

# A Generalized Focused Information Criterion for GMM with Applications to Panel Data Models\*

Minsu Chang    Francis J. DiTraglia<sup>†</sup>

University of Pennsylvania

This Version: February 1, 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.

Update this!

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

**JEL Codes:** C23, C52

## 1 Introduction

Update references, add additional references and tweak intro to refer to the new material.

---

\*We thank Manuel Arellano, Otilia Boldea, Bruce Hansen, Frank Kleibergen, and seminar participants at the 2013 Latin American Workshop in Econometrics, the 2014 Midwest Econometrics Group meetings, Tilburg, the Tinbergen Institute, and the University of Wisconsin for helpful comments and suggestions.

<sup>†</sup>Corresponding Author: [fditra@sas.upenn.edu](mailto:fditra@sas.upenn.edu), 3718 Locust Walk, Philadelphia, PA 19104

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  $\mu$  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  $\mu$ . But the very fact that the baseline assumptions do not raise eyebrows suggests that they may not be especially informative about  $\mu$ . On the other hand are one or more stronger controversial “suspect” assumptions. These stronger assumptions are expected to be much more informative about  $\mu$ . 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 (2016) 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 (2016) 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 (2016) 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.<sup>1</sup> 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 incorporating 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 (2016) to the more general setting of this paper. Section 5.1 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

---

<sup>1</sup>See Smith (1992) for an approach to GMM model selection based on non-nested hypothesis testing.

the two examples and Section ??? concludes. Proofs appear in the Appendix.

## 2 Notation and Asymptotic Framework

After Minsu adds the high-level sufficient conditions, put them into a numbered assumption and update the theorems etc to refer to them

Let  $f(\cdot, \cdot)$  be a  $(p + q)$ -vector of moment functions of a random vector  $Z$  and an  $(r + s)$ -dimensional parameter vector  $\beta$ . 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  $\gamma$  is an  $r$ -vector and  $\theta$  an  $s$ -vector of parameters. The model selection problem is to decide which if any of the elements of  $\gamma$  to estimate, and which to set equal to the corresponding elements of  $\gamma_0$ , an  $r$ -vector of known constants. The parameters contained in  $\theta$  are those that we always estimate, the “protected” parameters. Any specification that does not estimate the full parameter vector  $\beta$  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  $\{Z_{ni} : 1 \leq i \leq n, n = 1, 2, \dots\}$  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)  $\{f(Z_{ni}, \gamma_n, \theta_0) : 1 \leq i \leq n, n = 1, 2, \dots\}$  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  $\delta$  an unknown  $r$ -vector of constants and  $\tau_n = n^{-1/2}\tau$  with  $\tau$  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  $n \rightarrow \infty$ . Unless some elements of  $\delta$  are zero, any estimator that restricts  $\gamma$  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  $\beta_n$  depends on the unknown constant vector  $\tau$ , 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  $\beta$  are correctly specified. In the limit, however, *every* estimator is correctly specified, regardless of which elements of  $\gamma$  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 from Assumption 2.1, although it is not strictly necessary.

**Assumption 2.2.**  $\{Z_{ni} : 1 \leq i \leq n, n = 1, 2, \dots\}$  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  $\gamma$  we have chosen to estimate. When  $b = 1_r$ , where  $1_m$  represents an  $m$ -vector of ones, we estimate both  $\theta$  and the full vector  $\gamma$ . When  $b = 0_r$ , where  $0_m$  denotes an  $m$ -vector of zeros, we estimate only  $\theta$ , setting  $\gamma = \gamma_0$ . More generally, we estimate  $|b|$  components of  $\gamma$  and set the others equal to the corresponding elements of  $\gamma_0$ . Let  $\gamma^{(b)}$  be the  $|b|$ -dimensional subvector of  $\gamma$  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  $\gamma$  that are *not* estimated. Analogously, let  $c = (c'_g, c'_h)'$  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  $\Xi_b$  and  $\Xi_c$ . Multiplying  $\beta$  by the  $(|b| + s) \times (r + s)$  *model selection matrix*  $\Xi_b$  extracts the elements corresponding to  $\theta$  and the subset of  $\gamma$  indicated by the model selection vector  $b$ . Thus  $\Xi_b \beta = (\gamma^{(b)'}, \theta')'$ . Similarly, multiplying a vector by the  $|c| \times (p + q)$  moment selection matrix  $\Xi_c$  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,

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

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

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

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

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

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:

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

Because it is correctly specified both for finite  $n$  and in the limit, the valid estimator contains the information we use to identify  $\tau$  and  $\delta$ , 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  $\theta_0$  under standard, high level regularity conditions.

Add high-level sufficient conditions! These should be an explicit assumption.

Essential differences arise, however, when we consider their respective asymptotic distributions. Let

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

partitioned according to

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

and define

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

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  $(\gamma_0, \theta_0)$ . Finally, let  $F(b, c) = \Xi_c F \Xi'_c$  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  $\delta$  and  $\tau$  induce a bias term in the limiting distribution of  $\sqrt{n} \left( \widehat{\beta}(b, c) - \beta_0^{(b)} \right)$ . The key result is as follows.

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

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

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

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

and

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

Because it employs the correct specification, the valid estimator of  $\theta$  shows no asymptotic bias. Moreover, the valid estimator of  $\gamma$  has an asymptotic distribution that is centered around  $\delta$ , 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  $W_{gg} = \text{plim } \widetilde{W}_{gg}$ .

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  $\varphi$  is a real-valued, almost surely continuous function of the



underlying model parameters  $\theta$  and  $\gamma$ . Let  $\mu_n = \varphi(\gamma_n, \theta_0)$  and define  $\mu_0$  and  $\hat{\mu}(b, c)$  analogously. By Theorem 2.1 and the delta method, we have the following result.

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

$$\sqrt{n}(\hat{\mu}(b, c) - \mu_0) \xrightarrow{d} -\nabla_{\beta}\varphi'_0\Xi'_bK(b, c)\Xi_c \left( \mathcal{N} + \begin{bmatrix} 0 \\ \tau \end{bmatrix} - F_{\gamma}\delta \right)$$

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

The true value of  $\mu$ , however, is  $\mu_n$  rather than  $\mu_0$  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,*

$$\sqrt{n}(\hat{\mu}(b, c) - \mu_n) \xrightarrow{d} -\nabla_{\beta}\varphi'_0\Xi'_bK(b, c)\Xi_c \left( \mathcal{N} + \begin{bmatrix} 0 \\ \tau \end{bmatrix} - F_{\gamma}\delta \right) - \nabla_{\gamma}\varphi'_0\delta$$

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

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

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

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

where  $\hat{\mu}_v = \varphi(\hat{\theta}_v, \hat{\gamma}_v)$ . Thus, the valid estimator  $\hat{\mu}_v$  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  $\hat{\mu}(b, c)$  is as follows,

$$\text{AMSE}(\hat{\mu}(b, c)) = \text{AVAR}(\hat{\mu}(b, c)) + \text{BIAS}(\hat{\mu}(b, c))^2 \quad (11)$$

where

$$\text{AVAR}(\hat{\mu}(b, c)) = \nabla_{\beta} \varphi_0' \Xi_b' K(b, c) \Omega_c K(b, c)' \Xi_b \nabla_{\beta} \varphi_0 \quad (12)$$

$$\text{BIAS}(\hat{\mu}(b, c)) = -\nabla_{\beta} \varphi_0' M(b, c) \begin{bmatrix} \delta \\ \tau \end{bmatrix} \quad (13)$$

and

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

The idea behind the GFIC is to construct an estimate  $\widehat{\text{AMSE}}(\hat{\mu}(b, c))$  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  $\hat{\beta}(b, c)$ , the usual sample analogues provide consistent estimators of  $K(b, c)$  and  $F_{\gamma}' = (G_{\gamma}', H_{\gamma}')$  under Assumption 2.1, and  $\varphi(\hat{\theta}_v, \gamma_0)$  is consistent for  $\varphi_0$ . Consistent estimators of  $\Omega$  are also readily available under local mis-specification although the best choice may depend on the situation.<sup>2</sup> Since  $\gamma_0$  is known, as are  $\Xi_b$  and  $\Xi_c$ , only  $\delta$  and  $\tau$  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  $\tau$  and  $\delta$  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  $\tau$  and  $\delta$ . Corollary 2.1 provides the required estimator for  $\delta$ , namely  $\hat{\delta} = \sqrt{n}(\hat{\gamma}_v - \gamma_0)$ .

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

$$\hat{\delta} = \sqrt{n}(\hat{\gamma}_v - \gamma_0) \xrightarrow{d} \delta - K_v^{\gamma} \mathcal{N}_g$$

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

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

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

$$\hat{\tau} = \sqrt{n}h_n(\hat{\beta}_v) \xrightarrow{d} \tau - HK_v \mathcal{N}_g + \mathcal{N}_h$$

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

---

<sup>2</sup>We discuss this in more detail for our examples in Section 5 below.

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,*

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

where

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

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

From Equation 13,

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

Thus, the bias parameters  $\tau$  and  $\delta$  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  $\tau$  and  $\delta$ , 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  $\Psi$  and  $\Omega$ . Then,*

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

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  $\mu$  under each specification  $(b, c) \in \mathcal{BC}$

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

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

Re-vamp this section a bit. Say that the main focus of the paper is selection but we also extend the inference results of DiTraglia 2016. Add proof of the main result. Format this section similarly to the earlier paper.

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 ? to the more general setting considered in this paper, one that allows for simultaneous model and moment selection.<sup>3</sup> Consider an estimator of the form

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

where  $\hat{\mu}$  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  $\hat{\omega}(b,c)$  denotes a collection of data-dependent weights satisfying the following assumption.

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

- (a)  $\sum_{(b,c) \in \mathcal{BC}} \hat{\omega}(b,c) = 1$
- (b)  $\hat{\omega}(b,c) \xrightarrow{d} \psi(\mathcal{N}, \delta, \tau | b, c)$  jointly for all  $(b,c) \in \mathcal{BC}$  where  $\psi$  is a function of the normal random vector  $\mathcal{N}$ , the bias parameters  $\delta$  and  $\tau$ , 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  $\hat{\mu}$  as follows.

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

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

---

<sup>3</sup>Because the conceptual issues are largely the same as in the case where one considers only moment selection, we direct the reader to ? for more discussion.

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  $\tau$  and  $\delta$  were known, the story would end here. We could simply substitute consistent estimators of  $K$  and  $M$ , and then repeatedly draw  $\mathcal{N} \sim N(0, \hat{\Omega})$ , where  $\hat{\Omega}$  is a consistent estimator of  $\Omega$ , and thus tabulate the distribution of  $\Lambda$  to arbitrary precision. Unfortunately, no consistent estimators of  $\tau$  or  $\delta$  exist: all we have at our disposal are asymptotically unbiased estimators. Simply plugging in these estimators  $\hat{\tau}$  and  $\hat{\delta}$  and proceeding with the simulation is not guaranteed to lead to valid confidence intervals.<sup>4</sup> 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  $\hat{\mu}$ ).

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

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

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

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

3. Define

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

---

<sup>4</sup>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 ?.

4. The confidence interval for  $\mu$  is given by

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

**Theorem 4.1** (Simulation-based Confidence Interval for  $\hat{\mu}$ ). *Let  $\nabla_{\beta}\hat{\varphi}_0$ ,  $\hat{\psi}(\cdot|b, c)$ ,  $\hat{K}(b, c)$  and  $\hat{M}(b, c)$  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  $(\delta, \tau)$  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  $J, n \rightarrow \infty$ .*

## 5 Examples of the GFIC in Panel Data Models

Add short paragraph introducing the examples.

### 5.1 Random Effects versus Fixed Effects Example

In this section we consider a simple example in which 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

$$y_{it} = \beta x_{it} + v_{it} \tag{18}$$

$$v_{it} = \alpha_i + \varepsilon_{it} \tag{19}$$

for  $i = 1, \dots, n$ ,  $t = 1, \dots, T$  where  $\varepsilon_{it}$  is iid across  $i, t$  with  $Var(\varepsilon_{it}) = \sigma_{\varepsilon}^2$  and  $\alpha_i$  is iid across  $i$  with  $Var(\alpha_i) = \sigma_{\alpha}^2$ . Stacking observations for a given individual over time in the usual way, let  $\mathbf{y}_i = (y_{i1}, \dots, y_{iT})'$  and define  $\mathbf{x}_i, \mathbf{v}_i$  and  $\boldsymbol{\varepsilon}_i$  analogously. Our goal in this example is to estimate  $\beta$ , 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  $\alpha_i$ . 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

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

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

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

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

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

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

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

$$\sum_{t=1}^T E[x_{it} \alpha_i] = \frac{\tau}{\sqrt{n}} \quad (23)$$

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

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

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  $(\mathbf{x}_{ni}, \alpha_{ni}, \boldsymbol{\epsilon}_{ni})$  be an iid triangular array of random variables such that  $Var(\boldsymbol{\epsilon}_i | \mathbf{x}_{ni}, \alpha_{ni}) \rightarrow \sigma_\epsilon^2 I_T$ ,  $E[\mathbf{x}_{it}' Q \boldsymbol{\epsilon}_{it}] = 0$ ,*

and  $E[\alpha_i \mathbf{l}'_T \mathbf{x}_{it}] = \tau/\sqrt{n}$  for all  $n$ . Then, under standard regularity conditions,

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

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

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

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

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



where we take the maximum over  $\hat{\tau} - \hat{\sigma}^2$  and zero so that  $\hat{\omega}^*$  is between zero and one.

## 5.2 Dynamic Panel Example

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

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

where  $i = 1, \dots, n$  indexes individuals and  $t = 1, \dots, T$  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.<sup>5</sup> The unobserved error  $\eta_i$  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  $|\gamma| < 1$  where  $\gamma = (\gamma_1, \dots, \gamma_k)'$ . Our goal is to estimate one of the following two target parameters with minimum MSE:

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

where  $\mu_{SR}$  denotes the short-run effect and  $\mu_{LR}$  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  $\mu_{SR}$  or  $\mu_{LR}$ . 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  $\eta_i$  we work in first differences. Defining  $\Delta$  in the usual

---

<sup>5</sup>Alternatively, we can simply de-mean and project out any time-varying exogenous covariates after taking first-differences.

way, so that  $\Delta y_{it} = y_{it} - y_{it-1}$  and so on, we can write Equation 25 as

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

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

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

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  $\ell$  as a placeholder for the lag length used in estimation. If  $\ell = 0$ ,  $\mathbf{z}'_{it}(0, P) = x_{it-1}$  and  $\mathbf{z}'_{it}(0, S) = (x_{it-1}, x_{it})$ . 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  $Z_i(\ell, P) = \text{diag} \{ \mathbf{z}'_{it}(\ell, P) \}_{t=\ell+2}^T$  and  $Z_i(\ell, S) = \text{diag} \{ \mathbf{z}'_{it}(\ell, S) \}_{t=\ell+2}^T$ . 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  $\mu_{SR}$  and  $\mu_{LR}$  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  $k^{\text{th}}$  lag local to zero. Let  $\boldsymbol{\gamma}' = (\gamma_1, \dots, \gamma_{k-1}, \gamma_k)$  denote the full vector of lag coefficients that  $\boldsymbol{\gamma}_r = (\gamma_1, \dots, \gamma_r)$  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  $\beta$  that excludes the  $k^{\text{th}}$  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  $\Delta \mathbf{y}_i = [\Delta y_{i,k+2}, \dots, \Delta y_{iT}]'$  and  $\Delta \mathbf{y}_i^+ = [\Delta y_{i,k+1}, \Delta y_{i,k+2}, \dots, \Delta y_{iT}]'$  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  $L^k \Delta \mathbf{y}_i^+ = [\Delta y_{i1}, \Delta y_{i2}, \dots, \Delta y_{iT-k}]'$  where  $L^k$  denotes the element-wise application of the  $k^{\text{th}}$  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:

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

Note that  $W_i^{+'}(r)$  contains one more row than  $W_i(k)$  but  $W_i(k)$  contains one more column than  $W_i^{+'}(r)$ : removing the  $k^{\text{th}}$  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  $W_\ell^+$  analogously, where  $\ell$  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

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

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

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

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  $\widehat{\beta}(k, P)$ ,  $\widehat{\beta}(k, S)$ ,  $\widehat{\beta}(r, P)$ , and  $\widehat{\beta}(r, S)$  which we will use to construct the GFIC.

**Theorem 5.2** (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,*

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

where  $k = r - 1$ ,  $\beta' = (\theta, \gamma_1, \dots, \gamma_r, 0)$ ,  $\beta'_r = (\theta, \gamma_1, \dots, \gamma_r)$ ,  $\mathcal{V}(k, \cdot) = \text{Var}[Z_i(k, \cdot) \Delta \mathbf{v}_i]$ ,  $\mathcal{V}(r, \cdot) = \text{Var}[Z_i(r, \cdot) \Delta \mathbf{v}_i^+]$ ,  $\widehat{Q}(\ell, \cdot) \rightarrow_p Q(\ell, \cdot)$ ,  $\boldsymbol{\psi}_P = E[\mathbf{z}_{it}(r, P) \Delta y_{it-k}]$ ,  $\psi_S = E[x_{it} \Delta y_{it-k}]$ ,  $\mathbf{z}_{it}(\ell, \cdot)$  is as in Equation 28,  $Z_i(\ell, \cdot) = \text{diag}\{\mathbf{z}'_{it}(\ell, \cdot)\}_{t=\ell+2}^T$  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 5.2. To estimate  $Q(k, P)$ ,  $Q(k, S)$ ,  $Q(r, P)$ , and  $Q(r, S)$  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  $\mathcal{V}(k, P)$ ,  $\mathcal{V}(k, S)$ ,  $\mathcal{V}(r, P)$ ,  $\mathcal{V}(r, S)$  under local mis-specification. In our simulations and empirical example below, we employ the usual heteroskedasticity-consistent, panel-robust variance matrix estimator. Because  $E[\mathbf{z}_{it}(\ell, S) \Delta v_{it}] \neq 0$ , we center our estimators of  $\mathcal{V}(\ell, S)$  by subtracting the sample analogue of this expectation when calculating the sample variance. We estimate  $\boldsymbol{\psi}_P$  and  $\psi_S$  as follows

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

using our assumption of stationarity from above. The only remaining quantities we need to construct the GFIC involve the bias parameters  $\delta$  and  $\tau$ . We can read off an asymptotically unbiased estimator of  $\delta$  directly from Theorem 5.2, namely  $\widehat{\delta} = \sqrt{n} \widehat{\gamma}_k(k, P)$  the estimator of  $\gamma_k$  based on the instrument set that assumes only that  $x$  is pre-determined rather than strictly exogenous. To construct an asymptotically unbiased estimator of  $\tau$ , we use the residuals from the specification that uses *both* the correct moment conditions and the correct lag specification, specifically

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

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

**Theorem 5.3** (Joint Limit Distribution of  $\widehat{\delta}$  and  $\widehat{\tau}$ ). *Under the conditions of Theorem 5.2,*

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

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,

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

$\boldsymbol{\xi}' = E \left\{ x_{it} \begin{bmatrix} \Delta x_{it} & L \Delta y_{it} & \cdots & L^k \Delta y_{it} \end{bmatrix} \right\}$ , the variance matrix  $\mathcal{V}(k, S)$  is as defined in Theorem 5.2, 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  $\Pi_2 = I_{T-k-1} \otimes \begin{bmatrix} \mathbf{0}'_{k+1} & 1 \end{bmatrix}$ ,  $\boldsymbol{\nu}_d$  is a  $d$ -vector of ones and  $I_d$  the  $(d \times d)$  identity matrix.

To provide asymptotically unbiased estimators of the quantities  $\tau^2$ ,  $\delta^2$  and  $\tau\delta$  that appear in the AMSE expressions for our estimators, we apply a bias correction to the asymptotically unbiased estimators of  $\delta$  and  $\tau$  from Theorem 5.3.

**Corollary 5.1.** *Let  $\widehat{\Psi}$  be a consistent estimator of  $\Psi$ , defined in Theorem 5.3, and  $\widehat{\mathcal{V}}(k, S)$  be a consistent estimator of  $\mathcal{V}(k, S)$ , defined in Theorem 5.2. Then, the elements of*

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

provide asymptotically unbiased estimators of  $\delta^2$ ,  $\tau^2$  and  $\tau\delta$ , where  $\Pi$  is the permutation matrix defined in Theorem 5.3.

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

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

is consistent for  $\boldsymbol{\xi}'$ . We now have all the quantities needed to construct the GFIC for  $\mu_{SR}$ , 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 5.2. For the long-run effect  $\mu_{LR}$ , however, we need to formally apply the Delta-method and account for the fact that the true value of  $\gamma_k$  is  $\delta/\sqrt{n}$ . Expressed as a function  $\varphi$  of the underlying model parameters,

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

and the derivatives of  $\varphi$  are

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

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

### 5.3 Slope Heterogeneity Example

This section needs to be tightened up and clarified. It's not very clear what the point of it is and why this example is interesting. There are also no references.

We now consider a simple panel model with heterogeneous effects, namely

$$y_{it} = \beta_i x_{it} + \epsilon_{it} \tag{33}$$

$$\beta_i = \beta + \eta_i, \quad \eta_i \sim \text{iid}(0, \sigma_\eta^2) \tag{34}$$

where  $i = 1, \dots, n$  indexes individuals and  $t = 1, \dots, T$  indexes time periods. As above we assume without loss of generality that all random variables are mean zero and any exogenous controls have been projected out. To simplify the exposition we further assume that  $\epsilon_{it}$  is iid over both  $i$  and  $t$  with variance  $\sigma_\epsilon^2$ .

The individual effect  $\eta_i$  for slope heterogeneity is i.i.d over  $i$  with mean 0 and variance  $\sigma_\eta^2$ . 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  $\mathbf{y}_i = (y_{i1}, \dots, y_{iT})'$  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.

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

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

$$\hat{\beta}_{MG} = \frac{1}{n} \sum_{i=1}^n \hat{\beta}_i = \frac{1}{n} \sum_{i=1}^n \left( \mathbf{x}_i' \mathbf{x}_i \right)^{-1} \left( \mathbf{x}_i' \mathbf{y}_i \right) \tag{36}$$

where  $\hat{\beta}_i$  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.<sup>6</sup> In this example, the local mis-specification assumption takes the form

$$\sum_{t=1}^T E[x_{it}^2 \eta_i] = \frac{\tau}{\sqrt{n}} \quad (37)$$

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

$$\hat{\tau} = \frac{1}{\sqrt{n}} \sum_{i=1}^n \mathbf{x}_i' (\mathbf{y}_i - \mathbf{x}_i \hat{\beta}_{MG}) \quad (38)$$

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

**Theorem 5.4** (OLS versus Mean Group Limit Distributions). *Under standard regularity conditions,*

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

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

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

and

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

.

---

<sup>6</sup>Different from the example of fixed effects and random effects, the OLS estimator does not always have lower variance than the mean group estimator.

**Proof.** By expanding, we have

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

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  $\hat{\beta}_{OLS}$ ,  $\hat{\beta}_{MG}$  and  $\hat{\tau}$  is derived to be

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

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

We see from Theorem 5.4 that

$$AMSE(\hat{\beta}_{OLS}) = c^2 \tau^2 + c^2 \sigma_\eta^2 E[\mathbf{x}'_i \mathbf{x}_i \mathbf{x}'_i \mathbf{x}_i] + c \sigma_\epsilon^2 \quad (39)$$

$$AMSE(\hat{\beta}_{MG}) = \sigma_\eta^2 + \sigma_\epsilon^2 E[(\mathbf{x}'_i \mathbf{x}_i)^{-1}]. \quad (40)$$

Also,  $\hat{\tau}^2 - \hat{\sigma}_\tau^2$  provides an asymptotically unbiased estimator of  $\tau^2$ , where  $\hat{\sigma}_\tau^2$  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  $\sigma_\epsilon^2$  and  $\sigma_\eta^2$ . We use the following estimators:

$$\hat{\sigma}_\epsilon^2 = \frac{1}{NT-1} \sum_{i=1}^N \sum_{t=1}^T (y_{it} - x_{it} \hat{\beta}_{OLS})^2 \quad (41)$$

$$\hat{\sigma}_\eta^2 = \frac{S_b}{n-1} - \frac{1}{n} \sum_{i=1}^n \hat{\sigma}_\epsilon^2 (\mathbf{x}'_i \mathbf{x}_i)^{-1} \quad (42)$$

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



## 6 Simulation Results

We now evaluate the performance of the GFIC in a number of simulation experiments based on the Examples from Section 5.

### 6.1 Fixed vs. Random Effects Example

We employ a simulation design similar to that used by [Guggenberger \(2010\)](#), namely

$$y_{it} = 0.5x_{it} + \alpha_i + \varepsilon_{it}$$

where

$$\begin{bmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{iT} \\ \alpha_i \end{bmatrix} \stackrel{\text{iid}}{\sim} N \left( \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \rho & \dots & \rho & \gamma \\ \rho & 1 & \dots & \rho & \gamma \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \rho & \rho & \dots & 1 & \gamma \\ \gamma & \gamma & \dots & \gamma & 1 \end{bmatrix} \right)$$

independently of  $(\varepsilon_{i1}, \dots, \varepsilon_{iT})' \sim \text{iid } N(0, \sigma_\varepsilon^2 \mathbf{I}_T)$ . In this design,  $\gamma$  controls the correlation between  $x_{it}$  and the individual effects  $\alpha_i$ , while  $\rho$  controls the persistence of  $x_{it}$  over time. Larger values of  $\gamma$  correspond to larger violations of the assumption underlying the random effects estimator, increasing its bias. Larger values of  $\rho$ , on the other hand, decrease the amount of variation within individuals, thus *increasing* the variance of the fixed effects estimator. Figures 1 and 2 present the RMSE under this simulation design of the random effects GLS estimator and the fixed effects estimator along with those of the post-GFIC and averaging estimators described above in Section 5.1 over a grid of values for  $T$ ,  $\gamma$ ,  $\rho$ , and sample size  $N$ . All calculations are based on 10,000 simulation replications.<sup>7</sup> In the interest of space, we present only results for  $\sigma_\varepsilon^2 = 2.5$  and our “coarse” parameter grid for  $\rho$  here. Additional simulation results are available upon request.

We see from Figures 1 and 2 that, regardless of the configuration of the other parameter values, there is always a range of values for  $\gamma$  for which the random effects estimator has a smaller RMSE than the fixed effects estimator. The width of this range increases as either the number of individuals  $N$  or number of time periods  $T$  decrease. It also increases as the persistence  $\rho$  of  $x_{it}$  increases. Indeed, when  $N$  and  $T$  are relatively small and  $\rho$  is relatively large, the individual effects  $\alpha_i$  can be *strongly* correlated with  $x_{it}$  and still result in a random effects estimator with a lower RMSE than the fixed effects estimator. The

---

<sup>7</sup>When  $T = 5$ , setting  $\rho = 0.3$  violates positive definiteness so we take  $\rho = 0.4$  as our smallest value in this case.

post-GFIC estimator essentially “splits the difference” between the random and fixed effects estimators. While it cannot provide a uniform improvement over the fixed effects estimator, the post-GFIC estimator performs well. When  $\gamma$  is not too large it can yield a substantially lower RMSE than the fixed effects estimator. The gains are particularly substantial when  $x_{it}$  is relatively persistent and  $T$  relatively small, as is common in micro-panel datasets. The averaging estimator performs even better, providing a nearly uniform improvement over the post-GFIC estimator. Only at very large values of  $\gamma$  does it yield a higher RMSE, and these are points in the parameter space where the fixed effects, post-GFIC and averaging estimators are for all intents and purposes identical in RMSE.

## 6.2 Dynamic Panel Example

We now consider two simulation experiments based on section 5.2, applying the GFIC to a dynamic panel model. For both experiments our data generating process is similar to that of Andrews and Lu (2001), specifically

$$\begin{bmatrix} x_i \\ \eta_i \\ v_i \end{bmatrix} \sim \text{iid } N \left( \begin{bmatrix} \mathbf{0}_T \\ 0 \\ \mathbf{0}_T \end{bmatrix}, \begin{bmatrix} I_T & \sigma_{x\eta}\iota_T & \sigma_{xv}\Gamma_T \\ \sigma_{x\eta}\iota_T' & 1 & \mathbf{0}_T' \\ \sigma_{xv}\Gamma_T' & \mathbf{0}_T & I_T \end{bmatrix} \right) \quad (43)$$

where  $\mathbf{0}_m$  denotes an  $m$ -vector of zeros,  $I_m$  the  $(m \times m)$  identity matrix,  $\iota_m$  an  $m$ -vector of ones, and  $\Gamma_m$  an  $m \times m$  matrix with ones on the subdiagonal and zeros elsewhere, namely

$$\Gamma_m = \begin{bmatrix} \mathbf{0}_{m-1}' & 0 \\ I_{m-1} & \mathbf{0}_{m-1} \end{bmatrix}. \quad (44)$$

Under this covariance matrix structure  $\eta_i$  and  $v_i$  are uncorrelated with each other, but both are correlated with  $x_i$ :  $E[x_{it}\eta_i] = \sigma_{x\eta}$  and  $x_{it}$  is predetermined but not strictly exogenous with respect to  $v_{it}$ . Specifically,  $E[x_{it}v_{it-1}] = \sigma_{xv}$ , while  $E[x_{it}v_{is}] = 0$  for  $s \neq t-1$ . We initialize the pre-sample observations of  $y$  to zero, the mean of their stationary distribution, and generate the remaining time periods according to Equation 25 with  $\theta = 0.5$  and  $\sigma_{x\eta} = 0.2$ . The true lag length differs in our two examples as does the target parameter, so we explain these features of the simulation designs below. Unlike Andrews and Lu (2001) we do not generate extra observations to keep the time dimension fixed across estimators with different lag specifications. This is for two reasons. First, in real-world applications such additional observations would not be available. Second, we are explicitly interested in trading off the efficiency gain from including additional time periods in estimation against the bias that arises from estimating an incorrect lag specification.

### 6.2.1 Long-run versus Short-run Effects

Consider two different researchers who happen to be working with the same panel dataset. One wishes to estimate the short-run effect of  $x$  on  $y$  while the other wishes to estimate the long-run effect. Should they use the same model specification? We now present an example showing that the answer, in general, is no. Suppose that the true model is

$$y_{it} = \theta x_{it} + \gamma_1 y_{it-1} + \gamma_2 y_{it-2} + \eta_i + v_{it}$$

where  $i = 1, \dots, n = 250$  and  $t = 1, \dots, T = 5$  and the regressor, individual effect and error term are generated according to Equation 43, as described in the preceding section. Our model selection decision in this example is whether to set  $\gamma_2 = 0$  and estimate a specification with one lag only. We denote this one-lag specification by L1 and the true specification, including both lags, by L2. To focus on the model selection decision, we fix the instrument set in this experiment to  $\mathbf{z}_{it}(\ell, P)$ , defined in Equation 28. Because this instrument set is valid when  $x$  is pre-determined, it does not introduce bias into our estimation. Thus, bias only emerges if we estimate L1 when  $\gamma_2 \neq 0$ . Our simulation design takes  $\theta = 0.5$ ,  $\gamma_1 = 0.4$ ,  $\sigma_{x\eta} = 0.2$ , and  $\sigma_{xv} = 0.1$  and varies  $\gamma_2$  over the range  $\{0.10, 0.11, \dots, 0.19, 0.20\}$ .

Table 1 presents the results of the simulation, based on 1000 replications at each grid point. Because they are based on *ratios* of estimators of  $\theta$  and  $\gamma_1, \gamma_2$ , estimators of the long-run effect may not have finite moments, making finite-sample MSE undefined. The usual solution to this problem in simulation settings is to work with so-called “trimmed” MSE by discarding observations that fall outside, say, a range  $[-M, M]$  before calculating MSE.<sup>8</sup> Because there is no clear way to set the trimming constant  $M$ , it can be difficult to interpret results based on trimmed MSE unless they consider multiple values of  $M$ . To avoid this issue, Table 1 reports simulation results for median absolute deviation (MAD). Results for trimmed MSE with different choices of  $M$  are similar and are available upon request.

The columns of Table 1 labeled L1 and L2 give the MAD of estimators that fix the lag length to one and two, while those labeled GFIC give the MAD of an estimator that selects lag length via the GFIC. Notice that throughout the table  $\gamma_2 \neq 0$  so that L1 is *mis-specified*. Nonetheless, L1 yields lower MAD estimators of both the short-run and long-run effects when  $\gamma_2$  is sufficiently small and the difference can be substantial. When  $\gamma_2 = 0.2$ , for example, MAD for is 0.582 for the long-run effect estimator based on L1 versus 0.801 for that based

---

<sup>8</sup>Note that *asymptotic* MSE remains well-defined even for estimators that do not possess finite-sample moments so that GFIC comparisons remain meaningful. By taking the trimming constant  $M$  to infinity, one can formalize the notion that asymptotic MSE comparisons can be used to “stand in” for finite-sample MSE even when the latter does not exist. For more details, See Hansen (2015) and online appendix C of DiTraglia (2016).

$\gamma_2$	Short-run Effect			Long-run Effect		
	L2	L1	GFIC	L2	L1	GFIC
0.10	0.231	<b>0.141</b>	0.173	0.801	<b>0.582</b>	0.688
0.11	0.237	<b>0.156</b>	0.181	0.834	<b>0.633</b>	0.716
0.12	0.240	<b>0.174</b>	0.193	0.850	<b>0.685</b>	0.752
0.13	0.238	<b>0.187</b>	0.201	0.870	<b>0.729</b>	0.787
0.14	0.220	<b>0.198</b>	0.203	0.870	<b>0.764</b>	0.808
<b>0.15</b>	<b>0.201</b>	0.219	0.211	0.844	<b>0.822</b>	0.839
<b>0.16</b>	<b>0.205</b>	0.223	0.210	0.883	<b>0.856</b>	0.862
0.17	<b>0.181</b>	0.242	0.204	<b>0.860</b>	0.911	0.897
0.18	<b>0.162</b>	0.258	0.189	<b>0.835</b>	0.959	0.891
0.19	<b>0.161</b>	0.265	0.181	<b>0.866</b>	0.997	0.917
0.20	<b>0.143</b>	0.288	0.162	<b>0.858</b>	1.054	0.910

Table 1: Comparisons of mean absolute deviation (MAD) for estimators of the Short-run and Long-run effects of  $x$  on  $y$  in the simulation experiment described in Section 6.2.1. The columns labeled L1 and L2 give the MAD of estimators that fix the lag length to one and two, while the columns labeled GFIC give the MAD of an estimator that selects lag length via the GFIC. Results are based on 1000 simulation replications from the DGP described in Section 6.2 with  $\gamma_1 = 0.4$ , using the estimators described in Section 5.2 and the instrument set  $\mathbf{z}_{it}(\ell, P)$  from Equation 28.

on L2. Note moreover that point at which  $\gamma_2$  becomes large enough for L2 to be preferred depends on which effect we seek to estimate. When  $\gamma_2$  equals 0.15 or 0.16, L1 gives a lower MAD for the short-run effect while L2 gives a lower MAD for the long-run effect. Because it chooses between L1 and L2 and is subject to random model selection errors, the GFIC can never outperform the oracle estimator that uses L1 when it is optimal in terms of MAD and L2 otherwise. Instead, the GFIC represents a compromise between two extremes: its MAD is never as large as that of the worst specification and never as small as that of the best specification. When there are large MAD differences between L1 and L2, however, GFIC is generally quite close to the optimum.

### 6.2.2 Simultaneous Model and Moment Selection for the Short-run Effect

This needs to be cleaned up. Introduce it by saying something like: now a more substantial simulation that focuses on the short-run effect as in our empirical example below. Also need to add the new tables and figures that I made.

In the simulation we take  $\theta = 0.5$ ,  $\sigma_{x\eta} = 0.2$  and vary  $\gamma$ ,  $\sigma_{xv}$ ,  $T$  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  $\theta$  based on specifications LW, LS, W, and S (see Section 5.2) changes with  $\gamma$  and  $\sigma_{xv}$ . Figures

3 and 4 present RMSE comparisons for these four estimators over a simulation grid with  $\gamma, \sigma_{xv} \in \{0, 0.005, 0.01, \dots, 0.195, 0.2\}$ ,  $N \in \{250, 500\}$ ,  $T \in \{4, 5\}$ .<sup>9</sup> For each point in the parameter space, the color in Figure 3 indicates the estimator of  $\theta$  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; lighter indicates a smaller advantage. While Figure 3 indicates *which* estimator is best, Figure 4 indicates how much of an advantage in RMSE can be gained over the correct specification, LW. These plots indicate that, provided  $\gamma$  and  $\sigma_{xv}$  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.
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 chosen 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.

---

<sup>9</sup>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.

To compare selection procedures we use the same simulation grid as above, namely  $\gamma$  and  $\sigma_{xv}$ , namely  $\gamma, \sigma_{xv} \in \{0, 0.005, 0.01, \dots, 0.195, 0.20\}$ . Again, each point on the simulation grid is calculated from 2000 simulation replications. Tables 2 and 3 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 3 gives average and maximum, i.e. worst-case, RMSE over the parameter space for  $\gamma, \sigma_{xv}$  while Table 2 gives *relative* RMSE comparisons. Specifically, the values in the panel “Average” of Table 2 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 2 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.

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 2: 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 3: Average and Worst-case RMSE.



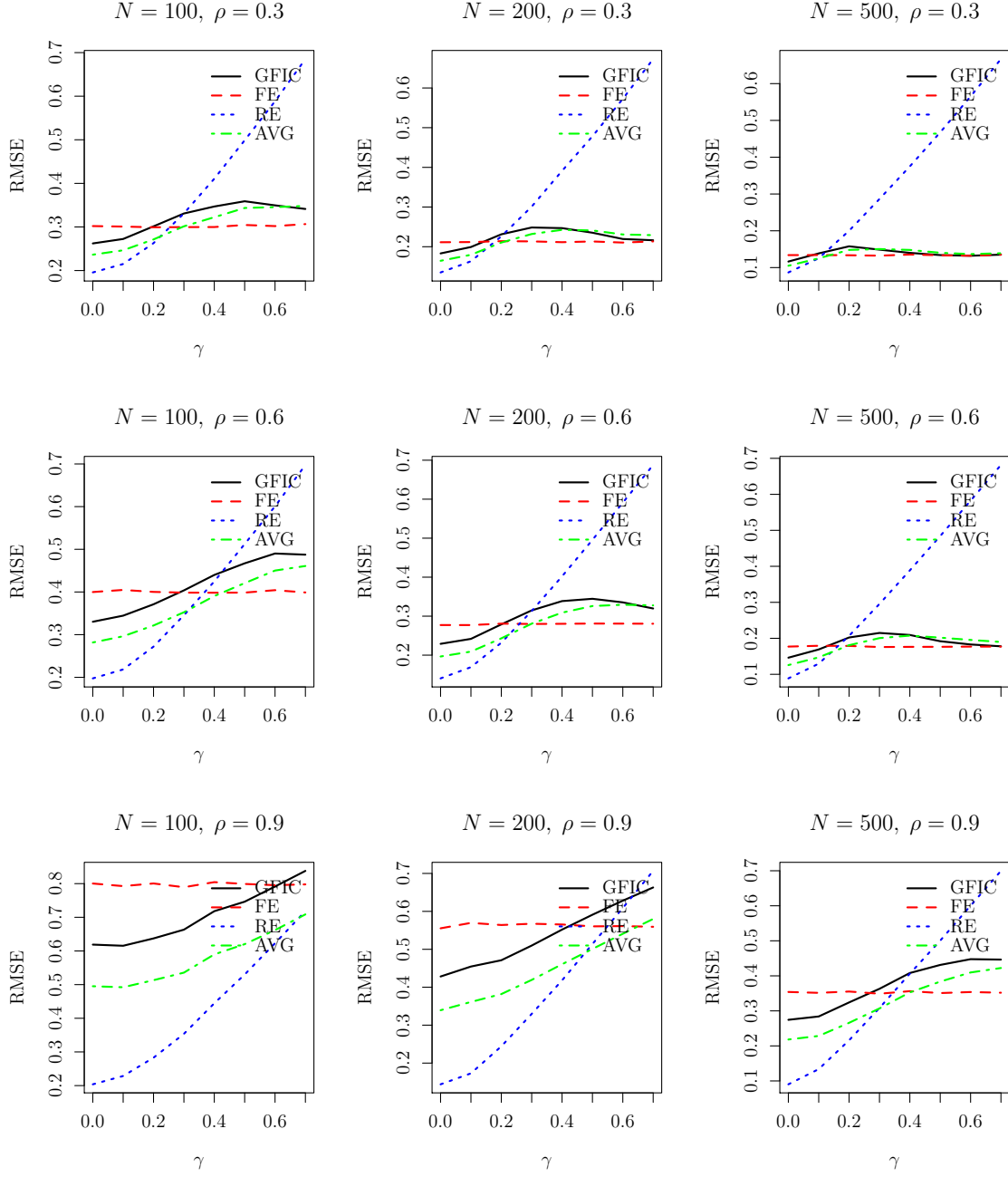


Figure 1: Random vs. Fixed effects example:  $T = 2, \sigma_\epsilon^2 = 2.5$

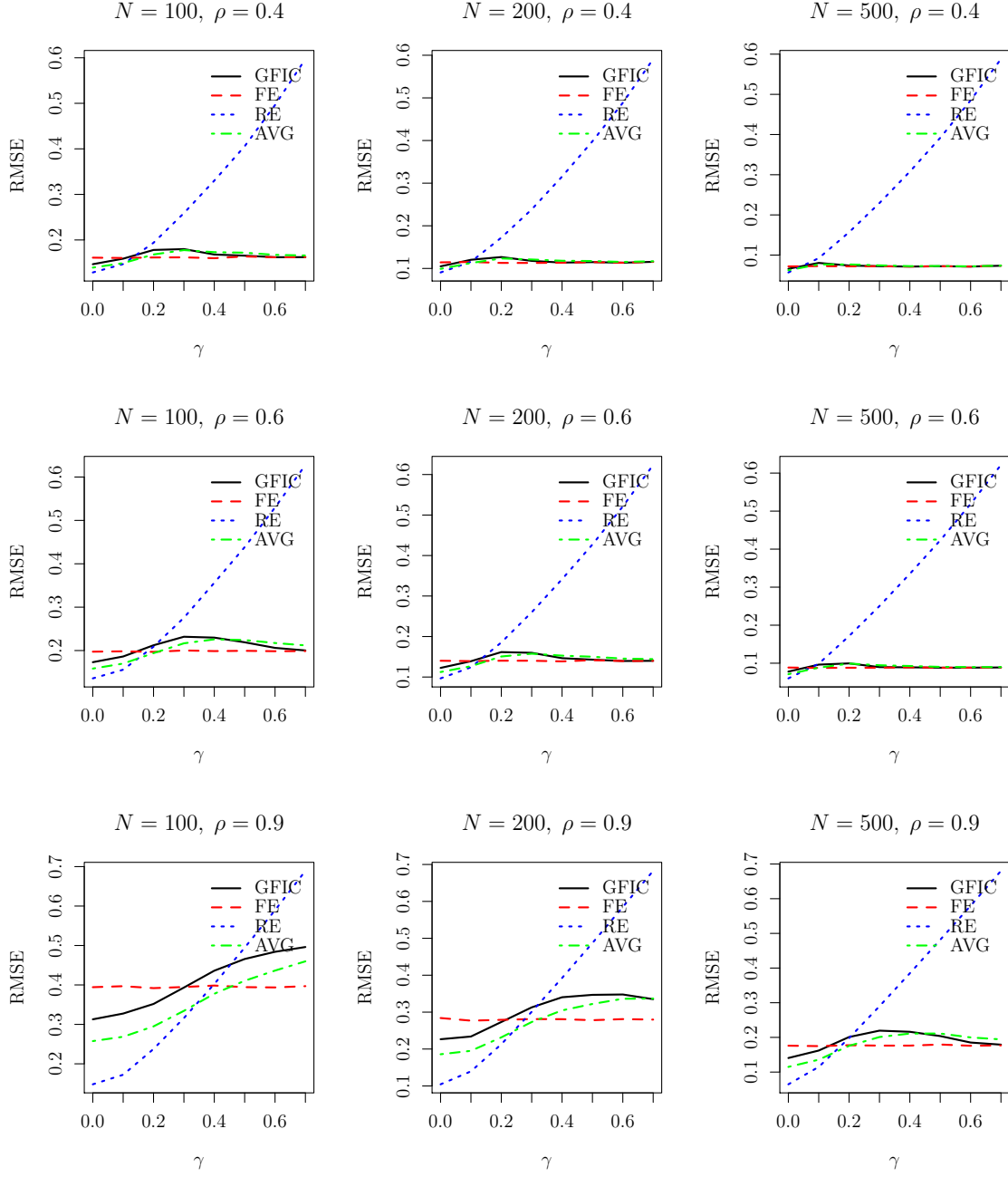


Figure 2: Random vs. Fixed effects example:  $T = 5, \sigma_\epsilon^2 = 2.5$

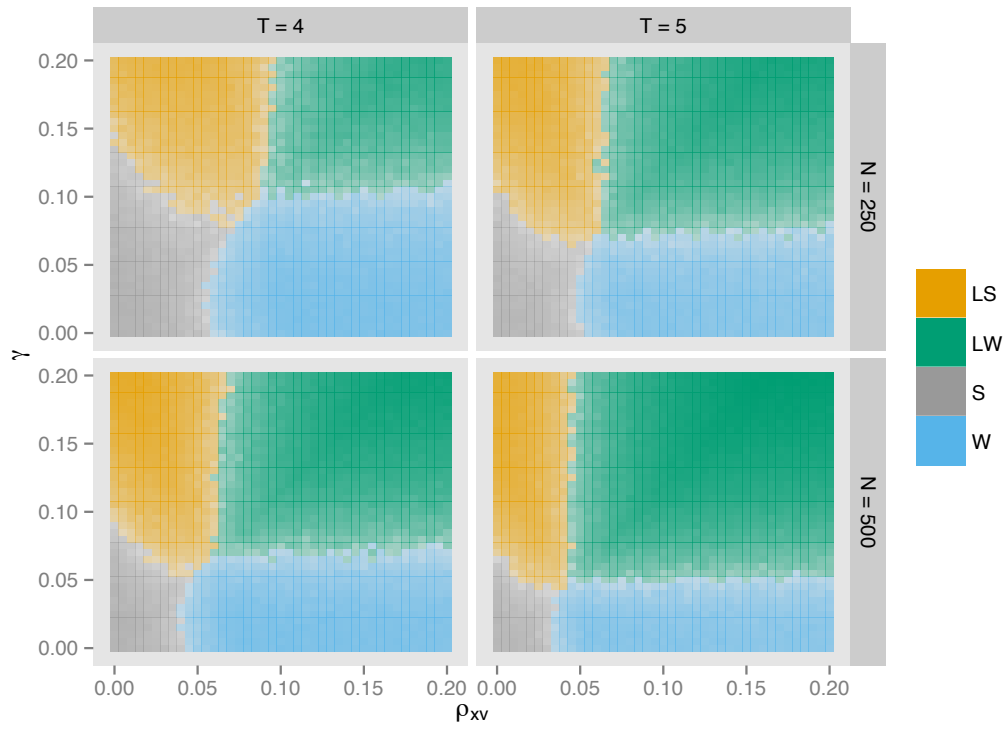


Figure 3: Minimum RMSE Specification at each combination of parameter values. Shading gives RMSE relative to second best specification.

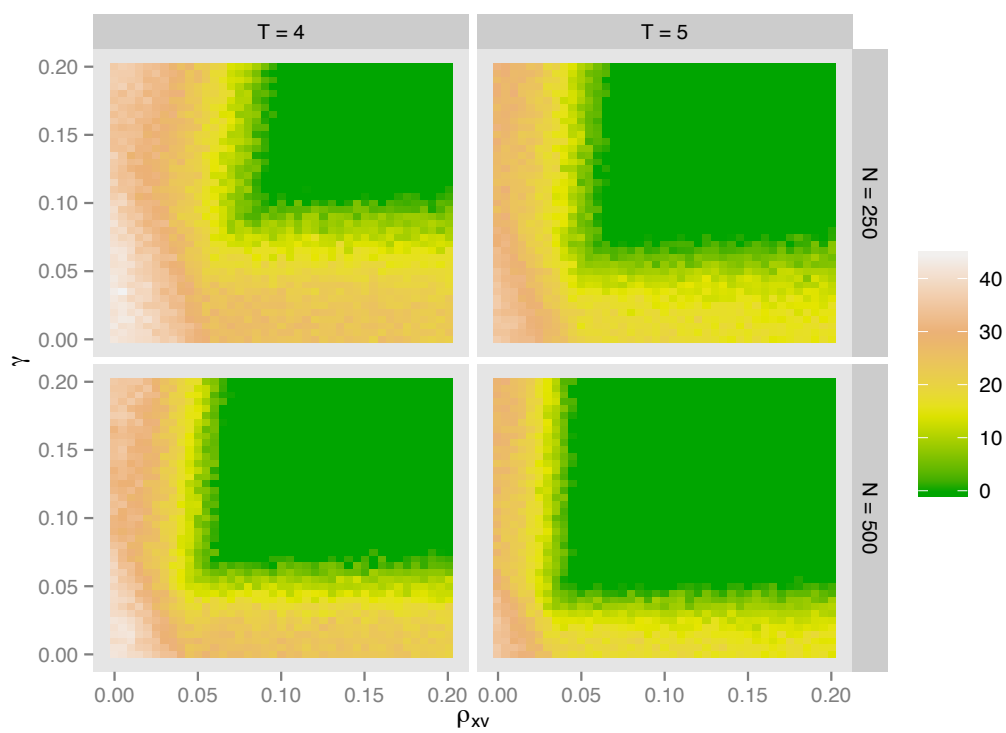


Figure 4: % RMSE Advantage of Best Specification (vs. LW)

## 7 Empirical Example: The Demand for Cigarettes

Still need to edit this section, add the reference to Baltagi, etc.

The empirical example is based on the panel data in Baltagi et al. (2000). The data is suitable to analyze the cigarette demand of 46 American states over 30 years (1963 - 1992). In this section, we consider the model and moment selection problem together. The demand model in Baltagi et al. (2000) is as follows (all prices in real terms):

$$\ln C_{it} = \gamma \ln C_{i,t-1} + \theta \ln P_{it} + \alpha W_{it} + \eta_i + v_{it}$$

where  $C_{it}$  is per capita sales of cigarettes,  $P_{it}$  is the average retail price of a pack of cigarettes, and  $\eta_i$  captures state fixed effect. The vector of all the remaining regressors is denoted as  $W_{it}$  above. It includes the minimum price of cigarettes in any neighboring state  $Pn_{it}$ , per capita disposable income  $Y_{it}$ , and year dummies. After first-differencing, we obtain

$$\Delta \ln C_{it} = \gamma \Delta \ln C_{i,t-1} + \theta \Delta \ln P_{it} + \alpha \Delta W_{it} + \Delta v_{it}$$

Suppose we are interested in whether to use a dynamic specification of cigarette demand to estimate price elasticity. Then we can project out  $\Delta W_{it}$  to rewrite the model as

$$\tilde{C}_{it} = \gamma \tilde{C}_{i,t-1} + \theta \tilde{P}_{it} + \tilde{v}_{it}$$

where  $\tilde{C}_{it}$  and  $\tilde{P}_{it}$  are the residuals of  $\Delta \ln C_{it}$  and  $\Delta \ln P_{it}$  respectively, after projecting out  $\Delta W_{it}$ .

The model selection decision is whether or not to set  $\gamma = 0$ . The moment selection is whether or not to use  $\ln P_{it}$  (after projecting out  $\Delta W_{it}$ ) as an instrument for period  $t$ . There are four specifications: LW, LS, W, and S.<sup>10</sup> Both LW and LS include a lagged dependent variable. LW and W designate the letter W for weak exogeneity assumption. Our aim is to use the GFIC to choose between competing 2SLS estimators of  $\theta$  on the basis of AMSE.

---

<sup>10</sup>The specification LW we consider is different from FD-2SLS in Baltagi et al. (2000). Baltagi et al. (2000) consider the following specification without time dummies:

$$\Delta \ln C_{it} = \beta_1 \Delta \ln C_{i,t-1} + \beta_2 \Delta \ln P_{it} + \beta_3 \Delta \ln Pn_{it} + \beta_4 \Delta \ln Y_{it} + \Delta v_{it}.$$

Furthermore, they instrument  $\Delta \ln C_{i,t-1}$  by the lagged values of exogenous variables  $\ln P_{it}, \ln Pn_{it}, \ln Y_{it}, \ln P_{i,t-1}, \ln Pn_{i,t-1}, \ln Y_{i,t-1}, \ln P_{i,t-2}, \ln Pn_{i,t-2}, \ln Y_{i,t-2}$ . Instead, our specification LW use instrument  $\ln C_{i,t-2}$  for  $\Delta \ln C_{i,t-1}$  as suggested in Anderson and Hsiao (1982).

Table 4: The Estimates of Price Elasticity (Shortened data, 1975 - 1985, T=11)

	LW	LS	W	S
$\theta$	-0.304	-0.259	-0.380	-0.280
GFIC value	0.0553	2.2200	0.2664	4.3398
Bias <sup>2</sup>	0.0	2.2137	0.2202	4.3348
Variance	0.0553	0.0063	0.0462	0.0050

Table 5: The Estimates of Price Elasticity (Shortened data, 1975 - 1980, T=6)

	LW	LS	W	S
$\theta$	-0.677	-0.317	-0.284	-0.366
GFIC value	0.158	0.022	1.172	3.097
Bias <sup>2</sup>	0.0	0.0	1.1004	3.0834
Variance	0.1582	0.0221	0.0720	0.0134

## 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  $(\gamma_0, \theta_0)$ ,

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

and by the Lindeberg-Feller CLT,  $\sqrt{n} f_n(\gamma_0, \theta_0) - \sqrt{n} E[f(Z_{ni}, \gamma_0, \theta_0)] \xrightarrow{d} (\mathcal{N}'_g, \mathcal{N}'_h)'$ . Now,

by a mean-value expansion around  $\gamma_n$ ,

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

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.  $\square$

**Proof of Corollary 2.1.** Since  $\Xi_c$  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,  $\sqrt{n}(\hat{\beta}_v - \beta_0) \xrightarrow{d} -K_v(\mathcal{N}_g - G_\gamma\delta)$ . Finally, by the definition of a matrix inverse,

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

$\square$

**Proof of Corollary 3.2.** For some  $\bar{\gamma}$  between  $\gamma_0$  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,  $\sqrt{n}(\mu_n - \mu_0) = \nabla_\gamma \varphi(\bar{\gamma}, \theta_0)'\delta \rightarrow \nabla_\gamma \varphi(\gamma_0, \theta_0)'\delta$ . The result follows since  $\sqrt{n}(\hat{\mu}(b, c) - \mu_n) = \sqrt{n}(\hat{\mu}(b, c) - \mu_0) - \sqrt{n}(\mu_n - \mu_0)$ .  $\square$

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

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

$\square$

**Proof of Lemma 3.1.** By a mean-value expansion around  $\beta_0 = (\gamma'_0, \theta'_0)'$ ,

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

Now, since

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

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

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

□

**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,

$$\begin{bmatrix} \hat{\delta} \\ \hat{\tau} \end{bmatrix} \begin{bmatrix} \hat{\delta}' & \hat{\tau}' \end{bmatrix} \xrightarrow{d} \begin{bmatrix} U \\ V \end{bmatrix} \begin{bmatrix} U' & V' \end{bmatrix}$$

The result follows since

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

□

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

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

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

**Proof of Theorem 5.2.** This proof is standard, so we provide only a sketch. Expanding,

$$\begin{aligned} \sqrt{n} [\hat{\beta}(k, \cdot) - \beta] &= (0, \mathbf{0}'_{k-1}, \delta)' + \hat{Q}(k, \cdot) [n^{-1/2} Z'(k, \cdot) \Delta \mathbf{v}] \\ \sqrt{n} [\hat{\beta}(r, \cdot) - \beta_r] &= \hat{Q}(r, \cdot) \{ \delta [n^{-1} Z'(r, \cdot) L^k \Delta \mathbf{y}^+] + [n^{-1/2} Z'(r, \cdot) \Delta \mathbf{v}^+] \} \end{aligned}$$

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. □

**Proof of Theorem 5.3.** Expanding  $\hat{\beta}(k, P)$  from Equation 30

$$\begin{aligned} n^{-1/2} X' [\Delta \mathbf{y} - W(k) \hat{\beta}(k, P)] &= n^{-1/2} X' [\Delta \mathbf{v} - W(k) \hat{Q}(k, P) n^{-1} Z'(k, P) \Delta \mathbf{v}] \\ &= \begin{bmatrix} -n^{-1} X' W(k) \hat{Q}(k, P) & I_{T-k-1} \end{bmatrix} \begin{bmatrix} n^{-1/2} Z'(k, P) \Delta \mathbf{v} \\ n^{-1/2} X' \Delta \mathbf{v} \end{bmatrix} \end{aligned}$$



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

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

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

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

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

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

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

## 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., 2016. Using invalid instruments on purpose: Focused moment selection and averaging for GMM. *Journal of Econometrics* 192 (2), 187–208.
- Guggenberger, P., 2010. The impact of a Hausman pretest on the size of a hypothesis test: The panel data case. *Journal of Econometrics* 156, 337–343.
- Hall, A. R., Peixe, F. P., 2003. A consistent method for the selection of relevant instruments in linear models. *Econometric Reviews* 22, 269–288.
- Hansen, B. E., 2015. Efficient shrinkage in parametric models, University of Wisconsin.
- 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 Planning 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.