UNIVERSITY OF CALIFORNIA
RIVERSIDE

Essays on Panel Data and System of Equations under Model Uncertainty

A Dissertation submitted in partial satisfaction
of the requirements for the degree of

Doctor of Philosophy

in

Economics

by

Ali Mehrabani

June 2021

Dissertation Committee:

Professor Aman Ullah, Chairperson
Professor Tae-Hwy Lee
Professor Gloria Gonzalez-Rivera

The Dissertation of Ali Mehrabani is approved:

Committee Chairperson

University of California, Riverside

Acknowledgments

I would like to express my most sincere gratitude to all those who provided me the support and encouragements during my Ph.D. program. I am grateful to my dissertation supervisor professor Aman Ullah, and my dissertation committee members professor Gloria Gonzalez-Rivera, and professor Tae-Hwy Lee for invaluable guidance and continuous support. I would like to give my special thanks to professor Marcelle Chauvet for her support during my Ph.D. program, and to all my instructors, professors, and supervisors who throughout my educational career have supported and encouraged me to do my best.

I will forever be thankful to my family. My special thanks to my parents and my wife for their support, encouragement, companionship, and love. I consider myself nothing without them.

To my parents and my wife.

ABSTRACT OF THE DISSERTATION

Essays on Panel Data and System of Equations under Model Uncertainty

by

Ali Mehrabani

Doctor of Philosophy, Graduate Program in Economics
University of California, Riverside, June 2021
Professor Aman Ullah, Chairperson

This dissertation consists of four chapters that study estimation and inference in system of equations and panel data under model uncertainty. In Chapter 2, I consider model uncertainty in a panel data model, and introduce a Stein-like shrinkage estimator that is a weighted average of an unrestricted estimator and a restricted estimator. The restricted estimator represents a belief about where the parameters of the model are likely to be close.

Chapter 3 considers the estimation uncertainty from choosing different number of lagged dependent variables as instruments in dynamic panel data models. Generalized method of moments (GMM), the typical estimation method, can produce efficient estimators when all lagged dependent variables are used as instruments. However, estimation using all instruments can cause substantial bias. Conversely, the GMM estimators that use one lag as instrument are asymptotically unbiased under forward demeaning transformation, but at the cost of losing efficiency. Therefore, I introduce an averaging estimator which is a weighted average of the two GMM estimators where the averaging weight is proportional to a quadratic loss function that minimizes the asymptotic risk.

In Chapter 4, I consider simultaneous equations models, and develop an estimator to deal with the model uncertainty about the magnitude of endogeneity. Ordinary least squares (OLS) estimators are the most efficient estimators, however, may suffer from substantial bias when the degree of endogeneity is substantial. On the contrary, two-stage least squares (2SLS) and Limited Information Maximum Likelihood (LIML) estimators are consistent but not as efficient. Therefore, I consider a Stein-like shrinkage estimator which is a weighted average of the OLS and 2SLS/LIML estimators, where the weight is inversely related to a Wu-Hausman statistic that measures the magnitude of the endogeneity.

Chapter 5 considers latent group structures to model uncertainty resulting from unobserved heterogeneity in panel data models. Basically, I consider a panel data model where the slope parameters are heterogenous across groups but homogenous within a group, and the group identity is unknown. I provide a framework for estimation and identification of the latent group structure using a pairwise fusion penalized approach.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

In the theory and practice of econometrics, there are typically a large amount of uncertainty about model specifications that are unobservable to the practitioners. In practice, estimation and inference are often conducted under a selected model specification without considering the model uncertainty. This can lead to many difficulties, including inconsistent estimation and misleading inference. For example, panel data usually cover observations from workers, firms, or countries that differ in many dimensions, so an undeniable feature of the data is its heterogeneity, but much of which is simply unobserved. Therefore, empirical researchers face a trade-off between using approaches that allow for model uncertainty (e.g. unobserved heterogeneity), and building restricted specifications that are adapted to the empirical data at hand. A way to deal with this issue is to build flexible yet parsimonious approaches that allow for the uncertainty (for example latent group structures considered in chapter 5). An alternative way is to use model averaging techniques where a weighted average of the candidate models is considered. In this dissertation, I

investigate and develop methods to deal with different types of uncertainty with the aim of providing robust and superior estimators.

For example, in chapter 2, I present a Stein-like shrinkage method for estimating the slope coefficients in heterogeneous panel data models with cross-section dependence, when the cross-section dimension is fixed while the time dimension is allowed to increase without bounds. The shrinkage estimator is a weighted average of a feasible generalized least-squares (FGLS) estimator and a feasible restricted generalized least-squares estimator. The restricted estimator belongs to a set of restricted parameter space, where the restrictions represent possible model specifications. The shrinkage weight is inversely proportional to a Wald statistic that measures the importance of the restrictions. The asymptotic properties of the shrinkage estimator are given. Further, it is shown that the shrinkage estimator is robust, and uniformly superior, in terms of asymptotic risks, relative to the FGLS estimator. Additionally, the shrinkage estimator achieves the lowest possible asymptotic risk in a high-dimensional large sample framework. A major advantage of this shrinkage method is that it is generalized to allow for the limitations of the existing model averaging techniques. For instance, the shrinkage method developed here is generalized to allow for any patterns of correlations in errors, is not confined to specific restricted estimators, its superiority conditions hold for any weighted mean squared error where the weight matrix is symmetric positive definite, and achieves the lowest possible risk bound. The finite sample performance of the proposed estimation method is evaluated via extensive simulation studies, that support the theoretical findings. As an empirical illustration, the method is applied to forecast the output growth rate of 33 advanced and emerging economies in the

global economy using a set of macroeconomic and financial variables by allowing potential parameter heterogeneity structures in the slope coefficients. This methodology has two major advantages over the existing studies: it considers the classification uncertainty about the potential heterogeneity, and allows for general correlation patterns across the errors in the cross-section equations of output growth. The results indicate that the shrinkage estimation forecast outperform the fixed effects and individual estimation forecasts.

In chapter 3, I consider dynamic panel data models with fixed effects and multiple exogenous regressors. The typical estimator in this framework is the Arellano-Bond generalized method of moments (GMM). One can gain in efficiency of the GMM by estimating the parameter values using all lagged dependent variables as instruments. However, estimation based on instrumenting all lagged dependent variables may suffer from substantial bias. On the other hand, the GMM estimators that use one lag (or fixed number of lags) as instruments are asymptotically unbiased under forward demeaning transformation, but not as efficient as the former one. In this chapter, I introduce an averaging estimator which is a weighted average of the GMM estimator using all lags as instruments, and the GMM estimator using one lagged dependent variable as instruments to balance this trade-off between the bias and variance efficiency. The averaging weight is proportional to a quadratic loss function that minimizes the asymptotic risk. In addition, the optimality, and the dominance conditions of the averaging estimator are derived. Furthermore, monte carlo simulations are provided to examine the finite sample performance of the proposed estimator.

In chapter 4, I consider a simultaneous equations system, and develop an estimator to deal with a typical model uncertainty that arises due to unknown magnitude of endogeneity. When the magnitude of endogeneity is weak, one can largely gain in efficiency by estimating the parameter values using the ordinary least squares (OLS) estimator. However, the OLS estimator ignores the potential endogeneity and may suffer from substantial bias. Alternatively the two-stage least squares (2SLS) or Limited Information Maximum Likelihood (LIML) frameworks can be used, which control the endogeneity and hence ar consistent, but the consistency comes at the cost of losing efficiency. This shows the typical bias-variance trade-off that needs to be considered carefully by practitioners in choosing models. To balance the trade-off, I consider two Stein-like shrinkage estimators which are weighted averages of the OLS and 2SLS/LIML estimators, and the weight is inversely related to a Wu-Hausman statistic that measures the magnitude of the endogeneity. I derive the dominance conditions of the proposed estimators relative to the 2SLS/LIML estimators for any size of endogeneity, and any weighted mean squared errors. I further investigate the finite sample performance of the estimation method through Monte Carlo simulations. The results show that the proposed estimators perform well, and support the theoretical findings.

Chapter 5 considers a long-existing model uncertainty issue in panel data analysis, referred by econometricians to as "to pool or not to pool", on which there is still no consensus. The issue is on how to model potential parameter heterogeneity across individual units. To deal with this issue, I model individual heterogeneity via latent group structures such that the slope parameters are heterogenous across groups but homogenous within

a group, and the group identity is unknown. In particular, this model setup faces the uncertainty resulting from the unobserved heterogeneity by allowing flexible forms of heterogeneity while remaining parsimonious. In this chapter, I provide a framework for estimation and identification of the latent group structure using a pairwise fusion penalized approach. I develop a penalized least squares (PLS) approach for models with exogenous regressors, and a penalized generalized method of moments (PGMM) for endogenous or dynamic models. This framework automatically partitions the individuals into groups. Therefore, it asymptotically identifies the true structure while estimating the model parameters consistently. Both of the estimators achieve the desirable property of classification consistency. Further, the PLS estimator achieves the oracle property, while the oracle property of the PGMM estimator holds under some restrictions. I have developed an alternating direction method of multipliers algorithm to implement the proposed approach. The method is further evaluated by monte carlo simulations, and illustrated by two empirical analysis of unemployment dynamics at the U.S. state level, and forecasting output growth of 33 countries using macroeconomic and financial variables.

Chapter 6 concludes and some technical results are provided in the appendix.

Chapter 2

Efficient Shrinkage Estimation in Heterogeneous Panel Data Models

2.1 Introduction

Estimation and forecasting under model uncertainty has been one of the fundamental issues in econometrics. In recent years, a large body of literature has been concerned with advancing a number of different approaches to address a variety of model uncertainty problems. The two most common approaches are model selection and model averaging. Model selection aims to find, among the set of models under consideration, the best approximate model for the unknown true data generating process. In this method, investigators typically first select the best performing model based on diagnostic statistics (like Wald test, F test, t-ratios, R-squared, information criteria, etc.) and then carry out inference according to the selected model. This popular approach (also known as

"pre-testing") is subject to many problems (Magnus (1999), Magnus and Durbin (1999), Danilov and Magnus (2004a), Danilov and Magnus (2004b)). The most important problem is that the model selection and estimation are completely separated such that the uncertainty of the initial model selection step is ignored throughout the parameter estimation and inference, see for example Magnus (2002) and Leeb and Pötscher (2003), Leeb and Pötscher (2006), among others, who show the initial model selection step may have non-negligible effects on the statistical properties of the resulting estimators. Taking the above problems into consideration, model averaging is introduced as an alternative to the model selection. In model averaging, the uncertainty is taken into consideration by averaging (weighted) over the set of candidate models. Model averaging methods are distinct in two main strands based on whether the estimation of each candidate model and the choice of the associated weighting scheme are developed along frequentist or Bayesian paradigms. Shrinkage estimation methods, similar to model averaging, allow for uncertainty emerging from both model selection and estimation (see Hansen (2014), Hansen (2016)). In addition, as shown by Hansen (2016), Stein-type shrinkage estimation methods, unlike recent model averaging techniques (such as focused information criterion of Claeskens and Hjort (2003), the plug-in estimator of Liu (2015), and the focused moment selection criterion of DiTraglia (2016)), have the minimax efficiency properties.

This chapter investigates a Stein-like shrinkage estimation method in linear heterogeneous panel data models to deal with uncertainty issues about the slope parameters. We allow for cross-section dependence and to estimate the contemporaneous error covariances freely, it is assumed that the cross-section dimension is small and the time series

dimension is large. The shrinkage estimator shrinks a feasible generalized least-squares (FGLS) estimator (the standard approach in this setup, see Zellner (1962)) towards a shrinkage direction, or equivalently a set of parameter restrictions. The restrictions are not necessarily believed to be true, but instead represent a belief about where the parameters of the model are likely to be close. Therefore, the proposed estimator is a weighted average of the FGLS estimator and a feasible restricted generalized least-squares estimator that belongs to the restricted parameter space. The shrinkage weight is inversely related to a Wald statistic that measures the weighted distance of the FGLS estimator and the restricted estimator. The asymptotic properties of our proposed estimator are derived under some mild conditions. Furthermore, we show the dominance properties of the Stein-like shrinkage estimator in terms of risk, which ensures that our proposed estimator is robust against arbitrary deviations from the restrictions. A major advantage of the shrinkage method introduced in this paper is that, unlike most of the existing model averaging methods, it allows for heteroskedasticity, and cross-section dependence of errors. These cross-sectional correlations could be due to omitted common effects, spatial effects, or could arise as a result of interactions within socioeconomic networks. In addition, the presence of some forms of cross-sectional correlation of errors in panel data applications in economics is likely to be the rule rather than the exception. Ignoring the cross-sectional correlations can have serious consequences such that conventional panel estimators can result in misleading inference and even inconsistent estimators, depending on the extent of the cross-sectional dependence, and whether the sources generating the cross-sectional dependence (such as an unobserved common shock) is correlated with regressors (Phillips and Sul (2003),

Phillips and Sul (2007), Andrews (2005), Sarafidis and Robertson (2009), and see a survey by Chudick and Pesaran (2015)).

In Monte Carlo simulations, we compare the small sample performance of our shrinkage estimator with the FGLS estimator and a restricted estimator where the restrictions impose slope parameter homogeneity across cross-sections. The results show that the shrinkage estimator generally produces a smaller risk than the restricted estimator, and the FGLS estimator. As an empirical illustration, we apply our estimator to forecast the output growth rate of 33 advanced and emerging economies in the global economy using a set of macroeconomic and financial variables by allowing potential parameter homogeneity structures in the slope coefficients.

The literature on shrinkage estimation is substantial, which mainly was initiated by a seminal paper by Stein (1956). In that paper, Stein showed that the maximum likelihood estimator (MLE) for the mean of a multivariate normal distribution is inadmissible. This means that it is possible to construct an estimator with a smaller risk than the MLE for the entire parameter space. James and Stein (1961) exhibited an estimator whose risk is uniformly smaller than that of the MLE. Paradoxically, the James-Stein estimator is itself inadmissible and can be dominated by another inadmissible estimate like its positive part (Baranchick (1964)). Judge and Bock (1978) developed this method for most of econometric estimators. Maddala et al. (2001) and recently Hansen (2016) use shrinkage estimation methods to deal with model uncertainty between two candidate models. The shrinkage estimation method in this chapter is similar to that of Hansen (2016) and Maddala et al. (2001). The main difference is that the shrinkage weight in Hansen (2016)

is inversely related to a weighted quadratic loss function, hence is subject to rotations of the coefficient vector, unless investigators are interested in minimizing a mean squared prediction error. However, the one considered in this paper is proportional to a Wald statistic which is an excellent choice as it is invariant to these rotations. Also, Hansen (2016) considers a homoscedastic likelihood framework, but this paper considers linear panel data models and allows for both heteroscedasticity and cross-section dependence in errors. The difference between the method used here and the one in Maddala et al. (2001) is that they use small-disturbance approximations to study the performance of their estimator, which cannot be applied to a model with unknown error cross-section dependence and variance-heteroscedasticity considered in this chapter.

Penalized methods are alternatives to shrinkage estimations for dealing with the uncertainty of covariate selection in regression models, which is arguably the most pervasive situation in economics. Methods that simultaneously select variables and shrink coefficients by minimizing some penalized loss functions include, among others, the least absolute shrinkage and selection operator (LASSO) of Tibshirani (1996), the smoothly clipped absolute deviation (SCAD) penalty of Fan and Li (2001), and the minimax concave penalty (MCP) of Zhang (2010). LASSO-type methods have been shown to be particularly effective in high-dimensional settings with a true small-dimensional structure, or when the number of predictors exceeds the sample size (see, e.g., Fan and Lv (2010); Chernozhukov et al. (2015); Belloni et al. (2017)). However, shrinkage methods do not exploit sparsity, and can work well even when there are many (but less than the sample size) non-zero parameters.

This chapter is also related to a long-existing issue in the panel data analysis, referred by econometricians to as "to pool or not to pool", on which there is still no consensus. The issue is on how to model potential parameter heterogeneity across individual units. On one hand, parameter heterogeneity results in consistent estimation and violation of this assumption causes misleading estimates, see, for example, Robertson and Symons (1992), Pesaran and Smith (1995), Su and Chen (2013), Durlauf et al. (2001), and Browning and Carro (2007). On the other hand, parameter homogeneity causes higher variance efficiency, but at the cost of estimation bias and inconsistency of the associated estimators, which is supported by an increasing number of studies due to a better forecast performance of these estimators, see, for example, Maddala (1991), Maddala and Hu (1996), Baltagi and Griffin (1984), Baltagi et al. (2000), and Hoogstrate et al. (2000). This shows the typical bias-variance trade-off that needs to be considered in choosing parameter specifications. In the literature, there are several ways to address this parameter heterogeneity such as the random coefficient model of Swamy (1970), the pooled mean group estimator of Pesaran et al. (1999), and various group estimators, see for example Lin and Ng (2012), Sarafidis and Weber (2015), Bonhomme and Manresa (2015), Su et al. (2016), among others. These estimators are reasonable choices when investigators are interested in the average effect or know the true specification of the heterogeneity structure or the number of groups. However, researchers are often more interested in the individual parameters, and in most cases the true specification is unknown. As a result, a more useful approach could be model averaging and shrinkage estimation methods. Maddala et al. (2001) show the superior

properties of shrinkage estimators among single-equation estimators and various averaging estimators in a heterogeneous panel data model under error homoscedasticity framework. Wang et al. (2019) propose a Mallow pooling averaging estimator for heterogeneous panel data models and conclude that the pooling estimator is preferred when the panel is heterogeneous and the signal-to-noise ratio is moderate or large. The Mallow model averaging estimator, however, is not asymptotically optimal in our framework since the condition (C.3) of Wang et al. (2019) does not hold here. The condition requires that there is no model for which the bias is zero, which does not hold in our framework since the FGLS estimator is unbiased.

This chapter is mainly concerned with point estimation and does not address the challenging issue of inference with shrinkage estimators. As a preliminary step in this direction, we study the mean squared errors of various estimators. However, since the distribution of shrinkage estimators are non-Gaussian, it is still unclear how to use this knowledge to construct confidence intervals. We leave the full treatment of this nontrivial, interesting and important issue to a follow-up paper.

The paper is organized as follows. Section 2.2 describes the model and the estimators. In section 2.3, we study properties of the estimators. In section 2.4, the asymptotic bias, asymptotic MSE matrix and asymptotic risk of the shrinkage estimator are presented. Monte Carlo results are given in section 2.5. Results from our empirical example are given in section 2.6. Conclusions are given in section 2.7. Proofs and detailed calculations are listed in Appendix B.

Notation: Throughout the paper we adopt the following notation. For an $m \times n$ real matrix A we write the transpose A'. When A is symmetric, we use $\varrho_{max}(A)$ and $\varrho_{min}(A)$ to denote the largest and smallest eigenvalues, respectively. I_p and $0_{p\times q}$ denote the $p \times p$ identity matrix and $p \times q$ matrix of zeros. The operator $\xrightarrow{p}$ denotes convergence in probability, $\xrightarrow{d}$ denotes convergence in distribution, and *plim* denotes probability limit.

2.2 The Model and Notation

Consider the following linear panel data model with heterogeneous slopes

$$y_{it} = x'_{it}\beta_i + u_{it}, \qquad i = 1, \ldots, N, \text{ and } t = 1, \ldots, T, \tag{2.1}$$

where y_{it} is the dependent variable, $x_{it} = (x_{it,1}, \ldots, x_{it,k})'$ is a $k \times 1$ vector of the regressors including the intercept[1] for unit i, and u_{it} is the unobserved error term, where T is the time dimension, and N is the cross-section dimension. The heterogeneous regression coefficients β_i is a $k \times 1$ vector of unknown coefficients of interest.

Stacking the observations over N units, can be expressed as

$$y_{.t} = X_{.t}\,\boldsymbol{\beta} + u_{.t}, \qquad t = 1, \ldots, T, \tag{2.2}$$

where $y_{.t} = (y_{1t}, \ldots, y_{NT})'$ is a $N \times 1$ vector of observations on the dependent variables at time t, $X_{.t} = diag(x'_{1t}, \ldots, x'_{Nt})$ is a $N \times Nk$ matrix of observations at time t on the regressors, $u_{.t} = (u_{1t}, \ldots, u_{NT})'$ is a $N \times 1$ vector of disturbances for $t = 1, \ldots, T$, and $\boldsymbol{\beta} = (\beta'_1, \ldots, \beta'_N)'$ is a $Nk \times 1$ vector of the unknown slope coefficients.

[1] The first element of x_{it} can take value one ($x_{it,1} = 1$) for all $i = 1, \ldots, N$, and $t = 1, \ldots, T$, which allows for fixed effects. Also, note that we do not assume that x_{it}s are the same, nor do we assume they are different across equations. In other words, our model supports complete heterogeneity, partial heterogeneity, and complete homogeneity of regressors.

Alternatively, stacking the observations over t, we can express the model in (2.1) as

$$y_i = X_i\beta_i + u_i, \qquad i = 1, \ldots, N, \tag{2.3}$$

where $y_i = (y_{i1}, \ldots, y_{iT})'$ is a $T \times 1$ vector of observations on the dependent variable, $X_i = (x_{i1}, \ldots, x_{iT})'$ is a $T \times k$ matrix of observations on the regressors, and $u_i = (u_{i1}, \ldots, u_{iT})'$ is a $T \times 1$ vector of disturbances for $i = 1, \ldots, N$. In a matrix form, we can write the model as

$$y = \mathbb{X}\boldsymbol{\beta} + u, \tag{2.4}$$

where the $NT \times 1$ vector $y = (y_1', \ldots, y_N')'$, $u = (u_1', \ldots, u_N')'$, and

$$\underset{NT\times Nk}{\mathbb{X}} = \begin{bmatrix} X_1 & 0 & \ldots & 0 \\ 0 & X_2 & & \\ \vdots & \ddots & \ddots & \\ 0 & \ldots & 0 & X_N \end{bmatrix}$$

We make the following classical linear system of equations assumptions.

Assumption 2.1 *(i)* $\mathbb{E}(u_{.t}|X_{.1}, \ldots, X_{.T}) = 0$.

(ii) $\mathbb{E}(u_{.t}u_{.t}'|X_{.1}, \ldots, X_{.T}) = \underset{N\times N}{\Sigma}$ *is positive definite, and* Σ^{-1} *is finite.*

Assumption 2.1 (i) requires that the regressors are strictly exogenous and it excludes regressors like lagged dependent variables. This assumption may be restrictive in many applications, but this assumption is needed as a technical regularity condition that is required for proving the asymptotic properties of the estimators when $T \to \infty$. The second

condition requires that the disturbances are uncorrelated across the time dimension, but can be correlated across the cross-sections. In this case $\mathbb{E}(uu'|\,\mathbb{X}) = \Omega = \Sigma \otimes I_T$, where

$$\underset{N\times N}{\Sigma} = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \dots & \sigma_{1N} \\ \sigma_{21} & \sigma_{22} & \dots & \sigma_{2N} \\ & & \vdots & \\ \sigma_{N1} & \sigma_{N2} & \dots & \sigma_{NN} \end{bmatrix}.$$

Therefore, the model of equation (2.1) can also be viewed as a system of seemingly unrelated regressions.

2.3 Estimation

2.3.1 Unrestricted Estimator

The standard estimator of $\boldsymbol{\beta}$ in equation (2.4), is the feasible generalized least-squares (FGLS) estimator of Zellner (1962). This estimator is defined as

$$\hat{\boldsymbol{\beta}} = (\mathbb{X}'\,\hat{\Omega}^{-1}\,\mathbb{X})^{-1}\,\mathbb{X}'\,\hat{\Omega}^{-1}\,y = \boldsymbol{\beta} + \Big(\sum_{t=1}^{T} X'_{.t}\hat{\Sigma}^{-1}X_{.t}\Big)^{-1} X'_{.t}\hat{\Sigma}^{-1}u_{.t}, \tag{2.5}$$

where $\hat{\Omega} = \hat{\Sigma} \otimes I_T$, and $\hat{\Sigma}$ is a consistent estimator of Σ. The (i,j)th element of $\hat{\Sigma}$ is s_{ij} which estimates σ_{ij} using single-equation least-squares estimator of β_i, denoted by $\breve{\beta}_i = (X'_iX_i)^{-1}X'_iy_i$, for $i = 1, 2, ..., N$. Hence

$$s_{ij} = (y_i - X_i\breve{\beta}_i)'(y_j - X_j\breve{\beta}_j)/T = u'_iM_iM_ju_j/T, \tag{2.6}$$

where $M_i = I_T - X_i(X'_iX_i)^{-1}X_i$.

2.3.2 Restricted Estimator

Because of a belief that the true parameter values may be close to a restricted parameter space $\Theta_0 = \{\boldsymbol{\beta} \in \mathbb{R}^{Nk} : \boldsymbol{r}(\boldsymbol{\beta}) = \mathbf{0}\}$ where $\boldsymbol{r}(\boldsymbol{\beta}) = R\boldsymbol{\beta} : \mathbb{R}^{Nk} \to \mathbb{R}^d$, we want to shrink $\hat{\boldsymbol{\beta}}$ towards the restriction space Θ_0. The purpose of the restrictions can be a specification, a structural model, a set of exclusion restrictions, parameter symmetry (like pooling), or any other restrictions that are often tested by means of hypothesis testing to improve the estimation efficiency.

Remark 2.2 *A common restricted parameter space, Θ_0, of particular interest in this setup is the homogeneity restriction of slope parameters across cross-sections, known as pooling. In this case, we would form the restriction as*

$$R\boldsymbol{\beta} = \begin{bmatrix} I_k & \mathbf{0} & \dots & \mathbf{0} & -I_k \\ \mathbf{0} & I_k & \dots & \mathbf{0} & -I_k \\ \vdots & \vdots & & \vdots & \\ \mathbf{0} & \mathbf{0} & \dots & I_k & -I_k \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_N \end{bmatrix} = \begin{bmatrix} \beta_1 - \beta_N \\ \beta_2 - \beta_N \\ \vdots \\ \beta_{(N-1)} - \beta_N \end{bmatrix} = \mathbf{0}, \tag{2.7}$$

this specifies a total of $d = (N-1)k$ restrictions on the $Nk \times 1$ vector of slope parameters.

Remark 2.3 *Another restricted parameter space, Θ_0, which is common in applied economics will take the form of an exclusion restriction for each cross-section equation. For example, if we partition*

$$\beta_i = \begin{bmatrix} \beta_{i,1} \\ \beta_{i,2} \end{bmatrix}, i = 1, 2, \dots, N, \tag{2.8}$$

where $\beta_{i,1}$, $(k - d_i) \times 1$, represents the slopes of the core regressors, and $\beta_{i,2}$, $d_i \times 1$, includes the slopes of included auxiliary regressors that are included in the model for robustness but

may or may not be included in the model. Therefore an exclusion restriction takes the form

$$R\boldsymbol{\beta} = \begin{bmatrix} R_1 & \mathbf{0} & \dots & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & R_2 & \dots & \mathbf{0} & \mathbf{0} \\ \vdots & \vdots & & \vdots & \\ \mathbf{0} & \mathbf{0} & \dots & \mathbf{0} & R_N \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_N \end{bmatrix} = \begin{bmatrix} \beta_{1,2} \\ \beta_{2,2} \\ \vdots \\ \beta_{N,2} \end{bmatrix} = \mathbf{0}, \tag{2.9}$$

where R is a matrix of $d \times Nk$, with $R_i = (\mathbf{0}_{d_i \times (k-d_i)}, I_{d_i})$, for $i = 1, 2, \dots, N$, and $d = \sum_{i=1}^{N} d_i$.

The restricted generalized least-squares estimator is obtained as the solution to the following minimization

$$\underset{\text{s.t. } \boldsymbol{\beta}}{\text{Minimize}} \ (y - \mathbb{X}\boldsymbol{\beta})'\Omega(y - \mathbb{X}\boldsymbol{\beta}) \qquad \text{subject to } \boldsymbol{r}(\boldsymbol{\beta}) = 0, \tag{2.10}$$

and the solution can be formulated as the feasible restricted generalized least-squares estimator in below

$$\tilde{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}} - (\mathbb{X}'\,\hat{\Omega}^{-1}\,\mathbb{X})^{-1}R'\Big[R(\mathbb{X}'\,\hat{\Omega}^{-1}\,\mathbb{X})^{-1}R'\Big]^{-1}R\hat{\boldsymbol{\beta}}. \tag{2.11}$$

Restrictions are often tested using hypothesis testing. The hypothesis to be tested is $H_0 : \boldsymbol{r}(\boldsymbol{\beta}) = 0$ against the alternative, $H_1 : \boldsymbol{r}(\boldsymbol{\beta}) \neq 0$. A conventional test static that has a limiting chi-squared distribution with d degrees of freedom when the null hypothesis is true is

$$F = \hat{\boldsymbol{\beta}}'R'\Big[R(\mathbb{X}'\,\hat{\Omega}^{-1}\,\mathbb{X})^{-1}R'\Big]^{-1}R\hat{\boldsymbol{\beta}} = \Big(\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}\Big)'\,\mathbb{X}'\,\hat{\Omega}^{-1}\,\mathbb{X}\,\Big(\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}\Big), \tag{2.12}$$

which can be recognized as a Wald statistic (Greene (2008)) and measures the weighted distance between $\hat{\boldsymbol{\beta}}$ and $\tilde{\boldsymbol{\beta}}$ [2].

2.3.3 Shrinkage Estimator

We use the restrictions and the test statistic to construct a shrinkage estimator, and show that the proposed estimator improves estimation efficiency and makes an appropriate trade-off between bias due to possible incorrect restrictions and variance efficiency gains from imposing the restrictions.

Our proposed shrinkage estimator of $\boldsymbol{\beta}$ is a weighted average of the FGLS estimator and the restricted estimator

$$\hat{\boldsymbol{\beta}}_s = \omega\, \hat{\boldsymbol{\beta}} + (1-\omega)\, \tilde{\boldsymbol{\beta}}, \tag{2.13}$$

where the weight takes the form

$$\omega = (1 - \frac{\tau}{F(\,\hat{\boldsymbol{\beta}}\,,\,\tilde{\boldsymbol{\beta}}\,)}), \tag{2.14}$$

such that, τ is a positive shrinkage parameter that controls the degree of shrinkage. We will defer describing the optimal choice for this parameter in the following sections. Alternatively, ω can be replaced by its positive part, $(\omega)_+ = \omega\, \mathbb{1}(\omega \geq 0)$, as it can be easily verified that the risk of the estimator with the positive part is smaller. However, it will not affect the results in the following sections, so for simplicity we do not impose it at this stage. Nevertheless, the Monte Carlo results and empirical results are reported using the positive part weight.

[2] The last equality in equation (2.12) holds because
$F = \hat{\boldsymbol{\beta}}' R' \Big[R(\mathbb{X}'\,\hat{\Omega}^{-1}\,\mathbb{X})^{-1} R'\Big]^{-1} R(\mathbb{X}'\,\hat{\Omega}^{-1}\,\mathbb{X})^{-1}(\mathbb{X}'\,\hat{\Omega}^{-1}\,\mathbb{X})(\mathbb{X}'\,\hat{\Omega}^{-1}\,\mathbb{X})^{-1} R' \Big[R(\mathbb{X}'\,\hat{\Omega}^{-1}\,\mathbb{X})^{-1} R'\Big]^{-1} R\hat{\boldsymbol{\beta}}.$

The shrinkage estimator defined above, shrinks the FGLS estimator towards the restricted estimator by the ratio $\tau/F(\hat{\boldsymbol{\beta}}, \tilde{\boldsymbol{\beta}})$, such that when the difference between these two estimators is small (the Wald statistic is small, so $(1-\omega)$ is large), the shrinkage estimator gives a large weight to the restricted estimator, as it is the most efficient estimator. However, when the difference between the two estimators is substantial or high ($F(\hat{\boldsymbol{\beta}}, \tilde{\boldsymbol{\beta}}) > \tau$), the bias of the restricted estimator could be more than its variance efficiency gain, so the shrinkage estimator becomes a weighted average of the restricted and FGLS estimators, while giving a larger weight to the consistent FGLS estimator.

2.4 Asymptotic Properties of the Shrinkage Estimator

In this section, we discuss the asymptotic properties, the asymptotic bias, MSE matrix and risk of the shrinkage estimator defined in (2.13) under a general local asymptotic framework (Assumption 2.5 below), when the time horizon $T \to \infty$ while the cross-section dimension (N) is fixed. We make the following standard set of regulatory assumptions.

Assumption 2.4 *(i)* $\{(X_{.t}, u_{.t}), t = 1, \ldots, T\}$ *are independent and identically distributed.*

(ii) $\mathbb{E}\,|x_{it,h}u_{it}|^2 < \infty$, *for* $i = 1, \ldots, N$, $t = 1, \ldots, T$, $h = 1, \ldots, k$, *and* $\mathbb{E}(X'_{.t}\Sigma^{-1}X_{.t})$ *is positive definite.*

(iii) $\mathbb{E}\,|x_{it,h}|^2 < \infty$, *for* $i = 1, \ldots, N$, $t = 1, \ldots, T$, $h = 1, \ldots, k$, *and* $\mathbb{E}(X'_{.t}X_{.t})$ *is positive definite.*

(iv) W_T *is a* $Nk \times Nk$ *positive definite matrix and tends to finite positive definite matrix* W*, as* $T \to \infty$.

Assumption 2.4 (i)–(iii) require that observations are independent and identically distributed across the time dimension, and give some standard moment conditions to ensure the central limit theorem validity. W_T is a weight matrix in the risk of the estimators. Condition (iv) is automatically satisfied if one sets $W_T = I_{Nk}$.

Assumption 2.5 *We assume that*

$$\beta_i = \bar{\beta}_i + \alpha_i, \quad i = 1, 2, \ldots, N,$$

and

$$\alpha_i = T^{-\kappa}\delta_i, \quad \text{where } \kappa > 0 \text{ , and, } \delta_i \in \mathbb{R}^k,$$

where $\bar{\beta}_i$ *is a centering value which belongs to the restricted parameter space* Θ_0*,* $\delta_i \in \mathbb{R}^k$ *is a localizing parameter which shows the difference between the unrestricted and restricted parameter space, and* κ *is the speed by which the localizing parameter converges to zero. In a matrix form we can write the equations above as*

$$\boldsymbol{\beta} = \bar{\boldsymbol{\beta}} + \boldsymbol{\alpha} = \bar{\boldsymbol{\beta}} + T^{-\kappa}\,\boldsymbol{\delta},$$

where $\underset{Nk\times 1}{\boldsymbol{\alpha}} = (\alpha_1', \alpha_2', \ldots, \alpha_N')'$ *and* $\underset{Nk\times 1}{\boldsymbol{\delta}} = (\delta_1', \delta_2', \ldots, \delta_N')'$.

Remark 2.6 *When the restricted parameter space exhibits the parameter symmetry restrictions across the cross-sections (see Remark 2.2),* $\bar{\beta}_i$ *represents a common mean, i.e.* $\bar{\beta}_i = \bar{\beta}$ *for* $i = 1, 2, \ldots, N$.

Remark 2.7 *For the restricted parameter space in Remark 2.3 that exhibits the exclusion restriction parameter space, the centering parameter takes the form* $\bar{\beta}_i = (\beta'_{i,1}, \mathbf{0}'_{d_i \times 1})'$.

Assumption 2.5 controls the magnitude of the difference between the restricted and unrestricted parameter space. We need this assumption to ensure that the distance between these two parameter space diminish as the sample size increases. Because otherwise, the magnitude of the bias and the risk of the restricted estimator increase with the sample size, and there is no gain of shrinking the unrestricted estimator toward the restricted parameter space. We will discuss in detail how different values of κ affect the bias and risk of the shrinkage estimator in Theorem 2.9.

Theorem 2.8 *Under assumptions 2.1–2.4, the asymptotic distribution of the FGLS estimator is*

$$\sqrt{T}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{d} Z \sim N(\mathbf{0}, V), \quad \text{where } V \equiv \Big[\mathbb{E}(X'_{.t}\Sigma^{-1}X_{.t})\Big]^{-1}, \tag{2.15}$$

and together with assumption 2.5, the asymptotic distribution of the restricted estimator is

$$\sqrt{T}(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{d} Z - VR'(RVR')^{-1}R(Z + \sqrt{T}\,\boldsymbol{\alpha}). \tag{2.16}$$

Further by using the above equations,

$$F(\,\hat{\boldsymbol{\beta}}\,,\,\tilde{\boldsymbol{\beta}}\,) \xrightarrow{d} (Z + \sqrt{T}\,\boldsymbol{\alpha})'R'(RVR')^{-1}R(Z + \sqrt{T}\,\boldsymbol{\alpha}) \equiv \xi(Z) \sim \boldsymbol{\chi}^2_d(\boldsymbol{\theta}'\,A\,\boldsymbol{\theta}), \tag{2.17}$$

where [3] $\boldsymbol{\theta} = V^{-1/2}\sqrt{T}\,\boldsymbol{\alpha}$, $A = V^{1/2}R'(RVR')^{-1}RV^{1/2}$ *is an idempotent matrix, and for the shrinkage weight, we have*

$$(1 - \frac{\tau}{F(\,\hat{\boldsymbol{\beta}}\,,\,\tilde{\boldsymbol{\beta}}\,)}) \xrightarrow{d} (1 - \frac{\tau}{\xi(Z)}) \equiv \omega(Z). \tag{2.18}$$

[3] $\chi^2_p(q)$ represents a chi-squared distribution with p degrees of freedom and a non-centrality parameter q.

Therefore, the asymptotic distribution of the shrinkage estimator is

$$\sqrt{T}(\hat{\boldsymbol{\beta}}_s - \boldsymbol{\beta}) \xrightarrow{d} \omega(Z)Z + (1-\omega(Z))(Z - VR'(RVR')^{-1}R(Z + \sqrt{T}\,\boldsymbol{\alpha})) \equiv Z_s. \tag{2.19}$$

Proof: Appendix B, (See page 150).

Theorem 2.8 gives the asymptotic distribution of the shrinkage estimator, which is a non-linear function of a normal distribution, and will be used to approximate the moments of the shrinkage estimator. Moreover, since the shrinkage estimator is a non-linear function of random variables, obtaining its bias, mean squared error matrix (MSEM), and risk is difficult. Hence, as a useful approximation, we study the truncated moments of the shrinkage estimator as $T \to \infty$. We define, the asymptotic bias of the shrinkage estimator as

$$\text{ABias }(\hat{\boldsymbol{\beta}}_s) = \lim_{\gamma\to\infty}\lim_{T\to\infty} \mathbb{E}\left[T_n\, \mathbb{1}\Big(|T_n| \leq \gamma\Big)\right] = \mathbb{E}(Z_s),$$

where $T_n = \sqrt{T}(\hat{\boldsymbol{\beta}}_s - \boldsymbol{\beta})$. The asymptotic MSEM of the shrinkage estimator is defined as,

$$\text{AMSEM}(\hat{\boldsymbol{\beta}}_s) = \lim_{\gamma\to\infty}\lim_{T\to\infty} \mathbb{E}\left[T_n\, T_n'\, \mathbb{1}\Big(|T_n| \leq \gamma\Big)\right] = \mathbb{E}(Z_s Z_s'),$$

and the asymptotic risk of the shrinkage estimator for a weight matrix W_T satisfying Assumption 2.4 (iv), is defined as

$$\text{ARisk }(\hat{\boldsymbol{\beta}}) = \lim_{\gamma\to\infty}\lim_{T\to\infty} \mathbb{E}\left[T_n' W_T T_n\, \mathbb{1}\Big(|T_n| \leq \gamma\Big)\right] = \mathbb{E}(Z_s' W Z_s) = \text{tr}\left(W\, \mathbb{E}(Z_s Z_s')\right).$$

The last equality in the above equalizations hold, because the truncated moments are continuous and bounded functions of the shrinkage estimator. Consequently, we can approximate the truncated moments, as the sample size increases with negligible trimming, using the asymptotic distribution of the shrinkage estimator [4]. Furthermore, W_T is a

[4]This holds because for every bounded continuous real-valued function f, $Z_T \xrightarrow{d} Z$ if and only if $\mathbb{E}(f(Z_T)) \xrightarrow{p} \mathbb{E}(f(Z))$, see theorem 1.8.8 of Lehmann and Casella (1998). In our case, $f(T_n) = T_n\, \mathbb{1}(|T_n| \leq \gamma) + \gamma \mathbb{1}(|T_n| > \gamma)$ for the asymptotic bias, and is similarly defined for the asymptotic MSE matrix and risk.

positive definite weight matrix that satisfies condition (iv) in Assumption 2.4. Two arbitrary choices of W_T are I_{Nk} and $T^{-1}(\mathbb{X}'\,\hat{\Omega}\,\mathbb{X})^{-1}$, where the former one in the risk, provides an unweighted mean squared error (MSE), and the latter gives the mean squared forecast (prediction) error (MSFE).

Theorem 2.9 *Under assumptions 2.1–2.5, the asymptotic bias of the shrinkage estimator is*

$$ABias\,(\hat{\boldsymbol{\beta}}_s) = -T^{(1-2\kappa)/2}\,\frac{\tau}{d}\,VR'(RVR')^{-1}R\,\boldsymbol{\delta}\;e^{(-T^{1-2\kappa}\lambda)}\;{}_1F_1\Big(\frac{d}{2}\,,\,\frac{d}{2}+1;\,T^{1-2\kappa}\,\lambda\Big), \tag{2.20}$$

and the asymptotic MSEM of the shrinkage estimator when $d > 2$ is

$$\begin{aligned} AMSEM\,(\hat{\boldsymbol{\beta}}_s) = V + \frac{\tau}{d}\,VR'(RVR')^{-1}RV\;e^{(-T^{1-2\kappa}\lambda)} \\ \Bigg[\frac{\tau}{d-2}\;{}_1F_1\Big(\frac{d}{2}-1\,,\,\frac{d}{2}+1\,;\,T^{1-2\kappa}\lambda\Big) - 2\;{}_1F_1\Big(\frac{d}{2}\,,\,\frac{d}{2}+1\,;\,T^{1-2\kappa}\lambda\Big)\Bigg] \\ +\,\tau\,T^{1-2\kappa}\,VR'(RVR')^{-1}R\,\boldsymbol{\delta}\,\boldsymbol{\delta}'\,R'(RVR')^{-1}RV\;e^{(-T^{1-2\kappa}\lambda)} \\ \Bigg[\frac{\tau}{d(d+2)}\;{}_1F_1\Big(\frac{d}{2}\,,\,\frac{d}{2}+2\,;\,T^{1-2\kappa}\lambda\Big) - 2\Big[\frac{1}{d+2}\;{}_1F_1\Big(\frac{d}{2}+1\,,\,\frac{d}{2}+2\,;\,T^{1-2\kappa}\lambda\Big) \\ -\,\frac{1}{d}\;{}_1F_1\Big(\frac{d}{2}\,,\,\frac{d}{2}+1\,;\,T^{1-2\kappa}\lambda\Big)\Big]\Bigg], \end{aligned} \tag{2.21}$$

where $\lambda = \boldsymbol{\delta}'\,R'(RVR')^{-1}R\,\boldsymbol{\delta}\,/2$, and ${}_1F_1(.,.;.)$ denotes the confluent hypergeometric function [5].

[5]The confluent hypergeometric function is given by

$${}_1F_1(a,c;x) = \sum_{n=0}^{\infty}\frac{(a)_n x^n}{(c)_n n!}$$

where $(a)_n = a(a+1)\ldots(a+n-1)$, $(a)_0 = 1$. Also $(a)_n = \Gamma(a+n)/\Gamma(a)$ for positive a.

Proof: Appendix B,(See page 153).

Corollary 2.10 *Under assumptions 2.1–2.5, and $d > 2$, we have the followings:*

(i) If $0 < \kappa < 1/2$, the shrinkage estimator is asymptotically unbiased, consistent, and very close to the FGLS estimator

$$ABias\ (\hat{\boldsymbol{\beta}}_s) = -T^{-(1-2\kappa)/2}\,\frac{\tau}{2\lambda} VR'(RVR')^{-1}R\,\boldsymbol{\delta}\,\Big[1 + O(T^{-(1-2\kappa)})\Big], \tag{2.22}$$

and the asymptotic MSEM is

$$\begin{aligned} AMSEM\ (\hat{\boldsymbol{\beta}}_s) = V - T^{-(1-2\kappa)}\,\frac{\tau}{\lambda}\,VR'(RVR')^{-1}RV \; + T^{-(1-2\kappa)}\,\Big[\frac{\tau^2}{4\lambda^2} + \frac{\tau}{\lambda^2}\Big] \\ VR'(RVR')^{-1}R\,\boldsymbol{\delta}\,\boldsymbol{\delta}'\,R'(RVR')^{-1}RV + O(T^{-2(1-2\kappa)}), \end{aligned} \tag{2.23}$$

(ii) Local Asymptotic: If $\alpha_i = \delta_i T^{-1/2}$, $i = 1, 2, \ldots, N$, i.e. $\kappa = 1/2$, then the shrinkage estimator has an asymptotic bias of order $O(T^{-1/2})$, is consistent and we have

$$ABias\ (\hat{\boldsymbol{\beta}}_s) = -\frac{\tau}{d}\,e^{-\lambda}\,VR'(RVR')^{-1}R\,\boldsymbol{\delta}\;\,{}_1F_1\Big(\frac{d}{2}\,,\,\frac{d}{2}+1\,;\,\lambda\Big), \tag{2.24}$$

and

$$\begin{aligned} AMSEM\ (\hat{\boldsymbol{\beta}}_s) = V + \frac{\tau}{d}\,e^{-\lambda}\Bigg[\frac{\tau}{d-2}\,{}_1F_1\Big(\frac{d}{2}-1\,,\,\frac{d}{2}+1\,;\,\lambda\Big) \\ -\,2\,{}_1F_1\Big(\frac{d}{2}\,,\,\frac{d}{2}+1\,;\,\lambda\Big)\Bigg]VR'(RVR')^{-1}RV \; + e^{\lambda}\,\tau\,\Bigg[\frac{\tau}{d(d+2)}\,{}_1F_1\Big(\frac{d}{2}\,,\,\frac{d}{2}+2\,;\,\lambda\Big) \\ -\,2\Big[\frac{1}{d+2}\,{}_1F_1\Big(\frac{d}{2}+1\,,\,\frac{d}{2}+2\,;\,\lambda\Big) - \frac{1}{d}\,{}_1F_1\Big(\frac{d}{2}\,,\,\frac{d}{2}+1\,;\,\lambda\Big)\Big]\Bigg] \\ VR'(RVR')^{-1}R\,\boldsymbol{\delta}\,\boldsymbol{\delta}'\,R'(RVR')^{-1}RV. \end{aligned} \tag{2.25}$$

Proof: Appendix B,(See page 156).

Remark 2.11 *The asymptotic MSEM of the shrinkage estimator in (2.25) can be rewritten as follows* [6]

$$
\begin{aligned}
AMSEM\,(\hat{\boldsymbol{\beta}}_s) &= V + \frac{1}{d(d-2)}\, e^{-\lambda}\, {}_1F_1\Big(\frac{d}{2}-1\,,\,\frac{d}{2}+1\,;\,\lambda\Big)\Big[\tau^2 - 2\tau(d-2)\Big] VR'(RVR')^{-1}RV \\
&+ \frac{1}{d(d+2)}\, e^{-\lambda}\, {}_1F_1\Big(\frac{d}{2}\,,\,\frac{d}{2}+2\,;\,\lambda\Big)\Bigg[\Big(\tau^2+4\tau\Big) VR'(RVR')^{-1}R\,\boldsymbol{\delta}\,\boldsymbol{\delta}'\,R'(RVR')^{-1}RV \\
&- 4\,\tau\,\lambda\, VR'(RVR')^{-1}RV\Bigg].
\end{aligned}
\tag{2.26}
$$

In the following corollary, we give our recommended value of τ that minimizes the risk of the shrinkage estimator under the local asymptotic condition.

Corollary 2.12 *Under assumptions 2.1–2.5, when* $\kappa = 1/2$, $tr(C)/\varrho_{max}(C) > 2$, *and* $0 < \tau \leq 2\Big[\, tr(C)/\varrho_{max}(C) - 2\Big]$, *then*

$$
\begin{aligned}
ARisk\,(\hat{\boldsymbol{\beta}}_s) \leq ARisk\,(\hat{\boldsymbol{\beta}}) &- \frac{e^{-\lambda}}{d}\Bigg[2\tau\Big(\frac{tr(C)}{\varrho_{max}(C)} - 2\Big) - \tau^2\Bigg] \\
&\Bigg[\frac{tr(C)}{d-2}\, {}_1F_1\Big(\frac{d}{2}-1\,,\,\frac{d}{2}+1\,;\,\lambda\Big) + \frac{2\lambda_W}{d+2}\, {}_1F_1\Big(\frac{d}{2}\,,\,\frac{d}{2}+2\,;\,\lambda\Big)\Bigg],
\end{aligned}
\tag{2.27}
$$

where $ARisk\,(\hat{\boldsymbol{\beta}}) = tr(W\,V)$, $C = AV^{1/2}WV^{1/2}A$, $\lambda_W = \boldsymbol{\delta}'\,V^{-1/2}CV^{-1/2}\,\boldsymbol{\delta}\,/2$. *The above result shows the superiority of the shrinkage estimator relative to the FGLS estimator. The optimal shrinkage parameter that minimizes the risk is*

$$
\tau_{opt} = tr(C)/\varrho_{max}(C) - 2. \tag{2.28}
$$

Proof: Appendix B, (See page 156).

[6] The result holds by using the following identities

$$
\begin{aligned}
&(c-a-1)\, {}_1F_1(a,c;x) = (c-1)\, {}_1F_1(a,c-1;x) - a\, {}_1F_1(a+1,c;x), \\
&{}_1F_1(a,c;x) = {}_1F_1(a+1,c;x) - \frac{x}{c}\, {}_1F_1(a+1,c+1;x),
\end{aligned}
$$

See Lebedev (1972), pp. 271.

Remark 2.13 *As the optimal shrinkage parameter depends on* Ω*, which is unknown, it can be estimated. That is one can replace* Ω *by its consistent estimator* $\hat{\Omega}$*, and use*

$$\hat{\tau}_{opt} = tr(\hat{C})/\varrho_{max}(\hat{C}) - 2. \tag{2.29}$$

In this case as $T \to \infty$*,* $\hat{\tau}_{opt} \xrightarrow{p} \tau_{opt}$*, and the results of corollary 2.12 will still hold.*

Corollary 2.14 *Under assumptions 2.1–2.5, when* $\kappa = 1/2$*,* $d > 2$*, and* $0 < \tau \leq 2[d-2]$*, the MSFE of the shrinkage estimator* $(W = V^{-1})$ *is*

$$MSFE\ (\hat{\boldsymbol{\beta}}_s) = MSFE\ (\hat{\boldsymbol{\beta}}) - e^{-\lambda}\ \frac{1}{d-2}\ {}_1F_1\Big(\frac{d}{2} - 1\,,\ \frac{d}{2}\,;\ \lambda\Big)\Big[2\tau(d-2) - \tau^2\Big], \tag{2.30}$$

where $MSFE\ (\hat{\boldsymbol{\beta}}) = I_{Nk}$*. The value of* τ *that minimizes the MSFE of the shrinkage estimator is*

$$\tau_{F,opt} = d - 2, \tag{2.31}$$

and the MSFE of the optimal shrinkage estimator is

$$MSFE\ (\hat{\boldsymbol{\beta}}_{s,opt}) = MSFE\ (\hat{\boldsymbol{\beta}}) -\ e^{-\lambda}\ (d-2)\ {}_1F_1\Big(\frac{d}{2} - 1\,,\ \frac{d}{2}\,;\ \lambda\Big). \tag{2.32}$$

Proof: Appendix B,(See page 157).

Corollary 2.15 *Under assumptions 2.1–2.5, when* $\kappa = 1/2$*, and* $d > 2$*, if* $\lambda \to \infty$ [7] *then*

$$ARisk\ (\hat{\boldsymbol{\beta}}_{s,opt}) = ARisk\ (\hat{\boldsymbol{\beta}}) + O(\frac{1}{\lambda}). \tag{2.33}$$

Proof: Appendix B,(See page 156).

[7] Equivalently when $\boldsymbol{\delta} \to \infty$.

The result in corollary 2.15 suggests that if the bias of the restricted estimator is very large (the restricted parameter space is too far from the true parameter space), the shrinkage estimator is asymptotically very close to the FGLS estimator, and achieves the global minimax efficiency bond of van der Vaart (1998). This condition assures that even for very large values of $\boldsymbol{\delta}$ in assumption 2.5, the shrinkage estimator remains asymptotically consistent and efficient by giving a weight one to the FGLS estimator.

2.4.1 High Dimensional Shrinkage

In this section, we study the performance of our estimator in a high dimensional case where the number of restrictions increases without bound. The asymptotic properties of our estimator is given in the following theorem using a sequential approximations by letting first the sample size, and then the number of restrictions, tent to infinity.

Theorem 2.16 *Under assumptions 2.1–2.5, when $\kappa = 1/2$, if as $d \to \infty$, $\lim_{d\to\infty} \lambda/d \to 0$ then*

$$\lim_{d\to\infty} \frac{ARisk\ (\hat{\boldsymbol{\beta}}_{s,opt})}{ARisk\ (\hat{\boldsymbol{\beta}})} \leq 1-\rho, \qquad \rho = \lim_{d\to\infty} \frac{tr(C)}{tr(WV)}, \tag{2.34}$$

where $0 \leq \rho \leq 1$. If $W = V^{-1}$, then we have

$$\lim_{d\to\infty} \frac{MSFE\ (\hat{\boldsymbol{\beta}}_{s,opt})}{MSFE\ (\hat{\boldsymbol{\beta}})} = 1 - \lim_{d\to\infty} \frac{d}{NK}. \tag{2.35}$$

Also, in the expressions above the optimal shrinkage estimator can be replaced with any shrinkage estimator in which the shrinkage parameter, τ, satisfies the condition below

$$\lim_{d\to\infty} \frac{\tau}{\tau_{opt}} \to 1. \tag{2.36}$$

Proof: Appendix B,(See page 158).

The right hand side of equation (2.34) is equal to the local minimax efficiency bound given in Theorem 5 of Hansen (2016), which specifies that our proposed estimator asymptotically achieves the local minimax bound, while the FGLS estimator does not. Therefore, the shrinkage estimator proposed in this paper is locally the most efficient estimator. Consequently, there is no need to find alternative methods (like model averaging) to balance between the bias and variance efficiency. A major advantage of our proposed shrinkage method relative to the Stein-type shrinkage estimator considered in Hansen (2016) is that, the risk of our estimator does not depend on the bound size of localizing parameters, as a result the gain of our proposed estimator relative to the FGLS estimator can be quantified.

2.5 Monte Carlo Simulation

The results below are the simulation results of the model of section 2.2, where $x_{it,1} = 1$ and the remaining regressors are independently generated from the standard normal distributions. The sample size varies from $T \in \{50, 100, 200\}$, $N \in \{3, 5\}$, $k \in \{4, 6\}$, leading to twelve combinations of N, T and k. u_1 is generated as $i.i.d\, N(0, 1)$, while $u_i = cu_1 + v_i$, for $i = 2, \ldots, N$, where $v_i \sim i.i.d\, N(0, 1)$ and $c = 0.25$. We consider two DGPs for generating β_i, the first one is under a complete heterogeneity in coefficients where we assume that

DGP1: $\beta_i = \bar{\beta} + (i \times \delta)/N,\ i = 1, 2, \ldots, N,$

with $\bar{\beta} = (1, 1, \ldots, 1)'$, and the second DGP is under a partial heterogeneity where we assume that

$$\text{DGP2:} \quad \beta_{i1}, \beta_{i2} = \begin{cases} 1 + (i \times \delta)/N, & \text{if } i = 1, \ldots, [N/2] \\ 1.2, & \text{if } i = [N/2] + 1, \ldots, N \end{cases}, \beta_{il} = 2,\ l \in \{3, \ldots, k\},$$

where $[N/2]$ denotes the nearest integer value that is smaller than $N/2$, and δ takes values on a 10-point grid on $[0, 1]$.

The results of $1,000$ monte carlo simulations are given in Figures 2.1–2.8, where the vertical axis measure the relative mean squared error (RMSE) of the FGLS estimator, the restricted estimator, a pre-test estimator, and the optimal shrinkage estimator, to the FGLS estimator. The horizontal axis measure the degree of heterogeneity (δ) which is set between zero and one with 0.1 grid value.

The Monte Carlo results support our theoretical findings of the previous section. The figures show that the RMSE of the shrinkage estimator for the whole parameter heterogeneity is below that of the FGLS estimator. This shows the superiority of our proposed shrinkage estimator relative to the FGLS estimator.

The RMSE of the shrinkage estimator in DGP1 of a complete heterogeneous panel data model, is smaller than that of the restricted estimator except for very small values of parameter heterogeneity. This is expected because as δ takes higher values, the bias of the restricted estimator increases, which then increases its MSE. Also, when the sample size is larger, the RMSE of the shrinkage estimator dominates that of the restricted estimator for most values of δ. In DGP2 where the model is characterized by some degrees of homogeneity, the RMSE of the restricted estimator remains smaller than that of the FGLS estimator for

even larger values of δ. In this case, the FGLS estimator can be inferior to the restricted estimator even with the presence of weak degrees of heterogeneity. This is because although the FGLS estimator is unbiased, it is inefficient, especially under small sample sizes, and high number of regressors. In contrast, the restricted estimator properly makes use of cross-section variation and thus provides a more accurate results.

In general, we find that the shrinkage estimator performs robustly well in heterogeneous panel data models with various degrees of heterogeneity. When there is a strong heterogeneity, the shrinkage estimator prevails. When there is a relatively weak heterogeneity, the shrinkage estimator tends to gain more from the efficiency of the restricted estimator by assigning a larger weight to this estimator, and thus still remains one of the best choices.

2.6 Application: Forecasting Cross-Country Output Growth

In this section, we present an empirical application that highlights the utility of the shrinkage estimator in forecasting. In particular, we forecast the output growth rate of 33 advanced and emerging economies in the global economy using a set of macroeconomic and financial variables by allowing potential parameter heterogeneity structures in the slope coefficients. This allows us to shrink the slope parameters of the countries with close response variables which can improve the forecasts. As pointed out by Pesaran et al. (2009) the unobserved heterogeneity is an important issue that practitioners face when constructing forecasting models which is still an open discussion. We consider a panel data model with an uncertainly about the parameter heterogeneity structures which adds to the current and

ongoing literature of forecasting economic and financial variables across countries including Dees et al. (2007,a), Dees et al. (2007,b), and Pesaran et al. (2009), among others.

The data set is taken from the Global VAR (GVAR) dataset[8]. We use quarterly macroeconomic and financial variables including log real GDP (y_{it}), the rate of inflation (π_{it}), short-term interest rate (r_{it}), long-term interest rate (lr_{it}), and log real equity prices (q_{it}) for $N = 33$ economies from 1979Q2 to 2016Q4.

We are interested in forecasting h quarters ahead rate of log real GDP, with the predictors in $z_{it} = (\Delta r_{i,t} - \Delta\pi_{it}, \Delta lr_{i,t} - \Delta\pi_{it}, \Delta q_{i,t} - \Delta\pi_{it})$ and $z^*_{it} = (\Delta y^*_{i,t}, \Delta r^*_{i,t} - \Delta\pi^*_{it}, \Delta lr^*_{i,t} - \Delta\pi^*_{it}, \Delta q^*_{i,t} - \Delta\pi^*_{it})$, where z^*_{it} is the country-specific foreign variables. The foreign variables are constructed using rolling three year moving averages of the annual trade weights which are computed as shares of exports and imports for each country [9].

Therefore, we consider the following equation

$$\Delta_h y_{i,t+h} = \eta_i + \beta_i' z_{it} + \beta_i^{*'} z^*_{it} + u_{it}, \quad i = 1, \ldots, N, \text{ and } t = 1, \ldots, T, \tag{2.37}$$

where $\Delta_h y_{i,t+h} = y_{i,t+h} - y_{it}$ for the forecast horizon h, and the slope parameters, β^*_i, admit a possible parameter heterogeneity structure, while η_i and β_i are heterogenous across countries. We estimate the slope parameters using the shrinkage estimation method developed in the previous sections. In our analysis, we consider up to $h = 4$ (four quarters ahead) and report results for one quarter ahead ($h = 1$) and one year ahead ($h = 4$). The forecasts are constructed using expanding windows of 15 ($T = 60$), 20 ($T = 80$), and 25 ($T = 100$) years time periods for the initial estimation window. When $T = 60$, this leaves

[8]The data is available at the GVAR Toolbox webpage https://sites.google.com/site/gvarmodelling/data

[9]For example the trade weights of year 2016 is based on the average trade flows computed over the three years 2013–2015. Because the trade flows observations start at 1980, the process of computing time-varying trade weights was initialized by using the same set of weights for the first four years of the sample period.

us with the last $H_1 = 83$ out-of-sample evaluation periods, 1996Q2-2016Q4 for $h = 1$, and $H_2 = 79$ out-of-sample evaluation periods, 1997Q2-2016Q4 for $h = 4$.

We evaluate the forecasting performance of our method, with individual equations forecasts, and a fixed effect approach using the root mean squared forecast error (RMSFE) of any given model, which is averaged across the N countries as below

$$RMSFE(h, H) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\frac{1}{H_h}\sum_{t=T}^{T+H_h-1}\hat{e}_{it}^2(h)}, \quad h = 1, 4, \tag{2.38}$$

where $\hat{e}_{it}(h) = \Delta_h y_{i,t+h} - \widehat{\Delta_h y}_{i,t+h|t}$ is the $h-$quarter ahead forecast error, with $\Delta_h y_{i,t+h}$ being the actual value, and $\widehat{\Delta_h y}_{i,t+h|t}$ the corresponding forecast formed at time t. RMSFE and relative RMSFE statistics for the one-quarter and one-year ahead forecasts of output growth rate are reported in Table 2.1.

Diebold and Mariano (1995) (DM) test statistics for testing $H_0 : \mathbb{E}\left(\hat{x}_{it,m}(h)\right) = 0$, where $\hat{x}_{it,m}(h) = \hat{e}^2_{it,\text{shrinkage}}(h) - \hat{e}^2_{it,m}(h)$ is the difference between the $h-$quarter ahead squared forecasting errors of our shrinkage method and method m (fixed effect or individual equations models) for country i. Specifically, by assuming serially uncorrelated h-step-ahead forecasting errors, we have

$$DM_{i,m}(h) = \sqrt{H_h}\,\frac{\bar{\hat{x}}_{i,m}(h)}{\hat{\sigma}_{i,m}(h)}, \quad i = 1, \ldots, N, \text{ and } h = 1, 4, \tag{2.39}$$

where $\bar{\hat{x}}_{i,m}(h) = \frac{1}{H_h}\sum_{t=T+1}^{T+H_h}\hat{x}_{it,m}(h)$ is the sample mean of $\hat{x}_{it,m}(h)$, and

$$\hat{\sigma}^2_{i,m}(h) = \frac{1}{H_h - 1}\sum_{t=T+1}^{T+H_h}\left(\hat{x}_{it,m}(h) - \bar{\hat{x}}_{it,m}(h)\right)^2. \tag{2.40}$$

To compare the forecasts across the countries, we compute the panel version of the DM test which is proposed in Pesaran et al. (2009) to statistically test the panel forecasts across countries against each method for a given forecast horizon. The panel DM $(\overline{DM})$ statistic

under assuming serially and cross-sectionally uncorrelated h-step-ahead forecasting errors is defined as

$$\overline{DM}_m = \frac{\bar{x}_m(h)}{\sqrt{V\Big(\bar{x}_m(h)\Big)}}, \quad h = 1, 4, \tag{2.41}$$

where $\bar{x}_m(h) = \frac{1}{N}\sum_{i=1}^{N} \bar{\hat{x}}_{i,m}(h)$ and $V\Big(\bar{x}_m(h)\Big) = \frac{1}{NT}\Big(\frac{1}{N}\sum_{i=1}^{N} \hat{\sigma}^2_{i,m}(h)\Big)$. The panel $\overline{DM}$ test results are reported in Table 2.2 for one-quarter and one-year ahead forecasts.

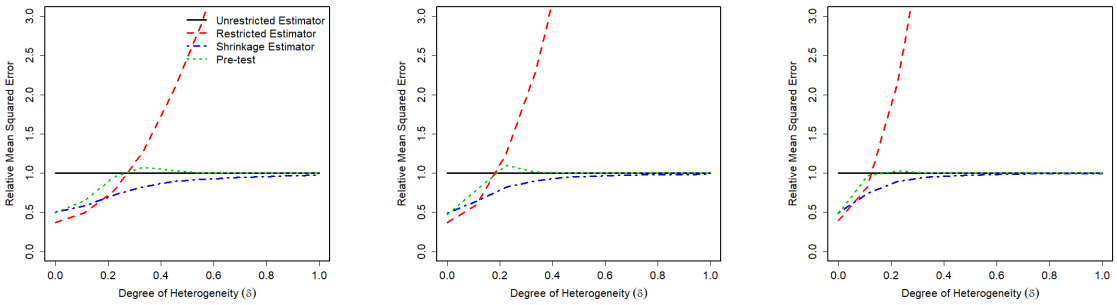
We note that one quarter ahead shrinkage estimation forecasts perform better than the fixed effects and individual estimators in all cases and the panel $\overline{DM}$ tests are significant. For the one-year ahead forecasts our proposed shrinkage estimation forecasts outperforms the other two methods.

2.7 Conclusion

We introduce a new method of estimation and forecasting in heterogeneous panel data models under cross section-dependence and heteroscedasticity of the errors to address the problem of model uncertainty. This method has four main advantages relative to the other model averaging and shrinkage estimation methods. First, it allows for heteroscedasticity and cross-section dependence of error terms which is essential in most of the panel data model applications. Second, the dominance and optimality of the shrinkage estimator proposed here is not limited to MSFE and holds for any weighted quadratic loss function where the weight is positive definite and symmetric. Third, the shrinkage weight is proportional to a Wald statistics that controls for rotations of the coefficient vectors, hence provides a shrinkage estimator with a uniformly lowest risk. Lastly, the framework

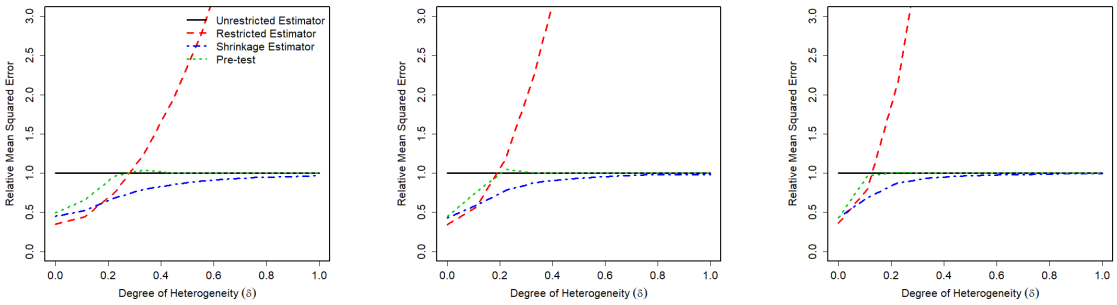
considered here is not limited to the local misspecification, and the dominance properties of the shrinkage estimator is given against a set of deviations from the restrictions.

Moreover, this chapter contributes to the long-existing issue in the panel data analysis referred by econometricians to as "to pool or not to pool". We compare the performance of our proposed estimator with the single-equation and pooling estimators, and show the reliability of the estimation results under our shrinkage estimator. Moreover, we apply our method to forecast the output growth rate of 33 advanced and emerging economies in the global economy using a set of macroeconomic and financial variables by allowing potential parameter heterogeneity structures in the slope coefficients. Our method has two advantages over the method considered in the literature. First, it allows for correlation among the error terms across the courtiers. This correlation could be due to omitted common effects, or could arise as a result of interactions within socioeconomic networks. Second, as there is a model specification uncertainty issue about the parameter heterogeneity of the output growth rates, our method, unlike previous studies, considers the uncertainty of model selection and estimation jointly. Therefore, the results are more robust and reliable than the single-equation or pooling estimators mostly considered in the literature.



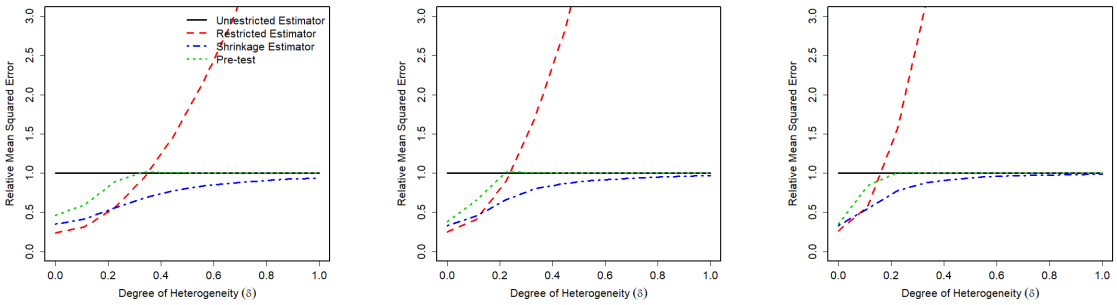
$T = 50$ $T = 100$ $T = 200$

Figure 2.1: Relative MSE of Unrestricted, Restricted, Pre-test, and Shrinkage Estimators, for DGP1, $N = 3$, $k = 4$



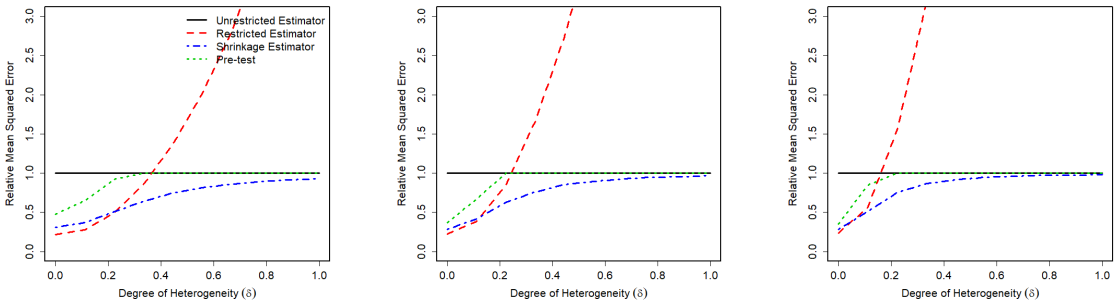
$T = 50$ $T = 100$ $T = 200$

Figure 2.2: Relative MSE of Unrestricted, Restricted, Pre-test, and Shrinkage Estimators, for DGP1, $N = 3$, $k = 6$



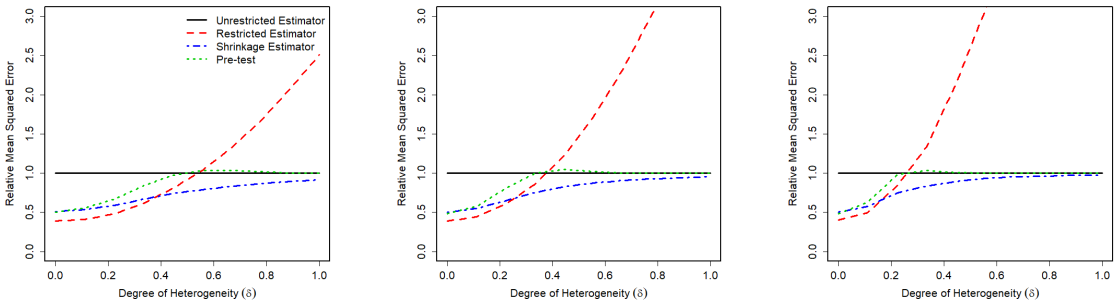
$T = 50$ $T = 100$ $T = 200$

Figure 2.3: Relative MSE of Unrestricted, Restricted, Pre-test, and Shrinkage Estimators, for DGP1, $N = 5$, $k = 4$



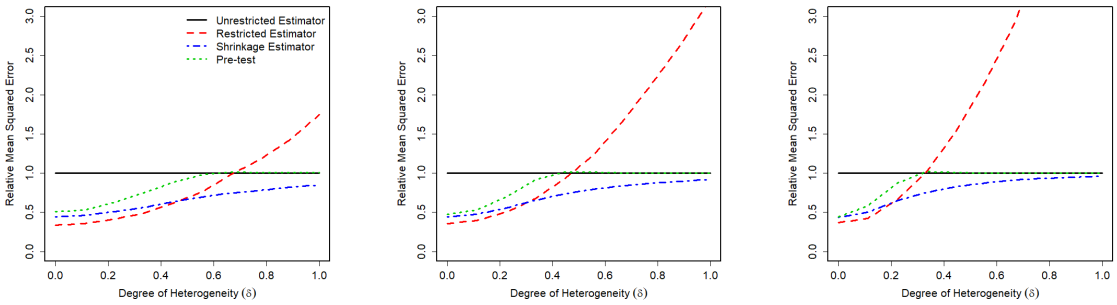
$T = 50$ $T = 100$ $T = 200$

Figure 2.4: Relative MSE of Unrestricted, Restricted, Pre-test, and Shrinkage Estimators, for DGP1, $N = 5$, $k = 6$



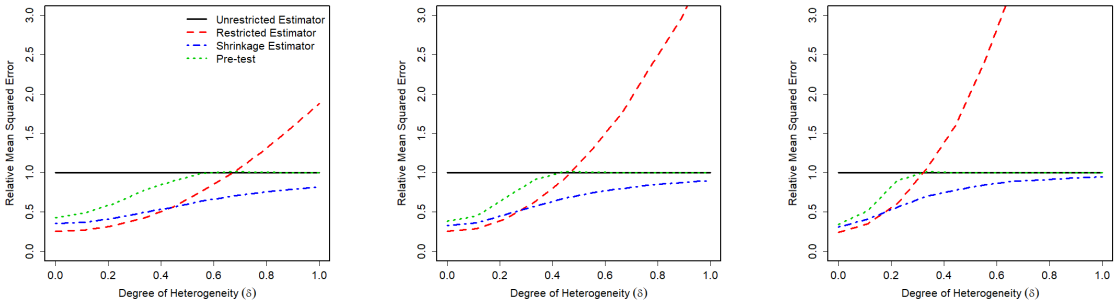
$T = 50$ $T = 100$ $T = 200$

Figure 2.5: Relative MSE of Unrestricted, Restricted, Pre-test, and Shrinkage Estimators, for DGP2, $N = 3$, $k = 4$



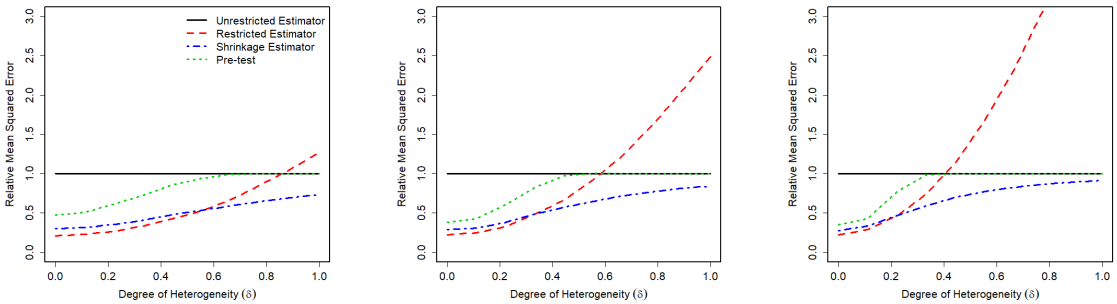
$T = 50$ $T = 100$ $T = 200$

Figure 2.6: Relative MSE of Unrestricted, Restricted, Pre-test, and Shrinkage Estimators, for DGP2, $N = 3$, $k = 6$



$T = 50$ $T = 100$ $T = 200$

Figure 2.7: Relative MSE of Unrestricted, Restricted, Pre-test, and Shrinkage Estimators, for DGP2, $N = 5$, $k = 4$



$T = 50$ $T = 100$ $T = 200$

Figure 2.8: Relative MSE of Unrestricted, Restricted, Pre-test, and Shrinkage Estimators, for DGP2, $N = 5$, $k = 6$

Table 2.1: RMSFE performance of the shrinkage estimation, individual estimators, and fixed effect methods for one quarter ahead ($h = 1$) and one year (four quarters, $h = 4$) ahead output growth forecasts across 33 countries

Models	$T = 60$		$T = 80$		$T = 100$	
	RMSFE ($\times 100$)	Relative RMSFE	RMSFE ($\times 100$)	Relative RMSFE	RMSFE ($\times 100$)	Relative RMSFE
			$h = 1$			
Shrinkage Est.	1.300	1.000	1.234	1.000	1.250	1.000
Fixed Effects	1.358	1.013	1.245	1.010	1.270	1.012
Individual Est.	1.308	1.001	1.239	1.004	1.255	1.003
			$h = 4$			
Shrinkage Est.	3.216	1.000	3.080	1.000	3.103	1.000
Fixed Effects	3.312	1.030	3.177	1.031	3.186	1.001
Individual Est.	3.220	1.001	3.083	1.001	3.106	1.027

Note: RMSFE is computed using an expanding forecasting scheme with an initial window of 60, 80, and 100 observations.

Table 2.2: Panel DM statistics for one quarter ahead ($h = 1$) and one year (four quarters, $h = 4$) ahead shrinkage estimation forecasts of real output growth relative to fixed effects and individual estimators as benchmarks for the $T = 60, 80$ and 100.

Benchmark Models	$T = 60$	$T = 80$	$T = 100$
		$h = 1$	
Fixed Effects	-1.350^{*}	-1.432^{*}	-2.717^{***}
Individual Est.	-2.463^{***}	-3.821^{***}	-1.500^{*}
		$h = 4$	
Fixed Effects	-3.266^{***}	-2.927^{***}	-1.896^{**}
Individual Est.	-2.251^{**}	-1.843^{**}	-1.620^{*}

Note: The results represent a one sided test, thus the 1% (***), 5% (**) and 10% (*) critical values are -2.326, -1.645, and -1.282, respectively. A positive value of the panel DM statistic represents evidence against the shrinkage estimation forecasts.

Chapter 3

Using All Lags or One Lag as Instruments: an Averaging Estimator in Dynamic Panel Data Models

3.1 Introduction

Analysis of linear dynamic panel data models where the time dimension (T) is not negligible relative to the cross section dimension (N), has recently received large attentions in applied microeconomics as a result of increasing availability of micro-panels. Due to the endogeneity problem, the estimation of dynamic panels with individual effect is carried out predominantly by the Generalized Method of Moments (GMM)

after first differencing (FD) or forward-demeaning (FOD). Several GMM estimation methods have been proposed in the literature, including Anderson and Hsiao (1981) and Anderson and Hsiao (1982), Arellano and Bond (1991), Ahn and Schmidt (1995), Arellano and Bover (1995), Blundell and Bond (1998), and Hayakawa (2012), among others. One main reason for the popularity of the GMM estimation approach is that they may provide asymptotically efficient inference employing a relatively minimal set of statistical assumptions. However, despite its optimal asymptotic properties, the performance of GMM estimators can be poor, specially when T is large, due to abundance of moment conditions.

Because of the over-identification, an important practical issue in such models is how many moment conditions to use. In practice, as it is shown in the literature (see, e.g., Bekker (1994)), numerous instruments can overfit endogenous variables in finite samples, resulting in a trade-off between bias and efficiency. This has resulted in a substantial theoretical work on the overfitting bias of the GMM estimators in panel data models. Alvarez and Arellano (2003) analyze a panel autoregressive model of order one, and show that although GMM remains consistent for $T/N \to c$, so long as $0 \leq c \leq 2$, for $c > 0$ the estimator exhibits a bias in its asymptotic distribution that is of order $1/N$. Bun and Kiviet (2006) show that in comparison with the GMM estimators that employ all available instruments, reducing the set of instruments by order T decreases the bias by an order smaller in magnitude by a factor T. Hsiao and Zhou (2017) show the asymptotic properties of the GMM estimators that are based on FD or FOD can be different. They show that when all available instruments are used, the two differencing methods of the GMM

estimation methods are biased of order $\sqrt{c}$ as $(N,T) \to \infty$. However, if only a fixed number of instruments are used, the GMM based on FD remains asymptotically biased of order $\sqrt{c}$, while the GMM based on FOD is asymptotically unbiased even $c \neq 0$ as $(N,T) \to \infty$. Ziliak (1997) examines the bias/efficiency trade-off issue using bootstrap algorithms in an empirical application to life cycle labor supply under uncertainty. Ziliak (1997) shows that the downward bias in GMM becomes larger as the number of moment conditions expands, where the bias is due to the nonzero correlation between the sample moments used in estimation and the estimated weight matrix. Windmeijer (2005) in a monte carlo simulation reported that for the two step FD GMM, using only two lags of the dependent variable as instruments appeared to decrease the average bias by 40% relative to the estimator that made use of the full set of instruments, although the standard deviation of the estimator increased by about 7.5%. Roodman (2009) compared two popular approaches for limiting the number of instruments: (i) the use of (up to) certain lags instead of all available lags and (ii) combining instruments into smaller sets. His results show that the bias in system GMM based on the first approach is similar to the bias when using the full set of instruments. However, there is clear bias reduction under the second approach. This is while, Hayakawa (2009) shows that in panels with large unobserved heterogeneity the bias in FD GMM can actually be larger when using a smaller set of instruments.

This chapter contributes to the GMM literature by introducing the idea of model averaging and shrinkage estimation in selecting the number of moments. Essentially, we introduce an averaging estimator which is a weighted average of the GMM estimator using all available lags, and the GMM estimator using the most recent lag as instruments. The

weights are similar to a minimum mean squared error estimator weights, which measure the weighted distance of the two GMM estimators. We derive the first order approximate bias, mean squared error matrix (MSEM) and risk of the averaging estimator, and show its robustness and efficiency.

The reminder of this chapter is organized as follows. Section 3.2 describes the model and the assumptions. In section 3.3, we introduce and study properties of the estimators. We give the bias, mean squared error matrix, and the risk of the averaging estimator using asymptotic expansions in section 3.4. Monte Carlo results are given is section 3.5. Conclusions are given is section 3.6. Proofs and detailed calculations are listed in Appendix C.

Notations: Let A be a $n \times n$ symmetric matrix, $\text{tr}(A)$ denotes the trace of matrix A, $\varrho_{max}(A)$ and $\varrho_{min}(A)$ denote the maximum and minimum eigenvalues of matrix A.

3.2 The Model

Consider the following first-order linear dynamic panel data model with multiple regressors

$$y_{it} = \gamma y_{i,t-1} + x_{it}'\beta + u_{it}, \qquad i = 1, 2, \ldots, N,\ t = 1, 2, \ldots, T, \tag{3.1}$$

where x_{it} is a $(k-1) \times 1$ vector of observations on the regressors, β is a $(k-1) \times 1$ vector of unknown coefficients, and the disturbance term contains two error components, an unobserved individual specific effect η_i, and a general disturbance term ϵ_{it}, i.e. $u_{it} = \eta_i + \epsilon_{it}$.

We assume that the time-variant regressor x_{it} is correlated with η_i, and is strictly exogenous with respect to ϵ_{it}, i.e.

$$\mathbb{E}(x_{it}\epsilon_{js}) = 0, \qquad i, j = 1, \ldots, N, \text{ and } t, s = 1, \ldots, T. \tag{3.2}$$

We assume mutual independence of the cross-section units and serial independence of the disturbances, i.e. for $i = 1, 2, \ldots, N$, $t = 1, 2, \ldots, T$,

$$\eta_i \sim i.i.d.(0, \sigma_\eta^2), \tag{3.3}$$

$$\epsilon_{it} \sim i.i.d.N(0, \sigma_\epsilon^2), \tag{3.4}$$

We assume the two error components are uncorrelated and all N initial observations y_{i0} are uncorrelated with all disturbances for $t > 0$, i.e.

$$\mathbb{E}(\eta_i\epsilon_{jt}) = 0, \qquad \forall i, j, t, \tag{3.5}$$

$$\mathbb{E}(y_{i0}\epsilon_{jt}) = 0, \qquad \forall i, j, t > 0. \tag{3.6}$$

Furthermore, we suppose that the model is dynamically stable, that is we assume that for the model in equation (3.1), $|\gamma| < 1$.

As in Kiviet (1995), we decompose y_{it} and x_{it} into zero mean relevant random components, denoted by a tilde, and irrelevant random plus deterministic components, denoted by a bar. These expressions make fully explicit how all observations on y_{it} and x_{it} depend on both error components, which then makes it possible to obtain the approximation results. The relevant random components are those which are related to the individual effects, η_i, and the disturbance terms, ϵ_{it}. Hence, in the asymptotic approximations we

condition on $\bar{y}_{it} = y_{it} - \tilde{y}_{it}$, and $\bar{x}_{it} = x_{it} - \tilde{x}_{it}, \forall i, t$. Therefore, we decompose x_{it}, for $i = 1, \ldots, N$, $t = 1, \ldots, T$, as

$$x_{it} = \bar{x}_{it} + \tilde{x}_{it}, \tag{3.7}$$

$$\tilde{x}_{it} = \pi \eta_i, \tag{3.8}$$

where $\mathbb{E}(\bar{x}_{it}\eta_j) = 0$ and $\mathbb{E}(\bar{x}_{it}\epsilon_{js}) = 0$ for all $i, j = 1, \ldots, N$ and $t, s = 1, \ldots, T$, and the $(k-1) \times 1$ parameter π allows for the correlation between the regressors and the individual effects.

Regarding y_{it}, for the relevant random component, $\tilde{y}_{it}$, and the irrelevant component, $\bar{y}_{it}$, for $i = 1, 2, \ldots, N$, $t = 1, 2, \ldots, T$, we have

$$\tilde{y}_{it} = \gamma \tilde{y}_{i,t-1} + \tilde{x}'_{it}\beta + \eta_i + \epsilon_{it}, \tag{3.9}$$

$$\bar{y}_{it} = \gamma \bar{y}_{i,t-1} + \bar{x}'_{it}\beta. \tag{3.10}$$

In order to further decompose the relevant random components of $\tilde{y}_{it}$ into the two error components η_i, and ϵ_{it}, we need to make an assumption on the accumulated size of the individual effect in y_{i0}. For simplicity, we assume that

$$\mathbb{E}(\tilde{y}_{it}|\eta_i) = \alpha \eta_i, \qquad i = 1, 2, \ldots, N, \; t = 0, 1, \ldots, T, \tag{3.11}$$

where $\alpha = (1 + \pi'\beta)/(1 - \gamma)$. The above equation implies that the long-run impact of the individual effect on y_{it} is already present in y_{i0}, so we have mean-stationarity for y_{it}. Further, we assume

$$\tilde{y}_{i0} = \alpha \eta_i + \omega \epsilon_{i0}, \qquad i = 1, 2, \ldots, N, \tag{3.12}$$

where as Kiviet (1995) we choose $\omega = 0$, because when x_{it} is strongly exogenous, ϵ_{i0} is an irrelevant random component and should not be included in $\tilde{y}_{i0}$.

Stacking observations over time, for $i = 1, \ldots, N$, we get

$$y_i = \gamma y_{i(-1)} + X_i\beta + \eta_i \iota_T + \epsilon_i, \tag{3.13}$$

$$X_i = \overline{X}_i + \widetilde{X}_i = \overline{X}_i + \eta_i \iota_T \pi', \tag{3.14}$$

where $y_i = (y_{i1}, \ldots, y_{i,T})'$, $y_{i(-1)} = (y_{i0}, \ldots, y_{i,T-1})'$, $\epsilon_i = (\epsilon_{i1}, \ldots, \epsilon_{i,T})'$, $X_i = (x_{i1}, \ldots, x_{i,T})'$, $\overline{X}_i = (\bar{x}_{i1}, \ldots, \bar{x}_{i,T})'$, and $\iota_T = (1, \ldots, 1)'$, is a $T \times 1$ vector of ones. From the above it follows that, for $i = 1, \ldots, N$,

$$\tilde{y}_i = \gamma \tilde{y}_{i(-1)} + (\pi'\beta + 1)\eta_i \iota_T + \epsilon_i = \gamma(L_T \tilde{y}_i + \tilde{y}_{i0} e_{T,1}) + (\pi'\beta + 1)\eta_i \iota_T + \epsilon_i, \tag{3.15}$$

where L_T is a $T \times T$ matrix with ones on the first lower sub-diagonal and zeros elsewhere, and $e_{i,j}$ is an $i \times 1$ vector of zeros with jth element equal to one. Using equation (3.12), the relevant random part of y_i can be written as

$$\tilde{y}_i = \alpha \eta_i \iota_T + \Gamma_T \epsilon_i, \tag{3.16}$$

where $\Gamma_T = (I_T - \gamma L_T)^{-1}$, and the irrelevant part of y_i can be written as

$$\bar{y}_i = \Gamma_T \overline{X}_i \beta. \tag{3.17}$$

Stacking the T observations per individual over all N individuals yields

$$y = W\delta + u, \tag{3.18}$$

where $\delta = (\gamma, \beta')'$, y and u are $NT \times 1$, $W = (y_{(-1)}, X)$ is $NT \times k$, $u = S\eta + \epsilon$, with $S = I_N \otimes \iota_T$ and $\eta = (\eta_1, \ldots, \eta_N)'$. Therefore, the random part of y can be written as

$$\tilde{y} = \alpha S \eta + \Gamma \epsilon, \tag{3.19}$$

where $\Gamma = I_N \otimes \Gamma_T$, and the irrelevant part can be written as

$$\bar{y} = \Gamma \overline{X} \beta. \tag{3.20}$$

Also, we decompose W in a relevant and irrelevant components as below

$$\overline{W} = (\bar{y}_{(-1)}, \overline{X}) = (L\Gamma \overline{X}\beta, \overline{X}), \tag{3.21}$$

$$\widetilde{W} = (\tilde{y}_{(-1)}, \widetilde{X}) = S\eta a + L\Gamma \epsilon e'_{k,1}, \tag{3.22}$$

where $a = (\alpha, \pi')$ is $1 \times k$, and

$$\widetilde{X} = S\eta\pi'. \tag{3.23}$$

3.3 Estimation

The assumptions made on the stochastic structure of the model in the previous section contain a set of linear and non-linear moment conditions for each individual unit, see Ahn and Schmidt (1995). In this study we will focus on method of moments implementations using only linear moment conditions and we will not exploit any moment conditions associated with the homoscedasticity of ϵ_{it}. To eliminate the time invariant individual effects, we employ the forward demeaning (FOD) transformation method proposed by Arellano and Bover (1995). We define $P = I_N \otimes P_T$ to be the FOD transformation, where P_T is a $(T-1) \times T$ upper-triangular matrix with rank $T-1$ and $P_T \iota_T = 0$, which transforms y_{it} as

$$y^*_{it} = c_t \Big[y_{it} - \frac{1}{(T-t)} (y_{i,t+1} + \cdots + y_{iT}) \Big], \tag{3.24}$$

with $c_t^2 = (T-t)/(T-t+1)$. Since $P_T P_T' = I_{T-1}$, independence of ϵ_{it} is preserved in the transformed model which is the advantage of the FOD transformation method. Premultiplying model in equation (3.18) by P, will result in

$$Py = PW\delta + P\epsilon. \tag{3.25}$$

For the above model, we consider GMM estimations exploiting $m_1 = k(T-1) = O(T)$ and $m_2 = kT(T-1)/2 = O(T^2)$ moment conditions for each individual i. These can be expressed as $\mathbb{E}(Z'_{li} P_T \epsilon_i) = 0$, for $l = 1, 2$, where Z_{li} is a $(T-1) \times m_j$ instrument matrix with variables in levels, defined as

$$Z_{2i} = \begin{pmatrix} y_{i0} & x'_{i1} & 0 & 0 & \mathbf{0} & \mathbf{0} & 0 & \dots & 0 & \dots & 0 & \mathbf{0} & \dots & \mathbf{0} \\ 0 & \mathbf{0} & y_{i0} & y_{i1} & x'_{i1} & x'_{i2} & 0 & \dots & 0 & & & & & \mathbf{0} \\ 0 & \mathbf{0} & 0 & 0 & \mathbf{0} & \mathbf{0} & \dots & & \vdots & & & & & \vdots \\ \vdots & & & & & \vdots & & \dots & 0 & \dots & 0 & \mathbf{0} & \dots & \mathbf{0} \\ 0 & \mathbf{0} & 0 & 0 & \mathbf{0} & \mathbf{0} & \dots & \dots & y_{i0} & \dots & y_{i,T-2} & x'_{i1} & \dots & \boldsymbol{x}'_{i,T-1} \end{pmatrix}, \tag{3.26}$$

and

$$Z_{1i} = \begin{pmatrix} y_{i0} & x'_{i1} & 0 & \mathbf{0} & 0 & \dots & 0 & \mathbf{0} \\ 0 & \mathbf{0} & y_{i1} & x'_{i2} & 0 & & & \mathbf{0} \\ \vdots & \vdots & & & \dots & & & \vdots \\ 0 & \mathbf{0} & 0 & \mathbf{0} & & \dots & 0 & \mathbf{0} \\ 0 & \mathbf{0} & 0 & \mathbf{0} & \dots & \mathbf{0} & y_{i,T-2} & x'_{i,T-1} \end{pmatrix}. \tag{3.27}$$

Stacking over individuals the moment conditions can be written as

$$\mathbb{E}(Z_l' P\epsilon) = 0, \qquad l = 1, 2, \tag{3.28}$$

where $Z_l = (Z'_{l1}, \ldots, Z'_{lN})'$ is $N(T-1) \times m_l$. Provided $P'Z_l$ has full column rank, the Arellano and Bond (1991) type GMM is to find $\widehat{\delta}$ that minimizes the quadratic form

$$(\frac{1}{N}\sum_{i=1}^{N} Z'_{li}P_T\epsilon_i)'(\frac{1}{N^2}\sum_{i=1}^{N} Z'_{li}P_T\epsilon_i\epsilon'_iP'_TZ_{li})(\frac{1}{N}\sum_{i=1}^{N} Z'_{li}P_T\epsilon_i), \quad l = 1, 2. \tag{3.29}$$

3.3.1 GMM Estimator Using All Lags as Instruments

The solution to the minimization in (3.29) when all lags are used as instruments (instruments matrix Z_2), yields to an optimal one-step GMM estimator which can be formulated as

$$\widehat{\delta}_{GMM,2} = \left(W'P'M_2PW\right)^{-1}W'P'M_2Py = \delta + \left(W'P'M_2PW\right)^{-1}W'P'M_2P\epsilon, \tag{3.30}$$

where $M_2 = Z_2(Z'_2Z_2)^{-1}Z'_2$ is an $N(T-1) \times N(T-1)$ idempotent matrix.

Theorem 3.1 *The bias of the GMM estimator using all lags as instruments up to order* $O(\frac{1}{\sqrt{NT}})$ *is*

$$Bias(\widehat{\delta}_{GMM,2}) = \mathbb{E}(\widehat{\delta}_{GMM,2} - \delta) = \sigma^2_\epsilon \, tr(H_2)Q_2e_{k,1} = O(\frac{1}{N}), \tag{3.31}$$

and the MSE matrix of the estimator up to order $O(\frac{1}{NT})$ *is*

$$MSE\ (\widehat{\delta}_{GMM,2}) = \sigma^4_\epsilon \, tr(H_2)^2Q_2e_{k,1}e'_{k,1}Q_2 + \sigma^2_\epsilon Q_2, \tag{3.32}$$

where $Q_2 = \left[\overline{W}'P'M_2P\overline{W} + \sigma^2_\epsilon e_{k,1}e'_{k,1}\, tr(H_2)\right]^{-1} = O(1/NT)$, $H_2 = P'M_2PL\Gamma$, *and* $tr(H_2) = -Tk/(1-\gamma) + O(1)$.

Proof: Appendix C,(See page 161).

3.3.2 GMM Estimator Using One Lag as Instruments

When Z_1, which includes a subset of Z_2, is used in the minimization in (3.29), the optimal one-step GMM estimator can be formulated as

$$\widehat{\delta}_{GMM,1} = \left(W'P'M_1PW\right)^{-1}W'P'M_1Py = \delta + \left(W'P'M_1PW\right)^{-1}W'P'M_1P\epsilon, \tag{3.33}$$

where $M_1 = Z_1(Z_1'Z_1)^{-1}Z_1'$ is an $N(T-1) \times N(T-1)$ idempotent matrix.

Theorem 3.2 *The bias of the GMM estimator using one lag as instruments up to order* $O(\frac{1}{\sqrt{NT}})$ *is*

$$Bias(\widehat{\delta}_{GMM,1}) = \mathbb{E}(\widehat{\delta}_{GMM,1} - \delta) = 0, \tag{3.34}$$

and the MSE matrix of the estimator up to order $O(\frac{1}{NT})$ *is*

$$MSE\ (\widehat{\delta}_{GMM,1}) = Var(\widehat{\delta}_{GMM,1}) = \sigma_\epsilon^2 Q_1, \tag{3.35}$$

where $Q_1 = \left[\overline{W}'P'M_1P\overline{W} + \sigma_\epsilon^2 e_{k,1}e_{k,1}'\, tr(H_1)\right]^{-1} = O(1/NT)$, *and* $H_1 = P'M_1PL\Gamma$.

Proof: Appendix C,(See page 162).

3.3.3 Averaging Estimator

To make an appropriate trade-off between bias due to many instruments and variance efficiency resulting from exploiting all of the moment conditions, we introduce an averaging estimator which is a weighted average of the GMM estimators introduced in the previous sections. Our averaging estimator of δ is

$$\widehat{\delta}_A = \frac{\tau}{F}\widehat{\delta}_{GMM,2} + (1 - \frac{\tau}{F})\widehat{\delta}_{GMM,1}, \tag{3.36}$$

where

$$F = \Big(\widehat{\delta}_{GMM,1} - \widehat{\delta}_{GMM,2}\Big)' \Big(\widehat{V}_1 - \widehat{V}_2\Big)^{-1} \Big(\widehat{\delta}_{GMM,1} - \widehat{\delta}_{GMM,2}\Big), \tag{3.37}$$

and $\widehat{V}_1$ and $\widehat{V}_2$ denote the conventional covariance matrix estimators for $\widehat{\delta}_{GMM,1}$ and $\widehat{\delta}_{GMM,2}$, which are defined as below

$$\widehat{V}_1 = \frac{1}{\hat{\sigma}^2_{\epsilon,1}} \Big(W'P'M_1PW\Big)^{-1} \tag{3.38}$$

$$\widehat{V}_2 = \frac{1}{\hat{\sigma}^2_{\epsilon,2}} \Big(W'P'M_2PW\Big)^{-1}, \tag{3.39}$$

where

$$\hat{\sigma}^2_{\epsilon,l} = \frac{1}{N(T-1)-k} (Py - PW\widehat{\delta}_{GMM,1})'(Py - PW\widehat{\delta}_{GMM,l}), \quad l = 1, 2. \tag{3.40}$$

In equation (3.36), τ is a positive parameter which measure the degree of significance. We will defer describing our recommended optimal choice for this parameter in the following sections. Since it can be easily verified that the risk of the estimator with positive part weights is smaller, alternatively, one could replace the weights by their positive part, i.e. $(1 - \tau/F)_+ = (1 - \tau/F)\, \mathbb{1}((1 - \tau/F) \geq 0)$, where $\mathbb{1}(.)$ denotes an indicator function. However, it will not affect the derivations of the approximations below, so for simplicity we do not impose it at this stage. Nevertheless, the Monte Carlo results are reported using the positive part weight.

Notice that when the difference between the two GMM estimators is small (F is small), the bias of the GMM using all lag instruments is relatively small, therefore the averaging estimator gives a large weight to this estimator, as it is the most efficient estimator. However, when the difference between the two GMM estimators is substantial or high ($F > \tau$), the

bias of the GMM estimator using all lags could be more than its variance efficiency gain, so the averaging estimator is a weighted average of the two GMM estimators, with more weight on the GMM estimator using one lag as instrument.

3.4 Finite Sample Approximation

In this section, we obtain the approximation bias, MSE matrix and risk of the averaging estimator.

Theorem 3.3 *The bias of the averaging estimator up to order* $O(\frac{1}{\sqrt{NT}})$ *is*

$$Bias\ (\widehat{\delta}_A) = \mathbb{E}(\widehat{\delta}_A - \delta) = \sigma_\epsilon^2 \frac{\tau}{k} e^{-\lambda/2}\ tr(H_2) Q_2 e_{k,1}\ {}_1F_1\Big(\frac{k}{2}\,,\ \frac{k}{2}+1;\ \lambda/2\Big), \tag{3.41}$$

and the MSE matrix of the averaging estimator up to order $O(\frac{1}{NT})$*, given* $k > 2$ *is*

$$\begin{aligned} MSE\ (\widehat{\delta}_A) = MSE\ (\widehat{\delta}_{GMM,1}) &+ \frac{\tau}{k}\ e^{-\lambda/2}(V_1 - V_2) \\ &\left[\frac{\tau}{k-2}\ {}_1F_1\Big(\frac{k}{2}-1\,,\ \frac{k}{2}+1\,;\ \lambda/2\Big) - 2\ {}_1F_1\Big(\frac{k}{2}\,,\ \frac{k}{2}+1\,;\ \lambda/2\Big)\right] \\ &+ \sigma_\epsilon^4 \tau\ e^{-\lambda/2}\ \ tr(H_2)^2 Q_2 e_{k,1} e_{k,1}' Q_2 \left[\frac{\tau}{k(k+2)}\ {}_1F_1\Big(\frac{k}{2}\,,\ \frac{k}{2}+2\,;\ \lambda/2\Big)\right. \\ &\left. - 2\Big[\frac{1}{k+2}\ {}_1F_1\Big(\frac{k}{2}+1\,,\ \frac{k}{2}+2\,;\ \lambda/2\Big) - \frac{1}{k}\ {}_1F_1\Big(\frac{k}{2}\,,\ \frac{k}{2}+1\,;\ \lambda/2\Big)\Big]\right], \end{aligned} \tag{3.42}$$

where $\lambda = \sigma_\epsilon^4\ tr(H_2)^2 e_{k,1}' Q_2 (V_1 - V_2)^{-1} Q_2 e_{k,1}$.

Proof: Appendix C,(See page 163).

Remark 3.4 *The asymptotic MSEM of the averaging estimator in (3.42) can be rewritten as follows* [1]

$$
\begin{aligned}
&MSE\,(\widehat{\delta}_A) = MSE\,(\widehat{\delta}_{GMM,1}) \\
&+ \frac{1}{k(k-2)}\, e^{-\lambda/2}\, {}_1F_1\Big(\frac{k}{2}-1\,,\,\frac{k}{2}+1\,;\,\lambda/2\Big)\Big[\tau^2 - 2\tau(k-2)\Big](V_1 - V_2) \\
&+ \frac{1}{k(k+2)}\, e^{-\lambda/2}\, {}_1F_1\Big(\frac{k}{2}\,,\,\frac{k}{2}+2\,;\,\lambda/2\Big)\Bigg[\Big(\tau^2+4\tau\Big)\sigma_\epsilon^4\ tr(H_2)^{\,2} Q_2 e_{k,1} e'_{k,1} Q_2 \\
&- 4\,\tau\,\lambda\,(V_1 - V_2)\Bigg].
\end{aligned}
\tag{3.43}
$$

In the following corollary we give our recommended value of τ that minimizes the risk of the averaging estimator.

Corollary 3.5 *When* $tr(C)/\varrho_{max}(C) > 2$, *and* $\quad 0 < \tau \leq 2\Big[\,tr(C)/\varrho_{max}(C) - 2\Big]$, *then the risk of the averaging estimator, for a positive definite weight matrix* D *whose elements are of order* $O(1)$, *is*

$$
\begin{aligned}
Risk\,(\widehat{\delta}_A) \leq Risk\,(\widehat{\delta}_{GMM,1}) - e^{-\lambda/2}\ \frac{1}{k}\ \Big[2\tau\Big(\frac{tr(C)}{\varrho_{max}(C)} - 2\Big) - \tau^2\Big] \\
\Big[\frac{tr(C)}{k-2}\, {}_1F_1\Big(\frac{k}{2}-1\,,\,\frac{k}{2}+1\,;\,\lambda/2\Big) + \frac{2\lambda_D}{k+2}\, {}_1F_1\Big(\frac{k}{2}\,,\,\frac{k}{2}+2\,;\,\lambda/2\Big)\Big],
\end{aligned}
\tag{3.44}
$$

where $Risk\,(\widehat{\delta}_{GMM,1}) = tr(D\,V_1)$, $C = (V_1 - V_2)^{1/2} D (V_1 - V_2)^{1/2}$, *and* $\lambda_D = \sigma_\epsilon^4\, tr(H_2)^2 e'_{k,1} Q_2 D Q_2 e_{k,1}$. *The above result shows the superiority of the averaging estimator relative to the GMM estimator using one lag. The optimal* τ *that minimizes the risk is*

$$
\tau_{opt} = tr(C)/\varrho_{max}(C) - 2. \tag{3.45}
$$

[1] The result holds by using the following identities

$$(c-a-1)\, {}_1F_1(a,c;x) = (c-1)\, {}_1F_1(a,c-1;x) - a\, {}_1F_1(a+1,c;x),$$

$${}_1F_1(a,c;x) = {}_1F_1(a+1,c;x) - \frac{x}{c}\, {}_1F_1(a+1,c+1;x),$$

See Lebedev (1972), pp. 271.

Proof: Appendix C,(See page 166).

Corollary 3.6 *When* $D = (V_1 - V_2)^{-1}$, $k > 2$, *and* $0 < \tau \leq 2[k-2]$, *the risk of the averaging estimator is*

$$Risk\ (\widehat{\delta}_A) = Risk\ (\widehat{\delta}_{GMM,1}) - e^{-\lambda/2}\ \frac{1}{k-2}\ {}_1F_1\Big(\frac{k}{2} - 1\,,\ \frac{k}{2}\,;\ \lambda/2\Big)\Big[2\tau(k-2) - \tau^2\Big]. \tag{3.46}$$

The value of τ *that minimizes the risk of the averaging estimator is*

$$\tau_{opt} = k - 2, \tag{3.47}$$

and the risk of the optimal averaging estimator is

$$Risk\ (\widehat{\delta}_{A,opt}) = Risk\ (\widehat{\delta}_{GMM,1}) -\ e^{-\lambda/2}\ (k-2)\ {}_1F_1\Big(\frac{k}{2} - 1\,,\ \frac{k}{2}\,;\ \lambda/2\Big). \tag{3.48}$$

Proof: Appendix C,(See page 167).

Corollary 3.7 *When* $k > 2$, *if* $\lambda \to \infty$ [2] *then*

$$Risk\ (\widehat{\delta}_{A,opt}) = Risk\ (\widehat{\delta}_{GMM,1}) + O(\frac{1}{\lambda}). \tag{3.49}$$

Proof: Appendix C,(See page 167).

The result in corollary 3.7 suggests that if the bias of the GMM estimator using all lags is very large, the averaging estimator is approximately very close to the GMM estimator using one lag as instrument. This condition assures that even for dynamic panels with λ close to one, the averaging estimator remains asymptotically consistent and efficient by giving a weight one to the GMM estimator using one lag as instrument.

[2] Equivalently when $\gamma \to 1$.

3.5 Monte Carlo Simulation

In this section, we investigate the finite sample properties of the averaging estimator of δ for dynamic panel data models. Data for the dependent variable y is generated according to equation (3.1), where the explanatory variables, x_{it} for $i = 1, \ldots, N$, and $t = 1, \ldots, T$, are generated as below

$$x_{it,j} = \bar{x}_{it,j} + \pi\eta_i,$$

$$\bar{x}_{it,j} = \rho\bar{x}_{i,t-1,j} + \xi_{it,j},$$

for $j = 1, \ldots, k$, where $\xi_{it,j} \sim i.i.d.N(0, \sigma_\xi^2)$ independent from $\epsilon_{it} \sim i.i.d.N(0, \sigma_\epsilon^2)$ with $\sigma_\epsilon = 1$, and these two are independent of $\eta_i \sim i.i.d.N(0, \sigma_\eta^2)$, also we set $|\rho| < 1$ to make $\bar{x}_{it}$ a stationary AR(1) process.

We set $\gamma = \{0.25, 0.75\}$, and choose $\beta = (1-\gamma)\iota_k$, so that the long-run effect of X on y is equal to a unit vector. Further, we choose $\rho = \{0, 0.5\}$, which yield stationary regressors, and $\pi = \{-\iota_k, \mathbf{0}\}$.

Similar to Bun and Kiviet (2006), we set

$$\sigma_\eta^2 = \mu^2 \frac{1-\gamma}{(1+\gamma)(1+\pi'\beta)^2}, \tag{3.50}$$

so that the impact of the variances of η_i and ϵ_{it} on $Var(y_{it})$ has a ratio of μ^2, which we set to be from $\{0, 1\}$. The parameter σ_ξ^2 is determined by controlling the signal-to-noise ratio (ϑ) of the model, and we choose $\vartheta = \{3, 9\}$. As it has been shown in Appendix B of Bun and Kiviet (2006), this requires

$$\sigma_\xi^2 = \frac{1}{\beta'\beta}\Big[\vartheta - \frac{\gamma^2}{1-\gamma^2}\Big]\frac{(1-\gamma^2)(1-\rho^2)(1-\gamma\rho)}{1+\gamma\rho}. \tag{3.51}$$

We choose $k = 3$, $T = 20$, and to allow different convergence rates between T and N, we consider $kT/N = \{0.25, 0.5, 0.75, 1\}$. It should be noted that, $kT/N \leq 2$, so the GMM estimators are identified.

The results of $1,000$ monte carlo simulations are given in Tables 3.1–3.4, where the values are relative mean squared error (RMSE) of the GMM using one lag estimator, the GMM using all lags estimator, and the optimal averaging estimator, to the mean squared error of the GMM using one lag estimator. The tables consist of 32 designs which represent different specifications.

The Monte Carlo results support our theoretical findings of the previous section. The table results show that the RMSE of the averaging estimator for the whole parameter specification is less than that of the GMM using one lag estimator. This shows the superiority of our proposed estimator relative to the alternative estimators.

In designs where $\mu = 1$, $\gamma = 0.75$, and $\pi = 0$, it implies $\sigma_\eta^2 = \sigma_\epsilon^2/7$. We see larger RMSE's when x_{it} is smoother. However, as γ increases, $\hat{\delta}_{GMM,1}$ is more efficient than $\hat{\delta}_{GMM,2}$ and hence the averaging estimator is dominant which is in agreement with our theoretical findings. As expected, the RMSE of $\hat{\delta}_{GMM,2}$ increases substantially when the ratio of kT/N increases or equivalently when N decreases.

In general, we find that the averaging estimator performs robustly well in dynamic panel data models with various degrees of specification. When there is a large number of cross-sections relative to the number of observations, the averaging estimator prevails. When there is a relatively small difference between these two, the averaging estimator tends

to gain more from the efficiency of the GMM using all lags estimator by assigning a high weight to this estimator, and thus still remains one of the best choices.

3.6 Conclusion

We introduce a new method of estimation and forecasting in dynamic panel data models when both the time dimension and cross-section dimension are large. We apply the idea of shrinkage estimation to the estimation of dynamic panels with individual effects, a lagged dependent variable, and multiple exogenous regressors. The proposed (averaging) estimator balances the trade-off between the bias and variance efficiency of GMM estimators using all of the moments and GMM estimators using one (a few) moments. The idea of averaging in dynamic panels opens new exciting research avenues. This idea can be considered in other setting, including dynamic panels with weak exogenous regressors, models that allow for cross-section correlation, and spatial panel data models. Last but not least, we have left the topic of constructing confidence intervals to future research.

Table 3.1: Relative MSE of GMM estimator using one lag instrument ($\widehat{\delta}_{GMM,1}$), GMM estimator by instrumenting all lags ($\widehat{\delta}_{GMM,2}$), and the averaging estimator ($\widehat{\delta}_A$), for $T = 20$, $\gamma = 0.25$, $\pi = -\iota_k$

Design	$c = \frac{kT}{N}$	$\widehat{\delta}_{GMM,1}$	$\widehat{\delta}_{GMM,2}$	$\widehat{\delta}_A$
1		$\gamma = 0.25$, $\pi = -\iota_k$, $\rho = 0$, $\vartheta = 3$, $\mu = 0$		
	0.25	1.000	1.034	0.998
	0.5	1.000	1.067	0.996
	0.75	1.000	1.086	0.993
	1	1.000	1.111	0.993
2		$\gamma = 0.25$, $\pi = -\iota_k$, $\rho = 0$, $\vartheta = 3$, $\mu = 1$		
	0.25	1.000	0.640	0.899
	0.5	1.000	0.657	0.916
	0.75	1.000	0.711	0.927
	1	1.000	0.713	0.924
3		$\gamma = 0.25$, $\pi = -\iota_k$, $\rho = 0$, $\vartheta = 9$, $\mu = 0$		
	0.25	1.000	0.984	0.995
	0.5	1.000	0.984	0.989
	0.75	1.000	0.961	0.984
	1	1.000	0.985	0.987
4		$\gamma = 0.25$, $\pi = -\iota_k$, $\rho = 0$, $\vartheta = 9$, $\mu = 1$		
	0.25	1.000	0.795	0.927
	0.5	1.000	0.838	0.940
	0.75	1.000	0.823	0.938
	1	1.000	0.807	0.934
5		$\gamma = 0.25$, $\pi = -\iota_k$, $\rho = 0.5$, $\vartheta = 3$, $\mu = 0$		
	0.25	1.000	1.018	0.996
	0.5	1.000	1.065	0.995
	0.75	1.000	1.066	0.991
	1	1.000	1.071	0.987
6		$\gamma = 0.25$, $\pi = -\iota_k$, $\rho = 0.5$, $\vartheta = 3$, $\mu = 1$		
	0.25	1.000	0.619	0.888
	0.5	1.000	0.652	0.903
	0.75	1.000	0.675	0.904
	1	1.000	0.692	0.915
7		$\gamma = 0.25$, $\pi = -\iota_k$, $\rho = 0.5$, $\vartheta = 9$, $\mu = 0$		
	0.25	1.000	0.980	0.993
	0.5	1.000	0.963	0.988
	0.75	1.000	0.980	0.984
	1	1.000	0.968	0.980
8		$\gamma = 0.25$, $\pi = -\iota_k$, $\rho = 0.5$, $\vartheta = 9$, $\mu = 1$		
	0.25	1.000	0.791	0.920
	0.5	1.000	0.829	0.928
	0.75	1.000	0.771	0.917
	1	1.000	0.802	0.926

Table 3.2: Relative MSE of GMM estimator using one lag instrument ($\widehat{\delta}_{GMM,1}$), GMM estimator by instrumenting all lags ($\widehat{\delta}_{GMM,2}$), and the averaging estimator ($\widehat{\delta}_A$), for $T = 20$, $\gamma = 0.75$, $\pi = -\iota_k$

Design	$c = \frac{kT}{N}$	$\widehat{\delta}_{GMM,1}$	$\widehat{\delta}_{GMM,2}$	$\widehat{\delta}_A$
9		$\gamma = 0.75$, $\pi = -\iota_k$, $\rho = 0$, $\vartheta = 3$, $\mu = 0$		
	0.25	1.000	1.160	1.000
	0.5	1.000	1.237	1.000
	0.75	1.000	1.293	0.998
	1	1.000	1.310	0.998
10		$\gamma = 0.75$, $\pi = -\iota_k$, $\rho = 0$, $\vartheta = 3$, $\mu = 1$		
	0.25	1.000	0.282	0.952
	0.5	1.000	0.367	0.951
	0.75	1.000	0.405	0.945
	1	1.000	0.448	0.941
11		$\gamma = 0.75$, $\pi = -\iota_k$, $\rho = 0$, $\vartheta = 9$, $\mu = 0$		
	0.25	1.000	1.656	1.000
	0.5	1.000	2.010	1.000
	0.75	1.000	2.241	1.000
	1	1.000	2.416	1.000
12		$\gamma = 0.75$, $\pi = -\iota_k$, $\rho = 0$, $\vartheta = 9$, $\mu = 1$		
	0.25	1.000	0.884	0.968
	0.5	1.000	1.037	0.969
	0.75	1.000	1.255	0.971
	1	1.000	1.316	0.972
13		$\gamma = 0.75$, $\pi = -\iota_k$, $\rho = 0.5$, $\vartheta = 3$, $\mu = 0$		
	0.25	1.000	1.054	0.999
	0.5	1.000	1.055	0.997
	0.75	1.000	1.081	0.995
	1	1.000	1.036	0.992
14		$\gamma = 0.75$, $\pi = -\iota_k$, $\rho = 0.5$, $\vartheta = 3$, $\mu = 1$		
	0.25	1.000	0.210	0.948
	0.5	1.000	0.242	0.945
	0.75	1.000	0.285	0.943
	1	1.000	0.283	0.938
15		$\gamma = 0.75$, $\pi = -\iota_k$, $\rho = 0.5$, $\vartheta = 9$, $\mu = 0$		
	0.25	1.000	1.251	1.000
	0.5	1.000	1.386	1.000
	0.75	1.000	1.463	0.999
	1	1.000	1.551	1.000
16		$\gamma = 0.75$, $\pi = -\iota_k$, $\rho = 0.5$, $\vartheta = 9$, $\mu = 1$		
	0.25	1.000	0.437	0.953
	0.5	1.000	0.501	0.952
	0.75	1.000	0.567	0.947
	1	1.000	0.636	0.946

Table 3.3: Relative MSE of GMM estimator using one lag instrument ($\widehat{\delta}_{GMM,1}$), GMM estimator by instrumenting all lags ($\widehat{\delta}_{GMM,2}$), and the averaging estimator ($\widehat{\delta}_{A}$), for $T = 20$, $\gamma = 0.25$, $\pi = 0$

Design	$c = \frac{kT}{N}$	$\widehat{\delta}_{GMM,1}$	$\widehat{\delta}_{GMM,2}$	$\widehat{\delta}_{A}$
17		$\gamma = 0.25$, $\pi = 0$, $\rho = 0$, $\vartheta = 3$, $\mu = 0$		
	0.25	1.000	1.023	0.998
	0.5	1.000	1.074	0.997
	0.75	1.000	1.109	0.995
	1	1.000	1.105	0.992
18		$\gamma = 0.25$, $\pi = 0$, $\rho = 0$, $\vartheta = 3$, $\mu = 1$		
	0.25	1.000	0.956	0.974
	0.5	1.000	0.996	0.981
	0.75	1.000	1.009	0.974
	1	1.000	1.029	0.984
19		$\gamma = 0.25$, $\pi = 0$, $\rho = 0$, $\vartheta = 9$, $\mu = 0$		
	0.25	1.000	1.003	0.996
	0.5	1.000	1.011	0.995
	0.75	1.000	0.995	0.989
	1	1.000	0.996	0.988
20		$\gamma = 0.25$, $\pi = 0$, $\rho = 0$, $\vartheta = 9$, $\mu = 1$		
	0.25	1.000	0.972	0.985
	0.5	1.000	0.971	0.982
	0.75	1.000	0.971	0.977
	1	1.000	0.989	0.977
21		$\gamma = 0.25$, $\pi = 0$, $\rho = 0.5$, $\vartheta = 3$, $\mu = 0$		
	0.25	1.000	1.023	0.997
	0.5	1.000	1.051	0.995
	0.75	1.000	1.074	0.991
	1	1.000	1.092	0.990
22		$\gamma = 0.25$, $\pi = 0$, $\rho = 0.5$, $\vartheta = 3$, $\mu = 1$		
	0.25	1.000	0.961	0.972
	0.5	1.000	0.999	0.974
	0.75	1.000	1.011	0.969
	1	1.000	1.048	0.980
23		$\gamma = 0.25$, $\pi = 0$, $\rho = 0.5$, $\vartheta = 9$, $\mu = 0$		
	0.25	1.000	0.988	0.994
	0.5	1.000	0.987	0.990
	0.75	1.000	0.982	0.986
	1	1.000	0.997	0.980
24		$\gamma = 0.25$, $\pi = 0$, $\rho = 0.5$, $\vartheta = 9$, $\mu = 1$		
	0.25	1.000	0.949	0.977
	0.5	1.000	0.960	0.975
	0.75	1.000	0.985	0.976
	1	1.000	0.937	0.963

Table 3.4: Relative MSE of GMM estimator using one lag instrument ($\widehat{\delta}_{GMM,1}$), GMM estimator by instrumenting all lags ($\widehat{\delta}_{GMM,2}$), and the averaging estimator ($\widehat{\delta}_{A}$), for $T = 20$, $\gamma = 0.75$, $\pi = 0$

Design	$c = \frac{kT}{N}$	$\widehat{\delta}_{GMM,1}$	$\widehat{\delta}_{GMM,2}$	$\widehat{\delta}_{A}$
25		$\gamma = 0.75$, $\pi = 0$, $\rho = 0$, $\vartheta = 3$, $\mu = 0$		
	0.25	1.000	1.166	1.000
	0.5	1.000	1.226	0.999
	0.75	1.000	1.299	0.999
	1	1.000	1.332	0.998
26		$\gamma = 0.75$, $\pi = 0$, $\rho = 0$, $\vartheta = 3$, $\mu = 1$		
	0.25	1.000	1.096	1.000
	0.5	1.000	1.191	0.999
	0.75	1.000	1.263	0.998
	1	1.000	1.243	0.998
27		$\gamma = 0.75$, $\pi = 0$, $\rho = 0$, $\vartheta = 9$, $\mu = 0$		
	0.25	1.000	1.654	1.000
	0.5	1.000	2.000	1.000
	0.75	1.000	2.303	1.000
	1	1.000	2.320	1.000
28		$\gamma = 0.75$, $\pi = 0$, $\rho = 0$, $\vartheta = 9$, $\mu = 1$		
	0.25	1.000	1.414	1.000
	0.5	1.000	1.693	1.000
	0.75	1.000	1.858	1.000
	1	1.000	1.884	1.000
29		$\gamma = 0.75$, $\pi = 0$, $\rho = 0.5$, $\vartheta = 3$, $\mu = 0$		
	0.25	1.000	1.045	0.998
	0.5	1.000	1.085	0.997
	0.75	1.000	1.071	0.995
	1	1.000	1.043	0.992
30		$\gamma = 0.75$, $\pi = 0$, $\rho = 0.5$, $\vartheta = 3$, $\mu = 1$		
	0.25	1.000	1.028	0.994
	0.5	1.000	1.029	0.988
	0.75	1.000	1.045	0.989
	1	1.000	1.008	0.977
31		$\gamma = 0.75$, $\pi = 0$, $\rho = 0.5$, $\vartheta = 9$, $\mu = 0$		
	0.25	1.000	1.235	1.000
	0.5	1.000	1.362	0.999
	0.75	1.000	1.476	1.000
	1	1.000	1.530	1.000
32		$\gamma = 0.75$, $\pi = 0$, $\rho = 0.5, \vartheta = 9$, $\mu = 1$		
	0.25	1.000	1.167	1.000
	0.5	1.000	1.280	0.998
	0.75	1.000	1.352	0.999
	1	1.000	1.356	1.001

Chapter 4

A Modified Stein-Like Estimator for Coefficients of A Single-Equation In Simultaneous Equations

4.1 Introduction

Simultaneous equations models (SEM) which arise from economic theory in terms of operations of markets and the simultaneous determination of economic variables through an equilibrium model, are one of the many developments in econometrics. The study of estimating coefficients of a single equation in a complete system of simultaneous structural equations has provided many estimation methods, including the ordinary least

squares (OLS), the two-stage least squares (2SLS) and the limited information maximum likelihood (LIML) which are the most commonly used ones. Because of the presence of endogeneity in the model, the OLS estimator is biased and inconsistent, however, the 2SLS and LIML estimators under appropriate general conditions are consistent (see e.g. Anderson and Rubin (1949)). Since these estimators are available, numerous articles have focused on the finite-sample properties of these estimators and their modifications.

One direction of modifying these estimators in the hope that the modified estimation method may improve the existing estimators, have been made by linearly combining these estimators. Sawa (1973 a) and Sawa (1973 b) proposed a combined estimator, which is a simple linear combination of the OLS and 2SLS estimators, to eliminate the bias of the 2SLS estimator. The coefficients of this combined estimator depends on the sample size and the numbers of included and excluded variables from the relevant equation, and the estimator is unbiased to a certain order. Similarly, Morimune (1978) proposed a set of combined estimators which are convex linear combinations of the LIML estimator and fixed $k-$class estimators of Theil (1961). The aim of this method is eliminating the small-disturbance asymptotic bias of the LIML estimator to construct improved estimators which are unbiased to a certain order. Morimune showed the inadmissibility of the LIML estimator in terms of asymptotic mean squared error, in other words, he showed that the combined estimators dominate LIML (See also, Morimune and Kunitomo (1980) for the same method in the problem of functional relationships). A comparison of the above modified estimators has been given by Anderson et al. (1986).

Another type of modified estimators considers a nonlinear function of the estimators. Stein (1956) is the pioneer of this method. Stein showed that the maximum likelihood estimator (MLE) for the mean of a multivariate normal distribution does not have the smallest risk, in other words, MLE is inadmissible. Later on this issue, James and Stein (1961) suggested a biased estimator which dominates the MLE estimator in the sense that its risk is smaller than that of the former, provided at least three parameters are to be estimated. In the context of a single equation estimation in a linear simultaneous equations system, Zellner and Vandaele (1975) considered Stein-type estimators under a general quadratic loss function and obtained the minimum risk estimator by applying 2SLS method. However, the resulting estimator is unavailable in applications as it involves certain unknown parameters. To face this issue, Ullah and Srivastava (1988) present a Stein-type estimator and analyze its properties and conditions under which the resulting estimator dominates the 2SLS estimator.

In reduced form estimation, Maasoumi (1978) constructed a modified Stein-like estimator which is the weighted average of Least Squares (LS) and Three-Stage-Least-Squares (3SLS) of the reduced form coefficients in a linear simultaneous equations system, in which the weight depends on the inverse of an over-identification test statistic. Maasoumi shows that this estimator has a few advantages over the LS and 3SLS estimators as it has finite moments, thinner tails, and has the edge on the LS estimator as it is asymptotically equivalent to the 3SLS estimator. Following Maasoumi (1978), in the context of single equation instrumental variable models, Hansen (2017) constructs a Stein-like estimator which is a weighted average of the OLS and 2SLS estimators for estimating the structural

coefficients of the model. The weight is defined similar to Maasoumi (1978), while the Wu-Hausman (1978) specification test statistic is used. Using asymptotic theory, Hansen shows that the asymptotic risk of the combined estimator is strictly less than that of 2SLS estimator when the number of included endogenous variables are more than 2.

For the purpose of comparing the estimators and the modified ones, there are several ways in the literature. One approach is to derive the exact distributions of the estimators (see e.g. Anderson and Sawa (1979) and Phillips (1984)). However, the analytical expressions of the distributions are usually too complicated to permit meaningful general conclusion. An alternative approach is to approximate each distribution by one or more terms in an asymptotic expansion of the distribution. One term, most of the time, is not enough as it is the common term between several estimators, but three terms serve to distinguish between the estimators (See e.g. Rothenberg (1984)).

The asymptotic expansions have been derived on the basis of limits as an index tends towards a value. In the large-sample theory, the number of observations increase without bound. In this context, Nagar (1959) noted that $k-$class estimators in simultaneous equations models can be expanded in formal series where the successive terms are increasing powers of $T^{1/2}$, where T is the number of observation for each dependent variable. Nagar calculated the moments of the truncated series by keeping the first few terms in the expansion. These moments can be interpreted as the moments of a statistic which serves to approximate the estimator, while Sargan (1974) showed that, under some conditions, these moments can be interpreted as approximations to the actual moments of the estimator. In the small-disturbance theory, initiated by Kadane (1971), it is suggested that it might

be more natural to consider a sequence indexed by the error variance. In this analysis, the reduced-form error-covariance matrix is written as $\sigma\Omega$, and while the sample size, and the matrix Ω are held fixed, σ approaches zero. The large-sample and small-disturbance theories can be related by the effect of them on the non-centrality parameter, which goes to infinity in both cases while in the small-disturbance theory, the sample size stays constant, however in the large-sample theory, the sample size and the non-centrality parameter both go to infinity at the same speed (Anderson (1977)).

In this chapter, we propose two Stein-like estimators for coefficients of a single equation in a complete system of simultaneous equations. The estimators are weighted averages of the OLS and 2SLS (or LIML) estimators where the weights are inspired by the weights in Hansen (2017). We study the bias and mean squared error (MSE) of the estimator using small-disturbance theory of Kadane (1971) and show the conditions under which the Stein-like estimators dominate the 2SLS and LIML estimators.

There are two related papers in the literature that similar to this chapter consider combining the 2SLS and OLS estimators. The first one is Sawa (1973 a) which gives fixed weights to the OLS and 2SLS estimators in order to create an unbiased estimator. The wights are $w_{S,OLS} = -(K-N-1)/(T-K)$ and $w_{S,2SLS} = (T-N-1)/(T-K)$ where N and K are the number of equations and the number of excluded regressor, respectively. Sawa shows that the combined estimator is dominated by the 2SLS estimator in terms of having smaller MSE when the condition $(T-K-2)(K-N-7) \leq 12$ holds. Under the local endogeneity assumption considered in this chapter, it is easy to show that Sawa's combined estimator is always dominated by the 2SLS estimator. Hence, the MSE of the combined

estimator proposed by Sawa (1973 a) is strictly greater than the Stein-like estimator in this chapter. The other related paper in the literature is Hansen (2017) which considers a similar Stein-like estimator in IV regression models and derives the conditions for dominance of the Stein-like estimator over the 2SLS estimator by minimizing the truncated asymptotic weighted risk of the Stein-like estimator using asymptotic distributions of the estimators. There are several limitations in Hansen (2017) which are not the case in this chapter. First, the method considered in this studies the approximate moments, and distribution, however the analysis in Hansen's paper is dealing with asymptotically minimizing a truncated risk. Second, Hansen (2017) minimizes a weighted risk where the weight matrix is set equal to the inverse of the difference of the asymptotic variances of the 2SLS and OLS estimators which might not be practical in most of the empirical applications. However, we derive the MSE matrix which allows for deriving a weighted risk with any positive definite weight matrix. Third, Hansen (2017) only considers shrinking the 2SLS toward the OLS estimator, while, we consider two Stein-like estimators one shrinks 2SLS and the other shrinks the LIML estimator. This is important as under weak instruments scenario the 2SLS estimator is biased in the direction of the OLS estimator, while the LIML estimator needs weaker conditions for the consistency.

Morimune (1978) similar to this chapter considers combining the LIML with the OLS estimator. Morimune (1978) uses fixed weights with the purpose of removing the higher order bias of the LIML estimator and shows that while Sawa (1973 a)'s combined estimator is dominated by the 2SLS estimator, combining the LIML estimator with the OLS estimator dominates the LIML estimator when $K - N > 0$ and $T > K + 2$. Although, the main goal

of Morimune (1978) is different from this chapter, but comparing the MSEs of the two estimators under the local endogeneity assumption shows that the Stein-like estimator in this chapter significantly performs better than that of Morimune (1978)'s estimator when the sample size is large enough [1].

The rest of this chapter is organized as follows. Section 4.2 describes the model and the estimators. In sections 4.3 and 4.4, we give the estimators and their approximations using Small-Disturbance theory. The approximate distributions, bias, and mean squared errors of the estimators are given in section 4.5. Monte-Carlo results and Conclusions are given in sections 4.6 and 4.7. Proofs and detailed calculations are listed in Appendix D.

4.2 The Model

Consider the following complete simultaneous equations model

$$Y_{T\times(N+1)}B_{(N+1)\times(N+1)} + X_{T\times K}\Gamma_{K\times(N+1)} = \sigma U_{T\times(N+1)}, \tag{4.1}$$

where in the system above, there are $N+1$ equations and $N+1$ endogenous variables, denoted by $Y = (y_1, y_2, \ldots, y_{(N+1)})$. There are K exogenous variables, $X = (x_1, x_2, \ldots, x_K)$. B is a nonsingular matrix of parameters with first column $(-1, \beta')'$, where β is a $N\times 1$ vector of unknown coefficients of interest in the first equation. Finally, $U = (u_1, u_2, \ldots, u_{(N+1)})$ are the structural disturbances. The subscript t is used to index observations, $t = 1, \ldots, T$, and σ is a (small) positive number.

[1] For example when $T \geq 2(K+2)$

The first equation of the above system, by assuming for simplicity that it includes no exogenous variables, may be written as

$$y_1 = Y_2\beta + \sigma u_1, \tag{4.2}$$

where y_1 is the first column of Y, and $Y_2 = (y_2, \ldots, y_{(N+1)})$ is $T \times N$, that contains the rest of the columns of Y and is the included endogenous variables.

Assumption 4.1 *The rows of U are independently normally distributed with mean zero and variance-covariance matrix Σ, that is for all t and t' in $\{1, \ldots, T\}$ and i and $j \in \{1, \ldots, N\}$,*

$$\mathbb{E}(u_{it}) = 0,$$

$$Cov(u_{it}, u_{jt'}) = \begin{cases} \sigma_{ij} & \text{if } t = t' \\ 0 & \text{otherwise } (t \neq t'), \end{cases}$$

and $\sigma_{11} = 1$, or in matrix form

$$\mathbb{E}(U) = \underset{T\times(N+1)}{\mathbf{0}},$$

$$\frac{1}{T}\,\mathbb{E}(U'U) = \underset{(N+1)\times(N+1)}{\Sigma} = \begin{bmatrix} 1 & \sigma_{12} & \cdots & \sigma_{1(N+1)} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2(N+1)} \\ & & \vdots & \\ \sigma_{(N+1)1} & \sigma_{(N+1)2} & \cdots & \sigma_{(N+1)(N+1)} \end{bmatrix} \equiv \begin{bmatrix} \sigma_1 & \cdots & \sigma_{N+1} \end{bmatrix}.$$

We assume that B is nonsingular, hence the reduced form of the structural equation (4.1) may be written as

$$Y = -X\Gamma B^{-1} + \sigma U B^{-1} \equiv X\Pi + \sigma V, \tag{4.3}$$

where $\Pi = -\Gamma B^{-1}$ and $V = UB^{-1}$ and $\Pi_{K\times(N+1)} = \underset{K\times 1}{\pi_1} \quad \underset{K\times N}{\Pi_2}$, and $V_{T\times(N+1)} = \underset{T\times 1}{v_1} \quad \underset{T\times N}{V_2}$.

Further, if we partition B^{-1} as

$$B^{-1} = \left[\dot{\beta}_{(N+1)\times 1} \quad \dot{B}_{(N+1)\times N}\right],$$

the reduced form system of equations above can be written in partition as below

$$y_1 = -X\Gamma\dot{\beta} + \sigma U\dot{\beta} = X\pi_1 + \sigma v_1, \tag{4.4}$$

and

$$Y_2 = -X\Gamma\dot{B} + \sigma U\dot{B} = X\Pi_2 + \sigma V_2 \equiv W + \sigma V_2, \tag{4.5}$$

where we define $W = X\Pi_2$.

Assumption 4.2 *Identification: Rank*$(\Pi_2) = N$.

Assumption 4.2 is the rank condition which ensures the identification of the system.

The reduced form error is also normally distributed with

$$\mathbb{E}(V) = \mathbf{0},$$

$$\frac{1}{T}\mathbb{E}(V'V) = \frac{1}{T}\mathbb{E}(B'^{-1}U'UB^{-1}) = B'^{-1}\Sigma B^{-1} = \underset{(N+1)\times(N+1)}{\Omega} \equiv \begin{bmatrix} \underset{1\times 1}{\varpi_{11}} & \underset{1\times N}{\boldsymbol{\varpi_{12}}} \\ \underset{N\times 1}{\boldsymbol{\varpi_{21}}} & \underset{N\times N}{\Omega_{22}} \end{bmatrix}.$$

Following Nagar (1959), we define $\underset{N\times T}{\Psi'} = V_2' - qu_1'$, where the normally distributed matrix Ψ consists of residuals from the population regression of V_2 on u_1. Hence Ψ and u_1 are uncorrelated by construction. Further,

$$\underset{N\times 1}{q} = \frac{Cov(V_2, u_1)}{Var(u_1)} = \frac{\mathbb{E}(V_2'u_1)}{T} = \dot{B}'\sigma_1, \tag{4.6}$$

and define

$$\begin{aligned} \underset{N\times N}{C_1} &= qq', \\ \underset{N\times N}{C_2} &= \frac{\mathbb{E}(\Psi'\Psi)}{T} = \dot{B}'\Sigma\dot{B} - qq', \end{aligned} \tag{4.7}$$

so that it can be shown that $\Omega_{22} = C_1 + C_2$.

4.3 Estimators

We consider three members of the $k-$class estimator of β, which are the OLS, 2SLS, and LIML estimators, and respectively correspond to k equal to zero, one, and λ where λ is the smallest root of the determinantal equation $|Y'Y - kY'M_XY| = 0$, where $M_X = I_T - P_X$ is the projection onto the space orthogonal to the columns of X, with $P_X = X(X'X)^{-1}X'$, and I_T is the identity matrix. Moreover, we consider two types of Stein-like estimators which are a weighted average of the 2SLS and the OLS estimators, and a weighted average of the LIML estimator and the OLS estimator.

4.3.1 $k-$Class Estimators

The $k-$class estimator is defined as

$$\hat{\beta}(k) = (Y_2'H_kY_2)^{-1}Y_2'H_ky_1 = \beta + \sigma(Y_2'H_kY_2)^{-1}Y_2'H_ku_1, \tag{4.8}$$

where $H_k = I_T - kM_X$.

4.3.2 Stein-Like Estimator

Following Maasoumi (1978) and Hansen (2017), we define the Stein-like estimators as the weighted average of a first-order consistent $k-$class estimator (we consider 2SLS estimator

with $k = 1$ and LIML with $k = \lambda$) with the OLS estimator ($k = 0$), where the weights similar to Hansen (2017) are inversely related to the Wu-Hausman (1978) misspecification test statistic. Hence, the Stein-like estimators are defined as

$$\hat{\beta}_{c,k} = \omega_k \hat{\beta}(0) + (1 - \omega_k)\hat{\beta}(k), \qquad \text{for } k = 1, \lambda \tag{4.9}$$

where $\omega_k = \tau / F_{k,WH}$, τ is a positive characterizing scalar which will be determined later, and $F_{k,WH}$ is the Wu-Hausman statistic test, defined as

$$F_{k,WH} = \Big(\hat{\beta}(k) - \hat{\beta}(0)\Big)' \, \mathbb{R}_k \, \Big(\hat{\beta}(k) - \hat{\beta}(0)\Big), \tag{4.10}$$

and $\mathbb{R}_k$ is

$$\mathbb{R}_k = \sigma^2 \Big((Y_2' H_k Y_2)^{-1} - (Y_2' Y_2)^{-1}\Big)^{-1}. \tag{4.11}$$

4.4 Small-Disturbance Asymptotic Expansions

We use Kadane (1971) small-disturbance method to derive the asymptotic expansions of the estimators. Then, we report the bias and mean squared error matrix (MSEM) of the estimators up to orders σ^2 and σ^4, respectively.

4.4.1 $k-$class Estimators

Employing equation (4.5) in equation (4.8), we have

$$\begin{aligned}\hat{\beta}(k)-\beta &= \Big[\Big(W+\sigma V_2\Big)' H_k \Big(W+\sigma V_2\Big)\Big]^{-1}\Big(W+\sigma V_2\Big)' H_k \sigma u_1 \\ &= \Big(W'H_kW+\sigma W'H_kV_2+\sigma V_2'H_kW+\sigma^2 V_2'H_kV_2\Big)^{-1}\Big(\sigma W'H_ku_1+\sigma^2 V_2'H_ku_1\Big) \\ &= \Big(I_N+\sigma QW'H_kV_2+\sigma QV_2'H_kW+\sigma^2 QV_2'H_kV_2\Big)^{-1} Q\Big(\sigma W'H_ku_1+\sigma^2 V_2'H_ku_1\Big) \\ &= \Big(I_N+\sigma QS+\sigma^2 QV_2'H_kV_2\Big)^{-1} Q\Big(\sigma W'u_1+\sigma^2 V_2'H_ku_1\Big),\end{aligned} \tag{4.12}$$

where $Q_{N\times N}=(W'W)^{-1}$, and $S=V_2'W+W'V_2$, and the use has been made of $W'H_k=W'$.

Using the standard geometric expansion for the inverse of a matrix[2], the above equation may be written as

$$\begin{aligned}\hat{\beta}(k)-\beta &= \sigma QW'u_1+\sigma^2 Q\Big(V_2'H_ku_1-SQW'u_1\Big) \\ &\quad +\sigma^3 Q\Big(SQSQW'u_1-V_2'H_kV_2QW'u_1-SQV_2'H_ku_1\Big)+O_p(\sigma^4).\end{aligned} \tag{4.13}$$

Theorem 4.3 *Under assumptions 4.1 and 4.2, the bias of the $k-$class estimators up to order σ^2, is*

$$\mathbb{E}(\hat{\beta}(k)-\beta)=\sigma^2 Qq(L_k-1), \quad \textit{for fixed } k, \tag{4.14}$$

$$\mathbb{E}(\hat{\beta}(\lambda)-\beta)=-\sigma^2 Qq, \quad \textit{for LIML estimator,} \tag{4.15}$$

and the mean squared error matrix up to order σ^4 is

$$\begin{aligned}\mathbb{E}(\hat{\beta}(k)-\beta)(\hat{\beta}(k)-\beta)' &= \sigma^2 Q+\sigma^4\big\{(3-2L_k)tr(C_1Q)Q+tr(QC_2)Q \\ &\quad +QC_1Q((L_k-2)^2+2+2s_k)+QC_2Q(2+s_k-L_k)\big\},\end{aligned} \tag{4.16}$$

[2] $(I+A)^{-1}=I-A+A^2-A^3+\ldots$

$$\mathbb{E}(\hat{\beta}(\lambda)-\beta)(\hat{\beta}(\lambda)-\beta)' = \sigma^2 Q + \sigma^4\big\{3tr(C_1Q)Q + tr(QC_2)Q + 6QC_1Q + QC_2Q[\frac{(L_1+2)(T-K+L_1-2)}{T-K-2}]\big\}, \tag{4.17}$$

where $L_k = (1-k)T + kK - N$ *and* $s_k = k(k-1)(T-K)$.

Proof: See Kadane (1971).

4.4.2 Stein-Like Estimator

In what follows from this section, we analyze the bias and MSE of the proposed Stein-like estimators. We start by expanding $\mathbb{R}_k$ defined in equation (4.11).

Using equation (4.13), we have

$$(Y_2'HY_2)^{-1} = \Big(I_N - \sigma QS - \sigma^2 QV_2'HV_2 + \sigma^2 QSQS + \sigma^3 QSQV_2'HV_2 + \sigma^3 QV_2'HV_2QS\Big)Q + O_p(\sigma^4),$$

$$(Y_2'Y_2)^{-1} = \Big(I_N - \sigma QS - \sigma^2 QV_2'V_2 + \sigma^2 QSQS + \sigma^3 QSQV_2'V_2 + \sigma^3 QV_2'V_2QS\Big)Q + O_p(\sigma^4).$$

Hence the difference of the expressions above may be written as

$$(Y_2'HY_2)^{-1} - (Y_2'Y_2)^{-1} = \sigma^2 k\Big[QV_2'M_XV_2Q - \sigma QSQV_2'M_XV_2Q - \sigma QV_2'M_XV_2QSQ\Big] + O_p(\sigma^4). \tag{4.18}$$

Further, by using equation (4.18) in equation (4.11),

$$\begin{aligned}\mathbb{R}_k &= \sigma^2\Big((Y_2'HY_2)^{-1} - (Y_2'Y_2)^{-1}\Big)^{-1} \\ &= \frac{1}{k}Q^{-1}\Big[I_N - \sigma(V_2'M_XV_2)^{-1}SQV_2'M_XV_2 - \sigma QS + O_p(\sigma^2)\Big]^{-1}(V_2'M_XV_2)^{-1}Q^{-1} \\ &= \frac{1}{k}Q^{-1}\Big[(V_2'M_XV_2)^{-1} + \sigma(V_2'M_XV_2)^{-1}SQ + \sigma QS(V_2'M_XV_2)^{-1} + O_p(\sigma^2)\Big]Q^{-1}.\end{aligned} \tag{4.19}$$

In addition from equation (4.13), we have

$$\hat{\beta}(0) - \hat{\beta}(k) = k\sigma^2\Big[QV_2'M_Xu_1 - \sigma QV_2'M_XV_2QW'u_1 - \sigma QSQV_2'M_Xu_1\Big] + O_p(\sigma^4). \quad (4.20)$$

Employing equations (4.19) and (4.20) in equation (4.10),

$$F_{k,WH} = k\Big[u_1'M_XV_2(V_2'M_XV_2)^{-1}V_2'M_Xu_1 - 2\sigma u_1'M_XV_2QW'u_1 + O_p(\sigma^2)\Big]. \quad (4.21)$$

Therefore, we have the following expression

$$\frac{1}{F_{k,WH}} = \frac{1}{ku_1'M_XV_2(V_2'M_XV_2)^{-1}V_2'M_Xu_1}\Big(1 + \frac{2\sigma u_1'M_XV_2QW'u_1}{u_1'M_XV_2(V_2'M_XV_2)^{-1}V_2'M_Xu_1} + O_p(\sigma^2)\Big). \quad (4.22)$$

Using equations (4.13), and (4.22) in equation (4.9), we can write the Stein-like estimators as

$$\begin{aligned}
\hat{\beta}_{c,k} - \beta &= (\hat{\beta}(k) - \beta) + \frac{\tau}{F_{k,WH}}\Big((\hat{\beta}(0) - \beta) - (\hat{\beta}(k) - \beta)\Big) \\
&= \sigma QW'u_1 + \sigma^2 Q\Big(V_2'H_ku_1 - SQW'u_1\Big) + \sigma^3 Q\Big(SQSQW'u_1 - V_2'H_kV_2QW'u_1 - SQV_2'H_ku_1\Big) \\
&+ \tau\sigma^2\frac{1}{u_1'M_XV_2(V_2'M_XV_2)^{-1}V_2'M_Xu_1}\Big[QV_2'M_Xu_1 - \sigma QV_2'M_XV_2QW'u_1 - \sigma QSQV_2'M_Xu_1 \\
&+ \frac{2\sigma}{u_1'M_XV_2(V_2'M_XV_2)^{-1}V_2'M_Xu_1}u_1'M_XV_2QW'u_1QV_2'M_Xu_1\Big] + O_p(\sigma^4).
\end{aligned} \quad (4.23)$$

The above equation has the product of normally distributed and correlated terms in the denominator, which make the moments calculations complicated. So, we make the following local endogeneity assumption, and then revise the asymptotic expansion of the Stein-Like estimator. We derive the bias and MSE of the estimator under this assumption in the next section.

Assumption 4.4 *Local Endogeneity:* $q = Cov(V_2, u_1)/T = \sigma\delta, \quad where \quad \delta_{N\times 1} \in R^N$ [3].

[3]Note that the local Endogeneity assumption here is similar to the local asymptotic considered in Hansen (2017), when σ is replaced by $1/\sqrt{T}$

Using Assumption 4.4, in equation (4.22), it is equal to the following expression

$$\frac{1}{F_{k,WH}} = \frac{1}{k}\frac{1}{u_1'M_X\Psi(\Psi'M_X\Psi)^{-1}\Psi'M_Xu_1}\Big[1+2\sigma u_1'M_X\Psi(\Psi'M_X\Psi)^{-1}\delta + \frac{2\sigma}{u_1'M_X\Psi(\Psi'M_X\Psi)^{-1}\Psi'M_Xu_1}\Big(u_1'M_X\Psi QW'u_1 - u_1'M_X\Psi(\Psi'M_X\Psi)^{-1}\delta u_1'M_Xu_1\Big)\Big] + O_p(\sigma^2). \tag{4.24}$$

Using equation (4.24) in the Stein-like estimator expression (equation (4.9)), we have

$$\begin{aligned}\hat{\beta}_{c,k} - \beta &= (\hat{\beta}(k)-\beta) + \frac{\tau}{F_{k,WH}}\Big((\hat{\beta}(0)-\beta)-(\hat{\beta}(k)-\beta)\Big) = \sigma QW'u_1 + \sigma^2Q(\Psi'H_ku_1 - S_\Psi QW'u_1)\\ &+ \sigma^3Q\Big[qu_1'H_ku_1 - (qu_1'W + W'u_1q')QW'u_1 + S_\Psi QS_\Psi QW'u_1 - \Psi'H_k\Psi QW'u_1 - S_\Psi Q\Psi_2'H_ku_1\Big]\\ &+ \frac{\tau\sigma^2}{u_1'M_X\Psi(\Psi'M_X\Psi)^{-1}\Psi'M_Xu_1}\Bigg[\Big(1+2\sigma u_1'M_X\Psi(\Psi'M_X\Psi)^{-1}\delta\Big)Q\Psi'M_Xu_1 + \sigma Qqu_1'M_xu_1\\ &- \sigma Q\Psi'M_X\Psi QW'u_1 - \sigma QS_\Psi Q\Psi'M_Xu_1 + \frac{2\sigma}{u_1'M_X\Psi(\Psi'M_X\Psi)^{-1}\Psi'M_Xu_1}\Big(u_1'M_X\Psi QW'u_1\\ &- u_1'M_X\Psi(\Psi'M_X\Psi)^{-1}\delta u_1'M_Xu_1\Big)Q\Psi'M_Xu_1\Bigg] + O_p(\sigma^4).\end{aligned} \tag{4.25}$$

We first derive the approximate distribution of the Stein-like estimator under Assumption 4.4 in the next section, then we give the bias and MSEM of the estimator.

4.5 The Approximate Distribution Functions of The Estimators

In this section, the approximate density functions of the estimators are derived for the statistics

$$\hat{e}_k = \frac{1}{\sigma}(\hat{\beta}_k - \beta), \quad \text{for } k = 1, \lambda, \tag{4.26}$$

and

$$\hat{e}_{c,k} = \frac{1}{\sigma}(\hat{\beta}_{c,k} - \beta), \qquad \text{for } k = 1, \lambda, \tag{4.27}$$

as σ goes to zero, where $\hat{e}$ for the 2SLS, and LIML are denoted by $\hat{e}_1$, $\hat{e}_\lambda$ and for the Stein-like estimators it is denoted by $\hat{e}_{c,k}$, $\quad k = 1, \lambda$.

Theorem 4.5 *Under assumptions 4.1–4.4, the asymptotic expansion of the density function of $\hat{e}_1$, and $\hat{e}_\lambda$ as σ goes to zero is given by*

$$\begin{aligned} f_{2SLS}(\xi) = \phi_Q(\xi) \Bigg[1 + \sigma^2 \delta' \xi \Big(N + 1 + L_1 - \xi' Q^{-1} \xi \Big) + \frac{\sigma^2}{2} \Big[L_1 \, tr(C_2 Q) \\ - \xi' C_2 \xi \Big(N + 2 + L_1 - \xi' Q^{-1} \xi \Big) \Big] \Bigg] + O(\sigma^3), \end{aligned} \tag{4.28}$$

$$\begin{aligned} f_{LIML}(\xi) = \phi_Q(\xi) \Bigg[1 + \sigma^2 \delta' \xi \Big(N + 1 - \xi' Q^{-1} \xi \Big) + \frac{\sigma^2}{2} \Big[- \frac{L_1 (T - N)}{T - K - 2} tr(C_2 Q) \\ - \xi' C_2 \xi \Big(N + 2 - \frac{L_1 (T + N)}{T - K - 2} - \xi' Q^{-1} \xi \Big) \Big] \Bigg] + O(\sigma^3), \end{aligned} \tag{4.29}$$

where ξ is an $N \times 1$ vector and $\phi_Q(\xi)$ is the multivariate normal density function with mean **0** *and covariance matrix Q.*

Proof: See Anderson et al. (1986).

Theorem 4.6 *Under assumptions 4.1–4.4, the asymptotic expansion of the density functions of $\hat{e}_{c,k}$ for $k = 1, \lambda$, as σ goes to zero are*

$$f_{c,1}(\xi) = f_{2SLS}(\xi) + \phi_Q(\xi) \sigma^2 \tau \left[\alpha_1 \delta' \xi + \frac{1}{2} \Big[\xi' C_2 \xi - tr(C_2 Q) \Big] \Big(\tau \alpha_2 - 2 \alpha_1 \Big) \right] + O(\sigma^3), \tag{4.30}$$

$$f_{c,\lambda}(\xi) = f_{LIML}(\xi) + \phi_Q(\xi) \sigma^2 \tau \left[\alpha_1 \delta' \xi + \frac{1}{2} \Big[\xi' C_2 \xi - tr(C_2 Q) \Big] \Big(\tau \alpha_3 - 2 \alpha_1 \Big) \right] + O(\sigma^3), \tag{4.31}$$

where $\alpha_1 = \frac{(T-K)}{N}$, $\alpha_2 = \frac{(T-K)}{N(N-2)}$, $\alpha_3 = \alpha_2 + c$, *and* $c \in \Big(-(T-2N)\alpha_2, \quad 0\Big)$.

Proof: Appendix D,(See page 168).

In the next theorem, the first and the second moments of the Stein-like estimators based on the approximate expansions of their distribution are given.

Theorem 4.7 *Under assumptions 4.1–4.4, the approximate bias of the Stein-like estimator* $\hat{\beta}_{c,k}$ *for* $k = 1, \lambda$ *as* σ *goes to zero is given by*

$$ABias(\hat{\beta}_{c,k}) = \mathbb{E}\left(\frac{1}{\sigma}(\hat{\beta}_{c,k} - \beta)\right) = 0 + O(\sigma^2), \tag{4.32}$$

and the approximate MSEM is

$$AMSE(\hat{\beta}_{c,1}) = \mathbb{E}\left(\frac{1}{\sigma^2}(\hat{\beta}_{c,1} - \beta)(\hat{\beta}_{c,1} - \beta)'\right) = AMSE(\hat{\beta}(1)) + \tau\sigma^2\Big[\tau\alpha_2 - 2\alpha_1\Big]QC_2Q + O(\sigma^3), \tag{4.33}$$

$$AMSE(\hat{\beta}_{c,\lambda}) = \mathbb{E}\left(\frac{1}{\sigma^2}(\hat{\beta}_{c,\lambda} - \beta)(\hat{\beta}_{c,\lambda} - \beta)'\right) \leq AMSE(\hat{\beta}(\lambda)) + \tau\sigma^2\Big[\tau\alpha_2 - 2\alpha_1\Big]QC_2Q + O(\sigma^3), \tag{4.34}$$

where from Theorem 4.5 (or equation (4.28)), it can be derived that

$$AMSE(\hat{\beta}(1)) = \mathbb{E}\left(\frac{1}{\sigma^2}(\hat{\beta}(1) - \beta)(\hat{\beta}(1) - \beta)'\right) = Q + \sigma^2\, tr(QC_2)Q + \sigma^2 QC_2Q\Big(2 - L_1\Big) + O(\sigma^3). \tag{4.35}$$

$$\begin{aligned} AMSE(\hat{\beta}(\lambda)) &= \mathbb{E}\left(\frac{1}{\sigma^2}(\hat{\beta}(\lambda) - \beta)(\hat{\beta}_\lambda - \beta)'\right) \\ &= Q + \sigma^2\, tr(QC_2)Q + \sigma^2 QC_2Q\frac{(L_1+2)(T-N-2)}{T-K-2} + O(\sigma^3). \end{aligned} \tag{4.36}$$

Proof: Appendix D,(See page 171).

Corollary 4.8 *Under assumptions 4.1–4.4, we have*

$$AMSE(\hat{\beta}_{c,1}) - AMSE(\hat{\beta}(1)) = \tau\sigma^2 QC_2Q\Big[\tau\alpha_2 - 2\alpha_1\Big] + O(\sigma^3), \tag{4.37}$$

$$AMSE(\hat{\beta}_{c,\lambda}) - AMSE(\hat{\beta}(\lambda)) \leq \tau\sigma^2 QC_2Q\Big[\tau\alpha_2 - 2\alpha_1\Big] + O(\sigma^3), \tag{4.38}$$

where the right-hand side of the above equations are negative when $N > 2$ and

$$0 < \tau < 2(N-2).$$

Therefore, the Stein-like estimators dominate the 2SLS, and LIML estimators in terms of their MSEs when the number of endogenous variables is more than 2. The optimal value of the shrinkage parameter that minimizes the MSE of the Stein-like estimator is

$$\tau_{opt} = N - 2.$$

As a comparison of the probability of concentration around the true β, we compute

$$P(||Q^{-1/2}\hat{e}_{c,k}|| < z) - P(||Q^{-1/2}\hat{e}_k|| < z) = \int\cdots\int_{||Q^{-1/2}\xi||<z} (f_{c,k}(\xi) - f_k(\xi))\, d\xi, \tag{4.39}$$

where $||\xi|| = \max\{|\xi_1|, \ldots, |\xi_N|\}$. Using Theorem 4.5 and Theorem 4.6 the next theorem follows.

Theorem 4.9 *Under assumptions 4.1–4.4,*

$$P(||Q^{-1/2}\hat{e}_{c,1}|| < z) - P(||Q^{-1/2}\hat{e}_1|| < z) = \sigma^2[\Phi(z) - \Phi(-z)]^N z\tilde{\phi}(z)\, tr(QC_2)d + O(\sigma^3), \tag{4.40}$$

$$P(||Q^{-1/2}\hat{e}_{c,\lambda}|| < z) - P(||Q^{-1/2}\hat{e}_\lambda|| < z) \geq \sigma^2[\Phi(z) - \Phi(-z)]^N z\tilde{\phi}(z)\, tr(QC_2)d + O(\sigma^3), \tag{4.41}$$

where $\tilde{\phi}(z) = \phi(z)/[\Phi(z) - \Phi(-z)]$, $d = \tau(2\alpha_1 - \tau\alpha_2)$, *and* $\Phi(.)$ *and* $\phi(.)$ *are, respectively, the standard normal distribution and density functions.*

Proof: Appendix D,(See page 172).

Corollary 4.10 *Using theorem 4.9 , and provided* $0 < \tau < 2(N-2)$, *and* $N > 2$, *we have*

$$P(||Q^{-1/2}\hat{e}_{c,k}|| < z) \geq P(||Q^{-1/2}\hat{e}_k|| < z) + O(\sigma^3), \quad k = 1, \lambda, \tag{4.42}$$

and the optimal value of τ *that maximizes the concentration probability of the Stein-like estimator is*

$$\tau_{opt} = N - 2.$$

4.6 Monte-Carlo Simulation

Our simulation experiment uses a design similar to that used by Hansen (2017), where $T \in \{100, 200\}$, $N \in \{3, 5, 8\}$. The observations are generated by the process

$$y_1 = Y_2\beta + \sigma u_1,$$

$$Y_2 = X\Pi_2 + V_2,$$

where u_1 has a standard normal distribution, V_2 and X have a multivariate normal distribution with mean zero, and variance-covariance matrix I_N, and I_K, respectively. We set the correlation between u_1 and the rows of V_2 equal to $\rho/\sqrt{N}$, where ρ takes values on a 40-point grid on $[0, 0.975]$. We set β to zero, $\Pi_2 = cI_m$, where $c = \sqrt{R^2/(1-R^2)}$, hence R^2 is the reduced form population R^2 for each endogenous variable. This is important because

R^2 measures the strength of the instrument. We consider two cases for the reduced form population R^2, which are $\{0.4, 0.8\}$.

The results of $1,000$ monte carlo simulations are given in Figures 4.1-4.6, where the vertical axis measure the relative mean squared error of OLS, 2SLS/LIML, and Stein-like estimators which is the mean squared errors of these estimators divided by the mean squared error of 2SLS/LIML estimator, and the horizontal axis measures the degree of endogeneity (ρ).

The monte carlo results support our theoretical findings of the previous sections. They show that the relative mean squared errors of the Stein-like estimators for the whole parameter space is below that of the 2SLS/LIML estimator. Further, the relative mean squared errors of the Stein-like estimators are smaller than that of OLS estimator except for very small size of endogeneity.

We note that when R^2 is relatively small (weak instruments) the OLS estimator performs better than the 2SLS/LIML estimator up to mild degrees of endogeneity. This is because the 2SLS/LIML estimator has high dispersion, so that the OLS estimator has smaller mean squared error. In this case, the Stein-like estimator tends to gain from the efficiency of the OLS estimator by assigning a larger weight to this estimator, and prevails. However, when R^2 is relatively large the 2SLS/LIML estimator performs better that the OLS estimator except for very small size of endogeneity, and the Stein-like estimator by giving more weight to the 2SLS/LIML estimator dominates the OLS estimator. Moreover, when the number of endogenous variables increases the OLS estimator gains from a higher efficiency and its mean squared error remains less than the 2SLS/LIML estimator even when the degree of

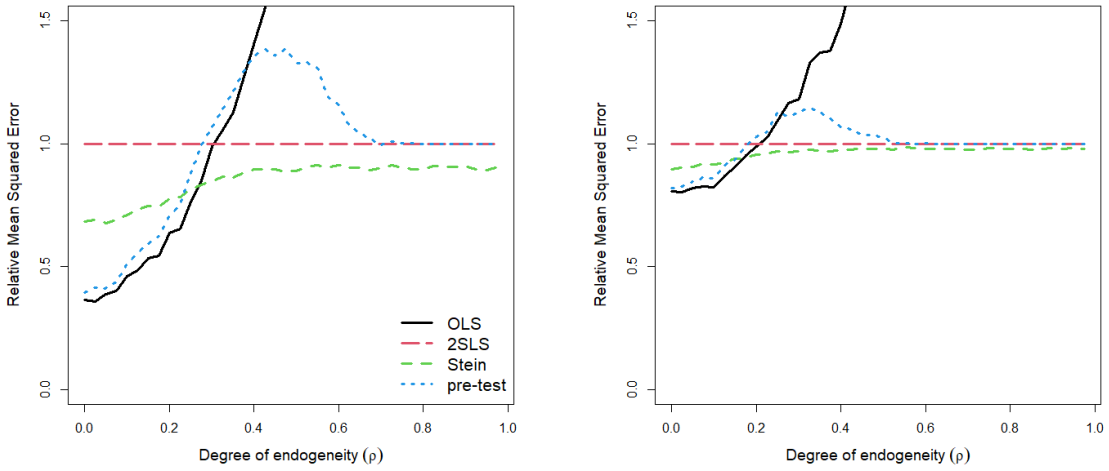
endogeneity is moderate. Similarly, the Stein-like estimator gains from the efficiency of the OLS estimator. We also report the results of the pre-test estimator which tests the null of endogeneity and assigns weight zero or one to the OLS or 2SLS/LIML estimator based on the test results under 5% critical value. The mean squared error of the pre-test estimator is small when the degree of endogeneity is small, but is very high for moderate degrees of endogeneity.

In general, we find that the Stein-like estimators perform robustly well in simultaneous equation models with various degrees of endogeneity. When there is a strong degree of endogeneity or the sample size is large, the Stein-like estimators prevails. When there is a relatively weak degree of endogeneity or weak instruments, the Stein-like estimators tend to gain more from the efficiency of the OLS estimator by assigning a larger weight to this estimator, and thus still remains one of the best choices.

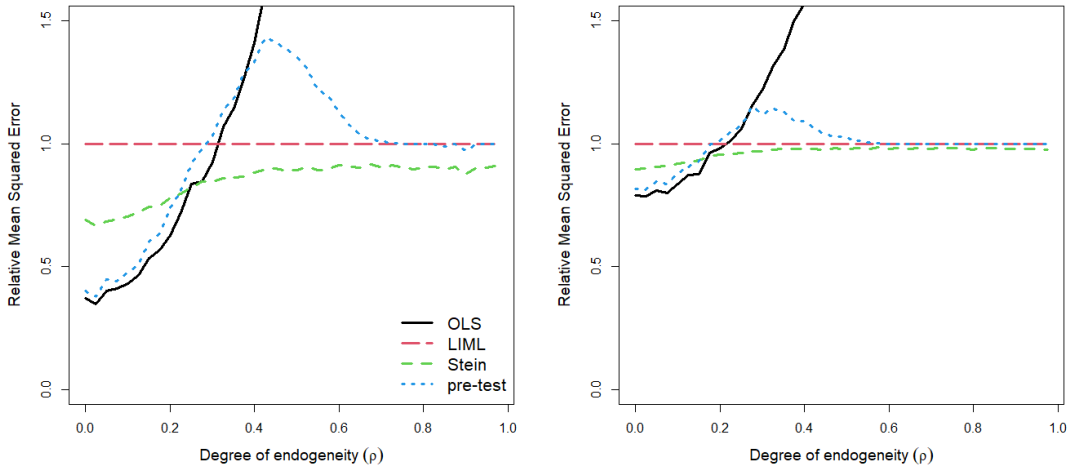
4.7 Conclusion

In this chapter, we introduce two Stein-like estimators for estimating the structural parameters of a Simultaneous Equations Model. The estimators are weighed averages of the 2SLS/LIML and the OLS estimators where the weight is inversely related to a Wu-Hausman test statistic. The approximate distribution, bias, and MSEM of the Stein-like estimators using Small-Disturbance approximations of Kadane (1971) are derived. The proposed method has several advantages relative to the existing methods. First, it allows us to study the performance of the weighted averages of any k-class estimators with the OLS estimator. This is important because under weak instruments the 2SLS estimator is

biased towards the OLS estimator, and an alternative consistent estimator is required to allow balancing between the bias and variance efficiency of the OLS estimator. Second, the dominance and optimality of the Stein-like estimators proposed here are not limited to a specific MSE and hold for any weighted quadratic loss function where the weight is positive definite and symmetric. Lastly, the framework considered here allows for studying the higher order terms, which is critical here because k-class estimators tend to have higher order bias.

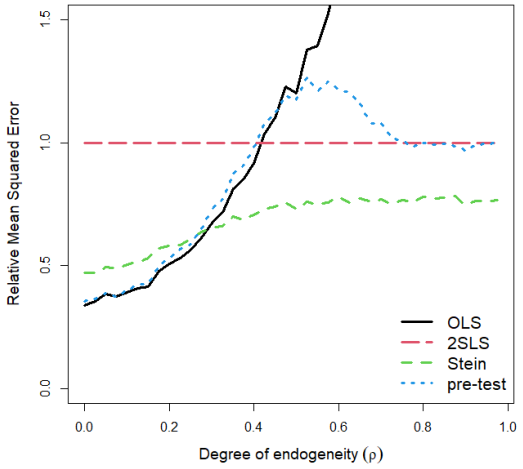


$R^2 = 0.4$ $R^2 = 0.8$

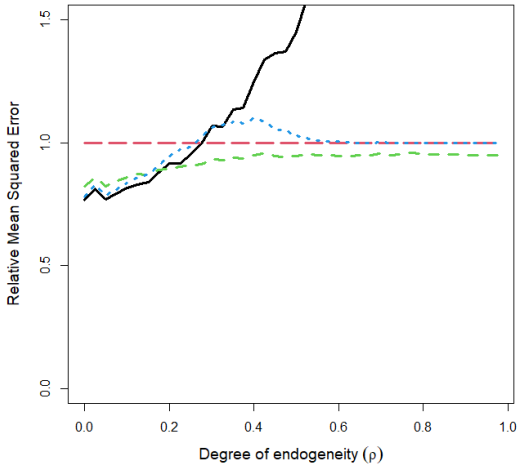


$R^2 = 0.4$ $R^2 = 0.8$

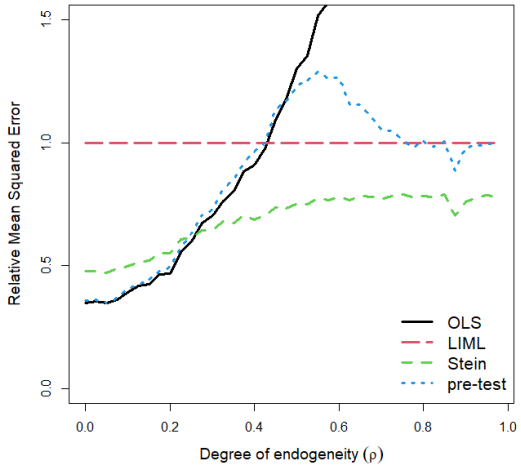
Figure 4.1: Relative mean squared error of OLS, 2SLS, LIML, Stein estimators, and pre-test, for $T = 100$, $N = 3$. The top two figures represent the Stein with 2SLS and OLS, and the bottom ones represent the Stein with OLS and LIML. Note: the pre-test estimator uses the Wu-Hausman test static under 5% critical value to choose between the estimators.



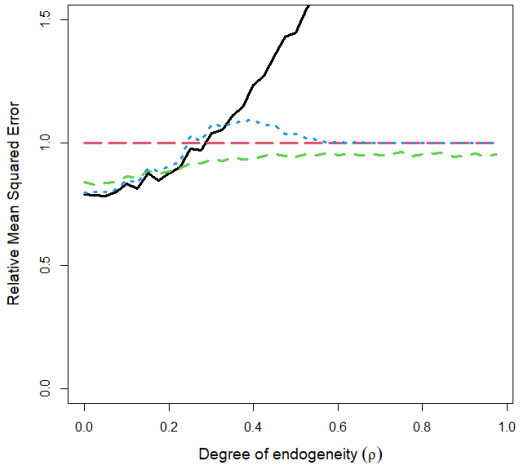
$R^2 = 0.4$



$R^2 = 0.8$

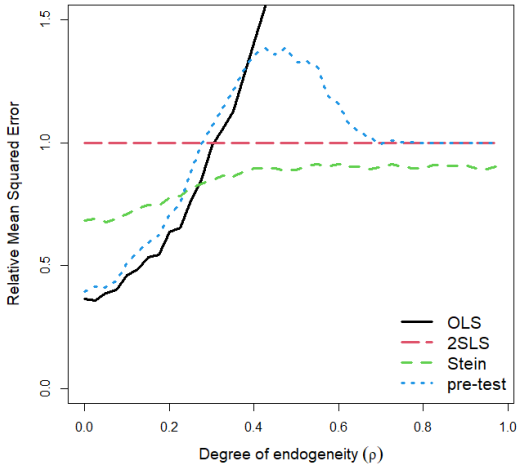


$R^2 = 0.4$

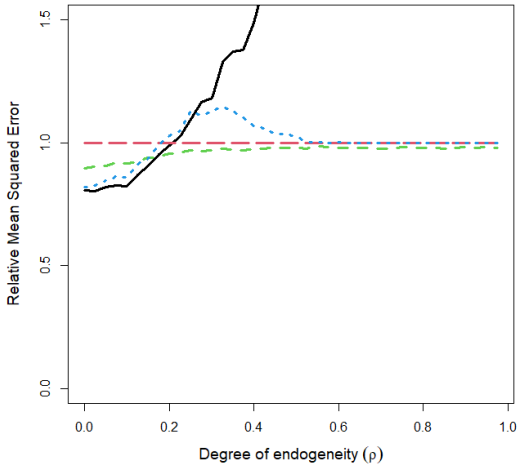


$R^2 = 0.8$

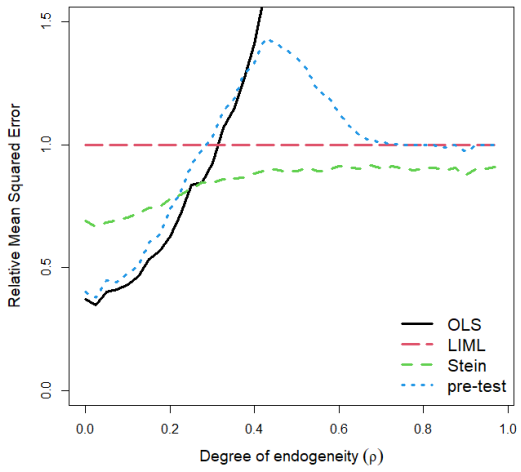
Figure 4.2: Relative mean squared error of OLS, 2SLS, LIML, Stein estimators, and pre-test, for $T = 100$, $N = 5$. The top two figures represent the Stein with 2SLS and OLS, and the bottom ones represent the Stein with OLS and LIML.



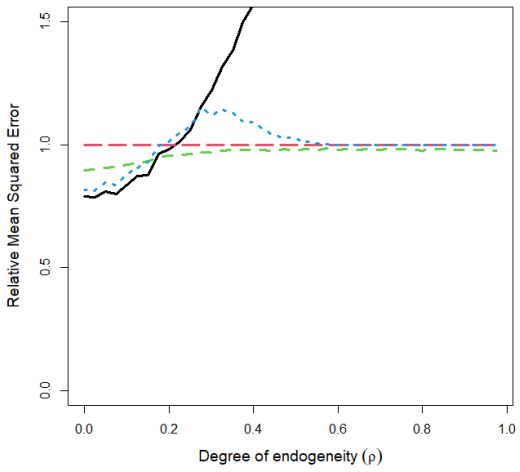
$R^2 = 0.4$



$R^2 = 0.8$

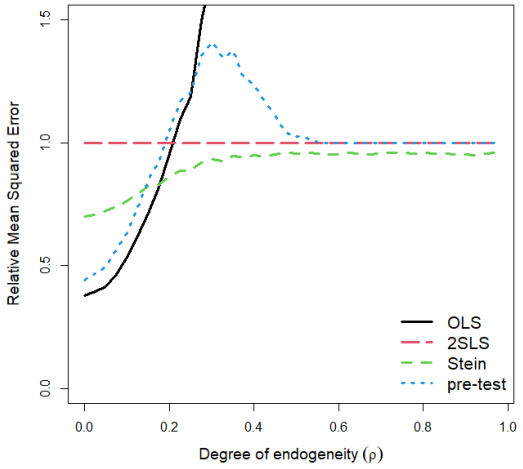


$R^2 = 0.4$

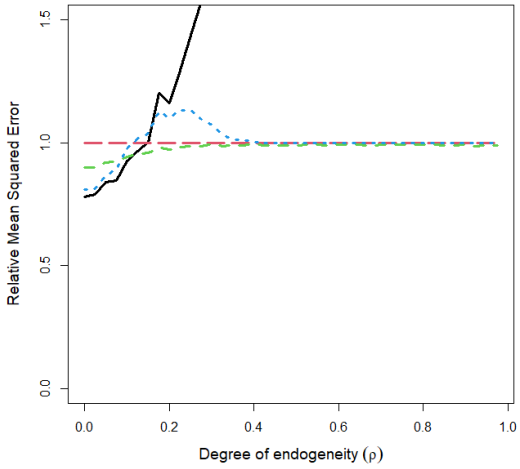


$R^2 = 0.8$

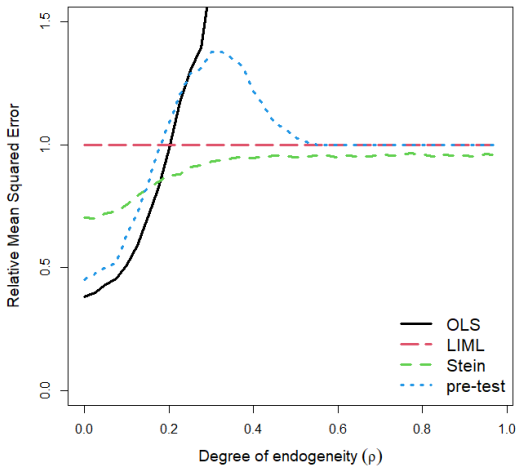
Figure 4.3: Relative mean squared error of OLS, 2SLS, LIML, Stein estimators, and pre-test, for $T = 100$, $N = 8$. The top two figures represent the Stein with 2SLS and OLS, and the bottom ones represent the Stein with OLS and LIML.



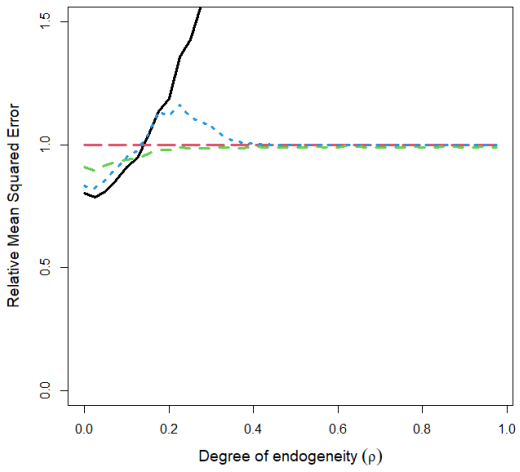
$R^2 = 0.4$



$R^2 = 0.8$

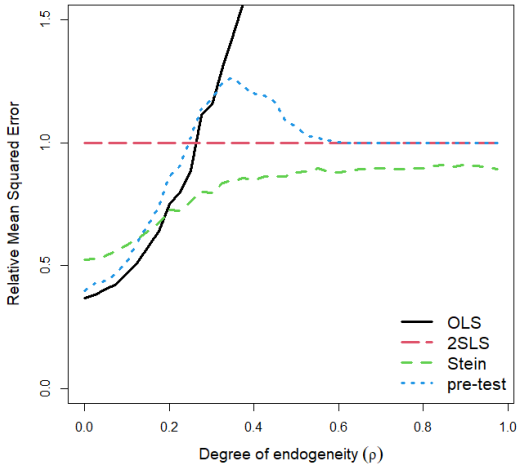


$R^2 = 0.4$

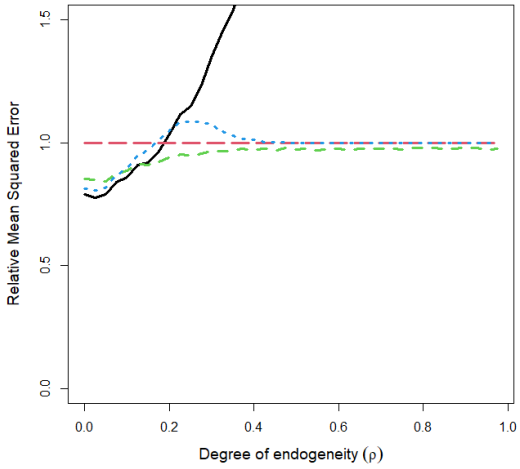


$R^2 = 0.8$

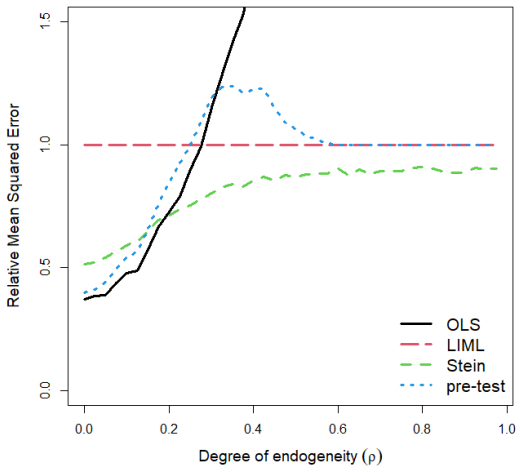
Figure 4.4: Relative mean squared error of OLS, 2SLS, LIML, Stein estimators, and pre-test, for $T = 200$, $N = 3$. The top two figures represent the Stein with 2SLS and OLS, and the bottom ones represent the Stein with OLS and LIML.



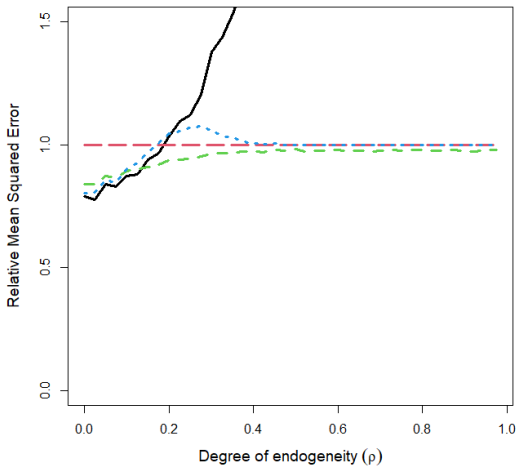
$R^2 = 0.4$



$R^2 = 0.8$



$R^2 = 0.4$



$R^2 = 0.8$

Figure 4.5: Relative mean squared error of OLS, 2SLS, LIML, Stein estimators, and pre-test, for $T = 200$, $N = 5$. The top two figures represent the Stein with 2SLS and OLS, and the bottom ones represent the Stein with OLS and LIML.

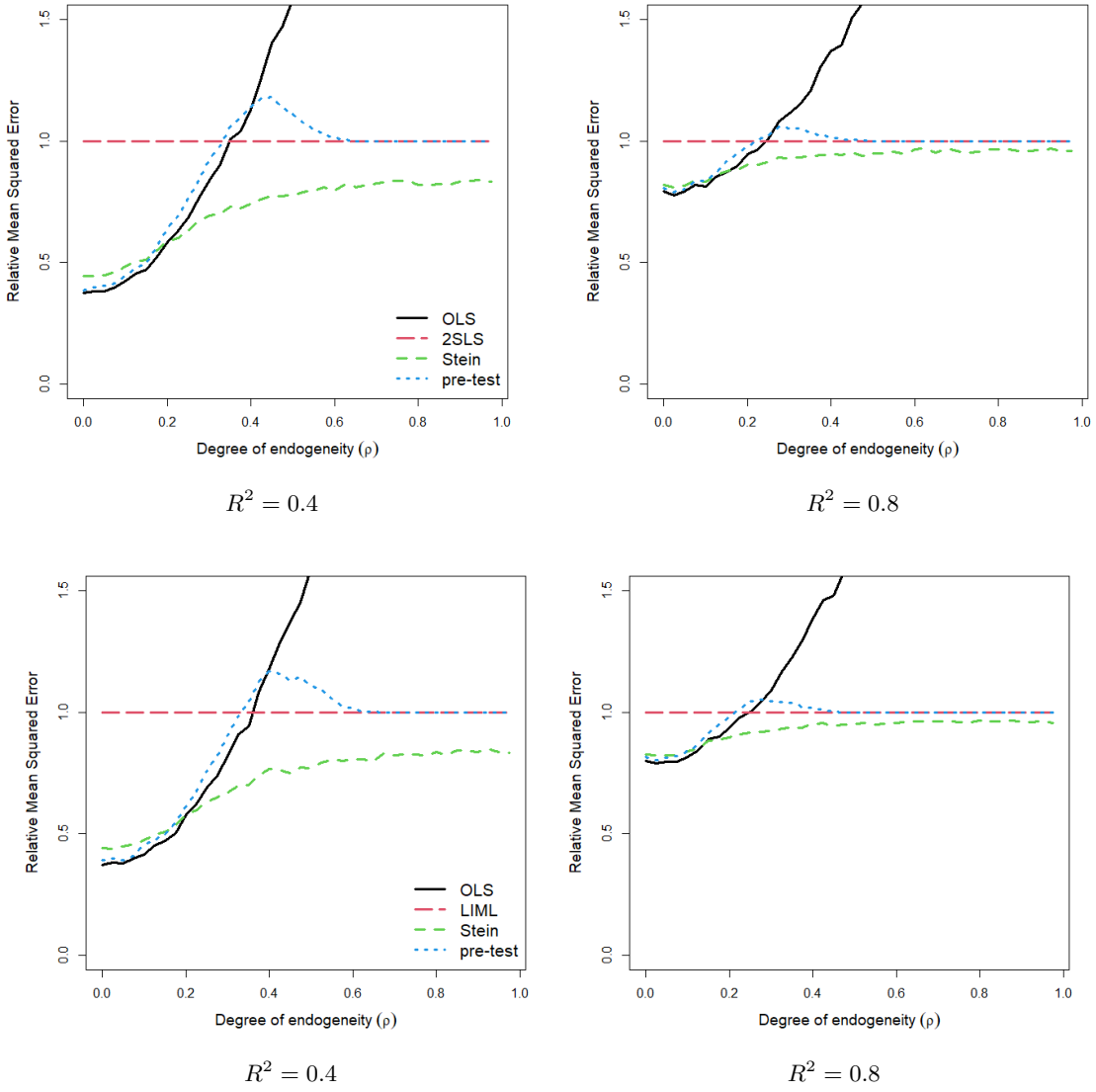


Figure 4.6: Relative mean squared error of OLS, 2SLS, LIML, Stein estimators, and pre-test, for $T = 200$, $N = 8$. The top two figures represent the Stein with 2SLS and OLS, and the bottom ones represent the Stein with OLS and LIML.

Chapter 5

Estimation and Identification of Latent Group Structures in Panel Data

5.1 Introduction

Panel data offer great opportunities in empirical research. Nevertheless, in practice, they typically involve aggregate data from various units (such as workers, firms, countries) that are different in some unobservable aspects to researchers. Accordingly, the researchers face a trade-off between using flexible methods to model the unobservable heterogeneity, and using pooled models that avoid the heterogeneity by assuming to some extent homogeneous coefficients for all individual units. To overcome this challenge, recently, latent group structures in panel data literature have received considerable attention. The most important

advantage of the latent group structure is that unlike completely heterogenous or fully homogenous models, it allows panel units to be classified into groups, where the individuals within a group share the same slope parameters, while heterogeneity exists across the groups. This chapter inspired by the literature introduces a simple and fast method to jointly identify and estimate latent group structures in panel data models when the number of groups and the individuals' group identities are both unknown.

A common approach to model heterogeneity in econometric analysis is to assume complete slope heterogeneity. This assumption avoids misspecification, but does not gain from working with panel data, and could result in imprecise estimates even if the time dimension is long (see, Baltagi and Griffin (1997)). Nonetheless, conventional panel data models often avoid the heterogeneity and assume the regression parameters are the same across individuals, and unobserved heterogeneity is modeled through individual-specific effects (fixed effect and random effect models). This assumption exploits cross-section averaging and causes higher efficiency, but at the cost of estimation bias and inconsistency, which is supported by an increasing number of studies due to a better forecast performance of the associated estimators (see for example, Baltagi et al. (1989), Maddala (1991), Maddala and Hu (1996), Baltagi and Griffin (1997), and Hoogstrate et al. (2000)). In spite of a better forecast performance, it is often difficult to justify the slope homogeneity assumption in the empirical work, as pointed out by Hsiao and Tahmiscioglu (1997), Phillips and Sul (2007), Browning and Carro (2007), and Su and Chen (2013). This discussion motivated much of the recent research on the latent group structures in panel data analysis including

Sun (2005), Lin and Ng (2012), Deb and Trivedi (2013), Bonhomme and Manresa (2015), Sarafidis and Weber (2015), Ando and Bai (2016), Bester and Hansen (2016), Su et al. (2016), Lu and Su (2017), Su and Ju (2018), Wang et al. (2018), Su et al. (2019), Gu and Volgushev (2019), Liu et al. (2020), and Wang and Su (2020), among others. Moreover, the group structure has sound foundations in game theory or macroeconomic models where multiplicity of Nash equilibria is expected (Hahn and Moon (2010)). The latent group structure models partition individuals in different groups and allow the within group individuals share common coefficients, while the groups are assumed to have slope heterogeneity. Since the group membership and the number of groups are unknown in these models, the determination of the true number of groups and each individual's group identity are the key questions. Several approaches have been proposed to address these questions. Sun (2005), Kasahara and Shimotsu (2009), and Browning and Carro (2007) consider finite mixture models. Su et al. (2016) develop a new variant of the Lasso (least absolute shrinkage and selection operator) procedure, called classifier-Lasso (C-Lasso), to achieve classification in panel structure models where the penalty takes an additive-multiplicative form. The C-Lasso method of Su et al. (2016) has been extended to allow for two-way component errors, interactive fixed effects, non-stationary regressors, and semi-parametric specification, respectively, in Lu and Su (2017), Su and Ju (2018), Huang et al. (2020), and Su et al. (2019). Lin and Ng (2012) and Sarafidis and Weber (2015) extend the K-means algorithm to the panel regression framework with latent group structures, but the asymptotic properties of the estimators and the procedures are not provided. Bonhomme and Manresa (2015) and Ando and Bai (2016) modify the K-means algorithm

to estimate the time-varying grouped patterns of heterogeneity and unobserved group interactive fixed effects, respectively. Wang et al. (2018) extend the CARDS (clustering algorithm in regression via data-driven segmentation) method of Ke et al. (2015) to panel structure models where the latent group structures exist in vectors of slope parameters. Recently, Liu et al. (2020) extend the modified K-means algorithm of Bonhomme and Manresa (2015) to estimate and identify the latent group structures in panel data. Wang and Su (2020) extend the sequential binary segmentation algorithm of Bai (1997) for break detection from the time series setup to the panel data framework to identify the latent group structures.

While these methods make important contributions by empirically estimating the group identities, they face the following limitations. First, to implement them, one often needs to determine the number of groups first. Consequently, the estimation error often accumulates across the two steps and leads to suboptimal performance. Second, C-Lasso procedure of Su et al. (2016) is not a convex problem[1], requires the number of groups to be fixed, and may leave some individuals unclassified. Third, K-means algorithm has been shown to be NP-hard, can get trapped in suboptimal local minima, and is sensitive to the choice of initial estimators. Fourth, the CARDS method of Wang et al. (2018) relies on the specification of at least three tuning parameters, thus the consistency results are sensitive to the choice of the tuning parameters. Fifth, Wang and Su (2020) and Wang et al. (2018) rely on ordered segmentations to identifying the latent group structure and construct the Lasso-type penalties, respectively, which are sensitive to the choice of initial estimators, and often it may be difficult to construct one. The objective of this chapter is to provide a new

[1]However, the numerical solution can be transformed into a sequence of convex problems.

framework free of the above limitations to jointly estimate and identify the latent group structures without a priori knowledge of classification or a natural basis for separating slope coefficients into groups.

Inspired by the adaptive group fused Lasso of Qian and Su (2016), and the pairwise fusion concave penalty of Ma and Huang (2017), we propose a penalized procedure with a pairwise fusion penalty to automatically estimate and identify homogenous groups where both the number of groups and the individual group identities are unknown. Our method and mainly our model is different from theirs in several important aspects. Qian and Su (2016) consider estimation and inference of common structural breaks in panel data models using an adaptive group fused Lasso. Their method cannot be used to classify individuals into different groups because there is no natural ordering across individuals, also a different algorithm to locate common individuals is required. Ma and Huang (2017) consider the problem of identifying subgroups among observations, using a concave pairwise fusion penalty. Clearly, their model is different from the model considered here to estimate and identify the latent group structures. Besides, the penalty term in Ma and Huang (2017) is imposed through concave penalties such as the SCAD (smoothly clipped absolute deviations penalty) of Fan and Li (2001) and the MCP (minimax concave penalty) of Zhang (2010), but our penalty is imposed through an adaptive group fused Lasso. The other main difference of our penalty from theirs lies in two aspects: 1) we impose the penalty on slope vector differences, whereas their method applies the penalty on the intercepts, 2) we assign different weights $\{\dot{w}_{ij}\}$, based on preliminary estimates of the slope parameters to penalize different coefficient differences, however these weights are not feasible in their

study. Since our proposed framework utilizes a pairwise adaptive group fused Lasso penalty, we denote our estimation procedure as PAGFL. To implement our method, we derive an ADMM (alternating direction method of multipliers) algorithm (Boyd et al. (2011)), and show the convergence properties of our ADMM algorithm.

We develop two classes of estimators for panel structure models to estimate the slope parameters: penalized least squares (PLS) and penalized generalized method of moments (PGMM). The PLS can be applied to static or dynamic panel models without endogenous regressors, while the PGMM is suitable for panel models with endogeneity or dynamic structures. We show that the PLS method is an oracle procedure (using the language of Fan and Li (2001)), in the sense that the PLS estimator classifies the right individuals in the right groups (classification consistency), and asymptotically is equivalent to the oracle estimator. The oracle estimator is obtained from least squares regression by assuming that the true group structure is known. Similarly, our PGMM estimator satisfies the classification consistency, but its oracle property does not hold generally. Our asymptotic results hold under $(N, T) \to \infty$ jointly, but T can pass to infinity at a slower rate, where T is the time series dimension, and N is the cross-section dimension. Moreover, our proposed method, compared to the existing methods in the literature, has several advantages in the following characteristics. First, the major contribution of our method is that it asymptotically identifies the true structure while estimating the model parameters consistently without relying on correct initial estimates of the number of groups. This implies that our estimation and classification consistency results hold without relying on correct estimation of the number of groups. It is of crucial importance, as in most empirical

research the number of groups is often unknown to practitioners. Second, unlike the K-means algorithm and C-Lasso method, our proposed approach allows the number of groups and the number of individuals within each group to be either divergent or fixed. This makes our method applicable to a large body of applications. Third, unlike K-means algorithm and C-Lasso method, our method admits a simple and fast iterative algorithm that is guaranteed to converge to the unique global minimizer. Therefore, the computation burden of our approach is not as much as the K-means algorithm and the C-Lasso. Fourth, unlike the CARDS method, our approach only requires a tuning parameter and does not rely on the ordered segmentations. Fifth, our method continues to perform well even if the number of groups is allowed to increase with the number of cross-sections, N.

The reminder of this chapter is organized as follows. Section 5.2 describes our fixed effect panel model, PLS and PGMM estimation methods depending on whether the regressors are endogenous. Sections 5.3 and 5.4 analyze the asymptotic properties of PLS and PGMM estimators, respectively. Section 5.5 presents the computation and algorithm. Monte Carlo results are given is section 5.6. In Section 5.7, we apply the estimators to a simple model of inter-temporal dynamics of the unemployment rate in the U.S. states, and to forecasting quarterly output growth rates across 33 countries using macro and financial variables. Conclusions and final remarks are given is section 5.8. Proofs and detailed calculations are listed in Appendices E–G.

A brief word on notation: For an $m \times n$ real matrix A, we write the transpose A', the Frobenius norm $\|A\| = (\text{tr}(AA'))^{1/2}$. When A is symmetric, we use $\mu_{max}(A)$ and $\mu_{min}(A)$ to denote the largest and smallest eigenvalues, respectively. I_p and $\mathbf{0}_{p\times 1}$ denote $p \times p$ identity

matrix and $p \times 1$ vector of zeros. $\mathbf{1}(.)$ denotes the indicator function and "p.d." abbreviates "positive definite". The operators $\xrightarrow{p}$, $\xrightarrow{D}$, and *plim* denote respectively, convergence in probability, convergence in distribution, and probability limit. We use $(N,T) \to \infty$ to signify that N and T pass jointly to infinity.

5.2 Model and Penalized Estimation

In this section, we consider a linear panel structure model with unknown group membership.

5.2.1 The Model

Consider the following linear panel data model

$$y_{it} = \beta_i^{0\prime} x_{it} + \eta_i + u_{it}, \qquad i = 1, \ldots, N, \quad t = 1, \ldots, T, \tag{5.1}$$

where y_{it} is the dependent variable, x_{it} is a $p \times 1$ vector of regressors explaining y_{it}, η_i is the individual fixed effect that may be correlated with the regressors, u_{it} is the idiosyncratic error term with zero mean, T is the number of observations, and N is the number of individual units. We assume that β_i^0 is a $p \times 1$ vector of slope parameters that admits a possible grouping structure of the form

$$\beta_i^0 = \begin{cases} \alpha_1^0, & \text{if } i \in G_1^0 \\ \vdots & \vdots \\ \alpha_{K_0}^0, & \text{if } i \in G_{K_0}^0, \end{cases} \tag{5.2}$$

where $\alpha_l^0 \neq \alpha_k^0$ for any $l, k = 1, \ldots, K_0$, and $l \neq k$, and $G = \{G_1^0, G_2^0, \ldots, G_{K_0}^0\}$ forms a partition of $\{1, 2, \ldots, N\}$. Let N_k be the number of individual units in G_k^0, and the $pK_0 \times 1$ matrix of $\boldsymbol{\alpha}$, and the $pN \times 1$ matrix $\boldsymbol{\beta}$ be defined as

$$\boldsymbol{\alpha} = (\alpha_1', \alpha_2', \ldots, \alpha_{K_0}')' \quad \text{and} \quad \boldsymbol{\beta} = (\beta_1', \beta_2', \ldots, \beta_N')', \tag{5.3}$$

where $\boldsymbol{\alpha}^0$ and $\boldsymbol{\beta}^0$ denote the true values of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. In practice, the number of groups, K_0, is unknown. However, it is usually reasonable to assume that K_0 is smaller than N. Our goal is to estimate the regression coefficient $\boldsymbol{\alpha}^0$ and identify the latent group structure.

We consider two cases about the exogeneity or endogeneity of the regressors:

(a) $\mathbb{E}(x_{is} u_{it}) = 0$, for all $1 \leq s \leq t \leq T$;

(b) $\mathbb{E}(x_{it} u_{it}) \neq 0$, for $t = 1, \ldots, T$.

The first case occurs when the regressors are weakly exogenous and allows for lagged values of y_{it} to be included in x_{it}, so that least squares criteria are appropriate. The second case happens when the regressors contain either lagged dependent variables or endogenous regressors that are correlated with the error term. In this case, we assume there exists a $q \times 1$ vector of instruments z_{it} with $q \geq p$.

Since the individual effects, η_i, are not of main interest, in case (a), we concentrate them out and obtain the following equation

$$\tilde{y}_{it} = \beta_i^{0\prime} \tilde{x}_{it} + \tilde{u}_{it}, \qquad i = 1, \ldots, N, \quad t = 1, \ldots, T, \tag{5.4}$$

where, e.g., $\tilde{x}_{it} = x_{it} - T^{-1} \sum_{t=1}^{T} x_{it}$. In case (b), to eliminate the effect of μ_i in the estimation procedure, we consider the first-differenced equation

$$\Delta y_{it} = \beta_i^{0\prime} \Delta x_{it} + \Delta u_{it}, \tag{5.5}$$

where, e.g., $\Delta y_{it} = y_{it} - y_{i,t-1}$ for $i = 1, \ldots, N$, and $t = 1, ..., T$, by assuming that we have observations on y_{i0} and x_{i0}.

5.2.2 Penalized Least Squares (PLS) Estimation

To estimate the model in (5.4) under case (a), we propose minimizing the following objective function

$$Q_{1,NT}(\boldsymbol{\beta}, \lambda_1) = \frac{1}{T}\sum_{i=1}^{N}\sum_{t=1}^{T}(\tilde{y}_{it} - \beta_i'\tilde{x}_{it})^2 + \frac{\lambda_1}{N}\sum_{1\leq i<j\leq N}\sum\dot{w}_{ij}\|\beta_i - \beta_j\|, \tag{5.6}$$

where $\lambda_1 \geq 0$ is a tuning parameter, and $\dot{w}_{ij}$ is a data-driven weight defined by

$$\dot{w}_{ij} = \|\dot{\beta}_i - \dot{\beta}_j\|^{-\kappa}, \quad \text{for } i, j = 1, \ldots, N, \tag{5.7}$$

$\dot{\beta}_i$ and $\dot{\beta}_j$ are preliminary estimates of β_i and β_j, respectively, and κ is a user-specified positive constant that usually takes value 2 in the literature of adaptive Lasso.

To obtain the adaptive weights $\{\dot{w}_{ij} : i, j \in \{1, \ldots, N\}\}$, we propose to obtain the preliminary estimate $\dot{\boldsymbol{\beta}}$ by minimizing the first term in equation (5.6) which results in the ordinary least squares. Thus for the i-th element of $\dot{\boldsymbol{\beta}}$, we have

$$\dot{\beta}_i = \Big(\sum_{t=1}^{T}\tilde{x}_{it}\tilde{x}_{it}'\Big)^{-1}\sum_{t=1}^{T}\tilde{x}_{it}\tilde{y}_{it}. \tag{5.8}$$

The objective function in (5.6) is related to the literature on adaptive Lasso (Zou (2006)), group Lasso (Yuan and Lin (2006)), fused Lasso (Tibshirani et al. (2005)) and group fused Lasso (Qian and Su (2016)). Qian and Su (2016) determine the unknown number of structural breaks which is different from the purpose of this chapter. The other listed papers above aim at determining the nonzero coefficients from the zero ones, and are not applicable here because our aim is to determine the unknown group structure.

It is worth emphasizing that minimization of (5.6) is a convex optimization problem, and thus it does not suffer from the multiple local minima issue, and its global minimizer can be efficiently solved. The penalty shrinks some of the pairs $\beta_i - \beta_j$ to zero, so that we can partition the slope parameters into groups. In practice, let $\hat{\lambda}_1$ be the value of the tuning parameter that we select based on a variant of the Bayesian information criterion (BIC), further let $\{\hat{\alpha}_1, \ldots, \hat{\alpha}_{\hat{K}}\}$ be the distinct values of the PLS estimator $\hat{\boldsymbol{\beta}} \equiv \hat{\boldsymbol{\beta}}(\hat{\lambda}_1) = \arg\min Q_{1,NT}(\boldsymbol{\beta}, \hat{\lambda}_1)$, then $\{\hat{G}_1, \ldots, \hat{G}_{\hat{K}}\}$ forms a partition of $\{1, 2, \ldots, N\}$, where $\hat{G}_k = \{i : \hat{\beta}_i = \hat{\alpha}_k, 1 \leq i \leq N\}$, $1 \leq k \leq \hat{K}$.

5.2.3 Penalized GMM (PGMM) Estimation

In case (b), we propose to estimate $\boldsymbol{\beta}$ by minimizing the following objective function

$$\begin{aligned} Q_{2,NT}(\boldsymbol{\beta}, \lambda_2) = &\sum_{i=1}^{N} \Big[\frac{1}{T}\sum_{t=1}^{T} z_{it}(\Delta y_{it} - \beta_i' \Delta x_{it})\Big]' W_{i,NT} \Big[\frac{1}{T}\sum_{t=1}^{T} z_{it}(\Delta y_{it} - \beta_i' \Delta x_{it})\Big] \\ &+ \frac{\lambda_2}{N} \sum_{1 \leq i < j \leq N} \sum \ddot{w}_{ij} \|\beta_i - \beta_j\|, \end{aligned} \tag{5.9}$$

where $\lambda_2 \geq 0$ is a tuning parameter, $W_{i,NT}$ is a $q \times q$ p.d. matrix, and $\ddot{w}_{ij}$ is a data-driven weight defined by

$$\ddot{w}_{ij} = \|\ddot{\beta}_i - \ddot{\beta}_j\|^{-\kappa}, \quad \text{for } i, j = 1, \ldots, N, \tag{5.10}$$

$\ddot{\beta}_i$ and $\ddot{\beta}_j$ are preliminary estimates of β_i and β_j, respectively, and κ is a user-specified positive constant that usually takes value 2 in the literature.

To obtain the adaptive weights $\{\ddot{w}_{ij} : i, j \in \{1, \ldots, N\}\}$, we propose to obtain the preliminary estimate $\ddot{\boldsymbol{\beta}}$ by minimizing the first term in equation (5.9). Thus, for the i-th element of $\ddot{\boldsymbol{\beta}}$, we have

$$\ddot{\beta}_i = \Big[\Big(\frac{1}{T}\sum_{t=1}^{T}\Delta x_{it} z_{it}'\Big) W_{i,NT} \Big(\frac{1}{T}\sum_{t=1}^{T} z_{it}\Delta x_{it}'\Big)\Big]^{-1} \Big(\frac{1}{T}\sum_{t=1}^{T}\Delta x_{it} z_{it}'\Big) W_{i,NT} \Big(\frac{1}{T}\sum_{t=1}^{T} z_{it}\Delta y_{it}\Big). \tag{5.11}$$

The first term in the definition of the objective function in (5.9) is different from the usual GMM objective function in the panel setting where only one weight matrix is needed and the double summation $\sum_{i=1}^{N}\sum_{t=1}^{T}$ occurs twice, one before the weight and the other after the weight matrix. The reason is that because the true group membership of individual units is unknown, we cannot apply the usual GMM objective function here.

It is worth emphasizing that minimization of (5.9) is also a convex optimization problem, hence it does not suffer from the multiple local minima issue, and its global minimizer can be efficiently solved. The penalty shrinks some of the pairs $\beta_i - \beta_j$ to zero, so that we can partition the slope parameters into groups. In practice, let $\tilde{\lambda}_2$ be the value of the tanning parameter that we select based on a variant of BIC, further let $\{\tilde{\alpha}_1, \ldots, \tilde{\alpha}_{\tilde{K}}\}$ be the distinct values of the penalized generalized method of moments (PGMM) estimator $\tilde{\boldsymbol{\beta}} \equiv \tilde{\boldsymbol{\beta}}(\tilde{\lambda}_2) = \arg\min Q_{2,NT}(\boldsymbol{\beta}, \tilde{\lambda}_2)$, then $\{\tilde{G}_1, \ldots, \tilde{G}_{\tilde{K}}\}$ forms a partition of $\{1, 2, \ldots, N\}$, where $\tilde{G}_k = \{i : \tilde{\beta}_i = \tilde{\alpha}_k, 1 \le i \le N\}$, $1 \le k \le \tilde{K}$.

5.3 Asymptotic properties of the PLS estimators

In this section, we address the asymptotic properties of the PLS estimator and the associated post-Lasso estimator.

5.3.1 Assumptions

Let $\hat{Q}_{i,\tilde{x}\tilde{x}} = \frac{1}{T}\sum_{t=1}^{T}\tilde{x}_{it}\tilde{x}_{it}'$ and $\hat{Q}_{i,\tilde{x}\tilde{u}} = \frac{1}{T}\sum_{t=1}^{T}\tilde{x}_{it}\tilde{u}_{it}'$, and define $J_{min} = \min_{1\leq l\&k\leq K^0}\|\alpha_l^0 - \alpha_k^0\|$ which denotes the minimum degree of heterogeneity in the slope coefficients between groups.

To study the asymptotic properties of the PLS estimator, denoted by $\hat{\boldsymbol{\beta}}$, we make the following assumptions.

Assumption 5.1 *(i)* $\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\tilde{x}_{it}\tilde{u}_{it} = O_p(1)$ *for each* $i = 1,\ldots,N$.

(ii) $\hat{Q}_{i,\tilde{x}\tilde{x}} \xrightarrow{P} Q_{i,\tilde{x}\tilde{x}} > 0$ *for each* $i = 1,\ldots,N$. *There exists a positive constant* $\underline{c}_{\tilde{x}\tilde{x}}$ *such that* $\lim_{(N,T)\to\infty}\min_{1\leq i\leq N}\mu_{min}(\hat{Q}_{i,\tilde{x}\tilde{x}}) \geq \underline{c}_{\tilde{x}\tilde{x}}$.

(iii) $\frac{1}{N}\sum_{i=1}^{N}\|\hat{Q}_{i,\tilde{x}\tilde{u}}\|^2 = O_p(T^{-1})$.

(vi) $N_k/N \to \tau_k \in [0,1)$ *for each* $k = 1,\ldots,K_0$ *as* $N\to\infty$.

Assumption 5.2 *(i)* $T^{1/2}J_{min} \to c_J \in (0,\infty]$ *as* $(N,T)\to\infty$.

(ii) $plim_{(N,T)\to\infty}NT^{1/2}\lambda_1 J_{min}^{-\kappa} = c \in [0,\infty)$.

(iii) $plim_{(N,T)\to\infty}N_k T^{(\kappa+1)/2}\lambda_1/N = \infty$.

Assumption 5.3 *(i) For each* $k = 1,\ldots,K_0$, $\bar{\Phi}_k \equiv \frac{1}{N_k T}\sum_{i\in G_k^0}\sum_{t=1}^{T}\tilde{x}_{it}\tilde{x}_{it}' \xrightarrow{P} \Phi_k > 0$ *as* $(N,T)\to\infty$.

(ii) For each $k = 1,\ldots,K_0$, $\frac{1}{\sqrt{N_k T}}\sum_{i\in G_k^0}\sum_{t=1}^{T}\tilde{x}_{it}\tilde{u}_{it} - \mathbb{B}_{k,NT} \xrightarrow{D} N(0,\Psi_k)$ *as* $(N,T)\to\infty$, *where* $\mathbb{B}_{k,NT} = \frac{1}{\sqrt{N_k T}}\sum_{i\in G_k^0}\sum_{t=1}^{T}\mathbb{E}(\tilde{x}_{it}\tilde{u}_{it})$ *is either zero or of order* $O(\sqrt{N_k/T})$ *depending on whether* x_{it} *is strictly exogenous.*

Assumption 5.1(i) will be mostly satisfied in large dimensional panel data models with weakly exogenous regressors and can be replaced with sufficient or primitive conditions on the process $\{(x_{it}, u_{it}), t \geq 1\}$ that ensure the central limit theory. Note that this assumption allows both conditional heteroscedasticity and serial correlation in $\{u_{it}, t \geq 1\}$. Also, Assumption 5.1(iii) can be easily verified from this assumption. The first part of Assumption 5.1(ii) is standard in the literature, but the second one imposes restriction on the moments of x_{it}, the dependence structure on the regressors processes, and the relative rates at which N and T pass to infinity. Su et al. (2016) give details on sufficient and primitive conditions that ensure this assumption. Assumption 5.1(iv) implies that as $N \to \infty$, the number of individuals within each group can be either asymptotically non-negligible or tend to infinity but at a rate slower than N. Assumption 5.2 mainly specifies conditions on J_{min}, λ_1, N, and T. We use the probability limit in 5.2(ii)-(iii) because we allow λ_1 to be data-driven and hence random. We assume the minimum degree of heterogeneity size, J_{min}, to shrink to zero as $T \to \infty$, but at a rate slower than $T^{-1/2}$. We make Assumption 5.3 to provide conditions to ensure the asymptotic normality of the Lasso estimators, but it can be replaced with various commonly primitive conditions.

5.3.2 Consistency

The following theorem establishes the consistency of $\hat{\beta}_i$ for $i = 1, \ldots, N$.

Theorem 5.4 *Suppose that Assumption 5.1 holds. Then for* $i = 1, \ldots, N$,

(i) $\hat{\beta}_i - \beta_i^0 = O_p(T^{-1/2})$,

(ii) $\frac{1}{N} \sum_{i=1}^{N} \|\hat{\beta}_i - \beta_i^0\|^2 = O_p(T^{-1})$.

Proof: Appendix E, (See page 174).

Theorem 5.4 (i) and (ii), respectively, establish the pointwise and mean square convergence rates of $\hat{\beta}_i$. Given the PLS estimate, $\hat{\boldsymbol{\beta}}$, we can obtain the estimated groups by classifying individuals with the same coefficient estimate, $\hat{\beta}_i$, into the same group. Let $\hat{G}_k$, $k = 1, \ldots, \hat{K}$ denote the $\hat{K}$ estimated groups. Let $\hat{\alpha}_k$, $k = 1, \ldots, \hat{K}$ denote the group-specific estimated slope coefficients. Then by definition:

$$\hat{G}_k = \{i \in 1, \ldots, N : \hat{\beta}_i = \hat{\alpha}_k\}, \text{ for } k = 1, \ldots, \hat{K}. \tag{5.12}$$

The following theorem establishes the classification consistency.

Theorem 5.5 *Suppose that Assumptions 5.1 and 5.2 hold. Then*

$$P\Big(\|\hat{\beta}_i - \hat{\beta}_j\| = 0 \text{ for all } i \,\&\, j \in G_k^0,\ k \in \{1, \ldots, K_0\}\Big) \to 1, \text{ as } T \to \infty.$$

Proof: Appendix E, (See page 175).

Theorem 5.5 says that with probability approaching one all the zero vectors in $\{\|\beta_i - \beta_j\|, 1 \leq i, j \leq N\}$ must be estimated as exactly zero by the PLS method so that the estimated number of groups cannot be large than K_0 when T is sufficiently large. These results together with the consistency results in Theorem 5.4 imply that the PAGFL has the ability to identify the true group structure with the correct number of individual units within each group consistently when the minimum group size J_{min} does not shrink to zero too fast.

Corollary 5.6 *Suppose that Assumptions 5.1 and 5.2 hold with $c_J = \infty$ in assumption 5.2(i). Then*

i) $\lim_{N\to\infty} P(\hat{K} = K_0) = 1$,

ii) $\lim_{N\to\infty} P(\hat{G}_1 = G_1^0, \ldots, \hat{G}_K = G_K^0) = 1$.

Proof: Appendix E, (See page 177).

The above corollary implies that, as long as the minimum degree of heterogeneity, J_{min}, remains fixed or shrinks to zero at a rate slower than $T^{-1/2}$ as $T \to \infty$, we can determine the correct number of groups.

5.3.3 Limiting Distribution and The Oracle Property of PLS

In this section we study the asymptotic distribution of the PLS and post-Lasso estimators.

Note that if each individual's group membership is known, the oracle estimator is the within group estimator of α_k^0 which can be formulated as $\bar{\alpha}_k = \left(\sum_{i\in G_k^0}\sum_{t=1}^{T}\tilde{x}_{it}\tilde{x}_{it}'\right)^{-1}\sum_{i\in G_k^0}\sum_{t=1}^{T}\tilde{x}_{it}\tilde{y}_{it}$. Also, note that under assumption 5.3, $\sqrt{N_kT}(\bar{\alpha}_k - \alpha_k^0) - \bar{\Phi}_k^{-1}\mathbb{B}_{k,NT} \xrightarrow{D} N(0, \Phi_k^{-1}\Psi_k\Phi_k^{-1})$.

The following theorem reports the limiting distribution of the PAGFL estimator, $\hat{\alpha}_k$, which is derived from the PLS estimates of $\hat{\boldsymbol{\beta}}$ after the classification.

Theorem 5.7 *Suppose that Assumptions 5.1–5.3 hold with $c_J = \infty$ in assumption 5.2(i). Then, conditional on $\hat{K} = K_0$,*

$$\sqrt{N_kT}(\hat{\alpha}_k - \alpha_k^0) - \bar{\Phi}_k^{-1}\mathbb{B}_{k,NT} \xrightarrow{D} N(0, \Phi_k^{-1}\Psi_k\Phi_k^{-1}), \text{ for } k = 1, \ldots, K_0.$$

Proof: Appendix E, (See page 177).

Theorem 5.7 indicates that the PLS estimator of $\hat{\alpha}_k$ achieves the same limiting distribution as the oracle within group estimator, therefore we say that the PLS estimator has the asymptotic oracle property.

Given the fact that we estimate the group structure, we can define the post-Lasso estimator of $\hat{\alpha}_k$ as

$$\hat{\alpha}_{\hat{G}_k} = \Big(\sum_{i \in \hat{G}_k} \sum_{t=1}^{T} \tilde{x}_{it} \tilde{x}_{it}' \Big)^{-1} \sum_{i \in \hat{G}_k} \sum_{t=1}^{T} \tilde{x}_{it} \tilde{y}_{it}. \tag{5.13}$$

The following theorem reports the asymptotic distribution of $\hat{\alpha}_{\hat{G}_k}$.

Theorem 5.8 *Suppose that Assumptions 5.1–5.3 hold with $c_J = \infty$ in assumption 5.2(i). Then, conditional on $\hat{K} = K_0$,*

$$\sqrt{N_k T}(\hat{\alpha}_{\hat{G}_k} - \alpha_k^0) - \bar{\Phi}_k^{-1} \mathbb{B}_{k,NT} \xrightarrow{D} N(0, \Phi_k^{-1} \Psi_k \Phi_k^{-1}), \text{ for } k = 1, \ldots, K_0.$$

Proof: Appendix E, (See page 178).

The above theorem holds using the classification consistency results in Theorem 5.5, and says that the post-Lasso estimator has the asymptotic oracle property. Although, the Lasso and post-Lasso estimators are asymptotically equivalent, it is well known that the post-Lasso estimator typically performs better than the Lasso estimator in terms of faster rates of convergency (see Belloni and Chernozhukov (2013)), thus it is recommended for practical use. Moreover, $\mathbb{B}_{k,NT}$ is not equal to zero in case of dynamic panel data models, in fact it is well known in the literature that the fixed effect estimator has asymptotically bias of order $O(1/T)$. This suggests that in dynamic panel models $\mathbb{B}_{k,NT} = O(\sqrt{N_k/T})$ and bias correction is required, unless the rate at which T goes to infinity is faster than that of N_k. There are various methods proposed in the literature to estimate the bias term

such as Kiviet (1995), Hahn and Kuersteiner (2002), Phillips and Sul (2007), Lee (2012), Gourieroux et al. (2010) and Han et al. (2014), among others, and we refer the readers to these papers.

5.4 Asymptotic properties of the PGMM estimators

In this section we address the asymptotic properties of the PGMM estimator and the associated post-Lasso estimator.

5.4.1 Assumptions

Let $\tilde{Q}_{i,z\Delta x} = \frac{1}{T}\sum_{t=1}^{T} z_{it}\Delta x_{it}'$, $\tilde{Q}_{i,z\Delta y} = \frac{1}{T}\sum_{t=1}^{T} z_{it}\Delta y_{it}$, $\bar{Q}_{i,z\Delta x} = \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}(z_{it}\Delta x_{it}')$, and $\bar{Q}_{i,z\Delta y} = \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}(z_{it}\Delta y_{it})$. Let $\xi_{it} = (\Delta y_{it}, (\Delta x_{it})', z_{it}')'$, $\rho(\xi_{it},\beta_i) = z_{it}(\Delta y_{it} - \beta_i'\Delta x_{it})$, and $\bar{\rho}_{i,T}(\beta_i) = \frac{1}{\sqrt{T}}\sum_{t=1}^{T}\Big[\rho(\xi_{it},\beta_i) - \mathbb{E}(\rho(\xi_{it},\beta_i))\Big]$. Also, for each group $k = 1,\ldots,K_0$, let $W_{NT}^{(k)}$ be a $d\times d$ p.d. matrix, $Q_{z\Delta x,NT}^{(k)} = \Big(\frac{1}{N_k T}\sum_{i\in G_k^0}\sum_{t=1}^{T} z_{it}(\Delta x_{it})'\Big)$.

To study the asymptotic properties of the PGMM estimator, denoted by $\tilde{\boldsymbol{\beta}}$, we make the following assumptions.

Assumption 5.9 *(i)* $\mathbb{E}(\rho(\xi_{it},\beta_i^0)) = 0$, *for each* $i = 1,\ldots,N$ *and* $t = 1,\ldots,T$.

(ii) $\sup_{\beta_i}\|\bar{\rho}_{i,T}(\beta_i)\| = O_p(1)$, *and* $\frac{1}{N}\sum_{i=1}^{N}\|\bar{\rho}_{i,T}(\beta_i)\|^2 = O_p(1)$, *for any* β_i *and* $i = 1,\ldots,N$.

(iii) $\tilde{Q}_{i,z\Delta x} \xrightarrow{P} \bar{Q}_{i,z\Delta x} > 0$, *for each* $i = 1,\ldots,N$. *There exists a positive constant* $\underline{c}_{\bar{Q}}$ *such that* $\lim_{(N,T)\to\infty}\min_{1\le i\le N}\mu_{min}(\bar{Q}_{i,z\Delta x}'\bar{Q}_{i,z\Delta x}) = \underline{c}_{\bar{Q}}$.

(iv) *There exist non-random matrices* W_i *such that* $\max_{1\leq i\leq N}\|W_{i,NT}-W_i\|=o_p(1)$, *and* $\liminf_{(N,T)\to\infty}\min_{1\leq i\leq N}\mu_{min}(W_i)=\underline{c}_W>0$.

(v) $N_k/N\to\tau_k\in[0,1)$, *for each* $k=1,\ldots,K_0$ *as* $N\to\infty$.

Assumption 5.10 *(i)* $T^{1/2}J_{min}\to c_J\in(0,\infty]$ *as* $(N,T)\to\infty$.

(ii) $plim_{(N,T)\to\infty}NT^{1/2}\lambda_1 J_{min}^{-\kappa}=c\in[0,\infty)$.

(iii) $plim_{(N,T)\to\infty}N_kT^{(\kappa+1)/2}\lambda_2/N=\infty$.

Assumption 5.11 *(i)* *For each* $k=1,\ldots,K_0$, $\bar{\Phi}_k\equiv\frac{1}{N_k}\sum_{i\in G_k^0}\|\tilde{Q}_{i,z\Delta x}-\bar{Q}_{i,z\Delta x}\|^2=o_p(1)$, *and* $W_{i,NT}\xrightarrow{p}W_i>0$ *for* $i\in G_k^0$.

(ii) *For each* $k=1,\ldots,K_0$, $\bar{A}_k\equiv\frac{1}{N_k}\sum_{i\in G_k^0}\bar{Q}'_{i,z\Delta x}W_{i,NT}\bar{Q}_{i,z\Delta x}\to A_k>0$ *as* $(N,T)\to\infty$.

(iii) *For each* $k=1,\ldots,K_0$, $\frac{1}{\sqrt{N_kT}}\sum_{i\in G_k^0}\tilde{Q}'_{i,z\Delta x}W_{i,NT}\sum_{t=1}^T z_{it}\Delta u_{it}-\mathbb{B}_{k,NT}\xrightarrow{D}N(0,C_k)$ *as* $(N,T)\to\infty$.

Assumption 5.12 *(i)* *For each* $k=1,\ldots,K_0$, $W_{NT}^{(k)}\xrightarrow{p}W^{(k)}>0$ *as* $(N,T)\to\infty$.

(ii) $Q_{z\Delta x,NT}^{(k)}\xrightarrow{p}Q_{z\Delta x}^{(k)}$ *which has rank* p.

(iii) $\frac{1}{\sqrt{N_kT}}\sum_{i\in G_k^0}\sum_{t=1}^T z_{it}\Delta u_{it}\xrightarrow{D}N(0,V_k)$.

Assumption 5.9 (i) specifies moment conditions to identify β_i^0. 5.9 (ii) is needed as we do not specify the data structure. 5.9 (iii) together with 5.9 (i) provide a rank condition for the identification. 5.12 is a standard assumption in GMM estimation literature. The rest of the assumptions parallel Assumptions 5.1–5.3.

5.4.2 Consistency

The following theorem establishes the consistency of the PGMM estimator, $\tilde{\beta}_i$ for $i = 1, \ldots, N$.

Theorem 5.13 *Suppose that Assumption 5.9 holds. Then*

(i) $\tilde{\beta}_i - \beta_i^0 = O_p(T^{-1/2})$ *for* $i = 1, \ldots, N$,

(ii) $\frac{1}{N}\sum_{i=1}^{N} \|\tilde{\beta}_i - \beta_i^0\|^2 = O_p(T^{-1})$.

Proof: Appendix F, (See page 180).

Theorem 5.13 (i) and (ii), respectively, establish the pointwise and mean square convergence rates of $\{\tilde{\beta}_i : i = 1, \ldots, N\}$. Given the PGMM estimates, $\tilde{\boldsymbol{\beta}}$, we can obtain the estimated groups by classifying individuals with the same coefficient estimate $\tilde{\beta}_i$ into the same group. Let $\hat{G}_k$, $k = 1, \ldots, \tilde{K}$ denote the $\tilde{K}$ estimated groups. Let $\tilde{\alpha}_k$, $k = 1, \ldots, \tilde{K}$ denote the group-specific estimated slope coefficients. Then, by definition:

$$\tilde{G}_k = \{i \in 1, \ldots, N : \tilde{\beta}_i = \tilde{\alpha}_k\}, \text{ for } k = 1, \ldots, \tilde{K}. \tag{5.14}$$

The following theorem establishes the classification consistency.

Theorem 5.14 *Suppose Assumptions 5.9 and 5.10 hold. Then*

$$P\Big(\|\tilde{\beta}_i - \tilde{\beta}_j\| = 0 \text{ for all } i \,\&\, j \in G_k^0,\ k \in \{1, \ldots, K_0\}\Big) \to 1, \text{ as } T \to \infty.$$

Proof: Appendix F, (See page 181).

Theorem 5.14 says that with probability approaching one all the zero vectors in $\{\|\beta_i - \beta_j\|, 1 \leq i, j \leq N\}$ must be estimated as exactly zero by the PGMM method so that the

estimated number of groups cannot be different from K_0 when T is sufficiently large. These results together with the consistency results in Theorem 5.13 imply that the PAGFL has the ability to identify the true group structure with the correct number of individual units within each group consistently when the minimum degree of heterogeneity, J_{min}, does not shrink to zero too fast.

Corollary 5.15 *Suppose that Assumptions 5.9 and 5.10 hold with $c_J = \infty$ in Assumption 5.10(i). Then*

(i) $\lim_{N\to\infty} P(\tilde{K} = K_0) = 1$,

(ii) $\lim_{N\to\infty} P(\tilde{G}_1 = G_1^0, \ldots, \tilde{G}_K = G_K^0) = 1$.

Proof: Appendix F, (See page 183).

The above corollary implies that, as long as J_{min} remains fixed or shrinks to zero at a rate slower than $T^{-1/2}$ as $T \to \infty$, we can determine the correct number of groups.

5.4.3 Limiting Distribution of PGMM

In this section we study the asymptotic distribution of the PGMM and post-Lasso estimators.

Note that if each individual's group membership is known, the oracle estimator is the solution to a usual GMM objective function which can be formulated as below

$$\breve{\alpha}_k = \left[Q_{z\Delta x,NT}^{(k)\prime} W_{NT}^{(k)} Q_{z\Delta x,NT}^{(k)} \right]^{-1} Q_{z\Delta x,NT}^{(k)\prime} W_{NT}^{(k)} Q_{z\Delta y,NT}^{(k)}, \tag{5.15}$$

where $Q_{z\Delta x,NT}^{(k)} = \left(\frac{1}{N_k T} \sum_{i\in G_k^0} \sum_{t=1}^{T} z_{it}(\Delta x_{it})' \right)$, $Q_{z\Delta y,NT}^{(k)} = \left(\frac{1}{N_k T} \sum_{i\in G_k^0} \sum_{t=1}^{T} z_{it}(\Delta y_{it}) \right)$ and $W_{NT}^{(k)}$ is a $q \times q$ symmetric p.d. matrix for each $k = 1, \ldots, K_0$. Apparently, the

PGMM estimator does not have the same asymptotic distribution under general conditions. However, under the further assumptions that for each $i \in G_k^0$, $W_{i,NT} = W_{NT}^{(k)}$, $\bar{Q}_{i,z\Delta x} = Q_{z\Delta x,NT}^{(k)}$, and $\mathbb{B}_{k,NT} = 0$ the PGMM estimator will have the oracle property.

The following theorem reports the limiting distribution of the PAGFL estimator, $\tilde{\alpha}_k$, which is derived from the PGMM estimates, $\tilde{\boldsymbol{\beta}}$, after the classification.

Theorem 5.16 *Suppose Assumptions 5.9–5.11 hold with $c_J = \infty$ in assumption 5.10(i). Then, conditional on $\tilde{K} = K_0$,*

$$\sqrt{N_k T}(\tilde{\alpha}_k - \alpha_k^0) - \bar{A}_k^{-1}\mathbb{B}_{k,NT} \xrightarrow{D} N(0, A_k^{-1}C_k A_k^{-1}), \text{ for } k = 1, \ldots, K_0.$$

Proof: Appendix F, (See page 183).

Given the fact that we estimate the grouping structure, we can define the post-Lasso estimator of $\tilde{\alpha}_k$ as

$$\tilde{\alpha}_{\tilde{G}_k} = \left(\tilde{Q}_{z\Delta x}^{(k)\prime} W_{NT}^{(k)} \tilde{Q}_{z\Delta x}^{(k)}\right)^{-1} \tilde{Q}_{z\Delta x}^{(k)\prime} W_{NT}^{(k)} \tilde{Q}_{z\Delta y}^{(k)}, \tag{5.16}$$

where $\tilde{Q}_{z\Delta x}^{(k)} = \frac{1}{N_k}\sum_{i\in\tilde{G}_k} \tilde{Q}_{i,z\Delta x}$ and $\tilde{Q}_{z\Delta y}^{(k)} = \frac{1}{N_k}\sum_{i\in\tilde{G}_k} \tilde{Q}_{i,z\Delta y}$. The following theorem reports the asymptotic distribution of $\tilde{\alpha}_{\tilde{G}_k}$.

Theorem 5.17 *Suppose Assumptions 5.9–5.12 hold with $c_J = \infty$ in assumption 5.10(i). Then, conditional on $\tilde{K} = K_0$,*

$$\sqrt{N_k T}(\tilde{\alpha}_{\tilde{G}_k} - \alpha_k^0) \xrightarrow{D} N(0, \Omega_k), \text{ for } k = 1, \ldots, K_0,$$

where $\Omega_k = \left[Q_{z\Delta x}^{(k)\prime} W^{(k)} Q_{z\Delta x}^{(k)}\right]^{-1} Q_{z\Delta x}^{(k)\prime} W^{(k)} V_k W^{(k)} Q_{z\Delta x}^{(k)} \left[Q_{z\Delta x}^{(k)\prime} W^{(k)} Q_{z\Delta x}^{(k)}\right]^{-1}$.

Proof: Appendix F, (See page 184).

The above theorem says that the post-Lasso GMM estimator $\tilde{\alpha}_{\tilde{G}_k}$ asymptotically has the same limiting distribution as the infeasible estimator $\breve{\alpha}_k$, which further indicates that the post-lasso GMM estimator has the oracle property.

5.5 Computation and Algorithm

The objective functions in (5.6) and (5.9) are not separable in β_i, which makes it difficult to compute the estimates directly. Thus, we define a new set of parameters $\delta_{ij} = \beta_i - \beta_j$ and reparameterize the criterion functions separately for PLS and PGMM and describe the implementation below.

5.5.1 PLS Computation

Reparametrizing the objective function in (5.6), is equivalent to the constraint optimization problem below

$$\min S_1(\boldsymbol{\beta}, \boldsymbol{\delta}) = \frac{1}{2}\sum_{i=1}^{N}\sum_{t=1}^{T}(\tilde{y}_{it} - \beta_i'\tilde{x}_{it})^2 + \lambda_1 \sum_{1\le i<j\le N}\sum \dot{w}_{ij}\|\delta_{ij}\|,$$

$$\text{subject to } \beta_i - \beta_j - \delta_{ij} = 0,$$

where $\boldsymbol{\delta} = \{\delta_{ij}, i < j\}'$. By the augmented Lagrangian method, the estimates of the parameters can be obtained by minimizing

$$L_1(\boldsymbol{\beta}, \boldsymbol{\delta}, \boldsymbol{\nu}) = S_1(\boldsymbol{\beta}, \boldsymbol{\delta}) + \sum_{1\le i<j\le N}\sum \nu_{ij}'(\beta_i - \beta_j - \delta_{ij}) + \frac{\vartheta}{2}\sum_{1\le i<j\le N}\sum \|\beta_i - \beta_j - \delta_{ij}\|^2,$$

where $\boldsymbol{\nu} = \{\nu_{ij}', i < j\}'$ are lagrange multipliers and ϑ is the penalty parameter. Therefore, we can obtain the estimates of $(\boldsymbol{\beta}, \boldsymbol{\delta}, \boldsymbol{\nu})$ through iterations by the ADMM.

The minimizer of $L_1(\boldsymbol{\beta}, \boldsymbol{\delta}, \boldsymbol{\nu})$ with respect to δ_{ij}, for given $(\boldsymbol{\beta}, \boldsymbol{\nu})$, has a closed form solution and is unique. In practice, for given $(\boldsymbol{\beta}, \boldsymbol{\nu})$, the minimization problem with respect to δ_{ij} is equivalent to the following minimization

$$\min \frac{\vartheta}{2} \sum_{1 \le i < j \le N} \sum \|\zeta_{ij} - \delta_{ij}\|^2 + \lambda_1 \sum_{1 \le i < j \le N} \sum \dot{w}_{ij} \|\delta_{ij}\|,$$

where $\zeta_{ij} = \beta_i - \beta_j + \vartheta^{-1}\nu_{ij}$. Thus, the closed form solution is

$$\hat{\delta}_{ij} = ST(\zeta_{ij}, \lambda_1/\vartheta), \tag{5.17}$$

where $ST(\boldsymbol{a}, b) = (1 - b/\|\boldsymbol{a}\|)_+ \boldsymbol{a}$ is the groupwise soft thresholds rule, and $(c)_+ = 1(c > 0)c$.

Implementation

In this part, we describe the computational algorithm for minimizing the objective function in (5.6) using the ADMM. The iteration process consists of updating $\boldsymbol{\beta}, \boldsymbol{\delta}$ and $\boldsymbol{\nu}$ iteratively. For a given $(\boldsymbol{\delta}, \boldsymbol{\nu})$, we obtain the updates of $\boldsymbol{\beta}$ by setting the derivative $\partial L_1(\boldsymbol{\beta}, \boldsymbol{\delta}, \boldsymbol{\nu})/\partial \boldsymbol{\beta}$ to zero, where

$$L_1(\boldsymbol{\beta}, \boldsymbol{\delta}, \boldsymbol{\nu}) = \frac{1}{2}\|\tilde{\boldsymbol{y}} - \tilde{\boldsymbol{X}}\boldsymbol{\beta}\|^2 + \frac{\vartheta}{2}\|\Lambda\boldsymbol{\beta} - \boldsymbol{\delta} + \vartheta^{-1}\boldsymbol{\nu}\|^2 + C,$$

and C is a constant independent of $\boldsymbol{\beta}$, $\tilde{\boldsymbol{y}} = (\tilde{y}_1', \ldots, \tilde{y}_N')'$, $\tilde{y}_i = (\tilde{y}_{i1}, \ldots, \tilde{y}_{iT})'$ for each $i = 1, \ldots, N$, $\tilde{\boldsymbol{X}} = diag(\tilde{X}_1, \ldots, \tilde{X}_N)$, $\tilde{X}_i = (\tilde{x}_{i1}, \ldots, \tilde{x}_{iT})'$ for each $i = 1, \ldots, N$. Besides, $\Lambda = \nabla \otimes I_p$, where $\nabla = \{(e_i - e_j), 1 \le i < j \le N\}'$ and e_i is an $N \times 1$ vector whose ith element is one and the remaining ones are zero. We track the progress of the ADMM based on the primal residual at step m, $r^{(m)} = \Lambda\boldsymbol{\beta}^{(m)} - \boldsymbol{\delta}^{(m)}$, and stop the algorithm when $\|r^{(m)}\| < \epsilon$. The algorithm can be summarized in below:

PLS Algorithm:

1. **Initialization:** Find initial estimates of $\beta_i^{(0)}$ by minimizing the first term of (5.6) for all $i = 1, \ldots, N$. Let the initial values of $\boldsymbol{\nu}^{(0)} = 0$, and $\delta_{ij}^{(0)} = \beta_i^{(0)} - \beta_j^{(0)}$.

2. **Iterations:** At iteration $m \geq 1$, for given $\boldsymbol{\delta}^{(m-1)}$ and $\boldsymbol{\nu}^{(m-1)}$,

   (a) update $\boldsymbol{\beta}^{(m)}$ which is the minimizer of $L_1(\boldsymbol{\beta}, \boldsymbol{\delta}^{(m)}, \boldsymbol{\nu}^{(m)})$ as below

   $$\boldsymbol{\beta}^{(m)} = \left[\tilde{\boldsymbol{X}}'\tilde{\boldsymbol{X}} + \vartheta\Lambda'\Lambda\right]^{-1}\left[\tilde{\boldsymbol{X}}'\tilde{\boldsymbol{y}} + \vartheta\Lambda'\left(\boldsymbol{\delta}^{(m-1)} - \vartheta^{-1}\boldsymbol{\nu}^{(m-1)}\right)\right];$$

   (b) update the value of δ_{ij} at the (m)th iteration by (5.17), after replacing

   $$\zeta_{ij} = \beta_i^{(m)} - \beta_j^{(m)} + \vartheta^{-1}\nu_{ij}^{(m-1)};$$

   (c) update the value ν_{ij} by

   $$\nu_{ij}^{(m)} = \nu_{ij}^{(m-1)} + \vartheta(\beta_i^{(m)} - \beta_j^{(m)} - \delta_{ij}^{(m)});$$

   (d) terminate the algorithm if the stopping rule $\|r^{(m)}\| < \epsilon$ is met at step m. Then, $(\boldsymbol{\beta}^{(m)}, \boldsymbol{\delta}^{(m)}, \boldsymbol{\nu}^{(m)})$ are the PAGFL estimates $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\delta}}, \hat{\boldsymbol{\nu}})$.

Proposition 5.18 *The primal residual* $r^{(m)} = \Lambda\boldsymbol{\beta}^{(m)} - \boldsymbol{\delta}^{(m)}$ *and the dual residual* $s^{(m)} = \vartheta\Lambda(\boldsymbol{\beta}^{(m)} - \boldsymbol{\beta}^{(m-1)})$ *of the ADMM satisfy the following conditions:*

i) $\lim_{m\to\infty} \|r^{(m)}\|^2 = 0$,

ii) $\lim_{m\to\infty} \|s^{(m)}\|^2 = 0$.

Proof: Appendix G, (See page 185).

Proposition 5.18 shows that both the primal and dual feasibility are achieved by the algorithm. Further, as the objective function in (5.6) is convex, therefore the algorithm converges to an optimal point.

5.5.2 PGMM Computation

Similarly, by reparametrizing the objective function in (5.9), the minimization is equivalent to the constraint optimization problem below

$$\min S_2(\boldsymbol{\beta}, \boldsymbol{\delta}) = \frac{1}{2}\sum_{i=1}^{N}\Big[\frac{1}{T}\sum_{t=1}^{T} z_{it}(\Delta y_{it} - \beta_i'\Delta x_{it})\Big]' W_{i,NT}\Big[\frac{1}{T}\sum_{t=1}^{T} z_{it}(\Delta y_{it} - \beta_i'\Delta x_{it})\Big]$$
$$+ \lambda_2 \sum_{1\le i<j\le N}\sum \ddot{w}_{ij}\|\delta_{ij}\|, \text{ subject to } \beta_i - \beta_j - \delta_{ij} = 0,$$

where $\boldsymbol{\delta} = \{\delta_{ij}, i < j\}'$. By the augmented Lagrangian method, the estimates of the parameters can be obtained by minimizing

$$L_2(\boldsymbol{\beta}, \boldsymbol{\delta}, \boldsymbol{\nu}) = S_2(\boldsymbol{\beta}, \boldsymbol{\delta}) + \sum_{1\le i<j\le N}\sum \nu_{ij}'(\beta_i - \beta_j - \delta_{ij}) + \frac{\vartheta}{2}\sum_{1\le i<j\le N}\sum \|\beta_i - \beta_j - \delta_{ij}\|^2,$$

where $\boldsymbol{\nu} = \{\nu_{ij}', i < j\}'$ are lagrange multipliers and ϑ is the penalty parameter. Therefore, we can obtain the estimates of $(\boldsymbol{\beta}, \boldsymbol{\delta}, \boldsymbol{\nu})$ through iterations by the ADMM.

The minimizer of $L_2(\boldsymbol{\beta}, \boldsymbol{\delta}, \boldsymbol{\nu})$ with respect to δ_{ij}, for given $(\boldsymbol{\beta}, \boldsymbol{\nu})$, has a closed form solution and is unique. In practice, for given $(\boldsymbol{\beta}, \boldsymbol{\nu})$, the minimizer problem with respect to δ_{ij} is equivalent to the following minimization

$$\frac{\vartheta}{2}\sum_{1\le i<j\le N}\sum \|\zeta_{ij} - \delta_{ij}\|^2 + \lambda_1 \sum_{1\le i<j\le N}\sum \ddot{w}_{ij}\|\delta_{ij}\|,$$

where $\zeta_{ij} = \beta_i - \beta_j + \vartheta^{-1}\nu_{ij}$. Thus, the closed form solution is

$$\tilde{\delta}_{ij} = ST(\zeta_{ij}, \lambda_1/\vartheta). \tag{5.18}$$

Implementation

Now, we describe the computational algorithm for minimizing the objective function in (5.9) using the ADMM. The iteration process consists of updating $\boldsymbol{\beta}, \boldsymbol{\delta}$ and $\boldsymbol{\nu}$ iteratively.

For a given $(\boldsymbol{\delta}, \boldsymbol{\nu})$, we obtain the updates of $\boldsymbol{\beta}$ by setting the derivative $\partial L_2(\boldsymbol{\beta}, \boldsymbol{\delta}, \boldsymbol{\nu})/\partial \boldsymbol{\beta}$ to zero, where

$$L_2(\boldsymbol{\beta}, \boldsymbol{\delta}, \boldsymbol{\nu}) = \frac{1}{2}(\Delta \boldsymbol{y} - \Delta \boldsymbol{X}\boldsymbol{\beta})' \boldsymbol{Z}' \boldsymbol{W} \boldsymbol{Z} (\Delta \boldsymbol{y} - \Delta \boldsymbol{X}\boldsymbol{\beta}) + \frac{\vartheta}{2} \|\Lambda \boldsymbol{\beta} - \boldsymbol{\delta} + \vartheta^{-1} \boldsymbol{\nu}\|^2 + C,$$

where C is a constant independent of $\boldsymbol{\beta}$, $\Delta \boldsymbol{y} = (\Delta y_1', \ldots, \Delta y_N')'$, $\Delta y_i = (\Delta y_{i1}, \ldots, \Delta y_{iT})'$, $\boldsymbol{Z} = diag(Z_1, \ldots, Z_N)$, $Z_i = (z_{i1}, \ldots, z_{iT})'$, $\Delta \boldsymbol{X} = diag(\Delta X_1, \ldots, \Delta X_N)$, $\Delta X_i = (\Delta x_{i1}, \ldots, \Delta x_{iT})'$ for each $i = 1, \ldots, N$, and $\boldsymbol{W} = diag(W_{1,NT}, \ldots, W_{N,NT})$. Similarly, we track the progress of the ADMM based on the primal residual at step m, $r^{(m)} = \Lambda \boldsymbol{\beta}^{(m)} - \boldsymbol{\delta}^{(m)}$, and stop the algorithm when $\|r^{(m)}\| < \epsilon$. The algorithm can be summarized in below:

PGMM Algorithm:

1. **Initialization:** Find initial estimates of $\beta_i^{(0)}$ by minimizing the first term of (5.9) for all $i = 1, \ldots, N$. Let the initial values of $\boldsymbol{\nu}^{(0)} = 0$, and $\delta_{ij}^{(0)} = \beta_i^{(0)} - \beta_j^{(0)}$.

2. **Iterations:** At iteration $m \geq 1$, for given $\boldsymbol{\delta}^{(m-1)}$ and $\boldsymbol{\nu}^{(m-1)}$,

   (a) update $\boldsymbol{\beta}^{(m)}$ which is the minimizer of $L_2(\boldsymbol{\beta}, \boldsymbol{\delta}^{(m)}, \boldsymbol{\nu}^{(m)})$ as below

   $$\boldsymbol{\beta}^{(m)} = \Big[\Delta \boldsymbol{X}' \boldsymbol{Z}' \boldsymbol{W} \boldsymbol{Z} \Delta \boldsymbol{X} + \vartheta \Lambda' \Lambda\Big]^{-1} \Big[\Delta \boldsymbol{X}' \boldsymbol{Z}' \boldsymbol{W} \boldsymbol{Z} \Delta \boldsymbol{y} + \vartheta \Lambda' \Big(\boldsymbol{\delta}^{(m-1)} - \vartheta^{-1} \boldsymbol{\nu}^{(m-1)}\Big)\Big];$$

   (b) update the value of δ_{ij} at the (m)th iteration by (5.18), after replacing

   $$\zeta_{ij} = \beta_i^{(m)} - \beta_j^{(m)} + \vartheta^{-1} \nu_{ij}^{(m-1)};$$

   (c) update the value ν_{ij} by

   $$\nu_{ij}^{(m)} = \nu_{ij}^{(m-1)} + \vartheta(\beta_i^{(m)} - \beta_j^{(m)} - \delta_{ij}^{(m)});$$

(d) terminate the algorithm if the stopping rule $\|r^{(m)}\| < \epsilon$ is met at step m. Then, $(\boldsymbol{\beta}^{(m)}, \boldsymbol{\delta}^{(m)}, \boldsymbol{\nu}^{(m)})$ are the PAGFL estimates $(\tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\delta}}, \tilde{\boldsymbol{\nu}})$.

Proposition 5.19 *The primal residual* $r^{(m)} = \Lambda\boldsymbol{\beta}^{(m)} - \boldsymbol{\delta}^{(m)}$ *and the dual residual* $s^{(m)} = \vartheta\Lambda(\boldsymbol{\beta}^{(m+1)} - \boldsymbol{\beta}^{(m)})$ *of the ADMM satisfy the following conditions:*

i) $\lim_{m\to\infty} \|r^{(m)}\|^2 = 0$,

ii) $\lim_{m\to\infty} \|s^{(m)}\|^2 = 0$.

Proof: Appendix G, (See page 187).

Proposition 5.19 shows that both the primal and dual feasibility are achieved by the algorithm. Further, as the objective function in (5.9) is convex, therefore the algorithm converges to an optimal point.

5.6 Monte Carlo Simulation

In this section, we investigate the finite sample properties of the PAGFL and associated post-Lasso estimators. We consider two similar data generating processes (DGP) to Su et al. (2016) that cover static and dynamic panels. The fixed effects and the idiosyncratic errors follow the standard normal distribution and are mutually independent across i and t for both DGPs. The observations in each DGP are drawn from three groups with the proportion $N_1/N = 0.4$, $N_2/N = 0.3$ and $N_3/N = 0.3$. We consider all combinations of (N, T) with $N = (100, 200)$ and $T = (40, 80)$.

1. **DGP 1 (Static panel with two exogenous regressors):** The regressors $(x_{it,1}, x_{it,2})'$ are generated as $x_{it,1} = 0.2\eta_i + e_{it,1}$ and $x_{it,2} = 0.2\eta_i + e_{it,2}$ where $e_{it,1}$ and $e_{it,2}$ are both $i.i.d\, N(0,1)$ and mutually independent. y_{it} is then generated from the panel structure model (5.1). The true coefficients are

$$(\alpha_1^0, \alpha_2^0, \alpha_3^0) = \left(\begin{pmatrix} 0.4 \\ 1.6 \end{pmatrix}, \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \begin{pmatrix} 1.6 \\ 0.4 \end{pmatrix} \right). \tag{5.19}$$

2. **DGP 2 (Dynamic Panel AR(1) with two exogenous regressors):** The model is generated from the following equation

$$y_{it} = \beta_{i1}^0 y_{i,t-1} + \beta_{i2}^0 x_{it,1} + \beta_{i3}^0 x_{it,2} + \eta_i(1 - \beta_{i1}^0) + u_{it}, \tag{5.20}$$

where the exogenous regressors $x_{it,1}$ and $x_{it,2}$ follow the standard normal distributions, mutually independent, and are independent of the error term. To make each individual's time series strictly stationary with mean η_i, the initial values take the form $y_{i0} = \beta_{i2}^0 x_{it,1} + \beta_{i3}^0 x_{it,2} + \eta_i + u_{i0}$. The true coefficients are

$$(\alpha_1^0, \alpha_2^0, \alpha_3^0) = \left(\begin{pmatrix} 0.8 \\ 0.4 \\ 0.4 \end{pmatrix}, \begin{pmatrix} 0.6 \\ 1 \\ 1 \end{pmatrix}, \begin{pmatrix} 0.4 \\ 1.6 \\ 1.6 \end{pmatrix} \right). \tag{5.21}$$

We use the modified BIC (Wang et al. (2009)) for high-dimensional data settings to select the tanning parameters, λ_1 and λ_2, by minimizing

$$BIC(\lambda_j) = Q_{j,NT}(\hat{\boldsymbol{\beta}}(\lambda_j)) + C_{NT}\frac{\log(NT)}{NT}(p\hat{K}(\lambda_j)), \text{ for } j = 1, 2, \tag{5.22}$$

with respect to λ_j, where C_{NT} is a positive number that can depend on the number of observations. Wang et al. (2009) used $C_{NT} = \log(\log(d))$ where d is the number of

predictors in their simulations and diverges with the sample size. Our findings indicate that this specific choice of C_{NT} is too small for the latent group specification. We experimented different alternatives, and found that $C_{NT} = 0.05\sqrt{NT}$, works fairly well. Further, we use a fixed value for ϑ in the ADMM algorithm.

As the goal of this chapter is to consistently identify the group memberships, and estimate the regression coefficients and the number of groups, we report and compare the finite sample performance of our proposed estimation methods with the C-Lasso of Su et al. (2016) by considering the following three criteria:

(i) Estimation Consistency: We report the Root Mean Squared Errors (RMSE) which is defined as

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N} \|\hat{\beta}_i - \beta_i^0\|^2}. \tag{5.23}$$

(ii) Consistency of $\hat{K}$: We report the selection consistency as the empirical percentage of selecting the true number of groups. For example in our simulation designs we measure the percentage of the number of time $\hat{K} = K^0 = 3$.

(iii) Classification Consistency: We measure the percentage of correct classification of the group membership of individuals, by calculating $\sum_{i=1}^{N} 1(\hat{g}_i = g_i^0)$, where g_i^0 denotes the true group membership of individual i, and $\hat{g}_i$ denotes the estimated one.

The simulation results of DGPs 1–2 for 200 monte carlo simulations are presented in Table 5.1–5.3. The summary of the simulation results is as below.

1. Table 5.1 provides the RMSE of the proposed PAGFL[2], and compare it with C-Lasso of Su et al. (2016). We observe that the RMSE of all of the estimators decreases as the sample size increases while at the same time the RMSE our estimator is slightly smaller than the others.

2. Table 5.2 summarizes the empirical probability that a particular group size from 1 to 5 is selected using our approach. From Table 5.2, we can observe that in both designs, the PAGFL performance is fairly well and when the sample size is large enough PAGFL always chooses the correct the number of groups.

3. Table 5.3 reports the classification consistency. As expected, when the sample size is large enough and the difference between the slope parameters across the groups is relatively large, classification of PAGFL is accurate.

4. Finally, the PAGFL unlike C-Lasso does not rely on a prior specification of the group numbers, and at the same time performs comparably better than the other methods.

In conclusion, we can claim that the simulation results confirm our theoretical findings of the previous sections regarding estimation and identification of latent structure.

5.7 Illustrations

We now illustrate the PAGFL estimation and identification in two empirical applications.

[2]The PLS estimator for DGP2 is bias-corrected by using the Split-panel jackknife method of Dhaene and Jochmans (2015).

5.7.1 Unemployment Dynamics at the U.S. State Level

In this application, we apply the PAGFL estimation and identification procedures to a model of unemployment dynamics at the U.S. state level. Bun and Carree (2005) studied this subject using a dynamic panel data model that relates each of the states' current unemployment rate (U_{it}) to the unemployment rate and economic growth rate (G_{it}) in the previous year. In addition to capture state specific effects, their model includes both state individual intercepts η_i, and time effect θ_t. Their model can be written as below

$$U_{it} = \gamma U_{i,t-1} + \beta G_{i,t-1} + \eta_i + \theta_t + \epsilon_{it}, \tag{5.24}$$

or equivalently

$$U_{it} - U_{i,t-1} = (\gamma - 1)(U_{i,t-1} - \alpha_i) + \beta(G_{i,t-1} - \delta) + \theta_t + \epsilon_{it}, \tag{5.25}$$

where $(1-\gamma)\alpha_i - \beta\delta = \eta_i$. The model in (5.25) shows that changes in unemployment rate are determined by two observable components: first, the adjustment of the unemployment rate toward a "natural" or "equilibrium" rate of unemployment, α_i, which is allowed to vary across states, second, the deviation of the economic growth rate around a constant equilibrium. In addition, in the model above, $1-\gamma$ denotes the speed of adjustment of the unemployment rate toward the "natural" or "equilibrium" rate, further it is expected to have $\beta < 0$ because a state that has relatively high economic growth is more likely to have reduced unemployment rates compared with states in which the economy is growing more slowly.

The model above imposes the assumption of heterogeneous "intercepts" and homogeneous "slope coefficients" across states, and as pointed out by Campello et al. (2019),

estimation methods of such models can result in severely biased parameters and incorrect inferences. To avoid this issue, alternatively, we consider the following latent group structure model

$$U_{it} = \gamma_{g_i} U_{i,t-1} + \beta_{g_i} G_{i,t-1} + \eta_i + \epsilon_{it}, \tag{5.26}$$

where g_i denotes group membership of state i. The model above equivalently can be written as

$$U_{it} - U_{i,t-1} = (\gamma_{g_i} - 1)(U_{i,t-1} - \alpha_i) + \beta_{g_i}(G_{i,t-1} - \delta_{g_i}) + \epsilon_{it}. \tag{5.27}$$

The data for the unemployment rate are taken from the U.S. Bureau of Labor Statistics for the 1976–2019 period, and the data for the state gross product are per capita personal income (thousands of dollars) which are obtained from the U.S. Bureau of Economic Analysis deflated by annual implicit price deflator. The economic growth rate is taken to be the relative growth of the state product. Therefore, in our application $N = 51$, all U.S. states and Washington, DC, and $T = 43$ because year 1976 is taken as the starting observation.

The PAGFL divides the states in three groups, where the group memberships are presented in Figure 5.1. Table 5.4 reports the estimated coefficient estimates based on full sample and three groups with their corresponding standard deviations. All the estimated coefficients of γ are highly significant among the four models under 1% level. The value of γ in full sample and group 1 are almost the same and equal to 0.8, which implies an adjustment rate of around 20% per year. The adjustment rate in group 3 is smaller around 14% and that of group 2 is faster around 28%. The value of the full sample estimate of β

equals -0.261, whereas the value of the estimate in group 1 and group 2 are -0.716, and -0.567 and all are significant under 1% level. This implies a somewhat stronger effect of economic growth on the change in unemployment than other states in group 3.

5.7.2 Forecasting Output Growth of 33 Countries

In this section, we present an empirical application that highlights the utility of PAGFL in forecasting. In particular, we forecast the output growth rate of a large number of countries in the global economy using a set of macroeconomic and financial variables by allowing latent group structures in the slope coefficients. This allows us to aggregate the countries with close response variables which can improve the forecasts. As pointed out by Pesaran et al. (2009) this is an important issue that practitioners face when constructing forecasting models which is still an open discussion. We consider a panel data model with latent group structures which adds to the current and ongoing literature of forecasting economic and financial variables across countries including Dees et al. (2007,a), Dees et al. (2007,b), and Pesaran et al. (2009), among others.

The data set is taken from the Global VAR (GVAR) dataset[3]. We use quarterly macroeconomic and financial variables including log real GDP (y_{it}), the rate of inflation (π_{it}), short-term interest rate (r_{it}), long-term interest rate (lr_{it}), and log real equity prices (q_{it}) for $N = 33$ economies from 1979Q2 to 2016Q4.

We are interested in forecasting h quarters ahead rate of log real GDP, with the predictors in $z_{it} = (\Delta y_{i,t-1}, \Delta r_{i,t} - \Delta \pi_{it}, \Delta lr_{i,t} - \Delta r_{it}, \Delta q_{i,t} - \Delta \pi_{it})$ and $z_{it}^* = (\Delta y_{i,t}^*, \Delta r_{i,t}^* - \Delta \pi_{it}^*, \Delta lr_{i,t}^* - \Delta r_{it}^*, \Delta q_{i,t}^* - \Delta \pi_{it}^*)$, where z_{it}^* is the country-specific foreign variables. The foreign variables

[3]The data is available at the GVAR Toolbox webpage: https://sites.google.com/site/gvarmodelling/data.

are constructed using rolling three year moving averages of the annual trade weights which are computed as shares of exports and imports for each country [4].

Therefore, we consider the following equation

$$\Delta_h y_{i,t+h} = \eta_{i,h} + \beta'_{i,h} x_{it} + u_{it}, \quad i = 1, \ldots, N, \text{ and } t = 1, \ldots, T, \tag{5.28}$$

where $\Delta_h y_{i,t+h} = y_{i,t+h} - y_{it}$ for the forecast horizon h, $x_{it} = (z_{it}, z^*_{it})$ and the slope parameters, $\beta_{i,h}$, admit a possible grouping structure of the form (5.2). We estimate and identify the slope parameters and the underlying group structure using the PAGFL method developed in the previous sections. In our analysis, we consider up to $h = 4$ (four quarters ahead) and report results for one quarter ahead ($h = 1$) and one year ahead ($h = 4$). The forecasts are constructed using both rolling windows and expanding windows of 15 years time periods, or $T = 60$ for the rolling window, which leaves us with the last $H_1 = 83$ out-of-sample evaluation periods, 1996Q2-2016Q4 for $h = 1$, and $H_2 = 79$ out-of-sample evaluation periods, 1997Q2-2016Q4 for $h = 4$.

In addition, to allow for possible structural breaks, following the suggestion of Pesaran and Timmermann (2007) and Pesaran and Pick (2011), we repeat the above forecasting process by changing the estimation window. Specifically, the start date of the estimation sample is moved forward by one quarter till the observations left for the estimation is at least twice the number of regressors, and the process of out-of-sample forecasting is repeated as before. This estimation process is repeated for each of the three models. Thus, for each model, and for each out-of-sample forecast date, there are $T - 2p + 1$ windows yielding a total of $T - 2p + 1$ forecasts to be averaged. We denote the average

[4] For example the trade weights of year 2016 is based on the average trade flows computed over the three years 2013–2015. Because the trade flows observations start at 1980, the process of computing time-varying trade weights was initialized by using the same set of weights for the first four years of the sample period.

forecast from a particular model estimated over different estimation windows by "Average Windows".

We evaluate the forecasting performance of our method, with individual equations forecasts, and a fixed effect approach using the root mean squared forecast error (RMSFE) of any given model, which is averaged across the N countries as below

$$RMSFE(h,H)=\sqrt{\frac{1}{N}\sum_{i=1}^{N}\frac{1}{H_h}\sum_{t=T}^{T+H_h-1}\hat{e}_{it}^2(h)},\quad h=1,4, \tag{5.29}$$

where $\hat{e}_{it}(h)=\Delta_h y_{i,t+h}-\widehat{\Delta_h y}_{i,t+h|t}$ is the $h-$quarter ahead forecast error, with $\Delta_h y_{i,t+h}$ being the actual value, and $\widehat{\Delta_h y}_{i,t+h|t}$ the corresponding forecast formed at time t. RMSFE and relative RMSFE statistics for the one-quarter and one-year ahead forecasts of output growth rate are reported in Table 5.5 and Table 5.6.

Diebold and Mariano (1995) (DM) test statistics for testing $H_0:\mathbb{E}\left(\hat{z}_{it,m}(h)\right)=0$, where $\hat{z}_{it,m}(h)=\hat{e}_{it,\text{PAGFL}}^2(h)-\hat{e}_{it,m}^2(h)$ is the difference between the $h-$quarter ahead squared forecasting errors of our PAGFL method and method m (fixed effect or individual equations models) for country i. Specifically, by assuming serially uncorrelated h-step-ahead forecasting errors, we have

$$DM_{i,m}(h)=\sqrt{H_h}\,\frac{\bar{\hat{z}}_{i,m}(h)}{\hat{\sigma}_{i,m}(h)},\quad i=1,\ldots,N,\ \text{and}\ h=1,4, \tag{5.30}$$

where $\bar{\hat{z}}_{i,m}(h)=\frac{1}{H_h}\sum_{t=T+1}^{T+H_h}\hat{z}_{it,m}(h)$ is the sample mean of $\hat{z}_{it,m}(h)$, and

$$\hat{\sigma}_{i,m}^2(h)=\frac{1}{H_h-1}\sum_{t=T+1}^{T+H_h}\left(\hat{z}_{it,m}(h)-\bar{\hat{z}}_{it,m}(h)\right)^2. \tag{5.31}$$

To compare the forecasts across the countries, we compute the panel version of the DM test which is proposed in Pesaran et al. (2009) to statistically test the panel forecasts across countries against each method for a given forecast horizon. The panel DM ($\overline{DM}$) statistic

under assuming serially and cross-sectionally uncorrelated h-step-ahead forecasting errors is defined as

$$\overline{DM}_m = \frac{\bar{z}_m(h)}{\sqrt{V\Big(\bar{z}_m(h)\Big)}}, \quad h = 1, 4, \tag{5.32}$$

where $\bar{z}_m(h) = \frac{1}{N}\sum_{i=1}^{N} \bar{\hat{z}}_{i,m}(h)$ and $V\Big(\bar{z}_m(h)\Big) = \frac{1}{NT}\Big(\frac{1}{N}\sum_{i=1}^{N} \hat{\sigma}^2_{i,m}(h)\Big)$. The panel $\overline{DM}$ test results are reported in Table 5.7 and Table 5.8 for one-quarter and one-year ahead forecasts.

We note that one quarter ahead PAGFL forecasts perform better than the fixed effects and individual estimators in all cases and the panel $\overline{DM}$ tests are significant. For the one-year ahead forecasts, under expanding windows, PAGFL outperforms the other two methods, however, under the rolling windows, the difference between PAGFL method and individual estimators is very small but both perform better than fixed effects. It is worth mentioning that we examined whether there exist obvious structural breaks by employing the recently developed break detection method by Baltagi et al. (2016) that allows for heterogeneous slope coefficients. Although, it did not detect any structural breaks, the performances of the forecasts under average windows are better than using the full sample.

5.8 Conclusion

The present chapter suggests a simple and computationally efficient way to jointly estimate and identify latent group structures in panel data, by developing pairwise fusion penalized least squares (PLS) and GMM (PGMM) methods. We develop theoretical results on consistent group structure estimation and discuss the asymptotic properties of the

estimators. The PLS estimator asymptotically achieves the oracle property, but the PGMM oracle property is confined to some restrictive assumptions. Monte Carlo simulations are conducted to examine the finite sample properties of the proposed method which show that the approach has good finite-sample performance. Our first empirical application on the unemployment dynamics in the U.S. state level finds strong evidence that the slope coefficients are heterogenous and can be conveniently classified into three distinct groups. In addition, our second application in forecasting output growth of 33 countries using macroeconomic and financial variables shows that our PAGFL framework outperforms other candidate methods.

There are several directions that we plan to explore in the future. First, our model is focused on linear panels, and it can be extended to include both linear and nonlinear models. Second, our method can be extended to non-stationary panels where panel unit and cointegrating relationships may possess latent group structures. Third, it may be appealing to consider a model with interactive fixed effects. Lastly, our approach can be applied to extend the panel data quantile regression of Gu and Volgushev (2019) to allow for latent group structure in the slope coefficients.

Table 5.1: RMSE of DGP1 and DGP2

	N	T	PAGFL	Post-Lasso	C-Lasso	Oracle
DGP 1	100	40	0.086	0.083	0.091	0.040
	100	80	0.027	0.026	0.028	0.026
	200	40	0.064	0.061	0.088	0.026
	200	80	0.024	0.021	0.023	0.020
DGP 2	100	40	0.043	0.041	0.051	0.032
	100	80	0.018	0.015	0.029	0.010
	200	40	0.031	0.030	0.036	0.021
	200	80	0.014	0.014	0.022	0.010

Table 5.2: Frequency of Selecting $K = 1, \ldots, 5$ Groups when $K_0 = 3$

		DGP 1					DGP 2				
N	T	1	2	3	4	≥ 5	1	2	3	4	≥ 5
100	40	0	0	0.995	0.005	0	0	0	0.998	0.002	0
100	80	0	0	1	0	0	0	0	1	0	0
200	40	0	0	1	0	0	0	0	1	0	0
200	80	0	0	1	0	0	0	0	1	0	0

Table 5.3: Percentage of Correct Classification

		DGP 1		DGP 2	
N	T	PAGFL	C-Lasso	PAGFL	C-Lasso
100	40	0.991	0.870	0.997	0.921
100	80	1.000	0.995	1.000	0.997
200	40	0.987	0.975	0.990	0.988
200	80	1.000	1.000	1.000	1.000

Table 5.4: Estimation results of the Unemployment-Growth Model

		PAGFL		
	Full Sample	Group 1	Group 2	Group 3
$\hat{\gamma}$	0.800^{***}	0.796^{***}	0.720^{***}	0.852^{***}
	(0.020)	(0.032)	(0.030)	(0.035)
$\hat{\beta}$	-0.261^{***}	-0.716^{***}	-0.567^{***}	0.028^{*}
	(0.011)	(0.018)	(0.017)	(0.019)

Note: *** 1% significant, * 10% significant.

Table 5.5: RMSFE performance of the PAGFL, individual estimators, and fixed effect methods for one quarter ahead output growth forecasts across 33 countries over the period 1969Q2-2016Q4)

Models	Full Sample		Average Windows	
	RMSFE (×100)	Relative RMSFE	RMSFE (×100)	Relative RMSFE
	Rolling Window			
PAGFL	0.856	0.947	0.839	0.932
Fixed Effects	0.871	0.964	0.852	0.946
Individual Est.	0.904	1.000	0.900	1.000
	Expanding Window			
PAGFL	0.870	0.947	0.863	0.945
Fixed Effects	0.888	0.967	0.879	0.963
Individual Est.	0.919	1.000	0.913	1.000

Note: RMSFE is computed using both a rolling and an expanding forecasting scheme with an initial window of 60 observations. To consider potential structural breaks, we average the forecasts across different estimation windows, the results are presented under the "Average Windows" columns.

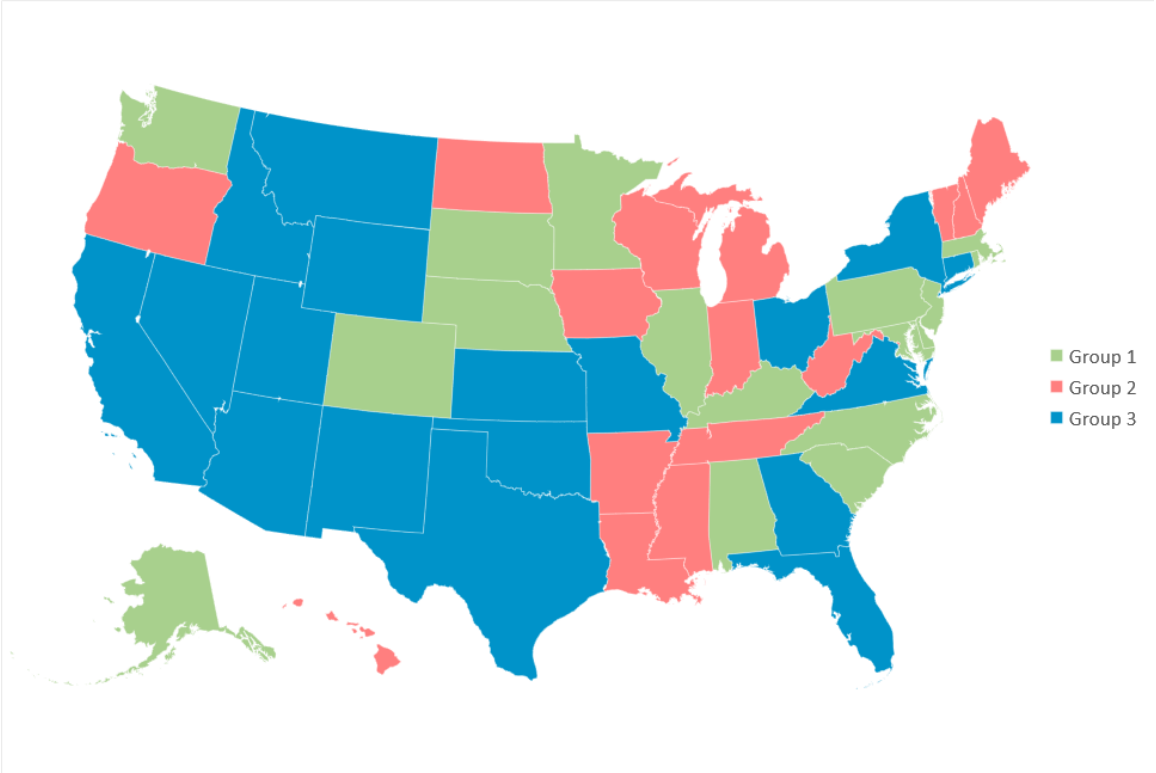


Figure 5.1: Group Membership of States

Table 5.6: RMSFE performance of the PAGFL, individual estimators, and fixed effect methods for one year (four quarters) ahead output growth forecasts across 33 countries over the period 1997Q2-2016Q4)

Models	Full Sample		Average Windows	
	RMSFE (×100)	Relative RMSFE	RMSFE (×100)	Relative RMSFE
Rolling Window				
PAGFL	1.987	1.019	1.702	1.016
Fixed Effects	2.107	1.081	1.928	1.151
Individual Est.	1.950	1.000	1.675	1.000
Expanding Window				
PAGFL	2.105	0.997	2.030	0.998
Fixed Effects	2.203	1.044	2.167	1.065
Individual Est.	2.110	1.000	2.034	1.000

Note: RMSFE is computed using both a rolling and an expanding forecasting scheme with an initial window of 60 observations. To consider potential structural breaks, we average the forecasts across different estimation windows, the results are presented under the "Average Windows" columns.

Table 5.7: Panel DM statistics for one quarter ahead PAGFL forecasts of real output growth over the period 1969Q2-2016Q4 relative to fixed effects and individual estimators as benchmarks.

Benchmark Models	Full Sample	Average Windows
	Rolling Window	
Fixed Effects	-4.337^{***}	-3.339^{***}
Individual Est.	-1.974^{**}	-2.229^{**}
	Expanding Window	
Fixed Effects	-5.861^{***}	-5.389^{***}
Individual Est.	-2.356^{***}	-2.019^{**}

Note: The results represent a one sided test, thus the 1% (***) and 5% (**) critical values are -2.326 and -1.645, respectively. A positive value of the panel DM statistic represents evidence against the PAGFL forecasts.

Table 5.8: Panel DM statistics for one year (four quarters) ahead PAGFL forecasts of real output growth over the period 1997Q2-2016Q4 relative to fixed effects and individual estimators as benchmarks.

Benchmark Models	Full Sample	Average Windows
	Rolling Window	
Fixed Effects	-2.759^{***}	-4.421^{***}
Individual Est.	3.167	2.678
	Expanding Window	
Fixed Effects	-3.260^{***}	-3.072^{***}
Individual Est.	-0.835	-0.531

Note: The results represent a one sided test, thus the 1% (***) and 5% (**) critical values are -2.326 and -1.645, respectively. A positive value of the panel DM statistic represents evidence against the PAGFL forecasts.

Chapter 6

Conclusions

This dissertation contributes to the estimation and inference of panel data and system of equations under model uncertainty. Several types of model uncertainty is considered which consist of : (1) uncertainty about a set of restrictions on the slope parameters in panel data or seemingly unrelated regressions, (2) uncertainty from choosing different number of lagged dependent variables as instruments in dynamic panel data models, (3) uncertainty about the magnitude of endogeneity in simultaneous equations models or instrumental variable regressions, (4) uncertainty resulting from unobserved heterogeneity in panel data models. The asymptotic properties of the proposed estimators of slope parameters are established. For each model, various Monte Carlo experiments are done to show the good finite sample performance of the proposed estimators. In empirical applications, the methods are employed to show how the methods perform in dealing with economic applications.

In the future research, there are several directions that I plan to explore my research to investigate empirical economic problems. First, I plan to extend the method developed in chapter 5 to other econometric models such as non-stationary panels, panel models with latent structures in both slope parameters and interactive fixed effects, and panel threshold models. Second, I plan to extend the shrinkage and model averaging methods devolved in the other chapters to multi-period forecasting of vector auto-regressions, panel data models with multi-factor error structures, and spatial models.

Bibliography

Abadir, K. M., J. R. Magnus, 2005. *Matrix Algebra* New York: Cambridge University Press.

Ahn, S. C. and P. Schmidt (1995). Efficient estimation of models for dynamic panel data. *Journal of Econometrics* 68, 5–27.

Alvarez, J. and M. Arellano (2003). The time series and cross-sectional asymptotics of dynamic panel data estimators. *Econometrica* 71, 1121–1159.

Anderson, T. W. (1977). Asymptotic expansions of the distributions of estimates in simultaneous equations for alternative parameter sequence. *Econometrica,* 45, 506-518.

Anderson, T. W. and C. Hsiao, (1981). Estimation of dynamic models with error components. *Journal of the American Statistical Association* 76: 598–606.

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

Anderson, T.W., Naoto Kunitomo and Kimio Morimune (1986). Comparing Single-Equation Estimators in a Simultaneous Equation System. *Econometric Theory*,2 (1), 1-32.

Anderson, T. W. and H. Rubin (1949). Anderson, T. W. and H. Rubin (1949). "Estimation of the parameters of a Single Equation in a Complete System of Stochastic Equations, *Annals of Mathematical Statistics,* 20, 1, 46-63.

Anderson, T. W. and T. Sawa. Evaluation of the distribution function of the two-stage least squares estimate. *Econometrica,* 47: 163-182.

Ando, T. and J. Bai (2016). Panel data models with grouped factor structure under unknown group membership. *Journal of Applied Econometrics.* 31, 163–191.

Andrews, D., 2005. Cross section regression with common shocks. *Econometrica* 73: 1551–85.

Arellano, M. and S. Bond (1991). Some tests of speciÖcation for panel data: Monte Carlo evidence and an application to employment equations. *The Review of Economic Studies* 58, 277–297.

Arellano, M. and O. Bover (1995). Another look at the instrumental variable estimation of errorcomponents models. *Journal of Econometrics* 68, 29–51.

Bai, J. (1997). Estimating multiple breaks one at a time. *Econometric Theory.* 13, 315–352.

Baltagi BH, Feng Q, Kao C. (2016). Estimation of heterogeneous panels with structural breaks. *Journal of Econometrics*, 191, 176–195.

Baltagi, B. H., S. Garvin, S. Kerman, 1989. Further Monte Carlo Evidence on Seemingly Unrelated Regressions with Unequal Number of Observations. *Annales D'Economie et de Statistique* 14: 103–15.

Baltagi, B. H., J. M. Griffin, W. Xiong, 2000. To pool or not to pool: Homogeneous versus heterogeneous estimations applied to cigarette demand. *The Review of Economics and Statistics* 82(1): 117–26.

Baltagi, B. H., J. M. Griffin, 1984. Short and long run effects in pooled models. *International Economic Review* 25(3): 631–45.

Baltagi, B. H. and Griffin, J. M. (1997). Pooled estimators vs. their heterogeneous counterparts in the context of dynamic demand for gasoline. *Journal of Econometrics.* 77, 303–327.

Baranchick, A., 1964. Multiple Regression and Estimation of the Mean of A Multivariate Normal Distribution, Technical Report No. 51. *Department of Statistics, Stanford University.*

Bekker, P.A. (1994). Alternative approximations to the distributions of Instrumental Variable estimators. *Econometrica* 62, 657–681.

Belloni, A., V. Chernozhukov (2013). Least Squares after Model Selection in High-Dimensional Sparse Models. *Bernoulli* 19, 521–547.

Belloni, A., Chernozhukov, V., Fernandez-Val, I., Hansen, C., 2017. Program evaluation and causal inference with high-dimensional data. *Econometrica* 85: 233–298.

Bertsekas, D. (1995). *JNonlinear Programming.* Athena Scientific, Belmont, MA.

Bester, C. A. and C. B. Hansen (2016). Grouped effects estimators in fixed effects models. *Journal of Econometrics.* 190, 197–208.

Billingsley, P., 1986. Convergence of Probability Measures. *John Wiley, New York.*

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

Bonhomme, S., Manresa, E., 2015. Grouped patterns of heterogeneity in panel data. *Econometrica* 83: 1147–84.

Boyd, S., Parikh, N., Chu, E., Peleato, B., and Eckstein, J. (2011). Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers. *Foundations and Trends in Machine Learning.* 3, 1–122.

Browning, M., J. Carro, 2007. Heterogeneity and microeconometrics modeling. In *Advances in Economics and Econometrics*. Edited by R. Blundell, W. Newey and T. Persson. Cambridge: Cambridge University Press, vol. 3, pp. 47–74.

Bun, M. J. G., and Carree, M. A. (2005). Bias-Corrected Estimation in Dynamic Panel Data Models. *Journal of Business & Economic Statistics.* 23, 200—210.

Bun, M.J.G. and J.F. Kiviet (2006). The effects of dynamic feedbacks on LS and MM estimator accuracy in panel data models. *Journal of Econometrics* 132, 409–444.

Campello, M., Galvao, A. F., and Juhl, T. (2019). Testing for Slope Heterogeneity Bias in Panel Data Models. *Journal of Business & Economic Statistics.* 37, 749–760.

Chernozhukov, V., Hansen, C., Spindler, M., 2015. Post-selection and post-regularization inference in linear models with many controls and instruments. *American Economic Review* 105: 486–90.

Chudick, A., M. H. Pesaran, 2015. Large Panel Data Models With Cross-Sectional Dependence A Survey *The Oxford Handbook Of Panel Data, Edited By Badi H. Baltagi* New York: Oxford University Press.

Claeskens, G., N. L. Hjort, 2003. The focused information criterion. *Journal of American Statistical Association* 98: 900–45.

Danilov, D., J. R. Magnus, 2004a. On the harm that ignoring pretesting can cause. *Journal of Econometrics* 122:27–46.

Danilov, D., J. R. Magnus, 2004b. Forecast accuracy after pretesting with an application to the stock market. *Journal of Forecasting* 23:251–274.

Deb, P. and P. K. Trivedi (2013). Finite mixture for panels with fixed effects. *Journal of Econometric Methods.* 2, 35–51.

Dees, S., di Mauro, F., Pesaran, M. H., & Smith, L. V. (2007). Exploring the international linkages of the euro area: A global VAR analysis. *Journal of Applied Econometrics*, 22, 1–38.

Dees, S., Holly, S., Pesaran, M. H., & Smith, L. V. (2007). Long run macroeconomic relations in the global economy. *The Open-Access, Open-Assessment E-Journal*, 2007-3.

Diebold, F. X., and Mariano, R. (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics*, 13, 253–263.

DiTraglia, F. J., 2016. Using invalid instruments on purpose: Focused moment selection and averaging for GMM. *Journal of Econometrics* 195: 187–208.

Dhaene, G., and Jochmans, K. (2015). Split-panel jackknife estimation of fixed-effect models. Review of Economic Studies. *Review of Economic Studies.* 82(3), 991-–1030

Durlauf, S. N., A. Kourtellos, A. Minkin, 2001. The local Solow growth model. *European Economic Review* 45(4–6): 928–40.

Fan, J., R. Li, 2001. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of American Statistical Association* 96: 1348–60.

Fan, J., J. Lv, 2010. A selective overview of variable selection in high dimensional feature space. *Statistical Sinica* 20: 102–148.

Greene, W. H. (2008). *Econometric Analysis* (7th Edition) Upper Saddle River, NJ: Prentice Hall.

Gourieroux, C., P. C. B. Phillips, And J. Yu (2010). Indirect Inference For Dynamic Panel Models. *Journal Of Econometrics.* 157, 68–77.

Gu, J. and S. Volgushev (2019). Panel Data Quantile Regression with Grouped Fixed Effects. *Journal of Econometrics.* 213, 68–91.

Hahn, J., And G. Kuersteiner (2002). Asymptotically Unbiased Inference For A Dynamic Panel Model With Fixed Effects When Both N And T Are Large. *Econometrica.* 70, 1639–1657.

Hahn, J., and H. R. Moon (2010). Panel Data Models With Finite Number of Multiple Equilibria. *Econometric Theory.* 26, 863–881.

Han, C., P. C. B. Phillips And D. Sul (2014). X-Differencing And Dynamic Panel Model Estimation. *Econometric Theory.* 30, 201–251.

Han, C., P. C. B. Phillips And D. Sul (2014). X-Differencing And Dynamic Panel Model Estimation. *Econometric Theory.* 30, 201–251.

Hansen, B. E. 2014. Model averaging, asymptotic risk, and regressor groups. *Quantitative Economics* 5: 495–530.

Hansen, B. E., 2016. Efficient Shrinkage in Parametric Models. *Journal of Econometrics* 190(1): 115–32.

Hansen, B. (2017). Stein-like 2SLS estimator. *Econometric Reviews.*

Hausman,J. A. (1978). "Specification Tests in Econometrics," *Econometrica*, 46, 1251-1271.

Hayakawa, K. (2009). On the effect of mean-nonstationarity in dynamic panel data models. *Journal of Econometrics* 153, 133–135.

Hayakawa, K. (2012). GMM estimation of short dynamic panel data models with interactive fixed effects. *Journal of the Japan Statistical Society* 42, 109–123.

Hoogstrate, A. J., F. C. Palm, G. A. Pfann, 2000. Pooling in dynamic panel-data models: An application to forecasting GDP growth rates. *Journal of business and Economic Studies* 18(3): 274–83.

Hsiao, C., And A. K. Tahmiscioglu (1997). A Panel Analysis Of Liquidity Constraints And Firm Investment*Journal Of The American Statistical Association* 92, 455–465.

Hsiao, C. and Q. Zhou, (2017). First difference or forward demeaning: Implications for the method of moments estimators *Econometric Reviews* 36:6-9, 883–897.

Huang, W., Jin, S., Su, L., (2020). Panel cointegration with latent group structures and an application to the PPP theory. *Econometric Theory forthcoming.*

James, W., C. Stein, 1961. Estimation with quadratic loss. *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability* 1: 361–80.

Judge, G., M.E. Bock, 1978. The Statistical Implications of Pre-test and Stein- rule Estimators in Econometrics. *North-Holland.*

Kadane, J. B. (1970), Testing Overidentifying Restrictions When the Disturbances are Small, *Journal of the American Statistical Association*, 65:329, 182-185.

Kadane, J. B. (1971), Comparison of K-Class Estimators When the Disturbances Are Small, *Econometrica*. 39, 723-737.

Kasahara, H., And K. Shimotsu (2009). Nonparametric Identification Of Finite Mixture Models Of Dynamic Discrete Choices. *Econometrica.* 77, 135–175.

Ke, T., Fan, J., and Wu, Y. (2015). Homogeneity in Regression. Journal of the American Statistical Association. *Journal of the American Statistical Association.* 110, 175–194.

Kiviet, J. F. (1995). On Bias, Inconsistency, And Efficiency Of Various Estimators In Dynamic Panel Data Models. *Journal Of Econometrics.* 68, 53–78.

Lebedev, N.N., 1972. *Special Functions and their Applications.* Dover, New York.

Lee, Y. (2012). Bias In Dynamic Panel Models Under Time Series Misspecification. *Journal Of Econometrics.* 169, 54–60.

Leeb, H. Pötscher, B. M., 2003. The finite-sample distribution of post-model-selection estimators and uniform versus nonuniform approximations. *Econometric Theory* 19:100–142.

Leeb, H. Pötscher, B. M., 2006. Can one estimate the conditional distribution of post-model-selection estimators? *The Annals of Statistics* 34:2554–2591.

Lehmann, E.L., Casella, G., 1998. Theory of Point Estimation. second edition, Springer, New York..

Lin, C.-C., And S. Ng (2012). Estimation Of Panel Data Models With Parameter Heterogeneity When Group Membership Is Unknown *Journal Of Econometric Methods.* 1, 42–55.

Liu, C., 2015. Distribution theory of the least squares averaging estimator. *Journal of Econometrics* 186: 142–59.

Liu, R., Schick, A., Shang, Z., Zhang, Y., Zhou, Q., (2020). Identification and estimation in panel models with overspecified number of groups. *Journal of Econometrics.* 2, 574-590.

Lu, X., and Su, L., (2017). Determining the number of groups in latent panel structures with an application to income and democracy. *Quantitative Economics.* 8, 729–760.

Ma, S., Huang, J., (2017). A Concave Pairwise Fusion Approach to Subgroup Analysis. *Journal of the American Statistical Association.* 517, 410–423.

Maddala, G. S., 1991. To pool or not to pool: That is the question. *Journal of Quantitative Economics* 7: 255–64.

Maddala, G. S., W. Hu, 1996. The pooling problem. In *The Econometrics of Panel Data.* Edited by L. Matyas and P. Sevestre. Advanced Studies in Theoretical and Applied Econometrics, vol 33. Springer, Dordrecht, pp. 307–22.

Maddala, G. S., Li H., V. K. Srivastava, 2001. A Comparative Study of Different Shrinkage Estimators for Panel Data Models. *Annals of Economics and Finance* 2: 1–30.

Magnus, J. R., 1999. The traditional pretest estimator. *Theory of Probability and Its Applications* 44:293–308.

Magnus, J. R., 2002. Estimation of the mean of a univariate normal distribution with known variance. *Econometrics Journal* 5:225–236.

Magnus, J. R., J. Durbin, 1999. Estimation of regression coefficients of interest when other regression coefficients are of no interest. *Econometrica* 67:639–643.

Maasoumi, E. (1978). A modified Stein-like estimator for the reduced form coefficients of simultaneous equations. *Econometrica*. 46, 695-703.

Morimune. K. (1978), Improving the Limited Information Maximum Likelihood Estimator When the Disturbances are Small, *Journal of the American Statistical Association,* 73:364, 867-871.

Morimune, K. and N. Kunitomo (1980), Improving the maximum likelihood estimate in linear functional relationships for alternative parameter sequences. *Journal of the American Statistical Association,* 75 : 230-237.

Nagar, A.L., 1959. The Bias and Moment Matrix of the General K-Class Estimators of the Parameters in Simultaneous Equations. *Econometrica* 27: 575–95.

Pesaran, M. H., and Pick, A. (2011). Forecast combinations across estimation windows. *Journal of Business & Economic Statistics*, 29, 307–318.

Pesaran, M. H., Y. Shin, R. P. Smith, 1999. Pooled mean group estimation of dynamic heterogeneous panels. *Journal of the American Statistical Association.* 94: 621–34.

Pesaran, M. H., R. Smith, 1995. Estimating long-run relationships from dynamic heterogeneous panels. *Journal of Econometrics* 68(1): 79–113.

Pesaran, M. H., Schuermann, T., and Smith, L. V. (2009). Forecasting economic and financial variables with global VARs. *International journal of forecasting*, 25(4), 642–675.

Pesaran, M. H., and Timmermann, A. (2007). Selection of estimation window in the presence of breaks. *Journal of Econometrics*, 137, 134–161.

Phillips, P.C.B. (1984), The exact distribution of LIML, *International Economic Review* 25-1, 249-261.

Phillips, P. C. B. and D. Sul, 2003. Dynamic panel estimation and homogeneity testing under cross section dependence. *The Econometrics Journal* 6: 217–59.

Phillips, P. C. B. and D. Sul, 2007. Bias in dynamic panel estimation with fixed effects, incidental trends and cross section dependence. *Journal of Econometrics* 137: 162–88.

Qian, J., and Su, L., (2016). Shrinkage estimation of common breaks in panel data models via adaptive group fused Lasso. *Journal of Econometrics.* 191, 86–109.

Robertson, D., J. Symons, 1992. Some strange properties of panel data estimators. *Journal of Applied Econometrics* 7(2): 175–89.

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

Rothenberg, T. (1984), 'Approximating the Distribution of Econometric Estimators and Test Statistics. *Handbook of Econometrics,* Volume II, Edited by Z. Griliehes and M.D. lntriligator.

Sarafidis, V. D. Robertson, 2009. On the impact of error cross-sectional dependence in short dynamic panel estimation. *The Econometrics Journal* 162: 62–81.

Sarafidis, V. and N. Weber (2015). A partially heterogeneous framework for analyzing panel data. *Oxford Bulletin of Economics and Statistics.* 77, 274–296.

Sargan J. D. (1974), The Validity of Nagar's Expansion for the Moments of Econometric Estimators, *Econometrica,* 42, 1, 169-176.

Sawa, T. (1973 a), Almost Unbiased Estimators in Simultaneous Equations Systems, *International Economic Review,* 14, 97-106.

Sawa, T. (1973 b), The Mean Square Error of A Combined Estimator and Numerical Comparison with The TSLS Estimator, *Journal of Econometrics,* 1, 115-132.

Slater, L.J., 1960. Confluent Hypergeometric Functions. *Cambridge University Press, London.*

Srivastava, J. N., and R. Tiwari, 1976. Evaluation of Expectations of Products of Stochastic Matrices. *Scandinavian Journal of Statistics* 3: 135–38.

Stein C., 1956.Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. *Proceedings of Berkeley Symposium on Mathematical Statistics and Probability* 1: 197–206.

Su, L., and Q. Chen, 2013. Testing homogeneity in panel data models with interactive fixed effects. *Econometric Theory* 29(6): 1079–135.

Su, L. and G. Ju (2018). Identifying latent grouped patterns in panel data models with interactive fixed effects. *Journal of Econometrics.* 206, 554–573.

Su, L., Z. Shi, P. C. B. Phillips, 2016. Identifying latent structures in panel data. *Econometrica* 84: 2215–64.

Su, L., Wang, X., Jin, S., (2019). Sieve estimation of time-varying panel data models with latent structures. *Journal of Business & Economics Statistics.* 37(2), 334–349.

Sun, Y. (2005). Estimation And Inference In Panel Structure Models. *Working Paper, Dept. Of Economics, Ucsd*

Swamy, P. A. V. B., 1970. Efficient inference in a random coefficient regression model. *Econometrica* 38: 311–23.

Tibshirani, R., 1996. *Journal of the Royal Statistical Society* 58: 267–88.

Tibshirani, R., M. Saunders, S. Rosset, J. Zhu, And K. Knight (2005). Sparsity And Smoothness Via The Fused Lasso. *Journal Of The Royal Statistical Society, Series B.* 67, 91–108.

Theil, H. (1961). *Economic Forecast and Policy*, 2nd ed. Amsterdam: North Holland.

Tseng, P. (2001). Convergence of a block coordinate descent method for nondifferentiable minimization. *Journal of Optimization Theory and Applications.* 109, 475–494.

Ullah, A., 1974. On the Sampling Distributions of Improved Estimators for Coefficients in Linear Regression. *Journal of Econometrics* 2: 143–50.

Ullah, A., 2004. *Finite Sample Econometrics*. Oxford: Oxford University Press.

Ullah, A. and V.K. Srivastava (1988). On the Improved Estimation of Structural Coefficients. *Sankhya: The Indian Journal of Statistics*. 50 , 111-118.

van der Vaart, A. W., 1998. Asymptotic Statistics. *Cambridge University Press, New York.*

Wang, H., Li, B., and Leng, C., (2009). Shrinkage Tuning Parameter Selection With a Diverging Number of Parameters. *Journal of Royal Statistical Society.* Series B, 71, 671-683.

Wang, W., Phillips, P.C., Su, L., (2018). Homogeneity pursuit in panel data models: Theory and application. *Journal of Applied Econometrics.* 33, 797–815.

Wang, W., Su, L., (2020). Identifying latent group structures in nonlinear panels. *Journal of Econometrics forthcoming.*

Wang, W., X. Zhang, R. Paap, 2019. To Pool or Not to Pool: What is a Good Strategy? *Journal of Applied Econometrics* 34(5): 724–45.

Windmeijer, F. (2005). A finite sample correction for the variance of linear efficient two step GMM estimators. *Journal of Econometrics* 126, 25–51.

Wu, D. M. (1973). Alternative Tests of Independence Between Stochastic Regressors and Disturbances, *Econometrica,* 41, 733-750.

Yuan, M., And Y. Lin (2006). Model Selection And Estimation In Regression With Grouped Variables. *Journal Of The Royal Statistical Society, Series B.* 68, 49–67.

Zellner, Arnold. 1962. An Efficient Method of Estimating Seemingly Unrelated Regressions and Tests for Aggregation Bias. *Journal of the American Statistical Association* 57(298): 348–68.

Zellner, A., and Vandaele, W. (1975). Bayes-Stain estimators for k-means, regression and simultaneous equation models . In *Studies in Bayesian Econometrics and Statistics* (eds. S. E. Fienberg and A. Zellner), North-Holland Publishing Company, Amsterdam, 627-653.

Ziliak, J.P. (1997). Efficient estimation with panel data when instruments are predetermined: An empirical comparison of moment-condition estimators. *Journal of Business & Economic Statistics* 15, 419–431.

Zhang, C.H., 2010. Nearly unbiased variable selection under minimax concave penalty. *Annals of Statistics* 38: 894–942.

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

A

Appendix A

The lemmas in Appendix A are used without special comments in the proofs of theorems in the next appendices.

Lemma A.1 *Let B_1 and B_2 be arbitrary $T \times T$ matrices. Let the $T \times 1$ random vector ϵ be such that $\epsilon \sim N(0, \sigma^2 \boldsymbol{I}_T)$, then the following results hold:*

$$\mathbb{E}\Big[(\epsilon' B_1 \epsilon)(\epsilon' B_2 \epsilon)\Big] = \sigma^4 \Big[\, tr(B_1)\, tr(B_2) + tr(B_1 B_2) + tr(B_1 B_2')\Big];$$

$$\mathbb{E}\Big[\epsilon\epsilon' B_1 \epsilon\epsilon'\Big] = \sigma^4 \Big[\, tr(B_1)\boldsymbol{I}_T + B_1 + B_1'\Big];$$

$$\begin{aligned}\mathbb{E}(\epsilon\epsilon' A \epsilon\epsilon' B \epsilon\epsilon') = \sigma^6 \Bigg[& \Big[\, tr(B_1)\, tr(B_2) + tr(B_1 B_2) + tr(B_1 B_2')\Big] I_T \\ & + tr(B_1)B_2 + tr(B_1)B_2' + tr(B_2)B_1 + tr(B_2)B_1' \\ & + B_1 B_2 + B_1' B_2 + B_1 B_2' + B_1' B_2' + B_2 B_1 + B_2' B_1 + B_2 B_1' + B_2' B_1'\Bigg],\end{aligned}$$

Proof: see Ullah (2004).

■

Lemma A.2 *Let A be a square constant matrix, and Ψ is $T \times N$ where its rows are independently normally distributes as $N(0, C_2)$. Then,*

i) $\mathbb{E}(\Psi' A \Psi) = tr(A) C_2$

ii) $\mathbb{E}(\Psi A \Psi') = tr(C_2 B) I_T$

iii) $\mathbb{E}(\Psi A \Psi) = A' C_2$

Proof: See Kadane (1971), Lemmas B1-B3. ■

Lemma A.3 *Let the $J \times 1$ vector ν is distributed normally with mean vector θ and covariance matrix I_J, and A is any $J \times J$ idempotent matrix. Also assume $\phi(.)$ is a Borel measurable function. Then*

$$\mathbb{E}\left[\phi(\nu' A \nu)\nu\right] = \mathbb{E}\left[\phi(\chi^2_{r+2}(\theta' A \theta/2))\right] A\theta + \mathbb{E}\left[\phi(\chi^2_r(\theta' A \theta/2))\right]\Big(I_J - A\Big)\theta,$$

where $r = rank(A) = tr(A)$.

Proof: Let P be an orthogonal matrix such that

$$PAP' = D = \begin{bmatrix} d_1 & 0 & \dots & 0 \\ 0 & d_2 & & \\ & & \vdots & \\ 0 & \dots & 0 & d_J \end{bmatrix} = \begin{bmatrix} I_r & \mathbf{0} \\ \mathbf{0} & \mathbf{0}_{J-r} \end{bmatrix}; \quad d_i \in \{0, 1\}.$$

Define the $J \times 1$ vector $\omega = (\omega_1, \dots, \omega_J)' = P\nu$, which has a $N(P\theta, I_J)$ distribution. Therefore

$$\mathbb{E}\left[\phi(\nu' A \nu)\nu\right] = \mathbb{E}\left[\phi(\omega' D \omega) P' \omega\right] = P' \mathbb{E}\left[\phi(\omega' D \omega)\omega\right].$$

Note that

$$\mathbb{E}\Big[\phi(\omega' D\omega)\omega\Big] = \Bigg[\mathbb{E}\Big[\mathbb{E}[\phi\Big(d_1\omega_1^2 + \sum_{j=2}^{J} d_j\omega_j^2\Big)\omega_1|\omega_j,\, j\neq 1]\Big], \\ \ldots, \mathbb{E}\Big[\mathbb{E}[\phi\Big(d_J\omega_J^2 + \sum_{j=1}^{J-1} d_j\omega_j^2\Big)\omega_J|\omega_j,\, j\neq J]\Big]\Bigg]'.$$

We now derive the expectation of the ith elements of the above equation,

$$\mathbb{E}\Big[\phi(\omega' D\omega)\omega_i\Big] = p_i'\theta\,\mathbb{E}\left[\mathbb{E}\Big[\phi\Big(d_i\chi_3^2((p_i'\theta)^2/2) + \sum_{j\neq i}\omega_j^2 d_j\Big)|\omega_j,\, j\neq i\Big]\right]$$

$$= \begin{cases} p_i'\theta\,\mathbb{E}\Big[\phi(\chi_{r+2}^2(\theta' A\theta/2))\Big], & \text{if } d_i = 1 \\ p_i'\theta\,\mathbb{E}\Big[\phi(\chi_r^2(\theta' A\theta/2))\Big], & \text{if } d_i = 0 \end{cases}$$

where the first equality holds by Lemma 1 of Appendix B.1 Judge and Bock (1978). Hence,

$$\begin{aligned}\mathbb{E}\Big[\phi(\nu' A\nu)\nu\Big] &= P'\,\mathbb{E}\Big[\phi(\omega' D\omega)\omega\Big] \\ &= P'DP\theta\,\mathbb{E}\Big[\phi(\chi_{r+2}^2(\theta' A\theta/2))\Big] + P'(I-D)P\theta\,\mathbb{E}\Big[\phi(\chi_r^2(\theta' A\theta/2))\Big] \\ &= A\theta\,\mathbb{E}\Big[\phi(\chi_{r+2}^2(\theta' A\theta/2))\Big] + (I-A)\theta\,\mathbb{E}\Big[\phi(\chi_r^2(\theta' A\theta/2))\Big],\end{aligned}$$

which completes the proof. ■

Lemma A.4 *Let the $J\times 1$ vector ν is distributed normally with mean vector θ and covariance matrix I_J, and A is any $J\times J$ idempotent matrix. Also assume $\phi(.)$ is a Borel measurable function. Then*

$$\begin{aligned}\mathbb{E}\Big[\phi(\nu' A\nu)\nu\nu'\Big] &= \mathbb{E}\Big[\phi(\chi_{r+2}^2(\theta' A\theta/2))\Big]A + \mathbb{E}\Big[\phi(\chi_r^2(\theta' A\theta/2))\Big]\Big(I_J - A\Big) \\ &\quad + \mathbb{E}\Big[\phi(\chi_{r+4}^2(\theta' A\theta/2))\Big]A\theta\theta' A + \mathbb{E}\Big[\phi(\chi_r^2(\theta' A\theta/2))\Big]\Big(I_J - A\Big)\theta\theta'\Big(I_J - A\Big) \\ &\quad + \mathbb{E}\Big[\phi(\chi_{r+2}^2(\theta' A\theta/2))\Big]\Big(\theta\theta' A + A\theta\theta' - 2A\theta\theta' A\Big),\end{aligned}$$

where $r = rank(A) = tr(A)$.

Proof: Let P be an orthogonal matrix such that

$$PAP' = D = \begin{bmatrix} d_1 & 0 & \dots & 0 \\ 0 & d_2 & & \\ & & \vdots & \\ 0 & \dots & 0 & d_J \end{bmatrix} = \begin{bmatrix} I_r & \mathbf{0} \\ \mathbf{0} & \mathbf{0}_{J-r} \end{bmatrix}; \quad d_i \in \{0,1\}.$$

Define the $J \times 1$ vector $\omega = (\omega_1, \dots, \omega_J)' = P\nu$, which has a $N(P\theta, I_J)$ distribution. Therefore

$$\mathbb{E}\left[\phi(\nu' A\nu)\nu\nu'\right] = \mathbb{E}\left[\phi(\omega' D\omega)P'\omega\omega' P\right] = P'\,\mathbb{E}\left[\phi(\omega' D\omega)\omega\omega'\right]P.$$

We first determine the diagonal and off-diagonal elements of $\mathbb{E}[\phi(\omega' D\omega)\omega\omega']$. The diagonal elements are of the form

$$\begin{aligned}
\mathbb{E}\left[\phi\Big(\sum_{j=1}^{J} d_j\omega_j^2\Big)\omega_i^2\right] &= \mathbb{E}\left[\mathbb{E}\left[\phi\Big(d_i\omega_i^2 + \sum_{j\neq i}\omega_j^2\Big)\omega_i^2 | \omega_j^2, j \neq i\right]\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\phi\Big(d_i\chi_3^2((P_i'\theta)^2/2) + \sum_{j\neq i}\omega_j^2\Big)|\omega_j^2, j\neq i\right]\right] \\
&\quad + (P_i'\theta)^2\,\mathbb{E}\left[\mathbb{E}\left[\phi\Big(d_i\chi_5^2((P_i'\theta)^2/2) + \sum_{j\neq i}\omega_j^2\Big)|\omega_j^2, j\neq i\right]\right] \\
&= \begin{cases} \mathbb{E}\left[\phi(\chi_{r+2}^2(\theta' A\theta/2)\right] + (P_i'\theta)^2\,\mathbb{E}[\phi(\chi_{r+4}^2(\theta' A\theta/2)\Big], & \text{if } d_i = 1 \\ \mathbb{E}\left[\phi(\chi_r^2(\theta' A\theta/2)\right] + (P_i'\theta)^2\,\mathbb{E}[\phi(\chi_r^2(\theta' A\theta/2)\Big], & \text{if } d_i = 0 \end{cases}
\end{aligned}$$

where the second equality holds by Lemma 1 of Appendix B.1 Judge and Bock (1978). Hence, the matrix form of the diagonal elements can be written as

$$\begin{aligned}
&D\,\mathbb{E}\left[\phi(\chi_{r+2}^2(\theta' A\theta/2))\right] + \mathbb{E}\left[\phi(\chi_{r+4}^2(\theta' A\theta/2))\right] diag(DP\theta\theta' P' D) \\
&+ (I_J - D)\,\mathbb{E}\left[\phi(\chi_r^2(\theta' A\theta/2))\right] + \mathbb{E}\left[\phi(\chi_r^2(\theta' A\theta/2))\right] diag((I_J - D)P\theta\theta' P'(I_J - D)).
\end{aligned}$$

For any $i \neq j$, the off-diagonal element is as follows

$$\mathbb{E}\Big[\phi\Big(\sum_{k=1}^{J} d_k\omega_k^2\Big)\omega_i\omega_j\Big] = \mathbb{E}\left[\omega_j\,\mathbb{E}\left[\phi\Big(d_i\omega_i^2 + \sum_{k\neq i} d_k\omega_k^2\Big)\omega_i|\omega_k, k\neq i\right]\right]$$

$$= \mathbb{E}\left[\omega_j P_i'\theta\,\mathbb{E}\left[\phi\Big(d_i\chi_3^2((P_i'\theta)^2/2) + \sum_{k\neq i} d_k\omega_k^2\Big)|\omega_k, k\neq i\right]\right]$$

$$= \mathbb{E}\left[\omega_j P_i'\theta_i\,\mathbb{E}\left[\phi\Big(d_i\chi_3^2((P_i'\theta)^2/2) + d_j\omega_j^2 + \sum_{k\neq i\&j} d_k\omega_k^2\Big)|\chi_3^2((P_i'\theta)^2/2), \omega_k, k\neq i\&j\right]\right]$$

$$= P_i'\theta P_j'\theta\,\mathbb{E}\left[\phi\Big(d_i\chi_3^2((P_i'\theta)^2/2) + d_j\chi_3^2((P_j'\theta)^2/2) + \sum_{k\neq i\&j} d_k\omega_k^2\Big)\right]$$

$$= P_i'\theta P_j'\theta \begin{cases} \mathbb{E}[\phi(\chi_{r+4}(\theta' A\theta/2))], & \text{if } d_i = d_j = 1 \\ \mathbb{E}[\phi(\chi_{r+2}(\theta' A\theta/2))], & \text{if } d_i = 1 \text{ and } d_j = 0 \\ \mathbb{E}[\phi(\chi_{r}(\theta' A\theta/2))], & \text{if } d_i = d_j = 0 \end{cases}$$

where the second equality holds by lemma 2 of Appendix B.1 Judge and Bock (1978). Hence, the off-diagonal matrix can be written as

$$\mathbb{E}\Big[\phi(\chi_{r+4}^2(\theta' A\theta/2))\Big](DP\theta\theta' P'D - diag(DP\theta\theta' P'D))$$

$$+\,\mathbb{E}\Big[\phi(\chi_{r}^2(\theta' A\theta/2))\Big]((I_J - D)P\theta\theta' P'(I_J - D) - diag((I_J - D)P\theta\theta' P'(I_J - D)))$$

$$+\,\mathbb{E}\Big[\phi(\chi_{r+2}^2(\theta' A\theta/2))\Big](P\theta\theta' P' - DP\theta\theta' P'D - (I_J - D)P\theta\theta' P'(I_J - D)).$$

Therefore, combining the diagonal and off-diagonal components, completes the proof. ■

Lemma A.5 *Let $\chi_\alpha^2(\lambda)$ denote a non-central chi-square random variable with noncentrally parameter λ and α degree of freedom. Also let α denote a positive integer such that $\alpha > 2p$. Then*

$$\mathbb{E}\left[\Big(\chi_\alpha^2(\lambda)\Big)^{-p}\right] = 2^{-p}\,e^{-\lambda}\,\frac{\Gamma(\frac{\alpha}{2} - p)}{\Gamma(\frac{\alpha}{2})}\,{}_1F_1\Big(\frac{\alpha}{2} - p\,,\,\frac{\alpha}{2}\,;\,\lambda\Big).$$

Proof: See Ullah (1974). ■

Lemma A.6 *If x is bounded and suppose $a, c \to \infty$ such that $\lim_{a,c\to\infty} \frac{(c-a)x}{c} = 0$. Then*

$$ {}_1F_1(a; c; x) = \exp(x)\Big[\sum_{j=0}^{p-1} \frac{(c-a)_j(-x)^j}{(c)_j j!} + O(|c|^{-p})\Big].$$

Proof: See Slater (1960), pp. 12, 65-66.

■

Lemma A.7 *Let M_1 and M_2 be two $T \times T$ idempotent matrices where $M_1 M_2 = \mathbf{0}$, and the $T \times 1$ vector $u \sim N(0, I_T)$. Then*

$$\mathbb{E}\left(u\frac{u'M_1u}{(u'M_2u)^2}u'\right) = \frac{tr(M_1)+2}{(tr(M_2)-2)(tr(M_2)-4)}M_1 + \frac{tr(M_1)}{tr(M_2)(tr(M_2)-2)}M_2 + \frac{tr(M_1)}{(tr(M_2)-2)(tr(M_2)-4)}, (I_T - M_1 - M_2)$$

when $tr(M_2) > 4$.

Proof: Since M_1 and M_2 commute, let the orthogonal matrix Γ simultaneously diagonalize them such that

$$\Gamma' M_1 \Gamma = \begin{bmatrix} I_{N_1} & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} = D_1, \text{ and } \Gamma' M_2 \Gamma = \begin{bmatrix} 0 & 0 & 0 \\ 0 & I_{N_2} & 0 \\ 0 & 0 & 0 \end{bmatrix} = D_1,$$

where $N_1 = tr(M_1)$ and $N_2 = tr(M_2)$. Further, define $\nu = \Gamma u$, and $\nu = (\nu_1', \nu_2', \nu_3')'$ be partitioned conformably with D_1 and D_2. Then

$$\begin{aligned}\mathbb{E}\Big(u\frac{u'M_1u}{(u'M_2u)^2}u'\Big) &= \Gamma\,\mathbb{E}\left(\nu\frac{\nu_1'\nu_1}{(\nu_2'\nu_2)\nu}\Big)\Gamma'\right. \\ &= \Gamma\left[\frac{N_1+2}{(N_2-2)(N_2-4)}D_1 + \frac{N_1}{N_2(N_2-2)}D_2 + \frac{N_1}{(N_2-2)(N_2-4)}(I_T - D_1 - D_2)\right]\Gamma',\end{aligned}$$

where the use has been made of Lemma A.1, Lemma A.3–A.5.

■

B

Appendix B

Proof of Theorem 2.8 :

First, we show that the single equation least-squares estimators are consistent, so that $\hat{\Sigma}$ is a consistent estimator of Σ.

Let us denote the vector of the single equation least-squares estimators as $\breve{\boldsymbol{\beta}} = (\breve{\beta}_1', \ldots, \breve{\beta}_N')'$, hence

$$\sqrt{T}(\breve{\boldsymbol{\beta}} - \boldsymbol{\beta}) = (\frac{1}{T}\sum_{t=1}^{T} X_{.t}' X_{.t})^{-1} \frac{1}{\sqrt{T}} \sum_{t=1}^{T} X_{.t}' u_{.t}. \tag{B.1}$$

Under Assumption 2.4 (i) and (iii), by weak law of large numbers (WLLN), we have $\frac{1}{T}\sum_{t=1}^{T} X_{.t}' X_{.t} \xrightarrow{p} \mathbb{E}(X_{.t}' X_{.t})$, and by Slutsky's theorem and the second part of Assumption 2.4 (iii), $(\frac{1}{T}\sum_{t=1}^{T} X_{.t}' X_{.t})^{-1} - [\mathbb{E}(X_{.t}' X_{.t})]^{-1} = o_p(1)$. Moreover, under Assumption 2.4 (i) and (ii), $\frac{1}{\sqrt{T}}\sum_{t=1}^{T} X_{.t}' u_{.t} = O_p(1)$. Therefore, we have $\breve{\boldsymbol{\beta}} - \boldsymbol{\beta} = o_p(1)$. Consequently, it is easy to show that $\widehat{\Sigma} = \frac{1}{T}\sum_{t=1}^{T} \hat{u}_{.t}\hat{u}_{.t}' = \Sigma + o_p(1)$, and by Slutsky's theorem and Assumption 2.1 (ii), we have $\widehat{\Sigma}^{-1} - \Sigma^{-1} = o_p(1)$.

Now, we show that the FGLS estimator and the infeasible GLS estimators are asymptotically equivalent, i.e. $\sqrt{T}(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{GLS}) = o_p(1)$, where the infeasible GLS estimator takes the form

$$\hat{\boldsymbol{\beta}}_{GLS} - \boldsymbol{\beta} = (\frac{1}{T}\sum_{t=1}^{T} X_{.t}'\Sigma^{-1}X_{.t})^{-1}\frac{1}{\sqrt{T}}\sum_{t=1}^{T} X_{.t}'\Sigma^{-1}u_{.t}. \tag{B.2}$$

Note that,

$$\begin{aligned}
\sqrt{T}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) &= \sqrt{T}(\hat{\boldsymbol{\beta}}_{GLS} - \boldsymbol{\beta}) \\
&+ \left[(\frac{1}{T}\sum_{t=1}^{T} X_{.t}'\widehat{\Sigma}^{-1}X_{.t})^{-1} - (\frac{1}{T}\sum_{t=1}^{T} X_{.t}'\Sigma^{-1}X_{.t})^{-1}\right]\frac{1}{\sqrt{T}}\sum_{t=1}^{T} X_{.t}'\Sigma^{-1}u_{.t} \\
&+ (\frac{1}{T}\sum_{t=1}^{T} X_{.t}'\Sigma^{-1}X_{.t})^{-1}\left[\frac{1}{\sqrt{T}}\sum_{t=1}^{T} X_{.t}'\widehat{\Sigma}^{-1}u_{.t} - \frac{1}{\sqrt{T}}\sum_{t=1}^{T} X_{.t}'\Sigma^{-1}u_{.t}\right] \\
&+ \left[(\frac{1}{T}\sum_{t=1}^{T} X_{.t}'\widehat{\Sigma}^{-1}X_{.t})^{-1} - (\frac{1}{T}\sum_{t=1}^{T} X_{.t}'\Sigma^{-1}X_{.t})^{-1}\right]\left[\frac{1}{\sqrt{T}}\sum_{t=1}^{T} X_{.t}'\widehat{\Sigma}^{-1}u_{.t} - \frac{1}{\sqrt{T}}\sum_{t=1}^{T} X_{.t}'\Sigma^{-1}u_{.t}\right] \\
&= \sqrt{T}(\hat{\boldsymbol{\beta}}_{GLS} - \boldsymbol{\beta}) + o_p(1)O_p(1) + O_p(1)o_p(1) + o_p(1)o_p(1)
\end{aligned} \tag{B.3}$$

where the last equality holds because by WLLN

$$\frac{1}{T}\sum_{t=1}^{T} X_{.t}'\widehat{\Sigma}^{-1}X_{.t} = \frac{1}{T}\sum_{t=1}^{T} X_{.t}'\Sigma^{-1}X_{.t} + \frac{1}{T}\sum_{t=1}^{T} X_{.t}'\Big(\widehat{\Sigma}^{-1} - \Sigma^{-1}\Big)X_{.t} = \frac{1}{T}\sum_{t=1}^{T} X_{.t}'\Sigma^{-1}X_{.t} + o_p(1), \tag{B.4}$$

and

$$\frac{1}{\sqrt{T}}\sum_{t=1}^{T} X_{.t}'\widehat{\Sigma}^{-1}u_{.t} = \frac{1}{\sqrt{T}}\sum_{t=1}^{T} X_{.t}'\Sigma^{-1}u_{.t} + \frac{1}{\sqrt{T}}\sum_{t=1}^{T} X_{.t}'\Big(\widehat{\Sigma}^{-1} - \Sigma^{-1}\Big)u_{.t} = \frac{1}{\sqrt{T}}\sum_{t=1}^{T} X_{.t}'\Sigma^{-1}u_{.t} + o_p(1). \tag{B.5}$$

Now, we derive the asymptotic distribution of the infeasible GLS estimator. But, first we note that by Assumption 2.4 (i)– (iii) and WLLN $\frac{1}{\sqrt{T}}\sum_{t=1}^{T} X_{.t}'\Sigma^{-1}X_{.t} \xrightarrow{p} \mathbb{E}(X_{.t}'\Sigma^{-1}X_{.t}) \equiv V^{-1}$, and by Slutsky's theorem $(\frac{1}{\sqrt{T}}\sum_{t=1}^{T} X_{.t}'\Sigma^{-1}X_{.t})^{-1} - V = o_p(1)$. hence

$$\begin{aligned}\sqrt{T}(\hat{\boldsymbol{\beta}}_{GLS} - \boldsymbol{\beta}) &= (\frac{1}{T}\sum_{t=1}^{T} X_{.t}'\Sigma^{-1}X_{.t})^{-1}\frac{1}{\sqrt{T}}\sum_{t=1}^{T} X_{.t}'\Sigma^{-1}u_{.t} \\ &= V\frac{1}{\sqrt{T}}\sum_{t=1}^{T} X_{.t}'\Sigma^{-1}u_{.t} + \Big[(\frac{1}{T}\sum_{t=1}^{T} X_{.t}'\Sigma^{-1}X_{.t})^{-1} - V\Big]\frac{1}{\sqrt{T}}\sum_{t=1}^{T} X_{.t}'\Sigma^{-1}u_{.t} \\ &= V\frac{1}{\sqrt{T}}\sum_{t=1}^{T} X_{.t}'\Sigma^{-1}u_{.t} + o_p(1).\end{aligned} \tag{B.6}$$

Further, under assumptions 2.1 and 2.4 and by the central limit theorem, we have $\sqrt{T}(\hat{\boldsymbol{\beta}}_{GLS} - \boldsymbol{\beta}) \xrightarrow{d} N(0, V)$. Consequently by the asymptotic equivalence of the FGLS and infeasible GLS estimators, equation (2.15) follows and $\sqrt{T}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{d} N(0, V)$.

Similarly, for the restricted estimator, we have

$$\begin{aligned}\sqrt{T}(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}) &= \sqrt{T}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) - VR'(RVR')^{-1}R\Big[\sqrt{T}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) + \sqrt{T}\,\boldsymbol{\beta}\Big] + o_p(1) \\ &= \sqrt{T}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) - VR'(RVR')^{-1}R\Big[\sqrt{T}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) + \sqrt{T}\boldsymbol{\alpha}\Big] + o_p(1) \\ &\xrightarrow{d} Z - VR'(RVR')^{-1}R(Z + \sqrt{T}\,\boldsymbol{\alpha}),\end{aligned} \tag{B.7}$$

where $Z \sim N(0, V)$, and the second equality holds by Assumption 2.5. Moreover, from the above equation we have

$$\sqrt{T}(\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}) \xrightarrow{d} VR'(RVR')^{-1}R(Z + \sqrt{T}\,\boldsymbol{\alpha}), \tag{B.8}$$

thus

$$F(\,\hat{\boldsymbol{\beta}}\,,\,\tilde{\boldsymbol{\beta}}\,) = T(\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}})V^{-1}(\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}) + o_p(1) \xrightarrow{d} (Z + \sqrt{T}\,\boldsymbol{\alpha})'R'(RVR')^{-1}R(Z + \sqrt{T}\,\boldsymbol{\alpha}),$$

(B.9)

which implies (2.17). The results of (2.18) and (2.19) follow by applying the continuous mapping theorem to the above results, and we omit their derivations to save space.

■

Proof of Theorem 2.9 :

Let $\zeta = \boldsymbol{\theta}' A \boldsymbol{\theta} /2$, and we note that since A is idempotent, $rank(A) = \text{tr}(A) = d$. Recall from equation (2.19) the asymptotic distribution of the shrinkage estimator is

$$\sqrt{T}(\hat{\boldsymbol{\beta}}_s - \boldsymbol{\beta}) \xrightarrow{d} Z_s = \omega(Z)Z + (1-\omega(Z))(Z - VR'(RVR')^{-1}R(Z + \sqrt{T}\,\boldsymbol{\alpha})),$$

hence, the asymptotic bias of $\sqrt{T}(\hat{\boldsymbol{\beta}}_s - \boldsymbol{\beta})$ is

$$\begin{aligned}
\text{ABias}\,(\hat{\boldsymbol{\beta}}_s) = \mathbb{E}(Z_s) &= \mathbb{E}(Z) - \tau VR'(RVR')^{-1}R\,\mathbb{E}\left(\frac{Z+\sqrt{T}\,\boldsymbol{\alpha}}{\xi(Z)}\right) \\
&= -\frac{\sqrt{T}\,\tau}{d}\, e^{-\zeta}\, VR'(RVR')^{-1}R\,\boldsymbol{\alpha}\ {}_1F_1\Big(\frac{d}{2}, \frac{d}{2}+1; \zeta\Big) \\
&\quad - \tau VR'(RVR')^{-1}RV^{1/2}(I_d - A)\,\boldsymbol{\theta}\,\mathbb{E}\left[(\chi^2_d(\theta' A\theta/2))^{-1}\right] \\
&= -\frac{\sqrt{T}\,\tau}{d}\, e^{-\zeta}\, VR'(RVR')^{-1}R\,\boldsymbol{\alpha}\ {}_1F_1\Big(\frac{d}{2}, \frac{d}{2}+1; \zeta\Big), \qquad \text{(B.10)}
\end{aligned}$$

where the last equality holds because by using Lemma A.3 and Lemma A.5

$$\mathbb{E}\left(\frac{Z+\sqrt{T}\,\boldsymbol{\alpha}}{\xi(Z)}\right) = V^{1/2}\,\mathbb{E}\left[V^{-1/2}\frac{Z+\sqrt{T}\,\boldsymbol{\alpha}}{(Z+\sqrt{T}\,\boldsymbol{\alpha})'R'(RVR')^{-1}R(Z+\sqrt{T}\,\boldsymbol{\alpha})}\right] = V^{1/2}\,\mathbb{E}\left(\frac{\nu}{\nu' A \nu}\right)$$

$$= V^{1/2}\left[\frac{1}{2}A\boldsymbol{\theta}\ e^{-\zeta}\,\frac{\Gamma(\frac{d}{2})}{\Gamma(\frac{d}{2}+1)}{}_1F_1(\frac{d}{2}, \frac{d}{2}+1; \zeta) + \mathbb{E}\left[(\chi^2_d(\theta' A\theta/2))^{-1}\right](I_d - A)\,\boldsymbol{\theta}\right]$$

$$= \frac{\sqrt{T}}{d} VR'(RVR')^{-1}R\,\boldsymbol{\alpha}\ e^{-\zeta}\,{}_1F_1(\frac{d}{2}, \frac{d}{2}+1; \zeta) + \mathbb{E}\left[(\chi^2_d(\theta' A\theta/2))^{-1}\right]V^{1/2}(I_d - A)\,\boldsymbol{\theta},$$

(B.11)

where $\nu = V^{-1/2}(Z + \sqrt{T}\,\boldsymbol{\alpha}) \sim N(\boldsymbol{\theta}, I_k)$.

Now we derive the expression for the asymptotic MSEM of the shrinkage estimator.

$$\text{AMSEM } (\hat{\boldsymbol{\beta}}_s) = \mathbb{E}(Z_s Z_s') = \mathbb{E}\left[\bar{\Pi}_1 - \bar{\Pi}_2 - \bar{\Pi}_2' + \bar{\Pi}_3\right], \tag{B.12}$$

where

$$\bar{\Pi}_1 = ZZ',$$

$$\bar{\Pi}_2 = \frac{\tau}{\xi(Z)} VR'(RVR')^{-1}R(Z + \sqrt{T}\,\boldsymbol{\alpha})Z',$$

$$\bar{\Pi}_3 = \frac{\tau^2}{\xi^2(Z)} VR'(RVR')^{-1}R(Z + \sqrt{T}\,\boldsymbol{\alpha})(Z + \sqrt{T}\,\boldsymbol{\alpha})'R'(RVR')^{-1}RV.$$

Now, in the following, we derive the expectations of $\bar{\Pi}_1$, $\bar{\Pi}_2$, and $\bar{\Pi}_3$,

$$\mathbb{E}(\bar{\Pi}_1) = \mathbb{E}(ZZ') = V, \tag{B.13}$$

$$\mathbb{E}(\bar{\Pi}_2) = \tau VR'(RVR')^{-1}R\,\mathbb{E}\left(\frac{(Z + \sqrt{T}\,\boldsymbol{\alpha})Z'}{\xi(Z)}\right)$$

$$= \tau VR'(RVR')^{-1}RV^{1/2}\,\mathbb{E}\left[\frac{V^{-1/2}(Z + \sqrt{T}\,\boldsymbol{\alpha})(Z + \sqrt{T}\,\boldsymbol{\alpha})'V^{-1/2}V^{1/2}}{\xi(Z)}\right]$$

$$- \tau VR'(RVR')^{-1}RV^{1/2}\,\mathbb{E}\left[\frac{V^{-1/2}(Z + \sqrt{T}\,\boldsymbol{\alpha})\sqrt{T}\,\boldsymbol{\alpha}'}{\xi(Z)}\right]$$

$$= \tau VR(RVR')^{-1}R\Big[V^{1/2}\,\mathbb{E}\left[(\nu' A\nu)^{-1}\nu\nu'\right]V^{1/2} - V^{1/2}\,\mathbb{E}\left[(\nu' A\nu)^{-1}\nu\right]\sqrt{T}\,\boldsymbol{\alpha}'\Big]$$

$$= \frac{\tau}{2} VR'(RVR')^{-1}R\,e^{-\zeta}\left[V^{1/2}AV^{1/2}\,\frac{\Gamma(\frac{d}{2})}{\Gamma(\frac{d}{2}+1)}\,{}_1F_1\Big(\frac{d}{2}, \frac{d}{2}+1; \zeta\Big)\right.$$

$$\left. + (V - V^{1/2}AV^{1/2})\,\mathbb{E}\left[(\chi^2_d(\theta' A\theta/2))^{-1}\right] + V^{1/2}A\,\boldsymbol{\theta}\,\boldsymbol{\theta}'\,AV^{1/2}\,\frac{\Gamma(\frac{d}{2}+1)}{\Gamma(\frac{d}{2}+2)}\,{}_1F_1\Big(\frac{d}{2}+1, \frac{d}{2}+2; \zeta\Big)\right.$$

$$+ V^{1/2}(I-A)\,\boldsymbol{\theta}\,\boldsymbol{\theta}'(I-A)V^{1/2}\,\mathbb{E}\left[(\chi^2_d(\theta' A\theta/2))^{-1}\right]$$

$$+ V^{1/2}(\boldsymbol{\theta}\,\boldsymbol{\theta}'\,A + A\,\boldsymbol{\theta}\,\boldsymbol{\theta}' - 2A\,\boldsymbol{\theta}\,\boldsymbol{\theta}'\,A)V^{1/2}\,\frac{\Gamma(\frac{d}{2})}{\Gamma(\frac{d}{2}+1)}\,{}_1F_1\Big(\frac{d}{2},\frac{d}{2}+1;\zeta\Big)$$

$$- V^{1/2}A\,\boldsymbol{\theta}\,\sqrt{T}\,\boldsymbol{\alpha}'\,\frac{\Gamma(\frac{d}{2})}{\Gamma(\frac{d}{2}+1)}\,{}_1F_1\Big(\frac{d}{2},\frac{d}{2}+1;\zeta\Big) - V^{1/2}(I-A)\,\boldsymbol{\theta}\,\sqrt{T}\,\boldsymbol{\alpha}'\,\mathbb{E}\left[(\chi^2_d(\theta' A\theta/2))^{-1}\right]\Big]$$

$$= \tau\,\frac{1}{d}\,e^{-\zeta}\,{}_1F_1\Big(\frac{d}{2},\frac{d}{2}+1;\zeta\Big)VR'(RVR')^{-1}RV + \tau T\,e^{-\zeta}\,VR'(RVR')^{-1}R\,\boldsymbol{\alpha}\,\boldsymbol{\alpha}'\,R'(RVR')^{-1}RV$$

$$\Big[\frac{1}{d+2}\,{}_1F_1\Big(\frac{d}{2}+1,\frac{d}{2}+2;\zeta\Big) - \frac{1}{d}\,{}_1F_1\Big(\frac{d}{2},\frac{d}{2}+1;\zeta\Big)\Big], \tag{B.14}$$

where the use has been made of Lemma A.3–Lemma A.5, and finally, provided $d > 2$

$$\mathbb{E}(\bar{\Pi}_3) = \tau^2 VR'(RVR')^{-1}R\,\mathbb{E}\left[\frac{(Z+\sqrt{T}\,\boldsymbol{\alpha})(Z+\sqrt{T}\,\boldsymbol{\alpha})'}{\xi^2(Z)}\right]R'(RVR')^{-1}RV$$

$$= \tau^2 VR'(RVR')^{-1}RV^{1/2}\,\mathbb{E}\left[(\nu' A\nu)^{-2}\nu\nu'\right]V^{1/2}R'(RVR')^{-1}RV$$

$$= \tau^2\,\frac{1}{4}\,e^{-\zeta}\,VR'(RVR')^{-1}R\Bigg[V^{1/2}AV^{1/2}\,\frac{\Gamma(\frac{d}{2}-1)}{\Gamma(\frac{d}{2}+1)}\,{}_1F_1\Big(\frac{d}{2}-1,\frac{d}{2}+1;\zeta\Big)$$

$$+ (V - V^{1/2}AV^{1/2})\,\mathbb{E}\left[(\chi^2_d(\theta' A\theta/2))^{-2}\right] + V^{1/2}A\,\boldsymbol{\theta}\,\boldsymbol{\theta}'\,AV^{1/2}\,\frac{\Gamma(\frac{d}{2})}{\Gamma(\frac{d}{2}+2)}\,{}_1F_1\Big(\frac{d}{2},\frac{d}{2}+2;\zeta\Big)$$

$$+ V^{1/2}(I-A)\,\boldsymbol{\theta}\,\boldsymbol{\theta}'(I-A)V^{1/2}\,\mathbb{E}\left[(\chi^2_d(\theta' A\theta/2))^{-2}\right]$$

$$+ V^{1/2}(\boldsymbol{\theta}\,\boldsymbol{\theta}'\,A + A\,\boldsymbol{\theta}\,\boldsymbol{\theta}' - 2A\,\boldsymbol{\theta}\,\boldsymbol{\theta}'\,A)V^{1/2}\,\frac{\Gamma(\frac{d}{2}-1)}{\Gamma(\frac{d}{2}+1)}\,{}_1F_1\Big(\frac{d}{2}-1,\frac{d}{2}+1;\zeta\Big)\Bigg]R'(RVR')^{-1}RV$$

$$= \tau^2\,e^{-\zeta}\,\frac{1}{d(d-2)}\,VR'(RVR')^{-1}RV\,{}_1F_1\Big(\frac{d}{2}-1,\frac{d}{2}+1;\zeta\Big)$$

$$+ \tau^2\,e^{-\zeta}\,\frac{1}{d(d+2)}\,T\,VR'(RVR')^{-1}R\,\boldsymbol{\alpha}\,\boldsymbol{\alpha}'\,R'(RVR')^{-1}RV\,{}_1F_1\Big(\frac{d}{2},\frac{d}{2}+2;\zeta\Big), \tag{B.15}$$

where the use has been made of Lemma A.4 and Lemma A.5.

Using the results in equations (B.13)–(B.15), the MSEM of the shrinkage estimator is obtained.

■

Proof of Corollary 2.10 and 2.15 :

The results hold by noting that if $x > 0$ and $a, c > 0$, then as $x \to \infty$,

$$ {}_1F_1(a, c; x) = \frac{\Gamma(c)}{\Gamma(a)}\, e^x\, x^{-(c-a)} \left[\sum_{j=0}^{p-1} \frac{(c-a)_j (1-a)_j}{j!} x^{-j} + O(|x|^{-p}) \right]. $$

See Lebedev (1972), pp. 271.

■

Proof of Corollary 2.12 :

We first note that we have

$$ \begin{aligned} \varrho_{min}(V^{1/2}BVWVBV^{1/2}) \leq \frac{\lambda_W}{\lambda} &= \frac{\boldsymbol{\delta}'\, BV^{1/2}V^{1/2}BVWVBV^{1/2}V^{1/2}B\,\boldsymbol{\delta}}{\boldsymbol{\delta}'\, BV^{1/2}V^{1/2}B\,\boldsymbol{\delta}} \\ &\leq \varrho_{max}(V^{1/2}BVWVBV^{1/2}), \end{aligned} \tag{B.16} $$

where $B = R'(RVR')^{-1}R$, and $BVB = B$. [1]

[1] The inequality holds by noting that for any symmetric $n \times n$ matrix Q, we have

$$ \varrho_{min}(Q) \leq \frac{\theta' Q \theta}{\theta'\theta} \leq \varrho_{max}(Q), $$

see Abadir and Magnus (2005)-Pages 181-182.

From Remark 2.11, we have

$$\begin{aligned}
\text{ARisk}\,(\hat{\boldsymbol{\beta}}_s) &= \text{ARisk}\,(\hat{\boldsymbol{\beta}}) + e^{-\lambda}\,\frac{\text{tr}(C)}{d(d-2)}\,{}_1F_1\Big(\frac{d}{2}-1\,,\,\frac{d}{2}+1\,;\,\lambda\Big)\Big[\tau^2-2\tau(d-2)\Big] \\
&\quad + e^{-\lambda}\,\frac{2\lambda_W}{d(d+2)}\,{}_1F_1\Big(\frac{d}{2}\,,\,\frac{d}{2}+2\,;\,\lambda\Big)\Big[\tau^2-2\tau\Big(\frac{\text{tr}(C)\lambda}{\lambda_W}-2\Big)\Big] \\
&\le \text{ARisk}\,(\hat{\boldsymbol{\beta}}) + e^{-\lambda}\,\frac{\text{tr}(C)}{d(d-2)}\,{}_1F_1\Big(\frac{d}{2}-1\,,\,\frac{d}{2}+1\,;\,\lambda\Big)\Big[\tau^2-2\tau(d-2)\Big] \\
&\quad + e^{-\lambda}\,\frac{2\lambda_W}{d(d+2)}\,{}_1F_1\Big(\frac{d}{2}\,,\,\frac{d}{2}+2\,;\,\lambda\Big)\Big[\tau^2-2\tau\Big(\frac{\text{tr}(C)}{\varrho_{max}(C)}-2\Big)\Big],
\end{aligned} \tag{B.17}$$

where the inequality holds by (B.16). Moreover, since $d \ge \text{tr}(C)/\varrho_{max}(C)$, the result in (2.27) follows straightforwardly.

■

Proof of Corollary 2.14 :

Using equation (2.25), and the following identities (see Lebedev (1972), pp. 271)

$$(c-a-1)\,{}_1F_1(a,c;x) = (c-1)\,{}_1F_1(a,c-1;x) - a\,{}_1F_1(a+1,c;x),$$

$${}_1F_1(a,c;x) = {}_1F_1(a+1,c;x) - \frac{x}{c}\,{}_1F_1(a+1,c+1;x),$$

$${}_1F_1(a,c;x) = \frac{c-a}{c}\,{}_1F_1(a,c+1;x) + \frac{a}{c}\,{}_1F_1(a+1,c+1;x),$$

when $W = V^{-1}$, we have

$$\begin{aligned}
\text{Risk}\,(\hat{\boldsymbol{\beta}}_s) &= \text{tr}(W\,V) + e^{-\lambda}\,\frac{1}{d-2}\,{}_1F_1\Big(\frac{d}{2}-1\,,\,\frac{d}{2}\,;\,\lambda\Big)\Big[\frac{\lambda_W}{\lambda}\tau^2-2\tau\Big(\text{tr}(WVBV)-2\frac{\lambda_W}{\lambda}\Big)\Big] \\
&\quad - e^{-\lambda}\,\frac{1}{d-2}\,{}_1F_1\Big(\frac{d}{2}-1\,,\,\frac{d}{2}+1\,;\,\lambda\Big)\Big(\frac{\lambda_W}{\lambda}-\frac{\text{tr}(WVBV)}{d}\Big)\Big[\tau^2+4\tau\Big] \\
&= \text{tr}(I_{Nk}) + e^{-\lambda}\,\frac{1}{d-2}\,{}_1F_1\Big(\frac{d}{2}-1\,,\,\frac{d}{2}\,;\,\lambda\Big)\Big[\tau^2-2\tau(d-2)\Big],
\end{aligned} \tag{B.18}$$

where the last equality holds because the third term on the right hand side of the first equality is zero, $\text{tr}(BV) = d$, and $\lambda_W = \lambda$. ■

Proof of Theorem 2.16 :

Using the results of Corollary 2.12 and Lemma A.6, and noting that $\tau/d \to 1$ as $d \to \infty$, we have

$$\text{ARisk } (\hat{\boldsymbol{\beta}}_{s,opt}) \leq \text{tr}(W\,V) - \tau_{opt}[1 + O(\frac{1}{d})] \tag{B.19}$$

and by dividing both sides by the asymptotic risk of the FGLS estimator, we have

$$\lim_{d\to\infty} \frac{\text{ARisk } (\hat{\boldsymbol{\beta}}_s)}{\text{ARisk } (\hat{\boldsymbol{\beta}})} \leq 1 - \rho[1 + O(\frac{1}{d})], \qquad \rho = \lim_{d\to\infty} \frac{tr(C)}{tr(WV)}. \tag{B.20}$$

■

C

Appendix C

Lemma C.1 *Let* $M_l = Z_l(Z_l'Z_l)^{-1}Z_l'$, *for* $l = 1, 2$ *be an* $N(T-1) \times N(T-1)$ *idempotent matrix where* $Z_l = (Z_{l1}', \ldots, Z_{lN}')'$ *is* $N(T-1) \times m_l$, *with* $m_1 = k(T-1)$ *and* $m_2 = kT(T-1)/2$. *Also define* $H_l = P'M_lPL\Gamma$, *where* $P = I_N \otimes P_T$, P_T *is a* $(T-1) \times T$ *upper-triangular matrix with rank* $T-1$ *and* $P_TP_T' = I_{T-1}$, $P_T'P_T = R_T = I_T - \frac{1}{T}\iota_T\iota_T'$. *We have*

$$i) \quad tr(H_l) = \begin{cases} -\frac{k}{1-\gamma}\left[1 - \frac{1}{T}\frac{1-\gamma^T}{1-\gamma}\right], & \text{if } l = 1, \\ \\ -\frac{k}{1-\gamma}T + O(1), & \text{if } l = 2; \end{cases}$$

$$ii) \quad tr(H_l^2) = O(1), \qquad l = 1, 2;$$

$$iii) \quad tr(H_l'H_l) = \frac{NT}{1-\gamma^2} + O(N), \qquad l = 1, 2.$$

Proof: Let B *be an* $N(T-1) \times N(T-1)$ *orthogonal* ($B'B = I_{N(T-1)}$) *permutation matrix, which changes the order of the rows of* Z_l *such that* $T-1$ *sub-matrices of* N

rows are put together, instead of $T-1$ *rows for the separate individuals. That is,* $BZ_l = diag\Big[Z'_{l,1},\ldots,Z'_{l,T-1}\Big]$, *where* $Z_{l,t}$ *is* $N \times m_{l,t}$ *with* $m_{1,t} = kt^{l-1}$ *for* $l = 1, 2$. *Hence*

$$BM_lB' = diag\Big[Z_{l,1}(Z'_{l,1}Z_{l,1})^{-1}Z_{l,1},\ldots,Z_{l,T-1}(Z'_{l,T-1}Z_{l,T-1})^{-1}Z_{l,T-1}\Big].$$

Therefore, we have

$$tr(H_l) = tr(P'B'BM_lBB'PL\Gamma) = \sum_{t=1}^{T} tr(M^*_{l,t})\, tr(p_tp'_tL_T\Gamma_T) = tr(L_T\Gamma_T\sum_{t=1}^{T} m_{l,t}p_tp'_t)$$

$$= \begin{cases} k\, tr(L_T\Gamma_TR_T) = -\frac{k}{1-\gamma}\Big[1-\frac{1}{T}\frac{1-\gamma^T}{1-\gamma}\Big], & = \text{if } l=1, \\ k\sum_{s=2}^{T} tr(L_T\Gamma_TJ_sR_sJ'_s) = k\sum_{s=2}^{T} tr(L_s\Gamma_sR_s) = -\frac{k}{1-\gamma}T + O(1), & \text{if } l=2, \end{cases}$$

where $J_s = (0, I_2)'$ *is a* $T \times s$ *selection matrix, and* p'_t *is the* t*th row of* P_T. *This completes the proof of* (i), *for* (ii) *we have*

$$tr(H_l^2) = tr(P'M_lPL\Gamma P'M_lPL\Gamma) \le \lambda^2_{max}(M_l)\, tr(P'PL\Gamma P'PL\Gamma) = tr(RL\Gamma RL\Gamma)$$

$$= tr(L\Gamma RL\Gamma) + O(1) = 0 + O(1),$$

where $R = I_N \otimes R_T$, *and* $\lambda_{max}(M_l)$ *denotes the maximum eigenvalue of* M_l, *which is equal to one, as it is an idempotent matrix. Also, the last equality holds because, the diagonal elements of* $L\Gamma$ *are all zero.*

For (iii), *note that we have*

$$tr(H'_lH_l) = tr(L'\Gamma'P'M_lPP'M_lPL\Gamma) = tr(\Gamma'L'P'M_lPL\Gamma) \le \lambda_{max}(M_l)\, tr(\Gamma'L'RL\Gamma)$$

$$= N\, tr(R_TL_T\Gamma_TL_T\Gamma_T) = N\Big[\sum_{t=1}^{T-1}(T-i)(\gamma^2)^{t-1} + O(1)\Big] = \frac{NT}{1-\gamma^2} + O(N).$$

■

Proof of Theorem 3.1 :

Let $D_2 = \left[W'P'M_2PW\right]^{-1}$, then

$$Q_2^{-1} \equiv \mathbb{E}(D_2^{-1}) = \mathbb{E}\left(\bar{W}'P'M_2P\bar{W}\right) + e_{k,1}e'_{k,1}\sigma_\epsilon^2 \operatorname{tr}(H_2'H_2) = O(NT), \tag{C.1}$$

hence we have

$$D_2 = \left[Q_2^{-1} + \left(D_2^{-1} - Q_2^{-1}\right)\right]^{-1} = Q_2\left[I_k + \left(D_2^{-1} - Q_2^{-1}\right)Q_2\right]^{-1} = Q_2 + o_p(\frac{1}{NT}). \tag{C.2}$$

therefore, equation (3.30) can be written as

$$\widehat{\delta}_{GMM,2} - \delta = Q_2W'P'M_2P\epsilon + o_p(\frac{1}{\sqrt{NT}}). \tag{C.3}$$

The bias of the estimator up to order $O(\frac{1}{\sqrt{NT}})$ is then

$$\begin{aligned}\mathbb{E}(\widehat{\delta}_{GMM,2} - \delta) &= Q_2\,\mathbb{E}(W'P'M_2P\epsilon) = Q_2\,\mathbb{E}(\bar{W}'P'M_2P\epsilon) + Q_2e_{k,1}\,\mathbb{E}(\epsilon'H_2\epsilon) \\ &= \sigma_\epsilon^2Q_2e_{k,1}\operatorname{tr}(H_2) = O_p(\frac{1}{N}),\end{aligned} \tag{C.4}$$

where the last equality holds by using the results in Lemma C.1.

The MSEM of the estimator up to order $O(\frac{1}{NT})$ is

$$\begin{aligned}\text{MSE }(\widehat{\delta}_{GMM,2}) &= \mathbb{E}\left[(\widehat{\delta}_{GMM,2} - \delta)(\widehat{\delta}_{GMM,2} - \delta)'\right] \\ &= Q_2\,\mathbb{E}\left[\left[\bar{W}'P'M_2P\epsilon + e_{k,1}\epsilon'H_2\epsilon\right]\left[\bar{W}'P'M_2P\epsilon + e_{k,1}\epsilon'H_2\epsilon\right]'\right]Q_2 \\ &= \sigma_\epsilon^4Q_2e_{k,1}e'_{k,1}Q_2\operatorname{tr}(H_2'H_2) + \sigma_\epsilon^2Q_2,\end{aligned} \tag{C.5}$$

where the use has been made of Lemma A.3, and note that $\operatorname{tr}(H_2'H_2) = O(NT)$, in light of Lemma C.1.

■

Proof of Theorem 3.2 :

Let $D_1 = \left[W'P'M_1PW\right]^{-1}$, then

$$Q_1^{-1} \equiv \mathbb{E}(D_1^{-1}) = \mathbb{E}\left(\bar{W}'P'M_1P\bar{W}\right) + e_{k,1}e'_{k,1}\sigma_\epsilon^2 \operatorname{tr}(H_1'H_1) = O(NT), \tag{C.6}$$

hence we have

$$D_1 = \left[Q_1^{-1} + \left(D_1^{-1} - Q_1^{-1}\right)\right]^{-1} = Q_1\left[I_k + \left(D_1^{-1} - Q_1^{-1}\right)Q_1\right]^{-1} = Q_1 + o_p(\frac{1}{NT}). \tag{C.7}$$

therefore, equation (3.33) can be written as

$$\widehat{\delta}_{GMM,1} - \delta = Q_1W'P'M_1P\epsilon + o_p(\frac{1}{\sqrt{NT}}) \tag{C.8}$$

The bias of the estimator up to order $O(\frac{1}{\sqrt{NT}})$ is then

$$\begin{aligned}\mathbb{E}(\widehat{\delta}_{GMM,1} - \delta) &= Q_1\,\mathbb{E}(W'P'M_1P\epsilon) = Q_1\,\mathbb{E}(\bar{W}'P'M_1P\epsilon) + Q_1e_{k,1}\,\mathbb{E}(\epsilon'H_1\epsilon) \\ &= \sigma_\epsilon^2Q_1e_{k,1}\operatorname{tr}(H_1) = 0 + O(\frac{1}{NT}),\end{aligned} \tag{C.9}$$

because by using Lemma C.1, $\operatorname{tr}(H_1) = O(1)$ and the MSEM of the estimator up to order $O(\frac{1}{NT})$ is

$$\begin{aligned}\text{MSE }(\widehat{\delta}_{GMM,1}) &= \mathbb{E}\left[(\widehat{\delta}_{GMM,1} - \delta)(\widehat{\delta}_{GMM,1} - \delta)'\right] \\ &= Q_1\,\mathbb{E}\left[\left[\bar{W}'P'M_1P\epsilon + e_{k,1}\epsilon'H_1\epsilon\right]\left[\bar{W}'P'M_1P\epsilon + e_{k,1}\epsilon'H_1\epsilon\right]'\right]Q_1 \\ &= \sigma_\epsilon^4Q_1e_{k,1}e'_{k,1}Q_1\operatorname{tr}(H_1'H_1) + \sigma_\epsilon^2Q_1 = O(\frac{1}{NT}),\end{aligned} \tag{C.10}$$

where the last equality holds by using Lemma C.1, and Lemma A.3.

■

Proof of Theorem 3.3 :

Using equations (C.3) and (C.8), we have

$$\begin{pmatrix} \widehat{\delta}_{GMM,2} - \delta \\ \\ \widehat{\delta}_{GMM,1} - \delta \end{pmatrix} = \begin{pmatrix} A_1\epsilon \\ \\ A_2\epsilon \end{pmatrix} + o_p(\frac{1}{\sqrt{NT}}) \equiv \zeta + o_p(\frac{1}{\sqrt{NT}}), \tag{C.11}$$

where $A_l = Q_l W' P' M_l P, \quad l = 1, 2$. Because ϵ has a normal distribution, then $\zeta \sim N(b, V)$ where

$$b = \begin{pmatrix} \sigma_\epsilon^2 Q_2 e_{k,1} \operatorname{tr}(H_2) \\ \\ 0 \end{pmatrix} \quad V = \begin{pmatrix} V_2 & V_2 \\ \\ V_2 & V_1 \end{pmatrix},$$

where V_1 and V_2 represent the variances of $\widehat{\delta}_{GMM,1}$ and $\widehat{\delta}_{GMM,2}$. Also, define $\nu = V^{-1/2}\zeta \sim N(\theta, I_{2k})$, where $\theta = V^{1/2}b$.

Note that, it is easily verified that

$$\hat{V}_l = \sigma_\epsilon^2 Q_l + o_p(\frac{1}{NT}), \tag{C.12}$$

hence we have

$$(\hat{V}_1 - \hat{V}_2)^{-1} = (V_1 - V_2)\Big[I_k + o_p(\frac{1}{NT})\Big]. \tag{C.13}$$

Therefore, F can be written as

$$F = \zeta' G' (V_1 - V_2)^{-1} G\zeta + o_p(1), \tag{C.14}$$

where $G = (-I_{2k}, I_{2k})$.

Using the above results, the averaging estimator can be written as

$$\widehat{\delta}_A - \delta = G_2\zeta - \frac{\tau}{\zeta' A \zeta} G\zeta + o_p(\frac{1}{NT}),$$

where $A = V^{1/2}G'(V_1 - V_2)^{-1}GV^{1/2}$ is an idempotent matrix, and $G_2 = (0, I_k)$.

Therefore, the approximate bias of the average estimator up to order $O_p(\frac{1}{\sqrt{NT}})$ is

$$\mathbb{E}\Big[(\widehat{\delta}_A - \delta)\Big] = G_2\,\mathbb{E}(\zeta) - \tau G V^{1/2}\,\mathbb{E}\left(\frac{\nu}{\nu' A \nu}\right) = -\frac{\tau}{k}\,e^{-\lambda/2}\,GV^{1/2}\,{}_1F_1\Big(\frac{k}{2}, \frac{k}{2}+1; \lambda/2\Big), \tag{C.15}$$

the equality above holds by using Lemma A.4 and Lemma A.5, and $\lambda = \theta' A \theta$.

Now we derive the expression for the asymptotic MSE of the averaging estimator.

$$\mathbb{E}\Big[(\widehat{\delta}_A - \delta)(\widehat{\delta}_A - \delta)'\Big] = \mathbb{E}\Big[\,\bar{\Pi}_1 - \bar{\Pi}_2 - \bar{\Pi}_2' + \bar{\Pi}_3\,\Big], \tag{C.16}$$

where

$$\bar{\Pi}_1 = G_2 \zeta \zeta' G_2',$$
$$\bar{\Pi}_2 = \frac{\tau}{\nu' A \nu} G V^{1/2} \nu \nu' V^{1/2} G_2',$$
$$\bar{\Pi}_3 = \frac{\tau^2}{(\nu' A \nu)^2} G V^{1/2} \nu \nu' V^{1/2} G'.$$

Now, in the following, we derive the expectations of $\bar{\Pi}_1 - \bar{\Pi}_3$,

$$\mathbb{E}(\bar{\Pi}_1) = G_2 V G_2' = V_1, \tag{C.17}$$

$$\mathbb{E}(\bar{\Pi}_2) = \tau G V^{1/2}\,\mathbb{E}\left(\frac{\nu\nu'}{\nu' A \nu}\right) V^{1/2} G_2'$$

$$= \frac{\tau}{2} G\, e^{-\lambda/2}\left[V^{1/2} A V^{1/2}\,\frac{\Gamma(\frac{k}{2})}{\Gamma(\frac{k}{2}+1)}\,{}_1F_1\Big(\frac{k}{2}, \frac{k}{2}+1; \lambda/2\Big)\right.$$

$$+ (V - V^{1/2} A V^{1/2})\;\mathbb{E}\Big[(\chi^2_k(\theta' A \theta/2))^{-1}\Big]$$

$$\left. + V^{1/2} A \theta \theta' A V^{1/2}\,\frac{\Gamma(\frac{k}{2}+1)}{\Gamma(\frac{k}{2}+2)}\,{}_1F_1\Big(\frac{k}{2}+1, \frac{k}{2}+2; \lambda/2\Big)\right.$$

$$
\begin{aligned}
&+ V^{1/2}(I_{2k} - A)\theta\theta'(I_{2k} - A)V^{1/2}\, \mathbb{E}\left[(\chi^2_k(\theta' A\theta/2))^{-1}\right] \\
&+ V^{1/2}(\theta\theta' A + A\theta\theta' - 2A\theta\theta' A)V^{1/2}\, \frac{\Gamma(\frac{k}{2})}{\Gamma(\frac{k}{2}+1)}\, {}_1F_1\Big(\frac{k}{2}, \frac{k}{2}+1; \lambda/2\Big)\Big] G_2' \\
&= \tau\, \frac{1}{k}\, e^{-\lambda/2}\, {}_1F_1\Big(\frac{k}{2}, \frac{k}{2}+1; \lambda/2\Big) GVG' \\
&+ \tau\, e^{-\lambda/2}\, GV^{1/2}\theta\theta' V^{1/2} G' \Big[\frac{1}{k+2}\, {}_1F_1\Big(\frac{k}{2}+1, \frac{k}{2}+2; \lambda/2\Big) - \frac{1}{k}\, {}_1F_1\Big(\frac{k}{2}, \frac{k}{2}+1; \lambda/2\Big)\Big],
\end{aligned}
\tag{C.18}
$$

where the fourth equality holds by using Lemma A.3–Lemma A.5.

$$
\begin{aligned}
\mathbb{E}(\bar{\Pi}_3) &= \tau^2 GV^{1/2}\, \mathbb{E}\left[\frac{\nu\nu'}{\nu' A\nu}\right] V^{1/2} G' \\
&= \tau^2\, \frac{1}{4}\, e^{-\lambda/2}\, G \Bigg[V^{1/2} A V^{1/2}\, \frac{\Gamma(\frac{k}{2}-1)}{\Gamma(\frac{k}{2}+1)}\, {}_1F_1\Big(\frac{k}{2}-1, \frac{k}{2}+1; \lambda/2\Big) \\
&+ (V - V^{1/2} A V^{1/2})\, \mathbb{E}\left[(\chi^2_k(\theta' A\theta/2))^{-2}\right] \\
&+ V^{1/2} A\theta\theta' A V^{1/2} \frac{1}{4}\, \frac{\Gamma(\frac{k}{2})}{\Gamma(\frac{k}{2}+2)}\, {}_1F_1\Big(\frac{k}{2}, \frac{k}{2}+2; \lambda/2\Big) \\
&+ V^{1/2}(I_{2k} - A)\theta\theta'(I_{2k} - A)V^{1/2}\, \mathbb{E}\Big[(\chi^2_k(\theta' A\theta/2))^{-2} \\
&+ V^{1/2}(\theta\theta' A + A\theta\theta' - 2A\theta\theta' A)V^{1/2}\, \frac{\Gamma(\frac{k}{2}-1)}{\Gamma(\frac{k}{2}+1)}\, {}_1F_1\Big(\frac{k}{2}-1, \frac{k}{2}+1; \lambda/2\Big)\Big] G' \\
&= \tau^2\, e^{-\lambda/2}\, \frac{1}{k(k-2)}\, GVG'\, {}_1F_1\Big(\frac{k}{2}-1, \frac{k}{2}+1; \lambda/2\Big) \\
&+ \tau^2\, e^{-\lambda/2}\, \frac{1}{k(k+2)}\, T\, GV^{1/2}\theta\theta' V^{1/2} G'\, {}_1F_1\Big(\frac{k}{2}, \frac{k}{2}+2; \lambda/2\Big),
\end{aligned}
\tag{C.19}
$$

where the third equality holds by using Lemma A.4 and Lemma A.5.

Using the results in equations (C.17)–(C.19), the MSE of the shrinkage estimator is obtained.

■

Proof of Corollary 3.5 :

Note that we have

$$\varrho_{min}\Big((V_1 - V_2)^{1/2} D (V_1 - V_2)^{1/2}\Big) \leq \frac{\lambda_D}{\lambda} = \frac{\theta' V^{1/2} G' D G V^{1/2} \theta}{\theta' V^{1/2} G' (V_1 - V_2)^{-1} G V^{1/2} \theta} \leq \varrho_{max}\Big((V_1 - V_2)^{1/2} D (V_1 - V_2)^{1/2}\Big), \tag{C.20}$$

where ϱ_{max} and ϱ_{min} denote the maximum and minimum eigenvalues. [1]

From Remark 3.4, we have

$$\begin{aligned}
\text{Risk } (\widehat{\delta}_A) &= \text{Risk } (\widehat{\delta}_{GMM,1}) + e^{-\lambda/2} \frac{\text{tr}(C)}{k(k-2)} {}_1F_1\Big(\frac{k}{2} - 1\,,\, \frac{k}{2} + 1\,;\, \lambda/2\Big)\Big[\tau^2 - 2\tau(k-2)\Big] \\
&\quad + e^{-\lambda/2} \frac{2\lambda_D}{k(k+2)} {}_1F_1\Big(\frac{k}{2}\,,\, \frac{k}{2} + 2\,;\, \lambda/2\Big)\Big[\tau^2 - 2\tau\Big(\frac{\text{tr}(C)\lambda}{\lambda_D} - 2\Big)\Big] \\
&\leq \text{Risk } (\widehat{\delta}_{GMM,1}) + e^{-\lambda/2} \frac{\text{tr}(C)}{d(d-2)} {}_1F_1\Big(\frac{k}{2} - 1\,,\, \frac{k}{2} + 1\,;\, \lambda/2\Big)\Big[\tau^2 - 2\tau(k-2)\Big] \\
&\quad + e^{-\lambda/2} \frac{2\lambda_D}{k(k+2)} {}_1F_1\Big(\frac{k}{2}\,,\, \frac{k}{2} + 2\,;\, \lambda/2\Big)\Big[\tau^2 - 2\tau\Big(\frac{\text{tr}(C)}{\varrho_{max}(C)} - 2\Big)\Big],
\end{aligned} \tag{C.21}$$

where the inequality holds by (C.20). Moreover, since $k \geq \text{tr}(C)/\varrho_{max}(C)$, the result in (3.44) follows straightforwardly.

■

[1] The inequality holds by noting that for any symmetric $n \times n$ matrix B, we have

$$\varrho_{min}(B) \leq \frac{\theta' B \theta}{\theta' \theta} \leq \varrho_{max}(B),$$

see Abadir and Magnus (2005)-Pages 181-182.

Proof of Corollary 3.6 :

Using equation (3.42), and the following identities (see Lebedev (1972), pp. 271)

$$(c-a-1)\,{}_1F_1(a,c;x) = (c-1)\,{}_1F_1(a,c-1;x) - a\,{}_1F_1(a+1,c;x),$$

$${}_1F_1(a,c;x) = {}_1F_1(a+1,c;x) - \frac{x}{c}\,{}_1F_1(a+1,c+1;x),$$

$${}_1F_1(a,c;x) = \frac{c-a}{c}\,{}_1F_1(a,c+1;x) + \frac{a}{c}\,{}_1F_1(a+1,c+1;x),$$

when $D = (V_1 - V_2)^{-1}$, we have

$$\begin{aligned}
\text{Risk }(\widehat{\delta}_A) = \text{Risk }(\widehat{\delta}_{GMM,1})
&+ e^{-\lambda/2}\,\frac{1}{k-2}\,{}_1F_1\Big(\frac{k}{2}-1\,,\,\frac{k}{2}\,;\,\lambda/2\Big)\Big[\frac{\lambda_D}{\lambda}\tau^2 - 2\tau\Big(\text{tr}(D(V_1-V_2)) - 2\frac{\lambda_D}{\lambda}\Big)\Big] \\
&- e^{-\lambda/2}\,\frac{1}{k-2}\,{}_1F_1\Big(\frac{k}{2}-1\,,\,\frac{k}{2}+1\,;\,\lambda/2\Big)\Big(\frac{\lambda_D}{\lambda} - \frac{\text{tr}(D(V_1-V_2))}{k}\Big)\Big[\tau^2+4\tau\Big] \\
&= \text{Risk }(\widehat{\delta}_{GMM,1}) + e^{-\lambda/2}\,\frac{1}{k-2}\,{}_1F_1\Big(\frac{k}{2}-1\,,\,\frac{k}{2}\,;\,\lambda/2\Big)\Big[\tau^2 - 2\tau(k-2)\Big], \quad \text{(C.22)}
\end{aligned}$$

where the last equality holds because the third term on the right hand side of the first equality is zero, $\text{tr}(D(V_1 - V_2)) = k$, and $\lambda_D = \lambda$.

■

Proof of Corollary 3.7:

The results hold by noting that if $x > 0$ and $a, c > 0$, then as $x \to \infty$,

$${}_1F_1(a,c;x) = \frac{\Gamma(c)}{\Gamma(a)}\,e^x\,x^{-(c-a)}\left[\sum_{j=0}^{p-1}\frac{(c-a)_j(1-a)_j}{j!}x^{-j} + O(|x|^{-p})\right].$$

See Lebedev (1972), pp. 271.

■

D

Appendix D

Theorem 4.6 : From equation (4.25), we have

$$\frac{1}{\sigma}(\hat{\beta}_{c,k} - \beta) = e_k^{(0)} + \sigma\left(e_k^{(1)} + e_c^{(1)}\right) + \sigma^2\left(e_k^{(2)} + e_c^{(2)}\right) + O(\sigma^3), \tag{D.1}$$

where $e_k^{(i)}, \quad i = 0, 1, 2$ are terms with order σ^i of $\frac{1}{\sigma}(\hat{\beta}(k) - \beta)$ and $e_c^{(i)}, \quad i = 0, 1, 2$ are the other terms with order σ^i which are defined below

$$e_k^{(0)} = QW'u_1,$$

$$e_k^{(1)} = Q(\Psi H_k^{(0)} u_1 - S_\Psi e_k^{(0)}),$$

$$e_k^{(2)} = Q\Psi' H_k^{(1)} u_1 + Q\delta u_1' H_k^{(0)} u_1 - Q(\delta u_1' W + W' u_1 \delta')e_k^{(0)}$$

$$+ QS_\Psi QS_\Psi e_k^{(0)} - Q\Psi' H_k^{(0)} \Psi e_k^{(0)} - QS_\Psi Q\Psi' H_k^{(0)} u_1,$$

$$e_c^{(1)} = \frac{\tau}{u_1' P_\Psi u_1} Q\Psi' M_x u_1,$$

$$e_c^{(2)} = \frac{\tau}{u_1' P_\Psi u_1}\Big[Q\delta u_1' M_x u_1 + 2u_1' M_x \Psi(\Psi' M_x \Psi)^{-1}\delta Q \Psi' M_x u_1$$
$$- Q\Psi' M_x \Psi e_k^{(0)} - Q S_\Psi Q \Psi' M_x u_1$$
$$+ \frac{2}{u_1' P_\Psi u_1}\Big[u_1' M_x \Psi e_k^{(0)} Q\Psi' M_x u_1 - u_1' M_x \Psi(\Psi' M_x \Psi)^{-1}\delta u_1' M_x u_1 Q\Psi' M_x u_1\Big]\Big],$$

where $P_\Psi = M_x\Psi(\Psi' M_x \Psi)^{-1}\Psi' M_x$, and if k is fixed $H_k = H_k^{(0)} = P_x$, so $H_k^{(1)} = 0$, and if $k = \lambda$, since λ is random, we have $H_k^{(0)} = I_T - \lambda_0 M_x$, and $H_k^{(1)} = -\lambda_1 M_x$, because by Kadane (1970)

$$\lambda = \frac{u_1' M_W u_1}{u_1' M_X u_1} + 2\sigma\frac{(u_1' W Q V_2' M_X u_1)(u_1' M_W u_1) - (u_1' W Q V_2' M_W u_1)(u_1' M_X u_1)}{(u_1' M_X u_1)^2} + O_p(\sigma^2)$$
$$\equiv \lambda_0 + \sigma\lambda_1 + O_p(\sigma^2)$$

where the definition of $\lambda_i, i = 0, 1$ should be apparent.

We derive the approximate expansions of the density function of $\hat{e}_{c,k}$ by inverting its characteristic function up to order σ^2. Using (D.1) the characteristic function of $\hat{e}_{c,k}$ can be expressed as

$$C_{c,k}(\theta) = C_k(\theta) + \sigma\,\mathbb{E}(i\theta'\,\mathbb{E}(e_c^{(1)}|e_k^{(0)})\exp(i\theta' e_k^{(0)})) + \sigma^2\,\mathbb{E}(i\theta'\,\mathbb{E}(e_c^{(2)}|e_k^{(0)})\exp(i\theta' e_k^{(0)}))$$
$$+ \frac{\sigma^2}{2}\,\mathbb{E}(i^2\theta'\,\mathbb{E}(e_c^{(1)} e_c^{(1)'}|e_k^{(0)})\theta\exp(i\theta' e_k^{(0)})) + \frac{\sigma^2}{2}\,\mathbb{E}(i^2\theta'\,\mathbb{E}(e_c^{(1)} e_k^{(1)'}|e_k^{(0)})\theta\exp(i\theta' e_k^{(0)}))$$
$$+ \frac{\sigma^2}{2}\,\mathbb{E}(i^2\theta'\,\mathbb{E}(e_k^{(1)} e_c^{(1)'}|e_k^{(0)})\theta\exp(i\theta' e_k^{(0)})) + O(\sigma^3), \tag{D.2}$$

where θ is a $N \times 1$ vector, $C_k(\theta)$ is the characteristic function of the k-class estimator, and $\mathbb{E}(.|e_k^{(0)})$ denotes the conditional expectation given $e_k^{(0)}$. The conditional expectations given the first-order term $e_k^{(0)}$ are calculated below.

$$\mathbb{E}(e_c^{(1)}|e_k^{(0)}) = 0, \tag{D.3}$$

as it is the product of odd numbers of normal distribution.

$$\mathbb{E}(e_c^{(2)}|e_k^{(0)}) = \frac{\tau(T-K)}{N}\Big[Q\delta - QC_2e_k^{(0)}\Big], \tag{D.4}$$

$$\mathbb{E}(e_c^{(1)}e_c^{(1)'}|e_k^{(0)}) = \quad = \frac{\tau^2(T-K)}{N(N-2)}QC_2Q, \tag{D.5}$$

$$\mathbb{E}(e_c^{(1)}e_k^{(1)'}|e_k^{(0)}) = \begin{cases} 0, & \text{if } k=1 \\ cQC_2Q, & \text{if } k=\lambda, \end{cases} \tag{D.6}$$

where $c \in \Big(-(T-2N)(T-K)/N(N-2), 0\Big)$.

Now, we invert the terms of the characteristic function of the Stein-like estimator in (D.2) term by term. The inverse transformation of the first term in (D.2) is

$$\mathfrak{F}^{-1}[C_k(\theta)] = f_k(\xi), \tag{D.7}$$

where $f_{2SLS}(\xi)$ is the approximate distribution of the k-class estimators given in Theorem 4.5. The inverse transformation of the rest of the terms in (D.2) are given below [1].

$$\mathfrak{F}^{-1}[(i\theta)'\,\mathbb{E}(\mathbb{E}(e_c^{(1)}|e_k^{(0)})\exp(i\theta' e_k^{(0)}))] = -\frac{\partial}{\partial\xi'}\{\mathbb{E}(e_c^{(1)}|e_k^{(0)}=\xi)\phi_Q(\xi)\} = 0, \tag{D.8}$$

$$\begin{aligned} &\mathfrak{F}^{-1}[(i\theta)'\,\mathbb{E}(\mathbb{E}(e_c^{(2)}|e_k^{(0)})\exp(i\theta' e_k^{(0)}))] = -\frac{\partial}{\partial\xi'}\{\mathbb{E}(e_c^{(2)}|e_k^{(0)}=\xi)\phi_Q(\xi)\} \\ &= \tau\phi_Q(\xi)\left[\alpha_1\delta'\xi + \alpha_1\Big[\operatorname{tr}(QC_2) - \xi'C_2\xi\Big]\right], \end{aligned} \tag{D.9}$$

[1] Note that, for any polynomial $g(.)$,

$\mathfrak{F}^{-1}[h(-i\theta)\,\mathbb{E}(g(x)\exp(i\theta' x))] = h(\frac{\partial}{\partial\xi})g(\xi)\phi_Q(\xi)$,

where $h(.)$ is any polynomial, and $\partial/\partial\xi' = (\partial/\partial\xi_1, ..., \partial/\partial\xi_N)$.

where $\alpha_1 = (T-K)/N$, and

$$\begin{aligned}\mathfrak{F}^{-1}[i^2\theta' \mathbb{E}(\mathbb{E}(e_c^{(1)} e_c^{(1)'} | e_k^{(0)})\theta \exp(i\theta' e_k^{(0)}))] &= \frac{\partial}{\partial \xi'}\{\mathbb{E}(e_c^{(1)} e_c^{(1)'} | e_k^{(0)} = \xi)\phi_Q(\xi)\}\frac{\partial}{\partial \xi} \\ &= \tau^2 \alpha_2 \Big[\xi' C_2 \xi - \mathrm{tr}(QC_2)\Big]\phi_Q(\xi)\end{aligned} \tag{D.10}$$

where $\alpha_2 = (T-K)/N(N-2)$. Furthermore, the inverse transformation of the last term is

$$\begin{aligned}\mathfrak{F}^{-1}[i^2\theta' \mathbb{E}(\mathbb{E}(e_c^{(1)} e_k^{(1)'} | e_k^{(0)})\theta \exp(i\theta' e_k^{(0)}))] &= \frac{\partial}{\partial \xi'}\{\mathbb{E}(e_c^{(1)} e_k^{(1)'} | e_k^{(0)} = \xi)\phi_Q(\xi)\}\frac{\partial}{\partial \xi} \\ &= \begin{cases} 0, & \text{if } k = 1 \\ \tau c \Big[\xi' C_2 \xi - \mathrm{tr}(QC_2)\Big], & \text{if } k = \lambda. \end{cases}\end{aligned} \tag{D.11}$$

Summation of the terms in equations (D.8)–(D.11) will provide the results in the theorem.

■

Theorem 4.7 :

Using (4.30), the approximate bias of the Stein-like estimator up to order of interest is equal to

$$\mathbb{E}(\frac{1}{\sigma}(\hat{\beta}_{c,k} - \beta)) = \mathbb{E}(\frac{1}{\sigma}(\hat{\beta}(k) - \beta)) + O(\sigma^2) = 0, \tag{D.12}$$

where the last equality holds by Theorem 4.3. The approximate MSEM of the Stein-like estimator up to the order of interest is

$$\begin{aligned}&\mathbb{E}\left(\frac{1}{\sigma^2}(\hat{\beta}_{c,1} - \beta)(\hat{\beta}_{c,1} - \beta)'\right) = \mathbb{E}\left(\frac{1}{\sigma^2}(\hat{\beta}(1) - \beta)(\hat{\beta}(1) - \beta)'\right) + \tau\sigma^2\alpha_1 \int \xi\xi'\delta'\xi\phi_Q(\xi)d\xi \\ &+ \frac{1}{2}\tau\sigma^2\Big[\tau\alpha_2 - 2\alpha_1\Big]\int \Big(\xi\xi' C_2 \xi\xi' - \mathrm{tr}(QC_2)\xi\xi'\Big)\phi_Q(\xi)d\xi \\ &= \mathbb{E}\left(\frac{1}{\sigma^2}(\hat{\beta}(1) - \beta)(\hat{\beta}(1) - \beta)'\right) + \tau\sigma^2\Big[\tau\alpha_2 - 2\alpha_1\Big]QC_2Q,\end{aligned} \tag{D.13}$$

similarly,

$$\mathbb{E}\left(\frac{1}{\sigma^2}(\hat{\beta}_{c,\lambda}-\beta)(\hat{\beta}_{c,\lambda}-\beta)'\right) \leq \mathbb{E}\left(\frac{1}{\sigma^2}(\hat{\beta}(\lambda)-\beta)(\hat{\beta}(\lambda)-\beta)'\right)+\tau\sigma^2\Big[\tau\alpha_2-2\alpha_1\Big]QC_2Q. \tag{D.14}$$

■

Theorem 4.9 :

To derive

$$\int_{||Q^{-1/2}\xi||<z}\cdots\int (f_{c,k}(\xi)-f_k(Q^{1/2}\xi))\,d\xi,$$

we take the integral of each term of the difference of the approximate distributions below.

$$\int_{||\zeta||<z}\cdots\int \tau\alpha_1\delta' Q^{1/2}\zeta\phi_I(\zeta)d\zeta=0, \tag{D.15}$$

$$\int_{||\zeta||<z}\cdots\int \frac{\tau}{2}\Big[\tau\alpha_2-2\alpha_1\Big]\,\mathrm{tr}(QC_2)\phi_I(\zeta)d\zeta=\frac{\tau}{2}\Big[\tau\alpha_2-2\alpha_1\Big]\,\mathrm{tr}(QC_2)\Big[\Phi(z)-\Phi(-z)\Big]^N, \tag{D.16}$$

$$\begin{aligned}&\int_{||\zeta||<z}\cdots\int \frac{\tau}{2}\Big[\tau\alpha_2-2\alpha_1\Big]\zeta' Q^{1/2}C_2Q^{1/2}\zeta\phi_I(\zeta)d\zeta\\ &\qquad=\frac{\tau}{2}\Big[\tau\alpha_2-2\alpha_1\Big]\,\mathrm{tr}(QC_2)\Big\{-2z\phi(z)\Big[\Phi(z)-\Phi(-z)\Big]^{N-1}+\Big[\Phi(z)-\Phi(-z)\Big]^N\Big\},\end{aligned} \tag{D.17}$$

where the last equality holds by using

$$\int_{|x|<z} x^2\phi(x)dx=-2z\phi(z)+\Phi(z)-\Phi(-z).$$

The results follow by adding the right-hand side of equations (D.15)–(D.17).

■

E

Appendix E

Lemma E.1 *Suppose that assumptions 5.1(i)-(ii) hold. Then,*

$$\sqrt{T}(\dot{\beta}_i - \beta_i^0) = O_p(1), \tag{E.1}$$

for each $i = 1, \ldots, N$. *Equivalently,* $\dot{\beta}_i - \beta_i^0 = \frac{1}{\sqrt{T}}\dot{\upsilon}_i = O_p(T^{-1/2})$. *Therefore, for any* i *and* j *in* $\{1, \ldots, N\}$,

$$\dot{w}_{ij}^{-\frac{1}{\kappa}} = \begin{cases} \frac{1}{\sqrt{T}}\|\dot{\upsilon}_i - \dot{\upsilon}_j\| = O_p(T^{-1/2}), & \text{if } i\&j \in G_k^0 \\ \|\beta_i^0 - \beta_j^0\| + \frac{1}{\sqrt{T}}\|\dot{\upsilon}_i - \dot{\upsilon}_j\| - c, & \text{if } i \in G_k^0 \,\&\, j \in G_l^0, \end{cases} \tag{E.2}$$

for any $k, l \in \{1, \ldots, K_0\}$ *and* $l \neq k$, *and* $c \geq 0$.

■

Proof of Theorem 5.4 :

i) Let $Q_{1,NT,i}(\beta_i) = \frac{1}{T}\sum_{t=1}^{T}(\tilde{y}_{it} - \beta_i'\tilde{x}_{it})^2$ and $Q_{1,NT,i}(\boldsymbol{\beta}, \lambda_1) = Q_{1,NT,i}(\beta_i) + \frac{\lambda_1}{2N}\sum_{j=1}^{N}\dot{w}_{ij}\|\beta_i - \beta_j\|$. Define $b_i = \beta_i - \beta_i^0$ and $\hat{b}_i = \hat{\beta}_i - \beta_i^0$. We have

$$Q_{1,NT,i}(\beta_i) - Q_{1,NT,i}(\beta_i^0) = \frac{1}{T}\sum_{t=1}^{T}(\tilde{u}_{it} - b_i'\tilde{x}_{it})^2 - \frac{1}{T}\sum_{t=1}^{T}\tilde{u}_{it}^2 = b_i'\hat{Q}_{i,\tilde{x}\tilde{x}}b_i - 2b_i'\hat{Q}_{i,\tilde{x}\tilde{u}}. \tag{E.3}$$

Given the fact that $Q_{1,NT,i}(\hat{\beta}_i, \lambda_1) - Q_{1,NT,i}(\beta_i^0, \lambda_1) \leq 0$, for any $k \in \{1, \ldots, K_0\}$, and any $i \in G_k^0$, we have

$$\begin{aligned}
0 &\geq Q_{1,NT,i}(\hat{\beta}_i, \lambda_1) - Q_{1,NT,i}(\beta_i^0, \lambda_1) \\
&= \hat{b}_i'\hat{Q}_{i,\tilde{x}\tilde{x}}\hat{b}_i - 2\hat{b}_i'\hat{Q}_{i,\tilde{x}\tilde{u}} + \frac{\lambda_1}{2N}\sum_{j=1}^{N}\dot{w}_{ij}\Big[\|\hat{\beta}_i - \hat{\beta}_j\| - \|\beta_i^0 - \beta_j^0\|\Big] \\
&\geq \hat{b}_i'\hat{Q}_{i,\tilde{x}\tilde{x}}\hat{b}_i - 2\hat{b}_i'\hat{Q}_{i,\tilde{x}\tilde{u}} - \frac{\lambda_1}{2N}\sum_{j=1}^{N}\dot{w}_{ij}\Big[\|(\hat{\beta}_i - \beta_i^0) - (\hat{\beta}_j - \beta_j^0)\|\Big] \\
&\geq \hat{b}_i'\hat{Q}_{i,\tilde{x}\tilde{x}}\hat{b}_i - 2\hat{b}_i'\hat{Q}_{i,\tilde{x}\tilde{u}} - \frac{\lambda_1}{2N}\sum_{j=1}^{N}\dot{w}_{ij}\Big[\|\hat{\beta}_i - \beta_i^0\| + \|\hat{\beta}_j - \beta_j^0\|\Big]
\end{aligned} \tag{E.4}$$

where the second and third inequalities hold by the triangle inequality. Note that by Lemma E.1, $\max_{i\in G_k^0, j\notin G_k^0}\dot{w}_{ij} = O_p(J_{min}^{-\kappa})$. Averaging the above term over all of the individuals, and employing assumptions 5.1(i)-(ii), we have

$$\begin{aligned}
0 &\geq \frac{1}{N}\sum_{i=1}^{N}\Big(\hat{b}_i'\hat{Q}_{i,\tilde{x}\tilde{x}}\hat{b}_i\Big) - 2\frac{1}{N}\sum_{i=1}^{N}\Big(\hat{b}_i'\hat{Q}_{i,\tilde{x}\tilde{u}}\Big) - \frac{\lambda_1}{2N}\sum_{i=1}^{N}\sum_{j=1}^{N}\dot{w}_{ij}\Big[\|\hat{\beta}_i - \beta_i^0\| + \|\hat{\beta}_j - \beta_j^0\|\Big] \\
&\geq \frac{1}{N}\sum_{i=1}^{N}\Big(\hat{b}_i'\hat{Q}_{i,\tilde{x}\tilde{x}}\hat{b}_i\Big) - 2\frac{1}{N}\sum_{i=1}^{N}\Big(\hat{b}_i'\hat{Q}_{i,\tilde{x}\tilde{u}}\Big) - \lambda_1\big(\max_{i\in G_k^0, j\notin G_k^0}\dot{w}_{ij}\big)\sum_{i=1}^{N}\|\hat{b}_i\| \\
&\geq \frac{1}{NT}\sum_{i=1}^{N}\Big[T\|\hat{b}_i\|^2 c_{\tilde{x}\tilde{x}} - 2T\|\hat{b}_i\|\|\hat{Q}_{i,\tilde{x}\tilde{u}}\| - O_p(\sqrt{T}\lambda_1 J_{min}^{-\kappa})\sqrt{T}\|\hat{b}_i\|\Big].
\end{aligned} \tag{E.5}$$

Under assumption 5.2(ii), $N\sqrt{T}\lambda_1 J_{min}^{-\kappa} = O_p(1)$, thus the above equation implies that $\sqrt{T}\|\hat{b}_i\| = O_p(1)$, because otherwise, the above term cannot be negative. Therefore, $\hat{\beta}_i - \beta_i^0 = O_p(T^{-1/2})$ for $i = 1, \ldots, N$.

ii) Let $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}^0 + T^{-1/2}\boldsymbol{\psi}$, where the $p \times N$ matrix $\boldsymbol{\psi} = (\psi_1', \ldots, \psi_N')'$. Similar to part (i), we have

$$
\begin{aligned}
0 \geq\ & T\Big[Q_{1,NT}(\hat{\boldsymbol{\beta}}, \lambda_1) - Q_{1,NT}(\boldsymbol{\beta}^0, \lambda_1)\Big] \\
\geq\ & \frac{1}{N}\sum_{i=1}^{N} \psi_i' \hat{Q}_{i,\tilde{x}\tilde{x}} \psi_i - 2\frac{\sqrt{T}}{N}\sum_{i=1}^{N} \psi_i' \hat{Q}_{i,\tilde{x}\tilde{u}} - \sqrt{T}\lambda_1 (\max_{i\in G_k^0, j\in G_l^0, l\neq k} \dot{w}_{ij}) \sum_{i=1}^{N} \|\psi_i\| \\
\geq\ & \underline{c}_{\tilde{x}\tilde{x},NT} \frac{1}{N}\sum_{i=1}^{N} \|\psi_i\|^2 - 2\Big[\frac{1}{N}\sum_{i=1}^{N} \|\psi_i\|^2\Big]^{1/2} \Big[\frac{T}{N}\sum_{i=1}^{N} \|\hat{Q}_{i,\tilde{x}\tilde{u}}\|^2\Big]^{1/2} \\
& - 2O_p(\sqrt{NT}\lambda_1 J_{min}^{-\kappa})\Big[\frac{1}{N}\sum_{i=1}^{N} \|\psi_i\|^2\Big]^{1/2},
\end{aligned}
\tag{E.6}
$$

By assumption 5.1(ii), $\underline{c}_{\tilde{x}\tilde{x},NT}$ is bounded below by $\underline{c}_{\tilde{x}\tilde{x}} > 0$, by assumption 5.1(iii) $\frac{T}{N}\sum_{i=1}^{N} \|\hat{Q}_{i,\tilde{x}\tilde{u}}\|^2 = O_p(1)$, and by assumption 5.2(ii), $\sqrt{NT}\lambda_1 J_{min}^{-\kappa} = o_p(1)$. Thus, if $\frac{1}{N}\sum_{i=1}^{N} \|\psi_i\|^2 = L$, for sufficiently large values of L, the first term dominates the second term in (E.6). In other words, for sufficiently large L, $0 \leq T\Big[Q_{1,NT}(\hat{\boldsymbol{\beta}}, \lambda_1) - Q_{1,NT}(\boldsymbol{\beta}^0, \lambda_1)\Big]$, and $Q_{1,NT}(\boldsymbol{\beta}^0, \lambda_1)$ cannot be minimized. Therefore, we must have $\frac{1}{N}\sum_{i=1}^{N} \|\hat{b}_i\|^2 = O_p(T^{-1})$.

■

Proof of Theorem 5.5 :

Take $k \in \{1, \ldots, K_0\}$. By the consistency results in theorem 5.4, we have $\hat{\beta}_i - \hat{\beta}_j \xrightarrow{p} \beta_i^0 - \beta_j^0 \neq 0$ for all $i \in G_k^0$ and $j \notin G_k^0$. Thus, $\|\hat{\beta}_i - \hat{\beta}_j\| \neq 0$ for all $i \in G_k^0$ and $j \notin G_k^0$.

By contrary, suppose that there exist $i \in G_k^0$ such that $\|\hat{\theta}_{ij}\| \equiv \|\hat{\beta}_i - \hat{\beta}_j\| \neq 0$, for any $j \in G_k^0$. There exists $r \in \{1, \ldots, p\}$ such that for the rth element of $\hat{\theta}_{ij}$, we have $|\hat{\theta}_{ij,r}| = \max\{|\hat{\theta}_{ij,l}| : l = 1 \ldots, p\}$. Without lose of generality assume that $r = p$, then it can be easily verified that $|\hat{\theta}_{ij,p}|/\|\hat{\theta}_{ij}\| \geq 1/\sqrt{p}$. Then, the first order condition (FOC) of the objective function in 5.4 with respect to the pth element of β_i, denoted by $\beta_{i,p}$ is

$$\begin{aligned}
\mathbf{0} &= \frac{-2}{\sqrt{T}} \sum_{t=1}^{T} \tilde{x}_{it,p}(\tilde{y}_{it} - \tilde{x}'_{it}\hat{\beta}_i) + \frac{\lambda_1 \sqrt{T}}{N} \sum_{j=1}^{N} \dot{w}_{ij} e_{ij,p} \\
&= \frac{-2}{\sqrt{T}} \sum_{t=1}^{T} \tilde{x}_{it,p}\tilde{u}_{it} + \frac{2}{\sqrt{T}} \sum_{t=1}^{T} \tilde{x}_{it,p}\tilde{x}'_{it}(\hat{\beta}_i - \beta_i^0) + \frac{\lambda_1 \sqrt{T}}{N} \sum_{j=1}^{N} \dot{w}_{ij} e_{ij,p} \\
&= \frac{-2}{\sqrt{T}} \sum_{t=1}^{T} \tilde{x}_{it,p}\tilde{u}_{it} + \frac{1}{T} \sum_{t=1}^{T} \tilde{x}_{it,p}\tilde{x}'_{it}\sqrt{T}(\hat{\beta}_i - \beta_i^0) + \frac{\sqrt{T}\lambda_1}{N} \sum_{j \in G_k^0} \dot{w}_{ij} \frac{\hat{\theta}_{ij,p}}{\|\hat{\theta}_{ij}\|} + \frac{\sqrt{T}\lambda_1}{N} \sum_{j \notin G_k^0} \dot{w}_{ij} \frac{\hat{\theta}_{ij,p}}{\|\hat{\theta}_{ij}\|} \\
&\equiv \hat{B}_{i1} + \hat{B}_{i2} + \hat{B}_{i3} + \hat{B}_{i4}.
\end{aligned} \tag{E.7}$$

where $e_{ij} = (\hat{\beta}_i - \hat{\beta}_j)/\|\hat{\beta}_i - \hat{\beta}_j\|$ if $\|\hat{\beta}_i - \hat{\beta}_j\| \neq 0$, and $\|e_{ij}\| \leq 1$ otherwise. By assumption 5.1(i), $\hat{B}_{i1} = O_p(1)$, by theorem 5.4 and assumption 5.1(ii) $\hat{B}_{i2} = O_p(1)$. By Assumption 5.2(ii), theorem 5.4, and given $\max_{i \in G_k^0, j \notin G_k^0} \dot{w}_{ij} = O_p(J_{min}^{-\kappa})$, $|\hat{B}_{i4}| \leq N\sqrt{T}\lambda_1 \max_{i \in G_k^0, j \notin G_k^0} \dot{w}_{ij} \sum_{j \notin G_k^0} = O_p(N\sqrt{T}\lambda_1 J_{min}^{-\kappa}) = O_p(1)$. In view of the fact that for $i\&j \in G_k^0$, $\dot{w}_{ij} = O_p(T^{-\kappa/2})$, $2\sqrt{p}|\hat{B}_{i3}| \geq \sqrt{T}\lambda_1 \sum_{j \in G_k^0} \dot{w}_{ij} = O_p(\tau_k \sqrt{T}\lambda_1 T^{\kappa/2})$, which is explosive in probability under assumption 5.2(iii). Consequently, $|\hat{B}_{i3}| \gg |\hat{B}_{i1} + \hat{B}_{i2} + \hat{B}_{i4}|$, such that (E.7) cannot hold for sufficiently large (N, T). Thus, we conclude that with probability approaching one, $\hat{\theta}_{ij}$ for all $i\&j \in G_k^0$ must be in a position where $\|\hat{\theta}_{ij}\|$ is not differentiable and $\sqrt{T}\lambda_1 \dot{w}_{ij} e_{ij} = O_p(1)$ in order for the FOC to hold. Furthermore, this implies that for all $i\&j \in G_k^0$, $P(\|\hat{\theta}_{ij}\| = 0) \to 1$ as $(N, T) \to \infty$.

■

Proof of Corollary 5.6 :

For any $l\&k \in \{1,\ldots,K_0\}$, where $l \neq k$, and any $i\&j \in \{1,\ldots,N\}$, we consider two cases: (i) $i\&j \in G_k^0$, and (ii) $i \in G_k^0$ and $j \in G_l^0$. In case (i), theorem 5.5 implies that asymptotically $\|\hat{\beta}_i - \hat{\beta}_j\| = 0$, thus $\hat{K} \leq K$. Under case (ii), we want to show that $\|\hat{\beta}_i - \hat{\beta}_j\| \neq 0$. Note that for any such i and j, $\|\beta_i^0 - \beta_j^0\| \geq J_{min}$. Now, suppose by controversy that there exist $i \in G_k^0$ and $j \in G_l^0$ that $\|\hat{\beta}_i - \hat{\beta}_j\| = 0$, besides by the consistency results in Theorem 5.4, we have $\hat{\beta}_i - \hat{\beta}_j = \beta_i^0 - \beta_j^0 + O_p(T^{-1/2})$, hence $\|\beta_i^0 - \beta_j^0\| = O_p(T^{-1/2})$. But, this contradicts assumption 5.2(i) that $T^{1/2} J_{min} \to \infty$ as $\|\beta_i^0 - \beta_j^0\| \geq J_{min}$

■

Proof of Theorem 5.7 :

Following Su et al. (2016) and Bertsekas (1995) Appendix B.5, we study the oracle property by utilizing conditions from sub-differential calculus. From the FOC of the objective function in (5.4) with respect to β_i evaluated at $\hat{\beta}_i$, for each $i \in \{1,\ldots,N\}$, we have

$$\mathbf{0} = \frac{-2}{T} \sum_{t=1}^{T} \tilde{x}_{it}(\tilde{y}_{it} - \hat{\beta}_i' \tilde{x}_{it}) + \frac{\lambda_1}{N} \sum_{j=1}^{N} \dot{w}_{ij} \hat{e}_{ij}, \tag{E.8}$$

where $\hat{e}_{ij} = \frac{\hat{\beta}_i - \hat{\beta}_j}{\|\hat{\beta}_i - \hat{\beta}_j\|}$ if $\|\hat{\beta}_i - \hat{\beta}_j\| \neq 0$ and $\|\hat{e}_{ij}\| \leq 1$ otherwise, and let $\hat{e}_{ii} = 0$.

Let $i \in \hat{G}_k$, for a $k \in \{1, \ldots, \hat{K}\}$. Note that by definition $\hat{\beta}_j = \hat{\alpha}_k$ for any $j \in \hat{G}_k$. Summing the FOC over the individual units in $\hat{G}_k$, we have

$$\begin{aligned} \mathbf{0} &= \frac{-2}{T} \sum_{i \in \hat{G}_k} \sum_{t=1}^{T} \tilde{x}_{it}(\tilde{y}_{it} - \hat{\beta}_i' \tilde{x}_{it}) + \frac{\lambda_1}{N} \sum_{i \in \hat{G}_k} \sum_{j=1}^{N} \dot{w}_{ij} \hat{e}_{ij} \\ &= \frac{-2}{T} \sum_{i \in \hat{G}_k} \sum_{t=1}^{T} \tilde{x}_{it}(\tilde{y}_{it} - \hat{\alpha}_k' \tilde{x}_{it}) + \frac{\lambda_1}{N} \sum_{i \in \hat{G}_k} \sum_{j \notin \hat{G}_k} \dot{w}_{ij} \hat{e}_{ij}. \end{aligned} \tag{E.9}$$

For the second term on the right hand side of the above equation, we have

$$\sqrt{N_k T} \| \frac{\lambda_1}{N N_k} \sum_{i \in \hat{G}_k} \sum_{j \notin \hat{G}_k} \dot{w}_{ij} \hat{e}_{ij} \| \leq \frac{\sqrt{T} \lambda_1}{N \sqrt{N_k}} \sum_{i \in \hat{G}_k} \sum_{j \notin \hat{G}_k} \| \dot{w}_{ij} \hat{e}_{ij} \| \leq \sqrt{N_k T} \lambda_1 \max_{i \in \hat{G}_k, j \notin \hat{G}_k} \dot{w}_{ij}, \tag{E.10}$$

which is of order $O_p(\sqrt{N_k T} \lambda_1 J_{min}^{-\kappa}) = o_p(1)$, under assumption 5.2(ii).

Thus, it follows that

$$\hat{\alpha}_k = \Big(\frac{1}{N_k T} \sum_{i \in \hat{G}_k} \sum_{t=1}^{T} \tilde{x}_{it} \tilde{x}_{it}' \Big)^{-1} \frac{1}{N_k T} \sum_{i \in \hat{G}_k} \sum_{t=1}^{T} \tilde{x}_{it} \tilde{y}_{it}. \tag{E.11}$$

By corollary 5.6, with probability approaching one $\hat{\alpha}_k = \bar{\alpha}_k$, and the limiting distribution follows.

■

Proof of Theorem 5.8 :

By corollary 5.6, it follows that $\hat{\alpha}_{\hat{G}_k} = \bar{\alpha}_k$, and the limiting distribution follows.

■

F

Appendix F

Let us define some notations which will be used in the proof of the results below. Define $V_{i,NT}(\beta_i) = \left[\frac{1}{T}\sum_{t=1}^{T}\rho(\xi_{it},\beta_i)\right]' W_{i,NT}\left[\frac{1}{T}\sum_{t=1}^{T}\rho(\xi_{it},\beta_i)\right]$, and $\bar{V}_i(\beta_i) = \left[\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}(\rho(\xi_{it},\beta_i))\right]' W_{i,NT}\left[\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}(\rho(\xi_{it},\beta_i))\right]$, and let $R_{i,T}(\beta_i)\left[\frac{1}{T}\sum_{t=1}^{T}\left[\rho(\xi_i,\beta_i) - \mathbb{E}(\rho(\xi_i,\beta_i))\right]\right]' W_{i,N}\left[\frac{1}{T}\sum_{t=1}^{T}\left[\rho(\xi_i,\beta_i) - \mathbb{E}(\rho(\xi_i,\beta_i))\right]\right]$.

Lemma F.1 *Suppose that assumptions 5.9(i)-(ii) hold. Then,*

$$\sqrt{T}(\ddot{\beta}_i - \beta_i^0) = O_p(1), \tag{F.1}$$

for each $i = 1,\ldots,N$. Equivalently, $\ddot{\beta}_i - \beta_i^0 = \frac{1}{\sqrt{T}}\ddot{v}_i = O_p(T^{-1/2})$. Therefore, for any i and j in $\{1,\ldots,N\}$,

$$\ddot{w}_{ij}^{-\frac{1}{\kappa}} = \begin{cases} \frac{1}{\sqrt{T}}\|\ddot{v}_i - \ddot{v}_j\| = O_p(T^{-1/2}), & \text{if } i\&j \in G_k^0 \\ \|\beta_i^0 - \beta_j^0\| + \frac{1}{\sqrt{T}}\|\ddot{v}_i - \ddot{v}_j\| - \ddot{c}, & \text{if } i \in G_k^0 \,\&\, j \in G_l^0, \end{cases} \tag{F.2}$$

for any $k,l \in \{1,\ldots,K_0\}$ and $l \neq k$, and $\ddot{c} \geq 0$.

■

Lemma F.2 *Suppose that assumptions 5.9(iv) holds. Then, for all $\beta_i \in \mathcal{B}_i$ with probability approaching one,*

$$\underline{c}\Big[\frac{1}{2}\bar{V}_i(\beta_i) - R_{i,T}(\beta_i)\Big] \leq V_{i,NT}(\beta_i) \leq \bar{c}[2\bar{V}_i(\beta_i) + 2R_{i,T}(\beta_i)\Big], \tag{F.3}$$

where $\underline{c}$ and $\bar{c}$ are some generic positive constants that do not depend on i with $0 < \underline{c} < 1 < \bar{c} < \infty$.

Proof: See Su et al. (2016). ■

Proof of Theorem 5.13 :

(i) Note that $Q_{2,NT,i}(\tilde{\boldsymbol{\beta}}, \lambda_2) - Q_{1,NT,i}(\boldsymbol{\beta}^0, \lambda_2) \leq 0$, we find an lower bound for this difference term. Assume individual $i \in G_k^0$ for any $k = 1, \ldots, K_0$, then we have

$$\begin{aligned}
0 &\geq Q_{2,NT,i}(\tilde{\boldsymbol{\beta}}, \lambda_2) - Q_{2,NT,i}(\boldsymbol{\beta}^0, \lambda_2) \\
&= V_{i,NT}(\tilde{\beta}_i) - V_{i,NT}(\beta_i^0) + \frac{\lambda_2}{2N}\sum_{j=1}^{N} \ddot{w}_{ij}\Big[\|\tilde{\beta}_i - \tilde{\beta}_j\| - \|\beta_i^0 - \beta_j^0\|\Big] \\
&\geq \underline{c}\Big[\frac{1}{2}\bar{V}_i(\tilde{\beta}_i) - \tilde{R}_{i,T}\Big] - 2\bar{c}R_{i,T}^0 + \frac{\lambda_2}{2N}\sum_{j\notin\tilde{G}_k} \ddot{w}_{ij}\Big[\|\tilde{\beta}_i - \tilde{\beta}_j\| - \|\beta_i^0 - \beta_j^0\|\Big]
\end{aligned} \tag{F.4}$$

where the first inequality holds by Lemma B.1 of Su et al. (2016). By averaging the above term over all of the individuals, and employing assumptions 5.9(i)-(ii), we have

$$\begin{aligned}
0 &\geq \frac{\underline{c}}{N}\sum_{i=1}^{N}\Big[\frac{1}{2}\bar{V}_i(\tilde{\beta}_i) - \tilde{R}_{i,T}\Big] - 2\bar{c}\sum_{i=1}^{N}\bar{c}R_{i,T}^0 + \frac{\lambda_2}{2N}\sum_{i=1}^{N}\sum_{j\notin\tilde{G}_k}\ddot{w}_{ij}\Big[\|\tilde{\beta}_i - \tilde{\beta}_j\| - \|\beta_i^0 - \beta_j^0\|\Big] \\
&\geq \frac{\underline{c}\underline{c}_{1,NT}}{N}\sum_{i=1}^{N}\|\tilde{\beta}_i - \beta_i^0\|^2 - \frac{1}{N}\sum_{i=1}^{N}\Big[\underline{c}\tilde{R}_{i,T} + 2\bar{c}R_{i,T}^0\Big] - \frac{\lambda_2}{N}\max_{1\leq i\&j\leq N}\ddot{w}_{ij}\sum_{i=1}^{N}\|\tilde{\beta}_i - \beta_i^0\|
\end{aligned} \tag{F.5}$$

$$\geq \frac{1}{NT}\sum_{i=1}^{N}\Big[\underline{c}\underline{c}_{1,NT}T\|\tilde{\beta}_i-\beta_i^0\|^2-\underline{c}\bar{\rho}_{i,T}(\tilde{\beta})-2\bar{c}\bar{\rho}_{i,T}(\beta^0)$$

$$-2O_p(N\sqrt{T}\lambda_2 J_{min}^{-\kappa})\sqrt{T}\|\tilde{\beta}_i-\beta_i^0\|\Big],$$

where the use has been made of

$$\max_{1\leq i\leq N}\bar{V}_i(\tilde{\boldsymbol{\beta}})=\max_{1\leq i\leq N}(\tilde{\beta}_i-\beta_i^0)'\bar{Q}_{i,z\Delta x}'W_i\bar{Q}_{i,z\Delta x}(\tilde{\beta}_i-\beta_i^0)\geq \underline{c}_{1,NT}\max_{1\leq i\leq N}\|\tilde{\beta}_i-\beta_i^0\|^2,$$

and $\underline{c}_{1,NT}=\min_{1\leq i\leq N}\mu_{min}(\bar{Q}_{i,z\Delta x}'W_i\bar{Q}_{i,z\Delta x})\geq \underline{c}_W\underline{c}_{\bar{Q}}>0$. The result above implies that $\sqrt{T}\|\tilde{\beta}_i-\beta_i^0\|=O_p(1)$, because otherwise, the above term cannot be negative. Therefore, $\tilde{\beta}_i-\beta_i^0=O_p(T^{-1/2})$ for $i=1,\ldots,N$.

(ii) From the last line in (F.5), let $c^*=O_p(1)$, then we have

$$\begin{aligned}\frac{\underline{c}\underline{c}_{1,NT}}{N}\sum_{i=1}^{N}\|\tilde{\beta}_i-\beta_i^0\|^2 \leq \underline{c}\frac{1}{NT}\sum_{i=1}^{N}\bar{\rho}_{i,T}(\tilde{\beta})+\frac{2}{NT}\bar{c}\sum_{i=1}^{N}\bar{\rho}_{i,T}(\beta^0)\\ +2c^*\frac{1}{N\sqrt{T}}\sum_{i=1}^{N}\|\tilde{\beta}_i-\beta_i^0\|=O_p(\frac{1}{T}),\end{aligned} \tag{F.6}$$

which proves the result.

■

Proof of Theorem 5.14 :

Take $k\in\{1,\ldots,K_0\}$. By the consistency results in theorem 5.13, we have $\tilde{\beta}_i-\tilde{\beta}_j\xrightarrow{p}\beta_i^0-\beta_j^0\neq 0$ for all $i\in G_k^0$ and $j\notin G_k^0$. Thus, $\|\tilde{\beta}_i-\tilde{\beta}_j\|\neq 0$ for all $i\in G_k^0$ and $j\notin G_k^0$.

By contrary, suppose that there exist $i\in G_k^0$ such that $\|\tilde{\theta}_{ij}\|\equiv\|\tilde{\beta}_i-\tilde{\beta}_j\|\neq 0$, for any $j\in G_k^0$. There exists $r\in\{1,\ldots,p\}$ such that for the rth element of $\tilde{\theta}_{ij}$, we have

$|\tilde{\theta}_{ij,r}| = \max\{|\tilde{\theta}_{ij,l}| : l = 1\ldots, p\}$. Without lose of generality assume that $r = p$, then it can be easily verified that $|\tilde{\theta}_{ij,p}|/\|\tilde{\theta}_{ij}\| \geq 1/\sqrt{p}$. Then, the first order condition (FOC) of the objective function in (5.4) with respect to the pth element of β_i, denoted by $\beta_{i,p}$ is

$$\begin{aligned}
\mathbf{0} &= \frac{-2}{\sqrt{T}}\tilde{Q}'_{i,z\Delta x,p}W_{i,NT}\sum_{t=1}^{T}\tilde{z}_{it}(\Delta y_{it} - \tilde{\beta}_i'\Delta x_{it}) + \frac{\lambda_2\sqrt{T}}{N}\sum_{j=1}^{N}\ddot{w}_{ij}e_{ij,p} \\
&= -2\tilde{Q}'_{i,z\Delta x}W_{i,NT}\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\tilde{z}_{it}\Delta u_{it} + 2\tilde{Q}'_{i,z\Delta x,p}W_{i,NT}\tilde{Q}_{i,z\Delta x}\sqrt{T}(\tilde{\beta}_i - \tilde{\beta}_j) \\
&+ \frac{\sqrt{T}\lambda_2}{N}\sum_{j\in G_k^0}\ddot{w}_{ij}\frac{\tilde{\theta}_{ij,p}}{\|\tilde{\theta}_{ij}\|} + \frac{\sqrt{T}\lambda_2}{N}\sum_{j\notin G_k^0}\ddot{w}_{ij}\frac{\tilde{\theta}_{ij,p}}{\|\tilde{\theta}_{ij}\|} \\
&\equiv -\tilde{B}_{i1} + \tilde{B}_{i2} + \tilde{B}_{i3} + \tilde{B}_{i4}.
\end{aligned} \tag{F.7}$$

where $e_{ij} = (\tilde{\beta}_i - \tilde{\beta}_j)/\|\tilde{\beta}_i - \tilde{\beta}_j\|$ if $\|\tilde{\beta}_i - \tilde{\beta}_j\| \neq 0$, and $\|e_{ij}\| \leq 1$ otherwise. By assumption 5.9(i), $\tilde{B}_{i1} = O_p(1)$, by Theorem 5.13 and assumption 5.9(ii) $\tilde{B}_{i2} = O_p(1)$. By Assumption 5.10(ii), Theorem 5.13, and given $\max_{i\in G_k^0, j\notin G_k^0}\dot{w}_{ij} = O_p(J_{min}^{-\kappa})$, $|\tilde{B}_{i4}| \leq N\sqrt{T}\lambda_2 \max_{i\in G_k^0, j\notin G_k^0}\ddot{w}_{ij}\sum_{j\notin G_k^0} = O_p(N\sqrt{T}\lambda_2 J_{min}^{-\kappa}) = O_p(1)$. In view of the fact that for $i\&j \in G_k^0$, $\ddot{w}_{ij} = O_p(T^{-\kappa/2})$, $2\sqrt{p}|\tilde{B}_{i3}| \geq \sqrt{T}\lambda_2\sum_{j\in G_k^0}\ddot{w}_{ij} = O_p(\tau_k\sqrt{T}\lambda_2 T^{\kappa/2})$, which is explosive in probability under assumption 5.10(iii). Consequently, $|\tilde{B}_{i3}| \gg |\tilde{B}_{i1} + \tilde{B}_{i2} + \tilde{B}_{i4}|$, such that (F.7) cannot hold for sufficiently large (N, T). Thus, we conclude that with probability approaching one, $\tilde{\theta}_{ij}$ for all $i\&j \in G_k^0$ must be in a position where $\|\tilde{\theta}_{ij}\|$ is not differentiable and $\sqrt{T}\lambda_2\ddot{w}_{ij}e_{ij} = O_p(1)$ in order for the FOC to hold. Furthermore, this implies that for all $i\&j \in G_k^0$, $P(\|\tilde{\theta}_{ij}\| = 0) \to 1$ as $(N, T) \to \infty$.

■

Proof of Corollary 5.15 :

For any $l\&k \in \{1,\ldots,K_0\}$, where $l \neq k$, and any $i\&j \in \{1,\ldots,N\}$, we consider two cases: (i) $i\&j \in G_k^0$, and (ii) $i \in G_k^0$ and $j \in G_l^0$. In case (i), Theorem 5.14 implies that asymptotically $\|\tilde{\beta}_i - \tilde{\beta}_j\| = 0$, thus $\tilde{K} \leq K$. Under case (ii), we want to show that $\|\tilde{\beta}_i - \tilde{\beta}_j\| \neq 0$. Note that for any such i and j, $\|\beta_i^0 - \beta_j^0\| \geq J_{min}$. Now, suppose by controversy that there exist $i \in G_k^0$ and $j \in G_l^0$ that $\|\tilde{\beta}_i - \tilde{\beta}_j\| = 0$, besides by the consistency results in Theorem 5.4, we have $\tilde{\beta}_i - \tilde{\beta}_j = \beta_i^0 - \beta_j^0 + O_p(T^{-1/2})$, hence $\|\beta_i^0 - \beta_j^0\| = O_p(T^{-1/2})$. But, this contradicts assumption 5.10(i) that $T^{1/2} J_{min} \to \infty$ as $\|\beta_i^0 - \beta_j^0\| \geq J_{min}$

■

Proof of Theorem 5.16 :

Similar to the proof of Theorem 5.7, we study the oracle property by utilizing conditions from sub-differential calculus. From the FOC of the objective function in (5.9) with respect to β_i evaluated at $\tilde{\beta}_i$, we have

$$\mathbf{0} = -2\tilde{Q}'_{i,z\Delta x} W_{i,NT} \frac{1}{NT} \sum_{t=1}^{T} \tilde{z}_{it}(\Delta y_{it} - \tilde{\beta}_i' \Delta x_{it}) + \frac{\lambda_2}{2N} \sum_{j=1}^{N} \ddot{w}_{ij} \tilde{e}_{ij}, \tag{F.8}$$

where $\tilde{e}_{ij} = \frac{\tilde{\beta}_i - \tilde{\beta}_j}{\|\tilde{\beta}_i - \tilde{\beta}_j\|}$ if $\|\tilde{\beta}_i - \tilde{\beta}_j\| \neq 0$, $\tilde{e}_{ii} = 0$, and $\|\tilde{e}_{ij}\| \leq 1$ otherwise.

Averaging the above equation over the individuals in $\tilde{G}_k$, for any $k \in \{1,\ldots,K_0\}$, we have

$$\begin{aligned} \mathbf{0} &= \frac{-2}{N_k T} \sum_{i \in \tilde{G}_k} \tilde{Q}'_{i,z\Delta x} W_{i,NT} \sum_{t=1}^{T} \tilde{z}_{it}(\Delta y_{it} - \tilde{\beta}_i' \Delta x_{it}) + \frac{\lambda_2}{2NN_k} \sum_{i \in \tilde{G}_k} \sum_{j=1}^{N} \ddot{w}_{ij} \tilde{e}_{ij} \\ &= \frac{-2}{N_k T} \sum_{i \in \tilde{G}_k} \tilde{Q}'_{i,z\Delta x} W_{i,NT} \sum_{t=1}^{T} \tilde{z}_{it}(\Delta y_{it} - \tilde{\beta}_i' \Delta x_{it}) + \frac{\lambda_2}{2NN_k} \sum_{i \in \tilde{G}_k} \sum_{j \notin \tilde{G}_k} \ddot{w}_{ij} \tilde{e}_{ij}, \end{aligned} \tag{F.9}$$

together with the definition of $\tilde{\alpha}_k$, it follows that

$$\begin{aligned}\tilde{\alpha}_k = &\Big(\frac{1}{N_k}\sum_{i\in\tilde{G}_k}\tilde{Q}'_{i,z\Delta x}W_{i,NT}\tilde{Q}_{i,z\Delta x}\Big)^{-1}\frac{1}{N_kT}\sum_{i\in\tilde{G}_k}\tilde{Q}'_{i,z\Delta x}W_{i,NT}\sum_{t=1}^{T}\tilde{z}_{it}\Delta u_{it}\\ &+\Big(\frac{1}{N_k}\sum_{i\in\tilde{G}_k}\tilde{Q}'_{i,z\Delta x}W_{i,NT}\tilde{Q}_{i,z\Delta x}\Big)^{-1}\frac{\lambda_2}{2NN_k}\sum_{i\in\tilde{G}_k}\sum_{j\notin\tilde{G}_k}\ddot{w}_{ij}\tilde{e}_{ij}.\end{aligned} \tag{F.10}$$

Note that

$$\begin{aligned}\sqrt{N_kT}\|\frac{\lambda_2}{NN_k}\sum_{i\in\tilde{G}_k}\sum_{j\notin\tilde{G}_k}\ddot{w}_{ij}\tilde{e}_{ij}\| &\le \frac{\sqrt{T}\lambda_2}{N\sqrt{N_k}}\sum_{i\in\tilde{G}_k}\sum_{j\notin\tilde{G}_k}\|\ddot{w}_{ij}\tilde{e}_{ij}\|\\ &\le \sqrt{N_kT}\lambda_2\max_{i\in\tilde{G}_k,j\notin\tilde{G}_k}\ddot{w}_{ij}=o_p(1),\end{aligned} \tag{F.11}$$

thus, we have

$$\sqrt{N_kT}(\tilde{\alpha}_k-\alpha_k^0)=\Big(\frac{1}{N_k}\sum_{i\in G_k^0}\tilde{Q}'_{i,z\Delta x}W_{i,NT}\tilde{Q}_{i,z\Delta x}\Big)^{-1}\frac{1}{\sqrt{N_kT}}\sum_{i\in G_k^0}\tilde{Q}'_{i,z\Delta x}W_{i,NT}\sum_{t=1}^{T}\tilde{z}_{it}\Delta u_{it}+o_p(1). \tag{F.12}$$

Besides, by assumptions 5.9(iv) and 5.11(i)-(ii), we have

$$\frac{1}{N_k}\sum_{i\in G_k^0}\tilde{Q}'_{i,z\Delta x}W_{i,NT}\tilde{Q}_{i,z\Delta x}=\frac{1}{N_k}\sum_{i\in G_k^0}\bar{Q}'_{i,z\Delta x}W_i\bar{Q}_{i,z\Delta x}+o_p(1)=A_k+o_p(1).$$

Employing the result above and assumption 5.11(iii), the limiting distribution follows from the slutsky theorem.

■

Proof of Theorem 5.17 :

By corollary 5.15, it follows that $\sqrt{N_kT}(\tilde{\alpha}_{\tilde{G}_k}-\alpha_k^0)=\sqrt{N_kT}(\breve{\alpha}_k-\alpha_k^0)+o_p(1)$, and the limiting distribution follows from assumption 5.12.

■

G

Appendix G

Proof of Corollary 5.18 :

Following the analysis in Ma and Huang (2017), we prove the proposition.

(i) Note that from $\boldsymbol{\beta}^{(m+1)}$ definition, for any $\boldsymbol{\delta}$ we have

$$L_1(\boldsymbol{\beta}^{(m+1)}, \boldsymbol{\delta}^{(m+1)}, \boldsymbol{\nu}^{(m)}) \leq L_1(\boldsymbol{\beta}^{(m+1)}, \boldsymbol{\delta}, \boldsymbol{\nu}^{(m)}). \tag{G.1}$$

Also, we have

$$\begin{aligned} L_1(\boldsymbol{\beta}^{(m+1)}, \boldsymbol{\delta}^{(m+1)}, \boldsymbol{\nu}^{(m)}) &\leq \inf_{A\boldsymbol{\beta}^{(m+1)} - \boldsymbol{\delta} = 0} \left[\|\tilde{\boldsymbol{y}} - \tilde{\boldsymbol{X}}\boldsymbol{\beta}\|^2 + \frac{\lambda_1}{N} \sum_{i<j} \dot{w}_{ij} \|\delta_{ij}\| \right] \\ &= \inf_{A\boldsymbol{\beta}^{(m+1)} - \boldsymbol{\delta} = 0} L_1(\boldsymbol{\beta}^{(m+1)}, \boldsymbol{\delta}, \boldsymbol{\nu}^{(m)}) \equiv f^{(m+1)}, \text{ say.} \end{aligned} \tag{G.2}$$

For any integer n, by the definition $\boldsymbol{\nu}^{(m+n-1)} = \boldsymbol{\nu}^{(m)} + \vartheta \sum_{i=1}^{n-1}(A\boldsymbol{\beta}^{(m+i)} - \boldsymbol{\delta}^{(m+i)})$. Thus,

$$\begin{aligned} L_1(\boldsymbol{\beta}^{(m+n)}, \boldsymbol{\delta}^{(m+n)}, \boldsymbol{\nu}^{(m+n-1)}) &= \|\tilde{\boldsymbol{y}} - \tilde{\boldsymbol{X}}\boldsymbol{\beta}^{(m+n)}\|^2 + \boldsymbol{\nu}^{(m+n-1)'}(\Lambda\boldsymbol{\beta}^{(m+n)} - \boldsymbol{\delta}^{m+n}) \\ &+ \frac{\vartheta}{2}\|\Lambda\boldsymbol{\beta}^{(m+n)} - \boldsymbol{\delta}^{m+n}\|^2 + \frac{\lambda_1}{N}\sum_{i<j}\dot{w}_{ij}\|\delta_{ij}^{(m+n)}\| \\ &= \|\tilde{\boldsymbol{y}} - \tilde{\boldsymbol{X}}\boldsymbol{\beta}^{(m+n)}\|^2 + \boldsymbol{\nu}^{(m)'}(\Lambda\boldsymbol{\beta}^{(m+n)} - \boldsymbol{\delta}^{m+n}) \\ &+ \vartheta\sum_{i=1}^{n-1}(\Lambda\boldsymbol{\beta}^{(m+i)} - \boldsymbol{\delta}^{(m+i)})'(\Lambda\boldsymbol{\beta}^{(m+n)} - \boldsymbol{\delta}^{(m+n)}) \\ &+ \frac{\vartheta}{2}\|\Lambda\boldsymbol{\beta}^{(m+n)} - \boldsymbol{\delta}^{m+n}\|^2 + \frac{\lambda_1}{N}\sum_{i<j}\dot{w}_{ij}\|\delta_{ij}^{(m+n)}\| \\ &\leq f^{(m+n)}. \end{aligned} \tag{G.3}$$

Given $L_1(\boldsymbol{\beta}, \boldsymbol{\delta}, \boldsymbol{\nu})$ is differentiable with respect to $\boldsymbol{\beta}$ and convex with respect to $\boldsymbol{\delta}$, therefore the sequence $(\boldsymbol{\beta}^{(m)}, \boldsymbol{\delta}^{(m)}, \boldsymbol{\nu}^{(m)})$ has a limiting point by Tseng (2001). Let us denote this point by $(\boldsymbol{\beta}^*, \boldsymbol{\delta}^*, \boldsymbol{\nu}^*)$, so for any $n \geq 0$, we have

$$f^* = \lim_{m\to\infty} f^{(m+n)} = \inf_{\Lambda\boldsymbol{\beta}^* - \boldsymbol{\delta}=0}\left[\|\tilde{\boldsymbol{y}} - \tilde{\boldsymbol{X}}\boldsymbol{\beta}^*\|^2 + \frac{\lambda_1}{N}\sum_{i<j}\dot{w}_{ij}\|\delta_{ij}\|\right], \tag{G.4}$$

also we have

$$\begin{aligned} \lim_{m\to\infty} L_1(\boldsymbol{\beta}^{(m+n)}, \boldsymbol{\delta}^{(m+n)}, \boldsymbol{\nu}^{(m+n-1)}) &= \|\tilde{\boldsymbol{y}} - \tilde{\boldsymbol{X}}\boldsymbol{\beta}^*\|^2 + \frac{\lambda_1}{N}\sum_{i<j}\dot{w}_{ij}\|\delta_{ij}^*\| \\ &+ \lim_{m\to\infty}\boldsymbol{\nu}^{(m)'}(\Lambda\boldsymbol{\beta}^* - \boldsymbol{\delta}^*) + (n - \frac{1}{2})\vartheta\|\Lambda\boldsymbol{\beta}^* - \boldsymbol{\delta}^*\|^2 \\ &\leq f^*. \end{aligned} \tag{G.5}$$

The above result implies that $\|r^{(m)}\|^2 = \|\Lambda\boldsymbol{\beta}^* - \boldsymbol{\delta}^*\|^2 = 0$.

(ii) By the definition, we have

$$\begin{aligned} 0 &= \frac{\partial L_1(\boldsymbol{\beta}^{(m+1)}, \boldsymbol{\delta}^{(m)}, \boldsymbol{\nu}^{(m-1)})}{\partial \boldsymbol{\beta}} \\ &= 2\tilde{\boldsymbol{X}}'\tilde{\boldsymbol{X}}\boldsymbol{\beta}^{(m+1)} - 2\tilde{\boldsymbol{X}}'\tilde{\boldsymbol{y}} + \Lambda'\boldsymbol{\nu}^{(m)} + \Lambda'\vartheta(\Lambda\boldsymbol{\beta}^{(m+1)} - \boldsymbol{\delta}^{(m+1)}) \\ &= 2\tilde{\boldsymbol{X}}'\tilde{\boldsymbol{X}}\boldsymbol{\beta}^{(m+1)} - 2\tilde{\boldsymbol{X}}'\tilde{\boldsymbol{y}} + \Lambda'\Big[\boldsymbol{\nu}^{(m)} + \vartheta(\Lambda\boldsymbol{\beta}^{(m+1)} - \boldsymbol{\delta}^{(m+1)})\Big] \\ &= 2\tilde{\boldsymbol{X}}'\tilde{\boldsymbol{X}}\boldsymbol{\beta}^{(m+1)} - 2\tilde{\boldsymbol{X}}'\tilde{\boldsymbol{y}} + \Lambda'\boldsymbol{\nu}^{(m+1)} + \vartheta\Lambda'(\boldsymbol{\delta}^{(m+1)} - \boldsymbol{\delta}^{(m)}), \end{aligned} \tag{G.6}$$

which implies that

$$s^{(m+1)} = \vartheta\Lambda'(\boldsymbol{\delta}^{(m+1)} - \boldsymbol{\delta}^{(m)}) = -\Big[2\tilde{\boldsymbol{X}}'\tilde{\boldsymbol{X}}\boldsymbol{\beta}^{(m+1)} - 2\tilde{\boldsymbol{X}}'\tilde{\boldsymbol{y}} + \Lambda'\boldsymbol{\nu}^{(m+1)}\Big]. \tag{G.7}$$

Also, by part (i), we have $\|\Lambda\boldsymbol{\beta}^* - \boldsymbol{\delta}^*\|^2 = 0$, hence as m tends to infinity, from (G.6) we have

$$\begin{aligned} 0 = \lim_{m\to\infty} \frac{\partial L_1(\boldsymbol{\beta}^{(m+1)}, \boldsymbol{\delta}^{(m)}, \boldsymbol{\nu}^{(m)})}{\partial \boldsymbol{\beta}} &= \lim_{m\to\infty} \Big[2\tilde{\boldsymbol{X}}'\tilde{\boldsymbol{X}}\boldsymbol{\beta}^{(m+1)} - 2\boldsymbol{X}'\boldsymbol{y} + \Lambda'\boldsymbol{\nu}^{(m+1)}\Big] \\ &= 2\tilde{\boldsymbol{X}}'\tilde{\boldsymbol{X}}\boldsymbol{\beta}^* - 2\tilde{\boldsymbol{X}}'\tilde{\boldsymbol{y}} + \Lambda'\boldsymbol{\nu}^*. \end{aligned} \tag{G.8}$$

Employing the results in equations (G.7)–(G.8), we have $\lim_{m\to\infty} \|s^{(m+1)}\|^2 = 0$.

■

Proof of Corollary 5.19 :

The proof follows from that of proposition 5.19, and we omit it.

■