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

Minsu Chang Francis J. DiTraglia†
University of Pennsylvania

This Version: December 18, 2016 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 mis-specified moment conditions and parameter restrictions to minimize the mean-squared error of a user-specified target parameter. In addition to presenting the general theory, we specialize the GFIC to the problem of choosing between random and fixed effects estimators, and propose a novel averaging estimator that combines the two. In addition, we consider an application to a dynamic panel model, in which the user wishes to select over both the exogeneity assumptions used to generate instruments, and the lag specification. The GFIC performs well in simulations for both examples.

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

JEL Codes: C23, C52

1 Introduction

An econometric model is a tool for answering a particular research question: different questions may suggest different models for the same data. And the fact that a model is wrong, as the old saying goes, does not prevent it from being useful. This paper proposes a novel selection criterion for GMM estimation that takes both of these points to heart: the generalized focused information criterion (GFIC). Rather than attempting to identify the correct specification, the GFIC chooses from a set of potentially mis-specified moment conditions and parameter restrictions to yield the smallest mean squared error (MSE) estimator of a user-specified scalar target parameter. We derive the GFIC under local mis-specification,

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

†Corresponding Author: fditra@sas.upenn.edu, *3718 Locust Walk, Philadelphia, PA 19104*

using asymptotic mean squared error (AMSE) to approximate finite-sample MSE. In this framework mis-specification, while present for any fixed sample size, disappears in the limit so that asymptotic variance and squared bias remain comparable. GMM estimators remain consistent under local mis-specification but their limit distributions show an asymptotic bias. Adding an additional moment condition or imposing a parameter restriction generally reduces asymptotic variance but, if incorrectly specified, introduces a source of bias. The GFIC trades off these two effects in the first-order asymptotic expansion of an estimator to approximate its finite sample behavior.

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

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

As its name suggests, the GFIC is related to the focused information criterion (FIC) of Claeskens and Hjort (2003), a model selection procedure for maximum likelihood estimators that uses local mis-specification to approximate the MSE of a target parameter. The idea of targeted, risk-based model selection has proved popular in recent years, leading to a number of interesting extensions. Hjort and Claeskens (2006), for example, propose an FIC for the Cox proportional hazards model while Claeskens and Carroll (2007) extend the FIC more generally to problems in which the likelihood involves an infinite-dimensional parameter but selection is carried out over the parametric part. More recently, Zhang and Liang (2011) extend the FIC to generalized additive partially linear models and Behl et al. (2012) develop an FIC for quantile regression.

While MSE is a natural risk-function for asymptotically normal estimators, different applications of model selection may call for different risk functions. Claeskens et al. (2006),

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

Like the FIC and related proposals, the GFIC uses local mis-specification to derive a risk-based selection criterion. Unlike them, however, the GFIC provides both moment and model selection for general GMM estimators. The focused moment selection criterion (FMSC) of DiTraglia (2015) represents a special case of the GFIC in which model specification is fixed and selection carried out over moment conditions only. Thus, the GFIC extends both the FIC and the FMSC. Comparatively few papers propose criteria for simultaneous GMM model and moment selection under mis-specification.[1] Andrews and Lu (2001) propose a family of selection criteria by adding appropriate penalty and "bonus" terms to the J-test statistic, yielding analogues of AIC, BIC, and the Hannan-Quinn information criterion. Hong et al. (2003) extend this idea to generalized empirical likelihood (GEL). The principal goal of both papers is consistent selection: they state conditions under which the correct model and all correct moment conditions are chosen in the limit. As a refinement to this approach, Lai et al. (2008) suggest a two-step procedure: first consistently eliminate incorrect models using an empirical log-likelihood ratio criterion, and then select from the remaining models using a bootstrap covariance matrix estimator. The point of the second step is to address a shortcoming in the standard limit theory. While first-order asymptotic efficiency requires that we use all available correctly specified moment conditions, this can lead to a deterioration in finite sample performance if some conditions are only weakly informative. Hall and Peixe (2003) make a similar point about the dangers of including "redundant" moment conditions while Caner (2009) proposes a lasso-type GMM estimator to consistently remove redundant parameters.

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

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

compromise estimator. Unlike the GFIC, however, the CM estimator is not targeted to a particular research goal.

The remainder of this paper is organized as follows. Section 2 derives the asymptotic distribution of GMM estimators under locally mis-specified moment conditions and parameter restrictions. Section 3 uses this information to calculate the AMSE of a user-specified target parameter and provides asymptotically unbiased estimators of the required bias parameters, yielding the GFIC. Section 4 extends the results on averaging estimators and post-selection inference from DiTraglia (2015) to the more general setting of this paper. Section 5 specializes the GFIC to the problem of choosing between random effects and fixed effects estimators, and proposes an estimator that optimally averages the two while Section 6 considers a dynamic panel example. Section 12 presents the results of simulation studies for each of the two examples and Section 13 concludes. Proofs appear in the Appendix.

2 Notation and Asymptotic Framework

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

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

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

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

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

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

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

Under Assumption 2.1, the true parameter vector $\beta_n = (\gamma_n', \theta_0')'$, changes with sample size but converges to $\beta_0 = (\gamma_0', \theta_0')'$ as $n \to \infty$. Unless some elements of δ are zero, any estimator that restricts γ is mis-specified for fixed n. In the limit, however, the restriction

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

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

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

Before defining the estimators under consideration, we require some further notation. Let b be a *model selection vector*, an r-vector of ones and zeros indicating which elements of γ we have chosen to estimate. When $b = 1_r$, where 1_m represents an m-vector of ones, we estimate both θ and the full vector γ. When $b = 0_r$, where 0_m denotes an m-vector of zeros, we estimate only θ, setting $\gamma = \gamma_0$. More generally, we estimate $|b|$ components of γ and set the others equal to the corresponding elements of γ_0. Let $\gamma^{(b)}$ be the $|b|$-dimensional subvector of γ corresponding to those elements selected for estimation. Similarly, let $\gamma_0^{(-b)}$ denote the $(r - |b|)$-dimensional subvector containing the values to which we set those components of γ that are *not* estimated. Analogously, let $c = \left(c_g', c_h'\right)'$ be a *moment selection vector*, a $(p+q)$-vector of ones and zeros indicating which of the moment conditions we have chosen to use in estimation. We denote by $|c|$ the total number of moment conditions used in estimation. Let $\mathcal{BC}$ denote the collection of all model and moment selection pairs (b, c) under consideration.

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

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

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

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

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

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

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

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

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

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

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

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

partitioned according to

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

and define

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

Notice that each of these expressions involves the limiting random variable Z rather than Z_{ni}. Thus, the corresponding expectations are taken with respect to a distribution for which all moment conditions have expectation zero evaluated at (γ_0, θ_0). Finally, let $F(b,c) = \Xi_c F \Xi_b'$ and similarly define $\Omega_c = \Xi_c \Omega \Xi_c'$ and $W_c = \Xi_c W \Xi_c'$ where W is the positive definite probability limit of $\widetilde{W}$. Under Assumption 2.1, both δ and τ induce a bias term in the limiting distribution of $\sqrt{n}\left(\widehat{\beta}(b,c) - \beta_0^{(b)}\right)$. The key results is as follows.

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

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

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

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

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

and

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

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

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

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

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

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

3 The GFIC

The GFIC chooses among potentially incorrect moment conditions and parameter restrictions to minimize estimator AMSE for a scalar target parameter. Denote this target parameter by $\mu = \varphi(\gamma, \theta)$, where φ is a real-valued, almost surely continuous function of the underlying model parameters θ and γ. Let $\mu_n = \varphi(\gamma_n, \theta_0)$ and define μ_0 and $\widehat{\mu}(b,c)$ analogously. By Theorem 2.1 and the delta method, we have the following result.

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

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

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

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

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

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

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

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

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

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

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

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

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

where

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

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

and

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

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

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

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

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

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

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

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

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

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

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

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

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

where

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

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

From Equation 13,

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

Thus, the bias parameters τ and δ enter the AMSE expression in Equation 11 as outer products: $\tau\tau'$, $\delta\delta'$ and $\tau\delta'$. Although $\widehat{\tau}$ and $\widehat{\delta}$ are asymptotically unbiased estimators of τ and δ, it does *not* follow that $\widehat{\tau}\widehat{\tau}$, $\widehat{\delta}\widehat{\delta}$ and $\widehat{\tau}\widehat{\delta}'$ are asymptotically unbiased estimators of $\tau\tau'$, $\delta\delta'$, and $\tau\delta'$. The following result shows how to adjust these quantities to provide the required asymptotically unbiased estimates.

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

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

is an asymptotically unbiased estimator of the squared bias matrix

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

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

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

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

4 Averaging and Post-Selection Inference

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

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

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

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

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

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

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

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

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

where

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

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

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

estimators. Simply plugging in these estimators $\widehat{\tau}$ and $\widehat{\delta}$ and proceeding with the simulation is not guaranteed to lead to valid confidence intervals.[5] In contrast, the following two-step procedure, is guaranteed to yield confidence intervals with asymptotic coverage probability *no less than* $1-(\alpha_1+\alpha_2)$.

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

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

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

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

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

3. Define

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

4. The confidence interval for is μ given by

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

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

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

5 Random Effects versus Fixed Effects Example

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

We see from Theorem 5.1 that

$$AMSE(\widehat{\beta}_{RE}) = c^2\tau^2 + \eta^2 \tag{25}$$

$$AMSE(\widehat{\beta}_{FE}) = c^2\sigma^2 + \eta^2 \tag{26}$$

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

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

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

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

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

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

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

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

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

6 Dynamic Panel Example

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

$$y_{it} = (\gamma_1 y_{it-1} + \cdots + \gamma_p y_{it-p}) + \theta x_{it} + \eta_i + v_{it} \tag{27}$$

where $i = 1, \ldots, n$ indexes individuals and $t = 1, \ldots, T$ indexes time periods. For simplicity, and without loss of generality, we suppose that there are no exogenous time-varying regressors, including a constant, so that the error terms are mean-zero.[6] The unobserved error term η_i is a correlated individual effect: $\sigma_{x\eta} \equiv \mathbb{E}[x_{it}\eta_i]$ may not equal zero. The endogenous regressor x_{it} is assumed to be predetermined but not necessarily strictly exogenous: $\mathbb{E}[x_{it}v_{is}] = 0$ for all $s \geq t$ but may be nonzero for $z < t$. We assume throughout that y_{it} is stationary, which requires both x_{it} and u_{it} to be stationary and $|\boldsymbol{\gamma}| < 1$ where $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_p)'$. Our goal is to estimate a given target parameter with minimum MSE: either $\mu_{SR} \equiv \theta$, the short-run effect of x on y, or $\mu_{LR} \equiv \theta/\left[1 - (\gamma_1 + \cdots + \gamma_p)\right]$, the long-run effect. The question is which assumptions should we use in estimation? As we discuss below, the answer may depend on whether our target is μ_{SR} or μ_{LR}.

Our first decision is what assumption to impose on the relationship between x_{it} and v_{it}. This is the *moment selection* decision. We assumed above that x is predetermined. Imposing the stronger assumption of strict exogeneity gives us more and stronger moment conditions, but using these in estimation introduces a bias if x is not in fact strictly exogenous. Our second decision is how many lags of y to use in estimation. This is the *model selection* decision. The true model contains p lags of y. If we estimate only $\ell < p$ 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 massive loss of

[6] Alternatively, we can simply project out any time-varying exogenous covariates $\mathbf{w}_{it}$ after taking first-differences.

information. At the same time, unless $\gamma_{\ell+1} = \cdots = \gamma_p = 0$, failing to include all p lags in the model introduces a bias. In this example, the GFIC simultaneously chooses over exogeneity assumptions for x and lag length for y to optimally trade off bias and variance.

To avoid many and weak and many instruments problems, we consider 2SLS estimators similar to those suggested by Anderson and Hsiao (1982).[7] Specifically,

6.1 Models and Moment Conditions

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

$$y_{it} = \gamma y_{it-1} + \theta x_{it} + u_{it} \tag{28}$$

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

$$u_{it} = \eta_i + v_{it} \tag{29}$$

with idiosyncratic component v_{it} and individual effect η_i. The individual effect η_i is correlated with x_{it} according to $E[x_i t \eta_i] = \sigma_{x\eta}$. Under the true DGP, x_{it} is predetermined but may not be strictly exogenous. That is, $E[x_{it}v_{is}] = 0$ for all $s \geq t$ but $E[x_{it}v_{is}]$ may be nonzero for $s < t$. To remove the correlated individual effects, we take first differences, yielding

$$\Delta y_{it} = \gamma \Delta y_{it-1} + \theta \Delta x_{it} + \Delta v_{it}. \tag{30}$$

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

Because we observe only $t = 1, \ldots, T$, estimation in differences with a lagged dependent variable uses information from $T - 2$ time periods: $t = 3, \ldots, T$.

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

Taking these considerations together, we consider four specifications: LW, LS, W, and S. Both LW and LS include a lagged dependent variable – hence the designation "L" – while W and S do not. LW and W assume only that x_{it} is predetermined – hence the designation

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

"W" for "weak exogeneity assumption" – while LS and S impose the stronger assumption of *strict* exogeneity. Thus, LW and LS estimate the correct model while LW and W use the correct instrument sets. The correct specification is LW.

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

$$E\left[\begin{pmatrix} y_{i,t-2} \\ x_{i,t-1} \end{pmatrix}(\Delta y_{it} - \gamma \Delta y_{i,t-1} - \theta \Delta x_{it})\right] = 0, \text{ for } t = 3, \ldots, T \tag{31}$$

to which LS adds

$$E\left[x_{it}\left(\Delta y_{it} - \gamma \Delta y_{i,t-1} - \theta \Delta x_{it}\right)\right] = 0, \text{ for } t = 3, \ldots, T \tag{32}$$

for a total of $3(T-2)$ moment conditions. The additional $T-2$ conditions in Equation 32, however, may be incorrect: $E[x_{it}\Delta v_{it}] = -E[x_{it}v_{it-1}]$ since x_{it} is only predetermined. Since it is the only violation of strict exogeneity that is relevant for the specifications under consideration, we let $E[x_{it}v_{it-1}] = \sigma_{xv}$. When $\sigma_{xv} \neq 0$, the moment conditions in Equation 32 are mis-specified.

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

$$E\left[x_{i,t-1}\left(\Delta y_{it} - \theta \Delta x_{it}\right)\right] = 0, \text{ for } t = 2, \ldots, T \tag{33}$$

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

$$E\left[x_{it}\left(\Delta y_{it} - \theta \Delta x_{it}\right)\right] = 0, \text{ for } t = 2, \ldots, T \tag{34}$$

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

$$E\left[\begin{pmatrix} x_{i,t-1} \\ x_{it} \end{pmatrix}(\Delta y_{it} - \theta \Delta x_{it})\right] = \begin{bmatrix} \gamma E[x_{it-1}\Delta y_{it-1}] \\ \gamma E[x_{it}\Delta y_{it-1}] - \sigma_{xv} \end{bmatrix} \tag{35}$$

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

6.2 Estimators and Local Mis-specification

Our aim is to use the GFIC to choose between competing estimators of θ on the basis of AMSE. To do so we must first specify the appropriate form of local mis-specification by analogy with Assumption 2.1. In this example, the parameters γ and σ_{xv} control the degree of mis-specification present in LS, W and S. When $\gamma = 0$, both models, with and without a lag, are correctly specified; when $\sigma_{xv} = 0$ all instruments under consideration are valid. Accordingly, we let $\gamma = \delta/\sqrt{n}$ and $-\sigma_{xv} = \tau/\sqrt{n}$ so that, in the limit, all four specifications are correct. In this framework the true parameter vector is $\beta_n = (\delta/\sqrt{n}, \theta_0)'$ which converges to $\beta_0 = (0, \theta_0)'$.

Assumption 6.1 (Local Mis-specification for Dynamic Panel Example)**.** *Assume that* $\gamma = \delta/\sqrt{n}$ *and* $-\sigma_{xv} = \tau/\sqrt{n}$ *where* δ *and* τ *are unknown constants.*

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

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

$$X_L = \left[\begin{array}{cc} \Delta y_{-1} & \Delta x \end{array} \right] \tag{36}$$

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

$$X_L^+ = \left[\begin{array}{cc} \Delta y_{-1}^+ & \Delta x^+ \end{array} \right]. \tag{37}$$

Although X_L^+ is not observed, we use it in the derivations that follow as it allows us to represent the true data generating process in matrix form. Specifically,

$$\Delta y = X_L \beta_n + \Delta v \tag{38}$$

$$\Delta y^+ = X_L^+ \beta_n + \Delta v^+ \tag{39}$$

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

$$\{z_t\}_{t=k}^T = (z_k, z_{k+1}, \ldots, z_{T-1}, z_T)' \tag{40}$$

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

$$D\{z_t\}_{t=k}^T = \left[\begin{array}{ccc} z_k & & 0 \\ & \ddots & \\ 0 & & z_T \end{array} \right]. \tag{41}$$

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

$$Z(y_{i,-2}) = D\{y_{i,t-2}\}_{t=3}^T \tag{42}$$

$$Z(x_{i,-1}) = D\{x_{i,t-1}\}_{t=3}^T \tag{43}$$

$$Z(x_i) = D\{x_{it}\}_{t=3}^T \tag{44}$$

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

$$Z(x_{i,-1}^+) = D\{x_{i,t-1}\}_{t=2}^T \tag{45}$$

$$Z(x_i^+) = D\{x_{it}\}_{t=2}^T. \tag{46}$$

As above, the symbol "+" used as a superscript indicates the addition of an additional time period. Combining these, define

$$Z_{LS,i} = (Z(y_{i,-2}), Z(x_{i,-1}), Z(x_i))' \tag{47}$$

$$Z_{LW,i} = (Z(y_{i,-2}), Z(x_{i,-1}))' \tag{48}$$

$$Z_{S,i} = \left(Z(x_{i,-1}^+), Z(x_i^+)\right)' \tag{49}$$

$$Z_{W,i} = Z(x_{i,-1}^+). \tag{50}$$

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

$$\widehat{K} = \left[\left(\frac{X'Z}{n} \right) \left(\frac{Z'Z}{n} \right)^{-1} \left(\frac{Z'X}{n} \right) \right]^{-1} \left(\frac{X'Z}{n} \right) \left(\frac{Z'Z}{n} \right)^{-1}. \tag{51}$$

Using this notation, our four estimators are:

$$\widehat{\beta}_{LS} = \widehat{K}_{LS} \left(\frac{Z'_{LS}\Delta y}{n} \right) \tag{52}$$

$$\widehat{\beta}_{LW} = \widehat{K}_{LW} \left(\frac{Z'_{LW}\Delta y}{n} \right) \tag{53}$$

$$\widehat{\theta}_{S} = \widehat{K}_{S} \left(\frac{Z'_{S}\Delta y^{+}}{n} \right) \tag{54}$$

$$\widehat{\theta}_{W} = \widehat{K}_{W} \left(\frac{Z'_{W}\Delta y^{+}}{n} \right). \tag{55}$$

which can be expanded as

$$\sqrt{n}(\widehat{\beta}_{LS} - \beta_0) = \sqrt{n} \begin{bmatrix} \widehat{\gamma}_{LS} \\ \widehat{\theta}_{LS} - \theta_0 \end{bmatrix} = \begin{bmatrix} \delta \\ 0 \end{bmatrix} + \widehat{K}_{LS} \left(\frac{Z'_{LS}\Delta v}{n^{1/2}} \right) \tag{56}$$

$$\sqrt{n}(\widehat{\beta}_{LW} - \beta_0) = \sqrt{n} \begin{bmatrix} \widehat{\gamma}_{LW} \\ \widehat{\theta}_{LW} - \theta_0 \end{bmatrix} = \begin{bmatrix} \delta \\ 0 \end{bmatrix} + \widehat{K}_{LW} \left(\frac{Z'_{LW}\Delta v}{n^{1/2}} \right) \tag{57}$$

and

$$\sqrt{n}\left(\widehat{\theta}_{S} - \theta_0\right) = \widehat{K}_{S} \left[\delta \left(\frac{Z'_{S}\Delta y^{+}_{-1}}{n} \right) + \left(\frac{Z'_{S}\Delta v^{+}}{n^{1/2}} \right) \right] \tag{58}$$

$$\sqrt{n}\left(\widehat{\theta}_{W} - \theta_0\right) = \widehat{K}_{W} \left[\delta \left(\frac{Z'_{W}\Delta y^{+}_{-1}}{n} \right) + \left(\frac{Z'_{W}\Delta v^{+}}{n^{1/2}} \right) \right] \tag{59}$$

by substituting Equation 38. Combining these expressions with the Lindeberg-Feller central limit theorem and standard regularity conditions gives

$$\sqrt{n} \begin{bmatrix} \widehat{\gamma}_{LS} \\ \widehat{\theta}_{LS} - \theta_0 \end{bmatrix} \xrightarrow{d} \begin{bmatrix} \delta \\ 0 \end{bmatrix} + K_{LS} \left\{ \begin{bmatrix} 0_2 \\ \tau \end{bmatrix} \otimes \iota_{T-2} + N\left(0, \mathcal{V}_{LS}\right) \right\} \tag{60}$$

$$\sqrt{n} \begin{bmatrix} \widehat{\gamma}_{LW} \\ \widehat{\theta}_{LW} - \theta_0 \end{bmatrix} \xrightarrow{d} \begin{bmatrix} \delta \\ 0 \end{bmatrix} + K_{LW}\ N\left(0, \mathcal{V}_{LW}\right) \tag{61}$$

and

$$\sqrt{n}\left(\widehat{\theta}_{S} - \theta_0\right) \xrightarrow{d} K_{S} \left[\left(\delta \begin{bmatrix} \psi_0 \\ \psi_1 \end{bmatrix} + \begin{bmatrix} 0 \\ \tau \end{bmatrix} \right) \otimes \iota_{T-1} + N\left(0, \mathcal{V}_{S}\right) \right] \tag{62}$$

$$\sqrt{n}\left(\widehat{\theta}_{W} - \theta_0\right) \xrightarrow{d} K_{W} \left[\delta\psi_0 \otimes \iota_{T-1} + N\left(0, \mathcal{V}_{W}\right) \right]. \tag{63}$$

where K denotes the probability limit of $\widehat{K}$ (see Equation 51) and

$$\psi_0 = E[x_{it}\Delta y_{it}] \tag{64}$$

$$\psi_1 = E[x_{it}\Delta y_{it-1}] \tag{65}$$

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

6.3 GFIC for the Dynamic Panel Example

To operationalize the GFIC, we need estimates of the unknowns in Equations 60–63. To estimate K_{LS}, K_{LW}, K_W and K_S we use $\widehat{K}_{LS}$, $\widehat{K}_{LW}$, $\widehat{K}_W$ and $\widehat{K}_S$, which remain consistent under local mis-specification. There are many consistent estimators of the variance matrices $\mathcal{V}_{LS}$, $\mathcal{V}_{LW}$, $\mathcal{V}_S$ and $\mathcal{V}_W$ under local mis-specification. For robustness, we use the centered, panel robust estimator that allows for heteroscedasticity. We do not center the estimator for LW because this specification is assumed correct, and this yields a more efficient estimator. Using the assumption of stationarity,

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

provide consistent estimators of ψ_0 and ψ_1. The only remaining quantities needed to calculate the GFIC involve the bias parameters τ and δ. As described above, no consistent estimators of these quantities exist under local mis-specification. It remains possible, however, to construct asymptotically unbiased estimators. We can read off an asymptotically unbiased estimator of δ directly from Equation 61, namely $\widehat{\delta} = \sqrt{n}\ \widehat{\gamma}_{LW}$. To construct an asymptotically unbiased estimator of τ, we define $Z'(x) = (Z(x_1), \ldots, Z(x_n))$, see Equation 44, and expand the quantity $n^{-1/2}(\Delta y - X_L\widehat{\beta}_{LW})$ as follows:

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

Now, by the Lindeberg-Feller Central Limit Theorem (c.f. Equation 60) we have

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

and by a Law of Large Numbers,

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

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

$$n^{-1/2}(\Delta y - X_L\widehat{\beta}_{LW}) \xrightarrow{d} \tau \otimes \iota_{T-2} + \left[\begin{array}{cc} \Psi & I \end{array}\right] N(0, \mathcal{V}_{LS}) \tag{66}$$

where

$$\Psi = -E\left[\begin{array}{cc} x_{it}\Delta y_{it-1} & x_{it}\Delta x_{it} \end{array}\right] \otimes \iota_{T-2}K_{LW} \tag{67}$$

Using stationarity to gain efficiency we take the time average

$$\widetilde{\tau} = \left(\frac{\iota'_{T-2}}{T-2}\right) n^{-1/2} Z'(x)(\Delta y - X_L \widehat{\beta}_{LW}) \tag{68}$$

as our estimator of τ. It follows from above that

$$\widetilde{\tau} \xrightarrow{d} \tau + \left(\frac{\iota'_{T-2}}{T-2}\right) \left[\begin{array}{cc} \Psi & I \end{array}\right] N(0, \mathcal{V}_{LS}) \tag{69}$$

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

$$\left[\begin{array}{c} \widehat{\delta} \\ \widetilde{\tau} \end{array}\right] \xrightarrow{d} \left[\begin{array}{c} \widehat{\delta} \\ \widetilde{\tau} \end{array}\right] + \left[\begin{array}{cc} K_{LW}^{\gamma} & 0 \\ \left(\frac{\iota'_{T-2}}{T-2}\right)\Psi & \left(\frac{\iota'_{T-2}}{T-2}\right) I \end{array}\right] N(0, \mathcal{V}_{LS}) \tag{70}$$

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

$$\delta^2: \quad \widehat{\delta}^2 - \widehat{\sigma}_{\delta}^2 \tag{71}$$

$$\tau^2: \quad \widetilde{\tau}^2 - \widehat{\sigma}_{\tau}^2 \tag{72}$$

$$\tau\delta: \quad \widetilde{\tau}\widehat{\delta} - \widehat{\sigma}_{\tau\delta} \tag{73}$$

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

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

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

$$\omega_1 = E[x_{it}\Delta y_{it-1}] \tag{74}$$

$$\omega_2 = E[x_{it}\Delta x_{it}] \tag{75}$$

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

$$\widehat{\omega}_1 = \frac{1}{n(T-2)} \sum_{t=3}^{T} \sum_{i=1}^{n} x_{it}\Delta y_{it-1} \tag{76}$$

$$\widehat{\omega}_2 = \frac{1}{n(T-1)} \sum_{t=2}^{T} \sum_{i=1}^{n} x_{it}\Delta x_{it} \tag{77}$$

Substituting these estimators into the AMSE expressions implied by Equations 60–63 yields the GFIC.

7 Dynamic Panel Example (Short-run Effect v.s. Long-run Effect)

The true data generating process is

$$y_{it} = \alpha_1 y_{it-1} + \alpha_2 y_{it-2} + \beta x_{it} + u_{it} \tag{78}$$

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

$$u_{it} = \eta_i + v_{it} \tag{79}$$

with idiosyncratic component v_{it} and individual effect η_i. We consider the model with two lagged dependent variables, since we are interested in the different estimators related to short-run and long-run effect. (So the focus is whether α_2 is zero or not.)

$$\text{short-run effect} : \quad \beta$$
$$\text{long-run effect} : \frac{\beta}{1 - \alpha_1 - \alpha_2}$$

We assume that x_{it} is predetermined. That is, $E[x_{it} v_{is}] = 0$ for all $s \geq t$ but $E[x_{it} v_{is}]$ may be nonzero for $s < t$. To remove the correlated individual effects, we take first differences, yielding

$$\Delta y_{it} = \alpha_1 \Delta y_{it-1} + \alpha_2 \Delta y_{it-2} + \beta \Delta x_{it} + \Delta v_{it}. \tag{80}$$

Under the true data generating process, x_{it-1}, and y_{it-2}, y_{it-3} are valid instruments for period t.[8]

Because we observe only $t = 1, \ldots, T$, estimation in differences with a lagged dependent variable uses information from $T - 3$ time periods: $t = 4, \ldots, T$.
In contrast, estimation without a lagged variable y_{it-2} uses information from $T - 2$ time periods: $t = 3, \ldots, T$. When T is small, as in many micro-data applications, including an unnecessary lagged variable could result in a huge loss in information, substantially increasing the variance of our estimate of β. On the other hand, unless α_2 is zero, failing to include a lagged variable y_{it-2} will bias our estimates. If α_2 is *nearly* zero, however, this bias may be small compared to the reduction in variance achieved by using an additional time period and estimating one fewer parameter. Our model selection decision is whether or not to set $\alpha_2 = 0$.

Taking these considerations together, we consider two specifications: LW (including y_{it-2}) and W (excluding y_{it-2}). LW is the correct specification in contrast to W.
Estimation based on LW uses the $3(T - 3)$ moment conditions

$$E\left[\begin{pmatrix} y_{i,t-2} \\ y_{i,t-3} \\ x_{i,t-1} \end{pmatrix} (\Delta y_{it} - \alpha_1 \Delta y_{i,t-1} - \alpha_2 \Delta y_{i,t-2} - \beta \Delta x_{it})\right] = 0, \text{ for } t = 4, \ldots, T \tag{81}$$

[8] We consider the just-identified case. Under homoskedasticity, 2SLS estimator is the same as efficient GMM estimator. We consider 2SLS estimator in this example. In Arellano & Bond (1991), all the previous lags $y_{t-2}, \ldots$ are used as instruments and efficient GMM is computed.

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

$$E\left[\begin{pmatrix} y_{i,t-2} \\ x_{i,t-1} \end{pmatrix}(\Delta y_{it} - \alpha_1 \Delta y_{i,t-1} - \beta \Delta x_{it})\right] = 0, \text{ for } t = 3, \ldots, T \tag{82}$$

Because specifications W uses the wrong model, however, these moment conditions are mis-specified:

$$E\left[\begin{pmatrix} y_{i,t-2} \\ x_{i,t-1} \end{pmatrix}(\Delta y_{it} - \alpha_1 \Delta x_{it} - \beta \Delta x_{it})\right] = \begin{bmatrix} \alpha_2 E[y_{i,t-2}\Delta y_{i,t-2}] \\ \alpha_2 E[x_{i,t-1}\Delta y_{i,t-2}] \end{bmatrix} \tag{83}$$

which are non-zero unless $\alpha_2 = 0$.

(a) Estimators and Local Mis-specification

Our aim is to use the GFIC to choose between competing estimators for short-run and long-run effect. To do so we must first specify the appropriate form of local mis-specification. In this framework the true parameter vector is $\theta_n = (\alpha_1, \delta/\sqrt{n}, \beta)'$ which converges to $\theta = (\alpha_1, 0, \beta)'$.

Assumption 7.1 (Local Mis-specification for Dynamic Panel Example)**.** *Assume that $\alpha_2 = \delta/\sqrt{n}$ where δ is unknown constant.*

For additional notations, the symbol "+" used as a superscript indicates the inclusion of the extra time period $t = 3$ that becomes available when we exclude the lagged dependent variable. Using this convention, let $\Delta y_i = (\Delta y_{i4}, \ldots, \Delta y_{iT})'$ and $\Delta y_i^+ = (\Delta y_{i3}, \ldots, \Delta y_{iT})'$. Define $\Delta x_i, \Delta x_i^+$ and $\Delta\epsilon, \Delta\epsilon^+$ analogously. Similarly, let $\Delta y_{i,-1} = (\Delta y_{i3}, \ldots, \Delta y_{i,T-1})'$ and $\Delta y_{i,-1}^+ = (\Delta y_{i2}, \ldots, \Delta y_{i,T-1})'$. Likewise, $\Delta y_{i,-2} = (\Delta y_{i2}, \ldots, \Delta y_{i,T-1})'$ and $\Delta y_{i,-2}^+ = (\Delta y_{i1}, \ldots, \Delta y_{i,T-1})'$ Note that the first element of $\Delta y_{i,-2}^+$ is not observed as $t = 1$ is the first available time period. Stacking over individuals in the usual way, define $\Delta y = (\Delta y_1', \ldots, \Delta y_n')'$ and so on.

The specifications LW has a design matrix as:

$$X_{LW} = \begin{bmatrix} \Delta y_{-1} & \Delta y_{-2} & \Delta x \end{bmatrix} \tag{84}$$

Similarly, let

$$X_W^+ = \begin{bmatrix} \Delta y_{-1}^+ & \Delta x^+ \end{bmatrix}. \tag{85}$$

Then we have:

$$\Delta y = X_{LW}\theta_n + \Delta v \tag{86}$$
$$\Delta y^+ = X_W^+\theta_n + \Delta v^+ \tag{87}$$

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

$$\{z_t\}_{t=k}^T = (z_k, z_{k+1}, \ldots, z_{T-1}, z_T)' \tag{88}$$

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

$$D\left\{z_t\right\}_{t=k}^{T}=\left[\begin{array}{ccc} z_k & & 0 \\ & \ddots & \\ 0 & & z_T \end{array}\right]. \tag{89}$$

To construct the instrument matrices, first define the $(T-3)\times(T-3)$ submatrices

$$\begin{aligned} Z(y_{i,-2}) &= D\left\{y_{i,t-2}\right\}_{t=4}^{T} && (90)\\ Z(y_{i,-3}) &= D\left\{y_{i,t-3}\right\}_{t=4}^{T} && (91)\\ Z(x_{i,-1}) &= D\left\{x_{i,t-1}\right\}_{t=4}^{T} && (92) \end{aligned}$$

and the $(T-2)\times(T-2)$ submatrices

$$\begin{aligned} Z(y_{i,-2}^{+}) &= D\left\{y_{i,t-2}\right\}_{t=3}^{T} && (93)\\ Z(x_{i,-1}^{+}) &= D\left\{x_{i,t-1}\right\}_{t=3}^{T}. && (94) \end{aligned}$$

As above, the symbol "+" used as a superscript indicates the addition of an additional time period. Combining these, define

$$\begin{aligned} Z_{LW,i} &= \left(Z(y_{i,-2}), Z(y_{i,-3}), Z(x_{i,-1})\right)' && (95)\\ Z_{W,i} &= \left(Z(y_{i,-2}^{+}), Z(x_{i,-1}^{+})\right)' && (96) \end{aligned}$$

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

$$\widehat{K}=\left[\left(\frac{X'Z}{n}\right)\left(\frac{Z'Z}{n}\right)^{-1}\left(\frac{Z'X}{n}\right)\right]^{-1}\left(\frac{X'Z}{n}\right)\left(\frac{Z'Z}{n}\right)^{-1}. \tag{97}$$

Using this notation, our two estimators are:

$$\begin{aligned} \widehat{\theta}_{LW} &= \widehat{K}_{LW}\left(\frac{Z_{LW}'\Delta y}{n}\right) && (98)\\ \widehat{\theta}_{W} &= \widehat{K}_{W}\left(\frac{Z_{W}'\Delta y^{+}}{n}\right). && (99) \end{aligned}$$

which can be expanded as

$$\sqrt{n}(\widehat{\theta}_{LW}-\theta)=\sqrt{n}\left[\begin{array}{c} \widehat{\alpha}_{1,LW}-\alpha_1 \\ \widehat{\alpha}_{2,LW} \\ \widehat{\beta}_{LW}-\beta \end{array}\right] = \left[\begin{array}{c} 0 \\ \delta \\ 0 \end{array}\right]+\widehat{K}_{LW}\left(\frac{Z_{LW}'\Delta v}{n^{1/2}}\right) \tag{100}$$

and

$$\sqrt{n}\left(\widehat{\theta}_{W}-\theta\right)=\sqrt{n}\left[\begin{array}{c} \widehat{\alpha}_{1,W}-\alpha_1 \\ \widehat{\beta}_{W}-\beta \end{array}\right] = \widehat{K}_{W}\left[\delta\left(\frac{Z_{W}'\Delta y_{-2}^{+}}{n}\right)+\left(\frac{Z_{W}'\Delta v^{+}}{n^{1/2}}\right)\right] \tag{101}$$

Combining these expressions with the Lindeberg-Feller central limit theorem and standard regularity conditions gives

$$\sqrt{n}\begin{bmatrix} \widehat{\alpha}_{1,LW} - \alpha_1 \\ \widehat{\alpha}_{2,LW} \\ \widehat{\beta}_{LW} - \beta \end{bmatrix} \xrightarrow{d} \begin{bmatrix} 0 \\ \delta \\ 0 \end{bmatrix} + K_{LW}\, N\left(0, \mathcal{V}_{LW}\right) \tag{102}$$

and

$$\sqrt{n}\begin{bmatrix} \widehat{\alpha}_{1,W} - \alpha_1 \\ \widehat{\beta}_{W} - \beta \end{bmatrix} \xrightarrow{d} K_W \left[\delta \begin{bmatrix} \psi_0 \\ \psi_1 \end{bmatrix} \otimes \iota_{T-2} + N\left(0, \mathcal{V}_W\right)\right]. \tag{103}$$

where K denotes the probability limit of $\widehat{K}$ and

$$\psi_0 = E[y_{i,t-2}\Delta y_{i,t-2}] \tag{104}$$
$$\psi_1 = E[x_{i,t-1}\Delta y_{i,t-2}] \tag{105}$$

with the expectations taken with respect to the limiting DGP, in which all specifications are correct.

(b) Estimators for Short-run Effect

Suppose we are interested in short-run effect of x_{it} on y_{it}, i.e. β. Then by $\Delta-$method,

$$\sqrt{n}(\widehat{\beta}_{LW} - \beta) \xrightarrow{d} N\left(0, \quad \begin{bmatrix} 0 & 0 & 1 \end{bmatrix} K_{LW}\mathcal{V}_{LW}K'_{LW} \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}\right) \tag{106}$$

and

$$\sqrt{n}(\widehat{\beta}_{W} - \beta) \xrightarrow{d} \begin{bmatrix} 0 & 1 \end{bmatrix} K_W \left[\delta \begin{bmatrix} \psi_0 \\ \psi_1 \end{bmatrix} \otimes \iota_{T-2} + N\left(0, \mathcal{V}_W\right)\right]. \tag{107}$$

(c) Estimators for Long-run Effect

Suppose we are interested in long-run effect which is expressed as $h(\theta) = h(\alpha_1, \alpha_2, \beta) = \frac{\beta}{1-\alpha_1-\alpha_2}$ and $h_w(\theta) = \frac{\beta}{1-\alpha_1}$. Note that $h'(\theta) = \begin{bmatrix} \frac{\beta}{(1-\alpha_1-\alpha_2)^2} & \frac{\beta}{(1-\alpha_1-\alpha_2)^2} & \frac{1}{1-\alpha_1-\alpha_2} \end{bmatrix}_{\alpha_2=0} = \begin{bmatrix} \frac{\beta}{(1-\alpha_1)^2} & \frac{\beta}{(1-\alpha_1)^2} & \frac{1}{1-\alpha_1} \end{bmatrix}$. Also, $h'_w(\theta) = \begin{bmatrix} \frac{\beta}{(1-\alpha_1-\alpha_2)^2} & \frac{1}{1-\alpha_1-\alpha_2} \end{bmatrix}_{\alpha_2=0} = \begin{bmatrix} \frac{\beta}{(1-\alpha_1)^2} & \frac{1}{1-\alpha_1} \end{bmatrix}$. From $\Delta-$method, we can get

$$\sqrt{n}(h(\widehat{\theta}_{LW}) - h(\theta)) \xrightarrow{d} N\left(0,\ h'(\theta)K_{LW}\mathcal{V}_{LW}K'_{LW}h'(\theta)\right) \tag{108}$$

and

$$\sqrt{n}(h_w(\widehat{\theta}_{W}) - h_w(\theta)) \xrightarrow{d} h'_w(\theta)K_W \left[\delta \begin{bmatrix} \psi_0 \\ \psi_1 \end{bmatrix} \otimes \iota_{T-2} + N\left(0, \mathcal{V}_W\right)\right]. \tag{109}$$

Consider centering expression (31) to $h(\theta_n)$ where $\alpha_2 = \frac{\delta}{\sqrt{n}}$. Using Taylor expansion,

$$h(\theta_n) = h(\theta) + \bigtriangledown_{\alpha_2} h(\bar{\theta})' \frac{\delta}{\sqrt{n}}.$$

Hence, we have

$$\sqrt{n}(h(\widehat{\theta}_{LW}) - h(\theta_n)) \quad \xrightarrow{d} \quad N\left(0,\ h'(\theta) K_{LW} \mathcal{V}_{LW} K'_{LW} h'(\theta)\right) - \bigtriangledown_{\alpha_2} h(\theta)' \delta \tag{110}$$

We can use $\widehat{\delta} = \sqrt{n}\ \widehat{\alpha}_{2,LW}$ which is asymptotically unbiased estimator of δ.

$$\widehat{\delta} \quad \xrightarrow{d} \quad \delta + (0,1,0) K_{LW} \mathcal{V}_{LW} K'_{LW} (0,1,0)' \tag{111}$$

8 Simulation Study: Dynamic Panel Example

We now evaluate the performance of the GFIC in a simulation experiment based on the dynamic panel example from the preceding section. We are interested in comparing LW (including y_{it-2}) and W (excluding y_{it-2}) cases. It is a model selection problem of whether $\alpha_2 = 0$. The simulated covariates and error terms are jointly normal with mean zero and unit variance. Specifically,

$$\begin{bmatrix} x_i \\ \eta_i \\ v_i \end{bmatrix} \sim \text{iid } N\left(\begin{bmatrix} 0_T \\ 0 \\ 0_T \end{bmatrix}, \begin{bmatrix} I_T & \sigma_{x\eta} \iota_T & \sigma_{xv} \Gamma_T \\ \sigma_{x\eta} \iota'_T & 1 & 0'_T \\ \sigma_{xv} \Gamma'_T & 0_T & I_T \end{bmatrix} \right) \tag{112}$$

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

$$\Gamma_m = \begin{bmatrix} 0'_{m-1} & 0 \\ I_{m-1} & 0_{m-1} \end{bmatrix}. \tag{113}$$

Under this covariance matrix structure, η_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 presample observations y_{i0} to zero, the mean of their stationary distribution, and generate the remaining time periods according to

$$y_{it} = \alpha_1 y_{it-1} + \alpha_2 y_{it-2} + \beta x_{it} + \eta_i + v_{it}$$

In the simulation we take

$$\alpha_1 = 0.4, \quad \beta = 0.5, \quad \sigma_{x\eta} = 0.2, \quad \sigma_{xv} = 0.1, \quad N = 250, \quad T = 5$$

and vary α_2 over a grid. Each grid point is based on 1000 simulation replications.

The question is how the finite sample "MAD" (Median Absolute Deviation)[9] of the 2SLS estimators of i) β (short run effect of x_{it} on y_{it}) and ii) $\beta/(1-\alpha_1-\alpha_2)$ (long run effect) changes with different parameter specification, especially α_2.

[9]We choose to use MAD instead of MSE since there appear cases where MSE is infinity. MAD is robust to outliers, and less arbitrary than truncated MSE where researcher should decide truncation point.

The following shows MAD comparisons for the short-run and long-run estimators over α_2 values. Column LW or W show MAD of the estimators when the specification corresponding to column name is chosen always. On the contrary, column GFIC shows MAD of the estimators whose specification is chosen according to GFIC.

Parameter	Short-run Effect			Long-run Effect		
α_2	LW	W	GFIC	LW	W	GFIC
0.10	0.231	**0.141**	0.173	0.801	**0.582**	0.688
0.11	0.237	**0.156**	0.181	0.834	**0.633**	0.716
0.12	0.240	**0.174**	0.193	0.850	**0.685**	0.752
0.13	0.238	**0.187**	0.201	0.870	**0.729**	0.787
0.14	0.220	**0.198**	0.203	0.870	**0.764**	0.808
0.15	**0.201**	0.219	0.211	0.844	**0.822**	0.839
0.16	**0.205**	0.223	0.210	0.883	**0.856**	0.862
0.17	**0.181**	0.242	0.204	**0.860**	0.911	0.897
0.18	**0.162**	0.258	0.189	**0.835**	0.959	0.891
0.19	**0.161**	0.265	0.181	**0.866**	0.997	0.917
0.20	**0.143**	0.288	0.162	**0.858**	1.054	0.910

Table 1: MAD Comparison of Short-run and Long-run Estimators ($\alpha_1 = 0.4$)

As expected, specification LW including y_{it-2} renders lower MAD compared to specification W as α_2 increases. Interestingly, there are cases where different specification is favored depending on the time horizon of interest given α_2. For $\alpha_2 = 0.15$ and 0.16, we can see that LW gives lower MAD for short-run effect whereas it gives higher MAD for long-run effect. Throughout α_2 values considered, GFIC works well since its MAD always lies between the best and the worst specification.

9 Slope Heterogeneity Example

$$\begin{aligned} y_{it} &= \beta_i x_{it} + \epsilon_{it} \\ &= (\beta + \eta_i) x_{it} + \epsilon_{it} \\ &= \beta x_{it} + \eta_i x_{it} + \epsilon_{it} \end{aligned}$$

where

$$\begin{aligned} \epsilon_{it} \text{ i.i.d. over } i,t, \quad & var(\epsilon_{it}) = \sigma_\epsilon^2 \\ \eta_i \text{ i.i.d. over } i, \quad & var(\eta_i) = \sigma_\eta^2 \\ E(\epsilon_i | \mathbf{x}_i) = 0, \quad & E(\eta_i \epsilon_i | \mathbf{x}_i) = 0 \end{aligned}$$

For simplicity, assume everything is mean zero, and $\beta \in \mathbb{R}$, then

$$E[y_{it}] = \beta E[x_{it}] + E[\eta_i x_{it}] + E[\epsilon_{it}] = 0.$$

(a) OLS Estimator

$$\begin{aligned}
\widehat{\beta}_{OLS} &= \left(\sum_{i=1}^{N}\mathbf{x}_i'\mathbf{x}_i\right)^{-1}\left(\sum_{i=1}^{N}\mathbf{x}_i'\mathbf{y}_i\right) \\
&= \left(\sum_{i=1}^{N}\mathbf{x}_i'\mathbf{x}_i\right)^{-1}\left(\sum_{i=1}^{N}\mathbf{x}_i'(\mathbf{x}_i\beta+\mathbf{x}_i\eta_i+\epsilon_i)\right) \\
&= \beta + \left(\sum_{i=1}^{N}\mathbf{x}_i'\mathbf{x}_i\right)^{-1}\left(\sum_{i=1}^{N}\mathbf{x}_i'\mathbf{x}_i\eta_i\right) + \left(\sum_{i=1}^{N}\mathbf{x}_i'\mathbf{x}_i\right)^{-1}\left(\sum_{i=1}^{N}\mathbf{x}_i'\epsilon_i\right)
\end{aligned}$$

Hence,

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

Let's consider local misspecification in the form of

$$\sum_{t=1}^{T}E[x_{it}^2\eta_i] = \frac{\delta}{\sqrt{N}}, \qquad \text{where } \delta \neq 0$$

Then the asymptotic distribution of $\widehat{\beta}_{OLS}$ is derived as

$$\sqrt{N}(\widehat{\beta}_{OLS}-\beta) \to_d N\Big(\big(E[\mathbf{x}_i'\mathbf{x}_i]\big)^{-1}\delta,\quad \sigma_\eta^2\big(E[\mathbf{x}_i'\mathbf{x}_i]\big)^{-1}E[\mathbf{x}_i'\mathbf{x}_i\mathbf{x}_i'\mathbf{x}_i]\big(E[\mathbf{x}_i'\mathbf{x}_i]\big)^{-1}+\sigma_\epsilon^2\big(E[\mathbf{x}_i'\mathbf{x}_i]\big)^{-1}\Big)$$

(b) Mean Group Estimator

Mean group (MG) estimator is the average of the OLS estimator $\widehat{\beta}_i$ across cross-section,

$$\begin{aligned}
\widehat{\beta}_{MG} &= \frac{1}{N}\sum_{i=1}^{N}\widehat{\beta}_i \\
&= \frac{1}{N}\sum_{i=1}^{N}\left(\mathbf{x}_i'\mathbf{x}_i\right)^{-1}\left(\mathbf{x}_i'\mathbf{y}_i\right) \\
&= \frac{1}{N}\sum_{i=1}^{N}\left(\mathbf{x}_i'\mathbf{x}_i\right)^{-1}\left(\mathbf{x}_i'(\mathbf{x}_i\beta+\mathbf{x}_i\eta_i+\epsilon_i)\right) \\
&= \beta + \frac{1}{N}\sum_{i=1}^{N}\eta_i + \frac{1}{N}\sum_{i=1}^{N}\left(\mathbf{x}_i'\mathbf{x}_i\right)^{-1}\left(\mathbf{x}_i'\epsilon_i\right)
\end{aligned}$$

The asymptotic distribution of $\widehat{\beta}_{MG}$ is

$$\sqrt{N}(\widehat{\beta}_{MG} - \beta) \to_d N\left(0,\ \sigma_\eta^2 + \sigma_\epsilon^2 E\big([\mathbf{x}_i'\mathbf{x}_i]^{-1}\big)\right)$$

(1) Since $var[\mathbf{x}_i'\mathbf{x}_i] = E[\mathbf{x}_i'\mathbf{x}_i\mathbf{x}_i'\mathbf{x}_i] - (E[\mathbf{x}_i'\mathbf{x}_i])^2 \geq 0$, the first term of asymptotic variance of $\widehat{\beta}_{OLS}$ is greater than σ_η^2.

(2) From set-up I, we checked that $E\left(\frac{1}{\sum_{t=1}^T x_{it}^2}\right) = E\big([\mathbf{x}_i'\mathbf{x}_i]^{-1}\big) \geq \big(E[\mathbf{x}_i'\mathbf{x}_i]\big)^{-1} = \frac{1}{E(\sum_{t=1}^T x_{it}^2)}$

$\Rightarrow$ It is not clear which is bigger between $AVAR(\widehat{\beta}_{OLS})$ and $AVAR(\widehat{\beta}_{MG})$.

(c) Estimators

Consistent estimator for σ_ϵ^2:

$$\widehat{\sigma}_\epsilon^2 = \frac{\mathbf{y}' M_{\mathbf{X}} \mathbf{y}}{NT - 1}$$

Consistent estimator for σ_η^2:

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

where $S_b = \sum_{i=1}^N \widehat{\beta}_i\widehat{\beta}_i' - \frac{1}{N}\sum_{i=1}^N \widehat{\beta}_i \sum_{i=1}^N \widehat{\beta}_i'$. As noted in the footnote of Swamy (1970), there is no guarantee that $\widehat{\sigma}_\eta^2$ is positive. The following shows the assumptions taken in Swamy (1970): For $i, j = 1, 2, \ldots, N$,

(a) $E\epsilon_i = 0, \quad E\epsilon_i\epsilon_j' = \begin{cases} \sigma_{\epsilon,ii} I & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases}$

(b) $E\beta_i = \beta$

(c) $E(\beta_i - \beta)(\beta_j - \beta)' = \begin{cases} \Delta & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases}$

(d) β_i and ϵ_j are independent

(e) β_i and β_j for $i \neq j$ are independent

Asymptotically unbiased estimator of δ:

$$\widehat{\delta} = \frac{1}{\sqrt{N}}\sum_{i=1}^N \mathbf{x}_i'(\mathbf{y}_i - \mathbf{x}_i\widehat{\beta}_{MG})$$

We can get the asymptotic distribution of $\widehat{\delta}$ as follows:

$$\begin{aligned}
\widehat{\delta} &= \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \mathbf{x}_i'(\mathbf{y}_i - \mathbf{x}_i \widehat{\beta}_{MG}) \\
&= \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \mathbf{x}_i'\big(\mathbf{y}_i - \mathbf{x}_i(\widehat{\beta}_{MG} - \beta) - \mathbf{x}_i\beta\big) \\
&= \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \mathbf{x}_i'\big(\mathbf{x}_i\eta_i + \epsilon_i - \mathbf{x}_i(\widehat{\beta}_{MG} - \beta)\big) \\
&= \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \mathbf{x}_i'(\mathbf{x}_i\eta_i + \epsilon_i) - \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \mathbf{x}_i'\mathbf{x}_i(\widehat{\beta}_{MG} - \beta) \\
&= \underbrace{\begin{bmatrix} 1 & 1 & -\frac{1}{N}\sum_{i=1}^{N} \mathbf{x}_i'\mathbf{x}_i \end{bmatrix}}_{term(1)} \underbrace{\begin{bmatrix} \frac{1}{\sqrt{N}}\sum_{i=1}^{N} \mathbf{x}_i'\mathbf{x}_i\eta_i \\ \frac{1}{\sqrt{N}}\sum_{i=1}^{N} \mathbf{x}_i'\epsilon_i \\ \sqrt{N}(\widehat{\beta}_{MG} - \beta) \end{bmatrix}}_{term(2)}
\end{aligned}$$

$$term(1) \to_p \begin{bmatrix} 1 & 1 & -E[\mathbf{x}_i'\mathbf{x}_i] \end{bmatrix}$$

$$term(2) = \begin{bmatrix} \frac{1}{\sqrt{N}}\sum_{i=1}^{N} \mathbf{x}_i'\mathbf{x_i}\eta_i \\ \frac{1}{\sqrt{N}}\sum_{i=1}^{N} \mathbf{x}_i'\epsilon_i \\ \frac{1}{\sqrt{N}}\sum_{i=1}^{N} \eta_i + \frac{1}{\sqrt{N}}\sum_{i=1}^{N} \left(\mathbf{x}_i'\mathbf{x}_i\right)^{-1}\mathbf{x}_i'\epsilon_i \end{bmatrix} \to_d N\Bigg(\begin{bmatrix} \delta, & 0, & 0 \end{bmatrix}', \Omega_\delta \Bigg)$$

Under the current assumptions, the variance-covariance matrix becomes

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

$$\therefore \widehat{\delta} \to_d N\Bigg(\delta, \quad \begin{bmatrix} 1 & 1 & -E[\mathbf{x}_i'\mathbf{x}_i] \end{bmatrix} \Omega_\delta \begin{bmatrix} 1 & 1 & -E[\mathbf{x}_i'\mathbf{x}_i] \end{bmatrix}' \Bigg)$$

Asymptotically unbiased estimator of δ^2 is $\widehat{\delta}^2 - \widehat{\sigma}_\delta^2$.

(d) Comparison of Asymptotic Variances and FMSC Criteria

Assume $x_{it} \sim N(0, \sigma_x^2)$ and $\epsilon_{it} \sim N(0, \sigma_\epsilon^2)$, both i.i.d. across t. With $x_{it} \in \mathbb{R}$,

$$AVAR(\widehat{\beta}_{OLS}) = \underbrace{\frac{\sigma_\eta^2 E\big[\sum_{t=1}^{T} x_{it}^2 \sum_{t=1}^{T} x_{it}^2\big]}{E\big[\sum_{t=1}^{T} x_{it}^2\big] E\big[\sum_{t=1}^{T} x_{it}^2\big]}}_{(***)} + \frac{\sigma_\epsilon^2}{E\big[\sum_{t=1}^{T} x_{it}^2\big]}$$

$$
\begin{aligned}
(***) &= \frac{\sigma_\eta^2 E\big[\sum_{t=1}^T x_{it}^2 \sum_{t=1}^T x_{it}^2\big]}{E\big[\sum_{t=1}^T x_{it}^2\big] E\big[\sum_{t=1}^T x_{it}^2\big]} \\
&= \sigma_\eta^2 \frac{TE[x_{it}^4] + T(T-1)E[x_{it}^2]E[x_{it}^2]}{T^2 E[x_{it}^2]E[x_{it}^2]} \\
&= \sigma_\eta^2 \frac{(3T + T(T-1))E[x_{it}^2]^2}{T^2 E[x_{it}^2]^2} \\
&= \sigma_\eta^2 \frac{T^2 + 2T}{T^2}
\end{aligned}
$$

The second and third equality come from Isserlis' theorem.

$$
\begin{aligned}
&(i) \quad E[x_{it}^4] = 3E[x_{it}^2]^2 \\
&(ii) \quad E[x_{ij}^2 x_{ik}^2] = E[x_{ij}^2]E[x_{ik}^2] + \underbrace{2E[x_{ij}x_{ik}]^2}_{=0 \text{ under i.i.d. across t}} \qquad \text{for } j \neq k
\end{aligned}
$$

Note that $E\big[\sum_{t=1}^T x_{it}^2\big] = \sum_{t=1}^T Var(x_{it}) = T\sigma_x^2$.

$$
\therefore AVAR(\widehat{\beta}_{OLS}) = \sigma_\eta^2 \frac{T^2 + 2T}{T^2} + \frac{\sigma_\epsilon^2}{T\sigma_x^2}
$$

Let's assume $\sigma_x = 1$, thereby $x_{it} \sim N(0,1)$ and $\sum_{t=1}^T x_{it}^2 \sim \chi^2(T)$. We can write

$$
\begin{aligned}
AVAR(\widehat{\beta}_{MG}) &= \sigma_\eta^2 + \sigma_\epsilon^2 E\left(\frac{1}{\sum_{t=1}^T x_{it}^2}\right) \\
&= \sigma_\eta^2 + \frac{\sigma_\epsilon^2}{T-2}
\end{aligned}
$$

The second equality is from expectation of inverse chi-squared random variable. Now we can compare the asymptotic variance of $\widehat{\beta}_{OLS}$ and $\widehat{\beta}_{MG}$.

$$
\begin{aligned}
\sigma_\eta^2 \frac{T^2 + 2T}{T^2} + \frac{\sigma_\epsilon^2}{T} &> \sigma_\eta^2 + \frac{\sigma_\epsilon^2}{T-2} \\
\sigma_\eta^2 \frac{2}{T} &> \frac{2\sigma_\epsilon^2}{T(T-2)} \\
\sigma_\eta^2 &> \frac{\sigma_\epsilon^2}{T-2}
\end{aligned}
$$

As T increases and σ_η^2 increases, $\widehat{\beta}_{MG}$ is preferred in terms of having lower variance.

$$
\begin{aligned}
AMSE(\widehat{\beta}_{OLS}) &= Bias^2 + AVAR \\
&= \left(\big(E[\mathbf{x}_i'\mathbf{x}_i]^{-1}\delta\big)^2 + \sigma_\eta^2 \frac{T^2 + 2T}{T^2} + \frac{\sigma_\epsilon^2}{T}\right. \\
&= \frac{\delta^2}{T^2} + \sigma_\eta^2 \frac{T^2 + 2T}{T^2} + \frac{\sigma_\epsilon^2}{T}
\end{aligned}
$$

$$\begin{aligned} AMSE(\widehat{\beta}_{MG}) &= Bias^2 + AVAR \\ &= 0 + \sigma_\eta^2 + \frac{\sigma_\epsilon^2}{T-2} \end{aligned}$$

Note that $\widehat{\sigma}_\delta^2 = \sigma_\eta^2 \cdot 2T - \sigma_\epsilon^2 T + \frac{T^2\sigma_\epsilon^2}{T-2}$. So FMSC can be derived as choosing $\widehat{\beta}_{MG}$ if

$$\frac{\widehat{\delta}^2 - \widehat{\sigma}_\delta^2}{T^2} + \sigma_\eta^2 \frac{T^2+2T}{T^2} + \frac{\sigma_\epsilon^2}{T} > \sigma_\eta^2 + \frac{\sigma_\epsilon^2}{T-2}$$

What is x_{it} are serially correlated? Suppose x_{it} is stationary MA(1) across t such that

$$x_{it} = u_{it} + \theta u_{i,t-1}, \quad u_{it} \sim N(0, \sigma_u^2)$$

$$\begin{aligned} E[x_{it}] &= 0 \\ Var[x_{it}] &= (1+\theta^2)\sigma_u^2 \\ Cov(x_{it}, x_{i,t-1}) &= \theta\sigma_u^2 \\ Cov(x_{it}, x_{i,k}) &= 0 \quad \text{for } |k| > 1 \end{aligned}$$

Note that $E(\sum_{t=1}^T x_{it}^2) = \sum_{t=1}^T Var[x_{it}] = T(1+\theta^2)\sigma_u^2$. Also,

$$\begin{aligned} E\big(\sum_{t=1}^T x_{it}^2 \sum_{t=1}^T x_{it}^2\big) &= E\Big[\big(x_{i1}^2 + x_{i2}^2 + \ldots + x_{iT}^2\big)\big(x_{i1}^2 + x_{i2}^2 + \ldots + x_{iT}^2\big)\Big] \\ &= T \cdot E[x_{it}^4] + \sum_{j=1}^T \sum_{j \neq k} E[x_{ij}^2 x_{ik}^2] \end{aligned}$$

By Isserlis' theorem,

$$E[x_{ij}^2 x_{ik}^2] = E[x_{ij}^2]E[x_{ik}^2] + 2E[x_{ij}x_{ik}]^2 = E[x_{ij}^2]E[x_{ik}^2] + 2\gamma(|j-k|)^2$$

where $\gamma(d)$ is the covariance with displacement d. Hence,

$$E\big(\sum_{t=1}^T x_{it}^2 \sum_{t=1}^T x_{it}^2\big) = 3T(1+\theta^2)^2\sigma_u^4 + T(T-1)(1+\theta^2)^2\sigma_u^4 + 4(T-1)\theta^2\sigma_u^4$$

since all the covariance terms with displacement greater than 1 are zero.

$$\begin{aligned} AVAR(\widehat{\beta}_{OLS}) &= \sigma_\eta^2 \frac{(T^2+2T)(1+\theta^2)^2 + 4(T-1)\theta^2}{T^2(1+\theta^2)^2} + \frac{\sigma_\epsilon^2}{(1+\theta^2)\sigma_u^2} \cdot \frac{1}{T} \\ AVAR(\widehat{\beta}_{MG}) &= \sigma_\eta^2 + \sigma_\epsilon^2 E\left(\frac{1}{\sum_{t=1}^T x_{it}^2}\right) \end{aligned}$$

Recall that

$$\begin{aligned} x_{it} &= u_{it} + \theta u_{i,t-1} \\ x_{it} &\sim N(0, (1+\theta^2)\sigma_u^2) \\ \frac{x_{it}}{\sqrt{(1+\theta^2)\sigma_u^2}} &\sim N(0,1) \end{aligned}$$

Then,

$$E\left(\frac{1}{\sum_{t=1}^T x_{it}^2}\right) = E\left(\frac{\frac{1}{(1+\theta^2)\sigma_u^2}}{\sum_{t=1}^T \left(\frac{x_{it}}{\sqrt{(1+\theta^2)\sigma_u^2}}\right)^2}\right) = \frac{1}{(1+\theta^2)\sigma_u^2} \cdot \frac{1}{T-2}$$

$$AVAR(\widehat{\beta}_{OLS}) = \sigma_\eta^2\left(1 + \frac{2}{T} + \frac{4(T-1)\theta^2}{T^2(1+\theta^2)^2}\right) + \frac{\sigma_\epsilon^2}{(1+\theta^2)\sigma_u^2} \cdot \frac{1}{T}$$

$$AVAR(\widehat{\beta}_{MG}) = \sigma_\eta^2 + \frac{\sigma_\epsilon^2}{(1+\theta^2)\sigma_u^2} \cdot \frac{1}{T-2}$$

10 Simulation Results

10.1 Slope Heterogeneity Example

We use the following data generating process which allows for the slope heterogeneity:

$$\begin{aligned} y_{it} &= \beta_i x_{it} + \epsilon_{it} = (\beta + \eta_i)x_{it} + \epsilon_{it} \\ &= \beta x_{it} + \eta_i x_{it} + \epsilon_{it} \end{aligned}$$

We consider the case where x_{it} is serially correlated to be MA(1),[10] i.e.,

$$x_{it} = u_{it} + \theta_u u_{i,t-1}, \quad u_{it} \sim N(0, \sigma_u^2).$$

For simplicity, we fix the variance of x_{it} to be 1. Once θ_u is chosen, σ_u is determined since $var(x_{it}) = (1+\theta_u^2)\sigma_u^2$. To construct FMSC criterion, we need an estimate for θ_u and σ_u. Since we assume θ_u is homogeneous across i, we take the first individual's time series of x_{1t} and estimate θ_u, σ_u. All calculations are based on 1,000 simulation replications. For the following figures, we use $\beta = 0.5, T = 20, \sigma_\epsilon = 4$. We vary θ_u (the persistence of x_{it}) and σ_η (the degree of slope heterogeneity). We present the results where we adopt "coarse" parameter grid for the persistence parameter $\theta_u \in \{0.3, 0.6, 0.9\}$. In the first figure, we consider the case where the degree of slope heterogeneity is small. Throughout the figures, we can check that the OLS estimator provides the lower than the MG estimator. The RMSE of the post-GFIC estimator lies between the OLS and the MG RMSEs. The second figure covers the case where the slope heterogeneity is substantial having value between 0.5 and 1.5. We can see that as σ_η increases, the RMSE of OLS becomes bigger than that of MG. Similar to the previous figures, the post-GFIC estimator splits the different between the OLS and MG estimators.

[10] We also consider the case without any serial correlation where $x_{it} \sim N(0,1)$. We do not include the results in the paper but they are available upon request.

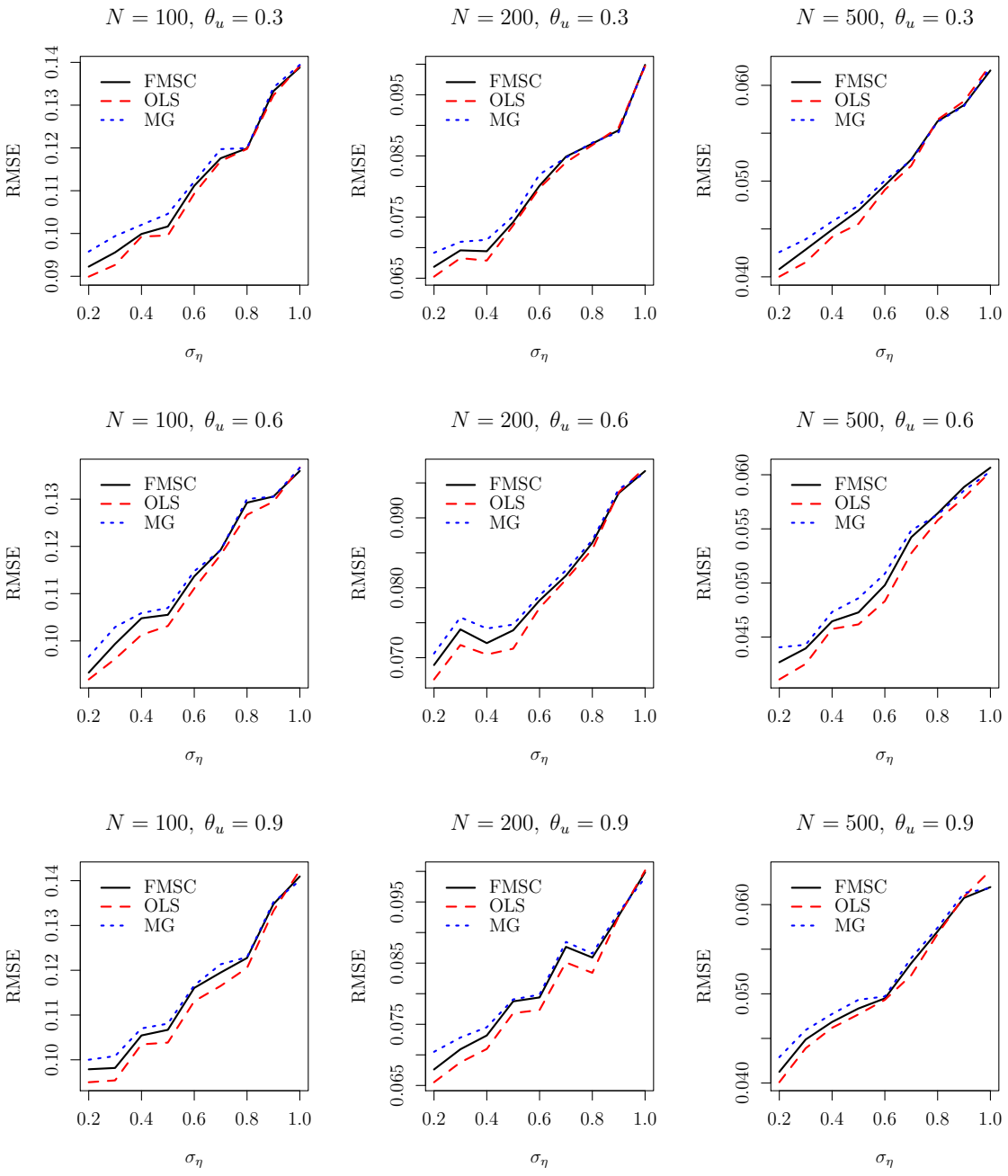


Figure 1: Slope heterogeneity: AR

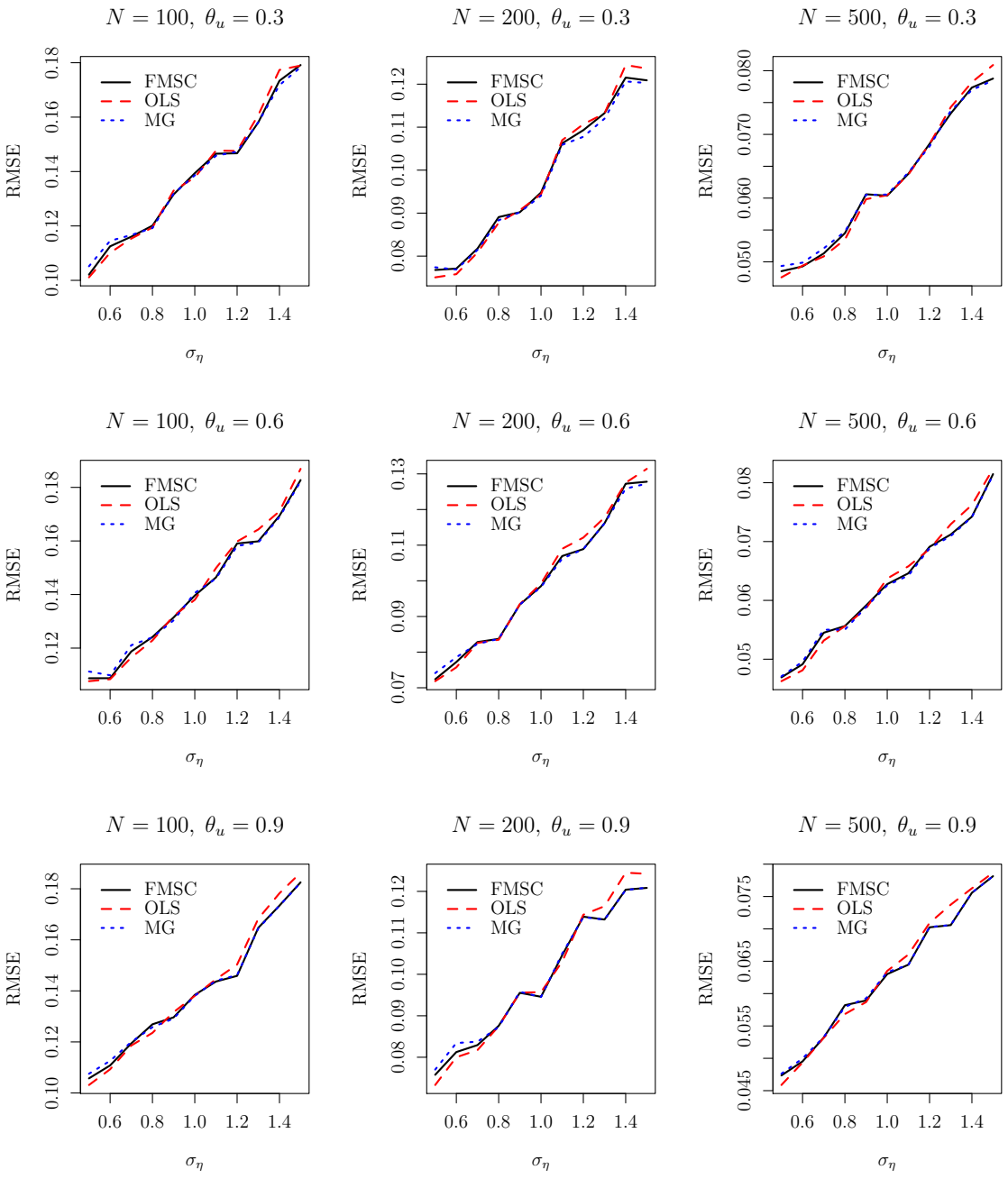


Figure 2: Slope heterogeneity: AR

11 Simulation: Heterogeneous Slopes

- Using the following DGP:

$$\begin{aligned} y_{it} &= \beta_i x_{it} + \epsilon_{it} \\ &= (\beta + \eta_i) x_{it} + \epsilon_{it} \\ &= \beta x_{it} + \eta_i x_{it} + \epsilon_{it} \end{aligned}$$

- If x is serially correlated to be MA(1),

$$x_{it} = u_{it} + \theta_u u_{i,t-1}, \quad u_{it} \sim N(0, \sigma_u^2).$$

- Without x_{it}'s serial correlation, I assume $x_{it} \sim N(0,1)$. Also, I fix $var(x_{it}) = 1$ for MA(1) case by using $var(x_{it}) = (1 + \theta_u^2)\sigma_u^2$. Once θ_u is chosen, σ_u is set accordingly.

- From the specification 1 to 3, $x_{it} \sim N(0,1)$. For the specification 4 and 5, x_{it} follows MA(1).

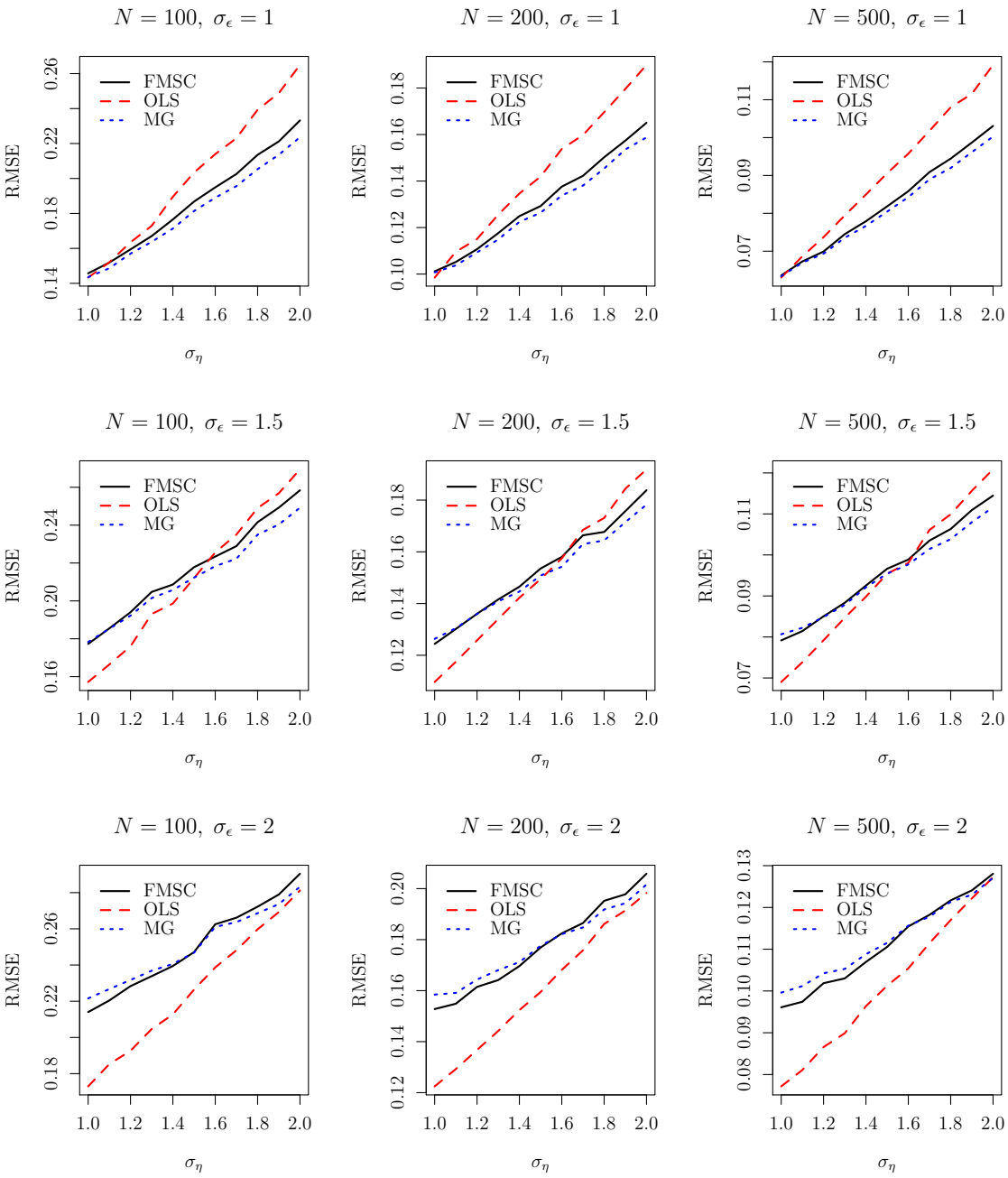
- To construct FMSC criterion, I need an estimate for θ_u and σ_u. Since θ_u is homogeneous across i, I take the first individual's time series of x_{1t} and estimate θ_u, σ_u. I use forecast package in R to get these estimates.

- For the case of $x_{it} \sim N(0,1)$, I set n.rep=10000. With MA(1) x_{it}, n.rep=1000.

- For $x_{it} \sim N(0,1)$, I vary σ_ϵ and σ_η. For MA(1) x_{it}, I vary θ_u (persistence of x_{it}) and σ_η (degree of heterogeneity in slope).

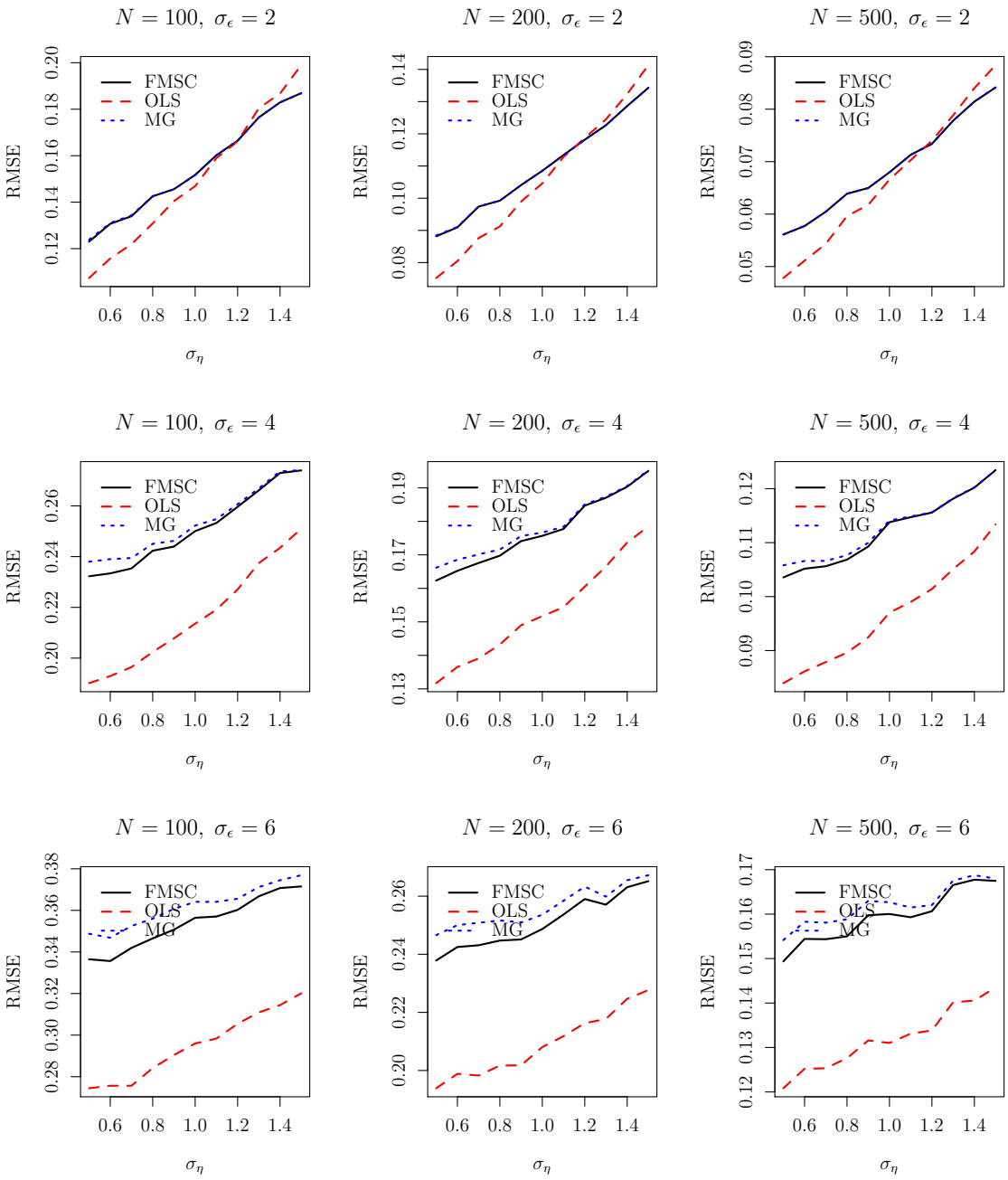
Specification 1

- $x_{it} \sim N(0,1)$, $\beta = 0.5$, $T = 3$
- σ_ϵ fine grid: seq(1, 2, 0.1), coarse grid: c(1, 1.5, 2)
- σ_η fine grid: seq(1, 2, 0.1), coarse grid: c(1, 1.5, 2)



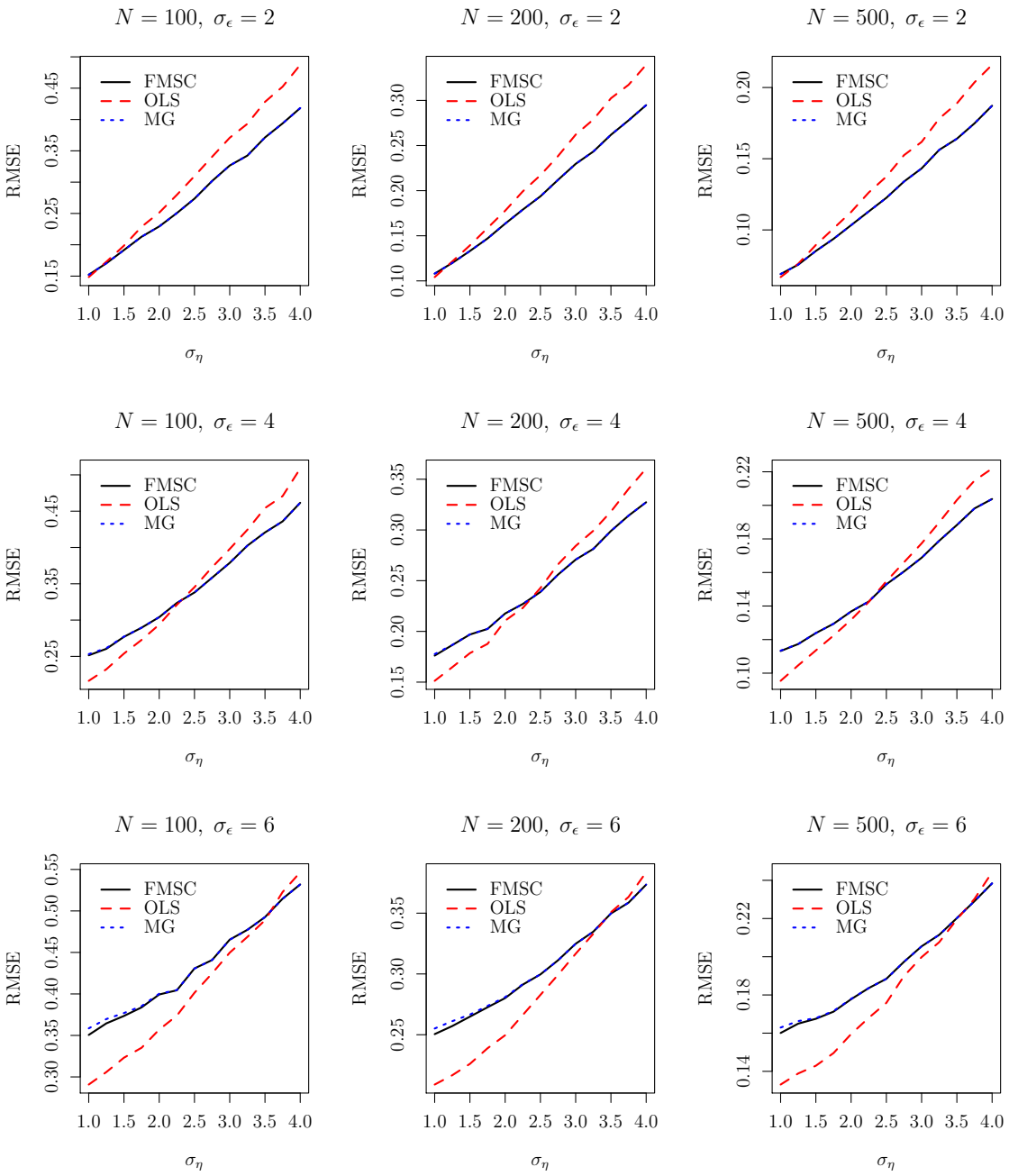
Specification 2

- $x_{it} \sim N(0,1)$, $\beta = 0.5$, $T = 5$
- σ_ϵ fine grid: seq(2, 6, 0.25), coarse grid: c(2, 4, 6)
- σ_η fine grid: seq(0.5, 1.5, 0.1), coarse grid: c(0.5, 1, 1.5)



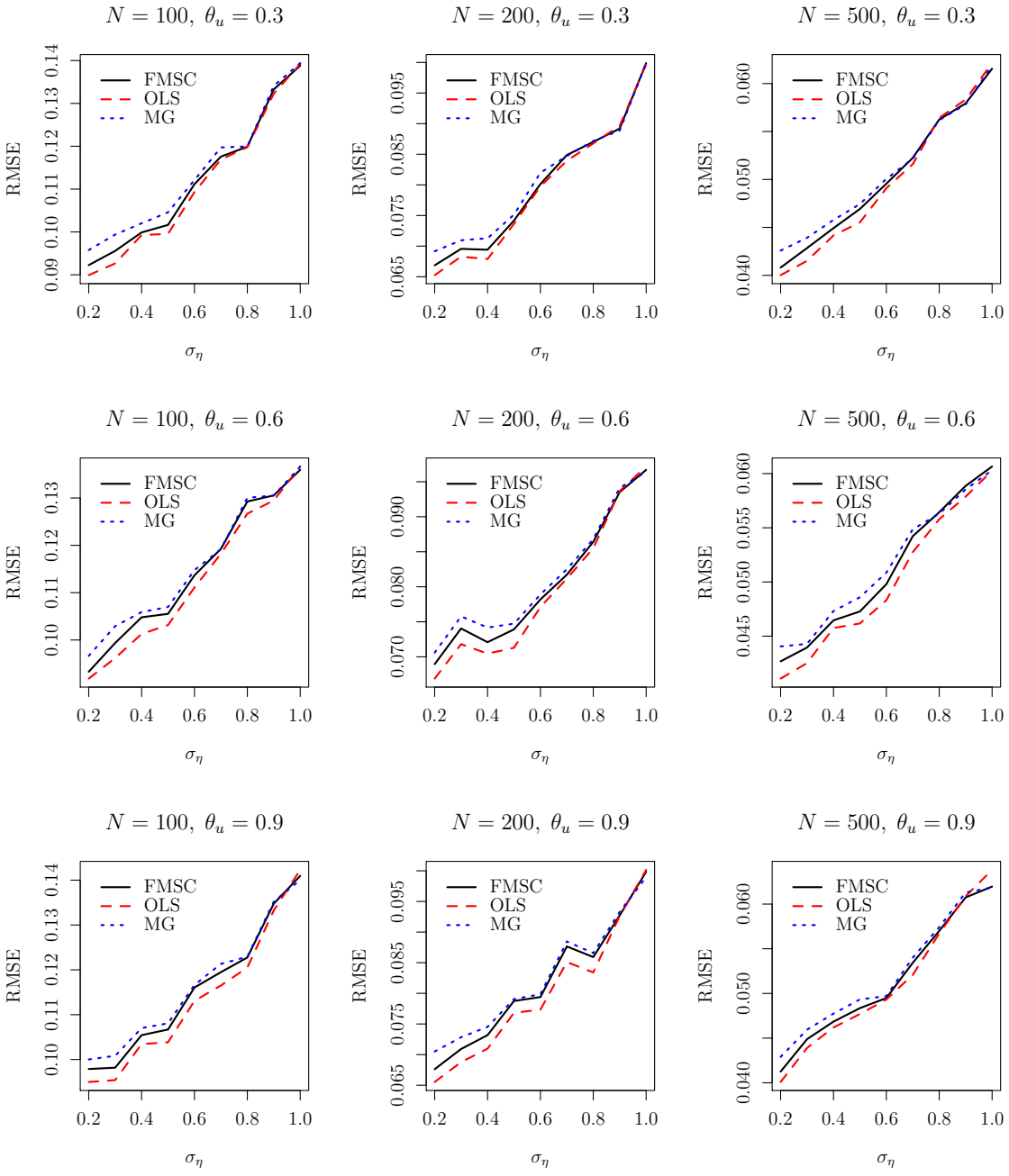
Specification 3

- $x_{it} \sim N(0,1)$, $\beta = 0.5$, $T = 5$
- σ_ϵ fine grid: seq(2, 6, 0.25), coarse grid: c(2, 4, 6)
- σ_η fine grid: seq(1, 4, 0.25), coarse grid: c(1, 2, 4)



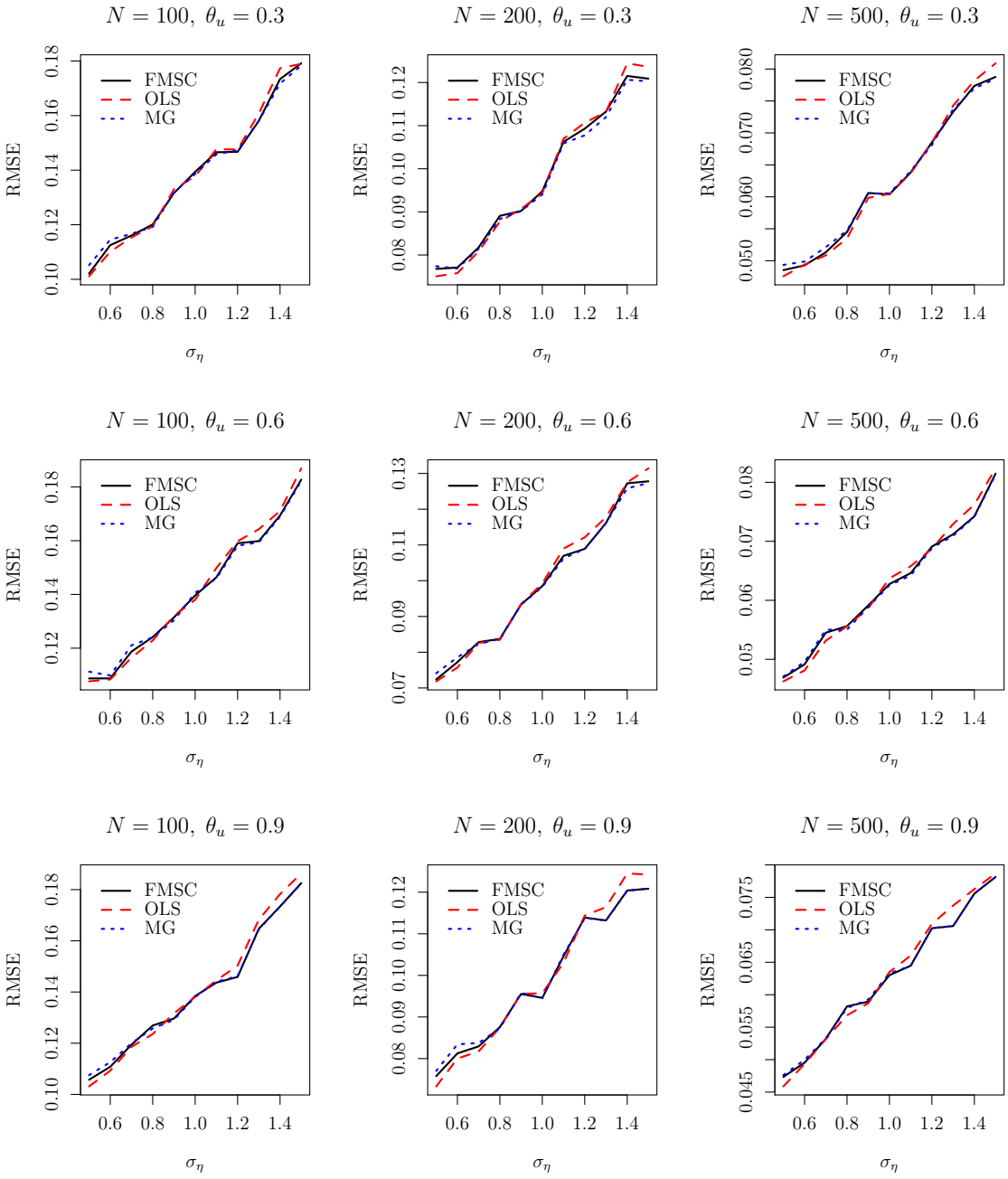
Specification 4

- with MA(1) $x_{it} = u_{it} + \theta_u u_{i,t-1}, u_{it} \sim N(0, \sigma_u^2)$
- $\beta = 0.5$, $T = 20$, $\sigma_\epsilon = 4$, n.rep $= 1000$
- θ_u fine grid: seq(0.3, 0.9, 0.1), coarse grid: c(0.3, 0.6, 0.9)
- σ_η fine grid: seq(0.2, 1, 0.1), coarse grid: c(0.2, 0.6, 1.0)



Specification 5

- with MA(1) $x_{it} = u_{it} + \theta_u u_{i,t-1}, u_{it} \sim N(0, \sigma_u^2)$
- $\beta = 0.5$, $T = 20$, $\sigma_\epsilon = 4$, n.rep $= 1000$
- θ_u fine grid: seq(0.3, 0.9, 0.1), coarse grid: c(0.3, 0.6, 0.9)
- σ_η fine grid: seq(0.5, 1.5, 0.1), coarse grid: c(0.5, 1.0, 1.5)



12 Simulation Results

We now evaluate the performance of the GFIC in two simulation experiment: one based on the fixed versus random effects example from Section 5 and another based on the dynamic panel example from Section 6.

12.1 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

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

independently of $(\varepsilon_{i1}, \dots, \varepsilon_{iT})' \sim$ iid $N(0, \sigma_\varepsilon^2 \mathbf{I}_T)$. In this design, γ controls the correlation between x_{it} and the individual effects α_i, while ρ controls the persistence of x_{it} over time. Larger values of γ correspond to larger violations of the assumption underlying the random effects estimator, increasing its bias. Larger values of ρ, on the other hand, decrease the amount of variation within individuals, thus *increasing* the variance of the fixed effects estimator. Figures 3 and 4 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 over a grid of values for T, γ, ρ, and sample size N. All calculations are based on 10,000 simulation replications.[11] In the interest of space, we present only results for $\sigma_\varepsilon^2 = 2.5$ and our "coarse" parameter grid for ρ here. Additional simulation results are available upon request.

We see from Figures 3 and 4 that, regardless of the configuration of the other parameter values, there is always a range of values for γ 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 ρ of x_{it} increases. Indeed, when N and T are relatively small and ρ is relatively large, the individual effects α_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 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 γ 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

[11] 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. Only at very large values of γ 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.

12.2 Dynamic Panel Example

The details of this simulation are similar those of Andrews and Lu (2001).[12] The simulated covariates and error terms are jointly normal with mean zero and unit variance. Specifically,

$$\begin{bmatrix} x_i \\ \eta_i \\ v_i \end{bmatrix} \sim \text{iid } N\left(\begin{bmatrix} 0_T \\ 0 \\ 0_T \end{bmatrix}, \begin{bmatrix} I_T & \sigma_{x\eta}\iota_T & \sigma_{xv}\Gamma_T \\ \sigma_{x\eta}\iota_T' & 1 & 0_T' \\ \sigma_{xv}\Gamma_T' & 0_T & I_T \end{bmatrix} \right) \tag{114}$$

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

$$\Gamma_m = \begin{bmatrix} 0_{m-1}' & 0 \\ I_{m-1} & 0_{m-1} \end{bmatrix}. \tag{115}$$

Under this covariance matrix structure, η_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 presample observations y_{i0} to zero, the mean of their stationary distribution, and generate the remaining time periods according to

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

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

The first question is how the finite sample MSE of the 2SLS estimators of θ based on specifications LW, LS, W, and S (see Section 6) changes with γ and σ_{xv}. Figures 5 and 6 present RMSE comparisons for these four estimators over a simulation grid with $\gamma, \sigma_{xv} \in \{0, 0.005, 0.01, \ldots, 0.195, 0.2\}$, $N \in \{250, 500\}$, $T \in \{4, 5\}$.[13] For each point in the parameter space, the color in Figure 5 indicates the estimator of θ with the *lowest* finite sample RMSE. The saturation of the color indicates the relative difference in RMSE of the best estimator at that point measured against the second-best estimator: darker indicates a larger advantage; lighter indicates a smaller advantage. While Figure 5 indicates *which* estimator is best, Figure 6 indicates how much of an advantage in RMSE can be gained over the correct specification, LW. These plots indicate that, provided γ and σ_{xv} are not too large, there are potentially large gains to be had by intentionally using an incorrectly specified estimator. The question remains, can the GFIC identify such situations?

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

[13] 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 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 chosed arbitrarily rather than with a view towards some kind of selection optimality. We also consider the GMM model and moment selection criteria of Andrews and Lu (2001):

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

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

To compare selection procedures we use the same simulation grid as above, namely γ and σ_{xv}, namely $\gamma, \sigma_{xv} \in \{0, 0.005, 0.01, \ldots, 0.195, 0.20\}$. Again, each point on the simulation grid is calculated from 2000 simulation replications. Tables 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 γ, σ_{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.

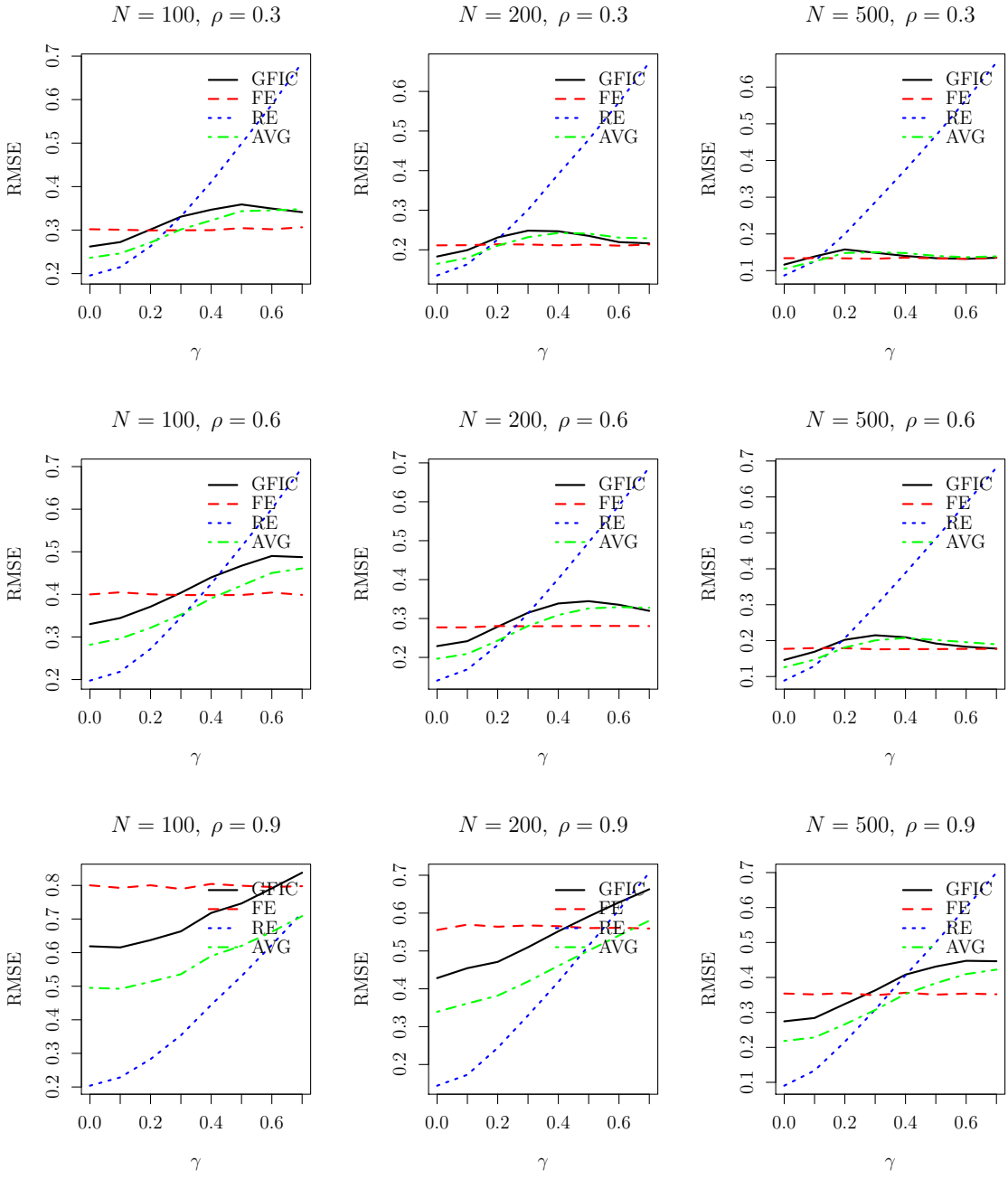


Figure 3: Random vs. Fixed effects example: $T = 2, \sigma_\varepsilon^2 = 2.5$

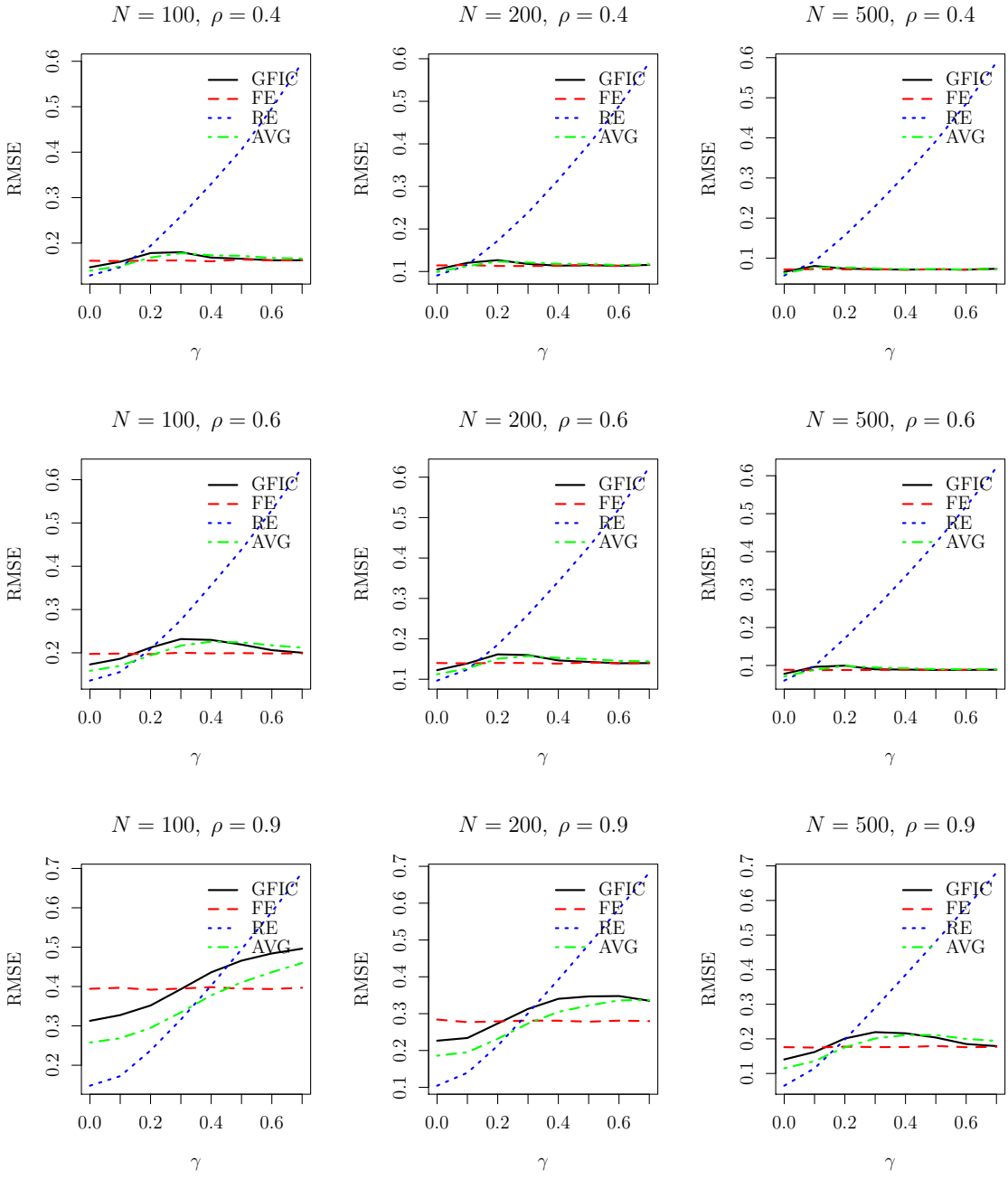


Figure 4: Random vs. Fixed effects example: $T = 5, \sigma_\varepsilon^2 = 2.5$

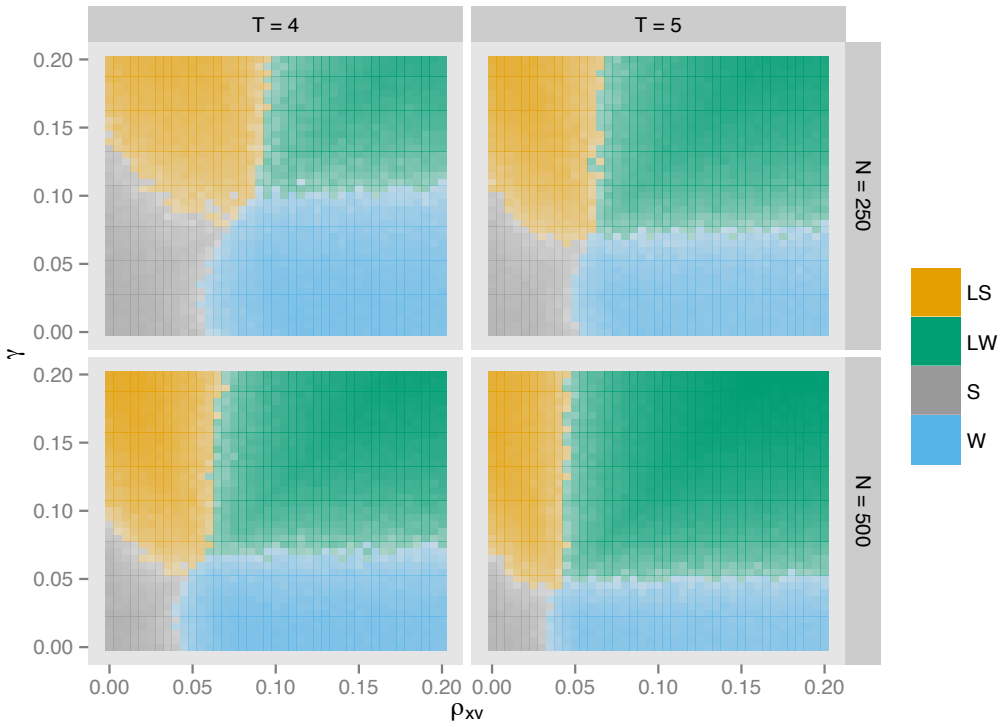


Figure 5: Minimum RMSE Specification at each combination of parameter values. Shading gives RMSE relative to second best specification.

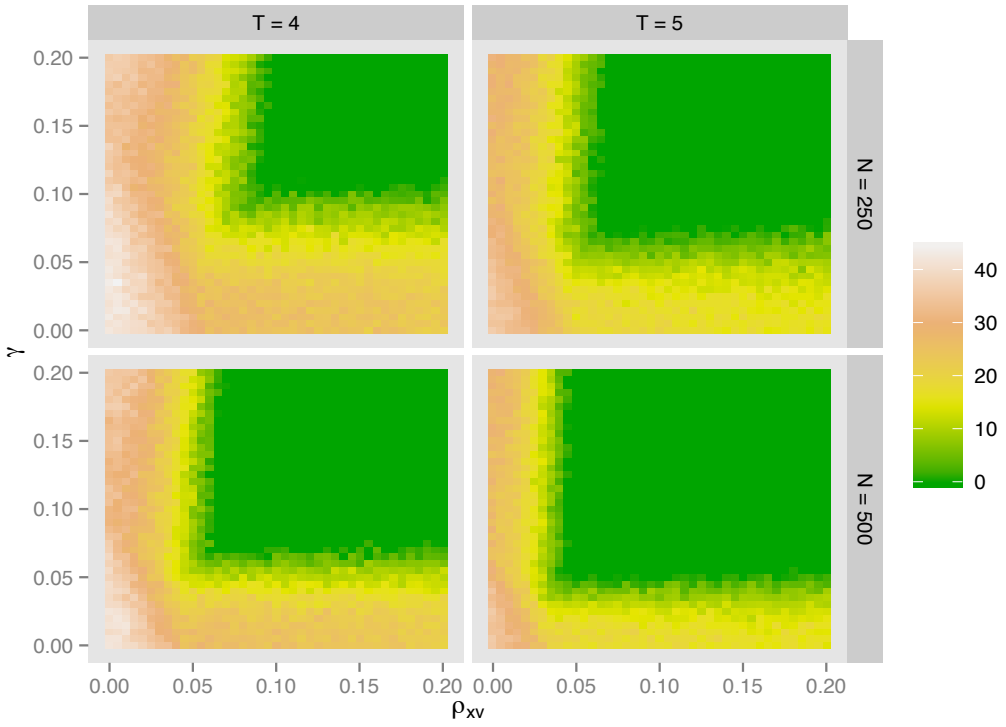


Figure 6: % RMSE Advantage of Best Specification (vs. LW)

13 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).

Results from Section 2

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

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

and by the Lindeberg-Feller central limit theorem,

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

Now, by a mean-value expansion around γ_n,

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

Hence,

$$\sqrt{n}f_n(\gamma_0,\theta_0) \xrightarrow{d} \begin{bmatrix} \mathscr{N}_g \\ \mathscr{N}_h \end{bmatrix} + \begin{bmatrix} 0 \\ \tau \end{bmatrix} - F_\gamma\delta \tag{116}$$

and the result follows by the continuous mapping theorem. □

Proof of Corollary 2.1. For the valid estimator,

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

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

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

Finally

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

by the definition of the matrix inverse. □

Results from Section 3

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

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

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

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

The result follows since

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

□

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

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

□

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

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

Now, since

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

we have

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

Substituting,

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

as claimed. □

Proof of Corollary 3.5. Define

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

By the Continuous Mapping Theorem and Theorem 3.1,

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

The result follows since

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

□

References

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

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

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

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

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

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

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

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

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

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

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.

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

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

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

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

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

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

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

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

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

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