Essays on Cross-Sectionally Dependent Panel Data with an Application to Fiscal Policy in the European Monetary Union

BART VAN-ARK

Essays on Cross-Sectionally Dependent Panel Data with an Application to Fiscal Policy in the European Monetary Union

Bart van Ark

A Thesis Submitted in Fulfilment of the Requirements for the Degree of Doctor of Philosophy in Economics at Durham University Business School

Durham University

September 2018

Executive Summary

In the past two decades, macroeconomists have used panel data to study the merits of fiscal policy for economic stabilisation. The datasets considered in these studies typically consist of a small number of time series corresponding to countries. This configuration does not match with the archetypical survey-style panel dataset for which a large literature concerning estimation and hypothesis testing exists. This PhD develops an estimation methodology that is catered towards macroeconomists: in four self-contained chapters, we develop a methodology for the estimation of dynamic models in the small N, large T framework in the presence of cross-sectional dependence in the error term.

In the first chapter we examine the effect of factors on the point estimates of several commonly-used estimators in the empirical literature and we find that these estimators are inconsistent. We also propose an estimator that is consistent for the parameters for of the model studied in that chapter.

In the second chapter we develop consistent quasi-difference GMM estimators and inferential procedures for the small N, large T dynamic panel data model with factor error structures. We also prove consistency and mixed-normality of the estimator when the number of factors is over-estimated.

In the third chapter we consider the large N, large T framework and show the first eigenvalues of the covariance matrix of an approximate factor model are dominated by the factors whereas the remainder is controlled by the residual noise. We show that this result is the basis for any consistent inferential procedure about R and continues to hold when R grows large, when the factors are weak and, importantly, in the large N, large T interactive fixed effects model.

In the fourth chapter we study fiscal policy using the methods developed in the thesis. We estimate vector autoregressions from European countries and restrict the impulse-response functions to adhere to the Stability and Growth Pact. We find that this one-size-fits-all approach is not appropriate for stabilization of the European economy.

Acknowledgements

This thesis is dedicated to my fiancée, my grandparents and my mother, in no particular order.

I owe a great deal of thanks to Dr. Hugo Kruiniger for his supervisory support and expert guidance throughout this process.

Financial support from the ESRC is kindly appreciated, as is the assistance of the Business School since the start of this project.

Contents

Chapter 1

Introduction and Literature Review

1.1 Introduction

A number of studies have employed panel data methods in the fiscal policy literature in recent years. Examples of such studies include Ilzetski (2011), Ilzetski et al (2013), Bénétrix and Lane (2013) and Beetsma and Guilidori (2010a, b, 2011). All these studies make use of econometric techniques that are designed for microeconomic datasets with large cross-sectional (N) and small time (T) dimensions. However, when considering country data, large N likely leads to significant cross-sectional heterogeneity because the fundamentals of the countries considered in the panel can differ substantially. For example, stabilization policy in Finland, where government spending consists of approximately sixty percent of GDP, can be expected to have different short and long-term effects than in the United States, where the corresponding figure is closer to thirty-five. As a result, estimation results that impose slope parameter homogeneity in large N country studies should be interpreted with caution and any efficiency advantages offered by a large dataset can reduce the interpretability of the results when the dataset contains countries with very different characteristics. In other words, careful selection of the countries combined in a panel data set is necessary for the interpretability of the results *a priori* and this argument restricts the size of N.

A second particularity of many macroeconomic variables is that they share common effects when considering them as vector processes rather than individual series. Often, a large portion of the variability of such a vector process can be modelled by a small and fixed number of unobserved variables and these unobservable variables are commonly referred to as factors (Bai, 2003). As an example, consider the traditional Cobb-Douglas production function of a set of countries indexed by $i = 1, \ldots, N$ at time t:

$$y_{i,t} = \alpha k_{i,t} + (1-\alpha) l_{i,t} + u_{i,t},$$

in the above, $y_{i,t}$ is the logarithm of output, $k_{i,t}$ is the logarithm of capital, $l_{i,t}$ the logarithm of labour, α is the share of capital and $u_{i,t}$ is total factor productivity (TFP) particular to country i that is typically considered to be random. For economically similar countries, i.e. those with similar levels of development, social welfare

or as members of a monetary and/or fiscal union, it may then be reasonable to assume α is constant over individuals and time (See, i.e., Barro and Sala-i-Martin, 1992). On the other hand, it seems unreasonable that TFP is distinctly specific for each country at each t. Rather, for economically similar countries, it is likely that TFP is instead a composite function of a common component and an individual specific component, i.e.

$$u_{i,t} = \lambda_i F_t + \varepsilon_{i,t}.$$

In this composite TFP representation, the vector $F_t = [f_{1,t}, \ldots, f_{R,t}]'$ consists of R variables which vary over time and hit each $y_{i,t}$ with different intensity $\lambda_i = [\lambda_{1,i}, \ldots, \lambda_{R,i}]$. It is important to note that under this specification, $E(u_{i,t}u_{j,t}) \neq 0$ and there is now cross-sectional dependence between the composite error terms of countries i and j. We may think of the F_t as macro-shocks: they can be global business cycle events such as financial crises that impact each country with different intensity, whilst the $\varepsilon_{i,t}$ are shocks at the country-level, for example the outcome of local elections or wage negotiations. As is customary in econometrics, the $u_{i,t}$ are unobservable and estimation of α is now complicated by the unobservable F_t. The literature has proposed alternative methodologies to account for factors and the approaches depend on the goal of the study: in some studies, factors are introduced to capture variables that cannot be observed, and therefore cannot be included directly in a regression model. In those studies, often an argument is made to use control variables and whiten the regression residual $\hat{u}_{i,t}$. In other studies, factors are used to reduce the curse of dimensionality by reducing a vector of variables to one of lower dimension. Yet other authors take the latter approach one step further and work only with the factor representation of data in an effort to combat misspecification. In our opinion, the factor methodology provides a captivating way of dealing with the problems of cross-sectional dependence, misspecification and unobservable variables found in traditional economic time series.

In light of the above it is perhaps surprising that suitable estimation methods for small N, large T panels are only sparsely developed and may further explain why workhorse macroeconomic techniques such as vector autoregression (VAR) analysis are only rarely considered in a panel data framework. This in turn suggests scope for the development of longitudinal methods designed specifically for macroeconomics, both under the limitations of an acceptable homogeneity assumption and in the presence of unobservable factors.

The overall objective of the thesis is therefore to develop comprehensive estimation and inference procedures that can be applied to the parameters of the small N, large T dynamic panel data model with cross-sectionally dependent errors. More specifically, in Chapter 2 we identify the problems with panel data econometrics when factors are present in real data. We find that factors cause omitted variable bias in the point estimates of certain popular estimators unless some very stringent assumptions are made. As a result, we argue that empirical results based on panel data need to be interpreted with caution if no suitable adjustment is made for this omitted variable bias. A small Monte Carlo experiment shows that this bias can be severely misleading, even pushing the coefficient estimate of a simple autoregressive model to unity and thus falsely implying the presence of unit roots in the data.

In Chapter 3 we develop a comprehensive Method of Moments methodology that is robust to factor

errors in dynamic models in the large T framework. We derive consistency and asymptotic normality of the estimators, procedures to determine the correct number of factors and a battery of basic statistical tests intended for the empirical researcher to apply to macroeconomic problems. This methodology is based on the work of Ahn et al (2013) in the large N framework and the econometric theory developed in this chapter offers formal justification for, and extensions to, results in that paper. Importantly, we derive a limit theory for the estimator when the number of factors included in the procedure is too large. This limit theory shows that inference about the slope parameters of the regression model remains valid even when the number of factors is overestimated.

We then consider the problem of selecting the correct number of factors in large N, T models in Chapter 4. This is motivated by the fact that popular methodologies based on information criteria have only asymptotic justification and are known to be very imprecise in small samples. We instead focus on results from Random Matrix Theory and apply these to the problem of selecting the correct number of factors in large dimensional panel data models. Under the assumptions maintained in that chapter, the order of magnitude of the eigenvalue distribution of the factors and the error covariance matrix differs. This fact can thus be used to distinguish the eigenvalues related to the factors from the eigenvalues related to the usual residuals. Furthermore, it allows us to place several consistent estimators of the number of factors in a class we call "Scree plot" estimators. The chapter subsequently uses these arguments to determine the number of factors in traditional factor models and interactive fixed effects models. It also shows that eigenvalue separation continues to hold in so-called weak factor models and models where the number of factors goes to infinity along with N and T under certain conditions.

The final chapter takes the methods developed in the thesis to the data and adds some new insights to the fiscal VAR literature. More specifically, we provide a step-by-step description of how a panel data VAR can be estimated in the presence of factor residuals. We then estimate three panel VARs from Eastern, Southern and Western European panel data and identify structural shocks based on a new methodology that incorporates both zero and sign-restrictions. The impulse-response functions of these VARs are further identified to obey a structure with and without the Stability and Growth Pact imposed. With these results, we contribute to the debate on the Stability and Growth Pact and its impact on the ability of fiscal stabilisation in an environment where monetary policy is no longer catered to the needs of an individual country. Since the content of Chapter 5 is applied macroeconomics, it is self-contained and includes a review of the empirical literature on fiscal policy and its implications for economic growth.

The remainder of this introductory chapter is devoted to an extensive review of the econometrics literature related to panel data with cross-sectional dependence. There are several other reviews dealing with cross-sectionally dependent data, such as Sarafidis and Wansbeek (2012) and Bai and Wang (2016); since this thesis focuses on cross-sectionally dependent data with small N, large T, this review has a different scope and it is intended to be complementary to the former reviews. Based on this review, we identify three gaps in the literature which are as follows. First, whilst the consistency of conventional estimators has been thoroughly examined in the case of large N and large N, T panel data models, consistency of conventional estimators applied to the large T dynamic panel data model has not. This is problematic because several

macroeconomists have employed such panel data methods anyway in recent years. Second, we conclude that estimation of large T dynamic panel data with a small and fixed number of time series, i.e. countries, cannot be directly undertaken with existing methodologies, although certain GMM estimators can be adapted to the large T model. Third, even though determination of the number of factors is crucial for consistent estimation of models with factor errors, the literature so far has focused only on determination of the number of factors in particular types of models under very specific assumptions. As a result there is scope to make a contribution to the analysis of the determination of the number of factors in other models.

1.2 Estimation of Models with Factor Errors

Model and Assumptions

Our review of the econometric literature is based on the following panel data model with multiple time-varying effects:

$$\begin{aligned} y_{i,t} &= \beta' x_{i,t} + u_{i,t}; \\ u_{i,t} &= \lambda_i F_t + \varepsilon_{i,t}, \text{for } i = 1,\ldots,N, t = 1,\ldots,T, \end{aligned} \tag{1.1}$$

where either N or T or both may be large depending on the context. In this thesis, we will refer to (1.1) as an *interactive fixed effects model*. In model (1.1), $y_{i,t}$ is the dependent variable and $x_{i,t}$ is a P-vector of regressors with representative element $x_{p,i,t}$; in macroeconomic models, we expect that $x_{i,t}$ will contain lagged dependent variables. Furthermore, $\varepsilon_{i,t}$ is an individual-specific disturbance term. The source of the cross-sectional dependence is the interaction term $\lambda_i F_t$ in the composite error $u_{i,t}$, which cannot be observed: the vector $F_t = [f_{1,t},\ldots,f_{R,t}]'$ consists of R variables which vary over time and hit each $y_{i,t}$ with different intensity $\lambda_i = [\lambda_{1,i},\ldots,\lambda_{R,i}]$. Throughout this thesis, the time series components will be referred to as factors; the intensity parameters as loadings and their product as interactive fixed effects. Across the literature, assumptions on the functional form and distribution of the λ_i and F_t vary and we purposefully leave them unspecified to accommodate this, although with macroeconomic models in mind, we expect them to be dynamic. Note further that we have not explicitly included a constant in model (1.1): this is because, in our view, factor errors constitute a generalization of individual-specific constants. Clearly, a constant can be included in F_t and as a result, the well-known individual-effects model is covered by model (1.1) as a special case. Similarly, a time-effects component can be accommodated by setting $\lambda_{1,i} = \overline{\lambda}_1$ for all $i = 1,\ldots,N$.

For now it is important to note that the λ_i and F_t are unobserved and estimation of β is thus complicated by the presence of the factor structure, which has unknown dimension R. This is a classic omitted variable problem, in the sense that if $\text{cov}(x_{i,t}, F_t) \neq 0$, an OLS estimate of β will be inconsistent. If $\text{cov}(x_{i,t}, F_t) = 0$, OLS applied to (a transformation of) the data will be consistent and Generalized Least Squares (GLS) is efficient. Much of the GLS literature dealing with factor errors is concerned with this specific setup, but we

will ignore this specification because it is less general than when $\text{cov}(x_{i,t}, F_t) \neq 0$.[1] When $\text{cov}(x_{i,t}, F_t) \neq 0$, a particularly interesting problem arises for the theorist because the omitted variable will dominate the variability of $y_{i,t}$, particularly when N, T are large. Moreover, estimation techniques often depend on a consistent estimate of F_t either implicitly or explicitly, which further complicates estimation, especially when both dimensions of the data grow to infinity.

The fact that the interactive fixed effects are unobservable adds yet another complication: letting $\Lambda = [\lambda_1', \ldots, \lambda_N']'$and $F = [F_1, \ldots, F_T]$, we have $\Lambda F = \Lambda C C^{-1} F$ for any invertible $R \times R$ matrix C. In other words, there is an identification problem as a result of $R \times R$ free parameters in ΛF. These free parameters necessitate the use of a normalization of the interactive fixed effects to achieve identification of any parameter estimated from model (1.1). Bai and Ng (2013) discuss two such commonly used normalizations, although one can conceive many more identification strategies. The first strategy consists of restricting $FF'/T = I_R$, yielding $R(R+1)/2$ restrictions; a further $R(R-1)/2$ restrictions are imposed on (a function of) Λ. The second methodology takes an $R \times R$ sub-matrix of either F or Λ and fixes this at I_R. It is important to note that such normalizations are only required for technical reasons and are generally immaterial for the interpretation of estimation results. However, the adopted identification strategy does require being supported by the data: for example, applying the second strategy is impossible if the sub-matrix designated to be fixed at I_R is singular and it is not difficult to envision similar objections to the first identification strategy in certain contexts.

It is also important to note that if $\beta = 0$, model (1.1) collapses to a "pure" factor model:

$$y_{i,t} = \lambda_i F_t + \varepsilon_{i,t},\ i = 1, \ldots, N, t = 1, \ldots, T. \tag{1.2}$$

Such models further separate into (i) the *strict factor model* if $\text{cov}(\lambda_i \varepsilon_{i,t}) = 0$ and $\text{cov}(F_t \varepsilon_{i,t}) = 0$ and (ii) the *approximate factor model* if $\text{cov}(\lambda_i \varepsilon_{i,t}) \neq 0$ and/or $\text{cov}(F_t \varepsilon_{i,t}) \neq 0$. Historically, model (1.2) has dominated the attention of the literature and whilst this thesis is primarily concerned with model (1.1), in Chapter 4 we necessarily devote attention to model (1.2). Finally, a third generalization of (1.2) is the *dynamic factor model*:

$$y_{i,t} = \lambda_i(L) F_t + \varepsilon_{i,t},\ i = 1, \ldots, N, t = 1, \ldots, T. \tag{1.3}$$

where the "L" in (1.3) signifies a possibly infinite lag operator operating on F_t. In the remainder of this review, we use model (1.2) to illustrate the evolution of estimation techniques capable of dealing with parameter β in model (1.1) and the question of how to determine the dimension of F_t consistently. Moreover, we mostly ignore model (1.3) in the remainder of this thesis, although we refer to the dynamic factor model at various points because certain methodologies were initially designed for (1.3).

To make a meaningful comparison of existing estimation methods in what follows, it is necessary to present a set of basic assumptions as a reference point and discuss estimation of β and subsequently R with reference to these assumptions. Let $c < \infty$ be a generic constant that may differ depending on the context.

[1] See Sarafidis and Wansbeek (2012), section 4.1 for details.

The following assumptions amalgamate those made in Bai (2003, 2009) and Moon and Weidner (2017):

ASSUMPTION 1: (i) $R \leq c$, (ii) $E\|F_t\|^4 \leq c$ and $T^{-1}FF' \to_p \Sigma_F > 0$ *and* (iii) $E\|\lambda_i\|^4 \leq c$ *and* $N^{-1}\Lambda'\Lambda \to_p \Sigma_\Lambda > 0$ *for* $R \times R$ *matrices* Σ_F *and* Σ_Λ.

ASSUMPTION 2: (i) *let* $E(\varepsilon_{i,t}) = 0$, $E|\varepsilon_{i,t}|^8 \leq c$ *for all* i *and* t, (ii) *let* $\varepsilon_t = [\varepsilon_{1,t}, \ldots, \varepsilon_{N,t}]'$, $\varepsilon = [\varepsilon_1, \ldots, \varepsilon_T]'$ *and* $\Omega_\varepsilon = p\lim\left(N^{-1}\varepsilon\varepsilon' \vee T^{-1}\varepsilon'\varepsilon : \Omega_\varepsilon > 0\right)$ *and* $(NT)^{-1}\operatorname{tr}(\varepsilon\varepsilon') \leq c$, (iii) *for all* i, j, t *and* s, $p\lim T^{-1}\sum_{t=1}^T \varepsilon_{i,t}\varepsilon_{j,s} \leq c$ *and* $p\lim N^{-1}\sum_{i=1}^N \varepsilon_{i,t}\varepsilon_{j,s} \leq c$.

ASSUMPTION 3.1: $\varepsilon_{i,t}$ *is independent of* F_s, λ_j *and* $x_{p,j,s}$ *for all* i, j, p, s *and* t.

ASSUMPTION 3.2: (i) $E\left\|N^{-1/2}\sum_{i=1}^N \varepsilon_{i,t}\lambda_i\right\|^2 \leq c$ *and* $N^{-1/2}\sum_{i=1}^N \lambda_i'\varepsilon_{i,t} \to_d N(0, V)$ *for all* t *and* (ii) $E\left\|(NT)^{-1/2}F\varepsilon'\right\|^2$ *and* $T^{-1/2}\sum_{t=1}^T F_t\varepsilon_{i,t} \to_d N(0, V)$ *for all* i.

Assumption 1 states that the number of factors is finite and that the factors and their loadings have finite fourth moments, in addition to requiring positive definite covariance matrices for either. This latter requirement implies (i) stationarity of the factors and (ii) that no factor can be written as a linear combination of the others. In other words, the latter condition is analogous to ruling out collinearity of covariates in standard regression theory. Assumption 2 is a set of moment conditions on the model error $\varepsilon_{i,t}$ which permits weak dependence in the cross-section and time dimensions. Assumption 3 is a composite assumption and the choice of application depends on consideration of model (1.1) or (1.2). Assumption 3.2 permits weak correlation between the errors and the factors and loadings and ensures an appropriate CLT applies depending on the dimension of the data in the approximate factor model. For model (1.1), stricter assumptions are required on the permitted correlation between the error and the (un-) observables for consistent estimation of β and these are given in Assumption 3.1. Note that Assumption 3.1 implies strict exogeneity of the regressors *and* the factors and loadings and this seems to be the maintained assumption throughout much of the estimation literature dealing with interactive fixed effects models.

Estimation Methods

Inconsistencies in Traditional Estimators

Before we discuss estimators that are specifically designed to estimate model permitting some form of factor structure, we note that there is a long tradition of investigating the consistency of more traditional estimators under various conditions, both in terms of the specification of the model and the configuration of the panel. Important contributions in this literature for dynamic panel data models are Nickel (1981) and Kiviet (1995), who show that the Within estimator in the AR(1) panel data model is inconsistent for large N and fixed T when individual-specific constants are present. These investigations have also spawned several studies into the effects of cross-sectional dependence on the estimator $\hat{\beta}$. Since this literature is generally concerned with large N panel data models, in all cases, the model includes individual constants that give rise to so-called "Nickel-bias." For example, Phillips and Sul (2007) study the consistency of the pooled OLS estimator of β in an autoregressive version of model (1.1) with $R = 1$ as $N \to \infty$ and $N, T \to \infty$. In this paper, the factor

is assumed to be i.i.d. over t; the loadings are i.i.d. over i and the $\varepsilon_{i,t}$ are weakly exogenous, leading to a simplification of our model (1.1). The conclusion is nonetheless that the factor error induces a random inconsistency that does not vanish with N or T unless $E(\lambda_i) = 0$. Similarly, Sarafidis and Robertson (2008) study the inconsistency of the instrumental variable (IV) estimator of β proposed by Anderson and Hsiao (1981) in a stationary AR(1) model with an error structure that includes $R = 1$ autoregressive factor as $N \to \infty$. They find that the IV estimator is inconsistent and that the inconsistency cannot be eliminated by using further lags to instrument the regressors. For some special cases however, they show that although this operation cannot remove the inconsistency entirely, it can be reduced by cross-sectional demeaning of the data at each t as long as $E(\lambda_i) \neq 0$.

Given that traditional estimators such as OLS, Within-Transformation and GMM estimators have been found to be inconsistent in the presence of factor errors under large N or large N, T asymptotics, we believe it is sensible to examine the consistency properties of such estimators in the fixed N, large T dynamic panel data model. In Chapter 2 of this thesis, we investigate the inconsistency of the OLS, cross-sectionally demeaned and CCE estimators with $R = 1$ in a dynamic panel data model. Whilst this model undoubtedly is a simplification, we leave the factor process of (1.1) unspecified beyond a summability condition on the autocovariance, thereby constituting a considerable generalization of some of the literature concerned with inconsistencies of dynamic panel data estimators which assume the factor is white noise. As we will see, all estimators under scrutiny are inconsistent in this regime unless some very stringent assumptions are placed on the processes constituting the interactive fixed effect ΛF. As a result, these traditional estimators are unsuitable for dynamic policy analysis using (fixed N) country data if factors are present.

Maximum Likelihood Estimation of Model (1.2)

From the presentation of the models in Section 3.1, it should be clear that (1.1), (1.2) and (1.3) are intimately related. Historically however, the literature has focused on the estimation of pure factor models, i.e., model (1.2) and the estimation of the remaining models constitutes a more recent development. Therefore, to give a full account of the issues involved in estimating factor models, we start by reviewing classical and recent methods for the estimation of model (1.2). In classical factor models, it is typically assumed that N is fixed and that only $T \to \infty$ and solutions to this estimation problem go back to at least Anderson and Rubin (1956). Estimation and inference are based on the following simplifying assumptions: $\Sigma_F = I_R$; $F_t \sim N(0, I_R)$ and that $\varepsilon_t \sim N(0, \Omega_\varepsilon)$, see Anderson (2013), Anderson and Rubin (1956) and Lawley and Maxwell (1971). Under these assumptions, $y_t = [y_{1,t}, \ldots, y_{N,t}]'$ is normally distributed with covariance matrix:

$$\begin{aligned} E\left(y_t y_t'\right) &:= \Sigma \\ &= \Lambda\Lambda' + \Omega_\varepsilon \end{aligned}$$

and the concentrated likelihood of the strict factor model is:

$$\mathcal{L}(\Lambda, \Omega_\varepsilon) = -N^{-1}\log\left|\Lambda\Lambda' + \Omega_\varepsilon\right| - N^{-1}\mathrm{tr}\left[\hat{\Sigma}\left(\Lambda\Lambda' + \Omega_\varepsilon\right)^{-1}\right],$$

where $\hat{\Sigma} = T^{-1}\sum_{t=1}^{T} y_t y_t'$. $\mathcal{L}(\Lambda, \Omega_\varepsilon)$ is then jointly maximized with respect to Λ and Ω_ε. It should be noted that the maximum likelihood (ML) estimators do not have an explicit solution and that iterative procedures are required to find the corresponding estimators $\hat{\Lambda}, \hat{\Omega}_\varepsilon$. Now if it is further assumed that $\Lambda\Omega_\varepsilon^{-1}\Lambda'$ is diagonal and $\hat{\Sigma} \to_p \Sigma$, then it can be shown that $\hat{\Lambda}$ and $\hat{\Omega}_\varepsilon$ are consistent, although a consistent estimate of F_t is not available with fixed N. Furthermore, if $\sqrt{T}\left(\hat{\Sigma} - \Sigma\right)$ satisfies a CLT, then the ML estimators have a limiting normal distribution.

An important generalization of the ML estimator is that normality in the errors or the factor component is not necessary and non-normality can be accommodated by interpreting $\mathcal{L}(\Lambda, \Omega_\varepsilon)$ as a quasi-maximum likelihood (QML). For example, in the large N, T framework, Bai and Liao (2012) consider estimation of a model with non-diagonal Ω_ε by QML using regularization of the covariance matrix. Similarly, Doz et al (2011, 2012) model F_t are a finite-order vector autoregression and estimate a large N, T version of model (1.2) using the Kalman Filter.

Principal Components Estimation of Model (1.2)

An alternative estimation procedure for model (1.2) to (Q)ML is the asymptotic principal components (PC) estimator. Consider minimization of the following objective function:

$$\begin{aligned} Q(R) &= (NT)^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T}(y_{i,t} - \lambda_i F_t)^2 \\ &= (NT)^{-1}\,\mathrm{tr}\left[(Y - \Lambda F)(Y - \Lambda F)'\right], \end{aligned} \tag{1.4}$$

where $Y = [y_1, \ldots, y_T]$. Under the normalization $FF'/T = I_R$, it can be shown that the estimate $\hat{F}$ is equal to $\sqrt{T}$ times the eigenvectors corresponding to the R largest eigenvalues of YY' and $\hat{\Lambda} = Y\hat{F}'/T$ is the least-squares projection on the estimated factors. The normalization $FF'/T = I_R$ is useful when $N > T$ and a symmetric normalization and computation exists for $\Lambda'\Lambda/N = I_R$ when $T > N$. A convenient feature of PC is that estimation is computationally less involved than ML because it only involves solving an eigenvalue problem, which is straightforward on modern computers. Of course this does imply that the covariance matrix of ε_t must now be estimated from residuals, requiring more moments of the error processes. For fixed T and large N, Chamberlain and Rothschild (1983) and Connor and Korajczyk (1986) prove that the PC estimator of F is consistent up to a rotation and requires neither diagonal error components Ω_ε nor $\mathrm{cov}(\lambda_i\varepsilon_{i,t}) = 0$ and $\mathrm{cov}(F_t\varepsilon_{i,t}) = 0$, i.e., a generalization to the approximate factor model. However, when $N, T \to \infty$ jointly, since (i) the smallest dimension of the panel now grows to infinity and (ii) the factors and loadings are identified only up to a rotation H, proving consistency and normality requires non-standard machinery. In this scenario, under Assumptions 1, 2 and 3.2, Bai (2003) proves that as $N, T \to \infty$ and $\sqrt{N}/T \to 0$, the PC estimator has the following limit distribution for each $t = 1, \ldots, T$:

$$\sqrt{N}\left(\hat{F}_t - HF_t\right) \to_d N(0, V_F),$$

where V_F is an estimable covariance matrix and H is the rotation matrix induced by the normalization $FF'/T = I_R$. Instead if $\sqrt{N}/T \to c$, for a constant $c > 0$,

$$T\left(\hat{F}_t - HF_t\right) = O_p(1),$$

which gives a rate on the consistency of the estimate of HF_t. When instead $\sqrt{T}/N \to 0$, Bai (2003) further shows that:

$$\sqrt{T}\left(\hat{\lambda}_i - \lambda_i H^{-1}\right) \to_d N(0, V_\Lambda),$$

for each $i = 1, \ldots, N$ and a covariance matrix V_Λ. Similarly, if $\sqrt{T}/N \to c$,

$$N\left(\hat{\lambda}_i - \lambda_i H^{-1}\right) = O_p(1) \tag{1.5}$$

for H following $\Lambda'\Lambda/N = I_R$. These results should be interpreted as follows: whilst it is not possible to estimate consistently the loadings and the factors, it is possible to estimate the factor-loadings *space*, which is practically equivalent to the former for forecasting purposes as the following result shows. That is, as $N, T \to \infty$, regardless of the relative rates, the PC estimator satisfies:

$$\min\left(\sqrt{N}, \sqrt{T}\right)\left(\hat{\lambda}_i\hat{F}_t - \lambda_i F_t\right) \to_d N(0, 1)$$

and this is also the best possible rate, showing that consistency depends on the smallest dimension of the panel (Bai, 2003). In terms of forecasting quality, Tanaka and Kurozumi (2012) present simulation evidence suggesting that PC and MLE of pure factor models perform well even when N is small as measured by the model coefficient of determination, thus offering no clear preference for either estimation procedure.

There are several extensions to the basic limit theory of the PC estimator: first of all, Bai and Ng (2013) derive limiting distributions for other rotation matrices H and inferential procedures about the estimators with the new identification strategies imposed. Second, Choi (2012) notes that the PC estimator is efficient only if Ω_ε is a scalar multiple of the identity matrix and appropriately rescales $Q(R)$ to accommodate non-spherical errors to a "generalized" PC. Under certain restrictions on N and T, Choi (2012) shows that this generalized estimator is more efficient than the ordinary PC estimator and has smaller forecasting errors. Further limit theory extending instead the generating mechanisms of the model include estimation of a factor model with non-stationary factors in Bai (2004); where $R \to \infty$ along with $\min(N, T)$ in Li et al (2017) using the normalization $\Lambda/\sqrt{R}$ and weakly influential factors in De Mol et al (2008) and Onatski (2012, 2015). In these latter papers, Assumption 1 is adjusted and the factor loadings are normalized such $N^\alpha\Lambda'\Lambda \to D$, for a fixed matrix D. Onatski (2012, 2015) considers PC estimation of this model when $\alpha = 0$ and finds that as long as the signal-to-noise is not too low, a consistent estimator exists. Similarly, de Mol et al (2008) consider PC and Bayesian estimators for the model with $0 < \alpha < 1$. Both these papers however restrict attention to the pure factor model and $\varepsilon_{i,t}$ is assumed to be white noise over i and t. Finally, Forni and Lippi (2001) and Forni et al (2000, 2004) present an estimation procedure for the dynamic factor model (1.3)

by dynamic PC. This method is based on PC estimation of the eigenvalues of the spectral density matrix $\Sigma(\theta)$ of $(NT)^{-1}YY'$ at each frequency $\theta \in [-\pi, \pi]$. For the large N, T dynamic factor model, they show that under certain conditions on the lag order L, the Dynamic Principal Components method of Brillinger (1982) remains valid and can be used to filter the data matrix Y. Whilst the factors are assumed to satisfy a version of Assumption 1 at each frequency θ, these authors assume strict exogeneity of the error process with respect to the dynamic factors and $\varepsilon_{i,t}$ is white noise, offering a trade-off in model complexity versus permitted error assumptions.

ML Estimation of Model (1.1)

The development of the PC estimator for the approximate factor model also spurred results in the direction of the interactive fixed effects model: realising that the residual of (1.1) constitutes an approximate factor model, it is intuitive to first estimate β and then extract PC estimates of the factors and/or their loadings from the covariance matrix of $\hat{U} = [\hat{u}_1, \ldots, \hat{u}_N]'$, where $\hat{u}_i = [\hat{u}_{i,1}, \ldots, \hat{u}_{i,T}]'$ is the estimated residual vector of the i-th individual. Following this procedure, Coakley et al (2002) propose augmenting a second-stage regression of the $y_{i,t}$ on the $X_{i,t}$ with the factors extracted from $(NT)^{-1}\hat{U}\hat{U}'$ in a first stage. However, whilst this methodology is straightforward to implement, as Pesaran (2006) notes, if the regressors are correlated with the factors and/or the loadings, it is not possible to consistently estimate the first-stage β, leading to biased estimates of the factors and thus a biased second-stage regression as well. Bai (2009) presents a consistent estimator of β in the presence of correlation between the factors and the regressors, for a model with strictly exogenous regressors, factors and loadings. His estimator jointly solves (1.4) for the factors and their loadings and the following least-squares estimator for β:

$$\hat{\beta}_{IPC} = \left(\sum_{i=1}^{N} X_i' M_{\hat{F}} X_i \right)^{-1} \sum_{i=1}^{N} X_i' M_{\hat{F}} y_i,$$

where X_i is a $T \times P$ matrix of regressors stacked over the time dimension; y_i is a T-vector and $M_{\hat{F}}$ is the orthogonal projection on the space of the estimated factors, which are obtained through PC applied to $(NT)^{-1}\hat{U}\hat{U}'$. Numerically, this method is iterated from several starting points and for this reason we refer to it as Iterative PC (IPC). Under Assumptions 1, 2 and 3.1 and the additional assumptions that $\sum_{i=1}^{N} X_i' M_{\hat{F}} X_i$ has full rank and:

$$(NT)^{1/2} \sum_{i=1}^{N} X_i' M_{\hat{F}} \varepsilon_i \to_d N(0, V),$$

for some variance matrix V, Bai (2009) shows that whilst the estimator for Λ and F preserve the distributional properties of the PC estimator,

$$\sqrt{NT}\left(\hat{\beta}_{IPC} - \beta\right) \to_d N\left(c^{1/2} b_1 + c^{-1/2} b_2, V_{IPC}\right), \tag{1.6}$$

as $N, T \to \infty$ and $T/N \to c$, where b_1 and b_2 are constant vectors and V_{IPC} is a variance matrix. Thus, although the estimator of β is consistent, it is not centred at zero when scaled with $\sqrt{NT}$ if the errors admit heteroskedasticity over i and t and/or auto- or cross-sectional correlation. Bai further shows that the estimator is centred at zero only if either of these correlations is absent and in addition $T/N \to 0$; If $T/N \to c$, Westerlund and Urbain (2015) show that the IPC estimator is biased regardless of any present error correlations. Clearly, the consistency of the estimator depends crucially on the projection $M_{\hat{F}}$, which in turn depends on knowledge of R. As Bai (2009) points out (Remark 4), the $\sqrt{NT}$-consistency is obtained by controlling the space of the factors and this suggests that over-estimating R should not affect the distributional result, although he does not produce theoretical support for this claim. The latter result is obtained by Moon and Weidner (2015) who use eigenvalue-perturbation theory to show that the distribution of β_{IPC} is independent of the number of factors, as long as this number is no smaller than R.

The generalization to the inclusion of weakly exogenous regressors, e.g. lagged dependent variables in X_i, is considered in Moon and Weidner (2017): these authors shows that a third bias term in the spirit of Nickel (1981) enters the distributional result (1.6) additively, i.e. as $c^{1/2}b_3$. They replace Assumption 3.1 with a weak exogeneity assumption that conditions on the sigma-algebra generated by the interactive fixed effects and the history of $\varepsilon_{i,t}$. Bai et al (2009) consider estimation of model (1.1) using a bias-corrected version of the IPC estimator when the F_t and $x_{i,t}$ are (cointegrated) $I(1)$ variables. Song (2013) further extends the IPC estimator to dynamic models with heterogeneous slope parameters and obtains $\sqrt{T}$-consistency of $\beta_{IPC,i}$ for each of $i = 1, \ldots, N$ sets of slope parameters as long as $T/N^2 \to 0$. Finally, Greenaway-Mcgrevy et al (2012) stay closer to the method of Coakley et al (2002): these authors modify the IPC estimator by projecting away not just the factors in u_i, but also those in X_i. As they show, this approach gives consistency of $\hat{\beta}$ but also leads to more stringent conditions on the relative rates of N and T relative to which equation (1.6) holds. In Chapter 4 of this thesis, we will see that in the selection of the number of factors in the interactive fixed effects model, the dimension of the factor space corresponding to the X_i is indeed important, particularly if the estimate of β is inconsistent.

Common Correlated Effects Estimation of Model (1.1)

An alternative approach to IPC for estimation of model (1.1) is the Common Correlated Effects (CCE) estimator of Pesaran (2006). The CCE estimator makes, in our simplified notation, the following parametric assumption on the $x_{i,t}$:

$$x_{i,t} = \Gamma_i F_t + v_{i,t}, \quad i = 1, \ldots, N, t = 1, \ldots, T$$

where Γ_i is the $P \times R$ matrix of factor loadings corresponding to the X_i and $v_{i,t}$ a P-vector of corresponding disturbances. Moreover, Pesaran (2006) considers both heterogeneous and common slope coefficient estimators but we will only consider the latter for brevity.[2] Pesaran endows λ_i and Γ_i with a random coefficient

[2]Pesaran (2006) also uses individual specific constants in both the $x_{i,t}$ and the $y_{i,t}$. As we have argued before, our notation can accommodate this by adding a unit-vector to F_t.

assumption. These parametric assumptions imply that the system:

$$\mathcal{Y}_{i,t} := \begin{bmatrix} y_{i,t} \\ x_{i,t} \end{bmatrix} = \begin{bmatrix} 1 & \beta' \\ 0 & I_P \end{bmatrix} \begin{bmatrix} \lambda_i \\ \Gamma_i \end{bmatrix} F_t + \begin{bmatrix} 1 & \beta' \\ 0 & I_P \end{bmatrix} \begin{bmatrix} \varepsilon_{i,t} \\ v_{i,t} \end{bmatrix},$$

obeys an approximate factor model. Furthermore, under Assumptions 2 and 3.1, if $E(\varepsilon_{i,t}) = E(v_{i,t}) = 0$, this implies that the cross-sectional average of $\mathcal{Y}_{i,t}$ at each t:

$$\begin{aligned} \overline{\mathcal{Y}}_t :=& N^{-1} \sum_{i=1}^{N} \mathcal{Y}_{i,t} = \begin{bmatrix} 1 & \beta' \\ 0 & I_P \end{bmatrix} \begin{bmatrix} \overline{\lambda} \\ \overline{\Gamma} \end{bmatrix} F_t + \begin{bmatrix} 1 & \beta' \\ 0 & I_P \end{bmatrix} \begin{bmatrix} \overline{\varepsilon}_t \\ \overline{v}_t \end{bmatrix} \\ :=& DF_t + o_p(1) \text{ when } N \to \infty \end{aligned}$$

can be used to proxy for the factor space. The averages can be modified to weighted averages subject to some technical conditions. For practical purposes however, the simple average is the obvious choice in all subsequently cited articles. Note that these proxies thus require large N and that no corresponding large-T method exists. That is, the following projection:

$$F_t - \left(D'D\right)^{-1} D' \overline{\mathcal{Y}}_t \to_p 0, \text{ as } N \to \infty. \tag{1.7}$$

is valid as long as the $\text{rank}(D) = R \leq P + 1$. This condition implies that one must have at least $R - 1$ regressors to control for the factors using $\overline{\mathcal{Y}}_t$. Furthermore, 1.7 combined with the Frisch-Waugh Theorem suggests that a consistent estimate of β is available by augmenting a regression of $y_{i,t}$ on $x_{i,t}$ with $\overline{\mathcal{Y}}_t$. The (pooled) CCE estimator is then defined as:

$$\hat{\beta}_{CCE} = \left(\sum_{i=1}^{N} X_i' M_{\overline{\mathcal{Y}}} X_i \right)^{-1} \sum_{i=1}^{N} X_i' M_{\overline{\mathcal{Y}}} y_i,$$

where $M_{\overline{\mathcal{Y}}}$ is the orthogonal projection on the cross-sectional averages at each $t = 1, \ldots, T$ using the generalized inverse of $\overline{\mathcal{Y}}'\overline{\mathcal{Y}}$. Under Assumptions 1, 2 and 3.1 and in addition that the errors of $x_{i,t}$ and $y_{i,t}$ are i.i.d. over i, Pesaran (2006) shows that as $N, T \to \infty$ and $T/N \to 0$:

$$\sqrt{N}\left(\hat{\beta}_{CCE} - \beta\right) \to_d N(0, V_{CCE}),$$

where V_{CCE} is a fixed matrix. It is important to note that this rate is slower than the corresponding rate of $\hat{\beta}_{IPC}$ although consistency of the estimator requires only $N \to \infty$. If T is fixed however, the covariance matrix V_{CCE} depends on nuisance parameters and the bootstrap is required for inference about $\hat{\beta}_{CCE}$. Moreover, when $N, T \to \infty$, Pesaran claims that the rank condition on D is not necessary.

The basic CCE estimator requires independence of the $\varepsilon_{i,t}$ and $v_{i,t}$ and several authors have analysed its properties when this assumption is changed to weakly exogeneous regressors instead. For example, Harding and Lamarche (2011) propose IV estimation based on the CCE estimator with endogenous regressors. De

Groote and Everaert (2016) show that lagged dependent variables induce a Nickel bias in the fixed-T regime, but that the CCE estimator remains consistent when both N and T are large. In De Vos and Everaert (2016), a bias-correction procedure is proposed for the CCE that is valid for T fixed. A general distribution theory for the CCE under weakly exogenous regressors, including lagged dependent variables, with N, T large is given in Chudik and Pesaran (2013), whilst spatial correlation of the errors, but not cross-correlation of the $\varepsilon_{i,t}$ and $v_{i,t}$, is considered in Pesaran and Tosetti (2011). Kapetanios et al (2011) prove consistency and normality as $N, T \to \infty$ when (some of) the F_t contain unit roots.

The CCE has subsequently received a lot of attention because (i) it is easy to compute and, in contrast with IPC-based methods, apparently does not require knowledge of R when $N, T \to \infty$. The estimator is not without its criticism however: first of all, Westerlund and Urbain (2015) show that the CCE estimator is actually biased when $T/N \to c$ under assumptions comparable to those presented here, as is the case with $\hat{\beta}_{IPC}$. More crucially, Urbain and Westerlund (2013) show that the rank condition on D is more critical than initially perceived by Pesaran (2006): if the rank condition is violated, an additional assumption is required that rules out correlation between λ_i and Γ_j for all i and j. This argument implies that one needs enough regressors to control the factor space and, as a result, an estimate of R is at least implicitly required. These authors further make the point that the CCE, under comparable assumptions as the IPC, can only allow for more factors if more regressors are included in the regression equation and that these may not be available or useful depending on the application under consideration. Compared to estimation through IPC, where with enough data of a given set of regressors, one may just extract an additional principal component, the requirement for the CCE may actually be *more* stringent. Karabyik et al (2017) also criticize the CCE for use of the generalized inverse in $M_{\overline{Y}}$. These authors show that the proof techniques on which the consistency and normality of the CCE estimator are based are incorrect when $P > R-1$ because the generalized inverse is not a continuous transformation of the data. They show that in this case additional bias terms persist when $T/N \to c$, although consistency is unaffected whenever $P = R-1$ exactly, thus further stipulating the point that the CCE does require (implicit) knowledge of R.

GMM Estimation of Model (1.1)

The fixed-T GMM literature has also spawned several consistent estimators of β in model (1.1). In the spirit of designing well-behaved moment functions pioneered by Anderson and Hsiao (1981) and Arellano and Bond (1991), these methods first transform the error $u_{i,t}$ to remove the factor structure. That is, one seeks a matrix M such that

$$Mu_i = M\varepsilon_i,\ i = 1, \ldots N.$$

This transformation is known as quasi-differencing and the $(T-R) \times T$ matrix M depends on the context, although the method assumes a transformation between time periods purges the error of the factor structure at the cost of losing estimable equations. Furthermore, M will introduce additional nuisance parameters required for the correction of the error u_i and the dimension of these grows with R and T, so that the method is not suitable for N, T large. In the current setup, these nuisance parameters will be (functions of)

the F_t at each $t = 1, \ldots, T$. As before, a normalization of the factor-loadings space is required to achieve identification. Ahn et al (2013) use the following normalisation:

$$\begin{aligned} F &= \left[F_+', F_-'\right]' \\ &:= \left[F^{\star\prime}, I_R\right]'. \end{aligned} \tag{1.8}$$

As an example, consider the error term of model (1.1) with $R = 1$:

$$\begin{aligned} Mu_i =& \lambda_i \left(f_+/f_- - f_+/f_-\right) + \left(\varepsilon_{i,+} - \varepsilon_{i,-} f_+/f_-\right) \\ :=& \left[I_{T-1}, -f^\star\right] \varepsilon_i := M\varepsilon_i,\ i = 1, \ldots N. \end{aligned}$$

where ε_i is partitioned conformably with F in the second line above. Note how M has removed the factor structure but also introduced temporal dependence and reduced the dimension of the error, so that in this example $T \geq 2$ is required for the transformation to be possible. After transforming the error term, a suitable instrument vector z_i correlated with x_i but orthogonal to Mu_i at each $i = 1, \ldots, N$ is sought to form moment conditions. Assuming an S-vector of strictly exogenous instruments at each t exists, we can form the following moment conditions:

$$E\left(m_i\right) := E\left(z_i \otimes Mu_i\right) = 0_{S(T-R)\times 1}. \tag{1.9}$$

It is then assumed that the empirical analogues of the moment conditions, appropriately scaled, satisfy a Central Limit Theorem:

$$N^{-1/2} \sum_{i=1}^{N} \left(z_i \otimes Mu_i\right) \to_d N\left(0, V\right), \tag{1.10}$$

where V is some full-rank matrix. Equations (1.9) and (1.10) imply that the parameter β may be recovered by GMM. Letting $\phi = \left[\beta', \text{vec}\left(F^\star\right)'\right]'$, a GMM estimator of ϕ solves:

$$Q_{ALS}\left(\phi | R\right) = \underset{\phi \in \Phi}{\text{argmin}} \left[N^{-1} \left(\sum_{i=1}^{N} m_i \right)' \hat{W} \sum_{i=1}^{N} m_i \right], \tag{1.11}$$

where Φ is the parameter space of ϕ, $\hat{W}$ is a weight matrix and the residual vector u_i is computed according to (1.1) above. Ahn et al (2013) propose an algorithm to solve (1.11) in the case of strictly exogenous covariates with homoskedastic $\varepsilon_{i,t}$ which jointly solves for the parameter ϕ using an eigenvalue decomposition. Note however that computing $\hat{\phi}$ as the minimizer of (1.11) with $\hat{W}_1 = I$ and $\hat{V}$ as the empirical variance of (1.10) in a first step and then solving (1.11) again using $\hat{W}_2 = \hat{V}^{-1}$ also leads to an asymptotically efficient quasi-difference GMM (QDGMM) estimator. Ahn et al (2013) argue that $\hat{\phi}$ is consistent and asymptotically normal as $N \to \infty$ by standard arguments in the GMM-estimator literature, although they do not formally prove this proposition and use a different set of assumptions from Assumptions 1-3.1 above. Apart from

a rank condition on the quasi-differenced gradient of $E(m_i)$ and the requirement that $R < N$, they assume that $(\varepsilon_{i,t}, x_{i,t} | \Lambda)$ are cross-sectionally i.i.d. with finite fourth moments over i, allowing for random sampling conditional on the interactive fixed effects. This i.i.d. assumption is not strictly necessary and, as long as the conditioning argument holds, the errors can exhibit (weak) time-series correlation if this structure is known or can be estimated. They further point out that the $x_{i,t}$ may be weakly exogenous, contain lagged dependent variables and even unit roots with T fixed.

So far, the exposition has followed Ahn et al (2013), which is the most general version of QDGMM by allowing estimation in the presence of R interactive fixed effects, although the methodology is by no means new: for example Holtz-Eakin et al (1988) use the method to remove one interactive fixed effect from a large N panel-VAR models; Nuages and Thomas (2003) consider dynamic panel data model estimation with $R = 1$ by QDGMM and Ahn et al (2001) consider an efficient QDGMM estimator with $R = 1$, a special case of Ahn et al (2013). Robertson and Sarafidis (2015) present an alternative Factor IV estimator (FIVE) for the parameters of model (1.1) as $N \to \infty$: instead of quasi-differencing the error term, they propose to estimate the parameters of the following covariance structure

$$E(z_{i,t}\lambda_i) F_t := \delta_t F_t$$

along with β at each t, so that $\Delta = [\delta_1, \ldots, \delta_T]$ and $\delta_t = [\delta_{1,t}, \ldots, \delta_{R,t}]'$ and in our previous notation: $\phi = \left[\beta', \text{vec}(F)', \text{vec}(\Delta)'\right]'$.[3] Using a normalization similar to Ahn et al (2013), they show that this Unidentified FIVE is asymptotically equivalent to QDGMM because it exploits the same of information: where the QDGMM deletes R estimating equations to remove the factors, the FIVE estimator instead estimates additional parameters Δ, so that the degrees of freedom of both estimators are equivalent. Robertson and Sarafidis (2015) further show that $\hat{\beta}$ obtained from FIVE with normalization restrictions imposed has asymptotic distribution equivalent to an estimate without such restrictions imposed. How the resulting generalized inverse of the unrestricted covariance matrix can have a non-degenerate distribution is not discussed however, which, for the same reasons as with the CCE above, is a problem that merits further exploration. Furthermore, under the stronger assumption that $E(\lambda_i \varepsilon_{i,t}) = 0$ for each i and t, these authors also propose an Identified FIVE which, by substitution of a constraint on Δ, yields a smaller estimable parameter vector. As the exploited information remains the same as in the unidentified estimator but the parameter is of lower dimension, it is easy to see that this I-FIVE is more efficient than FIVE and QDGMM. For this reason, the authors further argue that I-FIVE is the most efficient in the class of estimators that make use of second moments. Despite this, FIVE is also not without its problems: for example, Ahn (2015) argues that the estimator of Δ may not be consistent in the presence of cross-sectional heteroskedasticity over the individual Δ_i, say. In that case, consistently estimating Δ_i requires large T, thus inducing an incidental parameter problem which the QDGMM estimator is free of. Ahn (2015) also raises the concern that if F_- in equation (1.8) is a singular matrix and instruments are weakly exogenous, the spanning argument of equivalence between restricted and unrestricted U-FIVE of Robertson and Sarafidis (2015) no longer holds. Whilst a consistent

[3] As with QDGMM, this method is only valid for fixed T, because the dimension of the parameter F diverges with T.

estimate of β is still available, obviously the point estimate of $\delta_t F_t$ in that case is inconsistent. However, this argument is equally valid for the QDGMM estimator, suggesting that further research on the issue is required.

The large N, T estimators of Bai (2009) and Pesaran (2006) generally require strictly exogenous covariates for consistency and asymptotic normality. With dynamic panel data models in mind specifically, these estimators therefore do not suffice unless a bias-correction approach is adopted which invariably depends on knowledge of the structure of the inconsistency. Alternatively, taking a heterogeneous slope parameter approach in the spirit of Pesaran (2006) or Song (2013) with a large N estimator will not do because there is not enough cross-country data available to warrant such an estimation strategy and expect efficiency; nor is a spanning argument parallel to equation (1.7) available with small N and large T. As a result, we take the GMM-route in this thesis: in Chapter 2, we present a simple GMM estimator in the spirit of the CCE that uses the razor-edge transformation of quasi-differencing with cross-sectional averages in the case of a large T dynamic panel data model with a single factor, under the assumption that the data is stationary-ergodic. In Chapter 3, we extend the QDGMM estimator of Ahn et al (2013) to the general fixed N, large T dynamic panel data model with a multi-factor error structure. We show that the QDGMM approach can be used to estimate models with homogeneous and heterogeneous slope parameters and provide a general $\sqrt{T}$-consistency and asymptotic normality theory that requires (a) suitable mixing of the instrument-error process, (b) independence of the factors and the errors and (c) fourth moments of certain underlying processes. In line with the criticism of Ahn (2015) for the FIVE of Robertson and Sarafidis (2015), and as just argued to hold equally for QDGMM in the paragraph immediately above, we provide a limit theory that is in general a variance-mixture of normals whenever $\hat{R} > R$ quasi-differences are used and normally distributed in the special case of when exactly $\hat{R} = R$ quasi-differences are applied to the error. In either case, we show that $\hat{\beta}$ is $\sqrt{T}$-consistent but asymptotic normality of the (full) parameter vector $\hat{\phi}$ obtains only in the latter. This mixed-normality of the QDGMM estimator is new to the literature and can be adapted straightforwardly to hold in the situation studied in the original paper by Ahn et al (2013). The mixed-normality further implies that, if one is unsure about R, standard inference about $\hat{\beta}$ remains valid as long as $\hat{R} \geq R$. As a tangent, our analysis also shows that the part of the Monte Carlo experiment in the original paper by Ahn et al (2013) dealing with inference about $\hat{\beta}$ is invalid because it ignores the impact of the estimated loadings on the joint distribution of the random variable $\sqrt{N}\left(\hat{\phi} - \phi\right)$ and the mixed normality when $\hat{R} > R$.

1.3 Determination of the Number of Factors

Detecting Cross-Sectional Dependence

As we have seen in the previous section, in a model with factor errors, any consistent estimator of β depends crucially on knowledge of R. However, a first step is to detect the presence of cross-sectional dependence.

Sarafidis et al (2009) neatly summarize the problem with the following decision rule:

$$H_0 : E\left(u_{i,t}u_{j,t}\right) = 0 \forall t, i \neq j,$$

versus:

$$H_1 : E\left(u_{i,t}u_{j,t}\right) \neq 0 \text{for some } t \text{and} i \neq j.$$

The literature has spawned several tests that can be used to verify this hypothesis, for example the LM test of Breusch and Pagan (1980):

$$LM = T\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}\hat{\rho}_{i,j}^2,$$

where:

$$\hat{\rho}_{i,j} = \hat{\rho}_{j,i} = \frac{\sum_{t=1}^{T}\hat{u}_{i,t}\hat{u}_{j,t}}{\left(\sum_{t=1}^{T}\hat{u}_{i,t}^2\right)^{1/2}\left(\sum_{t=1}^{T}\hat{u}_{j,t}^2\right)^{1/2}}$$

is the Pearson correlation coefficient calculated from OLS estimates of $u_{i,t}$. Under the null hypothesis of no cross-sectional dependence, the statistic LM is distributed as $\chi^2\left\{N\left(N-1\right)/2\right\}$ as $T \to \infty$ and N fixed and diverges under H_1. Similarly, Pesaran (2004) uses this test in the large N, T framework and shows that:

$$CD = \sqrt{\frac{2T}{N\left(N-1\right)}}\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}\hat{\rho}_{i,j} \to_d N\left(0,1\right)$$

under the null hypothesis of cross-sectional independence. Pesaran (2004) shows that this test holds under a variety of assumptions and model specifications and Monte Carlo simulations shows that the small sample performance of CD is acceptable. However, note that Pesaran's generalization will lose power when N is large if (i) the λ_i are i.i.d. with mean-zero over i, or (ii) if the loadings are i.i.d. with unspecified mean and time effects are added to the regression model.

Sarafidis et al (2009) propose a test for cross-sectional independence based on Sargan's difference test in the GMM framework with $N \to \infty$ and T fixed. For difference and system-GMM estimators, c.f. Anderson and Hsiao (1981) and Arellano and Bond (1991), these authors specify two sets of moment conditions using transformations of the data which hold under the null of no cross-sectional correlation. Under the alternative hypothesis, only one of the sets of moment conditions hold so that the test has power against the alternative. They show that standard J-tests and difference-in-J-tests follow a chi-squared distribution under the null. To nest the tests in the null hypothesis however, these tests crucially rely on the assumption that the instruments, i.e. lags of $y_{i,t}$ and $x_{i,t}$, become cross-sectionally uncorrelated after demeaning these variables over the cross-section. This requirement is equivalent to stating that $\text{cov}\left(\Lambda_i, x_{j,t}\right) = 0$ for all i, j and this seems a rather artificial additional requirement over Assumption 3.1. A simpler approach to testing for cross-sectional dependence can instead be based on the fact that $Q_{ALS}\left(\phi|R\right)$ is asymptotically distributed as chi-squared with $\left(S-R\right)\left(N-R\right)$ degrees of freedom. That is, one can easily test if the GMM estimator with no quasi-differences rejects the null of no cross-sectional correlation, *viz.* acceptance of the test after

estimating β with $\hat{R}$ quasi-differences, holding constant the instruments in both setups.

Determination of the Number of Factors

After detecting the presence of cross-sectional dependence, a researcher has to make a decision on R: although in certain cases the number of factors may be known, we expect that in general R has to be estimated. There is now a large literature on consistent estimation of R particularly for large dimensional factor version of model (1.2), yet there is very little on the same problem in the interactive fixed effects model (1.1). The problem of determining R in a pure factor model goes back to Cattel (1966), who is credited with the invention of the Scree plot.[4] Cattel's recommendation is that one should retain only the $\hat{R}$ eigenvalues to the left of the inflection point of the Scree plot. Of course, determination of R based on inspection of the Scree plot is informal and subject to researcher scrutiny. However, as we will see in Chapter 4, this method is nonetheless a consistent way of determining R in a factor model with both N and T large, although formal proofs of the method are a much more recent phenomenon.

Under the assumption of a strict factor model with either N or T fixed, it is also possible to examine the likelihood of a model with k and $k+1$ estimated factors and construct corresponding likelihood-ratio tests, see Anderson (1984), Section 14.3.2. However, as is pointed out in Connor and Korajczyk (1993), such tests depend too strongly on the normality assumption of the strict factor model and therefore may have poor properties in practice. For the approximate factor model with T fixed, these authors instead assume that the disturbance $\varepsilon_{i,t}$ satisfies a cross-sectional mixing condition but is i.i.d. in the time dimension. They then use a sub-sampling scheme to split their data and show that under the null of k factors, as $N \to \infty$,

$$CK = 2N^{1/2}T^{-1}\sum_{t=1}^{T/2}\sum_{i=1}^{N}\left(\hat{u}_{i,2t}^2 - \hat{u}_{i,2t-1}^2\right) \to_d N(0,V),$$

where the $\hat{u}_{i,t}$ are computed after subtracting k principal components from the data. Under the alternative hypothesis, the statistic diverges to a positive constant. However, the problem with CK is that the support of the normal distribution is unbounded and as a result, with T finite it is possible to give examples where the influence of the k-th factor is not detectable using CK.

More recently the literature has cast its attention to the estimation of large dimensional factor models and several authors have proposed consistent estimators of R in the approximate factor model (1.2). The information criteria (IC) of Bai and Ng (2002) have proven to be a particularly popular method in empirical work. Bai and Ng (2002) show that we can determine R consistently by minimizing:

$$\hat{R}_{IC}: \begin{cases} PC_{\hat{R}} = \underset{k}{\operatorname{argmin}}\, Q(k) + \hat{\sigma}^2 k p(N,T); \\ IC_{\hat{R}} = \underset{k}{\operatorname{argmin}} \log\left[Q(k)\right] + k p(N,T), \end{cases}$$

[4] The Scree plot is a graphical device that denotes the eigenvalues of a covariance matrix ordered from largest to smallest.

where $Q(k)$ is the objective function defined in equation (1.4), calculated using a generic number of factors k; $\hat{\sigma}^2$ is an estimate of the model variance $(NT)^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T}\varepsilon_{i,t}^2$ and $p(N,T)$ is a penalty function that depends only on N,T. The consistency of these IC follows from a result by Pötscher (1983), which links the penalty function to the convergence rate of chi-squared random variables. Pötscher shows that necessary and sufficient conditions on the penalty function are, as $N,T\to\infty$, that (i) $p(N,T)\to 0$ and (ii) $\min(N,T)\times p(N,T)\to\infty$, where the latter condition is adapted to the large N, T framework following Theorem 2 in Bai and Ng (2002). Provided that the penalty function satisfies these conditions, the resulting probabilistic argument then yields a situation in which the probability of under- and overestimating R converges to zero with the sample size. That is, the penalty function approximates the mean of the chi-squared distribution as the significance level α_T goes to zero with T. This machinery is remarkably general and applied in virtually all circumstances of model selection to justify consistency of IC, for example: ARMA-order selection in Pötscher's original paper; rank estimation in Cragg and Donald (1997); instrument selection in GMM models in Andrews and Lu (2001) and, indeed, determination of the number of factors in Ahn, Lee and Schmidt (2014) and Robertson and Sarafidis (2015) in the large N, small T setting. It should be noted that this consistency result is asymptotic and that any $p(N,T)$ satisfying Pötscher's condition is valid. In the paper of Bai and Ng, the penalty term is either $(N+T)/(NT)\log[(NT)/(N+T)]$ or $(N+T)/(NT)\log[\min(N,T)]$. The logarithmic penalty originates from Schwarz (1978), where an IC was derived as a second-order expansion to the likelihood of an AR model with uninformative prior. Note that *viz.* the former arguments, in a factor model, this choice appears to be arbitrary other than that it apparently works well in practice. However, especially with $IC_{\hat{R}}$, the log penalty does have a theoretical justification in the sense that $Q(k)$ is a scaled sum of $\min(N,T)-k$ chi-squared random variables under the assumption of the strict factor model with unitary variances. Then, since the limit of $\exp[p(N,T)]$ as specified in Bai and Ng (2002) is unity, the penalty function corresponds on average to the correct critical value required for the removal of a chi-squared variable going from R to $R+1$ factors. When the strict factor assumption does not hold however, this argument breaks down because $Q(k)$ is now a non-central chi-squared variable. This, together with the fact that any penalty function satisfying Pötscher's condition is consistent, is likely the reason why the performance of the IC of Bai and Ng (2002) is often found to be less than satisfactory in Monte Carlo studies. This problem can at least in part be amended in the case where $\max\left(\sqrt{N},\sqrt{T}\right)/\min(N,T)\to 0$ by using the objective function of the Generalized Principal Components estimator of Choi (2012) instead of (1.4), as the objective function of the Generalized Principal Components is normalized to a sum of unitary variances. To our knowledge however, this strategy has not been investigated further in the literature.

As an extension to the IC of Bai and Ng (2002), several authors have proposed group lasso algorithms to determination of R in the approximate factor model. Hirose and Konishi (2013) present a group-lasso estimator for fixed T, large N in the strict factor model but do not prove consistency of the method. For the large N,T model, Caner and Han (2014) develop a group bridge algorithm that simultaneously estimates the factor space and the number of factors. For the objective function (1.4) with normalization $FF'=I_R$,

their penalty function is of the form:

$$p(N,T) = \frac{\psi}{\min(N,T)} \sum_{k=1}^{R_{\max}} N^{-1} \left(\sum_{i=1}^{N} \lambda_{r,i}^2 \right)^c,$$

where ψ is a constant and $0 < c < 1/2$ that depends on the data. Note that this particular $p(N,T)$ is a soft-threshold operator, forcing the loadings of the k-th factor to zero whenever a function of their norm is less than ψ. Under Assumptions 1, 2, 3.2 and the additional regularity assumptions on the penalty term that $\psi/\min\left(\sqrt{N},\sqrt{T}\right) \to 0$ as $N,T \to \infty$ and $\psi/\min(N^c,T^c) \to \infty$ as $N \to \infty$, Caner and Han (2014) show that the group bridge estimator satisfies (i) $\hat{R} \to_p R$ and (ii) that the estimated $\lambda_{r,i}$, $\lambda_{r,i}^\star$ satisfy:

$$N^{-1} \left\| \lambda_{r,i} - \lambda_{r,i}^\star \right\|^2 = O_p\left[\max\left(N^{-1},T^{-1}\right) \right].$$

Comparing the above to equation (1.5), we see that estimation by the group bridge estimator preserves the convergence rate of Bai and Ng (2002) when R is known, a manifestation of the so-called "oracle property" of shrinkage-estimation. Presumably, by appropriately choosing ψ, the aforementioned problem with the IC of Bai and Ng (2002) is alleviated, but it is not clear how the prescribed choice of ψ accomplishes this in practice. Similarly, smudging of c by the researcher is possible.

Hallin and Liška (2007) provide an alternative solution to the unsatisfactory performance of the IC of Bai and Ng: their key insight is that for any generic consistent IC, a correspondingly consistent *IC* can be defined as:

$$IC = \underset{k}{\operatorname{argmin}}\, Q(k) + kcp(N,T),$$

where c is some constant. For the dynamic factor model (1.3), they show that R can be consistently estimated by means of sub-sampling: using partitions of the data over N and T, they propose to conduct a grid-search over c to find so-called "stability regions." Such stability regions correspond to values of c where $\hat{R}$ is constant as measured by the empirical variance of $\hat{R}$ over the sub-samples. Hallin and Liška show that the first stability region beyond $k = R_{\max}$, for $R_{\max} \gg R$, corresponds to the true number of factors and Alessi et al (2010) subsequently apply this method to the approximate factor model (1.2) but do not prove consistency.

As we will see in Chapter 4, there is a Random Matrix Theory argument to justify the method of Hallin and Liška (2007) and several other authors have exploited a version of this argument to justify their estimators of R. In these studies, Assumption 2 is replaced by an assumption that the model error is generated as:

$$\varepsilon = G^{1/2} u H^{1/2},$$

where, u is an $N \times T$ matrix consisting of i.i.d. random variables $u_{i,t}$ with variance σ_u^2 and bounded fourth moments and G and H are (possibly random) matrices controlling cross-sectional and time series dependence with finite maximal eigenvalues. Under these assumptions the eigenvalues of $\min\left(N^{-1},T^{-1}\right) uu'$ will tend to the Marchenko-Pastur (1967, MP) distribution for which closed-form expressions are known, whilst

the distribution of the eigenvalues of the general error process is not known. Letting $\xi_1 \geq \cdots \geq \xi_{\min(N,T)}$ denote the eigenvalues of $(NT)^{-1} YY'$ ordered from largest to smallest, it can be shown that the contribution of the error covariance matrix to the first R eigenvalues is negligible, whilst the remaining $R+1, \ldots, \min(N, T)$ eigenvalues are fully determined by the covariance matrix of the ε as $N, T \to \infty$. Furthermore, under Assumptions 1 and 3 and the functional form of the error matrix ε replacing Assumption 2, it can be shown that $\|\Lambda F\|^2 = O_p(NT)$, whilst the $\|\varepsilon\|^2 = o_p(NT)$ as $N, T \to \infty$. As we show in Chapter 4, this result describes the conditions under which the eigenvalues of a factor model exactly separate as originally foreseen by Catell in the Scree plot.

Several authors have used this separation result to develop consistent estimators of R. For example, Onatski (2010) adds the additional assumption that the eigenvalues of the matrices G and H satisfy non-degenerate distributions with bounded support. Using this further assumption he shows that the distribution of the eigenvalues of $\min\left(N^{-1}, T^{-1}\right) \varepsilon\varepsilon'$ almost surely is non-degenerate and has bounded support. Since the number of eigenvalues in the error covariance matrix is large, he argues that the distance between any eigenvalue $\delta = \xi_k - \xi_{k+1}$ for $k = R+1, \ldots, \min(N, T) - 1$ should be stable in the error eigenvalues whilst this distance diverges when $k = R$. Onatski then proposes an algorithm that can be used to estimate this cut-off value δ based on OLS regression of a number of adjacent eigenvalues on the constant and the cardinality of eigenvalues raised to a certain power. The algorithm consists of iterating the former OLS regression starting at $R_{\max}$ and stopping whenever:

$$\hat{R}_{EDGE} = \underset{k}{\operatorname{argmax}} \left\{ k \leq R_{\max} : \hat{\xi}_k - \hat{\xi}_{k+1} \geq \hat{\delta} \right\}.$$

Onatski (2010) shows that this algorithm is consistent under the strong factor regime if $R_{\max}/\min(N, T)$ $\to 0$ and Onatski (2012) argues that it may be consistent under a weak factor structure too. We note that it is however not clear if the OLS regression estimate of the edge is consistent and Onatski does not motivate it further than being based on the "square-root behaviour" of the eigenvalues at the edge of the support of the spectral distribution. Irrespective of this caveat, the estimator is shown to perform well in Monte Carlo studies under a variety of conditions on the factor process and the error process, although it is somewhat sensitive to the implementation of the method as can be seen in the Monte Carlo studies in Ahn and Horenstein (2013), for example. These implementation issues are due to the choice of $R_{\max}$ in the algorithm and as a result the estimator can get stuck at some $k > R$.

Ahn and Horenstein (2013) also use the facts that $\|\Lambda F\|^2 = O_p(NT)$, whilst the $\|\varepsilon\|^2 = o_p(NT)$ as $N, T \to \infty$ without making distributional assumptions on G and H to show consistency of the following

tests:

$$\hat{R}_{ER} = \underset{k}{\operatorname{argmax}} \left\{ \hat{\xi}_k / \hat{\xi}_{k+1} \right\};$$

$$\hat{R}_{GR} = \underset{k}{\operatorname{argmax}} \left\{ \frac{\log\left(1+\hat{\xi}_k/Q(k)\right)}{\log\left(1+\hat{\xi}_{k+1}/Q(k+1)\right)} \right\}, k = 1,\ldots,R_{\max}.$$

Intuitively, both these tests operate on the observation that the ratio of the last eigenvalue of the factor matrix to the first eigenvalue of the residual matrix will diverge, whilst all other ratios are expected to be constant. Monte Carlo experiments show that these simple tests prove to be remarkably robust, although it is possible to conceive examples where the tests do not work in finite samples: for example, if one or more factors have extremely low or high variance so that the ratio tests explode at some $k < R$ in finite samples. Similarly, it is possible that the test cannot distinguish a maximal ratio at R when the signal-to-noise ratio is very low.

Consistent estimation of R is an area of ongoing research and we make several additions to this literature in this thesis. In Chapter 3, we follow Ahn et al (2013) and Robertson and Sarafidis (2015) by showing that IC based on the objective function (1.9) above are consistent for the QDGMM estimator applied to the large T dynamic panel data model, with or without efficient weight matrices. We make the critical observation that the problem is more complicated than just the determination of R and in fact entails jointly testing all aspects of the GMM model. That is, in addition to the number of factors, also the validity of the regressors and the included instruments. We show that this more general problem is however supported by the usual consistency argument of IC. Moreover, Monte Carlo evidence verifies that especially IC based on Bayes Criterion, albeit scaled with some data-dependent constant, have good finite sample properties. We also show that rank tests as in Cragg and Donald (1993), Kleibergen and Paap (2006) and Al-Sadoon (2017) can be fruitfully applied to obtain a consistent estimate of R. On the other hand, in Chapter 4 we abandon the fixed N model in favour of the large N, large T model and explore determination of R in both models (1.1) and (2.1) above. In that chapter we show that consistent estimation of R corresponds to the problem of finding the elbow in the Scree plot. We show that this methodology is consistent under more general circumstances than just the approximate factor model under Assumptions 1, 2 and the adjusted functional form of ε above and derive formally eigenvalue separation phenomena in appropriately scaled weak factor models and in a model where in addition to N, T, also $R \to \infty$ such that $R/\min(N,T) \to c$. Moreover, we show that the methodology of Hallin and Liška (2007) also fits in the class of estimators based on the elbow of the Scree plot. We then present two new estimators of R and a Lasso-algorithm in the spirit of Caner and Han (2014), that sets the constant ψ equal to a bound on the largest singular value of the error covariance matrix $(NT)^{-1}\varepsilon\varepsilon'$, thus nesting the Lasso in the class of separation estimators rather than only offering an asymptotic justification. Finally, we consider determination of R in the interactive fixed effects model (1.1) with N and T large. To the best of our knowledge, this problem has been so far overlooked in the literature and we show that eigenvalue separation continues to hold in this case, even when the estimate of β is inconsistent. In that case, estimating R becomes equivalent to testing $\max(R_Y, R_X)$, where the distinction

depends on which of the $x_{i,t}$ and $y_{i,t}$ variables contain the most factors. Since many estimators of β are consistent whenever $\hat{R} \geq R$, this leads to an iterative testing procedure that will consistently determine R regardless of whether the initial estimate of β is consistent or not.

1.4 Conclusions

As we have seen, the econometric literature has spent much attention on developing estimation and inferential procedures for factor models and regression models with interactive fixed effects, particularly in the large N and large N, T frameworks. On the other hand, whilst a number of macroeconomists have used panel data to analyse fiscal policy inquiries, the consistency properties of traditional estimators in the presence of cross-sectional dependence are not well understood in large T panels. This thesis is intended to begin and fill this void: in the following three chapters, we first verify that traditional estimators of the parameters of an interactive fixed effects model are inconsistent in the large T model; we present a comprehensive estimation procedure for the dynamic interactive fixed effects model and develop several tests that can be use to consistently estimate the number of factors both in the large N and large N, T models. In passing, we make several contributions to the econometric literature and we mention the most important ones: first, the econometric theory of Chapter 3 verifies results anticipated in the original paper by Ahn et al (2013) in addition to presenting some inconsistencies in that paper. Importantly, we show that inference about $\hat{\beta}$ continues to be standard if the number of factors is over-estimated. Second, in Chapter 4, we show that any consistent estimator of R based on eigenvalue separation belongs in a class of estimators we refer to as "Scree plot estimators." In that paper, we also show that Scree plot estimators can be applied to the interactive fixed effects model.

Bibliography

[1] Ahn, S. (2015): " Comment on 'IV Estimation of Panels with Factor Residuals'," *Journal of Econometrics*, Vol. 185, pp. 542-544, Elsevier North Holland.

[2] Ahn, S. and Horenstein, A. (2013): "Eigenvalue Ratio Test for the Number of Factors," *Econometrica*, Vol. 83, No. 3, pp. 1203-1227, Econometric Society, Wiley-Blackwell.

[3] Ahn, S. Lee, H. and Schmidt, P. (2001): "GMM Estimation of Linear Panel Data Models with Time-Varying Individual Effects," *Journal of Econometrics*, Vol. 101, Issue 2, pp. 219–255, Elsevier North Holland.

[4] Ahn, S. Lee, H. and Schmidt, P. (2014): "Panel Data Models with Multiple Time-Varying Individual Effects," *Journal of Econometrics*, Vol. 1174, Issue 1, pp. 1-14, Elsevier North Holland.

[5] Alessi, L., Barigozzi, M. and Capasso, M. (2010): "Improved Penalization for Determining the Number of Factors in Approximate Factor Models," *Statistics and Probability Letters*, Vol. 80 Issue 23-24, #1-15, pp. 1806-1813, Elsevier North Holland.

[6] Al-Sadoon, M. (2017): "A Unifying Theory of Tests of Rank," *Journal of Econometrics*, Vol. 199, Issue 1, pp. 47-62, Elsevier North Holland.

[7] Anderson, T. (1984): "*An Introduction to Multivariate Statistical Analysis*," Wiley, New York.

[8] Anderson, T. W., and Hsiao C. (1981): "Estimation of Dynamic Models with Error Components," *Journal of the American Statistical Association*, Vol. 76, pp. 598–606, American Statistical Association.

[9] Anderson, T. and Rubin, H. (1956): "Statistical Inference in Factor Analysis," in: *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*: *Contributions to the Theory of Statistics*, University of California Press.

[10] Andrews, D. and Lu, B. (2001): "Consistent Model and Moment Selection Procedures for GMM Estimation with Application to Dynamic Panel Data Models," *Journal of Econometrics*, Vol. 101, Issue 1, pp. 123–164, Elsevier North Holland.

[11] Arellano, M. and Bond, S. (1991): Some Tests of Specification for Panel Data: Monte Carlo Evidence and an Application to Employment Equations," *Review of Economic Studies*, Vol. 58 pp. 277–297, Oxford University Press.

[12] Bai, J. (2003): "Inferential Theory for Factor Models of Large Dimensions," *Econometrica*, Vol. 71 Issue, 1 pp.135–172, Econometric Society, Wiley-Blackwell.

[13] Bai, J. (2004), "Estimating Cross-Section Common Stochastic Trends in Non-Stationary Panel Data," *Journal of Econometrics*, Vol. 122, pp.137-183, Elsevier North Holland.

[14] Bai, J. (2009): "Panel Data Models with Interactive Fixed Effects," *Econometrica*, Vol. 77, Issue 4, pp. 1229-1279, Econometric Society, Wiley-Blackwell.

[15] Bai, J., Kao, C. and Ng, S. (2009): "Panel Cointegration with Global Stochastic Trends," *Journal of Econometrics*, Vol. 149, Issue 1, pp. 82-99, Elsevier North Holland.

[16] Bai, J. and Liao, Y. (2012): "Efficient Estimation of Approximate Factor Models via Regularized Maximum Likelihood," mimeo.

[17] Bai J. and Ng, S. (2002): "Determining the Number of Factors in Approximate Factor Models," *Econometrica*, Vol. 70, No. 1, pp. 191-221, Econometric Society, Wiley-Blackwell.

[18] Bai, J, and Ng, S. (2013): “Principal Components Estimation and Identification of Static Factors,” *Journal of Econometrics*, Vol. 176, Issue 1, pp. 18–29, Elsevier North Holland.

[19] Bai, J. and Wang, P. (2016): “Econometric Analysis of Large Factor Models,” *Annual Review of Economics*, Vol. 8 Issue 1, pp.53-80, Annual Reviews.

[20] Beetsma, R. and Guiliodori, M. (2010a): "The Macroeconomic Costs and Benefits of the EMU and Other Monetary Unions: An Overview of Recent Research," *Journal of Economic Literature*, Vol. 48, pp. 603-641, Elsevier.

[21] Beetsma, R. and Guiliodori, M. (2010b): "Fiscal Adjustment to Cyclical Developments in the OECD: an Empirical Analysis based on Real-Time Data," *Oxford Economic Papers 62*, pp. 419-441, Oxford University Press.

[22] Beetsma, R. and Giuliodori, M. (2011): "The Effects of Government Purchases Shocks: Review and Estimates for the EU," *Economic Journal*, Vol. 121 #550, pp. F4-F32, Royal Economic Society, Wiley-Blackwell.

[23] Bénétrix, A. S. and Lane, P. (2013): "Fiscal Cyclicality and EMU," *Journal of International Money and Finance*, Vol. 34, pp. 164-176, Elsevier

[24] Breusch, T. and Pagan, A. (1980): “The Lagrange Multiplier Test and its Application to Model specifications in Econometrics,” *Review of Economic Studies*, Vol. 47, pp. 239–253, The Review of Economic Studies Ltd.

[25] Brillinger, R. (1982): *Time Series: Data Analysis and Theory*, Holt, Rinehart, and Winston, New York.

[26] Caner, M. and Han, X. (2014): "Selecting the Correct Number of Factors," *Journal of Business and Economic Statistics,* Vol. 32, No. 3 pp. 359-374, Taylor & Francis Group.

[27] Cattell, R (1966). "The Scree Test for the Number of Factors," *Multivariate Behavioural Research*, Vol. 1, pp. 245-276, Taylor & Francis Group .

[28] Chamberlain, G. and Rothschild, M. (1983): “Arbitrage, Factor Structure and Mean-Variance Analysis on Large Asset Markets,” *Econometrica*, Vol. 51 No. 5, pp 1281-1304, Wiley-Blackwell.

[29] Choi, I. (2012): “Efficient Estimation of Factor Models,” *Econometric Theory*, Vol. 28, pp. 274-308, Cambridge University Press.

[30] Chudik, A. and Pesaran, M. H. (2013): "Common Correlated Effects Estimation of Heterogeneous Dynamic Panel Data Models with Weakly Exogenous Regressors," *Globalization and Monetary Policy Institute Working Paper*, No. 146, Federal Reserve Bank of Dallas.

[31] Coakley, J., Fuertes, A. and Smith, R. (2002): "A Principal Components Approach to Cross-Section Dependence in Panels," *10th International Conference on Panel Data*, Berlin, International Conferences on Panel Data.

[32] Connor G. and Korajczyk, R. (1986): "Performance Measurement with the Arbitrage Pricing Theory," *Journal of Financial Economics* Vol. 15, pp. 373-394, Elsevier North Holland.

[33] Connor, G. and Korajczyk, R. (1993): "A Test for the Number of Factors in an Approximate Factor Model." *Journal of Finance*, Vol. 48, pp. 1263-1291, American Finance Association, Wiley-Blackwell.

[34] Cragg, G. and Donald, S. (1993): "Testing Identifiability and Specification in Instrumental Variables models," *Econometric Theory*, Vol. 9, pp. 222-240, Cambridge University Press.

[35] Cragg, G. and Donald, S. (1997): "Inferring the Rank of a Matrix," *Journal of Econometrics*, Vol. 76, Issue 1-2, pp. 223-250, Elsevier North Holland.

[36] De Groote T. and Everaert, G. (2016): "Common Correlated Effects Estimation of Dynamic Panels with Cross-Sectional Dependence," *Econometric Reviews*, Vol. 35, pp. 428-463, Taylor and Francis.

[37] De Mol, C., Giannone, D. and Reichlin, L. (2008): "Forecasting Using a Large Number of Predictors: Is Bayesian Shrinkage a Valid Alternative to Principal Components?" *Journal of Econometrics*, Vol. 146 , pp. 318–328, Elsevier North Holland.

[38] De Vos, I. and Everaert, G. (2016): "Bias-Corrected Common Correlated Effects Pooled Estimation in Homogeneous Dynamic Panels," *Working Papers of Faculty of Economics and Business Administration*, Ghent University.

[39] Doz, C., Giannone, D. and Reichlin, L. (2011): "A Two-Step Estimator for Large Approximate Dynamic Factor Models Based on Kalman Filtering," *Journal of Econometrics* Vol. 164 Issue 1 pp. 188–205, Elsevier North Holland.

[40] Doz, C., Giannone, D. and Reichlin, L. (2012): "A Quasi-Maximum Likelihood Approach for Large Approximate Dynamic Factor Models," *The Review of Economics and Statistics*, Vol. 94 Issue 4, pp. 1014– 1024, MIT Press.

[41] Forni, M., and Lippi, M. (2001): "The Generalized Dynamic Factor Model: Representation Theory," *Econometric Theory*, Vol. 17, Cambridge University Press.

[42] Forni, M., Hallin, M., Lippi, M. and Reichlin, L. (2000): "The Generalized Dynamic Factor Model: Identification and Estimation," *The Review of Economics and Statistics*, Vol. 82 Issue 4 pp. 540–554, MIT Press.

[43] Forni, M., Hallin, M., Lippi, M. and Reichlin, L. (2004): The Generalized Dynamic Factor Model: Consistency and Rates," *Journal of Econometrics*, Vol. 119, pp. 231–255, Elsevier North Holland.

[44] Greenaway-McGrevy, R., Han, C. and Sul, D. (2012): "Asymptotic Distribution of Factor Augmented Estimators for Panel Regression," *Journal of Econometrics*, Vol. 169, pp. 48-53, Elsevier North Holland.

[45] Hallin, M. and Liška, R. (2007): "Determining the Number of Factors in the General Dynamic Factor Model," *Journal of the American Statistical Association,* Vol. 102, No. 478, pp. 603-617, Taylor & Francis Group.

[46] Harding, M. and Lamarche, C. (2011): "Least Squares Estimation of a Panel Data Model with Multifactor Error Structure and Endogenous Covariates", *Economics Letters*, Vol. 111, pp. 197-199, Elsevier North Holland.

[47] Hirose, K. and Konishi, S. (2013): "Variable Selection via the Weighted Group LASSO for Factor Analysis Models," *The Canadian Journal of Statistics*, Vol. 40, pp. 345-361, Wiley-Blackwell.

[48] Holtz-Eakin, D., Newey, W. and H. S. Rosen (1988): "Estimating Vector Autoregressions with Panel data," *Econometrica*, Vol. 56 Issue 6, pp. 1371–1395,Econometric Society, Wiley-Blackwell.

[49] Ilzetzki, E. (2011): "Fiscal Policy and Debt dynamics in Developing Countries," *Policy Research Working Paper Series*, No. 5666, The World Bank.

[50] Ilzetzki, E., Mendoza, E. G. and Végh, C. A. (2013): "How Big (Small?) are Fiscal Multipliers?," *Journal of Monetary Economics*, Vol. 60 #2, pp. 239-254, Elsevier North Holland

[51] Kapetanios, G., Pesaran, M. H. and Yamagata, T. (2011): "Panels with Non-Stationary Multifactor Error Structures," *Journal of Econometrics*, Vol. 160 Issue 2, Pp. 326-348, Elsevier North Holland.

[52] Karabiyik, H., Reese, S. and Westerlund, J. (2017): "On the Role of the Rank Condition in CCE Estimation of Factor-Augmented Regressions," *Journal of Econometrics*, Vol. 197, pp. 60-64, Elsevier North Holland.

[53] Kiviet, J. (1995): "On Bias, Inconsistency, and Efficiency of Various Estimators in Dynamic Panel Data Models," *Journal of Econometrics,* Vol. 68, pp. 53-78, Elsevier North Holland.

[54] Kleinbergen, F. and Paap, R. (2006): "Generalized Reduced Rank Tests using the Singular Value Decomposition," *Journal of Econometrics*, Vol. 133, Issue 1, pp. 97-126, Elsevier North Holland.

[55] Lawley D. and Maxwell, A. (1971): *Factor Analysis as a Statistical Method*, American Elsevier Publishing Company, New York.

[56] Li, H., Li, Q. and Shi, Y. (2017): "Determining the Number of Factors when the Number of Factors can Increase with Sample Size," *Journal of Econometrics*, Vol. 197, pp. 76-86, Elsevier North Holland.

[57] Marchenko,V. and Pastur, L. (1967): "Distribution of eigenvalues for some sets of random matrices", *Mathematics of the USSR-Sbornik*, Vol. 72 No. (1-4), pp. 507–536.

[58] Moon, H. R. and Weidner, M. (2015): "Linear Regression for Panel with Unknown Number of Factors as Interactive Fixed Effects," *Econometrica*, Vol. 83, Issue 4, pp. 1543-1579, Econometric Society, Wiley-Blackwell.

[59] Moon, H. R. and Weidner, M. (2017): "Dynamic Linear Panel Regression Models with Interactive Fixed Effects," *Econometric Theory*, Vol. 33, Issue 1, pp. 158-195, Cambridge University Press.

[60] Nickel, S., (1981): "Biases in Dynamic Models with Fixed Effects," *Econometrica*, Vol. 49, Issue 6, pp. 1417-26, Econometric Society, Wiley-Blackwell.

[61] Nauges, C. and Thomas, A. (2003): "Consistent Estimation of Dynamic Panel Data Models with Time-Varying Individual Effects," *Annales d'Economie et de Statistique*, Issue 70, pp. 53-75, ENSAE.

[62] Onatski, A. (2010): "Determining the Number of Factors from Empirical Distribution of Eigenvalues," *The Review of Economics and Statistics*, Vol. 92 No. 4, pp. 1004–1016, MIT Press.

[63] Onatski, A. (2012): "Asymptotics of the Principal Components Estimator of Large Factor Models of Weakly Influential Factors," *Journal of Econometrics*, Vol. 168, pp. 244-258, Elsevier North Holland.

[64] Onatski, A. (2015): "Asymptotic Analysis of the Squared Estimation Error in Misspecified Factor Models," *Journal of Econometrics*, Vol. 186 Issue 2, pp. 388–406, Elsevier North Holland.

[65] Pesaran, M. (2004): "General Diagnostic Tests for Cross-Section Dependence in Panels," mimeo *University of Cambridge*.

[66] Pesaran M. H. (2006): "Estimation and Inference in Large Heterogeneous Panels with a Multifactor Error Structure," *Econometrica*, Vol. 74, pp. 967-1012, Econometric Society, Wiley-Blackwell.

[67] Pesaran, M. H, and Tosetti, E. (2011): "Large Panels with Common Factors and Spatial Correlation," *Journal of Econometrics*, Vol. 161 Issue 2, Pp. 182-202, Elsevier North Holland.

[68] Phillips, P. C.B. and Sul, D. (2007) "Bias in Dynamic Panel Estimation with Fixed Effects, Incidental Trends and Cross-section Dependence," *Journal of Econometrics,* Vol. 137 Issue 1, pp 162-188, Elsevier North Holland.

[69] Pötscher, B. (1983): "Order estimation in ARMA-models by Lagrangian multiplier tests," *Annals of Statistics,* Vol. 11, pp. 872-885, Institute of Mathematical Statistics.

[70] Robertson, D. and Sarafidis, V. (2015): "IV Estimation of Panels with Factor Residuals," *Journal of Econometrics*, Vol. 185, pp. 526-541, Elsevier North Holland.

[71] Barro, R. and Sala-i-Martin, X. (1992): “Convergence,” *Journal of Political Economy*, Vol. 100, Issue 2, pp. 223-251, University of Chicago.

[72] Sarafidis, V. and Robertson, D. (2009): "On the Impact of Cross-sectional Dependence in Short Dynamic Panel Estimation, *Econometrics Journal,* Vol. 12 Issue 1, pp 149-161, Elsevier North Holland.

[73] Sarafidis, V. and Wansbeek, T. (2010): “Cross-Sectional Dependence in Panel Data Analysis,” *Econometric Reviews,* Vol. 31 Issue 5, pp. 483-531, Taylor and Francis.

[74] Sarafidis, V., Yamagata, T., and Robertson, D. (2009): “A Test of Cross-Section Dependence for a Linear Dynamic Panel Model with Regressors,” *Journal of Econometrics*, Vol. 148, pp. 149-161, Elsevier North Holland.

[75] Schwarz, G. (1978): "Estimating the dimension of a model," *Annals of Statistics*, Vol. 6 No. 2: pp. 461–464, Institute of Mathematical Statistics.

[76] Song, M., (2013): "Asymptotic Theory for Dynamic Heterogeneous Panels with Cross-Sectional Dependence and its Applications," *Working Paper*, Columbia University.

[77] Tanaka, S. and Kurozumi, E. (2012): "Investigating Finite Sample Properties of Estimators for Approximate Factor Models when *N* is Small," *Economics Letters*, Vol. 116, Issue 3, pp. 465-468, Elsevier North Holland.

[78] Urbain, J. and Westerlund, J. (2013): "On the Estimation and Inference in Factor-Augmented Panel Regressions with Correlated Loadings," *Economics Letters*, Vol. 119, Issue 3, pp 247–250, Elsevier North Holland.

[79] Westerlund, J. and Urbain, J. (2015): “Cross-Sectional Averages Versus Principal Components,” *Journal of Econometrics*, Vol. 185, pp. 372-377, Elsevier North Holland.

Chapter 2

Fixed N, Large T Panel Data Estimation with Cross-sectional Dependence

In this Chapter we study large T and sequential large T, N inconsistencies of parameter estimates of several popular panel data estimators as a result of factor error structures in the data generating process. We show that common estimators in the empirical literature are inconsistent unless very stringent assumptions are placed on the autocovariance of the factors and the covariance of the factor loadings. We also present a consistent estimator based on quasi-differencing that can be used to estimate a dynamic panel data model with a single factor in the error.

2.1 Introduction

Improved regulation of statistical bureaus around the world has aided the collection of macroeconomic time series since the Second World War. As a result, macroeconomic models are increasingly estimated using panel data in an effort to exploit efficiency advantages offered by large datasets. Important examples are the studies of economic growth in Barro and Sala-i-Martin (1993) using a sample of 48 contiguous U.S. states and Islam (1995) using the Summers-Heston dataset. More recently, Beetsma and Guiliadori (2011), Bénétrix and Lane (2013) undertake panel data estimation to study fiscal multipliers and the stabilizing effects of fiscal policy in the European Monetary Union. Similarly, Ilzetski et al (2013) study fiscal multipliers in sets of countries with relatively similar underlying fundamentals. In these latter studies, the number of countries N, is fixed whereas the number of observations per country T, is large relative to the number of countries. However, the estimators employed in these studies are designed for data with large N and fixed T asymptotics, which raises concerns about the consistency of these estimators in these circumstances.

At the same time, phenomena such as global business cycles or preferences that vary over time are often argued to be present in macroeconomic panels. When these phenomena are unobservable, they can lead to cross-sectional dependence in the error and this is often modelled as a factor structure: in addition to the usual equation-specific disturbance, a factor structure links individual cross-sectional equations by an

unobserved process, albeit with different intensities. For example, in the Heston-Summers dataset, a shock to the numeraire currency would affect all country-specific purchasing power parity measures with differing intensity, as a result of proximity to, and trade with, the country holding the numeraire currency (Phillips and Sul, 2003). When both the dependent variable and the regressors of a panel data regression are affected by the factor, special attention is required to avoid inconsistency of the estimated regression coefficients.

There is a long history of studying the consistency of dynamic panel data estimators for models with large N and fixed T, starting with Nerlove (1971): in that paper, simulation evidence is presented on the bias of the coefficient estimates obtained by OLS, GLS and the Within estimator for a dynamic model with correlated unobservable effects in both the time and cross-sectional dimension. Nickell (1981) presents analytical results for the inconsistency of the autoregressive parameter obtained from the Within estimator when N is large and T is small. Nickell's results were extended by Kiviet (1995), who derives the bias of the autoregressive parameter when T is small and obtains a bias-correction formula based on a second-order expansion of the bias of the Within estimator applied to a model with correlated effects.

In recent years, several authors have investigated the inconsistency of a variety of estimators with specific reference to the impact of cross-sectional dependence. For example, Philips and Sul (2007) derive the inconsistency of the pooled OLS estimator for a dynamic panel data model under large N and fixed T asymptotics when cross-sectional dependence is present, both with and without the presence of incidental trends and/or unit roots. These authors model cross-sectional dependence as a factor structure consisting of white noise loadings and factors and they conclude that cross-sectional dependence creates an additional (random) source of bias in the OLS estimator. Similarly, Robertson and Sarafidis (2008) study several popular IV and GMM estimators applied to a dynamic panel data model with individual-specific constants, a single autoregressive factor error component and large N. They show that these estimators are always inconsistent in this specification, regardless of the depth of the lags used as instruments in the moment conditions of the model. Although the estimators are inconsistent, they suggest the bias can be reduced in applications by using cross-sectionally demeaned data for estimation, provided that the mean of the factor loadings is non-zero. Finally, Everaert and de Groote (2016) study the consistency properties of the Pooled Common Correlated Effects (CCE) estimator of Pesaran (2006). The CCE estimator was originally designed for static panel data models with cross-sectional dependence and these authors study the consistency when the CCE estimator is applied instead to a dynamic panel data model with a single autoregressive factor in addition to individual-specific constants. Under similar assumptions as Robertson and Sarafidis (2008), they derive an expression for the inconsistency of the CCE estimator when T is fixed and show that this inconsistency only vanishes when, in addition to N, also T passes to infinity.

Although cross-sectional dependence in the large N, fixed T (dynamic) panel data model has attracted considerable attention, the same cannot be said for the fixed N, large T model. Motivated by the recent surge in macroeconomic studies employing panel data, this paper analyses the consistency of several commonly employed estimators in the large T environment. The estimators we consider are the pooled ordinary least squares estimator (OLS), the OLS estimator with Time Effects (TE) and Pesaran's CCE estimator. For these estimators, we derive first-order asymptotic expansions of the inconsistency in both a static and a dynamic

model as $T \to \infty$ and $T, N \to \infty$ sequentially. In each case, we find that bias stems from the variance of the factor, the variance of the independent variable and the autocovariance of the factor. Furthermore, analogously to the Within estimator in the large N panel data model, the bias is always proportional to the variance (covariance) of the factor loadings if the model is dynamic (static). Based on these conclusions, we find that these estimators are consistent only possible under very strong assumptions on the underlying processes which are unobserved. Moreover, the direction of the bias cannot be predicted without further knowledge of the underlying processes. These problems are serious and imply that if data is suspected to be cross-sectionally dependent, estimation techniques other than the ones studied here have to be considered. Several consistent estimators have been proposed for the factor error model with either large N or large N, T asymptotics such as Bai and Ng (2003), Bai (2009) and Ahn, Lee and Schmidt (2014), but not for the large T case with correlated cross-sectional effects. We therefore provide a simple consistent GMM estimator for the small N, large T framework based on quasi-differencing the data.

The paper is organised as follows: In Section 2.2, we present a simple model and a set of basic assumptions comparable to those of Philips and Sul (2007), Robertson and Sarafidis (2009) and Everaert and de Groote (2016). In Section 2.3, we derive asymptotic bias expressions for dynamic and static versions of the model described in Section 2.2. In Section 2.4, we present a $\sqrt{T}$-consistent estimator that can be interpreted as a restricted CCE estimator. Section 2.5 presents simulation evidence in support of the bias equations and the consistency of the estimator of Section 2.4. Section 2.6 concludes. All proofs are in Appendix 2.1.

2.2 Basic Model and Assumptions

We consider the following dynamic panel data model with one additional covariate and one unobservable interactive fixed effect:

$$\begin{aligned} y_{i,t} &= \rho y_{i,t-1} + \beta x_{i,t} + \lambda_i f_t + \varepsilon_{i,t}, \\ x_{i,t} &= \gamma_i f_t + v_{i,t}, \end{aligned} \tag{2.1}$$

where $i = 1, \ldots, N$ and $t = 1, \ldots T$. We assume that N is small and T is large, so that we work in the large T asymptotic framework and initial conditions are negligible. The $\varepsilon_{i,t}$ and $v_{i,t}$ are errors and the ρ, and β are slope coefficients. The model contains one unobservable factor f_t which we have purposefully left unspecified. We view the factor structure as a generalization of the Fixed Effects model and we therefore do not include an individual-specific constant in (2.1): such an effect would constitute a second factor with zero temporal variation and this would unnecessarily divert attention from the problem at hand. The factor enters into the i-th time series through the loading parameter λ_i directly and indirectly through γ_i and we study the case of a single interactive fixed effect to avoid unnecessary complication of the derivations. To focus on the impact of the interactive fixed effect we will set either $\rho = 0$ or $\beta = 0$ so that the model reduces to a static or a dynamic panel data model with large T and fixed N.

We make the following assumptions on process (2.1):

ASSUMPTION 1: $\{f_t\}$ *is covariance stationary with* $E(f_t) = 0$, $E\left(f_t^2\right) = \sigma_f^2$ *and* $E\left(f_t f_{|t-s|}\right) = \sigma_{f,s}$ *for all* $s \neq t$, *with* $\sum_{s=1}^{T} \left|\sigma_{f,s}\right| < \infty$ *and* $\sigma_{f,s} > 0$ *for at least one* s.

ASSUMPTION 2: *The loadings are i.i.d. over* i *with* $E(\lambda_i) = \mu_\lambda$, $E(\gamma_i) = \mu_\gamma$ *and finite (co)-variances* $\text{var}(\lambda_i) = \sigma_\lambda^2$, $\text{var}(\gamma_i) = \sigma_\gamma^2$ *and* $\text{cov}(\lambda_i, \gamma_i) = \sigma_{\lambda\gamma}$.

ASSUMPTION 3: *The* $\varepsilon_{i,t}$ *and* $v_{i,t}$ *are* (i) *mean-zero i.i.d. with finite second moments* σ_ε^2, σ_v^2 *and* (ii) *independent of* f_s, λ_j *and* γ_j *for all* i, j, t *and* s.

ASSUMPTION 4: $\{y_{i,t}\}$ *satisfies the stability condition* $|\rho| < 1$.

Assumptions 1-4 are designed for macroeconomic applications where strict stationarity of the composite process $y_{i,t}$ is deemed to be too strong. Assumption 1 imposes only mean-zero covariance stationarity with absolutely summable autocovariances on the factor process. The requirement that at least one autocovariance is non-zero avoids the trivial case where the OLS estimator is consistent with large T when $\beta = 0$.[1] Absolute summability of the autocovariance process allows for convenient simplification of functions involving the dynamic properties of the factor that aid the exposition. Similarly, the mean-zero assumption is for convenience and weakening it does not substantially alter the results in what follows. The functional form of the factor is further left unrestricted and can for example be moving average or autoregressive. Assumption 2 states that the loadings are independently and identically distributed and that the loadings of the $x_{i,t}$ and $y_{i,t}$ are finitely correlated. In analogy with the requirement that at least one autocovariance of the factor is zero, the latter condition is required to avoid the case where OLS is consistent if $\rho = 0$ under large N, T asymptotics. Furthermore, Assumption 3 requires that the $\varepsilon_{i,t}$ and $v_{i,t}$ are mutually independent i.i.d. with finite second moments across all $i = 1, \ldots, N$ and the $\varepsilon_{i,t}$ and $v_{i,t}$ are independent of the factor and the individual-specific loadings. These rather strong assumptions are imposed to avoid distracting from the problem at hand, which is the impact of the factor on panel data estimation with large T. We could remove the i.i.d. and zero-mean assumptions at the cost of higher moment conditions and more complicated derivations in what follows. Moreover, since $y_{i,t}$ is mean-zero by Assumptions 1-3, the Within Group estimator of $[\beta, \rho]'$ is equivalent to the OLS estimator to first order. For that reason, the Within Group estimator is not studied further in this paper. Finally, Assumption 4, is used to avoid unit-root asymptotics. Under the requirement that $y_{i,t}$ satisfies the stability condition, with covariance stationary f_t and i.i.d. errors over t, this further implies that $y_{i,t}$ is also a covariance stationary process.

[1] But not efficient of course, one should use the GLS estimator.

2.3 Inconsistency of OLS, Time Effects and Common Correlations Estimators

It will be convenient to stack model (2.1) over the time dimension and write:

$$\begin{aligned} y_i &= \rho y_{i,-1} + \beta x_i + \lambda_i f + \varepsilon_i, \\ x_i &= \gamma_i f + v_i, \end{aligned}$$

where suppression of the time index now indicates T-vectors. We will analyse the inconsistency of three estimators applied to the slope parameters of model (2.1) with either $\rho = 0$ or $\beta = 0$ as $T \to \infty$ with N fixed and as first T and then $N \to \infty$, denoted as "$T, N \to_{\text{seq}} \infty$". Let $w_i = x_i \vee y_{i,-1}$ depending on whether we analyse a static or dynamic model; let $\alpha = \beta \vee \rho$ depending on setting either $\rho = 0$ or $\beta = 0$ in model (2.1) and let $\hat{\alpha}_{(\cdot)}$ be the corresponding estimator. We will examine following estimators:

$$\begin{aligned} \hat{\alpha}_{OLS} &= \frac{\sum_{i=1}^N w_i' y_i}{\sum_{i=1}^N w_i' w_i}, \\ \hat{\alpha}_{TE} &= \frac{\sum_{i=1}^N \tilde{w}_i' \tilde{y}_i}{\sum_{i=1}^N \tilde{w}_i' \tilde{w}_i}, \\ \hat{\alpha}_{CCE} &= \frac{\sum_{i=1}^N w_i' \overline{M} y_i}{\sum_{i=1}^N w_i' \overline{M} w_i}, \end{aligned}$$

$\hat{\alpha}_{OLS}$ is the pooled OLS estimator, $\hat{\alpha}_{TE}$ is the cross-sectionally demeaned, i.e., Time Effects (TE) estimator with $\tilde{a}_i = a_i - \overline{a}$, and $\overline{a} = N^{-1} \sum_{i=1}^N a_i$ is the T-vector of cross-sectional averages of the matrix of time series a. Finally, $\hat{\alpha}_{CCE}$ is the pooled CCE estimator of Pesaran 2006 with:

$$\overline{M} = I_T - \overline{Z} \left(\overline{Z}' \overline{Z} \right)^{-1} \overline{Z}',$$

where $\overline{Z} = [\overline{w}, \overline{y}]$.

Both the OLS and TE are widely used in empirical macro studies[2] and as a result, obtaining expressions of the inconsistency of these estimator in the presence of cross-sectional correlation is useful even in a simple model such as (2.1). On the other hand, the CCE estimator of Pesaran was designed for static large N, T data, but has been shown to be consistent for dynamic panel data models as long as N, $T \to \infty$ sequentially by Everaert and de Groote (2016). The CCE estimator exploits the observation that cross-sectional averages proxy for the factors and can thus be used to project the factors from the estimator of α as long as the number of factors is smaller than one plus the number of regressors. For example, under Assumptions 1-3

[2] Some recent examples include Bénétrix and Lane (2013) and Beetsma and Guiliadori (2010).

when $\rho = 0$

$$\begin{aligned}\overline{Z} &= [\overline{x},\, \overline{y}] \\ &= f\left[\overline{\gamma},\, \beta\overline{\gamma}+\overline{\lambda}\right] + O_{\mathrm{p}}\left(N^{-1}\right) \\ &:= f\overline{c} + O_p\left(N^{-1}\right).\end{aligned}$$

We thus know *a priori* that the CCE estimator will be inconsistent for fixed N because of the $O_p\left(N^{-1}\right)$ term which implies that the cross-sectional averages cannot proxy for the factor perfectly. Note however that $\text{rank}\left(\overline{c}\right) = 1$, so that the rank condition of Pesaran (2006) is satisfied. Finally, we have excluded the Within Group estimator because a bias expression for that estimator would be equal to the bias of the OLS estimator as there are no individual constants included in model (2.1) and $E\left(f_t\right) = 0$. We are now in a position to analyse the inconsistency of all three estimators with first $\rho = 0$ and then $\beta = 0$ in (2.1) by means of first-order asymptotic expansions.

Inconsistencies of $\hat{\beta}$ when Model (2.1) is Static

Inconsistencies arising from cross-sectional dependence have serious consequences for applied work: since interactive fixed effects are unobservable, a practitioner may fail to detect the effects and adequately correct for them. As a result, it is not possible to conduct valid inference on the estimated model in the presence of cross-sectional dependence. To understand the nature of such asymptotic biases in applied work, and to understand exactly what assumptions may be required on the nature of cross-sectional dependence, we now present expressions for the inconsistencies as propositions when $\rho = 0$ in model (2.1):

PROPOSITION 2.1. Inconsistency of OLS in Static Model: *Under Assumptions 1-3, the inconsistency of the OLS estimator with* $\rho = 0$ *is:*

$$\operatorname*{plim}_{T\to\infty}\left(\hat{\beta}_{OLS} - \beta\right) = \frac{\sigma_f^2 N^{-1}\sum_{i=1}^{N}\lambda_i\gamma_i}{\sigma_f^2 N^{-1}\sum_{i=1}^{N}\gamma_i^2 + \sigma_v^2}$$

and the sequential limit is:

$$\operatorname*{plim}_{T,N\underset{\text{seq}}{\to}\infty}\left(\hat{\beta}_{OLS} - \beta\right) = \frac{\sigma_f^2\left(\sigma_{\lambda\gamma} + \mu_\lambda\mu_\gamma\right)}{\sigma_f^2\left(\sigma_\gamma^2 + \mu_\gamma^2\right) + \sigma_v^2}.$$

The OLS estimator is inconsistent when $\rho = 0$ under cross-sectional dependence as long as (i) the dependent and explanatory variables are correlated with the interactive fixed effect and (ii) the limits of $N^{-1}\sum_{i=1}^{N}\lambda_i\gamma_i$, $N^{-1}\sum_{i=1}^{N}\lambda_i$ and $N^{-1}\sum_{i=1}^{N}\gamma_i$ are non-zero. The first source is the usual omitted variable bias: the economet-

rician cannot observe the properties of the interactive fixed effect and therefore does not correct for them appropriately. The second source of the inconsistency follows from the fixed effect portion of the interactive fixed effects: even if the factor has zero variance but instead consists of a vector of ones, the numerator consists of the sum of the product of the i factor loadings of the dependent and independent variables. When N is small and fixed, even if this covariance is indeed zero in addition to the means, any realization of the quantity $N^{-1}\sum_{i=1}^{N}\lambda_i\gamma_i$ is equal to zero with probability zero for any continuous probability distribution and this subsequently results in inconsistency of the OLS estimator. Only under the assumption of zero correlation between the loadings of the dependent and independent variables and zero means of the loadings would the OLS estimator be consistent when N is large in addition to T. Do note however that the magnitude of the bias is negatively related to the variance of $v_{i,t}$: the larger σ_v^2 is relative to σ_f^2, the smaller the bias will be.

Similar conclusions follow for the TE estimator:

PROPOSITION 2.2. **Inconsistency of TE in Static Model:** *Under Assumptions 1-3, the inconsistency of the TE estimator with* $\rho = 0$ *is:*

$$\operatorname*{plim}_{T\to\infty}\left(\hat{\beta}_{TE}-\beta\right) = \frac{\sigma_f^2 N^{-1}\sum_{i=1}^{N}\left(\lambda_i-\bar{\lambda}\right)(\gamma_i-\bar{\gamma})}{\sigma_f^2 N^{-1}\sum_{i=1}^{N}(\gamma_i-\bar{\gamma})^2+(1-N^{-1})\sigma_v^2},$$

and the sequential limit is:

$$\operatorname*{plim}_{\substack{T,N\to\infty\\ \text{seq}}}\left(\hat{\beta}_{TE}-\beta\right) = \frac{\sigma_f^2\sigma_{\lambda\gamma}}{\sigma_f^2\sigma_\gamma^2+\sigma_v^2}.$$

The inconsistency of the TE estimator with $\rho = 0$ is very similar to the inconsistency of the OLS estimator and equal whenever μ_λ and μ_γ are identically equal to zero in the sequential large T, N-case. Of course, if either the λ_i or γ_i or both are constant over i, then the TE estimator is consistent.[3] Moreover, the direction of the inconsistency depends crucially on the covariance between the loadings λ_i and γ_i, and will be non-zero for fixed N due to $N^{-1}\sum_{i=1}^{N}\left(\lambda_i-\bar{\lambda}\right)(\gamma_i-\bar{\gamma})$ in the numerator. With $T, N \underset{\text{seq}}{\to} \infty$ however, we can give the following characterisation of the sign of the bias of the TE estimator with $\rho = 0$:

$$-\frac{\sigma_f^2\left(\sigma_\gamma^2\right)^2}{\sigma_f^2\sigma_\gamma^2+\sigma_v^2} \leq \operatorname*{plim}_{\substack{T,N\to\infty\\ \text{seq}}}\left(\hat{\beta}_{TE}-\beta\right) \leq \frac{\sigma_f^2\left(\sigma_\gamma^2\right)^2}{\sigma_f^2\sigma_\gamma^2+\sigma_v^2}. \tag{2.2}$$

Therefore, as either σ_f^2 or σ_γ^2 increases without bound, the inconsistency will be proportional to $\left(-\sigma_\gamma^2,\ \sigma_\gamma^2\right)$ and the empirical researcher has no hope of determining the sign of the inconsistency. Clearly, it can be exactly zero if the covariance of the loadings is zero, but such an assumption is likely too strong in many

[3] This is the situation that the TE estimator is designed for.

applications.

It is also interesting to contrast the conclusions of Propositions 3.1.1 and 3.1.2 with the finding of Urbain and Westerlund (2013) for the CCE estimator: in that paper it is shown that, unless $\sigma_{\lambda\gamma} = 0$, the CCE estimator of β is inconsistent under large N, T asymptotics whenever the rank condition on $\bar{c}$ is violated. Under that assumption however, the OLS and TE estimators are consistent and do not require a rank condition. For example in a model with R factors, let F be the $T \times R$ matrix containing all factors at each $t = 1, \ldots, T$; Let $\Lambda_i = [\lambda_{1,i}, \ldots, \lambda_{R,i}]'$ and $\Lambda = [\Lambda_1, \ldots, \Lambda_N]$ collect R mean-zero loadings for N individuals and similarly define the loadings of the independent variable as Γ_i and Γ. Then the numerator of the inconsistency of the TE and OLS estimators reads:

$$(NT)^{-1} \text{trace}\left(\Gamma' F' F \Lambda\right) \leq N^{-1} \text{trace}\left(\Lambda\Gamma'\right) \times T^{-1} \text{trace}\left(F'F\right).$$

By assumption, the first term on the right vanishes and the OLS and TE estimators are consistent for $R > 1$ without requiring a comparable rank condition. This argument raises questions on the practical usefulness of the CCE estimator in the static model: R is generally unknown and a sufficiently large number of regressors is required to span the space of the factors. On the other hand, for unknown R if one assumes that the loadings are uncorrelated, we might as well use OLS, TE or, indeed, GLS estimators, which are consistent under the same assumptions.

We finish the section with a proposition on the inconsistency of the CCE estimator for the static model:

PROPOSITION 2.3. Inconsistency of CCE in Static Model: *Under Assumptions 1-3, the inconsistency of the CCE estimator with* $\rho = 0$ *is:*

$$\operatorname*{plim}_{T\to\infty}\left(\hat{\beta}_{CCE} - \beta\right) = \frac{\left(N^{-1}\sigma_\varepsilon^2\right)\sigma_f^2 N^{-1}\sum_{i=1}^{N}\left(\lambda_i - \bar{\lambda}\right)\left(\gamma_i - \bar{\gamma}\right)}{\sigma_f^2\left(\sigma_v^2\bar{\lambda}^2 + \sigma_\varepsilon^2\bar{\gamma}^2\right) + N^{-1}\left(\sigma_\varepsilon^2\sigma_v^2 - \sigma_f^2\sigma_v^2\bar{\lambda}^2 + \sigma_\varepsilon^2\sigma_f^2\left[N^{-1}\sum_{i=1}^{N}\gamma_i^2 - 2\bar{\gamma}^2\right]\right) - N^{-2}\sigma_\varepsilon^2\sigma_v^2},$$

and $\operatorname*{plim}_{\substack{T,N\to\infty \\ \text{seq}}}\left(\hat{\beta}_{CCE} - \beta\right) = 0.$

When N is fixed, the inconsistency of the CCE estimator is again directly proportional to the covariance of the loadings λ_i and γ_i. Similarly to the TE estimator, if the loadings are constant for all i on either the dependent or the independent variables or both, the CCE estimator is consistent with fixed N. However, irrespective of these quantities, when $N \to \infty$ too, the estimator is also consistent due to the term $\left(\sigma_\varepsilon^2/N\right)$ in the numerator. Consequently, for moderate N, we expect the estimator to perform better than either the OLS or the TE estimators under our assumptions. Finally, as with the OLS and TE, the denominator is always positive and the sign of the bias is determined by the covariance of the factor loadings. As a result,

it is impossible in applied work to characterize the direction of the inconsistency without knowledge of this quantity.

Inconsistencies of $\hat{\rho}$ when Model (2.1) is Dynamic

Having seen that when $\rho = 0$, inconsistency of the OLS, TE and CCE estimators depends crucially on the (assumptions placed on the) covariance of the factor loadings of $x_{i,t}$ and $y_{i,t}$, we now analyse the impact of the interactive fixed effect on the dynamic model with $\beta = 0$. To simplify notation, let us first introduce the following quantity:

$$\pi = \operatorname*{plim}_{T\to\infty} \sum_{s=1}^{T} \rho^s \sigma_{f,s}$$

where we note that under Assumption 1 and 4, $\pi < \infty$ by the Cauchy-Schwarz inequality.

We can now analyse the inconsistency of the OLS estimator:

PROPOSITION 2.4. Inconsistency of OLS in Dynamic Model: *Under Assumptions 1-4, the inconsistency of the OLS estimator with* $\beta = 0$ *is:*

$$\operatorname*{plim}_{T\to\infty} (\hat{\rho}_{OLS} - \rho) = \frac{1-\rho^2}{\rho} \times \frac{\pi N^{-1} \sum_{i=1}^{N} \lambda_i^2}{\sigma_\varepsilon^2 + \left(\sigma_f^2 + 2\pi\right) N^{-1} \sum_{i=1}^{N} \lambda_i^2},$$

and the sequential limit is:

$$\operatorname*{plim}_{T,N\underset{\text{seq}}{\to}\infty} (\hat{\rho}_{OLS} - \rho) = \frac{1-\rho^2}{\rho} \times \frac{\pi \left(\sigma_\lambda^2 + \mu_\lambda^2\right)}{\sigma_\varepsilon^2 + \left(\sigma_f^2 + 2\pi\right) \left(\sigma_\lambda^2 + \mu_\lambda^2\right)}.$$

The inconsistency of the OLS estimator in the dynamic version of model (2.1) depends on (i) the autocovariance of the factor and (ii) the sum of squares of the factor loadings. That is, the OLS estimator is consistent if the factor has zero autocovariance, although not necessarily efficient. However, it does seem rather restrictive to assume that only the interactive fixed effect is static when the model itself is dynamic. Moreover, it is not possible to determine the direction of the inconsistency in the dynamic model as in both the numerator and the denominator, the interplay of ρ and the autocovariance makes their respective signs ambiguous.

As with the static model, the TE estimator is subject to a comparable inconsistency:

PROPOSITION 2.5. Inconsistency of TE in Dynamic Model: *Under Assumptions 1-4, the inconsistency of the TE estimator with* $\beta = 0$ *is:*

$$\underset{T\to\infty}{\text{plim}}\left(\hat{\rho}_{TE}-\rho\right) \quad = \frac{1-\rho^2}{\rho}\times \quad \frac{\pi N^{-1}\sum_{i=1}^{N}\left(\lambda_i-\bar{\lambda}\right)^2}{\left(1-N^{-1}\right)\sigma_\varepsilon^2+\left(\sigma_f^2+2\pi\right)N^{-1}\sum_{i=1}^{N}\left(\lambda_i-\bar{\lambda}\right)^2},$$

and the sequential limit is:

$$\underset{T,N\underset{\text{seq}}{\to}\infty}{\text{plim}}\left(\hat{\rho}_{TE}-\rho\right) \quad = \frac{1-\rho^2}{\rho}\times \quad \frac{\pi\sigma_\lambda^2}{\sigma_\varepsilon^2+\left(\sigma_f^2+2\pi\right)\sigma_\lambda^2}.$$

Compared to the inconsistency of OLS, the difference is that if the factor loadings are constant, the inconsistency vanishes with N. As mentioned before, this is the scenario for which the TE estimator is designed, but notice how the analogy with the Within Group estimator in a small T dynamic panel data model does not hold: where the latter would be inconsistent under small T, large N when equation-specific constants are included, the TE estimator does not suffer from the analogous problem. Indeed, the TE estimator is consistent if $\lambda_i=\bar{\lambda}$ for all i even when N is small.

Finally, we present the inconsistency of the CCE estimator in the dynamic panel data model:

PROPOSITION 2.6. Inconsistency of CCE in Dynamic Model: *Under Assumptions 1-4, the inconsistency of the CCE estimator with* $\beta=0$ *is:*

$$\underset{T\to\infty}{\text{plim}}\left(\hat{\rho}_{CCE}-\rho\right)= \frac{N^{-1}\rho\left(1-\rho^2\right)\sigma_\varepsilon^2\pi\sum_{i=1}^{N}\left(\lambda_i-\bar{\lambda}\right)^2}{\kappa_1\left[\left(N-1\right)\bar{\lambda}^2+\sum_{i=1}^{N}\left(\lambda_i-\bar{\lambda}\right)^2\right]+\rho^2\left(N^{-1}\sigma_\varepsilon^2\right)\left[\kappa_2 N+\left(\kappa_3+N^{-1}\sigma_\varepsilon^2\right)\left(N-1\right)\right]},$$

where:

$$\begin{aligned}
\kappa_1 &= \left[\rho\sigma_f^2+\left(\rho-1\right)\pi\right]\times\left[\rho\sigma_f^2+\left(\rho+1\right)\pi\right]\bar{\lambda}^2,\\
\kappa_2 &= \left(\sigma_f^2+2\pi\right)N^{-1}\sum_{i=1}^{N}\left(\lambda_i-\bar{\lambda}\right)^2,\\
\kappa_3 &= 2\left(\sigma_f^2+\pi\right)\bar{\lambda}^2
\end{aligned}$$

and $\underset{T,N\underset{\text{seq}}{\to}\infty}{\text{plim}}\left(\hat{\rho}_{CCE}-\rho\right)=0.$

The CCE estimator is inconsistent when N is fixed as $T \to \infty$ but consistent when $T, N \to \infty$ sequentially. This latter property is, similar to the static case, because the numerator is of lower order than the denominator: the expression that has κ_1 as leading term is $O_p(N)$, whilst those with κ_2 and κ_3 as leading terms are $O_{\mathrm{p}}(1)$. For this reason, we expect the CCE estimator to perform better in samples with moderately large N than the OLS and TE estimators. Furthermore, note that the CCE estimator inherits the property of the TE estimator that if the $\lambda_i = \overline{\lambda}$ for all i, the estimator is consistent. As before, the sign of the bias depends directly on quantities that may be positive or negative and therefore we will not try to sign the bias explicitly.

In summary, when a model exhibits cross-sectional dependence of the sort induced by interactive fixed effects, all estimators we have studied are inconsistent as $T \to \infty$, both in the static and dynamic case. The expressions for the inconsistency depend crucially on the properties of the variances and covariances of the factor loadings and disappears only when in addition to T, $N \to \infty$ in the case of the CCE estimator. However, N is typically small and the number of variables usually limited in macroeconomic applications and as a result, it is unlikely that the interactive fixed effect is removed by the CCE. It thus seems difficult to justify inference based on the estimators examined above, unless a set of very restrictive assumptions are imposed on the correlation structure of the factors and the loadings. Moreover, the requirement of uncorrelated factor loadings not only allows consistency of the CCE estimator, but also of the TE estimator in the static case when $T, N \underset{\text{seq}}{\to} \infty$.

2.4 $\sqrt{T}$-Consistent Estimation of Dynamic Panel Data Models with an Interactive Fixed Effect

As we have seen in the last section, the CCE estimator is inconsistent when N is fixed and $T \to \infty$ because the averages cannot proxy for the factor exactly. Instead of using the sample averages to proxy for the space of the factor however, we can also use them to exactly remove the factor parametrically from model (2.1):

$$\begin{aligned} y_{i,t} - \lambda_i^{\star}\overline{y}_t =& \rho\left(y_{i,t-1} - \lambda_i^{\star}\overline{y}_{t-1}\right) + \beta\left(x_{i,t} - \lambda_i^{\star}\overline{x}_t\right) + u_{i,t}, \\ u_{i,t} :=& \varepsilon_{i,t} - \lambda_i^{\star}\overline{\varepsilon}_t,\ i = 1,\ldots,N, t = 1,\ldots,T, \end{aligned} \tag{2.3}$$

and $\lambda_i^{\star} := \lambda_i/\overline{\lambda}$. The transformed model (2.3) exploits the fact that $\lambda_i f_t - \lambda_i^{\star} f_t = 0$ although we now have to estimate the factor loading $\lambda_i^{\star}$ to remove the factor from $u_{i,t}$. This quasi-difference transformation can be achieved for every cross-sectional unit in (2.1) and we thus have to estimate N loadings in addition to β and ρ. However, since the f_t are unobservable, we cannot separately identify the $\lambda_i^{\star}$ from the f_t without a normalization and we set $\overline{\lambda} = 1$ in equation (2.3) to achieve identification. Moreover, we drop the N-th equation to avoid a singular estimation problem and we thus estimate the parameters using only the first $N-1$ cross-sectional units.

We will estimate the parameters of (2.3) using the following non-linear moment conditions:

$$E\left(z_{i,t}u_{i,t}\right)=0_S, \tag{2.4}$$

where $z_{i,t}$ is an $S\times 1$ vector of instruments for all $i=1,\ldots,N-1$ errors of (2.3) at each t. With scalar ρ and β, identification requires that $S(N-1)\geq 2+(N-1)$. Valid moment conditions for (2.4) can be generated using $y_{i,t-p}$, $x_{i,t-(p-1)}$, $\bar{y}_{t-p}$ and $\bar{x}_{t-(p-1)}$ for all i and $p\geq 1$ and it should be clear from the transformed model (2.3) that $\bar{y}_t$ is not a valid instrument for (2.4). To conserve on notation, we now write $\alpha=[\beta,\ \rho]'$ and $\Lambda^{\star}=\left[\lambda_1^{\star},\ldots,\lambda_{N-1}^{\star}\right]'$ and define the grand parameter set as $\phi=[\alpha',\ \Lambda^{\star\prime}]'$ with parameter space Φ.

We will estimate the $S(N-1)$ population moments (2.4) jointly using the sample moment functions:

$$T^{-1}\sum_{t=1}^{T}m_t\left(\phi\right)=T^{-1}\sum_{t=1}^{T}\left[\left(z_{1,t}u_{1,t}\right)',\ldots,\ \left(z_{N-1,t}u_{N-1,t}\right)'\right]'$$

and a GMM estimator solves $\hat{\phi}=\underset{\phi\in\Phi}{\operatorname{argmin}}\,Q\left(\phi\right)$, where:

$$Q\left(\phi\right):=T^{-1}\sum_{t=1}^{T}m_t\left(\phi\right)'\left[\sum_{t=1}^{T}m_t\left(\tilde{\phi}\right)m_t\left(\tilde{\phi}\right)'\right]^{-1}\sum_{t=1}^{T}m_t\left(\phi\right)$$

and $\tilde{\phi}$ is an initial estimate of ϕ. We will refer to this estimator as the Quasi-difference CCE (QDCCE) estimator to make explicit the quasi-difference operation underlying the moment conditions.

We require the following additional assumptions for consistency and asymptotic normality of the QD-CCE estimator:

ASSUMPTION 4.A: (i) $\{\varepsilon_{i,t},\,f_t,\,x_{i,t}\}$ *is an ergodic stationary process for each* $i=1,\ldots,N$, (ii) $E\left(z_{i,t}\varepsilon_{j,t}|\,\varepsilon_{i,t-1},\,f_{t-1},\,x_{i,t-1},\,\varepsilon_{i,t-2},\,\ldots\right)=0$ *is an adapted martingale difference sequence* for all $i,\ j=1,\ldots,N$ *and* $t=1,\ldots,T$ *and* (iii) $E\left(\varepsilon_{i,t}|\,f_t,\,x_{i,t-1},\,\varepsilon_{i,t-1},\,f_{t-1},\ldots\right)=0$ for all $i=1,\ldots,N$ *and* $t=1,\ldots,T$.

ASSUMPTION 5: *The parameter space* Φ (i) *is a compact and* (ii) *excludes an open ball* $B\left(\overline{\lambda}=0,\ \eta\right)$ with arbitrarily small radius η.

ASSUMPTION 6: $V:=E\left[m_t\left(\phi\right)m_t'\left(\phi\right)\right]$ *is positive definite and* $G:=E\left(\frac{\partial m_t(\phi)}{\partial\phi'}\right)$ *has full rank.*

Assumption 4.A is made for convenience and is standard in textbook treatments of GMM in time series models such as Newey and McFadden (1994). The same is true for Assumption 6 and the first part of Assumption 5, which is required for uniform convergence of the (derivatives of the) sample moments to the population moments. The second part of Assumption 5 is necessary to quasi-difference model (2.1) using sample averages and is not overly restrictive: unless $\lambda_i=0$ for all $i=1,\ldots,N$, in a small sample, the cross-sectional average of the factor loadings will not equal zero exactly with high probability even if its expectation is zero when generated by some distribution. The exclusion of B thus avoids this situation, whilst

maintaining the compactness requirement of Φ. Note that this assumption is in line with the bias-reduction procedure in GMM as presented in Robertson and Sarafidis (2008).

We can now present consistency and asymptotic normality results for the QDCCE as propositions:

PROPOSITION 2.7. Distributional Result for QDCCE: *Let Assumptions 1-6 hold with 4.A instead of 4 and suppose we estimate model* (2.1) *by QDCCE. Then as* $T \to \infty$*, the QDCCE estimator is consistent and*

$$\sqrt{T}\left(\hat{\phi}-\phi\right) \to_d N\left(\mathbf{0}, \left(G'WG\right)^{-1}\left(G'WVWG\right)\left(G'WG\right)^{-1}\right),$$

where W *is a weight matrix and* $V = E\left[m_t\left(\phi\right)m_t'\left(\phi\right)\right]$.

We will not prove Proposition 4 but note that this follows from Theorem 2.6 of Newey and McFadden (1994): that is, with compactness, ergodic stationarity and finite second moments, there exists a functional with finite expectation that dominates $m_t\left(\phi\right)$. Moreover, in the next chapter we will prove a version of the proposition under more general conditions. Under the amended Assumptions 1-6 however, the conclusions of Proposition 4 follow straightforwardly from the derivations in that chapter.

REMARK 4.1: As is customary with GMM problems, setting $W = V^{-1}$ results in efficient GMM. By standard arguments, efficient QDCCE has variance equal to $\left(G'V^{-1}G\right)^{-1}$, which can be approximated by a two-step procedure: in a first step, W is set to an identity matrix or the covariance matrix of the instruments; in a second step, a weight matrix is constructed based on the first-step QDCCE. Under Assumption 4.A this weight matrix can be consistently estimated by $T^{-1}\sum_{t=1}^{T} m_t\left(\phi\right)m_t'\left(\phi\right)$.

REMARK 4.2: Compared to the conventional CCE estimator, a benefit of QDCCE can be that the number of parameters is increased only by the parameter λ_i. This adds $N-1$ parameters to the total parameter count instead of augmenting the regression with a cross-sectional average for each of the independent variables. In a VAR framework for example, this may help to reduce the curse of dimensionality.

REMARK 4.3: Everaert and de Groote (2016) define a similar estimator which they call "Restricted CCEp". However, because they do not take into account the endogeneity introduced by augmenting the regression with the cross-sectional average in the composite residual u_i, their estimator is inconsistent for small T. Their estimator could in principle be made consistent for small T by introducing an instrumental variables approach. In a recent paper, de Vos and Everaert (2016) instead provide a bias-corrected CCE estimator that is consistent for the dynamic panel data model with a factor error structure under large N and fixed T.

2.5 Monte Carlo Experiments

In this section we examine the finite sample properties of the OLS, TE, CCE and QDCCE estimators in detail. As our derivations suggest, we are particularly interested in the impact of the (autoco-)variance of the factor and the (co-)variance of the loading component of the interactive fixed effect. As we know that all estimators save the QDCCE estimator are inconsistent, the experiments thus indicate on exactly how much bias one should expect when estimating models without taking into account the interactive fixed effects.

We simulate 5000 Monte Carlo replications of model (2.1) above and include a single autoregressive interactive fixed effect, i.e:

$$f_t = \xi f_{t-1} + w_t.$$

The baseline parameter values are:

$$\begin{aligned} \beta = 1, \quad \rho = 0.5, \quad \xi = 0.5, \\ \sigma_\varepsilon^2 = 1, \quad \sigma_v^2 = 1, \quad \sigma_w^2 = 1, \\ \sigma_\lambda^2 = 1, \quad \sigma_\gamma^2 = 1, \quad \sigma_{\lambda\gamma} = 0.5, \\ \mu_\lambda = 1, \quad \mu_\gamma = 1, \end{aligned}$$

and as in the analysis we set either $\rho = 0$ or $\beta = 0$. We set $N = 5$, 10 and $T = 100$, 500 for the dimensions of the panel: from a macroeconomic perspective, we think $N = 5$ would constitute a small panel whilst $N = 10$ would be a relatively large panel. Furthermore, we expect the performance of the CCE estimator to improve over this increase in cross-sectional units. We set $T = 100$ because it is a typical size of macroeconomic time series, constituting some 25 years worth of quarterly data. On the other hand, we expect that with $T = 500$, the size of the time series is sufficient for our bias equations to be relatively accurate, as should be our QDCCE estimator. The experiments entail varying the variance of the factor from 1 to 5 in the static model and autoregressive parameter of the factor ξ to 0.9 in the dynamic model. In addition, we increase the variance of each λ_i from 1 to 5. For the QDCCE estimator we use as instruments the contemporaneous independent variable of all N time series for the static model whilst the first lag of all N time series is used in the dynamic model. We now discuss the simulation results in turn.

Static Model, $\rho = 0$

The coverage of the bias expressions is always close to the true values for the simple OLS and TE estimators regardless of the dimensions of the panels. For the CCE estimator, when $T = 100$, the bias equations underestimate but approach the bias when $T = 500$ and $N = 10$. This implies that higher order effects are present which are overlooked by our first-order expansions. The OLS and TE estimators always overestimate β quite severely, whereas the CCE estimator is somewhat less biased: in the baseline model, the bias is some 25 percent of the true value of unity for the OLS and TE estimators, where the CCE bias is 15 percent. As

expected, as the covariance of the factor loadings increases, the bias worsens to over 40 percent for the OLS and TE estimators and to 20 percent for the CCE estimator regardless of the dimensions of the panel. Similarly, as the variance of each λ_i is increased, and thus the correlation of the loadings reduced, the bias reduces to some 11 percent for the OLS and TE estimators and 4.5 percent for the CCE estimator. These results are in line with the predictions of the asymptotic expansions of the bias for the static model and thus cast doubt on any empirical work where one may expect factors to be present in the data. However, as expected, the QDCCE estimator is essentially unbiased in all specifications of the simulation exercise. In terms of consistency and RMSE, it always outperforms by a wide margin the other estimators, regardless of the model specifications.

Dynamic Model, $\beta = 0$

For the dynamic model, the conclusions drawn regarding the bias equations from the simulation exercise are largely similar to the static model: as T increases, the coverage of the bias formulae is essentially exact for the TE and OLS estimators. The coverage of the CCE estimator also approaches the true value but again there are higher order effects present that only vanish as T becomes large and the bias tends towards the quantity derived in Section 2.3.2.

Interestingly, increasing the variance of the loading and/or the autocorrelation of the interactive fixed effect pushes the estimate of ρ by OLS or TE estimation towards unity, which opens the possibility that standard unit root tests will mistake a model with interactive fixed effects for an integrated process in practice. This possibility is left for future research.

The QDCCE estimator now is not clearly the best estimator in terms of bias and RMSE in all configurations of the model and dimensions of the panel and the CCE estimator performs similarly or even better when $N = 10$. This finding is in line with Everaert and De Groote (2016), who find that the time dimension does not matter for the small sample properties of the CCE estimator.

2.6 Conclusions

Recent research in panel data econometrics has yielded a wealth of information regarding the applicability of panel data estimators under cross-sectional dependence in both the large N and large N, T setups but has so far overlooked the fixed N, large T paradigm. Despite this shortcoming of the literature, several authors have used estimators designed for large N in macroeconomic studies where the opposite asymptotics apply. This paper has investigated the bias of such estimators in detail in simple static and dynamic models of cross-sectional dependence: we derived expressions for the first-order bias of the OLS, TE and CCE estimators under large T asymptotics only and shown not only that bias exists but that it may be quite severe. We have subsequently designed a consistent and asymptotically normal estimator with the flavour of the CCE estimator that uses quasi-differencing to remove the interactive fixed effect.

An extensive Monte Carlo study then verified our theoretical results: for a static model, the bias is

proportional to the variance of the factor and the covariance of the factor loadings of the dependent and independent variables. For a dynamic model by contrast, we find that the bias is proportional to the autocovariance of the factor and the variance of the factor loadings. Furthermore, we have found that the QDCCE estimator typically outperforms all other estimators when T is sufficiently large, but that the CCE estimator of Pesaran (2006) is a clear competitor whenever N is not too small: Indeed, when N is as small as 10, it has reasonable finite sample performance, sometimes even outperforming the $\sqrt{T}$-consistent QDCCE estimator. Of course, in models with both weakly and strictly exogenous regressors, it is to be expected that this conclusion regarding the CCE estimator no longer holds.

In conclusion, if interactive fixed effects are present in macroeconomic data, as they likely are, either through unobserved components such as trends or business cycles or through omitted variable bias, then the use of OLS and TE estimators is very difficult to justify as the biases can be quite large. The same conclusion obtains for the CCE estimator of Pesaran (2006), when N is very small, or if the model has a high parameter count so that degrees of freedom do not permit augmenting the model with the required cross-sectional averages to proxy for the factors. As such, estimators that are designed for fixed N, large T specifically should be used, such as the QDCCE estimator proposed in this paper.

Appendix 2.1 Proofs for Chapter 2

For the proof of the various propositions, we first compile some lemmas of the asymptotic representations of second moments involved in the bias equations. After that, the proofs follow from combination and simplification of these representations.

LEMMA A: *Let Assumptions 1-3 hold, then when* $\rho = 0$*:*

- A1

$$\operatorname*{plim}_{T\to\infty} \frac{1}{T}\sum_{i=1}^{N} x_i' x_i \quad = \quad \sigma_f^2 \sum_{i=1}^{N} \gamma_i^2 + \sigma_v^2.$$

- A2

$$\operatorname*{plim}_{T\to\infty} \frac{1}{T}\sum_{i=1}^{N} x_i' f \lambda_i \quad = \quad \sigma_f^2 \sum_{i=1}^{N} \lambda_i \gamma_i.$$

Proof:

Consider A1: Since $E\left(f_t v_{i,t}\right) = 0$ for all i and t, we have:

$$\begin{aligned}
\operatorname*{plim}_{T\to\infty}\frac{1}{T}\sum_{i=1}^{N}x_i'x_i &= \operatorname*{plim}_{T\to\infty}\frac{1}{T}\sum_{i=1}^{N}(\gamma_i f+v_i)'(\gamma_i f+v_i)\\
&= \sum_{i=1}^{N}\gamma_i^2\operatorname*{plim}_{T\to\infty}\frac{1}{T}f'f+\operatorname*{plim}_{T\to\infty}\frac{1}{T}\sum_{i=1}^{N}v_i'v_i\\
&= \sigma_f^2\sum_{i=1}^{N}\gamma_i^2+\sigma_v^2.
\end{aligned}$$

A2 follows in a similar fashion. □

Proof of Proposition 2.1:

Let:

$$E\left(\hat{\beta}_{OLS}-\beta\right) = \left[\operatorname*{plim}_{T\to\infty}\frac{1}{T}\sum_{i=1}^{N}x_i'x_i\right]^{-1}\left[\operatorname*{plim}_{T\to\infty}\frac{1}{T}\sum_{i=1}^{N}x_i'(x_i\beta+f\lambda_i+\varepsilon_i)\right]$$

The proof of Proposition 2.1 follows immediately from combining A1 and A2 and noting that $\operatorname*{plim}_{T\to\infty}\frac{1}{T}\sum_{i=1}^{N}x_i'\varepsilon_i = 0$ by assumption. Dividing both terms by N and taking limits w.r.t. N then yields the desired result. □

LEMMA B: *Let Assumptions 1-3 hold, then with* $\rho = 0$*:*

- B1

$$\operatorname*{plim}_{T\to\infty}\frac{1}{T}\sum_{i=1}^{N}\tilde{x}_i'\tilde{x}_i = \sigma_f^2\sum_{i=1}^{N}\tilde{\gamma}_i^2+\left(1-\frac{1}{N}\right)\sigma_v^2.$$

- B2

$$\operatorname*{plim}_{T\to\infty}\frac{1}{T}\sum_{i=1}^{N}\tilde{x}_i'f\tilde{\lambda}_i = \sigma_f^2\sum_{i=1}^{N}\tilde{\lambda}_i\tilde{\gamma}_i.$$

Proof:

We again only show B1 and note that B2 follows similarly. Let:

$$\begin{aligned}
\operatorname*{plim}_{T\to\infty}\frac{1}{T}\sum_{i=1}^{N}\tilde{x}_i'\tilde{x}_i &= \operatorname*{plim}_{T\to\infty}\frac{1}{T}\sum_{i=1}^{N}\left(\tilde{\gamma}_i f+\tilde{v}_i\right)'\left(\tilde{\gamma}_i f+\tilde{v}_i\right)\\
&= \sum_{i=1}^{N}\tilde{\gamma}_i^2\operatorname*{plim}_{T\to\infty}\frac{1}{T}f'f+\operatorname*{plim}_{T\to\infty}\frac{1}{T}\sum_{i=1}^{N}\tilde{v}_i'\tilde{v}_i\\
&= \sigma_f^2\sum_{i=1}^{N}\tilde{\gamma}_i^2+\operatorname*{plim}_{T\to\infty}\frac{1}{T}\sum_{i=1}^{N}\left(v_i-\overline{v}\right)'\left(v_i-\overline{v}\right)\\
&= \sigma_f^2\sum_{i=1}^{N}\tilde{\gamma}_i^2+\left(1-\frac{1}{N}\right)\sigma_v^2,
\end{aligned}$$

As $E\left(v_i\right)=0$, $E\left(v_i^2\right)=\sigma_v^2$ and $E\left(v_iv_j\right)=0$ when $T\to\infty$, since:

$$\begin{aligned}
\operatorname*{plim}_{T\to\infty}\frac{1}{T}\sum_{i=1}^{N}\left(v_i-\overline{v}\right)'\left(v_i-\overline{v}\right) &= \operatorname*{plim}_{T\to\infty}\frac{1}{T}\sum_{i=1}^{N}v_i'v_i+\operatorname*{plim}_{T\to\infty}\frac{1}{T}\overline{v}'\overline{v}\\
&\quad -2\operatorname*{plim}_{T\to\infty}\frac{1}{T}\sum_{i=1}^{N}\left(v_i'\frac{1}{N}\sum_{i=1}^{N}v_i\right)\\
&= \operatorname*{plim}_{T\to\infty}\frac{1}{T}\sum_{i=1}^{N}v_i'v_i+\operatorname*{plim}_{T\to\infty}\frac{1}{T}\frac{1}{N}\sum_{i=1}^{N}v_i'\frac{1}{N}\sum_{i=1}^{N}v_i\\
&\quad -2\frac{1}{N}\operatorname*{plim}_{T\to\infty}\frac{1}{T}\sum_{i=1}^{N}v_i'v_i\\
&= \left(1-\frac{1}{N}\right)\sigma_v^2.
\end{aligned}$$

□

Proof of Proposition 2.2:

Similarly to the proof of Proposition 2.1, combination of B1 and B2 yields the result immediately after noticing that $\operatorname*{plim}_{T\to\infty}\frac{1}{T}\sum_{i=1}^{N}\tilde{x}_i'\tilde{\varepsilon}_i=0$. The sequential result follows by dividing both numerator and denominator by N and noting that $\operatorname*{plim}_{N\to\infty}\frac{1}{N}\sum_{i=1}^{N}\tilde{\lambda}_i\tilde{\gamma}_i=\sigma_{\lambda\gamma}$.□

Proof of Equation (2.2):

The probability limit of the bias of the TE estimator is:

$$\operatorname*{plim}_{T,N\to\infty}\left(\hat{\beta}_{TE}-\beta\right) = \frac{\sigma_f^2\sigma_{\lambda\gamma}}{\sigma_f^2\sigma_\gamma^2+\sigma_v^2}.$$

Replacing $\sigma_{\lambda\gamma}$ with the definition of the associated correlation and remembering that the correlation coefficient is bound in $(-1,\ 1)$, i.e. when $\lambda_i = -\gamma_i$ and $\lambda_i = \gamma_i$ for all $i = 1, \ldots,\ N$, we have:

$$\begin{array}{rcl} \dfrac{\sigma_f^2\sigma_{\lambda\gamma}}{\sigma_f^2\sigma_\gamma^2+\sigma_v^2}, & = & \dfrac{\sigma_f^2 \mathrm{corr}\,(\lambda_i,\gamma_i)\,\sigma_\lambda^2\sigma_\gamma^2}{\sigma_f^2\sigma_\gamma^2+\sigma_v^2}, \\ -\dfrac{\sigma_f^2\left(\sigma_\gamma^2\right)^2}{\sigma_f^2\sigma_\gamma^2+\sigma_v^2} & \wedge & \dfrac{\sigma_f^2\left(\sigma_\gamma^2\right)^2}{\sigma_f^2\sigma_\gamma^2+\sigma_v^2}. \square \end{array}$$

LEMMA C: *Let Assumptions 1-3 hold, then:*

- C1

$$\begin{array}{rcl} \underset{T\to\infty}{\mathrm{plim}}\dfrac{1}{T}x_i'x_i & = & \sigma_f^2\gamma_i^2+\sigma_v^2, \\ \underset{T\to\infty}{\mathrm{plim}}\dfrac{1}{T}x_i'\bar{x} & = & \sigma_f^2\gamma_i\bar{\gamma}+\dfrac{1}{N}\sigma_v^2, \\ \underset{T\to\infty}{\mathrm{plim}}\dfrac{1}{T}\bar{x}'\bar{x} & = & \sigma_f^2\bar{\gamma}^2+\dfrac{1}{N}\sigma_v^2. \end{array}$$

- C2

$$\underset{T\to\infty}{\mathrm{plim}}\frac{1}{T}\bar{y}'\bar{y} = \sigma_f^2\left(\beta\bar{\gamma}+\bar{\lambda}\right)^2+\frac{1}{N}\sigma_v^2\beta^2+\frac{1}{N}\sigma_\varepsilon^2.$$

- C3

$$\begin{array}{rcl} \underset{T\to\infty}{\mathrm{plim}}\dfrac{1}{T}x_i'\bar{y} & = & \gamma_i\left(\beta\bar{\gamma}+\bar{\lambda}\right)\sigma_f^2+\dfrac{1}{N}\sigma_v^2\beta, \\ \underset{T\to\infty}{\mathrm{plim}}\dfrac{1}{T}\bar{x}'\bar{y} & = & \bar{\gamma}\left(\beta\bar{\gamma}+\bar{\lambda}\right)\sigma_f^2+\dfrac{1}{N}\sigma_v^2\beta. \end{array}$$

- C4

$$\begin{array}{rcl} \underset{T\to\infty}{\mathrm{plim}}\dfrac{1}{T}\bar{x}'f\lambda_i & = & \lambda_i\bar{\gamma}\sigma_f^2, \\ \underset{T\to\infty}{\mathrm{plim}}\dfrac{1}{T}\bar{y}'f\lambda_i & = & \lambda_i\bar{\lambda}\sigma_f^2. \end{array}$$

- C5

$$\begin{array}{rcl} \underset{T\to\infty}{\mathrm{plim}}\dfrac{1}{T}\bar{x}'\varepsilon_i & = & 0, \\ \underset{T\to\infty}{\mathrm{plim}}\dfrac{1}{T}\bar{y}'\varepsilon_i & = & \dfrac{1}{N}\sigma_\varepsilon^2. \end{array}$$

Proof:

We will give a proof of one of each of the five sub-lemmas, the others follow straightforwardly by replacement of the single time series with the functional form of the cross-sectional average time series. Begin with C1. By assumption we have that $E\left(v_{i,t}v_{j,s}\right)=0$ and $E\left(v_{i,t}f_t\right)=0$ for all i, j, t and s, and thus:

$$\begin{aligned}\operatorname*{plim}_{T\to\infty}\frac{1}{T}x_i'\bar{x} &= \operatorname*{plim}_{T\to\infty}\frac{1}{T}\left(\gamma_i f+v_i\right)'\left(\bar{\gamma}f+\bar{v}\right)\\ &= \gamma_i\bar{\gamma}\operatorname*{plim}_{T\to\infty}\frac{1}{T}f'f+\operatorname*{plim}_{T\to\infty}\frac{1}{T}v_i\frac{1}{N}\sum_{i=1}^{N}v_i\\ &= \sigma_f^2\gamma_i\bar{\gamma}+\frac{1}{N}\sigma_v^2,\end{aligned}$$

since:

$$\begin{aligned}\operatorname*{plim}_{T\to\infty}\frac{1}{T}v_i\frac{1}{N}\sum_{i=1}^{N}v_i &= \operatorname*{plim}_{T\to\infty}\frac{1}{T}v_iv_i+\frac{1}{N}\sum_{j=1}^{N-1}v_iv_j\\ &= \operatorname*{plim}_{T\to\infty}\frac{1}{T}v_iv_i.\end{aligned}$$

Similarly, for C2 we have:

$$\begin{aligned}\operatorname*{plim}_{T\to\infty}\frac{1}{T}\bar{y}'\bar{y} &= \frac{1}{T}\left[\left(\beta\bar{\gamma}+\bar{\lambda}\right)f+\bar{v}\beta+\bar{\varepsilon}\right]'\left[\left(\beta\bar{\gamma}+\bar{\lambda}\right)f+\bar{v}\beta+\bar{\varepsilon}\right]\\ &= \sigma_f^2\left(\beta\bar{\gamma}+\bar{\lambda}\right)^2+\frac{1}{N}\sigma_v^2\beta^2+\frac{1}{N}\sigma_\varepsilon^2,\end{aligned}$$

since $E\left(\varepsilon_{i,t}f_t\right)=0$, $E\left(v_{i,t}f_t\right)=0$ and $E\left(\varepsilon_{i,t}v_{i,t}\right)=0$.

For C3 we have:

$$\begin{aligned}\operatorname*{plim}_{T\to\infty}\frac{1}{T}\bar{x}'\bar{y} &= \operatorname*{plim}_{T\to\infty}\frac{1}{T}\left(\bar{\gamma}f+\bar{v}\right)'\left[\left(\beta\bar{\gamma}+\bar{\lambda}\right)f+\bar{v}\beta+\bar{\varepsilon}\right]\\ &= \bar{\gamma}\left(\beta\bar{\gamma}+\bar{\lambda}\right)\operatorname*{plim}_{T\to\infty}\frac{1}{T}f'f+\beta\operatorname*{plim}_{T\to\infty}\frac{1}{T}\bar{v}'\bar{v}\\ &= \bar{\gamma}\left(\beta\bar{\gamma}+\bar{\lambda}\right)\sigma_f^2+\beta\operatorname*{plim}_{T\to\infty}\frac{1}{T}\left(\frac{1}{N}\sum_{i=1}^{N}v_i\right)'\left(\frac{1}{N}\sum_{i=1}^{N}v_i\right)\\ &= \bar{\gamma}\left(\beta\bar{\gamma}+\bar{\lambda}\right)\sigma_f^2+\beta\operatorname*{plim}_{T\to\infty}\frac{1}{T}\left(\frac{1}{N^2}\sum_{i=1}^{N}v_iv_i\right)\\ &= \bar{\gamma}\left(\beta\bar{\gamma}+\bar{\lambda}\right)\sigma_f^2+\frac{1}{N}\sigma_v^2\beta.\end{aligned}$$

In C4, we have:

$$\begin{aligned}\operatorname*{plim}_{T\to\infty}\frac{1}{T}\bar{x}'f\lambda_i &= \operatorname*{plim}_{T\to\infty}\frac{1}{T}\left(\bar{\gamma}f+\bar{v}\right)'f\lambda_i \\ &= \lambda_i\bar{\gamma}\sigma^2.\end{aligned}$$

Finally, for C5 clearly, since $E\left(\varepsilon_{i,t}f_t\right)=0$ and $E\left(\varepsilon_{i,t}v_{i,t}\right)=0$, the first equality follows. For the second equality:

$$\begin{aligned}\operatorname*{plim}_{T\to\infty}\frac{1}{T}\bar{y}'\varepsilon_i &= \operatorname*{plim}_{T\to\infty}\frac{1}{T}\left[\left(\beta\bar{\gamma}+\bar{\lambda}\right)f+\bar{v}\beta+\bar{\varepsilon}\right]'\varepsilon_i \\ &= \operatorname*{plim}_{T\to\infty}\frac{1}{T}\bar{\varepsilon}'\varepsilon_i \\ &= \operatorname*{plim}_{T\to\infty}\frac{1}{NT}\sum_{i=1}^{N}\varepsilon_i'\varepsilon_j,\end{aligned}$$

and the result follows because $E\left(\varepsilon_{i,t}\varepsilon_{j,s}\right)=0$ for all i, j, t and s. □

Proof of Proposition 2.3:

Start from the probability limit of the bias:

$$\begin{aligned}\operatorname*{plim}_{T\to\infty}\left(\hat{\beta}_{CCE}-\beta\right) &= \frac{\sum_{i=1}^{N}E\left(x_i'\overline{M}f\lambda_i\right)}{\sum_{i=1}^{N}E\left(x_i'\overline{M}x_i\right)}+\frac{\sum_{i=1}^{N}E\left(x_i'\overline{M}\varepsilon_i\right)}{\sum_{i=1}^{N}E\left(x_i'\overline{M}x_i\right)} \\ &:= \frac{A+B}{C},\end{aligned} \tag{2.5}$$

say. Note how all three terms in (2.5) consist of products of second moments. As T goes to infinity, the probability limit of the products in A, B and C is equal to the product of their probability limits under assumptions 1-3 and we can replace the quantities in (2.5) with their asymptotic counterparts. As an example consider A:

$$\begin{aligned}\sum_{i=1}^{N}E\left(x_i'\overline{M}f\lambda_i\right) &= \sum_{i=1}^{N}E\left\{\left(x_i'f\lambda_i\right)-\begin{bmatrix}x_i'\bar{x} & x_i'\bar{y}\end{bmatrix}\begin{bmatrix}\bar{x}'\bar{x} & \bar{x}'\bar{y}\\ \bar{y}'\bar{x} & \bar{y}'\bar{y}\end{bmatrix}^{-1}\begin{bmatrix}\bar{x}'f\lambda_i\\ \bar{y}'f\lambda_i\end{bmatrix}\right\} \\ &= \sum_{i=1}^{N}E\left(x_i'f\lambda_i\right)- \\ &\quad \sum_{i=1}^{N}\left\{\begin{bmatrix}E\left(x_i'\bar{x}\right) & E\left(x_i'\bar{y}\right)\end{bmatrix}\begin{bmatrix}E\left(\bar{x}'\bar{x}\right) & E\left(\bar{x}'\bar{y}\right)\\ E\left(\bar{y}'\bar{x}\right) & E\left(\bar{y}'\bar{y}\right)\end{bmatrix}^{-1}\begin{bmatrix}E\left(\bar{x}'f\lambda_i\right)\\ E\left(\bar{y}'f\lambda_i\right)\end{bmatrix}\right\}.\end{aligned}$$

The determinant associated with the projection matrix in curly brackets is a scalar constant and in anticipa-

tion of the rather tedious algebra involved in the coming proof, rewrite A, B and C slightly, for example:

$$\begin{aligned} A &= \sum_{i=1}^{N} E\left(x_i' \overline{M} f \lambda_i\right) \\ &= \left\{E\left(\overline{y}'\overline{y}\right)E\left(\overline{x}'\overline{x}\right) - E\left(\overline{x}'\overline{y}\right)E\left(\overline{y}'\overline{x}\right)\right\}\sum_{i=1}^{N} E\left(x_i' f \lambda_i\right) - \\ &\quad \sum_{i=1}^{N}\left\{\left[\begin{array}{cc} E\left(x_i'\overline{x}\right) & E\left(x_i'\overline{y}\right)\end{array}\right]\left[\begin{array}{cc} E\left(\overline{y}'\overline{y}\right) & -E\left(\overline{x}'\overline{y}\right) \\ -E\left(\overline{y}'\overline{x}\right) & E\left(\overline{x}'\overline{x}\right)\end{array}\right]\left[\begin{array}{c} E\left(\overline{x}' f \lambda_i\right) \\ E\left(\overline{y}' f \lambda_i\right)\end{array}\right]\right\}. \end{aligned}$$

Further expanding the quantity A, we find the following three asymptotic terms constituting A:

$$\begin{aligned} A1 &= E\left(\overline{x}'\overline{x}\right)E\left(\overline{y}'\overline{y}\right)\sum_{i=1}^{N} E\left(x_i' f \lambda_i\right) - E\left(\overline{y}'\overline{y}\right)\sum_{i=1}^{N} E\left(x_i'\overline{x}\right)E\left(\overline{x}' f \lambda_i\right), \\ A2 &= E\left(\overline{y}'\overline{x}\right)E\left(\overline{x}'\overline{y}\right)\sum_{i=1}^{N} E\left(x_i' f \lambda_i\right) - E\left(\overline{x}'\overline{y}\right)\sum_{i=1}^{N} E\left(x_i'\overline{y}\right)E\left(\overline{x}' f \lambda_i\right), \\ A3 &= -E\left(\overline{x}'\overline{y}\right)\sum_{i=1}^{N} E\left(x_i'\overline{x}\right)E\left(\overline{y}' f \lambda_i\right) + E\left(\overline{x}'\overline{x}\right)\sum_{i=1}^{N} E\left(x_i'\overline{y}\right)E\left(\overline{y}' f \lambda_i\right), \end{aligned}$$

such that $A = A1 - A2 - A3$. Now substituting the quantities from Lemma C, we find:

$$\begin{aligned} A1 &= \sigma_f^2\sigma_v^2\sum_{i=1}^{N}\lambda_i\left(\gamma_i - \overline{\gamma}\right)\left[\sigma_\varepsilon^2 + \beta^2\sigma_v^2 + N\sigma_f^2\left(\overline{\lambda} + \beta\overline{\gamma}\right)^2\right]\frac{1}{N^2}, \\ A2 &= \beta\sigma_f^2\sigma_v^2\sum_{i=1}^{N}\lambda_i\left(\gamma_i - \overline{\gamma}\right)\left[\beta\sigma_v^2 + N\sigma_f^2\left(\overline{\lambda}\overline{\gamma} + \beta\overline{\gamma}^2\right)\right]\frac{1}{N^2}, \\ A3 &= \left(\sigma_f^2\right)^2\sigma_v^2\sum_{i=1}^{N}\lambda_i\overline{\lambda}\left(\gamma_i - \overline{\gamma}\right)\left(\overline{\lambda} + \beta\overline{\gamma}\right)\frac{1}{N^2}. \end{aligned}$$

Similarly for B, we find:

$$\begin{aligned} B1 &= E\left(\overline{x}'\overline{x}\right)E\left(\overline{y}'\overline{y}\right)\sum_{i=1}^{N} E\left(x_i'\varepsilon_i\right) - E\left(\overline{y}'\overline{y}\right)\sum_{i=1}^{N} E\left(x_i'\overline{x}\right)E\left(\overline{x}'\varepsilon_i\right), \\ B2 &= E\left(\overline{y}'\overline{x}\right)E\left(\overline{x}'\overline{y}\right)\sum_{i=1}^{N} E\left(x_i'\varepsilon_i\right) - E\left(\overline{x}'\overline{y}\right)\sum_{i=1}^{N} E\left(x_i'\overline{y}\right)E\left(\overline{x}'\varepsilon_i\right), \\ B3 &= -E\left(\overline{x}'\overline{y}\right)\sum_{i=1}^{N} E\left(x_i'\overline{x}\right)E\left(\overline{y}'\varepsilon_i\right) + E\left(\overline{x}'\overline{x}\right)\sum_{i=1}^{N} E\left(x_i'\overline{y}\right)E\left(\overline{y}'\varepsilon_i\right), \end{aligned}$$

or, after replacing and summing over i:

$$
\begin{aligned}
B1 &= 0, \\
B2 &= 0, \\
B3 &= \sigma_\varepsilon^2 \sigma_f^2 \sigma_v^2 \lambda_i \left(\gamma_i - \bar{\gamma}\right) \left(\bar{\lambda} + \beta\bar{\gamma}\right) \frac{1}{N^2},
\end{aligned}
$$

so that the numerator of (2.5) is equal to $A - B3$ the numerator is:

$$
A + B := \frac{1}{N^2} \sigma_\varepsilon^2 \sigma_f^2 \sigma_v^2 \sum_{i=1}^{N} \left[\left(\gamma_i - \bar{\gamma}\right) \left(\lambda_i - \bar{\lambda}\right) \right]
$$

We skip a similar proof for the denominator of (2.5) as the algebra is tedious. The expression for the denominator is:

$$
\begin{aligned}
C := & \\
& \sigma_v^2 \left[\tfrac{1}{N} \sigma_f^2 \left(\bar{\lambda}^2 \sigma_v^2 + \bar{\gamma} \sigma_\varepsilon^2 \right) + \tfrac{1}{N^2} \left(\sigma_\varepsilon^2 \sigma_v^2 + \sigma_\varepsilon^2 \sigma_f^2 \textstyle\sum_{i=1}^{N} \gamma_i^2 - \bar{\lambda}^2 \sigma_f^2 \sigma_v^2 - 2\bar{\gamma}^2 \sigma_f^2 \sigma_v^2 \right) - \tfrac{1}{N^3} \sigma_\varepsilon^2 \sigma_v^2 \right].
\end{aligned}
$$

Using C, dividing and cancelling terms then leads to the result in Proposition 2.3. □

LEMMA D: *Let Assumptions 1-4 hold, then for* $\beta = 0$*:*

- D1:

$$
\begin{aligned}
\operatorname*{plim}_{T\to\infty} \frac{1}{T} \sum_{i=1}^{N} y_i' y_i &= \operatorname*{plim}_{T\to\infty} \frac{1}{T} \sum_{i=1}^{N} y_{i,-1}' y_{i,-1} = \\
\sum_{i=1}^{N} \operatorname{var}(y_i) &= \frac{1}{1-\rho^2} \sum_{i=1}^{N} \left[\lambda_i^2 \sigma_f^2 + 2\lambda_i^2 \pi + \sigma_\varepsilon^2 \right]
\end{aligned}
$$

- D2

$$
\operatorname*{plim}_{T\to\infty} \frac{1}{T} \sum_{i=1}^{N} y_{i,-1}' f \lambda_i = \frac{1}{\rho} \sum_{i=1}^{N} \lambda_i^2 \pi
$$

Proof:

Consider first D1:

$$\begin{aligned}
\operatorname*{plim}_{T\to\infty}\frac{1}{T}\sum_{i=1}^{N}y_i'y_i &= \operatorname*{plim}_{T\to\infty}\frac{1}{T}\sum_{i=1}^{N}\left(\rho y_{i,-1}+f\lambda_i+\varepsilon_i\right)'\left(\rho y_{i,-1}+f\lambda_i+\varepsilon_i\right)\\
\sum_{i=1}^{N}\operatorname{var}\left(y_i\right) &= \sum_{i=1}^{N}\left[\rho^2\operatorname{var}\left(y_{i,-1}\right)+\lambda_i^2\sigma_f^2+2\lambda_i^2\pi+\sigma_\varepsilon^2\right]\\
\sum_{i=1}^{N}\operatorname{var}\left(y_i\right) &= \frac{\sigma_\varepsilon^2+\left(\sigma_f^2+2\sum_{s=1}^{T}\rho^s E\left(f_t'f_{t-s}\right)\right)\sum_{i=1}^{N}\lambda_i^2}{1-\rho^2},
\end{aligned}$$

by assumption, $E\left(\varepsilon_{i,t}\varepsilon_{i,s}\right)=0$ for all $s\neq t$ and furthermore the cross-term:

$$\begin{aligned}
\operatorname*{plim}_{T\to\infty}\frac{1}{T}\rho y_{i,-1}'f\lambda_i &= \operatorname*{plim}_{T\to\infty}\frac{1}{T}\rho\left(\rho y_{i,-2}+f_{-1}\lambda_i+\varepsilon_{i-1}\right)'f\lambda_i\\
&= \operatorname*{plim}_{T\to\infty}\left(\frac{1}{T}\lambda_i^2\rho f_{-1}'f+\frac{1}{T}\lambda_i^2\rho^2 f_{-2}'f+\ldots\right)\\
&= \lambda_i^2\pi,
\end{aligned}$$

which is infinitely summable by assumption of absolutely summable covariance stationarity and independent of the summand of ρ, which itself is square-summable by assumption that $-1<\rho<1$. Cauchy-Schwarz then gives the $\pi<\infty$. D.2 is similarly derived. □

Proof of Proposition 2.4:

We have for the pooled OLS estimator:

$$\begin{aligned}
\operatorname*{plim}_{T\to\infty}\left(\hat{\rho}_{OLS}-\rho\right) &= \operatorname*{plim}_{T\to\infty}\left(\frac{1}{T}\sum_{i=1}^{N}y_{i,-1}'y_{i,-1}\right)^{-1}\operatorname*{plim}_{T\to\infty}\frac{1}{T}\sum_{i=1}^{N}y_{i,-1}'y_i\\
&= \left(\operatorname*{plim}_{T\to\infty}\frac{1}{T}\sum_{i=1}^{N}y_{i,-1}'y_{i,-1}\right)^{-1}\operatorname*{plim}_{T\to\infty}\frac{1}{T}\sum_{i=1}^{N}y_{i,-1}'\left(f\lambda_i+\varepsilon_i\right)\\
&= I\times\left(II+III\right).
\end{aligned}$$

The proofs of *I* and *II* follow directly from Lemma D. *III* is by assumption, i.e:

$$\begin{aligned}\operatorname*{plim}_{T\to\infty}\frac{1}{T}\sum_{i=1}^{N}y_{i,-1}'\varepsilon_i &= \operatorname*{plim}_{T\to\infty}\frac{1}{T}\sum_{i=1}^{N}\left(\rho y_{i,-1}+f\lambda_i+\varepsilon_{i,-1}\right)'\varepsilon_i \\ &= \frac{1}{\rho}\sum_{i=1}^{N}\sum_{s=1}^{T}E\left(\varepsilon_i'\varepsilon_{i,-s}\right) \\ &= 0,\end{aligned}$$

since $E\left(f_t\varepsilon_{i,s}\right)=0$ and $E\left(\varepsilon_{i,t}\varepsilon_{i,s}\right)=0$ for all $s\neq t$. The proposition then follows immediately from D1 and D2.□

LEMMA E: *Let Assumptions 1-4 hold, then for* $\beta=0$*:*

- E1:

$$\begin{aligned}\operatorname*{plim}_{T\to\infty}\frac{1}{T}\sum_{i=1}^{N}\tilde{y}_i'\tilde{y}_i &= \operatorname*{plim}_{T\to\infty}\frac{1}{T}\sum_{i=1}^{N}\tilde{y}_{i,-1}'\tilde{y}_{i,-1}= \\ \sum_{i=1}^{N}\operatorname{var}\left(\tilde{y}_i\right) &= \frac{1}{1-\rho^2}\sum_{i=1}^{N}\left[\tilde{\lambda}_i^2\sigma_f^2+2\tilde{\lambda}_i^2\pi+\sigma_{\tilde{\varepsilon}}^2\right]\end{aligned}$$

- E2

$$\operatorname*{plim}_{T\to\infty}\frac{1}{T}\sum_{i=1}^{N}\tilde{y}_{i,-1}'f\tilde{\lambda}_i = \frac{1}{\rho}\sum_{i=1}^{N}\tilde{\lambda}_i^2\pi$$

Proof:

Using $\tilde{a}=a_i-N^{-1}\sum_{i=1}^{N}a_i$ for a T-vector a_i as before, consider E1:

$$\begin{aligned}\operatorname*{plim}_{T\to\infty}\frac{1}{T}\sum_{i=1}^{N}\tilde{y}_i'\tilde{y}_i &= \operatorname*{plim}_{T\to\infty}\frac{1}{T}\sum_{i=1}^{N}\left(\rho\tilde{y}_{i,-1}+f\tilde{\lambda}_i+\tilde{\varepsilon}_i\right)'\left(\rho\tilde{y}_{i,-1}+f\tilde{\lambda}_i+\tilde{\varepsilon}_i\right) \\ \sum_{i=1}^{N}\operatorname{var}\left(\tilde{y}_i\right) &= \sum_{i=1}^{N}\left[\rho^2\operatorname{var}\left(\tilde{y}_{i,-1}\right)+\tilde{\lambda}_i^2\sigma_f^2+2\tilde{\lambda}_i^2\pi+\sigma_{\tilde{\varepsilon}}^2\right] \\ \sum_{i=1}^{N}\operatorname{var}\left(\tilde{y}_i\right) &= \frac{\sigma_\varepsilon^2+\left(\sigma_f^2+2\pi\right)\sum_{i=1}^{N}\tilde{\lambda}_i^2}{1-\rho^2},\end{aligned}$$

by assumption, $E\left(\varepsilon_{i,t}\varepsilon_{i,s}\right)=0$ for all $s\neq t$. E2 follows as D2, using the redefined variables in its place.□

Proof of Proposition 2.5:

The proof is identical to the proof of 2.5. The sequential limit follows from noticing that $\operatorname*{plim}_{N\to\infty}\frac{1}{N}\sum_{i=1}^{N}\tilde{\lambda}_i^2=\sigma_\lambda^2$.

LEMMA F: *Let Assumptions 1-4 hold, then for* $\beta=0$*:*

- F1

$$\begin{aligned}
\operatorname*{plim}_{T\to\infty}\frac{1}{T}y_{i,-1}'y_{i,-1} &= \frac{\sigma_\varepsilon^2+\lambda_i^2\left(\sigma_f^2+2\pi\right)}{1-\rho^2},\\
\operatorname*{plim}_{T\to\infty}\frac{1}{T}y_{i,-1}'\bar{y}_{-1} &= \frac{\frac{1}{N}\sigma_\varepsilon^2+\lambda_i\bar{\lambda}\left(\sigma_f^2+2\pi\right)}{1-\rho^2},\\
\operatorname*{plim}_{T\to\infty}\frac{1}{T}\bar{y}_{-1}'\bar{y}_{-1} &= \frac{\frac{1}{N}\sigma_\varepsilon^2+\bar{\lambda}^2\left(\sigma_f^2+2\pi\right)}{1-\rho^2}.
\end{aligned}$$

- F2

$$\begin{aligned}
\operatorname*{plim}_{T\to\infty}\frac{1}{T}y_{i,-1}'\bar{y} &= \frac{\rho}{1-\rho^2}\left[\frac{1}{N}\sigma_\varepsilon^2+\lambda_i\bar{\lambda}\sigma_f^2+\lambda_i\bar{\lambda}\left(1+\frac{1}{\rho^2}\right)\pi\right],\\
\operatorname*{plim}_{T\to\infty}\frac{1}{T}\bar{y}_{-1}'\bar{y} &= \frac{\rho}{1-\rho^2}\left[\frac{1}{N}\sigma_\varepsilon^2+\bar{\lambda}^2\sigma_f^2+\bar{\lambda}^2\left(1+\frac{1}{\rho^2}\right)\pi\right].
\end{aligned}$$

- F3

$$\begin{aligned}
\operatorname*{plim}_{T\to\infty}\frac{1}{T}y_{i,-1}'f\lambda_i &= \frac{1}{\rho}\lambda_i^2\pi,\\
\operatorname*{plim}_{T\to\infty}\frac{1}{T}\bar{y}_{-1}'f\lambda_i &= \frac{1}{\rho}\lambda_i\bar{\lambda}\pi,\\
\operatorname*{plim}_{T\to\infty}\frac{1}{T}\bar{y}'f\lambda_i &= \lambda_i\bar{\lambda}\pi+\lambda_i\bar{\lambda}\sigma_f^2.
\end{aligned}$$

- F4

$$\begin{aligned}
\operatorname*{plim}_{T\to\infty}\frac{1}{T}y_{i,-1}'\varepsilon_i &= 0,\\
\operatorname*{plim}_{T\to\infty}\frac{1}{T}\bar{y}_{-1}'\varepsilon_i &= 0,\\
\operatorname*{plim}_{T\to\infty}\frac{1}{T}\bar{y}'\varepsilon_i &= \frac{1}{N}\sigma_\varepsilon^2
\end{aligned}$$

Proof:

As in the proof of Lemma C, we only derive one asymptotic representation of each sub-lemma. The other representations can be easily derived using appropriate definitions of the variables involved. Focus first on

the second element of F1:

$$\begin{aligned}
\operatorname*{plim}_{T\to\infty}\frac{1}{T}y'_{i,-1}\overline{y}_{-1} &= \operatorname*{plim}_{T\to\infty}\frac{1}{T}\left(\rho y_{i,-2}+f_{-1}\lambda_i+\varepsilon_{i,-1}\right)'\left(\rho\overline{y}_{,-2}+f_{-1}\overline{\lambda}+\overline{\varepsilon}_{-1}\right)\\
&= \rho^2\operatorname*{plim}_{T\to\infty}\frac{1}{T}y'_{i,-2}\overline{y}_{-2}+2\lambda_i\overline{\lambda}\operatorname*{plim}_{T\to\infty}\frac{1}{T}\sum_{s=1}^{T}\beta^s\left(f'_t f_{t-s}\right)+\\
&\quad \lambda_i\overline{\lambda}\operatorname*{plim}_{T\to\infty}\frac{1}{T}f'f+\operatorname*{plim}_{T\to\infty}\frac{1}{T}\varepsilon'_i\overline{\varepsilon}.
\end{aligned}$$

Since $\operatorname*{plim}_{T\to\infty}\frac{1}{T}\varepsilon'_i\varepsilon_i=\operatorname*{plim}_{T\to\infty}\frac{1}{T}\varepsilon'_{i,-1}\varepsilon_{i,-1}$ and similarly for f_t and thus $y_{i,t}$.

For F2 we have:

$$\begin{aligned}
\operatorname*{plim}_{T\to\infty}\frac{1}{T}y'_{i,-1}\overline{y} &= \operatorname*{plim}_{T\to\infty}\frac{1}{T}\left(\rho y_{i,-2}+f_{-1}\lambda_i+\varepsilon_{i,-1}\right)'\left(\rho\overline{y}_{,-1}+f\overline{\lambda}+\overline{\varepsilon}\right)\\
&= \operatorname*{plim}_{T\to\infty}\frac{1}{T}\rho^2 y'_{i,-2}\overline{y}_{-1}+\overline{\lambda}\operatorname*{plim}_{T\to\infty}\frac{1}{T}\rho y'_{i,-2}f+\\
&\quad \lambda_i\rho\operatorname*{plim}_{T\to\infty}\frac{1}{T}f'_{-1}\overline{y}_{,-1}+\lambda_i\overline{\lambda}\operatorname*{plim}_{T\to\infty}\frac{1}{T}f'_{-1}f+\\
&\quad \rho\operatorname*{plim}_{T\to\infty}\frac{1}{T}\varepsilon'_{-1}\overline{y}_{,-1},
\end{aligned}$$

by the assumptions $E\left(\varepsilon_{i,s}f_s\right)=0$ and $E\left(\varepsilon_{i,t}\varepsilon_{i,s}\right)=0$ for all i, j, t and s, moreover,

$$\overline{\lambda}\operatorname*{plim}_{T\to\infty}\frac{1}{T}\rho y'_{i,-2}f \quad +\lambda_i\overline{\lambda}\operatorname*{plim}_{T\to\infty}\tfrac{1}{T}f'_{-1}f=\quad \frac{1}{\rho}\lambda_i\overline{\lambda}\pi,$$

$$\begin{aligned}
\lambda_i\rho\operatorname*{plim}_{T\to\infty}\frac{1}{T}f'_{-1}\overline{y}_{,-1} &= \lambda_i\overline{\lambda}\rho\operatorname*{plim}_{T\to\infty}\frac{1}{T}f'_{-1}f_{-1}+\\
&\quad \lambda_i\rho^2\operatorname*{plim}_{T\to\infty}\frac{1}{T}f'_{-1}\overline{y}_{,-2}\\
&= \lambda_i\overline{\lambda}\rho\sigma_f^2+\lambda_i\overline{\lambda}\rho\pi,
\end{aligned}$$

and:

$$\begin{aligned}
\rho\operatorname*{plim}_{T\to\infty}\frac{1}{T}\varepsilon'_{-1}\overline{y}_{,-1} &= \rho\operatorname*{plim}_{T\to\infty}\frac{1}{T}\varepsilon'_{-1}\left(\rho\overline{y}_{,-2}+f_{-1}\overline{\lambda}+\overline{\varepsilon}_{-1}\right)\\
&= \rho\operatorname*{plim}_{T\to\infty}\varepsilon'_{i,-1}\overline{\varepsilon}_{-1}\\
&= \rho\frac{1}{N}\sigma_\varepsilon^2
\end{aligned}$$

we then have finally:

$$\underset{T\to\infty}{\text{plim}}\frac{1}{T}y'_{i,-1}\bar{y} = \frac{\rho}{1-\rho^2}\left[\frac{1}{N}\sigma_\varepsilon^2+\lambda_i\bar{\lambda}\sigma_f^2+\lambda_i\bar{\lambda}\left(1+\frac{1}{\rho^2}\right)\pi\right].$$

Regarding F3, let:

$$\begin{aligned}\underset{T\to\infty}{\text{plim}}\frac{1}{T}\bar{y}'_{-1}f\lambda_i &= \underset{T\to\infty}{\text{plim}}\frac{1}{T}\left(\rho\bar{y}_{,-2}+f_{-1}\bar{\lambda}+\bar{\varepsilon}_{-1}\right)'f\lambda_i \\ &= \lambda_i\underset{T\to\infty}{\text{plim}}\frac{1}{T}\rho\bar{y}'_{,-2}f+\lambda_i\bar{\lambda}\underset{T\to\infty}{\text{plim}}\frac{1}{T}f'_{,-1}f=\frac{1}{\rho}\lambda_i\bar{\lambda}\pi.\end{aligned}$$

Similarly,

$$\begin{aligned}\underset{T\to\infty}{\text{plim}}\frac{1}{T}\bar{y}'f\lambda_i &= \underset{T\to\infty}{\text{plim}}\frac{1}{T}\left(\rho\bar{y}_{,-1}+f\bar{\lambda}+\bar{\varepsilon}\right)'f\lambda_i \\ &= \lambda_i\underset{T\to\infty}{\text{plim}}\frac{1}{T}\rho\bar{y}'_{,-1}f+\lambda_i\bar{\lambda}\underset{T\to\infty}{\text{plim}}\frac{1}{T}f'f \\ &= \lambda_i\bar{\lambda}\pi+\lambda_i\bar{\lambda}\sigma_f^2.\end{aligned}$$

Finally for F4, by assumption $E(\varepsilon_{i,s}f_s)=0$ and $E(\varepsilon_{i,t}\varepsilon_{i,s})=0$ for all i, j, t and s so that the first two terms are obvious. For the third, we note that $\underset{T\to\infty}{\text{plim}}\frac{1}{T}\varepsilon'_i\bar{\varepsilon}=\frac{1}{N}\sigma_\varepsilon^2$. □

Proof of Proposition 2.6:

The proof employs exactly the same strategy as with Proposition 2.3: We start with the probability limit of the bias of the CCE estimator and write:

$$\begin{aligned}E(\hat{\rho}_{CCE}-\rho) &= \frac{\sum_{i=1}^N E\left(y'_{i,-1}\overline{M}f\lambda_i\right)}{\sum_{i=1}^N E\left(y'_{i,-1}\overline{M}y_{i,-1}\right)}+\frac{\sum_{i=1}^N E\left(y'_{i,-1}\overline{M}\varepsilon_i\right)}{\sum_{i=1}^N E\left(y'_{i,-1}\overline{M}y_{i,-1}\right)} \\ &:= \frac{A+B}{C}.\end{aligned}$$

Again under assumptions 1-6, the probability limit of the product will equal the product of the asymptotic representations of the second moments included in A, B and C. We will therefore proceed to write the three

terms from expanding A as before:

$$\begin{aligned}
A1 &:= E\left(\overline{y}'\overline{y}\right)E\left(\overline{y}'_{-1}\overline{y}_{-1}\right)\sum_{i=1}^{N}E\left(y'_{i,-1}f\lambda_i\right)-E\left(\overline{y}'\overline{y}\right)\sum_{i=1}^{N}\left[E\left(y'_{i,-1}\overline{y}_{-1}\right)E\left(\overline{y}'_{-1}f\lambda_i\right)\right],\\
A2 &:= E\left(\overline{y}'_{-1}\overline{y}\right)E\left(\overline{y}'\overline{y}_{-1}\right)\sum_{i=1}^{N}E\left(y'_{i,-1}f\lambda_i\right)-E\left(\overline{y}'_{-1}\overline{y}\right)\sum_{i=1}^{N}\left[E\left(y'_{i,-1}\overline{y}\right)E\left(\overline{y}'_{-1}f\lambda_i\right)\right],\\
A3 &:= -E\left(\overline{y}'\overline{y}_{-1}\right)\sum_{i=1}^{N}E\left(y'_{i,-1}\overline{y}_{-1}\right)E\left(\overline{y}'f\lambda_i\right)+E\left(\overline{y}'_{-1}\overline{y}_{-1}\right)\sum_{i=1}^{N}\left[E\left(y'_{i,-1}\overline{y}\right)E\left(\overline{y}'f\lambda_i\right)\right].
\end{aligned}$$

Replacing the expectations with their limits and simplifying we find that:

$$\begin{aligned}
A1 &= \sum_{i=1}^{N}\left(\lambda_i^2-\overline{\lambda}^2\right)\frac{\left(\frac{1}{N}\sigma_\varepsilon^2\right)\pi\left[\left(\frac{1}{N}\sigma_\varepsilon^2\right)+\overline{\lambda}^2\left(\sigma_f^2+2\pi\right)\right]}{\rho\left(\rho^2-1\right)^2},\\
A2 &= -\sum_{i=1}^{N}\left(\lambda_i^2-\overline{\lambda}^2\right)\frac{\rho\left(\frac{1}{N}\sigma_\varepsilon^2\right)^2\pi}{\left(\rho^2-1\right)^2}-\\
&\quad\sum_{i=1}^{N}\left(\lambda_i^2-\overline{\lambda}^2\right)\frac{\overline{\lambda}^2\left(N^{-1}\sigma_\varepsilon^2\right)\pi\left[\left(1+\rho^2\right)\pi+\rho^2\right]}{\rho\left(\rho^2-1\right)^2},\\
A3 &= \sum_{i=1}^{N}\left(\lambda_i^2-\overline{\lambda}^2\right)\frac{\left(N^{-1}\sigma_\varepsilon^2\right)\pi}{\rho\left(\rho^2-1\right)^2}.
\end{aligned}$$

For B we find that:

$$\begin{aligned}
B1 &= E\left(\overline{y}'\overline{y}\right)E\left(\overline{y}'_{-1}\overline{y}_{-1}\right)\sum_{i=1}^{N}E\left(y'_{i,-1}\varepsilon_i\right)-E\left(\overline{y}'\overline{y}\right)\sum_{i=1}^{N}\left[E\left(y'_{i,-1}\overline{y}_{-1}\right)E\left(\overline{y}'_{-1}\varepsilon_i\right)\right],\\
B2 &= E\left(\overline{y}'_{-1}\overline{y}\right)E\left(\overline{y}'\overline{y}_{-1}\right)\sum_{i=1}^{N}E\left(y'_{i,-1}\varepsilon_i\right)-E\left(\overline{y}'_{-1}\overline{y}\right)\sum_{i=1}^{N}\left[E\left(y'_{i,-1}\overline{y}\right)E\left(\overline{y}'_{-1}\varepsilon_i\right)\right],\\
B3 &= -E\left(\overline{y}'\overline{y}_{-1}\right)\sum_{i=1}^{N}E\left(y'_{i,-1}\overline{y}_{-1}\right)E\left(\overline{y}'\varepsilon_i\right)+E\left(\overline{y}'_{-1}\overline{y}_{-1}\right)\sum_{i=1}^{N}\left[E\left(y'_{i,-1}\overline{y}\right)E\left(\overline{y}'\varepsilon_i\right)\right].
\end{aligned}$$

Since $E\left(\varepsilon'_{i,t}\varepsilon_{j,s}\right)=0$ for $t\neq s$ and $i\neq j$, it is straightforward to see that $B1$ and $B2$ are zero. It is not immediately straightforward to see that $B3$ is too, however note that since $\varepsilon_{i,t}$ is *i.i.d.*:

$$\begin{aligned}
B3 &= -E\left(\overline{y}'\overline{y}_{-1}\right)\sum_{i=1}^{N}E\left(y'_{i,-1}\overline{y}_{-1}\right)E\left(\overline{y}'\varepsilon_i\right)+E\left(\overline{y}'_{-1}\overline{y}_{-1}\right)\sum_{i=1}^{N}\left[E\left(y'_{i,-1}\overline{y}\right)E\left(\overline{y}'\varepsilon_i\right)\right]\\
&= -\frac{1}{N}\sigma_\varepsilon^2E\left(\overline{y}'\overline{y}_{-1}\right)E\left(\overline{y}'_{-1}\overline{y}_{-1}\right)+\frac{1}{N}\sigma_\varepsilon^2E\left(\overline{y}'_{-1}\overline{y}_{-1}\right)E\left(\overline{y}'\overline{y}_{-1}\right)\\
&= 0.
\end{aligned}$$

As a result, the numerator of the bias of the CCE estimator is equal to:

$$A1-A2-A3 \quad = \quad \frac{\left(\sigma_\varepsilon^2\right)^2 \pi N^{-1} \sum_{i=1}^{N} \left(\lambda_i - \overline{\lambda}\right)^2}{\rho\left(1-\rho^2\right)}.$$

The denominator is obtained in a similar fashion, but a proof is again omitted due to the tedious involved. The denominator C is:

$$C = \frac{\kappa_1\left(\sigma_\varepsilon^2/N\right)\left[(N-2)\overline{\lambda}^2+\sum_{i=1}^{N}\lambda_i^2\right]+\rho^2\left(\sigma_\varepsilon^2/N\right)^2\left[\kappa_2 N+\left(\left\{\sigma_\varepsilon^2/N\right\}+\kappa_3\right)(N-1)\right]}{\left\{\rho(1-\rho^2)\right\}^2},$$

with κ_1, κ_2 and κ_3 defined in the main body of the text. Division and simplification then yields the result in Proposition 2.3. □

Appendix 2.2 Monte Carlo Results

Table 2.1: Simulation results for static model (2.1), baseline:

	Bias of $\hat{\beta}_{OLS}$	*Predicted* $\hat{\beta}_{OLS}$ *bias*	*Bias of* $\hat{\beta}_{TE}$	*Predicted* $\hat{\beta}_{TE}$ *bias*	*Bias of* $\hat{\beta}_{CCE}$	*Predicted* $\hat{\beta}_{CCE}$ *bias*	*Bias of* $\hat{\beta}_{QDCCE}$
			$N = 5,$	$T = 100$			
Bias	0.2644	0.262	0.2619	0.2498	0.1504	0.0762	0.0271
% Bias	26.4393	26.1995	26.1946	24.9836	15.0447	7.6235	2.7119
RMSE	0.1419		0.1568	5	0.0632		0.029
			$N = 5,$	$T = 500$			
Bias	0.2685	0.2682	0.2595	0.2489	0.1462	0.1218	0.0097
% Bias	26.8475	26.8158	25.9451	24.8872	14.6218	12.18	0.9703
RMSE	0.139		0.149		0.0564		0.0104
			$N = 10,$	$T = 100$			
Bias	0.2767	0.2763	0.276	0.2705	0.1517	0.0768	0.0327
% Bias	27.6672	27.6328	27.5996	27.0533	15.1704	7.6837	3.2673
RMSE	0.1142		0.1173	5	0.0432		0.0352
			$N = 10,$	$T = 500$			
Bias	0.2752	0.2752	0.2729	0.2696	0.1547	0.1288	0.0133
% Bias	27.5173	27.5249	27.2948	26.9606	15.4667	12.8793	1.3317
RMSE	0.1112		0.1141	5	0.044		0.0116

Table 2.2: Simulation results for static model (2.1) with $\sigma_{\lambda\gamma} = 0.9$

	Bias of $\hat{\beta}_{OLS}$	*Predicted* $\hat{\beta}_{OLS}$ *bias*	*Bias of* $\hat{\beta}_{TE}$	*Predicted* $\hat{\beta}_{TE}$ *bias*	*Bias of* $\hat{\beta}_{CCE}$	*Predicted* $\hat{\beta}_{CCE}$ *bias*	*Bias of* $\hat{\beta}_{QDCCE}$
			$N = 5,$	$T = 100$			
Bias	0.3949	0.3941	0.3908	0.3909	0.2047	0.0806	0.0358
% Bias	39.495	39.4055	39.0778	39.0882	20.4682	8.0642	3.5777
RMSE	0.2388		0.2528		0.0906		0.0359
			$N = 5,$	$T = 500$			
Bias	0.3918	0.392	0.3882	0.39	0.2027	0.1554	0.0159
% Bias	39.1803	39.2032	38.8221	38.9953	20.2705	15.5414	1.5942
RMSE	0.2329		0.2466		0.0889		0.0159
			$N = 10,$	$T = 100$			
Bias	0.3987	0.3988	0.3967	0.3972	0.2026	0.0796	0.0481
% Bias	39.8656	39.8820	39.6718	39.7153	20.2571	7.9590	4.8103
RMSE	0.2046		0.2070		0.0715		0.0490
			$N = 10,$	$T = 500$			
Bias	0.4018	0.402	0.4001	0.4003	0.2045	0.1555	0.0216
% Bias	40.1766	40.1972	40.0117	40.0269	20.4463	15.5456	2.1646
RMSE	0.2029		0.2057		0.07		0.0213

Table 2.3: Simulation results for static model (2.1) with $\sigma_{\lambda}^{2} = 5$:

	Bias of $\hat{\beta}_{OLS}$	*Predicted* $\hat{\beta}_{OLS}$ *bias*	*Bias of* $\hat{\beta}_{TE}$	*Predicted* $\hat{\beta}_{TE}$ *bias*	*Bias of* $\hat{\beta}_{CCE}$	*Predicted* $\hat{\beta}_{CCE}$ *bias*	*Bias of* $\hat{\beta}_{QDCCE}$
			$N = 5,$	$T = 100$			
Bias	0.1187	0.1181	0.1124	0.1053	0.044	0.0224	0.0074
% Bias	11.8722	11.8083	11.2362	10.5277	4.4035	2.2381	0.7403
RMSE	0.3056		0.3713		0.0833		0.0251
			$N = 5,$	$T = 500$			
Bias	0.1056	0.1059	0.1031	0.0996	0.0415	0.0347	0.0028
% Bias	10.563	10.5865	10.3107	9.9636	4.1487	3.4696	0.2831
RMSE	0.3276		0.3919		0.0869		0.0149
			$N = 10,$	$T = 100$			
Bias	0.1236	0.1237	0.1207	0.1188	0.0465	0.0229	0.0032
% Bias	12.3617	12.3654	12.0657	11.8834	4.648	2.2867	0.3186
RMSE	0.1765		0.1925		0.0415		0.0322
			$N = 10,$	$T = 500$			
Bias	0.1137	0.1141	0.112	0.1088	0.0458	0.0379	0.0006
% Bias	11.3716	11.4102	11.1953	10.884	4.584	3.7883	0.0616
RMSE	0.177		0.1942		0.0424		0.0074

Table 2.4: Simulation results for dynamic model (2.1), baseline:

	Bias of $\hat{\rho}_{OLS}$	*Predicted* $\hat{\rho}_{OLS}$ *bias*	*Bias of* $\hat{\rho}_{TE}$	*Predicted* $\hat{\rho}_{TE}$ *bias*	*Bias of* $\hat{\rho}_{CCE}$	*Predicted* $\hat{\rho}_{CCE}$ *bias*	*Bias of* $\hat{\rho}_{QDCCE}$
			$N = 5,$	$T = 100$			
Bias	0.2238	0.2312	0.1778	0.1703	0.0077	0.0209	0.0142
% Bias	44.7512	46.2389	35.569	34.0696	1.5314	4.1786	2.8391
RMSE	0.0536		0.0366		0.0038		0.0052
			$N = 5,$	$T = 500$			
Bias	0.231	0.2323	0.1847	0.1718	0.0194	0.0214	0.0078
% Bias	46.1932	46.4623	36.9488	34.3501	3.8825	4.2788	1.5535
RMSE	0.0549		0.0373		0.0023		0.0017
			$N = 10,$	$T = 100$			
Bias	0.2294	0.2373	0.1893	0.1886	-0.0041	0.006	0.023
% Bias	45.8803	47.4605	37.8551	37.7247	-0.8281	1.2081	4.6039
RMSE	0.0554		0.0391		0.0012		0.0024
			$N = 10,$	$T = 500$			
Bias	0.2368	0.2389	0.1957	0.1905	0.0044	0.0063	0.0086
% Bias	47.3624	47.7709	39.1322	38.0923	0.8815	1.262	1.7291
RMSE	0.057		0.0398		0.0005		0.0005

Table 2.5: Simulation results for dynamic model (2.1), $\xi = 0.9$:

	Bias of $\hat{\rho}_{OLS}$	*Predicted* $\hat{\rho}_{OLS}$ *bias*	*Bias of* $\hat{\rho}_{TE}$	*Predicted* $\hat{\rho}_{TE}$ *bias*	*Bias of* $\hat{\rho}_{CCE}$	*Predicted* $\hat{\rho}_{CCE}$ *bias*	*Bias of* $\hat{\rho}_{QDCCE}$
			$N = 5,$	$T = 100$			
Bias	0.4317	0.4394	0.4018	0.3992	0.0159	0.0356	0.0118
% Bias	86.3377	87.8828	80.3549	79.8472	3.1744	7.1208	2.3629
RMSE	0.1877		0.1651		0.0076		0.0062
			$N = 5,$	$T = 500$			
Bias	0.4407	0.4421	0.4118	0.4032	0.0316	0.035	0.0045
% Bias	88.1406	88.4206	82.3505	80.6371	6.3111	7.0046	0.9082
RMSE	0.1947		0.172		0.0065		0.0019
			$N = 10,$	$T = 100$			
Bias	0.4362	0.4437	0.4137	0.417	-0.0085	0.0097	0.0287
% Bias	87.2444	88.7368	82.7306	83.4032	-1.7016	1.9339	5.742
RMSE	0.1912		0.173		0.0019		0.0026
			$N = 10,$	$T = 500$			
Bias	0.4443	0.4459	0.4241	0.4218	0.0051	0.0089	0.0067
% Bias	88.8607	89.1799	84.8145	84.3625	1.0258	1.7858	1.3446
RMSE	0.1976		0.1805		0.0009		0.0004

Table 2.6: Simulation results for dynamic model (2.1), $\sigma_{\lambda}^2 = 5$:

	Bias of $\hat{\rho}_{OLS}$	*Predicted* $\hat{\rho}_{OLS}$ *bias*	*Bias of* $\hat{\rho}_{TE}$	*Predicted* $\hat{\rho}_{TE}$ *bias*	*Bias of* $\hat{\rho}_{CCE}$	*Predicted* $\hat{\rho}_{CCE}$ *bias*	*Bias of* $\hat{\rho}_{QDCCE}$
			$N = 5,$	$T = 100$			
Bias	0.2605	0.269	0.2516	0.2527	0.0438	0.057	0.0424
% Bias	52.1014	53.806	50.3154	50.5399	8.7643	11.3905	8.4741
RMSE	0.0708		0.0668		0.011		0.0143
			$N = 5,$	$T = 500$			
Bias	0.2673	0.2692	0.2582	0.2526	0.0564	0.0583	0.028
% Bias	53.4503	53.8459	51.6457	50.5295	11.2855	11.6562	5.5998
RMSE	0.0724		0.0682		0.01		0.008
			$N = 10,$	$T = 100$			
Bias	0.2654	0.2745	0.2604	0.2664	0.0273	0.039	0.046
% Bias	53.0888	54.896	52.0732	53.2712	5.4538	7.8075	9.1931
RMSE	0.0731		0.0705		0.0074		0.0108
			$N = 10,$	$T = 500$			
Bias	0.2729	0.2749	0.2678	0.2668	0.0375	0.04	0.0268
% Bias	54.5878	54.9725	53.557	53.3513	7.5	7.9998	5.3538
RMSE	0.0751		0.0724		0.0064		0.0066

Bibliography

[1] Ahn, C., Lee, Y. H. and Schmidt, P. (2001): "GMM Estimation of Linear Panel Data Models with Time-varying Individual Effects," *Journal of Econometrics,* Vol. 101 Issue 2, pp. 219-255, Elsevier North Holland.

[2] Ahn, C., Lee, Y. H. and Schmidt, P. (2013): "Panel Data Models with Multiple Time-Varying Individual Effects," *Journal of Econometrics,* Vol. 174, pp. 1-14, Elsevier North Holland.

[3] Bai, J. (2009): "Panel Data Models with Interactive Fixed Effects," *Econometrica*, Vol. 77, Issue 4, pp. 1229-1279, Econometric Society, Wiley-Blackwell.

[4] Bai, J. and Ng, S. (2002): "Determining the Number of Factors in Approximate Factor Models," *Econometrica*, Vol. 70, Issue i, pp. 191-221, Econometric Society, Wiley-Blackwell.

[5] Barro, R. and Sala-i-Martin, X. (1992): "Convergence," *The Journal of Political Economy*, Vol. 100, Issue 2, pp. 223-251, Chicago University Press.

[6] Beetsma, R. and Guiliodori, M. (2010): "Fiscal Adjustment to Cyclical Developments in the OECD: an Empirical Analysis based on Real-Time Data," *Oxford Economic Papers,* Vol.62, pp. 419-441, Oxford University Press

[7] Beetsma, R. and Giuliodori, M. (2011): "The Effects of Government Purchases Shocks: Review and Estimates for the EU," *Economic Journal,* Vol. 121 Issue 550, pp. F4-F32, Royal Economic Society, Wiley-Blackwell.

[8] Bénétrix, A. S. and Lane, P. (2013): "Fiscal Cyclicality and EMU, " *Journal of International Money and Finance,* Vol. 34, pp. 164-176, Elsevier

[9] Bernanke, B., Boivin, J. and Eliasz, P. (2005): "Factor Augmented Vector Autoregression and the Analysis of Monetary Policy," *Quarterly Journal of Economics,* Vol. 120, pp. 387-422, Oxford University Press.

[10] De Groote T. and Everaert, G. (2016): "Common Correlated Effects Estimation of Dynamic Panels with Cross-Sectional Dependence," *Econometric Reviews*, Vol. 35, pp. 428-463, Taylor and Francis.

[11] De Vos, I. and Everaert, G. (2016): "Bias-corrected Common Correlated Effects Pooled estimation in homogeneous dynamic panels," *Working Papers of Faculty of Economics and Business Administration,* Ghent University.

[12] Ilzetzki, E., Mendoza, E. G. and Végh, C. A. (2013): "How Big (Small?) are Fiscal Multipliers?," *Journal of Econometrics,* Vol. 60 Issue 2, pp. 239-254, Elsevier North Holland.

[13] Islam, N. (1995): “Growth Empirics: A Panel Data Approach,” *The Quarterly Journal of Economics*, Vol. 110, No.4, pp. 1127-1170, The MIT Press.

[14] Kiviet, J. (1995): "On Bias, Inconsistency, and Efficiency of Various Estimators in Dynamic Panel Data Models," *Journal of Econometrics,* Vol. 68, pp. 53-78, Elsevier North Holland.

[15] Nerlove, M. (1971): "Further Evidence on the Estimation of Dynamic Economic Relations from a Time Series of Cross-sections," *Econometrica*, vol. 39 pp359-387, Econometric Society, Wiley-Blackwell.

[16] Newey, W. and McFadden, D. (1994): "Large Sample Estimation and Hypothesis Testing," in: Editors: Engle, R. and McFadden, D., *Handbook of Econometrics 4,* Chapter 36, pp. 2111-2245, Elsevier North Holland.

[17] Nickel, S., (1981): "Biases in Dynamic Models with Fixed Effects," *Econometrica,* Vol. 49, Issue 6, Pp. 1417-26, Econometric Society, Wiley-Blackwell.

[18] Pesaran M. H. (2006): “Estimation and Inference in Large Heterogeneous Panels with a Multifactor Error Structure,” *Econometrica,* Vol. 74, pp. 967-1012, Econometric Society, Wiley-Blackwell.

[19] Phillips, P. C.B. and Sul, D. (2003) "Dynamic Panel Estimation and Homogeneity Testing Under Cross-section Dependence," *Econometrics Journal,* Vol. 6, pp 217-259, Econometric Society, Wiley-Blackwell.

[20] Phillips, P. C.B. and Sul, D. (2007) "Bias in Dynamic Panel Estimation with Fixed Effects, Incidental Trends and Cross-section Dependence," *Journal of Econometrics,* Vol. 137 Issue 1, pp 162-188, Elsevier North Holland.

[21] Robertson, D. and Sarafidis, V. (2015): "IV estimation of panels with factor residuals," *Journal of Econometrics,* Vol. 185 Issue 2, pp 526-541, Elsevier North Holland.

[22] Sarafidis, V. and Robertson, D. (2009): "On the Impact of Cross-sectional Dependence in Short Dynamic Panel Estimation, *Econometrics Journal,* Vol. 12 #(1), pp 149-161, Elsevier North Holland.

[23] Urbain, J. P. and Westerlund, J. (2013): "On the Estimation and Inference in Factor-Augmented Panel Regressions with Correlated Loadings," *Economics Letters*, Vol. 119, Issue 3, pp 247–250, Elsevier North Holland.

Chapter 3

Quasi-Difference Estimation of Fixed N, Large T Panels with Multi-Factor Error Structures

In this chapter we develop fixed N, large T quasi-difference GMM estimators for dynamic panel data models with multiple time-varying individual effects. The estimators have a normal limiting distribution when the number of factors is correctly estimated but a mixed-normal limiting distribution when the number of factors is over-estimated. We also consider model selection and specification testing.

3.1 Introduction

In recent years the econometric literature has devoted much attention to panel data models with cross-sectional dependence in the error term. Such cross-sectional dependence is often modelled as interactive fixed effects by which 'individuals' respond with different intensities to unobserved common random processes. These time-varying individual effects generalize the time-invariant individual effects that can be found in the traditional panel data models and are known as factor structures. Economic theory gives rise to many cross-sectionally dependent phenomena. For example in studies of economic growth, e.g. Mankiw et al (1992), country data is cross-sectionally dependent as technology shocks are proliferated across the world. In tests of the Purchasing Power Parity hypothesis such as Froot and Rogoff (1995), internationally traded goods yield a dependence between countries. In empirical finance the processes that drive the returns of security portfolios are modelled as factors, see e.g. Ross (1976) and Roll and Ross (1980). Factors are also commonly used in macroeconomic applications such as business cycle modelling, cf. Sargent and Sims (1977); the analysis of monetary policy in Bernanke et al (2005); forecasting exchange rates in Engel et al (2014) and the transmission mechanism of country-wide shocks to intra-country sectors of production in

Foerster et al (2011).

The econometric literature has developed several estimators that are robust to cross-sectional dependence. Examples include the quasi-difference estimators of Holtz-Eakin et al (1988), Nauges and Thomas (2001) and Ahn, Lee and Schmidt (2001, 2013). These authors generalize the GMM estimator of Arellano and Bond (1991) for a dynamic panel data model with time-invariant individual effects to (multiple) time-varying effects within the large N, fixed T framework. Pesaran (2006) proposes an estimator for the large N, T time-varying individual effects model: the Common Correlated Effects (CCE) estimator is based on the observation that cross-sectional averages can proxy for the factors when N is large. Furthermore, the CCE estimator does not require knowledge of the number of effects and is simple to implement. These favourable properties have led to extensive analysis of the conditions under which the CCE estimator is consistent: Urbain and Westerlund (2013) and Westerlund and Urbain (2015) show that the CCE estimator is inconsistent when the number of factors is larger than the number of regressors. Moreover, Everaert and de Groote (2016) show that the CCE estimator is inconsistent when lagged dependent variables enter into the regression model and T is fixed, although the inconsistency vanishes if T is also allowed to grow large. Bai (2009) and Moon and Weidner (2015, 2017) present alternative estimators that are generalize the large N, T Principal Components estimator of Bai and Ng (2002): they obtain consistent and asymptotically normal Quasi-Maximum Likelihood estimators of the factors, loadings and slope parameters. Moon and Weidner (2015) further show that the estimator is consistent and asymptotically normal as long as the number of factors included in the model is at least as large as the true number of factors. However, these estimators are inconsistent when the idiosyncratic error is correlated in the time and cross-sectional dimensions and/or when lagged dependent variables enter the model. In either case, bias-correction methods are necessary.

Estimation techniques for panel data models with cross-sectionally dependent errors are thus typically designed for data with large N, fixed T and large N, large T dimensions, whilst fixed N, large T models have been largely ignored. This is relevant especially since improvements in the collection and quality of country-level time series have since created a surge in panel data studies by macroeconomists. For example, Beetsma and Giuliodori (2011) use the Within estimator to study fiscal multipliers in the European Monetary Union (EMU) by means of a vector autoregression (VAR). As the number of countries in EMU is small, the dimensions of their dataset are better represented by fixed N and large T. Moreover, the Within estimator is inconsistent when the data contain time-varying effects that are correlated with the regressors. Similarly, Ilzetski (2011) and Ilzetski et al (2013) compile data of approximately thirty countries and conduct VAR analysis using the Within estimator. Whilst the dimensions of this dataset are characterized by large N and T, the size of fiscal multipliers is analysed by splitting the sample along common characteristics of the countries. Again, these split samples fit better in the small N, large T framework and the Within estimator is inconsistent when the error is cross-sectionally dependent.

The need for estimation procedures that can handle cross-sectional dependence in fixed N, large T panels is growing as more and more panel datasets of high quality become available to macroeconomists. This chapter is intended to address that need: we present a GMM estimation procedure for large T dynamic panel data models with factor errors based on quasi-differencing as in Ahn, Lee and Schmidt (2001,

2013). Moreover, our estimators can accommodate individual-specific slope parameters. When the number of factors is known or estimated correctly, our theory delivers $\sqrt{T}$-consistent and asymptotically normal estimators of the slope parameters. We also study in detail the estimation problem when too many factors are assumed: we show that the quasi-difference GMM estimator of the slope parameters remains $\sqrt{T}$-consistent and develop a novel mixed normal limit theory for that case. Finally, we formulate tests for various hypotheses about the model such as the number of factors and poolability of the parameters, in addition to the usual tests for restrictions on parameters.

The paper is organized as follows: Section 2 describes the model, identification of the factor loadings and our assumptions. Section 3 describes in detail the estimation procedure, consistency and asymptotic normality of the GMM estimators when the number of factors is correctly specified. Section 4 then develops a novel mixed-normal limit theory for the case where the number of factors included in the estimation procedure is larger than the true number of factors, whereas Section 5 deals with model selection and specification testing. Section 6 presents an elaborate Monte Carlo study to evaluate the finite-sample properties of the various estimators and test procedures, whilst Section 7 concludes. Finally, on notation: ':=' denotes definitional equality. A column vector a is written in small script, whereas an $m \times n$ matrix A is in capital. The $n \times n$ identity matrix is I_n and the $n \times 1$ null-vector is 0_n. "*" denotes the *column-wise* Khatri-Rao product, "$\otimes$" is the Kronecker product, "$\oplus_{i=1}^{n} A_i$" is the direct product, i.e., $\oplus_{i=1}^{n} A_i = \text{diag}(A_1, A_2, \ldots, A_n)$and $||A||$ is the Frobenius norm of A, i.e.: $||A|| = \sqrt{\text{tr}(A'A)}$, or the corresponding Euclidean norm of a vector a. For a random matrix Ψ_t, we denote its expectation as $E(\Psi_t)$ and the sample analogue as $\hat{\Psi} := T^{-1}\sum_{t=1}^{T}\Psi_t$. Finally, the limit of some non-stochastic series is denoted " $\to$ " , " $\to_p$ " denotes convergence in probability and " $\to_d$ " convergence in distribution.

3.2 Basic Model and Assumptions

Basic Model

We will consider the following dynamic panel data model with individual specific regression coefficients and a factor error structure:

$$y_i = \mathbf{X}_i\beta_i + F\lambda_i + \varepsilon_i, \;\; i = 1, \ldots, N, \tag{3.1}$$

where $y_i = [y_{i,1}, \ldots, y_{i,T}]'$ is a T-vector of observations on the dependent variable for individual i, $\mathbf{X}_i = [Y_i, X_i]$ where $Y_i = [y_{-1,i}, \ldots, y_{-L,i}]$ is a $T \times L$ matrix whose columns correspond to the first L lags of y_i and $X_i = [x_{1,i}, \ldots, x_{K,i}]$ is a $T \times K$ matrix that corresponds to K regressors and β_i contains $K+L$ slope parameters for individual i. Furthermore, $F = [f_1, \ldots, f_R]$ is a $T \times R$ matrix containing R unobservable factors and λ_i is a vector of individual-specific factor loadings. The first column of F may be a vector of ones so that our model includes the well-known heterogeneous intercept model as a special case. Finally, ε_i is a T-vector of idiosyncratic disturbances. We are interested in estimating the slope parameters β_i for each $i = 1, \ldots, N$ and

determining the value of R. We will further assume that N is fixed and T is large so that we can only use fixed N, large T asymptotics.

Model (3.1) is quite general: for example, the model may describe N production functions, linked together by a common factor structure. This factor structure may correspond to global business cycles to which individual producers respond with different intensities. Alternatively, if we impose slope homogeneity, i.e., restrict the $\beta_i = \beta$ for all $i = 1, \ldots, N$, model (3.1) can be used to describe the k-th equation of the reduced form of a $K+1$-dimensional panel VAR with a cross-sectionally dependent error structure.

We now briefly describe the difficulty of estimating the slope parameters of (3.1) when factors are present in the error. Since the ε_i and F cannot be observed, we have the following composite error term:

$$u_i := F\lambda_i + \varepsilon_i, \ i = 1, \ldots, N$$

If the factors are static, i.e., not autocorrelated, and uncorrelated with all $x_{k,i,t}$, the slope parameters of the model can be estimated consistently and efficiently by GLS. However, since model (3.1) is dynamic, there is *a priori* no reason to assume that the factors are static. If they are not, then correlation between the $y_{i,t-1}$ and the $u_{i,t}$ will result in inconsistency of the OLS and GLS estimators for the slope parameters.[1] A similar argument applies if some of the $x_{k,i,t}$ are correlated with the factors.

It will be convenient to restate (3.1) as follows:

$$\begin{aligned} y &= (\beta * I_N)' \mathbf{X} + U, \\ U &:= \Lambda F' + \varepsilon. \end{aligned} \tag{3.2}$$

In (3.2), $y = [y_1, \ldots, y_N]'$ is an $N \times T$ matrix and $\beta = [\beta_1, \ldots, \beta_N]$ is $(K+L) \times N$. The $N(K+L) \times T$ matrix $\mathbf{X} = [Y', X']'$ collects all regressors and these are stacked first over the cross-section and then by their label, i.e. $X = \left[X^{(1)\prime}, \ldots, X^{(K)\prime}\right]'$ with k-th block $X^{(k)} = [x_{k,1}, \ldots, x_{k,N}]'$ is $N \times T$ for $k = 1, \ldots, K$ and likewise for Y. Furthermore, $\Lambda = [\lambda_1, \ldots, \lambda_N]'$ is an $N \times R$ matrix and $\varepsilon = [\varepsilon_1, \ldots, \varepsilon_N]'$ and $U = [u_1, \ldots, u_N]'$ are of dimension $N \times T$. Note that in the special case of slope homogeneity, β is a $(K+L) \times 1$ vector and model (3.1) may be written compactly as:

$$y = \left(\beta' \otimes I_N\right) \mathbf{X} + U. \tag{3.3}$$

We will assume that model (3.3) is the true model and sometimes use the subscript "0" on β, Λ and R to distinguish the true parameters from placeholders.

[1] As mentioned before, many studies nonetheless assume that the factors are not autocorrelated.

Restrictions on Λ

The Λ and F are unobservable and it is therefore not possible to identify them without imposing restrictions on the factor structure because $\Lambda F' = \Lambda C C^{-1} F'$ for any $R \times R$ invertible matrix C. This is the well-known "rotation"-problem. We require R^2 restrictions on Λ and F to identify these components, see e.g. Jöreskog and Goldberger (1975). Since N is fixed and T is large, we will estimate the Λ matrix and treat the F as unobserved regressors and the restrictions are therefore imposed on Λ only.

Many studies achieve identification by restricting the last $R \times R$ sub-matrix of the Λ to be an identity matrix, i.e., they partition $\Lambda = \left[\Lambda_+', \Lambda_-'\right]'$ and then restrict the loadings matrix as follows:

$$\Lambda^\star := \left[\Lambda_+^{\star\prime}, I_R\right]' = \Lambda\Lambda_-^{-1}, \tag{3.4}$$

where the asterisk signifies the normalized version.[2] As an example, consider applying this identification strategy to the factor loadings of a model with $R = 1$:

$$[\lambda_1^\star, \lambda_2^\star, \ldots, 1]' := \Lambda/\lambda_N = \left[\frac{\lambda_1}{\lambda_N}, \frac{\lambda_2}{\lambda_N}, \ldots, \frac{\lambda_N}{\lambda_N}\right]'. \tag{3.5}$$

As noted by Bai and Ng (2013) however, a certain structure of the loadings is assumed in identification strategy (3.4) which requires that $\det(\Lambda_-) \neq 0$. Note however that in empirical work it is possible that one or several of the cross-sectional units are unaffected by (a subset of) the factors. Clearly, if $\lambda_N = 0$ in (3.5), then the $\lambda_j^\star$ are not defined and this argument extends to the multi-factor case when Λ_- is singular.[3]

An alternative identification strategy to (3.4) is to force the loadings of each of the R factors to sum to unity. To illustrate this strategy, consider again a model where $R = 1$ and normalize the loadings so that

$$[\lambda_1^\star, \lambda_2^\star, \ldots, \lambda_N^\star]' := \Lambda/\sum_{j=1}^N \lambda_j = \left[\frac{\lambda_1}{\sum_{j=1}^N \lambda_j}, \frac{\lambda_2}{\sum_{j=1}^N \lambda_j}, \ldots, \frac{\lambda_N}{\sum_{j=1}^N \lambda_j}\right]$$

and $\lambda_N^\star = 1 - \sum_{j=1}^{N-1} \lambda_j^\star$. Note that this strategy allows $\lambda_j^\star = 0$ for any $j \in \{1, \ldots, N\}$ as long as $\sum_{j=1}^N \lambda_j^\star = 1$. When $R > 1$, normalizing each column of Λ to sum to unity delivers R restrictions and we therefore require $R(R-1)$ further restrictions. By restricting all off-diagonal elements of Λ_- to zero, we obtain the following generalization of the identification strategy for arbitrary R:

$$\Lambda_-^\star := \begin{bmatrix} 1 - \sum_{j=1}^{N-R} \lambda_{j,1}^\star & 0 & \ldots & 0 \\ 0 & \ddots & \ddots & \ldots \\ \ldots & \ddots & \ddots & 0 \\ 0 & \ldots & 0 & 1 - \sum_{j=1}^{N-R} \lambda_{j,R}^\star \end{bmatrix}, \tag{3.6}$$

[2] See for example: Robertson and Sarafidis (2015); Ahn, Lee and Schmidt (2013); PC3 in Bai and Ng (2013).

[3] In practice, a near-singular $R \times R$ unrestricted sub-matrix is problematic as well. Nonetheless, it may be possible to re-order the cross-sectional units so that Λ_- is not near-singular.

and $\Lambda_+^\star$ is further unrestricted. The restrictions in (3.6) above thus circumvent the problem of singularity of Λ_-, although this will come at the price of increased computational complexity as we will see in the next subsection.

Quasi-Difference Matrix *M*

In order to estimate the parameters of (3.1), i.e., the elements of β and $\Lambda^\star$, we will remove the factors from the composite error term by linearly combining the columns of U. That is, we apply a transformation matrix M to (3.1) such that:

$$M\Lambda = 0_{N-R,R}. \tag{3.7}$$

It should be noted that M is not uniquely defined by (3.7) and the researcher may desire to choose a transformation matrix that delivers computational convenience. When the restrictions in (3.4) are used to identify the model, it is convenient to define M as:

$$M := \left[I_{N-R}, -\Lambda_+^\star\right].$$

In the case of identification strategy (3.6) we have:

$$M := \left[\det\left(\Lambda_-^\star\right) \times I_{N-R}, -\Lambda_+ \left(\Lambda_-^\star\right)^\dagger\right],$$

where $\det\left(\Lambda_-^\star\right)$ and $\left(\Lambda_-^\star\right)^\dagger$ are the determinant and adjoint matrix of $\Lambda_-^\star$, respectively. Comparing both identification strategies, it is clear that (3.4) allows for a transformation matrix that is linear in the elements of $\Lambda^\star$, irrespective of the value of R. On the other hand, (3.6) entails non-linearity on M when $R \geq 2$.

To illustrate the different transformations, consider a model with $K = 0$, $L = 1$ and $\beta_i = \beta$ for all $i = 1, \ldots, N$ and $R = 1$ and initially without normalisation imposed on Λ:

$$y = \beta Y + \Lambda f' + \varepsilon.$$

As the identifying restriction will be imposed on Λ_-, that is, on λ_N, we define the following quasi-difference matrix:

$$M := \begin{bmatrix} \lambda_N & 0 & \ldots & 0 & -\lambda_1 \\ 0 & \lambda_N & \ddots & \ldots & -\lambda_2 \\ \ldots & \ddots & \ddots & 0 & \ldots \\ 0 & \ldots & 0 & \lambda_N & -\lambda_{N-1} \end{bmatrix}. \tag{3.8}$$

Examining the j-th row of $My = \beta MY + M\varepsilon$, for each $j = 1, \ldots N-1$, we find:

$$M_j y \quad = \quad \beta\left(\lambda_N y_{-1,j} - \lambda_j y_{-1,N}\right) + \lambda_N \varepsilon_j - \lambda_j \varepsilon_N.$$

We see that by using quasi-differencing, we have removed the unobservable factor from the model at the expense of introducing cross-sectional dependence (of a known form) in the error terms. Next, restricting λ_N to unity as per identification strategy (3.4) gives:

$$y_j - \lambda_j^\star y_N = \beta\left(y_{-1,j} - \lambda_j^\star y_{-1,N-1}\right) + \varepsilon_j - \lambda_j^\star \varepsilon_N. \tag{3.9}$$

On the other hand, identification strategy (3.6) yields:

$$\lambda_N^\star y_j - \lambda_j^\star y_N = \beta\left(\lambda_N^\star y_{-1,j} - \lambda_j^\star y_{-1,N-1}\right) + \lambda_N^\star \varepsilon_j - \lambda_j^\star \varepsilon_N, \tag{3.10}$$

with $\lambda_N^\star = 1 - \sum_{j=1}^{N-1} \lambda_j^\star$. Comparing the quasi-differenced models (3.9) and (3.10), it is clear that even in the case of $R = 1$, identification strategy (3.6) yields a more complicated structure with M as above, although both transformed models are linear in the loadings.

When $R > 1$ however, the use of (3.6) with a transformation matrix as defined in (3.8) involves polynomials of the form $\Pi_{r=1}^{R}\left(1 - \sum_{j=1}^{N-R} \lambda_{j,r}^\star\right)$ and this quickly leads to computational complications.

Finally, note that transformation matrices that conform to (3.8) are chosen for convenience but we can conceive of other transformation matrices that remove the factors from the data parametrically. For example, for the AR(1) process above with $R = 1$, consider:

$$M := \begin{bmatrix} \lambda_2 & -\lambda_1 & 0 & \ldots & \ldots & 0 \\ 0 & \lambda_3 & -\lambda_2 & 0 & \ldots & \ldots \\ \ldots & 0 & \ddots & \ddots & \ddots & \ldots \\ \ldots & \ldots & \ddots & \ddots & \ddots & 0 \\ 0 & \ldots & \ldots & 0 & \lambda_N & -\lambda_{N-1} \end{bmatrix}.$$

Compared to the above specifications, this transformation matrix is more symmetric by taking quasi-differences in neighbours. On the other hand, a generalization of this structure becomes complicated when $R > 1$ and makes the problems associated with normalization of R intractable.

Assumptions

Before we can state our assumptions, we have to introduce some additional notation: let $\mathbf{X}_t = [Y_t', X_t']$ be an $N \times (K+L)$ matrix that stacks the observations on all regressors for all N individuals at time t and let $x_{i,t}$ be a K-vector with observations on all covariates $x_{k,i,t}$ for the i-th individual at time t and $k = 1, \ldots, K$; furthermore, let y_t, u_t and ε_t be N-vectors at each t consisting of elements $y_{i,t}$, $u_{i,t}$ and $\varepsilon_{i,t}$ for $i = 1, \ldots, N$

and note that f_t is the $1 \times R$ vector of observations on the factors with representative element $f_{t,r}$; we also let $z_{i,t} = [z_{1,i,t}, \ldots, z_{S,i,t}]'$ denote an S-vector of instruments for each $i = 1, \ldots, N$ and $t = 1, \ldots, T$ and Z_t an $N \times S$ matrix with all instruments of all individuals at time t. Finally, depending on whether one considers model (3.2) or (3.3), denote the parameter vector as $\phi = \left[\text{vec}(\beta)', \text{vec}\left(\Lambda_+'\right)'\right]'$ or $\phi = \left[\beta', \text{vec}\left(\Lambda_+'\right)'\right]'$ and define the parameter space $\Phi \subset \mathbb{R}^{\dim(\phi)}$ such that $\phi \in \Phi$.

Recall that it is not necessary to remove *all* factors from U in model (3.2). That is, any static factor in u_t that is uncorrelated with the X_t can be absorbed in ε_t. Similarly, note that factors that affect X_t but are uncorrelated with u_t cannot be removed from model (3.2) by quasi-differencing but do cause the $y_{i,t}$ to be cross-sectionally dependent. Let $\mathcal{F}_t$ be the σ-field generated by all random unobservable factors. Then

$$\mathcal{F}_0 \subset \mathcal{F}_1 \subset \cdots \subset \mathcal{F}_{t-1} \subset \mathcal{F}_t$$

is the history of all logically distinct factors that affect the N equations of the model either directly or indirectly up to time t.

We impose the following assumptions on model (3.2):

ASSUMPTION 1: $\left\{\left(\text{vec}(Z_t)', \text{vec}(X_t)', \varepsilon_t', f_t\right)'\right\}$ *is strong mixing of size* $-d/(d-2)$, $d > 2$ *for all t.*

ASSUMPTION 2: (i) $E(z_{i,t}\varepsilon_{j,t} | \mathcal{F}_t, \mathcal{G}_{t-\tau}) = 0$ *for all* $i,\ j = 1, \ldots, N$, *all* $\tau > P+1 < \infty$ *and adapted* σ*-fields* $\mathcal{G}_t = \sigma(\mathcal{X}_t, \mathcal{X}_{t-1}, \ldots)$ *where* $\mathcal{X}_t := \left(\text{vec}(Z_t)', \text{vec}(X_t)', \varepsilon_t'\right)'$, $t = 1, 2, \ldots$; (ii) $E(\varepsilon_{i,t} | \mathcal{F}_t, \mathcal{G}_{t-\tau}) = 0$ *for all* $i = 1, \ldots, N$, *all* $\tau > P+1 < \infty$ *and* $t = 1, 2, \ldots$; (iii) $E\left|z_{s,i,t}\varepsilon_{j,t}\right|^{d+\delta} < \triangle < \infty$ *for some* $\delta > 0$ *and all* $s = 1, \ldots, S$, $i, j = 1, \ldots, N$ *and* $t = 1, 2, \ldots$; (iv) $E\left|z_{s,i,t}x_{k,j,t}\right|^{d+\delta} < \triangle < \infty$ *for some* $\delta > 0$ *and all* $k = 1, \ldots, K$, $s = 1, \ldots, S$, $i, j = 1, \ldots, N$ *and* $t = 1, 2, \ldots$; (v) $E\left|z_{s,i,t}f_{t,r}\right|^{d+\delta} < \triangle < \infty$ *for some* $\delta > 0$ *and all* $r = 1, \ldots, R$, $s = 1, \ldots, S$, $i = 1, \ldots, N$ *and* $t = 1, 2, \ldots$

ASSUMPTION 3: *The parameter space* Φ *is totally bounded.*

ASSUMPTION 4: *The matrices* $T^{-1}\sum_{t=1}^{T}\left[E(z_{1,t}M_1\mathbf{X}_t)', \ldots, E(z_{N-R,t}M_{N-R}\mathbf{X}_t)'\right]'$ *and* $T^{-1}\sum_{t=1}^{T}\left[E(z_{1,t}f_t)', \ldots, E(z_{N-R,t}f_t)'\right]'$ *have full column rank.*

ASSUMPTION 5: $\text{rank}\left[E(f_t'f_t) - E(f_t')E(f_t)\right] = R < N$.

Our assumptions are designed with macroeconomic applications in mind and allow for data that exhibit both dependence and heterogeneity over time. In particular, by assuming that the individual time series are strong mixing, Assumption 1 allows the data to be heterogeneously distributed across both individuals and time. Furthermore, Assumption 2(i) allows for "P-dependence" of the products of the instruments and the errors. Assumption 2(ii) requires mean independence of the errors and the factors. Note that although the sigma-fields generated by the factors f_t and those generated by the vectors $\left(\text{vec}(Z_t)', \text{vec}(X_t)', \varepsilon_t'\right)'$ have been defined separately, there will in general be correlation between the factors and (some of) the regressors. The moment requirements on products of the $f_{t,r}$, $x_{k,i,t}$, $z_{s,i,t}$ and $\varepsilon_{i,t}$ in Assumption 2(iii)-(v) are needed for consistent estimation of the optimal weight matrix used by GMM under P-dependence. Existence of $2d + \delta$-th order moments of the elements of the vectors $\left(\text{vec}(Z_t)', \text{vec}(X_t)', \varepsilon_t'\right)'$ is sufficient for

Assumption 2(iii)-(v) to hold. These moment requirements can be reduced to (i) finite fourth moments under ergodic stationarity and (ii) finite second moments by assuming both ergodic stationarity and conditional homoskedasticity. On the other hand, the P-dependence in Assumption 2(i) can be relaxed at the cost of making stronger moment assumptions, cf. White (2001, Section 6.4). Finally, note that cross-sectional dependence of the $\varepsilon_{i,t}$ is permitted under Assumptions 1 and 2.

Assumption 3 is a technical condition required for the uniform convergence of the objective function used in GMM estimation of model (3.1), cf. Andrews (1992). This assumption implies that $\|\phi\| < \infty$ for all $i = 1, \ldots, N$ which is a natural requirement for an interpretable model. Note that total boundedness is a weaker condition for uniform convergence than the standard requirement of compactness in the case of stationary or i.i.d. data, cf. Newey and McFadden (1994).[4]

Assumption 4 is similar to Assumption BA.4 in Ahn et al (2013) and consists of two identification conditions: it requires that the part of the gradient matrix corresponding to β has rank $K+L$ and that $T^{-1}\sum_{t=1}^{T}\left[E\left(z_{1,t}f_t\right)', \ldots, E\left(z_{N-R,t}f_t\right)'\right]'$ has rank R. The former condition is needed to identify β, whilst the latter requires correlation between the factors in f_t and (some of) the instruments in Z_t which allows us to identify and interpret R. Note however that Assumption 4 is restrictive and rules out certain types of exact multicollinearity. First, Assumption 4 can be violated if time-invariant regressors are included and the model contains unobservable time-invariant individual effects. As an example consider l_Γ regressors Γ that vary over individuals but not over time. Then, since $\text{rank}\left(\Lambda^\star\right) = R$ and the largest dimension of $\Lambda^\star$ is N, identification of the parameters requires that $\text{rank}\left[\Lambda^\star, \Gamma\right] \leq N$, implying that at most $l_\Gamma \leq N - R$ time-invariant regressors Γ are permissible to pin down $\Lambda^\star$ uniquely. Now partition $\Gamma = [\Gamma_1, \Gamma_2]$ so that $\text{rank}\left[\Lambda^\star, \Gamma\right] = N$ whereas the column-dimension of $\left[\Lambda^\star, \Gamma\right]$ is $l_{\Gamma_1} + l_{\Gamma_2} + R > N$ and assume that $l_{\Gamma_1} + R = N$. In that case, the matrix $\left[\Lambda^\star, \Gamma\right]$ contains l_{Γ_2} linearly dependent columns and $\text{rank}\left(M\Gamma\right) = \text{rank}\left(M\Gamma_1\right)$, thus violating the first component of Assumption 4. Of course, R is unknown and has to be estimated and it is *a priori* not clear how many-time invariant regressors are permissible in the model. As a result, we recommend to either leave them out entirely, or to add them after the number of factors is determined. Similarly, if the regressors include a variable $\bar{x}_t$ that varies over time but not over individuals and a column of Λ is a multiple of the unit vector, then any quasi-difference matrix M that removes the factor component from u_t will also remove $\bar{x}_t$ from $\mathbf{X}_t$. The result is that the first matrix in Assumption 4 is not of full rank. Note that this argument holds for both the homogeneous and heterogeneous parameter model and for that reason "observable factors" are not permitted as regressors.

Finally, Assumption 5 is an identifying assumption that requires that the true number of factors that are correlated with the instruments, R, is smaller than the number of cross-sectional units. The assumption further requires that the factors are not perfectly collinear, which is comparable to the no-collinearity assumption in estimation problems with observable regressors.

[4]Recall that a compact space is bounded and closed, but that total boundedness is stronger than boundedness. Boundedness requires a finite η-cover of the space, whereas total boundedness requires a finite number of η-covers of differing radii η. For this reason, total boundedness is also referred to as pre-compact.

3.3 Quasi-Difference GMM Estimation when $\hat{R} = R$

In this section we discuss estimation of models (3.2) and (3.3) using GMM under the assumption that the number of factors R is known. We first present the QDGMM moment conditions and objective function and then derive consistency and asymptotic normality. We also consider an alternating least squares algorithm that can be used estimate the model.

Moment Conditions and Objective Function

The GMM estimator of ϕ_0 will exploit moment conditions based on the quasi-differenced error of system (3.2):

$$M\left[Y - (\beta * I_N)'\mathbf{X}\right] = MU,$$

which, if the model is correctly specified, implies that

$$Mu_t = M\varepsilon_t,\ t = 1,\ldots,T.$$

We now introduce an S-vector of instruments $z_{j,t}$ for all $j = 1,\ldots,N-R$ quasi-differences $M_j\varepsilon_t$ for which the following moment conditions hold:

$$\begin{aligned} E\left[m_t\left(\phi_0|R\right)\right] :=& E\left[\left(z_{1,t}M_1u_t\right)',\ldots,\left(z_{N-R,t}M_{N-R}u_t\right)'\right]' \\ =& 0_{S(N-R)\times 1}, \end{aligned} \tag{3.11}$$

where the dependence on correct specification is made explicit. Note that model (3.2) involves $N(K+L)+R(N-R)$ parameters and identification requires that the following order condition is satisfied:

$$S(N-R) \geq N(K+L)+R(N-R).$$

If instead slope parameter homogeneity is imposed, identification of the parameters requires at least $(K+L)+R(N-R)$ instruments. It should be clear from the structure of (3.11) that the quasi-difference operation does not affect the time series properties of (3.11) and that instruments that valid for quasi-differencing once are also valid for quasi-differencing R times. Moreover, relevant instruments for (3.11) depend on the length of the time series, the strength of the autocorrelation in F and y and slope parameter homogeneity. This implies that valid moment conditions for the j-th quasi-difference can be generated using current and lagged values of $X_{j,t}$ and $y_{j,t-l}$ at each $t = 1,\ldots,T$. Furthermore, identification of the factor loadings requires that the matrix $T^{-1}\sum_{t=1}^{T} z_{j,t}f_t$ has full rank for all j and this implies that *all* current and lagged values of $\mathbf{X}_t$ are valid instruments for each of the $M_j\varepsilon_t$. However, since T is large, the number of instruments based on lags can diverge rapidly and must be truncated to avoid a singular covariance matrix. To limit the number of instruments, we therefore recommend using (i) current and lagged regressors $\mathbf{X}_{j,t}$ for the j-th quasi-difference

and (ii) the regressors of those individuals whose error is linearly combined through M_j as instruments.

A quasi-difference GMM estimator then solves the following minimization program:

$$\hat{\phi} = \underset{\phi \in \Phi}{\operatorname{argmin}}\, \hat{Q}(\phi|R),$$

where:

$$\hat{Q}(\phi|R) := T\left[T^{-1}\sum_{t=1}^{T} m_t(\phi|R)\right]' \hat{W}\left[T^{-1}\sum_{t=1}^{T} m_t(\phi|R)\right]$$

and $\hat{W}$ is a positive semi-definite weight matrix. We will construct an efficient two-step GMM estimator by following Newey and McFadden (1994, section 6): first, an initial consistent estimator $\tilde{\phi}$ is obtained using a conformable positive semi-definite matrix $\hat{W} = \tilde{W}$ and this estimator is referred to as the one-step QDGMM estimator. Suitable one-step weight matrices are the identity matrix or the weight matrix that is optimal when $R = 0$ and the $\varepsilon_{i,t}$ are a scalar multiple of the identity matrix. In that case the optimal weight matrix is the inverse of:

$$\oplus_{j=1}^{N-R}\left(T^{-1}\sum_{t=1}^{T} z_{j,t}z_{j,t}'\right).$$

The efficient two-step QDGMM estimator $\hat{\phi}$ is then obtained by minimizing $\hat{Q}(\phi|R)$ with $\hat{W}\left(\tilde{\phi}\right)$ computed from $\tilde{\phi}$. It is also possible to optimize $\hat{Q}(\phi|R)$ directly using $\hat{W}\left(\tilde{\phi}\right) := W(\phi)$ as in Hansen, Heaton and Yaron (1996) by continuous updating GMM. However, such a procedure would be numerically unattractive due to the non-linear nature of the objective function and for this reason we consider multi-step estimators only.

It is important to note that the moment conditions are correlated up to lag P by Assumption 4(i) and the covariance matrix of $T^{-1}\sum_{t=1}^{T} m_t(\phi_0|R)$ thus consists of a sum of P autocovariance matrices. A suitable estimator of the optimal weight matrix under this assumption is the inverse of:

$$\begin{aligned} T^{-1}\sum_{t=1}^{T}\left[m_t\left(\tilde{\phi}|R\right)m_t\left(\tilde{\phi}|R\right)'\right] \\ +T^{-1}\sum_{p=1}^{P}\sum_{t=p+1}^{T}\left[m_t\left(\tilde{\phi}|R\right)m_{t-p}\left(\tilde{\phi}|R\right)' + m_{t-p}\left(\tilde{\phi}|R\right)m_t\left(\tilde{\phi}|R\right)'\right]. \end{aligned} \tag{3.12}$$

The weight matrix (3.12) is positive semi-definite by construction for small P. Moreover, $\hat{W}\left(\tilde{\phi}\right)$ is valid for cross-sectional dependence and autocorrelation in the ε, but may be simplified if such dependence is known. If the weight matrix is not positive semi-definite it can also be estimated by the methods of White (2001, pp. 153-154) and Newey and West (1988) by introducing a weighting scheme in the second term of (3.12) at each p.

Identification, Consistency and Asymptotic Normality

We will henceforth specialize M to the identification strategy (3.4) and we note that analogous results can be obtained for strategy (3.6) at the cost of more complicated notation. Using (3.4), the moment function can be written compactly as a function of $\text{vec}(\Lambda_0^\star)$:

$$\hat{m}(\phi_0|R) := T^{-1}\sum_{t=1}^{T} m_t(\phi|R) := \hat{m}_{ZU} - \hat{G}_\Lambda \text{vec}(\Lambda_0^\star),$$

where:

$$\hat{m}_{ZU} := T^{-1}\sum_{t=1}^{T}\left[(z_{1,t}u_{1,t})', \ldots, (z_{N-R,t}u_{N-R,t})'\right]';$$
$$\hat{G}_\Lambda := T^{-1}\sum_{t=1}^{T}\left[\oplus_{j=1}^{N-R}(z_{j,t}u_{N-R+1,t}), \ldots, \oplus_{j=1}^{N-R}(z_{j,t}u_{N,t})\right],$$

and $\hat{m}_{ZU}$ and $\hat{G}_\Lambda$ are of dimensions $S(N-R)\times 1$ and $S(N-R)\times R(N-R)$ respectively. Equivalently, the moment functions can be written as a function of β_0:

$$\hat{m}(\phi_0|R) := \hat{m}_{ZY} - \hat{G}_\beta \beta_0,$$

with $S(N-R)\times 1$ vector:

$$\hat{m}_{ZY} := T^{-1}\sum_{t=1}^{T}\left[(z_{1,t}M_1 y_t)', \ldots, (z_{N-R,t}M_{N-R}y_t)'\right]'$$

and in the case of the homogeneous parameter model (3),

$$\hat{G}_\beta := T^{-1}\sum_{t=1}^{T}\left[(z_{1,t}M_1\mathbf{X}_t)', \ldots, (z_{N-R,t}M_{N-R}\mathbf{X}_t)'\right]',$$

of dimension $S(N-R)\times(K+L)$ and the appropriate definition of $\hat{G}_\beta$ for the heterogeneous parameter model is contained in Lemma 1 in the Appendix. Under identification strategy (3.4) we thus see that the moment functions permit a dual representation:

$$\hat{m}_t(\phi_0|R) := \hat{m}_{ZY} - \hat{G}_\beta\beta_0 = \hat{m}_{ZU} - \hat{G}_\Lambda \text{vec}(\Lambda_0^\star) \tag{3.13}$$

which implies that the objective function $\hat{Q}(\phi|R)$ is bi-convex in ϕ.[5] Now letting:

$$G := -E\left(T^{-1}\sum_{t=1}^{T}\frac{\partial m_t(\phi|R)}{\partial \phi'}\bigg|_{\phi=\phi_0}\right) := E\left[\hat{G}_\beta, \hat{G}_\Lambda\right],$$

local identification of ϕ follows from the dual representation above as the following proposition makes precise:

PROPOSITION 3.1: Local Identification of QDGMM: *Suppose Assumptions 1-5 hold and normalize* $\Lambda_0^\star = \left[\Lambda_{0+}^{\star\prime}, I_R\right]'$. *Then*

$$\begin{aligned} E\left[m_t\left(\beta, \Lambda_{0+}^\star|R\right)\right] = 0 \quad iff \quad \beta = \beta_0 \quad and \\ E\left[m_t\left(\beta_0, \Lambda|R\right)\right] = 0 \quad iff \quad \text{vec}(\Lambda) = \text{vec}\left(\Lambda_{0+}^\star\right) \end{aligned}$$

if G has full rank.

Proof: See Appendix.

REMARK 3.1: The identification condition holds only when the number of factors estimated $\hat{R} = R$. When $\hat{R} < R$, the GMM estimator cannot remove the factors completely. As a result, the corresponding objective function $Q\left(\phi|\hat{R}\right)$ diverges to infinity. When $\hat{R} > R$, the system is under-determined and infinitely many solutions to the moment conditions exist, implying that $E\left(\hat{G}_\Lambda\right)$ cannot be full rank. In this situation, a consistent estimate of β_0 is however still available.

Given identification, we need measurability of the moment function with respect to the data $\mathbf{X}_t$ and $\hat{W}$ to be positive semi-definite for consistency:

PROPOSITION 3.2. Consistency of QDGMM Estimators when $\hat{R} = R$: *Suppose Assumptions 1-5 hold. Assume further that the weight matrix* $\hat{W} \to_p W$ *is positive semi-definite. Then* $\hat{\phi} \to_p \phi_0$ *as* $T \to \infty$.

Proof: See Appendix.

REMARK 3.2: The proof follows Andrews (1992) by showing that a stochastic Lipschitz condition holds on the moment conditions, which implies stochastic equicontinuity. Stochastic equicontinuity then takes pointwise weak convergence of the objective function $\hat{Q}(\phi|R)$ to uniform weak convergence. The result of Andrews (1992) covers weaker assumptions than techniques which rely on a dominance condition on the moment function such as in Theorem 2.6 of Newey and McFadden (1994), whereas the stated result also

[5] For a set $C \subset A \times B$, a function $\xi(a,b) : C \to \mathbb{R}$ is bi-convex if ξ is convex in a holding constant b and *vice versa*, see Gorski et al (2007).

covers i.i.d. and/or strictly stationarity processes.

Asymptotic normality of QDGMM estimators readily follows from standard results in the M-estimator literature:

PROPOSITION 3.3. Asymptotic Normality of QDGMM Estimators when $\hat{R} = R$: *Let Assumptions 1-5, let* $E\left[\sup_{\phi\in\Phi}||\sum_{t=1}^{T}\frac{\partial m_t(\phi|R)}{\partial\phi'}||\right] < \infty$, *G have full rank,* $\hat{W} \to_p W$ *and* $\frac{1}{\sqrt{T}}\sum_{t=1}^{T} m_t(\phi_0|R) \to_d N\left(0_{S(N-R)\times 1}, V\right)$ *for a positive definite matrix V. Then:*

1.
$$\sqrt{T}\left(\hat{\phi}-\phi_0\right) \to_d N\left(\mathbf{0}, \left(G'WG\right)^{-1}G'WVWG\left(G'WG\right)^{-1}\right),$$

where

$$V := E\left[T^{-1}\sum_{p=-P}^{P}\sum_{t=|p|+1}^{T} m_t(\phi_0|R)\, m_{t-p}(\phi_0|R)'\right].$$

2. *Furthermore, if* $W = V^{-1}$, *then:*

$$\hat{Q}\left(\hat{\phi}|R\right) \to_d \chi^2\left\{(S-R)(N-R) - \dim(\beta)\right\}.$$

Proof: See Appendix.

REMARK 3.3: As is typical for GMM estimators, setting $W = V^{-1}$ yields the efficient QDGMM estimator by standard arguments. The variance estimator V, corresponds to the cross-product matrix of the quasi-differenced residuals $Mu_t = M\varepsilon_t$ and the instrument matrix Z_t over the lag window $p = 1,\ldots,P$. When the same instruments are used for all equations so that z_t is an S-vector and $P = 1$ however, the covariance matrix can be simplified to $M\Omega M' \otimes E(z_t z_t')$, where $\Omega = \text{var}(\varepsilon_t)$. These matrices can be consistently estimated by sample moments.

The moment conditions (3.13) are non-linear in parameters and numerical optimization is required to find argmin $\hat{Q}(\phi|R)$ over ϕ. We will exploit the bi-convexity of $\hat{Q}(\phi|R)$ to solve for the QDGMM estimator. That is, the dual representation of the moment conditions implies that the quadratic form $\hat{Q}(\phi|R)$ is strictly convex in β, holding fixed vec(Λ) and *vice versa*. This property can be used to compute estimates of β_0 and vec$(\Lambda_0^\star)$ respectively as the following set of conditional least squares solutions of argmin $\hat{Q}(\phi|R)$:

$$\hat{\beta}|\Lambda = \left[\left(\hat{G}_\beta|\Lambda\right)'\hat{W}\left(\hat{G}_\beta|\Lambda\right)\right]^{-1}\left(\hat{G}_\beta|\Lambda\right)'\hat{W}\left(\hat{m}_{ZY}|\Lambda\right), \tag{3.14}$$

$$\text{vec}\left(\hat{\Lambda}\right)|\beta = \left[\left(\hat{G}_\Lambda|\beta\right)'\hat{W}\left(\hat{G}_\Lambda|\beta\right)\right]^{-1}\left(\hat{G}_\Lambda|\beta\right)'\hat{W}\left(\hat{m}_{ZU}|\beta\right) \tag{3.15}$$

for any conformable weight matrix $\hat{W}$. The conditional strict convexity of the solutions (3.14) and (3.15) to $\hat{Q}(\phi|R)$ motivate the use of an alternating least squares (ALS) algorithm to solve the optimization problem.[6] Such an algorithm is attractive because it is simple to implement and fast compared to more elaborate optimization routines. An ALS algorithm for the homogeneous parameter model using $\tilde{W}$ in the first step is:

ALGORITHM 3.1. Alternating Least Squares Computation of QDGMM Estimator:

1. For $b = 1, \ldots, B$

   (a) If $b = 1$ set $\hat{W} = \tilde{W}$; if $b \geq 2$ compute $\hat{W}$ from formula (3.12) using $\hat{U}^{b-1} = Y - \left(\hat{\beta}^{b-1} \otimes I_N\right)' \mathbf{X}$.

   (b) For $c = 1, \ldots, C$

      i. draw random numbers to initialize the ALS algorithm: for example, $\hat{\beta}^0 \sim U(\varsigma_2 - \varsigma_1)$ in a user-specified interval $[\varsigma_1, \varsigma_2]$;

      ii. using $\hat{\beta}^0$, compute vec $\left(\hat{\Lambda}^1\right)$ according to equation (3.15); then, using vec $\left(\hat{\Lambda}^1\right)$, compute $\hat{\beta}^1$ according to equation (3.14);

      iii. continue step (ii) D times and collect $\hat{\phi}^c := \left[\hat{\beta}^{D\prime}, \text{vec}\left(\hat{\Lambda}^D\right)'\right]'$ and the corresponding $\hat{Q}^c$.

   (c) Set $\hat{Q}^b = \underset{c}{\text{argmin}}\left(\hat{Q}^c\right)$ and $\hat{\phi}^b$ corresponding to $\hat{Q}^b$.

2. Repeat step 1 B times.

REMARK 3.4: The ALS algorithm requires choosing the number of steps of the GMM estimator, B; the number of initializations of the algorithm, C; the distributional parameters of the initial conditions, i.e., ς_1, ς_2 and finally the maximum number of iterations D of each of the initializations C. We recommend setting both C and D large to allow for the exploration of the (local) minima of $\hat{Q}(\phi|R)$. As an alternative to using D large one can also specify a convergence criterion such as $\left|\hat{\phi}_d - \hat{\phi}_{d-1}\right| / \dim\left(\hat{\phi}\right) < \eta$ for η small, which can speed up the algorithm considerably. To initialize the algorithm, we recommend exploiting the properties of the parameter space: for example, to estimate an AR(1) model, we would restrict the starting value of the parameter β to be within the unit circle in accordance with the mixing of Assumption 1.

REMARK 3.5: It is important to note that solutions found by iterating Algorithm 3.1 do not necessarily correspond to a global minimum, as can be seen from the Hessian of $\hat{Q}\left(\hat{\phi}|R\right)$:

$$\left.\frac{\partial^2 \hat{Q}(\phi|R)}{\partial\phi'\partial\phi}\right|_{\phi=\hat{\phi}} = 2 \times \begin{bmatrix} \hat{G}_\beta' \hat{W} \hat{G}_\beta & \hat{G}_\beta' \hat{W} \hat{G}_\Lambda + \xi(\hat{m})' \\ \hat{G}_\Lambda' \hat{W} \hat{G}_\beta + \xi(\hat{m}) & \hat{G}_\Lambda' \hat{W} \hat{G}_\Lambda \end{bmatrix}, \tag{3.16}$$

[6] Defining the solutions as in (3.14) and (3.15) above is more insightful than the solution method of Ahn, Lee and Schmidt (2013), who would use the Rayleigh coefficient to solve for $\Lambda^\star$. The Rayleigh coefficient formulation has no tractable form and thus masks the fact that the $\hat{\Lambda}$ are estimated parameters, which is clearly inconvenient. In contrast, solution (3.15) can be used directly to obtain point and variance estimates of the $\hat{\Lambda}^\star$, which will be useful in what follows.

where $\xi(\hat{m})$ is a conformable matrix whose columns consist of partial derivatives of $\hat{G}$ that are proportional to the moment functions. The matrices on the block-diagonal are clearly positive semi-definite (PSD) if $\hat{G}$ is full rank and $\hat{W}$ is PSD for any choice of ϕ. However, as can be seen from the Schur complement of (3.16), the Hessian matrix can be negative semi-definite depending on the terms proportional to $\hat{m}$ on the off-diagonal blocks, implying that the objective function is not globally convex in ϕ. On the other hand, by Propositions 3.1 and 3.2, if $\hat{\phi} \to_p \phi_0$ and thus $\hat{m} \to_p 0$, the Schur-complement of the Hessian is proportional to:

$$\left(\hat{G}_\Lambda' \hat{W}^{1/2} M_{\hat{W}^{1/2}\hat{G}_\beta} \hat{W}^{1/2} \hat{G}_\Lambda\right),$$

where $M_{\hat{W}^{1/2}\hat{G}_\beta}$ is the orthogonal projection on the space of $\hat{W}^{1/2}\hat{G}_\beta$. This matrix is clearly PSD since $M_{(\cdot)}$ is idempotent and $\hat{G}$ is full rank by Proposition 3.1, so that the *global* minimum indeed occurs at ϕ_0. Moreover, bi-convexity of $\hat{Q}(\phi|R)$ and total boundedness of Φ guarantee that all sequences of iterates $\hat{\phi}^d = \left[\hat{\beta}^{d\prime}, \text{vec}\left(\hat{\Lambda}^d\right)'\right]'$ will converge to a local critical point (cf. Grippo and Sciandrone (2000), Propositions 4, 5). The results of Grippo and Sciandrone (2000) imply that repetition of the algorithm D times with D large should in practice yield a reasonable approximation to the global minimizer of $\hat{Q}(\phi|R)$ as long as the resulting approximation has a Hessian matrix that is PSD. If concerns about the solution exist, we can further follow Andrews (1997) and use the J-statistic to determine if the estimated minimum is close to the global minimum. That is, according to Proposition 3.3, $\hat{Q}\left(\hat{\phi}|R\right)$ is chi-squared distributed in a $\sqrt{T}$-neighbourhood of ϕ_0 and this rules out very large values of the objective function.

REMARK 3.7 Hayakawa (2016) claims that the QDGMM estimator has identification problems and that the algorithm proposed by Ahn et al (2013) converges to critical points not necessary equal to the global minimum. We believe this criticism is unsubstantiated: first, GMM objective functions are in general not convex except in some special cases, see for example Hayashi (2000, p. 468). As a result, this criticism is not exclusive to the QDGMM estimator above nor Algorithm 3.1 or that of Ahn et al (2013). Moreover, in light of the analysis of the Hessian in Remark 3.4, the problem can be avoided entirely if the definiteness of the Hessian is evaluated at all critical points found by the QDGMM estimator. Practically, this can be done by computing the condition number of the Hessian at each critical point, i.e., the ratio of the largest to the smallest eigenvalue, and consider only critical points corresponding to relatively small condition numbers as candidates for the global minimizer.[7] Indeed, applying this analysis to the examples presented in Hayakawa (2016) reveals that the incorrect critical points correspond to the Hessian having (near-) zero eigenvalues, whereas the condition number of the correct critical points is several orders of magnitude smaller than that of the inflection points, thus separating inflection points and minima very clearly.

[7] "Relatively small" can be qualified by the researcher based on the observed Hessian.

3.4 Quasi-Difference GMM Estimation when $\hat{R} > R$

In the previous section we have assumed that R is known and this allowed us to derive a limit theory for QDGMM in the standard non-linear GMM framework. Of course, in practice R is unknown and must be replaced by an estimate $\hat{R}$. We have argued in Remark 3.1 that a consistent estimator of ϕ does not exist when $\hat{R} < R$, although a consistent estimator of β_0 does exist when $\hat{R} > R$. The objective of this section is therefore to extend the results of Section 3 to the case of $\hat{R} > R$.[8]

When $\hat{R} > R$, the Rank-Nullity theorem implies that the matrix $M_{\hat{R}>R}$ no longer constitutes a basis for the (left) nullity of Λ_0. As a result the QDGMM estimator is not identified and an infinite number of $\left(N-\hat{R}\right) \times N$ quasi-difference matrices $M_{\hat{R}>R}$ exist which belong to the set:

$$\mathcal{M} = \left\{ M_{\hat{R}>R} | M_{\hat{R}>R}\Lambda_0 = 0 \wedge \text{rank}\left(M_{\hat{R}>R}\right) = N-\hat{R} \right\}.$$

However, as the following intermediate result makes precise, β_0 can still be consistently estimated using a QDGMM estimator $\hat{\beta}$ for any given $M_{\hat{R}>R} \in \mathcal{M}$:

Proposition 3.4: Consistency of QDGMM with $\hat{R} > R$ and fixed $M_{\hat{R}>R} \in \mathcal{M}$: *Suppose Assumptions 1-5 hold. Assume further that some $M_{\hat{R}>R} \in \mathcal{M}$ is known and the weight matrix $\hat{W} \to_p W$ is symmetric positive semi-definite. Then $\hat{\beta} \to_p \beta_0$ as $T \to \infty$.*

Proof: Given $M_{\hat{R}>R} \in \mathcal{M}$, identification requires $E\left(\hat{G}_\beta\right)$ being full rank. Proposition 3.2 then immediately follows by repeating the steps of the proof with M replaced by some fixed $M_{\hat{R}>R} \in \mathcal{M}$. □

Proposition 3.4 covers the case of fixed $M_{\hat{R}>R}$, but not when $\Lambda_0^\star$ is estimated along with β_0 and $\hat{R} > R$. The proposition does however imply that a consistent estimate of G_Λ is available, although the structure of G_Λ depends on the chosen identification strategy. We will analyse the QDGMM estimator which uses the 'identifying restrictions' $\underline{\Lambda}^\star = \left[\underline{\Lambda}_+^{\star\prime}, I_{\hat{R}>R}\right]'$, where $\underline{\Lambda}_+^\star$ is of dimension $\left(N-\hat{R}\right) \times \hat{R}$ and the underscore signifies $\hat{R} > R$. Similarly partitioning the true loadings as $\underline{\Lambda}_0 = \left[\underline{\Lambda}_{0+}', \underline{\Lambda}_{0-}'\right]'$ so that $\underline{\Lambda}_{0+}$ is of dimension $\left(N-\hat{R}\right) \times R$, we observe that incorrect restrictions lead to the estimation of $\left(\hat{R}-R\right)\left(N-\hat{R}\right)$ excess loadings. In other words, over-estimating R can be thought of as padding $\underline{\Lambda}_{0+}$ with a zero matrix of the same dimension. The identification problem then manifests in the gradient of the moment function corresponding to vec $\left(\underline{\Lambda}_+^\star\right)$:

$$\begin{aligned}
\hat{G}_\Lambda :=& T^{-1}\sum_{t=1}^{T}\left[\oplus_{j=1}^{N-\hat{R}}\left(z_{j,t}u_{N-\hat{R}+1,t}\right), \ldots, \oplus_{j=1}^{N-\hat{R}}\left(z_{j,t}u_{N,t}\right)\right] \\
=& T^{-1}\sum_{t=1}^{T}\left[\oplus_{j=1}^{N-\hat{R}}\left(z_{j,t}f_t\lambda_{N-\hat{R}+1}\right), \ldots, \oplus_{j=1}^{N-\hat{R}}\left(z_{j,t}f_t\lambda_N\right)\right] + O_p\left(T^{-1/2}\right).
\end{aligned}$$

[8] Analogous results for the large N, fixed T model of Ahn, Lee and Schmidt (2013) can be obtained straightforwardly.

Note that $\hat{G}_\Lambda$ is now linearly dependent with $\text{rank}(G_\Lambda) = R\left(N-\hat{R}\right)$ since $\underline{\Lambda}_{0-}$ is of dimension $\hat{R}\times R$ and thus $\underline{\Lambda}_{0-}'v = 0$ for some $v \neq 0$. As a result the matrix $G'WG$ is singular and Proposition 3.1 cannot hold.

To circumvent the linear dependence of $\hat{G}_\Lambda$ we define an $\hat{R}\left(N-\hat{R}\right)\times\hat{R}\left(N-\hat{R}\right)$ matrix H such that (i) $\hat{G}_H := \hat{G}_\Lambda H$ and $G_H := E\left(\hat{G}_H\right)$ have full rank and (ii) the resulting estimator of $\text{vec}(\underline{\Lambda}_H) := H^{-1}\text{vec}(\underline{\Lambda}^\star)$ satisfies $\text{span}(\underline{\Lambda}_H) = \text{span}(\underline{\Lambda}_{0+})$. To this end we re-define the moment functions as:

$$\begin{aligned}\hat{m}\left(\underline{\phi},\hat{R}\right) =& \hat{m}_{ZU} - \hat{G}_\Lambda H H^{-1}\text{vec}(\underline{\Lambda}^\star)\\ :=& \hat{m}_{ZU} - \hat{G}_H\text{vec}(\underline{\Lambda}_H),\end{aligned} \tag{3.17}$$

where $\underline{\phi} := \left[\underline{\beta}', \text{vec}(\underline{\Lambda}^\star)'\right]'$ is of dimension $\dim(\beta) + \hat{R}\left(N-\hat{R}\right)$. Under these conditions the factors can be removed from U whilst the design matrix of the QDGMM estimator can be inverted. This implies that $E\left[\hat{m}\left(\underline{\phi},\hat{R}\right)\right] = 0$ and the QDGMM estimator is defined. It should be clear that many choices of H are available since $\underline{\Lambda}^\star$ is not identified and we will henceforth construct H based on the identifying restrictions (3.4) applied to the bottom $R\times R$ block of $\underline{\Lambda}_{0-}$. In particular, we normalize $\underline{\Lambda}^\star_{0-} = \left[\Lambda^{\star\prime}_{0\times}, I_R\right]'$ and use $\Lambda^\star_{0\times}$ to denote the $\left(\hat{R}-R\right)\times R$ partition of the (true) loadings matrix which is used to quasi-difference too often. Imposing this normalization then allows us to focus on scaling matrices $H := H\left(\Lambda^\star_{0\times}\right)$ with the following structure:

$$H := \begin{bmatrix} \sqrt{T}I_{\hat{R}-R} & 0_{(\hat{R}-R)\times R} \\ -\sqrt{T}\Lambda^{\star\prime}_{0\times} & I_R \end{bmatrix} \otimes I_{N-\hat{R}} := [H_1, H_2]\otimes I_{N-\hat{R}},$$

where H_1 and H_2 are of dimension $\hat{R}\times\left(\hat{R}-R\right)$ and $\hat{R}\times R$ respectively.[9] Although we cannot observe either Λ_0 or $\underline{\Lambda}_{0-}$, such a normalization is possible regardless because the matrix H is immaterial for the moment condition (3.17).

Post-multiplying the gradient $\hat{G}_\Lambda$ by H and taking limits yields:

$$\begin{aligned}G_H =& \Big[\oplus_{j=1}^{N-\hat{R}}\psi_{N-\hat{R}+1,j},\ldots,\oplus_{j=1}^{N-\hat{R}}\psi_{N-R,j},\\ &\qquad \oplus_{j=1}^{N-\hat{R}}E\left(z_{j,t}f_t\lambda_{N-R+1}\right),\ldots,\oplus_{j=1}^{N-\hat{R}}E\left(z_{j,t}f_t\lambda_N\right)\Big]\\ :=& [G_\times, G_-],\end{aligned}$$

where $G_\times$ has dimensions $S\left(N-\hat{R}\right)\times\left(\hat{R}-R\right)\left(N-\hat{R}\right)$, G_- has dimensions $S\left(N-\hat{R}\right)\times R\left(N-\hat{R}\right)$ and our specification of H has normalized the collinear columns of G_Λ to be in $G_\times$. Moreover, H has quasi-differenced and scaled the linearly dependent columns of G_Λ and instead formed random S-vectors $\psi_{r,j}$ for each $r = N-\hat{R}+1,\ldots,N-R$ and $j = 1,\ldots,N-\hat{R}$. Letting $H_{1,r}$ be the r-th column of H_1 and $\underline{\varepsilon}_t$ the last

[9] Note that this matrix is unique up to a sign change in H_1 under identifying restrictions (3.4).

$\hat{R} \times 1$ sub-vector of ε_t, the random vectors are defined as:

$$\psi_{r,j} = T^{-1/2} \sum_{t=1}^{T} z_{j,t} H_{1,r}' \underline{\varepsilon}_t \to_d N \left[0_S, \text{var} \left(T^{-1/2} \sum_{t=1}^{T} z_{j,t} H_{1,r}' \underline{\varepsilon}_t \right) \right], \tag{3.18}$$

by the CLT for heterogeneous dependent variables, cf. White (2001, page 130). Since normally distributed random variables are zero with probability equal to zero, this implies that the matrix $G_\times$ has full rank almost surely. By contrast, G_- is unaffected by H and corresponds to the linearly independent columns of the original matrix G, e.g., the $R \times \left(N - \hat{R}\right)$ true factor loadings.

Now letting $\mathcal{H} = \text{diag}\left[I_{\dim(\beta)}, H\right]$, the re-scaled gradient of the full parameter vector

$$G_{\mathcal{H}} := E\left(\hat{G}\mathcal{H}\right)$$

thus has full rank almost surely because the dependent columns of G_Λ are replaced with a matrix consisting of normally distributed random variables. We are now in a position to analyse the QDGMM estimator when $\hat{R} > R$ and we denote the scaled estimator by $\underline{\hat{\phi}}_H = \left[\underline{\hat{\beta}}_H', \text{vec}\left(\underline{\hat{\Lambda}}_H\right)'\right]'$. Of course, $\underline{\hat{\phi}}_H$ is infeasible because we cannot observe H. The infeasible estimator is however useful in analysing the feasible unscaled estimator $\underline{\hat{\phi}} = \left[\underline{\hat{\beta}}', \text{vec}\left(\underline{\hat{\Lambda}}\right)'\right]'$. We further partition the $\left(N - \hat{R}\right) \times \hat{R}$ matrix of estimated factor loadings *without* rescaling as $\underline{\hat{\Lambda}} = \left[\underline{\hat{\Lambda}}_\times, \underline{\hat{\Lambda}}_+\right]$ where the excess loadings $\underline{\hat{\Lambda}}_\times$ are of dimension $\left(N - \hat{R}\right) \times \left(\hat{R} - R\right)$ and $\underline{\hat{\Lambda}}_H$ is partitioned conformably. Finally, the true loadings under normalization (3.4) are $\underline{\Lambda}_0^\star = \left[\underline{\Lambda}_{0+}^{\star\prime}, \Lambda_{0\times}^{\star\prime}, I_R\right]'$ and we extend the dimension of the (true) parameter space by padding $\underline{\Lambda}_{0+}^\star$ with an $\left(N - \hat{R}\right) \times \left(\hat{R} - R\right)$ matrix of zeros on the left to match the dimension of the estimators of $\underline{\hat{\Lambda}}$ and $\underline{\hat{\Lambda}}_H$. As a result the model parameters are extended to $\underline{\phi}_0 = \left[\beta_0', 0', \text{vec}\left(\underline{\Lambda}_{0+}^\star\right)'\right]'$. The following presents a limit theory for QDGMM when $\hat{R} > R$:

THEOREM 3.1: Mixed Normal Limit Theory of QDGMM when $\hat{R} > R$: *Let Assumptions 1-5 hold and assume that $\hat{W} \to_p W$ is positive definite. Then conditionally on $H\left(\Lambda_{0\times}^\star\right)$:*

1. T*he infeasible QDGMM estimator $\underline{\hat{\phi}}_H$ is consistent:*

$$\underline{\hat{\phi}}_H := \begin{bmatrix} \underline{\hat{\beta}}_H \\ \text{vec}\left(\underline{\hat{\Lambda}}_{H\times}\right) \\ \text{vec}\left(\underline{\hat{\Lambda}}_{H+}\right) \end{bmatrix} \to_p \begin{bmatrix} \beta_0 \\ 0 \\ \text{vec}\left(\underline{\Lambda}_{0+}^\star\right) \end{bmatrix}$$

and mixed normal:

$$\sqrt{T}\left(\underline{\hat{\phi}}_H - \underline{\phi}_0\right) \to MN\left[0, \left(G_{\mathcal{H}}' W G_{\mathcal{H}}\right)^{-1} G_{\mathcal{H}}' W V_R W G_{\mathcal{H}} \left(G_{\mathcal{H}}' W G_{\mathcal{H}}\right)^{-1}\right],$$

where:

$$V_R := \text{var}\left[T^{-1/2}\sum_{t=1}^{T} m_t\left(\underline{\phi}_0|R\right)\right].$$

2. *The feasible QDGMM estimator of the slope parameter is consistent:* $\hat{\underline{\beta}} \to_p \beta_0$*;* $\text{vec}\left(\hat{\underline{\Lambda}}\right)$ *is inconsistent and degenerate:*

$$\begin{bmatrix} \text{vec}\left(\hat{\underline{\Lambda}}_{\times}\right) \\ \text{vec}\left(\hat{\underline{\Lambda}}_{+}\right) \end{bmatrix} \to_d \begin{bmatrix} 0 \\ \text{vec}\left(\underline{\Lambda}_{0+}\right) \end{bmatrix} + \begin{bmatrix} I_{\hat{R}-R} \otimes I_{N-\hat{R}} \\ -\Lambda_{0\times}^{\star\prime} \otimes I_{N-\hat{R}} \end{bmatrix} \Psi_{\Lambda},$$

where Ψ_{Λ} *is defined in the Appendix. The joint distribution of* $\hat{\underline{\phi}}$ *is singular mixed normal:*

$$\text{diag}\left[\sqrt{T} I_{\dim(\beta)}, I_{\hat{R}\left(N-\hat{R}\right)}\right] \times \hat{\underline{\phi}} \to_d MN\left(\underline{\phi}_0, \mathscr{G} V_R \mathscr{G}'\right),$$

where:

$$\mathscr{G} := \begin{bmatrix} \left(G_{\beta}' W^{1/2} M_{W^{1/2} G_{\mathscr{H}\setminus\beta}} W^{1/2} G_{\beta}\right)^{-1} G_{\beta}' W^{1/2} M_{W^{1/2} G_{\mathscr{H}\setminus\beta}} W^{1/2} \\ H_{\Lambda_{\times},1}\left(G_{\times}' W^{1/2} M_{W^{1/2} G_{\mathscr{H}\setminus\Lambda_{\times}}} W^{1/2} G_{\times}\right)^{-1} G_{\times}' W^{1/2} M_{W^{1/2} G_{\mathscr{H}\setminus\Lambda_{\times}}} W^{1/2} \end{bmatrix};$$

V_R *is as above;* $M_{W^{1/2} G_{\mathscr{H}}\setminus(\cdot)}$ *is the orthogonal projection on the columns of* $W^{1/2} G_{\mathscr{H}}$ *excluding the* $(\cdot)$*-th block and* $H_{\Lambda_{\times},1} = T^{-1/2} H_1 \otimes I_{N-\hat{R}}$ *is a fixed matrix of dimension* $\hat{R}\left(N-\hat{R}\right) \times \left(\hat{R}-R\right)\left(N-\hat{R}\right)$.

3. *The J-statistic with* $\hat{W} = \hat{V}_R^{-1}$ *satisfies:*

$$\hat{Q}\left(\hat{\underline{\phi}}|\hat{R}\right) \to_d \chi^2\left\{\left(S-\hat{R}\right)\left(N-\hat{R}\right) - \dim\left(\beta_0\right)\right\}.$$

Proof: See Appendix.

REMARK 4.1: Theorem 4.1 states that $\hat{\underline{\phi}}_H$ is $\sqrt{T}-$consistent even when $\hat{R} > R$. Of course $\hat{\underline{\phi}}_H$ is less efficient than the QDGMM estimator with $\hat{R} = R$ because it estimates an excess of $\left(\hat{R}-R\right)\left(N-\hat{R}\right)$ zero loadings and this leads to mixed normality. Moreover, since H is unobservable, $\hat{\underline{\phi}}_H$ is infeasible and the researcher estimates $\hat{\underline{\phi}}$ instead when $\hat{R} > R$. In that case the parameters of the model display mixing rates of convergence and even though $\hat{\underline{\beta}}$ remains $\sqrt{T}$-consistent, the $\hat{\underline{\Lambda}}$ converge slower than $\hat{\underline{\beta}}$. Crucially, the $\hat{\underline{\Lambda}}$ now converge in distribution to a singular combination of the true loadings $\underline{\Lambda}_{0+}$ and a mixed normal random vector Ψ_{Λ}. The lack of identification of $\hat{\underline{\Lambda}}_{\times}$ thus implies that the joint distribution of $\hat{\underline{\phi}}$ is singular mixed normal because the matrix $H_{\Lambda_{\times},1}$ is of larger row dimension than $G_{\times}$.

REMARK 4.2: It is also interesting to note that the distribution of the QDGMM estimator with $\hat{R} > R$ depends on moment conditions based on only R quasi-differences as if we had started the estimation problem with a dataset consisting of only $N - \left(\hat{R}+R\right)$ dependent variables. Lemma 6 in the Appendix however

shows that $\hat{V}$ is consistent for V_R, even in the case of estimation with $\hat{R} > R$. The combination of Lemma 5 and 6 in the Appendix then implies that the J-statistic has the usual chi-squared distribution when $\hat{R} > R$ if the optimal weight matrix is used. Furthermore, inference about $\underline{\hat{\beta}}$ when $\hat{R} > R$ is feasible even if we over-estimate R because the conditional distributions of $\sqrt{T}\left(\underline{\hat{\beta}} - \beta_0\right)$ and $\sqrt{T}\left(\underline{\hat{\beta}}_H - \beta_0\right)$ are equivalent in the limit, provided that the optimal weight matrix is used.

REMARK 4.3: Theorem 3.1 is related to results of Caner (2008) for nearly-singular GMM estimation: in that paper the covariance matrix of the moment conditions is assumed to be singular and therefore cannot be inverted. Caner assumes that a rescaled version of V can be inverted and obtains asymptotic normality, albeit at a reduced rate. By contrast, in our paper, the non-invertibility of the design matrix stems from linear dependence in the columns of the gradient. Our results are closer to the partially identified simultaneous equation model of Phillips (1985) and especially collinear regression in Phillips (2016). In that latter paper, the distribution of IV estimators based on asymptotically collinear regressors is derived and shown to exhibit diverging rates and mixed normality. Theorem 3.1 can thus be seen as a generalization of the results of Phillips (2016) to panel GMM with unobservable regressors.

Example

We now make the implications of the Theorem 3.1 more concrete by presenting an example with $R = 1$ but $\hat{R} = 2$. The gradient is $G_{\mathcal{H}} := \left[G_\beta, G_\times, G_-\right]$, where the $S(N-2) \times (N-2)$ partitions corresponding to the loadings are defined as:

$$\begin{aligned} G_\times &:= \left[\oplus_{j=1}^{N-2} \psi_{N-1,j}\right]; \\ G_- &:= \left[\oplus_{j=1}^{N-2} E\left(z_{j,t} f_t\right)\right], \end{aligned}$$

where the $\psi_{N-1,j}$ are as in equation (3.18) above and $G_{\mathcal{H}}$ is thus of full rank. Since $\underline{\hat{\beta}}_H$ is consistent by Proposition 4, the point estimates of $\text{vec}\left(\underline{\hat{\Lambda}}_H\right)$ are:

$$\begin{aligned} \begin{bmatrix} \text{vec}\left(\underline{\hat{\Lambda}}_{H\times}\right) \\ \text{vec}\left(\underline{\hat{\Lambda}}_{H+}\right) \end{bmatrix} &= \begin{bmatrix} \left(\hat{G}_\times' \hat{W}^{1/2} M_{\hat{W}^{1/2}\hat{G}_{\mathcal{H}\setminus\Lambda_\times}} \hat{W}^{1/2} \hat{G}_\times\right)^{-1} & 0 \\ 0 & \left(\hat{G}_-' \hat{W}^{1/2} M_{\hat{W}^{1/2}\hat{G}_{\mathcal{H}\setminus\Lambda_-}} \hat{W}^{1/2} \hat{G}_-\right)^{-1} \end{bmatrix} \times \\ &\quad \begin{bmatrix} \hat{G}_\times' \hat{W}^{1/2} M_{\hat{W}^{1/2}\hat{G}_{\mathcal{H}\setminus\Lambda_\times}} \hat{W}^{1/2} \hat{G}_- \text{vec}\left(\underline{\Lambda}_{0+}^\star\right) \\ \hat{G}_-' \hat{W}^{1/2} M_{\hat{W}^{1/2}\hat{G}_{\mathcal{H}\setminus\Lambda_-}} \hat{W}^{1/2} \hat{G}_- \text{vec}\left(\underline{\Lambda}_{0+}^\star\right) \end{bmatrix} + O_p\left(T^{-1/2}\right) \\ &\to_p \begin{bmatrix} 0 \\ \text{vec}\left(\underline{\Lambda}_{0+}^\star\right) \end{bmatrix}, \end{aligned} \tag{3.19}$$

where $M_{\hat{W}^{1/2}\hat{G}_{\mathcal{H}\setminus\Lambda_\times}}$ is the orthogonal projection on all columns of $G_{\mathcal{H}}$ except those in $G_\times$ and similarly for $M_{\hat{W}^{1/2}\hat{G}_{\mathcal{H}\setminus\Lambda_-}}$. Equation (3.19) makes explicit that scaling the loadings with an appropriate matrix H^{-1}

ensures that $\underline{\hat{\Lambda}}_{H\times}$ converge to a zero vector of length $(N-R)$ and that the $\underline{\hat{\Lambda}}_{H+}$ converge to the first $(N-2)$ elements of $\underline{\Lambda}_0^\star = \left[\underline{\Lambda}_{0+}^{\star\prime}, \Lambda_{0\times}^{\star'}, I_R\right]'$.

In practice we can only form $\underline{\hat{\phi}}$ and we now analyse the singular distribution of the unscaled loadings when $\hat{R}=2$ and $R=1$. Pre-multiplying $\underline{\hat{\Lambda}}_H$ by H we find the limit of $\underline{\hat{\Lambda}}$ as:

$$\begin{bmatrix} \text{vec}\left(\underline{\hat{\Lambda}}_\times\right) \\ \text{vec}\left(\underline{\hat{\Lambda}}_+ - \underline{\Lambda}_{0+}^\star\right) \end{bmatrix} \to_d \left(\begin{bmatrix} 1 \\ -\lambda_{N-1} \end{bmatrix} \otimes I_{N-2} \right) \Psi_\Lambda.$$

The $\underline{\hat{\Lambda}}$ thus converge in distribution to a singular combination of a random vector Ψ_Λ and the $\underline{\Lambda}_{0+}^\star$ without appropriate scaling. Furthermore, using the optimal weight matrix in the proof of Theorem 3.1 shows that Ψ_Λ is a covariance mixture of normals:

$$\Psi_\Lambda \sim MN\left[0, \left(G_\times' V_1^{-1/2} M_{V_1^{-1/2} G_{\mathcal{H}\setminus\Lambda_\times}} V_1^{-1/2} G_\times\right)^{-1}\right].$$

Note that $M_{(\cdot)}$ in the above is a constant matrix and mixed normality comes solely from the occurrence of the matrix $G_\times$ on the left and right sides of the variance formula. However, since both the moment condition $T^{1/2}\sum_{t=1}^T m_t\left(\underline{\hat{\phi}}, R\right)$ and $G_\times$ are quasi-differenced using the last R individuals, there is correlation between these random variables and inference based on the wald principle is not valid for $\underline{\hat{\Lambda}}$.

On the other hand, we know from Proposition 3.4 that $\underline{\hat{\beta}}$ is consistent if $M_{\hat{R}>R} \in \mathcal{M}$ when $\hat{R} > R$. The latter condition is easily seen to hold by evaluating the quasi-difference operation on $\Lambda_0^\star$:

$$M_{\hat{R}>R}\Lambda_0^\star = \left[I_{N-2}, -\Psi_\Lambda, -\left(\underline{\Lambda}_{0+}^\star - \lambda_{N-1}^\star \Psi_\Lambda\right)\right]\Lambda_0^\star = 0,$$

with the normalization $\lambda_N^\star = 1$. Furthermore, $\underline{\hat{\beta}}$ is actually conditionally normally distributed:

$$\sqrt{T}\left(\underline{\hat{\beta}} - \beta_0\right) \to_d N\left[0, \left(G_\beta' V_1^{-1/2} M_{V_1^{-1/2} G_{\mathcal{H}\setminus\beta}} V_1^{-1/2} G_\beta\right)^{-1}\right].$$

That is, $G_\times$ occurs only in the annihilator $M_{(\cdot)}$ and idempotent matrices have eigenvalues strictly equal to zero and unity. As a result, the randomness of $G_\times$ does not affect affect the conditional distribution of $\underline{\hat{\beta}}$ despite correlation with $T^{1/2}\sum_{t=1}^T m_t\left(\underline{\hat{\phi}}, R\right)$. Moreover, the distribution of $\underline{\hat{\beta}}$ is equivalent to that of $\underline{\hat{\beta}}_H$ because H only operates on the blocks of the gradient corresponding to the factor loadings. These latter arguments imply that standard Wald and t-statistics constructed from $\underline{\hat{\beta}}$ are valid even when $\hat{R} > R$.

3.5 Inferential Procedures for the Quasi-Difference GMM Estimator

The following section provides inferential procedures about the QDGMM estimator. These involve testing basic (non-) linear hypotheses about the estimated parameters $\hat{\phi}$ and testing poolability of the model

parameters. After a brief discussion of standard hypothesis testing, we then move on to the question of estimating the correct number of factors R. We will henceforth use R_0 to denote the true number of factors and use R as a placeholder.

When R_0 is known, the QDGMM estimator is a standard non-linear GMM problem and simple linear hypotheses about $\hat{\phi}$ can be tested by the t-ratio. Under $H_0 : \hat{\phi}_j = \overline{\phi}_j$, using a hypothetical parameter value $\overline{\phi}_j$ for some $j = 1, \ldots, \dim\left(\hat{\phi}\right)$, the t-ratio is defined as:

$$\frac{\sqrt{T}\left(\hat{\phi}_j - \overline{\phi}_j\right)}{\sqrt{\left(\hat{G}'\hat{V}^{-1}\hat{G}\right)^{-1}_{j,j}}} \to_d N(0,1).$$

The proof is by standard GMM theory and follows from Proposition 3.3 applied to an efficient QDGMM estimator with $\hat{W} = \hat{V}^{-1}$. Similarly, (non-) linear q-composite hypotheses

$$H_0 : \ a\left(\hat{\phi}\right) = 0,$$

can be tested using the Wald principle and we assume that $\partial a\left(\hat{\phi}\right)/\partial\hat{\phi}' = A$ where A is a full rank matrix of dimension $q \times \dim\left(\hat{\phi}\right)$ with rank equal to the number of hypotheses. Under the null-hypothesis, the Wald-statistic:

$$Ta\left(\hat{\phi}\right)'\left[A\left(\hat{G}'\hat{V}^{-1}\hat{G}\right)^{-1}A'\right]^{-1}a\left(\hat{\phi}\right) \to_d \chi^2(q)$$

and diverges under the alternative. The Wald-test is again standard in GMM theory and the distributional result under the null follows from the Continuous Mapping Theorem applied to the transformed random variable $a\left(\hat{\phi}\right)$. Theorem 3.1 further implies that inference about $\underline{\hat{\beta}}$ is valid when $R > R_0$ and we make this precise for the Wald-statistic in the following proposition:

Proposition 3.5: Inference about $\underline{\hat{\beta}}$ when $R > R_0$: *Let Assumptions 1-5 hold, let $R > R_0$ and assume:*

$$H_0 : \ a\left(\underline{\hat{\phi}}\right) = \mathcal{A}\left(\underline{\hat{\beta}}\right) = 0$$

holds. Then:

$$T\mathcal{A}\left(\underline{\hat{\beta}}\right)'\left[A\left(\hat{G}_\beta'\hat{V}^{-1/2}M_{\hat{V}^{-1/2}\hat{G}_\Lambda}\hat{V}^{-1/2}\hat{G}_\beta\right)^{-1}A'\right]^{-1}\mathcal{A}\left(\underline{\hat{\beta}}\right) \to_d \chi^2(q)$$

as $T \to \infty$.

Proof: See Appendix.

REMARK 5.1: Valid inference about $\underline{\hat{\beta}}$ when $R > R_0$ is facilitated by the fact that the conditional distributions of $\underline{\hat{\beta}}_H$ and $\underline{\hat{\beta}}$ are equivalent under the null hypothesis and this implies that inference on $\underline{\hat{\beta}}$ can commence as if H were known. This result is reassuring because the researcher is typically interested in

β_0 alone. On the other hand, when R_0 is known, the QDGMM estimator is a standard non-linear GMM estimator and equivalent Lagrange Multiplier and Likelihood Ratio tests will exist under the null-hypothesis $a\left(\hat{\phi}\right)$, cf. Proposition 7.11 in Hayashi (2000).

REMARK 5.2: The above tests are asymptotic and small sample performance may be improved through a bootstrap methodology as in Bun (2004) and Giersbergen and Kiviet (2002). It is not difficult to see that a parametric bootstrap, using estimates of the quasi-differenced residuals $MU = M\varepsilon$, is feasible as long as any temporal and/or cross-sectional dependence in ε is known, so that appropriate blocking can be implemented.

From an empirical point of view, poolability tests are particularly important in simultaneous equation models because if they hold true, they allow for simplification of the empirical model and thus a reaping of efficiency rewards (Zellner, 1962). Two relevant poolability considerations for QDGMM can be framed as Wald-tests: first, slope homogeneity of the β_0 can be tested with the following null hypothesis:

$$H_0 : A\hat{\beta} = 0,$$

where

$$A = \begin{bmatrix} I_{K+L} & -I_{K+L} & 0 & \dots & \dots & 0 \\ 0 & I_{K+L} & -I_{K+L} & 0 & \ddots & \dots \\ \dots & 0 & \ddots & \ddots & \ddots & \dots \\ \dots & \ddots & \ddots & \ddots & \ddots & 0 \\ 0 & \dots & \dots & 0 & I_{K+L} & -I_{K+L} \end{bmatrix}.$$

This leads to the following Wald-test:

$$T\left(A\hat{\beta}\right)'\left[A\left(\hat{G}_{\beta}'\hat{V}^{-1/2}M_{\hat{V}^{-1/2}\hat{G}_{\Lambda}}\hat{V}^{-1/2}\hat{G}_{\beta}\right)^{-1}A'\right]^{-1}A\hat{\beta} \rightarrow_d \chi^2\left\{\dim\left(\hat{\beta}\right)(N-1)\right\}.$$

Similarly, the question of whether the model is generated by a Time Effects model can also be answered by means of the following null-hypothesis:

$$H_0 : A\text{vec}\left(\hat{\Lambda}\right) = 0$$

with $(N-2)\times(N-1)$ restrictions matrix:

$$A=\begin{bmatrix} 1 & -1 & 0 & \dots & \dots & 0 \\ 0 & 1 & -1 & 0 & \ddots & \dots \\ \dots & 0 & \ddots & \ddots & \ddots & \dots \\ \dots & \ddots & \ddots & \ddots & \ddots & 0 \\ 0 & \dots & \dots & 0 & 1 & -1 \end{bmatrix}.$$

Under the null, there is only one factor and $\lambda_i=\lambda$ for all $i=1,\dots,N$, so that the loading parameter is fixed at unity and although the QDGMM estimator is consistent, more efficient estimators available for the Time-Effects model. The corresponding Wald-test for equality of the loadings vector is:

$$T\left[A\text{vec}\left(\hat{\Lambda}\right)\right]'\left[A\left(\hat{G}_{\Lambda}'\hat{V}^{-1/2}M_{\hat{V}^{-1/2}\hat{G}_{\beta}}\hat{V}^{-1/2}\hat{G}_{\Lambda}\right)^{-1}A'\right]^{-1}\left[A\text{vec}\left(\hat{\Lambda}\right)\right]\rightarrow_d\chi^2\left(N-2\right),$$

Although it is clearly not necessary to estimate the model with more than $R_0=1$ factors included to test this hypothesis, the case of $R>1$ is estimable and asymptotically chi-squared distributed under the null as a special case:

PROPOSITION 3.6: Poolability of the Factor Loadings when $R>1$: *Let Assumptions 1-5 hold, let $R>R_0$ and let*

$$H_0:\ A\text{vec}\left(\underline{\hat{\Lambda}}\right)=0,$$

hold, where:

$$A=\mathbf{1}_{1\times R}\otimes\begin{bmatrix} 1 & -1 & 0 & \dots & \dots & 0 \\ 0 & 1 & -1 & 0 & \ddots & \dots \\ \dots & 0 & \ddots & \ddots & \ddots & \dots \\ \dots & \ddots & \ddots & \ddots & \ddots & 0 \\ 0 & \dots & \dots & 0 & 1 & -1 \end{bmatrix}$$

is of dimension $(N-R-1)\times R(N-R-1)$. Then the Wald-test satisfies:

$$T\left[A\text{vec}\left(\hat{\Lambda}\right)\right]'\left[A\left(\hat{G}_{\Lambda}'\hat{V}^{-1/2}M_{\hat{V}^{-1/2}\hat{G}_{\beta}}\hat{V}^{-1/2}\hat{G}_{\Lambda}\right)^{-1}A'\right]^{-1}\left[A\text{vec}\left(\hat{\Lambda}\right)\right]\rightarrow_d\chi^2\left(N-R-1\right)$$

as $T\rightarrow\infty$.

Proof: See Appendix.

An important condition for asymptotic normality of $\hat{\phi}$ is that the number of factors R_0 in models (3.2) and

(3.3) is estimated correctly. In what follows, we will provide a treatment of the estimation of R_0 following Ahn, Lee and Schmidt (2013) and generalize their results through a series of remarks. Their methodology revolves around the properties of the QDGMM objective function $\hat{Q}(\phi|R)$, which is the J-Statistic of Hansen (1982). Using Proposition 3.3 and Theorem 3.1 the asymptotic behaviour of $\hat{Q}(\phi|R)$ is summarized as follows:

1. $\hat{Q}\left(\hat{\phi}|\hat{R}\right) \underset{p}{\to} \infty$ if $\hat{R} < R_0$;
2. $\hat{Q}\left(\hat{\phi}|\hat{R}\right) \underset{d}{\to} \chi^2\left\{\left(S-\hat{R}\right)\left(N-\hat{R}\right) - \dim(\beta)\right\}$ if $\hat{R} \geq R_0$.

In the above enumeration, item (1) is due to under-identification which makes it impossible to entirely remove the factors from U. As a result, $Q\left(\hat{\phi},|\hat{R}\right)$ diverges to infinity for any $\hat{\phi}$. On the other hand, item (2) is the well-known result of Hansen (1982), which states that an efficient GMM objective function of a correctly specified model converges to a chi-squared distribution, extended to the case where $\hat{R} > R_0$ by Theorem 3.1.

These characteristics of $\hat{Q}\left(\hat{\phi},|\hat{R}\right)$ suggest that we may estimate R_0 by sequentially evaluating the J-Test at $R = 0, 1, 2, ...R_{\max}$, where $R_{\max} \leq N$.[10] By (2), the number of factors is then estimated at the first $\hat{R}$ where $Q\left(\phi|\hat{R}\right) \leq c\left(\chi^2, \hat{R}, \alpha_T\right)$ and $c(\cdot)$ is the critical value associated with a chi-squared random variable with degrees of freedom equal to $\left(S-\hat{R}\right)\left(N-\hat{R}\right) - \dim(\beta)$ at the significance level α_T. However, since the test is sequential, the significance level α_T must vary with the sample size to account for type-I errors. The following proposition yields weak consistency for the sequential J-Test if certain conditions on the significance level are met:

PROPOSITION 3.7. Consistency of the Sequential *J*-Test: *Under Assumptions 1-5, the sequential J-Test yields* $\hat{R} \to_p R_0$ *if* (1) *the significance level* $\alpha_T \to 0$ *and* (2) $-T^{-1}\log(\alpha_T) \to 0$ *as* $T \to \infty$.

Proof: See Appendix.

REMARK 5.3: Proposition 3.7 is based on an asymptotic result due to Pötscher (1983). This result links the significance level α_T to the convergence rate of a chi-squared variable $Q(\phi|R)$ and this is necessary to bound the probability of $\hat{R} > R$ at zero. Note however that the proposition is silent about the functional form of α_T and this implies that different choices of α_T may result in different $\hat{R}$ in finite samples.

REMARK 5.4: It should be noted that instead of estimating $\hat{R}$ only, the test also depends on the specification of the slope parameter model and the relevance of included instruments $z_{j,t}$ for all t and $j = 1, \ldots, N-\hat{R}$. To make this precise, let $\mathcal{L}$ be an index set of all possible (mixed) lag order specifications ranging from zero to $L_{\max}$; Let $\mathcal{Z}$ be a set containing all distinct combinations of candidate instruments and let $\mathcal{R} := 0, 1, \ldots R_{\max}$ be the set of possible R we are willing to test. Finally, define by $\mathcal{T} := \{\mathcal{L} \cup \mathcal{R} \cup \mathcal{Z}\}_{\neq}$ the set that contains

[10] We set $M = I_N$ when $R = 0$, resulting in a pooled GMM procedure without quasi-differencing.

the union of all logically distinct elements of the individual components. Then, using the definition of $\mathcal{T}$, an easy corollary to Proposition 3.7 follows from repeating the proof with $\hat{R}$ replaced by $\hat{\mathcal{T}}$. As a result, testing only $\hat{R}$ is conditional on the others being correctly specified. On the other hand, treating the specification aspects simultaneously through $\mathcal{T}$ makes the test difficult to interpret.

We can also estimate $\hat{R}$ using information criteria based on the weak consistency result in Theorem 3.1 of Cragg and Donald (1997). To this end define a family of information criteria as:

$$IC_T(R) := Q(\phi|R) + \pi(R) \times \theta(T), \tag{3.20}$$

where $\pi(R)$ and $\theta(T)$ are functions of the estimated number of factors and T respectively. The interpretation of $IC_T(R)$ is the usual one: the reduction in $Q(\phi|R)$ from adding additional parameters is compensated by increases in the penalty function $\pi(R) \times \theta(T)$. As a result, model parsimony is preferred unless the sum of the aforementioned effects is negative. Application of these criteria is computationally straightforward: for $R = 0, 1, \ldots, R_{\max}$, compute the QDGMM estimators and their corresponding criteria and estimate $\hat{R}$ corresponding to $\operatorname{argmin} IC_T(R)$ over R. The family of possible $IC_T(R)$ is large, although several familiar functional forms are available for the problem at hand. For example:

- BIC: $\theta(T) = \log(T)$ and $\pi(R) = -\triangle \times \{(S-R)(N-R) - \dim(\beta)\}$
- HIC: $\theta(T) = \log\log(T)$ and $\pi(R) = -\triangle \times \{(S-R)(N-R) - \dim(\beta)\}$

and $\triangle$ is a user-specified constant. Note that we have specified $\pi(R)$ to depend on the degrees of freedom of the GMM estimator rather than just the number of estimated parameters. This is common when information criteria are applied to GMM problems because the objective function $Q(\phi|\hat{R})$ is a chi-squared random variable with mean equal to the degrees of freedom under correct specification of the model. This form of $\pi(R)$ is inconsequential when the number of instruments is fixed when searching over $\mathcal{R}$, but not when the instrument set depends on the number of factors: for example, if the number of instruments of each equation grows with R, then it may be that certain sequences of $\pi(R)$ have (i) extrema and/or (ii) duplicates over the search space $\mathcal{R}$. Such information criteria will always favour less parsimonious models. These considerations lead to weak consistency of $IC_T(\hat{R})$ under certain conditions:

PROPOSITION 3.8. Consistency of $IC_T(\hat{R})$: *Let* (1) $\theta(T) \to \infty$, (2) $T^{-1}\theta(T) \to 0$ *as* $T \to \infty$ *and* (3) $\partial\pi(R)/\partial R > 0$ *for all* $R = 1, \ldots, R_{\max} \in \mathcal{R}$, *then under Assumptions 1-5* $\hat{R} = \underset{R}{\operatorname{argmin}}\, IC_T(R) \to_p R_0$.

Proof: See Appendix.

REMARK 5.5: Both remarks 5.3 and 5.4 apply equally to Proposition 3.8. Specifically, Proposition 3.8 holds for any choice of $\triangle$ and this may result in different finite sample estimates $\hat{R}$. Similarly, an easy corollary obtains for the problem of searching over $\mathcal{T}$ instead of just $\mathcal{R}$. It is also interesting to note that

the consistency proof techniques are very similar for both estimation procedures, with the correspondence between the significance level and the critical value of the chi-squared distribution now played by $\theta(T)$.

REMARK 5.6: In contrast to Proposition 3.7, determining R_0 through $IC_T\left(\hat{R}\right)$ does not require an efficient weight matrix in $Q(\phi|R)$ and information criteria can be applied to one-step estimators as was observed by Ahn, Lee and Schmidt (2013) and stated as a Theorem in Sarafidis and Robertson (2015). This offers flexibility to information criteria over test procedures that depend on the chi-squared approximation. Furthermore, it is possible that estimation of R_0 is actually easier with one-step estimators, because differences in $\hat{Q}(\phi|R)$ as a function of R are much more pronounced in the region $R < R_0$.

REMARK 5.7: Note that Akaike's information criterion (AIC) is not a member of the family of information criteria defined in Proposition 3.8. That is, in AIC, $\theta(T) := \theta = 2$ does not depend on T and therefore does not satisfy condition (1) of the proposition.

We can also cast estimation of R_0 in the rank-testing framework of Al-Sadoon (2017) and construct a test based the gradient of the factor loadings. This is convenient because if $\hat{\beta} \to_p \beta_0$,

$$\hat{G}_\Lambda = T^{-1}\sum_{t=1}^{T}\left[\oplus_{j=1}^{N-R^\star}\left(z_{j,t}f_t\lambda_{N-R^\star+1}\right),\ldots,\oplus_{j=1}^{N-R^\star}\left(z_{j,t}f_t\lambda_N\right)\right]+O_p\left(T^{-1/2}\right)$$

is consistent for any $R^\star \geq R_0$ and does not directly depend on the identification status of the $\underline{\Lambda}_{0+}$. However, $\hat{G}_\Lambda$ consists of $R^\star$ block diagonal matrices and we would like to reduce this sparsity. We will reduce the dimension of $\hat{G}_\Lambda$ by averaging over the instruments of each quasi-difference equation:

$$B_1\hat{G}_\Lambda B_2 := T^{-1}\sum_{t=1}^{T}\left[\bar{z}_t\hat{u}_{N-R^\star+1,t},\ldots,\overline{Z}_t\hat{u}_{N,t}\right],$$

where:

$$\begin{aligned}B_1 &= 1'_{N-R^\star\times 1}\otimes I_S,\\ B_2 &= I_{R^\star}\otimes 1_{(N-R^\star)\times 1}\left(N-R^\star\right)^{-1}\end{aligned}$$

and $\bar{z}_t$ is the cross-sectional average over each $z_{j,t}$ for $j = 1,\ldots,N-R^\star$. Now note that:

$$B_1\hat{G}_\Lambda B_2 = T^{-1}\sum_{t=1}^{T}\left(\bar{z}_t f_t\Lambda'_-\right)+O_p\left(T^{-1/2}\right)$$

is of dimension $S\times R^\star$ because Λ_- is of dimension $R^\star\times R_0$. Estimating R_0 is now equivalent to testing for

any R:

$$H_0 : \text{rank}\left(B_1\hat{G}_\Lambda B_2\right) = R \text{ against } H_1 : \text{rank}\left(B_1\hat{G}_\Lambda B_2\right) > R.$$

That is, although the column dimension of $B_1\hat{G}_\Lambda B_2$ is $R^\star$, the number of non-zero singular values under H_0 is R whilst the remaining $R^\star - R$ singular values tend to zero. This hypothesis can be verified by constructing $(S-R)\times S$ left and $R^\star \times (R^\star - R)$ right null space estimators $\hat{\Xi}_1^{(R)}$ and $\hat{\Xi}_2^{(R)}$ such that:

$$\hat{\Xi}_1^{(R)}\left(B_1\hat{G}_\Lambda B_2\right)\hat{\Xi}_2^{(R)} \to_p \Xi_1^{(R)}\left(B_1 G_\Lambda B_2\right)\Xi_2^{(R)} = 0_{(S-R)\times(R^\star-R)}. \tag{3.21}$$

Pre- and post-multiplication by the null space estimators thus amounts to demeaning $\hat{G}_\Lambda$ by annihilating the factors and Al-Sadoon (2017) shows that the null space estimators can be obtained from various matrix decompositions such as the SVD, LU and spectral decompositions applied to $(B_1 G_\Lambda B_2)$. Furthermore, since the components of $B_1\hat{G}_\Lambda B_2$ consist of sample averages, we can appeal to a suitable CLT to construct a test statistic based on the vectorization of (3.21). Before we can construct a test statistic based on $B_1\hat{G}_\Lambda B_2$ however, note that $\hat{G}_\Lambda$ is calculated using a preliminary estimator of β_0 and we must make a correction. To see this let $u_{-,t}$ be the last $R^\star$ elements of the N-vector u_t and let $\mathbf{X}_{-,t}$ correspond to the bottom $R^\star \times (K+L)$ sub-matrix of $\mathbf{X}_t$, then:

$$B_1\hat{G}_\Lambda B_2 = T^{-1}\sum_{t=1}^{T}\overline{z}_t u'_{-,t} - T^{-1}\sum_{t=1}^{T}\overline{z}_t\left(\hat{\beta}-\beta_0\right)'\mathbf{X}'_{-,t}.$$

The second term above corresponds to the QDGMM sampling error of $\hat{\beta}$ and must be taken into account in any test statistic based on $\sqrt{T}B_1\hat{G}_\Lambda B_2$. Proposition 3.3 and Theorem 3.1 imply that we can estimate the sampling error of $\hat{\beta}$ consistently and a correction can be made on the variance of $\sqrt{T}\text{vec}\left(B_1\hat{G}_\Lambda B_2\right)$ or directly on $B_1\hat{G}_\Lambda B_2$. We will use the latter so that for each $t = 1,\ldots,T$:

$$\begin{aligned}\hat{G}^\star_{\Lambda,t} = &\left[\oplus_{j=1}^{N-R^\star} z_{j,t}\hat{u}_{N-R^\star+1,t},\ldots,\oplus_{j=1}^{N-R^\star} z_{j,t}\hat{u}_{N,t}\right] + \\ &\left[\oplus_{j=1}^{N-R^\star} z_{j,t}\left(\hat{\beta}-\beta_0\right)'\mathbf{X}'_{N-R^\star+1,t},\ldots,\oplus_{j=1}^{N-R^\star} z_{j,t}\left(\hat{\beta}-\beta_0\right)'\mathbf{X}'_{N,t}\right]\end{aligned}$$

and thus $\hat{G}^\star_\Lambda = T^{-1}\sum_{t=1}^{T}\hat{G}^\star_{\Lambda,t}$, where we note that $\hat{G}^\star_\Lambda = G_\Lambda + O_p\left(T^{-1/2}\right)$. Now re-defining the null space estimators to $B_1\hat{G}^\star_\Lambda B_2$, we will determine R_0 using the random variable:

$$\sqrt{T}\text{vec}\left(\hat{\Xi}_1^{(R)}B_1\hat{G}^\star_\Lambda B_2\hat{\Xi}_2^{(R)}\right) = \sqrt{T}\left(\hat{\Xi}_2^{(R)\prime}\otimes\hat{\Xi}_1^{(R)}\right)\text{vec}\left(B_1\hat{G}^\star_\Lambda B_2\right). \tag{3.22}$$

To construct a test statistic based on (3.22) we will also need an estimate of the variance:

$$\hat{V}_{\hat{G}^{\star}_{\Lambda}} = T^{-1}\sum_{t=1}^{T}\left[\text{vec}\left(B_1\hat{G}^{\star}_{\Lambda,t}B_2\right)\text{vec}\left(B_1\hat{G}^{\star}_{\Lambda,t}B_2\right)'\right] + T^{-1}\sum_{p=1}^{P}\sum_{t=p+1}^{T}\left[\text{vec}\left(B_1\hat{G}^{\star}_{\Lambda,t}B_2\right)\text{vec}\left(B_1\hat{G}^{\star}_{\Lambda,t-p}B_2\right)' + \text{vec}\left(B_1\hat{G}^{\star}_{\Lambda,t-p}B_2\right)\text{vec}\left(B_1\hat{G}^{\star}_{\Lambda,t}B_2\right)'\right]. \quad (3.23)$$

The proof of the following proposition shows that (3.23) is consistent for the variance of the random variable $\sqrt{T}\text{vec}\left(B_1\hat{G}^{\star}_{\Lambda}B_2\right)$.

Proposition 3.9. Rank Test of $H_0 : \text{rank}\left(B_1\hat{G}^{\star}_{\Lambda}B_2\right) = R_0$ **against** $H_1 : \text{rank}\left(B_1\hat{G}^{\star}_{\Lambda}B_2\right) > R_0$**:** *Let Assumptions 1-5 hold and let* $\hat{\beta}_0 \to_p \beta_0$*. Then under* H_0*:*

1. $$\sqrt{T}\left(\hat{\Xi}_2^{(R)\prime}\otimes\hat{\Xi}_1^{(R)}\right)\text{vec}\left(B_1\hat{G}^{\star}_{\Lambda}B_2\right) \to_d N\left[0,\left(\Xi_2^{(R)\prime}\otimes\Xi_1^{(R)}\right)V_{G^{\star}_{\Lambda}}\left(\Xi_2^{(R)}\otimes\Xi_1^{(R)\prime}\right)\right],$$

   where $V_{G^{\star}_{\Lambda}}$ *is the limit of* (3.23)*.*

2. *The Rank Test*

$$RK_{R_0} := T\left[\left(\hat{\Xi}_2^{(R)\prime}\otimes\hat{\Xi}_1^{(R)}\right)\text{vec}\left(B_1\hat{G}^{\star}_{\Lambda}B_2\right)\right]'\left[\left(\hat{\Xi}_2^{(R)\prime}\otimes\hat{\Xi}_1^{(R)}\right)\hat{V}_{G^{\star}_{\Lambda}}\left(\hat{\Xi}_2^{(R)}\otimes\hat{\Xi}_1^{(R)\prime}\right)\right]^{-1}\times \left(\hat{\Xi}_2^{(R)\prime}\otimes\hat{\Xi}_1^{(R)}\right)\text{vec}\left(B_1\hat{G}^{\star}_{\Lambda}B_2\right) \to_d \chi^2\left\{(S-R_0)(R^{\star}-R_0)\right\}.$$

   and diverges under the alternative.

Proof: See Appendix.

The Rank Test thus constructed is a Wald-test and Proposition 3.9 shows that it approaches the chi-squared distribution when $\text{rank}\left(B_1\hat{G}^{\star}_{\Lambda}B_2\right) = R_0$ and diverges when it is under-estimated. To operationalize a procedure that can be applied in practice however, we have to develop a sequential version of the Rank Test:

PROPOSITION 3.10. Consistency of the Sequential Rank Test: *Let Assumptions 1-5 hold. Then the sequential Rank Test is weakly consistent if* (1) $\hat{\beta} \to_p \beta_0$ *can be calculated using some* $R^{\star} \geq R_0$*,* (2) *the significance level* $\alpha_T \to 0$ *and* (3) $-T^{-1}\log(\alpha_T) \to 0$ *as* $T \to \infty$*.*

Proof: Immediate from Proposition 3.9 and the proof of Proposition 3.7.

REMARK 5.8: The consistency of the sequential Rank Test derives from the same mechanic as the sequential *J*-Test but requires an extra assumption. As a result, the test will loose power when the approximation of $\hat{\beta}$ is poor because the sampling error inflates both the variance and the numerator. This consideration is also the reason to bias-correct the numerator instead of the denominator of the test: inflating the denominator

would result in a further loss of power if the sampling error is large in any finite sample realization of the test.

REMARK 5.9: One can construct the null space estimators using any appropriate method in the literature and these include but are not limited to: SVD as in Kleibergen and Paap (2006); LU in Cragg and Donald (1996) and by finding the nearest rank R matrix by norm minimization as in Cragg and Donald (1997). We have experimented with several of these and found little qualitative differences other than computation time. For that reason we will apply the SVD in what follows.

REMARK 5.10: Several algorithms can be conceived to estimate R_0 based on Propositions 3.9 and 3.10. For example, we can form G_Λ with $R^\star = 1$ and then test $R = 0$, if the test rejects we calculate G_Λ with $R^\star = 2$ and test $R = 1$. We then estimate $\hat{R}$ as the first time the test cannot reject at significance level α_T. We can also estimate G_Λ with $R^\star$ close to N and then testing up as before holding $R^\star$ constant. In all cases however, it is advisable to use $R^\star$ large enough and hope to estimate β_0 precisely: after an estimate of the sampling error of β_0 is obtained, one can always re-compute the gradient using a smaller $R^\star$.

3.6 Monte Carlo Experiments

We will now examine the small sample properties of the one-step, efficient two-step QDGMM estimators and the test procedures of Section 5 applied to a simple AR(1) model with factor error residuals. We first compare the performance of the QDGMM estimators in terms of bias and RMSE with several popular estimators in the empirical macroeconomics literature, namely the Pooled OLS estimator, the Fixed Effects (FE) estimator and the Correlated Common Effects (CCE) estimator of Pesaran (2006). The finite sample behaviour of these estimators further illustrates the degree of bias in the point estimates when the data contains factor residuals. We now briefly describe the alternative estimators. The Pooled OLS estimator is defined as:

$$\hat{\beta}_{OLS} = \left(\sum_{i=1}^{N} \mathbf{X}_i' \mathbf{X}_i \right)^{-1} \sum_{i=1}^{N} \mathbf{X}_i' y_i.$$

The OLS estimator is constructed by averaging N individual components in the numerator and denominator and we know *a priori* that $E(\mathbf{X}_i' F) \neq 0$ unless F is not dynamic and therefore $p\lim\left(\hat{\beta}_{OLS} - \beta_0\right) \neq 0$ as $T \to \infty$. We also consider the Fixed Effects estimator:

$$\hat{\beta}_{FE} = \left(\sum_{i=1}^{N} \mathbf{X}_i' M_\iota \mathbf{X}_i \right)^{-1} \sum_{i=1}^{N} \mathbf{X}_i' M_\iota y_i,$$

with the within transformation $M_\iota = I_T - \iota(\iota'\iota)^{-1}\iota' = I_T - T^{-1}\iota\iota'$ and ι is a T-vector of ones. The FE estimator thus corrects each observation by subtracting a time series average from each cross-sectional equation. As with Pooled OLS, $E(X_i' M_\iota F) \neq 0$ for all $i = 1, \ldots, N$ unless F has zero temporal variation and we thus anticipate a bias of the coefficients generated by this estimator in a model with factor error

components. Finally, the CCE estimator is defined as:

$$\hat{\beta}_{CCE} = \left(\sum_{i=1}^{N} \mathbf{X}_i' M_{\overline{X}} \mathbf{X}_i \right)^{-1} \sum_{i=1}^{N} \mathbf{X}_i' M_{\overline{X}} y_i,$$

where $M_{\overline{X}} = I_T - \overline{X} \left(\overline{X}' \overline{X} \right)^{-1} \overline{X}'$ and $\overline{X} = \left[\overline{y}, \overline{X} \right]$ is a matrix of cross-sectional averages of the dependent and independent variables, i.e., $\overline{y} = N^{-1} \sum_{i=1}^{N} y_i$ and similar for the components of $\mathbf{X}$. We study the CCE estimator because it is related to QDGMM when $R_0 = 1$ and, as Everaert and de Groote (2016) show, applying the CCE to the parameters of a dynamic panel data model with factor errors yields consistent parameter estimates when both N and T are large. Moreover, their simulations further suggest that the bias does not depend much on N or T when either N or T are fixed.

The basic model we consider is a dynamic panel data model with common coefficients and a factor error structure:

$$y_{i,t} = y_{i,t-1}\beta + f_t \lambda_i + \varepsilon_{i,t}, \;\; i = 1, \ldots, N, \; t = 1, \ldots, T, \tag{3.24}$$

where we fix $\beta = 0.5$, the $\varepsilon_{i,t}$ are standard normal over all i and t and the loadings are generated as $\lambda_i \sim N(\iota_R, I_R)$ for each $i = 1, \ldots, N$. Although the loadings are estimable parameters and could be held fixed, we deliberately disadvantage the QDGMM estimator by allowing for the possibility of the normalization on $\Lambda_{0-}^{\star}$ to be ill-behaved. This structure is chosen to make the simulation more realistic. The factors in (3.24) are generated as:

$$f_t' = \vartheta f_{t-1}' + \nu_t, \;\; t = 1, \ldots, T,$$

where $\vartheta = 0.5$, $\nu_t \sim N(0, I_R)$ and we only consider $R = 1$ or $R = 2$ for brevity. As the focus of the study is on large T samples with small cross-sections, we begin the study by using combinations of $N \in \{R+1, 5, 10\}$ and $T \in \{50, 100, 200, 400\}$. The smallest cross-section we can study is 2 by Assumption 5, whereas in the largest we have a configuration of 10×400 observations where we expect our large sample results to hold. In all experiments we generate a sample of 1000 observations and only use the last T observations as our sample to allow for a sufficiently long burn-in period and guarantee stationarity of the process (3.24).

For the QDGMM estimators, we use the following instrument set to identify each of the $j = 1, \ldots, N-R$ equations:

$$z_{j,t} = \left[y_{j,t-1}, y_{N-R+1,t-1}, \ldots, y_{N,t-1} \right]'.$$

The argument for using these instruments specifically is that they provide a compromise between the many instruments problem on the one hand and identification of the factors on the other, cf. the discussion in Section 3.1. Note that when $N \leq 3$, this choice of $z_{j,t}$ implies that the model is just-identified, whereas a degree of over-identification is present when $N > 3$. We study the performance of one-step QDGMM estimators with $W = I_{S(N-R)}$ and efficient two-step estimators. By homoskedasticity of the $\varepsilon_{j,t}$, the optimal

weight matrix is estimated as the inverse of the empirical counterpart of:

$$V = \begin{bmatrix} M_1 \Omega M_1' \times E\left(Z_{1,t} Z_{1,t}'\right) & \dots & M_1 \Omega M_{N-R}' \times E\left(Z_{1,t} Z_{1,N-R}'\right) \\ \dots & & \dots \\ M_{N-R} \Omega M_1' \times E\left(Z_{N-R,t} Z_{1,t}'\right) & \dots & M_{N-R} \Omega M_{N-R}' \times E\left(Z_{N-R,t} Z_{N-R,t}'\right) \end{bmatrix},$$

since Ω is diagonal. Furthermore, identification of the factor loadings is based on the identification strategy (3.6) in Section 2.2 when $R = 1$ and on the identification strategy (3.4) when $R > 1$. The point estimates are obtained by iterating Algorithm 1 using 500 random initializations $\beta^1 \sim U(-1, 1)$ at each of the two QDGMM steps and every iteration is terminated whenever $\left|\phi^d - \phi^d\right| / [1 + R(N-R)] < 10^{-5}$ or $D = 200$ iterations occurred. We compute the condition number of the Hessian at each critical point ϕ^c and the estimator $\hat{\phi}$ is chosen as the smallest value of $Q^c\left(\hat{\phi}, \hat{R}\right)$ over all $c = 1, \dots C$ with condition number smaller than 100.

We first evaluate the average bias and RMSE of the estimators described above based on 5000 Monte Carlo replications for the process (3.24) and Appendix 2 summarizes the simulation results. As expected, in the simulations of $R = 1$, both the OLS and FE estimators are severely biased, with RMSE of over forty percent of the value of β_0. OLS is less biased than the FE estimator and the mean estimate is strictly less than unity, but the latter estimator suggests that β_0 is approximately equal to unity and thus that the model is non-stationary. This finding has important implications for empirical modelling where this estimator wrongly suggests the presence of unit roots whilst the process is actually stationary, but has a factor residual. On the other hand, the CCE estimator in the one-factor case is essentially unbiased with very low RMSE, although this bias is increasing in T. This finding is surprising because the CCE is a quasi-difference estimator with an invalid instrument set when N is small and we therefore expected the point estimates to be biased. Reassuringly, the one and two-step specifications of the QDGMM estimator perform very well in terms of bias and RMSE: for any N and any T, they outperform the CCE estimator by quite a large margin apart from the case of $N = 10$ and $T = 100$, where we observe very large RMSE of the one-step estimator. The one-step QDGMM estimator typically outperforms the theoretically efficient QDGMM estimator by a small amount as measured by mean bias, although they perform equally well in terms of RMSE. This difference is typically less than one percent of the value of β_0 and due to finite sample bias in the first step spilling into the second-step QDGMM estimator.

In the experiment with $R = 2$, both OLS and FE are again severely biased in all specifications of N and T: the OLS estimator exhibits more bias and RMSE than when $R = 1$, whereas the FE estimator remains stable with mean bias centred on, or very close to, unity, again suggesting a non-stationary model. As expected, since the CCE estimator is a quasi-difference estimator for a model with $R = 1$, the CCE is now severely biased upwards when $R = 2$. We have also simulated one and two-step QDGMM estimators using $\hat{R} = 1$ quasi-differences as a benchmark, which are denoted with (*) in the table. These estimators, similar to the CCE, are severely biased. On the other hand, the finite sample performance of the one and two-step QDGMM estimators using $R = 2$ quasi-differences is very good in all circumstances, provided that $T > 50$

or $N > 3$. Moreover, as in the experiment with $R = 1$, the one-step estimator typically performs marginally better than the two-step estimator, presumably because of first-step biases entering W. We further observe that the largest bias of one-step QDGMM is approximately eight percent when $N = 3$ and $T = 50$, whilst the bias of the two-step estimator is always smaller. In conclusion, the QDGMM estimators behave very well in all specifications we have considered here of the simulated process (3.24) and strongly outperform alternative estimators in terms of both mean bias and RMSE.

We also simulated the behaviour of Wald-tests about the slope parameter β_0 when the true number of factors is one but we incorrectly allow for two factors in the QDGMM procedure. Table 3.6 in appendix 3.2.2 summarizes the results when $N = 5$. As can be seen from the table, the null-hypothesis of $\overline{\beta} = 0$ is almost always rejected, whilst the null-hypothesis of $\overline{\beta} = \beta_0$ often is accepted. We do note however that empirical size does not approach the theoretical size in the sample sizes we consider. It is likely that this is a manifestation of the identification problem at hand and that, as a result, larger samples are required. As in the experiments in Ahn et al (2013), we find that the estimator of β_0 is consistent and suitably close in mean to the estimators using the correct number of factors.

The next experiments evaluate the small sample performance of the poolability tests of Section 3.5. First, we have simulated a Time Effects model with $\lambda_i = \overline{\lambda}$ for all $i = 1, \ldots, N$ and generate $\overline{\lambda}$ randomly from $N(1, 1)$. Note that in order to identify the model, this is equivalent to setting $\overline{\lambda} = 1$, regardless of the realization of $\overline{\lambda}$. Clearly, to compute the Wald-test for poolability of Λ, the smallest $N = 3$, which allows the test to distinguish between two loadings. We thus set $N \in \{3, 5, 10\}$ and further leave the experimental setup as in the previous simulations. Appendix 3 summarizes 5000 Monte Carlo replications of the Wald-test for the null hypothesis that the model is generated by a Time Effects model. The reported acceptance rates show that regardless of the dimensions of the panel data model, the test behaves very well: even when $T = 50$, the reported acceptance rates are always at least as large as the nominal size and tend towards unity when T becomes large.

To evaluate the poolability test of the β_i, we first apply heterogeneous parameter QDGMM to the homogeneous parameter model (3.24) with $R = 1$ and identification strategy (3.5). This results in an inefficient QDGMM estimator as the poolability hypothesis is true, but the average coefficient bias and RMSE's are of separate interest in showing the consistency of heterogeneous parameter QDGMM estimation. The experiments are based on $N \in \{2, 5, 10\}$ and we must therefore extend the instrument set to accommodate the additional parameters and ensure over-identification of the moment conditions. To this end we include one additional lag of the instruments of each of the j cross-sectional units and the last equation used to quasi-difference. This yields the following instrument set for each $j = 1, \ldots, N-1$ quasi-differences at each t:

$$z_{j,t} = [y_{j,t-1}, y_{j,t-2}y_{N,t-1}, y_{N,t-2}]'.$$

To conserve space, the first panel of Appendix 4 only reports the first, the last and, where appropriate, the middle coefficients, i.e., β_1, β_N and β_{middle} respectively. The simulations show that, as T gets large, the estimator is again virtually unbiased as long as $N > 2$. The price of estimating heterogeneous parameters

is of course that the resulting RMSE is now high compared to the pooled parameter estimator for the first and middle coefficients, but not the last, which gains efficiency from pooling, *viz.* $\hat{G}_{\beta,2}$ in Lemma 1 in the appendix. The second panel of Appendix 4 reports empirical acceptance rates of the poolability tests against theoretical size of the chi-squared distribution. Since the number of estimable parameters excluding the Λ under the homogeneity restriction is one, whereas the heterogeneous model has N parameters, the number of restrictions of the Wald-test is $N-1$. When T is large, the Wald-statistic is practically exactly sized, regardless of the size of N. Only when T is small, and $N > 2$ is the Wald-statistic somewhat oversized. However, even when $N = 2$ and $T = 50$, the Wald-statistic is still remarkably well-sized.

In the remainder of the section we test the adequacy of the model selection procedures in finite samples. Following the discussion in Remark 5.4, we would like our Monte Carlo experiment to focus only on the number of factors and not on the validity of the instruments and the factors jointly. To accomplish this we extend the number of instruments for each individual moment function to:

$$z_{j,t} = [y_{1,t-1}, \ldots, y_{N,t-1}]'.$$

which ensures that the number of instruments is fixed for each moment condition as a function of R. We generate 5000 replications of model (3.24) with T as above but fix $N = 5$ to conserve space. The model is estimated with $R = 0, \ldots, 4$ and we compute one and two-step versions of AIC, BIC and HIC using the following balancing constants:

$$\begin{aligned} \triangle_{\text{one-step}} &= \log\,(N) \times N^{\nu_1} \\ \triangle_{\text{two-step}} &= \log\,(N) \times N^{\nu_2} \end{aligned}$$

Specifically, for AIC we set $v_1 = 1$ and $v_2 = 0$ in the penalty function; for BIC $v_1 = 0$ and $v_2 = -1$ and finally for HIC $v_1 = 1$ and $v_2 = 0$. The results can be summarized as follows: although no information criteria selects $R_0 = 2$ exactly even as the sample size gets large, the results show that all information criteria approach $\hat{R} \geq R_0$ with T. This result is surprising particularly for AIC, which does not satisfy the conditions of Proposition 3.8. Another interesting observation is that one-step information criteria typically outperform two-step information criteria when the sample size is relatively small. This is due to the fact that the objective function of one-step GMM is strictly larger than the objective function of two-step GMM, implying that it becomes harder to find significant differences in $\hat{Q}(\cdot)$ as a function of R.

We also construct the sequential RK-test using $\hat{\beta}$ computed at $R = 3$ and subsequently form the gradients at $R^\star = 1, \ldots, 4$. This thus allows us to test the hypotheses $R = 0$, 1, 2, 3 in a sequential fashion for some α_T satisfying Proposition 3.10. To demonstrate the different finite sample results anticipated in Remark 5.3 we generate the significance level of the sequential J and RK-tests according to the following two functions:

$$\begin{aligned} \alpha_T &= 4N/T, \\ \alpha_T &= 0.5\sqrt{N/T} \end{aligned}$$

and the critical values for the J and RK-tests correspond to the $(1-\alpha_T)$-th quantile of the chi-squared distribution with $(S-R)(N-R)$ and $(S-R)(R^\star-R)$ degrees of freedom. Note that these functions satisfy the conditions of Propositions 3.7 and 3.10 and both are calibrated to approximately five percent when $T = 400$. The performance of neither the RK nor the J-test is competitive relative to the information criteria unless the sample size is very large. In that case, the J-test estimates R_0 correctly with the highest proportion, despite still under-estimating more often than the information criteria. Furthermore, the RK-test always performs worse than the J-test although the proportion does rise with the sample size. In either case, the convergence appears to be too slow to be of practical use, and we recommend using BIC in moderate samples.

3.7 Conclusions

A plethora of macroeconomic phenomena can be modelled as dynamic systems of individual time series, linked together by an unobserved factor structure in the residual. This factor structure captures phenomenon that are typically difficult to approximate empirically, but when present, will distort point estimates obtained by classical econometric methods. Furthermore, the empirical literature has spawned several papers that use panel data in macroeconomics which employ classical methods known to be inconsistent when a factor residual is present in the data and the results obtained are therefore difficult to interpret.

This paper develops an estimation and testing procedure for macro-panels when factors are (expected to be) present in the data, without making strong assumptions on the specification and distribution of the factor structure. To this end, we adapt the quasi-difference methodology of Holtz-Eakin et al (1989), Nauges and Thomas (2001) and Ahn, Lee and Schmidt (2001, 2013) to small N and large T dynamic panel data models with both homogeneous and heterogeneous parameter structures. We develop an extensive QDGMM limit theory for the case when the number of factors is known and also for the case when the number of factors is unknown and too many factors are included in the model. The limit theory for the case of over-estimating the number of factors is entirely new to the literature and supports the simulation results of Ahn, Lee and Schmidt (2013), which suggest that consistent estimation of the slope parameters is still possible even when $\hat{R} > R_0$. Moreover, our results show that inference about the slope parameters remains valid when the number of factors is over-estimated is re-assuring. We also provide basic inferential procedures such as model-selection and specification testing procedures that, as our Monte Carlo experiments show, work reasonably well.

Future research in this topic could develop along several directions. First, it is possible to cast the estimation procedure in the Lasso-framework: by penalizing the ALS algorithm on the blocks corresponding to excess loadings, it is expected that the approximation to the minimizer of the objective function can be improved. Second, since R_0 is unknown to the researcher, it would be prudent to conduct inference about the parameters using the bootstrap to approximate the finite sample distribution of the model when $\hat{R} \geq R_0$. Further useful extensions to this chapter can be an analysis of structural breaks in the factor model estimated by QDGMM and the behaviour of the QDGMM estimator when unit roots are present in the data, either in

the factors or the quasi-differenced model would be interesting extensions to the large T dynamic panel data with factor error residuals.

Appendix 3.1 Proofs

LEMMA 1. Definition of $\hat{G}_\beta$ for the heterogeneous Parameter Model:

Let $\hat{G}_\beta := \left[\hat{G}_{\beta,1}, -\hat{G}_{\beta,2}\right]$, where $\hat{G}_{\beta,2} := T^{-1}\sum_{t=1}^{T}\left[\oplus_{j=1}^{N-R} z_{j,t}\mathbf{X}'_{j,t}\right]$ with:

$$\begin{aligned}\mathbf{X}_{j,t} &:= \left[X'_{j,t}, Y'_{j,t-1}, \ldots, Y'_{j,t-L}\right]'; j = 1, \ldots, N \quad \text{and} \\ \mathbf{X}_{-,t} &:= \left[\mathbf{X}'_{N-R+1,t}, \ldots, \mathbf{X}'_{N,t}\right]'.\end{aligned}$$

Furthermore,

$$\begin{aligned}\hat{G}_{\beta,2} &:= T^{-1}\sum_{t=1}^{T}\left[\left(z'_{1,t}, \ldots, z'_{N-R,t}\right)' \times \left(\Lambda_+ * \mathbf{X}'_{-,t}\right)\right] \\ &= T^{-1}\sum_{t=1}^{T}\begin{bmatrix} \lambda_{1,1}z_{1,t}\mathbf{X}'_{N-R+1,t} & \cdots & \lambda_{1,R}z_{1,t}\mathbf{X}'_{N,t} \\ \cdots & & \cdots \\ \lambda_{N-R,1}z_{N-R,t}\mathbf{X}'_{N-R+1,t} & \cdots & \lambda_{N-R,R}z_{N-R,t}\mathbf{X}'_{N,t}\end{bmatrix}\end{aligned}$$

and the Khatri-Rao product in the second-last line is understood to operate on the blocks corresponding to all regressors of each of the last R cross-sectional units.

Proof of Proposition 3.1:

Proving local identification is immediate from the bi-convexity of the problem: given β_0, $E(m_t) = 0$ iff $\Lambda = \Lambda_0^\star$ and similarly for β holding fixed $\Lambda_0^\star$. Furthermore, a solution ϕ is a minimum only if the Hessian of $Q(\phi|R)$ evaluated at $E(m_t) = 0$ is positive definite and this holds whenever $G := \left[G_\beta, G_\Lambda\right]$ is full rank for any PSD weight matrix. □

LEMMA 2. Generic Uniform Convergence of $Q\left(\hat{\phi}, .\right)$ to Q_0 (Andrews 1992, Theorem 1):

If (i) $Q\left(\hat{\phi}\right) \underset{p}{\to} E\left[Q\left(\hat{\phi}\right)\right]$, (ii) $\phi \in \Phi$ with Φ totally bounded and (iii) $Q(\cdot)$ is stochastically equicontinuous, then:

$$\sup_{\phi\in\Phi}|Q(\phi) - E\left[Q(\phi)\right]| \underset{p}{\to} 0$$

and:

$$\left|\hat{\phi}-\phi_0\right| \underset{p}{\to} 0$$

Proof: See Andrews (1992). □

LEMMA 3. Stochastic Equicontinuity (Andrews 1992, Definition on page 244):

Let $B(T)$ be a function of the data and T alone, $p(\Phi)$ a function of the parameter space and $\phi, \phi^\star \in \Phi$. If:

$$\sup_{\phi,\phi^\star\in\Phi} |Q(\phi)-Q(\phi^\star)| \leq B(T)\, p(\phi,\phi^\star) < \infty$$

then $Q(\phi)$ is stochastically equicontinuous.

Proof: See Andrews (1992). □

Proof of Proposition 3.2:

To prove consistency of QDGMM with $\hat{R}=R$, we follow Theorem 1 of Andrews (1992). Using Lemmas 2-3 and given Assumptions 1-5, we first show pointwise convergence of $\hat{Q}\left(\hat{\phi}|R\right) \underset{p}{\to} Q\left(\hat{\phi}|R\right)$ and then uniform convergence of $\hat{Q}[(\cdot)|R]$ by means of stochastic equicontinuity. Given those conditions and Proposition 1, the convergence of ϕ to ϕ_0 then occurs uniformly for any $\phi \in \Phi$.

To prove pointwise convergence, let $m(\cdot) := E[\hat{m}(\cdot)]$ and write for any PSD matrix $\hat{W}$:

$$\begin{aligned}
\sup_{\phi\in\Phi}\left|\hat{Q}(\phi)-E\left[\hat{Q}(\phi)\right]\right| \leq & 2\left\|[\hat{m}(\phi)-m(\phi)]'\hat{W}[\hat{m}(\phi)-m(\phi)]\right\| + \\
& 2\left\|m(\phi)'\hat{W}[\hat{m}(\phi)-m(\phi)]\right\| \\
\leq & 2\sup_{\phi\in\Phi}\|\hat{m}(\phi)-m(\phi)\|^2\left\|\hat{W}\right\| + \\
& 2\sup_{\phi\in\Phi}\|\hat{m}(\phi)-m(\phi)\|\left\|\hat{W}\right\|\sup_{\phi\in\Phi}\|m(\phi)\|,
\end{aligned}$$

where the first inequality follows from the Triangle-inequality and the second by Cauchy-Schwartz. Note

that:

$$\begin{aligned}\hat{m}(\phi) =& \hat{m}(\phi_0) - \hat{G}^{\dagger}(\phi - \phi_0) \\ =& \hat{m}(\phi_0) - G(\phi - \phi_0) + o_p(1) \\ \text{and} & \\ m(\phi) =& m(\phi_0) - G^{\dagger}(\phi - \phi_0) \\ =& m(\phi_0) - \hat{G}(\phi - \phi_0) + o_p(1),\end{aligned}$$

where $G^{\dagger}$ is the gradient evaluated at an intermediate value $\phi^{\dagger}$ that may differ from row to row. Further note that since G and $\hat{G}$ are differentiable at ϕ_0, for any $\|\phi - \phi_0\| < \eta$ with $\eta \to 0$:

$$\begin{aligned}\sup_{\phi \in \Phi, \|\phi - \phi_0\| < \eta} \left\|\hat{G} - G\right\| &= \sup_{\phi \in \Phi, \|\phi - \phi_0\| < \eta} \sum_{i=1}^{\dim(\phi)} \left\| \frac{\partial}{\partial \phi'} G_i^{\dagger}(\phi - \phi_0) \right\| \\ &\leq \sum_{i=1}^{\dim(\phi)} \left\| \frac{\partial}{\partial \phi'} G_i^{\dagger} \right\| \|\phi - \phi_0\| = o_p(1), \end{aligned} \tag{3.25}$$

where $\frac{\partial}{\partial \phi'} G_i^{\dagger}$ is the derivative of the i-th column of G with respect to ϕ for all $i = 1, \ldots, \dim(\phi)$ and $\frac{\partial}{\partial \phi'} G_i^{\dagger}$ does not depend on ϕ. As a result, we have that:

$$\|\hat{m}(\phi) - m(\phi)\| = \|\hat{m}(\phi_0) - m(\phi_0)\| + o_p(1) = o_p(1)$$

because $\hat{m}(\phi_0) \to_p m(\phi_0) = 0$ as $T \to \infty$ by Proposition 3.1. Furthermore,

$$\|m(\phi)\| = \|m(\phi_0) - G(\phi - \phi_0)\| \leq \|G\| \|\phi - \phi_0\| = o_p(1),$$

if $\|G\|$ is bounded. Noting that $G = \left[G_{\beta}, G_{\Lambda}\right]$, we have

$$\|G\| \leq \left\|G_{\beta}\right\| + \|G_{\Lambda}\|$$

and letting $\mathbf{X}_{k,t}$ be the k-column of $\mathbf{X}_t$ we have:

$$\left\|G_{\beta}\right\| = \sum_{k=1}^{K+LN-R} \sum_{i=1}^{S} \sum_{s=1}^{T} \sum_{t=1} |z_{s,i,t} M_i \mathbf{X}_{k,t}|^2 = O_p(1)$$

by Assumption 2.(iv) and total boundedness of $\Lambda_+^{\star}$ in Assumption 3. Similarly,

$$\|G_{\Lambda}\| = \sum_{r=1}^{R} \sum_{i=1}^{N-R} \sum_{s=1}^{S} \sum_{t=1}^{T} |z_{s,i,t} f_{t,r}|^2 = O_p(1).$$

As a consequence of (3.25) and the above, a uniform weak law of large number will go through on G since

the components have sufficient moments for mixing random variables. Combining results, this implies that:

$$\sup_{\phi\in\Phi}\left|\hat{Q}(\phi)-Q(\phi)\right|=o_p(1). \tag{3.26}$$

To obtain uniform convergence of the objective function, we next show stochastic equicontinuity of $\hat{Q}(\cdot)$. Let $\tilde{\phi}$, $\phi^\star\in\Phi$ and proceed as above:

$$\begin{aligned}\left|\hat{Q}\left(\tilde{\phi}\right)-\hat{Q}(\phi^\star)\right| \leq& \left\|\left[\hat{m}\left(\tilde{\phi}\right)-\hat{m}(\phi^\star)\right]'\hat{W}\left[\hat{m}\left(\tilde{\phi}\right)-\hat{m}(\phi^\star)\right]\right\|+\\ &2\left\|\hat{m}(\phi^\star)'\hat{W}\left[\hat{m}\left(\tilde{\phi}\right)-\hat{m}(\phi^\star)\right]\right\|\\ \leq& \left\|\hat{m}\left(\tilde{\phi}\right)-\hat{m}(\phi^\star)\right\|^2\left\|\hat{W}\right\|+\\ &2\left\|\hat{m}\left(\tilde{\phi}\right)-\hat{m}(\phi^\star)\right\|\left\|\hat{W}\right\|\left\|\hat{m}(\phi^\star)-\hat{m}(\phi_0)\right\|+\\ &2\left\|\hat{m}\left(\tilde{\phi}\right)-\hat{m}(\phi^\star)\right\|\left\|\hat{W}\right\|\left\|\hat{m}(\phi_0)\right\|,\end{aligned}$$

for any PSD matrix $\hat{W}$. We now have:

$$\begin{aligned}\left\|\hat{m}\left(\tilde{\phi}\right)-\hat{m}(\phi^\star)\right\| =& \left\|\hat{G}^\dagger\left(\tilde{\phi}-\phi^\star\right)\right\| \leq \left\|\hat{G}^\dagger\right\|\left\|\tilde{\phi}-\phi^\star\right\|\\ =& \|G\|\left\|\tilde{\phi}-\phi^\star\right\|+o_p(1),\\ \left\|\hat{m}(\phi^\star)-\hat{m}(\phi_0)\right\| =& \left\|\hat{m}(\phi_0)-G^\dagger(\phi^\star-\phi_0)-\hat{m}(\phi_0)\right\|\\ =& \|G\|\left\|\phi^\star-\phi_0\right\|+o_p(1)\end{aligned}$$

and

$$\begin{aligned}\left\|\hat{m}(\phi_0)\right\| =& T^{-1}\mathrm{tr}\left[\sum_{t=1}^{T}\hat{m}_t(\phi_0)\hat{m}_t(\phi_0)'\right]:=\mathrm{tr}\left(\hat{V}\right)\\ \|G\| =& O_p(1),\end{aligned}$$

given Assumptions 1 and 2. This gives for any $\eta\to 0$:

$$\begin{aligned}\sup_{\left\|\tilde{\phi}-\phi^\star\right\|<\eta}&\left|\hat{Q}\left(\tilde{\phi}\right)-\hat{Q}(\phi^\star)\right| \leq\\ &\|G\|\left\|\hat{W}\right\|\left\{\mathrm{tr}(V_T)+\|G\|\left\|\tilde{\phi}-\phi^\star\right\|+\|G\|\left|\phi^\star-\phi_0\right|\right\}\left\|\tilde{\phi}-\phi^\star\right\|\\ &:=B_T\left\|\tilde{\phi}-\phi^\star\right\|,\end{aligned} \tag{3.27}$$

establishing Lemma 3 since $B_T=O_p(1)$.

Given uniform convergence of $\hat{Q}(\cdot)$ to $Q(\cdot)$, consistency of $\hat{\phi}$ to ϕ_0 now follows from the first order condition:

$$-\hat{G}\hat{W}\hat{m}\left(\hat{\phi}\middle|R\right)=0.$$

Substituting a mean-value expansion of $\hat{m}$ at the true parameters gives:

$$\begin{aligned}\left(\hat{\phi}-\phi_0\right) =& \left(\hat{G}W\hat{G}\right)^{-1}\hat{G}\hat{W}\hat{m}\left(\phi_0|R\right)\\ =& o_p\left(1\right)\end{aligned}$$

since $E\left[\hat{m}\left(\phi_0|R\right)\right]=0$ by Proposition 3.1 and the gradient evaluated at a mean-value $\hat{\phi}<\phi^{\dagger}<\phi_0$ converges to G. □

LEMMA 4. Consistency of Sample Covariance Matrix: *Under Assumptions 1-5 and for fixed P, the sample covariance matrix:*

$$\begin{aligned}\hat{V}\left(\hat{\phi}|R\right) =& T^{-1}\sum_{t=1}^{T}\hat{m}_t\left(\hat{\phi}\right)\hat{m}_t\left(\hat{\phi}\right)' +\\ T^{-1}\sum_{p=1}^{P}\sum_{t=p+1}^{T}&\left[\hat{m}_t\left(\hat{\phi}\right)m_{t-p}\left(\hat{\phi}\right)' + m_{t-p}\left(\hat{\phi}\right)\hat{m}_t\left(\hat{\phi}\right)'\right]\\ \rightarrow_p V\left(\phi_0|R\right) =& E\left[\hat{m}_t\left(\phi_0\right)\hat{m}_t\left(\phi_0\right)'\right] +\\ &\sum_{p=1}^{P}\left\{E\left[\hat{m}_t\left(\phi_0\right)m_{t-p}\left(\phi_0\right)'\right] + E\left[m_{t-p}\left(\phi_0\right)\hat{m}_t\left(\phi_0\right)'\right]\right\},\end{aligned}$$

Proof:

A mean-value expansion of $\hat{m}_t$ around ϕ_0 gives

$$\hat{m}_t\left(\hat{\phi}\right)=\hat{m}_t\left(\phi_0\right)-\hat{G}_t^{\dagger}\left(\hat{\phi}-\phi_0\right)=\hat{m}_t\left(\phi_0\right)-\hat{G}_t\left(\hat{\phi}-\phi_0\right)+o_p\left(1\right),$$

where $\hat{G}_t$ and $\hat{G}_t^{\dagger}$ are the gradients evaluated at each t. Then we can write:

$$\begin{aligned}T^{-1}\sum_{t=1}^{T}&\left[\hat{m}_t\left(\hat{\phi}\right)\hat{m}_t\left(\hat{\phi}\right)'\right] =\\ T^{-1}\sum_{t=1}^{T}&\left[\hat{m}_t\left(\phi_0\right)-\hat{G}_t\left(\hat{\phi}-\phi_0\right)\right]\times\left[\hat{m}_t\left(\phi_0\right)-\hat{G}_t\left(\hat{\phi}-\phi_0\right)\right]' =\\ &T^{-1}\sum_{t=1}^{T}\hat{m}_t\left(\phi_0\right)\hat{m}_t\left(\phi_0\right)' - T^{-1}\sum_{t=1}^{T}\hat{m}_t\left(\phi_0\right)\left[\hat{G}_t\left(\hat{\phi}-\phi_0\right)\right]' - T^{-1}\sum_{t=1}^{T}\hat{G}_t\left(\hat{\phi}-\phi_0\right)\hat{m}_t\left(\phi_0\right)'\\ &+T^{-1}\sum_{t=1}^{T}\hat{G}_t\left(\hat{\phi}-\phi_0\right)\left[\hat{G}_t\left(\hat{\phi}-\phi_0\right)\right]' := I+II+II'+III.\end{aligned}$$

This expansion implies

$$T^{-1}\sum_{t=1}^{T}\left[\hat{m}_t\left(\hat{\phi}\right)\hat{m}_t\left(\hat{\phi}\right)'\right]\rightarrow_p E\left[\hat{m}_t\left(\phi_0\right)\hat{m}_t\left(\phi_0\right)'\right],$$

if *II* and *III* vanish. Writing:

$$T^{-1}\text{vec}\sum_{t=1}^{T}\hat{m}_t\left(\phi_0\right)\left[\hat{G}_t\left(\hat{\phi}-\phi_0\right)\right]' = T^{-1}\sum_{t=1}^{T}\left[\hat{G}_t\otimes\hat{m}_t\left(\phi_0\right)\right]\left(\hat{\phi}-\phi_0\right),$$
$$T^{-1}\text{vec}\sum_{t=1}^{T}\hat{G}_t\left(\hat{\phi}-\phi_0\right)\left[\hat{G}_t\left(\hat{\phi}-\phi_0\right)\right]' = T^{-1}\sum_{t=1}^{T}\left[\hat{G}_t\otimes\hat{G}_t\right]\text{vec}\left[\left(\hat{\phi}-\phi_0\right)\left(\hat{\phi}-\phi_0\right)'\right],$$

we note that both terms go to zero in probability by consistency of $\hat{\phi}$ provided that

$$T^{-1}\sum_{t=1}^{T}\left[\hat{G}_t\otimes\hat{m}_t\left(\phi_0\right)\right] = O_p\left(1\right)$$
$$T^{-1}\sum_{t=1}^{T}\left[\hat{G}_t\otimes\hat{G}_t\right] = O_p\left(1\right).$$

By Cauchy-Schwartz, for the first line above:

$$\sum_{t=1}^{T}\left[\hat{G}\otimes\hat{m}_t\left(\phi_0\right)\right] \leq \sqrt{T^{-1}\sum_{t=1}^{T}\left|\hat{G}_t\right|^2}\otimes\sqrt{T^{-1}\sum_{t=1}^{T}\left|\hat{m}_t\left(\phi_0\right)\right|^2},$$

where the absolute value and power are understood as operating element-wise.

Similarly, for fixed P and if $P/T\to 0$,

$$\begin{aligned}
T^{-1}\left(T-P\right)\left(T-P\right)^{-1}&\sum_{t=p+1}^{T}\left[\hat{m}_t\left(\hat{\phi}\right)m_{t-p}\left(\hat{\phi}\right)'\right] = \\
&\left(T-P\right)^{-1}\sum_{t=p+1}^{T}\left[\hat{m}_t\left(\phi_0\right)-\hat{G}\left(\hat{\phi}-\phi_0\right)\right]\times\left[m_{t-p}\left(\phi_0\right)-G_{t-p}\left(\hat{\phi}-\phi_0\right)\right]' = \\
&\left(T-P\right)^{-1}\sum_{t=p+1}^{T}\left[\hat{m}_t\left(\phi_0\right)m_{t-p}\left(\phi_0\right)'\right]-\left(T-P\right)^{-1}\sum_{t=p+1}^{T}\hat{m}_t\left(\phi_0\right)\left[G_{t-p}\left(\hat{\phi}-\phi_0\right)\right]' \\
-\left(T-P\right)^{-1}&\sum_{t=p+1}^{T}G_{t-p}\left(\hat{\phi}-\phi_0\right)m_{t-p}\left(\phi_0\right)'+\left(T-P\right)^{-1}\sum_{t=p+1}^{T}\hat{G}\left(\hat{\phi}-\phi_0\right)\left[G_{t-p}\left(\hat{\phi}-\phi_0\right)\right]' \\
&:= I+II+II'+III.
\end{aligned}$$

and *II* and *III* vanish as before:

$$\left(T-P\right)^{-1}\text{vec}\sum_{t=1}^{T}\hat{m}_t\left(\phi_0\right)\left[G_{t-p}\left(\hat{\phi}-\phi_0\right)\right]' = \left(T-P\right)^{-1}\sum_{t=1}^{T}\left[G_{t-p}\otimes\hat{m}_t\left(\phi_0\right)\right]\left(\hat{\phi}-\phi_0\right),$$
$$\left(T-P\right)^{-1}\text{vec}\sum_{t=1}^{T}\hat{G}_t\left(\hat{\phi}-\phi_0\right)\left[G_{t-p}\left(\hat{\phi}-\phi_0\right)\right]' = \left(T-P\right)^{-1}\sum_{t=1}^{T}\left[G_{t-p}\otimes\hat{G}_t\right]\text{vec}\left[\left(\hat{\phi}-\phi_0\right)\left(\hat{\phi}-\phi_0\right)'\right],$$

implying that:

$$T^{-1}\sum_{t=1}^{T}\left[\hat{m}_t\left(\hat{\phi}\right)m_{t-p}\left(\hat{\phi}\right)'\right]\rightarrow_p E\left[\hat{m}_t\left(\phi_0\right)m_{t-p}\left(\phi_0\right)'\right].\ \square$$

Proof of Proposition 3.3:

Distributional theory for correctly specified QDGMM is obtained in a standard way, cf. Theorem 3.2 of Newey and McFadden (1984). The distribution of the vector of moment conditions follows from a CLT applied to the random sums $T^{-1/2}\sum_{t=1}^{T}z_{j,t}M_j\varepsilon_t$ for each $j=1,\ldots N-R$ because f_t is removed at the true parameters.[11] Under Assumptions 1, 2, 3 and 4, a CLT for mixing variables goes through because measurable functions of mixing variables are also mixing and we assume sufficient moments are finite, cf. Theorem 5.20 of White (2001). The distribution of the moment vector is then:

$$\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\hat{m}_t\left(\phi_0|R\right)\rightarrow_d N\left(0_{S(N-R)\times 1},V\right)$$

with V defined in Proposition 3.3. Conclusion (2) now follows immediately by standard arguments: since $T^{-1/2}V^{-1/2}\sum_{t=1}^{T}\hat{m}_t\left(\phi\right)$ converges to a multivariate standard normal of the same dimension, the corresponding quadratic form converges to a chi-squared random variable as $T\rightarrow\infty$. This implies that setting $W=V^{-1}$, the objective function $\hat{Q}\left(\phi|R\right)$ follows a chi-squared distribution, with degrees of freedom equal to the number of instruments *sans* the parameter count of the homogeneous or heterogeneous slope parameter model, i.e., $(N-R)(S-R)-\dim(\beta)$.

For conclusion (1) we use standard GMM limit theory by substituting the mean-value expansion:

$$\hat{m}\left(\hat{\phi}\right)=\hat{m}\left(\phi_0|R\right)-\hat{G}^{\dagger}\left(\hat{\phi}-\phi_0\right),$$

where $\hat{G}^{\dagger}$ is the gradient evaluated at $\hat{\phi}\leq\phi^{\dagger}\leq\phi_0$, into the first order conditions:

$$\sqrt{T}\left(\hat{\phi}-\phi_0\right)=\left(\hat{G}\hat{W}\hat{G}^{\dagger}\right)^{-1}G^{\dagger\prime}\hat{W}\sqrt{T}\hat{m}\left(\phi_0|R\right).$$

Since $\hat{\phi}$ is consistent, the gradients evaluated at $\hat{\phi}$ and $\phi^{\dagger}$ will converge in probability to G and the stated result follows. $\square$

LEMMA 5: *Suppose Assumptions 1-5 hold and $\hat{R}>R$. Then there exist sequences of matrices $\mathcal{H}_T$ such that:*

$$\hat{m}\left(\hat{\phi}|\hat{R}\right)=\hat{m}\left(\phi_0|R\right)-\hat{G}^{\dagger}_{\mathcal{H}}\left(\underline{\hat{\phi}}_{\mathcal{H}}-\underline{\phi}_0\right),$$

where $\underline{\phi}_0=\left[\beta',0',\mathrm{vec}\left(\underline{\Lambda}^{\star}_{0+}\right)'\right]'$ *and* $\hat{G}^{\dagger}_{\mathcal{H}}$ *is full rank.*

[11] Note that the normality results under weaker assumptions on the dependence of the model also readily follow.

Proof:

The result follows from a mean-value expansion about $\underline{\phi}_0$:

$$\begin{aligned}\hat{m}\left(\hat{\underline{\phi}}|\hat{R}\right) =& \hat{m}\left(\underline{\phi}_0|R\right) - \hat{G}^{\dagger}\left(\hat{\underline{\phi}} - \underline{\phi}_0\right) \\ =& \hat{m}\left(\underline{\phi}_0|R\right) - \hat{G}\mathcal{H}\mathcal{H}^{-1}\left(\hat{\underline{\phi}} - \underline{\phi}_0\right) + o_p(1) \\ =& \hat{m}\left(\underline{\phi}_0|R\right) - \hat{G}^{\dagger}_{\mathcal{H}}\left(\hat{\underline{\phi}}_{\mathcal{H}} - \underline{\phi}_0\right),\end{aligned}$$

where $\hat{G}^{\dagger}$ and $\hat{G}^{\dagger}_{\mathcal{H}}$ are evaluated at intermediate values of $\underline{\phi}_0$ which may differ from row to row and it is noted that $G^{\dagger}_{\mathcal{H}}$ is of full rank. □

LEMMA 6: *Suppose Assumptions 1-5 hold, $\hat{R} > R$ and P fixed. Then the optimal weight matrix is consistent if $\hat{\underline{\phi}}_{\mathcal{H}} \rightarrow_p \underline{\phi}_0$:*

$$\begin{aligned}\hat{V}\left(\hat{\underline{\phi}}|\hat{R}\right) :=& T^{-1}\sum_{t=1}^{T} m_t\left(\hat{\underline{\phi}}|\hat{R}\right) m_t\left(\hat{\underline{\phi}}|\hat{R}\right)' + \\ T^{-1}\sum_{p=1}^{P}\sum_{t=p+1}^{T}&\left[m_t\left(\hat{\underline{\phi}}|\hat{R}\right) m_{t-p}\left(\hat{\underline{\phi}}|\hat{R}\right)' + m_{t-p}\left(\hat{\underline{\phi}}|\hat{R}\right) m_t\left(\hat{\underline{\phi}}|\hat{R}\right)'\right] \rightarrow_p \\ V\left(\underline{\phi}_0|R\right) :=& E\left[m_t\left(\underline{\phi}_0|R\right) m_t\left(\underline{\phi}_0|R\right)'\right] + \\ &\sum_{p=1}^{P}\left\{E\left[m_t\left(\underline{\phi}_0|R\right) m_{t-p}\left(\underline{\phi}_0|R\right)'\right] + E\left[m_{t-p}\left(\underline{\phi}_0|R\right) m_t\left(\underline{\phi}_0|R\right)'\right]\right\}.\end{aligned}$$

Proof:

The proof makes use of Lemma 5 and is very similar to the proof of lemma 4. As a result we only show consistency of the leading term.

Using the expansion in Lemma 4 at each t, then squaring, summing and dividing by T gives:

$$\begin{aligned}T^{-1}\sum_{t=1}^{T} m_t\left(\hat{\underline{\phi}}|\hat{R}\right) m_t\left(\hat{\underline{\phi}}|\hat{R}\right)' =& T^{-1}\sum_{t=1}^{T} m_t\left(\underline{\phi}_0|R\right) m_t\left(\underline{\phi}_0|R\right)' - \\ T^{-1}\sum_{t=1}^{T} m_t\left(\underline{\phi}_0|R\right)\left(\hat{\underline{\phi}}_{\mathcal{H}} - \underline{\phi}_0\right)'\hat{G}'_{\mathcal{H},t} - T^{-1}\sum_{t=1}^{T}\hat{G}_{\mathcal{H},t}\left(\hat{\underline{\phi}}_{\mathcal{H}} - \underline{\phi}_0\right) m_t\left(\underline{\phi}_0|R\right)' + \\ T^{-1}\sum_{t=1}^{T}\hat{G}_{\mathcal{H},t}\left(\hat{\underline{\phi}}_{\mathcal{H}} - \underline{\phi}_0\right)\left(\hat{\underline{\phi}}_{\mathcal{H}} - \underline{\phi}_0\right)'\hat{G}'_{\mathcal{H},t} :=& I + II + II' + IV.\end{aligned}$$

Since $II = II'$, the claim follows if we can show that $II, IV = o_p(1)$.

For II we have:

$$\text{vec}\left[T^{-1}\sum_{t=1}^{T} m_t\left(\underset{\sim}{\phi}_0|R\right)\left(\hat{\underset{\sim}{\phi}}_{\mathcal{H}} - \underset{\sim}{\phi}_0\right)'\hat{G}'_{\mathcal{H},t}\right] = T^{-1}\sum_{t=1}^{T}\left[\hat{G}_{\mathcal{H},t}\otimes m_t\left(\underset{\sim}{\phi}_0|R\right)\right]\times \text{vec}\left[\left(\hat{\underset{\sim}{\phi}}_{\mathcal{H}} - \underset{\sim}{\phi}_0\right)'\right].$$

So II vanishes by consistency of $\hat{\underset{\sim}{\phi}}_{\mathcal{H}}$ if the first term is $O_p(1)$. As in the proof of Lemma 4 we have by element-wise Cauchy-Schwartz:

$$T^{-1}\sum_{t=1}^{T}\left[\hat{G}_{\mathcal{H},t}\otimes\hat{m}\left(\underset{\sim}{\phi}_0|R\right)\right] \leq \sqrt{T^{-1}\sum_{t=1}^{T}\left|\hat{G}_{\mathcal{H},t}\right|^2}\otimes\sqrt{T^{-1}\sum_{t=1}^{T}\left|m_t\left(\underset{\sim}{\phi}_0|R\right)\right|^2},$$

which are finite by Assumptions 1 and 2.

For IV we similarly have:

$$T^{-1}\sum_{t=1}^{T}\left[\hat{G}_{\mathcal{H},t}\otimes\hat{G}'_{\mathcal{H},t}\right]\times\text{vec}\left[\left(\hat{\underset{\sim}{\phi}}_{\mathcal{H}} - \underset{\sim}{\phi}_0\right)\left(\hat{\underset{\sim}{\phi}}_{\mathcal{H}} - \underset{\sim}{\phi}_0\right)'\right] = o_p(1)$$

We conclude that:

$$T^{-1}\sum_{t=1}^{T} m_t\left(\hat{\underset{\sim}{\phi}}|\hat{R}\right)m_t\left(\hat{\underset{\sim}{\phi}}|\hat{R}\right)' \to_p E\left[m_t\left(\underset{\sim}{\phi}_0|R\right)m_t\left(\underset{\sim}{\phi}_0|R\right)'\right].$$

The P autocovariance terms are derived similarly, thus proving the consistency of $\hat{V}\left(\hat{\underset{\sim}{\phi}}|\hat{R}\right)$. □

Proof of Theorem 3.1:

To establish uniform convergence of $\hat{Q}\left(\underset{\sim}{\phi}|\hat{R}\right)$ for all $\underset{\sim}{\phi}\in\Phi$, we proceed as in Proposition 3.2 by following Lemma 1 and 2. By replacing $\hat{G}$ and G by $\hat{G}_{\mathcal{H}}$ and $G_{\mathcal{H}}$, pointwise convergence readily follows since $\|G_\times\| = O_p(1)$ by Assumption 2. Thus establishing stochastic equicontinuity as in equations (3.26) and (3.27) in the proof of Proposition 3.2.

Consistency of $\hat{\underset{\sim}{\phi}}_{\mathcal{H}}$ then follows from the first-order condition of the QDGMM optimization program and substitution of the expansion in Lemma 4:

$$\left(\hat{\underset{\sim}{\phi}}_{\mathcal{H}} - \underset{\sim}{\phi}_0\right) = \left(\hat{G}'_{\mathcal{H}}\hat{W}\hat{G}_{\mathcal{H}}\right)^{-1}\hat{G}'_{\mathcal{H}}\hat{W}\hat{m}\left(\underset{\sim}{\phi}_0|R\right) = O_p\left(T^{-1/2}\right).$$

Since $\hat{G}_{\mathcal{H}} = \left[\hat{G}'_\beta, \hat{G}'_\times, \hat{G}'_-\right]'$, continuity of the normal distribution implies that $\hat{G}_\times$ consists of realizations

exactly equal to zero with probability zero and thus that the matrix $\hat{G}_{\mathcal{H}}$ has full rank almost surely. As a result, the solution above exists and is consistent. Note further that:

$$\begin{aligned}
\hat{m}_{ZU} =& \left[\oplus_{j=1}^{N-\hat{R}} T^{-1}\sum_{t=1}^{T}(z_{j,t}f_{1,t}),\ldots,\oplus_{j=1}^{N-\hat{R}} T^{-1}\sum_{t=1}^{T}(z_{j,t}f_{R,t})\right]\text{vec}\left(\underline{\Lambda}_{0+}^{\star}\right)+\\
&\left[T^{-1}\sum_{t=1}^{T}(z_{1,t}\varepsilon_{1,t})',\ldots,T^{-1}\sum_{t=1}^{T}\left(z_{N-\hat{R},t}\varepsilon_{N-\hat{R},t}\right)'\right]'+O_p\left(T^{-1/2}\right)\\
=&\hat{G}_{-}\text{vec}\left(\underline{\Lambda}_{0+}^{\star}\right)+\left[T^{-1}\sum_{t=1}^{T}(z_{1,t}\varepsilon_{1,t})',\ldots,T^{-1}\sum_{t=1}^{T}\left(z_{N-\hat{R},t}\varepsilon_{N-\hat{R},t}\right)'\right]'-\\
&\left[\oplus_{j=1}^{N-\hat{R}} T^{-1}\sum_{t=1}^{T}\left(z_{j,t}\varepsilon_{N-\hat{R}+1,t}\right),\ldots,\oplus_{j=1}^{N-\hat{R}} T^{-1}\sum_{t=1}^{T}(z_{j,t}\varepsilon_{N,t})\right]\text{vec}\left(\underline{\Lambda}_{0+}^{\star}\right)+O_p\left(T^{-1/2}\right)\\
=&\hat{G}_{-}\text{vec}\left(\underline{\Lambda}_{0+}^{\star}\right)+\hat{m}\left(\underline{\phi}_0|R\right),
\end{aligned}$$

where we have used the normalization $\underline{\Lambda}_{-}=I_{R_0}$ in the first and second lines. We then have by the block-inverse formula that:

$$\begin{aligned}
\text{vec}\left(\hat{\underline{\Lambda}}_{H\times}\right)=&\left(\hat{G}_{\times}'\hat{W}^{1/2}M_{\hat{W}^{1/2}\hat{G}_{\mathcal{H}/\times}}\hat{W}^{1/2}\hat{G}_{\times}\right)^{-1}\hat{G}_{\times}'\hat{W}^{1/2}M_{\hat{W}^{1/2}\hat{G}_{\mathcal{H}/\times}}\hat{W}^{1/2}\hat{m}_{ZU}\\
=&0_{(\hat{R}-R)(N-\hat{R})\times 1}+O_p\left(T^{-1/2}\right)\qquad(3.28)\\
&\text{and}\\
\text{vec}\left(\hat{\underline{\Lambda}}_{H+}\right)=&\left(\hat{G}_{-}'\hat{W}^{1/2}M_{\hat{W}^{1/2}\hat{G}_{\mathcal{H}/-}}\hat{W}^{1/2}\hat{G}_{-}\right)^{-1}\hat{G}_{-}'\hat{W}^{1/2}M_{\hat{W}^{1/2}\hat{G}_{\mathcal{H}/-}}\hat{W}^{1/2}\hat{m}_{ZU}\\
=&\text{vec}\left(\underline{\Lambda}_{0+}^{\star}\right)+O_p\left(T^{-1/2}\right),
\end{aligned}$$

where $M_{\hat{W}^{1/2}\hat{G}_{\mathcal{H}/\Lambda_{\times}}}$ is the orthogonal projection on the columns of $\hat{W}^{1/2}\left[\hat{G}_{\beta}',\hat{G}_{-}'\right]'$ and similarly for $M_{\hat{W}^{1/2}\hat{G}_{\mathcal{H}/\Lambda_{-}}}$. The above makes explicit (i) the loss of information due to overestimating R and (ii) that overestimating still removes the factors, facilitating a consistent estimate of β_0.

For the distribution of $\left(\hat{\underline{\phi}}_H-\underline{\phi}_0\right)$, note that:

$$\sqrt{T}\hat{m}\left(\underline{\phi}_0|R\right)\sim N\left[0,V\left(\underline{\phi}_0|R\right)\right]$$

and therefore as $T\to\infty$:

$$\begin{aligned}
&\sqrt{T}\left(\hat{\underline{\phi}}_{\mathcal{H}}-\underline{\phi}_0\right)\sim\\
&\qquad MN\left[0,\left(G_{\mathcal{H}}'WG_{\mathcal{H}}\right)^{-1}G_{\mathcal{H}}'WV\left(\underline{\phi}_0|R\right)WG_{\mathcal{H}}\left(G_{\mathcal{H}}'WG_{\mathcal{H}}\right)^{-1}\right]
\end{aligned}$$

The mixed normality in the above is due to the occurrence of $G_{\times}$ in $G_{\mathcal{H}}$, which is correlated with $m(\cdot)$.

For part (2) of the Theorem, first note that:

$$H = \begin{bmatrix} I_{\hat{R}-R} \otimes I_{N-\hat{R}} & 0_{(\hat{R}-R)\times R} \otimes I_{N-\hat{R}} \\ -\Lambda_{0\times}^{\star\prime} \otimes I_{N-\hat{R}} & I_R \otimes I_{N-\hat{R}} \end{bmatrix} \begin{bmatrix} \sqrt{T} I_{\hat{R}-R} \otimes I_{N-\hat{R}} & 0_{(\hat{R}-R)\times R} \otimes I_{N-\hat{R}} \\ 0_{R\times(N-\hat{R})} \otimes I_{N-\hat{R}} & I_R \otimes I_{N-\hat{R}} \end{bmatrix}$$
$$:= H_{\Lambda_\times} \times H_{\sqrt{T}}.$$

In light of (3.28), the partition of H implies:

$$H_{\sqrt{T}} H^{-1} \text{vec}\left(\underline{\hat{\Lambda}}\right) = \begin{bmatrix} \sqrt{T}\text{vec}\left(\underline{\hat{\Lambda}}_{H\times}\right) \\ \text{vec}\left(\underline{\hat{\Lambda}}_{H+}\right) \end{bmatrix}$$

and it follows that:

$$HH^{-1}\text{vec}\left(\underline{\hat{\Lambda}}\right) = H_{\Lambda_\times} H_{\sqrt{T}} H^{-1}\text{vec}\left(\underline{\hat{\Lambda}}\right) \to_d \begin{bmatrix} \Psi_\Lambda \\ \text{vec}\left(\Lambda_{0+}^{\star}\right) - \left(\underline{\Lambda}_{0\times}^{\star\prime} \otimes I_{N-\hat{R}}\right)\Psi_\Lambda \end{bmatrix},$$

where

$$\Psi_\Lambda \sim \left(G_\times' W^{1/2} M_{W^{1/2} G_{\mathcal{H}/\times}} W^{1/2} G_\times\right)^{-1} G_\times' W^{1/2} M_{W^{1/2} G_{\mathcal{H}/\times}} W^{1/2} \sqrt{T}\hat{m}\left(\underline{\phi}_0 | R\right)$$

is a mixed normal random variable exhibiting correlation between $G_\times$ and the limit of $\sqrt{T}\hat{m}(\cdot)$. The joint distribution of $\underline{\hat{\phi}}$ in partitioned form is obtained as follows. Let:

$$H_{\Lambda_\times} = \left[H_{\Lambda_\times,1}, H_{\Lambda_\times,2}\right],$$
$$\hat{\mathcal{G}} = \begin{bmatrix} \left(\hat{G}_\beta' \hat{W}^{1/2} M_{\hat{W}^{1/2}\hat{G}_{\mathcal{H}/\beta}} \hat{W}^{1/2} \hat{G}_\beta\right)^{-1} & 0 \\ 0 & H_{\Lambda_\times,1}\left(\hat{G}_\times' \hat{W}^{1/2} M_{\hat{W}^{1/2}\hat{G}_{\mathcal{H}/\times}} \hat{W}^{1/2} \hat{G}_\times\right)^{-1} \end{bmatrix} \times$$
$$\begin{bmatrix} \hat{G}_\beta' \hat{W}^{1/2} M_{\hat{W}^{1/2}\hat{G}_{\mathcal{H}/\beta}} \hat{W}^{1/2} \\ \hat{G}_\times' \hat{W}^{1/2} M_{\hat{W}^{1/2}\hat{G}_{\mathcal{H}/\times}} \hat{W}^{1/2} \end{bmatrix}.$$

then:

$$\text{diag}\left[\sqrt{T} I_{\dim(\beta)}, I_{\hat{R}(N-\hat{R})}\right]\left(\underline{\hat{\phi}} - \underline{\phi}_0\right) = \hat{\mathcal{G}}\sqrt{T}\hat{m}\left(\underline{\phi}_0 | R\right) \to_d MN\left[0, \mathcal{G}V\left(\underline{\phi}_0 | R\right)\mathcal{G}'\right].$$

It should be clear from the dimensions of $H_{\Lambda_\times,1}$ that the unscaled parameter vector has a degenerate random limit and this is due to the fact that we are using more parameters than necessary to estimate the (span of the) loadings.

Finally, we show that the J-statistic continues to have a chi-squared distribution even when $\hat{R} > R$. Using

the first order conditions in the expansion of Lemma 4, we have:

$$\begin{aligned}\hat{m}\left(\underset{\sim}{\hat{\phi}}|\hat{R}\right) =& \hat{m}\left(\underset{\sim}{\phi}_0|R\right) - \hat{G}_{\mathcal{H}}\left(\hat{G}'_{\mathcal{H}}\hat{W}\hat{G}_{\mathcal{H}}\right)^{-1}\hat{G}'_{\mathcal{H}}\hat{W}\hat{m}\left(\underset{\sim}{\phi}_0|R\right)\\ =& \left[I - \hat{G}_{\mathcal{H}}\left(\hat{G}'_{\mathcal{H}}\hat{W}\hat{G}_{\mathcal{H}}\right)^{-1}\hat{G}'_{\mathcal{H}}\hat{W}\right]\hat{m}\left(\underset{\sim}{\phi}_0|R\right)\\ =& \hat{W}^{-1/2}M_{\hat{W}^{1/2}\hat{G}_{\mathcal{H}}}\hat{W}^{1/2}\hat{m}\left(\underset{\sim}{\phi}_0|R\right).\end{aligned}$$

Then by Lemma 5 setting $\hat{W} = \hat{V}^{-1}$, the quadratic form

$$\begin{aligned}\hat{m}\left(\underset{\sim}{\hat{\phi}}|\hat{R}\right)'\hat{V}^{-1}\hat{m}\left(\underset{\sim}{\hat{\phi}}|\hat{R}\right) =& \hat{m}\left(\underset{\sim}{\phi}_0|R\right)'\hat{V}^{-1/2}M_{\hat{V}^{-1/2}\hat{G}_{\mathcal{H}}}\hat{V}^{-1/2}\hat{m}\left(\underset{\sim}{\phi}_0|R\right)\\ &\rightarrow_d \chi^2\left\{\left(S-\hat{R}\right)\left(N-\hat{R}\right) - \dim(\beta)\right\},\end{aligned}$$

since (i), $\hat{V} \rightarrow V$, (ii) $\hat{V}^{-1/2}\hat{m}\left(\underset{\sim}{\phi}_0|R\right) \rightarrow_d N(0,1)$ and (iii) $M_{\hat{W}^{1/2}\hat{G}_{\mathcal{H}}}$ is idempotent.

Proof of Proposition 3.5:

Let $\partial a/\partial\phi = A = [A,0]$, let $a\left(\underset{\sim}{\hat{\phi}}\right) = \mathcal{A}\left(\underset{\sim}{\hat{\beta}}\right)$ and note that $A\mathcal{H} = A$. Then by Theorem 3.1, under H_0, a mean-value expansion gives

$$\begin{aligned}\sqrt{T}a\left(\underset{\sim}{\hat{\phi}}\right) =& \sqrt{T}a\left(\underset{\sim}{\phi}_0\right) + \sqrt{T}A\mathcal{H}\mathcal{H}^{-1}\left(\underset{\sim}{\hat{\phi}} - \underset{\sim}{\phi}_0\right)\\ =& \sqrt{T}A\left(\underset{\sim}{\hat{\phi}}_H - \underset{\sim}{\phi}_0\right) \sim MN\left[0, A\left(\hat{G}'_{\mathcal{H}}\hat{V}^{-1}\hat{G}_{\mathcal{H}}\right)^{-1}A'\right]\\ =& N\left[0, A\left(\hat{G}'_{\beta}\hat{V}^{-1/2}M_{\hat{V}^{-1/2}\hat{G}_H}\hat{V}^{-1/2}\hat{G}_{\beta}\right)^{-1}A'\right],\end{aligned}$$

since (i) $a\left(\underset{\sim}{\phi}_0\right) = 0$ and (ii) the random matrix $G_{\times}$ is in the annihilator $M_{(\cdot)}$ and projection matrices have eigenvalues equal to zero and unity. As a result, the quadratic form in the proposition is chi-squared distributed with rank (A) degrees of freedom. □

Proof of Proposition 3.6:

In the special case that $\lambda_i = \overline{\lambda}$, we may apply the normalization that $\lambda_i = 1$ for all $i = 1,\ldots,N$. We have by Theorem 3.1:

$$\text{vec}\left(\underset{\sim}{\hat{\Lambda}}\right) \rightarrow_d \begin{bmatrix} \Psi_{\Lambda} \\ \mathbf{1}_{N-R} - \left(\mathbf{1}'_{R-R_0} \otimes I_{N-R}\right)\Psi_{\Lambda} \end{bmatrix},$$

and thus $A\text{vec}\left(\hat{\underline{\Lambda}}\right) = 0_{(N-R-1)\times 1}$. To obtain the distributional result, note that:

$$A\left(H^{-1}H\right)\left(\hat{G}_{\Lambda}'\hat{V}^{-1/2}M_{\hat{V}^{-1/2}\hat{G}_{\beta}}\hat{V}^{-1/2}\hat{G}_{\Lambda}\right)^{+}\left(HH^{-1}\right)'\hat{G}_{\Lambda}'\hat{V}^{-1/2}M_{\hat{V}^{-1/2}\hat{G}_{\beta}}\hat{V}^{-1/2}\hat{m}_{ZU} =$$
$$AH\left(\hat{G}_{H}'\hat{V}^{-1/2}M_{\hat{V}^{-1/2}\hat{G}_{\beta}}\hat{V}^{-1/2}\hat{G}_{H}\right)^{-1}\hat{G}_{H}'\hat{V}^{-1/2}M_{\hat{V}^{-1/2}\hat{G}_{\beta}}\hat{V}^{-1/2}\hat{m}_{ZU},$$

where now

$$H = \begin{bmatrix} I_{R-1}\otimes I_{N-R} & 0_{(R-1)\times 1}\otimes I_{N-R} \\ -1_{1,R-1}\otimes I_{N-R} & I_{N-R} \end{bmatrix}.$$

As a result we have

$$AH = \left[0_{1\times(R-1)},1\right]\otimes\begin{bmatrix} 1 & -1 & 0 & \dots & \dots & 0 \\ 0 & 1 & -1 & 0 & \ddots & \dots \\ \dots & 0 & \ddots & \ddots & \ddots & \dots \\ \dots & \ddots & \ddots & \ddots & \ddots & 0 \\ 0 & \dots & \dots & 0 & 1 & -1 \end{bmatrix},$$

so that

$$A\text{vec}\left(\hat{\underline{\Lambda}}\right) = AHH^{-1}\text{vec}\left(\hat{\underline{\Lambda}}\right) = \begin{bmatrix} 1 & -1 & 0 & \dots & \dots & 0 \\ 0 & 1 & -1 & 0 & \ddots & \dots \\ \dots & 0 & \ddots & \ddots & \ddots & \dots \\ \dots & \ddots & \ddots & \ddots & \ddots & 0 \\ 0 & \dots & \dots & 0 & 1 & -1 \end{bmatrix}\text{vec}\left(\hat{\underline{\Lambda}}_{H+}\right) = 0,$$

since $\lambda_j^{\star} = 1$ for all $j = 1,\dots,N-R$ and we arrive at the distributional result:

$$A\text{vec}\left(\hat{\underline{\Lambda}}\right) \to_d N\left[0, A\left(G_{-}'V_{R_0}^{-1/2}M_{\hat{V}^{1/2}\hat{G}_{\mathcal{H}/-}}V_{R_0}^{-1/2}G_{\Lambda}\right)^{-1}A'\right],$$

so that the quadratic form

$$\left[A\text{vec}\left(\hat{\underline{\Lambda}}\right)\right]'\left[A\left(G_{-}'V_{R_0}^{-1/2}M_{\hat{V}^{1/2}\hat{G}_{\mathcal{H}/\Lambda_{-}}}V_{R_0}^{-1/2}G_{\Lambda}\right)^{-1}A'\right]^{-1}A\text{vec}\left(\hat{\underline{\Lambda}}\right) \to_d \chi^2\left(N-R-1\right).$$

□

Proof of Proposition 3.7:

We follow Theorem 5 of Cragg and Donald (1997). These authors note that it is not possible to bound the probability of overestimating without conditions (1) and (2) on the significance level function α_T. These requirements are from Pötscher (1983), who derives a convergence rate on the difference between chi-squared variables with k and l degrees of freedom where $k < l$.

Let E_R be the event of rejecting the null hypothesis of R factors based on the J-test and let $\bar{E}_R$ the probability of accepting a null hypothesis of R factors. The probability of estimating any R factors is then:

$$P(E_1 \cap \cdots \cap E_{R-1} \cap \bar{E}_R) = P(E_1 \cap \cdots \cap E_{R-1}) P(\bar{E}_R | E_1 \cap \cdots \cap E_{R-1}),$$

so that for $R < R_0$,

$$P(R < R_0) \leq P(\bar{E}_R) = 1 - P(E_R),$$

and

$$P(E_R) = P\left[Q(\phi, R) > c\left(\chi^2, R, \alpha_T\right)\right],$$

where for $R < R_0$, $Q(\phi, R) = O_p(T)$ since $\|\Lambda F' v\| = O_p(T)$ for any conformable vector v and, using $-T^{-1}\log(\alpha_T) \to 0$ of the proposition, $T^{-1}c\left(\chi^2, R, \alpha_T\right) \to_p 0$, leading to:

$$P(E_R) = P(O_p(1) > o_p(1)) \to_p 1,$$

yielding that $\lim P(R < R_0) = 0$.

For $R > R_0$, we have

$$P(R > R_0) \leq P(E_{R_0}) = P\left[\hat{Q}\left(\hat{\phi}, R_0\right) > c\left(\chi^2, R_0, \alpha_T\right)\right],$$

which is the type-I error. To bound this, fix $\eta > 0$ and let $\triangle > 0$ be some constant. Then, as $T \to \infty$, $\alpha_T \to 0$ and $c\left(\chi^2, R_0, \alpha_T\right) \to \infty$, there will be some $T > \overline{T}$ for which $\triangle < c\left(\chi^2, R_0, \alpha_T\right)$ and:

$$P\left[\chi^2\{(S - R_0)(N - R_0) - \dim(\beta)\} > \triangle\right] = \eta.$$

But $\hat{Q}\left(\hat{\phi}, R_0\right) \to_d \chi^2\{(S - R_0)(N - R_0) - \dim(\beta)\}$ so by construction there is a $T > \tilde{T}$ such that

$$\left|P\left[Q\left(\hat{\phi}, R_0\right) > \triangle\right] - P\left[\chi^2\{(S - R_0)(N - R_0) - \dim(\beta)\} > \triangle\right]\right| \leq \eta,$$

which implies that

$$P(R > R_0) \leq P\left[Q\left(\hat{\phi}, R_0\right) > \triangle\right] \leq 2\eta$$

for $T > \max\left(\overline{T}, \tilde{T}\right)$ by construction since the second term in the display in the previous line is bounded at η. Since η is arbitrary and fixed, we have $\lim P(R > R_0) = 0$ and thus $\lim P\left(\hat{R} = R_0\right) = 1$. □

Proof of Proposition 3.8:

We follow Theorem 3 of Cragg and Donald (1997).

Start with the case when $R < R_0$:

$$P(R < R_0) = P[IC_T(R) \leq IC_T(R_0)]$$

Replacing $IC_T(R)$ and $IC_T(R_0)$ with their specifications and rearranging, we obtain:

$$P\left[\hat{Q}\left(\hat{\phi}|R\right) - \hat{Q}\left(\hat{\phi}|R_0\right) \leq -\theta(T)\{\pi(R) - \pi(R_0)\}\right].$$

Here, $\hat{Q}\left(\hat{\phi}|R\right) \to_p \infty$ and $\hat{Q}(\phi|R_0) \to_d \chi^2$, so that $\hat{Q}\left(\hat{\phi}|R\right) - \hat{Q}\left(\hat{\phi}|R_0\right) \to \infty$ as $T \to \infty$, but note that $T^{-1}\left\{\hat{Q}\left(\hat{\phi}|R\right) - \hat{Q}(\phi|R_0)\right\} = O_p(1)$ by the argument in the proof of Proposition 5. On the other hand, if $R_0 < R$ and $\theta(T) \to \infty$ as $T \to \infty$, we have that $T^{-1}\theta(T)\{\pi(R) - \pi(R_0)\} \to 0$, so that $P(R < R_0) = P(O_p(1) < 0)$ and thus $\lim P(R < R_0) = 0$. Note that this argument applies not only to BIC and HIC, but also AIC.

Next consider the case where $R > R_0$:

$$P(R > R_0) = P[IC_T(R_0) > IC_T(R)].$$

Similarly, expand and rearrange:

$$P\left[\hat{Q}(\phi|R_0) - \hat{Q}(\phi|R) \geq \theta(T)\{\pi(R) - \pi(R_0)\}\right],$$

then since $\hat{Q}\left(\hat{\phi}|R_0\right) \underset{p}{\to} Q_0(\phi_0|R_0) = O_p(1)$ uniformly under Proposition 2 and $\hat{Q}\left(\hat{\phi}|R\right) = O_p(1)$ by Theorem 1, we have that $Q(\phi|R_0) - Q\left(\underline{\phi}|R\right) = O_p(1)$. Furthermore, since π is strictly increasing in R, $\pi(R) - \pi(R_0) > 0$ so that $\theta(T)[\pi(R) - \pi(R_0)] \to \infty$ and thus $\lim P(R > R_0) = 0$. □

Proof Proposition 3.9:

To prove the proposition note first that $B_1\hat{G}^{\star}_{\Lambda}B_2$ consists of sample averages and therefore, by Theorem 5.20 of White (2001), a CLT will operate on:

$$\sqrt{T}\text{vec}\left(B_1\hat{G}^{\star}_{\Lambda}B_2\right) \to_d N\left[\left(B_2' \otimes B_1\right)\text{vec}(G_{\Lambda}), V_{G_{\Lambda}}\right],$$

since $E\left(B_1\hat{G}^{\star}_{\Lambda}B_2\right) = (B_1 G_{\Lambda} B_2)$. We will estimate the variance $\sqrt{T}\text{vec}\left(B_1\hat{G}^{\star}_{\Lambda}B_2\right)$ as:

$$\begin{aligned}\hat{V}_{G_{\Lambda}} = T^{-1}\sum_{t=1}^{T}\Big[&\left(B_2' \otimes B_1\right)\text{vec}\left(\hat{G}_{\Lambda,t}\right)\text{vec}\left(\hat{G}_{\Lambda,t}\right)'\left(B_2' \otimes B_1\right)'\Big] + \\ T^{-1}\sum_{p=1}^{P}\sum_{t=p+1}^{T}\Big[&\left(B_2' \otimes B_1\right)\text{vec}\left(\hat{G}^{\star}_{\Lambda,t}\right)\text{vec}\left(\hat{G}^{\star}_{\Lambda,t-p}\right)'\left(B_2' \otimes B_1\right)' \\ &+ \left(B_2' \otimes B_1\right)\text{vec}\left(\hat{G}^{\star}_{\Lambda,t-p}\right)\text{vec}\left(\hat{G}^{\star}_{\Lambda,t}\right)'\left(B_2' \otimes B_1\right)\Big].\end{aligned}$$

The first term is:

$$\begin{aligned}\left(B_2'\otimes B_1\right)T^{-1}\sum_{t=1}^{T}\left[\operatorname{vec}\left(\hat{G}_{\Lambda,t}\right)\operatorname{vec}\left(\hat{G}_{\Lambda,t}\right)'\right] =& T^{-1}\sum_{t=1}^{T}\operatorname{vec}\left(\bar{z}_t\hat{u}_{-,t}'\right)\operatorname{vec}\left(\bar{z}_t\hat{u}_{-,t}'\right)' = \\ & T^{-1}\sum_{t=1}^{T}\operatorname{vec}\left(\bar{z}_t u_{-,t}'\right)\operatorname{vec}\left(\bar{z}_t u_{-,t}'\right)' - T^{-1}\sum_{t=1}^{T}\operatorname{vec}\left(\bar{z}_t u_{-,t}'\right)\operatorname{vec}\left(\bar{z}_t\left(\hat{\beta}-\beta_0\right)'\mathbf{X}_{-,t}'\right)' \\ -T^{-1}\sum_{t=1}^{T}\operatorname{vec}\left(\bar{z}_t\left(\hat{\beta}-\beta_0\right)'\mathbf{X}_{-,t}'\right)\operatorname{vec}\left(\bar{z}_t u_{-,t}'\right)' + & T^{-1}\sum_{t=1}^{T}\operatorname{vec}\left(\bar{z}_t\left(\hat{\beta}-\beta_0\right)'\mathbf{X}_{-,t}'\right)\operatorname{vec}\left(\bar{z}_t\left(\hat{\beta}-\beta_0\right)'\mathbf{X}_{-,t}'\right)' \\ & = I + II + II' + III.\end{aligned}$$

Now I converges to a constant matrix by a weak law of large numbers for mixing variables in light of assumptions 1 and 2. For II and III, observe that

$$\operatorname{vec}\left(\bar{z}_t\left(\hat{\beta}-\beta_0\right)'\mathbf{X}_{-,t}'\right) = \left(\mathbf{X}_{-,t}\otimes\bar{z}_t\right)\left(\hat{\beta}-\beta_0\right)$$

and thus II:

$$\begin{aligned}\operatorname{vec}\left\{T^{-1}\sum_{t=1}^{T}\operatorname{vec}\left(\bar{z}_t u_{-,t}'\right)\left[\left(\mathbf{X}_{-,t}\otimes\bar{z}_t\right)\left(\hat{\beta}-\beta_0\right)\right]'\right\} = \\ T^{-1}\sum_{t=1}^{T}\left[\left(\mathbf{X}_{-,t}\otimes\bar{z}_t\right)\otimes\operatorname{vec}\left(\bar{z}_t u_{-,t}'\right)\right]\left(\hat{\beta}-\beta_0\right)\end{aligned}$$

converges to zero in probability since for each $l=1,\ldots,\dim(\beta)$, each $r=N-R^{\star}+1,\ldots,N$ and each $s=1,\ldots,S$, $E\left|x_{l,r,t}\bar{z}_{s,t}\right|$ and $E\left|u_{r,t}\bar{z}_{s,t}\right|$ are finite by Assumptions 1 and 2 and $\hat{\beta}$ is consistent. Similarly, for III we have:

$$\begin{aligned}\operatorname{vec}T^{-1}\sum_{t=1}^{T}\left(\bar{z}_t\mathbf{X}_{-,t}\otimes\right)\left(\hat{\beta}-\beta_0\right)\left(\hat{\beta}-\beta_0\right)'\left[\bar{z}_t\left(\hat{\beta}-\beta_0\right)'\mathbf{X}_{-,t}'\right]' = \\ T^{-1}\sum_{t=1}^{T}\left[\left(\mathbf{X}_{-,t}\otimes\bar{z}_t\right)'\otimes\left(\mathbf{X}_{-,t}\otimes\bar{z}_t\right)\right]\operatorname{vec}\left[\left(\hat{\beta}-\beta_0\right)\left(\hat{\beta}-\beta_0\right)'\right] = \\ T^{-1}\sum_{t=1}^{T}\left(\mathbf{X}_{-,t}'\mathbf{X}_{-,t}\otimes\bar{z}_t'\bar{z}_t\right)\operatorname{vec}\left[\left(\hat{\beta}-\beta_0\right)\left(\hat{\beta}-\beta_0\right)'\right] =,\end{aligned}$$

and III converges to zero in probability since for each $l=1,\ldots,\dim(\beta)$, each $r=N-R^{\star}+1,\ldots,N$ and each $s=1,\ldots,S$, $E\left|x_{l,r,t}\bar{z}_{s,t}\right|$ is finite. This shows that the first term in the variance formula $\hat{V}_{G_\Lambda}$ is consistent and by similar method the remaining P terms can be also shown to be consistent. The result is that $\hat{V}_{\hat{G}_\Lambda}\to_p V_{G_\Lambda}$.

Now note that consistency implies that $B_1\hat{G}_{\Lambda}^{\star}B_2 - B_1G_\Lambda B_2\to_p 0$ and that $B_1\hat{G}_{\Lambda}^{\star}B_2 - B_1G_\Lambda B_2 = O_p\left(T^{-1/2}\right)$ by the above. Furthermore, since $\operatorname{rank}\left(B_1G_\Lambda B_2\right)=R_0$ but the column dimension is $R^{\star}$, there exist decom-

positions of $B_1 G_\Lambda B_2$ and $B_1 \hat{G}^\star_\Lambda B_2$ such that:

$$\Xi_1^{(R_0)} B_1 G_\Lambda B_2 \Xi_2^{(R_0)} = 0_{(S-R_0)\times(R^\star-R_0)},$$
$$\hat{\Xi}_1^{(R_0)} B_1 \hat{G}^\star_\Lambda B_2 \hat{\Xi}_2^{(R_0)} = 0_{(S-R_0)\times(R^\star-R_0)}$$

and by Theorem 13.5.1 of Gohberg et al (2006):

$$\max\left[\left\|\hat{\Xi}_1^{(R_0)} - \Xi_1^{(R_0)}\right\|, \left\|\hat{\Xi}_2^{(R_0)} - \Xi_2^{(R_0)}\right\|\right] \leq \triangle \left\|B_1 \hat{G}^\star_\Lambda B_2 - B_1 G_\Lambda B_2\right\|,$$

for a constant $\triangle$ that depends only on $B_1 \hat{G}^\star_\Lambda B_2$. Now since $B_1 \hat{G}^\star_\Lambda B_2 - B_1 G_\Lambda B_2 \to_p 0$, the above inequality implies that the null space estimators at R_0 are consistent and thus that:

$$\hat{\Xi}_1^{(R_0)} B_1 \hat{G}^\star_\Lambda B_2 \hat{\Xi}_2^{(R_0)} \to_p \Xi_1^{(R_0)} B_1 G_\Lambda B_2 \Xi_2^{(R_0)} = 0_{(S-R_0)\times(R^\star-R_0)}.$$

This further implies that the variance of $\text{vec}\left(\hat{\Xi}_1^{(R_0)} B_1 \hat{G}^\star_\Lambda B_2 \hat{\Xi}_2^{(R_0)}\right)$ at $R = R_0$,

$$\left(\hat{\Xi}_2^{(R_0)\prime} \otimes \hat{\Xi}_1^{(R_0)}\right) \hat{V}_{G_\Lambda} \left(\hat{\Xi}_2^{(R_0)} \otimes \hat{\Xi}_1^{(R_0)\prime}\right) \to_p \left(\Xi_2^{(R_0)\prime} \otimes \Xi_1^{(R_0)}\right) V_{G_\Lambda} \left(\Xi_2^{(R_0)} \otimes \Xi_1^{(R_0)\prime}\right),$$

is of full rank and its smallest eigenvalue is bounded away from zero in probability. As a result, the distributional results under H_0 in items (1) and (2) of Proposition 3.9 hold.

To show that RK diverges under H_1 we have to show that:

$$\hat{\Xi}_1^{(R)} B_1 \hat{G}^\star_\Lambda B_2 \hat{\Xi}_2^{(R)} \to_p \triangle$$

for some $\triangle > 0$ and that the smallest eigenvalue of:

$$\left(\hat{\Xi}_2^{(R)\prime} \otimes \hat{\Xi}_1^{(R)}\right) \hat{V}_{G_\Lambda} \left(\hat{\Xi}_2^{(R)} \otimes \hat{\Xi}_1^{(R)\prime}\right)$$

is bounded away from zero in probability so that the inverse is continuous. For the first requirement note that since $\text{rank}\,(B_1 G_\Lambda B_2) = R_0$ it has R_0 non-vanishing singular values and the matrix can be factored by SVD as:

$$B_1 G_\Lambda B_2 = \begin{bmatrix} \Xi_{1,\perp}^{(R_0-j)\prime}, \Xi_{1,\perp}^{\prime(j)}, & \Xi_1' \end{bmatrix} \begin{bmatrix} \Theta & \\ & 0 \end{bmatrix} \begin{bmatrix} \Xi_{2,\perp}^{(R_0-j)\prime} \\ \Xi_{2,\perp}^{(j)\prime} \\ \Xi_2' \end{bmatrix},$$

where Θ is an $R_0 \times R_0$ diagonal matrix consisting of the singular values of $B_1 G_\Lambda B_2$ and the rows of $\Xi_{(\cdot)}$ are unitary and therefore orthogonal. Now setting $\hat{\Xi}_1^{(R_0-j)} = \left[\hat{\Xi}_{1,\perp}^{\prime(j)}, \hat{\Xi}_1'\right]'$ and $\hat{\Xi}_2^{(R_0-j)} = \left[\hat{\Xi}_{2,\perp}^{\prime(j)}, \hat{\Xi}_2'\right]'$ we have:

$$\hat{\Xi}_1^{(R_0-j)} \left(B_1 \hat{G}^\star_\Lambda B_2\right) \hat{\Xi}_2^{(R_0-j)} \to_p \sum_{i=1}^{j} \Xi_{1,\perp}^{\prime(i)} \theta_i \left(B_1 G_\Lambda B_2\right) \Xi_{2,\perp}^{\prime(i)} = O_p(1)$$

since $B_1\hat{G}^{\star}_{\Lambda}B_2$ is consistent with R_0 singular values θ_i. For the second requirement, note that:

$$\begin{aligned}\left(\hat{\Xi}_2^{(R)\prime}\otimes\hat{\Xi}_1^{(R)}\right)\text{vec}\left(\bar{z}_{j,t}u'_{-,t}\right) &= \left(\hat{\Xi}_2^{(R)\prime}\otimes\hat{\Xi}_1^{(R)}\right)\text{vec}\left(\overline{Z}_t f_t\Lambda'_-\right)+\left(\hat{\Xi}_2^{(R)\prime}\otimes\hat{\Xi}_1^{(R)}\right)\text{vec}\left(\bar{z}_t\varepsilon'_{-,t}\right)\\ &= \left(\hat{\Xi}_2^{(R)\prime}\otimes\hat{\Xi}_1^{(R)}\right)\left(\Lambda_-\otimes I_S\right)\text{vec}\left(\bar{z}_t f_t\right)+\left(\hat{\Xi}_2^{(R)\prime}\otimes\hat{\Xi}_1^{(R)}\right)\text{vec}\left(\bar{z}_t\varepsilon'_{-,t}\right)\\ &= \begin{bmatrix}\left(\hat{\Xi}_2^{(R)\prime}\Lambda_-\otimes\hat{\Xi}_1^{(R)}\right) & 0\\ 0 & \left(\hat{\Xi}_2^{(R)\prime}\otimes\hat{\Xi}_1^{(R)}\right)\end{bmatrix}\text{vec}\left[\bar{z}_t f_t,\bar{z}_t\varepsilon'_{-,t}\right].\end{aligned}$$

Note that the $S(R_0+R^{\star})$-vector $\text{vec}\left[\bar{z}_t f_t,\bar{z}_t\varepsilon'_{-,t}\right]$ is non-degenerate normally distributed and that the block matrix pre-multiplying $\text{vec}\left[\bar{z}_t f_t,\bar{z}_t\varepsilon'_{-,t}\right]$ is an affine transformation if it is of full rank smaller or equal than $S(R+R^{\star})$. Using the facts that for a block diagonal matrix $A=\text{diag}(A_{11},A_{22})$, $\text{rank}(A)=\text{rank}(A_{11})+\text{rank}(A_{22})$; that $\text{rank}(A_{11}\otimes A_{22})=\text{rank}(A_{11})\,\text{rank}(A_{22})$ and that orthogonal matrices are full rank, we have that the $(S-R)(R^{\star}-R)\times SR^{\star}$ matrix $\left(\hat{\Xi}_2^{(R)\prime}\otimes\hat{\Xi}_1^{(R)}\right)$ in the second block is full rank. For the first block, note that $\text{rank}\left(\hat{\Xi}_2^{(R)\prime}\Lambda_-\right)=\min(R^{\star}-R,R_0)$, but that the dimension of $\left(\hat{\Xi}_2^{(R)\prime}\Lambda_-\otimes\hat{\Xi}_1^{(R)}\right)$ is $(R^{\star}-R)(S-R)\times R_0S$. This means that the matrix pre-multiplying $\text{vec}\left[\bar{z}_t f_t,\bar{z}_t\varepsilon'_{-,t}\right]$ can only be an affine transformation if $(R^{\star}-R)\leq R_0$. In that case,

$$\left(\hat{\Xi}_2^{(R)\prime}\otimes\hat{\Xi}_1^{(R)}\right)V_{G_{\Lambda}}\left(\hat{\Xi}_2^{(R)\prime}\otimes\hat{\Xi}_1^{(R)}\right)$$

is symmetric positive definite and thus invertible and since an invertible matrix has all eigenvalues positive, it implies that its inverse must be $O_p(1)$. Now since the numerator and denominator are bounded, multiplying the quadratic form of the Rank Test at $\hat{R}<R_0$ by T implies it diverges. □

Appendix 3.2. Simulation Results

Appendix 3.2.1. Monte Carlo Results: Bias and RMSE of QDGMM

Table 3.1: Monte Carlo results for model (3.24) with $R = 1$ and $\beta_0 = 0.5$, mean and RMSE over 5000 replications:

N	T[1]		*OLS*	*FE*	*CCE*	*QD 2-step*	*QD 1-step*
2	50	mean	0.696	0.983	0.506	0.488	0.479
		RMSE	0.220	0.514	0.138	0.053	0.026
	100	mean	0.700	0.991	0.524	0.504	0.498
		RMSE	0.217	0.510	0.107	0.034	0.034
	200	mean	0.707	0.995	0.531	0.507	0.504
		RMSE	0.220	0.505	0.089	0.060	0.060
	400	mean	0.710	0.998	0.535	0.508	0.506
		RMSE	0.221	0.503	0.079	0.028	0.028
N	T		*OLS*	*FE*	*CCE*	*QD 2-step*	*QD 1-step*
5	50	mean	0.718	0.990	0.495	0.530	0.530
		RMSE	0.232	0.503	0.077	0.028	0.043
	100	mean	0.724	0.997	0.510	0.529	0.508
		RMSE	0.232	0.504	0.062	0.034	0.040
	200	mean	0.728	0.999	0.515	0.514	0.501
		RMSE	0.233	0.502	0.053	0.027	0.008
	400	mean	0.729	0.998	0.518	0.507	0.500
		RMSE	0.233	0.500	0.049	0.038	0.035
N	T		*OLS*	*FE*	*CCE*	*QD 2-step*	*QD 1-step*
10	50	mean	0.723	0.992	0.485	0.539	0.500
		RMSE	0.235	0.502	0.051	0.039	0.052
	100	mean	0.731	0.997	0.495	0.531	0.501
		RMSE	0.237	0.502	0.035	0.056	0.126
	200	mean	0.735	0.998	0.501	0.519	0.501
		RMSE	0.239	0.501	0.026	0.021	0.003
	400	mean	0.737	0.999	0.503	0.507	0.500
		RMSE	0.240	0.500	0.021	0.006	0.008

[1] OLS is pooled OLS, FE is the Fixed Effects Estimator, CCE is the Common Correlations Estimator and finally QD 1- and 2-step are QDGMM estimators.

Table 3.2: Monte Carlo results for model (3.24) with $R = 2$ and $\beta_0 = 0.5$, mean and RMSE over 5000 replications.

N	T[2]		*OLS*	*FE*	*CCE*	*QD 2-step**	*QD 1-step**	*QD 2-step*	*QD 1-step*
3	50	mean	0.812	1.011	0.668	0.589	0.582	0.469	0.456
		RMSE	0.328	0.511	0.231	0.144	0.140	0.136	0.136
	100	mean	0.817	1.002	0.686	0.610	0.600	0.498	0.490
		RMSE	0.331	0.502	0.234	0.145	0.137	0.029	0.029
	200	mean	0.825	0.996	0.696	0.619	0.608	0.509	0.504
		RMSE	0.336	0.496	0.238	0.145	0.135	0.019	0.020
	400	mean	0.823	0.999	0.699	0.621	0.612	0.516	0.512
		RMSE	0.334	0.499	0.240	0.146	0.136	0.039	0.039

N	T		*OLS*	*FE*	*CCE*	*QD 2-step**	*QD 1-step**	*QD 2-step*	*QD 1-step*
5	50	mean	0.819	0.969	0.698	0.607	0.591	0.510	0.481
		RMSE	0.333	0.469	0.244	0.147	0.134	0.088	0.084
	100	mean	0.824	0.993	0.711	0.626	0.607	0.521	0.500
		RMSE	0.336	0.493	0.249	0.153	0.134	0.097	0.096
	200	mean	0.828	1.012	0.717	0.633	0.612	0.515	0.502
		RMSE	0.337	0.512	0.252	0.155	0.134	0.011	0.035
	400	mean	0.828	0.999	0.717	0.635	0.611	0.509	0.501
		RMSE	0.337	0.499	0.251	0.155	0.132	0.006	0.007

N	T		*OLS*	*FE*	*CCE*	*QD 2-step**	*QD 1-step**	*QD 2-step*	*QD 1-step*
10	50	mean	0.723	0.827	0.973	0.718	0.617	0.518	0.476
		RMSE	0.235	0.339	0.473	0.258	0.152	0.109	0.030
	100	mean	0.731	0.829	1.000	0.726	0.635	0.524	0.492
		RMSE	0.237	0.339	0.500	0.261	0.158	0.061	0.017
	200	mean	0.735	0.829	0.998	0.733	0.642	0.516	0.499
		RMSE	0.239	0.339	0.498	0.266	0.161	0.019	0.018
	400	mean	0.737	0.829	1.004	0.734	0.644	0.506	0.498
		RMSE	0.240	0.337	0.504	0.266	0.162	0.013	0.013

[2]OLS corresponds to pooled OLS, FE is the Fixed Effects Estimator, CCE is the Common Correlations Estimator, and finally QD 1- and 2-step are QDGMM estimators, where (*) denotes estimation with $R = 1$.

Appendix 3.2.2 Monte Carlo Results: Wald-Tests for Time Effects Model

Table 3.6: Wald-Test for Poolability of the factor loadings, acceptance rates for true null of Time Effects model, 5000 replications of model (3.24):

Wald-Statistic		*Acceptance rates*[1]					*Acceptance rates*		
N	*T*	10%	5%	1%	*N*	*T*	10%	5%	1%
3	50	95.4	97.8	99.5	5	50	96.1	97.9	99.4
	100	95.8	98.3	99.8		100	97.5	98.9	99.7
	200	95.5	98.5	99.9		200	97.3	99.1	99.9
	400	94.6	97.8	99.7		400	96.4	98.4	99.8

		Acceptance rates		
N	*T*	10%	5%	1%
10	50	93.1	95.4	97.9
	100	97.5	98.7	99.7
	200	98.4	99.4	99.9
	400	98.2	99.3	99.9

Table 3.7: Acceptance rates for Wald-Tests about the slope parameter β_0 where $R_0 = 1$ but $R = 2$, 5000 replications of model (3.24):

Wald-Statistic acceptance rates[2]		$\overline{\beta} = 0$			$\overline{\beta} = 0.5$			
N	*T*	10%	5%	1%	10%	5%	1%	mean
5	50	4.21	6.92	15.12	82.5	89.0	95.9	0.468
	100	0.7	1.2	2.4	82.1	87.6	95.6	0.493
	200	0.3	0.5	0.8	88.8	89.3	95.1	0.499
	400	0.3	0.3	0.6	85.0	90.1	95.7	0.502

[1] Acceptance rates using critical values from the chi-squared distribution with $N-1$ degrees of freedom.

[2] Acceptance rates using critical values from the chi-squared distribution with 1 degree of freedom.

Appendix 3.2.3 Monte Carlo Results: Heterogeneous Parameter Model

Table 3.8: Monte Carlo results for heterogeneous specification of parameter model (3.24) with $R = 1$ and $\beta_i = 0.5$, 5000 replications:

N	T		β_1[1]	β_{middle}	β_{end}	N	T		β_1	β_{middle}	β_{end}
2	50	*mean*	0.464	0.464	0.463	5	50	*mean*	0.479	0.484	0.480
		RMSE	1.146	1.146	2.251			*RMSE*	0.644	0.470	0.122
	100	*mean*	0.505	0.505	0.442		100	*mean*	0.476	0.502	0.490
		RMSE	2.330	2.330	1.790			*RMSE*	1.851	2.229	0.080
	200	*mean*	0.536	0.536	1.353		200	mean	0.510	0.480	0.494
		RMSE	1.845	1.845	58.263			*RMSE*	0.384	1.770	0.053
	400	*mean*	0.533	0.533	0.487		400	*mean*	0.509	0.513	0.498
		RMSE	3.134	3.134	0.386			*RMSE*	0.578	0.546	0.037

N	T		β_1	β_{middle}	β_{end}
10	50	*mean*	0.499	0.502	0.491
		RMSE	0.526	0.585	0.137
	100	*mean*	0.501	0.510	0.496
		RMSE	0.624	0.433	0.080
	200	*mean*	0.520	0.515	0.498
		RMSE	0.512	0.319	0.050
	400	*mean*	0.505	0.506	0.499
		RMSE	0.248	0.231	0.035

Table 3.9: Wald-Test for Poolability, acceptance rates for true null hypothesis of parameter homogeneity $\beta_i = \overline{\beta}$, 5000 replications:

Wald-Statistic		*Acceptance rates*[2]					*Acceptance rates*		
N	T	10%	5%	1%	N	T	10%	5%	1%
2	50	90.9	95.7	99.5	5	50	94.1	97.6	99.8
	100	90.4	95.6	99.3		100	92.7	96.7	99.6
	200	90.9	95.7	99.1		200	91.1	96.0	99.4
	400	90.4	95.4	99.2		400	90.4	95.5	99.2

		Acceptance rates		
N	T	10%	5%	1%
10	50	98.5	99.5	100.0
	100	95.4	98.4	99.9
	200	93.6	97.2	99.6
	400	91.6	96.0	99.3

[1] $\beta_{(\cdot)}$ computed by QDGMM.

[2] Acceptance rates using critical values from the chi-squared distribution with $N-1$ degrees of freedom.

Appendix 3.2.4 Monte Carlo Results: Determination of R_0

Table 3.10: Determination of $R_0 = 2$ by information criteria and sequential test procedures. Reported are acceptance rates at $R = 0, \ldots, 4$ based on 5000 Monte Carlo replications and in all cases $N = 5$.

		One-step					*Two-step*				
	$T/\hat{R}$	0	1	2	3	4	0	1	2	3	4
AIC	50	4.4	24.5	47.9	19.7	3.5	14.1	63.4	18.9	3.4	0.2
	100	2.7	17.2	51.5	24.4	4.1	1.9	50.7	37.3	9.4	0.8
	200	0.8	9.2	56.2	28.4	5.4	0.0	16.8	58.3	22.8	2.1
	400	0.0	3.4	57.1	32.6	6.9	0.0	2.3	64.2	29.7	3.8
BIC	50	2.0	17.6	47.0	26.9	6.4	2.9	35.7	36.1	21.9	3.4
	100	2.1	15.5	51.0	26.7	4.7	0.5	33.9	43.1	19.6	2.8
	200	1.1	10.2	56.9	26.9	4.9	0.0	15.4	57.9	24.3	2.3
	400	0.2	4.8	61.0	28.7	5.3	0.0	3.7	68.8	24.9	2.6
HIC	50	8.9	32.6	44.3	12.3	1.9	39.5	57.3	3.2	0.1	0.0
	100	9.1	27.1	49.0	13.6	1.2	13.0	76.7	10.1	0.2	0.0
	200	6.2	20.4	56.6	15.5	1.4	0.7	60.7	36.8	1.7	0.0
	400	2.2	9.7	64.8	20.8	2.4	0.0	16.8	73.1	9.7	0.3

		$\alpha_T = 4\,(N/T)$				$\alpha_T = 0.5\sqrt{N/T}$			
	$T/\hat{R}$	0	1	2	3	0	1	2	3
J-Test	50	7.7	48.1	24.4	13.0	23.6	63.8	9.7	2.3
	100	3.0	57.8	28.8	8.0	5.3	70.4	20.2	3.5
	200	0.1	35.9	52.5	10.2	0.1	40.6	49.8	8.4
	400	0.0	8.6	75.3	14.9	0.0	8.1	75.3	15.3
RK-Test	50	19.0	24.0	25.5	23.2	39.4	29.4	20.7	8.6
	100	19.2	27.6	35.2	15.1	26.3	31.6	33.0	8.0
	200	10.6	28.3	47.8	11.9	11.5	30.5	47.3	9.5
	400	5.1	23.9	61.4	8.3	4.9	23.4	61.3	9.0

Bibliography

[1] Ahn, C., Lee, Y. H. and Schmidt, P. (2001): "GMM Estimation of Linear Panel Data Models with Time-varying Individual Effects," *Journal of Econometrics*, Vol. 101 Issue 2, pp. 219-255, Elsevier North Holland.

[2] Ahn, C., Lee, Y. H. and Schmidt, P. (2013): "Panel Data Models with Multiple Time-Varying Individual Effects," *Journal of Econometrics*, Vol. 174, pp. 1-14, Elsevier North Holland.

[3] Al-Sadoon, M. (2017): "A Unifying Theory of Tests of Rank," *Journal of Econometrics*, Vol. 199 Issue 1, pp. 49-62, , Elsevier North Holland.

[4] Anderson, T. W., and Hsiao C. (1981): "Estimation of Dynamic Models with Error Components," *Journal of the American Statistical Association*, Vol. 76, pp. 598–606, American Statistical Association.

[5] Andrews, D. W. K. (1992): "Generic Uniform Convergence," *Econometric Theory*, Vol. 8, No. 2, pp. 241-257, Cambridge University Press, Cambridge UK.

[6] Andrews, D. W. K. (1997): "A Stopping Rule for the Computation of Generalized Method of Moments Estimators," Econometrica, Vol. 65, No. 4, pp. 913-931, Wiley-Blackwell.

[7] Bai, J. (2009): "Panel Data Models with Interactive Fixed Effects," *Econometrica*, Vol. 77, Issue 4, pp. 1229-1279, Econometric Society, Wiley-Blackwell.

[8] Bai, J. and Ng, S. (2002): "Determining the Number of Factors in Approximate Factor Models," *Econometrica*, Vol. 70, Issue i, pp. 191-221, Econometric Society, Wiley-Blackwell.

[9] Bai, J. and Ng, S. (2013): "Principal Components Estimation and Identification of Static Factors," *Journal of Econometrics*, Vol. 176, Issue 1, pp. 18-29, Elsevier North Holland.

[10] Beetsma, R. and Giuliodori, M. (2011): "The Effects of Government Purchases Shocks: Review and Estimates for the EU," *Economic Journal*, Vol. 121 Issue 550, pp. F4-F32, Royal Economic Society, Wiley-Blackwell.

[11] Bernanke, B., Boivin, J. and Eliasz, P. (2005): "Factor Augmented Vector Autoregression and the Analysis of Monetary Policy," *Quarterly Journal of Economics*, Vol. 120, pp. 387-422, Oxford University Press.

[12] Bun, M. J. G. (2004): "Testing Poolability in a System of Dynamic Regressions with Non-Spherical Disturbances," *Empirical Economics*, Vol. 29 Issue (1), pp. 89-106, Springer.

[13] Caner, M. (2008): "Nearly-Singular Design in GMM and Generalized Empirical Likelihood Estimators," *Journal of Econometrics*, Vol. 144, pp. 511-523, Elsevier North Holland.

[14] Cragg, J. G. and Donald, S. (1996): "On the Asymptotic Properties of LDU-based Tests of the Rank of a Matrix," *Journal of the American Statistical Association*, Vol. 91 Issue 435, pp. 1301-1309, American Statistical Association.

[15] Cragg, J. G. and Donald, S. (1997): "Inferring the Rank of a Matrix," *Journal of Econometrics*, Vol. 76, Issue 1-2, pp. 223-250, Elsevier North Holland.

[16] Engel, C., Nelson, M. and West, K. (2014): "Factor Model Forecasts of Exchange Rates," *Econometric Reviews*, Vol. 0 Issue 0, pp. 1–24, Taylor & Francis Group.

[17] Foerster, A., Sarte, P.-D. and Watson, M. (2011): “Sectoral versus Aggregate Shocks: a Structural Factor Analysis of Industrial Production,” *Journal of Political Economy*, Vol. 119 No. 1, pp 1-38, University of Chicago Press.

[18] Froot, K. and Rogoff, K., (1995): “Perspectives on PPP and Long-Run Real Exchange Rates, in: *Handbook of International Economics*, Vol. 3, edited by: Rogoff, K., and Grossman, G., Elsevier North Holland.

[19] Giersbergen, N. and Kiviet, J. (2002): "How to Implement the Bootstrap in Static or Stable Dynamic Regression Models: Test Statistic versus Confidence Region Approach," *Journal of Econometrics*, Vol. 108 Issue (1), pp. 133-156, Elsevier North Holland.

[20] Gohberg, I., Lancaster, P. and Rodman, L. (1986) “Invariant Subspaces of Matrices with Applications,” Wiley and Sons Inc., New York.

[21] Gorski, J., Pfeuffer, F. and Klamroth, K. (2007): “Biconvex Sets and optimization with Biconvex Functions: A Survey and Extensions,” *Mathematical Methods of Operational Research*, Vol. 66, pp. 373-407, Springer.

[22] Grippo, L. and Sciandrone, M. (2000): "On the Convergence of the Block Non-linear Gauss-Seidel Method under Convex Constraints," *Operations Research Letters*, Vol. 26, Issue 3, pp. 127-136, Elsevier North Holland.

[23] De Groote T. and Everaert, G. (2016): "Common Correlated Effects Estimation of Dynamic Panels with Cross-Sectional Dependence," *Econometric Reviews*, Vol. 35, pp. 428-463, Taylor and Francis.

[24] Hansen, L. P. (1982): "Large Sample Properties of GMM Estimators," *Econometrica*, Vol. 50 (4), pp. 1029-1054, Econometric Society, Wiley-Blackwell.

[25] Hansen, L. P., Heaton, J. and Yaron, A. (1996): "Finite-Sample Properties of Some Alternative GMM Estimators," *Journal of Business and Economic Statistics*, Vol. 14, Issue 3, pp 262-280, American Statistical Association.

[26] Hayakawa, K. (2016): "Identification Problem of GMM Estimators for Short Panel Data Models with Interactive Fixed Effects," *Economics Letters*, Vol. 139, pp. 22-26, Elsevier North Holland.

[27] Hayashi, F. (2000): "Econometrics," Princeton University Press, Princeton US.

[28] Holtz-Eakin, D., Newey, W. and H. S. Rosen (1988): "Estimating Vector Autoregressions with Panel data," *Econometrica*, Vol. 56 Issue 6, pp. 1371–1395, Econometric Society, Wiley-Blackwell.

[29] Ilzetzki, E. (2011): "Fiscal Policy and Debt dynamics in Developing Countries," *Policy Research Working Paper Series* 5666, The World Bank.

[30] Ilzetzki, E., Mendoza, E. G. and Végh, C. A. (2013): "How Big (Small?) are Fiscal Multipliers?," *Journal of Monetary Economics*, Vol. 60 Issue 2, pp. 239-254, Elsevier North Holland.

[31] Jöreskog, K. G. and Goldberger, A. *Z*. (1975). "Estimation of a Model with Multiple Indicators and Multiple Causes of a Single Latent Variable". *Journal of the American Statistical Association*, Vol. 70 (351), pp. 631–639, American Statistical Association.

[32] Kleibergen, F. and Paap, R. (2006): "Generalized Reduced Rank Tests Using the Singular Value Decomposition," *Journal of Econometrics*, Vol. 133 Issue 1, pp. 97-126, Elsevier North Holland.

[33] Mankiw, N. Romer, D. and Weil, D. (1992): "A Contribution to the Empirics of Economic Growth, *Quarterly Journal of Economics*, Vol. 107, pp. 407-437, Oxford University Press.

[34] Moon, H. R. and Weidner, M. (2015): "Linear Regression for Panel with Unknown Number of Factors as Interactive Fixed Effects," *Econometrica*, Vol. 83, Issue 4, pp. 1543-1579, Econometric Society, Wiley-Blackwell.

[35] Moon, H. R. and Weidner, M. (2017): "Dynamic Linear Panel Regression Models with Interactive Fixed Effects," *Econometric Theory*, Vol. 33, Issue 1, pp. 158-195, Cambridge University Press.

[36] Nauges, C. and Thomas, A. (2003): "Consistent Estimation of Dynamic Panel Data Models with Time-varying Individual Effects," *Annales d'Economie et de Statistique*, Issue 70, pp. 53-75, ENSAE.

[37] Newey W.K. and McFadden, D. (1994): "Large Sample Estimation and Hypothesis Testing", in: *Handbook of Econometrics*, Vol. 4, edited by Robert F. Engle and Daniel L. McFadden, Elsevier North Holland.

[38] Newey, W. K. and West, K. (1987): "A Simple Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix," *Econometrica*, Vol. 55, pp. 703-708, Wiley-Blackwell.

[39] Pesaran M. H. (2006): "Estimation and Inference in Large Heterogeneous Panels with a Multifactor Error Structure," *Econometrica*, Vol. 74, pp. 967-1012, Econometric Society, Wiley-Blackwell.

[40] Pötscher, B. M. (1983): "Order Estimation in ARMA-Models by Lagrangian Multiplier Tests," *Annals of Statistics*, Vol. 11, No. 3, pp. 872-885, Institute of Mathematical Statistics.

[41] Robertson, D. and Sarafidis, V. (2015): "IV Estimation of Panels with Factor Residuals," *Journal of Econometrics*, Vol. 185, pp. 526-541, Elsevier North Holland.

[42] Roll, R. and Ross, S (1980): "An Empirical Investigation of the Arbitrage Pricing Theory," *Journal of Finance*, Vol. 35 (5), pp. 1073–1103, Wiley-Blackwell.

[43] Ross, S. (1976): "The Arbitrage Theory of Capital Asset Pricing," *Journal of Economic Theory*, Vol. 13 (3), pp. 341–360, Elsevier North Holland.

[44] Sargent, T. and Sims, C. (1977): “Business Cycle Modeling Without Pretending to Have Too Much A-Priori Economic Theory,” in: *New Methods in Business Cycle Research*, edited by Christopher Sims, Federal Reserve Bank of Minneapolis, Minneapolis.

[45] Urbain, J. and Westerlund, J. (2013): "On the Estimation and Inference in Factor-Augmented Panel Regressions with Correlated Loadings," *Economics Letters*, Vol. 119, Issue 3, pp 247–250, Elsevier North Holland.

[46] Westerlund, J. and Urbain, J. (2015): “Cross-Sectional Averages Versus Principal Components,” *Journal of Econometrics*, Vol. 185, pp. 372-377, Elsevier North Holland.

[47] White, H. (2001): "Asymptotic Theory for Econometricians," Academic Press, Orlando US.

[48] Zellner, A. (1962): "An Efficient Method of Estimating Seemingly Unrelated Regressions and Tests for Aggregation Bias," *Journal of the American Statistical Association*, Vol. 57, No. 298, pp. 348-368, American Statistical Association.

Chapter 4

Determination of the Number of Factors in Cross-Sectionally Dependent Data: Redemption for the Scree Plot

This paper presents a unifying framework for the determination of the number of factors in large dimensional factor models. We show that consistent estimation of the number of factors follows from separation of the eigenvalues of the data covariance matrix into a portion corresponding to the factor structure and a portion corresponding to the residual noise, formally justifying the Scree plot. We further show that eigenvalue separation continues to hold for the estimated error covariance matrix of regression models with interactive fixed effects and present several new consistent estimators of the number of factors in approximate factor models.

4.1 Introduction

Factor models have become a staple of applied economics: in financial economics, factors are used to model the exposure of security portfolios to the risk characteristics of the market (Chamberlain and Rothschild, 1983). In macroeconomics, factor models are utilized to reduce the curse of dimensionality in large datasets and are found to provide superior time series forecasts, see for example Bernanke et al (2005) or Engel et al (2014). Further examples of applications of factor models beyond economics are in genome analysis (Price et al, 2006) and by participants in the so-called Netflix Challenge.[1] Extensive statistical theory for the estimation of pure, approximate and dynamic factor models in addition to regression models with interactive fixed effects has been developed, see for example Bai (2009), Bai and Ng (2002, 2008). Typically, consistent estimation of the parameters of a factor model is conditional on knowledge of the number of factors in the

[1] In this public contest, the online video-streaming service Netflix tasked computer scientists with the objective of finding an efficient way to recommend video content to subscribers based on the ratings awarded to other videos by themselves and peer subscribers, see for example Feuerverger et al (2012).

model and for that reason, determination of the correct number of factors is still an active area of research. This is motivated further by the observation that popular methodologies based on information criteria as in Bai and Ng (2002) have poor finite sample properties in relevant data configurations, see for example Hallin and Liška (2007), Onatski (2010) and Ahn and Horenstein (2013) (AH). Onatski (2010) and AH develop alternative estimators that are considerably more successful at determining the correct number of factors. These methodologies model the error covariance matrix of a factor model as a sequence of large random matrices and inference is based on the eigenvalue structure of such matrices, making use of results from a field of mathematical statistics known as Random Matrix Theory (RMT).

RMT studies the properties of large random matrices and the seminal contribution is Marchenko and Pastur (1967) (MP). In that paper, the distribution of the eigenvalues of a large random matrix with independent and identically distributed mean-zero entries is characterised and since then, much work has been done on extending that result.[2] Relevant extensions for econometricians are matrices with dependent entries as studied in Bai and Zhou (2008), Pfaffel and Schlemm (2012), Liu et al (2015); whereas distributional results for autocovariance matrices of data ensembles with dependent entries are given in Li et al (2014) and Wang et al (2015). Important for the determination of the number of factors in a factor model are the extreme values of the support of the MP distribution: Yin et al (1988) and Bai et al (1988) derive the limit of the largest eigenvalue of the MP distribution, whilst Bai and Yin (1993) study the smallest eigenvalue. Another important result is the phenomenon of eigenvalue separation in Bai and Silverstein (1999) and Paul and Silverstein (2008): these authors show that if data is generated by some distribution, then the probability that the eigenvalues of the sample covariance matrix deviate from the eigenvalues of the population covariance matrix tends to zero with the sample size. The combination of these results is crucial for consistent estimation of the number of factors: if (the error term of) a statistical model consists of a factor and a noise component, then knowledge of the support of the eigenvalue distribution of the covariance matrix of the latter can be used to separate eigenvalues related to the factors from the noise eigenvalues estimated from the overall model covariance matrix. This argument is used by Onatski (2010) and AH in the econometrics literature. In the mathematical statistics literature, eigenvalue separation is applied to static factor models by Lam and Yao (2012) and to determine the lag order of dynamic factor models in Li et al (2014).

In this paper we use the eigenvalue separation result to derive new estimators of the number of factors in large dimensional factor models and these estimators are shown to have better small-sample properties than existing methodologies based on information criteria. In doing so, we make an argument for redemption of the Scree plot: we show that any estimator based on eigenvalue separation is a numerical implementation of the Scree plot. This argument also allows us to place the so-called "tuned" information criteria of Hallin and Liška (2007) and Alessi et al (2014) in the class of Scree plot estimators. We then show that the separation result holds in models where the number of factors goes to infinity along with the sample size and in so-called weak factor models with appropriate scaling. Importantly, we also derive a version of eigenvalue separation for the covariance matrix calculated from the estimated regression residuals of an interactive

[2]The interested reader is referred to the book of Bai and Silverstein (2010).

fixed effects model, which holds regardless of the consistency of the slope parameter estimator.

The paper is structured as follows: Section 4.2 discusses the basic approximate factor model and the assumptions we impose. Section 4.3 presents the main eigenvalue separation theorem and presents several extensions as corollaries. Section 4.4 presents several new estimators for the number of factors in approximate factor models that follow from eigenvalue separation, whilst Section 4.5 extends the eigenvalue separation result to the interactive fixed effects model. Section 4.6 explores the small-sample properties whilst Section 4.7 concludes. All proofs are in Appendix 4.1. Finally, on notation: a column vector a is written in small script, whereas an $m \times n$ matrix A is in capital. We write $A > 0$ to designate that A is positive definite, i.e. that all eigenvalues of A are positive, and $||A||$ is the Frobenius norm of A, i.e., $||A|| = \sqrt{\text{tr}\,(A'A)}$, or the corresponding Euclidean norm of a vector a. Finally, " $\to$ " is the limit of a non-stochastic sequence and " $\to_p$ " denotes convergence in probability.

4.2 Basic Model and Assumptions

Consider the following approximate factor model:

$$y_{i,t} = \Lambda_i F_t + \varepsilon_{i,t}; i = 1, \ldots, N, t = 1, \ldots, T,$$

where $y_{i,t}$ is the dependent variable, $F_t = [f_{1,t}, \ldots, f_{R,t}]'$ consists of R factors that can be both dynamic and static and the $\Lambda_i = [\lambda_{1,i}, \ldots, \lambda_{R,i}]$ are the factor loadings of the i-th individual. Neither the factors nor the loadings can be observed and $R \ll N$. Lastly, the $\varepsilon_{i,t}$ is an idiosyncratic error which can exhibit autocorrelation and/or cross-sectional dependence made precise below. Stacking the model over all $i = 1, \ldots, N$ variables and $t = 1, \ldots, T$ observations we obtain the following matrix representation:

$$Y = \Lambda F + \varepsilon, \tag{4.1}$$

where Y and ε are now matrices of dimension $N \times T$, Λ is of dimension $N \times R$ and F is of dimension $R \times T$. We assume that both N and T are large and we restrict the dimensions of the panel such that $N/T \to \gamma \in (0, 1]$. This is to conserve on notation and the results of the paper apply equally if the opposite restriction holds by a mirror argument.

We will estimate Λ and F in model (4.1) using the Principal Components estimator (PC) of Bai and Ng (2002) and this method involves solving an eigenvalue problem. When $N < T$ and R is known, it can be

shown that the PC estimator solves:

$$\begin{aligned} Q(R) :=& \min_{F,\Lambda} (NT)^{-1} \sum_{i=1}^{N} \sum_{t=1}^{T} \left(y_{i,t} - \sum_{r=1}^{R} \lambda_{r,i} f_{r,t} \right)^2 \text{ s. t. } \Lambda'\Lambda/N = I_R \\ =& \sum_{i=R+1}^{N} \hat{\xi}_i, \end{aligned}$$

where the $\hat{\xi}_{R+1} \geq \cdots \geq \hat{\xi}_N$ are the $N-R$ smallest eigenvalues of the covariance matrix $\hat{\Xi} := (NT)^{-1} YY'$. Since the factors are unobservable and consistent estimation depends crucially on knowledge of the rank of ΛF, the practitioner is posed with a more general problem. That is, for any fixed k, minimization of $Q(k)$ is equivalent to deleting the k largest eigenvalues from:

$$\begin{aligned} \text{tr}\left(\hat{\Xi}\right) :=& \text{tr}\left[YY'/(NT)\right] \\ =& \text{tr}\left[\Lambda FF'\Lambda'/(NT)\right] + \text{tr}\left[\varepsilon\varepsilon'/(NT)\right] + 2\text{tr}\left[\Lambda F\varepsilon'/(NT)\right] \\ =& \sum_{i=1}^{N} \hat{\xi}_i. \end{aligned}$$

The objective of the paper is therefore to determine the true number of factors from a set of candidates $\{0, 1, \ldots, R_{\max}\}$, where $R_{\max}$ is chosen by the practitioner. We will denote an estimate of the number of factors by $\hat{R}$.

The following assumptions are necessary to (i) consistently estimate model (4.1) by PC and (ii) obtain eigenvalue separation in $\hat{\Xi}$:

ASSUMPTION 1: (i) *R is finite and $N/T \to \gamma \in (0, 1]$ as $N, T \to \infty$;* (ii) *For $r = 1, \ldots, R$, $0 < v_r < \infty$ where $v_r\left[\Lambda FF'\Lambda'/(NT)\right]$ denote the eigenvalues of $\Lambda FF'\Lambda'/(NT)$ in descending order.*

ASSUMPTION 2: (i) *$T^{-1}\sum_{t=1}^{T} F_t F_t' \to_p \Sigma_F > 0$, $E\|F_t\|^4 < \infty$ for all t and $E\left\|F\varepsilon'/\sqrt{NT}\right\|^2 < \infty$;* (ii) *$N^{-1}\sum_{i=1}^{N} \Lambda_i'\Lambda_i \to_p \Sigma_\Lambda > 0$, $E\|\Lambda_i\|^4 < \infty$ for all i and $E\left\|\sum_{i=1}^{N} \varepsilon_{i,t}\Lambda_i/\sqrt{N}\right\|^2 < \infty$ for all t.*

ASSUMPTION 3: (i) *$\varepsilon = G^{1/2} u H^{1/2}$, where $G^{1/2}$ and $H^{1/2}$ are the symmetric square roots of $N \times N$ and $T \times T$ matrices $G > 0$ and $H > 0$ and u is $N \times T$;* (ii) *The $u_{i,t}$ are i.i.d. random variables over i and t with mean zero, variance σ_u^2 and finite fourth moments;* (iii) *The largest eigenvalues $\omega_1(G) < g_+$ and $\omega_1(H) < h_+$ and the smallest eigenvalues $\omega_N(G) > g_-$, $\omega_T(H) > h_-$, uniformly in N and T for some $0 < g_{(\cdot)} < \infty$ and $0 < h_{(\cdot)} < \infty$.*

Assumptions 1-3 are from Bai and Ng (2002) and AH and imply that Y obeys an approximate factor structure in the sense of Chamberlain and Rothschild (1983). In particular, since the rank of both F and Λ is R at most, Assumption 1 requires the covariance matrix of the factors and their loadings to jointly have finite second moments by restricting the largest eigenvalue of their product to be finite and the smallest to be non-

zero. This is strengthened in Assumption 2, by requiring that the individual covariance matrices of F_t and Λ_i converge to positive-definite matrices. The remainder of Assumption 2 defines the approximate factor model by accommodating (bounded) correlation of the $\varepsilon_{i,t}$ and the factors and loadings, resulting in the requirement of finite fourth moments of the factor structure. Assumptions 1-2 are sufficient for consistency of the PC estimator. Note that Assumptions 1-2 rule out (i) non-stationary factors and (ii) a scenario where the factors are perfectly collinear. Non-stationary factors could be accommodated by considering an appropriate scaling $\delta > 1$ such that $T^{-\delta}\sum_{t=1}^{T} F_t F_t' \to_p \Sigma_F > 0$ and similarly rescaling $\hat{\Xi}$, but we abstract from this complication for brevity. The second scenario cannot be accommodated because if $\Sigma_F \geq 0$, it is not possible to guarantee that the R-th eigenvalue of $\hat{\Xi}$ does not vanish with N and T. Furthermore, Assumptions 1 and 2 correspond to "strong" factors in the sense of Onatski (2010, 2015). Weak factors, that is, factors for which $T^{-\delta}\sum_{t=1}^{T} F_t F_t' \to_p \Sigma_F > 0$ using some reduced rate $0 \leq \delta < 1$, do not yield eigenvalue separation under the maintained definition of $\hat{\Xi}$. However, in the next section we show that if $0 < \delta < 1$, the weak factor model does yield eigenvalue separation if $\hat{\Xi}$ is rescaled appropriately and the above assumptions are modified accordingly. Finally, the conditions in Assumption 3 ensure that the error covariance matrix $\hat{\Omega} := T^{-1}\varepsilon\varepsilon'$ has well-defined limiting eigenvalues $\omega_i(\Omega)$ for which we can derive parametric bounds. To accommodate this we assume that the error matrix ε can be decomposed in a matrix that controls autocorrelation ($H^{1/2}$), a matrix that controls for cross-sectional dependence ($G^{1/2}$) and a random component u. Positive definiteness of the matrices G and H ensures that the grand covariance matrix is also positive definite, whilst the conditions on u allow for an appeal on RMT. Note however that Assumption 3 is restrictive because divergence of the eigenvalues under Assumptions 1-2 requires only that $\hat{\xi}_{R+1} = o_p(1)$. In this context Assumption 3 is used to bound the largest and smallest eigenvalues of Ω and subsequently the error covariance matrix.[3] Note also that the specification implied by Assumption 3 is weaker than that of Onatski (2010) who requires that both G and H have well-defined spectral distributions. The assumption does however allow the error to be generated by common examples of DGPs. For example, if the elements of ε are stationary, i.i.d over i and first-order autocorrelated, i.e $\varepsilon_{i,t} = \rho\varepsilon_{i,t-1} + u_{i,t}$, with common variance $\sigma_u^2 = 2$, then we have $G = I_N$ and H Toeplitz with:

$$H = \frac{1}{1-\rho^2}\begin{bmatrix} 1 & \rho & \rho^2 & \dots & \rho^{T-1} \\ \rho & 1 & \rho & \dots & \rho^{T-2} \\ \rho^2 & \rho & 1 & \dots & \rho^{T-3} \\ \dots & \dots & \dots & \dots & \dots \\ \rho^{T-1} & \rho^{T-2} & \rho^{T-2} & \dots & 1 \end{bmatrix},$$

and autoregressive parameter $\rho = 0.5$ and the variance of ε is thus $\text{var}(\varepsilon_{i,t}) = 8/3$ for all $i = 1, \dots, N$. Then, following Grenander and Szegö (1958), the largest eigenvalue of H converges to $\frac{1+\rho}{1-\rho}$, which is finite as long as $\rho \neq 1$.

[3] See also the discussion in Onatski (2010).

4.3 Eigenvalue Separation and the Link to Information Criteria

In this section we study the empirical distribution of the eigenvalues of $\hat{\Xi}$. We will see that under Assumptions 1-3, the eigenvalues of $\hat{\Xi}$ separate into disjoint intervals corresponding to the factor structure and the error. The remainder of the section is devoted to studying spectral separation under modified assumptions. All later results are corollary to the following theorem on the convergence of the eigenvalues of $\hat{\Xi}$:

THEOREM 4.1. Eigenvalue Separation in the Approximate Factor Model: *Under Assumptions* 1-3*, as* $N, T \to \infty$*:*

1. $\hat{\xi}_r = (NT)^{-1} v_r(\Lambda F F' \Lambda') + O_p\left(T^{-1/2}\right)$ *and* $v_r[\Lambda F F' \Lambda'/(NT)] = O_p(1)$ *for all* $r = 1, \ldots, R$*. Furthermore, the first eigenvalue of* $\hat{\Xi}$ *is bounded above:*

$$\begin{aligned}\hat{\xi}_1 &= (NT)^{-1} v_1\left(\Lambda F F' \Lambda'\right) + O_p\left(T^{-1/2}\right) \\ &\leq v_1(\Sigma_F)\, v_1(\Sigma_\Lambda) + O_p\left(T^{-1/2}\right)\end{aligned}$$

and the R-th eigenvalue of $\hat{\Xi}$ *is bounded below:*

$$\begin{aligned}\hat{\xi}_R &= (NT)^{-1} v_R\left(\Lambda F F' \Lambda'\right) + O_p\left(T^{-1/2}\right) \\ &\geq v_R(\Sigma_F)\, v_R(\Sigma_\Lambda) + O_p\left(T^{-1/2}\right).\end{aligned}$$

2. $N\hat{\xi}_i = \omega_i\left(\hat{\Omega}\right)$ *and* $\hat{\xi}_i = o_p(1)$ *for all* $i = R+1, \ldots, N$*. Moreover, the* $R+1$*th eigenvalue is bounded above:*

$$N\hat{\xi}_{R+1} \to_p \omega_1(\Omega) \leq \left(1 + \gamma^{1/2}\right)^2 \sigma_u^2 c_+ \tag{4.2}$$

for some $c_+ > 0$ *and the N-th eigenvalue is bounded below:*

$$N\hat{\xi}_N \to_p \omega_N(\Omega) \geq \left(1 - \gamma^{1/2}\right)^2 \sigma_u^2 c_-. \tag{4.3}$$

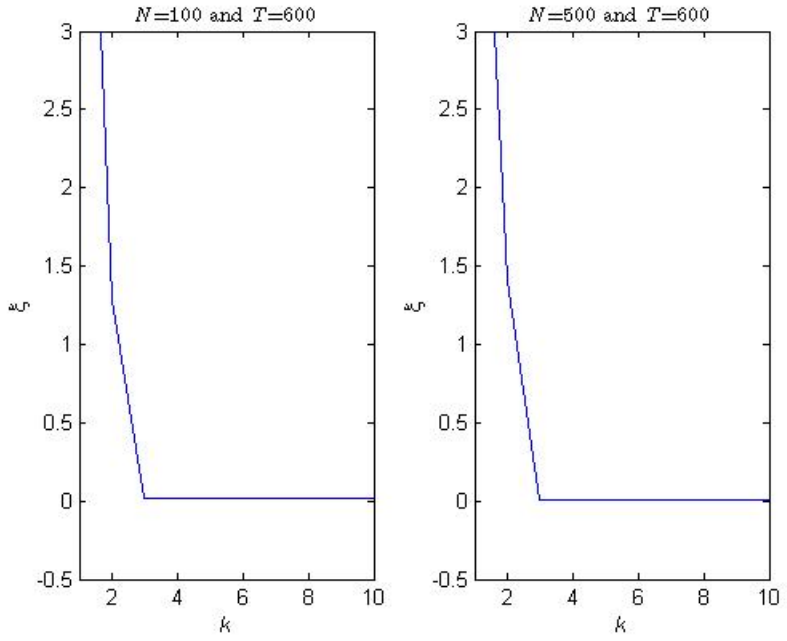
for some $c_- > 0$*.*

Proof: See Appendix.

The message of Theorem 4.1 is that the eigenvalues of the covariance matrix of a factor model will separate exactly into R non-vanishing eigenvalues and $N - R$ remaining eigenvalues that tend to zero as $N, T \to \infty$. An interesting implication of Theorem 4.1 is that the eigenvalues of the covariance matrix of a factor model

under Assumptions 1-3 exactly describe a Scree plot as envisioned by Cattell (1966): the Scree plot is a graphical device that plots the eigenvalues of a data matrix ordered from largest to smallest as a function of their position in the ordering, see for example figure 1.[4] Theorem 4.1 thus justifies why choosing the "elbow" of the Scree plot results in a consistent estimate of R under Assumptions 1-3: the inflection point corresponds to a threshold that separates the factor eigenvalues from a bulk of error eigenvalues. In fact, Theorem 4.1 makes precise that the *only* way to determine R consistently is to estimate the elbow of the Scree plot. For this reason, the consistent estimators of Onatski (2010), AH and the estimators below all depend on verification of (a version of) Theorem 4.1 and thus constitute various numerical implementations of the Scree plot. That is, they belong to a class of "Scree plot" estimators.

Figure 4.1: Eigenvalues of $\hat{\Omega}$, $\sigma_u^2 = 1$ and $R = 2$ AR(1) factors



Another message of Theorem 4.1 is that methods that do not explicitly capture the separation of the support of the eigenvalue distribution may be of little value to practitioners, even if an asymptotic justification for the converse exists. The Information Criteria (IC) of Bai and Ng (2002) fall in this category and their asymptotic validity follows from a theorem by Pötscher (1983), which has been used extensively to justify

[4]The Scree plot shown is computed from a particular realization of model (4.1) with $R = 2$, $N = 100$, $N = 500$ and $T = 600$ and autocorrelation in the factors: generated by $F_t = \kappa F_{t-1} + \upsilon_t$, where $\kappa = \text{diag}(0.9, 0.5)$, for each $t = 1, \ldots T$ and$\upsilon_t \sim N(0, I_R)$; the loadings are $\Lambda_r \sim N(0, I_N)$ for each $r = 1, \ldots, R$ and the individual specific residual is $\varepsilon_t \sim N(0, I_N)$ for each $t = 1, \ldots, T$.

IC methodologies in a variety of settings.[5] In the notation of our paper, IC's have the following structure:

$$\hat{R}_{IC} = \underset{k}{\operatorname{argmin}}\, Q(k) + k \times p(N,T), \tag{4.4}$$

where $Q(k)$ is defined above and $p(N,T)$ is a penalty function satisfying (i) $p(N,T) \to 0$ and (ii) $\min(N,T) \times p(N,T) \to \infty$ as $N, T \to \infty$. Setting $p(N,T) = \left(\frac{N+T}{NT}\right) \log\left(\frac{NT}{N+T}\right) \hat{\sigma}_{\varepsilon}^2$, we recover $\hat{R}_{PC_{p1}}$ of Bai and Ng (2002) and a feasible version minimizes:

$$\hat{R}_{PC_{p1}} = \underset{k}{\operatorname{argmin}} \sum_{i=k+1}^{N} \hat{\xi}_i + k \left(\frac{N+T}{NT}\right) \log\left(\frac{NT}{N+T}\right) \hat{\sigma}_{\varepsilon}^2,$$

where $\hat{\sigma}_{\varepsilon}^2$ is the ML estimate of the average error covariance and corresponds to $\sum_{i=R_{\max}+1}^{N} \hat{\xi}_i$. Bai and Ng (2002) also present a logarithmic version:

$$\hat{R}_{IC_{p1}} = \underset{k}{\operatorname{argmin}} \log\left(\sum_{i=k+1}^{N} \hat{\xi}_i\right) + k \left(\frac{N+T}{NT}\right) \log\left(\frac{NT}{N+T}\right).$$

It is important to note that Potscher's argument is designed for objective functions that tend to the chi-squared distribution. Similarly, the logarithmic penalty function has a theoretical justification through a Taylor expansion of the log-likelihood as shown in Schwarz (1978), which is also chi-squared when correctly specified. However, $Q(k)$ is non-central chi-squared unless G and H are identity matrices and, for general G and H, it is unlikely that scaling by the diagonal elements of $(NT)^{-1} \varepsilon\varepsilon'$ only will sufficiently purge $Q(k)$ towards the chi-squared distribution in $\hat{R}_{PC_{p1}}$. On the other hand, whilst the logarithmic transformation in $\hat{R}_{IC_{p1}}$ renders the variance estimate a negligible constant, the origin of the log-penalty function is unclear other than that it apparently works well in finite samples. In both cases, the use of the logarithmic penalty function is arbitrary and chosen only because it is (i) commonly used and (ii) subject to the requirements of Pötscher (1983). In fact, an infinite set of penalty functions will satisfy these requirements and it is therefore no surprise that $\hat{R}_{PC_{p1}}$ and $\hat{R}_{IC_{p1}}$ are found to severely overestimate R in Monte Carlo studies, see e.g. Onatski (2010) and AH. This indeterminacy is a result of the lack of structure on the penalty function and we know from Theorem 4.1 that any consistent IC should delete eigenvalues from $Q(k)$ until the deleted eigenvalues are no larger than the edge of the noise matrix (4.2). Comparing functional forms, it is clear that the logarithmic penalty function is not appropriate for this unless $Q(R)$ tends to the central chi-squared distribution.

The poor finite sample performance of IC based on Pötscher's asymptotic results are well known in the literature and Hallin and Liška (2007) develop a data-driven method to improve the accuracy of IC for the dynamic factor model. This "tuning" methodology was further adapted to the approximate factor model by

[5] For example, ARMA-order determination in Pötscher's original article; Instrument selection in GMM by Andrews and Lu (2001); Rank estimation in Cragg and Donald (1997) and, indeed, to justify IC methods in factor models as in Bai and Ng, Ahn et al (2013), Sarafidis and Robertson (2015) and Chapter 3 of this Thesis.

Alessi et al (2010). Compared to (4.4), the IC is modified as follows:

$$\hat{R}_{IC} = \underset{k}{\operatorname{argmin}}\, Q(k) + k \times c \times p(N,T), \tag{4.5}$$

where c is a constant that depends on the data. It should be clear that IC (4.5) is a special case of (4.3) above with $c = 1$ and that Pötscher's argument continues to hold for any finite $c > 0$. Hallin and Liška (2007) propose a cross-validation scheme where $\hat{R} = J^{-1}\sum_{j=1}^{J}\hat{R}_j$ is estimated according to (4.5) over sub-samples Y_j where $j = 1,\ldots,J$. That is, $Y_j = A_j Y$, where A_j is a selection matrix that truncates the original data matrix Y by deleting the last rows of Y incrementally into sub samples of length $N_1 < N_2 \cdots < N_J = N$. The key insight of Hallin and Liška (2007) is that there are choices of c where the variance of the estimator $\hat{R}$,

$$S = \frac{1}{J}\sum_{j=1}^{J}\left(\hat{R}_j - \frac{1}{J}\sum_{j=1}^{J}\hat{R}_j\right)^2,$$

is minimized over all R_j as calculated over the truncated samples Y_j. There are two boundary cases: first, if c is too small, too little penalization implies $R_{\max}$ factors will be selected with $S = 0$. Second, if c is too large, no factors will be estimated and S will also be zero. Hallin and Liška (2007) argue that a second intermediate stability region of S, as measured from $R_{\max}$, provides penalization that corresponds to estimating the number of factors correctly in finite samples. Theorem 1 can be used to formalize this "tuning" procedure and the following corollary thus provides a proof to the informal treatment in Alessi et al (2010):

COROLLARY 4.1. Eigenvalue Separation in Tuned Information Criteria: *Let Assumptions* 1-3 *hold and define constants* c_1-c_4 *that may depend on N and T such that:*

$$c_1 = \left(N^{-1/2} - T^{-1/2}\right)^2 \sigma_u^2 c_-; \quad c_2 = \left(N^{-1/2} + T^{-1/2}\right)^2 \sigma_u^2 c_+;$$
$$c_3 = v_R(\Sigma_F)\, v_R(\Sigma_\Lambda); \qquad c_4 = v_1(\Sigma_F)\, v_1(\Sigma_\Lambda).$$

Then, there exist families of $IC_{\hat{R}}$ *with penalty functions where if:*

$$c \times p(N,T) \in [0, c_1) \quad : S = 0, IC_{\hat{R}} = R_{\max}; \tag{4.6}$$
$$c \times p(N,T) \in [c_1, c_2] \quad : S > 0, IC_{\hat{R}} = \hat{R} > R; \tag{4.7}$$
$$c \times p(N,T) \in (c_2, c_3) \quad : S = 0, IC_{\hat{R}} = R; \tag{4.8}$$
$$c \times p(N,T) \in [c_3, c_4] \quad : S \geq 0, IC_{\hat{R}} = \hat{R} < R; \tag{4.9}$$
$$c \times p(N,T) \in (c_4, \infty) \quad : S = 0, IC_{\hat{R}} = 0 \tag{4.10}$$

and penalty functions calibrated to (c_2, c_3) *give Information Criteria that select* $IC_{\hat{R}} \to_p R$ *as* $N, T \to \infty$.

Proof: See Appendix.

Corollary 4.1 provides a link between tuned IC and the Scree plot: by bounding the eigenvalues of the sub-samples, it implies that tuning an IC towards the interval (c_2, c_3) is equivalent to finding the partition of the real line where the eigenvalues of the data covariance matrix separate at the end point of the factor covariance matrix on the one hand and the noise covariance matrix on the other. As a result, we find that tuned IC procedures also belong in the class of Scree plot estimators.

The remainder of the section is dedicated to proving eigenvalue seperation in the approximate factor model (4.1) under different assumptions than Assumptions 1-3. These extensions involve alternative models which were proposed to explan the empirical phenomenon that the distinction between factor and error eigenvalues is often not as sharp as predicted by Theorem 4.1. For example, Onatski (2012) analyses asymptotics for so-called weakly influential factors by modelling the factor loadings as converging to a matrix analogue of the Pitman drift, i.e. $\Lambda'\Lambda - \Sigma_\Lambda \to 0$. Similarly, De Mol et al (2008) consider a weak factor model in which $\Lambda'\Lambda = O_p\left(N^\delta\right)$ with $0 < \delta < 1$ in the situation where $N > T$.

To accommodate the asymptotic regime of the latter, we need to modify our basic assumptions to weak factors (W) as follows:

ASSUMPTION W1(i): *R is finite, $N = N(T)$ such that $N/T \to \gamma \in (0, 1]$ and there exists a constant $1/2 < \delta < 1$ such that $N/T^{1-\delta} \to \infty$ as $N, T \to \infty$.*
ASSUMPTION W2(i): *$T^{-\delta}\sum_{t=1}^{T} F_t F_t' \to_p \Sigma_F$, $E\left(T^{1-\delta}\|F_t\|^4\right) < \infty$ for all t, $E\left\|F\varepsilon'/\sqrt{NT}\right\|^2 < \infty$ and $\hat{\Xi}_\delta = YY'/\left(NT^\delta\right)$.*

Assumption W2(i) defines our weak factor regime. Since we assume that $N < T$, our model reduces the rate of the factors rather than the loadings. This particular regime is stronger than the corresponding regimes considered in De Mol et al (2008) and Onatski (2012) but maintains the permitted correlation of the errors and the factors. Weaker factors, as measured by δ, are permitted by appropriately restricting this correlation. For example, in both De Mol et al (2008) and Onatski (2012), the errors are restricted to be independent of the factors. Eigenvalue seperation in the weak factor regime then follows from rescaling $\hat{\Xi}_\delta$ and by restricting the relative rates of N and T in Assumption W1(i). That is, given δ, N must grow sufficiently fast to maintain a divergence of rates. With these modifications we can now specialize Theorem 4.1 to the weak factor regime:

COROLLARY 4.2. Eigenvalue Separation in Approximate Factor Models with Weakly Influential Factors: *Let Assumptions* 1-3 *hold with* 1(i) *and* 2(i) *replaced by* W1(i) *and* W2(i) *respectively, then as $N, T \to \infty$:*

1. *For $i = 1, \ldots, N$, the $\hat{\xi}_i\left(\hat{\Xi}_\delta\right)$ separate as in Theorem 4.1:*

$$\hat{\xi}_R\left(\hat{\Xi}_\delta\right) = v_R\left[\Lambda F F' \Lambda' / \left(NT^{-\delta}\right)\right] + O_p\left(T^{1/2-\delta}\right),$$
$$\left(N/T^{1-\delta}\right)\hat{\xi}_{R+1}\left(\hat{\Xi}_\delta\right) \to_p \omega_1(\Omega) \leq \left(1+\sqrt{\gamma}\right)^2 \sigma_u^2 c_+,$$

2. *The tuning result of Corollary 4.1 holds: for $p(N,T) \in (c_2, c_3)$, $\hat{R}_{IC} \to_p R$ and $S = 0$.*

Proof: See Appendix.

Corollary 4.2 shows that exact separation obtains in the weak factor model if and only if the conditions on δ apply and the covariance matrix is rescaled appropriately. In that case, the first R eigenvalues are strictly determined by the covariance matrix of the factor component and not by any terms involving products of the error and the factor components. The requirement that $\delta > 1/2$ is restrictive for the permissible weakness of the factors however and not necessary for the weaker result that $\hat{\xi}_k = O_p(1)$ for $k = 1, \ldots, R$. That is, the cross-terms do not vanish from the first R eigenvalues when $\delta < 1/2$, although they continue to have no effect on the $R+1, \ldots, N$ remaining eigenvalues since the cross-terms have rank R. On the other hand, δ can be arbitrarily close to zero if we follow De Mol et al (2008) Onatski (2010) and restrict the correlation between the factors and errors to zero. These considerations imply that separation obtains when considerably weaker factor processes are considered as long as $T^{1-\delta}/N = o(1)$ and this condition is satisfied when, for example, $N = \delta_1 T$ with $0 < \delta_1 < 1$. As a result, the condition that $T^{1-\delta}/N = o(1)$ is necessary and sufficient for eigenvalue separation because it ensures that the eigenvalues of the noise matrix are of smaller order than those of the factors. It also implies that progressively larger N is necessary to compensate the weakness of the factors and afford appropriate scaling to the eigenvalues of the error covariance matrix without violating $N < T$.

The second extension of model (4.1) was recently studied by Li et al (2017): these authors consider estimation of a factor model where in addition to $N, T \to \infty$, also $R \to \infty$ in a certain way. We will consider a version of the large-R (L) factor model by amending our assumptions as follows:

ASSUMPTION L1(i): *$R/T \to \tau \in (0, 1)$, $N/T \to \gamma \in (0, 1]$ and $\tau < \gamma$ as $R, N, T \to \infty$.*
ASSUMPTION L2(i): (a) *$F = G_F^{1/2} \nu H_F^{1/2}$, where ν is $R \times T$ a matrix consisting of zero-mean i.i.d. random variables $\nu_{r,t}$ over r and t with variance σ_ν^2 and finite fourth moments, G_F and H_F are $T \times T$ and $R \times R$ matrices respectively with $g_{F-} < v_t(G_F) < g_{F+}$ and $h_{F-} < v_r(H_F) < h_{F+}$ uniformly in r and t for some $0 < g_{(\cdot)} < \infty$, $0 < h_{(\cdot)} < \infty$;* (b) *$E(u_{i,t}\nu_{r,t}) < \infty$ such that $E\left\|F\varepsilon'/\sqrt{NT}\right\|^2 < \infty$.*

Assumptions L1(i) defines the rate at which R is allowed to grow in relation to N and T: clearly if R grows too fast, the factors would dominate the smallest dimension of the covariance matrix $\hat{\Xi}$ and eigenvalue separation is impossible. Assumption L2(i) imposes structure on the factor covariance matrix that is similar to Assumption 3 above. This allows us to bound the smallest factor eigenvalue even when $R \to \infty$

and $R/T \to \tau$.[6] The following corollary shows that the eigenvalue separation holds in the large-R model under these conditions:

COROLLARY 4.3. Eigenvalue Separation in the Large-R Approximate Factor Model: *Let Assumptions* 1-3 *hold with* 1(i) *and* 2(i) *replaced by* L1(i) *and* L2(i) *respectively. Then, as* $R, N, T \to \infty$*:*

1. *For* $i = 1, \ldots, N$, *the* $\hat{\xi}_i\left(\hat{\Xi}\right)$ *separate as in Theorem 4.1:*

$$\begin{aligned}\hat{\xi}_R =& v_R\left[(NT)^{-1}\Lambda F F'\Lambda'\right] + O_p\left(T^{-1/2}\right)\\ \geq& \left(1-\sqrt{\tau}\right)^2\sigma_{u_F}^2 c_{F-},\\ N\hat{\xi}_{R+1} \to_p& \omega_1(\Omega) \leq \left(1+\sqrt{\gamma}\right)^2\sigma_u^2 c_+.\end{aligned}$$

2. *The tuning result holds: for* $p(N,T) \in (c_2, c_3)$, $\hat{R}_{IC} \to_p R$ and $S = 0$.

Proof: See Appendix.

According to Corollary 4.3, the first R eigenvalues of $\hat{\Xi}$ are fully determined by the large-R factor structure and the remaining $N-R$ eigenvalues by the error process as in Theorem 4.1. However, although $\hat{\xi}_{R+1} = o_p(1)$ and bounded above by the quantity in the corollary, this bound is no longer tight when R is large. This is because $\hat{\xi}_{R+1}$ is sandwiched between the first and $R+1$-th eigenvalue of $(NT)^{-1}\varepsilon\varepsilon'$. On the other hand, the scaling of $\hat{\Xi}$ is exactly sufficient for the first R eigenvalues, although we note that these are now $o_p(R)$ by analogy. The former implies that any consistent estimator of R based on eigenvalue separation can be consistent if $R \to \infty$ as long as $R/T \to \tau \in (0,1]$ and $\tau < \gamma$. Furthermore, the first R eigenvalues are now are spaced with distance R^{-1} on average and form a smoothly declining sequence in the Scree plot.[7]

Corollary 4.3 can thus be used to explain why eigenvalue separation does not obtain as clearly as posited by the approximate factor model satisfying Assumptions 1-3 in empirical work. For example, consider a model with relatively low signal-to-noise ratio: if the number of factors grows with the sample size and the eigenspacing is relatively smooth in both the factor and error components, then it may be difficult to detect separation of the eigenvalues in a finite sample. This is shown in figures 2 and 3 below, where the model is as in figure 1 above but the number of factors is generated as $R = 0.1N$ and the autoregressive parameter for the r-th factor is generated as 0.9^r for $r = 1, \ldots R$: in figure 2 below, the signal-to-noise is low at $\text{var}(\varepsilon_i) := \sigma_\varepsilon^2 = 1$ for all $i = 1, \ldots, N$ and separation clearly obtains when both $N = 100$ or $N = 500$, but now the eigenvalues corresponding to the factors form a smoothly declining function. By contrast, figure 3 presents a Scree plot of the same factor model when the signal-to-noise is very high at $\sigma_\varepsilon^2 = 50$. The result is that the separation of the eigenvalues of $\hat{\Xi}$ is now much less clear, especially with N relatively small.

[6] Comments similar to those following Assumption 3 apply.

[7] Compare Figures 1 above and 3 below.

Figure 4.3: Eigenvalues of $\hat{\Xi}$ when $R = 0.1 \times N$ AR(1) factors and $\sigma_\varepsilon^2 = 50$:

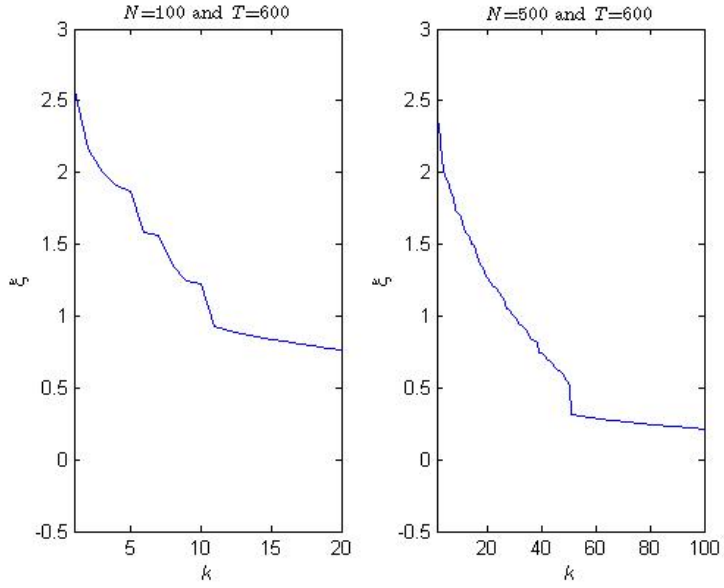
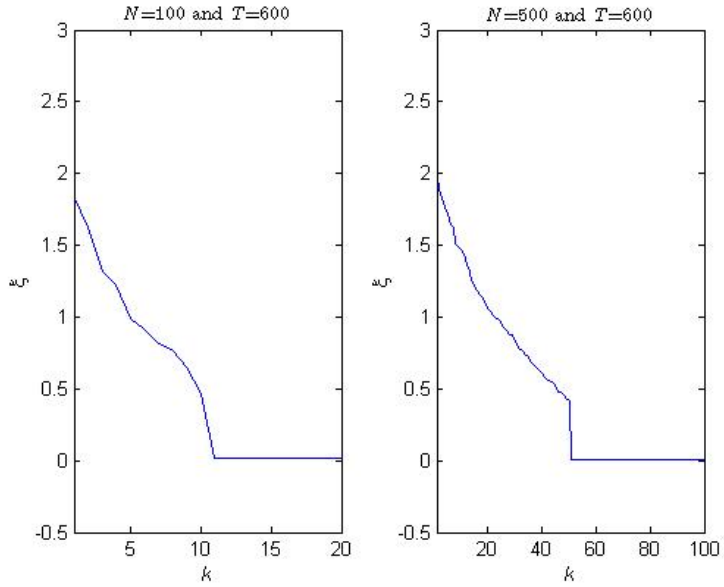


Figure 4.2: Eigenvalues of $\hat{\Xi}$ when $R = 0.1 \times N$ AR(1) factors and $\sigma_\varepsilon^2 = 1$:



In summary, the interplay of the number of factors, the strength of the factors and the signal-to-noise ratio may well provide a rationale for the empirical phenomenon that eigenvalue separation is often hard to detect.

4.4 Consistent Estimation of R based on Eigenvalue Separation

In Section 4.3 we have shown that consistent estimators of the number of factors in model (1) are numerical implementations of the Scree plot. However, despite the fact that all Scree plot estimators operate on the mechanism of eigenvalue separation, their finite sample performance will likely differ. For example, the AH test may have difficulties in finite samples when at least one eigenvalue relating to the factors is sufficiently larger or smaller than the others, causing their ratio to diverge at $\hat{R} < R$. Similarly, the threshold estimator of Onatski (2010) and the tuned IC of Hallin and Liška (2007) are known to be sensitive to the implementation of the procedure. As a result, it is useful to have more modes of comparison and we now present three further estimators to assist the practitioner in determining R.

We first present an estimator based on bound (4.1). Kapetanios (2004) also develops such an estimator and, in our notation, considers all eigenvalues larger than

$$N^{-1}\left(1+\sqrt{\gamma}\right)^2+\delta=\left(\frac{1}{\sqrt{N}}+\frac{1}{\sqrt{T}}\right)^2+\delta$$

to correspond to factors, using some δ set by the practitioner. However, given bound (4.2), it should be clear that this approach is only valid for G and H equal to identity matrices without an appropriate tuning mechanism for δ. Instead, we propose to estimate a feasible version of bound (4.2) for general σ_u^2, G and H and estimate R as the number of eigenvalues larger than this bound. The following algorithm can be used to estimate R:

ALGORITHM 4.1. Eigenvalue Difference Rule (EDR):

1. Order the N eigenvalues of $\hat{\Xi}$ from largest to smallest;
2. For the k-th eigenvalue of $\hat{\Xi}$ starting at $k = R_{\max}$, any eigenvalue satisfying the below is an eigenvalue corresponding to a factor:

$$\hat{\xi}_k > \hat{\xi}^+ := \left(\frac{1}{\sqrt{N}}+\frac{1}{\sqrt{T}}\right)^2 \hat{c}_k \hat{\sigma}_k^2 \text{ for } k = R_{\max}, R_{\max}-1\ldots, 1,$$

where:

$$\hat{\sigma}_k^2 = \frac{NT}{NT-k(N+T)}\sum_{i=k+1}^{N}\hat{\xi}_i,$$

is the ML estimate of σ_u^2 whenever $\hat{R} = R$ and:

$$\hat{c}_k = J^{-1}\sum_{j=k}^{J}\frac{\hat{\xi}_j}{\hat{\sigma}_j{}^2 N^{-1}\left(1+\sqrt{\gamma}\right)^2},$$

is a moving average estimate of c_+ at the edge of the spectrum using σ_j^2 analogously defined to σ_k^2 and J is a bandwidth parameter.

3. Repeat step 2 and stop the whenever:

$$\hat{\xi}_k - \hat{\xi}^+ \leq \delta,$$

for δ small.

EDR is a generalization of the method of Kapetanios (2004) and specializes the bound for general correlation matrices of the error G and H. Note that in addition to choosing $R_{\max}$, the practitioner must set the moving average bandwidth J and this is very similar to setting the number of eigenvalues on which the regression of the Edge estimator of Onatski (2010) is based. Furthermore, the mechanism of estimating the edge of the eigenvalues of the error covariance matrix is very similar to that of the Edge estimator and for this reason, we expect these estimators to behave similarly in finite samples. The consistency of EDR is established in the following proposition:

PROPOSITION 4.1. Consistency of EDR: *Let Assumptions 1-3 hold, let* $J/N \to 0$ *and* $\delta = O\left(N^{-(1+a)}\right)$ *with* $a > 0$ *and small. Then setting* $\hat{R}_{EDR}$ *according to Algorithm 1 gives* $\hat{R}_{EDR} \to_p R$.

Proof: See Appendix.

The consistency follows immediately from the fact that any eigenvalue approximated by the feasible version of the bound (4.2) at $k \geq R$ vanishes whilst for $k < R$, the approximated eigenvalue is $O_p(1)$. This is facilitated by the fact that the moving average estimate of c_+ operates as a smoother on the eigenvalue bound estimate and results in rejecting any non-smooth estimate of c_+ due to non-vanishing eigenvalues. As a result, as long as J is small and fixed, any factor eigenvalue in the estimated bound does not vanish with N, T.

Our second test is motivated by the shape of the Scree plot: from figure 1 above, notice that before $k = 2$, the Scree plot declines rather sharply whilst beyond $k = 2$, the Scree plot is relatively flat and slowly approaches zero as $N \to \infty$. This is because the number of eigenvalues of the error covariance matrix eigenvalues is large and their support shrinks to zero. As a result, the difference between any $\hat{\xi}_{R+k}$ and $\hat{\xi}_{R+k+1}$ is small, whilst the spacing between the factor eigenvalues is more erratic. These observations lead to the following estimator:

Minimum k-Test (MKT) :

$$\hat{R}_{MKT} = \underset{k}{\operatorname{argmin}}\, k \times \hat{\xi}_k \qquad \text{where } k \in \{1, \ldots, R_{\max}\}.$$

We know from Theorem 4.1 that a local maximum exists in $k \times \hat{\xi}_k$ whenever $k \leq R$ because $v_k = O_p(1)$ for all $k = 1, \ldots R$. However, at which k this maximum occurs depends on the data and Theorem 4.1 is of no help without further restrictions on the factor-loadings product. Another local maximum obtains at some $k > R$ because even though $\hat{\xi}_{R+k} = o_p(N)$ for $k = R+1, \ldots, N$, the atomistic spacing of consecutive eigenvalues

implies that the functional $k \times \hat{\xi}_k$ is increasing between neighbours as long as

$$\hat{\xi}_k > k\left(\hat{\xi}_{k+1} - \hat{\xi}_k\right)$$

and decreasing thereafter. The above inequality is reversed at some k by the same logic and this leads to a degenerate minimum at the end of the support of $\hat{\xi}_i$. This phenomenon is a consequence of the Courant-Fisher Theorem applied to partial traces, which states that the sum of the largest eigenvalues is a convex function whereas the sum of the smallest is concave. At the point where the $O_p(1)$ eigenvalues transition into $o_p(1)$, therefore, the estimator $k \times \hat{\xi}_k$ has a minimum over the range of eigenvalues, provided that we do not choose $R_{\max}$ too large. This implies that we may estimate the number of factors by finding the minimum of $k \times \hat{\xi}_k$:

PROPOSITION 4.2: Consistency of MKT: *Let Assumptions 1-3 hold and let* $R_{max}/N \to 0$. *Then* $\hat{R}_{MKT} = \underset{k}{\operatorname{argmin}}\left(k \times \hat{\xi}_k\right) - 1$ *and* $\hat{R}_{MKT} \to_p R$ *as* $N, T \to \infty$.

Proof: See Appendix.

MKT offers a consistent estimate of the elbow of the Scree plot which only requires the practitioner to choose $R_{\max}$ and this property makes MKT far simpler to implement than EDR or the Edge estimator. However, since the support of the error covariance matrix eigenvalues has an inflexion point, it might be necessary to tune $R_{\max}$ in order to avoid a corner solution at $R_{\max}$ in applications.

In the remainder of this section we develop a Group-Lasso algorithm in the spirit of Tibshirani (1996) and Yuan and Lin (2006). The Lasso is a modification of IC (4.5) above and implements a soft cut-off which sets to zero any coeficient that does not exceed a user-specified threshold. As a result, Lasso algorithms estimate and select models jointly. The Group-Lasso extends this method by instead of searching over individual variables, blocks of variables are either estimated or set to zero jointly. Selection of the number of factors naturally fits in this framework: for any $k \leq R$, we want both factors and loadings to exceed the threshold and be estimated, whilst for any $k > R$, we want to ignore them. The Group-Lasso has been applied to the factor model before by Hirose and Konishi (2012) and Caner and Han (2014), although the former authors do not derive consistency whilst the latter prove asymptotic consistency as an extension of the theorem attributed to Pötscher above. By contrast, we will base consistency of the Group Lasso on Theorem 4.1.

These considerations lead to the following Group-Lasso objective function:

$$V(k)_{\text{Lasso}} = \min_{\lambda_i, F_t} \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \left(y_{i,t} - \lambda_i F_t\right)^2 + 2\frac{\phi}{\sqrt{NT}} \sum_{r=1}^{k} N^{-1} \left\|\lambda_r f_r\right\|,$$

where ϕ is a user-selected threshold that sets to zero all factors and loadings for which $\|\lambda_r f_r\| < \phi$. Maximizing $V(k)_{\text{Lasso}}$ with respect to the loadings of the r-th factor and using the normalisation that $\Lambda'\Lambda/N = I_k$,

we get:

$$f_r' = \left(1 - \frac{\phi}{\left\|Y'\lambda_r/\sqrt{NT}\right\|}\right)_+ \frac{Y'\lambda_r}{N} \quad , \tag{4.11}$$

where $(a)_+ := \max(0,a)$ is the soft threshold operator. Equation (4.11) is a special case of the usual least-squares estimate of the r-th factor as in Bai and Ng (2002), which can be seen by setting $\phi = 0$.[8] Moreover, the quantity in the norm in the threshold operator is the k-th singular value of the covariance matrix $\hat{\Xi}$: from Theorem 4.1, we know that the eigenvalues related to the factors are $O_p(1)$ whilst the remaining $r+1$ are $o_p(1)$ and this carries over to the singular values straightforwardly. This suggests that setting $\phi = \sqrt{N^{-1}\omega_1(\Omega)}$ yields a consistent estimate of the number of factors. We can now concentrate out the estimated factors to obtain the following criterion function:

$$V\left(k\,|\,\hat{F}\right)_{\text{Lasso}} = -\text{tr}\left(\Lambda' YY'\Lambda\right) + 2\frac{\phi}{\sqrt{NT}}\sum_{k=1}^{R_{\max}}\left\|Y'\lambda_k\right\|. \tag{4.12}$$

Criterion function (4.12) is similar to the concentrated criterion function of Bai and Ng (2002), although the Lasso inflates the point estimate of the loadings of the r-th factor. This can be seen from the solution using a sequence of Lagrangian multipliers μ_r for each loading in the constraint $\Lambda'\Lambda/N = I_R$:

$$\begin{aligned}\lambda_r' YY'\lambda_r &= N^{-1}\lambda_r'\left[\mu_r + \phi\left\|Y'\lambda_r/\sqrt{NT}\right\|\right]\lambda_r \\ &:= N^{-1}\lambda_r'\psi_r\lambda_r.\end{aligned}$$

Here, ψ_r is an auxiliary eigenvalue that corresponds to the solution of the standard Principal Component estimator of Bai and Ng (2002). The Group-Lasso thus works *only* on the OLS estimates of the factors but leaves the estimates of the λ_r unaffected. We implement the Group-Lasso Estimator through the following algorithm.

ALGORITHM 4.2. Group-Lasso:

1. For $k = 1,\ldots,R^1_{\max}$ calculate the eigenvalues of $\hat{\Xi}$;
2. Estimate the upper bound $\hat{\xi}^+$, as in Algorithm 1 at $R^1_{\max}$;
3. Starting from $R^1_{\max}$, for any $s = R^1_{\max}, R^1_{\max}-1,\ldots$ for which $1 - \frac{\left(\hat{\xi}^+\right)^{1/2}}{\left\|Y'\lambda_k/\sqrt{NT}\right\|} \leq 0$, set $f_s' = 0$ and update $R^2_{\max} = R^1_{\max} - 1$;
4. Repeat steps 2-3 $1, 2, \ldots, m$ times until convergence.

Consistency of the Group-Lasso procedure is established in the following Proposition:

[8] Lemma 6 in the Appendix collects the steps involved in deriving equations (14) and (15) below.

PROPOSITION 4.3. Consistency of the Group-Lasso Estimator: *Under Assumptions 1-3 above, as N and $T \to \infty$,*

1. $\hat{R} \underset{p}{\to} R$;
2. *For any $k > R$, $f_k' = 0$.*

Proof: See Appendix.

Proving the consistency of the point estimate of the factors and the loadings can be done by the methods of Caner and Han (2014). Since this estimator jointly estimates the factors, the loadings and R based on a consistent estimate of the edge of the spectrum, it is clear that this procedure has Oracle properties. That is, it jointly estimates R and the parameters consistently. However, as we have provided an estimate for ϕ directly, we have negated the need for Lasso-type machinery for the consistency of the model selection portion of such a proof. As a result, the post-Lasso estimator, that is, the Principal Component estimator using the $\hat{R}$ eigenvalues of $\hat{\Xi}$ where $\hat{R}$ is computed using the Group-Lasso procedure, is consistent for any rotation of the factors and their loadings.

4.5 Eigenvalue Separation in Interactive Fixed Effects Models

In this section we extend the eigenvalue separation result to the interactive fixed effects model studied in Bai (2009) and Moon and Weidner (2015, 2017). This is important, because although determination of the number of factors in the approximate factor model has received a lot attention, the same cannot be said for the extension to traditional regression models with factor error structures. The interactive fixed effects model is defined as:

$$y_{i,t} = X_{i,t}\beta + \Lambda_i F_t + \varepsilon_{i,t},\ i = 1,\ldots,N, t = 1,\ldots,T. \tag{4.13}$$

Compared to model (4.1) above, the difference is the addition of the term $X_{i,t}\beta$: the vector $X_{i,t}$ involves L observable regressors $X_{i,t} = [x_{1,i,t},\ldots,x_{L,i,t}]$ and β collects L slope parameters which correspond to the $x_{l,i,t}$ for $l = 1,\ldots L$. The $X_{i,t}$ may include both strictly and weakly exogenous variables (i.e., lagged dependent variables) so that model (4.13) is a generalization of the well-known fixed effects panel data model in the large N, T framework. We will consider the matrix representation of (4.13) by stacking the model over all i and t:

$$Y = \left(I \otimes \beta'\right) X + \Lambda F + \varepsilon,$$

where $X = [X_1,\ldots,X_N]'$ is an $LN \times T$ matrix where each individual $T \times L$ matrix $X_i = [X_{1,i},\ldots,X_{L,i}]$ collects L regression variables at each $t = 1,\ldots,T$.

It is important to note that the focus of model (4.13) often differs from the approximate factor model (4.1) above: in applications, the researcher is typically interested in β alone and estimation of β is thus complicated by the presence of the factor error structure: if $\Lambda_i F_t$ has non-vanishing correlation with (some of) the

independent variables $X_{i,t}$, then the estimator $\hat{\beta}$ will be inconsistent if the factor structure is unaccounted for. For example, in a pure time series model with factor errors, an inconsistency arises if at least one factor is autocorrelated. On the other hand, if a model consists only of exogenous covariates, inconsistencies follow if the $X_{i,t}$ are correlated with a (sub-) set of the F_t. In fact, for all i and t, the $X_{i,t}$ may be correlated with a superset of F_t, denoted E_t and in that case we have $\text{rank}(E) = S > R$. In each case, inconsistent slope parameter estimates may well lead to incorrect conclusions in the application under consideration, whilst consistent estimation of the slope parameters typically depends on knowledge of R.

To study eigenvalue separation in the interactive fixed effects model, we will consider the estimated covariance matrix of the regression residual $\hat{\Psi} = \hat{W}\hat{W}'/(NT)$, where:

$$\hat{W} := Y - \left(I \otimes \hat{\beta}'\right) X = \Lambda F - \left(I \otimes \hat{\Upsilon}'\right) X + \varepsilon.$$

We thus assume that ignorance of (the dimension of) the factor structure leads to a linear inconsistency $\hat{\Upsilon}$ in the estimator $\hat{\beta}$ and that $p\text{lim}\left(\hat{\Upsilon}\right) = \Upsilon$. The literature has proposed several consistent estimators for β in model (4.13) and we will not worry ourselves with consistent estimation of β.[9] Instead we will impose some high level assumptions on the inconsistency of $\hat{\beta}$ and show that an eigenvalue separation result continues to hold for $\hat{\Psi}$ regardless of $\hat{R}$.

Before we can extend our assumptions to accommodate the covariates in X however, we must introduce some additional notation: let $X^{(l)} = [X_{l,1}, \ldots, X_{l,N}]$ collect all N time series on the l-th regressor for $l = 1, \ldots, L$, let y_i be the T-vector of dependent variables of the i-th individual and let $M_F = I_T - F(F'F)F'$ be the orthogonal projection on the column space of the factors.

We introduce the following additional assumptions:

ASSUMPTION 4: *As* $N, T \to \infty$, (i) $(NT)^{-1}\sum_{i=1}^{N} X_i'X_i > 0$, (ii) $(NT)^{-1}\sum_{i=1}^{N} X_i'M_F X_i > 0$ *and* (iii) *the factor component of* X, E *has* $\text{rank}(E) = \max(R, S)$, *with* R, S *finite.*

ASSUMPTION 5: *As* $N, T \to \infty$, (i) $\text{tr}\left[X^{(l)}(\Lambda F)'\right] = O_p(NT)$ *and* (ii) $\text{tr}\left(X^{(l)}\varepsilon'\right) = o_p(NT)$ *for all* $l = 1, \ldots, L$.

ASSUMPTION 6: $p\text{lim}\left(\hat{\beta}\right) = \beta + \Upsilon$ *with* (i) $\left\|\Lambda F - X\hat{\Upsilon}\right\|^2 = O_p(NT)$ *and* (ii), $\text{tr}\left[\left(I \otimes \hat{\Upsilon}'\right)X\varepsilon'\right] = o_p(NT)$ *as* $N, T \to \infty$.

Assumptions 4 and 5 amalgamate the assumptions in Bai (2009) and Moon and Weidner (2015, 2017). Assumption 4(i) is standard in the regression literature: it rules out multicollinearity in the X_i and ensures that the inverse of $(NT)^{-1}\sum_i^N X_i'X_i$ exists. Note that in the case of lagged dependent variables, this implies that model (4.13) must be ergodic stationary or mixing. The addition of 4(ii) rules out certain types of "low-rank" regressors. Examples of low-rank regressors are time-invariant or cross-sectionally invariant common

[9] For example: Pesaran (2006); Bai (2009) when R is known and Moon and Weidner (2015, 2017) when $\hat{R} \geq R$ under certain conditions. For consistent estimation of β in the case of fixed T, see Ahn, Lee and Schmidt (2014) and Robertson and Sarafidis (2015) when R is known. For the case of fixed N and $\hat{R} \geq R$, the previous chapter of this thesis.

regressors and these regressors do not have enough variation to separate them from the space spanned by the loadings and/or factors respectively. Our results can be adapted to account for these regressors by incorporating further high level assumptions of Moon and Weidner (2017), but we will ignore this complication for brevity. Assumption 4(iii) allows the $X_{i,t}$ to contain more unobservable factors than the composite error of the $y_{i,t}$ process by fixing the rank of the factor component of X at $\max(R,S)$. Typically, the literature assumes that $R=S$. Assumption 5 restricts the cross-correlation of the regressors and the composite error term: 5(i) requires that the correlation of each regressor and the factors is finite. Assumption 5(ii) restricts the error to be weakly dependent, so that lagged dependent variables are permitted as regressors. Finally, Assumption 6 is a new high level assumption which requires some justification. The assumption asserts that we consider estimators which contain a linear (first-order) inconsistency. The idea of the specification of Υ is twofold: first, it retains the properties of the correlation of the regressors with the error components. Second, the inconsistency is concentrated in the portion of the regression error associated with the factor component, not the idiosyncratic error. Note that Assumption 6 covers consistent estimators with $\hat{\Upsilon}=o_p(1)$ as a special case. Furthermore, Assumption 6 only makes claims about the first-order properties of $\hat{\beta}$. This is convenient because it is known that the estimators of Bai (2009) and Moon and Weidner (2015, 2017) are biased to order $(NT)^{1/2}$ even when $\hat{R}=R$ and by considering first-order effects we can avoid higher-order complications. Assumption 6 must be verified on a case-by-case basis and we now discuss two examples of situations where it holds. Consider first a pooled OLS regression of the parameters of model (4.13) under Assumptions 1-6:

$$\begin{aligned}
p\lim\hat{\beta}_{OLS} =& p\lim\left(\sum_{i=1}^{N}X_i'X_i\right)^{-1}p\lim\sum_{i=1}^{N}X_i'y_i \\
=&\beta+p\lim\left(\sum_{i=1}^{N}X_i'X_i\right)^{-1}p\lim\sum_{i=1}^{N}X_i'\left(\Lambda_iF\right)'+O_p\left(1/\sqrt{NT}\right) \\
:=&\beta+\Upsilon_{OLS}+o_p(1),
\end{aligned}$$

and Υ_{OLS} satisfies Assumption 6 as $N,T\to\infty$ which can be seen by squaring and summing the second term over i:

$$\begin{aligned}
&\sum_{i=1}^{N}\Lambda_iF\left(\Lambda_iF\right)'-2\sum_{i=1}^{N}\Lambda_iFX_i\Upsilon_{OLS}+\sum_{i=1}^{N}\left(X_i\Upsilon_{OLS}\right)'X_i\Upsilon_{OLS} \\
&\qquad=\|\Lambda F-X\Upsilon_{OLS}\|^2:=\|\Lambda FM_{\overline{X}}\|^2 \\
&\qquad\leq\|\Lambda F\|^2=O_p(NT),
\end{aligned}$$

where $M_{\overline{X}}$ is an orthogonal projection on the columns of $\overline{X}=\sum_{i=1}^{N}X_i$. As another example, Moon and Weidner (2015, 2017) show that their estimator $\hat{\beta}_{QML}\to_p\beta$ under Assumptions 1-5 whenever $\hat{R}\geq R$, so that one obtains $\Upsilon_{QML}=0_L$ and Assumption 6 is trivially satisfied. Furthermore, the OLS estimator can be interpreted as a special case of the QML estimator of Bai (2009) and Moon and Weidner (2015, 2017) with

$\hat{R} = 0$. It is therefore expected that the QML estimator of β with $0 < \hat{R} < R$ will similarly contain inconsistencies satisfying Assumption 6. The implications of the former for the eigenvalues of $\hat{\Psi}$ are summarized in the following theorem:

THEOREM 4.2. Eigenvalue Separation in the Interactive Fixed Effects Model: *Let Assumptions 1-6 hold. Then, as $N, T \to \infty$:*

1. $\xi_i\left(\hat{\Psi}\right) = O_p(1)$ *for* $i = 1, \ldots, max(R, S)$ *and* $\xi_i\left(\hat{\Psi}\right) = o_p(1)$ *for* $i = max(R, S) + 1, \ldots, N$.
2. *If* $\hat{\Upsilon} = O_p(1)$, *then* $\xi_i\left(\hat{\Psi}\right) \to_p v_i\left[\left(\Lambda F - (I \otimes \Upsilon')X\right)\left(\Lambda F - (I \otimes \Upsilon')X\right)'/(NT)\right]$ *for* $i = 1, \ldots, \max(R, S)$ *and* $N\xi_i\left(\hat{\Psi}\right) \to_p \omega_i(\Omega)$ *for* $i = \max(R, S) + 1, \ldots, N$.
3. *If* $\hat{\Upsilon} = o_p(1)$, *then* $\xi_i\left(\hat{\Psi}\right) \to_p v_i\left[\Lambda F F' \Lambda/(NT)\right]$ *for* $i = 1, \ldots, \max(R, S)$ *and* $N\xi_i\left(\hat{\Psi}\right) \to_p \omega_i(\Omega)$ *for* $i = \max(R, S) + 1, \ldots, N$.

Proof: As in Theorem 4.1, see appendix for details.

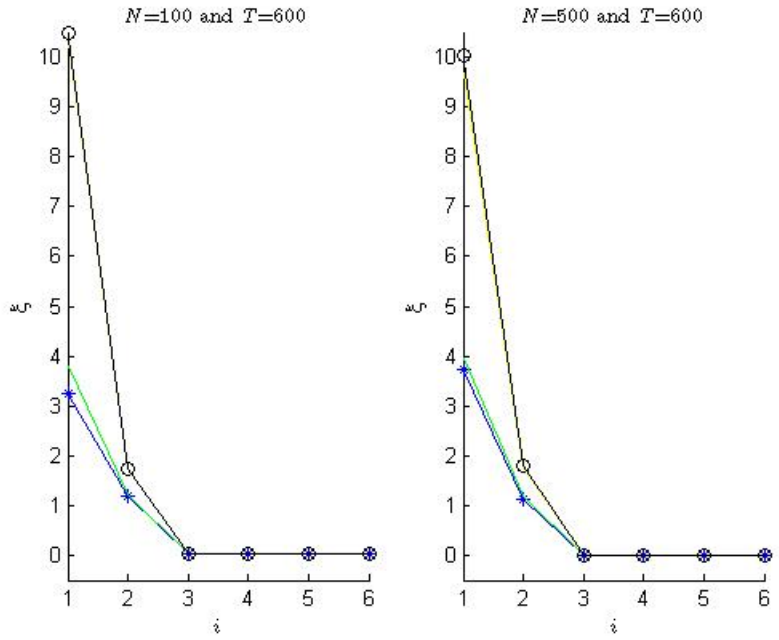
Theorem 4.2 states that the eigenvalues of the estimated covariance matrix of the residual $\hat{W}$ separate in exactly the same way in an interactive fixed effects model satisfying Assumptions 1-6 as in the approximate factor model. This result is to be expected for a model of which $\hat{\beta}$ is first-order consistent, because the error $\hat{W}$ itself constitutes an approximate factor model. Interestingly however, the separation result continues to hold if the first-order inconsistency does not vanish, although with the qualification that the inconsistency of $\hat{\beta}$ introduces the factors that enter the regressors into $\hat{\Psi}$. In that case, it is possible that the number of eigenvalues that separate from the bulk is larger than R if $\text{rank}(E) \geq \text{rank}(F)$, although the remaining eigenvalues eigenvalues $\xi_{\max(R,S)+i}$ vanish and those at the edge of the support converges weakly to the largest eigenvalue of $N^{-1}\Omega$.

An implication of Theorem 4.2 is that the Scree plot drawn from the eigenvalues of $\hat{\Psi}$ will continue to summarize the eigenvalue separation result as in Section 4.3. In fact, generating the Scree plot $R_{\max}$ times using the corresponding estimators $\hat{\beta}$ will yield a consistent estimate of R as long as $0 \leq R \leq R_{\max}$ even when the estimate of β is inconsistent at some $\hat{R}$. An example of the aforementioned is pictured in figure (4) below.

In this figure, we have generated model (4.13) above as an AR(1) model with $\beta = 0.5$ and the factor structure is as in figure (2), i.e. with $R = 2$ AR(1) factors, which are thus correlated with the regressor.[10] The individual Scree plots are displayed as follows: blue-starred corresponds to $\hat{R} = 0$; green to $\hat{R} = 1$; red to $\hat{R} = 2$; yellow to $\hat{R} = 3$ and finally black-circled corresponds to $\hat{R} = 4$. The figure can be summarized as follows: regardless of the choice of $\hat{R}$, the separation result clearly shows with two eigenvalues being substantially larger than the others which tend to zero as predicted by Theorem 4.2. Furthermore, the magnitude of the first R eigenvalues varies depending on whether $\hat{R} < R$ or $\hat{R} \geq R$: when $\hat{R} \geq R$, the estimator

[10] We have used the algorithm of Bai (2009) to compute an estimate of β and calculate the eigenvalues of $(NT)^{-1}\hat{W}_{\hat{R}}\hat{W}_{\hat{R}}'$ based on $\hat{\beta}_{QML}$ setting $\hat{R} = 0, 1, \ldots, 4$ using $T = 600$ and $N = 100$ and $N = 500$. See Section 6 for computational details.

Figure 4.4: Scree plot of $\hat{\Omega}$ of an AR(1) Interactive Fixed Effects Model with $R = 2$ AR(1) factors and $\sigma_\varepsilon^2 = 1$, and $\hat{R} = 0, 1, \ldots, 4$:



$\hat{\beta}$ filters $\hat{W}$ sufficiently to first order: as a result, $\hat{\Psi}$ is the sample covariance matrix of an approximate factor model and the conclusions of Theorem 4.1 holds. This argument corresponds to the three Scree plots clustering at approximately ten in figure 4. By contrast, when $\hat{R} < R$ and since $\beta > 0$, $\hat{W}$ contains the term $\Lambda F - \left(I \otimes \hat{\Upsilon}'\right) X < \Lambda F$ and Theorem 4.2 predicts that a product of this matrix will have correspondingly smaller eigenvalues than products based on ΛF as shown by the two Scree plots starting between three and four.

The implications for determining R in an interactive fixed effects model are therefore clear and summarized in the following proposition:

PROPOSITION 4.4: *Let Assumptions 1-6 hold, then any consistent method of determining the number of factors based on eigenvalue separation in $\hat{\Xi}$ is also consistent when applied to the eigenvalues of $\hat{\Psi}$ in the sense that $\hat{R} \to_p \max(R, S)$.*

Proof: Immediate from Theorem 4.2 and the proofs of the individual tests.

Following Proposition 4.4, the EDR and MKT estimators of Section 4 all estimate R consistently in the interactive fixed effects model, regardless of bias in $\hat{\beta}$. This argument further extends to the tests of AH, Hallin and Liška (2007), Alessi et al (2010) and Onatski (2010). Furthermore, an additional strategy for determining the number of factors is available in the interactive fixed effects model relative to the approximate

factor model: instead of estimating R only once, we can now compute $R_{\max}+1$ estimates of the covariance matrix $\hat{\Psi}$ and estimate $\hat{R}=\max(R,S)$ based on the eigenvalue distribution of each covariance matrix. By Theorem 4.2, each of these estimates yields a consistent estimate of $\max(R,S)$ and this procedure will converge to the correct number of factors because, as soon as $\hat{\beta}$ is consistent, $\hat{\Psi}$ is the covariance matrix of an approximate factor model with $\text{rank}(F)=R$.

4.6 Monte Carlo Experiments

In this section we study the finite sample performance of the proposed estimators of R in Section 4 and demonstrate eigenvalue separation in the interactive fixed effects model following Section 5. We first compare the newly presented estimators with several competing estimators, which we now introduce briefly. The first is the ratio estimator of AH:

$$\hat{R}_{AH}=\max_{k}\hat{\xi}_k/\hat{\xi}_{k+1} \text{ for } k=1,\ldots,R_{\max}$$

AH estimate the edge of the spectrum using Theorem 1: since the first R eigenvalues are $O_p(1)$ and the remaining $R+1,\ldots N$ eigenvalues are $o_p(1)$, the ratio at $k=R$ will explode whereas it is expected to be roughly constant at all other values of $k\neq R$. This simple idea is shown to be a good alternative to more complicated estimation strategies in finite samples and has the desirable property that it requires no tuning apart from the choice of $R_{\max}$.

The second estimator is IC_{p1} of Bai and Ng (2002), which we denote as BN:

$$\begin{aligned}\hat{R}_{BN}=&\underset{k}{\text{argmin}}\log\left(\sum_{i=k+1}^{N}\hat{\xi}_i\right)+k\left(\frac{N+T}{NT}\right)\log\left(\frac{NT}{N+T}\right)\\ &\text{over } k=0,1,\ldots,R_{\max}.\end{aligned}$$

Although several estimators are presented in their paper, BN is found to be most effective and therefore we only report results for this estimator. As discussed before, BN is a classical IC methodology applied to the determination of R and we know it tends to severely overestimate R in finite samples.

The third estimator is the Edge Estimator of Onatski (2010):

$$\hat{R}_{ON}=\max_{k}\left\{k\leq R_{\max}:\hat{\xi}_k-\hat{\xi}_{k+1}\geq\delta\right\},$$

where $\delta=2|\beta|$ is a threshold estimated from a regression of the eigenvalues at some value k on the constant and the cardinality of the eigenvalue raised to the power 2/3 yielding the slope parameter β.[11] Then, all $\hat{\xi}_k-\hat{\xi}_{k+1}<\delta$ are considered to belong to the error covariance matrix and the process is repeated at the

[11] See Onatski (2010) for details.

newly estimated edge of the spectrum until the process converges, that is, when no more changes in the edge occur. The ON estimator is shown to have very good properties but is also sensitive to tuning issues through the choice of $R_{\max}$ and the choice of how many eigenvalues should be used in the regression at each iteration. We follow the recommendation of Onatski (2010) by using five eigenvalues to estimate the spectral edge.

We now compare the former estimators with EDR and MKT. Since GL is an implementation of EDR, we do not consider this estimator in our finite sample comparison. Finally, in all cases we have set $R_{\max} = 20$ and in the case of EDR, $\delta = 1/N$, so that we estimate $\hat{R}$ whenever:

$$\hat{R}_{EDR} = \min_k \left\{ k \leq R_{\max} : \left(1/\sqrt{N} + 1/\sqrt{T}\right)^2 \hat{c}_k \hat{\sigma}_k^2 < \delta \right\}$$

and experimentation with other values of δ yields qualitatively similar results.

To evaluate the estimators in finite samples, we use a version of the factor model (4.1) as used by Bai and Ng (2002), Onatski (2010) and AH in their respective Monte Carlo experiments. The basic model is:

$$\begin{aligned} Y_{i,t} =& \Lambda_i F_t + \sqrt{\theta\eta}\varepsilon_{i,t},\ i = 1, \ldots N, t = 1, \ldots, T, \text{ where:} \\ \eta =& \frac{1-\rho^2}{1+2J\alpha^2},\ \varepsilon_{i,t} = \rho\varepsilon_{i,t-1} + v_{i,t} + \sum_{1 \leq |j| \leq J} \alpha v_{i-j,t}. \end{aligned}$$

Furthermore, $\lambda_{r,i}$, $f_{r,t}$ and $v_{i,t}$ are all distributed as $N(0,1)$ and α, θ, ρ and J are parameters we adjust to mimic different models. Following AH, the term η is used to normalize the variance of the noise component in $Y_{i,t}$ to unity regardless of the specification of the error term. Note that this specification implies that:

$$\begin{aligned} \text{tr}\left(Y'Y\right) =& \text{tr}\left(\Lambda'\Lambda FF'\right) + \theta\eta\,\text{tr}\left(\varepsilon'\varepsilon\right) \\ =& R \times (NT) + \theta \times (NT) \\ :=& \text{signal} + \text{noise}, \end{aligned}$$

so that θ controls the signal-to-noise ratio (SNR) for given R as R/θ. We will adjust the strength of the factors through θ, making it more difficult to estimate R when θ is larger. Furthermore the parameter ρ is used to introduce (first-order) autocorrelation in the error, whereas α and J introduce cross-sectional correlation. For every characteristic, specified through adjusting the model parameters, we simulate models with symmetric dimensions in N and T, ranging from $N, T = 50$ to $N, T = 400$ and intermediate cases with $N < T$, i.e. $N = 50, T = 50$; $N = 50, T = 100$; $N = 100, T = 100$ and so on. In all experiments, we centre the data by subtracting time and cross-sectional averages. Finally, the models are simulated using $R \in \{1, 5, 10\}$ on the basis of 1000 Monte Carlo trials, for which we report (i) the fraction of the number of times the estimator selected the correct number R over all trials, (ii) the mean R to measure consistency, and (iii) the root mean-squared error of the estimated number of factors relative to the true number. Our Monte Carlo is divided into into four parts: first, we study the performance of the estimators when there is no temporal or cross-sectional correlation in the error by reducing the approximate factor model to an exact

factor model and we subsequently analyse the impact of cross-sectional and time series correlation in the errors when $\theta = 1$. In the second part, we increase θ to ten and evaluate the estimators with and without correlation in the error. In the third and fourth parts, we adjust SNR by fixing the variance of one factor at ten or $1/6$th respectively and simultaneously introducing correlation structures in the error.

Table 4.1 summarizes the Monte Carlo experiments without temporal or cross-sectional correlation in the error and the results can be summarized as follows: regardless of the dimensions of the panel or the number of factors, all estimators behave similarly as long as R/N is not too large. When this is the case, MKT loses efficiency relative to the other estimators because it becomes harder to find the appropriate local maximum to the right of the edge of the spectrum. The other estimators however determine R correctly in essentially hundred percent of the cases, yielding very good finite sample properties whenever the exact factor model assumption holds. It is often argued that the pure factor model is too strong for empirical applications because the errors of the data exhibit some form of correlation. We now relax the pure factor model by introducing autocorrelation in the errors by setting $\rho = 0.5$ and repeat the experiments as reported in table 4.2. The experiments reveal that autocorrelated errors make it more difficult for the eigenvalue-based estimators to correctly determine the number of factors in small samples: whilst AH remains practically unbiased for all panel configurations and R, EDR and ON now require samples larger than $N, T = 50$ to estimate the edge of the spectrum with hundred percent certainty, especially when $R = 10$. This conclusion extends to BN and MKT more strongly, although MKT always outperforms BN and converges more rapidly to hundred percent success rate when the sample size becomes large and over-estimates R far less often, resulting in higher efficiency as measured by RMSE. In conclusion, all estimators suffer from the introduction of time series correlation in the error and we contrast this finding with the introduction of cross-sectional dependence in the error by setting $\alpha = 0.5$ and $J = \max(10, N/20)$. Note that this specification constitutes cross-sectional correlation that grows with N and this implies that Assumption 3 is violated because the maximal eigenvalue of G is now unbounded. However, we will follow the literature by examining this construct. The results in table 4.3 show that under cross-sectional correlation of the errors, we find even more difficulty to determine R consistently: whilst AH, ON and EDR tend to hundred percent success rates as N, T get large, BN and MKT are considerably less successful and actually fail to approach exact success rates in all samples with $R > 1$. In fact, with strong cross-sectional correlation, BN always estimates $R_{\max}$, reinforcing the point of Section 3 that the penalty term is inappropriate in finite samples when the error exhibits dependence in either N or T. On the other hand, the AH and EDR estimators now require at least $N, T = 100$ to approximate the correct number of factors whilst ON falls behind and requires even larger samples. These conclusions are reinforced when we set both $\alpha = 0.5$ with $J = \max(10, N/20)$ and $\rho = 0.5$ in the Monte Carlo experiment summarized in Table 4.4: it appears that the detrimental effect of correlation structures in the error on the performance of the estimators is driven mostly by the cross-sectional correlation. As before, AH, ON and EDR approach exactly correct determination of R as N and T grow large, whilst BN always strictly overestimates the number of factors in all configurations of the model. By contrast, although MKT suffers from the correlation structures when $R > 1$, it does not over-estimate as strongly as BN resulting in far smaller RMSE.

In practical situations, the separation of factors and residual eigenvalues is often far less pronounced as in the model with $\theta = 1$. To approximate this phenomenon, we now dramatically reduce the SNR in the following experiments by setting $\theta = 10$. Initially, we adhere to the assumptions of the pure factor model and these results are summarized in table 4.5. Compared to the baseline model in table 4.1, all estimators again reach exact identification of the number of factors as the sample size increases, now requiring instead $N, T = 100$ to approach exact identification. AH, ON and EDR all require the smallest panel dimensions to approach hundred percent correct estimation and it is not obvious which estimator is superior in this experiment. Furthermore, when R/N is large, both BN and MKT suffer as in the baseline model with the latter never reaching exactly hundred percent success rates. The excellent performance of AH, ON and EDR in the pure factor model with $\theta = 10$ is dramatically altered when we also introduce cross-sectional and time series correlation as before: as the results in table 4.6 show, all estimators now fail to consistently select R unless the sample size is very large at $N, T = 400$ and only then do AH and EDR reach at least ninety percent success at all choices of R; ON consistently estimates the factors with success rate approaching unity in the case of $R = 1, 5$, whilst MKT only approaches hundred percent in the case of $R = 5$. Finally, BN always overestimates at $R_{\max}$.

So far we have held the variance of the individual factors constant even though it is unlikely that this specification is realistic empirically. Moreover, this specification implicitly favours AH. For this reason we now change the variance of one factor to be substantially higher or lower than the remaining factors and assess the ability of the estimators to select R consistently. The results of these experiments are reported in tables 4.7-10 and we first set the variance of the first factor equal to ten, considering the baseline specification without correlation in the error. As expected, AH now suffers in the smallest configurations of the panels, whilst the other estimators have essentially exact determination percentages, apart from MKT, which requires R/N to not be too large. AH however only requires the sample size to be $N, T = 100$ to approach exact determination results. As we further introduce correlation in the error however, this conclusion no longer holds and AH requires very large samples to approach exact determination of R, exhibiting very poor performance in smaller samples. ON behaves better than AH, but still requires larger samples than EDR, which approaches exact determination from $N, T = 100$. Furthermore, with correlation in the error, as before, BN strongly over-estimates the number of factors, with zero acceptance and average rates of correct R. MKT is somewhere in between BN and AH, requiring large samples to correctly estimate R, otherwise severely over-estimating. When instead we fix the variance of one factor at $1/6$, a similar pattern occurs. Without correlation in the error, all estimators perform very well when $R = 1$. However, when $R > 1$, MKT and especially AH require larger samples to reach exact determination, whilst BN, ON and EDR continue to determine R correctly in essentially hundred percent of the experiments using the pure factor model. When we subsequently introduce correlation structures in the error, this effect is exacerbated: in line with the previous results, BN never selects R with high probability, whereas the other estimators approach hundred percent correct R only as N, T are very large in the case when $R = 1$. When $R > 1$, only EDR eventually reaches good finite sample properties.

In practice, the data is likely to be a mixture of the experiments considered in this section and the Monte

Carlo results stipulate that it is very important to analyse the properties of the data so that a reasonable estimate of R can be made. The new estimators, particularly EDR, are in all cases very competitive with AH and ON and often more robust to the complications examined in this section.

The remainder of Section 4.5 is dedicated to studying finite sample eigenvalue separation in the interactive fixed effects model and the determination of R in that situation. We initially study the following AR(1) model with AR(1) factor error components:

$$\begin{aligned} y_{i,t} =& \beta y_{i,t-1} + \Lambda_i F_t + \varepsilon_{i,t}, \\ F_t =& \kappa F_{t-1} + \upsilon_t,\ i = 1,\ldots,N,\ t = 1,\ldots,T \end{aligned}$$

where the autoregressive parameter $\beta = 0.5$, the loadings are distributed as $\lambda_i \sim N(1_R, I_R)$ and the errors are $\varepsilon_{i,t} \sim N(0,1)$. For the factors, we set $\kappa = 0.9$ and $\upsilon_t \sim N(0_R, I_R)$. Note that because F_t is dynamic, the composite error of $Y_{i,t}$, i.e. $\Lambda_i F_t + \varepsilon_{i,t}$, is autocorrelated and thus requires explicit treatment of the factors to obtain a consistent estimate of β. In the following experiments, we use the QML estimator of Bai (2009) for this purpose. To compute $\hat{\beta}$, we use the algorithm of Bai (2009) using twenty random starting values drawn from $U(-1,1)$ and iterate until $\left|\hat{\beta}^j - \hat{\beta}^{j-1}\right| < 0.01$, where j is the iteration index corresponding to each of the starting values. After twenty trials, we choose $\hat{\beta}$ corresponding to the smallest value of the objective function computed as $\mathrm{tr}\left(\hat{\Psi}_{\hat{R}}\right)$, where $\hat{\Psi}_{\hat{R}} = (NT)^{-1}\hat{W}_{\hat{R}}\hat{W}'_{\hat{R}}$ at some $\hat{R}$. The dimension of each panel under consideration is as above and we run 1000 Monte Carlo trials for each combination of N and T. Summary statistics of these experiments are printed in tables 4.11-14 where each contains two results: the first table summarizes consistency and RMSE statistics for $\hat{\beta}$; the second table reports results on determination of R. To conserve space, we report only AH and EDR applied to the eigenvalues of the resulting residual covariance matrix $\hat{\Psi}_{\hat{R}}$.[12]

Initially, we set $R = 2$ and $R_{\max} = 3$ and the results are reported in table 4.11. Regarding bias of the QML estimator: as long as $\hat{R} \geq 2$, the estimator is virtually unbiased with very low RMSE provided that $T \geq 100$. On the other hand, when $\hat{R} = 0$, the resulting OLS estimator suggests a non-stationary model with $\hat{\beta} \approx 1$. Furthermore, both AH and EDR determine R correctly with high probability in all samples considered, even when $\hat{R} = 0$. Only when $N, T = 50$ does EDR perform slightly worse than AH. This pattern is repeated when we set $R = 5$ and $R_{\max} = 7$: whilst both bias and RMSE are now larger in the smallest samples, the results are essentially the same as in the situation with $R = 2$ when $T \geq 200$. The estimators of R on the other hand behave qualitatively the same as before, with very good properties in every sample of our Monte Carlo study, selecting R correctly with success rates approaching unity rapidly.

In the pure AR(1) model of tables 4.11 and 4.12, the number of factors is necessarily equal in the dependent variables and the regressors. However, as we have seen in Theorem 4.2, the regressors may contain more or less factors than the dependent variable, and if the number of factors in the regressors is larger, these will dominate additional eigenvalues of $\hat{\Psi}$ beyond the first R when $\hat{\Upsilon} \neq o_p(1)$. To demonstrate

[12] When $\hat{R} = 0$, the QML estimator collapses to the OLS estimator, see Section 4.5 for discussion.

this complication, we extend the baseline model at the top of this section to the following static interactive fixed effects model:

$$\begin{aligned}
Y_{i,t} =& X_{i,t}\beta + \Lambda_i F_t + \sqrt{\theta\eta}\varepsilon_{i,t}, \text{ where:} \\
X_{i,t} =& \tau_{1,i} F_t + \tau_{2,i} Z_t + \upsilon_{i,t}, \\
\eta =& \frac{1-\rho^2}{1+2J\alpha^2}, \\
\varepsilon_{i,t} =& \rho\varepsilon_{i,t-1} + v_{i,t} + \sum_{1\leq|j|\leq J} \alpha v_{i-j,t}.
\end{aligned}$$

In this model, we have introduced an independent variable $X_{i,t}$ that is composed of the factors F_t and an additional factor Z_t. We will leave the distributional assumptions of the factors as at the top of this section and now generate $[F_t', Z_t']' \sim N(0_{R+1}, I_{R+1})$. Since the OLS estimator is consistent in the large N, T model if $\text{corr}(\Lambda_i, \tau_{1,i}) = 0$, we want to impose the need for QML estimation. The former is achieved by specifying perfectly correlated loadings of the $Y_{i,t}$ and $X_{i,t}$ processes as $\Lambda_i = 2\tau_{1,i}$, where $\left[\tau_{1,i}', \tau_{2,i}\right]' \sim N(0_{R+1}, I_{R+1})$. Finally, the error process $\varepsilon_{i,t}$ is defined as above, the $\upsilon_{i,t} \sim N(0,1)$ and $\beta = 1$.

We first study the finite sample performance based on the model with no correlation in the errors and the results are printed in table 4.13. Compared to the time series model, the QML estimator is now essentially unbiased with RMSE approaching zero at all panel configurations as long as $R \geq 2$. Furthermore, when we estimate β with $\hat{R} = 0$, the estimator is biased severely as anticipated by the introduction of correlation between the loadings of $Y_{i,t}$ and $X_{i,t}$. When $\hat{R} = 2, 3$ both AH and EDR determine R consistently at any sample size, only having small RMSE in the panel of size $N, T = 50$. When $\hat{R} = 0$ however, both estimators converge to $\hat{R} = 3$, as predicted by Theorem 4.2, although initially, larger values of $\hat{R}$ are observed with fairly high variation as measured by RMSE. However, since the estimators perform very well whenever $\hat{R} > R$, we expect that iterating the estimators over $\hat{R}$ will in practice result in consistent $\hat{\beta}$.

For the final experiments, we introduce cross-sectional and time series dependence by setting $\rho = 0.5$, $\alpha = 0.5$ and $J = 8$ as in the experiments with the approximate factor model at the top of this section. As can be seen from table 4.14, the QML estimator is also consistent in this specification of the model and essentially unbiased as long as $\hat{R} \geq R$ and $N, T > 50$, showing the robustness of the QML estimator to correlated errors. As for estimation of R, the impact of the correlation structure of the errors on the eigenvalues now makes it more difficult to correctly find R, especially so for AH: whilst EDR estimates R with probability approaching unity whenever $\hat{R} \geq R$ in panels of dimension $N, T \geq 100$, AH requires at least $N, T \geq 200$. On the other hand, when $\hat{R} = 0$, neither AH nor EDR consistently find $\hat{R} = 3$, with AH strictly under-estimating and EDR strictly overestimating R. As N, T become big however, the overestimation of EDR will eventually lead to a consistent estimate of the number of factors because it is consistent when $\hat{R} \geq R$, whilst AH falsely selects a smaller number of factors in this scenario.

4.7 Conclusions

In this paper we have made an argument for redemption of the Scree plot. We show that the eigenvalues of the covariance matrix of an approximate factor model that correspond to factors dominate the eigenvalues that correspond to the error process. As a result, the eigenvalues of the covariance matrix separate into distinct sets that differ in orders of magnitude. This mechanism is exactly captured by the Scree plot and thus any consistent estimator, including the tuned IC of Hallin and Liška (2007) but not the IC of Bai and Ng (2002), belongs to the class of Scree plot estimators. We have further shown that the separation result continues to hold in the weak factor model with appropriate scaling and in a model where the number of factors to goes to infinity with N and $R/T \to (0,1)$. Based on spectral separation, we develop two new estimators of R and a Group Lasso algorithm that estimates R and the parameters of the model jointly. Extensive Monte Carlo experimentation confirms that these estimators have competitive properties to several other estimators of R, with especially EDR coming out as very robust in relation to the competitors we considered.

Our paper also extends the eigenvalue separation result to the interactive fixed effects model, even when a point estimate of the slope parameter is inconsistent. In that case, separation of the eigenvalues of the covariance matrix of the regression error may depend on the rank of the factor component of the regressors, rather than the composite error of the dependent variable. On the other hand, the result of the approximate factor model goes through without further complications when the estimate of the slope parameters is consistent. This result is important, because although determination of R in the approximate factor model has received a lot of attention, the same cannot be said for this problem in the interactive fixed effects model. It also implies that estimators that depend on eigenvalue separation continue to be consistent in the interactive fixed effects model. This conclusion is reassuring, because it implies that there is no immediate need to develop new estimators for the problem of determining R in those models. Instead, one can estimate the model using an appropriate estimator $R_{\max}$ times and compute $\hat{R}$ using the point estimate at each of these. Theorem 4.2 then tells us that as long as the estimator of the slope parameters is consistent, this will subsequently lead to consistent determination of R and thus consistent estimation of β in the large N, T interactive fixed effects model, as long as at least one $\hat{R} \geq R$ is included in the algorithm.

Appendix 4.1 Proofs for Chapter 4

LEMMA 1: *Let U be an $N \times T$ random matrix with i.i.d. elements $u_{i,t}$. Assume $u_{i,t}$ has zero mean, variance σ^2 and finite fourth moments. Then as $N, T \to \infty$ and $N/T \to \gamma \in (0,1)$:*

$$\xi_1\left(T^{-1}UU'\right) \to_p (1+\sqrt{\gamma})^2\sigma^2;$$
$$\xi_1\left(T^{-1}UU'\right) \to_p (1-\sqrt{\gamma})^2\sigma^2$$

where $\xi_1 \geq \xi_2 \geq \cdots \geq \xi_N$ *are the eigenvalues of* $T^{-1}UU'$.

Proof: See Bai and Silverstein (2010, Theorem 5.11). □

LEMMA 2: *For symmetric positive semi-definite* $p \times p$ *matrices A and B*:

$$\begin{aligned} \xi_i(A) \leq & \xi_i(A+B); \\ \xi_{i+k-1}(A+B) \leq & \xi_i(A) + \xi_k(B) \\ & \text{for } i+k \leq p+1. \end{aligned}$$

Proof: See Anderson and Das Gupta (1963, Theorem 2.3). □

LEMMA 3: *For symmetric positive semi-definite* $p \times p$ *matrices A and B*:

$$\begin{aligned} \xi_i(AB) \leq & \xi_j(A)\xi_k(B); \\ \xi_{p-i+1}(AB) \geq & \xi_{p-j+1}(A)\xi_{p-k+1}(B) \\ & \text{for } j+k-1 \leq i. \end{aligned}$$

Proof: See Anderson and Das Gupta (1963, Corollary 2.2.1). □

LEMMA 4: *For a symmetric positive definite matrix A of dimension* $p \times p$:

$$\begin{aligned} \xi_1(A) \leq & \operatorname{tr}(A); \\ \xi_p(A) \geq & 1/\operatorname{tr}(A). \end{aligned}$$

Proof: For the first inequality, we have from Lemma 3

$$\xi_1(A) \leq \sum_{i=1}^{p} \xi_i(A) = \operatorname{tr}(A),$$

For the second inequality, note that symmetric positive semi-definite matrices can be diagonalized, i.e. $A = QDQ^{-1}$, where D is a matrix with the eigenvalues of A on the diagonal and Q is orthogonal such that $A^{-1} = QD^{-1}Q^{-1}$. Since A is positive definite, it has all eigenvalues greater than zero. This implies that the reciprocal of the largest eigenvalue of A^{-1} corresponds to the smallest eigenvalue of A and *vice versa*:

$$\xi_p(A) = \xi_1\left(A^{-1}\right) = 1/\xi_1(A),$$

using this and multiplying out the first inequality implies gives the desired result:

$$1/\mathrm{tr}\,(A) \leq 1/\xi_1\,(A) = \xi_p\,(A)\,.$$

□

LEMMA 5: *For symmetric $p \times p$ matrices A and B*,

$$\max_i |\xi_i\,(A) - \xi_i\,(B)|^2 \leq \|A-B\|^2 = \mathrm{tr}\,(A-B)\,(A-B)'\,.$$

Proof: See Hallin and Liska (2008), Lemma A.1.

LEMMA 6: Hoeffding's Lemma: *for a random variable x with bounded support (a, b) and mean μ,*

$$P\,(x-\mu > c) \leq \exp\left[2c^2/\,(b-a)\right].$$

Proof: See Hoeffding (1963).

LEMMA 7: *Derivation of equations* (4.11) *and* (4.12)*:*

To derive (4.11), note that the first order condition with respect to f_r' after concentrating out $\Lambda'\Lambda/N = I_R$ is:

$$\frac{\partial V}{\partial f_r'} = \left(N + \frac{\phi\sqrt{NT}}{\|f_r\|}\right) f_r' - Y'\lambda_r.$$

Holding fixed f_r,

$$f_r' = \left(N + \frac{\phi\sqrt{NT}}{\|f_r\|}\right)^{-1} Y'\lambda_r,$$

which implies that:

$$\begin{aligned} \|f_r\| &= \left(\frac{\|f_r\|}{N\|f_r\| + \phi\sqrt{NT}}\right) \left\|Y'\lambda_r\right\| \\ &\Leftrightarrow \\ N\|f_r\|^2 + \|f_r\|\,\phi\sqrt{NT} &= \|f_r\| \left\|Y'\lambda_r\right\| \\ \|f_r\| &= N^{-1}\left\|Y'\lambda_r\right\| - N^{-1}\phi\sqrt{NT}. \end{aligned}$$

The last line above can be rewritten to substitute $\left\|\frac{Y'\lambda^r}{\sqrt{NT}}\right\|$ for $\|f^r\|$ and yield (4.11), i.e.

$$
\begin{aligned}
Y'\lambda_r &= \left(N + \frac{\phi\sqrt{NT}}{N^{-1}\left\|Y'\lambda_r\right\| - N^{-1}\phi\sqrt{NT}}\right) f_r' \\
&= N\left(1 + \frac{\phi\sqrt{NT}}{\left\|Y'\lambda_r\right\| - \phi\sqrt{NT}}\right) f_r' \\
f_r' &= \frac{Y'\lambda_r}{N}\left(1 + \frac{\phi\sqrt{NT}}{\left\|Y'\lambda_r\right\| - \phi\sqrt{NT}}\right)^{-1}.
\end{aligned}
$$

For equation (4.12), note that the matrix containing all f_r' may be written as:

$$
\begin{aligned}
F' &= \frac{Y'\Lambda}{N}\begin{bmatrix}
\left(1 - \frac{\phi}{\left\|Y'\lambda^1/\sqrt{NT}\right\|}\right)_+ & 0 & \dots & 0 \\
0 & \dots & & \dots \\
\dots & & \dots & 0 \\
0 & \dots & 0 & \left(1 - \frac{\phi}{\left\|Y'\lambda^r/\sqrt{NT}\right\|}\right)_+
\end{bmatrix} \\
&:= \frac{Y'\Lambda}{N} - \frac{Y'\Lambda}{N}A.
\end{aligned}
$$

Then expanding the objective function using the normalisation $\Lambda'\Lambda/N = I_R$ and noting that trace $(A\Lambda'YY'\Lambda A)$ $= R \times \phi$, which is independent of the optimization variables, then yields the result.

Proof of Theorem 4.1:

We first show that the first R eigenvalues of $\hat{\Xi}$ are converge to $(NT)^{-1}\Lambda FF'\Lambda'$ in probability. To this end we present the conclusions of Lemma A.10 and A.11 of Ahn and Horenstein (2014). Under the normalization that $N \leq T$, the first $k = 1, \dots, R$ largest eigenvalues are bounded by:

$$
\begin{aligned}
\sum_{k=1}^{R}\hat{\xi}_k &\leq \sum_{k=1}^{R} v_k\left[FF'\Lambda\Lambda'/(NT)\right] + O_p\left(T^{-1}\right) + O_p\left(T^{-1/2}\right) \\
\sum_{k=1}^{R}\hat{\xi}_k &\geq \sum_{k=1}^{R} v_k\left[FF'\Lambda\Lambda'/(NT)\right] + O_p\left(T^{-1}\right) + O_p\left(T^{-1/2}\right).
\end{aligned}
$$

By Lemma 3, we have:

$$
\begin{aligned}
&v_R\left(FF'/T\right)v_R\left(\Lambda'\Lambda/N\right) \leq v_R\left[FF'\Lambda\Lambda'/(NT)\right] \\
&\quad \leq v_1\left[FF'\Lambda\Lambda'/(NT)\right] \leq v_1\left(FF'/T\right)v_1\left(\Lambda'\Lambda/N\right),
\end{aligned}
$$

which immediately implies that the $\hat{\xi}_k(YY') = O_p(NT)$ for all $k = 1, \ldots, R$, since by Lemma 4:

$$\begin{aligned} 1/\text{tr}\left[FF'\Lambda\Lambda'/(NT)\right] \leq & v_R\left[FF'\Lambda'\Lambda/(NT)\right] \\ \leq & v_1\left[FF'\Lambda'\Lambda/(NT)\right] \leq \text{tr}\left[FF'\Lambda\Lambda'/(NT)\right]. \end{aligned}$$

Furthermore, since the F_t and Λ_i have finite second moments, R is finite, $\text{tr}[FF'\Lambda\Lambda'/(NT)] = O_p(1)$ and thus $1/\text{tr}[FF'\Lambda\Lambda'/(NT)] = O_p(1)$. Since eigenvalues are continuous functions of the elements of the data matrix and $T^{-1}\sum_{t=1}^{T} F_t F_t' \to_p \Sigma_F > 0$ and similarly for Σ_Λ by Assumption 2, the continuous mapping theorem implies $\hat{\xi}_k \to_p v_k(\cdot)$ for $k = 1, \ldots, R$. This gives (4) and (5) of Theorem 1.

For $k = R+1, \ldots, N$, using Lemma 2, we can show:

$$\begin{aligned} \hat{\xi}_{R+1}\left[YY'/(NT)\right] \leq & \hat{\xi}_{R+1}\left[\Lambda FF'\Lambda'/(NT)\right] + \hat{\xi}_1\left[\varepsilon\varepsilon'/(NT)\right] \\ = & \hat{\xi}_1\left[\varepsilon\varepsilon'/(NT)\right], \end{aligned}$$

since $\text{rank}(\Sigma_F) = R$, $v_{R+j}(\Sigma_F\Sigma_\Lambda) = 0$ a.s. for $j = 1, 2, \ldots$. Similarly:

$$\begin{aligned} \xi_{R+1}\left[\hat{\Omega}/N\right] \leq & \xi_{R+1}\left[\Lambda FF'\Lambda'/(NT) + \varepsilon\varepsilon'/(NT)\right] \\ = & \xi_{R+1}\left(\hat{\Xi}\right) = \hat{\xi}_{R+1}\left[\varepsilon\varepsilon'/(NT)\right], \end{aligned}$$

which implies that:

$$\hat{\xi}_{R+1}\left[\varepsilon\varepsilon'/(NT)\right] \leq \xi_{R+1}\left(\hat{\Omega}/N\right) \leq \hat{\xi}_1\left[\varepsilon\varepsilon'/(NT)\right].$$

Alternatively we can show that:

$$\begin{aligned} \hat{\xi}_{R+i}\left[\varepsilon\varepsilon'/(NT)\right] = & \hat{\xi}_{R+i}\left[\varepsilon M_F\varepsilon'/(NT) + \varepsilon P_F\varepsilon'/(NT)\right] \\ \leq & \hat{\xi}_i\left[\varepsilon M_F\varepsilon'/(NT)\right] + \hat{\xi}_{R+i}\left[\varepsilon P_F\varepsilon'/(NT)\right] \\ = & \hat{\xi}_i\left[\varepsilon M_F\varepsilon'/(NT)\right], \end{aligned}$$

for matrices such that $P_F = F'(FF')^{-1}F$, $M_F = I - P_F$ and $\xi_{R+i}[\varepsilon P_F\varepsilon'/(NT)] = 0$ for $i \geq 1$ since the rank of P_F is R. Similarly for any $i > R$,

$$\begin{aligned} \hat{\xi}_i\left[\varepsilon M_F\varepsilon'/(NT)\right] \leq & \hat{\xi}_i\left[\varepsilon M_F\varepsilon'/(NT) + \varepsilon P_F\varepsilon'/(NT)\right] \\ = & \hat{\xi}_i\left[\varepsilon\varepsilon'/(NT)\right], \end{aligned}$$

so that we obtain the bound above again:

$$\hat{\xi}_{R+i}\left[\varepsilon\varepsilon'/(NT)\right] \leq \hat{\xi}_i\left[\varepsilon M_F\varepsilon'/(NT)\right] \leq \hat{\xi}_i\left[\varepsilon\varepsilon'/(NT)\right]$$

and for the N-th eigenvalue we have also:

$$\hat{\xi}_N\left[\varepsilon\varepsilon'/(NT)\right] \leq \hat{\xi}_N\left[\Lambda FF'\Lambda'/(NT)+\varepsilon\varepsilon'/(NT)\right] \leq \hat{\xi}_{N-R}\left[\varepsilon\varepsilon'/(NT)\right].$$

Then, since $R/N \to 0$, these bounds are tight and we have shown that all $R+i$th eigenvalues of $N\hat{\Xi}$ are bounded by those of Ω, i.e. for $i = 1,\ldots,N$, $\hat{\omega}_i\left[\varepsilon\varepsilon'/T\right]$. This result then leads immediately to $\hat{\xi}_{R+1}\left(N\hat{\Xi}\right) \to_p \omega_1(\Omega)$. Note that:

$$\begin{aligned}\omega_1\left[\varepsilon\varepsilon'/(NT)\right] =& \omega_1\left[G^{1/2}uHu'G^{1/2\prime}/(NT)\right] \\ \leq& N^{-1}\omega_1\left(uu'/T\right)\omega_1(G)\,\omega_1(H),\end{aligned}$$

by repeated use of Lemma 3. Now taking limits in this bound gives the required result:

$$\begin{aligned}N^{-1}\lim\omega_1\left(uu'/T\right)\omega_1(G)\,\omega_1(H) =& N^{-1}\left(1+\sqrt{\gamma}\right)^2\sigma_u^2 c_{G_+}c_{H_+} \\ :=& N^{-1}\left(1+\sqrt{\gamma}\right)^2\sigma_u^2 c_+ = O\left(N^{-1}\right)\end{aligned}$$

by Lemma 1 and Assumption 3.[13] For the smallest eigenvalue, similarly:

$$\begin{aligned}\omega_N\left[\varepsilon\varepsilon'/(NT)\right] =& \omega_N\left[G^{1/2}uHu'G^{1/2\prime}/(NT)\right] \\ \geq& N^{-1}\lim\left[\omega_N\left(uu'/T\right)\right]\omega_N(G)\,\omega_N(H) \\ :=& N^{-1}\left(1-\sqrt{\gamma}\right)^2\sigma_u^2 c_- = O\left(N^{-1}\right),\end{aligned}$$

and $c_- := c_{G-}c_{H-}$. The above inequalities thus bound the spectrum of $\varepsilon\varepsilon'/(NT)$ and prove that the eigenvalues vanish as $N, T \to \infty$, yielding (2) and (3) in the theorem. □

Proof of Corollary 4.1:

From Theorem 4.1 we have that for $k = 1,\ldots,R$,

$$\xi_k\left[YY'/(NT)\right] = v_k\left[FF'\Lambda\Lambda'/(NT)\right] + o_p(1),$$

[13] Lemma A.1 of AH is not strictly correct for the assumptions they provide: They posit that for an i.i.d. matrix, $\lambda_1(N\times\Omega) \underset{p}{\to} \left(1+\sqrt{\frac{N}{T}}\right)^2$. From Lemma 1, we can see that this is only true when $\sigma_u^2 = 1$.

and since $FF'\Lambda\Lambda'/(NT)$ is PD by assumption, all eigenvalues are non-negative. Then for any sample $Y^i \in Y^j$ with $N_i \approx N_j$, let A correspond to the matrix equal to the rows deleted from Y^j when going from sample j to i and $b_k^{(\cdot)}$ be the eigenvector corresponding to the k-th eigenvalue based on the $(\cdot)$-th sub-sample. By variational characterization of eigenvalues, the mini-max theorem then gives:

$$\begin{aligned} v_k^i\left[\left(FF'\Lambda\Lambda' - AA'\right)/(N_iT)\right] =& b_k^{i\prime}\left[\left(FF'\Lambda\Lambda' - AA'\right)/(N_iT)\right] b_k^i \\ \leq& b_k^{j\prime}\left[FF'\Lambda\Lambda'/(N_jT)\right] b_k^j - b_k^{j\prime}\left[AA'/(N_jT)\right] b_k^j, \\ \leq& v_k^j\left[FF'\Lambda\Lambda'/(N_jT)\right]. \end{aligned}$$

This gives immediately that $v_k^i(\cdot) \leq v_k^j(\cdot) \leq \cdots \leq v_k^J(\cdot)$ and, using Lemma 2, we have that $S = 0$ and $\hat{R} = 0$ for any penalty function larger than c_5 where

$$v_1^J\left[FF'\Lambda\Lambda'/(NT)\right] \leq v_1(\Sigma_F)\, v_1(\Sigma_\Lambda) < c_5$$

By the same argument we have that:

$$v_R(\Sigma_F)\, v_R(\Sigma_\Lambda) \leq v_R^1\left[FF'\Lambda\Lambda'/(NT)\right],$$

where $v_R^1(\cdot)$ is the last factor eigenvalue of the smallest sub-sample. Furthermore, by Lemma 5 we have for any i and j:

$$\begin{aligned} \max_{k=1,\ldots,R}\left|v_k^j - v_k^i\right|^2 \leq& \left\|FF'\left(\Lambda\Lambda'\right)_j/(N_jT) - FF'\left(\Lambda\Lambda'\right)_i/(N_iT)\right\|^2 \\ \leq& \left\|\left(FF'/T\right)\right\|^2\left\{\left\|\left(\Lambda\Lambda'\right)_j/N_j - \Sigma_\Lambda\right\|^2 + \left\|\Sigma_\Lambda - \left(\Lambda\Lambda'\right)_i/N_i\right\|^2\right\} \\ =& o_p(1), \end{aligned}$$

since (i) $\operatorname{rank}\left[FF'\Lambda\Lambda'/(NT)\right] = R$, (ii) $FF'/T \to_p \Sigma_F$ and $\left(\Lambda\Lambda'\right)_j/N_j \to_p \Sigma_\Lambda$ for any j as $\min_j N_j, T \to \infty$ and (iii), $\|A\|^2 = \operatorname{tr}(AA') \leq \operatorname{tr}(A)^2$ for square matrices A. Taking expectations in the above, the Markov inequality then implies that there is N_j and T such that for any $k = 1, \ldots, R$,

$$P\left(\left|v_k^j - v_k^i\right| > c\right) \to 0$$

for any fixed c independent of T. The former implies that for any penalty function in the interval (c_3, c_4) we have that $\hat{R} < R$ and $S \geq 0$, with strict equality in the limit.

For c_3, first note that by the Cauchy-interlacing Lemma, we have that $\omega_{k+1}^j(\cdot) \leq \omega_k^i(\cdot) \leq \omega_k^j(\cdot)$ for the principal minors of covariance matrices constructed from $Y^i \in Y^j \in \cdots \in Y^J$. This fact combined with the

bounds in Theorem 4.1 gives:

$$\begin{aligned} N_1^{-1}\left(1+\sqrt{\gamma_1}\right)^2\sigma_u^2 c_+ \leq & N_2^{-1}\left(1+\sqrt{\gamma_2}\right)^2\sigma_u^2 c_+ \\ & \leq \cdots < N_{J-1}^{-1}\left(1+\sqrt{\gamma_{J-1}}\right)^2\sigma_u^2 c_+ \leq N^{-1}\left(1+\sqrt{\gamma}\right)^2\sigma_u^2 c_+. \end{aligned}$$

Since each residual eigenvalue is $o_p(1)$, whilst the smallest factor eigenvalue is $O_p(1)$ regardless of the subsampling in the limit, the stability interval with $\hat{R}=R$ and $S=0$ is bounded by:

$$N^{-1}\left(1+\sqrt{\gamma}\right)^2\sigma_u^2 c_+ < c_3 < v_R^1\left[FF'\Lambda\Lambda'/(NT)\right].$$

By a similar argument to the above we have by Theorem 4.1 that:

$$\begin{aligned} 0 \leq c_1 < N^{-1}\left(1-\sqrt{\gamma}\right)^2\sigma_u^2 c_- \leq & N_{J-1}^{-1}\left(1-\sqrt{\gamma_{J-1}}\right)^2\sigma_u^2 c_- \\ & \leq \cdots \leq N_1^{-1}\left(1-\sqrt{\gamma_1}\right)^2\sigma_u^2 c_-, \end{aligned}$$

suggesting an interval where $\hat{R}=R_{\max}$ and $S=0$.

For c_2 note that the support of the residual eigenvalues shrinks as a result of the bounds above with the principal minor under consideration. Denote by $s_{(\cdot)}$ the supports corresponding to covariance matrices based on $Y^1 \in \cdots \in Y^j \in \cdots \in Y^J$ and let $\sigma = E\left(\omega_i^j\right)$ be the grand mean of the residual eigenvalues, where it is observed that $\sigma = \lim_{N,T\to\infty}\mathrm{tr}\left[\varepsilon\varepsilon'/\left(N^2T\right)\right] = O\left(N^{-1}\right)$ is independent of the sub-sample under consideration. By Hoeffding's Lemma, the probability that any i-th eigenvalue of the j-th sample is larger than c is bounded by:

$$P\left(\omega_i^j - \sigma > c\right) \leq \exp\left[-2c^2/s_j\right].$$

Note that the probability above goes to zero for any i, j and c because $s_j = O_p\left(N_j^{-1}\right)$ on account of the limits of the bounds of the eigenvalues of each sub-sample. This implies that in the limit, no instability interval c_2 exists. However, since $s_j = O_p\left(N^{-1}\right)$ we have furthermore that:

$$P\left(\sqrt{N_j}\left\{\omega_i^j(\cdot)-\sigma\right\} > c\right) \leq \exp\left[-2c^2/(N_j s_j)\right] := \eta_j$$

and observe that $\eta_1 \leq \cdots \leq \eta_J$ as a result of the scaled bound $N_j s_j$. Now for the result that $\hat{R} > R$ with $S=0$, we require the following probability over J sub-samples to approach unity for some N_j, T and c:

$$P\left(\sum_{j=1}^{J} I_{\sqrt{N_j}\left\{\omega_i^j(\cdot)-\sigma\right\}>c}\right) \leq \min_j P\left(\sqrt{N_j}\left\{\omega_i^j(\cdot)-\sigma\right\} > c\right) = \eta_j \neq 1$$

since c is fixed and $N_j s_j = O_p(1)$. □

Proof of Corollary 4.2:

The proof of the second claim follows if we can show that the first R eigenvalues are $O_p(1)$ and the remainder are $o_p(1)$. The first claim follows immediately if we can show that Lemmas A.8 and A.10 of AH hold using the appropriate scaling. For the latter, let B be an eigenvector of $\hat{\Xi}$, normalized so that $B'B = I \times N$, then we have for the first $k = 1, \ldots, R$ eigenvalues:

$$\begin{aligned}\left(N^2T^\delta\right)^{-1}\left|\mathrm{tr}\left(B'\Lambda F\varepsilon'B\right)\right| \leq & T^{1/2-\delta}\left\|B/N^{1/2}\right\|^2\left\|\Lambda/N^{1/2}\right\|\left\|F\varepsilon/(NT)^{1/2}\right\| \\ = & O_p\left(T^{1/2-\delta}\right).\end{aligned}$$

The first term is unity, and the second and third terms are bounded by Assumption 2, so the cross-term vanishes as long as $\delta > 1/2$. Furthermore we have:

$$\begin{aligned}\left(N^2T^\delta\right)^{-1}\mathrm{tr}\left(B'\varepsilon F'\left(FF'\right)^{-1}F\varepsilon'B\right) = & T^{1-2\delta}\mathrm{tr}\left[\left(BB'/N\right)\left(\varepsilon F'/\sqrt{NT}\right)O_p(1)\left(F\varepsilon'/\sqrt{NT}\right)\right] \\ \leq & T^{1-2\delta}\left\|B/N^{1/2}\right\|^2 O_p(1)\left\|F\varepsilon/(NT)^{1/2}\right\|^2 \\ = & O_p\left(T^{1-2\delta}\right),\end{aligned}$$

where the $O_p(1)$ in the middle of the first line is by assumption of the weak factor process, $T^{-\delta}\sum_{t=1}^{T}F_tF_t' \to_p \Sigma_F$. Note that both these cross-terms are rank R and will dominate the first R eigenvalues if $\delta \leq 1/2$ only, showing separation continues to hold in that situation.

For Lemma A.8, we have from the proof of Theorem 1 that:

$$\xi_{R+i}\left[\varepsilon\varepsilon'/\left(NT^\delta\right)\right] \leq \xi_i\left[\varepsilon M_F\varepsilon'/\left(NT^\delta\right)\right] \leq \xi_i\left[\varepsilon\varepsilon'/\left(NT^\delta\right)\right].$$

Setting $N = N(T)$ and $T^{1-\delta} = o(N)$, this implies that:

$$\begin{aligned}\xi_i\left[\varepsilon\varepsilon'/\left(NT^\delta\right)\right] = & T/\left(NT^\delta\right)\xi_i\left[\varepsilon\varepsilon'/T\right] \\ \leq & T^{1-\delta}/N\left(1+\sqrt{\gamma}\right)^2\sigma_u^2c_+ \\ = & O\left(T^{1-\delta}/N\right),\end{aligned}$$

so that the eigenvalues of the error covariance matrix vanish only when $T^{1-\delta} = o(N)$. □

Proof of Corollary 4.3:

For the first part, we proceed as in the proof of Theorem 4.1. We use the inequalities of Lemma 3 to

show convergence in probability of the factor eigenvalues to:

$$\hat{\xi}_1\left[\hat{\Xi}\right] = v_1\left[\Sigma_F \Sigma_\Lambda\right] + O_p\left(T^{-1/2}\right)$$

and

$$\hat{\xi}_R\left[\hat{\Xi}\right] = v_R\left[\Sigma_F \Sigma_\Lambda\right] + O_p\left(T^{-1/2}\right)$$

Since $\sqrt{N}\Lambda$ is an eigenvector of $YY'/(NT)$, without loss of generality we normalise it to be orthonormal, i.e., $\Lambda'\Lambda/N = I_N$. With this normalization it is immediate from the above that the first R eigenvalues of $\hat{\Xi}$ are completely determined by the eigenvalues of the $R-$dimensional covariance matrix FF'/T. Since $R \to \infty$, $R/T \to \tau$ and $F = G_F u_F H_F$ by assumption, the eigenvalues of the limiting matrix Σ_F are then bounded by:

$$v_k\left(T^{-1}FF'\right) = v_k\left(T^{-1}G_F^{1/2}u_F H_F u_F' G_F^{1/2\prime}\right) \gtreqless \left(1 \mp \sqrt{\tau}\right)^2 \sigma_{u_F}^2 c_{F\mp},$$

for $k = [R, 1]$, where we have used the trace inequality repeatedly and used Lemma 1 for $u_F u_F'/T \to_p \left(1 \mp \sqrt{\tau}\right)^2 \sigma_{u_F}^2$. Clearly, $v_k = O_p(1)$ for $k = [R, 1]$. On the other hand, the bound on the extremal eigenvalue of the noise matrix is as in the proof of Theorem 4.1 with the modification that, as $R \to \infty$, the $R+1$-th eigenvalue is no longer tightly bounded at the largest eigenvalue of Ω, but rather only bounded above by $N^{-1}\omega_1(\cdot)$, i.e.:

$$\hat{\xi}_{R+1}\left[\varepsilon\varepsilon'/(NT)\right] \leq N^{-1}\omega_{R+1}\left(\hat{\Omega}\right) \leq \hat{\xi}_1\left[\varepsilon\varepsilon'/(NT)\right].$$

Since $R/T \to \tau$ and $R < N < T$, the lower bound on $N\xi_{R+1}(\cdot)$ can be far detached from the edge $\omega_1(\Omega)$.

For the second claim, first note that since $R \to \infty$, but $R/T \to \tau$ and $N = N(T)$, we now have R eigenvalues in a finite support, so that the spacing between the eigenvalues shrinks to zero with R. This leads to an argument in congruence with the interval (c_1, c_2) in Theorem 4.1 for the factor portion of the eigenvalue structure and stability regions no longer exist. To show that a separating region exists, note that the R-th eigenvalue of $\hat{\Xi}$, corresponding to the factors, based on J subsamples is bounded below by the smallest subsample:

$$\begin{aligned}\left(1 - \sqrt{\tau}\right)^2 \sigma_{u_F}^2 c_{F-} &> \left(1 - \sqrt{\tau_{J-1}}\right)^2 \sigma_{u_F}^2 c_{F-} > \cdots > \left(1 - \sqrt{\tau_2}\right)^2 \sigma_{u_F}^2 c_{F-} \\ &> \left(1 - \sqrt{\tau_1}\right)^2 \sigma_{u_F}^2 c_{F-} > c_3 = O_p(1),\end{aligned}$$

where $\tau_j = R/T_j$ with $j = 1, \ldots, J$ since $T = N^{-1}(T)$ for any truncated sample. Applying Hoeffding's inequality again as in the proof of Corollary 4.1 gives the result for the interval (c_3, c_4) and $IC_{\hat{R}}$ with $p(N, T) \in (c_2, c_3)$ consistently estimates R as claimed. □

Proof of Proposition 4.1:

First note that the estimator of the bound can be deconstructed as:

$$\hat{\xi}^{+} = J^{-1}\left(\hat{\xi}_k + \frac{\hat{\sigma}_k^2}{\hat{\sigma}_{k+1}^2}\hat{\xi}_{k+1} + \frac{\hat{\sigma}_k^2}{\hat{\sigma}_{k+2}^2}\hat{\xi}_{k+2}\cdots + \frac{\hat{\sigma}_k^2}{\hat{\sigma}_{k+J}^2}\hat{\xi}_{k+J},\right)$$

and that the variance estimates are all $O_p(1)$. Consistency of $\hat{\sigma}_k^2$ for the trace element of the estimator is by standard arguments which means that c_+ is tautologically defined by the bound.

The proof then follows immediately from Theorem 4.1: for any $k < R$, we have that $\hat{\xi}_k - \hat{\xi}^{+} = O_p(1)$, whereas $\delta = o_p(1)$, giving that $P\left(\hat{\xi}_k - \hat{\xi}^{+} < \delta\right) \to 0$. For $k > R$, first note that $N\hat{\xi}_{R+1} \to_p \omega_1(\Omega) = O_p(1)$ and $N\hat{\xi}_N \to_p \omega_N(\Omega) = O_p(1)$ by Theorem 4.1. Now since $\delta = O\left(N^{-(1+a)}\right)$, we have that:

$$P\left(\hat{\xi}_{R+1} - \hat{\xi}^{+} > \delta\right) = P\left[N\left(\hat{\xi}_{R+1} - \hat{\xi}^{+}\right) > N\delta\right] = P\left[O_p(1) > O_p\left(N^{-a}\right)\right] \to 1$$

and thus $P\left(\hat{R} = R\right) \to 1$. □

Proof of Proposition 4.2:

From Theorem 4.1, we have for $k < R$ that $k \times \xi_k = O_p(1)$ and $k \times \xi_k = o_p(1)$ for $k > R$. Then by definition of argmin, we have that $P(k < R) \to 0$. For $k > R$, note that this implies that MKT is declining at a point larger than $R+1$, i.e. $(R+2)\xi_{R+2} - (R+1)\xi_{R+1} < 0$ or that:

$$N\xi_{R+1} < (R+2)N(\xi_{R+1} - \xi_{R+2}) \Rightarrow O_p(1) < (R+2)o_p(1),$$

since $N(\xi_{R+1} - \xi_{R+2}) = [N\xi_{R+1} - \omega_1(\Omega)] + [\omega_1(\Omega) - N\xi_{R+2}] < \delta$ for arbitrary δ as long as $R/N \to 0$ and this implies that $P(k > R) \to 0$. In other words, MKT must be increasing at $R+1$ and thus $P\left(\hat{R} = R\right) \to 1$. □

Proof of Proposition 4.3:

For the first part, first consider the case when $k > R$, we want to show that

$$P(k > R) \to 0$$

as $N, T \to \infty$. Let

$$\begin{aligned} P(k > R) =& P\left(\left[1 - \frac{\phi}{\left\|\lambda_{\hat{R}}' Y/\sqrt{NT}\right\|}\right]_+ > 0\right) \\ =& P\left(\left\|\lambda_k' Y/\sqrt{NT}\right\| > \phi\right), \end{aligned}$$

when $k = R + s > R$, using Lemma 3 we have that:

$$\xi_k \left[(NT)^{-1} \Lambda F' F \Lambda' + (NT)^{-1} \varepsilon\varepsilon' \right] \leq \xi_{R+1} \left[(NT)^{-1} \Lambda F' F \Lambda' \right] + \xi_s \left[(NT)^{-1} \varepsilon\varepsilon' \right]$$

for $s = 1, \ldots N$. Since $\operatorname{rank} \left[(NT)^{-1} \Lambda F' F \Lambda' \right] = R$, $\xi_{R+s} \left[(NT)^{-1} \Lambda F' F \Lambda' \right] = 0$, the k-th eigenvalue of $YY' / (NT)$ is no larger than the bound on $N^{-1} \omega_1 (\Omega)$ by Theorem 4.1. This implies that:

$$P \left(\xi_s > \xi^+ \right) \to 0.$$

as $N, T \to \infty$.

For the converse we want:

$$\begin{aligned} P(k < R) =& P \left(\left[1 - \frac{\phi}{\left\| \lambda_k' Y / \sqrt{NT} \right\|} \right]_+ \leq 0 \right) \\ =& P \left(\left\| \lambda_k' Y / \sqrt{NT} \right\| \leq \phi \right), \end{aligned}$$

as $N, T \to \infty$. Since $\left\| \lambda_k' Y / \sqrt{NT} \right\| = O_p(1)$, we have by Lemma 3:

$$\begin{aligned} \xi_k \left[(NT)^{-1} \Lambda F' F \Lambda' + (NT)^{-1} \varepsilon\varepsilon' \right] \leq& \xi_k \left[(NT)^{-1} \Lambda F' F \Lambda' \right] + \xi_1 \left[(NT)^{-1} \varepsilon\varepsilon' \right] \\ =& O_p(1) + o_p(1) \end{aligned}$$

when $k \leq R$, implying that

$$P(k < R) = P(O_p(1) < o_p(1)) \to 0.$$

The second part of the Proposition is then a consequence of Algorithm 4.2 in conjunction with the proof of part 1.□

Proof of Theorem 4.2:

We first expand the covariance matrix of $Y - \left(I \otimes \hat{\beta}' \right) X$:

$$\begin{aligned} (NT)^{-1} \hat{W} \hat{W}' =& (NT)^{-1} \left[\Lambda F - \left(I \otimes \hat{\Upsilon}' \right) X + \varepsilon \right] \left[\Lambda F - \left(I \otimes \hat{\Upsilon}' \right) X + \varepsilon \right]' \\ =& (NT)^{-1} \Lambda F (\Lambda F)' - (NT)^{-1} \Lambda F \left[\left(I \otimes \hat{\Upsilon}' \right) X \right]' + (NT)^{-1} \Lambda F \varepsilon' \\ & - (NT)^{-1} \left(I \otimes \hat{\Upsilon}' \right) X (\Lambda F)' + (NT)^{-1} \left(I \otimes \hat{\Upsilon}' \right) X \left[\left(I \otimes \hat{\Upsilon}' \right) X \right]' \\ & - (NT)^{-1} \left(I \otimes \hat{\Upsilon}' \right) X \varepsilon' + (NT)^{-1} \varepsilon (\Lambda F)' - (NT)^{-1} \varepsilon \left(I \otimes \hat{\Upsilon}' \right) X \\ & + (NT)^{-1} \varepsilon\varepsilon'. \end{aligned}$$

Taking traces, this simplifies to:

$$\begin{aligned}(NT)^{-1}\operatorname{tr}\left(\hat{W}\hat{W}'\right) &= (NT)^{-1}\operatorname{tr}\left(\Lambda'\Lambda FF'\right) + (NT)^{-1}\operatorname{tr}\left(\left[I\otimes\hat{\Upsilon}'\right]X\left[\left(I\otimes\hat{\Upsilon}'\right)X\right]'\right)\\ &\quad -2(NT)^{-1}\operatorname{tr}\left(\Lambda F\left[\left(I\otimes\hat{\Upsilon}'\right)X\right]'\right)\\ &\quad +(NT)^{-1}\operatorname{tr}\left(\varepsilon\varepsilon'\right)+2(NT)^{-1}\operatorname{tr}\left(\Lambda F\varepsilon'\right)-2(NT)^{-1}\operatorname{tr}\left[\left(I\otimes\hat{\Upsilon}'\right)X\varepsilon'\right]\\ &= (i)+(ii).\end{aligned}$$

Using Theorem 4.1 and Assumptions 5 and 6, (i) is $O_p(1)$, whereas (ii) are $o_p(1)$. The latter follows from the fact that $\operatorname{rank}\left(\hat{W}\hat{W}'\right) = N$ but that $\operatorname{rank}(i) = \max(R,S)$. This implies by Lemmas 3 and 4 above for $k = \max(R,S)+1,\ldots,N$ that:

$$(NT)^{-1}\xi_k\left(\hat{W}\hat{W}'\right) = (NT)^{-1}\xi_k\left(\varepsilon\varepsilon'+2\Lambda F\varepsilon'-2\left[I\otimes\hat{\Upsilon}'\right]X\varepsilon'\right),$$

by repeating the steps of the proof of Theorem 4.1. All these terms are $o_p(1)$ by a combination of Assumptions 5 and 6 and the proof of Theorem 4.1. This gives the first part.

For the second and third claims, note that the rank of the combined first three terms is equal to $\max(R,S)$ if $\hat{\Upsilon} = O_p(1)$ or R if $\hat{\Upsilon} = o_p(1)$. This can be seen by noting that for $k = 1,\ldots,\max(R,S)$:

$$\begin{aligned}(NT)^{-1}\xi_k\left(\hat{W}\hat{W}'\right) = (NT)^{-1}\xi_k\Big(&\Lambda'\Lambda FF' + \left[I\otimes\hat{\Upsilon}'\right]X\left[\left(I\otimes\hat{\Upsilon}'\right)X\right]'\\ &-2\Lambda F\left[\left(I\otimes\hat{\Upsilon}'\right)X\right]'\Big)+o_p(1),\end{aligned}$$

the first term is of rank R; the second term is rank S and the third term is rank $\min(R,S)$ by analogy of the argument in Theorem 4.1. As a result, the eigenvalues of the error covariance matrix are dominated by the eigenvalues corresponding to the factor and the factors in X that stem from the inconsistency of $\hat{\beta}$ induced by underestimating R leading to $\max(R,S)$ non-vanishing eigenvalues. Finally for the third claim, note that if $\hat{\Upsilon} = o_p(1)$, the W collapses to the approximate factor model as in Theorem 4.1, so that the result follows. □

Appendix 4.2. Simulation Results

Table 4.1: Monte Carlo Results for approximate factor model with $\alpha = 0$, $\rho = 0$ and $\theta = 1$.

N	T	R^*	*1*					*5*					*10*				
			AH	*BN*	*ON*	*MKT*	*EDR*	*AH*	*BN*	*ON*	*MKT*	*EDR*	*AH*	*BN*	*ON*	*MKT*	*EDR*
50	50	*%corr*	100	100	97.9	100	100	100	100	99.8	97.4	100	99.2	99.9	99.6	33.7	99.4
		$\hat{R}$	1.0	1.0	1.03	1.0	1.0	5.0	5.0	5.0	5.1	5.0	10.0	10.0	10.0	12.7	10.0
		MSE	0.0	0.0	0.04	0.0	0.0	0.0	0.0	0.0	1.4	0.0	0.0	0.0	0.0	15.0	0.0
	100	*%corr*	100	100	98.1	100	100	100	100	99.7	100	100	100	100	100	93.2	100
		$\hat{R}$	1.0	1.0	1.02	1.0	1.0	5.0	5.0	5.0	5.0	5.0	10.0	10.0	10.0	10.2	10.0
		MSE	0.0	0.0	0.03	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.7	0.0
100	100	*%corr*	100	100	98.9	100	100	100	100	99.7	100	100	100	100	99.9	99.9	100
		$\hat{R}$	1.00	1.00	1.01	1.00	1.0	5.0	5.0	5.0	5.0	5.0	10.0	10.0	10.0	10.0	10.0
		MSE	0.00	0.00	0.02	0.00	0.0	0.0	0.0	0.01	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	200	*%corr*	100	100	99.2	100	100	100	100	99.8	100	100	100	100	100	100	100
		$\hat{R}$	1.0	1.0	1.01	1.0	1.0	5.0	5.0	5.0	5.0	5.0	10.0	10.0	10.0	10.0	10.0
		MSE	0.0	0.0	0.01	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
200	200	*%corr*	100	100	99.1	100	100	100	100	100	100	100	100	100	100	100	100
		$\hat{R}$	1.0	1.0	1.01	1.0	1.0	5.0	5.0	5.0	5.0	5.0	10.0	10.0	10.0	10.0	10.0
		MSE	0.0	0.0	0.02	0.0	0.0	0.0	0.0	0.00	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	400	*%corr*	100	100	99.2	100	100	100	100	99.7	100	100	100	100	99.9	100	100
		$\hat{R}$	1.0	1.0	1.01	1.0	1.0	5.0	5.0	5.0	5.0	5.0	10.0	10.0	10.0	10.0	10.0
		MSE	0.0	0.0	0.01	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
400	400	*%corr*	100	100	99.1	100	100	100	100	99.8	100	100	100	100	100	100	100
		$\hat{R}$	1.0	1.0	1.01	1.0	1.0	5.0	5.0	5.0	5.0	5.0	10.0	10.0	10.0	10.0	10.0
		MSE	0.0	0.0	0.01	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

* *% corr* is the percentage of correct $\hat{R}$, $\hat{R}$ is the average estimate of R and *MSE* is the mean squared error over the Monte Carlo trials.

Table 4.2: Monte Carlo Results for approximate factor model with $\alpha = 0$, $\rho = 0.5$ and $\theta = 1$.

N	T	R^*	*1*					*5*					*10*				
			AH	*BN*	*ON*	*MKT*	*EDR*	*AH*	*BN*	*ON*	*MKT*	*EDR*	*AH*	*BN*	*ON*	*MKT*	*EDR*
50	50	*%corr*	100	0.0	82.9	75.7	95.7	98.4	0.0	91.8	13.1	81.3	91.9	0.0	91.3	5.1	72.7
		$\hat{R}$	1.0	8.7	1.4	2.3	1.0	5.0	15.3	5.2	8.7	4.9	9.9	18.9	9.9	9.7	9.7
		MSE	0.0	63.6	1.7	8.8	0.0	0.1	112.8	0.8	27.0	0.2	0.3	78.7	1.7	26.9	3.2
	100	*%corr*	100	23.1	92.0	100	100	99.9	8.4	97.8	38.8	97.4	99.4	1.5	99.1	9.8	99.5
		$\hat{R}$	1.0	2.7	1.1	1.0	1.0	5.0	7.8	5.0	7.9	5.0	10.0	15.5	10.0	11.5	10.0
		MSE	0.0	5.3	0.4	0.0	0.0	0.0	11.9	0.1	20.2	0.0	0.0	37.2	0.0	20.7	0.1
100	100	*%corr*	100	0.7	96.6	100	100	100	0.2	98.6	67.9	100	100	0.0	99.7	19.3	99.9
		$\hat{R}$	1.0	6.4	1.0	1.0	1.0	5.0	11.3	5.0	6.7	5.0	10.0	17.1	10.0	11.8	10.0
		MSE	0.0	33.0	0.0	0.0	0.0	0.0	44.3	0.0	13.0	0.0	0.0	53.0	0.0	15.6	0.0
	200	*%corr*	100	97.3	97.3	100	100	100	94.2	99.4	99.1	100	100	89.6	99.7	82.8	100
		$\hat{R}$	1.0	1.0	1.0	1.0	1.0	5.0	5.1	5.0	5.0	5.0	10.0	10.1	10.0	10.4	10.0
		MSE	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.1	0.0	0.0	0.1	0.0	1.3	0.0
200	200	*%corr*	100	62.9	97.7	100	100	100	55.3	99.7	100	100	100	42.0	99.7	98.6	100
		$\hat{R}$	1.0	1.4	1.0	1.0	1.0	5.0	5.6	5.0	5.0	5.0	10.0	10.9	10.0	10.0	10.0
		MSE	0.0	0.6	0.0	0.0	0.0	0.0	0.9	0.0	0.0	0.0	0.0	1.6	0.0	0.0	0.0
	400	*%corr*	100	100	98.2	100	100	100	100	99.8	100	100	100	100	99.9	100	100
		$\hat{R}$	1.0	1.0	1.0	1.0	1.0	5.0	5.0	5.0	5.0	5.0	10.0	10.0	10.0	10.0	10.0
		MSE	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
400	400	*%corr*	100	100	98.8	100	100	100	100	99.7	100	100	100	100	100	100	100
		$\hat{R}$	1.0	1.0	1.0	1.0	1.0	5.0	5.0	5.0	5.0	5.0	10.0	10.0	10.0	10.0	10.0
		MSE	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

* *% corr* is the percentage of correct $\hat{R}$, $\hat{R}$ is the average estimate of R and *MSE* is the mean squared error over the Monte Carlo trials.

Table 4.3: Monte Carlo Results for approximate factor model with $\alpha = 0.5$, $J = \max(10, N/20)$ $\rho = 0$ and $\theta = 1$.

N	T	R^*	*1*					*5*					*10*				
			AH	*BN*	*ON*	*MKT*	*EDR*	*AH*	*BN*	*ON*	*MKT*	*EDR*	*AH*	*BN*	*ON*	*MKT*	*EDR*
50	50	*%corr*	61.4	0.0	0.0	0.0	0.1	12.1	0.0	0.0	1.6	0.1	3.4	0.0	0.0	0.3	38.2
		$\hat{R}$	1.9	19.0	4.4	3.9	2.7	7.1	19.0	8.2	7.8	7.0	12.3	19.0	13.0	9.7	8.9
		MSE	2.2	323.9	12.3	8.8	3.1	5.1	195.9	10.9	10.0	4.0	5.7	81.0	9.6	20.8	5.0
	100	*%corr*	64.7	0.0	0.0	0.0	0.00	11.3	0.0	0.0	1.9	0.1	2.2	0.0	0.0	0.3	42.1
		$\hat{R}$	1.9	19.0	4.5	4.0	2.9	7.2	19.0	8.3	7.8	6.8	12.4	19.0	13.1	9.8	10.3
		MSE	2.2	324.0	13	9	3.6	5.7	196.0	11.5	10.2	4.5	6.3	81.0	9.8	20.0	1.1
100	100	*%corr*	100	0.0	0.0	50.7	99.7	95.7	0.0	0.1	14.9	90.7	74.5	0.0	85.3	1.7	94.5
		$\hat{R}$	1.0	11.5	8	4.1	1.0	5.3	16.2	11.9	9.8	5.2	11.5	19.0	9.3	12.2	9.9
		MSE	0.0	113.2	48.7	20.7	0.0	1.6	127.5	47.3	33.4	1.2	9.7	80.5	9.6	37.0	0.3
	200	*%corr*	100	0.0	0.0	52.0	100	99.2	0.0	0.0	16.4	94.8	90.9	0.0	97.3	2.9	99.1
		$\hat{R}$	1.0	14.4	8.1	4.2	1.0	5.1	18.4	12.0	10.0	5.2	10.6	19.0	9.9	12.4	10.0
		MSE	0.0	182.3	50.4	22	0.0	0.4	179.8	49.6	34.3	0.8	3.8	81.0	1.8	35.8	0.0
200	200	*%corr*	100	0.0	95.7	97.8	100	100	0.0	99.9	66.7	100	100	0.0	100.0	14.6	100
		$\hat{R}$	1.0	16.7	1.5	1.3	1.0	5.0	19.0	5.0	8.2	5.0	10.0	19.0	10.0	13.1	10.0
		MSE	0.0	248.1	7.0	4.0	0.0	0.0	196.0	0.0	43.2	0.0	0.0	81.0	0.0	40.4	0.00
	400	*%corr*	100	0.0	99.4	100	100	100	0.0	100	78.4	100	100	0.0	100	29.3	100
		$\hat{R}$	1.0	17.5	1.1	1.0	1.0	5.0	19.0	5.0	7.8	5.0	10.0	19.0	10.0	12.6	10.0
		MSE	0.0	271.2	0.9	0.0	0.0	0.0	196.0	0.0	39.0	0.0	0.0	81.0	0.0	33.3	0.0
400	400	*%corr*	100	0.0	99.1	100	100	100	0.0	100	76.5	100	100	0.0	100	33.4	100
		$\hat{R}$	1.0	19.0	1.1	1.0	1.0	5.0	19.0	5.0	8.1	5.0	10.0	19.0	10.0	12.9	10.0
		MSE	0.0	324.0	1.2	0.0	0.0	0.0	196.0	0.0	43.4	0.0	0.0	81.0	0.0	31.3	0.0

* *% corr* is the percentage of correct $\hat{R}$, $\hat{R}$ is the average estimate of R and *MSE* is the mean squared error over the Monte Carlo trials.

Table 4.4: Monte Carlo Results for approximate factor model with $\alpha = 0.5$, $J = \max(10, N/20)$ $\rho = 0.5$ and $\theta = 1$.

N	T	R^*	*1*					*5*					*10*				
			AH	*BN*	*ON*	*MKT*	*EDR*	*AH*	*BN*	*ON*	*MKT*	*EDR*	*AH*	*BN*	*ON*	*MKT*	*EDR*
50	50	*%corr*	60.2	0.0	0.0	0.1	2.3	16.7	0.0	0.0	2.0	0.9	4.0	0.0	0.0	0.6	32.6
		$\hat{R}$	1.9	19.0	4.5	3.9	2.5	6.9	19.0	8.2	7.7	6.3	12.1	19.0	13.0	9.4	9.0
		MSE	2.2	324.0	13.3	8.6	2.5	4.7	196.0	10.7	10.3	5.7	5.3	81.0	9.0	22.6	5.2
	100	*%corr*	63.2	0.0	0.0	0.0	0.2	13.4	0.0	0.0	2.1	1.6	4.6	0.0	0.0	0.4	44.5
		$\hat{R}$	1.9	19.0	4.5	4.0	2.7	7.1	19.0	8.3	7.9	6.9	12.3	19.0	13.1	9.8	10.3
		MSE	2.1	324.0	13.0	9.0	3.2	5.3	196.0	11.5	10.4	4.0	5.7	81.0	9.8	20.3	1.1
100	100	*%corr*	99.6	0.0	0.0	45.6	97.7	92.1	0.0	1.9	9.9	86.4	70.7	0.0	80.1	1.5	90.1
		$\hat{R}$	1.0	17.7	7.8	4.1	1.0	5.5	18.9	11.5	10.1	5.2	11.7	19.0	9.1	12.0	9.8
		MSE	0.2	282.6	46.0	19.8	0.0	2.7	193.6	42.9	34.3	1.3	10.0	81.0	12.9	36.7	0.5
	200	*%corr*	100	0.0	0.0	53.6	99.9	98.6	0.0	0.2	15.0	93.9	86.9	0.0	95.0	2.8	97.9
		$\hat{R}$	1.0	18.2	8.0	4.0	1.0	5.1	19.0	11.9	10.1	5.2	10.8	19.0	9.9	12.2	10.0
		MSE	0.0	296.6	49.1	20.2	0.0	0.5	195.1	48.0	35.1	0.8	5.1	81.0	2.6	35.7	0.1
200	200	*%corr*	100	0.0	90.8	93.5	100	100	0.0	99.2	52.1	100	100	0.0	99.9	7.1	100
		$\hat{R}$	1.0	18.9	2.0	1.8	1.0	5.0	19.0	5.0	8.7	5.0	10.0	19.0	10.0	13.3	10.0
		MSE	0.0	320.2	12.1	10.6	0.0	0.0	196.0	0.0	45.7	0.0	0.0	81.0	0.0	41.8	0.0
	400	*%corr*	100	0.0	97.4	98.7	100	100	0.0	100	70.7	100	100	0.0	100	20.3	100
		$\hat{R}$	1.0	19.0	1.3	1.2	1.0	5.0	19.0	5.0	8.2	5.0	10.0	19.0	10.0	12.9	10.0
		MSE	0.0	323.2	4.0	2.2	0.0	0.0	196.0	0.0	44.0	0.0	0.0	81.0	0.0	34.3	0.0
400	400	*%corr*	100	0.0	97.4	99.3	100	100	0.0	99.9	67.5	100	100	0.0	100	24.4	100
		$\hat{R}$	1.0	19.0	1.3	1.1	1.0	5.0	19.0	5.0	8.6	5.0	10.0	19.0	10.0	13.4	10.0
		MSE	0.0	324.0	3.9	1.2	0.0	0.0	196.0	0.0	48.3	0.0	0.0	81.0	0.0	32.2	0.0

* *% corr* is the percentage of correct $\hat{R}$, $\hat{R}$ is the average estimate of R and *MSE* is the mean squared error over the Monte Carlo trials.

Table 4.5: Monte Carlo Results for approximate factor model with $\alpha = 0$, $\rho = 0$ and $\theta = 10$.

N	T	R^*	*1*					*5*					*10*				
			AH	*BN*	*ON*	*MKT*	*EDR*	*AH*	*BN*	*ON*	*MKT*	*EDR*	*AH*	*BN*	*ON*	*MKT*	*EDR*
50	50	*%corr*	98.9	64.0	95.5	40.7	95.5	19.7	1.6	28.5	47.8	36.3	1.7	0.0	0.3	4.8	3.9
		$\hat{R}$	1.1	0.6	1.0	7.8	1.1	2.8	2.3	2.6	4.7	5.1	3.8	3.5	0.8	10.2	6.0
		MSE	0.5	0.4	0.1	89.0	0.1	6.9	8.2	9.9	5.7	3.9	45.7	45.3	89.1	44.7	22.6
50	100	*%corr*	100	85.0	98.5	80.5	100	63.9	9.7	86.3	86.8	64.5	21.4	0.1	17.5	22.2	34.8
		$\hat{R}$	1.0	0.9	1.0	3.6	1.0	4.2	3.3	4.7	4.8	5.6	5.8	5.0	2.7	8.4	8.3
		MSE	0.0	0.2	0.0	35.9	0.0	2.5	3.9	1.0	0.6	2.4	29.8	27.6	71.0	16.2	11.3
100	100	*%corr*	100	99.1	98.9	99.5	100	98.2	66.7	99.7	98.2	88.7	90.9	8.3	98.0	46.8	92.7
		$\hat{R}$	1.0	1.0	1.0	1.1	1.0	5.0	4.6	5.0	5.0	5.1	9.7	8.2	9.9	8.3	9.9
		MSE	0.0	0.0	0.0	1.3	0.0	0.1	0.4	0.0	0.1	0.2	1.5	4.4	1.1	9.7	0.8
100	200	*%corr*	100	100	98.0	100	100	99.9	98.0	99.7	98.7	98.7	100	75.7	99.9	62.5	100
		$\hat{R}$	1.0	1.0	1.0	1.0	1.0	5.0	5.0	5.0	5.0	5.0	10.0	9.7	10.0	8.5	10.0
		MSE	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.0	0.0	0.3	0.0	7.4	0.0
200	200	*%corr*	100	100	98.7	100	100	100	100	99.6	100	100	100	100	99.9	74.2	100
		$\hat{R}$	1.0	1.0	1.0	1.0	1.0	5.0	5.0	5.0	5.0	5.0	10.0	10.0	10.0	9.0	10.0
		MSE	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	4.6	0.0
	400	*%corr*	100	100	99.6	100	100	100	100	99.7	100	100	100	100	100	82.2	100
		$\hat{R}$	1.0	1.0	1.0	1.0	1.0	5.0	5.0	5.0	5.0	5.0	10.0	10.0	10.0	9.3	10.0
		MSE	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2.8	0.0
400	400	*%corr*	100	100	99.6	100	100	100	100	99.8	100	100	100	100	100	90.7	100
		$\hat{R}$	1.0	1.0	1.0	1.0	1.0	5.0	5.0	5.0	5.0	5.0	10.0	10.0	10.0	9.7	10.0
		MSE	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.3	0.0

* *% corr* is the percentage of correct $\hat{R}$, $\hat{R}$ is the average estimate of R and *MSE* is the mean squared error over the Monte Carlo trials.

Table 4.6: Monte Carlo Results for approximate factor model with $\alpha = 0.5$, $J = \max(10, N/20)$ $\rho = 0.5$ and $\theta = 10$.

N	T	R^*	*1*					*5*					*10*				
			AH	*BN*	*ON*	*MKT*	*EDR*	*AH*	*BN*	*ON*	*MKT*	*EDR*	*AH*	*BN*	*ON*	*MKT*	*EDR*
50	50	*%corr*	5.5	0.0	0.0	0.0	0.0	1.2	0.0	0.3	0.0	1.8	0.9	0.0	0.0	0.0	5.3
		$\hat{R}$	4.3	19.0	5.1	15.5	4.0	7.9	19.0	8.2	16.9	8.6	12.2	19.0	6.8	17.6	6.7
		MSE	11.9	324.0	17.0	214.6	9.6	13.6	196.0	16.0	143.8	14.5	20.5	81.0	54.2	58.4	18.6
50	100	*%corr*	0.5	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.1	0.1	0.0	0.0	0.0	6.8
		$\hat{R}$	4.8	19.0	5.1	14.4	4.2	8.9	19.0	9.0	16.5	8.9	13.7	19.0	12.2	17.5	8.0
		MSE	14.6	324.0	17.0	189.1	10.5	15.4	196.0	16.4	134.7	15.7	15.9	81.0	25.5	56.9	17.9
100	100	*%corr*	0.6	0.0	0.6	0.0	3.6	0.0	0.0	0.0	0.0	4.6	0.2	0.0	0.0	0.0	4.2
		$\hat{R}$	8.4	17.0	8.4	15.1	6.4	12.5	18.8	9.8	17.1	8.7	17.7	19.0	0.1	18.0	12.1
		MSE	56.6	260.2	62.4	203.1	37.3	58.7	191.4	55.7	147.1	22.1	61.0	81.0	97.6	63.9	25.1
100	200	*%corr*	0.2	0.0	0.0	0.0	1.8	0.0	0.0	0.0	0.0	0.5	0.0	0.0	0.0	0.0	0.9
		$\hat{R}$	8.9	16.4	9.6	13.7	8.1	13.0	18.8	12.6	16.7	10.4	18.1	19.0	0.1	18.0	15.8
		MSE	63.4	241.4	75.1	168.1	56.1	65.3	190.7	70.4	138.4	31.9	66.3	81.0	98.3	63.9	35.6
200	200	*%corr*	4.0	0.0	12.7	0.0	26.5	0.1	0.0	0.0	0.0	2.8	0.8	0.0	0.0	0.0	1.3
		$\hat{R}$	16.3	19.0	0.2	17.9	4.5	17.2	19.0	0.1	17.9	7.8	13.0	19.0	0.1	16.8	15.3
		MSE	248.6	324.0	0.9	287.0	25.8	166.7	196.0	24.2	166.8	34.0	62.8	81.0	98.7	57.2	55.5
200	400	*%corr*	4.4	0.0	28.1	0.0	7.3	0.2	0.0	0.0	0.0	0.6	1.0	0.0	0.0	1.0	0.1
		$\hat{R}$	17.3	19.0	0.3	18.0	8.1	17.7	19.0	0.1	17.9	12.0	13.7	19.0	0.0	15.7	19.3
		MSE	278.6	324.0	0.7	287.7	58.6	175.4	196.0	24.5	166.2	60.7	63.8	81.0	99.2	50.1	88.7
400	400	*%corr*	99.9	0.0	98.4	24.8	92.9	97.6	0.0	94.2	94.2	93.4	93.7	0.0	55.9	49.4	90.6
		$\hat{R}$	1.0	19.0	1.0	10.5	1.1	5.0	19.0	4.7	5.1	5.1	9.9	19.0	5.6	11.0	9.6
		MSE	0.0	324.0	0.0	129.0	0.1	0.1	196.0	1.3	0.1	0.1	0.8	81.0	43.6	3.3	3.7

* *% corr* is the percentage of correct $\hat{R}$, $\hat{R}$ is the average estimate of R and *MSE* is the mean squared error over the Monte Carlo trials.

Table 4.7: Monte Carlo Results for approximate factor model with $\alpha = 0$ $\rho = 0$ and $\sigma^2_{F_1} = 10$.

N	T	R^*	*1*					*5*					*10*				
			AH	*BN*	*ON*	*MKT*	*EDR*	*AH*	*BN*	*ON*	*MKT*	*EDR*	*AH*	*BN*	*ON*	*MKT*	*EDR*
50	50	*%corr*	100	100	97.0	100	100	66.0	100	99.6	98.1	100	48.2	100	99.9	35.7	100
		$\hat{R}$	1.0	1.0	1.0	1.0	1.0	3.6	5.0	5.0	5.1	5.0	5.3	10.0	10.0	12.7	10.0
		MSE	0.0	0.0	0.0	0.0	0.0	5.4	0.0	0.0	0.7	0.0	41.9	0.0	0.0	14.8	0.0
	100	*%corr*	100	100	97.5	100	100	90.7	100	99.7	100	100	83.3	100	99.9	93.7	99.7
		$\hat{R}$	1.0	1.0	1.0	1.0	1.0	4.6	5.0	5.0	5.0	5.0	8.5	10.0	10.0	10.1	10.0
		MSE	0.0	0.0	0.0	0.0	0.0	1.5	0.0	0.0	0.0	0.0	13.5	0.0	0.0	0.5	0.2
100	100	*%corr*	100	100	98.8	100	100	99.6	100	99.5	100	100	99.8	100	99.8	99.7	97.5
		$\hat{R}$	1.0	1.0	1.0	1.0	1.0	5.0	5.0	5.0	5.0	5.0	10.0	10.0	10.0	10.0	10.0
		MSE	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.2	0.0	0.0	0.0	0.0
	200	*%corr*	100	100	99.2	100	100	100	100	99.5	100	100	100	100	99.8	100	99.9
		$\hat{R}$	1.0	1.0	1.0	1.0	1.0	5.0	5.0	5.0	5.0	5.0	10.0	10.0	10.0	10.0	10.0
		MSE	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
200	200	*%corr*	100	100	99.6	100	100	100	100	99.7	100	100	100	100.0	99.9	100	100
		$\hat{R}$	1.0	1.0	1.0	1.0	1.0	5.0	5.0	5.0	5.0	5.0	10.0	10.0	10.0	10.0	10.0
		MSE	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	400	*%corr*	100	100	99.7	100	100	100	100	99.9	100	100	100	100	100	100	100
		$\hat{R}$	1.0	1.0	1.0	1.0	1.0	5.0	5.0	5.0	5.0	5.0	10.0	10.0	10.0	10.0	10.0
		MSE	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
400	400	*%corr*	100	100	98.5	100	100	100	100	100	100	100	100	100	99.9	100	100
		$\hat{R}$	1.0	1.0	1.0	1.0	1.0	5.0	5.0	5.0	5.0	5.0	10.0	10.0	10.0	10.0	10.0
		MSE	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

* *% corr* is the percentage of correct $\hat{R}$, $\hat{R}$ is the average estimate of R and *MSE* is the mean squared error over the Monte Carlo trials.

Table 4.8: Monte Carlo Results for approximate factor model with $\alpha = .5$, $J = 8$, $\rho = 0.5$ and $\sigma^2_{F_1} = 10$.

N	T	R^*	*1*					*5*					*10*				
			AH	*BN*	*ON*	*MKT*	*EDR*	*AH*	*BN*	*ON*	*MKT*	*EDR*	*AH*	*BN*	*ON*	*MKT*	*EDR*
50	50	*%corr*	100	0.0	0.0	0.0	14.3	2.6	0.0	0.0	0.0	69.0	1.7	0.0	0.2	0.0	44.2
		$\hat{R}$	1.0	19.0	5.1	4.9	2.3	1.3	19.0	8.9	9.0	4.5	2.3	19.0	13.7	14.2	8.5
		MSE	0.0	324.0	17.6	15.1	2.1	15.5	196.0	15.8	16.1	1.4	73.0	81.0	14.1	17.9	8.3
	100	*%corr*	100	0.0	0.0	0.0	5.2	1.8	0.0	0.0	0.0	84.0	0.8	0.0	0.0	0.0	64.5
		$\hat{R}$	1.0	19.0	5.1	5.0	2.5	1.2	19.0	9.0	9.0	4.7	1.9	19.0	14.0	14.1	10.2
		MSE	0.0	324.0	16.7	15.9	2.6	15.7	196.0	16.3	16.1	0.9	76.0	81.0	15.8	16.6	0.6
100	100	*%corr*	100	0.0	7.6	0.0	98.6	3.0	0.0	17.8	0.0	99.5	1.0	0.0	97.1	0.0	99.0
		$\hat{R}$	1.0	17.1	8.6	8.9	1.0	1.1	18.8	11.6	13.3	5.0	1.1	19.0	9.8	17.9	10.0
		MSE	0.0	264.1	62.8	62.7	0.0	15.5	192.0	54.4	69.8	0.0	80.2	81.0	1.6	62.4	0.0
	200	*%corr*	100	0.0	0.8	0.0	100	1.4	0.0	4.7	0.0	100	1.5	0.0	99.9	0.0	100
		$\hat{R}$	1.0	16.5	9.7	9.2	1.0	1.1	18.8	13.2	13.6	5.0	1.1	19.0	10.0	18.0	10.0
		MSE	0.0	244.3	75.7	67.9	0.0	15.8	190.9	71.1	74.7	0.0	79.8	81.0	0.1	63.5	0.0
200	200	*%corr*	100	0.0	97.7	0.0	100	38.9	0.0	99.7	0.1	100	30.1	0.0	99.4	0.0	100
		$\hat{R}$	1.0	19.0	1.0	15.5	1.0	2.6	19.0	5.0	16.9	5.0	3.7	19.0	10.0	17.1	10.0
		MSE	0.0	324.0	0.1	216.2	0.0	9.8	196.0	0.0	144.4	0.0	56.6	81.0	0.0	51.5	0.0
	400	*%corr*	100	0.0	99.0	0.0	100	56.7	0.0	100	0.2	100	54.8	0.0	100	1.2	100
		$\hat{R}$	1.0	19.0	1.0	16.1	1.0	3.3	19.0	5.0	17.2	5.0	5.9	19.0	10.0	16.8	10.0
		MSE	0.0	324.0	0.0	231.2	0.0	6.9	196.0	0.0	150.1	0.0	36.6	81.0	0.0	49.3	0.0
400	400	*%corr*	100	0.0	98.5	89.3	100	100	0.0	99.8	100	100	100	0.0	99.7	90.8	100
		$\hat{R}$	1.0	19.0	1.0	1.2	1.0	5.0	19.0	5.0	5.0	5.0	10.0	19.0	10.0	10.2	10.0
		MSE	0.0	324.0	0.0	1.1	0.0	0.0	196.0	0.0	0.0	0.0	0.0	81.0	0.0	0.4	0.0

* *% corr* is the percentage of correct $\hat{R}$, $\hat{R}$ is the average estimate of R and *MSE* is the mean squared error over the Monte Carlo trials.

Table 4.9: Monte Carlo Results for approximate factor model with $\alpha = 0$, $\rho = 0$ and $\sigma^2_{F_1} = 1/6$.

N	T	R^*	*1*					*5*					*10*				
			AH	*BN*	*ON*	*MKT*	*EDR*	*AH*	*BN*	*ON*	*MKT*	*EDR*	*AH*	*BN*	*ON*	*MKT*	*EDR*
50	50	*%corr*	99.9	97.1	97.3	94.7	99.6	32.2	96.8	98.6	97.6	97.6	45.6	95.8	94.2	36.3	92.4
		$\hat{R}$	1.0	1.0	1.0	1.1	1.0	4.3	5.0	5.0	5.1	5.0	9.4	10.0	9.9	12.7	9.9
		MSE	0.0	0.0	0.1	0.9	0.0	0.8	0.0	0.0	1.0	0.0	0.8	0.0	0.1	15.9	0.1
	100	*%corr*	100	99.8	98.5	100	99.9	43.9	99.5	99.4	99.9	99.2	62.1	99.7	99.6	93.0	97.6
		$\hat{R}$	1.0	1.0	1.0	1.0	1.0	4.4	5.0	5.0	5.0	5.0	9.6	10.0	10.0	10.1	10.0
		MSE	0.0	0.0	0.0	0.0	0.0	0.6	0.0	0.0	0.0	0.0	0.4	0.0	0.0	0.4	0.0
100	100	*%corr*	100	100	98.1	100	100	62.9	100	99.5	100	100	79.2	100	100	100	99.7
		$\hat{R}$	1.0	1.0	1.0	1.0	1.0	4.6	5.0	5.0	5.0	5.0	9.8	10.0	10.0	10.0	10.0
		MSE	0.0	0.0	0.0	0.0	0.0	0.4	0.0	0.0	0.0	0.0	0.2	0.0	0.0	0.0	0.0
	200	*%corr*	100	100	99.0	100	100	85.0	100	99.5	100	100	92.8	100	99.9	100	100
		$\hat{R}$	1.0	1.0	1.0	1.0	1.0	4.9	5.0	5.0	5.0	5.0	9.9	10.0	10.0	10.0	10.0
		MSE	0.0	0.0	0.0	0.0	0.0	0.2	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0
200	200	*%corr*	100	100	99.3	100	100	98.5	100	99.3	100	100	99.9	100	99.8	100	100
		$\hat{R}$	1.0	1.0	1.0	1.0	1.0	5.0	5.0	5.0	5.0	5.0	10.0	10.0	10.0	10.0	10.0
		MSE	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	400	*%corr*	100	100	99.3	100	100	100	100	99.8	100	100	100	100	100	100	100
		$\hat{R}$	1.0	1.0	1.0	1.0	1.0	5.0	5.0	5.0	5.0	5.0	10.0	10.0	10.0	10.0	10.0
		MSE	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
400	400	*%corr*	100	100	99.1	100	100	100	100	99.9	100	100	100	100	99.9	100	100
		$\hat{R}$	1.0	1.0	1.0	1.0	1.0	5.0	5.0	5.0	5.0	5.0	10.0	10.0	10.0	10.0	10.0
		MSE	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

* *% corr* is the percentage of correct $\hat{R}$, $\hat{R}$ is the average estimate of R and *MSE* is the mean squared error over the Monte Carlo trials.

Table 4.10: Monte Carlo Results for approximate factor model with $\alpha = .5$, $J = 8$, $\rho = 0.5$ and $\sigma^2_{F_1} = 1/6$.

N	T	R^*	*1*					*5*					*10*				
			AH	*BN*	*ON*	*MKT*	*EDR*	*AH*	*BN*	*ON*	*MKT*	*EDR*	*AH*	*BN*	*ON*	*MKT*	*EDR*
50	50	*%corr*	2.1	0.0	0.0	0.0	36.0	5.9	2.0	0.0	0.0	0.0	1.0	0.0	2.0	0.0	0.2
		$\hat{R}$	4.4	19.0	5.1	4.8	0.9	2.5	6.6	19.0	9.0	9.0	5.1	3.0	12.9	19.0	13.6
		MSE	12.2	324.0	17.1	14.9	1.2	2.8	7.7	196.0	15.8	15.9	3.6	4.4	10.7	81.0	14.5
	100	*%corr*	0.1	0.0	0.0	0.0	22.2	1.6	0.0	0.0	0.0	0.0	0.0	0.0	0.6	0.0	0.0
		$\hat{R}$	4.8	19.0	5.1	5.0	0.9	2.7	7.4	19.0	9.1	9.0	6.1	3.1	13.4	19.0	13.9
		MSE	14.5	324.0	17.1	15.9	1.5	3.4	10.7	196.0	17.1	16.1	5.0	4.0	13.2	81.0	15.8
100	100	*%corr*	1.0	0.0	0.3	0.0	7.8	0.0	0.0	0.0	1.0	0.0	2.0	0.0	0.5	0.0	1.8
		$\hat{R}$	8.2	17.2	8.5	8.9	1.9	4.5	4.1	18.9	11.0	13.3	4.0	3.5	9.9	19.0	8.7
		MSE	54.7	265.1	63.0	62.4	5.5	13.2	1.5	193.1	50.7	69.1	1.0	2.7	6.6	81.0	4.5
	200	*%corr*	0.4	0.0	0.1	0.0	5.2	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.1	0.0	0.6
		$\hat{R}$	8.8	16.4	9.6	9.1	2.7	4.8	4.0	18.8	13.0	13.7	4.0	3.7	9.2	19.0	9.0
		MSE	62.2	241.6	75.7	66.6	7.3	14.9	1.0	191.3	70.0	75.1	1.0	2.0	2.6	81.0	1.2
200	200	*%corr*	73.9	0.0	88.4	0.0	42.1	0.0	0.0	0.0	76.0	0.0	36.0	0.0	0.3	0.0	63.8
		$\hat{R}$	5.2	19.0	0.9	15.6	0.4	8.7	4.0	19.0	4.8	17.0	4.4	3.9	9.0	19.0	9.6
		MSE	69.0	324.0	0.1	218.2	0.6	61.5	1.0	196.0	0.2	145.1	0.6	1.4	1.0	81.0	0.4
	400	*%corr*	89.8	0.0	99.3	0.0	50.8	0.0	0.0	0.0	95.0	0.0	37.0	0.0	0.0	0.0	91.1
		$\hat{R}$	2.7	19.0	1.0	16.1	0.5	9.2	4.0	19.0	5.0	17.3	4.4	3.9	9.0	19.0	9.9
		MSE	29.3	324.0	0.0	230.8	0.5	68.9	1.0	196.0	0.1	152.2	0.6	1.5	1.0	81.0	0.1
400	400	*%corr*	100	0.0	98.1	100	100	0.0	0.0	0.0	100	100	78.0	0.0	4.6	0.0	100
		$\hat{R}$	1.0	19.0	1.0	1.0	1.0	17.3	4.0	19.0	5.0	5.0	4.8	4.0	9.0	19.0	10.0
		MSE	0.0	324.0	0.0	0.0	0.0	269.7	1.0	196.0	0.0	0.0	0.2	1.1	1.0	81.0	0.0

* *% corr* is the percentage of correct $\hat{R}$, $\hat{R}$ is the average estimate of R and *MSE* is the mean squared error over the Monte Carlo trials.

Table 4.11: Monte Carlo Results for the dynamic AR(1) interactive fixed effects model with $\beta = 0.5$, $R = 2$ AR(1) factors and white noise errors: $\sigma^2_{u_{i,t}} = 1$. The first table summarizes consistency results of the estimator of β, the second table summarizes the estimators of R.

N^*	T	$\hat{R}$	0	2	3
50	50	*mean* $\hat{\beta}$	0.95	0.43	0.43
		RMSE	0.21	0.06	0.05
	100	*mean* $\hat{\beta}$	0.96	0.47	0.48
		RMSE	0.21	0.05	0.05
100	100	*mean* $\hat{\beta}$	0.95	0.47	0.49
		RMSE	0.21	0.05	0.05
	200	*mean* $\hat{\beta}$	0.96	0.49	0.51
		RMSE	0.21	0.05	0.05
200	200	*mean* $\hat{\beta}$	0.96	0.50	0.51
		RMSE	0.21	0.05	0.05
	400	*mean* $\hat{\beta}$	0.96	0.49	0.50
		RMSE	0.21	0.05	0.05
400	400	*mean* $\hat{\beta}$	0.96	0.50	0.50
		RMSE	0.21	0.05	0.05

$N^\dagger$	T	$\hat{R}$	0		2		3	
			AH	*EDR*	*AH*	*EDR*	*AH*	*EDR*
50	50	*%corr*	97.3	93.2	96.9	92.6	97.4	91.7
		$\hat{R}$	2.0	2.1	2.0	2.2	2.0	2.2
		MSE	0.0	0.2	0.0	0.5	0.0	0.9
	100	*%corr*	99.8	100	99.9	99.2	99.8	98.8
		$\hat{R}$	2.0	2.0	2.0	2.0	2.0	2.0
		MSE	0.0	0.0	0.0	0.0	0.0	0.4
100	100	*%corr*	100	100	99.80	99.8	99.8	100
		$\hat{R}$	2.0	2.0	2.0	2.0	2.0	2
		MSE	0.0	0.0	0.0	0.0	0.0	0.0
	200	*%corr*	100	100	100	100	100	99.8
		$\hat{R}$	2.0	2.0	2.0	2.0	2.0	2.2
		MSE	0.0	0.0	0.0	0.0	0.0	0.0
200	200	*%corr*	100	100	100	100	100	100
		$\hat{R}$	2.0	2.0	2.0	2.0	2.0	2.2
		MSE	0.0	0.0	0.0	0.0	0.0	0.0
	400	*%corr*	100	100	100	100	100	100
		$\hat{R}$	2.0	2.0	2.0	2.0	2.0	2.2
		MSE	0.0	0.0	0.0	0.0	0.0	0.0
400	400	*%corr*	100	100	100	100	100	100
		$\hat{R}$	2.0	2.0	2.0	2.0	2.0	2.2
		MSE	0.0	0.0	0.0	0.0	0.0	0.0

* *mean* is the average $\hat{\beta}$ and *MSE* is the mean squared error over 1000 Monte Carlo trials.

† *% corr* is the percentage of correct $\hat{R}$, $\hat{R}$ is the average estimate of R and *MSE* is the mean squared error over the Monte Carlo trials.

Table 4.12: Monte Carlo Results for the static interactive fixed effects model with $\beta = 0.5$, $R = 5$, white noise errors: $\sigma^2_{u_{i,t}} = 1$. The first table summarizes consistency results of the estimator of β, the second table summarizes the estimators of R.

N^*	T	$\hat{R}$	0	5	7
50	50	*mean* $\hat{\beta}$	0.96	0.36	0.31
		RMSE	0.21	0.08	0.10
	100	*mean* $\hat{\beta}$	0.96	0.45	0.42
		RMSE	0.21	0.06	0.07
100	100	*mean* $\hat{\beta}$	0.96	0.44	0.43
		RMSE	0.21	0.06	0.06
	200	*mean* $\hat{\beta}$	0.96	0.48	0.46
		RMSE	0.21	0.05	0.06
200	200	*mean* $\hat{\beta}$	0.96	0.48	0.48
		RMSE	0.21	0.06	0.06
	400	*mean* $\hat{\beta}$	0.96	0.49	0.48
		RMSE	0.21	0.06	0.06
400	400	*mean* $\hat{\beta}$	0.96	0.49	0.48
		RMSE	0.21	0.05	0.06

$N^\dagger$	T	$\hat{R}$	0		5		7	
			AH	*EDR*	*AH*	*EDR*	*AH*	*EDR*
50	50	*%corr*	86.6	97	94.2	91.9	95.2	89.3
		$\hat{R}$	4.5	5.0	4.8	5.2	4.8	5.3
		MSE	2.1	0.1	0.9	2.0	0.7	1.8
	100	*%corr*	99	100	100	99.2	99.8	98.6
		$\hat{R}$	5.0	5.0	5.0	5.0	5.0	5.1
		MSE	0.2	0.0	0.0	0.5	0.0	1.1
100	100	*%corr*	100	100	100	100	100	100
		$\hat{R}$	5.0	5.0	5.0	5.0	5.0	5.0
		MSE	0.0	0.0	0.0	0.0	0.0	0.0
	200	*%corr*	100	100	100	99.9	100	99.9
		$\hat{R}$	5.0	5.0	5.0	5.0	5.0	5.0
		MSE	0.0	0.0	0.0	0.0	0.0	0.2
200	200	*%corr*	100	100	100	100	100	100
		$\hat{R}$	5.0	5.0	5.0	5.0	5.0	5.0
		MSE	0.0	0.0	0.0	0.0	0.0	0.0
	400	*%corr*	100	100	100	100	100	100
		$\hat{R}$	5.0	5.0	5.0	5.0	5.0	5.0
		MSE	0.0	0.0	0.0	0.0	0.0	0.0
400	400	*%corr*	100	100	100	100	100	100
		$\hat{R}$	5.0	5.0	5.0	5.0	5.0	5.0
		MSE	0.0	0.0	0.0	0.0	0.0	0.0

* *mean* is the average $\hat{\beta}$ and *MSE* is the mean squared error over 1000 Monte Carlo trials.

† *% corr* is the percentage of correct $\hat{R}$, $\hat{R}$ is the average estimate of R and *MSE* is the mean squared error over the Monte Carlo trials.

Table 4.13: Monte Carlo Results for the static interactive fixed effects model with $\beta = 1$, $R = 2$ and white noise errors: $\sigma^2_{u_{i,t}} = 1$. The first table summarizes consistency results of the estimator of β, the second table summarizes the estimators of R.

N^*	T	$\hat{R}$	0	2	3
50	50	*mean* $\hat{\beta}$	2.14	0.99	0.99
		RMSE	1.31	0.00	0.00
	100	*mean* $\hat{\beta}$	2.14	0.99	0.99
		RMSE	1.31	0.00	0.00
100	100	*mean* $\hat{\beta}$	2.14	0.99	0.99
		RMSE	1.30	0.00	0.00
	200	*mean* $\hat{\beta}$	2.14	0.99	0.99
		RMSE	1.31	0.00	0.00
200	200	*mean* $\hat{\beta}$	2.14	0.99	0.99
		RMSE	1.31	0.00	0.00
	400	*mean* $\hat{\beta}$	2.14	0.99	0.99
		RMSE	1.31	0.00	0.00
400	400	*mean* $\hat{\beta}$	2.14	0.99	0.99
		RMSE	1.31	0.00	0.00

$N^\dagger$	T	$\hat{R}$	0		2		3	
			AH	*EDR*	*AH*	*EDR*	*AH*	*EDR*
50	50	*%corr*	4.3	5	99.8	99.7	100	99.9
		$\hat{R}$	2.3	3.7	2.0	2.0	2.0	2.0
		MSE	1.0	12.7	0.0	0.6	0.0	0.0
	100	*%corr*	0.6	0.7	99.6	99.6	100	100
		$\hat{R}$	2.7	4.2	2.0	2.0	2.0	2.0
		MSE	1.0	18.8	0.0	0.7	0.0	0.0
100	100	*%corr*	0	0.1	99.9	99.9	100	100
		$\hat{R}$	3.0	3.7	2.0	2.0	2.0	2.0
		MSE	1.0	10.3	0.0	0.3	0.0	0.0
	200	*%corr*	0	0	99.8	99.8	100	100
		$\hat{R}$	3.0	3.4	2.0	2.0	2.0	2.0
		MSE	1.0	6.9	0.0	0.3	0.0	0.0
200	200	*%corr*	0	0	99.9	99.9	100	100
		$\hat{R}$	3.0	3.1	2.0	2.0	2.0	2.0
		MSE	1.0	2.9	0.0	0.0	0.0	0.0
	400	*%corr*	0	0	100	100	100	100
		$\hat{R}$	3.0	3.0	2.0	2.0	2.0	2.0
		MSE	1.0	1.3	0.0	0.0	0.0	0.0
400	400	*%corr*	0	0	99.9	99.9	100	100
		$\hat{R}$	3.0	3.0	2.0	2.0	2.0	2.0
		MSE	1.0	1.0	0.0	0.0	0.0	0.0

* *mean* is the average $\hat{\beta}$ and *MSE* is the mean squared error over 1000 Monte Carlo trials.

† *% corr* is the percentage of correct $\hat{R}$, $\hat{R}$ is the average estimate of R and *MSE* is the mean squared error over the Monte Carlo trials.

Table 4.14: Monte Carlo Results for the static interactive fixed effects model with $\beta = 1$, $R = 2$ and correlated errors: $\sigma^2_{u_{i,t}} = 1$, $\alpha = 0.5$, $J = 8$ and $\rho = 0.5$. The first table summarizes consistency results of the estimator of β, the second table summarizes the estimators of R.

N^*	T	$\hat{R}$	0	2	3
50	50	*mean* $\hat{\beta}$	2.13	1.06	1.01
		RMSE	1.30	0.02	0.00
	100	*mean* $\hat{\beta}$	2.14	1.04	1.01
		RMSE	1.30	0.01	0.00
100	100	*mean* $\hat{\beta}$	2.14	1.00	1.00
		RMSE	1.31	0.00	0.00
	200	*mean* $\hat{\beta}$	2.14	1.00	1.00
		RMSE	1.30	0.00	0.00
200	200	*mean* $\hat{\beta}$	2.14	1.00	0.99
		RMSE	1.31	0.00	0.00
	400	*mean* $\hat{\beta}$	2.15	0.99	1.00
		RMSE	1.31	0.00	0.00
400	400	*mean* $\hat{\beta}$	2.14	0.99	0.99
		RMSE	1.31	0.00	0.00

$N^\dagger$	T	$\hat{R}$	0		2		3	
			AH	*EDR*	*AH*	*EDR*	*AH*	*EDR*
50	50	*%corr*	3.5	16.6	22.4	57.4	23.4	55.6
		$\hat{R}$	1.6	1.9	1.9	3.0	1.9	2.9
		MSE	2.8	4.4	2.8	3.8	2.7	1.7
	100	*%corr*	1.1	11.8	18.7	59.8	18.3	59.5
		$\hat{R}$	1.5	2.4	2.0	2.9	2.0	2.8
		MSE	2.8	6.3	3.7	2.6	3.8	2.0
100	100	*%corr*	0.3	6.3	52.2	97.5	52.2	97.7
		$\hat{R}$	1.0	2.7	1.5	2.1	1.5	2.1
		MSE	1.2	7.9	0.6	0.1	0.6	0.4
	200	*%corr*	0	5.1	61.1	99.6	61.5	99.7
		$\hat{R}$	1.0	3.0	1.6	2.0	1.6	2.0
		MSE	1.0	9.2	0.4	0.7	0.4	0.0
200	200	*%corr*	0	1.7	96.1	100	96.2	100
		$\hat{R}$	1.0	3.3	2.0	2.0	2.0	2.0
		MSE	1.0	7.5	0.0	0.0	0.0	0.0
	400	*%corr*	0	0.5	99.2	99.9	99.5	100
		$\hat{R}$	1.0	3.4	2.0	2.0	2.0	2.0
		MSE	1.0	7.9	0.0	0.3	0.0	0.0
400	400	*%corr*	0	0.1	100	100	100	100
		$\hat{R}$	1.0	4.7	2.0	2.0	2.0	2.0
		MSE	1.0	14.0	0.0	0.0	0.0	0.0

* *mean* is the average $\hat{\beta}$ and *MSE* is the mean squared error over 1000 Monte Carlo trials.

† *% corr* is the percentage of correct $\hat{R}$, $\hat{R}$ is the average estimate of R and *MSE* is the mean squared error over the Monte Carlo trials.

Bibliography

[1] Ahn, S. and Horenstein, A. (2013): "Eigenvalue Ratio Test for the Number of Factors," *Econometrica*, Vol. 83, No. 3, pp. 1203-1227, Econometric Society, Wiley-Blackwell

[2] Ahn, S. Lee, H. and Schmidt, P. (2014): "Panel Data Models with Multiple Time-Varying Individual Effects," *Journal of Econometrics*, Vol. 101, Issue 1, pp. 123–164, Elsevier North Holland

[3] Alessi, L., Barigozzi, M. and Capasso, M. (2010): "Improved Penalization for Determining the Number of Factors in Approximate Factor Models," *Statistics and Probability Letters*, Vol. 80 Issue 23-24, #1-15, pp. 1806-1813, Elsevier North Holland

[4] Anderson, T. W. and Das Gupta, S. (1963): "Some Inequalities on Characteristic Roots of Matrices," *Biometrika*, Vol. 50, Issue 3-4 pp. 522–524, Oxford University Press

[5] Andrews, D. and Lu, B. (2001): "Consistent Model and Moment Selection Procedures for GMM Estimation with Application to Dynamic Panel Data Models," *Journal of Econometrics*, Vol. 101, Issue 1, pp. 123–164, Elsevier North Holland

[6] Bai, J. (2009): "Panel Data Models with Interactive Fixed Effects," *Econometrica*, Vol. 77, Issue 4, pp. 1229-1279, Econometric Society, Wiley-Blackwell

[7] Bai J. and Ng, S. (2002): "Determining the Number of Factors in Approximate Factor Models," *Econometrica*, Vol. 70, No. 1, pp. 191-221, Econometric Society, Wiley-Blackwell

[8] Bai, J. and S. Ng (2006), "Evaluating Latent and Observed Factors in Macroeconomics and Finance," *Journal of Econometrics*, Vol. 113 No. (1–2), pp. 507–537 , Elsevier North Holland

[9] Bai, Z and Silverstein, J. (1998): "No Eigenvalues Outside the Support of the Limiting Spectral Distribution of Large Dimensional Sample Covariance Matrices, *Annals of Probability*, Vol. 26 pp. 316-345, Institute of Mathematical Statistics

[10] Bai, Z and Silverstein, J. (1999): "Exact Separation of Eigenvalues of Large Dimensional Sample Covariance Matrices," *Annals of Probability*, Vol. 27 No. 3 pp. 1536-1555, Institute of Mathematical Statistics

[11] Bai, Z and Silverstein, J. (2010): "Spectral Analysis of Large Dimensional Random Matrices," *Springer*

[12] Bai, Z., Silverstein, J. and Yin, Q. (1988): "A Note on the largest Eigenvalue of a Large Dimensional Sample Covariance Matrix," *Journal of Multivariate Analysis*, Vol. 26, pp. 166-168, Elsevier North Holland

[13] Bai, Z. and Yin, Y. (1993): "Limit of the Smallest Eigenvalue of a large Dimensional Covariance Matrix," *Annals of Probability*, Vol. 21 No. 3 pp. 1275-1294, Institute of Mathematical Statistics

[14] Bai, Z. and Zhou, W. (2008): "Large Sample Covariance Matrices without Independence Structures in Columns," *Statistica Sinica*, Vol. 18, pp. 425-442

[15] Bernanke, B., Boivin, J. and Eliasz, P. (2005): "Factor Augmented Vector Autoregression and the Analysis of Monetary Policy," *Quarterly Journal of Economics*, Vol. 120, pp. 387-422, Oxford University Press

[16] Caner, M. and Han, X. (2014): "Selecting the Correct Number of Factors," *Journal of Business and Economic Statistics,* Vol. 32, No. 3 pp. 359-374, Taylor & Francis Group

[17] Cattell, R (1966). "The Scree Test for the Number of Factors," *Multivariate Behavioural Research*, Vol. 1, pp. 245-276, Taylor & Francis Group

[18] Cragg, G. and Donald, S. (1997): "Inferring the Rank of a Matrix," *Journal of Econometrics*, Vol. 76, Issue 1-2, pp. 223-250, Elsevier North Holland

[19] Chamberlain, G. and Rothschild, M. (1983): "Arbitrage, Factor Structure, and Mean-Variance Analysis on Large Asset Markets", *Econometrica*, Vol. 51, No. 5, pp. 1281-1304, Econometric Society, Wiley-Blackwell

[20] De Mol, C., Giannone, D. and Reichlin, L. (2008): "Forecasting Using a Large Number of Predictors: Is Bayesian Shrinkage a Valid Alternative to Principal Components?" *Journal of Econometrics*, Vol. 146 , pp. 318–328, Elsevier North Holland

[21] Engel, C., Nelson, M. and West, K. (2014): "Factor Model Forecasts of Exchange Rates," *Econometric Reviews*, Vol. 0 (0), pp. 1–24, Taylor & Francis Group

[22] Feuerverger, A., He, Y. and Khatri, S. (2012): "Statistical Significance of the Netflix Challenge," *Statistical Science*, Vol. 27, No. 2, pp. 202–231, Institute of Mathematical Statistics

[23] Grenander, U. and and Szegö, G. (1958): "Toeplitz Forms and Their Applications," University of California Press, Berkeley

[24] Hallin, M. and Liška, R. (2007): "Determining the Number of Factors in the General Dynamic Factor Model," *Journal of the American Statistical Association,* Vol. 102, No. 478, pp. 603-617, Taylor & Francis Group

[25] Hirose, K. and Konishi, S. (2013): "Variable Selection via the Weighted Group LASSO for Factor Analysis Models," *The Canadian Journal of Statistics*, Vol. 40, pp. 345-361, Wiley-Blackwell

[26] Hoeffding, W. (1963). "Probability Inequalities for Sums of Bounded Random Variables," *Journal of the American Statistical Association*. Vol. 58 Issue 301, pp. 13–30, Taylor and Francis

[27] Kapetanios, G. (2004): "A New Method for Determining the Number of Factors in Factor Models with Large Datasets," *Queen Mary University of London Working Paper*

[28] Lam, C. and Yao, Q. (2012): "Factor Modelling for High-Dimensional Time Series: Inference for the Number of Factors," *Annals of Statistics*, Vol. 29#2, pp 295-327, Institute of Mathematical Statistics

[29] Li, H., Li, Q. and Shi, Y. (2017): "Determining the Number of Factors when the Number of Factors can Increase with Sample Size," *Journal of Econometrics*, Vol. 197, pp. 76-86, Elsevier North Holland

[30] Li, Z., Pan, G. and Yao, J. (2014): On Singular Value Distribution of Large-Dimensional Autocovariance Matrices, mimeo

[31] Li, Z., Wang, W. and Yao, J. (2014): Identifying the Number of Factors from Singular Values of a Large Sample Auto-Covariance Matrix, mimeo

[32] Liu, H. Aue, A. and Paul, D. (2015): "On the Marcenko-Pastur Law for Linear Time Series," *Annals of Statistics*, Vol. 43, No. 2, pp. 675–712, Institute of Mathematical Statistics

[33] Marchenko,V. and Pastur, L. (1967): "Distribution of eigenvalues for some sets of random matrices", *Mathematics of the USSR-Sbornik*, Vol. 72 No. (1-4), pp. 507–536

[34] Moon, H. R. and Weidner, M. (2015): "Linear Regression for Panel with Unknown Number of Factors as Interactive Fixed Effects," *Econometrica*, Vol. 83, Issue 4, pp. 1543-1579, Econometric Society, Wiley-Blackwell.

[35] Moon, H. R. and Weidner, M. (2017): "Dynamic Linear Panel Regression Models with Interactive Fixed Effects," *Econometric Theory*, Vol. 33, Issue 1, pp. 158-195, Cambridge University Press.

[36] Nickel, S., (1981): "Biases in Dynamic Models with Fixed Effects," *Econometrica*, Vol. 49, Issue 6, Pp. 1417-26, Econometric Society, Wiley-Blackwell

[37] Onatski, A. (2010): "Determining the Number of Factors from Empirical Distribution of Eigenvalues," *The Review of Economics and Statistics*, Vol. 92 No. 4, pp. 1004–1016, MIT Press

[38] Onatski, A. (2012): "Asymptotics of the Principal Components Estimator of Large Factor Models of Weakly Influential Factors," *Journal of Econometrics*, Vol. 168, pp. 244-258, Elsevier North Holland

[39] Onatski, A. (2015): "Asymptotic Analysis of the Squared Estimation Error in Misspecified Factor Models," *Journal of Econometrics*, Vol. 186 Issue 2, pp. 388–406, Elsevier North Holland

[40] Paul, D. and Silverstein, J. (2008): "No Eigenvalues Outside the Support of Limiting Empirical Spectral Distribution of a Separable Covariance Matrix," mimeo

[41] Pesaran M. H. (2006): “Estimation and Inference in Large Heterogeneous Panels with a Multifactor Error Structure,” *Econometrica*, Vol. 74, pp. 967-1012, Econometric Society, Wiley-Blackwell

[42] Pfaffel, O. and Schlemm, E. (2012): Eigenvalue Distribution of Large Sample Covariance Matrices of Linear Processes, mimeo

[43] Pötscher, B. (1983): "Order estimation in ARMA-models by Lagrangian multiplier tests," *Annals of Statistics,* Vol. 11, pp. 872-885, Institute of Mathematical Statistics

[44] Price, A. L., Patterson, N. J. Plenge, R. M., Weinblatt, M. E. Shadick, N. A. and Reich, D. (2006): "Principal Components Analysis Corrects for Stratification in Genome-Wide Association Studies," *Nature Genetics*, Vol. 38, pp. 904-909

[45] Robertson, D. and Sarafidis, V. (2015): "IV Estimation of Panels with Factor Residuals," *Journal of Econometrics*, Vol. 185, pp. 526-541, Elsevier North Holland

[46] Schwarz, G. (1978), "Estimating the dimension of a model," *Annals of Statistics*, Vol. 6 No. 2: pp. 461–464, Institute of Mathematical Statistics

[47] Tibshirani, R. (1996): "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society B.*, Vol. 58, No. 1, pp. 267-288, Econometric Society, Wiley-Blackwell

[48] Wang, C., Jin, B., Bai, Z. Nair, K. and Harding, M. (2015): “Strong Limit of the Extreme Eigenvalues of a Symmetrized Auto-Cross Covariance Matrix,” *Annals of Applied Probability*, Vol. 25 #6, pp. 3624-3683, Institute of Mathematical Statistics

[49] Yin, Y., Bai, Z. and Krishnaiah, P. (1988): “On the limit of the Largest Eigenvalue of the Large Dimensional Sample Covariance Matrix,” *Probability Theory and Related Fields*, Vol 78, pp. 509-521, Springer

[50] Yaun, M. and Lin, Y. (2006): "Model Selection and Estimation in Regression with Grouped Variables," *Journal of the Royal Statistical Society B.*, Vol. 68, Part 1, pp. 49–67, Econometric Society, Wiley-Blackwell

Chapter 5

Fiscal Multipliers and the Stability and Growth Pact: A Panel-VAR Analysis of Europe

This chapter analyses the size of fiscal multipliers in three panels of European Countries. We develop a novel methodology to estimate vector autoregressions in the presence of factor errors and use a combination of sign, zero and equality restrictions to identify prototypical fiscal shocks. We find evidence for Ricardian Equivalence in all three panels in the sense that unless the government matches expenditures with increased taxation, the Euro-zone does not strongly or prolongedly benefit from fiscal stimuli.

5.1 Introduction

The European Monetary Union (EMU) is largely considered a success: monetary unification caused no disruptions in financial markets and the introduction of Euro coins and notes was surprisingly smooth despite the daunting logistic challenges. Furthermore, the common market, currency and monetary policy all had their share in the integration of the Euro zone since its conception (Beetsma and Guilidori, 2010a). Despite these successes, economists and politicians alike voiced concerns on the ability of European national governments to stabilise the downturns following the global mortgage crisis and the Hellenic debt crisis under the regulations of EMU. These concerns are a direct result of the consequences of monetary unification for the adopting member states: first of all, the adoption of a single currency means that national policy makers can no longer rely on currency depreciation to fuel exports of their national economies. Entwined with the common currency is the constitution of the supra-national European Central Bank (ECB) in 1998. The supra-nationality of the ECB is intended to guarantee independence of national interests in governing the common currency and price level, but also prevents national governments from stimulating investment in

downturns by reducing interest rates.

Monetary unification also has consequences for the fiscal autonomy of current and aspiring EMU members since the signing of the Maastricht Treaty and these are codified in the Stability and Growth Pact (SGP) of 1997. The SGP is a blueprint for responsible fiscal policy and consists of three provisions: the first provision requires all member states to accord to a public surveillance system that is designed to prevent countries from breaching the treaty and, as Buiter (2006) argues, is intended to 'name-and-shame' offending member states. However, even if such a peer-pressure mechanism is credible for aspiring EMU members, its potency is lost once a member has entered EMU because it cannot be evicted involuntarily. The second provision requires that members strive for medium term sustainable public finance goals, that is, on-balance or surplus net public finance and decreasing debt to a sustainable level of typically sixty percent of GDP. This provision is intended to enhance economic convergence and ensure the fiscal sustainability of member states, but is also largely homogeneous over heterogeneous member states across Europe: for example, relatively low-debt members can temporarily sustain higher deficits than others, whilst long-term economic growth and convergence to the EU average in certain member states can only be achieved by large public investment which increases in the stock of debt without supranational redistribution.

Finally, the third provision is intended to enforce the second provision and requires that the overall budget deficit of a member state must under no circumstance exceed three percent of GDP and if it does, immediate corrective action must be taken. If no corrective action is undertaken, the Excessive Debt Procedure (EDP) initiates sanctions until the deficit is within the bound of the SGP again, unless the cause of the deficit increase is outside the control of the government and/or real GDP fell by at least two percent in the reporting year. This provision is restrictive for the day-to-day operation of public finance and has therefore attracted much criticism for both political and economic reasons. For example Eichengreen and Wyplosz (1998) point out that no country experienced a fall of (more than) two percent of GDP during the conception period of the SGP and that the third provision is therefore too ambitious. This argument foreshadowed a revision of the original 1997 SGP: following a downturn in the start of the twenty-first century, fiscal policy in France and Germany had exceeded the conditions of the third provision of the SGP and should have elicited sanctions from the EDP, but none were imposed. Over the following year the EDP was suspended and, despite a court ruling against the suspension, new negotiations led to a revision of the SGP in 2005. Under the amended version, sanctions are suspended not just when economic growth is negative two percent, but also when an economy is below potential output for a prolonged period of time. Moreover, the key condition of the reasons for the deficit being outside the control of the affected member state is no longer included in the definition and it is therefore likely that fines will never be levied, especially since such levies would further increase the deficit (Buiter, 2006). The third provision of the SGP is also seen as being a one-size-fits-all cure for potential irresponsible fiscal policy of national governments regardless of past or expected economic conditions. This rigidity can make fiscal policy more pro-cyclical and thereby reduce the effectiveness of automatic stabilizers. In addition, since political decision makers are in office only for a short time, the SGP incentivizes a shift of public expenditure towards short term goals that may help the politician be re-elected when the fiscal constraints bind, but away from long run public investment projects.

Whilst such policy is in accordance with the SGP, it is expected to reduce long term economic growth and shifts wealth from future to current generations (Balassone and Franco, 2000).

A substantial body of evidence on the potential merits and demerits of EMU and the SGP has emerged since the conception of the Maastricht Treaty, but questions related to the ability of fiscal multipliers to stabilise recessionary periods and to which degree fiscal tightness affected economic growth in different areas of the Euro-area have so far been largely unaddressed. This paper intends to fill that gap: we estimate vector autoregressions (VAR) based on panel datasets compiled from Eastern, Northern and Southern European countries and analyse the dynamics of key macroeconomic variables induced by fiscal shocks with and without the SGP imposed. Using VARs to study fiscal multipliers has a long history in macroeconomics and our paper thus also adds to the evidence on the (im-)potency of fiscal policy of individual countries in a monetary union. Our results show that fiscal constraints have been especially tight in Eastern and Southern European member states, where fiscal stimulus is strong. On the other hand, we find that fiscal stimulus is relatively impotent in the West and stricter fiscal rules reduce this potency further. We also find that unless expenditure is matched by increases in taxation, the data suggests that fiscal policy does not make a lasting contribution to economic growth. This finding further points to complicated dynamics where commitment to higher expenditure and growth must be initiated by structurally higher taxes. In other words, a form of Ricardian equivalence.

The remainder of the paper is structured as follows: In Section 5.2, we present an extensive review of empirical studies on EMU and the size of fiscal multipliers in general. In Section 5.3 we present a methodology to study the implications of the SGP on the wider Euro-area, using panels of Eastern, Northern and Southern European countries and show how to obtain consistent estimates of a panel VAR model in the presence of common factors. Section 5.4 gives the empirical results and several robustness tests whilst Section 5.5 concludes.

5.2 Literature Review

When the Maastricht Treaty was first signed, no data on the empirical implications of the SGP was available and some authors have instead tried to learn from the experience of the US. For example, Bayoumi and Eichengreen (1995) argue that similarly to EMU member states, individual US states are subject to constitutional and statutory restrictions on the level of debt and deficits they hold. Stabilization policy in the US is however undertaken by the federal government, whereas such policy is undertaken primarily by individual members in the Euro-area and this asymmetry suggests that unification through EMU can destabilize the Euro-area if member states cannot exercise counter-cyclical fiscal policy under the restrictions of the SGP.

Gali and Perotti (2003) address this concern and explore the cyclicality of fiscal variables as EMU came into existence and contrasts these findings with a sample of OECD countries. They show that the automatic stabilisers of EMU members were not impeded by binding SGP-constraints but actually shifted from significantly pro-cyclical to counter-cyclical fiscal policy since the advent of EMU. However, even though EMU members moved towards counter-cyclicality, the transition lagged behind an OECD-wide trend and

the restrictions of the SGP thus hampered counter-cyclical fiscal policy at least relatively. More recently, Fatas and Mihov (2010) present cross-country comparisons of the fiscal policy of twenty-two OECD countries with data until 2007. These authors find that fiscal policy has not differed substantially across the OECD and that the introduction of the Euro has not led to significant change in fiscal policy in the Euro-area. The aforementioned is due to the fact that counter-cyclical adjustment in Europe is largely undertaken by automatic stabilisers, whilst the impact of discretionary pro-cyclical policy is negligible. On the other hand, Bénétrix and Lane (2013) measure the change in cyclicality of fiscal variables due to the conception of EMU using a panel of eleven EMU countries in the period of 1980 to 2007. Their analysis implies both a weakening of the feedback of outstanding debt and increasing pro-cyclicality of fiscal balances since the shift to EMU. This is extended in Beetsma and Guiliodori (2010b) who argue that deviations from fiscal planning show how policy makers deal with new information. These authors include fiscal forecasts in the study of cyclicality and show that substantial differences between plans and implementation exist and that such differences drive the cyclicality of fiscal variables, especially so in recessionary periods.

The cyclicality of fiscal variables is closely linked to the potency of fiscal policy to correct the cycle. That is, if the constraints of the SGP bind in down-swings, it is unlikely that fiscal policy can dampen recessions. Moreover, by constraining discretionary fiscal policy, impact multipliers of changes in the fiscal stance on macroeconomic indicators such as GDP are likely constrained too. The estimation of fiscal multipliers is a prickly issue in the profession because it is inherentily difficult to identify truly exogenous fiscal shocks. This is because changes to fiscal variables are typically debated in parliament and reported on in real-time. As a result, changes in the fiscal stance are internalized by agents almost immediately.

Economists have developed two competing methods to identify fiscal shocks: the first method is the so-called 'narrative' approach pioneered by Ramey and Shapiro (1998). In this study it is argued that defense spending in times of war is truly exogenous to the economy and pinpoint three such episodes for the US. They conclude that output rises in response to such defense shocks; consumption and investment initially rise but fall over time; interest rates decline but move back to baseline; input prices rise and finally product wages fall. These episodes are not without criticism themselves however: first of all, times of war are rare and the robustness of results based on the methodology with respect to less radical changes in the fiscal stance is questionable. Second, due to real-time media attention to events as rare as open war, the exogeneity of the episodes might not be as clear-cut as is presented by Ramey and Shapiro (Ravn et al, 2007). An alternative narrative method is presented in Romer and Romer (2010): this paper aggregates records of presidential speeches and reports by the House of Representatives to separate changes in the tax regime of the US in the post-war era from non-policy changes due to the business cycle. Romer and Romer (2010) find robust negative (positive) effects of increases (decreases) in taxes on US GDP as large as three percent.

The second method of identifying exogenous fiscal shocks is based on the VAR methodology and several identifying strategies have been put forward. The seminal contribution is Blanchard and Perotti (2002) (BP), who estimate a trivariate VAR based on US data that includes taxes, government spending and output as endogenous variables. Fiscal shocks are identified by the observation that taxes and government expenditure typically do not respond to outside shocks in the same quarter and by exploiting outside information on

elasticities of government spending and taxation with respect to the remaining variables to restrict the contemporaneous covariance matrix. BP find that increasing government spending raises output and increased taxes decrease output. Decomposing output further, they find that government spending raises private consumption, crowds out investment and decreases imports and exports. Importantly, they find that spending multipliers are typically smaller than unity. Perotti (2004) extends the BP VAR model with inflation and interest rates and compares estimates of five OECD countries. He finds that spending shocks have positive effects on GDP although considerable heterogeneity is present in impact multipliers among the countries. Similarly, spending has a positive effect on interest rates in the last decades of the sample, but fiscal variables have no effect on inflation. Perotti further finds that tax multipliers range from positive to negative impact on output (insignificant in the US). Apart from this heterogeneity of responses to fiscal shocks, Perotti also shows that the potency of fiscal policy weakened substantially over time in the OECD. Favero and Giavazzi (2007) and Chung and Leeper (2007) make the important observation that a VAR that does not include a feedback of net government spending to debt will yield an explosive debt sequence. They show that the BP identification strategy is improved by the inclusion of a feedback of government debt to ensure stability of the fiscal variables. With this specification, they show that fiscal variables have a large impact on the US interest rate through debt dynamics. Furthermore, they show that fiscal shocks have a significantly smaller impact on GDP relative to the original BP identifying restrictions because government debt is now non-explosive. Favero et al (2011) take the debt-augmented BP VAR to a panel of fifteen OECD countries and, in accordance with the results in Perotti (2004), find substantial heterogeneity in responses to fiscal shocks, depending on the degree of openness and debt dynamics in the individual countries. Ilzetski (2011) uses the BP specification to study fiscal multipliers in panels of developed and developing countries and concludes that developed countries have positive spending multipliers but insignificant effects of tax shocks. In contrast, this finding is reversed in developing countries: taxes have significantly negative effects, but spending has no effect on output.

An alternative method of identifying fiscal shocks in a VAR is through a recursive ordering: for example, Fatas and Mihov (2001) estimate fiscal shocks with quarterly data of US data with spending, a component of private consumption or wages, GDP, GDP deflator, taxes and the interest rate as endogenous variables. They find that a spending shock yields increases in GDP, consumption, investment, taxes, wages, employment and the interest rate follow whilst the price level falls. Responses to tax shocks are not studied. Similarly, Beetsma and Giuliodori (2005) study the effects of fiscal policy spillovers through the trade channel: using recursive VARs for France, Germany and Italy, they find evidence that tax reduction has a positive effect on GDP in Germany, whilst increasing spending is more effective in France and Italy. Ilzetski, Mendoza and Vegh (2013) use the Cholesky decomposition to identify fiscal shocks in a panel dataset consisting of quarterly data for forty-four countries. These authors use sample splits to show that the multiplier of government expenditure is larger in industrialized countries than in developing countries, large in closed economies but small in open economies, negative in high-debt countries and finally zero under flexible exchange rates. This exercise is repeated using annual data for the EU by Beetsma and Guiliodori (2011), who find a government multiplier of GDP of over one percent. Furthermore, their results show that trade

balances decline as a result of expenditure shocks in open economies and, as a result, the multiplier is therefore far weaker.

A third methodology of identifying fiscal shocks in VARs was developed in response to criticism on the practice of imposing zero restrictions on contemporaneous responses of fiscal variables. In this methodology sign restrictions are placed on the shape of the impulse response functions to preserve a degree of agnosticism with respect to the data generating process without commitment to a single identification strategy. For example, Mountford and Uhlig (2009) use sign-restrictions to identify deficit, revenue, and balanced budget shocks. The results they obtain for the US are largely complementary to the aforementioned studies but differ in that deficit-financed fiscal shocks, i.e. pure tax and expenditure shocks with no direct co-movement of the other fiscal variable, leave consumption and GDP unchanged within a year. Canova and Pappa (2007) study the impact of fiscal policy on price differentials in monetary unions and impose sign-restrictions on the contemporaneous covariance matrix of a panel VAR obtained from annual US state and EU country data. Their methodology also identifies pure tax, government and balanced budget shocks and shows positive output responses to spending shocks. On the other hand, output multipliers of tax shocks are found to be larger in the EU than in the US and balanced budget shocks yield negative responses. Similarly, Pappa (2009a) studies the impact of fiscal policy on the US labour market and finds strongly positive effects of fiscal expansions in both aggregate and disaggregated data. Pappa (2009b), again using sign-restrictions, presents a cross-country comparison of the impact of government spending shocks on key variables. She concludes that considerable heterogeneity is present in responses but that the signs typically are similar to the aforementioned studies.

Caldara and Camps (2008) compare the various identification schemes above and find that the signs of spending shocks are largely the same over all identification strategies as long as the specification of the VAR contains the same variables and the same country is studied with data at the same frequency. They argue that the observed differences in responses found from fiscal VARs is due mostly to the proportion of the fiscal variables that responds automatically to business cycle movements and not to the identification scheme applied: the larger these automatic stabilizers are, the more distortionary their effects are. This is echoed in Auerbach and Gorodnichenko (2012): using the BP-specification, they estimate a smooth transition time-varying VAR that uses a measure of the business cycle as an indicator for the system across states of the economy. Their results give a further qualifier that fiscal multipliers operate most strongly at the bottom of the cycle. This evidence corroborates the concerns of Bayoumi and Eichengreen (1995): if the SGP is too rigid, it would harm the ability of automatic stabilizers to operate in EMU. Moreover, since certain countries in EMU were hit harder than others, cross-country heterogeneity is expected to be exercerbated by the cycle when studying EMU fiscal multipliers.

5.3 Methodology

Data

We will study the effects of the SGP on the size and strength of fiscal multipliers in EMU using panel data methods. However, as the studies in Perotti (2004), Pappa (2009b) and Favero et al (2011) suggest, there is substantial heterogeneity present in VAR estimates of fiscal multipliers across OECD countries. This finding seriously limits the scope for the econometric benefits of panel data and implies that we must build datasets consisting of countries that we can reasonably expect to be similar *a priori*. We believe the greater Euro-area provides a case study where such pooling can be fruitfully undertaken: the Euro zone shares a common market for goods and services and, since the signing of the Maastricht Treaty, nineteen EMU countries adopted the Euro. Furthermore, convergence in the Euro-area has been impressive: by 2017, all EMU members are classified by the World Bank as high income whereas the peripheral countries are higher middle income (World Bank 2018). On the other hand, important differences do persist: whilst Northern Europe admits a culture where tax collection is accepted, Southern Europe struggles to collect taxes; Northern Europe consists of comprehensive welfare state regimes whilst the provisions for pensions are less generous in Southern Europe. Finally Southern European countries struggled under the debt crisis as investors were nervous about their ability to settle government debt whilst the North did not directly. These differences provide an argument for a case study that compares the impact of the SGP on Northern and Southern European economic growth and how these countries cope with the fiscal restrictions of the SGP, especially in the aftermath of the Hellenic Debt Crisis. We consider only those countries in Southern Europe who held the Euro since its inception, for we do not wish to murk the results with further transitory effects of adopting the Euro at a later point in time. Furthermore, Ireland is included in the Southern European block for reasons of financial instability, whilst we exclude Sweden, Denmark and the United Kingdom from the panel of Western European countries for their staggered ascendance to, or indeed, opting-out of, EMU. Moreover, since the conditions for admission into EMU carry over to use of the Euro directly, (aspiring) EMU members in Eastern Europe make a third candidate pool for comparison, particularly since all these countries have economic fundamentals deriving from membership of the Warsaw Pact during most of the build-up leading to EMU. Table 5.1 below summarizes the countries in the three samples under consideration:

Table 5.1: Samples:

	Eastern Europe	Northern Europe	Southern Europe
	Bulgaria	Austria	Greece
	Czech Republic	Belgium	Italy
	Hungary	Finland	Ireland
	Latvia	France	Portugal
	Lithuania	Germany	Spain
	Poland	The Netherlands	
Coverage:	2001Q1-2013Q4	2000Q1-2013Q4	2001Q1-2013Q4

For these samples, we collect quarterly data on total government expenditure G, total government revenues R, government net debt, the price level as defined by Europe's HICP, interest rates paid on long-term government bonds I and population numbers spanning at most the beginning of 2000 through the end 2013. The data is obtained from Eurostat with the exception of interest rates for certain Eastern European countries, which are from the World Bank Financial Statistics database. All data is seasonally adjusted using the X-12 Arima routine available from the Office of National Statistics and linearly de-trended. Finally, all variables are in logarithms after being transformed to per capita to allow for the best possible comparison and we subsequently construct the debt-to-GDP ratio D and inflation π as the first difference of the logarithm of the HICP. The only series that are not in logarithmic form are thus the debt-to-GDP ratio D and the interest rates which are in percentages.

Some justification regarding the fiscal variables and their frequency is required: quarterly data is preferred for two reasons. First, using quarterly over annual data makes the datasets large enough to apply classical econometric methods in a panel data context and more importantly, we follow the identifying assumption of Blanchard and Perotti (2003) that fiscal variables take time to respond to changes in the economic environment. That is, we assume that a fiscal variable does not respond to news about the other within the same quarter and occurs with implementation lags. Furthermore, our VARs are estimated using total government expenditure and revenues as response variables because the SGP is defined in terms of these variables, which means that fiscal variable shocks include automatic stabilisers in the data. Similarly, the inclusion of net debt and country-specific interest rates in the VAR is motivated by Favero and Giavazzi (2007): these variables will provide feedbacks that ensure that government expenditure cannot induce explosive, unsustainable, debt dynamics.

VAR Model

We will estimate the following VAR model for all three samples:

$$X_{i,t} = \sum_{p=1}^{P} A_p X_{i,t-p} + \sum_{p=1}^{P} B_p D_{i,t-p} + U_{i,t},\ i = 1,\ldots,N,\ t = 1,\ldots,T \tag{5.1}$$

where $X_{i,t} = [G_{i,t}, R_{i,t}, Y_{i,t}, \pi_{i,t}, I_{i,t}]'$, $X_{i,t-p}$ and $D_{i,t-p}$ are lags with coefficient matrices A_p and B_p for each $p = 1,\ldots,P$ and the K-vector $U_{i,t}$ contains equation-specific errors for each $i = 1,\ldots,N$. We also let $C_p = \left[A_p', B_p{}'\right]'$ and $W_{i,t-p} = \left[X_{i,t-p}', D_{i,t-p}'\right]'$ for notational brevity. In what follows, we set $P = 4$ because we have at most 54 observations for each country. Furthermore, whilst $X_{i,t}$ is endogenous, we will not estimate an equation for the debt-to-GDP ratio $D_{i,t}$: given $G_{i,t}$ and $T_{i,t}$, the debt-ratio $D_{i,t}$ is implied tautologically based on the past, the interest rate $I_{i,t}$, the growth rate of output and $\pi_{i,t}$. This means that the number of equations estimated is five, whilst we have twenty-four VAR parameters to estimate for each equation. We follow Favero and Giavazzi (2007) in using the intertemporal government budget constraint to ensure the

stability of fiscal balances. That is, the debt-to-GDP ratio in a country i at time t is defined as:

$$D_{i,t} = \frac{1+I_{i,t}}{(1+\pi_{i,t})(1+\triangle Y_{i,t})} D_{i,t-1} + \frac{\exp(G_{i,t}) - \exp(R_{i,t})}{\exp(Y_{i,t})},\ i = 1,\ldots,N,\ t = 1,\ldots,T,$$

with $D_{i,0}$ calibrated at the start of the sample. The stock of debt in country i at time t is thus equal to the sum of (i) the value of the debt at time $t-1$, corrected for the time t real interest rate and (ii) the time t budget deficit. The attractive feature of a methodology based on this identity is that we can simulate the stock of debt over time without having to estimate the underlying process in the presence of the aforementioned multicollinearity. Since the containment of government debt, or at least the stability thereof, is a primary concern of the provisions of ERM and the SGP, this debt-augmentation methodology provides a convenient tool to study the dynamics of debt under the SGP and variations thereof. However, since we use a panel data methodology, we are forced to use the average debt in the sample under consideration:

$$\overline{D}_t = \frac{1+\overline{I}_t}{(1+\overline{\pi}_t)\left(1+\triangle \overline{Y}_t\right)} \overline{D}_{t-1} + \frac{\exp\left(\overline{G}_t\right) - \exp\left(\overline{R}_t\right)}{\exp\left(\overline{Y}_t\right)},\ t = 1,\ldots,T,$$

where $\overline{D}_t$ is the cross-sectional average over all $D_{i,t}$ and similarly for the other variables. Using $\overline{D}_t$ will still give an indication of the stability of debt under various policy scenarios in the three samples under consideration.

Estimation

To estimate the reduced form (5.1), we make an argument about the underlying processes that govern the observed variables in the reduced form: we do not assume that our regression model is correctly specified in the sense that it contains all relevant variables. Instead we assume that the co-movement of the omitted variables can be summarized by unobserved factors which affect each cross-sectional element and equation possibly with different intensity. Recent contributions in macroeconomics have stipulated the relevance of factors in forecasting macroeconomic time series and forecasts based on these factors often outperform forecasts based on traditional time series models using autoregressive processes, see i.e. Stock and Watson (2002) and Banarjee et al (2008). The use of factor models for forecasting is now widespread and the factor model has also been combined with traditional time series analysis to obtain factor-augmented estimators (Bai and Ng, 2006). Furthermore, the Euro-area has by now sustained a substantial degree of convergence in terms of open markets, open borders, a common currency and supranational legislation. As a result, it is likely that the Euro-area is hit by the same outside shocks and a similar argument holds for the (Eastern-European) EMU candidates in various stages of ERM. Moreover, the empirical contributions on EMU and the SGP that exploit panel data by Beetsma and Guiliodori (2010b) explicitly mention the significance of time effects for their regression equations. These considerations suggest that the presence of factors in the data under consideration and we will assume that a realistic view of the economic system dictates that factors are present, although we are agnostic on their origin. Factors may correspond to unobserved shocks that hit

EMU from outside or stem from omitted variable bias because the VAR is under-specified.

In particular, we assume that the error of the k-th equation of (5.1) can be decomposed as follows:

$$U_{k,t} = \Lambda_k F_t + \varepsilon_{k,t} \; k = 1, \ldots, K, \; t = 1, \ldots, T, \tag{5.2}$$

where $U_{k,t} = [u_{k,1,t}, \ldots, u_{k,N,t}]'$ is the N-vector stacking the composite errors of all N individuals in the k-th equation, i.e. $u_{k,i,t}$; $\Lambda_k = \left[\lambda'_{k,1}, \ldots, \lambda'_{k,N}\right]'$ contains $N \times R$ factor loadings and F_t is an R-vector of factors for the k-th equation of (5.1). We will also be agnostic about the factors that affect the i-th individual of the k-th VAR equation. That is, we allow for the possibility that relative to the l-th composite error, there may be more or less factors in $U_{k,t}$ and these factors may or may not be correlated with those in $U_{l,t}$. Furthermore, note that if $F_t = 1$ for all t and some k, then we have a fixed effects model and similarly for a time effects model if $F_t = f_t$ with Λ_k fixed over i. In this sense, our model thus generalizes common panel data models and contains either as a special cases. Finally, the N-vector of errors $\varepsilon_{k,t} = [\varepsilon_{k,1,t}, \ldots, \varepsilon_{k,N,t}]'$ is assumed to be i.i.d. over t such that $E\left(\varepsilon_{k,i,t}\varepsilon_{k,j,s}\right) = 0$, but $E\left(\varepsilon_{k,i,t}\varepsilon_{l,i,t}\right) \neq 0$ for all for $i \neq j$, k, l and $t \neq s$. Furthermore, we let the $K \times K$ model covariance matrix $E\left(\varepsilon_{i,t}\varepsilon'_{i,t}\right) = \Omega_i$ for each i. We will further assume that the factors are persistent in time, that the factors and equation-specific idiosyncratic errors are ergodic stationary and that lagged dependent variables are exogenous to the individual specific error $\varepsilon_{k,i,t}$.

It is not possible to estimate the system subject to the error structure (5.2) consistently without a suitable correction if factors have persistent effects on the data. We will follow Ahn, Lee and Schmidt (2013) and particularly Chapter 3 of this thesis by applying a parametric correction to remove the factors from (5.2). That is, for each k we treat the loadings matrix Λ_k as estimable parameters and find an $(N-R) \times N$ matrix M_k such that:

$$M_k U_{k,t} = M_k \varepsilon_{k,t}.$$

Since $\Lambda_k F_t$ is not observed, we have to apply a normalization to separately identify its components. We will assume that $\Lambda_k = \left[\Lambda_k^{\star\prime}, I_R\right]'$ for each k. Such a normalization is immaterial and just serves to simplify the estimation procedure as long as the determinant of the bottom $R \times R$ block of Λ_k is non-zero. In that case M can be easily constructed from the nullity of Λ_k as $M_k = \left[I_R, -\Lambda_k^{\star}\right]$ which implies a linear transformation of $U_{k,t}$ irrespective of the rank of Λ_k. Letting $C_{k,p}$ be the k-th row of C_p at each $p = 1 \ldots, P$ and $C_k = [C_{k,1}, \ldots, C_{k,P}]'$, we can now use the method of moments to estimate the parameters of (5.1) by solving the following orthogonality conditions for each k:

$$E\left[\left(M_k \varepsilon_{k,t}\right) \otimes Z_{k,t}\right] = 0. \tag{5.3}$$

In equation (5.3), each of the $N-R$ elements of $M_k\varepsilon_{k,t}$ is multiplied by an S-vector of instruments and the moment vector of the k-th equation thus consists of $S(N-R)$ orthogonality conditions and the corresponding estimator $\hat{C}_k$ is known as the QDGMM estimator. Since $\varepsilon_{k,i,t}$ is assumed to be i.i.d. over t and the regressors are exogenous to the idiosyncratic error, this implies that all lags of all K equations are valid instruments

for (5.3). The intuition of equation (5.3) is that we have replaced the unobserved factors with products of the observed data, for which we can find instruments in a standard method of moments setting. This is convenient because the solution of (5.3) does not require estimation of the factors or imposition of structure on the factors beyond stationarity in the current specification, although we do have to estimate Λ_k. Note that we cannot estimate the factors and we therefore do not know what they are. However, since we are interested in dynamic responses of macroeconomic variables to fiscal shocks, questions relating to the former are beyond the scope of this study and this is not a downside in our view.

To estimate the model we use the ALS algorithm of Chapter 3. This algorithm is based on conditional minimisers of the blocks of slope parameters and factor loadings which allows estimation of the VAR equation-by-equation with little numerical difficulty. However, we cannot guarantee global convergence of such an algorithm and only local fixed points are guaranteed if the parameter space is convex.[1] A helpful numerical insight from experimenting with artificial data is that for the algorithm to work effectively, the VAR is required to be stable whilst bounds on the factor loadings are unknown. This implies we can choose starting values of all VAR slope parameters in the unit disk and apply the following algorithm for the k-th equation:

ALGORITHM 5.1:

1. Draw initial values for C_k from the uniform distribution on the interval $(-1, 1)$;

2. a) Using the initial values generated in 1, compute $\hat{\Lambda}_k^v$;
   b) Using $\hat{\Lambda}_k^v$ compute $\hat{C}_k^v$;
   c) Using the $\hat{C}_k^v$ compute $\hat{\Lambda}_k^{v+1}$;
   c) Repeat 2.a-c until the difference between $\hat{\Lambda}_k^v$ and $\hat{\Lambda}_k^{v+1}$, $\hat{C}_k^v$ and $\hat{C}_k^{v+1}$ is small in Euclidean norm and record the resulting value of the QDGMM objective function, J^w, say;

3. Repeat steps 1-2 many times and select the parameter set that corresponds to $\underset{w}{\text{argmin}}\,(J^w)$ from the set of J^w's with $w = 1, \ldots, \mathscr{W}$.

We will subsequently assume that the global minimum is found after sufficient repetitions of the algorithm and this is also the approach taken with many practical (non-convex, non-linear) GMM problems in the literature. For the problem at hand, we have set $\mathscr{W} = 5000$ for each equation of the VARs and typically find convergence to occur with differences less than 10^{-4} in Euclidean norm scaled by the total parameter count of the VAR equation both with simulated and real data. Finally, in accordance with Andrews (1997), we apply Algorithm 1 three times to obtain an approximation to the efficient GMM estimator. The first estimator uses a conformable identity matrix as weight matrix whereas the following two use the inverse covariance matrix of the previous estimators as weight matrices.

The moment conditions in (5.3) can hold only if we know the true number of factors in the data and we will use a generalization of the ratio estimator of Ahn and Horenstein (2013) to estimate this number: if

[1] See i.e Bertsekas (1999), section 2.7 or Grippo and Sciandrone (2000)

the estimator of C_p is consistent, then the residual of the first block of (5.1) stacked over all $i = 1, \ldots, N$ is the residual (5.2). The key insight of Ahn and Horenstein (2013) is that the factors component of (5.2) is of different order than the usual error component $\varepsilon_{k,t}$ as measured over $T^{-1}\sum U_{k,t}U'_{k,t}$ and as we know from Chapter 4, this property carries over to the ordered eigenvalues of the covariance matrix of the residual of a regression equation. We therefore evaluate the ratio λ_i/λ_{i+1} for each equation of the VAR of each sample and choose the number of factors where it is largest.

The final ingredient needed for impulse-response analysis is an estimate of the covariance matrix of the VAR (5.1) of each subsample. Direct estimation based on $\hat{U}_{k,t} = X_l - \sum_{p=1}^{P} C_{k,p}W_{-p}$ yields an estimate of the covariance matrix that paints an incorrect picture of the idiosyncratic component of the covariance matrix because the factors are not removed from that matrix. We will therefore estimate the average factor-corrected covariance matrix $\overline{\Omega}$ as follows: define $R_{\max} = \max[R_1, \ldots, R_K]$ to be the largest number of factors estimated across all K equations and define the orthogonal projector associated with the Λ_k of the $N - R_{\max}$ cross-sectional elements as M_k and the one associated with Λ_l of the $N - R_{\max}$ cross-sectional elements as M_l. Using these projectors and the $N - R_{\max}$ residuals $U_{k,t}$, we can get estimates of the k diagonal elements of $\overline{\Omega} = \sum_{i=1}^{N-R\max} \Omega_i$ as follows:

$$\begin{aligned} E\left\{M_k\left[\Lambda_k F + \varepsilon_{k,t}\right]\left(M_k\left[\Lambda_k F + \varepsilon_{k,t}\right]\right)'\right\} &= M_k E\left\{\varepsilon_{k,t}\varepsilon'_{k,t}\right\} M_k \\ &= M_k \Phi_k M_k, \end{aligned}$$

where $\Phi_k = \text{diag}\left(\Omega_{k,1}, \ldots, \Omega_{k,N}\right)$ is the cross-sectional covariance matrix of the k-th VAR equation which is diagonal by assumption. Taking the trace and dividing by $N - R_{\max}$, we find the k-th diagonal element of $\overline{\Omega}$ as:

$$\begin{aligned} \frac{1}{N - R_{\max}}\text{trace}\left(M_k \Phi_k M_k\right) &= \frac{1}{N - R_{\max}}\text{trace}\left(M_k \Phi_k\right) \\ &= \frac{N - 2R}{N - R_{\max}}\text{trace}\left(\Phi_k\right) \\ &= \left(N - 2R\right)\overline{\Omega}_{k,k}, \end{aligned}$$

and thus:

$$\overline{\Omega}_{k,k} = \frac{1}{N - 2R}\text{trace}\left(M_k\left[\frac{1}{N - R_{\max}}\Phi_k\right]M_k\right).$$

Off-diagonal elements are constructed in a similar fashion. Let $M_l = I_{N-R_{\max}} - P_l$ and let the cross-sectional

covariance matrix between the k-th and l-th equations be $\Phi_{k,l} = \text{diag}(\Omega_{k,l,1}, \ldots, \Omega_{k,l,N})$, then:

$$\begin{aligned}
\frac{1}{N - R_{\max}} \text{trace}(M_k \Phi_{k,l} M_l) &= \frac{1}{N - R_{\max}} \text{trace}(M_l M_k \Phi_{k,l}) \\
&= \frac{1}{N - R_{\max}} \text{trace}(\Phi_{k,l}) \, \text{trace}(M_l M_k) \\
&= \frac{1}{N - R_{\max}} \text{trace}(\Phi_{k,l}) \, \text{trace}([I_N - P_k][I_N - P_l]) \\
&= \frac{1}{N - R_{\max}} \text{trace}(\Phi_{k,l}) [N - R_k - R_l + \text{trace}(P_k P_l)],
\end{aligned}$$

so that

$$\overline{\Omega}_{k,l} = \frac{1}{[N - R_k - R_l + \text{trace}(P_k P_l)]} \text{trace}\left(M_k \left[\frac{1}{N - R_{\max}} \Phi_{k,l} \right] M_l \right).$$

We can estimate both the on and off-diagonal elements of the VAR covariance matrix from sample analogues and make use of the fact that the cross-sectional correlations do not matter. This is in line with estimation of $\overline{\Omega}$ based on standard OLS VARs. Note however that estimation of the covariance matrix in this fashion puts a further restriction on the number of factors one can estimate: where estimation of the VAR coefficients requires the number of factors to be smaller than the number of cross-sectional units, estimation of the covariance matrix requires $N > 2R_{\max}$.

Structural VAR Identification

Before we can discuss identification of the fiscal shocks we must first consider the implications of the SGP. Technically, the SGP imposes the following condition on the *annual* government deficit of each country:

$$\sum_{t=4\times s-3}^{4\times s} \frac{(G_{i,t} - R_{i,t})}{Y_{i,t}} \leq 0.03 \, \forall i = 1, \ldots, N, t = 1, \ldots, T \text{ and } s = 1, \ldots S, \tag{5.4}$$

where s is used to indicate the start of all S fiscal years in the sample. Since the SGP has been in operation in the sample of countries we consider, the model we estimate is conditional on being within the bounds of the SGP, especially when a debt-feedback is included. For that reason we will also examine a stronger version to the SGP. That is, we consider a strict version of the SGP by linearly interpolating the annual SGP over its corresponding quarters as averaged over N:

$$\frac{\overline{G}_t - \overline{R}_t}{\overline{Y}_t} \leq 0.0075 \, t = 1, \ldots, T \tag{5.5}$$

The crucial point is that strict interpretation of the SGP imposes a contemporaneous link between G and R at any time t and this has implications for the identification of impulse response functions of the estimated VAR. Similar studies seeking "kinked" dynamics are the monetary VAR models at the zero lower bound in Peersman (2011) and Schenkelberg and Watzka (2013). In these papers, (sign) identification in "non-

standard" times is derived from a theoretical model at an otherwise unobservable real zero-lower bound on the interest rate and compared with "normal times," where the difference stems from identification based on the dynamics of a model at, and away from, the zero lower bound. We offer a similar argument by treating the strict interpretation of the SGP (5.5) as "non-standard" times.

We have argued above that fiscal shocks are inherently difficult to identify and it is therefore problematic to impose a lot of structure on both the impact and the dynamic responses of the system to fiscal shocks. For this reason, we will follow Mountford and Uhlig (2009) by implementing a set of zero and sign-restrictions that deliver prototypical fiscal shocks to our model whilst maintaining the assumption that fiscal variables do not contemporaneously respond to one another. Table 5.2 summarises the identified shocks and restrictions:

Table 5.2: Shocks and sign-restrictions to identify model (5.5):

Shocks		G	R	Y	P	I	$\frac{G-R}{Y}$
Fiscal	Revenue Shock	0	<0	...	...	...	$\leq SGP$ for $t=1,2,\ldots,\tau$
	Balanced budget Shock	$G=R>0$	...	...	...	...	0 for $t=1:\leq SGP$ for $t>1$
	Expenditure Shock	>0	0	...	...	...	$\leq SGP$ for $t=1,2,\ldots,\tau$
Other	Liquidity Shock	...	...	...	...	>0	$\leq SGP$ for $t=1,2,\ldots,\tau$
	Business Cycle Shock	...	>0	>0	...	...	$\leq SGP$ for $t=1,2,\ldots,\tau$

The shocks in table 5.2 are directly adapted from Mountford and Uhlig (2009), apart from the liquidity shock. The other new restriction is the requirement that $\left(\overline{G}_t-\overline{R}_t\right)/\overline{Y}_t \leq SGP$ for $t=1,2,\ldots,\tau$: The SGP thus operates as an additional constraint on the impulse response functions over the full forecast horizon, and when $SGP=0.0075$, it mimics the quarterly approximation to the true *annual* SGP as discussed above. In all cases, the response variables that have signs imposed by a particular shock are restricted to maintain that sign for at least four quarters and are thereafter unrestricted, other than the SGP-constraint which binds over all forecasting periods up to τ. On the other hand, when dynamics are unrestricted, they are denoted by "...".

The pure revenue and expenditure shocks in table 5.2 correspond to archetypical fiscal shocks found in many studies employing conventional restrictions on the covariance matrix of model (5.1) but do not employ unrealistic contemporaneous zero-restrictions on any of the non-fiscal response variables Y, π and I by leaving both their impact and dynamic responses fully unrestricted. On the other hand the zero restriction on the mirror fiscal variable in the pure shocks is consistent with the argument of BP that one fiscal variable will not affect another fiscal variable in the same quarter. As such, this identification strategy harmonizes sign-restrictions with the BP-identification above and further separates them from the balanced budget shock: the balanced budget shock is another special case where taxes finance expenditure exactly in the first period, in addition to both fiscal variables being larger than zero for the first four periods. We consider such a shock to be prudent fiscal stimulus to the economy which is intended to show if net-expenditure stimulates the economy. We also identify two non-fiscal shocks because, as Mountford and Uhlig (2009) argue, non-fiscal shocks are required the fiscal shocks in order to identify the fiscal shocks and this means they must be causally prior. The first is a liquidity shock which yields a one-off increase in the interest rate on the

remainder of the system. Such a shock may result in acute refinancing problems and we imagine it has detrimental effects on a country under refinancing constraints. We also identify a business cycle shock to disentangle fiscal from "other" shocks hitting the system: the business cycle shock has direct impact on tax receipts and GDP, but has otherwise unspecified dynamics. Such a shock is consistent with the notion of automatic stabilization: when output increases, so do tax receipts.

Implementation of both sign- and zero-restrictions is done by the methodology of Arias et al (2018) which entails finding orthogonal matrices Q which are uniformly distributed with respect to the Haar Measure on $O(K)$. This technically necessary condition implies a unique mapping of measure one and is invariant under rotations and reflections, thus leaving the distributional properties of the data unchanged. Zero restrictions, or equality restrictions in the case of the balanced budget shock, are found by imposing linear restrictions on the random matrix Q and then projecting onto the null-space of the previous columns by means of the QR-decomposition, rather than imposing non-linear constraints on the system directly, which is a substantially more difficult problem. Finding such matrices is a generalization of the usual sign-restrictions algorithm as in Rubio-Ramirez, Waggoner and Zha (2010) and is straightforwardly implemented on the estimated covariance matrix of a VAR: let $\overline{\Omega}$ be the reduced form covariance matrix of equation (5.1) and $\overline{\Omega}^{1/2}$ its Cholesky factor so that $\overline{\Omega}^{1/2}\overline{\Omega}^{1/2\prime} = \overline{\Omega}$. Then, for conformable Q and by orthonormality, $QQ' = I_K$, so that $\overline{\Omega}^{1/2}Q\left(\overline{\Omega}^{1/2}Q\right)' = \overline{\Omega}^{1/2}$ and exploration of the space of Q matrices is equivalent to exploration of valid sign-restrictions on the impulse-response functions with zero-restrictions in place. Drawing many matrices Q and storing those satisfying the sign-restrictions thus yields an impression of the possible models that can be generated from the estimated VAR without committing to a fixed identification pattern like Cholesky or indeed the BP-VAR.

We now replace C_p, Λ_k and $\overline{\Omega}$ by estimates and assume the $\varepsilon_{k,i,t}$ are normally distributed. Then conditional on the $\hat{C}_p$, $\hat{\Lambda}_k$ and $\hat{\overline{\Omega}}$, the following Monte Carlo procedure is used to obtain a distribution about the impulse response functions:

ALGORITHM 5.2:

1. Draw residuals ε_t^v from the estimated and centred $\hat{\overline{\Omega}}$ and simulate the data using the starting values of the data, the exogenous variable D and the estimated $K \times (K \times (1+P)-1)$ VAR slope parameters $\hat{C}_p$ for $p = 1, \ldots, P$;
2. Estimate the reduced form VAR (5.1) by panel OLS and record the $\hat{\overline{\Omega}}^v$ and $\hat{C}_p^v$ for $p = 1, \ldots, P$;
3. Draw $w = 1, \ldots, \mathscr{W}$ random orthonormal matrices matrices Q that satisfy the zero restrictions and simulate impulse response functions, save those which satisfy the sign restrictions;
4. Repeat steps 1-3 $v = 1, \ldots, \mathscr{V}$ times to obtain a distribution about the impulse response functions with sign and zero restrictions imposed.

Algorithm 5.2 is a bootstrap analogue of the Bayesian methodology developed in Arias et al (2018) and is asymptotically equivalent to a Bayesian re-sampling scheme conditional on the QDGMM parameters using

a fixed prior for the covariance matrix $\overline{\Omega}$. Note that for each OLS repetition, we draw $\mathscr{W}$ orthonormal matrices Q rather than until a certain number of Q satisfy the sign restrictions. This means that the weight given to models with many Q satisfying the sign restrictions is high and, following Arias et al (2018), needed for correct posterior inference.

5.4 Results

On Estimation of the Model and the Number of Factors

We estimate the three VARs for Eastern, Southern and Western Europe using the tools developed in the previous section. Since we deal with quarterly data, the number of lags is set at $P = 4$ and we absorb constants and trends in the factor space. In estimating model (5.1), we have used all cross-sectional units of all equations as instruments for the GMM procedure after applying the quasi-difference procedure to the data. To keep the number of instruments as low as possible we have only used the first four lags of all cross-sectional units as instruments in each equation: this is the least required number of instruments to estimate the corresponding VAR parameters consistently. However, as a result of the choice of instruments, the weight matrices of the second and third step estimators are no longer positive semi-definite and we are forced to use a generalized inverse instead to weight the estimator. Estimation is subsequently undertaken using Algorithm 5.1, where we have started the algorithm 5000 times until convergence for each step of a three-step QDGMM estimator. We find the same parameter convergents over the search space with differences no larger than 10^{-4} in individual parameters using different seeds of the random number generator. All VARs we estimate are stable in the sense sense that no eigenvalues of the companion form constructed from the coefficients of the endogenous variables lie outside the unit circle.

We then apply Algorithm 5.2 with $\mathscr{V} = 5000$ and $\mathscr{W} = 10000$ to obtain a bootstrap distribution about the impulse-response functions. The yield of permissible models under the imposed sign-restrictions is typically in the range of one-hundredth to one percent of each bootstrap iteration and we thus obtain a distribution about the models that is larger than the number of bootstrap iterations. This means that we study jointly more than one identified model and the result is that some of the impulse-response functions are more noisy than those typically reported in the literature. Since all data apart from the interest rates on long government paper are in logs, the interpretation of the impulse responses is that the percentage change in one variable leads to percentage change in another variable.

Finally, tables 5.6-8 in Appendix 5.1 contain the number of factors we find in the data of all three VAR models using the specification of a testing-up procedure as discussed before. When the number of factors is unclear, we estimate the model again but quasi-differencing once more and then analyse the eigenvalue ratios of the covariance matrix $T^{-1}\sum_{t=1}^{T} U_{k,t}U_{k,t}'$ for the k-th equation and the largest eigenvalue ratio is then subsequently taken to be the number of factors. We find clear evidence of factors driving at least part of the residual of many of the VAR equations. Particularly, the I and π equations contain at least one factor in all three models, but up to two factors not uncommon in many models. In other VAR equations however, the

results are less clear as can be seen from the tables and the testing-up procedure is followed to the outcome of the number of factors in the final columns of the table for each equation in each sample.

Interpretation of Impulse-Response Functions without the Strict SGP Imposed

Appendix 5.2 contains the impulse-response functions based on the QDGMM estimator without the strict SGP imposed on the three panels. The impulse-response functions show the 16th, 50th and 84th percentiles of the distribution obtained from algorithm 5.2.

Table 5.3: Percentage QDGMM impact multipliers to fiscal shocks without strict SGP.

	Western Europe			*Southern Europe*			*Eastern Europe*		
	Y	*P*	*I*	*Y*	*P*	*I*	*Y*	*P*	*I*
Balanced budget shock	0.49*	-0.01	0.00	1.58*	-0.53*	-0.33	0.22	0.03	-0.01
Expenditure shock	0.09	-0.01	0.01	-0.02	0.00	-0.03*	0.21*	0.03	0.00
Revenue shock	-0.03	0.02	0.00	0.32	-0.18*	0.25	0.03	0.03	0.01

(*) implies the impact multiplier is within the confidence bands.

We consider first the balanced budget shock. The balanced budget shock yields strong sustained responses in all three models and the fiscal variables slowly move back to zero over a period of over three years. As can be seen from table 5.3 above, the response of GDP on impact is significant in Western and Southern Europe but not in Eastern Europe. The strength of the impact multiplier especially in Southern Europe is striking, as these countries suffered the most from the crisis in the period of the sample and suggest that responsible stimulus is not only desired but has strong effects on the economy. Over time, the response to a balanced budget shock is always positive and strongest in Western and Southern Europe where the positive effect on GDP persists for up to three years. In Eastern Europe, the admissible models are much wider and the balanced budget shock takes almost three years to become significant. Inflation responses are insignificant in Eastern and Western Europe but fall on impact in Southern Europe and rise as output increases. The response of the rate on government paper to a balanced budget shock is also insignificant in Western Europe, as we expect investors to not respond to the fiscal policy choices of well-perceived governments in Western Europe much. In Eastern Europe, by contrast, a small rise in the interest rate is perceived after about two years when the response of tax collection becomes insignificant and the spending component of the shock is no longer matched. The interest rate on Southern European debt visibly decreases over time when the balanced budget shock hits the system and both receipts and spending are significantly different from zero. This also suggests that investors in Southern government paper value responsible fiscal policy in the South of Europe strongly in a sample period that includes both the Financial and Debt crises. The debt-to-GDP ratios fall in the short run in all three regions although its effects are rather modest and return to parity within five quarters. The dynamics are characterised as follows: whilst the budget deficit is small and GDP grows, the stock of debt is discounted and its per capita value is reduced.

The own-responses of a pure government expenditure shock are very similar across the regions and the shock dies out within five and twelve quarters respectively. Only in Eastern Europe is the impact multiplier

of output significantly different from zero. In Southern and Western Europe, the output multiplier instead is insignificant and neither taxes nor output move on impact. In Western Europe, the expenditure shock is followed by a dip in taxes and output around the tenth quarter, whilst output and taxes rise with a delay of about five quarters in Sourthern Europe. By contrast, expenditure is matched by taxes in Eastern Europe and GDP rises for four quarters as a result. Eastern and Western Europe have insignificant responses of inflation to the government spending shock, whilst inflation increases with output in Southern Europe and persists for another ten quarters. The interest rate is insignificantly affected by the spending shock in Eastern and Western Europe, although it falls slightly as output rises in Southern Europe. Finally, it should be expected that the debt-to-GDP ratio rises directly as a result of the spending shock and does so most strongly in Western and Southern Europe as a result to a pure expenditure shock. This effect occurs with a delay in Eastern Europe as revenues rise with output in the short run.

An unanticipated tax shock dies out within five quarters in Southern and Western Europe and does not have a significant impact multiplier of output. On the other hand, a tax shock raises government expenditure with a time lag, which in turn stimulates the economy. This effect is strongest in the Eastern European countries and weakest in Southern Europe, where the effect is insignificant in all but the third quarter following the shock. Of all the fiscal shocks, indeed, Eastern Europe seems to be benefited most by fiscal policy which first raises a firm basis in tax collection before trying to stimulate the economy with fiscal spending. On the other hand, the effect of the revenue shock on inflation is small and insignificant in Eastern and Western Europe, only rising somewhat as output rises in response to tax-stimulated government expenditures. In the Southern region by contrast, inflation falls significantly in the first year following the shock and a slowdown of inflation appears to be linked to tax receipts. The interest rate is again insignificantly affected in the West, although the model now suggests that rising taxes point to falling costs of government borrowing, with the impulse response broadly mirroring that of the expenditure shock. In Eastern and Southern Europe the tax shock has opposite effects on the interest rate: in Eastern Europe, after initial fiscal balancing, the very strong and persistent effect on government expenditure pushes the interest rate up after six quarters for a period of roughly fifteen quarters. In Southern Europe by contrast, the interest rate lags behind the tax receipts and falls as a result. However, as the tax shock dies out and spending rises, the interest rate also moves back up to its initial value. The debt-to-GDP ratios of Eastern and Western Europe fall initially for up to twenty quarters following the tax shock. In Eastern Europe, the rising interest rate thereafter pushes the ratio to a level slightly higher than before the shock occurred although not significantly so. In Western Europe, the result is an indefinite reduction in the debt-to-GDP ratio: for the duration of the shock, the West runs a surplus, whilst the interest rates are not affected much and output growth and inflation experience a slight increase, thus eroding the debt. In Southern Europe by contrast, the debt-to-GDP ratio is unaffected by a revenue shock, which implies that the countries in the sample are debt-constrained: even as taxes rise, the debt does not fall significantly.

The effects of a business cycle shock has somewhat similar effects on all three regions under consideration: the increase in taxes brought about by the economic upturn persists for at least eight quarters whilst output also stays positive for at least six quarters. The effect on government expenditure is somewhat dif-

ferent across the blocks however: in Eastern Europe, government expenditure sees a small but persistent increase, which further amplifies the upswing on output and receipts. In Southern Europe by contrast, government expenditure initially falls, reflecting the times of crisis that the Southern region has been exposed to the in the past decade. In the West, government expenditure is initially unaffected by the cycle although it turns significantly positive for approximately two years afterwards. As expected, inflation moves with rising output in Eastern and Western Europe with a delay. In Southern Europe, we observe a temporary deflation linked to increased tax receipts. Regarding the interest rate in Western Europe, we find very little response of these variables to the business cycle. In Eastern Europe, we observe a temporary hike after approximately two years occurs in response to rising deficits. Significantly different dynamics are observed in Southern Europe: the combination of an initial government surplus on impact, coupled with positive economic growth yields a sharp and significant reduction in the rate on government paper. However, as soon as the growth declines and the surpluses move towards deficits, the interest rate actually rises to above its original level before returning to parity. As we would expect, in the short to medium run, debt-to-GDP ratios in all three regions decline but move up in the medium run to approximately their initial levels. The reason is clear: as deficits fall or become surpluses and economic growth erodes the cost of servicing debt, debt-to-GDP ratios fall.

Finally, regarding the shock to interest rates on government bonds, it is clear that the responses to this shock are poorly estimated in Eastern and Western Europe: either the confidence bands are very wide, suggesting that many models fit the estimated reduced form, or they are centred around zero, or both. The response of all variables to a liquidity shock is typically insignificant, both on impact and over the forecast horizon, which suggests that government expenditure does not respond to the cost of repaying the debt much in the short run. This is not unreasonable, as no countries in these samples have had re-financing difficulties over the course of the sample period. In the case of Southern Europe however, the story is quite different: as interest rates rise, both expenditures rise and revenues fall which result in an increase in output. The interest rate keeps rising however whereas the response of the fiscal variables dies out within a year and induces a surplus, after which the variables move back to their initial values over time. This result is puzzling, although it is consistent with the following arguments: first, the shock may not actually identify a liquidity shock but instead an irresponsible fiscal policy shock. That is, slashing revenues and raising expenditure stimulates the economy, but also makes investors nervous. The result is rising interest rates and as governments in the South fear they cannot service their debt, they quickly move into surpluses again. Second, as Southern European countries become increasingly unable to service their debt, the European Union offered collateral which gives breathing room to national governments in the South. This implies that rising interest rates actually operated as short-run stimuli to debt-constrained countries in the South, through the mechanism of guarantees by the European Union.

Discussion

It is difficult to directly compare our results with the existing literature: most studies focus solely on the US and the identification approach is often either recursive or using the BP approach. Moreover, the exact specification of the VARs is different and/or only a subset of the shocks and their responses is studied and, as we have seen before, these differences can be quite substantial. However, as we have noted, the fiscal shocks in the BP identification scheme can be interpreted as being either pure tax or government expenditure shocks. Similarly, the first shock of a recursive scheme with either taxes or expenditure ordered first can be interpreted as a pure shock in either variable. On the other hand such schemes lack a balanced budget shock, which we have identified as the most important shock for all three regions in Europe.

BP (2002) study the effects of tax and spending shocks in the US and their results show that a positive tax shock reduces output significantly and government spending falls slightly, whilst a spending shock raises taxes and eventually GDP. These findings are essentially echoed in Fatas and Mihov (2001) who study a recursive VAR of the US economy. Compared to our findings for the three European regions, these results are quite different however: instead, we find that government expenditure has positive (auto-) correlation with the tax shock and this interplay may actually increase output. Regarding the effect of a positive spending shock, we do not find an unequivocal increase in output in all European region. Rather, the strength and persistence of the shock are actually closely linked to the sustainability of the response as matched by sufficient tax receipts. This finding is in agreement with a form of Ricardian Equivalence: in countries with strong automatic stabilizers, the economy will not resond to fiscal stimulus much unless it does not come at the cost of future taxes. We offer two points that may help explain these differences: first of all, we have aggregated macroeconomic data of relatively short time series, a time of considerable uncertainty with two large area-wide crises and overall uncertainty about the sustainability of EMU where countries gave their national monetary policy abilities over to the supranational ECB. The second point is that drawing inference for economies across the world based on only the US might lead to rather incorrect conclusions. Indeed, in Perotti (2004) some nuance is added to these conclusions by studying several OECD countries individually and the effects of negative tax shocks on GDP can actually be negative. Comparing the results of a positive government expenditure shock in Ilzetzki et al (2012), we immediately run into the problem of the division of data: the countries in all three regions are classified as high-income in that paper, where by contrast we find substantially different dynamics between the regions, thus suggesting that the aggregation in Ilzetzki et al masks these differences. Moreover, whilst the impact multipliers of the spending shock are comparatively small, we do not find significantly positive responses in all three regions. Beetsma and Guiliodori (2011), using annual data, find strongly increasing output dynamics and falling taxes on impact in their recursive panel VAR based on fourteen original EU members, including Denmark, Sweden and the United Kingdom. We attribute the differences in estimation results to the use of annual data, which obscures quarterly effects, whilst the pooling of two completely different regions, as well adding in several countries that are not member states of EMU, will aggregate, or indeed, amplify the strongly positive effects of the spending shock in the South compared with insignificantly small effects in the West.

Regarding fiscal VARs that are identified with sign-restrictions, the balanced budget scenario as identified for the US in Mountford and Uhlig (2012) yields the exact opposite of our findings: Where we see the strongest increases in GDP, Mountford and Uhlig find a significantly negative effect, reflecting the fact that taxation is low and the taxation multiplier strongly negative, whilst the government multiplier is comparatively small in the US. A further point relates to their definition of government revenues and expenditures, which is not as comprehensive as our measure. Such findings are also broadly in line with Canova and Pappa (2005) and the results in Pappa (2009) for the US. Moreover, Pappa's results for an aggregated measure of EU macroeconomic data also confirm our findings regarding the fiscal shocks, although we do not distinguish between government consumption and investment but instead aggregate these series.

We have seen that the effect of fiscal shocks on the interest rate has a strikingly strong correlation with the ability to increase the tax revenues of the European periphery, whilst the interest rate of Western Europe is largely unaffected. It is not difficult to argue that the strength of the Western European economies makes their governments be perceived as credible and trustworthy borrowers, so that financial markets do not require compensation when unanticipated fiscal shocks occur in that region. This result is broadly similar to the results from Fatas and Mihov (2002), Pappa (2005) and Perotti (2004) for Europe, at least when spending shocks are considered. Giuliodori and Beetsma (2005) find similar results for France and Germany in their VAR results, specifically with insignificant responses of the interest rate variable. Contrarily, in Eastern and Southern Europe, we see the interest rates on government paper rise when the revenue basis for fiscal expansion erodes. The extreme example is Southern Europe: when taxes are increased, a sharp decline in the interest rate is observed. This has to do with the nervousness of creditors to those countries, whose perception of improvement fiscal stances comes with substantial refinancing rewards. However, as soon as the tax shock dies out, the interest rate hikes up again, sometimes to higher values than before because the revenue shock did little for repayment of outstanding debt. Again, Giuliodori and Beetsma (2005) find a similar result for Italy using their single-country VAR. This idea is also complimentary with the argument of Fernandez-Villaverde et al (2015), who explain such unexpected responses of interest rate and price variables by means of periods of fiscal uncertainty.

The result that inflation does not move significantly with fiscal shocks in Western Europe conforms with the findings of Giuliodori and Beetsma (2005) for Italy and Germany and Fatas and Mihov (2004) for the US. Interestingly, the sign restrictions of Mountford and Uhlig (2012) show insignificant responses of inflation to fiscal shocks only when they are anticipated. When they are surprise shocks, they have opposite effects: negative for expenditure and positive for taxation shocks. Furthermore, only when a positive business cycle shock hits the system does inflation move significantly in their specification. With monetary policy out of the hands of the individual countries of Western Europe yet possibly somewhat biased towards that region by the ECB, we would expect only minimal responses to fiscal shocks in the region. This finding is in line with the result of Mountford and Uhlig (2012) for the US, although we do only find a small peak at the fourth quarter but do not observe an eventual significant rise as they do after some ten quarters. The finding that inflation drops significantly and strongly as a result of rising taxes in Southern Europe is somewhat surprising at first glance. However, we must remember that the sample contains both the mortgage and debt

Table 5.4: Percentage OLS impact multipliers to fiscal shocks without strict SGP.

	Western Europe			*Southern Europe*			*Eastern Europe*		
	Y	*P*	*I*	*Y*	*P*	*I*	*Y*	*P*	*I*
Balanced budget Shock	0.57*	-0.06	0.02	0.72*	-0.16	-0.16	0.54*	0.04	-0.03
Expenditure Shock	-0.04	-0.02	0.01	-0.10	0.00	-0.09	0.51*	0.03	-0.01
Revenue Shock	0.08	0.01	-0.01	0.04	0.01	0.02	0.18*	0.02	-0.02

(*) implies the impact multiplier is within the confidence bands.

crises, which forced strong negative wage renegotiations in some of these countries. Thus, as taxes rise and disposable income falls, inflation must fall to compensate for bloating inventories. On the other hand, Southern Europe does not control its monetary policy anymore due to EMU and therefore could not slash interest rates to try to spur growth and inflation. In times of financial turmoil, strained means to fight these recessions thus yielded falling inflation when tax revenues are increased. Such fiscal volatility has been studied in Fernandez-Villaverde et al (2015) and indeed they find that increasing fiscal uncertainty yields falling prices.

Comparison with Panel OLS Results

We have also estimated the three VARs using panel OLS to compare with our QDGMM methodology: we know that factors are present in the data of all three VARs but we want to get a sense of the potential differences/bias from estimating the model when ignoring the factors. We have experimented with deterministic terms and found that there is little difference in the dynamics of models estimated with no constant, a constant or fixed effects. Therefore, to conserve both space and degrees of freedom, Appendix 5.3 only reports OLS results without any deterministic terms and as before we have set the number of lags equal to four and subsequently calibrated the debt-to-GDP ratio. The VARs for Eastern and Western Europe are stable as measured by the eigenvalues of the VAR companion matrix, whereas the VAR for Southern Europe is not.

As can be seen from table 5.4, impact multipliers of fiscal policy shocks on output are smaller and often non-significant with panel OLS, although all impact multipliers of output for Eastern Europe are now significant. Moreover, the confidence bands on OLS are quite wide for the fiscal variables and in general OLS responses have more often insignificant impulse-response functions than the QDGMM results and many more rotations fit the data. These findings show that not accounting for the factors leads to substantial bias in the results and that interpretation of such results is very prone to error.

Regarding Western Europe, we see that OLS responses are typically far less accurately estimated than the QDGMM responses. Compared to the OLS estimates, the QDGMM response of output is less persistent than the one obtained through OLS. This is because the own-response of the fiscal variables is less pronounced. As a result, the debt-to-GDP ratio is structurally higher in the OLS calibration. The spending shock never has significant effects on taxes and output in the OLS estimates and is very poorly estimated. A broadly similar effect is observed for the tax and business cycle shocks in both OLS and QDGMM responses, although the timing is different and more delayed in the QDGMM results. The response of the debt-to-GDP ratio derived

from the OLS estimator is always in the same direction as the QDGMM estimator, but larger for the OLS responses. Inflation is broadly similar in shape in both estimators, but the response of the interest rate is very different, although insignificant in both.

The signs of the fiscal shocks in Eastern Europe are also broadly similar in the OLS and QDGMM responses, although the QDGMM responses are far more persistent. Because the covariance matrix estimated from OLS responses is far larger in magnitude, the debt ratio behaves far more erratically, although the shape of the response is broadly similar. The OLS responses show increases in the interest rate after about six quarters in all fiscal shocks as the tax basis for fiscal expansion deteriorates. This is similar to the QDGMM responses although those are not significantly estimated as such. Broadly similar behaviour is found for the business cycle and interest rate shocks where we note that the business cycle shock yields far less persistent responses of all the variables in the system.

Since the panel OLS estimates yield an unstable VAR for Southern Europe, the responses of fiscal variables are far more persistent as compared to those obtained by QDGMM. OLS results suggest that a balanced budget shock yields strongly negative government expenditure after an initially positive spending response in addition to strongly positive tax receipts. This coupled with positive GDP growth yields an unrealistically large and sustained decline in the debt-to-GDP ratio and only the price response appears broadly similar. The government expenditure shock is much stronger on impact but appears to have the same shape and sustain. However, since taxes do not move in response to the spending shock, output is insignificant over the whole forecast horizon. This would suggest a similar dynamic as with the QDGMM estimates, although this finding is far less clear. Furthermore, since the median response of tax receipts is now negative (but insignificant), the debt-to-GDP ratio now explodes. A similar effect occurs with the pure tax shock, mirroring negatively the spending shock. The business cycle shock has an insignificant effect on impact for the spending variable which turns somewhat negative over time. This is very different from the negatively oscillating effect found with QDGMM. The effect of the output and tax variables is however somewhat similar and only for the restricted first four quarters are they positive. As a result, the debt-to-GDP ratio again violently drops over the entire forecast horizon. Finally, the interest rate shock has similar dynamics in both models, but the effect on output now is slightly negative after about two years. The fiscal variables display an initially positive spending and insignificant revenue effect, but revenues actually fall with the interest rate over time again. As a result of the falling revenues, now the debt again explodes.

Impulse-Response Function with the SGP Imposed

We have argued in Section 5.3 that the SGP is an additional constraint on the impulse-response functions, in addition to the sign and zero-restrictions in the previous section. This means that the impulse-response functions that satisfy the SGP and the sign-restrictions are a subset of the impulse-response functions that satisfy only the zero and sign-restrictions. As a result, imposing the SGP is straightforward and the comparison is direct. Studying that subset then allows us to study if 'responsible' fiscal policy as measured by an SGP-type rule suggests better stimuli and at the same time asses how the one-size-fits-all nature of the

Table 5.5: Percentage QDGMM impact multipliers to fiscal shocks with strict SGP imposed.

	Western Europe			*Southern Europe*			*Eastern Europe*		
	Y	*P*	*I*	*Y*	*P*	*I*	*Y*	*P*	*I*
Balanced budget Shock	0.58*	-0.01	0.02	1.83*	-0.59*	-0.34	0.19	0.06	-0.03
Expenditure Shock	0.05	-0.01	0.00	-0.02	0.00	-0.04	0.19*	0.00	0.00
Revenue Shock	0.04	0.03	0.01	0.35	-0.17*	0.23	0.10	0.01	0.03

(*) implies the impact multiplier is within the confidence bands.

policy works out in the regions. Another important point is that the SGP has been operational over the full length of the sample in the EMU countries and imposing the SGP based on equation (5.4) is expected to have little impact on the impulse response functions. This is indeed the case: when drawing the distribution and identification structures of the impulse-response functions in the previous section we find them to be no different than the impulse-response functions found with the SGP imposed. However, as one would expect, the strong interpretation of the SGP as a quarterly constraint does have some impact. Since the constraint implies either government expenditure or taxation is directly changed to satisfy the SGP, compensatory movement of the fiscal variables spread out over four quarters is no longer possible and we expect to see a reduction in the ability of fiscal policy to stimulate the economy; The impact multipliers are summarized in table 5.5 whereas impulse-response functions are in Appendix 5.4.

The subset of the impulse-response functions that are restricted to satisfy the strict SGP, lead to the following conclusions. First of all, the Southern European responses are qualitatively the same as those without the strict SGP imposed. This finding implies that Southern Europe has been forced into extremely prudent fiscal policy over the sample period as a result of being hit hardest by both the mortgage and debt crises as investors were nervous about their ability to service government debt in addition to a response through the EDP. These constraints further offer an explanation as to why both balanced budget and pure government expenditure shocks stimulate the economy so strongly in Southern Europe: since the scope for fiscal stimuli is limited, any stimulus improves the condition of the economy over a prolonged period of time and any shock that increases taxes reduces the interest rate, thus further alleviating immediate pressure on repayment of the government debt.

The impulse-response functions for Eastern Europe are also largely similar, reflecting their attempts to ascend into EMU in a decade that was riddled by crisis. Only in the case of the revenue shock does strictly adhering to the SGP yield a significantly different dynamic: when expenditure is restricted to be within the bound of the SGP, output becomes significantly positive in approximately two rather than five quarters and so does the rate on government paper. As we have seen before, the Ricardian equivalence argument works strongest in the East and it is therefore not surprising that prudence is valued in the region. This also has an effect on the debt-to-GDP ratio, which after an initial dip rises faster because government expenditure is higher in response to the initially higher taxes.

Finally, in Western Europe, the strict SGP makes balanced budget shocks far less capable of stimulating the economy. That is, although the impact multiplier is still positive and significant, as taxes fall after

approximately three years, the compensating fall in expenditure pushes output into the negative. Whilst the effect on the interest rate is still insignificant in the first two years, it rises marginally as the fall in tax receipts leads the fall in expenditure. This outcome has an important effect on the calibration of the debt-to-GDP ratio, which now rises more strongly with the balanced budget shock. Something opposite is found with the revenue shock: whilst revenue clearly moves first, the need to compensate subsequent expenditure with revenues reduces the power of fiscal stimuli on GDP than in a scenario where the strict SGP is not imposed and output thus rises far less. However, the erratic behaviour of the government expenditure variable reverses the shape of the median interest rate and, as a result, rather than observing a fall in the debt-to-GDP ratio, we find that it increases now. The responses of the government expenditure, business cycle and interest rate shocks actually differ very little in either case, showing how Western European governments apparently follow a very strict interpretation of the SGP when compensating variables in accordance with the SGP. The only difference between the two identification strategies is that the interest rate is either slightly higher or lower and as a result, so is the debt-ratio.

In summary, several remarks regarding a strict interpretation of the SGP are in order. Although the crisis has constrained Southern Europe to adherence to very strict fiscal policy rules, we find that strict rules impede on the ability of Western Europe to stimulate the economy. On the other hand, in Eastern Europe, we find that a stronger SGP may actually help strengthen the effects of fiscal stimulus, a result we could interpret as giving credibility to the policy makers there. These results suggest that one single rule to govern EMU fiscal policy may not be the best way to restrict local governments from overspending. Moreover, we have used the debt-to-GDP ratio to assess the degree to which the SGP brings thrift to the Euro zone. As we have seen, the effect of either a weak or strong SGP has ambiguous effects on the debt-ratio and may actually exacerbate the negative effects of rising debt-ratios, something that the SGP was designed to explicitly avoid.

5.5 Conclusions

In this paper we have given an extensive overview of the empirical evidence on fiscal policy and the SGP specifically. The empirical literature fears that the adoption of the SGP will deny the union from access to fiscal stabilization and will make fiscal policy more pro-cyclical. On the other hand, the VAR literature has spent considerable effort on finding "the" effect of fiscal shocks on the economy. Various identification strategies have been applied to VARs estimated from data of varying countries and periods. From a bird's eye perspective, results based on differing identification strategies are broadly similar in terms of signs and shapes of the responses, although substantial differences exist when comparing VARs of different countries and/or periods of time. These results are further complicated by the strength of automatic stabilisers in the country and period under scrutiny and the definition of fiscal variables in the study at hand. As has been noted by others, it is therefore difficult to conclusively characterise the strength and shape of fiscal impulse-response functions in the general case. We have subsequently developed a methodology to estimate panel VARs in the presence of factors: we assume, and subsequently find, that factors are present in the panel

data under consideration and correct omitted variable bias due to the factors by using a quasi-difference methodology, which is then estimated through non-linear GMM. Our stance towards the factors is therefore a purely practical one, regardless of what they are, their presence requires us to take them into account to improve the precision of our inference. Our estimator furthermore holds the cross-section fixed in an effort to reduce excessive cross-sectional heterogeneity and thereby justify pooling of the datasets. The results we then find are quite different from pooled OLS estimates: the impact multipliers of the shocks are larger for the QDGMM estimator and the shapes of the responses also differ substantially, suggesting that not correcting for the factors leads to rather wrong inference on the estimated VARs.

Our results show that there is relatively little evidence for the theoretical concern of intentional violation of the SGP by member states, at least on average. Instead, we find that Southern European countries have been under the threat of debt-default which results in a strict adherence to the SGP. In Eastern Europe by contrast, more responsible fiscal policy as measured by an SGP-type rule can actually improve the potency of fiscal stimuli. Finally, only in Western Europe do we observe a reduction in potency of fiscal policy for stabilizing the cycle. In all cases, such a rule does very little to reduce debt burdens in the Euro-zone. This finding is complementary to the view that one-size-fits-all is not the best way to constrain individual governments from overspending in EMU. We interpret these findings as corroborating the call for a supranational tax and transfer system in the Euro-zone. Regarding the debate on volatility and cyclicality of fiscal policy in EMU, we note that whilst the VAR methodology we have developed cannot account for such characteristics directly, it does appear to numb fiscal policy in Western Europe whilst Southern Europe is severely constrained in its policy choices by forces outside of the SGP. This result is taken as further evidence of sterilization of fiscal stimulus in Europe due to EMU, although other countries stand to gain from more credible fiscal policy as we have seen is the case in Eastern Europe. This observation together with the fact that our sample contains two large economic crises thus seems to suggest that the SGP has not been very helpful to European stability.

Placing our results in the wider body of fiscal VARs, we echo the observation that heterogeneity in results based on the countries studied is very clearly visible in our findings. Compared to the mostly US-based evidence in the literature, we find that in Europe the best recipe for fiscal stabilisation or indeed stimulation is when government increases taxes along with expenses to develop a strong basis for a subsequent expansion. In Eastern Europe, this effect is especially pervasive and an increase in taxes yields so strong a tax basis to finance government expenditure for a prolonged period of time and subsequently stimulates output substantially. Moreover, it would appear that a Ricardian Equivalence argument can be put forward to pure government expenditure shocks in the sense that the economy does not move much but instead anticipates either taxes to rise in response or that the shock is transitory. On the other hand, in Southern Europe and Ireland, a balanced budget and pure expenditure shocks yield extensive output growth. Since these countries were were hit most by the crisis and severely constrained in discretrionary fiscal policy by outside sources, it is therefore no surprise that their economies thus move accordingly even without a tax basis to finance expansionary fiscal policy.

As for future research, it would be very interesting to apply the QDGMM methodology to other VAR

problems. More specifically, we can imagine that panel VARs estimated from Atlantic, Asian and South American countries would be an important extension to the study undertaken in this paper and the comparison with the results in this paper would shed further light on the strength and persistence of fiscal multipliers in a variety of economic circumstances and systems. That is, working under the factor hypothesis with VARs estimated from suitably similar countries, using data collected with similar definitions could aid our understanding of what determines effective fiscal policy and if there are circumstances where fiscal policy is not.

Appendix 5.1 Eigenvalue Ratio Tests for the Number of Factors

Table 5.6: Western Europe

λ_i/λ_{i+1}						
Equation:	1	2	3	4	5	# Factors
G	2.29	2.06	2.56	2.01	1.17	1
R	2.14	1.23	1.77	1.44	1.41	1
Y	6.73	1.38	2.43	1.47	1.19	1
P	3.79	1.94	1.36	1.89	1.59	1
I	15.50	6.61	1.35	1.26	1.88	2

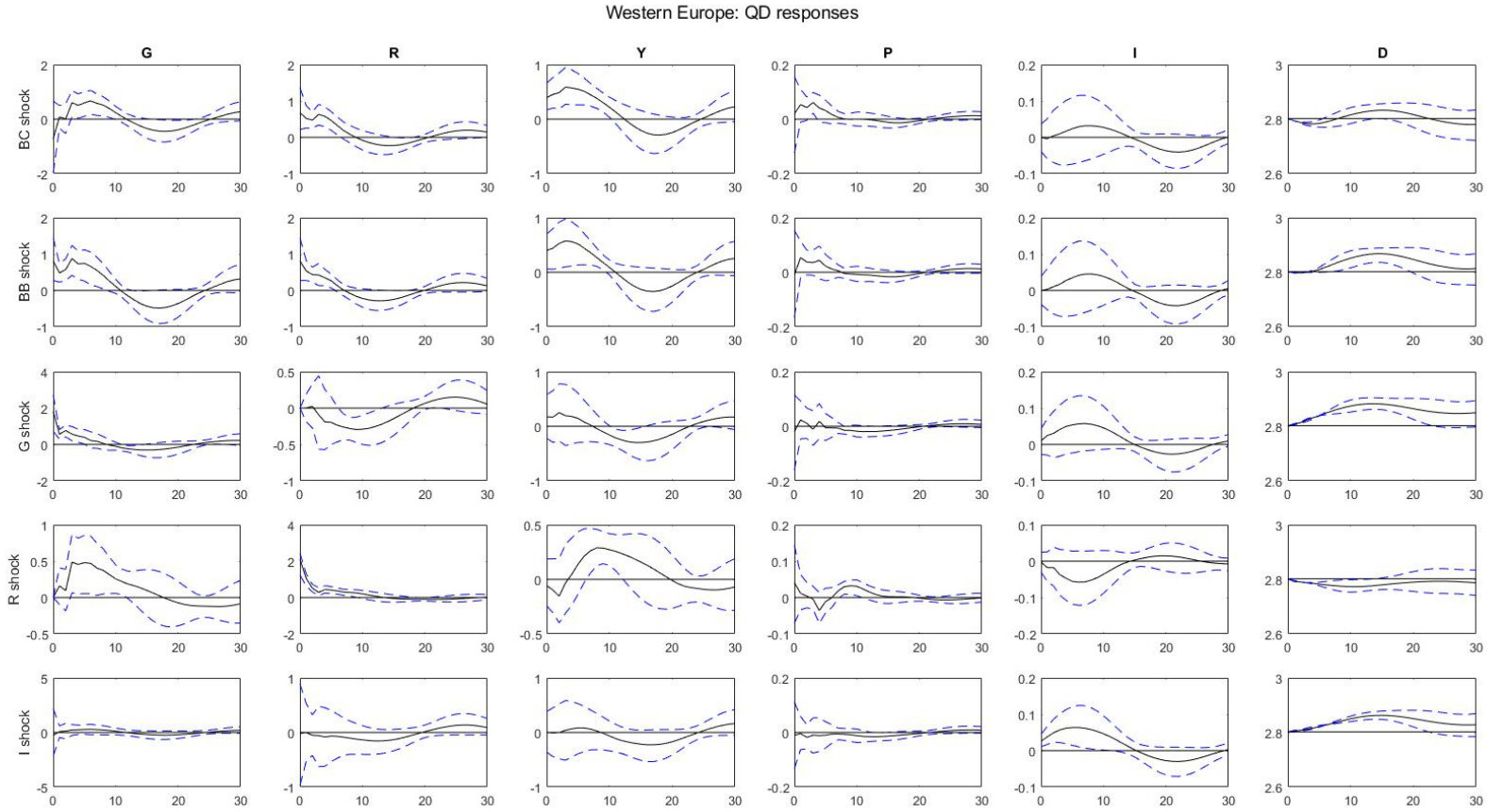
Table 5.7: Eastern Europe

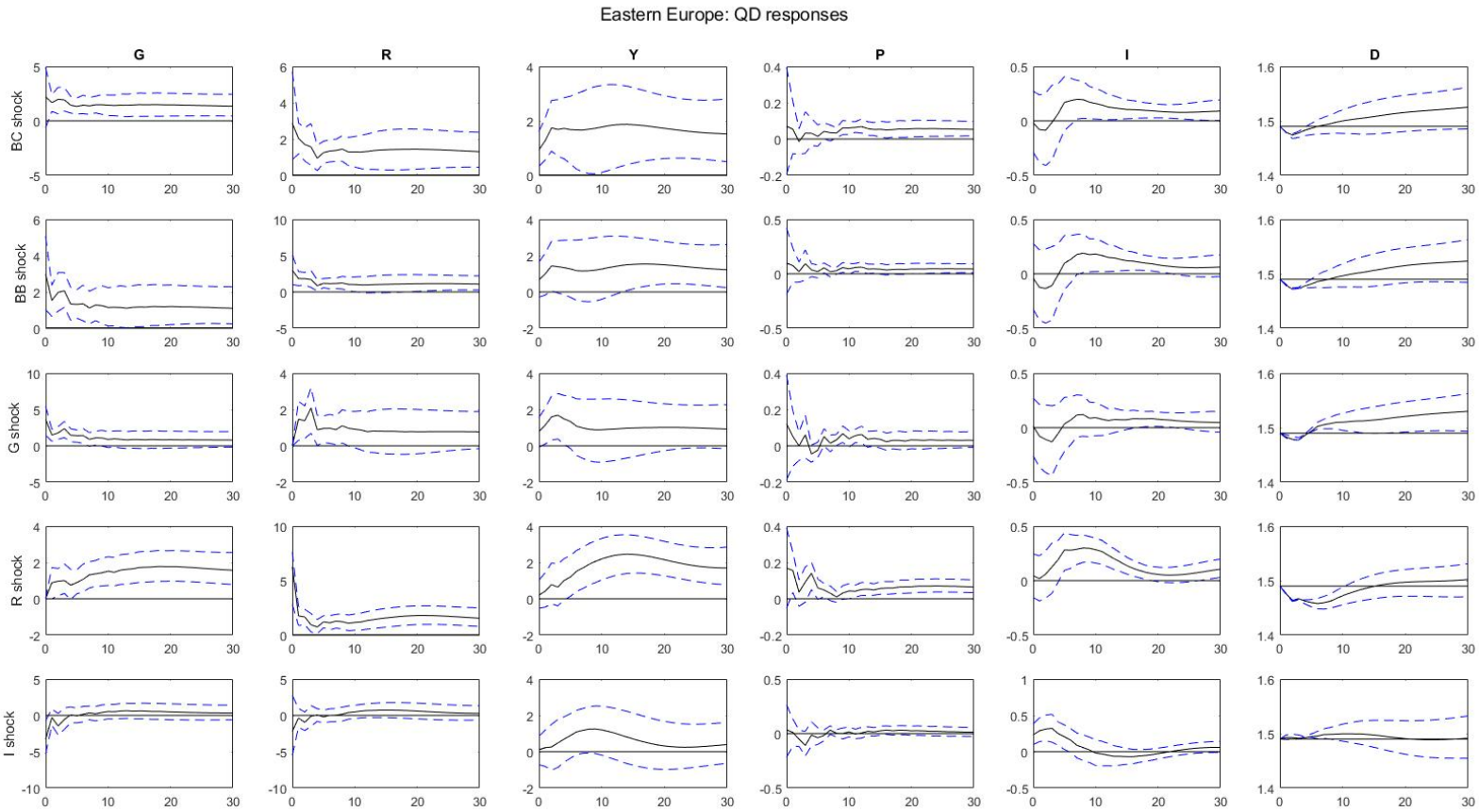
λ_i/λ_{i+1}						
Equation:	1	2	3	4	5	# Factors
G	1.43	1.60	1.26	2.28	1.64	2
R	1.85	1.63	1.46	1.30	1.19	2
Y	3.00	2.17	1.31	1.78	1.16	2
P	4.87	1.54	1.49	1.35	2.00	1
I	4.48	1.76	1.35	1.35	1.24	2

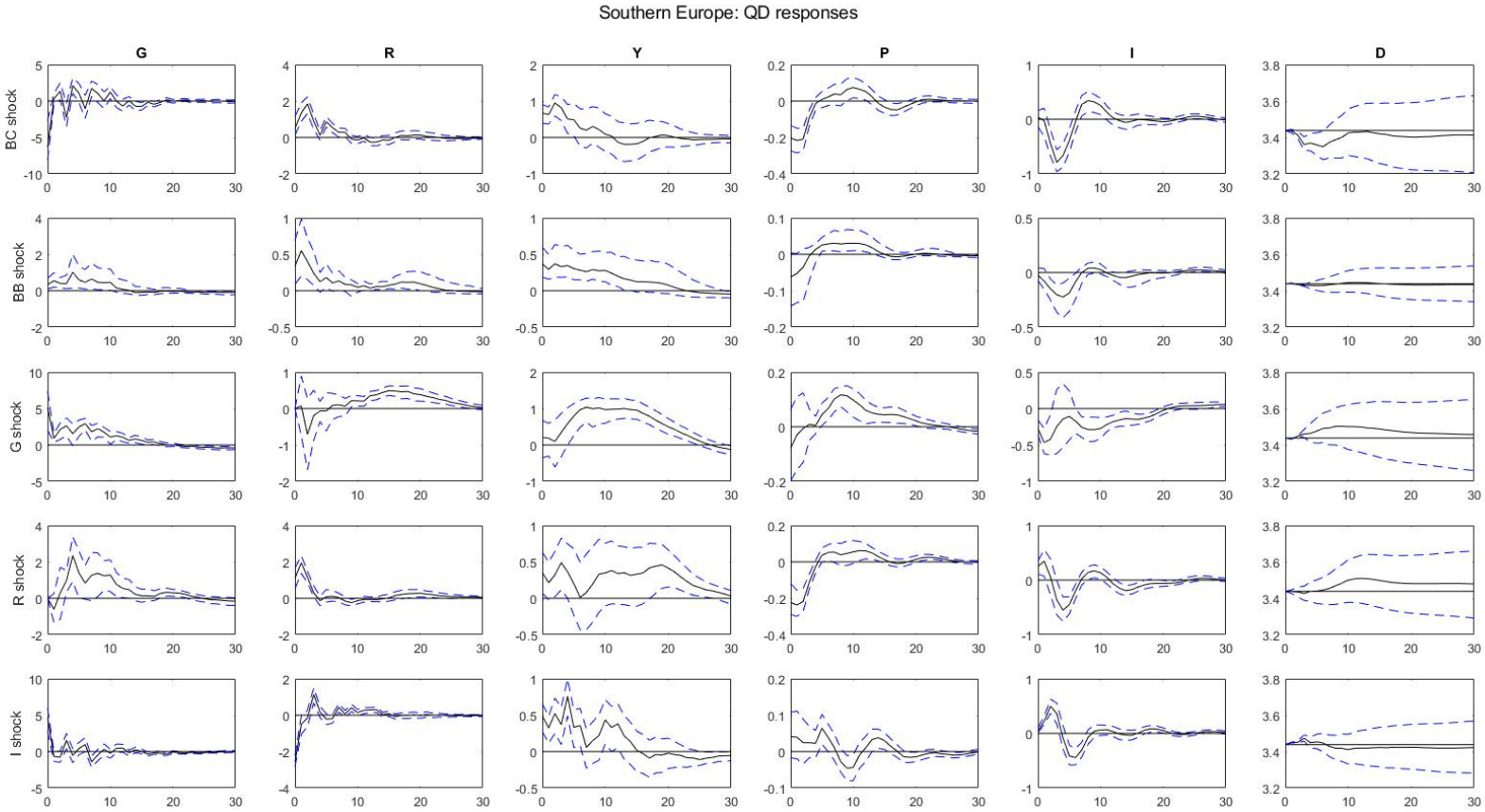
Table 5.8: Southern Europe

λ_i/λ_{i+1}					
Equation:	1	2	3	4	# Factors
G	1.28	1.72	2.35	2.31	2
R	1.07	2.54	3.03	1.01	2
Y	2.46	3.89	2.85	1.18	2
P	3.74	1.91	1.48	1.56	1
I	2.91	4.78	1.65	2.57	2

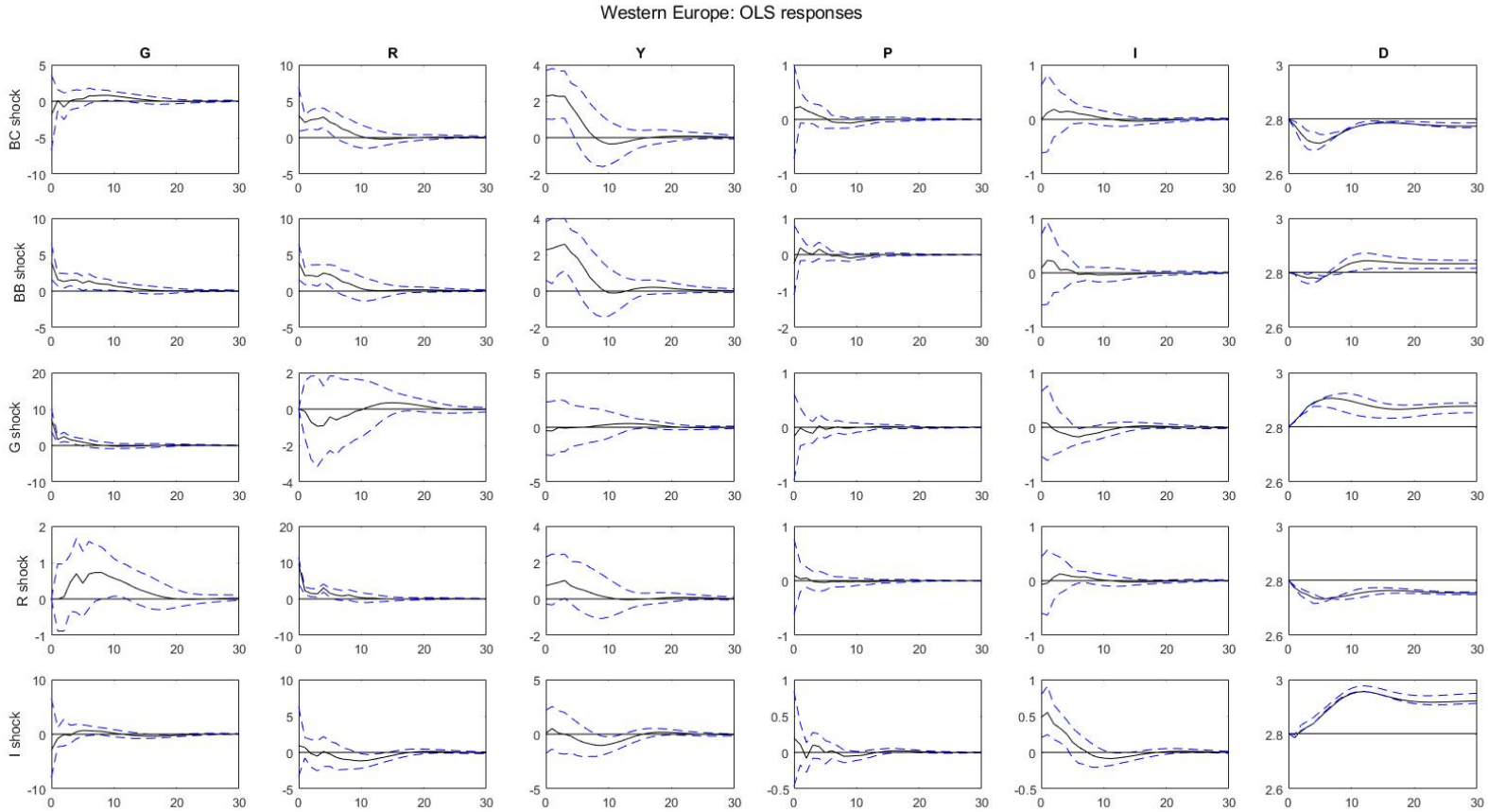
Appendix 5.2 Impulse-Response Functions for Quasi-Difference Vector Autoregressions

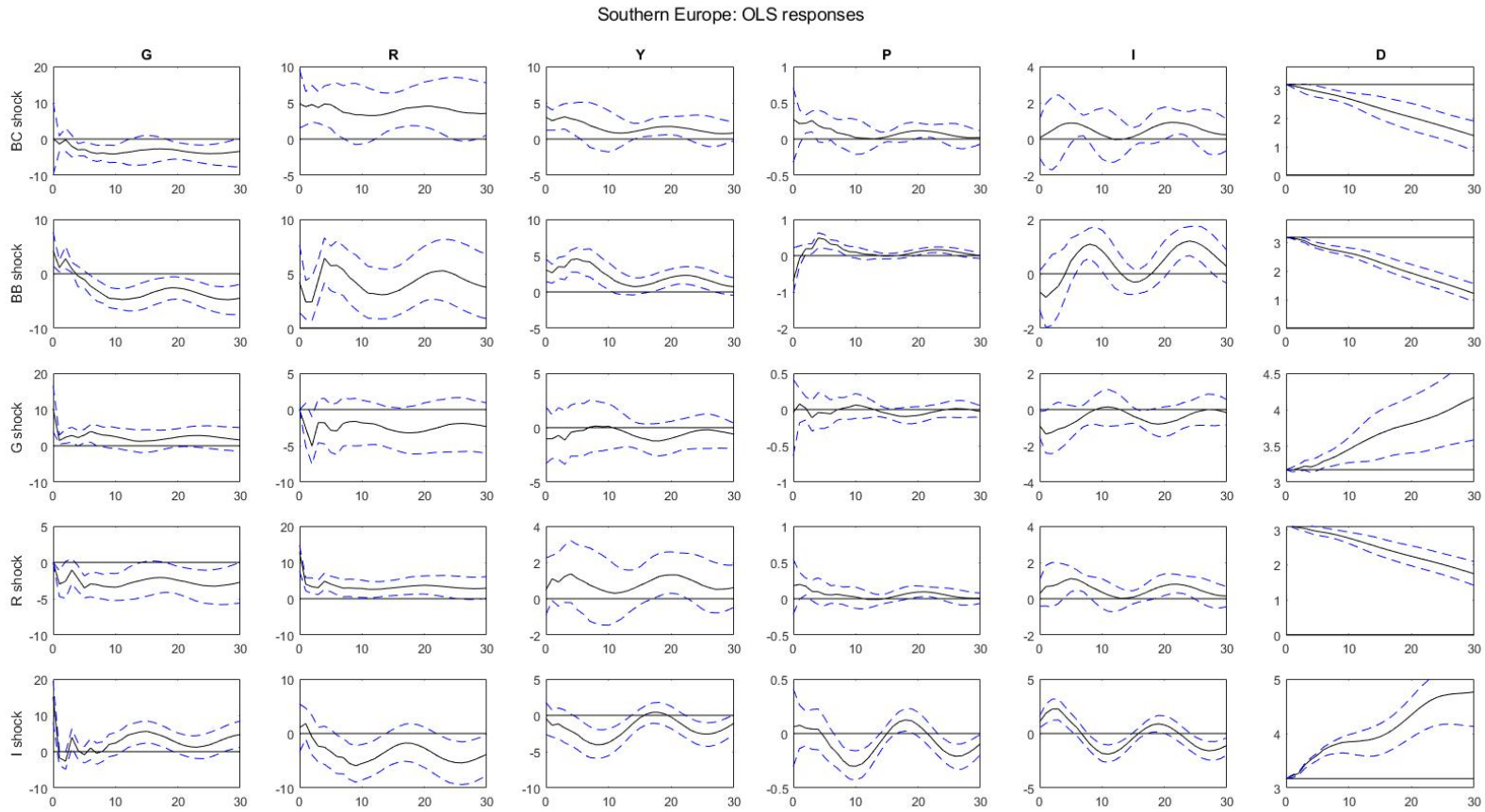


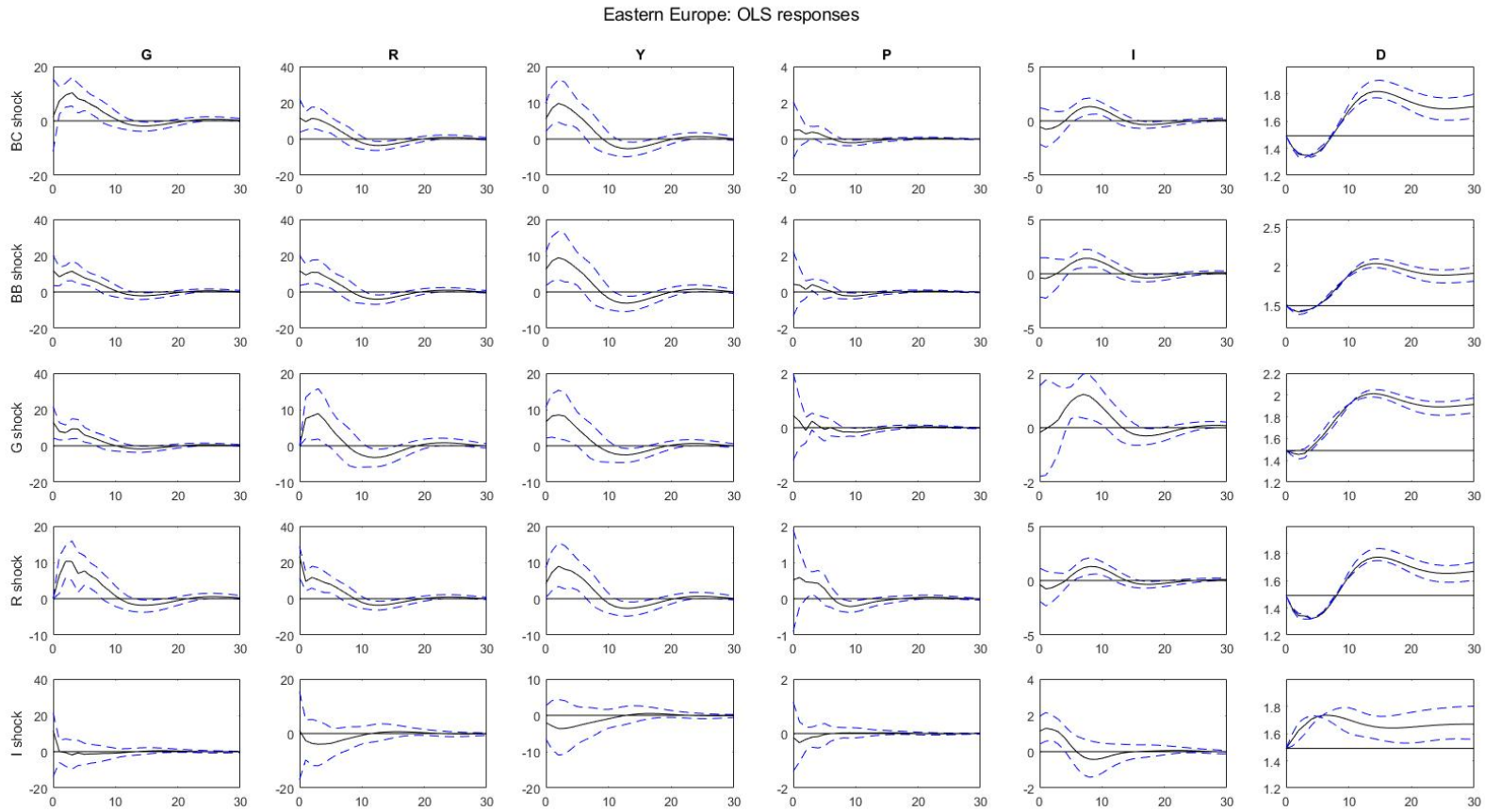

Eastern Europe: QD responses
G
R
Y
P
I
D
BC shock
BB shock
G shock
R shock
I shock


Southern Europe: QD responses
G
R
Y
P
I
D
BC shock
BB shock
G shock
R shock
I shock

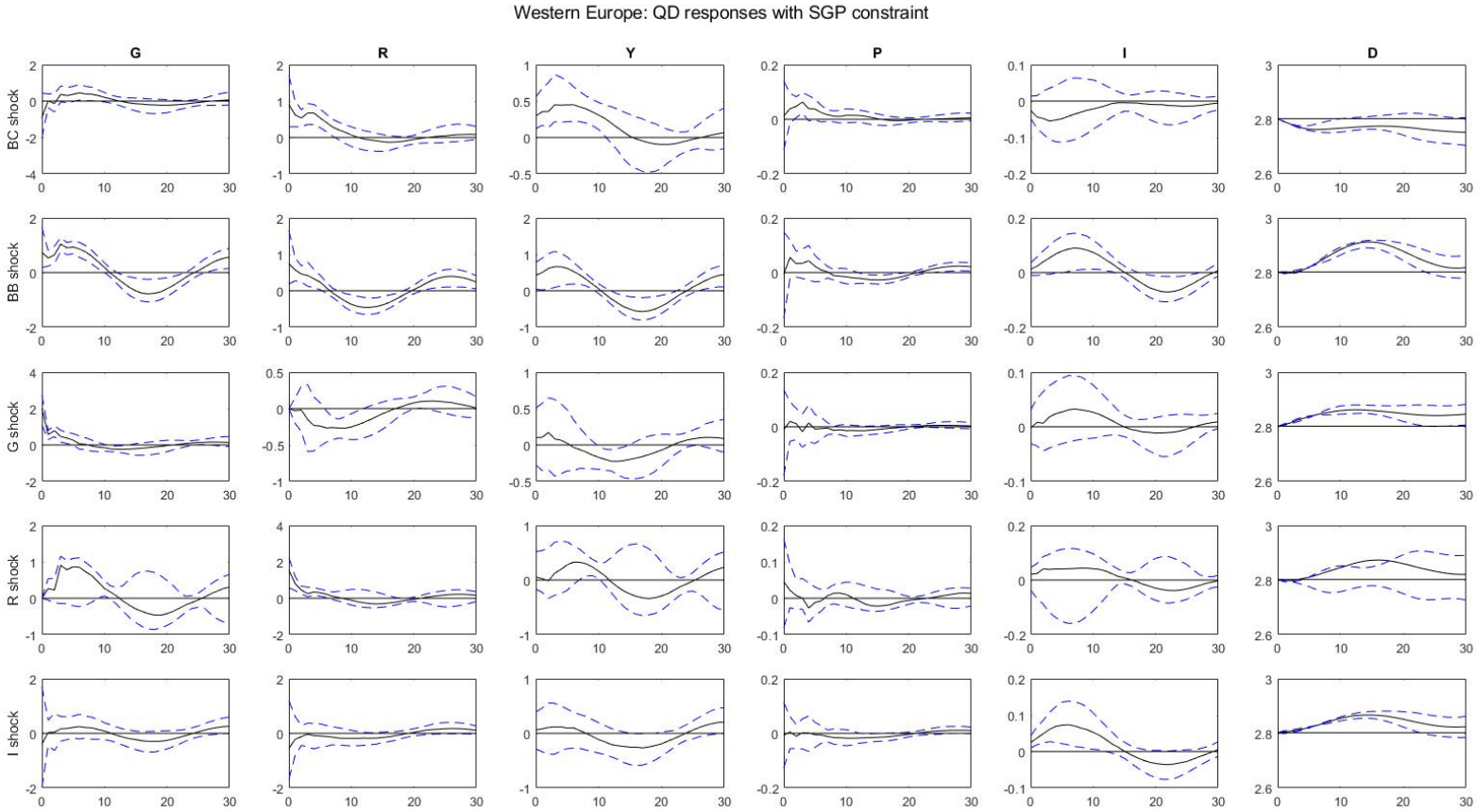
Appendix 5.3 Impulse-Response Functions for OLS Vector Autoregressions

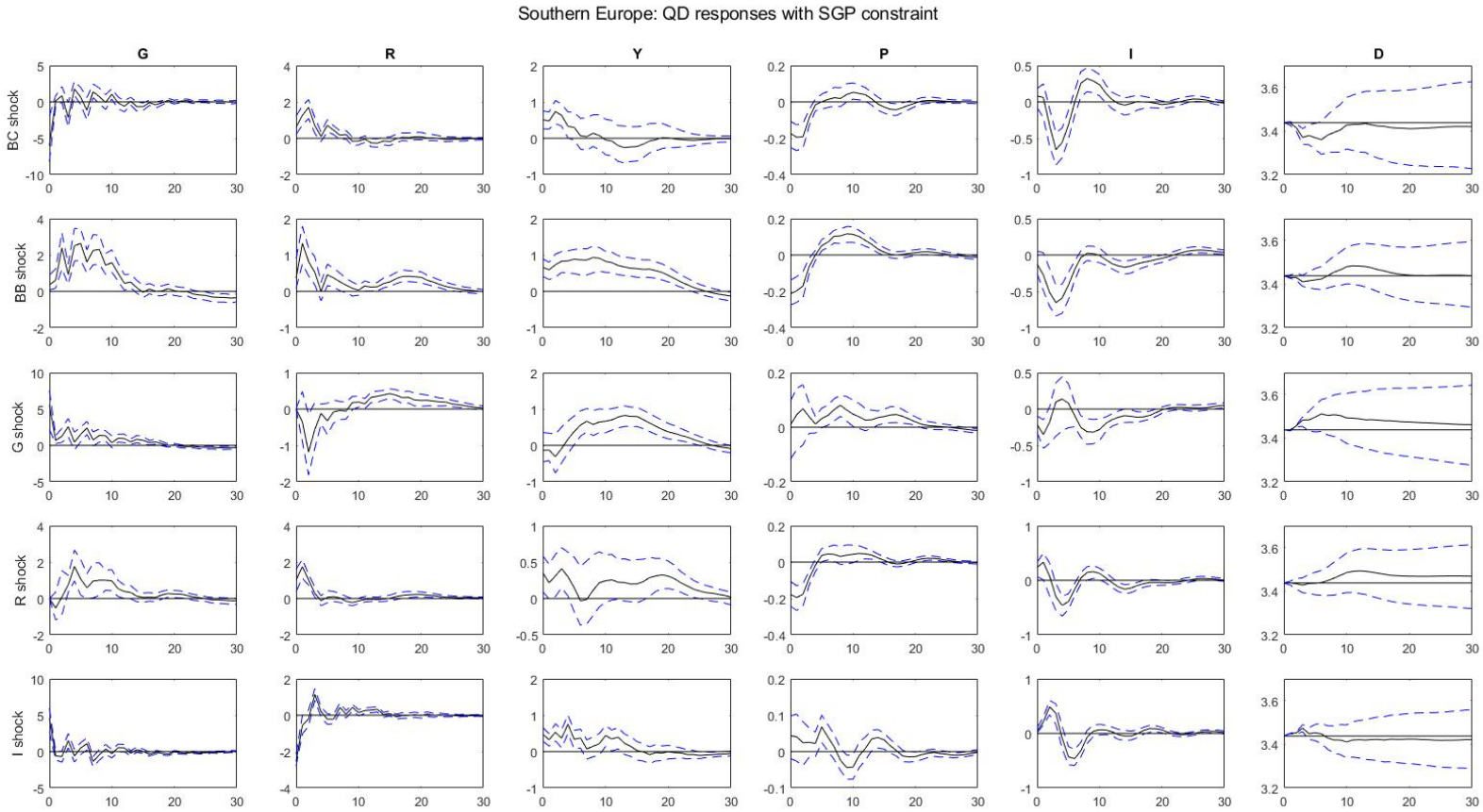


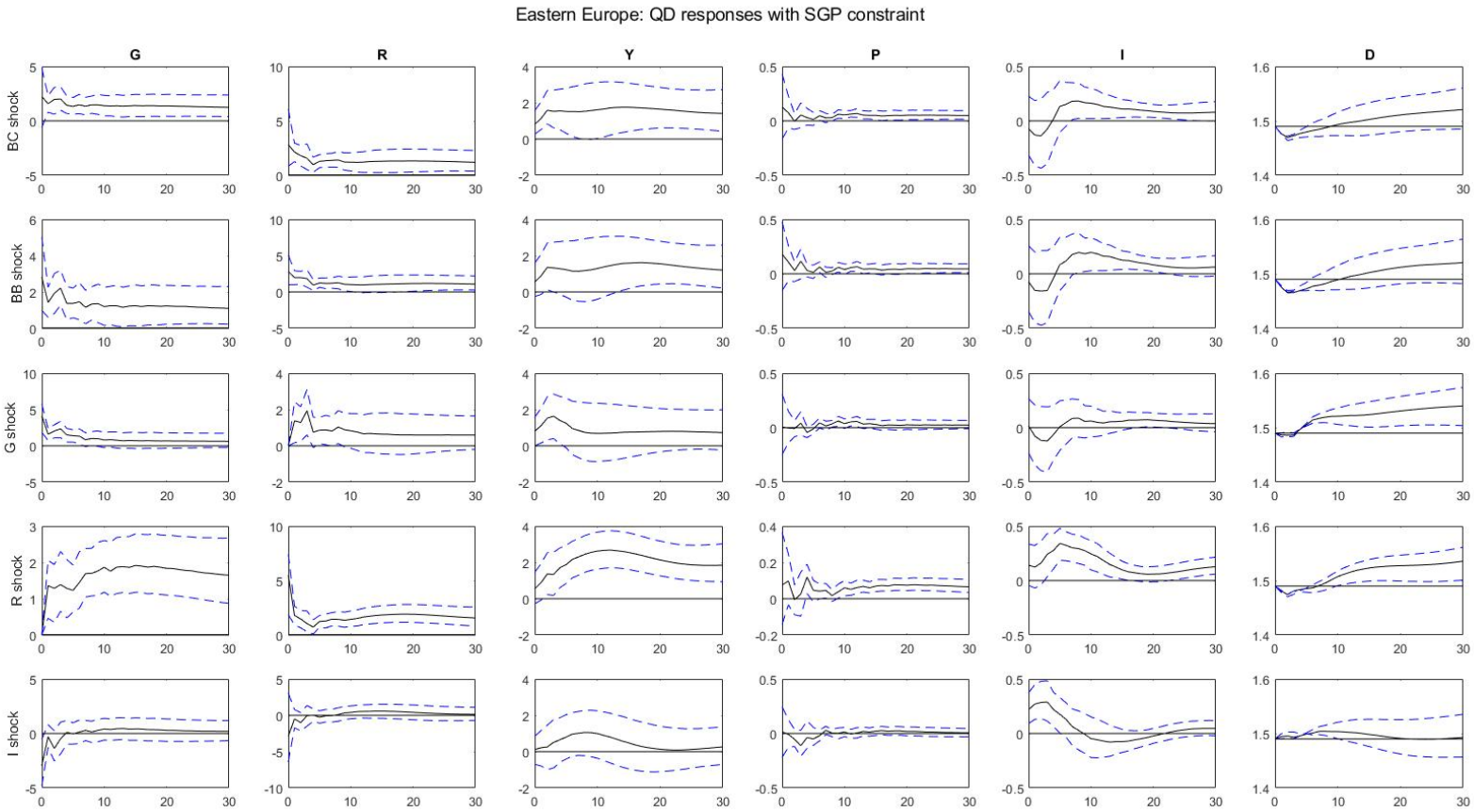

Southern Europe: OLS responses
G
R
Y
P
I
D
BC shock
BB shock
G shock
R shock
I shock


Eastern Europe: OLS responses
G
R
Y
P
I
D
BC shock
BB shock
G shock
R shock
I shock

Appendix 5.4 Impulse-Response Functions for Quasi-Difference Vector Autoregressions with SGP Imposed




Southern Europe: QD responses with SGP constraint
G
R
Y
P
I
D
BC shock
BB shock
G shock
R shock
I shock


Eastern Europe: QD responses with SGP constraint
G
R
Y
P
I
D
BC shock
BB shock
G shock
R shock
I shock

Bibliography

[1] Ahn, S. and Horenstein, A. (2013): "Eigenvalue Ratio Test for the Number of Factors," *Econometrica*, Vol.. 83, No. 3, pp. 1203-1227.

[2] Ahn, C., Lee, Y. H. and Schmidt, P. (2013): "Panel Data Models with Multiple Time-Varying Individual Effects," *Journal of Econometrics,* Vol.. 174, pp. 1-14, Elsevier North Holland.

[3] Andrews, D. (1992): "Generic Uniform Convergence," *Econometric Theory*, Vol.. 8 #2, pp. 241-257, Cambridge University Press.

[4] Andrews, D. (1997): "A Stopping Rule for the Computation of the Generalized Method of Moments Estimators," *Econometrica*, Vol.. 65 #4, pp. 913-931, Royal Econometric Society, Whiley-Blackwell.

[5] Arias, J., Rubio-Ramirez, J. and Waggoner, D. (2018): 'Inference based on Structural Vector Autoregressions with Sign and Zero Restrictions: Theory and Applications,' *Econometrica*, Vol.. 86, Issue 2, pp. 685-720, Wiley-Blackwell.

[6] Auerbach, A. and Gorodnichenko, Y. (2012): "Measuring the Output Responses to Fiscal Policy," *American Economic Journal: Economic Policy*, Vol. 4 No 2, pp. 1-27, American Economic Association.

[7] Bai, J. and Ng, S. (2006): "Confidence Intervals for Diffusion Index Forecasts and Inference for Factor Augmented-Augmented Regressions," *Econometrica*, Vol.. 74 # 4, pp. 1133–1150, Royal Econometric Society, Wiley-Blackwell.

[8] Balassone, F. and Franco, D. (2000): "Public Investment: the Stability Pact and the 'Golden Rule'," *Fiscal Studies*, Vol.. 28 #2, pp. 207-229, Institute for Fiscal Studies.

[9] Banerjee, A., Marcellino, M. and Masten, I. (2008): "Forecasting Macroeconomic Variables using Diffusion Indexes in Short Samples with Structural Change," Forthcoming in: *Forecasting in the Presence of Structural Breaks and Model Uncertainty*, edited by D. Rapach and M. Wohar, Elsevier.

[10] Bayoumi, T. and Eichengreen, B. (1995): "Restraining Yourself: Fiscal Rules and Stabilization," *IMF Staff Papers* Vol.. 42 #1, pp. 32-48, International Monetary Fund.

[11] Beetsma, R. and Guiliodori, M. (2010a): "Fiscal Adjustment to Cyclical Developments in the OECD: an Empirical Analysis based on Real-Time Data," *Oxford Economic Papers* 62, pp. 419-441, Oxford University Press.

[12] Beetsma, R. and Guiliodori, M. (2010b): 'The Macroeconomic Costs and Benefits of the EMU and Other Monetary Unions: An Overview of Recent Research,' *Journal of Economic Literature*, Vol.. 48 Issue 3, pp. 603-641, American Economic Association.

[13] Beetsma, R. and Guiliodori, M. (2011): 'The Effects of Government Purchases Shocks: Review and Estimates for the EU,' The Economic Journal, Vol.. 121, Issue 550, pp. F4-F32, Wiley-Blackwell.

[14] Bénétrix, A. S. and Lane, P. (2013): "Fiscal Cyclicality and EMU," *Journal of International Money and Finance*, Vol. 34, pp. 164-176, Elsevier.

[15] Bernanke, B., Boivin, J. and Eliasz, P.S. (2005): "Measuring the Effects of Monetary Policy: A Factor-Augmented Vector Autoregressive (FAVAR) Approach," *The Quarterly Journal of Economics* Vol.. 120, pp. 387-422, Oxford University Press.

[16] Bertsekas, D. (1999): "Non-Linear Programming," Athena Scientific.

[17] Blanchard, O. and Perotti, R. (2002): "An Empirical Characterization Of The Dynamic Effects Of Changes In Government Spending And Taxes On Output," *The Quarterly Journal of Economics*, Vol.. 117 #4, pp. 1329-1368, MIT Press.

[18] Buiter, W. H. (2006): "The 'Sense and Nonsense of Maastricht' Revisited: What Have we Learnt about Stabilization in EMU?," *Journal of Common Market Studies*, Vol.. 44 #4, pp. 687-710, Elsevier.

[19] Caldara, D. and Kamps, C. (2008): "What are the Effects of Fiscal Policy Shocks? A VAR-based Comparative Analysis," *ECB Working Paper Series*, No. 0877, European Central Bank.

[20] Canova, F. (2007): *"Methods for Applied Macroeconomic Research,"* Princeton University Press, Princeton, New Jersey.

[21] Canova F. and Pappa, E. (2007): "Price Differentials in Monetary Unions: The Role of Fiscal Shocks," *Economic Journal*, Vol.. 117 #520, pp. 713-737, Royal Economic Society.

[22] Chung, H and Leeper, E. M. (2007): "What has Financed Government Debt?" *NBER Working Papers* 13425, National Bureau of Economic Research, Inc.

[23] Eichengreen, B. and Wyplosz, C. (1998): "Stability Pact: More than a Nuisance?" *Economic Policy*, Vol.. 13 #26, pp. 65-113, Oxford University Press.

[24] Fatás, A. and Mihov, I. (2001): "The Effects of Fiscal Policy on Consumption and Employment: Theory and Evidence," *CEPR Discussion Papers* 2760, C.E.P.R. Discussion Papers.

[25] Fatás, A. and Mihov, I. (2010): "The Euro and Fiscal Policy," in: Alesina, A. and Giavazzi, F. (editors), "*Europe and the Euro*," University of Chicago Press, National Bureau of Economic Research, Inc.

[26] Favero, C. and Giavazzi, F. (2007): "Debt and the Effects of Fiscal Policy," *Working Papers* 317, IGIER (Innocenzo Gasparini Institute for Economic Research), Bocconi University.

[27] Fernandez-Villaverde, J., Guerron-Quintana, P., Kuester, K. and Rubio-Ramirez, J. (2015): *"Fiscal Volatility Shocks and Economic Activity," American Economic Review*, Vol.. 105 #11, pp. 3352-3384, American Economics Association.

[28] Galí, J. Perotti, R. (2003): "Fiscal Policy and Monetary Integration in Europe," *Economic Policy*, 2003, Vol..18 #37, pp. 533-572, Oxford University Press.

[29] Grippo, L. and Sciandrone, M. (2000): "On the Convergence of the Block Non-linear Gauss-Seidel Method under Convex Constraints," *Operations Research Letters* Vol.. 26, pp. 127-136, Elsevier North Holland.

[30] Giuliodori, M. and Beetsma, R. (2005): "What are the Trade Spill-Overs from Fiscal Shocks in Europe? An Empirical Analysis," *De Economist*, Vol.. 153 Issue 2, pp. 167-197, Springer.

[31] Ilzetzki, E. (2011): "Fiscal Policy and Debt dynamics in Developing Countries," *Policy Research Working Paper Series* 5666, The World Bank.

[32] Ilzetzki, E., Mendoza, E. G. and Végh, C. A. (2013): "How Big (Small?) are Fiscal Multipliers?" *Journal of Monetary Economics*, Vol.. 60 #2, pp. 239-254, Elsevier North Holland

[33] Mountford A. and Uhlig, H. (2009): "What are the Effects of Fiscal Policy Shocks?" *Journal of Applied Econometrics*, Vol.. 24 #6, pp. 960-992, John Wiley & Sons, Ltd.

[34] Pappa, E. (2009a): "The Effects Of Fiscal Shocks On Employment And The Real Wage," *International Economic Review*, Vol.. 50 #1, pp. 217-244, Department of Economics, University of Pennsylvania and Osaka University Institute of Social and Economic Research Association.

[35] Pappa, E. (2009b): "The Effects of Fiscal Expansions: an International Comparison," *Working Papers* 409, Barcelona Graduate School of Economics.

[36] Peersman, G. (2011): "Macroeconomic Effects of Unconventional Monetary policy in the Euro Area," *ECB Working Paper,* No. 1397.

[37] Perotti, R. (2004): "Estimating the Effects of Fiscal Policy in OECD Countries," *Working Papers 276*, IGIER (Innocenzo Gasparini Institute for Economic Research), Bocconi University

[38] Ramey, V. A. and Shapiro, M. D. (1998): "Costly Capital Reallocation and the Effects of Government Spending," *Carnegie-Rochester Conference Series on Public Policy*, Vol.. 48 # 1, pp. 145-194, Elsevier North Holland.

[39] Ravn, M., Schmitt-Grohé, S. and Uribe, M., (2007): "Explaining the Effects of Government Spending Shocks on Consumption and the Real Exchange Rate," *NBER working paper* 13328.

[40] Romer, C. and Romer, D. (2010): 'The Macroeconomic Effects of Tax Changes: Estimates Based on a New Measure of Fiscal Shocks,' *American Economic Review*, Vol.. 100, pp 763-801, American Economic Association.

[41] Rubio-Ramirez, J., Waggoner, D. and Zha, T. (2010): "Structural Vector Autoregressions: Theory of Identification and Algorithms for Inference," *Review of Economic Studies*, Vol.. 77 #2, pp. 665-696, Oxford Journals.

[42] Schenkelberg, H. and Watzka, S. (2013): "Real Effects of Quantitative Easing at the Zero Lower Bound: Structural VAR-based Evidence from Japan," *Journal of International Money and Banking*, Vol.. 33, pp. 327-357, Elsevier North Holland.

[43] Stock, J. H. and Watson, M. W. (2002): "Forecasting Using Principal Components From a Large Number of Predictors," *Journal of the American Statistical Association*, Vol.. 97, pp. 1167-1179, American Statistical Association.

[44] World Bank (2018): Updated Income Classifications, accessed via http://data.worldbank.org/news/2017-country-classifications