Shrinkage Estimation of Network Spillovers with Factor Structured Errors*

Ayden Higgins† Federico Martellosio‡
University of Surrey, UK University of Surrey, UK

December 4, 2020

Abstract

This paper explores the estimation of a panel data model with cross-sectional interaction that is flexible both in its approach to specifying the network of connections between cross-sectional units, and in controlling for unobserved heterogeneity. It is assumed that there are different sources of information available on a network, which can be represented in the form of multiple weights matrices. These matrices may reflect observed links, different measures of connectivity, groupings or other network structures, and the number of matrices may be increasing with sample size. A penalised quasi-maximum likelihood estimator is proposed which aims to alleviate the risk of network misspecification by shrinking the coefficients of irrelevant weights matrices to exactly zero. Moreover, controlling for unobserved factors in estimation provides a safeguard against the misspecification that might arise from unobserved heterogeneity. The estimator is shown to be consistent and selection consistent as both n and T tend to infinity, and its limiting distribution is characterised. Finite sample performance is assessed by means of a Monte Carlo simulation and the method is applied to study the prevalence of network spillovers in determining growth rates across countries.

Keywords: interactive fixed effects, high-dimensional estimation, panel models, penalised quasi-likelihood, social network models.
JEL classification: C13, C23, C51.

*We are grateful to three anonymous referees and to the Associate Editor for insightful comments that helped us improve the paper. We would also like to thank Xun Lu and Liangjun Su for sharing their data with us, as well as Valentina Corradi, João Santos Silva and Sorawoot Srisuma for their comments and suggestions.
†Email: a.higgins@surrey.ac.uk
‡Email: f.martellosio@surrey.ac.uk

1 Introduction

Increased attention is being given over to panel data models which take into account cross-sectional interaction. These models have proven to be empirically relevant in a diverse range of economic settings such as social interactions between individuals, business connections between firms, trading relations between nations, and dependencies between financial assets. At the heart of many econometric models of this kind lies a weights matrix, which summarises the network of connections between interacting cross-sectional units. Yet networks are rarely fully observed, and the uncertainty in how a weights matrix should be specified has been a common critique of this growing literature (see, e.g., Blume et al., 2015; Lewbel et al., 2019; de Paula et al., 2020). In practice, situations in which networks are partially observed are more frequent, with some information being available on cross-sectional links or their absence, as well as information on other network structures such as groupings. As an example, within a school one might observe family, friendship, classroom and cohort groupings, each of which provide information on the network of connections between different students. In other settings, such as international networks, there are multiple ways to quantify connectivity between nations, including economic measures such as trade and foreign direct investment flows, physical distance, and infrastructure links. Nevertheless, it is not usually obvious how these pieces of the jigsaw fit together, and this uncertainty inevitability increases the risk of model misspecification.

Typical methods to inform the choice of weights matrix include sequential specification testing, or model selection with reference to an information criterion (e.g., Zhang and Yu, 2018). These approaches have largely been focused on the problem of discerning a single best weights matrix from a set of mutually exclusive competitors. In contrast, there are many cases in which weights matrices manifest equally relevant, rather than competing, specifications and, in cases such as these, a model that includes multiple weights matrices may be preferable. This presents a more challenging model selection problem since prospective model specifications may be nested in one another, generating a large number of alternative models. In order to tackle this empirically important issue the current paper uses penalised estimation methods, which retain relevant weights matrix specifications, while at the same time shrinking the coefficients of irrelevant matrices to exactly zero.

A related concern in models of this kind is unobserved heterogeneity. Intuitively, there are likely to be many common factors which are unobserved, and yet have an influence on the outcomes of cross-sectional units; for example exposure to common shocks or a common environment. The presence of common factors can make the identification of model parameters difficult in the event that these are correlated with covariates. The typical approach in dealing with unobserved heterogeneity is to transform the model in a way which purges the unobserved factors (see, e.g., Yu et al., 2008; Lee and Yu, 2010). Nonetheless, a transform risks purging the very variation needed to identify network spillovers and therefore identification remains a delicate issue, with variation in the regressors, and the structure of the weights matrices and of the unobserved heterogeneity each having a part to play. An additional challenge in transforming the model is that prior knowledge on the nature of the unobserved heterogeneity is needed to specify a transform. Traditional examples of this include time, unit and group effect models in which case information on time, unit and group identities is used. Yet with a complex structure of cross-sectional interaction, it is desirable to go beyond these models and to allow for more general forms of heterogeneity. The present paper models a factor structure in the error, which provides this flexibility since common factors may vary across time and have a fully heterogeneous effect on the cross-section. By way of principal component methods, a transform to purge these factors is, in effect, estimated alongside model parameters, removing the reliance on prior knowledge to specify a transform. Taken together, multiple weights matrices, penalisation, and a factor structure error, provide a means of estimating various network spillovers which addresses some of the empirical concerns raised in models of cross-sectional interaction.

The present paper lies in the intersection of several literatures, including social and spatial network models, high-dimensional estimation, and models with factor structured errors. In the social network literature, estimation and identification of network spillovers has been extensively discussed; e.g., Lee (2007), Bramoullé et al. (2009) and Lee et al. (2010). These papers each devote attention to the challenges which may arise in the presence of unobserved heterogeneity, in models where a time dimension is absent. Elsewhere, panel data models which combine interaction and factor structures in the error term have been considered; see, for example, Shi and Lee (2017), Bai and Li (2018) and Kuersteiner and Prucha (2020). In a likelihood framework, Shi and Lee (2017) studies the estimation of a dynamic spatial model with interactive fixed effects and a single weights matrix. Bai and Li (2018) do similarly, though explicitly allowing for cross-sectional heteroskedasticity. The present paper also pursues likelihood based estimation, and generalises these papers to allow for multiple weights matrices and the possibility that the number may be increasing with sample size. Kuersteiner and Prucha (2020) consider estimation of a model with multiple potentially endogenous weights matrices alongside a factor structure in the error, by way of a method of moments estimator. The approach of Shi and Lee (2017) is partly inspired by Moon and Weidner (2015), who derive the properties of an estimator using an eigenvalue perturbation approach. On the other hand, Bai and Li (2018) more closely follow Bai (2009), who derives results using first order conditions as a starting point for analysis. In terms of theory, this paper follows the latter approach, and proceeds from first order conditions in similar fashion to Bai (2009).

In the high-dimensional estimation literature, Lu and Su (2016) examine a model with interactive fixed effects and an increasing number of parameters, but without cross-sectional interaction. They make use of the adaptive Lasso penalty of Zou (2006) to induce sparsity amongst both estimated coefficients and factor loadings, assuming that many of these are redundant. Their procedure yields efficiency gains when compared to estimating the model with the number of factors overestimated. The present paper also uses the adaptive Lasso, which penalises the ℓ_1 norm of the estimated parameter vector, encouraging sparsity amongst coefficient estimates. High-dimensional spatial models have also been studied elsewhere, such as Lam and Souza (2019), who consider a model which allows for an increasing number of spatial weights matrices, and also use the adaptive Lasso as a penalty, though do not consider unobserved heterogeneity beyond standard fixed effect approaches. Liu (2017) similarly uses penalised estimation in a cross-sectional model with many spatial weights matrices. Gupta and Robinson (2015, 2018) consider estimation of a cross-sectional spatial model, with the number of weights matrices increasing with sample size, by using instrumental variables and quasi-maximum likelihood respectively. The authors carefully study the asymptotic behaviour of these estimators, but do not pursue penalised estimation nor discuss unobserved heterogeneity.

Some recent works have also considered the case where the network is entirely unobserved, such as Lewbel et al. (2019) and de Paula et al. (2020). This situation is especially relevant in the context of social interactions, where connections between individuals might be particularly to observe or to quantify. The approach taken in de Paula et al. (2020) involves estimating an entire weights matrix using observations of the same set of individuals across multiple time periods. This can be seen as an extreme case of the current paper in which each weights matrix consists of a single nonzero element taking a value of one. Lewbel et al. (2019) takes a different perspective whereby multiple groups of individuals are observed, a special case of which is when each group consists of the same individuals observed in different time periods. In contrast, the focus of the present paper is on the case when the network is partially observed, which in practice may be quite common. Moreover, establishing identification of the entire weights matrix once a factor structure is introduced in the error may be a nontrivial matter.

Outline of the paper: The model of interest is introduced in Section 2, alongside some basic assumptions and the estimation method pursued. This is followed by asymptotic results in Section 3, and discussion of implementation in Section 4. In Section 5 finite sample performance is assessed be means of a small Monte Carlo study, followed by an empirical application of the method to consider whether network spillovers are prevalent in determining growth rates across countries. Section 6 concludes. Proofs of the main results can be found in Appendix A. For further discussion, proofs of lemmas and additional simulation output, see the Supplementary Material.

Notation: Throughout the paper, all vectors and matrices are real. For an $n \times 1$ vector $\boldsymbol{b}$ with elements b_i, $||\boldsymbol{b}||_1 := \sum_{i=1}^n |b_i|$, $||\boldsymbol{b}||_2 := \sqrt{\sum_{i=1}^n b_i^2}$, $||\boldsymbol{b}||_\infty := \max_{1\leq i\leq n} |b_i|$. Let $\boldsymbol{B}$ be an $n \times m$ matrix with elements B_{ij}. When $m = n$, and the eigenvalues of $\boldsymbol{B}$ are real, they are denoted by $\mu_n(\boldsymbol{B}) \leq ... \leq \mu_1(\boldsymbol{B})$. The following matrix norms are those induced by their vector counterparts: $||\boldsymbol{B}||_1 := \max_{1\leq j\leq m} \sum_{i=1}^n |B_{ij}|$ which is the maximum absolute column sum of $\boldsymbol{B}$, $||\boldsymbol{B}||_2 := \sqrt{\mu_1(\boldsymbol{B}'\boldsymbol{B})}$, and $||\boldsymbol{B}||_\infty := \max_{1\leq i\leq n} \sum_{j=1}^m |B_{ij}|$ which is the maximum absolute row sum of $\boldsymbol{B}$. The Frobenius norm of $\boldsymbol{B}$ is denoted $||\boldsymbol{B}||_F := \sqrt{\sum_{i=1}^n \sum_{j=1}^m B_{ij}^2} = \sqrt{\mathrm{tr}(\boldsymbol{B}'\boldsymbol{B})}$. Let $\boldsymbol{P}_B := \boldsymbol{B}(\boldsymbol{B}'\boldsymbol{B})^+\boldsymbol{B}'$ and $\boldsymbol{M}_B := \boldsymbol{I}_n - \boldsymbol{P}_B$, where $\boldsymbol{I}_m$ is the $m \times m$ identity matrix and the superscript $+$ denotes the Moore-Penrose generalised inverse. A sequence of $n \times n$ matrices $\boldsymbol{C}_n$ is said to be uniformly bounded in absolute row and column sums (UB) if both the sequences $||\boldsymbol{C}_n||_1$ and $||\boldsymbol{C}_n||_\infty$ are bounded. Throughout c, potentially indexed as c_i when there are many, is used to denote some arbitrary positive constant.

2 Model and Estimation

2.1 Model

The model considered in this paper supposes that, amongst n cross-sectional units in time period $t = 1, ..., T$, outcomes are generated according to

$$\boldsymbol{y}_t = \sum_{q=1}^{Q_{nT}} \rho_q \boldsymbol{W}_q \boldsymbol{y}_t + \sum_{k=1}^{K_{nT}} \beta_k \boldsymbol{x}_{kt} + \boldsymbol{\eta}_t, \tag{1}$$

where $\boldsymbol{y}_t, \boldsymbol{x}_{kt}$ and $\boldsymbol{\eta}_t$ are $n \times 1$ vectors of outcomes, covariates and error terms respectively, and $\boldsymbol{W}_q$ is an $n \times n$ weights matrix specified in advance. Both the number Q_{nT} of potentially relevant weights matrices and the number K_{nT} of potentially relevant regressors can increase with sample size. The covariates may be subdivided into various types, such that

$$\sum_{k=1}^{K_{nT}} \beta_k \boldsymbol{x}_{kt} = \sum_{\kappa=1}^{K_{nT}^*} \delta_\kappa \boldsymbol{x}_{\kappa t}^* + \phi_1 \boldsymbol{y}_{t-1} + \sum_{q=1}^{Q_{nT}} \phi_{q+1} \boldsymbol{W}_q \boldsymbol{y}_{t-1}. \tag{2}$$

The first K^*_{nT} regressors may be either primitive exogenous covariates, or formed by the interaction of weights matrices and primitive exogenous covariates. Moreover, lagged outcomes and the interaction of these with weights matrices can provide additional covariates of the form $\boldsymbol{W}_q\boldsymbol{y}_{t-1}$. It may be that many of the parameters ρ_q, δ_κ and ϕ_q are truly zero since many of the covariates or weights matrix specifications may be irrelevant. Such restrictions need not be imposed a priori, since penalised estimation induces the estimates of these parameters to take values of exactly zero.

The weights matrices $\boldsymbol{W}_q$ contain information about the connections between the cross-sectional units, with larger elements – positive or negative – measuring a stronger connection strength. The literature often assumes that the weights matrices have positive elements and are row normalised such that each of the rows of $\boldsymbol{W}_q$ sum to 1. These assumptions lend products of the form $\boldsymbol{W}_q\boldsymbol{b}$ the interpretation of a weighted average of the entries of a vector $\boldsymbol{b}$. While these two assumptions are not necessary in this paper, the assumption that the weights matrices have zero diagonals, which forbids self-links, is retained. The coefficients ρ_q on $\boldsymbol{W}_q\boldsymbol{y}_t$ capture endogenous spillovers; that is, the impact on the outcome of each unit, generated by the units that are neighbours according to the q-th weights matrix. Analogously, those δ_κ coefficients on covariates of the form $\boldsymbol{W}_q\boldsymbol{x}^*_{\kappa t}$ capture exogenous spillovers, also referred to as contextual effects in the social interaction literature. The coefficients ϕ_q on products $\boldsymbol{W}_q\boldsymbol{y}_{t-1}$ capture dynamic spillovers. Combined, the endogenous, exogenous and dynamic spillovers, allow model (1) to quantify a breadth of different network spillovers.

It is assumed that the error term has a factor structure of the form

$$\boldsymbol{\eta}_t = \boldsymbol{\Lambda}\boldsymbol{f}_t + \boldsymbol{\varepsilon}_t, \tag{3}$$

where $\boldsymbol{\Lambda}$ is an $n \times R$ matrix of time-invariant loadings, $\boldsymbol{f}_t$ is an $R \times 1$ vector of unit-invariant factors, and $\boldsymbol{\varepsilon}_t$ is an $n \times 1$ vector of idiosyncratic error terms. In addition, the rows of $\boldsymbol{\Lambda}$ are denoted by $\boldsymbol{\lambda}_i$, for $i = 1, ..., n$, and the factors are arranged in the $T \times R$ matrix $\boldsymbol{F} := (\boldsymbol{f}_1, ..., \boldsymbol{f}_T)'$. Following a fixed effects approach, both factors and loadings are treated as (nuisance) parameters in estimation. Thus, in the model, either is allowed to be arbitrarily correlated with covariates. The framework is very general; for instance $\boldsymbol{f}_t$ could be aggregate shocks affecting the entire network at time t, with a heterogenous effect on each individual. Moreover, this factor structure nests more traditional fixed effect models as special cases.

It is worth stressing that unobserved heterogeneity may arise from various sources. Consider, as a simple example, a model with a single exogenous regressor and no endogenous spillovers, i.e.,

$$\boldsymbol{y}_t = \beta^* \boldsymbol{x}_t^* + \sum_{q=1}^{Q_{nT}} \alpha_q \boldsymbol{W}_q \boldsymbol{x}_t^* + \alpha_{q+1} \boldsymbol{W}_L \boldsymbol{x}_t^* + \boldsymbol{\varepsilon}_t, \tag{4}$$

with $\boldsymbol{W}_q$ being the q-th observed weights matrix, and $\beta^*, \delta, \alpha_q$ being scalars. Suppose that $\boldsymbol{W}_L$ is either low rank or well approximated by a low rank matrix and represents, for example, low rank measurement error in some $\boldsymbol{W}_q$, or unobserved connections between cross-sectional units arising due to network sampling; see, for example, Wang (2018). Defining $\boldsymbol{\Lambda}^* \boldsymbol{f}_t^* = \alpha_{q+1} \boldsymbol{W}_L \boldsymbol{x}_t^*$, it is clear that (4) is nested in model (1) and highlights that the decomposition of the unobserved terms into factors $\boldsymbol{\Lambda}^*$ and loadings $\boldsymbol{f}_t^*$ is arbitrary; it is the low rank restriction on $\alpha_{q+1} \boldsymbol{W}_L \boldsymbol{x}_t^*$ that allows this term to be distinguished and controlled for.

Going forward, it is convenient to introduce some new notation. The subscript nT used previously is suppressed from Q_{nT}, K_{nT}, K_{nT}^*, and the following parameter vectors and covariate matrices are defined: $\boldsymbol{\rho} \coloneqq (\rho_1, ..., \rho_Q)'$, $\boldsymbol{\delta} \coloneqq (\delta_1, ..., \delta_{K^*})'$, $\boldsymbol{\phi} \coloneqq (\phi_1, ..., \phi_{Q+1})'$, $\boldsymbol{\beta} \coloneqq (\beta_1, ..., \beta_K)' \coloneqq (\boldsymbol{\delta}', \boldsymbol{\phi}')'$, $\boldsymbol{\theta} \coloneqq (\boldsymbol{\rho}', \boldsymbol{\beta}')'$, and $\boldsymbol{X}_t \coloneqq (\boldsymbol{X}_t^*, \boldsymbol{y}_{t-1}, \boldsymbol{W}_1 \boldsymbol{y}_{t-1}, ..., \boldsymbol{W}_{\mathrm{Q}} \boldsymbol{y}_{t-1})$, where $\boldsymbol{X}_t^* \coloneqq (\boldsymbol{x}_{1t}^*, ..., \boldsymbol{x}_{K^*t}^*)$, and $\boldsymbol{S}(\boldsymbol{\rho}) \coloneqq \boldsymbol{I}_n - \sum_{q=1}^{Q} \rho_q \boldsymbol{W}_q$. Given these, model (1) can be restated more succinctly as

$$\boldsymbol{S}(\boldsymbol{\rho}) \boldsymbol{y}_t = \boldsymbol{X}_t \boldsymbol{\beta} + \boldsymbol{\Lambda} \boldsymbol{f}_t + \boldsymbol{\varepsilon}_t. \tag{5}$$

Throughout, the superscript 0 is used to distinguish the true values of the factors, loadings, and parameters, as well as the true numbers of these, and $n, T \to \infty$ is used to indicate that n and T diverge simultaneously. The total number of parameters in the vector $\boldsymbol{\theta}$ is $P \coloneqq Q + K$, of which only P^0 are truly nonzero. In fact, one might often expect that the vector $\boldsymbol{\theta}$ is sparse in the sense that many of its components are zero, particularly in cases with a large number of weights matrices and covariates. Accordingly $\boldsymbol{\theta}$ may be reordered as $\boldsymbol{\vartheta} = (\boldsymbol{\theta}_{(1)}', \boldsymbol{\theta}_{(2)}')'$, where $\boldsymbol{\theta}_{(1)}$ is the $P^0 \times 1$ vector of nonzero parameters, and $\boldsymbol{\theta}_{(2)}^0 = \mathbf{0}_{(P-P^0)\times 1}$. Sparsity, however, is not necessary and indeed the results of this paper equally allow for the possibility that all of the weights matrices and covariates may be relevant. The $n \times T$ data matrix for the κ-th exogenous covariate is denoted $\boldsymbol{\mathcal{X}}_\kappa^* \coloneqq (\boldsymbol{x}_{\kappa 1}^*, ..., \boldsymbol{x}_{\kappa T}^*)$ for $\kappa = 1, ..., K^*$, and the $n \times T$ data matrix for the lagged outcomes is denoted $\boldsymbol{\mathcal{Y}}_{-1} \coloneqq (\boldsymbol{y}_{t-1}, ..., \boldsymbol{y}_{t-T})$. The data matrix for the generic k-th covariate of any type, $\boldsymbol{\mathcal{X}}_\kappa^*$, $\boldsymbol{\mathcal{Y}}_{-1}$ or $\boldsymbol{W}_q \boldsymbol{\mathcal{Y}}_{-1}$, is denoted $\boldsymbol{\mathcal{X}}_k \coloneqq (\boldsymbol{x}_{k1}, ..., \boldsymbol{x}_{kT})$, for $k = 1, ..., K$. Also, $\boldsymbol{A}(\boldsymbol{\rho}, \boldsymbol{\phi}) \coloneqq \boldsymbol{S}^{-1}(\boldsymbol{\rho})(\phi_1 \boldsymbol{I}_n + \sum_{q=1}^{Q} \phi_{q+1} \boldsymbol{W}_q)$, $\boldsymbol{A} \coloneqq \boldsymbol{A}(\boldsymbol{\rho}^0, \boldsymbol{\phi}^0)$ and $\boldsymbol{S} \coloneqq \boldsymbol{S}(\boldsymbol{\rho}^0)$.

2.2 Assumptions

The first set of assumptions concerns the idiosyncratic error term ε_{it}.

Assumption 1.

1.1 The errors ε_{it} are identically and independently distributed over i and t with $\mathbb{E}[\varepsilon_{it}] = 0$, $\mathbb{E}[\varepsilon_{it}^2] = \sigma_0^2$, with fourth moments uniformly bounded over i and t.

1.2 The errors ε_{it} are independent of the elements of the matrices $\boldsymbol{\Lambda}^0$, $\boldsymbol{F}^0$, and $\boldsymbol{\mathcal{X}}_\kappa^*$, for $\kappa = 1, ..., K^*$.

These assumptions have been employed across various papers. Cross-sectional homoskedasticity and independence is commonly assumed, though this can be relaxed by estimation of a more general $n \times n$ covariance matrix $\boldsymbol{\Sigma}^0$, at the expense of additional parameters; see for example Bai and Liao (2017) and Bai and Li (2018). Additional structure in the error term could also be considered as is commonplace throughout the spatial econometrics literature. Yet since the factor structure provides a mechanism for capturing such correlation, Assumption 1.1 assumes $\boldsymbol{\Sigma}^0 = \sigma_0^2 \boldsymbol{I}_n$. Differing assumptions concerning the relationship between the errors, the factors, and the loadings appear across the literature; these are comprehensively surveyed by Hsiao (2018). Assumption 1.2 imposes independence of the factors and the loadings from the error term as in Bai (2009).

Some additional assumptions are required regarding the other components of the model. Let $|\boldsymbol{B}|$ denote the entrywise absolute value of a matrix, and let the parameter space for $\boldsymbol{\theta}$ be denoted $\boldsymbol{\Theta}$, and the parameter spaces for $\boldsymbol{\rho}$ and $\boldsymbol{\phi}$ be denoted $\boldsymbol{\Theta}_\rho$ and $\boldsymbol{\Theta}_\phi$ respectively. Since the matrices $\boldsymbol{W}_1, ..., \boldsymbol{W}_{\mathrm{Q}}$ depend on n, and the number of weights matrices Q, the matrices $\boldsymbol{S}(\boldsymbol{\rho})$, $\boldsymbol{A}(\boldsymbol{\rho}, \boldsymbol{\phi})$, as well as the parameter spaces, depend on n and T, it is understood that the following set of assumptions hold for any n, T.

Assumption 2.

2.1 The parameter vector $\boldsymbol{\theta}^0$ is in the interior of $\boldsymbol{\Theta}$, with $\boldsymbol{\Theta}$ being a compact subset of $\mathbb{R}^P$.

2.2 The weights matrices $\boldsymbol{W}_1, ..., \boldsymbol{W}_Q$ are nonstochastic and UB *uniformly over q.*

2.3 For all $\boldsymbol{\rho} \in \boldsymbol{\Theta}_\rho$ and $\boldsymbol{\phi} \in \boldsymbol{\Theta}_\phi$, $\boldsymbol{S}(\boldsymbol{\rho})$ is invertible, $\boldsymbol{S}(\boldsymbol{\rho}), \boldsymbol{S}^{-1}(\boldsymbol{\rho})$ and $\sum_{h=1}^{\infty} |\boldsymbol{A}^h(\boldsymbol{\rho}, \boldsymbol{\phi})|$ are UB*, and $||\boldsymbol{A}(\boldsymbol{\rho}, \boldsymbol{\phi})||_2 < 1$. In addition, $\lim_{n,T\to\infty} \sup_{\boldsymbol{\rho}\in\boldsymbol{\Theta}_\rho, \boldsymbol{\phi}\in\boldsymbol{\Theta}_\phi} ||\boldsymbol{A}(\boldsymbol{\rho}, \boldsymbol{\phi})||_2 < 1$ and $\lim_{n,T\to\infty} \inf_{\boldsymbol{\rho}\in\boldsymbol{\Theta}_\rho} \det(\boldsymbol{S}(\boldsymbol{\rho})) \neq 0$. .*

2.4 The elements of the matrices $\boldsymbol{\mathcal{X}}_\kappa^*$ *have fourth moments uniformly bounded over* i, t *and* κ.

2.5 $||\boldsymbol{\beta}^0||_1 < c$.

2.6 The true number of factors R^0 *is constant.*

2.7 The elements of the matrices $\boldsymbol{F}^0$ *and* $\boldsymbol{\Lambda}^0$ *have eighth moments uniformly bounded over* i *and* t.

Assumption 2.1 considers a sequence of compact parameter spaces over which to maximise the objective function. The condition in Assumption 2.2 that the weights matrices are UB is standard and serves to limit interactions to a manageable degree. Here, uniform boundedness over q is also required, due to the possibility that Q increases with sample size. Assumption 2.3 ensures that the model admits a reduced form, and the dynamic process in stationary. Restrictions on the parameter space of $\boldsymbol{\rho}$ which ensure that $\boldsymbol{S}(\boldsymbol{\rho})$ is invertible have been discussed elsewhere in the literature, particularly in the case $Q = 1$. A general condition sufficient for the invertibility of $\boldsymbol{S}(\boldsymbol{\rho})$ is $||\sum_{q=1}^{Q} \rho_q \boldsymbol{W}_q|| < 1$ for some matrix norm $||\cdot||$, though with $Q > 1$ more informative conditions can be difficult to obtain outside of exceptional cases.[1] However, as noted by Gupta and Robinson (2018), even when it is possible to characterise inadmissible values of $\boldsymbol{\rho}$ and exclude these, the resulting parameter space is unlikely to be compact. It is therefore commonplace in the literature to restrict attention to a region around the origin in which $\boldsymbol{S}(\boldsymbol{\rho})$ can be guaranteed to be invertible. This is where $\sum_{q=1}^{Q} |\rho_q| < (\max_{1\leq q\leq Q} ||\boldsymbol{W}_q||)^{-1}$.[2] Yet while the set of $\boldsymbol{\rho}$ which satisfy this is bounded, it is also open. Therefore to ensure the existence of a maximiser over this space, a closed subset can be considered such that $\sum_{q=1}^{Q} |\rho_q| \leq (1-\tau)(\max_{1\leq q\leq Q} ||\boldsymbol{W}_q||)^{-1}$, with $\tau \in (0, 1)$. Row normalisation of the matrices $\boldsymbol{W}_q$ further simplifies this condition since it implies $\max_{1\leq q\leq Q} ||\boldsymbol{W}_q||_\infty = 1$. Model (5) can be rewritten as $\boldsymbol{y}_t = \boldsymbol{A}\boldsymbol{y}_{t-1} + \boldsymbol{S}^{-1}(\boldsymbol{X}_t^*\boldsymbol{\delta} + \boldsymbol{\Lambda}\boldsymbol{f}_t + \boldsymbol{\varepsilon}_t)$, or, after recursive substitution, $\boldsymbol{y}_t = \sum_{h=0}^{\infty} \boldsymbol{A}^h \boldsymbol{S}^{-1}(\boldsymbol{X}_{t-h}^*\boldsymbol{\delta} + \boldsymbol{\Lambda}\boldsymbol{f}_{t-h} + \boldsymbol{\varepsilon}_{t-h})$; Assumption 2.3 guarantees that this series converges. Further discussion of parameter restrictions ensuring convergence of this series can be found in Lee and Yu (2014) and Shi and Lee

[1]One such case is when the matrices $\boldsymbol{W}_1, ..., \boldsymbol{W}_Q$ are simultaneously diagonalisable and hence have common eigenvectors. This encompasses several notable cases, such as where the weights matrices consist of nonoverlapping blocks, or are powers of a single weights matrix.

[2]This inequality is obtained from the condition $||\sum_{q=1}^{Q} \rho_q \boldsymbol{W}_q|| < 1$ and the fact that $||\sum_{q=1}^{Q} \rho_q \boldsymbol{W}_q|| \leq \sum_{q=1}^{Q} |\rho_q| \max_{1\leq q\leq Q} ||\boldsymbol{W}_q||$ for any matrix norm $||\cdot||$.

(2017).[3] Assumption 2.4 ensures that $||\boldsymbol{\mathcal{X}}_k^*||_F = O_P(\sqrt{nT})$, for $k = 1, ..., K$. For Assumption 2.5, note that $\boldsymbol{W}_q\boldsymbol{S}^{-1}\boldsymbol{X}_t\boldsymbol{\beta}^0$ can be used as an instrument in the estimation of ρ_q^0.[4] With a diverging number of parameters, Assumption 2.5 assures that for these instruments $||\sum_{k=1}^{K^0}\beta_k^0\boldsymbol{W}_q\boldsymbol{S}^{-1}\boldsymbol{\mathcal{X}}_k||_F = O_P(\sqrt{nT})$, which follows by Hölder's inequality. Alternatively, this assumption could be replaced by one restricting the growth of K^0 and n, T, yet Assumption 2.5 is more convenient for theoretical analysis. Assumption 2.6 is common throughout the literature, but could be relaxed at the expense of slower rates of convergence. Several differing assumptions concerning the moments of the factors and the loadings appear in the literature. Given the possible presence of lagged outcomes as covariates, Assumption 2.7 serves the same purpose as Assumption 5(vi) in Moon and Weidner (2017), and ensures that the y_{it} has uniformly bounded fourth moments.

2.3 Objective Function

The estimation strategy employed in this paper is penalised quasi-maximum likelihood (PQML), using the multivariate standard normal distribution for the error term, i.e., $\varepsilon_{it} \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma_0^2)$, and following a fixed effects approach. Maximum likelihood estimation is a standard in the literature for models of this type, since the simultaneity in the determination of outcomes generates an endogeneity problem which results in least squares estimates being biased. The parameter of interest is $\boldsymbol{\theta}$, whereas $\boldsymbol{\Lambda}, \boldsymbol{F}, \sigma^2$ are treated as nuisance parameters. Since fixing $\boldsymbol{\theta}$ results in a pure factor model (and $\boldsymbol{\Lambda}, \boldsymbol{F}, \sigma^2$ are not penalised), the estimators of $\boldsymbol{\Lambda}$ and $\boldsymbol{F}$ for fixed $\boldsymbol{\theta}$ are a solution to a standard principal component problem (see, e.g., Bai, 2009; Shi and Lee, 2017). In this subsection R is fixed such that $R \geq R^0$; this is discussed in greater detail in Section 3.1. With R fixed, the average (quasi) log-likelihood is

$$\begin{aligned}\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\Lambda}, \boldsymbol{F}, \sigma^2) \coloneqq &-\frac{1}{2}\log(2\pi) + \frac{1}{n}\log(\det(\boldsymbol{S}(\boldsymbol{\rho}))) - \frac{1}{2}\log(\sigma^2) \\ &-\frac{1}{2\sigma^2}\frac{1}{nT}\sum_{t=1}^{T}(\boldsymbol{S}(\boldsymbol{\rho})\boldsymbol{y}_t - \boldsymbol{X}_t\boldsymbol{\beta} - \boldsymbol{\Lambda}\boldsymbol{f}_t)'(\boldsymbol{S}(\boldsymbol{\rho})\boldsymbol{y}_t - \boldsymbol{X}_t\boldsymbol{\beta} - \boldsymbol{\Lambda}\boldsymbol{f}_t)\end{aligned} \tag{6}$$

and its penalised counterpart is

$$\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\Lambda}, \boldsymbol{F}, \sigma^2) \coloneqq \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\Lambda}, \boldsymbol{F}, \sigma^2) - \varrho(\boldsymbol{\theta}, \boldsymbol{\gamma}, \boldsymbol{\zeta}), \tag{7}$$

[3]For example, where the matrices $\boldsymbol{W}_1, ..., \boldsymbol{W}_Q$ are simultaneously diagonalisable, $\sum_{q=1}^{Q}|\rho_q| + \sum_{q=1}^{Q+1}|\phi_q| \leq 1$ is sufficient for $||\boldsymbol{A}(\boldsymbol{\rho}, \boldsymbol{\phi})||_2 < 1$.

[4]Observing that $\boldsymbol{S}^{-1} = \boldsymbol{I}_n + \sum_{q=1}^{Q}\rho_q\boldsymbol{W}_q\boldsymbol{S}^{-1}$, then $\boldsymbol{y}_t = \boldsymbol{X}_t\boldsymbol{\beta}^0 + \sum_{q=1}^{Q}\rho_q^0\boldsymbol{W}_q\boldsymbol{S}^{-1}\boldsymbol{X}_t\boldsymbol{\beta}^0 + \boldsymbol{S}^{-1}\boldsymbol{\Lambda}^0\boldsymbol{f}_t^0 + \boldsymbol{S}^{-1}\boldsymbol{\varepsilon}_t$, which makes the role of $\boldsymbol{W}_q\boldsymbol{S}^{-1}\boldsymbol{X}_t\boldsymbol{\beta}^0$ as an instrument for ρ_q transparent.

where $\varrho(\boldsymbol{\theta}, \boldsymbol{\gamma}, \boldsymbol{\zeta})$ is a penalty function and $\boldsymbol{\gamma}$, $\boldsymbol{\zeta}$ are regularisation parameters. The specific form of penalty function is introduced in Section 2.4, and the choice of regularisation parameters is discussed in Section 4.1, however for the moment these are both also taken to be fixed alongside the number of factors. Concentrating out σ^2, as well as the factors, and dropping the constant in (7) yields the concentrated expression

$$\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\Lambda}) = \frac{1}{n}\log(\det(\boldsymbol{S}(\boldsymbol{\rho}))) - \frac{1}{2}\log\left(\hat{\sigma}^2(\boldsymbol{\theta}, \boldsymbol{\Lambda})\right) - \varrho(\boldsymbol{\theta}, \boldsymbol{\gamma}, \boldsymbol{\zeta}), \tag{8}$$

where $\hat{\sigma}^2(\boldsymbol{\theta}, \boldsymbol{\Lambda}) := \frac{1}{nT}\sum_{t=1}^{T} \boldsymbol{e}_t' \boldsymbol{M}_{\boldsymbol{\Lambda}} \boldsymbol{e}_t$ and $\boldsymbol{e}_t := \boldsymbol{S}(\boldsymbol{\rho})\boldsymbol{y}_t - \boldsymbol{X}_t\boldsymbol{\beta}$. Hereafter the terms likelihood and log-likelihood are used synonymously. In order to maximise (8) with respect to $\boldsymbol{\Lambda}$, note that

$$\begin{aligned}
\min_{\boldsymbol{\Lambda}\in\mathbb{R}^{n\times R}} \frac{1}{nT}\sum_{t=1}^{T} \boldsymbol{e}_t'\boldsymbol{M}_{\boldsymbol{\Lambda}}\boldsymbol{e}_t &= \frac{1}{nT}\sum_{t=1}^{T} \boldsymbol{e}_t'\boldsymbol{e}_t - \max_{\boldsymbol{\Lambda}\in\mathbb{R}^{n\times R}} \frac{1}{nT}\sum_{t=1}^{T} \boldsymbol{e}_t'\boldsymbol{P}_{\boldsymbol{\Lambda}}\boldsymbol{e}_t \\
&= \operatorname{tr}\left(\frac{1}{nT}\sum_{t=1}^{T}\boldsymbol{e}_t\boldsymbol{e}_t'\right) - \max_{\boldsymbol{\mathcal{V}}_{\boldsymbol{\Lambda}}\in\mathbb{R}^{n\times R}:\boldsymbol{\mathcal{V}}_{\boldsymbol{\Lambda}}'\boldsymbol{\mathcal{V}}_{\boldsymbol{\Lambda}}=\boldsymbol{I}_R} \operatorname{tr}\left(\frac{1}{nT}\sum_{t=1}^{T}\boldsymbol{\mathcal{V}}_{\boldsymbol{\Lambda}}'\boldsymbol{e}_t\boldsymbol{e}_t'\boldsymbol{\mathcal{V}}_{\boldsymbol{\Lambda}}\right) \\
&= \operatorname{tr}\left(\frac{1}{nT}\sum_{t=1}^{T}\boldsymbol{e}_t\boldsymbol{e}_t'\right) - \sum_{r=1}^{R}\mu_r\left(\frac{1}{nT}\sum_{t=1}^{T}\boldsymbol{e}_t\boldsymbol{e}_t'\right),
\end{aligned} \tag{9}$$

where the second line follows from the fact that any orthogonal projector $\boldsymbol{P}_{\boldsymbol{B}}$ can be written as $\boldsymbol{\mathcal{V}}_{\boldsymbol{B}}\boldsymbol{\mathcal{V}}_{\boldsymbol{B}}'$, with the columns of $\boldsymbol{\mathcal{V}}_{\boldsymbol{B}}$ forming an orthonormal basis for the column space of $\boldsymbol{B}$,[5] and the third line follows from a standard result (e.g., Horn and Johnson, 2012, Corollary 4.3.39). Hence, (9) can be used to concentrate out $\boldsymbol{\Lambda}$ in (8), whereby the PQML estimator of $\boldsymbol{\theta}^0$ is characterised as

$$\hat{\boldsymbol{\theta}} = \operatorname*{arg\,max}_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \mathcal{Q}(\boldsymbol{\theta}), \tag{10}$$

where

$$\mathcal{Q}(\boldsymbol{\theta}) = \frac{1}{n}\log(\det(\boldsymbol{S}(\boldsymbol{\rho}))) - \frac{1}{2}\log\left(\sum_{i=R+1}^{n}\mu_i\left(\frac{1}{nT}\sum_{t=1}^{T}\boldsymbol{e}_t\boldsymbol{e}_t'\right)\right) - \varrho(\boldsymbol{\theta}, \boldsymbol{\gamma}, \boldsymbol{\zeta}). \tag{11}$$

Here it is worth highlighting that both the factors and the loadings have been concentrated out without imposing any of the normalisations typically encountered in the wider factor literature. This is due to the treatment of both the factors and the loadings as nuisance

[5] For example, by the QR decomposition, $\boldsymbol{B} = \boldsymbol{\mathcal{V}}_{\boldsymbol{B}}\boldsymbol{\mathcal{R}}$ with $\boldsymbol{\mathcal{V}}_{\boldsymbol{B}} \in \mathbb{R}^{n\times m}$ having orthonormal columns and $\boldsymbol{\mathcal{R}} \in \mathbb{R}^{m\times m}$ being upper triangular. Since $\boldsymbol{B}$ has full column rank $\boldsymbol{\mathcal{R}}$ is invertible (e.g., Horn and Johnson, 2012, Theorem 2.1.14) and therefore $\boldsymbol{P}_{\boldsymbol{B}} := \boldsymbol{B}(\boldsymbol{B}'\boldsymbol{B})^{-1}\boldsymbol{B} = \boldsymbol{\mathcal{V}}_{\boldsymbol{B}}\boldsymbol{\mathcal{V}}_{\boldsymbol{B}}'$.

parameters, in which case only the space spanned by the loadings implicitly features in the objective function (11). It would, of course, be possible to consider estimators of the factors and the loadings, however the same fundamental indeterminacy issue would arise in separating the these as is encountered elsewhere in the factor literature, and therefore some normalisations would typically be required in order to do this.

2.4 Penalty

The present paper adopts the adaptive Lasso, which induces sparsity in parameter estimation by augmenting an objective function with a constraint on the ℓ_1 norm of the estimated parameter vector. A desirable feature of this method of penalisation is that it can achieve the oracle property; that is, perform consistent variable selection and, at the same time, possess an optimal rate of convergence. This is done by using an initial consistent estimate of the parameters to weight the penalty. In this way, as sample size increases, the penalty for zero coefficients increases and yet remains constant for nonzero coefficients. The cost of this is the need for an initial consistent estimate, which can be difficult to obtain in settings where the number of parameters is greater than the number of observations ($nT < P$ in the present case). This complication is not considered in this paper and attention is restricted to the $nT > P$ setting. Explicitly, the penalty function employed in this paper has the additive form

$$\varrho(\boldsymbol{\theta}, \boldsymbol{\gamma}, \boldsymbol{\zeta}) \coloneqq \sum_{p=1}^{P} \gamma_p \frac{1}{|\theta_p^\dagger|^{\zeta_p}} |\theta_p|, \tag{12}$$

with γ_p and ζ_p being positive regularisation parameters which potentially may vary across p, and $\theta_p^\dagger$ being some initial consistent estimate of θ_p^0. The regularisation parameter γ_p serves the purpose of weighting the effect of the penalty term on the overall objective function and varies with n, T. The parameter ζ_p adjusts to the scale of the initial estimates. The following properties are assumed.

Assumption 3.

3.1 $\max_{1 \leq p \leq P} \gamma_p \sqrt{\min\{n, T\}} |\theta_p^0|^{-\zeta_p} = O(1)$.

3.2 $||\boldsymbol{\theta}^\dagger - \boldsymbol{\theta}^0||_2 = O_P(c_{nT})$ *where* $c_{nT} \to 0$.

Assumption 3.1 requires the nonzero parameter values be sufficiently large that, as n, T increase, they can be correctly distinguished from zero. To see this, consider the case when $\gamma_1 = ... = \gamma_p =: \gamma$ and $\zeta_1 = ... = \zeta_p =: \zeta$. In such an event, Assumption 3.1 can be restated as $\gamma\sqrt{\min\{n,T\}}(\underline{\theta}^0)^{-\zeta} = O(1)$ with $\underline{\theta}^0 := \min_{1\leq p\leq P^0}|\theta_p^0|$. Requiring nonzero coefficients to be sufficiently separated from zero is indicative of the approach taken in this paper, which is to appeal to the model selection consistency of the procedure to justify use of the asymptotic expression derived in Section 3.3. This approach is not without its shortcomings, and these are discussed in greater detail in Section 3.3; see also Leeb and Pötscher (2005). Assumption 3.2 requires consistency of the initial estimate $\boldsymbol{\theta}^\dagger$ at some rate c_{nT}. In the following it is shown that the unpenalised likelihood can be used to produce a consistent initial estimate though other estimation procedures might equally be considered.

3 Asymptotic Results

3.1 Consistency

Mirroring Bai (2009), in this section a preliminary consistency result is established which will be improved upon later. Yet, before proceeding, it is worth providing a few remarks on the identification of model parameters. In the standard consistency argument for an extremum estimator, the essence of the idea is to show that *"the limit of the maximum $\hat{\boldsymbol{\theta}}$ should be the maximum of the limit"*, with the latter being unique (Newey and McFadden, 1994, p. 2120). In that argument, the role that identification plays is transparent with population parameters being 'identified' asymptotically. With identification established, uniform convergence of the sample objective function to the limiting objective function often then appeals to a uniform law of large numbers and consistency follows thereafter. Yet in models where the number of parameters, nuisance or otherwise, depends on the sample size, there is no fixed population distribution from which a sample is drawn, and therefore uniform convergence must be considered more carefully. In cases such as these, consistency is often shown directly, forgoing an explicit identification discussion, though identification is still implicit in the construction of $\boldsymbol{\theta}^0 = \text{plim}_{n,T\to\infty}\hat{\boldsymbol{\theta}}$. In this sense Assumption 4.2, given below, fulfils the role of an identification condition.[6]

Before formulating the next assumption, it is necessary to introduce some additional notation. Define the $n\times P$ matrix of instruments $\boldsymbol{Z}_t := (\boldsymbol{G}_1\boldsymbol{X}_t\boldsymbol{\beta}^0, ..., \boldsymbol{G}_Q\boldsymbol{X}_t\boldsymbol{\beta}^0, \boldsymbol{X}_t)$, with $\boldsymbol{G}_q(\boldsymbol{\rho}) := \boldsymbol{W}_q\boldsymbol{S}^{-1}(\boldsymbol{\rho})$ and $\boldsymbol{G}_q := \boldsymbol{G}_q(\boldsymbol{\rho}^0)$. The $n\times T$ data matrix for the instrument associated with some ρ_q is $\sum_{k=1}^{K^0}\beta_k^0\boldsymbol{G}_q\boldsymbol{\mathcal{X}}_k$. The generic $n\times T$ data matrix of either type, $\boldsymbol{\mathcal{X}}_k$ or $\sum_{k=1}^{K^0}\beta_k^0\boldsymbol{G_q}\boldsymbol{\mathcal{X}}_k$, is denoted $\mathcal{Z}_p := (\boldsymbol{z}_{p1}, ..., \boldsymbol{z}_{pT})$, where $\boldsymbol{z}_{pt}$ is the p-th column of $\boldsymbol{Z}_t$, for $p = 1, ..., P$. Finally, let $\boldsymbol{\mathcal{H}}_1 := \frac{1}{nT}\boldsymbol{\mathcal{Z}}'(\boldsymbol{M}_{\boldsymbol{F}^0}\otimes\boldsymbol{M}_{\boldsymbol{\Lambda}})\boldsymbol{\mathcal{Z}}$ and $\boldsymbol{\mathcal{H}}_2 := \frac{1}{nT}\boldsymbol{\mathcal{Z}}'\boldsymbol{\mathcal{Z}}$, where $\boldsymbol{\mathcal{Z}} := (\boldsymbol{Z}_1', ..., \boldsymbol{Z}_T')$ is a $P\times nT$ matrix.

[6] For further discussion see Appendix B in the Supplementary Material.

Assumption 4.

4.1 $R \geq R^0$.

4.2 *For* $i = 1, 2$, *there exist nonstochastic* $P \times P$ *matrices* $\boldsymbol{\mathcal{H}}_i^*$ *such that* $||\boldsymbol{\mathcal{H}}_i - \boldsymbol{\mathcal{H}}_i^*||_2 = o_P(1)$, $\inf_{\boldsymbol{\Lambda} \in \mathbb{R}^{n \times R}, \boldsymbol{F}^0 \in \mathbb{R}^{T \times R^0}} \mu_P(\boldsymbol{\mathcal{H}}_1^*) > 0$, *and* $\mu_1(\boldsymbol{\mathcal{H}}_2^*) < \infty$.

4.3 $\frac{P}{\min\{n,T\}} \to 0$.

Assumption 4.1 allows for the number of factors R^0 to be unknown, as long as the number of factors R used in estimation is no less than R^0; see (Moon and Weidner, 2015). Assumption 4.2 demands a certain level of variation in sample data after projecting out the true factors and arbitrary factor loadings. This condition is especially intuitive when considering the particular case of individual or time effects, in which case the projections perform between individual and between time period differences to the data. It is also worth noting that Assumption 4.2 implies

$$\sup_{\boldsymbol{\Lambda} \in \mathbb{R}^{n \times R}, \boldsymbol{F}^0 \in \mathbb{R}^{T \times R^0}} \mu_1 \left(\boldsymbol{\mathcal{H}}_1^* \right) < c < \infty \tag{13}$$

and

$$\mu_P \left(\boldsymbol{\mathcal{H}}_2^* \right) > c > 0, \tag{14}$$

which ensures both $\boldsymbol{\mathcal{H}}_1$ and $\boldsymbol{\mathcal{H}}_2$ are well defined asymptotically (see Appendix C in the Supplementary Material for details). Assumption 4.3 requires that the number of parameters does not grow too fast in relation to n and T. This is necessary since consistency is stated in terms of the ℓ_2 norm of a vector with increasing dimension. Unfettered growth in the number of parameters relative to the sample size could thus lead to inconsistency even in the event that an estimator converged pointwise.

Proposition 1 (Consistency)**.** *Under Assumptions 1–4, with probability approaching 1, there exist global maximisers of the unpenalised and penalised average likelihood functions,* $\tilde{\boldsymbol{\theta}}$ *and* $\hat{\boldsymbol{\theta}}$ *respectively, such that* $||\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^0||_2 = O_P\left(a_{nT}\right)$ *and* $||\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^0||_2 = O_P\left(a_{nT}\right)$, *with* $a_{nT} := \sqrt{\frac{P}{\min\{n,T\}}}$.

This preliminary result is an important step towards those which follow. Moreover, the result is of interest in and of itself since it applies provided that the number of factors is not underspecified, and irrespective of the relationship between n and T, as long as both diverge to infinity. Despite both the factors and the loadings having been concentrated out, the spaces spanned by both are implicitly estimated by their respective first order conditions, and as a result both n and T are required to diverge. Finally, the rate a_{nT} is in line with the existing literature; see for example Theorem 4.1 in Moon and Weidner (2015), where a preliminary $\sqrt{\min\{n,T\}}$-consistency rate is established for a fixed number of (non-nuisance) parameters.[7]

3.2 Selection Consistency

In addition to the consistency result established in Proposition 1, it is also desirable that the proposed estimator is selection consistent. This requires that, with probability approaching 1, the estimates of the truly zero coefficients are zero, while those of nonzero coefficients are nonzero.

Assumption 5. $\min_{1\leq p\leq P} \gamma_p c_{nT}^{-\zeta_p} \to \infty$ *as* $n, T \to \infty$.

Assumption 5 ensures selection consistency of the estimator by taking advantage of the singularity of the penalty term at zero. Under Assumption 5, the term $\gamma_p/|\theta_p^{\dagger}|^{\zeta_p}$ will be explosive in probability for those truly zero θ_p and as a result, asymptotically, the first order conditions cannot not be met unless $\hat{\theta}_p$ takes a value of exactly zero. For the following, recall from the end of Section 2.1 that $\boldsymbol{\theta}_{(2)}$ contains the truly zero θ_p.

Proposition 2 (Selection Consistency)**.** *Under Assumptions 1–5,*

$$\Pr\left(||\hat{\boldsymbol{\theta}}_{(2)}||_2 = 0\right) \to 1 \ \text{as } n, T \to \infty. \tag{15}$$

Proposition 2 demonstrates that the estimator will correctly set coefficients with a true value of zero to exactly zero with probability approaching 1. Moreover, the consistency result proved in Proposition 1 implies that, with probability approaching 1, the estimates of nonzero coefficients must be nonzero. Thus together, Propositions 1 and 2 indicate that, with an appropriate choice of regularisation parameters, the PQMLE is model selection consistent.

3.3 Asymptotic Distribution

An implication of the model selection consistency result obtained in Proposition 2 is that, for every θ, the asymptotic distribution of nonzero coefficient estimates coincides with that

[7] As a referee points out, by imposing sparsity, and, with a judicious and data specific choice of penalty parameter, it may be possible to obtain faster rates of convergence. This may be of particular significance in very high dimensional settings with potentially $P > nT$, though such results are not pursued in this paper.

of the infeasible 'oracle' estimator, which uses knowledge of which parameters are truly zero. The limiting distribution of the nonzero coefficients is derived appealing to this result, and, in keeping with the high dimensional literature, this is done indirectly, by considering arbitrary linear combinations of parameters. In adopting this approach, the results which are obtained are pointwise, and thereby it is implicitly assumed that the true parameters have fixed values that are, recalling Assumption 3.1, sufficiently separated from zero. A consequence of the lack of uniformity is that the finite sample distribution of the estimator may be quite different to that derived in Theorem 1; a point made clear by Leeb and Pötscher (2005). However, this is a broader issue in the literature and is particularly difficult to overcome in models of significant complexity, where obtaining uniform results is often challenging.

Assumption 6.

6.1 $\frac{QP^3}{\min\{n,T\}} \to 0$ *as* $n, T \to \infty$.

6.2 $\frac{1}{n}\mathbf{\Lambda}^{0\prime}\mathbf{\Lambda}^0 \xrightarrow{p} \mathbf{\Sigma}_{\mathbf{\Lambda}^0}$ *as* $n \to \infty$ *with* $\mathbf{\Sigma}_{\mathbf{\Lambda}^0}$ *being a* $R^0 \times R^0$ *positive definite matrix.*

6.3 $\frac{1}{T}\boldsymbol{F}^{0\prime}\boldsymbol{F}^0 \xrightarrow{p} \mathbf{\Sigma}_{\boldsymbol{F}^0}$ *as* $T \to \infty$ *with* $\mathbf{\Sigma}_{\boldsymbol{F}^0}$ *being a* $R^0 \times R^0$ *positive definite matrix.*

6.4 $\frac{T}{n} \to c$ *with* $0 < c < \infty$.

6.5 $R = R^0$.

6.6 $\max_{1\leq p\leq P} \gamma_p \sqrt{PnT}|\theta^0|^{-\zeta_p} = o(1)$.

Assumption 6.1 ensures that the estimation of the coefficients has a negligible effect on the estimation of the factors and the loadings. Lu and Su (2016), who consider estimation of a standard regression model without interactions, require $P^2/\min\{n,T\} \to 0$ for analogous purposes. A stronger condition is needed here to ensure that the estimators of the reduced form factors $\boldsymbol{S}^{-1}(\boldsymbol{\rho})\mathbf{\Lambda}$ converge sufficiently fast, since the reduced form is implicitly used in instrumenting the endogenous variables. As $\boldsymbol{S}(\boldsymbol{\rho}) = \boldsymbol{I}_n - \sum_{q=1}^{Q} \rho_q \boldsymbol{W}_q$ involves an increasing number of weights matrices, the number of these cannot be allowed to increase too quickly. Moreover the convergence of the $P \times P$ covariance matrix of $\hat{\boldsymbol{\theta}}$ requires further limits on the growth of P. Fan and Peng (2004) require $P^5/n \to 0$, which corresponds to Assumption 6.1 in a cross-sectional framework. The condition given in Liu (2017), in a cross-sectional spatial model without a factor structure error effects, also requires $P^5/n \to 0$. Assumptions 6.2 and 6.3 impose that the factors are strong, that is to say that the factors and loadings have a nonnegligible impact on the variance of the unobserved term $\boldsymbol{\eta} := (\boldsymbol{\eta}_1, ..., \boldsymbol{\eta}_T)$. Other authors consider models with weak factors however this is not pursued here. Assumption 6.4 requires n and T to grow in proportion. Similar asymptotic regimes are assumed in several papers in which biases arise in models with interactive fixed effects, and which use similar estimation approaches. Examples of these include Moon and Weidner (2017) and Shi and Lee (2017). Other papers, such as Bai (2009) and Lu and Su (2016), consider regimes where both $n/T^2, T/n^2 \to 0$, which provide similar limits on the relative growth rates of n and T. Assumption 6.5 requires the true number of factors to be known. Nonetheless, Proposition 1 shows that the PQML estimator remains consistent as long as the number of factors is not understated; that is $R \geq R^0$. In the absence of interaction, Moon and Weidner (2015) show that the asymptotic distribution of a least squares estimator is unaffected by overstatement of the number of factors, under certain conditions. It might, therefore, be expected that this extends to the present setting, however, since there may be significant complications in obtaining such results, the asymptotic distribution is derived under the assumption $R = R^0$. Section 4.2 shows how the number of factors can be chosen consistently with reference to an information criterion. Assumption 6.6 strengthens the restrictions on the penalty term.

Let $\mathcal{D}$ denote the sigma algebra generated by $\boldsymbol{\mathcal{X}}_1^*, ..., \boldsymbol{\mathcal{X}}_{K^0}^*$, $\boldsymbol{\Lambda}^0$ and $\boldsymbol{F}^0$. Define $\bar{\mathfrak{Z}}_p := \mathbb{E}[\mathfrak{Z}_p|\mathcal{D}]$, $\mathbb{Z}_p := \boldsymbol{M}_{\boldsymbol{\Lambda}^0}\bar{\mathfrak{Z}}_p\boldsymbol{M}_{\boldsymbol{F}^0} + (\mathfrak{Z}_p - \bar{\mathfrak{Z}}_p)$, $\mathbb{Z}_{(1)} := (\text{vec}(\mathbb{Z}_1), ..., \text{vec}(\mathbb{Z}_{P^0}))$, and $\boldsymbol{\mathcal{Z}}_{(1)} := (\text{vec}(\mathfrak{Z}_1), ..., \text{vec}(\mathfrak{Z}_{P^0}))$, that is, $\mathbb{Z}_{(1)}$ and $\boldsymbol{\mathcal{Z}}_{(1)}$ contain only covariates associated with nonzero parameters. Also, let

$$\boldsymbol{D} := \frac{1}{\sigma_0^2}\frac{1}{nT}\boldsymbol{\mathcal{Z}}_{(1)}'(\boldsymbol{M}_{\boldsymbol{F}^0} \otimes \boldsymbol{M}_{\boldsymbol{\Lambda}^0})\boldsymbol{\mathcal{Z}}_{(1)} + \begin{pmatrix} \boldsymbol{\Omega} & \boldsymbol{0}_{Q^0\times K^0} \\ \boldsymbol{0}_{K^0\times Q^0} & \boldsymbol{0}_{K^0\times K^0} \end{pmatrix}, \tag{16}$$

$$\boldsymbol{V} := \frac{\mathcal{M}_\varepsilon^3}{\sigma_0^4}(\boldsymbol{\Phi} + \boldsymbol{\Phi}') + \frac{\mathcal{M}_\varepsilon^4 - 3\sigma_0^4}{\sigma_0^4}\begin{pmatrix} \boldsymbol{\Xi} & \boldsymbol{0}_{Q^0\times K^0} \\ \boldsymbol{0}_{K^0\times Q^0} & \boldsymbol{0}_{K^0\times K^0} \end{pmatrix}, \tag{17}$$

where the matrices $\boldsymbol{\Omega}$ and $\boldsymbol{\Xi}$ are $Q^0 \times Q^0$ with elements $\Omega_{qq'} := \frac{1}{n}\text{tr}(\boldsymbol{G}_q(\boldsymbol{G}_{q'} + \boldsymbol{G}_{q'}')) - \frac{2}{n^2}\text{tr}(\boldsymbol{G}_q)\text{tr}(\boldsymbol{G}_{q'})$ and $\Xi_{qq'} := \sum_{t=1}^T\sum_{i=1}^n(\boldsymbol{G}_q^*)_{ii}(\boldsymbol{G}_{q'}^*)_{ii}$, respectively, for $q, q' = 1, ..., Q^0$, and with $\boldsymbol{G}_q^* := \boldsymbol{G}_q - \frac{1}{n}\text{tr}(\boldsymbol{G}_q)\boldsymbol{I}_n$. The matrix $\boldsymbol{\Phi}$ is $P^0 \times P^0$ and has the structure $\boldsymbol{\Phi} := (\bar{\boldsymbol{\Phi}}', \boldsymbol{0}_{P^0\times K^0})'$, with $\bar{\Phi}_{qp} := \sum_{t=1}^T\sum_{i=1}^n(\mathbb{Z}_p)_{it}(\boldsymbol{G}_q^*)_{ii}$, for $q = 1, ..., Q^0$ and $p = 1, ..., P^0$.

Assumption 7.

7.1 *For some fixed integer L, $\mathbb{S}$ is a nonstochastic $L \times P^0$ matrix such that $\mathbb{S}\mathbb{S}'$ converges to a (entrywise) nonnegative matrix with eigenvalues bounded away from zero and infinity as $n, T \to \infty$.*

7.2 There exist nonstochastic $P^0 \times P^0$ matrices $\mathbb{D} := \mathbb{E}[\boldsymbol{D}]$ and $\mathbb{V} := \mathbb{E}[\boldsymbol{V}]$ such that $||\boldsymbol{D} - \mathbb{D}||_2 = o_P(1), ||\boldsymbol{V} - \mathbb{V}||_2 = o_P(1)$, and the eigenvalues of $\mathbb{D}$, $\mathbb{V}$ and $\mathbb{D} + \mathbb{V}$ are bounded from below by zero and from above.

Since the limiting distribution of the estimator is difficult to derive directly, a selection matrix $\mathbb{S}$ is introduced with a finite dimension L. Assumption 7.1 sets out basic properties of this matrix. Assumption 7.2 ensures that the covariance matrix of the PQMLE $\hat{\boldsymbol{\theta}}$ is well defined asymptotically. Let $\mathcal{M}_\varepsilon^m$ denote the m-th raw moment of ε_{it}, $\boldsymbol{J}_h := (\boldsymbol{0}_{T\times(T-h)}, \boldsymbol{I}_T, \boldsymbol{0}_{T\times h})'$, are recall that $\boldsymbol{\theta}_{(1)}$ contains only those truly nonzero coefficients.

Theorem 1 (Asymptotic Normality)**.** *Under Assumptions 1–7,*

$$\sqrt{nT}\big(\mathbb{S}(\boldsymbol{D} + \boldsymbol{V})\mathbb{S}'\big)^{-\frac{1}{2}}\mathbb{S}\big(\boldsymbol{D}(\hat{\boldsymbol{\theta}}_{(1)} - \boldsymbol{\theta}^0_{(1)}) - \mathbb{b}\big) \xrightarrow{d} \mathcal{N}\big(\boldsymbol{0}_{L\times L}, \boldsymbol{I}_L\big), \tag{18}$$

with

$$\mathbb{b} := \begin{pmatrix} \mathbb{b}^{(1)} \\ \boldsymbol{0}_{K^0\times 1} \end{pmatrix} + \begin{pmatrix} \mathbb{b}^{(2)} \\ \boldsymbol{0}_{K^{*0}\times 1} \\ \mathbb{b}^{(3)} \end{pmatrix}, \tag{19}$$

where the vector $\mathbb{b}^{(1)}$ is $Q^0 \times 1$ with elements $\mathbb{b}_q^{(1)} := \sqrt{\frac{T}{n}}(\frac{R^0}{n}\mathrm{tr}(\boldsymbol{G}_q) - \mathrm{tr}(\boldsymbol{P}_{\boldsymbol{\Lambda}^0}\boldsymbol{G}_q))$, the vector $\mathbb{b}^{(2)}$ is $Q^0 \times 1$ with elements $\mathbb{b}_q^{(2)} := -\frac{1}{\sqrt{nT}}\sum_{h=1}^{T-1}\mathrm{tr}(\boldsymbol{J}_0\boldsymbol{P}_{\boldsymbol{F}^0}\boldsymbol{J}_h')\mathrm{tr}(\boldsymbol{W}_q\boldsymbol{A}^h\boldsymbol{S}^{-1})$ and the vector $\mathbb{b}^{(3)}$ is $(Q^0+1)\times 1$ with first element $\mathbb{b}_1^{(3)} := -\frac{1}{\sqrt{nT}}\sum_{h=1}^{T-1}\mathrm{tr}(\boldsymbol{J}_0\boldsymbol{P}_{\boldsymbol{F}^0}\boldsymbol{J}_h')\mathrm{tr}(\boldsymbol{A}^{h-1}\boldsymbol{S}^{-1})$ and remaining elements $\mathbb{b}_{q+1}^{(3)} := -\frac{1}{\sqrt{nT}}\sum_{h=1}^{T-1}\mathrm{tr}(\boldsymbol{J}_0\boldsymbol{P}_{\boldsymbol{F}^0}\boldsymbol{J}_h')\mathrm{tr}(\boldsymbol{W}_q\boldsymbol{A}^{h-1}\boldsymbol{S}^{-1})$.

Theorem 1 describes the asymptotic properties of the estimators for the nonzero coefficients, detailing the asymptotic covariance matrix and the bias terms which arise. Closer inspection reveals the bias $\mathbb{b}^{(1)}$ is of order $\sqrt{T/n}$, while $\mathbb{b}^{(2)}$ and $\mathbb{b}^{(3)}$ are of order $\sqrt{n/T}$. These biases are a consequence of the incidental parameters in both dimensions of the panel. The bias $\mathbb{b}^{(1)}$ depends on the resemblance between the loadings and the network structure; both are sources of cross-sectional dependence and therefore may be easily conflated. If the column space of $\boldsymbol{G}_q$ is orthogonal to the space of loadings, then $\boldsymbol{P}_{\boldsymbol{\Lambda}^0}\boldsymbol{G}_q = \boldsymbol{0}_{n\times n}$ and $\mathbb{b}^{(1)}$ does not feature. The second source of bias is characterised in $\mathbb{b}^{(2)}$ for the $\boldsymbol{\rho}$ coefficients, and in $\mathbb{b}^{(3)}$ for the $\boldsymbol{\phi}$ coefficients. In contrast, this bias is related to the relationship between the structure of time dependence and the factors.

Expression (18) can be simplified in the case when the number of parameters does not depend on n, T. To see this, set $\mathbb{S} = \boldsymbol{I}_{P^0}$. If both $\boldsymbol{D}$ and $\mathbb{D}$ are positive definite, then (18) becomes

$$\sqrt{nT}\big(\hat{\boldsymbol{\theta}}_{(1)} - \boldsymbol{\theta}^0_{(1)}\big) - \boldsymbol{D}^{-1}\mathbb{b} \xrightarrow{d} \mathcal{N}\big(\mathbf{0}_{P^0\times P^0}, \mathbb{D}^{-1}(\mathbb{D}+\mathbb{V})\mathbb{D}^{-1}\big). \tag{20}$$

Thus, with a fixed number of parameters, the covariance matrix has a standard sandwich form, where $\mathbb{V}$ reflects additional structure in the event that the errors are not normally distributed. In the case of error normality, $\mathcal{M}^3_\varepsilon = 0$ and $\mathcal{M}^4_\varepsilon = 3\sigma_0^4$ and so $\mathbb{V}$ is not present. The covariance matrix then reduces to $\mathbb{D}^{-1}$.

3.4 Bias Correction

Given the characterisation of the bias term in Theorem 1, it is shown in the following proposition that this can be consistently estimated and the limiting distribution of the estimates can be recentred. Let $\hat{\boldsymbol{D}}$ and $\hat{\mathbb{b}}$ denote the analogues of $\boldsymbol{D}$ and $\mathbb{b}$, respectively, with $\boldsymbol{\theta}^0, \boldsymbol{F}^0$, $\boldsymbol{\Lambda}^0$ and σ_0^2 replaced by their estimates.

Proposition 3 (Bias Correction)**.** *Under Assumptions 1–7,*

$$\sqrt{nT}\big(\mathbb{S}(\boldsymbol{D}+\boldsymbol{V})\mathbb{S}'\big)^{-\frac{1}{2}}\mathbb{S}\boldsymbol{D}\big(\hat{\boldsymbol{\theta}}^c_{(1)} - \hat{\boldsymbol{\theta}}^0_{(1)}\big) \xrightarrow{d} \mathcal{N}\big(\mathbf{0}_{L\times L}, \boldsymbol{I}_L\big), \tag{21}$$

with $\hat{\boldsymbol{\theta}}^c_{(1)} := \hat{\boldsymbol{\theta}}_{(1)} - \hat{\boldsymbol{D}}^{-1}\hat{\mathbb{b}}$ *being the bias corrected estimates.*

4 Implementation

The estimation procedure suggested in this paper requires input of the number of factors and the regularisation parameters. Several methods have been suggested elsewhere in the literature to inform these choices. This section shows how two of these methods, one for selecting the number of factors, and one for selecting the regularisation parameters, can be adapted to the present context. Ideally, it would be preferable to select both of these jointly, however, the approach adopted here is pragmatic, and relies on the fact that, by Proposition 1, the coefficients can be consistently estimated with knowledge only of an upper bound on the number of factors. Performing penalised estimation using this upper bound, one obtains consistent estimates of the coefficients from which a pure factor model can be obtained, and the true number of factors detected. Of course, this two step procedure neglects to account for uncertainty in the first stage, therefore additional Monte Carlo results are provided in Appendix J of the Supplementary Material in order to assess the possible impact of varying the number of factors on the properties of the estimator.

4.1 Choosing the Penalty Parameter

In this section the appropriate choice of the penalty parameter $\boldsymbol{\gamma}$ is considered. The choice of $\boldsymbol{\zeta}$, the other parameter in (12), is not discussed, because typically this can be appropriately chosen with reference to the rate of convergence of the initial estimate. For simplicity, attention is restricted to the case in which the penalty parameter γ_p is varied only between ρ and β; that is, it is assumed that $\gamma_1 = ... = \gamma_Q =: \gamma^{(1)}$ and $\gamma_{Q+1} = ... = \gamma_P =: \gamma^{(2)}$. Accordingly, redefine $\boldsymbol{\gamma}$ as the 2-dimensional vector $(\gamma^{(1)}, \gamma^{(2)})'$. An information criterion is considered to select $\boldsymbol{\gamma}$, analogous to that proposed in Lu and Su (2016), except that here the penalty parameter may vary across the two coefficient types, rather than varying between the regression coefficients and the loadings, as in that paper. This information criterion takes the form

$$\mathrm{IC}^*(\boldsymbol{\gamma}) := \hat{\sigma}^2(\boldsymbol{\gamma}) + \varrho_\rho|\mathcal{S}_\rho(\boldsymbol{\gamma})| + \varrho_\beta|\mathcal{S}_\beta(\boldsymbol{\gamma})|, \tag{22}$$

where the notation $\hat{\sigma}^2(\boldsymbol{\gamma})$ is used for $\hat{\sigma}^2$ to emphasise the dependence on $\boldsymbol{\gamma}$, ϱ_ρ and ϱ_β are some positive penalty functions of (n, T), and $\mathcal{S}_\rho(\boldsymbol{\gamma})$, $\mathcal{S}_\beta(\boldsymbol{\gamma})$ denote the index sets for the nonzero elements of the parameter estimates under $\boldsymbol{\gamma}$. Following closely the exposition in Lu and Su (2016), define $\mathcal{S}_{F,\rho} := \{1, ..., Q\}$ and $\mathcal{S}_{F,\beta} := \{1, ..., K\}$ as the index sets for the full set of weights matrices and for all covariates respectively. Analogous index sets $\mathcal{S}_{T,\rho} := \{1, ..., Q^0\}$ and $\mathcal{S}_{T,\beta} := \{1, ..., K^0\}$ are used to denote the relevant covariates and weights matrices. Next, define two closed intervals, $\Gamma_\rho := [0, \bar{\gamma}^{(1)}]$ and $\Gamma_\beta := [0, \bar{\gamma}^{(2)}]$, with $\Gamma_\rho, \Gamma_\beta \subset \mathbb{R}_+$ and where $\bar{\gamma}^{(1)}, \bar{\gamma}^{(2)}$ are two upper bounds beyond which all parameters would be set to zero. The space $\Gamma := \Gamma_\rho \times \Gamma_\beta$ can be subdivided into three regions:

$$\Gamma^0 := \{\boldsymbol{\gamma} \in \Gamma : \mathcal{S}_\rho(\boldsymbol{\gamma}) = \mathcal{S}_{T,\rho} \text{ and } \mathcal{S}_\beta(\boldsymbol{\gamma}) = \mathcal{S}_{T,\beta}\},$$

$$\Gamma^- := \{\boldsymbol{\gamma} \in \Gamma : \mathcal{S}_\rho(\boldsymbol{\gamma}) \not\supset \mathcal{S}_{T,\rho} \text{ or } \mathcal{S}_\beta(\boldsymbol{\gamma}) \not\supset \mathcal{S}_{T,\beta})\},$$

$$\Gamma^+ := \{\boldsymbol{\gamma} \in \Gamma : \mathcal{S}_\rho(\boldsymbol{\gamma}) \supset \mathcal{S}_{T,\rho}, \mathcal{S}_\beta(\boldsymbol{\gamma}) \supset \mathcal{S}_{T,\beta} \text{ and } |\mathcal{S}_\rho(\boldsymbol{\gamma})| + |\mathcal{S}_\beta(\boldsymbol{\gamma})| > |\mathcal{S}_{T,\rho}| + |\mathcal{S}_{T,\beta}|\},$$

where $|.|$ denotes the cardinality of a set. Respectively, these are the sets of $\boldsymbol{\gamma}$ in which the true model is selected, the model is underfitted and the model is overfitted. The following assumptions are made.

Assumption 8.

8.1 As $n, T \to \infty$, $(\sqrt{Q}a_{nT})^{-1}\varrho_\rho \to \infty, (\sqrt{Q}a_{nT})^{-1}\varrho_\beta \to \infty$, $Q^0\varrho_\rho \to 0$, and $K^0\varrho_\beta \to 0$.

8.2 For any $\boldsymbol{\gamma} \in \Gamma^-$, there exists σ_-^2 such that $\hat{\sigma}^2(\boldsymbol{\gamma}) \xrightarrow{p} \sigma_-^2 > \sigma_0^2$.

Assumption 8 is analogous to Assumptions A.7 and A.8 in Lu and Su (2016). Assumption 8.1 requires that the penalty functions ϱ_ρ and ϱ_β relax sufficiently fast as sample size increases. In practice, there may be many functions which satisfy Assumption 8.1, though these may have different impacts in finite samples; for further discussion see Bai and Ng (2002). Assumption 8.2 ensures that underfitted models yield a larger mean squared error than a correctly fitted model.

Proposition 4 (Information Criterion Consistency)**.** *Under Assumptions 1–8,*

$$\Pr\left(\inf_{\boldsymbol{\gamma}\in\Gamma^-\cup\Gamma^+} \mathrm{IC}^*(\boldsymbol{\gamma}) > \mathrm{IC}^*(\boldsymbol{\gamma}^0)\right) \to 1 \text{ as } n, T \to \infty, \tag{23}$$

for any $\boldsymbol{\gamma}^0 \in \Gamma^0$.

4.2 Choosing the Number of Factors

In order to select the number of factors to be used in estimation, it is suggested to first perform penalised estimation of the parameters with the number of factors overspecified. Assume an upper bound on the number of factors $R_{\max} \geq R^0$ is known, and then let $\check{\boldsymbol{\theta}} = (\check{\boldsymbol{\rho}}', \check{\boldsymbol{\beta}}')'$ be the estimates obtained under $R = R_{\max}$. A pure factor model can then be constructed as

$$\boldsymbol{S}(\check{\boldsymbol{\rho}})\boldsymbol{Y} - \sum_{k=1}^{K} \check{\beta}_k \boldsymbol{\mathcal{Z}}_k = \boldsymbol{\Lambda}^0 \boldsymbol{F}^{0\prime} + \check{\boldsymbol{\varepsilon}}, \tag{24}$$

with $\check{\boldsymbol{\varepsilon}} := \sum_{q=1}^{Q}(\rho_q^0 - \check{\rho}_q)\boldsymbol{G}_q\big(\sum_{k=1}^{K^0} \beta_k^0 \boldsymbol{\mathcal{X}}_k + \boldsymbol{\Lambda}^0\boldsymbol{F}^{0\prime} + \boldsymbol{\varepsilon}\big) + \sum_{k=1}^{K}(\beta_k^0 - \check{\beta}_k)\boldsymbol{\mathcal{X}}_k + \boldsymbol{\varepsilon}$. Existing information criteria can then be used to detect the number of factors, and this suggested number can be input into a second estimation step. For example, Shi and Lee (2017) considers information criteria of the form

$$\mathrm{IC}(R) := \log\left(\frac{1}{nT}\sum_{i=R+1}^{n} \mu_i\left(\left(\boldsymbol{\Lambda}^0\boldsymbol{F}^{0\prime} + \check{\boldsymbol{\varepsilon}}\right)\left(\boldsymbol{\Lambda}^0\boldsymbol{F}^{0\prime} + \check{\boldsymbol{\varepsilon}}\right)'\right)\right) + \varrho_f R, \tag{25}$$

with ϱ_f being a positive penalty function of (n, T). With minor modification to Theorem 5 in that paper, it can be shown that the information criterion in (25) is consistent in determining the number of factors, in the sense that $\lim_{n,T\to\infty} \Pr(R^* = R^0) = 1$, with $R^* := \arg\min_{0\leq R\leq R_{\max}} \mathrm{IC}(R)$ and under the additional assumption that the penalty function ϱ_f satisfies $\varrho_f \to 0$ and $a_{nT}\varrho_f \to \infty$, with a_{nT} being the preliminary rate established in Proposition 1.

5 Illustration

This section demonstrates the finite sample performance and practicability of the procedure through the use of a small Monte Carlo study and an empirical example.

5.1 Simulations

In the following design, the data are generated according to model (1), with the number of parameters and weights matrices increasing with sample size. The design is summarised in Table 1 with a little under half of the parameters taking a true value of 0 for each sample size. Dashes in the table indicate that a covariate is absent.

Table 1: True parameter values

n	T	ρ_1^0	ρ_2^0	ρ_3^0	ρ_4^0	ρ_5^0	δ_1^0	δ_2^0	δ_3^0	δ_4^0	δ_5^0	δ_{11}^0	δ_{12}^0	δ_{13}^0	δ_{14}^0	δ_{15}^0	ϕ_1^0	ϕ_2^0	ϕ_3^0	ϕ_4^0	ϕ_5^0
	25	0.2	0.2	0	-	-	3	0	-3	-	-	1	0	-1	-	-	0.15	0	-0.15	-	-
25	50	0.2	0.2	0	-	-	3	0	-3	0	-	1	0	-1	-	-	0.15	0	-0.15	-	-
	100	0.2	0.2	0	-	-	3	0	-3	0	3	1	0	-1	-	-	0.15	0	-0.15	-	-
	25	0.2	0.2	0	0.2	-	3	0	-3	-	-	1	0	-1	0	-	0.15	0	-0.15	0	-
50	50	0.2	0.2	0	0.2	-	3	0	-3	0	-	1	0	-1	0	-	0.15	0	-0.15	0	-
	100	0.2	0.2	0	0.2	-	3	0	-3	0	3	1	0	-1	0	-	0.15	0	-0.15	0	-
	25	0.2	0.2	0	0.2	0	3	0	-3	-	-	1	0	-1	0	1	0.15	0	-0.15	0	0
100	50	0.2	0.2	0	0.2	0	3	0	-3	0	-	1	0	-1	0	1	0.15	0	-0.15	0	0
	100	0.2	0.2	0	0.2	0	3	0	-3	0	3	1	0	-1	0	1	0.15	0	-0.15	0	0

The error term ε_{it}, the loadings λ_{ir}^0 and the factors f_{tr}^0 are generated as standard normal variables.[8] Primitive exogenous variables are generated according to $x_{\kappa it}^* = \nu + \sum_{r=1}^{R^0} \lambda_{ir}^0 f_{rt}^0 + e_{it}$ with ν being uniformly drawn from the integers $\{-10, ..., 10\}$ and $e_{it} \sim \mathcal{N}(0, 2)$. By design these are correlated with the factors and the loadings and have associated coefficients δ_κ^0. There are also additional covariates formed by interacting the q-th weights matrix with the first primitive exogenous regressor in the manner of (2). These covariates have associated coefficients δ_{1q}^0. The number of weights matrices is increasing with n, with the first weights matrix being constructed as if the cross-sectional units were arrayed on a line and connected only to the units immediately to the left and right. This is the simplest example of a path and produces a matrix with ones along the diagonals immediately above and below the main diagonal, and zeros elsewhere. The remaining matrices are specified in similar fashion, but now represent neighbours to the q-th degree. All matrices are then row normalised. Finally,

[8]For simplicity results are reported only for idiosyncratic errors generated according to a standard normal. Similar results can be obtained under alternative error distributions and additional simulation results are available in Appendix J in the Supplementary Material.

a lag of outcomes is included, as well as interactions of this lagged outcome and the weights matrices.[9]

Table 2 reports bias corrected estimates $\hat{\boldsymbol{\theta}}^c$, across various n and T, each with 1000 Monte Carlo replications, and where $R = R^0 = 3$.

Table 2: Bias of bias corrected estimates of nonzero parameters ($R = R^0$)

n	T	ρ_1	ρ_2	ρ_4	δ_1	δ_3	δ_5	δ_{11}	δ_{13}	δ_{15}	ϕ_1	ϕ_3
	25	0.0002	−0.0004	-	0.0008	−0.0014	-	−0.0027	0.0031	-	−0.0004	0.0004
25	50	0.0001	−0.0002	-	−0.0002	−0.0006	-	−0.0016	0.0026	-	−0.0002	0.0002
	100	0.0001	−0.0002	-	0.0001	−0.0005	0.0005	−0.0014	0.0017	-	−0.0001	0.0001
	25	0.0001	0	−0.0001	0.0005	0.0005	-	−0.0007	0.0005	-	−0.0002	0.0002
50	50	0.0002	−0.0003	0	0.0002	−0.0006	-	−0.0005	0.0013	-	−0.0001	0.0001
	100	0	−0.0001	0	−0.0001	−0.0003	0.0003	0	0.0005	-	−0.0002	0.0002
	25	−0.0001	−0.0002	0.0002	0.0003	−0.0011	-	−0.0004	0.0022	−0.0006	−0.0003	0.0003
100	50	0	0	0	0.0003	−0.0002	-	0.0001	0.0005	−0.0006	−0.0003	0.0002
	100	0.0001	−0.0001	0	0	−0.0001	0.0002	−0.0004	0.0007	−0.0001	−0.0002	0.0002

Table 2 shows that the biases are generally decreasing with both n and T and tend to be larger for the parameters δ_1, δ_3 and δ_5, as well as the exogenous spillovers δ_{11}, δ_{13} and δ_{15}. This is unsurprising since the covariates $\boldsymbol{\mathcal{X}}_\kappa^*$ are directly correlated with the loadings and the factors by design. The biases of the ρ_q parameters are lower since these implicitly use the instrument $\boldsymbol{W}_q\boldsymbol{S}^{-1}\boldsymbol{X}_t\boldsymbol{\beta}^0$, which may not itself be strongly correlated with the factors and the loadings. The same is true of the coefficients ϕ_1 and ϕ_3, since the lags $\boldsymbol{\mathcal{Y}}_{-1}$ and interactions $\boldsymbol{W}_q\boldsymbol{\mathcal{Y}}_{-1}$ are less directly correlated with the factors and the loadings. These biases can be favourably compared with Table 6 in Appendix J of the Supplementary Material which presents biases of the penalised maximum likelihood estimates without controlling for interactive effects, where there are large biases which persist with n and T.

Table 3 presents coverage probabilities of Wald confidence intervals based on Theorem 1 and with a nominal coverage of 95%. These generally improve with n and T, though due to the complexity of the design it is unsurprising that they do not do so monotonically. Table 4 shows the percentage of true zero parameters correctly estimated as such, with the procedure performing well and achieving near 100% accuracy across all n and T.

[9] Assumptions 1–8 are verified for this design in Appendix I of the Supplementary Material.

Table 3: Coverage of nonzero parameter estimates ($R = R^0$)

n	T	ρ_1	ρ_2	ρ_4	δ_1	δ_3	δ_5	δ_{11}	δ_{13}	δ_{15}	ϕ_1	ϕ_3
	25	0.901	0.902	-	0.885	0.908	-	0.891	0.897	-	0.904	0.907
25	50	0.906	0.922	-	0.921	0.924	-	0.922	0.928	-	0.916	0.922
	100	0.930	0.919	-	0.926	0.929	0.920	0.917	0.930	-	0.929	0.915
	25	0.920	0.932	0.931	0.924	0.927	-	0.927	0.927	-	0.913	0.920
50	50	0.939	0.935	0.931	0.936	0.926	-	0.932	0.917	-	0.926	0.930
	100	0.946	0.942	0.922	0.932	0.934	0.932	0.945	0.921	-	0.921	0.928
	25	0.929	0.929	0.923	0.930	0.921	-	0.926	0.916	0.932	0.934	0.931
100	50	0.937	0.935	0.947	0.941	0.926	-	0.920	0.939	0.939	0.931	0.934
	100	0.947	0.930	0.942	0.950	0.946	0.948	0.941	0.957	0.942	0.922	0.921

Table 4: Percentage of true zeros ($R = R^0$)

n	T	ρ_3	ρ_5	δ_2	δ_4	δ_{12}	δ_{14}	ϕ_2	ϕ_4	ϕ_5
	25	99.9	-	100	-	99.9	-	99.9	-	-
25	50	99.8	-	100	100	100	-	99.8	-	-
	100	99.8	-	100	100	100	-	99.9	-	-
	25	100	-	100	-	100	100	100	100	-
50	50	99.9	-	100	100	100	100	99.9	99.9	-
	100	99.6	-	100	100	100	100	99.6	99.6	-
	25	99.9	99.9	99.9	-	99.9	99.9	99.9	99.9	99.9
100	50	99.8	99.8	100	100	100	100	99.8	99.8	99.8
	100	99.7	99.8	100	100	100	100	99.7	99.7	99.7

As well as estimation where the correct number of factors is inputted ($R = R^0 = 3$), additional estimations are performed for each Monte Carlo draw, with the postulated number of factors R being $1, 2, 4$ and 5. For each number of factors, the information criterion (25) is then computed. Table 5 presents the number of times, as a percentage, that the true number of factors is found to minimise the information criterion. Three variants of this criterion are used (IC1, IC2 and IC3) which differ only in their choice of penalty function ϱ_f.[10] As sample size increases, the performance of all three variants improves, though there is significant variability between the three criteria.[11]

[10]The functions used in IC1, IC2 and IC3 are, respectively, $\log(\min\{n,T\})/\min\{n,T\}$, $((n+T)/(nT))\log(\min\{n,T\})$ and $((n+T)/(nT))\log((nT)/(n+T))$. For both ϱ_ρ and ϱ_β in IC*, $\log(\min\{n,T\})/\min\{n,T\}$ is used.

[11]The penalty function IC1 is smaller in magnitude than IC2 and IC3 across all samples sizes, and hence the penalisation for a larger R is also smaller which, in two cases, leads to an overestimation of the number of factors, resulting in its poor performance with small sample sizes.

Table 5: True number of factors is selected (%)

T	25			50			100		
n	IC1	IC2	IC3	IC1	IC2	IC3	IC1	IC2	IC3
25	14.1	79.5	58	51	85.2	74.5	73.3	91.3	85.8
50	48.6	83.8	71.1	44.2	96.1	86.1	72.9	96.8	93.3
100	74.4	92.4	87.5	73.9	96.9	93.3	60.2	99.6	97.5

To gauge the likely impact of the factors not being known, estimation results with the number factors misspecified are provided in Appendix J in the Supplementary Material. These results illustrate cases in which the correct number of factors R^0 remains fixed at 3, and yet $R = 1$, $R = 5$ and $R = 10$ are inputted in estimation. In line with the result in Proposition 1, when the number of factors is underestimated ($R = 1$) large biases persist, while the estimator remains consistent with the number of factors overestimated ($R = 5$), even significantly so ($R = 10$), though overestimation can result in considerable inefficiency.

5.2 Application

As an empirical demonstration, the method is applied to study the determinants of economic growth, using a panel data set where several countries are observed over multiple time periods. It is natural to suppose that economic growth might be influenced by unobserved shocks, as well as observable regressors, and in this spirit Lu and Su (2016) estimate a model of economic growth controlling for unobserved factors. In that paper, the authors focus, in particular, on applying shrinkage methods to determine an unknown number of factors. Extending their work to include interaction is well motivated, since one might reasonably expect the growth rates of different countries to be interrelated. Yet in such cases it can be difficult to specify weights matrices a priori. Indeed Durlauf et al. (2009) remark: *"Spatial methods may yet have an important role to play in growth econometrics. However, when these methods are adapted from the spatial statistics literature, they raise the problem of identifying the appropriate notion of space .... countries are perhaps best thought of as occupying some general socio-economic-political space defined by a range of factors; spatial methods then require a means to identify their locations"*. The present method may provide insight into growth rate determination, where uncertainty in specifying cross-national interactions provides an example of the type of uncertainty which the present methodology seeks to address.

The data are obtained from Lu and Su (2016), with additional data on income classifications from the World Bank. The outcome y_{it} is the growth rate (Grth) in real GDP per capita for one of a cross-section of 108 countries observed between the years 1970 − 2005. The same 9 primitive exogenous covariates are used as in Lu and Su (2016), which include variables such as life expectancy, population growth, and consumption, investment and government expenditure shares. A series of weights matrices are specified based on grouping countries according to four Word Bank classifications: high income ($\boldsymbol{W}_1$), upper-middle income ($\boldsymbol{W}_2$), lower-middle income ($\boldsymbol{W}_3$) and low income ($\boldsymbol{W}_4$) economies, and reflect the more general notion of a socio-economic space remarked upon on by Durlauf et al. (2009). Each of these weights matrices are constructed by setting the (ij)-th element to 1 if country i and j share the same income classification, and setting it equal to zero otherwise, before then row normalising each of the matrices.

Table 6: Estimation results without interaction.

R	Young	Fert	Life	Popu	Invpri	Con	Gov	Inv	Open	Lag1	IC1	IC2	IC3
0	0	0	0	−0.462	0	0	0	0.099	0	0.161	-	-	-
t-stat	0	0	0	−8.030	0	0	0	17.394	0	10.386	-	-	-
1	0	0	0	−0.474	0	0	−0.051	0.118	0	0.137	3.463	3.497†	3.486
t-stat	0	0	0	−7.317	0	0	−4.224	18.504	0	8.855	-	-	-
2	0	0.444	0	−0.489	0	0	−0.238	0.228	0	0	3.446	3.513	3.491
t-stat	0	4.804	0	−5.186	0	0	−9.424	19.112	0	0	-	-	-
3	0	0	0	−0.061	0	0	−0.170	0.228	0	0	3.418†	3.518	3.486†
t-stat	0	0	0	−0.690	0	0	−8.644	19.821	0	0	-	-	-
5	0.090	−0.986	0	−0.479	0	0	−0.222	0.224	0	0	3.420	3.586	3.532
t-stat	6.043	−5.118	0	−4.737	0	0	−8.472	18.227	0	0	-	-	-

Table 6 reports bias corrected estimates $\hat{\boldsymbol{\theta}}^c$ in the absence of interaction.[12] For $R = 3$, these can be compared to the results for the AgLasso (which selects $R = 3$) given in Table 7 of Lu and Su (2016). In this case coefficient estimates and t-statistics are similar. Three variants of the information criterion given in (25) are reported, each of which differ only in the choice of penalty function ϱ_f.[13] In two out of three cases, the information criteria suggest that the number of factors is 3, matching the number suggested in Lu and Su (2016).

[12]Note that, in the absence of interaction, the quasi-maximum likelihood estimator reduces to the usual principal component least squares estimator (e.g., Bai, 2009).

[13]These variants are the same as those used in simulations.

Table 7(a): Estimation results with endogenous interaction and temporal lags.

R	$\boldsymbol{W}_1\times$Grth	$\boldsymbol{W}_2\times$Grth	$\boldsymbol{W}_3\times$Grth	$\boldsymbol{W}_4\times$Grth	Young	Fert	Life	Popu	Invpri	Con	Gov	Inv	Open
0	0.210	0.150	0	0.258	0	0	0	-0.492	0	0	0	0.090	0
t-stat	3.167	1.297	0	3.795	0	0	0	-8.460	0	0	0	15.115	0
1	0.295	0.289	-0.192	0.345	0	-0.070	0	-0.443	0	0	-0.050	0.111	0
t-stat	3.688	4.060	-1.475	5.141	0	-1.239	0	-4.977	0	0	-4.511	16.372	0
2	0.100	0	-0.325	0.207	0	0.355	0	-0.477	0	0	-0.237	0.218	0
t-stat	1.323	0	-2.383	2.808	0	3.958	0	-5.107	0	0	-9.493	17.823	0
3	0.195	0	-0.305	0.227	0	-0.001	0	-0.095	0	0	-0.188	0.215	0
t-stat	2.603	0	-2.277	3.099	0	-0.016	0	-0.953	0	0	-8.129	18.055	0
5	0	0	-0.280	0	0.090	-0.957	0	-0.483	0	0	-0.217	0.224	0
t-stat	0	0	-2.591	0	6.057	-4.977	0	-4.787	0	0	-8.375	12.26	0

Table 7(b): Estimation results with endogenous interaction and temporal lags.

R	Lag1	$\boldsymbol{W}_1\times$Lag1	$\boldsymbol{W}_2\times$Lag1	$\boldsymbol{W}_3\times$Lag1	$\boldsymbol{W}_4\times$Lag1	IC1	IC2	IC3
0	0.159	0	0	0	0	-	-	-
t-stat	10.279	0	0	0	0	-	-	-
1	0.129	0.172	0	0.400	0	3.445	3.479†	3.468†
t-stat	8.145	1.730	0	2.695	0	-	-	-
2	0.031	0	0	0.177	0	3.440	3.506	3.485
t-stat	1.965	0	0	1.137	0	-	-	-
3	0.033	0	0	0.233	0	3.410†	3.510	3.478
t-stat	2.070	0	0	1.572	0	-	-	-
5	0	0	0	0	0	3.418	3.584	3.530
t-stat	0	0	0	0	0	-	-	-

Tables 7(a) and 7(b) report estimation results when endogenous interaction and dynamic interaction is added. Government spending and investments shares in particular remain highly significant. However there is also evidence to suggest that there are significant endogenous spillovers, especially between high income and low income countries. The results indicate that amongst these two groups of countries, growth rates are interrelated with a positive spillover. In addition, there is evidence to suggest the presence of dynamic spillovers, these being positive, between lower-middle income countries. Note also that, when interactions are modelled, a lower number of factors is suggested by the information criteria. Several works have recently explored the possibility of latent group structures in panel data, for instance Bonhomme and Manresa (2015) and Su et al. (2016), and it is likely that in this case the weights matrices capture features in the data that would otherwise appear in the error term. The result of this is a lower number of factors being detected.

6 Conclusion

To conclude, this paper considers the estimation of a model of cross-section interaction, whose salient features are the use of multiple weights matrices and a factor structure in the error term. A penalised quasi-maximum likelihood estimator is proposed, in order to perform inference on network spillovers of various kinds, and its asymptotic properties are studied. A small Monte Carlo study reports good finite sample performance, and an empirical application studying the determinants of economic growth finds positive spillovers between the growth rates of high income and low income countries.

This work could be extended in several directions. For instance, one might consider possible endogeneity of the weights matrices as in Shi and Lee (2018) and Kuersteiner and Prucha (2020), or extend the use of weights matrices to the error term. Since they are observed, the possibility of time varying weights matrices might also be of interest.[14] With some modified assumptions, the consistency result in Proposition 1 could be extended quite readily to this case, though additional work would be required to characterise the asymptotic distribution. Another prospect might be to consider higher dimensional settings, for example, one might consider an entirely unknown weights matrix, modelled in this framework as a series of weights matrices containing a single unitary element. However, identification in this setting would need to be carefully studied since including parameters which increase too quickly with n, alongside the factor loadings, may present complications. As a final thought, it might also be natural to allow the number of factors to increase with sample size. When the number of interacting cross-sectional units increases, and more units in a network are observed, it might be expected that additional latent structures in the error term would lead to an increase in the rank of the factor term.

Appendix A. Proofs of Main Results

This appendix provides proofs of the main results. Before those proofs, a series of lemmas are stated. The proof of the lemmas are in the Supplementary Material. The following facts are used repeatedly (proofs can be found, for instance, in Moon and Weidner, 2017). Let $\boldsymbol{A}$ and $\boldsymbol{B}$ be two conformable matrices. Then $||\boldsymbol{A}||_2 \leq ||\boldsymbol{A}||_F \leq \sqrt{\text{rank}(\boldsymbol{A})}||\boldsymbol{A}||_2$, $||\boldsymbol{A}||_2 \leq \sqrt{||\boldsymbol{A}||_1||\boldsymbol{A}||_\infty}$ and $||\boldsymbol{AB}||_F \leq ||\boldsymbol{A}||_F||\boldsymbol{B}||_2 \leq ||\boldsymbol{A}||_F||\boldsymbol{B}||_F$. Let the i-th row of an $n \times m$ matrix $\boldsymbol{B}$ be denoted $(\boldsymbol{B})_{i\bullet}$, and the j-th column be denoted $(\boldsymbol{B})_{\bullet j}$. Then

[14] We thank an anonymous referee for this suggestion.

$\big(\sum_{j=1}^{m}||\boldsymbol{B}_{\boldsymbol{\cdot}j}||_2^2\big)^{\frac{1}{2}} = \big(\sum_{i=1}^{n}||\boldsymbol{B}_{i\boldsymbol{\cdot}}||_2^2\big)^{\frac{1}{2}} = ||\boldsymbol{B}||_F$. Finally, under Assumption 1.1, $||\boldsymbol{\varepsilon}||_2 = O_P(\sqrt{\min\{n,T\}})$ (see Latala, 2005).

Estimated factors and loadings: It is clear that the maximiser of $\mathcal{Q}(\boldsymbol{\theta},\boldsymbol{\Lambda})$ with respect to $\boldsymbol{\Lambda}$ is not unique, since for any $\boldsymbol{\Lambda}^* = \boldsymbol{\Lambda}\boldsymbol{H}$, with $\boldsymbol{H}$ being an $R\times R$ invertible matrix, $\boldsymbol{M}_{\boldsymbol{\Lambda}} = \boldsymbol{M}_{\boldsymbol{\Lambda}^*}$. In order to achieve uniqueness of the estimators of $\boldsymbol{\Lambda}$ and $\boldsymbol{F}$, the normalisations that $\frac{1}{n}\boldsymbol{\Lambda}'\boldsymbol{\Lambda} = \boldsymbol{I}_R$ and $\boldsymbol{F}'\boldsymbol{F}$ is a diagonal matrix are adopted, see for example Bai (2009).[15] Under these normalisations, define

$$\hat{\boldsymbol{\Lambda}}(\boldsymbol{\theta}) \coloneqq \underset{\boldsymbol{\Lambda}:\frac{1}{n}\boldsymbol{\Lambda}'\boldsymbol{\Lambda}=\boldsymbol{I}_R}{\arg\min}\left\{\frac{1}{nT}\sum_{t=1}^{T}\boldsymbol{e}_t'\boldsymbol{M}_{\boldsymbol{\Lambda}}\boldsymbol{e}_t\right\} = \underset{\boldsymbol{\Lambda}:\frac{1}{n}\boldsymbol{\Lambda}'\boldsymbol{\Lambda}=\boldsymbol{I}_R}{\arg\max}\left\{\frac{1}{n}\mathrm{tr}\left(\boldsymbol{\Lambda}'\frac{1}{nT}\sum_{t=1}^{T}\boldsymbol{e}_t\boldsymbol{e}_t'\boldsymbol{\Lambda}\right)\right\}. \quad \text{(A.1)}$$

It can be shown that the columns of $\hat{\boldsymbol{\Lambda}}(\boldsymbol{\theta})$ are equal to R orthonormal eigenvectors of the matrix $\frac{1}{nT}\sum_{t=1}^{T}\boldsymbol{e}_t'\boldsymbol{e}_t$ associated with the R largest eigenvalues and are unique, up to a column-wise sign change. Hereafter let $\hat{\boldsymbol{\Lambda}} \coloneqq \hat{\boldsymbol{\Lambda}}(\hat{\boldsymbol{\theta}})$.

Additional notation: For matrices $\boldsymbol{B}$ and $\boldsymbol{B}^*$, $\boldsymbol{B} = \boldsymbol{B}^* + \boldsymbol{O}_P(a_{nT})$ means that $||\boldsymbol{B}-\boldsymbol{B}^*||_2 = O_P(a_{nT})$. Similarly $\boldsymbol{B} = \boldsymbol{B}^* + \boldsymbol{o}_P(a_{nT})$ means that $||\boldsymbol{B}-\boldsymbol{B}^*||_2 = o_P(a_{nT})$. The elements of the matrices $\boldsymbol{\mathcal{X}}_\kappa^*$, $\boldsymbol{\mathcal{X}}_k$, $\boldsymbol{\mathcal{Z}}_p$, $\boldsymbol{\varepsilon}$, $\boldsymbol{\Lambda}$ and $\boldsymbol{F}$ are respectively denoted $x_{\kappa it}^*$, x_{kit}, z_{pit}, ε_{it}, λ_{ir} and f_{tr}. For any other $n\times m$ matrix $\boldsymbol{B}$, the (i,j)-th element is denoted $(\boldsymbol{B})_{ij}$. For brevity, $\hat{\sigma}^2 \coloneqq \hat{\sigma}^2(\hat{\boldsymbol{\theta}},\hat{\boldsymbol{\Lambda}})$, $\boldsymbol{Z}_t^*$ denotes the $n\times P$ matrix $(\boldsymbol{W}_1\boldsymbol{y}_t,...,\boldsymbol{W}_Q\boldsymbol{y}_t,\boldsymbol{X}_t)$. The notation $\varrho_p(\theta_p,\gamma_p,\zeta_p)$ is used to denote $\gamma_p\frac{1}{|\theta_p^\dagger|^{\zeta_p}}|\theta_p|$; that is, the penalty term relevant to θ_p. Finally, the l-th raw moment of some random variable s is denoted $\mathcal{M}_s^l$.

Lemma A.1. *For any positive definite matrix* $\boldsymbol{B}$, $\det(\boldsymbol{B})^{\frac{1}{n}} \leq \frac{1}{n}\mathrm{tr}(\boldsymbol{B})$, *with equality if and only if* $\boldsymbol{B} = c\boldsymbol{I}_n$ *for some* $c>0$.

Lemma A.2. *Under Assumptions* 1–2,

(i) $\boldsymbol{S}(\boldsymbol{\rho})\boldsymbol{S}^{-1} = \boldsymbol{I}_n + \sum_{q=1}^{Q}(\rho_q^0-\rho_q)\boldsymbol{G}_q$;

(ii) $||\boldsymbol{\mathcal{Z}}_p||_2 \leq ||\boldsymbol{\mathcal{Z}}_p||_F = O_P(\sqrt{nT})$ *for* $p=1,...,P$;

(iii) $||\boldsymbol{\Lambda}^0||_2 \leq ||\boldsymbol{\Lambda}^0||_F = O_P(\sqrt{n}), ||\boldsymbol{F}^0||_2 \leq ||\boldsymbol{F}^0||_F = O_P(\sqrt{T})$;

(iv) $(\sum_{p=1}^{P}||\boldsymbol{\mathcal{Z}}_p||_2^2)^{\frac{1}{2}}, (\sum_{t=1}^{T}||\boldsymbol{Z}_t||_2^2)^{\frac{1}{2}} = O_P(\sqrt{PnT})$;

(v) $\mathbb{E}\big[\sum_{p=1}^{P}\big(\mathrm{tr}(\boldsymbol{\mathcal{Z}}_p'\boldsymbol{S}(\boldsymbol{\rho})\boldsymbol{S}^{-1}\boldsymbol{\varepsilon})^2\big] = O(PnT)$;

[15] It is straightforward to see that such a rotation exists. For example, by the singular value decomposition, decompose $\boldsymbol{\Lambda}\boldsymbol{F}^\top = \boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^\top$. Let $\check{\boldsymbol{\Lambda}}$ be the R columns of $\sqrt{n}\boldsymbol{U}$ associated with the nonzero singular values and $\check{\boldsymbol{F}}^\top$ be the corresponding R rows of $\boldsymbol{S}\boldsymbol{V}^\top/\sqrt{n}$. As the columns of $\boldsymbol{U}$ and $\boldsymbol{V}$ are orthogonal, and $\boldsymbol{S}$ is diagonal, it follows that $\check{\boldsymbol{\Lambda}}^\top\check{\boldsymbol{\Lambda}}/n = \boldsymbol{I}_R$, $\check{\boldsymbol{F}}^\top\check{\boldsymbol{F}}$ is diagonal and $\check{\boldsymbol{\Lambda}}\check{\boldsymbol{F}}^\top = \boldsymbol{\Lambda}\boldsymbol{F}^\top$.

(vi) $||\boldsymbol{\varepsilon}||_F = O_P(\sqrt{nT})$;

(vii) $(\sum_{t=1}^T ||\boldsymbol{X}_t\boldsymbol{\beta}^0||_2^2)^{\frac{1}{2}} = O_P(\sqrt{nT})$;

(viii) $||\boldsymbol{S}(\hat{\boldsymbol{\rho}})\boldsymbol{S}^{-1} - \boldsymbol{I}_n||_2 = O_P(\sqrt{Q}||\boldsymbol{\theta}^0 - \hat{\boldsymbol{\theta}}||_2)$.

Lemma A.3. *Under Assumptions* 1–4,

(i) $(\frac{1}{nT}\sum_{t=1}^T ||\boldsymbol{Z}_t(\boldsymbol{\theta}^0 - \hat{\boldsymbol{\theta}})||_2^2)^{\frac{1}{2}} \leq c||\boldsymbol{\theta}^0 - \hat{\boldsymbol{\theta}}||_2$;

(ii) $\hat{\sigma}^{-2}(\hat{\boldsymbol{\theta}}, \boldsymbol{\Lambda}) = O_P(1)$.

Lemma A.4. *Under Assumptions* 1–6,

$$\boldsymbol{D}\sqrt{nT}(\hat{\boldsymbol{\theta}}_{(1)} - \boldsymbol{\theta}^0_{(1)}) = \frac{1}{\sigma_0^2}\frac{1}{\sqrt{nT}}\boldsymbol{\mathcal{Z}}'_{(1)}(\boldsymbol{M}_{\boldsymbol{F}^0} \otimes \boldsymbol{M}_{\boldsymbol{\Lambda}^0})\mathrm{vec}(\boldsymbol{\varepsilon}) + \frac{1}{\sigma_0^2}\frac{1}{\sqrt{nT}}\begin{pmatrix} \mathrm{tr}\left((\boldsymbol{G}_1^*\boldsymbol{\varepsilon})'\boldsymbol{M}_{\boldsymbol{\Lambda}^0}\boldsymbol{\varepsilon}\boldsymbol{M}_{\boldsymbol{F}^0}\right) \\ \vdots \\ \mathrm{tr}\left((\boldsymbol{G}_{Q^0}^*\boldsymbol{\varepsilon})'\boldsymbol{M}_{\boldsymbol{\Lambda}^0}\boldsymbol{\varepsilon}\boldsymbol{M}_{\boldsymbol{F}^0}\right) \\ \boldsymbol{0}_{K^0\times 1} \end{pmatrix} + \boldsymbol{o}_P(1),$$

where the matrix $\boldsymbol{D}$ is defined in equation (16).

Lemma A.5. *Under Assumptions* 1–6,

(i) $||\boldsymbol{D}^{-1} - \hat{\boldsymbol{D}}^{-1}||_2 = O_P((Q^0)^{1.5}P^0||\boldsymbol{\theta}^0 - \hat{\boldsymbol{\theta}}||_2) + O_P\left(\frac{Q^0P^0}{\sqrt{\min\{n,T\}}}\right)$;

(ii) $\mathbb{E}\left[\sum_{q=1}^{Q^0}\left(\mathrm{tr}((\boldsymbol{G}_q^*\boldsymbol{\varepsilon})'\boldsymbol{P}_{\boldsymbol{\Lambda}^0}\boldsymbol{\varepsilon}) - \sigma_0^2T\mathrm{tr}(\boldsymbol{P}_{\boldsymbol{\Lambda}^0}\boldsymbol{G}_q^*)\right)^2\right] = O(Q^0T)$;

(iii) $\mathbb{E}\left[\sum_{q=1}^{Q^0}\left(\mathrm{tr}((\boldsymbol{G}_q^*\boldsymbol{\varepsilon})'\boldsymbol{P}_{\boldsymbol{\Lambda}^0}\boldsymbol{\varepsilon}\boldsymbol{P}_{\boldsymbol{F}^0}) - \sigma_0^2R^0\mathrm{tr}(\boldsymbol{P}_{\boldsymbol{\Lambda}^0}\boldsymbol{G}_q^*)\right)^2\right] = O(Q^0)$;

(iv) $\mathbb{E}\left[\sum_{q=1}^{Q^0}\left(\mathrm{tr}((\boldsymbol{G}_q^*\boldsymbol{\varepsilon})'\boldsymbol{\varepsilon}\boldsymbol{P}_{\boldsymbol{F}^0}) - \sigma_0^2R^0\mathrm{tr}(\boldsymbol{G}_q^*)\right)^2\right] = O(Q^0n)$;

(v)
$$\frac{1}{\sigma_0^2}\frac{1}{\sqrt{nT}}\begin{pmatrix} \mathrm{tr}\left((\mathbb{Z}_1 - \bar{\mathbb{Z}}_1)'(\boldsymbol{P}_{\boldsymbol{\Lambda}^0}\boldsymbol{\varepsilon} + \boldsymbol{M}_{\boldsymbol{\Lambda}^0}\boldsymbol{\varepsilon}\boldsymbol{P}_{\boldsymbol{F}^0})\right) \\ \vdots \\ \mathrm{tr}\left((\mathbb{Z}_{P^0} - \bar{\mathbb{Z}}_{P^0})'(\boldsymbol{P}_{\boldsymbol{\Lambda}^0}\boldsymbol{\varepsilon} + \boldsymbol{M}_{\boldsymbol{\Lambda}^0}\boldsymbol{\varepsilon}\boldsymbol{P}_{\boldsymbol{F}^0})\right) \end{pmatrix} = \begin{pmatrix} \mathbb{b}^{(2)} \\ \boldsymbol{0}_{K^{*0}\times 1} \\ \mathbb{b}^{(3)} \end{pmatrix} + \boldsymbol{o}_P(1).$$

Lemma A.6. *Under Assumptions* 1–7, $\frac{1}{\sqrt{nT}}\frac{1}{\sigma_0^2}\left(\mathbb{S}(\mathbb{D}+\mathbb{V})\mathbb{S}'\right)^{-\frac{1}{2}}\mathbb{S}\boldsymbol{c} \xrightarrow{d} \mathcal{N}(\mathbf{0}_{L\times L}, \boldsymbol{I}_L)$, *where* $\boldsymbol{c} := \mathbb{Z}_{(1)}'\mathrm{vec}(\boldsymbol{\varepsilon}) + (\mathrm{tr}(\boldsymbol{\varepsilon}'\boldsymbol{G}_1^*\boldsymbol{\varepsilon}), ..., \mathrm{tr}(\boldsymbol{\varepsilon}'\boldsymbol{G}_{Q^0}^*\boldsymbol{\varepsilon}), \mathbf{0}_{1\times K^0})'$ *and* $\mathbb{S}, \mathbb{D}$ *and* $\mathbb{V}$ *are defined in Assumptions 7.1 and 7.2.*

Proof of Proposition 1. Here only a sketch of the proof is provided. A detailed version can be found in Appendix D of the Supplementary Material. Due to the diverging number of parameters, this consistency proof follows the approach taken by Fan and Peng (2004). Let $\boldsymbol{u}$ be a $P \times 1$ vector, and $\mathcal{T}_{nT}(\boldsymbol{\theta}^0) := \{\boldsymbol{\theta}^0 + a_{nT}\boldsymbol{u} : ||\boldsymbol{u}||_2 \leq d\}$ be a closed ball centred at $\boldsymbol{\theta}^0$ with radius $a_{nT}d$. The objective is to show that for any $\epsilon > 0$ and sufficiently large n, T, there exists a large enough d such that

$$\Pr\left(\sup_{||\boldsymbol{u}||_2 = d} \mathcal{Q}(\boldsymbol{\theta}^0 + a_{nT}\boldsymbol{u}) < \mathcal{Q}(\boldsymbol{\theta}^0)\right) \geq 1 - \epsilon. \tag{A.2}$$

Because $\mathcal{T}_{nT}(\boldsymbol{\theta}^0)$ is compact, (A.2) implies that, as $n, T \to \infty$, there exists a local maximiser in the interior of $\mathcal{T}_{nT}(\boldsymbol{\theta}^0)$ with probability approaching 1, call this $\hat{\boldsymbol{\theta}}_L$, such that $||\hat{\boldsymbol{\theta}}_L - \boldsymbol{\theta}^0||_2 < a_{nT}||\boldsymbol{u}||_2 = O_P(a_{nT})$. First, however, the existence of an a_{nT}-consistent local maximiser of the unpenalised objective function, denoted $\tilde{\boldsymbol{\theta}}_L$, is demonstrated. This follows from showing

$$\Pr\left(\sup_{||\boldsymbol{u}||_2 = d} \mathcal{L}(\boldsymbol{\theta}^0 + a_{nT}\boldsymbol{u}) < \mathcal{L}(\boldsymbol{\theta}^0)\right) \geq 1 - \epsilon, \tag{A.3}$$

where the average concentrated quasi-likelihood is given by

$$\mathcal{L}(\boldsymbol{\theta}) := \sup_{\boldsymbol{\Lambda}\in\mathbb{R}^{n\times R}} \left\{\frac{1}{n}\log(\det(\boldsymbol{S}(\boldsymbol{\rho}))) - \frac{1}{2}\log\left(\hat{\sigma}^2(\boldsymbol{\theta}, \boldsymbol{\Lambda})\right)\right\}. \tag{A.4}$$

First, consider $\mathcal{L}(\boldsymbol{\theta}^0)$. A lower bound for this, denoted $\underline{\mathcal{L}}(\boldsymbol{\theta}^0)$, can be established by substituting in the true DGP, and using Assumptions 1.1 and 1.2,

$$\underline{\mathcal{L}}(\boldsymbol{\theta}^0) := \frac{1}{n}\log(\det(\boldsymbol{S})) - \frac{1}{2}\log\left(\sigma_0^2 + O_P\left(\frac{1}{\min\{n,T\}}\right)\right) \leq \mathcal{L}(\boldsymbol{\theta}^0). \tag{A.5}$$

Next, consider $\sup_{||\boldsymbol{u}||_2=d}\{\mathcal{L}(\boldsymbol{\theta}^0 + a_{nT}\boldsymbol{u})\}$. Let $\ddot{\boldsymbol{u}} := \arg\sup_{||\boldsymbol{u}||_2=d}\{\mathcal{L}(\boldsymbol{\theta}^0 + a_{nT}\boldsymbol{u})\}$. Partition $\ddot{\boldsymbol{u}}$ into two vectors, $\ddot{\boldsymbol{u}}_\rho$ and $\ddot{\boldsymbol{u}}_\beta$ with the former being $Q \times 1$ and the latter being $K \times 1$. Define $\ddot{\boldsymbol{\theta}} := \boldsymbol{\theta}^0 + a_{nT}\ddot{\boldsymbol{u}}$, $\ddot{\boldsymbol{\rho}} := \boldsymbol{\rho}^0 + a_{nT}\ddot{\boldsymbol{u}}_\rho$ and $\ddot{\boldsymbol{\beta}} := \boldsymbol{\beta}^0 + a_{nT}\ddot{\boldsymbol{u}}_\beta$. One then has

$$\mathcal{L}(\ddot{\boldsymbol{\theta}}) = \frac{1}{n}\log(\det(\boldsymbol{S}(\ddot{\boldsymbol{\rho}}))) - \frac{1}{2}\log\left(\inf_{\boldsymbol{\Lambda}\in\mathbb{R}^{n\times R}} \hat{\sigma}^2(\ddot{\boldsymbol{\theta}}, \boldsymbol{\Lambda})\right). \tag{A.6}$$

Using Lemmas A.2(i), A.2(iv), A.2(v), A.3(i) and Assumption 4.2, an upper bound for $\mathcal{L}(\ddot{\boldsymbol{\theta}})$, denoted $\bar{\mathcal{L}}(\ddot{\boldsymbol{\theta}})$, can also be established,

$$\mathcal{L}(\ddot{\boldsymbol{\theta}}) \leq \frac{1}{n}\log(\det(\boldsymbol{S}(\ddot{\boldsymbol{\rho}}))) - \frac{1}{2}\log\left(a_{nT}^2 c||\ddot{\boldsymbol{u}}||_2^2 + a_{nT}||\ddot{\boldsymbol{u}}||_2 O_P\left(\sqrt{\frac{P}{nT}}\right) + a_{nT}||\ddot{\boldsymbol{u}}||_2 O_P\left(\sqrt{\frac{P}{\min\{n,T\}}}\right)\right.$$

$$+\frac{\sigma_0^2}{n}\operatorname{tr}((\boldsymbol{S}(\ddot{\boldsymbol{\rho}})\boldsymbol{S}^{-1})'\boldsymbol{S}(\ddot{\boldsymbol{\rho}})\boldsymbol{S}^{-1}) + O_P\left(\frac{1}{\sqrt{nT}}\right) + O_P\left(\frac{1}{\min\{n,T\}}\right)\Bigg) =: \bar{\mathcal{L}}(\ddot{\boldsymbol{\theta}}). \tag{A.7}$$

Now, equation (A.3) is satisfied if $\mathcal{L}(\ddot{\boldsymbol{\theta}}) - \mathcal{L}(\boldsymbol{\theta}^0) < 0$ as $n, T \to \infty$. Since $\underline{\mathcal{L}}(\boldsymbol{\theta}^0) \leq \mathcal{L}(\boldsymbol{\theta}^0)$ and $\mathcal{L}(\ddot{\boldsymbol{\theta}}) \leq \bar{\mathcal{L}}(\ddot{\boldsymbol{\theta}})$, then equation (A.3) is equivalently satisfied if $\bar{\mathcal{L}}(\ddot{\boldsymbol{\theta}}) - \underline{\mathcal{L}}(\boldsymbol{\theta}^0) \leq 0$ as $n, T \to \infty$. Using the fact that

$$\begin{aligned}
&\frac{1}{n}\log(\det(\boldsymbol{S}(\ddot{\boldsymbol{\rho}}))) - \frac{1}{n}\det(\boldsymbol{S}) + \frac{1}{2}\log\left(\sigma_0^2 + O_P\left(\frac{1}{\min\{n,T\}}\right)\right) \\
&= \frac{1}{2}\log\left(\sigma_0^2\det((\boldsymbol{S}(\ddot{\boldsymbol{\rho}})\boldsymbol{S}^{-1})'\boldsymbol{S}(\ddot{\boldsymbol{\rho}})\boldsymbol{S}^{-1})^{\frac{1}{n}} + O_P\left(\frac{1}{\min\{n,T\}}\right)\right),
\end{aligned} \tag{A.8}$$

and ignoring dominated terms,

$$\begin{aligned}
\bar{\mathcal{L}}(\ddot{\boldsymbol{\theta}}) - \underline{\mathcal{L}}(\boldsymbol{\theta}^0) \leq &\frac{1}{2}\log\left(\sigma_0^2\det((\boldsymbol{S}(\ddot{\boldsymbol{\rho}})\boldsymbol{S}^{-1})'\boldsymbol{S}(\ddot{\boldsymbol{\rho}})\boldsymbol{S}^{-1})^{\frac{1}{n}} + O_P\left(\frac{1}{\min\{n,T\}}\right)\right) \\
&-\frac{1}{2}\log\left(a_{nT}^2 c||\ddot{\boldsymbol{u}}||_2^2 + \operatorname{tr}((\boldsymbol{S}(\ddot{\boldsymbol{\rho}})\boldsymbol{S}^{-1})'\boldsymbol{S}(\ddot{\boldsymbol{\rho}})\boldsymbol{S}^{-1})\right).
\end{aligned} \tag{A.9}$$

Recall that $c > 0$, and note that by Lemma A.1, $\sigma_0^2\det((\boldsymbol{S}(\ddot{\boldsymbol{\rho}})\boldsymbol{S}^{-1})'\boldsymbol{S}(\ddot{\boldsymbol{\rho}})\boldsymbol{S}^{-1})^{\frac{1}{n}} \leq \frac{\sigma_0^2}{n}\operatorname{tr}((\boldsymbol{S}(\ddot{\boldsymbol{\rho}})\boldsymbol{S}^{-1})'\boldsymbol{S}(\ddot{\boldsymbol{\rho}})\boldsymbol{S}^{-1})$. Then, by the monotonicity of the logarithm, as $n, T \to \infty$ and for sufficiently large d, the right-hand side of (A.9) is strictly negative. Therefore with probability approaching 1 there exists an a_{nT}-consistent local maximiser $\tilde{\boldsymbol{\theta}}_L$ of the unpenalised average likelihood function $\mathcal{L}(\boldsymbol{\theta})$. With the existence of a local maximiser established, consider next a global maximiser $\tilde{\boldsymbol{\theta}} := \arg\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \mathcal{L}(\boldsymbol{\theta})$. From (A.7), an upper bound for $\mathcal{L}(\tilde{\boldsymbol{\theta}})$ is given by

$$\begin{aligned}
\mathcal{L}(\tilde{\boldsymbol{\theta}}) \leq &\frac{1}{n}\log(\det(\boldsymbol{S}(\hat{\boldsymbol{\rho}}))) - \frac{1}{2}\log\Big(c||\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^0||_2^2 + O_P(a_{nT})||\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^0|| + O_P(a_{nT}^2) \\
&+\frac{\sigma_0^2}{n}\operatorname{tr}((\boldsymbol{S}(\hat{\boldsymbol{\rho}})\boldsymbol{S}^{-1})'\boldsymbol{S}(\hat{\boldsymbol{\rho}})\boldsymbol{S}^{-1})\Big) =: \bar{\mathcal{L}}(\tilde{\boldsymbol{\theta}}).
\end{aligned} \tag{A.10}$$

Since $\tilde{\boldsymbol{\theta}}$ is a global maximiser $\mathcal{L}(\boldsymbol{\theta}^0) \leq \mathcal{L}(\tilde{\boldsymbol{\theta}})$ and therefore $\underline{\mathcal{L}}(\boldsymbol{\theta}^0) \leq \bar{\mathcal{L}}(\tilde{\boldsymbol{\theta}})$. Combining this with (A.5) and (A.10) gives

$$\begin{aligned}
&\frac{1}{n}\log(\det(\boldsymbol{S})) - \frac{1}{2}\log\left(\sigma_0^2 + O_P\left(\frac{1}{\min\{n,T\}}\right)\right) \\
&\leq \frac{1}{n}\log(\det(\boldsymbol{S}(\hat{\boldsymbol{\rho}}))) - \frac{1}{2}\log\Big(c||\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^0||_2^2 + O_P(a_{nT})||\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^0|| + O_P(a_{nT}^2) \\
&\quad+\frac{\sigma_0^2}{n}\operatorname{tr}((\boldsymbol{S}(\hat{\boldsymbol{\rho}})\boldsymbol{S}^{-1})'\boldsymbol{S}(\hat{\boldsymbol{\rho}})\boldsymbol{S}^{-1})\Big).
\end{aligned} \tag{A.11}$$

Multiplying both sides of (A.11) by -2, exponentiating, and then noticing that, by Lemma A.1, $\sigma_0^2\det((\boldsymbol{S}(\hat{\boldsymbol{\rho}})\boldsymbol{S}^{-1})'\boldsymbol{S}(\hat{\boldsymbol{\rho}})\boldsymbol{S}^{-1})^{\frac{1}{n}} \leq \frac{\sigma_0^2}{n}\operatorname{tr}((\boldsymbol{S}(\hat{\boldsymbol{\rho}})\boldsymbol{S}^{-1})'\boldsymbol{S}(\hat{\boldsymbol{\rho}})\boldsymbol{S}^{-1})$, results in

$$0 \geq c||\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^0||_2^2 + O_P(a_{nT})||\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^0|| + O_P(a_{nT}^2). \tag{A.12}$$

Completing the square, $0 \geq (\sqrt{c}||\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^0||_2 + O_P(a_{nT}))^2 + O_P(a_{nT}^2)$ whereby it follows that $||\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^0||_2 = O_P(a_{nT})$. Combined with the existence of a local maximiser, this result demonstrates the existence of an a_{nT}-consistent global maximiser of the unpenalised likelihood. Moving to the penalised average likelihood $\mathcal{Q}(\boldsymbol{\theta})$, and using the same notation $\ddot{\boldsymbol{u}}$ to denote $\ddot{\boldsymbol{u}} := \arg\sup_{||\boldsymbol{u}||_2=d}\{\mathcal{Q}(\boldsymbol{\theta}^0 + a_{nT}\boldsymbol{u})\}$,

$$\mathcal{Q}(\ddot{\boldsymbol{\theta}}) - \mathcal{Q}(\boldsymbol{\theta}^0) \leq \mathcal{L}(\ddot{\boldsymbol{\theta}}) - \mathcal{L}(\boldsymbol{\theta}^0) - \sum_{p=1}^{P_0} \varrho_p(\ddot{\theta}_p, \gamma_p, \zeta_p) + \sum_{p=1}^{P_0} \varrho_p(\theta_p^0, \gamma_p, \zeta_p). \tag{A.13}$$

Using Assumptions 3.1 and 3.2, it can be shown that

$$-\sum_{p=1}^{P_0} \varrho_p(\ddot{\theta}_p, \gamma_p, \zeta_p) + \sum_{p=1}^{P_0} \varrho_p(\theta_p^0, \gamma_p, \zeta_p) = O_P(a_{nT}^2)||\ddot{\boldsymbol{u}}||_2, \tag{A.14}$$

whereby $\mathcal{Q}(\ddot{\boldsymbol{\theta}}) - \mathcal{Q}(\boldsymbol{\theta}^0) = \mathcal{L}(\ddot{\boldsymbol{\theta}}) - \mathcal{L}(\boldsymbol{\theta}^0) + O_P(a_{nT}^2)||\ddot{\boldsymbol{u}}||_2$. It has already been established that, for large enough d, $\mathcal{L}(\ddot{\boldsymbol{\theta}}) - \mathcal{L}(\boldsymbol{\theta}^0)$ is strictly negative as $n, T \to \infty$, therefore it follows from equation (A.2) that there exists a local maximiser of the average penalised likelihood, $\hat{\boldsymbol{\theta}}_L$, in the interior of the ball $\{\boldsymbol{\theta}^0 + a_{nT}\ddot{\boldsymbol{u}} : ||\ddot{\boldsymbol{u}}||_2 \leq d\}$ such that $||\hat{\boldsymbol{\theta}}_L - \boldsymbol{\theta}^0||_2 = O_P(a_{nT})$. By the same steps used to derive (A.12), it can be shown that a global maximiser $\hat{\boldsymbol{\theta}}$ of the unpenalised likelihood must be a_{nT}-consistent whereby both the existence and a_{nT}-consistency of the global maximum of the penalised likelihood is established. □

Proof of Proposition 2. The penalised QMLE $\hat{\boldsymbol{\theta}}$, the existence of which is established in Proposition 1, must solve the first order condition

$$\frac{\partial \mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\Lambda})}{\partial \boldsymbol{\theta}} = \frac{\partial \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\Lambda})}{\partial \boldsymbol{\theta}} - \frac{\partial \varrho(\boldsymbol{\theta}, \gamma, \zeta)}{\partial \boldsymbol{\theta}} = \mathbf{0}_{P \times 1}, \tag{A.15}$$

where

$$\frac{\partial \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\Lambda})}{\partial \boldsymbol{\theta}} = \begin{pmatrix} -\frac{1}{n}\mathrm{tr}(\boldsymbol{G}_1(\boldsymbol{\rho})) + \frac{1}{\hat{\sigma}^2(\boldsymbol{\theta}, \boldsymbol{\Lambda})}\frac{1}{nT}\sum_{t=1}^T (\boldsymbol{W}_1\boldsymbol{y}_t)'\boldsymbol{M}_{\boldsymbol{\Lambda}}(\boldsymbol{S}(\boldsymbol{\rho})\boldsymbol{y}_t - \boldsymbol{X}_t\boldsymbol{\beta}) \\ \vdots \\ -\frac{1}{n}\mathrm{tr}(\boldsymbol{G}_Q(\boldsymbol{\rho})) + \frac{1}{\hat{\sigma}^2(\boldsymbol{\theta}, \boldsymbol{\Lambda})}\frac{1}{nT}\sum_{t=1}^T (\boldsymbol{W}_q\boldsymbol{y}_t)'\boldsymbol{M}_{\boldsymbol{\Lambda}}(\boldsymbol{S}(\boldsymbol{\rho})\boldsymbol{y}_t - \boldsymbol{X}_t\boldsymbol{\beta}) \\ \frac{1}{\hat{\sigma}^2(\boldsymbol{\theta}, \boldsymbol{\Lambda})}\frac{1}{nT}\sum_{t=1}^T \boldsymbol{x}_{1t}'\boldsymbol{M}_{\boldsymbol{\Lambda}}(\boldsymbol{S}(\boldsymbol{\rho})\boldsymbol{y}_t - \boldsymbol{X}_t\boldsymbol{\beta}) \\ \vdots \\ \frac{1}{\hat{\sigma}^2(\boldsymbol{\theta}, \boldsymbol{\Lambda})}\frac{1}{nT}\sum_{t=1}^T \boldsymbol{x}_{Kt}'\boldsymbol{M}_{\boldsymbol{\Lambda}}(\boldsymbol{S}(\boldsymbol{\rho})\boldsymbol{y}_t - \boldsymbol{X}_t\boldsymbol{\beta}) \end{pmatrix}. \tag{A.16}$$

In the following it is shown that, as $n, T \to \infty$, this first order condition cannot hold unless the estimators of those θ_p which have a true value of zero also take a value of exactly zero with probability approaching 1. To reach a contradiction, suppose that there is some p, call this p^*, for which $\theta_p^0 = 0$ yet $\Pr(\hat{\theta}_p = 0)$ does not go to zero as $n, T \to \infty$. It is first shown that $\frac{\partial \mathcal{L}(\hat{\theta}_{p^*}, \mathbf{\Lambda})}{\partial \theta_p} = O_P(1)$, i.e., the first order condition evaluated at $\hat{\theta}_{p^*}$ is not explosive in probability. Since θ_{p^*} could be some ρ_q or β_k, both cases are examined in turn. Consider first the case where θ_{p^*} is some ρ_q. Substituting in the true data generating process, the element of $\frac{\partial \mathcal{L}(\hat{\boldsymbol{\theta}}, \mathbf{\Lambda})}{\partial \boldsymbol{\theta}}$ relating to ρ_q is equal to

$$
\begin{aligned}
& -\frac{1}{n}\mathrm{tr}(\boldsymbol{G}_q(\boldsymbol{\rho})) + \frac{1}{\hat{\sigma}^2(\boldsymbol{\theta}, \mathbf{\Lambda})}\frac{1}{nT}\sum_{t=1}^{T}(\boldsymbol{W}_q\boldsymbol{y}_t)'\boldsymbol{M}_{\mathbf{\Lambda}}(\boldsymbol{S}(\boldsymbol{\rho})\boldsymbol{y}_t - \boldsymbol{X}_t\boldsymbol{\beta}) \\
= & -\frac{1}{n}\mathrm{tr}(\boldsymbol{G}_q(\hat{\boldsymbol{\rho}})) + \frac{1}{\hat{\sigma}^2(\hat{\boldsymbol{\theta}}, \mathbf{\Lambda})}\frac{1}{nT}\sum_{t=1}^{T}(\boldsymbol{G}_q\boldsymbol{X}_t\boldsymbol{\beta}^0)'\boldsymbol{M}_{\mathbf{\Lambda}}\boldsymbol{Z}_t(\boldsymbol{\theta}^0 - \hat{\boldsymbol{\theta}}) \\
& + \frac{1}{\hat{\sigma}^2(\hat{\boldsymbol{\theta}}, \mathbf{\Lambda})}\frac{1}{nT}\sum_{t=1}^{T}(\boldsymbol{G}_q\boldsymbol{X}_t\boldsymbol{\beta}^0)'\boldsymbol{M}_{\mathbf{\Lambda}}\boldsymbol{S}(\hat{\boldsymbol{\rho}})\boldsymbol{S}^{-1}\mathbf{\Lambda}^0\boldsymbol{f}_t^0 + \frac{1}{\hat{\sigma}^2(\hat{\boldsymbol{\theta}}, \mathbf{\Lambda})}\frac{1}{nT}\sum_{t=1}^{T}(\boldsymbol{G}_q\boldsymbol{X}_t\boldsymbol{\beta}^0)'\boldsymbol{M}_{\mathbf{\Lambda}}\boldsymbol{S}(\hat{\boldsymbol{\rho}})\boldsymbol{S}^{-1}\boldsymbol{\varepsilon}_t \\
& + \frac{1}{\hat{\sigma}^2(\hat{\boldsymbol{\theta}}, \mathbf{\Lambda})}\frac{1}{nT}\sum_{t=1}^{T}(\boldsymbol{G}_q\mathbf{\Lambda}^0\boldsymbol{f}_t^0)'\boldsymbol{M}_{\mathbf{\Lambda}}\boldsymbol{Z}_t(\boldsymbol{\theta}^0 - \hat{\boldsymbol{\theta}}) + \frac{1}{\hat{\sigma}^2(\hat{\boldsymbol{\theta}}, \mathbf{\Lambda})}\frac{1}{nT}\sum_{t=1}^{T}(\boldsymbol{G}_q\mathbf{\Lambda}^0\boldsymbol{f}_t^0)'\boldsymbol{M}_{\mathbf{\Lambda}}\boldsymbol{S}(\hat{\boldsymbol{\rho}})\boldsymbol{S}^{-1}\mathbf{\Lambda}^0\boldsymbol{f}_t^0 \\
& + \frac{1}{\hat{\sigma}^2(\hat{\boldsymbol{\theta}}, \mathbf{\Lambda})}\frac{1}{nT}\sum_{t=1}^{T}(\boldsymbol{G}_q\mathbf{\Lambda}^0\boldsymbol{f}_t^0)'\boldsymbol{M}_{\mathbf{\Lambda}}\boldsymbol{S}(\hat{\boldsymbol{\rho}})\boldsymbol{S}^{-1}\boldsymbol{\varepsilon}_t + \frac{1}{\hat{\sigma}^2(\hat{\boldsymbol{\theta}}, \mathbf{\Lambda})}\frac{1}{nT}\sum_{t=1}^{T}(\boldsymbol{G}_q\boldsymbol{\varepsilon}_t)'\boldsymbol{M}_{\mathbf{\Lambda}}\boldsymbol{Z}_t(\boldsymbol{\theta}^0 - \hat{\boldsymbol{\theta}}) \\
& + \frac{1}{\hat{\sigma}^2(\hat{\boldsymbol{\theta}}, \mathbf{\Lambda})}\frac{1}{nT}\sum_{t=1}^{T}(\boldsymbol{G}_q\boldsymbol{\varepsilon}_t)'\boldsymbol{M}_{\mathbf{\Lambda}}\boldsymbol{S}(\hat{\boldsymbol{\rho}})\boldsymbol{S}^{-1}\mathbf{\Lambda}^0\boldsymbol{f}_t^0 + \frac{1}{\hat{\sigma}^2(\hat{\boldsymbol{\theta}}, \mathbf{\Lambda})}\frac{1}{nT}\sum_{t=1}^{T}(\boldsymbol{G}_q\boldsymbol{\varepsilon}_t)'\boldsymbol{M}_{\mathbf{\Lambda}}\boldsymbol{S}(\hat{\boldsymbol{\rho}})\boldsymbol{S}^{-1}\boldsymbol{\varepsilon}_t \\
=: & \; k_1 + \ldots + k_{10}. \qquad \text{(A.17)}
\end{aligned}
$$

Since $\boldsymbol{G}_q(\boldsymbol{\rho})$ is UB, terms $k_5, \ldots, k_{10}$ are $O_P(1)$ by the same arguments as for their counterparts in the proof of Lemma A.3(ii) (terms $l_2, \ldots, l_6$; see Supplementary Material), and using the result in that lemma (whereby $1/\hat{\sigma}^2(\hat{\boldsymbol{\theta}}, \mathbf{\Lambda}) = O_P(1)$). Since the rank of $\boldsymbol{G}_q(\boldsymbol{\rho})$ can be no more than n, using $|\mathrm{tr}(\boldsymbol{B})| \leq \mathrm{rank}(\boldsymbol{B})||\boldsymbol{B}||_2$ for some square matrix $\boldsymbol{B}$ (Moon and Weidner, 2017, Lemma S.4.1(v)), and that $\boldsymbol{S}^{-1}(\boldsymbol{\rho})$ and $\boldsymbol{W}_q$ are UB, one has

$$
|k_1| = \frac{1}{n}|\mathrm{tr}(\boldsymbol{G}_q(\hat{\boldsymbol{\rho}}))| \leq ||\boldsymbol{G}_q(\hat{\boldsymbol{\rho}})||_2 \leq ||\boldsymbol{S}^{-1}(\hat{\boldsymbol{\rho}})||_2||\boldsymbol{W}_q||_2 = O_P(1). \qquad \text{(A.18)}
$$

Using Lemmas A.2(vii), A.3(i) and A.3(ii), as well as Proposition 1, yields

$$
\begin{aligned}
|k_2| &\leq \frac{1}{\sqrt{nT}}\frac{1}{\hat{\sigma}^2(\hat{\boldsymbol{\theta}}, \mathbf{\Lambda})}||\boldsymbol{G}_q||_2||\boldsymbol{M}_{\mathbf{\Lambda}}||_2\left(\sum_{t=1}^{T}||\boldsymbol{X}_t\boldsymbol{\beta}^0||_2^2\right)^{\frac{1}{2}}\left(\frac{1}{nT}\sum_{t=1}^{T}||\boldsymbol{Z}_t(\boldsymbol{\theta}^0 - \hat{\boldsymbol{\theta}})||_2^2\right)^{\frac{1}{2}} \\
&= \frac{1}{\sqrt{nT}}O_P(\sqrt{nT})O_P(a_{nT}) = O_P(1). \qquad \text{(A.19)}
\end{aligned}
$$

The remaining terms, k_3 and k_4, follow similarly and are $O_P(1)$ using Lemmas A.2(iii), A.2(vi) A.2(vii) and A.3(ii). Next consider the case where θ_{p^*} is some β_k. The element of $\frac{\partial \mathcal{L}(\hat{\boldsymbol{\theta}},\boldsymbol{\Lambda})}{\partial \boldsymbol{\theta}}$ corresponding to β_k is

$$\begin{aligned}
&\frac{1}{\hat{\sigma}^2(\hat{\boldsymbol{\theta}},\boldsymbol{\Lambda})}\frac{1}{nT}\sum_{t=1}^T \boldsymbol{x}_{kt}'\boldsymbol{M}_{\boldsymbol{\Lambda}}(\boldsymbol{S}(\hat{\boldsymbol{\rho}})\boldsymbol{S}^{-1}(\boldsymbol{X}_t\boldsymbol{\beta}^0+\boldsymbol{\Lambda}^0\boldsymbol{f}_t^0+\boldsymbol{\varepsilon}_t)-\boldsymbol{X}_t\hat{\boldsymbol{\beta}}) \\
&= \frac{1}{\hat{\sigma}^2(\hat{\boldsymbol{\theta}},\boldsymbol{\Lambda})}\frac{1}{nT}\sum_{t=1}^T \boldsymbol{x}_{kt}'\boldsymbol{M}_{\boldsymbol{\Lambda}}\boldsymbol{Z}_t(\boldsymbol{\theta}^0-\boldsymbol{\theta}) + \frac{1}{\hat{\sigma}^2(\hat{\boldsymbol{\theta}},\boldsymbol{\Lambda})}\frac{1}{nT}\sum_{t=1}^T \boldsymbol{x}_{kt}'\boldsymbol{M}_{\boldsymbol{\Lambda}}\boldsymbol{S}(\hat{\boldsymbol{\rho}})\boldsymbol{S}^{-1}\boldsymbol{\Lambda}^0\boldsymbol{f}_t^0 \\
&\quad + \frac{1}{\hat{\sigma}^2(\hat{\boldsymbol{\theta}},\boldsymbol{\Lambda})}\frac{1}{nT}\sum_{t=1}^T \boldsymbol{x}_{kt}'\boldsymbol{M}_{\boldsymbol{\Lambda}}\boldsymbol{S}(\hat{\boldsymbol{\rho}})\boldsymbol{S}^{-1}\boldsymbol{\varepsilon}_t \\
&=: k_{11}+k_{12}+k_{13}.
\end{aligned}$$

Using Lemmas A.2(ii), A.2(iii), A.2(vi), A.3(i) and A.3(ii), one has

$$\begin{aligned}
|k_{11}| &\leq \frac{1}{nT}\frac{1}{\hat{\sigma}^2(\hat{\boldsymbol{\theta}},\boldsymbol{\Lambda})}||\boldsymbol{M}_{\boldsymbol{\Lambda}}||_2\left(\sum_{t=1}^T||\boldsymbol{x}_{kt}||_2^2\right)^{\frac{1}{2}}\left(\sum_{t=1}^T||\boldsymbol{Z}_t(\boldsymbol{\theta}^0-\hat{\boldsymbol{\theta}})||_2^2\right)^{\frac{1}{2}} \\
&= \frac{1}{\sqrt{nT}}\frac{1}{\hat{\sigma}^2(\hat{\boldsymbol{\theta}},\boldsymbol{\Lambda})}||\boldsymbol{M}_{\boldsymbol{\Lambda}}||_2||\boldsymbol{\mathcal{X}}_k||_F\left(\frac{1}{nT}\sum_{t=1}^T||\boldsymbol{Z}_t(\boldsymbol{\theta}^0-\hat{\boldsymbol{\theta}})||_2^2\right)^{\frac{1}{2}} \\
&= \frac{1}{\sqrt{nT}}O_P(\sqrt{nT})O_P(a_{nT}) = O_P(1), \qquad \text{(A.20)}
\end{aligned}$$

$$\begin{aligned}
|k_{12}| &\leq \frac{1}{nT}\frac{1}{\hat{\sigma}^2(\hat{\boldsymbol{\theta}},\boldsymbol{\Lambda})}||\boldsymbol{M}_{\boldsymbol{\Lambda}}||_2||\boldsymbol{S}(\hat{\boldsymbol{\rho}})\boldsymbol{S}^{-1}||_2||\boldsymbol{\Lambda}^0||_2\left(\sum_{t=1}^T||\boldsymbol{x}_{kt}||_2^2\right)^{\frac{1}{2}}\left(\sum_{t=1}^T||\boldsymbol{f}_t^0||_2^2\right)^{\frac{1}{2}} \\
&= \frac{1}{nT}\frac{1}{\hat{\sigma}^2(\hat{\boldsymbol{\theta}},\boldsymbol{\Lambda})}||\boldsymbol{M}_{\boldsymbol{\Lambda}}||_2||\boldsymbol{S}(\hat{\boldsymbol{\rho}})\boldsymbol{S}^{-1}||_2||\boldsymbol{\Lambda}^0||_2||\boldsymbol{\mathcal{X}}_k||_F||\boldsymbol{F}^0||_F \\
&= \frac{1}{nT}O_P(\sqrt{n})O_P(\sqrt{T})O_P(\sqrt{nT}) = O_P(1), \qquad \text{(A.21)}
\end{aligned}$$

and

$$\begin{aligned}
|k_{13}| &\leq \frac{1}{nT}\frac{1}{\hat{\sigma}^2(\hat{\boldsymbol{\theta}},\boldsymbol{\Lambda})}||\boldsymbol{M}_{\boldsymbol{\Lambda}}||_2||\boldsymbol{S}(\hat{\boldsymbol{\rho}})\boldsymbol{S}^{-1}||_2\left(\sum_{t=1}^T||\boldsymbol{x}_{kt}||_2^2\right)^{\frac{1}{2}}\left(\sum_{t=1}^T||\boldsymbol{\varepsilon}_t||_2^2\right)^{\frac{1}{2}} \\
&= \frac{1}{nT}\frac{1}{\hat{\sigma}^2(\hat{\boldsymbol{\theta}},\boldsymbol{\Lambda})}||\boldsymbol{M}_{\boldsymbol{\Lambda}}||_2||\boldsymbol{S}(\hat{\boldsymbol{\rho}})\boldsymbol{S}^{-1}||_2||\boldsymbol{\mathcal{X}}_k||_F||\boldsymbol{\varepsilon}||_F \\
&= \frac{1}{nT}O_P(\sqrt{nT})O_P(\sqrt{nT}) = O_P(1). \qquad \text{(A.22)}
\end{aligned}$$

Combining the previous results gives $\frac{\partial \mathcal{L}(\hat{\theta}_{p^*},\mathbf{\Lambda})}{\partial \theta_p} = O_P(1)$. Now the derivative of the penalty term for θ_{p^*}, evaluated at $\hat{\theta}_{p^*}$, is

$$\frac{\partial \varrho_{p^*}(\hat{\theta}_{p^*},\gamma_{p^*},\zeta_{p^*})}{\partial \theta_{p^*}} = -\gamma_{p^*}\frac{1}{|\theta^{\dagger}_{p^*}|^{\zeta_{p^*}}}\frac{\hat{\theta}_{p^*}}{|\hat{\theta}_{p^*}|}. \tag{A.23}$$

By Assumption 5, $\gamma_p/|\theta^{\dagger}_{p^*}|^{\zeta_p}$ is explosive in probability. As such, as $n, T \to \infty$, the first order condition cannot be satisfied since $\frac{\partial \mathcal{L}(\hat{\theta}_{p^*},\mathbf{\Lambda})}{\partial \theta_p} = O_P(1)$ and yet the derivative of the penalty term diverges. This contradicts $\hat{\boldsymbol{\theta}}$ being a maximiser of the objective function. Therefore, instead, it must be that $\hat{\theta}_{p^*} = 0$ with probability approaching 1 as $n, T \to \infty$ for the first order condition (A.15) to be satisfied. $\square$

Proof of Theorem 1. Starting with the expression obtained in Lemma A.4,

$$\begin{aligned}
\boldsymbol{D}\sqrt{nT}(\hat{\boldsymbol{\theta}}_{(1)} - \boldsymbol{\theta}^0_{(1)}) &= \frac{1}{\sigma_0^2}\frac{1}{\sqrt{nT}}\boldsymbol{\mathcal{Z}}'_{(1)}(\boldsymbol{M}_{\boldsymbol{F}^0} \otimes \boldsymbol{M}_{\mathbf{\Lambda}^0})\text{vec}(\boldsymbol{\varepsilon}) \\
&\quad + \frac{1}{\sigma_0^2}\frac{1}{\sqrt{nT}}\begin{pmatrix} \text{tr}\left((\boldsymbol{G}_1^*\boldsymbol{\varepsilon})'\boldsymbol{M}_{\mathbf{\Lambda}^0}\boldsymbol{\varepsilon}\boldsymbol{M}_{\boldsymbol{F}^0}\right) \\ \vdots \\ \text{tr}\left((\boldsymbol{G}_{Q^0}^*\boldsymbol{\varepsilon})'\boldsymbol{M}_{\mathbf{\Lambda}^0}\boldsymbol{\varepsilon}\boldsymbol{M}_{\boldsymbol{F}^0}\right) \\ \mathbf{0}_{K^0\times 1}\end{pmatrix} + \boldsymbol{o}_P(1) \\
&= \frac{1}{\sigma_0^2}\frac{1}{\sqrt{nT}}\boldsymbol{c} - \frac{1}{\sigma_0^2}\frac{1}{\sqrt{nT}}\begin{pmatrix} \text{tr}\left((\boldsymbol{G}_1^*\boldsymbol{\varepsilon})'(\boldsymbol{P}_{\mathbf{\Lambda}^0}\boldsymbol{\varepsilon} + \boldsymbol{M}_{\mathbf{\Lambda}^0}\boldsymbol{\varepsilon}\boldsymbol{P}_{\boldsymbol{F}^0})\right) \\ \vdots \\ \text{tr}\left((\boldsymbol{G}_{Q^0}^*\boldsymbol{\varepsilon})'(\boldsymbol{P}_{\mathbf{\Lambda}^0}\boldsymbol{\varepsilon} + \boldsymbol{M}_{\mathbf{\Lambda}^0}\boldsymbol{\varepsilon}\boldsymbol{P}_{\boldsymbol{F}^0})\right) \\ \mathbf{0}_{K^0\times 1}\end{pmatrix} \\
&\quad - \frac{1}{\sigma_0^2}\frac{1}{\sqrt{nT}}\begin{pmatrix} \text{tr}\left((\mathbb{Z}_1 - \bar{\mathbb{Z}}_1)'(\boldsymbol{P}_{\mathbf{\Lambda}^0}\boldsymbol{\varepsilon} + \boldsymbol{M}_{\mathbf{\Lambda}^0}\boldsymbol{\varepsilon}\boldsymbol{P}_{\boldsymbol{F}^0})\right) \\ \vdots \\ \text{tr}\left((\mathbb{Z}_{P^0} - \bar{\mathbb{Z}}_{P^0})'(\boldsymbol{P}_{\mathbf{\Lambda}^0}\boldsymbol{\varepsilon} + \boldsymbol{M}_{\mathbf{\Lambda}^0}\boldsymbol{\varepsilon}\boldsymbol{P}_{\boldsymbol{F}^0})\right)\end{pmatrix} + \boldsymbol{o}_P(1),
\end{aligned} \tag{A.24}$$

where $\boldsymbol{c} := \mathbb{Z}'_{(1)}\text{vec}(\boldsymbol{\varepsilon}) + (\text{tr}(\boldsymbol{\varepsilon}'\boldsymbol{G}_1^*\boldsymbol{\varepsilon}), ..., \text{tr}(\boldsymbol{\varepsilon}'\boldsymbol{G}_{Q^0}^*\boldsymbol{\varepsilon}), \mathbf{0}_{1\times K^0})'$, recalling the definition of $\mathbb{Z}_p$ and $\mathbb{Z}$ given just prior to the statement of Assumption 7. By expanding the second term on the right-hand side of (A.24) and applying Lemmas A.5(ii), A.5(iii), A.5(iv), and also applying Lemma A.5(v) to the third term on the right, one obtains

$$\boldsymbol{D}\sqrt{nT}(\hat{\boldsymbol{\theta}}_{(1)} - \boldsymbol{\theta}^0_{(1)}) = \frac{1}{\sigma_0^2}\frac{1}{\sqrt{nT}}\boldsymbol{c} + \mathbb{b} + \boldsymbol{o}_P(1). \tag{A.25}$$

Rearranging and premultiplying by $\big(\mathbb{S}(\boldsymbol{D}+\boldsymbol{V})\mathbb{S}'\big)^{-\frac{1}{2}}$ gives

$$\sqrt{nT}\big(\mathbb{S}(\boldsymbol{D}+\boldsymbol{V})\mathbb{S}'\big)^{-\frac{1}{2}}\mathbb{S}\boldsymbol{D}(\hat{\boldsymbol{\theta}}_{(1)}-\boldsymbol{D}^{-1}\mathbb{b}-\boldsymbol{\theta}^0_{(1)})=\big(\mathbb{S}(\boldsymbol{D}+\boldsymbol{V})\mathbb{S}'\big)^{-\frac{1}{2}}\mathbb{S}\frac{1}{\sqrt{nT}}\frac{1}{\sigma_0^2}\boldsymbol{c}+\boldsymbol{o}_P(1). \tag{A.26}$$

Finally, using Lemma A.6 and Assumption 7.2, $\big(\mathbb{S}(\boldsymbol{D}+\boldsymbol{V})\mathbb{S}'\big)^{-\frac{1}{2}}\mathbb{S}\frac{1}{\sqrt{nT}}\frac{1}{\sigma_0^2}\boldsymbol{c}\xrightarrow{d}\mathcal{N}(\boldsymbol{0}_{L\times L},\boldsymbol{I}_L)$, which yields the result. □

Proof of Proposition 3. In order to prove the result, it suffices to show that $||\boldsymbol{D}^{-1}\mathbb{b}-\hat{\boldsymbol{D}}^{-1}\hat{\mathbb{b}}||_2=o_P(1)$. Observe that

$$||\boldsymbol{D}^{-1}\mathbb{b}-\hat{\boldsymbol{D}}^{-1}\hat{\mathbb{b}}||_2\leq||\boldsymbol{D}^{-1}-\hat{\boldsymbol{D}}^{-1}||_2||\hat{\mathbb{b}}||_2+||\boldsymbol{D}^{-1}||_2||\mathbb{b}-\hat{\mathbb{b}}||_2. \tag{A.27}$$

It is straightforward to establish that $||\boldsymbol{D}^{-1}-\hat{\boldsymbol{D}}^{-1}||_2||\hat{\mathbb{b}}||_2=\boldsymbol{o}_P(1)$ using Lemma A.5(i). For the second term in (A.27), $||\mathbb{b}-\hat{\mathbb{b}}||_2=o_P(1)$ can be shown using Lemmas A.5(ii)–A.5(v), and the following two results. First, let $\hat{\boldsymbol{\rho}}_{(1)}$ denote the $Q^0\times 1$ vector containing the estimates of the truly nonzero parameters in $\boldsymbol{\rho}^0$. Then,

$$\begin{aligned}
||\boldsymbol{G}^*_q-\boldsymbol{G}^*_q(\hat{\boldsymbol{\rho}}_{(1)})||_2&=||\boldsymbol{G}_q-\boldsymbol{G}_q(\hat{\boldsymbol{\rho}}_{(1)})-\boldsymbol{I}_n\frac{1}{n}\mathrm{tr}(\boldsymbol{G}_q)+\frac{1}{n}\boldsymbol{I}_n\mathrm{tr}(\boldsymbol{G}_q(\hat{\boldsymbol{\rho}}_{(1)}))||_2\\
&\leq||\boldsymbol{G}_q-\boldsymbol{G}_q(\hat{\boldsymbol{\rho}}_{(1)})||_2+\frac{1}{n}|\mathrm{tr}(\boldsymbol{G}_q(\hat{\boldsymbol{\rho}}_{(1)})-\boldsymbol{G}_q)|\\
&\leq 2||\boldsymbol{G}_q-\boldsymbol{G}_q(\hat{\boldsymbol{\rho}}_{(1)})||_2\\
&=2||\boldsymbol{G}_q(\hat{\boldsymbol{\rho}}_{(1)})(\boldsymbol{I}_n-\boldsymbol{S}(\hat{\boldsymbol{\rho}}_{(1)})\boldsymbol{S}^{-1})||_2\\
&\leq 2||\boldsymbol{G}_q(\hat{\boldsymbol{\rho}}_{(1)})||_2||\boldsymbol{I}_n-\boldsymbol{S}(\hat{\boldsymbol{\rho}}_{(1)})\boldsymbol{S}^{-1}||_2\\
&=O_P(\sqrt{Q^0}||\boldsymbol{\theta}^0-\hat{\boldsymbol{\theta}}||_2)
\end{aligned}\tag{A.28}$$

using Lemma A.2(viii). Second,

$$\begin{aligned}
||\boldsymbol{A}-\boldsymbol{A}(\hat{\boldsymbol{\rho}}_{(1)},\hat{\boldsymbol{\phi}})||_2&=||\boldsymbol{S}^{-1}(\phi^0_1\boldsymbol{I}_n+\sum_{q=1}^{Q^0}\phi^0_{q+1}\boldsymbol{W}_{Q^0})-\boldsymbol{S}^{-1}(\hat{\boldsymbol{\rho}}_{(1)})(\hat{\phi}_1\boldsymbol{I}_n+\sum_{q=1}^{Q^0}\hat{\phi}_{q+1}\boldsymbol{W}_q)||_2\\
&\leq|\phi^0_1-\hat{\phi}_1|||\boldsymbol{S}^{-1}||_2+|\hat{\phi}_1-\phi^0_1|||\boldsymbol{S}^{-1}(\hat{\boldsymbol{\rho}}_{(1)})||_2||\boldsymbol{S}(\hat{\boldsymbol{\rho}}_{(1)})\boldsymbol{S}^{-1}-\boldsymbol{I}_n||_2\\
&\quad+|\phi^0_1|||\boldsymbol{S}^{-1}(\hat{\boldsymbol{\rho}}_{(1)})||_2||\boldsymbol{S}(\hat{\boldsymbol{\rho}}_{(1)})\boldsymbol{S}^{-1}-\boldsymbol{I}_n||_2+\sum_{q=1}^{Q^0}|\phi^0_{q+1}-\hat{\phi}_{q+1}|||\boldsymbol{S}^{-1}||_2||\boldsymbol{W}_q||_2\\
&\quad+\sum_{q=1}^{Q^0}|\phi^0_{q+1}|||\boldsymbol{S}(\hat{\boldsymbol{\rho}}_{(1)})||_2||\boldsymbol{S}(\hat{\boldsymbol{\rho}}_{(1)})\boldsymbol{S}^{-1}-\boldsymbol{I}_n||_2||\boldsymbol{W}_q||_2
\end{aligned}$$

$$+\sum_{q=1}^{Q^0}|\hat{\phi}_{q+1}-\phi_{q+1}^0|||\boldsymbol{S}(\hat{\boldsymbol{\rho}}_{(1)})||_2||\boldsymbol{S}(\hat{\boldsymbol{\rho}}_{(1)})\boldsymbol{S}^{-1}-\boldsymbol{I}_n||_2||\boldsymbol{W}_q||_2$$
$$=O_P(\sqrt{Q^0}||\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}^0||_2) \tag{A.29}$$

since $|\phi_1^0-\hat{\phi}_1|||\boldsymbol{S}^{-1}||_2 \leq ||\boldsymbol{\theta}^0-\hat{\boldsymbol{\theta}}||_2||\boldsymbol{S}^{-1}||_2 = O_P(\sqrt{Q^0}||\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}^0||_2)$, $|\hat{\phi}_1-\phi_1^0|||\boldsymbol{S}^{-1}(\hat{\boldsymbol{\rho}}_{(1)})||_2||\boldsymbol{S}(\hat{\boldsymbol{\rho}}_{(1)})\boldsymbol{S}^{-1}-\boldsymbol{I}_n||_2 = O_P(\sqrt{Q^0}||\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}^0||_2)$, $|\phi_1^0|||\boldsymbol{S}^{-1}(\hat{\boldsymbol{\rho}}_{(1)})||_2||\boldsymbol{S}(\hat{\boldsymbol{\rho}}_{(1)})\boldsymbol{S}^{-1}-\boldsymbol{I}_n||_2 \leq ||\boldsymbol{\beta}^0||_1|\boldsymbol{S}^{-1}(\hat{\boldsymbol{\rho}}_{(1)})||_2||\boldsymbol{S}(\hat{\boldsymbol{\rho}}_{(1)})\boldsymbol{S}^{-1}-\boldsymbol{I}_n||_2 = O_P(\sqrt{Q^0}||\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}^0||_2)$, $\sum_{q=1}^{Q^0}|\phi_{q+1}^0|||\boldsymbol{S}(\hat{\boldsymbol{\rho}}_{(1)})||_2||\boldsymbol{S}(\hat{\boldsymbol{\rho}}_{(1)})\boldsymbol{S}^{-1}-\boldsymbol{I}_n||_2||\boldsymbol{W}_q||_2 \leq ||\boldsymbol{\beta}^0||_1||\boldsymbol{S}(\hat{\boldsymbol{\rho}}_{(1)})||_2$ $||\boldsymbol{S}(\hat{\boldsymbol{\rho}}_{(1)})\boldsymbol{S}^{-1}-\boldsymbol{I}_n||_2 \max_{1\leq q\leq Q^0}||\boldsymbol{W}_{Q^0}||_2 = O_P(\sqrt{Q^0}||\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}^0||_2)$ and $\sum_{q=1}^{Q^0}|\hat{\phi}_{q+1}-\phi_{q+1}^0|||\boldsymbol{S}(\hat{\boldsymbol{\rho}}_{(1)})||_2$ $||\boldsymbol{S}(\hat{\boldsymbol{\rho}}_{(1)})\boldsymbol{S}^{-1}-\boldsymbol{I}_n||_2||\boldsymbol{W}_{Q^0}||_2 \leq ||\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}^0||_2||\boldsymbol{S}(\hat{\boldsymbol{\rho}}_{(1)})||_2||\boldsymbol{S}(\hat{\boldsymbol{\rho}}_{(1)})\boldsymbol{S}^{-1}-\boldsymbol{I}_n||_2\sqrt{Q^0}\max_{1\leq q\leq Q^0}||\boldsymbol{W}_{Q^0}||_2 = O_P(Q^0||\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}^0||_2^2)$, where the order of $||\boldsymbol{S}(\hat{\boldsymbol{\rho}}_{(1)})\boldsymbol{S}^{-1}-\boldsymbol{I}_n||_2$ is established using Lemma A.2(viii). The result then follows. □

Proof of Proposition 4. The proof largely follows the same structure as the proof of Theorem 3.5 in Lu and Su (2016). Details can be found in Appendix D in the Supplementary Material. □

References

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

Bai, J., Li, K., 2018. Dynamic spatial panel data models with common shocks. Working paper, Columbia University.

Bai, J., Liao, Y., 2017. Inferences in panel data with interactive effects using large covariance matrices. Journal of Econometrics 200 (1), 59–78.

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

Blume, L. E., Brock, W. A., Durlauf, S. N., Jayaraman, R., 2015. Linear social interactions models. Journal of Political Economy 123 (2), 444–496.

Bonhomme, S., Manresa, E., 2015. Grouped patterns of heterogeneity in panel data. Econometrica 83 (3), 1147–1184.

Bramoullé, Y., Djebbari, H., Fortin, B., 2009. Identification of peer effects through social networks. Journal of Econometrics 150 (1), 41–55.

de Paula, Á., Rasul, I., Souza, P., 2020. Identifying network ties from panel data: Theory and an application to tax competition. Working paper, CeMMAP.

Durlauf, S. N., Johnson, P. A., Temple, J. R., 2009. The methods of growth econometrics. In: Mills, T., Patterson, K. (Eds.), Palgrave Handbook of Econometrics. Vol. 2. Elsevier, Ch. 8, pp. 1119–1179.

Fan, J., Peng, H., 2004. Nonconcave penalized likelihood with a diverging number of parameters. The Annals of Statistics 32 (3), 928–961.

Gupta, A., Robinson, P. M., 2015. Inference on higher-order spatial autoregressive models with increasingly many parameters. Journal of Econometrics 186 (1), 19–31.

Gupta, A., Robinson, P. M., 2018. Pseudo maximum likelihood estimation of spatial autoregressive models with increasing dimension. Journal of Econometrics 202 (1), 92–107.

Horn, R. A., Johnson, C. R., 2012. Matrix Analysis, 2nd Edition. Cambridge University Press, New York, USA.

Hsiao, C., 2018. Panel models with interactive effects. Journal of Econometrics 206 (2), 645–673.

Kuersteiner, G. M., Prucha, I. R., 2020. Dynamic spatial panel models: Networks, common shocks, and sequential exogeneity. *forthcoming in* Econometrica.

Lam, C., Souza, P. C., 2019. Estimation and selection of spatial weight matrix in a spatial lag model. Journal of Business & Economic Statistics 38 (3), 693–710.

Latala, R., 2005. Some estimates of norms of random matrices. Proceedings of the American Mathematical Society 133 (5), 1273–1282.

Lee, L.-F., 2007. Identification and estimation of econometric models with group interactions, contextual factors and fixed effects. Journal of Econometrics 140 (2), 333–374.

Lee, L.-F., Liu, X., Lin, X., 2010. Specification and estimation of social interaction models with network structures. The Econometrics Journal 13 (2), 145–176.

Lee, L.-F., Yu, J., 2010. Estimation of spatial autoregressive panel data models with fixed effects. Journal of Econometrics (1), 165–185.

Lee, L.-F., Yu, J., 2014. Spatial panel data models. In: Baltagi, B. H. (Ed.), The Oxford Handbook of Panel Data. Oxford University Press, Ch. 12, pp. 363–401.

Leeb, H., Pötscher, B. M., 2005. Model selection and inference: Facts and fiction. Econometric Theory 21 (1), 21–59.

Lewbel, A., Qu, X., Tang, X., 2019. Social networks with misclassifed or unobserved links. Working paper.

Liu, T., 2017. Model selection and adaptive lasso estimation of spatial models. PhD Thesis, The Ohio State University.

Lu, X., Su, L., 2016. Shrinkage estimation of dynamic panel data models with interactive fixed effects. Journal of Econometrics 190 (1), 148–175.

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

Moon, H. R., Weidner, M., 2017. Dynamic linear panel regression models with interactive fixed effects. Econometric Theory 33 (1), 158–195.

Newey, W. K., McFadden, D., 1994. Large sample estimation and hypothesis testing. In: Engle, R. F., McFadden, D. L. (Eds.), Handbook of Econometrics. Vol. 4. Elsevier, Ch. 36, pp. 2111–2245.

Shi, W., Lee, L.-F., 2017. Spatial dynamic panel data models with interactive fixed effects. Journal of Econometrics 197 (2), 323–347.

Shi, W., Lee, L.-F., 2018. A spatial panel data model with time varying endogenous weights matrices and common factors. Regional Science and Urban Economics 72, 6–34.

Su, L., Shi, Z., Phillips, P. C. B., 2016. Identifying latent structures in panel data. Econometrica 84 (6), 2215–2264.

Wang, Y., 2018. Panel data with high-dimensional factors: Inference on treatment effects with an application to sampled networks. Working paper.

Yu, J., de Jong, R., Lee, L.-F., 2008. Quasi-maximum likelihood estimators for spatial dynamic panel data with fixed effects when both n and T are large. Journal of Econometrics 146 (1), 118–134.

Zhang, X., Yu, J., 2018. Spatial weights matrix selection and model averaging for spatial autoregressive models. Journal of Econometrics 203 (1), 1–18.

Zou, H., 2006. The adaptive lasso and its oracle properties. Journal of the American Statistical Association 101 (476), 1418–1429.