma}_\epsilon^2; \quad \widehat{\sigma}_\epsilon^2 = \frac{1}{N(T-1)-1}\sum_{i=1}^{N}\sum_{t=1}^{T}\widehat{\epsilon}_{it}^2; \quad \widehat{\sigma}_v^2 = \frac{1}{NT-1}\sum_{i=1}^{N}\sum_{t=1}^{T}\widehat{v}_{it}^2$$

Selection, of course, is a somewhat crude procedure: it is essentially an average that uses all-or-nothing weights. As a consequence, relatively small changes to the data could produce discontinuous changes in the weights, leading to a procedure with a high variance. Rather than selecting between the random effects and fixed effects estimators based on estimated AMSE, an alternative idea is to consider a more general weighted average of the form

$$\widetilde{\beta}(\omega) = \omega\widehat{\beta}_{FE} + (1-\omega)\widehat{\beta}_{RE}$$

and for $\omega \in [0,1]$ *optimize* the choice of ω to minimize AMSE. From Theorem 5.1 we see that the AMSE-minimizing value of ω is $\omega^* = (1 + \tau^2/\sigma^2)^{-1}$. Substituting our asymptotically unbiased estimator of τ^2 and our consistent estimator $\widehat{\sigma}^2$ of σ^2, we propose the following plug-in estimator of ω^*

$$\omega^* = \left[1 + \frac{\max\{\widehat{\tau}^2 - \widehat{\sigma}^2, 0\}}{\widehat{\sigma}^2}\right]$$

where we take the maximum over $\widehat{\tau}$ and zero to ensure that $\widehat{\omega}^*$ is between zero and one.

6 Dynamic Panel Example

We now specialize the GFIC to a dynamic panel model of the form

$$y_{it} = \theta x_{it} + \gamma_1 y_{it-1} + \cdots + \gamma_k y_{it-k} + \eta_i + v_{it} \tag{25}$$

where $i = 1, \ldots, n$ indexes individuals and $t = 1, \ldots, T$ indexes time periods. For simplicity, and without loss of generality, we suppose that there are no exogenous time-varying regressors and that all random variables are mean zero.[6] The unobserved error η_i is a correlated individual effect: $\sigma_{x\eta} \equiv \mathbb{E}[x_{it}\eta_i]$ may not equal zero. The endogenous regressor x_{it} is assumed to be predetermined but not necessarily strictly exogenous: $\mathbb{E}[x_{it}v_{is}] = 0$ for all $s \geq t$ but may be nonzero for $s < t$. We assume throughout that y_{it} is stationary, which requires both x_{it} and u_{it} to be stationary and $|\boldsymbol{\gamma}| < 1$ where $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_k)'$. Our goal is to estimate one

---

[6]Alternatively, we can simply de-mean and project out any time-varying exogenous covariates after taking first-differences.

of the following two target parameters with minimum MSE:

$$\mu_{\mathrm{SR}} \equiv \theta, \quad \mu_{\mathrm{LR}} \equiv \frac{\theta}{1 - (\gamma_1 + \cdots + \gamma_k)}. \tag{26}$$

where μ_{SR} denotes the short-run effect and μ_{LR} the long-run effect of x on y.

The question is which assumptions to use in estimation. Naturally, the answer may depend on whether our target is μ_{SR} or μ_{LR}. Our first decision is what assumption to impose on the relationship between x_{it} and v_{it}. This is the *moment selection* decision. We assumed above that x is predetermined. Imposing the stronger assumption of strict exogeneity gives us more and stronger moment conditions, but using these in estimation introduces a bias if x is not in fact strictly exogenous. Our second decision is how many lags of y to use in estimation. This is the *model selection* decision. The true model contains k lags of y. If we estimate only $r < k$ lags we not only have more degrees of freedom but more observations: every additional lag of y requires us to drop one time period from estimation. In the short panel datasets common in microeconomic applications, losing even one additional time period can represent a substantial loss of information. At the same time, unless $\gamma_{r+1} = \cdots = \gamma_k = 0$, failing to include all k lags in the model introduces a bias.

To eliminate the individual effects η_i we work in first differences. Defining Δ in the usual way, so that $\Delta y_{it} = y_{it} - y_{it-1}$ and so on, we can write Equation 25 as

$$\Delta y_{it} = \theta \Delta x_{it} + \gamma_1 \Delta y_{it-1} + \cdots + \gamma_k \Delta y_{it-k} + \Delta v_{it}. \tag{27}$$

For simplicity and to avoid many instruments problems – see e.g. Roodman (2009) – we focus here on estimation using the instrument sets

$$\mathbf{z}'_{it}(\ell, \mathrm{P}) \equiv \begin{bmatrix} y_{it-2} & \cdots & y_{it-(\ell+1)} & x_{it-1} \end{bmatrix} \qquad \mathbf{z}'_{it}(\ell, \mathrm{S}) \equiv \begin{bmatrix} \mathbf{z}'_{it}(\ell, \mathrm{P}) & x_{it} \end{bmatrix} \tag{28}$$

similar to Anderson and Hsiao (1982). Modulo a change in notation, one could just as easily proceed using the instrument sets suggested by Arellano and Bond (1991). Throughout this discussion we use ℓ as a placeholder for the lag length used in estimation. If $\ell = 0$, $\mathbf{z}'_{it}(0, \mathrm{P}) = x_{it-1}$ and $\mathbf{z}'_{it}(0, \mathrm{S}) = (x_{it-1}, x_{it})$. Given these instrument sets, we have $(\ell + 1) \times (T - \ell - 1)$ moment conditions if x is assumed to be predetermined versus $(\ell + 2) \times (T - \ell - 1)$ if it is assumed to be strictly exogenous, corresponding to the instrument matrices $Z_i(\ell, \mathrm{P}) = \mathrm{diag} \{\mathbf{z}'_{it}(\ell, \mathrm{P})\}_{t=\ell+2}^{T}$ and $Z_i(\ell, \mathrm{S}) = \mathrm{diag} \{\mathbf{z}'_{it}(\ell, \mathrm{S})\}_{t=\ell+2}^{T}$. To abstract for a moment from the model selection decision, suppose that we estimate a model with the true lag length: $\ell = k$. The only difference between the P and S sets of moment conditions is that the latter adds over-identifying information in the form of $E[x_{it} \Delta v_{it}]$. If x is strictly exogenous, this

expectation equals zero, but if x is only predetermined, then $E[x_{it}\Delta v_{it}] = -E[x_{it}v_{it-1}] \neq 0$ so the over-identifying moment condition is invalid. Given the instrument sets that we consider, this is the only violation of strict exogeneity that is relevant for our moment selection so we take $E[x_{it}v_{it-1}] = -\tau/\sqrt{n}$.

In the examples and simulations described below we consider two-stage least squares (TSLS) estimation of μ_{SR} and μ_{LR} using the instruments defined in Equation 28. For simplicity, we select between two lag length specifications: the first is correct, $\ell = k$, and the second includes one lag too few: $\ell = r$ where $r = k - 1$. Accordingly, we make coefficient associated with the kth lag local to zero. Let $\boldsymbol{\gamma}' = (\gamma_1, \cdots, \gamma_{k-1}, \gamma_k)$ denote the full vector of lag coefficients that $\boldsymbol{\gamma}_r = (\gamma_1, \cdots, \gamma_r)$ the first $r = k - 1$ lag coefficients. Then, the true parameter vector is $\beta_n = (\theta, \boldsymbol{\gamma}'_r, \delta/\sqrt{n})'$ which becomes, in the limit, $\beta = (\theta, \boldsymbol{\gamma}'_r, 0)'$. To indicate the subvector of β that excludes the kth lag coefficient, let $\beta_r = (\theta, \boldsymbol{\gamma}'_r)'$.

Because the two lag specifications we consider use different time periods in estimation, we require some additional notation to make this clear. First let $\Delta\mathbf{y}_i = [\Delta y_{i,k+2}, \cdots, \Delta y_{iT}]'$ and $\Delta\mathbf{y}_i^+ = [\Delta y_{i,k+1}, \Delta y_{i,k+2}, \cdots, \Delta y_{iT}]'$ where the superscript "+" indicates the inclusion of an additional time period: $t = k + 1$. Define $\Delta\mathbf{x}_i$, $\Delta\mathbf{x}_i^+$, $\Delta\mathbf{v}_i$, and $\Delta\mathbf{v}_i^+$ analogously. Next, define $L^k\Delta\mathbf{y}_i^+ = [\Delta y_{i1}, \Delta y_{i2}, \cdots, \Delta y_{iT-k}]'$ where L^k denotes the element-wise application of the kth order lag operator. Note that the first element of $L^k\Delta\mathbf{y}_i^+$ is unobserved since $\Delta y_{i1} = y_{i1} - y_{i0}$ but $t = 1$ is the first time period. Now we define the matrices of regressors for the two specifications:

$$W_i^{+'}(r) = \begin{bmatrix} \Delta\mathbf{x}_i^+ & L\Delta\mathbf{y}_i^+ & L^2\Delta\mathbf{y}_i^+ & \cdots & L^{k-1}\Delta\mathbf{y}_i^+ \end{bmatrix}$$

$$W_i'(k) = \begin{bmatrix} \Delta\mathbf{x}_i & L\Delta\mathbf{y}_i & L^2\Delta\mathbf{y}_i & \cdots & L^{k-1}\Delta\mathbf{y}_i & L^k\Delta\mathbf{y}_i \end{bmatrix}.$$

Note that $W_i^+(r)$ contains one more row than $W_i(k)$ but $W_i(k)$ contains one more column that $W_i^+(r)$: removing the kth lag from the model by setting $\ell = r = k - 1$ allows us to use an additional time period in estimation and reduces the number of regressors by one. Stacking over individuals, let $\Delta\mathbf{y} = [\Delta\mathbf{y}_1' \cdots \Delta\mathbf{y}_n']'$, $W_\ell = [W_1(\ell) \cdots W_n(\ell)]'$ and define $\Delta\mathbf{y}^+$ and W_ℓ^+ analogously, where ℓ denotes the lag length used in estimation. Finally, let $Z'(\ell, \cdot) = [Z_1'(\ell, \cdot) \cdots Z_n'(\ell, \cdot)]$ where $(\cdot)$ is P or S depending on the instrument set is in use. Using this notation, under local mis-specification the true model is

$$\Delta\mathbf{y} = W(k)\beta_n + \Delta\mathbf{v} \qquad\qquad \Delta\mathbf{y}^+ = W(k)^+\beta_n + \Delta\mathbf{v}^+ \qquad\qquad (29)$$

Using the shorthand $\widehat{Q} \equiv n[W'Z(Z'Z)^{-1}Z'W]^{-1}W'Z(Z'Z)^{-1}$ our candidate estimators are

$$\widehat{\beta}(k,\cdot) = \widehat{Q}(k,\cdot)\left[\frac{Z'(k,\cdot)\Delta\mathbf{y}}{n}\right] \qquad \widehat{\beta}(r,\cdot) = \widehat{Q}(r,\cdot)\left[\frac{Z'(r,\cdot)\Delta\mathbf{y}^+}{n}\right] \qquad (30)$$

where $(\cdot)$ is either P or S depending on which instrument set is used and $r = k - 1$, one lag fewer than the true lag length k. The following result describes the limit distribution of $\widehat{\beta}(k,\mathrm{P})$, $\widehat{\beta}(k,\mathrm{S})$, $\widehat{\beta}(r,\mathrm{P})$, and $\widehat{\beta}(r,\mathrm{S})$ which we will use to construct the GFIC.

Theorem 6.1 (Limit Distributions for Dynamic Panel Estimators). *Let $(y_{nit}, x_{nit}, v_{nit})$ be a triangular array of random variables that is iid over i, stationary over t, and satisfies Equation 27 with $\gamma_k = \delta/\sqrt{n}$. Suppose further that x_{it} is predetermined with respect to v_{it} but not strictly exogenous: $E[x_{it}\Delta v_{it}] = \tau/\sqrt{n}$. Then, under standard regularity conditions,*

$$\sqrt{n}\left[\widehat{\beta}(k,P) - \beta\right] \to^d \begin{bmatrix} 0 & \mathbf{0}'_{k-1} & \delta \end{bmatrix}' + Q(k,P)\, N(\mathbf{0}, \mathcal{V}(k,P))$$

$$\sqrt{n}\left[\widehat{\beta}(k,S) - \beta\right] \to^d \begin{bmatrix} 0 & \mathbf{0}'_{k-1} & \delta \end{bmatrix}' + Q(k,S)\left\{ \boldsymbol{\iota}_{T-(k+1)} \otimes \begin{bmatrix} \mathbf{0}_{k+1} \\ \tau \end{bmatrix} + N(\mathbf{0}, \mathcal{V}(k,S)) \right\}$$

$$\sqrt{n}\left[\widehat{\beta}(r,P) - \beta_r\right] \to^d Q(r,P)\left[\boldsymbol{\iota}_{T-k} \otimes \delta\boldsymbol{\psi}_P + N(\mathbf{0}, \mathcal{V}(r,P))\right]$$

$$\sqrt{n}\left[\widehat{\beta}(r,S) - \beta_r\right] \to^d Q(r,S)\left[\boldsymbol{\iota}_{T-k} \otimes \left(\delta\begin{bmatrix} \boldsymbol{\psi}_P \\ \psi_S \end{bmatrix} + \begin{bmatrix} \mathbf{0}_k \\ \tau \end{bmatrix}\right) + N(\mathbf{0}, \mathcal{V}(r,S))\right]$$

where $k = r - 1$, $\beta' = (\theta, \gamma_1, \ldots, \gamma_r, 0)$, $\beta'_r = (\theta, \gamma_1, \ldots, \gamma_r)$, $\mathcal{V}(k,\cdot) = Var[Z_i(k,\cdot)\Delta\mathbf{v}_i]$, $\mathcal{V}(r,\cdot) = Var[Z_i(r,\cdot)\Delta\mathbf{v}_i^+]$, $\widehat{Q}(\ell,\cdot) \to_p Q(\ell,\cdot)$, $\boldsymbol{\psi}_P = E[\mathbf{z}_{it}(r,P)\Delta y_{it-k}]$, $\psi_S = E[x_{it}\Delta y_{it-k}]$, $\mathbf{z}_{it}(\ell,\cdot)$ is as in Equation 28, $Z_i(\ell,\cdot) = diag\{\mathbf{z}'_{it}(\ell,\cdot)\}_{t=\ell+2}^{T}$ and $\boldsymbol{\iota}_d$ denotes a d-vector of ones.

To operationalize the GFIC, we need to provide appropriate estimators of all quantities that appear in Theorem 6.1. To estimate $Q(k,\mathrm{P})$, $Q(k,\mathrm{S})$, $Q(r,\mathrm{P})$, and $Q(r,\mathrm{S})$ we employ the usual sample analogues $\widehat{Q}(\cdot,\cdot)$ given above, which remain consistent under local mis-specification. There are many consistent estimators for the variance matrices $\mathcal{V}(k,\mathrm{P})$, $\mathcal{V}(k,\mathrm{S})$, $\mathcal{V}(r,\mathrm{P})$, $\mathcal{V}(r,\mathrm{S})$ under local mis-specification. In our simulations and empirical example below, we employ the usual heteroskedasticity-consistent, panel-robust variance matrix estimator. Because $E[\mathbf{z}_{it}(\ell,\mathrm{S})\Delta v_{it}] \neq 0$, we center our estimators of $\mathcal{V}(\ell,\mathrm{S})$ by subtracting the sample analogue of this expectation when calculating the sample variance. We estimate $\boldsymbol{\psi}_\mathrm{P}$ and ψ_S as follows

$$\widehat{\boldsymbol{\psi}}_\mathrm{P} = \frac{1}{n(T-k-1)}\sum_{t=k+2}^{T}\sum_{i=1}^{n}\mathbf{z}_{it}(r,\mathrm{P})\Delta y_{it-k} \qquad \widehat{\psi}_\mathrm{S} = \frac{1}{n(T-k-1)}\sum_{t=k+2}^{T}\sum_{i=1}^{n}x_{it}\Delta y_{it-k}$$

using our assumption of stationarity from above. The only remaining quantities we need to construct the GFIC involve the bias parameters δ and τ. We can read off an asymptotically unbiased estimator of δ directly from Theorem 6.1, namely $\widehat{\delta} = \sqrt{n}\ \widehat{\gamma}_k(k, \mathrm{P})$ the estimator of γ_k based on the instrument set that assumes only that x is pre-determined rather than strictly exogenous. To construct an asymptotically unbiased estimator of τ, we use the residuals from the specification that uses *both* the correct moment conditions and the correct lag specification, specifically

$$\widehat{\tau} = \left(\frac{\boldsymbol{\iota}'_{T-k-1}}{T-k-1} \right) n^{-1/2} X' \left[\Delta \mathbf{y} - W(k) \widehat{\beta}(k, \mathrm{P}) \right] \tag{31}$$

where $X' = [X_1 \cdots X_n]$ and $X_i = \mathrm{diag}\{x_{it}\}_{t=k+2}^{T}$. The following result gives the joint limiting behavior of $\widehat{\delta}$ and $\widehat{\tau}$, which we will use to construct the GFIC.

Theorem 6.2 (Joint Limit Distribution of $\widehat{\delta}$ and $\widehat{\tau}$). *Under the conditions of Theorem 6.1,*

$$\begin{bmatrix} \widehat{\delta} - \delta \\ \widehat{\tau} - \tau \end{bmatrix} \xrightarrow{d} \Psi N(\mathbf{0}, \Pi\, \mathcal{V}\,(k, S)\ \Pi')$$

where $\widehat{\delta} = \sqrt{n}[\mathbf{e}'_k\, \widehat{\beta}(k, P)]$, $\mathbf{e}_k = (0, \mathbf{0}'_{k-1}, 1)'$, $\widehat{\tau}$ is as defined in Equation 31,

$$\Psi = \begin{bmatrix} \left(\dfrac{\boldsymbol{\iota}'_{T-k-1}}{T-k-1} \right) \{\boldsymbol{\xi}'Q(k, P) \otimes \boldsymbol{\iota}'_{T-k-1}\} & \left(\dfrac{\boldsymbol{\iota}_{T-k-1}}{T-k-1} \right) \\ \mathbf{e}'_k Q(k, P) & \mathbf{0}'_{T-k-1} \end{bmatrix},$$

$\boldsymbol{\xi}' = E\left\{ x_{it} \begin{bmatrix} \Delta x_{it} & L\Delta y_{it} & \cdots & L^k \Delta y_{it} \end{bmatrix} \right\}$, *the variance matrix $\mathcal{V}(k, S)$ is as defined in Theorem 6.1, the permutation matrix $\Pi = \begin{bmatrix} \Pi'_1 & \Pi'_2 \end{bmatrix}'$ with $\Pi_1 = I_{T-k-1} \otimes \begin{bmatrix} I_{k+1} & \mathbf{0}_{k+1} \end{bmatrix}$ and $\Pi_2 = I_{T-k-1} \otimes \begin{bmatrix} \mathbf{0}'_{k+1} & 1 \end{bmatrix}$, $\boldsymbol{\iota}_d$ is a d-vector of ones and I_d the $(d \times d)$ identity matrix.*

To provide asymptotically unbiased estimators of the quantities τ^2, δ^2 and $\tau\delta$ that appear in the AMSE expressions for our estimators, we apply a bias correction to the asymptotically unbiased estimators of δ and τ from Theorem 6.2.

Corollary 6.1. *Let $\widehat{\Psi}$ be a consistent estimator of Ψ, defined in Theorem 6.2, and $\widehat{\mathcal{V}}(k, S)$ be a consistent estimator of $\mathcal{V}(k, S)$, defined in Theorem 6.1. Then, the elements of*

$$\begin{bmatrix} \widehat{\delta^2} & \widehat{\delta\tau} \\ \widehat{\tau\delta} & \widehat{\tau^2} \end{bmatrix} - \widehat{\Psi}\, \Pi\, \widehat{V}(k, S)\, \Pi'\, \widehat{\Psi}'$$

provide asymptotically unbiased estimators of of δ^2, τ^2 and $\tau\delta$, where Π is the permutation

matrix defined in Theorem 6.2.

We have already discussed consistent estimation of $\widehat{\mathcal{V}}(k, \mathrm{S})$. Since Π is a known permutation matrix, it remains only to propose a consistent estimator of Ψ. The matrix Ψ, in turn, depends only on $Q(k, \mathrm{P})$, and $\boldsymbol{\xi}'$. The sample analogue $\widehat{Q}(k, \mathrm{P})$ is a consistent estimator for $Q(k, \mathrm{P})$, as mentioned above, and

$$\widehat{\boldsymbol{\xi}}' = \frac{1}{n(T-k-1)} \sum_{t=k+2}^{T} \sum_{i=1}^{n} x_{it} \begin{bmatrix} \Delta x_{it} & L\Delta y_{it} & \cdots & L^k \Delta y_{it} \end{bmatrix} \tag{32}$$

is consistent for ξ'. We now have all the quantities needed to construct the GFIC for μ_{SR}, the short-run effect of x on y. Since $\mu_{SR} = \theta$, we can read off the AMSE expression for this parameter directly from Theorem 6.1. For the long-run effect μ_{LR}, however, we need to formally apply the Delta-method and account for the fact that the true value of γ_k is $\delta/\sqrt{n}$. Expressed as a function φ of the underlying model parameters,

$$\mu_{LR} = \varphi(\theta, \boldsymbol{\gamma}_r, \gamma_k) = \frac{\theta}{1 - \boldsymbol{\iota}_r' \boldsymbol{\gamma}_r - \gamma_k}$$

and the derivatives of φ are

$$\nabla \varphi(\theta, \boldsymbol{\gamma}_r, \gamma_k)' = \begin{bmatrix} \dfrac{\partial \varphi}{\partial \theta} & \dfrac{\partial \varphi}{\partial \boldsymbol{\gamma}_r'} & \dfrac{\partial \varphi}{\partial \gamma_k} \end{bmatrix} = \left(\frac{1}{1 - \boldsymbol{\iota}_r' \boldsymbol{\gamma}_r - \gamma_k} \right)^2 \begin{bmatrix} (1 - \boldsymbol{\iota}_r' \boldsymbol{\gamma}_r - \gamma_k) & \theta \, \boldsymbol{\iota}_r' & \theta \end{bmatrix}.$$

Using this notation, the limiting value of μ_{LR} is $\mu_{LR}^0 = \varphi(\theta, \boldsymbol{\gamma}_r, 0)$ while the true value is $\mu_{LR}^n = \varphi(\theta, \boldsymbol{\gamma}_r, \delta/\sqrt{n})$. Similarly, the limiting value of $\nabla \varphi$ is $\nabla \varphi_0 = \nabla \varphi(\theta, \boldsymbol{\gamma}_r, 0)'$, obtained by sputting zero in place of γ_r. We estimate this quantity consistently by plugging in the estimates from $\widehat{\beta}(k, \mathrm{P})$.

7 Slope Heterogeneity Example

We now specialize the GFIC to a model with slope heterogeneity. Our aim is to allow for individual specific coefficient β_i. The true data generating process is

$$y_{it} = (\beta + \eta_i)x_{it} + \epsilon_{it} \tag{33}$$

where $i = 1, \ldots, n$ indexes individuals and $t = 1, \ldots, T$ indexes time periods. We assume stationarity of x_{it} and ϵ_{it}. The error term ϵ_{it} is i.i.d over i and t with mean 0 and variance σ_ϵ^2. The individual effect η_i for slope heterogeneity is i.i.d over i with mean 0 and variance σ_η^2.

Homoscedasticity is assumed, that is, $E[\epsilon_{it}^2|x_{it}] = \sigma_\epsilon^2$ and $E[\eta_i^2|x_{it}] = \sigma_\eta^2$. Also, $E[\eta_i\epsilon_{it}|x_{it}] = 0$ is assumed. Stacking observations for a given individual over time in the usual way, let $\mathbf{y}_i = (y_{i1}, \ldots, y_{iT})'$ and define $\mathbf{x}_i$ analogously. For simplicity, I consider the case where $\beta \in \mathbb{R}$. First, we can consider the OLS estimator which does not take into account of slope heterogeneity.

$$\widehat{\beta}_{OLS} = \left(\sum_{i=1}^n \mathbf{x}_i'\mathbf{x}_i \right)^{-1} \left(\sum_{i=1}^n \mathbf{x}_i'\mathbf{y}_i \right) \tag{34}$$

If we know for certain that there is no slope heterogeneity, we would prefer to use the OLS estimator. However, when there is slope heterogeneity, the OLS estimator is biased. We can consider the "mean group" estimator given by

$$\widehat{\beta}_{MG} = \frac{1}{n} \sum_{i=1}^n \widehat{\beta}_i = \frac{1}{n} \sum_{i=1}^n \left(\mathbf{x}_i'\mathbf{x}_i \right)^{-1} \left(\mathbf{x}_i'\mathbf{y}_i \right) \tag{35}$$

where $\widehat{\beta}_i$ is the OLS estimator using individual i's data only. One should use the mean group estimator to achieve unbiasedness when there is slope heterogeneity. However, if the slope heterogeneity is sufficiently small, the OLS estimator could render the lower variance which compensates for its bias.[7] In this example, the local mis-specification assumption takes the form

$$\sum_{t=1}^T E[x_{it}^2\eta_i] = \frac{\tau}{\sqrt{n}} \tag{36}$$

where τ is fixed, unknown constant. In the limit, $\tau/\sqrt{n} \to 0$. However, this term is not zero for any finite sample size unless $\tau = 0$. An asymptotically unbiased estimator of τ for this example is given by

$$\widehat{\tau} = \frac{1}{\sqrt{n}} \sum_{i=1}^n \mathbf{x}_i'(\mathbf{y}_i - \mathbf{x}_i\widehat{\beta}_{MG}) \tag{37}$$

resulting in the following result, from which we will construct the GFIC.

Theorem 7.1 (OLS versus Mean Group Limit Distributions). *Under standard regularity conditions,*

$$\begin{bmatrix} \sqrt{n}(\widehat{\beta}_{OLS} - \beta) \\ \sqrt{n}(\widehat{\beta}_{MG} - \beta) \\ \widehat{\tau} \end{bmatrix} \xrightarrow{d} N\left(\begin{bmatrix} c\tau \\ 0 \\ \tau \end{bmatrix}, AVA' \right)$$

---

[7]Different from the example of fixed effects and random effects, the OLS estimator does not always have lower variance than the mean group estimator.

where $c = \left(E[\mathbf{x}_i'\mathbf{x}_i]\right)^{-1}$,

$$A = \begin{bmatrix} c & c & 0 \\ 0 & 0 & 1 \\ 1 & 1 & -c^{-1} \end{bmatrix}$$

and

$$V = \begin{bmatrix} \sigma_\eta^2 E[\mathbf{x}_i'\mathbf{x}_i\mathbf{x}_i'\mathbf{x}_i] & 0 & \sigma_\eta^2 E[\mathbf{x}_i'\mathbf{x}_i] \\ 0 & \sigma_\epsilon^2 E[\mathbf{x}_i'\mathbf{x}_i] & \sigma^2 \\ \sigma_\eta^2 E[\mathbf{x}_i'\mathbf{x}_i] & \sigma^2 & \sigma_\eta^2 + \sigma_\epsilon^2 E[(\mathbf{x}_i'\mathbf{x}_i)^{-1}] \end{bmatrix}$$

.

Proof. By expanding, we have

$$\sqrt{n}(\widehat{\beta}_{OLS} - \beta) = \left[\left(\frac{1}{n}\sum_{i=1}^n \mathbf{x}_i'\mathbf{x}_i\right)^{-1}, \left(\frac{1}{n}\sum_{i=1}^n \mathbf{x}_i'\mathbf{x}_i\right)^{-1}\right] \begin{bmatrix} n^{-1/2}\sum_{i=1}^n \mathbf{x}_i'\mathbf{x}_i\eta_i \\ n^{-1/2}\sum_{i=1}^n \mathbf{x}_i'\epsilon_i \end{bmatrix}$$

$$\sqrt{n}(\widehat{\beta}_{MG} - \beta) = n^{-1/2}\sum_{i=1}^n \eta_i + n^{-1/2}\sum_{i=1}^n (\mathbf{x}_i'\mathbf{x}_i)^{-1}\mathbf{x}_i'\epsilon_i$$

$$\widehat{\tau} = \begin{bmatrix} 1 & 1 & -\frac{1}{n}\sum_{i=1}^n \mathbf{x}_i'\mathbf{x}_i \end{bmatrix} \begin{bmatrix} n^{-1/2}\sum_{i=1}^n \mathbf{x}_i'\mathbf{x}_i\eta_i \\ n^{-1/2}\sum_{i=1}^n \mathbf{x}_i'\epsilon_i \\ n^{-1/2}\sum_{i=1}^n \eta_i + n^{-1/2}\sum_{i=1}^n (\mathbf{x}_i'\mathbf{x}_i)^{-1}\mathbf{x}_i'\epsilon_i \end{bmatrix}$$

The result follows, after some algebra, by applying the Lindeberg-Feller central limit theorem jointly to $n^{-1/2}\sum_{i=1}^n \mathbf{x}_i'\mathbf{x}_i\eta_i$, $n^{-1/2}\sum_{i=1}^n \mathbf{x}_i'\epsilon_i$, and $n^{-1/2}\sum_{i=1}^n \eta_i + n^{-1/2}\sum_{i=1}^n (\mathbf{x}_i'\mathbf{x}_i)^{-1}\mathbf{x}_i'\epsilon_i$. Then the joint distribution of $\widehat{\beta}_{OLS}, \widehat{\beta}_{MG}$ and $\widehat{\tau}$ is derived to be

$$\begin{bmatrix} \sqrt{n}(\widehat{\beta}_{OLS} - \beta) \\ \sqrt{n}(\widehat{\beta}_{MG} - \beta) \\ \widehat{\tau} \end{bmatrix} \to_d \begin{bmatrix} c & c & 0 \\ 0 & 0 & 1 \\ 1 & 1 & -c^{-1} \end{bmatrix} \left(\begin{bmatrix} \tau \\ 0 \\ 0 \end{bmatrix} + M\right)$$

where $c = (E[\mathbf{x}_i'\mathbf{x}_i])^{-1}$ and $M \sim N(0, V)$. $\qquad\square$

We see from Theorem 7.1 that

$$AMSE(\widehat{\beta}_{OLS}) = c^2\tau^2 + c^2\sigma_\eta^2 E[\mathbf{x}_i'\mathbf{x}_i\mathbf{x}_i'\mathbf{x}_i] + c\sigma_\epsilon^2 \tag{38}$$

$$AMSE(\widehat{\beta}_{MG}) = \sigma_\eta^2 + \sigma_\epsilon^2 E[(\mathbf{x}_i'\mathbf{x}_i)^{-1}]. \tag{39}$$

Also, $\widehat{\tau}^2 - \widehat{\sigma}_\tau^2$ provides an asymptotically unbiased estimator of τ^2, where $\widehat{\sigma}_\tau^2$ is the last diagonal element of AVA' above. For the variables in expectation such as $c^{-1} = E[\mathbf{x}_i'\mathbf{x}_i]$,

we can use sample mean as the consistent estimator. That is, $\widehat{c^{-1}} = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i'\mathbf{x}_i)$. We also need to construct consistent estimators for σ_ϵ^2 and σ_η^2. We use the following estimators:

$$\widehat{\sigma}_\epsilon^2 = \frac{1}{NT-1}\sum_{i=1}^{N}\sum_{t=1}^{T}(y_{it} - x_{it}\widehat{\beta}_{OLS})^2 \tag{40}$$

$$\widehat{\sigma}_\eta^2 = \frac{S_b}{n-1} - \frac{1}{n}\sum_{i=1}^{n}\widehat{\sigma}_\epsilon^2(\mathbf{x}_i'\mathbf{x}_i)^{-1} \tag{41}$$

where $S_b = \sum_{i=1}^{n}\widehat{\beta}_i\widehat{\beta}_i' - \frac{1}{n}\sum_{i=1}^{n}\widehat{\beta}_i\sum_{i=1}^{n}\widehat{\beta}_i'$.

8 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. In simulations for two panel examples, the GFIC performs well, as does our proposed averaging estimator that combines the random and fixed effects specifications. While we focus here on applications to panel data, 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 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).

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 CLT, $\sqrt{n}f_n(\gamma_0,\theta_0) - \sqrt{n}E\left[f(Z_{ni},\gamma_0,\theta_0)\right] \xrightarrow{d} (\mathcal{N}_g', \mathcal{N}_h')'$. Now, by a mean-value expansion around γ_n,

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

Hence, $\sqrt{n}f_n(\gamma_0,\theta_0) \xrightarrow{d} (\mathcal{N}_g', \mathcal{N}_h')' + (0', \tau')' - F_\gamma\delta$. The result follows by the continuous mapping theorem. $\quad\square$

Proof of Corollary 2.1. Since Ξ_c picks out only the components corresponding to g, For

the valid estimator,

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

Thus, $\sqrt{n}\left(\widehat{\beta}_v-\beta_0\right)\overset{d}{\to}-K_v\left(\mathcal{N}_g-G_\gamma\delta\right)$. Finally, by the definition of a matrix inverse,

$$K_vG_\gamma\delta=\begin{bmatrix}G_\gamma'W_{gg}G_\gamma & G_\gamma'W_{gg}G_\theta\\G_\theta'W_{gg}G_\gamma & G_\theta'W_{gg}G_\theta\end{bmatrix}^{-1}\begin{bmatrix}G_\gamma'W_{gg}G_\gamma\\G_\theta'W_{gg}G_\gamma\end{bmatrix}\delta=\begin{bmatrix}0\\\delta\end{bmatrix}.$$

$\square$

Proof of Corollary 3.2. For some $\bar{\gamma}$ between γ_0 and $\gamma_n=\gamma_0+\delta/\sqrt{n}$.

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

by a mean-value expansion. 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 $\sqrt{n}\left(\widehat{\mu}(b,c)-\mu_n\right)=\sqrt{n}\left(\widehat{\mu}(b,c)-\mu_0\right)-\sqrt{n}\left(\mu_n-\mu_0\right)$. $\square$

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

$$
\begin{aligned}
\sqrt{n}\left(\widehat{\mu}_v-\mu_n\right)\ &\overset{d}{\to}\ \nabla_\beta\varphi_0'\left\{\begin{bmatrix}0\\\delta\end{bmatrix}-K_v\mathcal{N}_g\right\}-\nabla_\gamma\varphi_0'\delta\\
&=\ -\nabla_\beta\varphi_0'K_v\mathcal{N}_g+\begin{bmatrix}\nabla_\theta\varphi_0' & \nabla_\gamma\varphi_0'\end{bmatrix}\begin{bmatrix}0\\\delta\end{bmatrix}-\nabla_\gamma\varphi_0'\delta\\
&=\ -\nabla_\beta(\gamma_0,\theta_0)'K_v\mathcal{N}_g.
\end{aligned}
$$

$\square$

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

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

we have $\sqrt{n}h_n(\gamma_0,\theta_0)\overset{d}{\to}\mathcal{N}_h+\tau-H_\gamma\delta$. Substituting,

$$\sqrt{n}h_n(\widehat{\beta}_v)\overset{d}{\to}\mathcal{N}_h+\tau-H_\gamma\delta+H\left(-K_v\mathcal{N}_g+\begin{bmatrix}0\\\delta\end{bmatrix}\right)=\tau-HK_v\mathcal{N}_g+\mathcal{N}_h.$$

$\square$

Proof of Corollary 3.5. Define $(U',V')'=(\delta',\tau')'+\Psi(\mathcal{N}_g',\mathcal{N}_h')'$. By the Continuous Mapping Theorem and Theorem 3.1,

$$\begin{bmatrix}\widehat{\delta}\\\widehat{\tau}\end{bmatrix}\begin{bmatrix}\widehat{\delta}' & \widehat{\tau}'\end{bmatrix}\overset{d}{\to}\begin{bmatrix}U\\V\end{bmatrix}\begin{bmatrix}U' & V'\end{bmatrix}$$

The result follows since

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

$\square$

Proof of Theorem 5.1. This proof is standard so we provide only a sketch. First, let $A_n = (n^{-1}\sum_{i=1}^n \mathbf{x}_i'\widehat{\Omega}^{-1}\mathbf{x}_i)$, $B_n = (n^{-1}\sum_{i=1}^n \mathbf{x}_i'Q\mathbf{x}_i)$, and $C_n = T\widehat{\sigma}_\alpha^2 + \widehat{\sigma}_\varepsilon^2$. Now, expanding $\widehat{\beta}_{FE}$, β_{RE}, and $\widehat{\tau}$ and re-arranging

$$\begin{bmatrix} \sqrt{n}(\widehat{\beta}_{RE} - \beta) \\ \sqrt{n}(\widehat{\beta}_{FE} - \beta) \\ \widehat{\tau} \end{bmatrix} = \begin{bmatrix} A_n^{-1} & 0 \\ 0 & B_n^{-1} \\ C_n & -C_n A_n^{-1} B_n^{-1} \end{bmatrix} \begin{bmatrix} n^{-1/2}\sum_{i=1}^n \mathbf{x}_i'\widehat{\Omega}\mathbf{v}_i \\ n^{-1/2}\sum_{i=1}^n \mathbf{x}_i'Q\mathbf{v}_i \end{bmatrix}.$$

The result follows by applying a law of large numbers to A_n, B_n, C_n, and $\widehat{\Omega}$ and the Lindeberg-Feller CLT jointly to $n^{-1/2}\sum_{i=1}^n \mathbf{x}_i'Q\mathbf{v}_i$ and $n^{-1/2}\sum_{i=1}^n \mathbf{x}_i'\Omega^{-1}\mathbf{v}_i$. $\square$

Proof of Theorem 6.1. This proof is standard, so we provide only a sketch. Expanding,

$$\sqrt{n}\left[\widehat{\beta}(k,\cdot) - \beta\right] = (0, \mathbf{0}_{k-1}', \delta)' + \widehat{Q}(k,\cdot)\left[n^{-1/2}Z'(k,\cdot)\Delta\mathbf{v}\right]$$

$$\sqrt{n}\left[\widehat{\beta}(r,\cdot) - \beta_r\right] = \widehat{Q}(r,\cdot)\left\{\delta\left[n^{-1}Z'(r,\cdot)L^k\Delta\mathbf{y}^+\right] + \left[n^{-1/2}Z'(r,\cdot)\Delta\mathbf{v}^+\right]\right\}$$

The result follows, after some algebra, by applying the Lindeberg-Feller CLT to $n^{-1/2}Z'(k,\cdot)\Delta\mathbf{v}$ and $n^{-1/2}Z'(r,\cdot)\Delta\mathbf{v}^+$ and an appropriate law of large numbers to $n^{-1}Z'(r,\cdot)L^k\Delta\mathbf{y}^+$, where $(\cdot)$ is either P or S depending on the instrument set used. $\square$

Proof of Theorem 6.2. Expanding $\widehat{\beta}(k,\mathrm{P})$ from Equation 30

$$n^{-1/2}X'\left[\Delta\mathbf{y} - W(k)\widehat{\beta}(k,\mathrm{P})\right] = n^{-1/2}X'\left[\Delta\mathbf{v} - W(k)\widehat{Q}(k,\mathrm{P})n^{-1}Z'(k,\mathrm{P})\Delta\mathbf{v}\right]$$

$$= \begin{bmatrix} -n^{-1}X'W(k)\widehat{Q}(k,\mathrm{P}) & I_{T-k-1} \end{bmatrix}\begin{bmatrix} n^{-1/2}Z'(k,\mathrm{P})\Delta\mathbf{v} \\ n^{-1/2}X'\Delta\mathbf{v} \end{bmatrix}$$

using $\Delta\mathbf{y} = W(k)\beta_n + \Delta\mathbf{v}$. Similarly, expanding $\widehat{\beta}(k,\mathrm{P})$ from Equation 30,

$$\sqrt{n}\left[\widehat{\beta}(k,\mathrm{P})\right] = \begin{bmatrix} 0 & \mathbf{0}_{k-1} & \delta \end{bmatrix} + \widehat{Q}(k,\mathrm{P})n^{-1/2}Z'(k,\mathrm{P})\Delta\mathbf{v}$$

and since $\widehat{\delta}$ is $\sqrt{n}$ times the kth element of $\widehat{\beta}(k,\mathrm{P})$ and the kth element β is zero, we have

$$\widehat{\delta} = \delta + \begin{bmatrix} 0 & \mathbf{0}_{k-1}' & 1 \end{bmatrix}\widehat{Q}(k,\mathrm{P})n^{-1/2}Z'(k,\mathrm{P})\Delta\mathbf{v}.$$

By a law of large numbers $\widehat{Q}(k,\mathrm{P}) \to_p Q(k,\mathrm{P})$ and $n^{-1}X'W(k) \to_p \boldsymbol{\xi}' \otimes \boldsymbol{\iota}_{T-k-1}$, hence

$$\begin{bmatrix} \widehat{\delta} - \delta \\ \widehat{\tau} \end{bmatrix} = \Psi\begin{bmatrix} n^{-1/2}Z'(k,\mathrm{P})\Delta\mathbf{v} \\ n^{-1/2}X'\Delta\mathbf{v} \end{bmatrix} + o_p(1).$$

The result follows, after some algebra, by applying the Lindeberg-Feller CLT jointly to $n^{-1/2}Z'(k,\mathrm{P})\Delta\mathbf{v}$ and $n^{-1/2}X'\Delta\mathbf{v}$ and noting that the permutation matrix Π maps the vector $[n^{-1/2}Z'(k,\mathrm{S})\Delta\mathbf{v}]$ to the vector $\left[\ \left\{n^{-1/2}Z'(k,P)\Delta\mathbf{v}\right\}'\ \left\{n^{-1/2}X'\Delta\mathbf{v}\right\}'\ \right]'$. $\qquad\square$

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.

Arellano, M., Bond, S., 1991. Some tests of specification for panel data: Monte carlo evidence and an application to employment equations. The review of economic studies 58 (2), 277–297.

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., August 2015. Using invalid instruments on purpose: Focused moment selection and averaging for GMM. Tech. rep., University of Pennsylvania.

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.