Likelihood and Order Selection for ARMA Dynamics in Panel Models with Interactive Effects

Jushan Bai and Pablo Mones*

March 23, 2026

Abstract

We study dynamic panel data models with short time dimension T and large cross-sectional dimension N in which outcomes follow an ARMA(p_0, q_0) process with interactive effects, extending the workhorse AR(1) dynamic specification. The model implies a parametric structure for the $T \times T$ covariance matrix of the outcome that captures general dynamic dependence through an ARMA process, common shocks through low-rank factor effects, and time-series heteroskedasticity in the idiosyncratic innovations. Exploiting these covariance restrictions, we develop estimation and inference based on Gaussian quasi-maximum likelihood and a second-moment GMM approach. Because identification and estimation are driven by the covariance restrictions rather than within transformation, the framework avoids incidental-parameter (Nickell) bias when T is fixed. Under regularity conditions, we establish consistency and asymptotic normality as $N \to \infty$ with fixed T, and provide feasible variance estimation. We also propose information criteria that consistently select the number of factors r_0 and jointly recover the true orders (p_0, q_0, r_0).

Keywords: Dynamic panel, Arellano-Bond estimator, Nickell bias, Heteroskedasticity, Covariance structure

*Bai: Department of Economics, Columbia University, jb3064@columbia.edu;
Mones: Department of Economics, Columbia University, pm3257@columbia.edu.

1 Introduction

Dynamic panel data models are widely used to study persistence, adjustment, and propagation in both microeconomic applications (e.g., Bloom et al. (2007), Gorbachev (2011), Carlsson et al. (2021)) and macroeconomic settings (e.g., Acemoglu et al. (2008), Acemoglu et al. (2019), Leduc and Liu (2024)). When the time dimension is short and the cross-sectional dimension is large, however, conventional fixed-effects (within) estimation of models with lagged dependent variables is biased; this is the well-known Nickell bias (Nickell, 1981). A leading response has been to use generalized method of moments (GMM) estimators based on internal instruments, with the difference-GMM estimator of Arellano and Bond (1991) as the most widely used approach.

A practical limitation of standard dynamic panel GMM implementations is that they are often tied to relatively restrictive dynamic and error specifications. In particular, empirical work frequently starts from an AR(1) dynamic with serially uncorrelated innovations. Allowing for richer forms of time-series dependence structures can be costly: accommodating higher-order persistence or moving-average components requires additional instruments, increases sensitivity to weak-instrument problems, and aggravates instrument proliferation. At the same time, many panels exhibit substantially richer time-series dependence than an AR(1) model, including moving-average behavior induced by measurement error or aggregation, and more flexible short-run propagation patterns captured by ARMA processes.

A complementary strand of the literature addresses cross-sectional dependence by introducing interactive effects (latent common factors). This approach has become increasingly prevalent in applied work (see Gobillon and Magnac (2016), Jardim et al. (2024), Gobillon et al. (2025), and Besley et al. (2026) to name a few examples). Recent likelihood and quasi-likelihood methods exploit the restrictions implied by interactive effects on the

second moments of each individual's time series. In particular, Bai (2013) and Bai (2024) show that, under a factor error structure, the panel system implies parametric relationships linking the mean and covariance matrix. These restrictions can be used for estimation in levels, without differencing the data. An attractive implication is that estimation can avoid the within transformation that generates Nickell bias when T is fixed.

This paper extends the likelihood-based approach of Bai (2013) and Bai (2024) to allow for general ARMA(p_0, q_0) dynamics. While the existing short-T panels with interactive effects often focus on AR(1) dynamics or use transformed likelihood methods (Hayakawa et al., 2023, e.g.,), our main contribution is to show that the covariance-restriction approach can be extended in a direct and implementable way to ARMA models. In contrast to transformed approaches, our method operates on the data in levels and does not require differencing.

The paper is also directly motivated by the global identification analysis for short-T dynamic panels with interactive effects under an AR(1) specification in Bai and Mones (2025). A key observation is that their identification argument extends naturally to general ARMA(p, q) idiosyncratic dynamics. In particular, the mapping from structural parameters to population second moments continues to be governed by the implied covariance structure, and the key uniqueness steps in the AR(1) case carry over with only minor modifications. Establishing global identification for ARMA models therefore provides a natural bridge from identification to estimation: once the covariance-implied representation is uniquely determined, it can be exploited directly for likelihood- and moment-based inference and for consistent order selection.

The model implies a structured $T \times T$ covariance structure that simultaneously accommodates (i) dynamic dependence through ARMA(p_0, q_0) dynamics, (ii) cross-sectional

dependence through common shocks captured by a low-rank factor component, and (iii) time-series heteroskedasticity in the idiosyncratic innovations. Estimating such a covariance structure is of independent interest in many applications and, indeed, follows a long tradition in the earnings-dynamics literature that treats intertemporal second moments as objects of direct economic and econometric interest (see, e.g., Moffitt and Gottschalk (2011)). In our setting, these covariance restrictions are also the key vehicle for identification and estimation.

We develop two complementary estimation strategies. First, we propose a Gaussian quasi-maximum likelihood estimator (QMLE) based on the implied covariance matrix, under $N \to \infty$ with fixed T. Second, we develop a second-moment GMM estimator that matches the sample covariance matrix to its parametric counterpart. Under regularity conditions, we establish consistency and asymptotic normality for both approaches and provide feasible variance estimation. Importantly, because our identification and estimation rely on covariance restrictions rather than within transformation or differencing, the resulting estimators do not suffer from Nickell bias when T is fixed.

Model selection is another central component of the paper. In empirical work, the ARMA orders and the number of factors are rarely known. We propose the concept of order minimality under the joint ARMA-factor structure, which rules out observationally equivalent lower-order representations. We then show that information criteria consistently determine the number of factors r_0 (extending the results of Bai and Ng (2002) to ARMA panels with short T) and jointly recover the true triple (p_0, q_0, r_0) as $N \to \infty$ with fixed T, which constitutes a novel contribution.

This paper contributes to the literature on dynamic panel data models and panel models with interactive effects. On the dynamic panel side, our work relates to the large body of

research on estimation with fixed T and lagged dependent variables, including the classic bias result of Nickell (1981) and the internal-instrument GMM framework developed by Arellano and Bond (1991), Arellano and Bover (1995), and Blundell and Bond (1998). Moreover, we contribute to the literature on ARMA panel models, such as Baltagi and Li (1994), Chen (2006), and Robinson and Velasco (2020).

Regarding cross-sectional dependence, our framework builds on the interactive-effects and factor-augmented panel literature, including Pesaran (2006), Bai (2009), Ahn et al. (2013), Moon and Weidner (2015), and Hong et al. (2023). It is also closely related to recent work combining dynamic specifications with latent factor structures in short panels, such as Hayakawa et al. (2023) and Bai (2024).

Relative to this work, our contribution has three main elements. First, we incorporate general ARMA dynamics into a short-T interactive-effects model while maintaining identification and estimation based on covariance restrictions. Second, we provide likelihood- and moment-based estimators that operate in levels and therefore avoid Nickell bias without relying on differencing or large-T asymptotics. Third, we develop information criteria that consistently select both the number of factors and the ARMA orders in a fixed-T environment. Together, these results deliver a unified framework for estimation and model selection in dynamic panels with flexible time-series and cross-sectional dependence.

Organization. Section 2 introduces the ARMA interactive-effects model and derives the implied covariance structure. Section 3 discusses estimation of model parameters. Subsection 3.1 presents QMLE and its asymptotic theory; Subsection 3.2 develops the second-moment GMM estimator. Section 4 proposes information criteria for estimating (p_0, q_0, r_0) and establishes consistency. Section 5 concludes.

2 Framework

2.1 The model

Throughout this paper we discuss the autoregressive moving average (ARMA) panel data model with interactive effects, focusing on the case when the number of cross-sectional units N is large and the time dimension T is fixed. The data generating process for the outcome y_{it} is given by:

$$y_{it} = \sum_{s=1}^{p_0} \phi_s y_{it-s} + \delta_t + \lambda_i' f_t + \varepsilon_{it} + \sum_{s=1}^{q_0} \theta_s \varepsilon_{it-s} \tag{1}$$

where $i = 1, 2, \ldots, N$ and $t = 1, 2, \ldots, T$. y_{it} is the outcome variable, $\{\phi_s\}_{s=1}^{p_0}$ is a collection of p_0 autoregressive coefficients, δ_t captures time fixed effects, f_t is an $r_0 \times 1$ vector containing the value of the r_0 latent factors at time t, and λ_i is an $r_0 \times 1$ vector of factor loadings for individual i. We assume $\mathbb{E}[\lambda_i' f_t] = 0$ without loss of generality.[1] $\{\theta_s\}_{s=1}^{q_0}$ is the collection of moving average coefficients and ε_{it} are the idiosyncratic shocks, which we assume to be i.i.d. over i and independent over t satisfying $\mathbb{E}[\varepsilon_{it}] = 0$ and $\mathbb{V}[\varepsilon_{it}] = \sigma_t^2$.

We denote by u_{it} the idiosyncratic error term defined as:

$$u_{it} = \varepsilon_{it} + \sum_{s=1}^{q_0} \theta_s \varepsilon_{it-s} \tag{2}$$

so u_{it} follows an MA(q_0) process. Equivalently, this model can be represented employing the lag operator and polynomial definitions:

$$\Phi(L) y_{it} = \delta_t + \lambda_i' f_t + \Theta(L) \varepsilon_{it}$$

where

$$\Phi(L) = 1 - \phi_1 L - \cdots - \phi_{p_0} L^{p_0}, \qquad \Theta(L) = 1 + \theta_1 L + \cdots + \theta_{q_0} L^{q_0}$$

[1] If the expected value were different from zero, then it can be absorbed into the time fixed effects as $\widetilde{\delta}_t = \delta_t + \mathbb{E}[\lambda_i' f_t]$ for each t, so we can just relabel the time fixed effect.

where we shall assume that the polynomials $\Phi(L)$ and $\Theta(L)$ do not share any roots, and that their roots lie outside the unit circle.

This model allows for flexible dynamics in the outcome and idiosyncratic error. It permits past values of the outcome to have an impact on the current outcome through an autoregressive structure and time dependence in the idiosyncratic errors through a moving average structure. The case where u_{it} also follows an ARMA$(\dot{p}, q_0)$ is nested in this model and corresponds to the case where y_{it} is an ARMA$(p_0 + \dot{p}, q_0)$ panel (see Appendix B.1 for a proof).

Moreover, the interactive effects structure permits common unobserved shocks f_t to have a heterogeneous impact on the outcomes weighted by λ_i. This type of structure nests the classical two-way fixed effects (TWFE) model, which corresponds to the case where $r_0 = 1$ and $f_t = 1$, yielding $\lambda_i' f_t = \lambda_i$. Thus, if interactive effects are present but one erroneously estimates an additive (TWFE) model, the estimators can be biased and inconsistent. Conversely, if the true effects are additive but one estimates the richer interactive-effects model, the estimates remain consistent (albeit less efficient).

Furthermore, we allow the variance of idiosyncratic shocks to vary over time by assuming $\mathbb{V}[\varepsilon_{it}] = \sigma_t^2$, so the model accommodates time-series heteroskedasticity. This nests the classical ARMA assumption where $\mathbb{V}[\varepsilon_{it}] = \sigma^2$ independent of time. Relaxing this assumption does not cause problems in terms of identification and estimation because we are working with a short panel (T fixed) and identification comes from the cross-sectional dimension (N large). Therefore, the number of time-dependent parameters does not grow asymptotically and they can, in general, be identified by exploiting cross-sectional variation.

Estimation of model parameters can be conducted via Gaussian QMLE or GMM (discussed in Section 3) based on the expected value of the sample mean and sample covariance

matrix of the individual outcomes $y_i = (y_{i1}, \ldots, y_{iT})'$, a $T \times 1$ vector. Define the sample mean of the stacked observations as:

$$\bar{y} = \frac{1}{N} \sum_{i=1}^{N} y_i \tag{3}$$

which is a $T \times 1$ vector. The expected value of $\bar{y}$ can be used to identify the time fixed effects δ_t once the autoregressive coefficients have been identified. Meanwhile, the sample covariance of y_i is a $T \times T$ matrix defined as:

$$\mathbf{S}_N = \frac{1}{N} \sum_{i=1}^{N} (y_i - \bar{y}) (y_i - \bar{y})' \tag{4}$$

Define the sample covariance matrix of the factor loadings and its expected value as:

$$\mathbf{\Psi}_N = \frac{1}{N} \sum_{i=1}^{N} \left(\lambda_i - \bar{\lambda}\right) \left(\lambda_i - \bar{\lambda}\right)', \qquad \mathbf{\Psi}_{(r_0)} = \mathbb{E}\left[\mathbf{\Psi}_N\right]$$

an $r_0 \times r_0$ positive definite (PD) matrix, with $\bar{\lambda} = \frac{1}{N} \sum_{i=1}^{N} \lambda_i$. Then, the expected value of the sample covariance of y_i is a $T \times T$ matrix and is a function of the structural parameters of the model:

$$\mathbf{\Sigma}_{(p_0,q_0,r_0)} \left(\vartheta^0\right) = \mathbb{E}\left[\mathbf{S}_N\right] \tag{5}$$

where $\vartheta^0 = \text{vec}\left(\{\phi_s\}_{s=1}^{p_0}, \{\theta_s\}_{s=1}^{q_0}, \{\sigma_t^2\}, \mathbf{\Psi}_{(r_0)}, \mathbf{F}_{(r_0)}, \Pi_{(p_0,q_0)}\right)$ denotes the true structural parameters of the model. $\mathbf{F}_{(r_0)}$ is a $T \times r_0$ matrix where each column j corresponds to the T realizations of the $j-$th latent factor. $\Pi_{(p_0,q_0)}$ is a set of nuisance parameters (properly defined in subsection 2.3) whose dimension depends on the ARMA order (p_0, q_0). The specific functional form of $\mathbf{\Sigma}_{(p_0,q_0,r_0)} \left(\vartheta^0\right)$ is discussed through subsection 2.3. The dimension of the parameter space $\Theta_{(p_0,q_0,r_0)}$ depends on the dynamic order of the model (p_0, q_0) and the true number of factors r_0, which are, in general, unknown. We denote $\dim(\vartheta^0) = d(p_0, q_0, r_0)$.

Estimation and identification of ϑ^0 relies on the functional form of $\mathbf{\Sigma}(\vartheta)$ as $N \to \infty$ when T is fixed. In subsection 2.2 we outline the assumptions needed for these derivations

and future parts of the paper. Then, in subsection 2.3 we show how to derive $\mathbf{\Sigma}(\vartheta)$ for any given order (p, q, r).

2.2 Assumptions

The following assumptions are sufficient for the parameters of the model to be globally identified, thus permitting a consistent and asymptotically normal estimator of ϑ^0.

Assumption A1. $\mathbf{\Psi}_{(r_0)}$ *is an unrestricted symmetric and PD matrix, and* $\mathbf{F}_{(r_0)} = (\mathbf{I}_{r_0}, \mathbf{F}_2')'$.[2]

Assumption A2. $T \geq 2(r_0 + p_0 + q_0 + 1)$.

Assumption A3. *The idiosyncratic shocks* $\{\varepsilon_{it}\}$*:*

1. *Are i.i.d. over* i *and independent over* t*, satisfying* $\mathbb{E}[\varepsilon_{it}] = 0$*,* $\mathbb{V}[\varepsilon_{it}] = \sigma_t^2 > 0$ *and* $\mathbb{E}[\varepsilon_{it}^4] < \infty$.

2. *Are independent of* $\{\lambda_i\}$ *and of* $\mathbf{F}_{(r_0)}$.

3. *For each* i*, the sequence* $\{\varepsilon_{it}\}$ *is defined at least for* $t = 1 - q_0, \ldots, T$.

Assumption A4. *The factor loadings* λ_i*:*

1. *Are either random or deterministic constants with* $\mathbf{\Psi}_N \to \mathbf{\Psi}_{(r_0)}$ *as* $N \to \infty$.

2. *If random:* $\mathbb{E}\|\lambda_i\|^4 < \infty$*. If deterministic:* $\sup_i \|\lambda_i\| < \infty$.

Assumption A5. $\{\phi_s\}_{s=1}^{p_0}, \{\theta_s\}_{s=1}^{q_0}, \{\sigma_t^2\}_{t=1-q_0}^{T}, \mathbf{\Psi}_{(r_0)}$*, and* $\Pi_{(p_0,q_0)}$ *are fixed population parameters.*

[2] This assumption comes from the fact that factor models (and thus interactive effects) are identifiable up to a $r_0 \times r_0$ matrix rotation, so a normalization needs to be imposed for identification. In particular r_0^2 restrictions must be imposed. A discussion of why this is the case can be found in Section 5 of Anderson and Rubin (1956). We work with this restriction for convenience, but any other normalization that imposes r_0^2 restrictions would also work.

Assumption A6. *Let* $\mathbf{F}_{2,(r_0)}$ *be the non-fixed elements of* $\mathbf{F}_{(r_0)}$*. Then:*

1. $\mathbf{F}_{2,(r_0)}$ *is assumed to be a continuously distributed random matrix with support in a full-dimensional subset of* $\mathbb{R}^{(T-r_0)\times r_0}$*.*

2. $\mathbb{E}[\|f_t\|^4] < \infty$*.*

Assumption A7. *Restrictions on the* $\Phi(L)$ *and* $\Theta(L)$ *polynomials:*

1. *The roots of the AR and MA polynomials* $\Phi(L)$ *and* $\Theta(L)$ *lie outside the unit circle.*

2. *The polynomials* $\Phi(L)$ *and* $\Theta(L)$ *do not have any common root.*

3. *The ARMA order is minimal:* $\theta_{q_0} \neq 0$ *if* $q_0 \geq 1$ *and* $\phi_{p_0} \neq 0$ *if* $p_0 \geq 1$*.*

Assumption A8. *The parameter space* $\Theta_{(p_0,q_0,r_0)}$ *is a compact subset of* $\mathbb{R}^{d(p_0,q_0,r_0)}$*, and the true parameter* ϑ^0 *lies in its interior.*

2.3 Deriving the covariance matrix $\mathbf{\Sigma}_{(p,q,r)}\left(\vartheta_{(p,q,r)}\right)$

In this subsection we derive (5) in closed form, which is a necessary step for estimation of ϑ^0 and the model order (p_0, q_0, r_0). As we are working with a short-T dynamic panel we project the initial p observations, which introduces a set of nuisance parameters that should also be estimated. We illustrate the procedure with a homoskedastic ARMA$(1,2)$ panel, then we discuss how to derive $\mathbf{\Sigma}$ for a homoskedastic ARMA(p,q) panel and then discuss how to do so when the variance of the idiosyncratic shocks is allowed to vary over time.

2.3.1 Example: Homoskedastic ARMA(1,2) panel

Consider an ARMA(1,2) panel with r interactive effects:

$$y_{it} = \phi_1 y_{it-1} + \delta_t + \lambda_i' f_t + u_{it} \tag{6}$$

where u_{it} follows a homoskedastic MA(2) process.

Because we are in a short-T panel with an autoregressive term, the initial condition y_{i1} is not generated by (6) (which only applies from $t = 2$ onward) as y_{i0} is not observable. We handle y_{i1} by projecting it onto the factor loadings and an intercept $[1, \lambda_i']$, which yields:

$$y_{i1} = \delta_1^* + \lambda_i' f_1^* + u_{i1}^*, \qquad \mathbb{E}[u_{i1}^* \mid 1, \lambda_i] = 0$$

As δ_1^* and f_1^* are free parameters, we can simply rename δ_1^*, f_1^*, as δ_1, f_1. Then, we can drop the * notation and write:

$$y_{i1} = \delta_1 + \lambda_i' f_1 + \eta_{i1}$$

where $\eta_{i1} = u_{i1}^*$ is the projected error, which does not necessarily follow a MA(2) process and is allowed to be arbitrarily correlated with the past history of ε_{it}, $t \leq 1$ but not with the future. Stack observations over t:

$$y_i = \begin{bmatrix} y_{i1} \\ y_{i2} \\ \vdots \\ y_{iT} \end{bmatrix}, \quad \delta = \begin{bmatrix} \delta_1 \\ \delta_2 \\ \vdots \\ \delta_T \end{bmatrix}, \quad \mathbf{F}_{(r)} = \begin{bmatrix} f_1' \\ f_2' \\ \vdots \\ f_T' \end{bmatrix}, \quad \xi_i = \begin{bmatrix} \eta_{i1} \\ u_{i2} \\ \vdots \\ u_{iT} \end{bmatrix}.$$

where y_i is a $T \times 1$ vector containing the value of the outcome for each period, δ is a $T \times 1$ vector of time fixed effects, $\mathbf{F}_{(r)}$ is a $T \times r$ matrix containing the value of the r factors in each time period, and ξ_i is a $T \times 1$ vector containing the idiosyncratic errors after projecting the initial observation. Then, we can write our model in matrix notation as:

$$\mathbf{B}_{(1)} y_i = \delta + \mathbf{F}_{(r)} \lambda_i + \xi_i \tag{7}$$

where

$$\mathbf{B}_{(1)} = \begin{bmatrix} 1 & 0 & 0 & \dots & 0 \\ -\phi_1 & 1 & 0 & \dots & 0 \\ 0 & -\phi_1 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & \dots & 0 & -\phi_1 & 1 \end{bmatrix}, \qquad \mathbf{\Gamma}_{(1)} \equiv \mathbf{B}_{(1)}^{-1} = \begin{bmatrix} 1 & 0 & 0 & \dots & 0 \\ \phi_1 & 1 & 0 & \dots & 0 \\ \phi_1^2 & \phi_1 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ \phi_1^{T-1} & \phi_1^{T-2} & \dots & \phi_1 & 1 \end{bmatrix}$$

Pre-multiply both sides of (7) by $\mathbf{\Gamma}_{(1)}$ to obtain:

$$y_i = \mathbf{\Gamma}_{(1)}\delta + \mathbf{\Gamma}_{(1)}\mathbf{F}_{(r)}\lambda_i + \mathbf{\Gamma}_{(1)}\xi_i \tag{8}$$

So the covariance matrix of y_i will depend on the covariance matrix of the projected errors ξ_i. The non-projected errors u_i follow a MA(2) process and have a well-known covariance matrix given by

$$\mathbf{\Sigma}_{u,(2)} \;=\; \mathbb{E}(u_i u_i') \;=\; \begin{bmatrix} \gamma_0 & \gamma_1 & \gamma_2 & & & \\ \gamma_1 & \gamma_0 & \gamma_1 & \ddots & & \\ \gamma_2 & \gamma_1 & \gamma_0 & \ddots & & \\ & \ddots & \ddots & \ddots & \gamma_1 & \gamma_2 \\ & & & \gamma_1 & \gamma_0 & \gamma_1 \\ & & & \gamma_2 & \gamma_1 & \gamma_0 \end{bmatrix} \tag{9}$$

where $\gamma_0 = \sigma^2(1 + \theta_1^2 + \theta_2^2),\ \gamma_1 = \sigma^2(\theta_1 + \theta_1\theta_2),\ \gamma_2 = \sigma^2\theta_2$. The rest of the elements are equal to zero, so the covariance matrix of u_i is sparse.

Instead, ξ_i corresponds to the idiosyncratic error after projecting the initial observation, so the covariance matrix of ξ_i differs from the one of u_i because the projected error η_{i1} may come from a totally different distribution. Let $\sigma_{\varepsilon_t,\eta} = \text{Cov}(\varepsilon_{it}, \eta_{i1})$. Then, we have: $E[\eta_{i1}^2] = \sigma_\eta^2$; $E[\eta_{i1}u_{i2}] = \theta_1\sigma_{\varepsilon_1,\eta} + \theta_2\sigma_{\varepsilon_0,\eta}$; $E[\eta_{i1}u_{i3}] = \theta_2\sigma_{\varepsilon_1,\eta}$; and $E[\eta_{i1}u_{it}] = 0$ for all $t \geq 4$.[3]

[3] ε_0 is well-defined because of assumption A3.

Then, the covariance matrix of ξ_i can be written as:

$$\mathbf{\Sigma}_{\xi,(1,2)} = \mathbb{E}(\xi_i \xi_i') = \begin{bmatrix} \sigma_\eta^2 & \theta_1\sigma_{\varepsilon_1,\eta} + \theta_2\sigma_{\varepsilon_0,\eta} & \theta_2\sigma_{\varepsilon_1,\eta} & 0 & \cdots & 0 \\ \theta_1\sigma_{\varepsilon_1,\eta} + \theta_2\sigma_{\varepsilon_0,\eta} & \gamma_0 & \gamma_1 & \gamma_2 & 0 & \cdots \\ \theta_2\sigma_{\varepsilon_1,\eta} & \gamma_1 & \gamma_0 & \gamma_1 & \gamma_2 & \ddots \\ 0 & \gamma_2 & \gamma_1 & \gamma_0 & \ddots & \ddots \\ \vdots & 0 & \gamma_2 & \ddots & \ddots & \gamma_1 \\ 0 & \cdots & & \ddots & \gamma_1 & \gamma_0 \end{bmatrix} \tag{10}$$

Note that because $\xi_i = (\eta_{i1}, u_{i2}, ..., u_{iT})'$ the covariance matrix of the errors starting from the second element up to the last one is known and given by (9), which implies that the submatrix corresponding to rows/columns 2 through T is equal to the corresponding submatrix of $\mathbf{\Sigma}_{u,(2)}$. Instead, the first row and column of $\mathbf{\Sigma}_{\xi,(1,2)}$ are characterized by the covariance between η_{i1} and u_{it}. Thus, due to the initial projection and the MA order, we introduce three new parameters $(\sigma_\eta^2, \sigma_{\varepsilon_0,\eta}, \sigma_{\varepsilon_1,\eta})$. These should be treated as nuisance parameters and are included in ϑ for the purpose of estimation and identification. We denote the nuisance parameters by $\Pi_{(1,2)} = \left(\sigma_\eta^2, \sigma_{\varepsilon_0,\eta}, \sigma_{\varepsilon_1,\eta}\right)$.

Then, the expected value of the sample covariance matrix of y_i defined in (5) is:

$$\mathbb{E}[\mathbf{S}_N] = \mathbf{\Sigma}_{(1,2,r)}\left(\vartheta_{(1,2,r)}\right) = \mathbf{\Gamma}_{(1)}(\mathbf{F}_{(r)}\mathbf{\Psi}_{(r)}\mathbf{F}_{(r)}^{'} + \mathbf{\Sigma}_{\xi,(1,2)})\mathbf{\Gamma}_{(1)}'$$

2.3.2 Homoskedastic ARMA(p,q) panel

Consider the general case of an ARMA(p, q) panel as in (1):

$$y_{it} = \sum_{s=1}^{p} \phi_s y_{it-s} + \delta_t + \lambda_i' f_t + \varepsilon_{it} + \sum_{s=1}^{q} \theta_s \varepsilon_{it-s}$$

where the idiosyncratic shocks ε_{it} are i.i.d. over t and i satisfying $\mathbb{E}[\varepsilon_{it}] = 0$, $\mathbb{V}[\varepsilon_{it}] = \sigma^2$. The idiosyncratic error is defined in (2). As the model follows an autoregressive process of

order p, we project the first p observations onto $[1, \lambda_i']$:[4]

$$y_{it} = \delta_t^* + \lambda_i' f_t^* + u_{it}^*, \qquad \mathbb{E}[u_{it}^* \mid 1, \lambda_i] = 0, \quad t \leq p$$

Because δ_t and f_t are free parameters, we can drop the asterisks without loss of generality. Relabel the projected residual as $\eta_{it} = u_{it}^*$, where η_{it} are allowed to be arbitrarily correlated with the whole history of ε_{it}, but not with the future: $\text{Cov}(\eta_{is}, \varepsilon_{it}) = 0, \quad \forall\, t > s$. Then, we stack observations over t and write

$$y_i = \begin{bmatrix} y_{i1} \\ \vdots \\ y_{ip} \\ y_{ip+1} \\ \vdots \\ y_{iT} \end{bmatrix}, \quad \delta = \begin{bmatrix} \delta_1 \\ \vdots \\ \delta_p \\ \delta_{p+1} \\ \vdots \\ \delta_T \end{bmatrix}, \quad \mathbf{F}_{(r)} = \begin{bmatrix} f_1' \\ \vdots \\ f_p' \\ f_{p+1}' \\ \vdots \\ f_T' \end{bmatrix}, \quad \xi_i = \begin{bmatrix} \eta_{i1} \\ \vdots \\ \eta_{ip} \\ \varepsilon_{ip+1} + \sum_{s=1}^{q} \theta_s \varepsilon_{ip+1-s} \\ \vdots \\ \varepsilon_{iT} + \sum_{s=1}^{q} \theta_s \varepsilon_{iT-s} \end{bmatrix}.$$

where y_i, δ, and $\mathbf{F}_{(r)}$ are defined as before. Instead, the first p observations in the error ξ_i correspond to the projected errors, which do not necessarily follow a MA(q) process. From observation $p+1$ onward each of the errors in ξ_i follows a MA(q) process as in (2). Therefore, we can write the model in matrix notation as:

$$\mathbf{B}_{(p)} y_i = \delta + \mathbf{F}_{(r)}\, \lambda_i + \xi_i \tag{11}$$

where $\mathbf{B}_{(p)}$ is a full-rank lower triangular matrix given by:

$$\mathbf{B}_{(p)} = \begin{bmatrix} \mathbf{I}_{(p)} & \mathbf{0}_{(p)} \\ \mathbf{A}_{(p)} & \mathbf{C}_{(p)} \end{bmatrix}.$$

[4] Alternatively, for each observation $2 \leq s \leq p$ we could have projected y_{is} onto $[1, \lambda_i', y_{i1}, \ldots, y_{is-1}]$ through an AR($s-1$) projection. However, this is equivalent to the projection presented in the main text. See Appendix B.2 for a proof.

Here we only use the subscript (p) because the structure of this matrix depends only on the autoregressive order. $\mathbf{I}_{(p)}$ is the $p \times p$ identity matrix coming from the projection of the initial p observations. $\mathbf{0}_{(p)}$ is a $p \times (T-p)$ matrix of zeros. $\mathbf{A}_{(p)}$ and $\mathbf{C}_{(p)}$ are $(T-p) \times p$ and $(T-p) \times (T-p)$ matrices defined respectively as follows. The matrix $\mathbf{A}_{(p)}$ has its first row equal to $(-\phi_p, -\phi_{p-1}, \ldots, -\phi_1)$, and each subsequent row is obtained by shifting the previous row one step to the right, with a zero filled in the first column. Rows beyond the $p-$th are entirely zero. The matrix $\mathbf{C}_{(p)}$ is a full rank lower triangular Toeplitz matrix (thus invertible) with ones on the main diagonal and $-\phi_k$ on the $k-$th subdiagonal for $k = 1, \ldots, p$; all other entries are zero. Visually:

$$\mathbf{A}_{(p)} = \begin{bmatrix} -\phi_p & -\phi_{p-1} & \ldots & -\phi_2 & -\phi_1 \\ 0 & -\phi_p & \ddots & \ddots & -\phi_2 \\ 0 & 0 & \ddots & \ddots & -\phi_3 \\ \vdots & \ddots & \ddots & \ddots & \ddots \\ 0 & \ldots & \ddots & \ddots & 0 \\ 0 & \ldots & \ddots & 0 & 0 \\ 0 & \ldots & \ddots & 0 & 0 \\ 0 & \ldots & \ddots & 0 & 0 \\ 0 & \ldots & \ddots & 0 & 0 \end{bmatrix}, \quad \mathbf{C}_{(p)} = \begin{bmatrix} 1 & 0 & 0 & 0 & \ldots & \ldots & \ldots & \ldots & 0 \\ -\phi_1 & 1 & 0 & 0 & \ldots & \ldots & \ldots & \ldots & 0 \\ -\phi_2 & -\phi_1 & 1 & 0 & \ldots & \ldots & \ldots & \ldots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & 0 \\ -\phi_{p-1} & -\phi_{p-2} & \ddots & \ddots & 1 & \ddots & \ddots & \ddots & 0 \\ -\phi_p & -\phi_{p-1} & \ddots & \ddots & \ddots & 1 & \ddots & \ddots & 0 \\ 0 & -\phi_p & \ddots & \ddots & \ddots & \ddots & 1 & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & \ldots & 0 & -\phi_p & -\phi_{p-1} & \ldots & -\phi_2 & -\phi_1 & 1 \end{bmatrix}$$

The inverse of $\mathbf{B}_{(p)}$ is well-defined and given by (see Lu and Shiou (2002)):

$$\mathbf{\Gamma}_{(p)} \equiv \mathbf{B}_{(p)}^{-1} = \begin{bmatrix} \mathbf{I}_{(p)} & \mathbf{0}_{(p)} \\ -\mathbf{C}_{(p)}^{-1}\mathbf{A}_{(p)} & \mathbf{C}_{(p)}^{-1} \end{bmatrix}$$

Thus, we can pre-multiply (11) by $\mathbf{\Gamma}_{(p)}$ and write the model in matrix notation as:

$$y_i = \mathbf{\Gamma}_{(p)}\, \delta + \mathbf{\Gamma}_{(p)} \mathbf{F}_{(r)}\, \lambda_i + \mathbf{\Gamma}_{(p)}\, \xi_i \tag{12}$$

Thus, in order to characterize the expected value of the sample covariance matrix of y_i we are left to compute the covariance matrix of the idiosyncratic errors after projection (ξ_i)

defined as:

$$\boldsymbol{\Sigma}_{\xi,(p,q)} = \mathbb{E}\left[\xi_i \xi_i'\right] = \begin{bmatrix} \boldsymbol{\Sigma}_{\eta,(p)} & \boldsymbol{\Sigma}_{u,\eta,(p,q)} \\ \boldsymbol{\Sigma}'_{u,\eta,(p,q)} & \boldsymbol{\Sigma}_{u,(p,q)} \end{bmatrix}. \tag{13}$$

Note that the structure of this matrix will in general depend both on the autoregressive and moving average order (p, q). Here, $\boldsymbol{\Sigma}_{\eta,(p)}$ is a $p \times p$ matrix of the covariances among the p projected initial-error terms.

$$\boldsymbol{\Sigma}_{\eta,(p)} = \begin{bmatrix} \sigma^2_{\eta 1} & \sigma_{\eta 12} & \dots & \sigma_{\eta 1 p} \\ \sigma_{\eta 12} & \sigma^2_{\eta 2} & \dots & \sigma_{\eta 2 p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{\eta 1 p} & \sigma_{\eta 2 p} & \dots & \sigma^2_{\eta p} \end{bmatrix}. \tag{14}$$

where all the $\frac{p(p+1)}{2}$ unique coefficients that characterize this matrix are nuisance parameters introduced by the projection of the initial p observations.

Furthermore, $\boldsymbol{\Sigma}_{u,(p,q)}$ is a $(T-p) \times (T-p)$ matrix containing the covariance between the $T - p$ (non-projected) idiosyncratic errors, each following a MA(q). Thus, it is analogous to (9) for a MA(q) model with $T - p$ observations (see Hamilton (2020) for a derivation).

Finally, $\boldsymbol{\Sigma}_{u,\eta,(p,q)}$ is a $p \times (T - p)$ matrix containing the cross-covariances between the $T - p$ non-projected idiosyncratic errors $u_{i,p+1}, \dots, u_{iT}$ and the first p projected errors $\eta_{i1}, \dots, \eta_{ip}$. By construction of the projection, the projected errors η_{is} are orthogonal to future innovations, that is

$$\mathrm{Cov}(\eta_{is}, \varepsilon_{it}) = 0 \qquad \forall\, t > s.$$

Since each u_{it} depends only on the current and past q innovations, it follows that

$$\mathrm{Cov}(\eta_{is}, u_{it}) = 0 \qquad \text{whenever } t > s + q.$$

As a result, $\boldsymbol{\Sigma}_{u,\eta,(p,q)}$ is sparse: for each row $s \leq p$, only the columns corresponding to $t = p + 1, \dots, \min(T, s + q)$ may be non-zero, while all remaining entries are identically

zero. Note that because of Assumption A2, the minimum is always equal to $s+q$ because $s+q \leq p+q < T$, so for each row s the only columns that may be non zero are $t = p+1, \ldots, s+q$ (with the usual convention that if $p+1 > s+q$ then all columns are zero). In particular, if $q = 0$ (pure AR model), then $\mathbf{\Sigma}_{u,\eta,(p,0)}$ is the zero matrix.

The non-zero entries of $\mathbf{\Sigma}_{u,\eta,(p,q)}$ can be parameterized in terms of the covariances between the projected errors and past shocks. Let

$$\sigma_{\eta_s \varepsilon_k} \equiv \text{Cov}(\eta_{is}, \varepsilon_{ik}), \qquad s = 1, \ldots, p, \ k \leq s.$$

Then, for $t = p+1, \ldots, T$,

$$\text{Cov}(\eta_{is}, u_{it}) = \sum_{\ell=0}^{q-(t-s)} \theta_{t-s+\ell}\, \sigma_{\eta_s \varepsilon_{s-\ell}}, \quad \text{if } p+1 \leq t \leq s+q, \ \text{ and } 1 \geq s \geq p \tag{15}$$

and $\text{Cov}(\eta_{is}, u_{it}) = 0$ if $t > s+q$. The $\sigma_{\eta_s \varepsilon_k}$ parameters appear only in $\mathbf{\Sigma}_{u,\eta,(p,q)}$ and nowhere else in $\mathbf{\Sigma}_{\xi,(p,q)}$ and should be treated as nuisance parameters. The total number of new parameters introduced by the cross-covariance block is therefore

$$K(p,q) = \begin{cases} \dfrac{q(q+1)}{2}, & q \leq p, \\ \\ p(q-p) + \dfrac{p(p+1)}{2}, & q > p, \end{cases} \tag{16}$$

Therefore, the $\mathbf{\Sigma}_{\xi,(p,q)}$ matrix has $p(p+1)/2 + K(p,q) + q + 1$ parameters: the first $p(p+1)/2$ come from the covariance between the projected errors $\mathbf{\Sigma}_{\eta,(p)}$, the next $K(p,q)$ come from covariance between the projected errors and the true idiosyncratic errors $\mathbf{\Sigma}_{u,\eta,(p,q)}$, and the last $q+1$ come from the covariance between the $MA(q)$ idiosyncratic errors, which are parameterized by q moving average coefficients $\{\theta_s\}_{s=1}^{q}$ and σ^2. We stack the nuisance parameters introduced into

$$\Pi_{(p,q)} = \left(\{\sigma^2_{\eta_s}\}_{s=1}^{p}, \ \{\sigma_{\eta_{ij}}\}_{1 \leq i < j \leq p}, \ \{\sigma_{\eta_s, \varepsilon_k} : s = \max\{1, p+1-q\}, \ldots, p, \ \ k = p+1-q, \ldots, s\} \right) \tag{17}$$

Overall, these results imply that the expected value of the sample covariance matrix of y_i is given by

$$\mathbb{E}[\mathbf{S}_N] = \mathbf{\Sigma}_{(p,q,r)}\left(\vartheta_{(p,q,r)}\right) = \mathbf{\Gamma}_{(p)}(\mathbf{F}_{(r)}\mathbf{\Psi}_{(r)}\mathbf{F}'_{(r)} + \mathbf{\Sigma}_{\xi,(p,q)})\mathbf{\Gamma}'_{(p)}$$

where under time-series homoskedasticity $\vartheta^0 = \text{vec}\left(\{\phi_s\}_{s=1}^{p_0}, \{\theta_s\}_{s=1}^{q_0}, \sigma^2, \mathbf{\Psi}_{(r_0)}, \mathbf{F}_{(r_0)}, \Pi_{(p_0,q_0)}\right)$.

2.3.3 Heteroskedastic ARMA(p,q) panel

While it simplifies notation and exposition, the assumption $\mathbb{V}(\varepsilon_{it}) = \sigma^2$ can be relaxed into a more general version where heteroskedasticity is allowed and $\mathbb{V}(\varepsilon_{it}) = \sigma_t^2$. In this setting, the idiosyncratic error u_{it} is a linear combination of the present and the past q shocks with MA coefficients θ_s as before. However, the shocks $\varepsilon_{i,t-s}$ are drawn from potentially different distributions with variances given by σ_{t-s}^2, though they remain independent over time. A model of this kind nests both the model discussed in the previous examples where $\mathbb{V}(\varepsilon_{it}) = \sigma^2$, and the model discussed in Bai (2024) and Bai and Mones (2025) where it was assumed that the autoregressive order was equal to one, the MA order was equal to zero, but the variance of the shocks was allowed to vary over time.

Allowing for heteroskedastic innovations affects only the specification of the non-projected innovation block of $\mathbf{\Sigma}_{\xi,(p,q)}$, previously denoted as $\mathbf{\Sigma}_{u,(p,q)}$, through the introduction of time-varying variances $\{\sigma_t^2\}$. The blocks $\mathbf{\Sigma}_{\eta,(p)}$ and $\mathbf{\Sigma}_{u,\eta,(p,q)}$ continue to be treated as nuisance components with the same dimension and sparsity restrictions as in the homoskedastic ARMA model of subsection 2.3.2, although their numerical values may differ. For this reason, the discussion about these matrices is not repeated, as it coincides with the one offered in the homoskedastic case.

To define the counterpart of $\mathbf{\Sigma}_{u,(p,q)}$ for the heteroskedastic case first stack shocks over

t and write

$$\varepsilon_i = (\varepsilon_{i,1-q}, \varepsilon_{i,2-q}, \ldots, \varepsilon_{i,T})', \qquad \mathbf{D} = \text{diag}(\sigma^2_{1-q}, \sigma^2_{2-q}, \ldots, \sigma^2_T),$$

where ε_i is a $(T+q) \times 1$ vector of idiosyncratic shocks with $\mathbb{V}(\varepsilon_{it}) = \sigma^2_t$ and covariance matrix defined by $\mathbf{D}$. The first element of ε_i is $\varepsilon_{i,1-q}$ as it is required in order for u_{i1} to be well-defined (see Assumption A3). Moreover, define $\mathbf{\Theta}_{(q)}$ as the $T \times (T+q)$ banded Toeplitz matrix generated by the MA coefficients $\{\theta_\ell\}_{\ell=0}^q$, $\theta_0 = 1$, such that

$$u_i = (u_{i1}, \ldots, u_{iT})' = \mathbf{\Theta}_{(q)}\, \varepsilon_i,$$

i.e. the t-th row of $\mathbf{\Theta}_{(q)}$ contains $(\theta_q, \ldots, \theta_1, 1)$ in columns $(t, \ldots, t+q)$ and zeros elsewhere.

Then, under heteroskedasticity the covariance matrix of the idiosyncratic errors u_i is defined by the following $T \times T$ matrix:

$$\mathbf{\Sigma}^+_{u,(q)} = \mathbf{\Theta}_{(q)}\, \mathbf{D}\, \mathbf{\Theta}'_{(q)}, \tag{18}$$

We denote the heteroskedastic covariance matrices with a superscript $+$. Under time-series homoskedasticity $\mathbf{D} = \sigma^2 \mathbf{I}_{T+q}$, so the definition collapses to the covariance matrix used in the previous subsection. Also, if $q = 0$ the covariance matrix is diagonal and collapses to the one employed in the baseline proof of Bai and Mones (2025).

In the definition of ξ_i the first p observations are projected errors, so the structure of $\mathbf{\Sigma}^+_{u,(q)}$ only applies to the $(p+1)$-th observation onward. Then, define $\mathbf{\Sigma}^+_{u,(p,q)}$ as the $(T-p) \times (T-p)$ submatrix of $\mathbf{\Sigma}^+_{u,(q)}$ corresponding to rows and columns $p+1$ to T. This matrix is q-banded, in the sense that its (t,s) element is equal to zero whenever $|t-s| > q$.

Note that $\mathbf{\Sigma}^+_{u,(p,q)}$ depends only on $\{\sigma^2_t\}_{t=p+1-q}^T$, so it effectively introduces $T-p+q-1$ new parameters with respect to the homoskedastic specification. In particular, the total number of parameters in $\mathbf{\Sigma}^+_{u,(p,q)}$ under heteroskedasticity is given by the MA coefficients

$\{\theta_s\}_{s=1}^{q}$ together with the $T-p+q$ innovation variances $\{\sigma_t^2\}_{t=p+1-q}^{T}$ that enter the covariance matrix of $(u_{i,p+1},\ldots,u_{i,T})'$. Observe that variances dated prior to $t=p+1-q$ affect only the first p observations of u_i and therefore do not appear in $\mathbf{\Sigma}^{+}_{u,(p,q)}$ so their values cannot be identified.

Replacing $\mathbf{\Sigma}_{u,(p,q)}$ with $\mathbf{\Sigma}^{+}_{u,(p,q)}$ in (13) yields the heteroskedastic version of $\mathbf{\Sigma}_{\xi,(p,q)}$ (we do not add the + from now on to avoid making the notation more burdensome but it should be clear that we always work with the heteroskedastic model as a baseline specification). Then, the expected value of the sample covariance matrix of y_i under heteroskedasticity is given by:

$$\mathbb{E}[\mathbf{S}_N] = \mathbf{\Sigma}_{(p,q,r)}\left(\vartheta_{(p,q,r)}\right) = \mathbf{\Gamma}_{(p)}(\mathbf{F}_{(r)}\mathbf{\Psi}_{(r)}\mathbf{F}'_{(r)} + \mathbf{\Sigma}_{\xi,(p,q)})\mathbf{\Gamma}'_{(p)} \tag{19}$$

where now $\vartheta^0 = \text{vec}\left(\{\phi_s\}_{s=1}^{p_0}, \{\theta_s\}_{s=1}^{q_0}, \{\sigma_t^2\}_{t=p_0+1-q_0}^{T}, \mathbf{\Psi}_{(r_0)}, \mathbf{F}_{(r_0)}, \Pi_{(p_0,q_0)}\right)$. This will be our working expression for the covariance matrix from now on. Overall, the total number of parameters in a model of order (p,q,r) is:

$$d(p,q,r) = p+q+(T-p+q)+r(r+1)/2+(T-r)r+p(p+1)/2+K(p,q)$$

3 Estimation and inference for ϑ^0

The dimension of the unknown parameters ϑ^0 depends on the true ARMA order and number of factors (p_0,q_0,r_0). If we think the researcher does not know the true order (p_0,q_0,r_0) then the dimension of the structural parameters ϑ^0 is also unknown. Thus, it is convenient to separate the problem of estimating ϑ^0 conditional on the true orders (p_0,q_0,r_0) from the problem of estimating the orders (p_0,q_0,r_0) *per-se*.

In this section, we consider the problem of estimating ϑ^0 using either Gaussian Quasi Maximum Likelihood (QMLE) or the Generalized Method of Moments (GMM), conditional on the model order being known and equal to (p_0,q_0,r_0). We show that under Assumptions

A1-A8 both methods yield consistent and asymptotically normal estimators of ϑ^0. Given this result, constructing confidence intervals for ϑ or subsets of the parameters follows immediately once we show how to derive a consistent estimator for the asymptotic variance of $\widehat{\vartheta}$. Model selection and estimation of (p_0, q_0, r_0) is discussed in Section 4.

Consider the model presented in Equations (1)-(2). We have shown that the expected value of the sample covariance matrix of y_i is given by (19). Once this matrix has been correctly parameterized we can discuss how to estimate the parameters ϑ^0 of the model via QMLE and GMM. Throughout this section, because (p_0, q_0, r_0) are known, we denote $\mathbf{\Sigma}(\vartheta)$ as a shorthand for $\mathbf{\Sigma}_{(p_0,q_0,r_0)}(\vartheta_{(p_0,q_0,r_0)})$ and $\mathbf{\Sigma}(\vartheta^0)$ as the covariance matrix evaluated at the true parameters to simplify notation. Moreover, we refer to $d(p_0, q_0, r_0)$ as d_0.

3.1 QMLE

Bai (2024) proposes a QMLE approach to estimate the parameters of an AR(1) panel for the case where $u_{it} \sim (0, \sigma_t^2)$ i.i.d. over i. We extend this method for the case where y_{it} follows an ARMA(p_0, q_0) process. Define $\delta^\dagger \equiv \mathbf{\Gamma}_{(p_0)}\delta$. The Gaussian quasi log-likelihood function is given by

$$\ell_N(\vartheta) = -\frac{N}{2}\ln|\mathbf{\Sigma}(\vartheta)| - \frac{1}{2}\sum_{i=1}^{N}(y_i - \delta^\dagger)'\mathbf{\Sigma}(\vartheta)^{-1}(y_i - \delta^\dagger)$$

Concentrating out $\delta^\dagger$, which is estimated by $\bar{y}$, yields:

$$\ell_N^{\text{conc}}(\vartheta) \;=\; -\frac{N}{2}\ln|\mathbf{\Sigma}(\vartheta)| - \frac{N}{2}\text{tr}\left[\mathbf{S}_N\mathbf{\Sigma}(\vartheta)^{-1}\right].$$

Dividing by N, we obtain[5]

$$\frac{1}{N}\ell_N^{conc}(\vartheta) = -\frac{1}{2}\ln|\mathbf{\Sigma}(\vartheta)| - \frac{1}{2}\text{tr}\left[\mathbf{S}_N\mathbf{\Sigma}(\vartheta)^{-1}\right]$$

[5]In the MLE, the definition of $\mathbf{S}_N$ needs a slight modification, namely, $1/(N-1)$ is replaced by $1/N$.

Then, the QMLE estimator of ϑ is defined as:

$$\widehat{\vartheta}_{QMLE} = \arg \max_{\vartheta \in \Theta_{(p_0,q_0,r_0)}} - \ln |\mathbf{\Sigma}(\vartheta)| - \operatorname{tr}\left[\mathbf{S}_N \mathbf{\Sigma}(\vartheta)^{-1}\right] \tag{20}$$

For a fixed T and as $N \to \infty$ by the law of large numbers, $\mathbf{S}_N \xrightarrow{p} \mathbf{\Sigma}(\vartheta^0)$, so we have:

$$\frac{2}{N}\ell_N(\vartheta) \xrightarrow{p} - \ln |\mathbf{\Sigma}(\vartheta)| - \operatorname{tr}\left[\mathbf{\Sigma}(\vartheta^0)\mathbf{\Sigma}(\vartheta)^{-1}\right] \tag{21}$$

where $\mathbf{\Sigma}(\vartheta)$ is the same as in (19) evaluated at an arbitrary $\vartheta \in \Theta_{(p_0,q_0,r_0)}$. Bai and Mones (2025) show that ϑ^0 is globally identified in this context, meaning that ϑ^0 is the unique maximizer of (21) conditional on correct specification of the model order (p_0, q_0, r_0).

Under Assumptions A1-A8 and correct specification of the model order (p_0, q_0, r_0), standard QMLE theory White (1982) implies that the QMLE is consistent and asymptotically normal:

Theorem 1. *Under Assumptions A1-A8 and correct order (p_0, q_0, r_0) specification, the QMLE estimator is consistent and asymptotically normal.*

$$\widehat{\vartheta}_{QMLE} \xrightarrow{p} \vartheta^0, \qquad \sqrt{N}\left(\widehat{\vartheta}_{QMLE} - \vartheta^0\right) \xrightarrow{d} \mathcal{N}(0, \mathbf{V}_{QMLE})$$

The asymptotic variance of the QMLE is given by

$$\mathbf{V}_{QMLE} = \mathbf{H}^{-1}\mathbf{J}\mathbf{H}^{-1},$$

where

$$\mathbf{H} = \mathbb{E}\left[\frac{\partial^2 \ell_i(\vartheta^0)}{\partial \vartheta \, \partial \vartheta'}\right], \qquad \mathbf{J} = \mathbb{E}\left[\frac{\partial \ell_i(\vartheta^0)}{\partial \vartheta}\frac{\partial \ell_i(\vartheta^0)}{\partial \vartheta}'\right]$$

and $\ell_i(\vartheta)$ denotes the per–individual log-likelihood contribution defined as:

$$\ell_i(\vartheta) = -\frac{1}{2}\ln|\mathbf{\Sigma}(\vartheta)| - \frac{1}{2}\left(y_i - \bar{y}\right)' \mathbf{\Sigma}(\vartheta)^{-1}\left(y_i - \bar{y}\right),$$

where $\bar{y} = N^{-1}\sum_{i=1}^{N} y_i$. A consistent estimator of $\mathbf{V}_{QMLE}$ is given by

$$\widehat{\mathbf{V}}_{QMLE} = \widehat{\mathbf{H}}^{-1}\widehat{\mathbf{J}}\widehat{\mathbf{H}}^{-1},$$

where

$$\widehat{\mathbf{H}} = \frac{1}{N}\sum_{i=1}^{N} \frac{\partial^2 \ell_i(\widehat{\vartheta})}{\partial \vartheta \, \partial \vartheta'}, \qquad \widehat{\mathbf{J}} = \frac{1}{N}\sum_{i=1}^{N} \left(\frac{\partial \ell_i(\widehat{\vartheta})}{\partial \vartheta}\right)\left(\frac{\partial \ell_i(\widehat{\vartheta})}{\partial \vartheta}\right)'.$$

3.2 GMM

Another alternative consists in estimating ϑ through the Generalized Method of Moments (GMM) using information contained in $\mathbf{\Sigma}(\vartheta)$. Let $\{y_i\}_{i=1}^N$ denote the stacked outcomes. For each $\vartheta \in \Theta_{(p_0,q_0,r_0)}$ (treating $\mathbf{F}$ as a fixed component of ϑ), the outcomes $\{y_i\}_{i=1}^N$ are independent across i conditional on $\mathbf{F}$ and satisfy

$$\mathbb{E}\left(\frac{1}{N}\sum_{i=1}^{N} y_i\right) = \mathbf{\Gamma}_{(p_0)}\delta, \qquad \mathbb{E}[\mathbf{S}_N] = \mathbf{\Sigma}(\vartheta^0),$$

where $\delta \in \mathbb{R}^T$ is unrestricted and $\mathbf{\Sigma}(\vartheta) \in \mathbb{R}^{T\times T}$ is defined in (19) and parameterized by $\vartheta \in \Theta_{(p_0,q_0,r_0)} \subset \mathbb{R}^{d_0}$. Define

$$s = \text{vech}(\mathbf{S}_N), \qquad \sigma(\vartheta) = \text{vech}(\mathbf{\Sigma}(\vartheta)), \qquad \dot{T} = \frac{T(T+1)}{2}.$$

The scalar $\dot{T}$ is the number of unique second-moment conditions implied by symmetry of $\mathbf{S}_N$ and $\mathbf{\Sigma}(\vartheta)$. Accordingly, s and $\sigma(\vartheta)$ are $\dot{T}\times 1$ vectors collecting the distinct elements of $\mathbf{S}_N$ and $\mathbf{\Sigma}(\vartheta)$, respectively. Define the i-specific moment function

$$g_i(\vartheta) = s_i - \sigma(\vartheta), \qquad s_i = \text{vech}\left((y_i - \bar{y})(y_i - \bar{y})'\right),$$

so that the sample moment vector is

$$\bar{g}_N(\vartheta) = \frac{1}{N}\sum_{i=1}^{N} g_i(\vartheta) = \frac{1}{N}\sum_{i=1}^{N} s_i - \sigma(\vartheta) = s - \sigma(\vartheta).$$

We therefore use the moment conditions

$$\bar{g}_N(\vartheta) = s - \sigma(\vartheta), \tag{22}$$

which satisfy $\mathbb{E}[\bar{g}_N(\vartheta^0)] = 0$ under correct specification of $\boldsymbol{\Sigma}(\vartheta)$. Let $\mathbf{W}_N \in \mathbb{R}^{\dot{T}\times\dot{T}}$ be a positive definite weighting matrix such that $\mathbf{W}_N \xrightarrow{p} \mathbf{W}$, where $\mathbf{W}$ is positive definite. The GMM objective function is

$$J_N(\vartheta) = N\,\bar{g}_N(\vartheta)'\mathbf{W}_N\,\bar{g}_N(\vartheta). \tag{23}$$

And the normalized GMM objective function is given by

$$Q_N(\vartheta) = \frac{1}{N} J_N(\vartheta). \tag{24}$$

Since T is fixed, the number of moment conditions $\dot{T}$ does not grow with N, and standard GMM asymptotics apply. Therefore, the GMM estimator is defined as:

$$\widehat{\vartheta}_{\text{GMM}} = \arg \min_{\vartheta\in\Theta_{(p_0,q_0,r_0)}} J_N(\vartheta) \tag{25}$$

Under Assumptions A1-A8 the GMM estimator $\widehat{\vartheta}_{\text{GMM}}$ is consistent and asymptotically normal (Newey and McFadden, 1994), which is formalized below:

Theorem 2. *Under Assumptions A1-A8 and correct order specification* (p_0, q_0, r_0) *the GMM estimator is consistent and asymptotically normal*

$$\widehat{\vartheta}_{GMM} \xrightarrow{p} \vartheta^0, \qquad \sqrt{N}(\widehat{\vartheta}_{GMM} - \vartheta^0) \xrightarrow{d} \mathcal{N}(0, \mathbf{V}_{GMM})$$

where:

$$\mathbf{V}_{\text{GMM}} = (\mathbf{G}'\mathbf{W}\mathbf{G})^{-1}\mathbf{G}'\mathbf{W}\boldsymbol{\Upsilon}\mathbf{W}\mathbf{G}(\mathbf{G}'\mathbf{W}\mathbf{G})^{-1} \tag{26}$$

with:

$$\mathbf{G} = \frac{\partial\sigma(\vartheta)}{\partial\vartheta'}\Big|_{\vartheta=\vartheta^0}, \qquad \boldsymbol{\Upsilon} = \mathbb{E}\left(g_i(\vartheta^0)g_i(\vartheta^0)'\right)$$

An optimal GMM estimator can be obtained using $\mathbf{W} = \boldsymbol{\Upsilon}^{-1}$, which implies that:

$$\mathbf{V}_{\text{Eff}} = (\mathbf{G}'\boldsymbol{\Upsilon}^{-1}\mathbf{G})^{-1} \tag{27}$$

While $\boldsymbol{\Upsilon}$ is unobserved, it can be consistently estimated either by:

$$\widehat{\boldsymbol{\Upsilon}} = \frac{1}{N}\sum_{i=1}^{N}(s_i - s)(s_i - s)'$$

Or by a two-step GMM using an arbitrary PD matrix $\mathbf{W}_N^{(1)}$, then obtaining an initial (consistent) GMM estimator $\widehat{\vartheta}_{GMM}^{(1)}$ and then estimating $\boldsymbol{\Upsilon}$ as:

$$\widehat{\boldsymbol{\Upsilon}} = \frac{1}{N}\sum_{i=1}^{N} g_i(\widehat{\vartheta}_{GMM}^{(1)}) g_i(\widehat{\vartheta}_{GMM}^{(1)})'$$

Under correct specification and overidentification ($\dot{T} > d_0$), the GMM objective function evaluated at $\widehat{\vartheta}_{\text{GMM}}$ with $\mathbf{W}_N = \widehat{\boldsymbol{\Upsilon}}^{-1}$ satisfies (Hansen, 1982):

$$J_N(\widehat{\vartheta}_{\text{GMM}}) \xrightarrow{d} \chi^2_{\dot{T}-d_0}.$$

3.3 Inference and Hypothesis Testing

We can use the results from the previous subsections to do inference and hypothesis testing within the framework of this model. While here we discuss just some of the possibilities, many of the classical testing procedures are available once we have established that our estimator is consistent and asymptotically normal.

For instance, t-tests for individual parameters can be carried out as usual for every element contained in ϑ. Let $\vartheta_k^0 \in \vartheta^0$ be a parameter of interest and ϑ_k a hypothesized value of this parameter. Consider the following hypothesis test $H_0 : \vartheta_k^0 = \vartheta_k$, $H_1 : \vartheta_k^0 \neq \vartheta_k$. Then, under H_0 the $t-$statistic is asymptotically normal:

$$T = \frac{\widehat{\vartheta}_k - \vartheta_k}{\widehat{s}(\widehat{\vartheta}_k)} \xrightarrow{d} \mathcal{N}(0,1)$$

where $\widehat{\vartheta}_k$ denotes a consistent estimator of ϑ_k and $\widehat{s}(\widehat{\vartheta}_k) = \sqrt{\widehat{\mathbf{V}}_{kk}/N}$ a consistent estimator for the standard error of $\widehat{\vartheta}_k$. Therefore, a test at significance level α rejects H_0 whenever $|T| > c_{1-\frac{\alpha}{2}}$, where $c_{1-\frac{\alpha}{2}}$ denotes the $1-\frac{\alpha}{2}$ left quantile of a standard normal distribution.

Confidence intervals for the model parameters can be constructed by inverting this $t-$test so that a $100(1-\alpha)\%$ confidence interval for ϑ_k^0 is given by:

$$\mathcal{C}_{1-\alpha}^k = \left\{ \widehat{\vartheta}_k - c_{1-\frac{\alpha}{2}} \widehat{s}(\widehat{\vartheta}_k) \ ; \ \widehat{\vartheta}_k + c_{1-\frac{\alpha}{2}} \widehat{s}(\widehat{\vartheta}_k) \right\}$$

Similarly, Wald tests for joint restrictions of the type $\mathbf{R}\vartheta = c$ (where $\mathbf{R}$ is a $\ell \times d_0$ matrix and c is a $\ell \times 1$ vector) can be conducted in the usual way. For instance, if a researcher is interested in testing whether $H_0 : \mathbf{R}\vartheta = c$ against $H_1 : \mathbf{R}\vartheta \neq c$, she can use the fact that

$$W = N \left(\mathbf{R}\widehat{\vartheta} - c \right)' \left[\mathbf{R}\widehat{\mathbf{V}}\mathbf{R}' \right]^{-1} \left(\mathbf{R}\widehat{\vartheta} - c \right) \xrightarrow{d} \chi_\ell^2.$$

Hence, the null hypothesis H_0 is rejected at significance level α whenever the statistic exceeds the left $(1-\alpha)$-quantile of a chi-square distribution with ℓ degrees of freedom, i.e.

$$W > \chi_{\ell,(1-\alpha)}^2.$$

Similarly, a joint confidence set for a subset of parameters of the model can be obtained by inverting the corresponding Wald test. Let ϑ_S denote a $\ell \times 1$ subvector of ϑ, and let $\mathbf{R}$ be a $\ell \times d_0$ selection matrix such that $\mathbf{R}\vartheta = \vartheta_S$. Then, a $100(1-\alpha)\%$ joint confidence set for ϑ_S^0 is given by

$$\mathcal{C}_{1-\alpha}^S = \left\{ \vartheta_S \in \mathbb{R}^\ell : N \left(\mathbf{R}\widehat{\vartheta} - \vartheta_S \right)' \left[\mathbf{R}\widehat{\mathbf{V}}\mathbf{R}' \right]^{-1} \left(\mathbf{R}\widehat{\vartheta} - \vartheta_S \right) \leq \chi_{\ell,(1-\alpha)}^2 \right\}.$$

Extension to nonlinear restrictions can be carried out as usual employing the Delta method.

4 Model selection

In this section we construct a consistent estimator of the model order, indexed by the triple of integers (p_0, q_0, r_0). We begin with the case in which the dynamic order (p_0, q_0) is known and only the number of factors r_0 is unknown. This setting provides useful intuition and encompasses many short-T dynamic panel models with interactive effects (e.g., Hayakawa

et al., 2023; Bai, 2024), where the dynamic specification (p_0, q_0) is taken as given. Following Bai and Ng (2002), we construct an information criterion that consistently selects the true number of factors.

We then consider the joint estimation of the model order (p_0, q_0, r_0). This problem is more demanding, as identification of the ARMA order typically relies on large-T asymptotics and frequency-domain arguments. In the present setting, however, we exploit the availability of cross-sectional variation together with the restriction that the dynamic coefficients are common across individuals. Under one additional assumption, we show that an information criterion can consistently select (p_0, q_0, r_0) even when the time dimension T is fixed, provided that the model order is sufficiently small relative to T.

4.1 Estimating r_0

Here we discuss how to estimate the true number of factors r_0 conditional on the true autoregressive and moving average order (p_0, q_0) being known. In this setting, the dimension of the parameter vector depends only on the number of factors used for estimation r so the expected value of the covariance matrix of y_i is indexed by $\vartheta_{(r)} \equiv \vartheta_{(p_0,q_0,r)}$ as follows:

$$\boldsymbol{\Sigma}(\vartheta_{(r)}) = \boldsymbol{\Gamma}_{(p_0)}\big(\mathbf{F}_{(r)}\boldsymbol{\Psi}_{(r)}\mathbf{F}'_{(r)} + \boldsymbol{\Sigma}_{\xi,(p_0,q_0)}\big)\boldsymbol{\Gamma}'_{(p_0)}$$

$$\vartheta_{(r)} = \text{vec}\left(\{\phi_s\}_{s=1}^{p_0}, \{\theta_s\}_{s=1}^{q_0}, \{\sigma_t^2\}_{t=p_0+1-q_0}^{T}, \boldsymbol{\Psi}_{(r)}, \mathbf{F}_{(r)}, \Pi_{(p_0,q_0)}\right)$$

Let $\mathcal{M}_{(r)}$ denote the class of covariance matrices that can be represented in this form.

$$\mathcal{M}_{(r)} \equiv \mathcal{M}_{(p_0,q_0,r)} = \left\{\boldsymbol{\Sigma}_{(p_0,q_0,r)}(\vartheta_{(r)}) : \vartheta_{(r)} \in \Theta_{(p_0,q_0,r)}\right\}.$$

where by definition $\boldsymbol{\Sigma}_{(p_0,q_0,r_0)}(\vartheta_{(r_0)}) \in \mathcal{M}_{(r_0)}$. The sequence $\{\mathcal{M}_{(r)}\}_{r\geq 0}$ is nested:

$$\mathcal{M}_{(0)} \subset \mathcal{M}_{(1)} \subset \cdots \subset \mathcal{M}_{(r_{\max})},$$

where $\mathcal{M}_{(0)}$ corresponds to the model without interactive effects and $r_{\max} = \lfloor \frac{T}{2} - p_0 - q_0 - 1 \rfloor$ denotes the maximum number of factors under which we are sure global identification holds (Assumption A2). Note that the nested models satisfy the following condition:

Lemma 1. *Under Assumptions A1-A8 the true covariance matrix* $\mathbf{\Sigma}_{(p_0,q_0,r_0)}(\vartheta_{(r_0)})$ *belongs to* $\mathcal{M}_{(r_0)}$ *but not to* $\mathcal{M}_{(r)}$ *for any* $r < r_0$.

The result is very intuitive: we cannot obtain an observationally equivalent representation of an ARMA(p_0, q_0) panel with r_0 interactive fixed effects using only $r < r_0$ interactive fixed effects. A proof of this result is available in Appendix A.1.

For each candidate rank r, denote the GMM estimator and the value of the objective function evaluated at the GMM estimator by[6]:

$$\widehat{\vartheta}_{(r)} \in \arg \min_{\vartheta_{(r)} \in \Theta_{(p_0,q_0,r)}} Q_N(\vartheta_{(r)}), \qquad \widehat{Q}_{N(r)} = Q_N(\widehat{\vartheta}_{(r)}), \tag{28}$$

where Q_N is defined as in (24). Moreover, let $d(r)$ denote the dimension of $\vartheta_{(r)}$. Using the results derived in Section 2.3 we know that for each r:

$$d(r) = p_0 + q_0 + (T - p_0 + q_0) + r(r+1)/2 + (T-r)r + p_0(p_0+1)/2 + K(p_0, q_0) \tag{29}$$

The proposed information criterion is

$$IC_N(r) = \widehat{Q}_{N(r)} + \kappa \, \frac{d(r) \log N}{N}, \qquad r = 0, 1, \ldots, r_{\max}, \tag{30}$$

where $\kappa > 0$ is a penalty parameter and $r_{\max}$ is fixed. The selected number of factors minimizes this information criterion:

$$\widehat{r} = \arg \min_{0 \le r \le r_{\max}} IC_N(r). \tag{31}$$

This permits consistently estimating the true number of factors, which is formalized below:

[6]Derivations using the QMLE estimator and objective function are analogous, thus omitted.

Theorem 3. *Under Assumptions A1-A8 and correct dynamic specification* (p_0, q_0)*:*

$$\lim_{N\to\infty} \mathbb{P}(\widehat{r} = r_0) = 1$$

The Proof of Theorem 3 is in the Appendix A.2

4.2 Joint estimation of (p_0, q_0, r_0)

In this subsection we discuss how to jointly estimate the model order (p_0, q_0, r_0). The expected covariance matrix of y_i depends on the true structural parameter vector $\vartheta_{(p_0,q_0,r_0)} \in \Theta_{(p_0,q_0,r_0)}$, whose dimension is unknown. For any triple (p, q, r), we parameterize the expected value of the sample covariance matrix of y_i as

$$\boldsymbol{\Sigma}(\vartheta_{(p,q,r)}) = \boldsymbol{\Gamma}_{(p)}\big(\mathbf{F}_{(r)}\boldsymbol{\Psi}_{(r)}\mathbf{F}'_{(r)} + \boldsymbol{\Sigma}_{\xi,(p,q)}\big)\boldsymbol{\Gamma}'_{(p)}$$

$$\vartheta_{(p,q,r)} = \mathrm{vec}\Big(\{\phi_s\}_{s=1}^{p}, \{\theta_s\}_{s=1}^{q}, \{\sigma_t^2\}_{t=p+1-q}^{T}, \boldsymbol{\Psi}_{(r)}, \mathbf{F}_{(r)}, \Pi_{(p,q)}\Big)$$

Similarly, denote by $\mathcal{M}_{(p,q,r)}$ the class of covariance matrices that admit such a representation,

$$\mathcal{M}_{(p,q,r)} = \big\{\boldsymbol{\Sigma}_{(p,q,r)}(\vartheta) : \vartheta \in \Theta_{(p,q,r)}\big\},$$

where we drop the (p, q, r) subscript in ϑ to simplify notation. The set of feasible model orders (those satisfying Assumption A2) is

$$\mathcal{G} = \big\{(p, q, r) \in \mathbb{N} :\ 2(p + q + r + 1) \le T\big\}.$$

Note that the sequence $\{\mathcal{M}_{(p,q,r)}\}_{(p,q,r)\in\mathcal{G}}$ is nested with respect to monotone index comparisons:

$$\mathcal{M}_{(p,q,r)} \subset \mathcal{M}_{(p',q',r')}, \qquad (p, q, r) < (p', q', r'),$$

where $(p, q, r) < (p', q', r')$ means that each component is weakly smaller and at least one is strictly smaller.

This nesting property, however, does not allow for comparison across all models. When the only unknown order was r_0, a model with r_0 factors trivially nests any model with $r < r_0$ factors. In the present setting, an analogous argument applies only when (p, q, r) and (p_0, q_0, r_0) are ordered componentwise. By contrast, models with nonmonotonic order differences (e.g., $p > p_0$ but $q < q_0$ and $r < r_0$) are not nested, so direct comparison is no longer straightforward. Consequently, a different approach is required.

Let $d(p, q, r)$ denote the effective dimension of the (p, q, r) model. From Section 2.3,

$$d(p, q, r) = p + q + (T - p + q) + r(r + 1)/2 + (T - r)r + p(p + 1)/2 + K(p, q)$$

To compare models we define a pre-order $\preceq$ on $\mathbb{N}^3$ as follows. For two triples (p, q, r) and (p', q', r'), we write

$$(p, q, r) \preceq (p', q', r') \iff d(p, q, r) \leq d(p', q', r')$$

It is important to note that $(p, q, r) \preceq (p', q', r')$ does not imply $(p, q, r) \leq (p', q', r')$ componentwise, while the converse holds because $d(p, q, r)$ is componentwise monotone. Moreover, $\preceq$ is not antisymmetric: it is possible that two distinct triples satisfy both $(p, q, r) \preceq (p', q', r')$ and $(p', q', r') \preceq (p, q, r)$. Finally, we define the associated strict order $\prec$ by

$$(p, q, r) \prec (p', q', r') \iff d(p, q, r) < d(p', q', r').$$

A sufficient condition for identification of the true model order (p_0, q_0, r_0) is that the true covariance matrix $\mathbf{\Sigma}_{(p_0,q_0,r_0)}(\vartheta^0)$ cannot be generated by any alternative specification $(p, q, r) \in \mathcal{G}$ different from (p_0, q_0, r_0) satisfying $(p, q, r) \preceq (p_0, q_0, r_0)$. In other words, no model with at least one strictly smaller order and weakly fewer parameters, $d(p, q, r) \leq d(p_0, q_0, r_0)$, is observationally equivalent to the true model. We refer to this

condition as *order minimality*. Formally, define the set

$$\mathcal{S}_{\Theta_0} = \left\{ \vartheta \in \Theta_{(p_0,q_0,r_0)} : \exists (p,q,r) \preceq (p_0,q_0,r_0),\ (p,q,r) \neq (p_0,q_0,r_0),\ \text{s.t. } \boldsymbol{\Sigma}_{(p_0,q_0,r_0)}(\vartheta) \in \mathcal{M}_{(p,q,r)} \right\}$$

We say that ϑ^0 is *order minimal* if $\vartheta^0 \notin \mathcal{S}_{\Theta_0}$.

Elements of $\mathcal{S}_{\Theta_0}$ correspond to parameter configurations for which there exists an observationally equivalent model with at least one strictly smaller component and weakly smaller effective dimension. Such equivalence requires exact algebraic cancellations among the AR, MA, and factor components and therefore constitutes a knife-edge phenomenon. This leads to the following sufficient condition for identification of the true model order (p_0, q_0, r_0):

Assumption A9. *The true parameter is order minimal:* $\vartheta^0 \notin \mathcal{S}_{\Theta_0}$.

That is, the true covariance matrix cannot be generated by any lower-order specification, which is something we should expect for most parameterizations, specially considering assumptions A1 and A7. Indeed, we conjecture that under Assumptions A1-A8 the set $\mathcal{S}_{\Theta_0}$ is either empty or has Lebesgue measure zero in $\Theta_{(p_0,q_0,r_0)}$, as it requires the parameters of the model to lie in a particular lower dimensional manifold in order for them to be interchangeable/redundant. However, providing a formal proof of this conjecture for a short T panel is challenging due to the non-linearity of the problem and the presence of nuisance parameters given the projection of initial observations.

Given Assumption A9, we now show that it is possible to construct an information criterion that consistently selects the true model order. For each candidate triple $(p, q, r) \in \mathcal{G}$, define the GMM estimator and associated objective value as

$$\widehat{\vartheta}_{(p,q,r)} \in \arg \min_{\vartheta_{(p,q,r)} \in \Theta_{(p,q,r)}} Q_N(\vartheta_{(p,q,r)}), \qquad \widehat{Q}_{(p,q,r)} = Q_N(\widehat{\vartheta}_{(p,q,r)}).$$

We then define the information criterion

$$IC_N(p,q,r) = \widehat{Q}_{N(p,q,r)} + \kappa \, \frac{d(p,q,r)\log N}{N} \tag{32}$$

where $\kappa > 0$ is a constant. The selected model orders are given by

$$(\widehat{p},\widehat{q},\widehat{r}) = \arg \min_{(p,q,r)\in\mathcal{G}} IC_N(p,q,r) \tag{33}$$

The following result establishes the consistency of this procedure.

Theorem 4. *Under Assumptions A1-A9,*

$$\lim_{N\to\infty} \mathbb{P}\big((\widehat{p},\widehat{q},\widehat{r}) = (p_0,q_0,r_0)\big) = 1.$$

The proof of Theorem 4 is provided in Appendix A.3.

5 Conclusion

This paper studies short-T, large-N dynamic panel models with interactive effects in which the idiosyncratic component follows a general ARMA (p_0, q_0) process. The central object is the implied $T \times T$ covariance structure of each individual time series, which combines (i) dynamic dependence through the ARMA component, (ii) common shocks through a low-rank factor structure, and (iii) time-series heteroskedasticity in the innovations. By exploiting the covariance restrictions implied by the model, we develop Gaussian QMLE and second-moment GMM estimators that operate in levels (without differencing) and are therefore free of incidental-parameter (Nickell) bias when T is fixed.

We establish consistency and asymptotic normality of both QMLE and second-moment GMM as $N \to \infty$ with fixed T, and we provide feasible variance estimation which permits conducting hypothesis testing and constructing confidence intervals for the structural parameters. We also develop procedures for model selection that consistently recover the

number of factors and the dynamic orders. In particular, we formalize an order-minimality principle under the joint ARMA–factor structure, which rules out observationally equivalent lower-order representations, and we propose information criteria that consistently select (p_0, q_0, r_0).

Taken together, these results provide a unified framework for estimation and model selection in dynamic panels with flexible time-series and cross-sectional dependence. The approach offers a practical alternative to transformation-based methods and opens the door to richer dynamic specifications in short panels without relying on large-T asymptotics.

Appendix A Proof of Main Results

A.1 Proof of Lemma 1

Suppose, by contradiction, that there exists an alternative parameterization of order (p_0, q_0, r) with $r < r_0$ such that $\mathbf{\Sigma}_{(p_0,q_0,r_0)}(\vartheta_{(r_0)}) \in \mathcal{M}_{(p_0,q_0,r)}$. Then there must exist parameters $\vartheta_{(r)}$ of order (p_0, q_0, r) satisfying

$$\mathbf{\Gamma}_0 \left(\mathbf{F}_0 \mathbf{\Psi}_0 \mathbf{F}_0' + \mathbf{\Sigma}_{\xi,0}\right) \mathbf{\Gamma}_0' = \mathbf{\Gamma}_{(p_0)} \left(\mathbf{F}_{(r)} \mathbf{\Psi}_{(r)} \mathbf{F}_{(r)}' + \mathbf{\Sigma}_{\xi,(p_0,q_0)}\right) \mathbf{\Gamma}_{(p_0)}'$$

Here the subscript 0 denotes the true parameters. The right-hand side corresponds to an alternative representation of order (p_0, q_0, r) with $r < r_0$. Although $\mathbf{\Gamma}_{(p_0)}$ has the same structural form as $\mathbf{\Gamma}_0$, its coefficients are free parameters $\{\phi_s\}_{s=1}^{p_0}$ and need not coincide with the true values $\{\phi_{s0}\}_{s=1}^{p_0}$. The same convention applies to all other parameters on the right-hand side.

Pre-multiplying by $\mathbf{B}_{(p_0)} = \mathbf{\Gamma}_{(p_0)}^{-1}$ and post-multiplying by $\mathbf{B}_{(p_0)}'$ yields

$$\mathbf{B}_{(p_0)} \mathbf{\Gamma}_0 \left(\mathbf{F}_0 \mathbf{\Psi}_0 \mathbf{F}_0' + \mathbf{\Sigma}_{\xi,0}\right) \mathbf{\Gamma}_0' \mathbf{B}_{(p_0)}' = \mathbf{F}_{(r)} \mathbf{\Psi}_{(r)} \mathbf{F}_{(r)}' + \mathbf{\Sigma}_{\xi,(p_0,q_0)}$$

Let

$$R = \{r_1, \ldots, r_{r+1}\}, \qquad C = \{c_1, \ldots, c_{r+1}\}$$

be subsets of $\{p_0 + 1, \ldots, T\}$ such that

$$|i - j| > q_0 \quad \text{for all } i \in R,\ j \in C.$$

We consider the $(r+1) \times (r+1)$ submatrix of $\mathbf{F}_{(r)} \mathbf{\Psi}_{(r)} \mathbf{F}_{(r)}' + \mathbf{\Sigma}_{\xi,(p_0,q_0)}$ indexed by (R, C). Because $|i-j| > q_0$ for all $i \in R$ and $j \in C$ and both indexes are subsets of $\{p_0+1, \ldots, T\}$, all corresponding entries of $\mathbf{\Sigma}_{\xi,(p_0,q_0)}$ lie outside the q_0-band and therefore vanish by sparsity. Hence,

$$\left(\mathbf{F}_{(r)} \mathbf{\Psi}_{(r)} \mathbf{F}_{(r)}' + \mathbf{\Sigma}_{\xi,(p_0,q_0)}\right)_{[R,C]} = \left(\mathbf{F}_{(r)} \mathbf{\Psi}_{(r)} \mathbf{F}_{(r)}'\right)_{[R,C]}.$$

Since $\mathbf{F}_{(r)}\mathbf{\Psi}_{(r)}\mathbf{F}'_{(r)}$ has rank r, every $(r+1)\times(r+1)$ minor of this matrix has determinant zero. Therefore,

$$\det\Big(\big(\mathbf{F}_{(r)}\mathbf{\Psi}_{(r)}\mathbf{F}'_{(r)}+\mathbf{\Sigma}_{\xi,(p_0,q_0)}\big)_{[R,C]}\Big)=\det\Big(\big(\mathbf{F}_{(r)}\mathbf{\Psi}_{(r)}\mathbf{F}'_{(r)}\big)_{[R,C]}\Big)=0.$$

The existence of such index sets R and C is guaranteed by Assumption A2, which ensures that there are sufficiently many indices outside the q_0-band.

This implies that for every admissible pair (R, C) constructed as above, the equality

$$\det\left(\Big(\mathbf{B}_{(p_0)}\mathbf{\Gamma}_0(\mathbf{F}_0\mathbf{\Psi}_0\mathbf{F}'_0+\mathbf{\Sigma}_{\xi,0})\mathbf{\Gamma}'_0\mathbf{B}'_{(p_0)}\Big)_{[R,C]}\right)=0$$

must hold. This argument is in the spirit of Bai and Mones (2025): we have derived a collection of necessary conditions that any alternative parameterization must satisfy in order for $\mathbf{\Sigma}_{(p_0,q_0,r)}(\vartheta_{(p_0,q_0,r)})=\mathbf{\Sigma}(\vartheta^0)$. Importantly, these conditions depend only on the true structural parameters ϑ^0 and on the alternative autoregressive coefficients $\phi_1,\ldots,\phi_{p_0}$, through $\mathbf{B}_{(p_0)}$.

To simplify the argument (which is already sufficient for our purposes), we restrict attention to a subset of admissible index pairs (R, C) consisting only of contiguous row and column blocks. Define the collection of contiguous subsets of $\{p_0+1,\ldots,T\}$ of cardinality $r+1$ as

$$\mathcal{I}_{r+1}^{T,p_0} \;=\; \Big\{\{a,a+1,\ldots,a+r\}\subseteq\{p_0+1,\ldots,T\}:\; a\in\{p_0+1,\ldots,T-r\}\Big\}.$$

Next, define the set of row-column index pairs generating an $(r+1)\times(r+1)$ contiguous minor outside the q_0-band as

$$\mathcal{C}(T,p_0,q_0,r) \;=\; \Big\{(R,C)\in\mathcal{I}_{r+1}^{T,p_0}\times\mathcal{I}_{r+1}^{T,p_0}:\; |r_i-c_j|>q_0 \quad \forall\, r_i\in R,\ \forall\, c_j\in C\Big\}.$$

Under Assumption A2, the number of distinct (up to symmetry) row-column contiguous submatrices satisfying these restrictions is

$$N_{\text{contig}}(T,p_0,q_0,r)=\frac{(T-p_0-2r-q_0)(T-p_0-2r-q_0-1)}{2},$$

see Appendix B.3 for a proof. Therefore, the conditions above generate the system

$$\det\left(\left(\mathbf{B}_{(p_0)}\mathbf{\Gamma}_0(\mathbf{F}_0\mathbf{\Psi}_0\mathbf{F}_0' + \mathbf{\Sigma}_{\xi,0})\mathbf{\Gamma}_0'\mathbf{B}_{(p_0)}'\right)_{[R,C]}\right) = 0, \qquad (R,C) \in \mathcal{C}(T, p_0, q_0, r), \tag{A.1}$$

which consists of $N_{\text{contig}}(T, p_0, q_0, r)$ polynomial equations in the free parameters $\phi = (\phi_1, \ldots, \phi_{p_0})$. To analyze this system, note first that since $r < r_0$ by assumption, the matrix

$$\mathbf{B}_{(p_0)}\mathbf{\Gamma}_0\mathbf{F}_0\mathbf{\Psi}_0\mathbf{F}_0'\mathbf{\Gamma}_0'\mathbf{B}_{(p_0)}'$$

has rank $r_0 \geq r + 1$. Using the identity (see Appendix B.4)

$$\mathbf{B}_{(p_0)}\mathbf{\Gamma}_0 = I_T + M_{(p_0)}\sum_{k=1}^{p_0}(\phi_{k0} - \phi_k)L_k,$$

it follows that, for each fixed $(R,C) \in \mathcal{C}(T, p_0, q_0, r)$, the determinant appearing in (A.1) is a polynomial function of $\phi_1, \ldots, \phi_{p_0}$ whose coefficients depend on the factor realizations $\mathbf{F}_0$. This polynomial is not identically zero. In particular, the polynomial determinant in (A.1) admits a decomposition in which the determinant of $(\mathbf{F}_0\mathbf{\Psi}_0\mathbf{F}_0')[R,C]$ appears as a nonvanishing summand (see Proposition 1 in Bai and Mones (2025) for a proof). Assumption A6 implies that this summand is nonzero with probability one, imposing that this constant term is nonzero with probability one.

Consequently, with probability one, every equation in (A.1) is not identically zero as a polynomial in $\phi_1, \ldots, \phi_{p_0}$. Since

$$p_0 < N_{\text{contig}}(T, p_0, q_0, r),$$

the system (A.1) contains more nontrivial polynomial equations than unknowns. Such a system can admit a solution only if the parameter vector $\phi = (\phi_1, \ldots, \phi_{p_0})$ lies in a proper algebraic subset of $\mathbb{R}^{p_0}$, that is, only if some of the equations are algebraically redundant.

By Assumption A6, the factor realizations $\mathbf{F}_0$ are continuously distributed on a full-dimensional support. Since the coefficients of the polynomials in (A.1) are continuous

functions of $\mathbf{F}_0$, the event that these polynomials are algebraically redundant occurs with probability zero.

Therefore, with probability one, the system (A.1) admits no solution. Equivalently, at least one determinant in (A.1) is nonzero, contradicting the existence of an alternative parameterization with $r < r_0$. □

A.2 Proof of Theorem 3:

The proof follows the same argument as in the proof of Theorem 4, but restricting attention to the subclass of models with the correct dynamic order $(p, q) = (p_0, q_0)$ and varying only the factor dimension r. That is, we minimize the information criterion over the reduced index set

$$\mathcal{G}_0 = \{(p_0, q_0, r) \in \mathcal{G} : \ r \in \mathbb{N}\}.$$

With this restriction, the candidate model classes are nested in r, and the three-case comparison in Theorem 4 simplifies to the standard underfitting/overfitting comparison in the factor dimension (Bai and Ng, 2002). Therefore, consistency of the selected factor number in Theorem 3 is obtained by the same steps as in Theorem 4.

A.3 Proof of Theorem 4

Proving Theorem 4 requires deriving some intermediate results that are essential for this proof but not for the general understanding of the paper. We state and prove these results directly here. The GMM objective function is:

$$Q_N(\vartheta_{(p,q,r)}) = \left\|\text{vech}\left(\mathbf{S}_N - \mathbf{\Sigma}_{(p,q,r)}(\vartheta_{(p,q,r)})\right)\right\|^2_{\mathbf{W}_N}, \qquad \widehat{Q}_{N,(p,q,r)} = \inf_{\vartheta_{(p,q,r)} \in \Theta_{(p,q,r)}} Q_N(\vartheta_{(p,q,r)}),$$

and the GMM population criterion

$$Q(\vartheta_{(p,q,r)}) = \left\|\text{vech}\left(\mathbf{\Sigma}_{(p_0,q_0,r_0)}\left(\vartheta^0\right) - \mathbf{\Sigma}_{(p,q,r)}(\vartheta_{(p,q,r)})\right)\right\|^2_{\mathbf{W}}, \qquad Q^*_{(p,q,r)} = \inf_{\vartheta_{(p,q,r)} \in \Theta_{(p,q,r)}} Q(\vartheta_{(p,q,r)}),$$

Lemma A.1. *For each triple* $(p,q,r) \in \mathcal{G}$ *such that* $\boldsymbol{\Sigma}_{(p_0,q_0,r_0)}(\vartheta^0) \in \mathcal{M}_{(p,q,r)}$,

$$\widehat{Q}_{N,(p,q,r)} \; = O_p\big(N^{-1}\big)\,.$$

Proof. Fix a triple (p,q,r) satisfying $\boldsymbol{\Sigma}_{(p_0,q_0,r_0)}(\vartheta^0) \in \mathcal{M}_{(p,q,r)}$. Define

$$u_N = \operatorname{vech}\left(\mathbf{S}_N - \boldsymbol{\Sigma}_{(p_0,q_0,r_0)}\left(\vartheta^0\right)\right), \qquad m(\vartheta_{(p,q,r)}) = \operatorname{vech}\left(\boldsymbol{\Sigma}_{(p_0,q_0,r_0)}\left(\vartheta^0\right) - \boldsymbol{\Sigma}(\vartheta_{(p,q,r)})\right).$$

Then $\operatorname{vech}\big(\mathbf{S}_N - \boldsymbol{\Sigma}(\vartheta_{(p,q,r)})\big) = u_N + m(\vartheta_{(p,q,r)})$ and therefore, for any $\vartheta_{(p,q,r)} \in \Theta_{(p,q,r)}$,

$$\begin{aligned} Q_N(\vartheta_{(p,q,r)}) &= (u_N + m(\vartheta_{(p,q,r)}))'\mathbf{W}_N(u_N + m(\vartheta_{(p,q,r)})) \\ &= u_N'\mathbf{W}_N u_N + 2u_N'\mathbf{W}_N m(\vartheta_{(p,q,r)}) + m(\vartheta_{(p,q,r)})'\mathbf{W}_N m(\vartheta_{(p,q,r)}) \\ &= u_N'\mathbf{W}_N u_N + 2u_N'\mathbf{W}_N m(\vartheta_{(p,q,r)}) + Q(\vartheta_{(p,q,r)}). \end{aligned} \tag{A.2}$$

Since T is fixed a multivariate central limit theorem implies:

$$\sqrt{N}\, u_N \;\Rightarrow\; \mathcal{N}(0, \boldsymbol{\Upsilon})$$

meaning that

$$\|u_N\| = O_p(N^{-1/2}),$$

and hence

$$u_N'\mathbf{W}_N u_N \;\le\; \|\mathbf{W}_N\|\,\|u_N\|^2 \;=\; O_p(N^{-1}), \tag{A.3}$$

where $\|\mathbf{W}_N\|$ denotes the spectral norm. Because by assumption $\boldsymbol{\Sigma}_{(p_0,q_0,r_0)}(\vartheta^0) \in \mathcal{M}_{(p,q,r)}$, there exists $\bar{\vartheta}_{(p,q,r)} \in \Theta_{(p,q,r)}$ such that

$$\boldsymbol{\Sigma}_{(p,q,r)}(\bar{\vartheta}_{(p,q,r)}) = \boldsymbol{\Sigma}_{(p_0,q_0,r_0)}\left(\vartheta^0\right).$$

Equivalently,

$$m(\bar{\vartheta}_{(p,q,r)}) = \operatorname{vech}\big(\boldsymbol{\Sigma}_{(p_0,q_0,r_0)}\left(\vartheta^0\right) - \boldsymbol{\Sigma}_{(p,q,r)}(\bar{\vartheta}_{(p,q,r)})\big) = 0,$$

so that

$$Q(\bar{\vartheta}_{(p,q,r)}) = 0.$$

Evaluating (A.2) at $\bar{\vartheta}_{(p,q,r)}$ yields

$$Q_N(\bar{\vartheta}_{(p,q,r)}) = u_N' \mathbf{W}_N u_N.$$

Since

$$\widehat{Q}_{N,(p,q,r)} = \inf_{\vartheta_{(p,q,r)} \in \Theta_{(p,q,r)}} Q_N(\vartheta_{(p,q,r)}),$$

it follows that

$$0 \leq \widehat{Q}_{N,(p,q,r)} \leq Q_N(\bar{\vartheta}_{(p,q,r)}) = u_N' \mathbf{W}_N u_N = O_p(N^{-1}).$$

Combining this bound with (A.3) gives

$$\widehat{Q}_{N,(p,q,r)} = O_p(N^{-1}),$$

as claimed.

□

Lemma A.2. *For any triple* $(p,q,r) \in \mathcal{G}$ *such that* $\mathbf{\Sigma}_{(p_0,q_0,r_0)}(\vartheta^0) \notin \mathcal{M}_{(p,q,r)}$,

$$\widehat{Q}_{N,(p,q,r)} \xrightarrow{p} Q^*_{(p,q,r)} > 0.$$

Proof. Fix such a triple. Recall from the previous Lemma

$$u_N = \text{vech}\left(\mathbf{S}_N - \mathbf{\Sigma}_{(p_0,q_0,r_0)}\left(\vartheta^0\right)\right), \qquad m(\vartheta_{(p,q,r)}) = \text{vech}\left(\mathbf{\Sigma}_{(p_0,q_0,r_0)}\left(\vartheta^0\right) - \mathbf{\Sigma}(\vartheta_{(p,q,r)})\right).$$

and that for any $\vartheta_{(p,q,r)} \in \Theta_{(p,q,r)}$,

$$Q_N(\vartheta_{(p,q,r)}) = u_N' \mathbf{W}_N u_N + 2u_N' \mathbf{W}_N m(\vartheta_{(p,q,r)}) + m(\vartheta_{(p,q,r)})' \mathbf{W}_N m(\vartheta_{(p,q,r)}).$$

The first two terms are $o_p(1)$ uniformly over $\Theta_{(p,q,r)}$ because $\|u_N\| = O_p(N^{-1/2})$ and $\sup_\vartheta \|m(\vartheta)\|$ is bounded (by compactness of $\Theta_{(p,q,r)}$ and continuity of m). Indeed,

$$\sup_\vartheta |u_N' \mathbf{W}_N m(\vartheta)| \leq \|u_N\| \cdot \|\mathbf{W}_N\| \cdot \sup_\vartheta \|m(\vartheta)\| = O_p(N^{-1/2}),$$

and similarly $u_N' \mathbf{W}_N u_N = O_p(N^{-1})$.

For the third term, note that by the uniform bound on $m(\vartheta)$,

$$\sup_{\vartheta} \left| m(\vartheta)' \mathbf{W}_N m(\vartheta) - m(\vartheta)' \mathbf{W} m(\vartheta) \right| \leq \|\mathbf{W}_N - \mathbf{W}\| \cdot \sup_{\vartheta} \|m(\vartheta)\|^2 = o_p(1).$$

Since the set $\Theta_{(p,q,r)}$ is compact and the function $m(\vartheta)' \mathbf{W} m(\vartheta)$ is continuous, the infimum $Q^*_{(p,q,r)}$ is attained and positive under the assumption $\mathbf{\Sigma}_0 \notin \mathcal{M}_{(p,q,r)}$. Therefore,

$$\inf_{\vartheta} m(\vartheta)' \mathbf{W}_N m(\vartheta) = \inf_{\vartheta} m(\vartheta)' \mathbf{W} m(\vartheta) + o_p(1) = Q^*_{(p,q,r)} + o_p(1).$$

Putting the three parts together,

$$\widehat{Q}_{N,(p,q,r)} = \inf_{\vartheta_{(p,q,r)} \in \Theta_{(p,q,r)}} Q_N(\vartheta) = Q^*_{(p,q,r)} + o_p(1),$$

which is the desired convergence in probability. □

Given these Lemmas we can finally prove Theorem 4:

Proof of Theorem 4. We aim to show that the selected model order

$$(\widehat{p}, \widehat{q}, \widehat{r}) = \arg \min_{(p,q,r) \in \mathcal{G}} IC_N(p,q,r)$$

satisfies

$$\lim_{N \to \infty} \mathbb{P}\left((\widehat{p}, \widehat{q}, \widehat{r}) = (p_0, q_0, r_0) \right) = 1.$$

Recall the criterion:

$$IC_N(p,q,r) = \widehat{Q}_{N(p,q,r)} + \kappa \, \frac{d(p,q,r) \log N}{N},$$

So that:

$$IC_N(p,q,r) - IC_N(p_0,q_0,r_0) = \left(\widehat{Q}_{N(p,q,r)} - \widehat{Q}_{N(p_0,q_0,r_0)} \right) + \kappa \left(d(p,q,r) - d(p_0,q_0,r_0) \right) \frac{\log N}{N}$$

Therefore, for every $(p,q,r) \in \mathcal{G}$ such that $(p,q,r) \neq (p_0,q_0,r_0)$, we must show

$$\lim_{N \to \infty} \mathbb{P}\left(IC_N(p,q,r) - IC_N(p_0,q_0,r_0) > 0 \right) = 1.$$

There are two different cases:

Case 1: Consider any triple $(p, q, r) \in \mathcal{G}$ different from (p_0, q_0, r_0) such that $(p, q, r) \preceq (p_0, q_0, r_0)$. By Assumption A9 $\boldsymbol{\Sigma}_{(p_0,q_0,r_0)}(\vartheta^0) \notin \mathcal{M}_{(p,q,r)}$, and therefore by Lemma A.2,

$$\widehat{Q}_{N(p,q,r)} \xrightarrow{p} Q^*_{(p,q,r)} > 0.$$

Hence,

$$\lim_{N\to\infty} IC_N(p, q, r) - IC_N(p_0, q_0, r_0) \xrightarrow{p} Q^*_{(p,q,r)} > 0.$$

This implies

$$\lim_{N\to\infty} \mathbb{P}\left(IC_N(p, q, r) - IC_N(p_0, q_0, r_0) > 0\right) = 1.$$

Case 2: Let $(p, q, r) \in \mathcal{G}$ satisfy $(p, q, r) \succ (p_0, q_0, r_0)$, that is, $d(p, q, r) > d(p_0, q_0, r_0)$. Two situations are possible.

(i) $\boldsymbol{\Sigma}_{(p_0,q_0,r_0)}(\vartheta^0) \in \mathcal{M}_{(p,q,r)}$. In this case, the true covariance matrix admits a representation under the (p, q, r) specification. By Lemma A.1,

$$\widehat{Q}_{N(p,q,r)} = O_p(N^{-1}), \qquad \widehat{Q}_{N(p_0,q_0,r_0)} = O_p(N^{-1}).$$

Let $\Delta_N = \widehat{Q}_{N(p,q,r)} - \widehat{Q}_{N(p_0,q_0,r_0)} = O_p(N^{-1})$. Then

$$IC_N(p, q, r) - IC_N(p_0, q_0, r_0) = \Delta_N + \kappa\big(d(p, q, r) - d(p_0, q_0, r_0)\big)\frac{\log N}{N}.$$

Since $d(p, q, r) > d(p_0, q_0, r_0)$ and $\kappa > 0$, the penalty difference is strictly positive and of order $\log N/N$, which dominates $\Delta_N = O_p(1/N)$. Therefore,

$$\lim_{N\to\infty} \mathbb{P}(IC_N(p, q, r) - IC_N(p_0, q_0, r_0) > 0) = 1.$$

(ii) $\boldsymbol{\Sigma}_{(p_0,q_0,r_0)}(\vartheta^0) \notin \mathcal{M}_{(p,q,r)}$. In this case, the (p, q, r) model is misspecified. By Lemma A.2,

$$\widehat{Q}_{N(p,q,r)} \xrightarrow{p} Q^*_{(p,q,r)} > 0.$$

Since $\widehat{Q}_{N(p_0,q_0,r_0)} = O_p(N^{-1})$, it follows immediately that

$$\lim_{N\to\infty} \mathbb{P}(IC_N(p,q,r) - IC_N(p_0,q_0,r_0) > 0) = 1.$$

Combining Cases 1 and 2, we conclude that for every $(p,q,r) \in \mathcal{G}$ with $(p,q,r) \neq (p_0,q_0,r_0)$,

$$\lim_{N\to\infty} \mathbb{P}(IC_N(p,q,r) - IC_N(p_0,q_0,r_0) > 0) = 1.$$

Hence, the information criterion selects the true model order with probability approaching one:

$$\lim_{N\to\infty} \mathbb{P}\big((\widehat{p},\widehat{q},\widehat{r}) = (p_0,q_0,r_0)\big) = 1.$$

□

Appendix B Additional Derivations

B.1 Autoregressive errors

We show that when y_{it} is an ARMA(p,q) panel and u_{it} has an autoregressive order $\dot{p}$ then the model is observationally equivalent to an ARMA$(p+\dot{p},q)$ where the AR coefficients are just a combination of the outcome and error AR coefficients.

Proof: Consider the model:

$$y_{it} = \sum_{s=1}^{p} \phi_s y_{it-s} + \delta_t + \lambda_i' f_t + u_{it}$$

where everything is defined as in (1) but now u_{it} follow an ARMA$(\dot{p},q)$ process:

$$u_{it} = \sum_{s=1}^{\dot{p}} \dot{\phi}_s u_{it-s} + \varepsilon_{it} + \sum_{s=1}^{q} \vartheta_s \varepsilon_{it-s}$$

where $\{\dot{\phi}_s\}_{s=1}^{\dot{p}}$ denote the autoregressive coefficients of the idiosyncratic error. Thus:

$$y_{it} = \sum_{s=1}^{p} \phi_s y_{it-s} + \delta_t + \lambda_i' f_t + \sum_{s=1}^{\dot{p}} \dot{\phi}_s u_{it-s} + \varepsilon_{it} + \sum_{s=1}^{q} \vartheta_s \varepsilon_{it-s}$$

Now, note that we can always write:

$$u_{it-s} = y_{it-s} - \sum_{k=1}^{p} \phi_k y_{it-s-k} - \delta_{t-s} - \lambda_i' f_{t-s}$$

Substitute this expression into the original model:

$$y_{it} = \sum_{s=1}^{p} \phi_s y_{it-s} + \delta_t + \lambda_i' f_t + \sum_{s=1}^{\dot{p}} \dot{\phi}_s \left(y_{it-s} - \sum_{k=1}^{p} \phi_k y_{it-s-k} - \delta_{t-s} - \lambda_i' f_{t-s} \right) + \varepsilon_{it} + \sum_{s=1}^{q} \vartheta_s \varepsilon_{it-s}$$

Which implies that the model allows an ARMA$(p + \dot{p}, q)$ representation in y_{it}:

$$y_{it} = \sum_{s=1}^{p+\dot{p}} \rho_s y_{it-s} + \widetilde{\delta}_t + \lambda_i' \widetilde{f}_t + \varepsilon_{it} + \sum_{s=1}^{q} \vartheta_s \varepsilon_{it-s}$$

where

$$\rho_s = \phi_s + \dot{\phi}_s - \sum_{k=1}^{s-1} \phi_k \dot{\phi}_{s-k}, \qquad \widetilde{\delta}_t = \delta_t - \sum_{s=1}^{\dot{p}} \dot{\phi}_s \delta_{t-s}, \qquad \widetilde{f}_t = f_t - \sum_{s=1}^{\dot{p}} \dot{\phi}_s f_{t-s},$$

□

B.2 Equivalent projection

We illustrate this point with an ARMA$(2, q)$ for simplicity, but it is immediate to see how the same logic applies to ARMA(p, q) models. Consider the model:

$$y_{it} = \phi_1 y_{it-1} + \phi_2 y_{it-2} + \delta_t + \lambda_i' f_t + u_{it}.$$

As the model follows an ARMA$(2, q)$ process, we need to project the first two observations. Consider the following projection:

$$y_{i1} = \delta_1^* + \lambda_i' f_1^* + u_{i1}^*, \qquad \mathbb{E}[u_{i1}^* \mid 1, \lambda_i] = 0,$$

$$y_{i2} = \phi_1^* y_{i1} + \delta_2^* + \lambda_i' f_2^* + u_{i2}^*, \qquad \mathbb{E}[u_{i2}^* \mid 1, \lambda_i, y_{i1}] = 0$$

Substitute the projection of the initial observation

$$\begin{aligned} y_{i2} &= \phi_1^* \left(\delta_1^* + \lambda_i' f_1^* + u_{i1}^*\right) + \delta_2^* + \lambda_i' f_2^* + u_{i2}^* \\ &= \phi_1^* \delta_1^* + \phi_1^* \lambda_i' f_1^* + \phi_1^* u_{i1}^* + \delta_2^* + \lambda_i' f_2^* + u_{i2}^* \\ &= (\phi_1^* \delta_1^* + \delta_2^*) + \lambda_i' (\phi_1^* f_1^* + f_2^*) + (\phi_1^* u_{i1}^* + u_{i2}^*) \\ &= \delta_2^{**} + \lambda_i' f_2^{**} + u_{i2}^{**} \end{aligned}$$

where

$$\delta_2^{**} = \phi_1^* \delta_1^* + \delta_2^*; \qquad f_2^{**} = \phi_1^* f_1^* + f_2^*; \qquad u_{i2}^{**} = \phi_1^* u_{i1}^* + u_{i2}^*$$

Finally, note that:

$$\mathbb{E}[u_{i2}^{**} | 1, \lambda_i] = \mathbb{E}[u_{i2}^* | 1, \lambda_i] + \phi_1^* \mathbb{E}[u_{i1}^* | 1, \lambda_i] = \mathbb{E}\left[\mathbb{E}[u_{i2}^* | 1, \lambda_i, y_{i1}] | 1, \lambda_i\right] + 0 = 0$$

As these are free parameters, we can just drop one asterisk without loss of generality, which corresponds to the representation in the main text. □

B.3 Counting contiguous exclusion minors

Consider contiguous index sets of cardinality $r + 1$,

$$R(a) = \{a, a+1, \ldots, a+r\}, \qquad C(b) = \{b, b+1, \ldots, b+r\},$$

where $a, b \in \{p_0 + 1, \ldots, T\}$. The submatrix of $\mathbf{\Sigma}_{\xi,(p_0,q_0)}$ indexed by $(R(a), C(b))$ lies completely outside the q-band if and only if

$$|i - j| > q_0 \quad \forall\, i \in R(a),\ \forall\, j \in C(b).$$

Since $R(a)$ and $C(b)$ are intervals, this condition is equivalent to

$$\max R(a) + q_0 < \min C(b) \quad \text{or} \quad \max C(b) + q_0 < \min R(a).$$

To count each minor only once up to symmetry, we restrict attention to the one-sided condition

$$\max R(a) + q_0 < \min C(b),$$

which enforces that $R(a)$ lies to the left of $C(b)$ and avoids double counting of $(R(a), C(b))$ and $(C(b), R(a))$. Now $\max R(a) = a + r$ and $\min C(b) = b$, so the one-sided condition becomes

$$a + r + q_0 < b \quad \Longleftrightarrow \quad b \geq a + r + q_0 + 1.$$

Moreover, $C(b) \subseteq \{p_0 + 1, \ldots, T\}$ requires $b \leq T - r$. Hence, for a fixed a, the admissible choices of b are

$$b \in \{a + r + q_0 + 1,\ a + r + q_0 + 2,\ \ldots,\ T - r\},$$

whose cardinality is

$$(T - r) - (a + r + q_0 + p_0 + 1) + 1 \ = \ T - p_0 - 2r - q_0 - a.$$

Therefore admissible b exist if and only if $T - p_0 - 2r - q_0 - a \geq 1$, i.e. $a \leq T - p_0 - 2r - q_0 - 1$. Summing over all such a gives

$$N_{\text{contig}}(T, p_0, q_0, r) = \sum_{a=1}^{T - p_0 - 2r - q_0 - 1} \big(T - p_0 - 2r - q_0 - a\big).$$

Let $D = T - p_0 - 2r - q_0$. Then the sum is

$$\sum_{a=1}^{D-1} (D - a) = \sum_{k=1}^{D-1} k = \frac{D(D-1)}{2} = \frac{(T - p_0 - 2r - q_0)(T - p_0 - 2r - q_0 - 1)}{2}.$$

So the number of different row/column contiguous minors yields

$$N_{\text{contig}}(T, p_0, q_0, r) = \frac{(T - p_0 - 2r - q_0)(T - p_0 - 2r - q_0 - 1)}{2},$$

whenever $T - p_0 - 2r - q_0 \geq 2$, which is automatically satisfied by Assumption A2. $\square$

B.4 Decomposing $B_{(p)}\Gamma_0$

We extend the results of Bai and Mones (2025) to models where the autoregressive order of the alternative parameterization p can differ from the true order p_0.

Define the one step lag matrix $\mathbf{J} \in \mathbb{R}^{T\times T}$ by

$$\mathbf{J}_{t,t-1} = 1 \quad (t = 2, \ldots, T), \qquad \text{all other entries } 0.$$

Then $\mathbf{J}^k$ has ones on the k-th subdiagonal. Moreover, let:

$$\mathbf{M}_{(p)} = \text{diag}(\underbrace{0, \ldots, 0}_{p}, \underbrace{1, \ldots, 1}_{T-p}).$$

Now, note that the baseline model (order p_0) can be expressed as:

$$\mathbf{\Phi}_0(\mathbf{J}) = \sum_{k=1}^{p_0} \phi_{k0}\mathbf{J}^k, \qquad \mathbf{B}_0 = \mathbf{I}_T - \mathbf{M}_{(p_0)}\mathbf{\Phi}_0(\mathbf{J}), \qquad \mathbf{\Gamma}_0 = \mathbf{B}_0^{-1}.$$

whereas the alternative model (order p) satisfies:

$$\mathbf{\Phi}(\mathbf{J}) = \sum_{k=1}^{p} \phi_k\mathbf{J}^k, \qquad \mathbf{B}_{(p)} = \mathbf{I}_T - \mathbf{M}_{(p)}\mathbf{\Phi}(\mathbf{J}).$$

Finally, let

$$\mathbf{L}_k = \mathbf{J}^k\mathbf{\Gamma}_0, \qquad k \geq 1$$

Combining these results we can see that

$$\mathbf{B}_{(p)}\,\mathbf{\Gamma}_0 = (\mathbf{B}_0 + (\mathbf{B}_{(p)} - \mathbf{B}_0))\mathbf{\Gamma}_0 = \mathbf{I}_T + (\mathbf{B}_{(p)} - \mathbf{B}_0)\mathbf{\Gamma}_0.$$

Since

$$\mathbf{B}_{(p)} - \mathbf{B}_0 = -\mathbf{M}_{(p)}\mathbf{\Phi}(\mathbf{J}) + \mathbf{M}_{(p_0)}\mathbf{\Phi}_0(\mathbf{J}),$$

we obtain

$$\mathbf{B}_{(p)}\,\mathbf{\Gamma}_0 = \mathbf{I}_T - \mathbf{M}_{(p)}\mathbf{\Phi}(\mathbf{J})\mathbf{\Gamma}_0 + \mathbf{M}_{(p_0)}\mathbf{\Phi}_0(\mathbf{J})\mathbf{\Gamma}_0.$$

Using $\mathbf{J}^k\mathbf{\Gamma}_0 = \mathbf{L}_k$, all expressions below are linear combinations of $\mathbf{L}_k$. Now, consider three different cases.

Case 1: $p = p_0$**.** Since $\mathbf{M}_{(p)} = \mathbf{M}_{(p_0)}$,

$$\mathbf{B}_{(p)} - \mathbf{B}_0 = \mathbf{M}_{(p)}(\mathbf{\Phi}_0(\mathbf{J}) - \mathbf{\Phi}(\mathbf{J})) = \mathbf{M}_{(p)} \sum_{k=1}^{p} (\phi_{k0} - \phi_k)\mathbf{J}^k.$$

Hence

$$\mathbf{B}_{(p_0)}\, \mathbf{\Gamma}_0 = \mathbf{I}_T + \mathbf{M}_{(p_0)} \sum_{k=1}^{p_0} (\phi_{k0} - \phi_k)\mathbf{L}_k.$$

Case 2: $p < p_0$

Define

$$\mathbf{Q}_{(p,p_0)} = \text{diag}(\underbrace{0, \ldots, 0}_{p}, \underbrace{1, \ldots, 1}_{p_0 - p}, \underbrace{0, \ldots, 0}_{T - p_0}),$$

so that

$$\mathbf{M}_{(p_0)} = \mathbf{M}_{(p)} - \mathbf{Q}_{(p,p_0)}.$$

Then

$$\mathbf{B}_{(p)} - \mathbf{B}_0 = -\mathbf{M}_{(p)}\mathbf{\Phi}(\mathbf{J}) + (\mathbf{M}_{(p)} - \mathbf{Q}_{(p,p_0)})\mathbf{\Phi}_0(\mathbf{J})$$

$$= \mathbf{M}_{(p)}(\mathbf{\Phi}_0(\mathbf{J}) - \mathbf{\Phi}(\mathbf{J})) - \mathbf{Q}_{(p,p_0)}\mathbf{\Phi}_0(\mathbf{J}).$$

Expand

$$\mathbf{\Phi}_0(\mathbf{J}) - \mathbf{\Phi}(\mathbf{J}) = \sum_{k=1}^{p} (\phi_{k0} - \phi_k)\mathbf{J}^k + \sum_{k=p+1}^{p_0} \phi_{k0}\mathbf{J}^k.$$

Thus

$$\mathbf{B}_{(p)}\, \mathbf{\Gamma}_0 = \mathbf{I}_T + \mathbf{M}_{(p)} \left[\sum_{k=1}^{p} (\phi_{k0} - \phi_k)\mathbf{L}_k + \sum_{k=p+1}^{p_0} \phi_{k0}\mathbf{L}_k \right] - \mathbf{Q}_{(p,p_0)} \sum_{k=1}^{p_0} \phi_{k0}\mathbf{L}_k.$$

Case 3: $p > p_0$

Define

$$\mathbf{Q}_{(p_0,p)} = \text{diag}(\underbrace{0, \ldots, 0}_{p_0}, \underbrace{1, \ldots, 1}_{p - p_0}, \underbrace{0, \ldots, 0}_{T - p}),$$

so that

$$\mathbf{M}_{(p)} = \mathbf{M}_{(p_0)} - \mathbf{Q}_{(p_0,p)}.$$

Then

$$\mathbf{B}_{(p)} - \mathbf{B}_0 = -(\mathbf{M}_{(p_0)} - \mathbf{Q}_{(p_0,p)})\mathbf{\Phi}(\mathbf{J}) + \mathbf{M}_{(p_0)}\mathbf{\Phi}_0(\mathbf{J})$$

$$= \mathbf{M}_{(p_0)}(\mathbf{\Phi}_0(\mathbf{J}) - \mathbf{\Phi}(\mathbf{J})) + \mathbf{Q}_{(p_0,p)}\mathbf{\Phi}(\mathbf{J}).$$

Expand

$$\mathbf{\Phi}_0(\mathbf{J}) - \mathbf{\Phi}(\mathbf{J}) = \sum_{k=1}^{p_0}(\phi_{k0} - \phi_k)\mathbf{J}^k - \sum_{k=p_0+1}^{p} \phi_k \mathbf{J}^k.$$

Therefore

$$\mathbf{B}_{(p)}\,\mathbf{\Gamma}_0 = \mathbf{I}_T + \mathbf{M}_{(p_0)}\left[\sum_{k=1}^{p_0}(\phi_{k0} - \phi_k)\mathbf{L}_k - \sum_{k=p_0+1}^{p} \phi_k \mathbf{L}_k\right] + \mathbf{Q}_{(p_0,p)}\sum_{k=1}^{p}\phi_k \mathbf{L}_k.$$

References

Acemoglu, D., Johnson, S., Robinson, J. A., and Yared, P. (2008). Income and democracy. *American economic review*, 98(3):808–842.

Acemoglu, D., Naidu, S., Restrepo, P., and Robinson, J. A. (2019). Democracy does cause growth. *Journal of political economy*, 127(1):47–100.

Ahn, S. C., Lee, Y. H., and Schmidt, P. (2013). Panel data models with multiple time-varying individual effects. *Journal of econometrics*, 174(1):1–14.

Anderson, T. and Rubin, H. (1956). Statistical inference in factor analysis. In *Proceedings of the Berkeley Symposium on Mathematical Statistics and Probability*, pages 111–149. University of California Press.

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

Arellano, M. and Bover, O. (1995). Another look at the instrumental variable estimation of error-components models. *Journal of econometrics*, 68(1):29–51.

Bai, J. (2009). Panel data models with interactive fixed effects. *Econometrica*, 77(4):1229–1279.

Bai, J. (2013). Fixed-effects dynamic panel models, a factor analytical method. *Econometrica*, 81(1):285–314.

Bai, J. (2024). Likelihood approach to dynamic panel models with interactive effects. *Journal of Econometrics*, 240(1):105636.

Bai, J. and Mones, P. (2025). Global identification of dynamic panel models with interactive effects. *arXiv preprint arXiv:2504.14354*.

Bai, J. and Ng, S. (2002). Determining the number of factors in approximate factor models. *Econometrica*, 70(1):191–221.

Baltagi, B. H. and Li, Q. (1994). Estimating error component models with general ma (q) disturbances. *Econometric Theory*, 10(2):396–408.

Besley, T., Dann, C., and Dray, S. (2026). Growth experiences and trust in government. *The Quarterly Journal of Economics*, page qjaf056.

Bloom, N., Bond, S., and Van Reenen, J. (2007). Uncertainty and investment dynamics. *The review of economic studies*, 74(2):391–415.

Blundell, R. and Bond, S. (1998). Initial conditions and moment restrictions in dynamic panel data models. *Journal of econometrics*, 87(1):115–143.

Carlsson, M., Messina, J., and Nordström Skans, O. (2021). Firm-level shocks and labour flows. *The Economic Journal*, 131(634):598–623.

Chen, W.-D. (2006). An approximate likelihood function for panel data with a mixed arma (p, q) remainder disturbance model. *Journal of Time series analysis*, 27(6):911–921.

Gobillon, L. and Magnac, T. (2016). Regional policy evaluation: Interactive fixed effects and synthetic controls. *Review of Economics and Statistics*, 98(3):535–551.

Gobillon, L., Magnac, T., and Roux, S. (2025). Life-cycle wages and human capital investments: Selection and missing data. *Review of Economic Studies*, page rdaf099.

Gorbachev, O. (2011). Did household consumption become more volatile? *American Economic Review*, 101(5):2248–2270.

Hamilton, J. D. (2020). *Time series analysis*. Princeton university press.

Hansen, L. P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica*, pages 1029–1054.

Hayakawa, K., Pesaran, M. H., and Smith, L. V. (2023). Short t dynamic panel data models with individual, time and interactive effects. *Journal of Applied Econometrics*, 38(6):940–967.

Hong, S., Su, L., and Jiang, T. (2023). Profile gmm estimation of panel data models with interactive fixed effects. *Journal of Econometrics*, 235(2):927–948.

Jardim, E., Long, M. C., Plotnick, R., Vigdor, J., and Wiles, E. (2024). Local minimum wage laws, boundary discontinuity methods, and policy spillovers. *Journal of Public Economics*, 234:105131.

Leduc, S. and Liu, Z. (2024). Automation, bargaining power, and labor market fluctuations. *American Economic Journal: Macroeconomics*, 16(4):311–349.

Lu, T.-T. and Shiou, S.-H. (2002). Inverses of 2× 2 block matrices. *Computers & Mathematics with Applications*, 43(1-2):119–129.

Moffitt, R. A. and Gottschalk, P. (2011). Trends in the covariance structure of earnings in the us: 1969–1987. *The Journal of Economic Inequality*, 9(3):439–459.

Moon, H. R. and Weidner, M. (2015). Linear regression for panel with unknown number of factors as interactive fixed effects. *Econometrica*, 83(4):1543–1579.

Newey, W. K. and McFadden, D. (1994). Large sample estimation and hypothesis testing. *Handbook of econometrics*, 4:2111–2245.

Nickell, S. (1981). Biases in dynamic models with fixed effects. *Econometrica*, pages 1417–1426.

Pesaran, M. H. (2006). Estimation and inference in large heterogeneous panels with a multifactor error structure. *Econometrica*, 74(4):967–1012.

Robinson, P. M. and Velasco, C. (2020). Estimation for dynamic panel data with individual effects. *Econometric Theory*, 36(2):185–222.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica*, pages 1–25.